# MODELO DE MINERAÇÃO DE DADOS PARA DETECÇÃO E PREVISÃO DE INTERAÇÕES MEDICAMENTOSAS POTENCIAIS

FELIPE FERRÉ

# MODELO DE MINERAÇÃO DE DADOS PARA DETECÇÃO E PREVISÃO DE INTERAÇÕES MEDICAMENTOSAS POTENCIAIS

Tese apresentada ao Programa de Pós-Graduação em Bioinformática dos Instituto de Ciências Exatas e Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais como requisito para a obtenção do grau de Doutor em Bioinformática.

ORIENTADOR: WAGNER MEIRA JÚNIOR

Belo Horizonte

18 de dezembro de 2013

Ferré, Felipe
043 Modelo de mineração de dados para detecção e previsão de interações medicamentosas potenciais [manuscrito] / Felipe Ferré. — Belo Horizonte, 2013

XLIII, 226 f. : il. ; 29cm

Orientador: Wagner Meira Júnior.

Tese (doutorado) — Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas.

1. Saúde pública - Teses. 2. Mineração de dados (Computação) - Teses. 3. Farmacoepidemiologia - Teses. 4. Medicamentos - Interações - Teses. 5. Bioinformática - Teses. I. Meira Júnior, Wagner, II. Universidade Federal de Minas Gerais, III. Título.

CDU 573:004

*Universidade Federal de Minas Gerais*
*Instituto de Ciências Biológicas*
*Programa de Pós-Graduação em Bioinformática*

*ATA DA DEFESA DE TESE*

41/2013
entrada
2°/2009
CPF:
305.340.918-65

# Felipe Ferré

Às quatorze horas do dia **18 de dezembro de 2013**, reuniu-se, no Instituto de Ciências Exatas da UFMG, a Comissão Examinadora de Tese, indicada pelo Colegiado de Programa, para julgar, em exame final, o trabalho intitulado: **"Modelo de mineração de dados para detecção e previsão de interações medicamentosas potenciais"**, requisito para obtenção do grau de Doutor em **Bioinformática.** Abrindo a sessão, o Presidente da Comissão, **Dr. Wagner Meira Junior**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

| Prof./Pesq. | Instituição | Indicação |
|---|---|---|
| Dr. Wagner Meira Junior | UFMG | Aprovado |
| Dra. Gisele Lobo Pappa | UFMG | APROVADO |
| Dr. Augusto Afonso Guerra Júnior | UFMG | Aprovado |
| Dr. Braulio Roberto Goncalves Marinho Couto | UNI/BH | Aprovado |
| Dr. José Pedrazzoli Júnior | USF/UNIFAG | APROVADO |

Pelas indicações, o candidato foi considerado: APROVADO
O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

**Belo Horizonte, 18 de dezembro de 2013.**

Dr. Wagner Meira Junior - Orientador _____

Dra. Gisele Lobo Pappa _____

Dr. Augusto Afonso Guerra Júnior _____

Dr. Braulio Roberto Goncalves Marinho Couto _____

Dr. José Pedrazzoli Júnior _____

*Dedicar um trabalho a outrem é o momento de deixar de lado o quanto foi dedicado na empreitada.*

*Não somente o autor desferiu golpes no destino, mas outros também deixaram neste marcas produtivas ou não.*

*Conspirações do tempo. Tudo culmina neste texto.*

*Inexorável fato de que jamais estará completo. Herança da mãe. Assumindo-o ou não como filho, a Ciência não passa de um Ecce Homo que cobra direitos autorais para que vejam diferente mais do mesmo.*

# Agradecimentos

Sou grato aos membros do Grupo de Pesquisa em Farmacoepidemiologia (Departamento de Farmácia Social, UFMG), em especial ao Cairon Costa, Cristina Mariano Ruas Brandão, Gustavo Laine, Juliana Costa, Lívia Lemos, Mariana Michel Barbosa, Marina Amaral de Ávila Machado, Matheus Henrique Sales, Vânia Eloísia de Araújo e Thiago Henrique Neves. Ao Francisco de Assis Acúrcio devo minha gratidão pelo espaço e ensinamentos de uma saúde que tange o humano.

Agradeço também aos membros do speed e do Laboratório de Bioinformática e Sistemas (Departamento de Ciências da Computação, UFMG), em especial ao Bruno Coutinho, Claudiane Fonseca, Valdete M. Gonçalves de Almeida e Walter dos Santos Filho.

Meus agradecimentos ao Grupo de Pesquisa em Economia da Saúde, em especial a Eli Iola Gurgel Andrade, Orozimbo Henrique Campos Neto e Tiago Lopes Coelho.

Agradeço à Sandhi Maria Barreto e Roberta Carvalho de Figueiredo (Medicina, UFMG).

Agradeço aos colaboradores da Superintendência de Assistência Farmacêutica da Secretaria de Saúde do Estado de Minas Gerais, em particular à Liziane Silva e Ana Alice Pandolfi.

Agradeço aos integrantes do Programa de Doutorado em Bioinformática pela estrutura, em particular ao Carlos E. F. Santos, Natália e Sheila Santana, aos docentes pelos ensinamentos; à FUNDEP, FAPEMIG, CAPES e CNPq pelos fomentos diretos e indiretos ao grupo de pesquisa, ao trabalho e à minha formação; à UFMG por oferecer um espaço profícuo a criatividade, entretenimento e relações. Pelo *background*, sou grato à Unifal-MG - Universidade Federal de Alfenas e ao CEFET Uned de Cubatão, à USMED, Santa Maria, em particular ao Bento, Cidinha, Gaspar, Luiz, Paulo e Soares. Agradeço fundamentalmente a todos os grandes mestres André Márcio do Nascimento, Eliseu César Miguel, Fábio de Barros Silva, Ilma Manso Vieira Mansur, Lúcia Helena Silveira Ávila Terra, Luciene Alves Moreira Marques, Paulo Bueno Guerra, Sandra Maria Oliveira Morais Veiga, Stephanie Hill Feodorow e Nelson de Campos Villela. Agradeço aos organizadores da ISPE - *International Society of Pharmacoepidemiology* por ter aberto um portal que volveu definitivamente minha escolha em trabalhar com farmacoepidemiologia dentre tantas interessantes áreas.

Agradeço à incipiente atlas, por exclusão lógica, ao Douglas Eduardo Valente e ao Fernando Carvalho.

Agradeço às ricas contribuições no período de minha qualificação ao professor Adriano

Veloso e aos membros da banca de qualificação e defesa Augusto Alfonso Guerra Júnior, Braulio Roberto Goncalves Marinho Couto, Cristiano Moura, Gisele Lobo Pappa, José Pedrazzoli Júnior e Raquel Minardi. Agradeço ao Marcelo Santoro que não pode comparecer, mas gentilmente concedeu valorosas opiniões.

Sou grato ao meu orientador pela franca confiança em meu trabalho. Confiança ofertada desde que adentrei em sua sala. Agradeço pelos sábios ensinamentos meirísticos, os quais pretendo levar e propagar como a boa nova.

Agradeço a Deus, esta força onipotente, onipresente, onisciente e online.

Agradeço ao Linus Torvalds por ser o pai do *kernel* do Linux, responsável por ter me tirado do "kernel" de Platão. Agradeço ao Sócrates, Platão, Aristóteles, Descartes, Espinosa, Bacon, Locke, Hobbes, Voltaire, Rousseau, Leibniz, Kant, Hegel, Schopenhauer, Nietzsche, Heidegger, Sartre e Foucault por me manterem vivo. Peço desculpas a Agostinho e Tomas de Aquino por ter-lhes pulado.

Agradeço ao SOAD, Gorillaz e aos LH por tirar-me metafisicamente do corpo em meio a burcas e campos marcianos.

Agradeço à família alfenense, sobretudo à vovó Geralda, Márcia, Marcos, Maristela, Lucas, Paulo, David e Davidson; à santista, em especial ao Ignácio, Cília, Luana, Matheus, vovô Ezequiel[1] e vovó Dina; também, à santista por adoção Sta, Tel, Sol, Nego, Leo, Van e Sil. Agradeço a minha mãe pelo rebentar e instigar à leitura no dia seguinte.

Agradeço, sobretudo, à @lisina, @almiscar, @cipo[2], @voltaire, @octopu's e @hidra's pelos dias e noites que trabalharam incessantemente por mim. Peço desculpas pelos dolorosos episódios de swapagem, pela poeira e castigos ao teclado. Agradeço também à geladeira do speed a qual ingratamente não possui nome, mas salvou minha vida em inúmeros domingos e feriados hostis.

[1] *in memoriam*
[2] *in memoriam*

*"E era bom. 'Não entender' era tão vasto que ultrapassava qualquer entender - entender era sempre limitado. Mas não-entender não tinha fronteiras e levava ao infinito, ao Deus. Não era um não-entender como um simples de espírito. O bom era ter uma inteligência e não entender. Era uma bênção estranha como a de ter loucura sem ser doida. Era um desinteresse manso em relação às coisas ditas do intelecto, uma doçura de estupidez."*

(Clarice Lispector)

# Resumo

Frequentemente, a interação medicamentosa, efeito diferenciado da combinação de dois ou mais fármacos em relação ao uso isolado, é documentada apenas após a manifestação em populações. Devido à complexidade da determinação clínica e epidemiológica métodos computacionais se colocam como complemento ou alternativa na busca por novas interações a partir de quantidades massivas de dados estruturados e informações da experiência tradicional expressa em linguagem natural. O presente trabalho apresenta um metamodelo dedutivo, holístico e heurístico, intitulado DataMInt, o qual conjuga técnicas de extração, engenharia, processamento e análise para gerar modelos preditivos de interações medicamentosas alimentados pela integração de bases de dados biológicos e populacionais com o espaço de hipóteses de combinações de fármacos. A partir da vetorização de dados na forma de texto, número ou ontologia, métricas de distância entre as instâncias são combinadas sob diversos tratamentos, filtros e métodos de seleção de dados, de modo a gerar modelos capazes de delinear o conhecimento latente que caracteriza uma interação medicamentosa. O metamodelo abriga o conceito ”entidade-atributo”, visto que as entidades são melhor caracterizadas conforme cresce o número de atributos e quanto mais entidades descritas, aumenta o poder informativo e discriminativo do atributo. Um espaço de hipóteses amplo possibilita às técnicas de aprendizado de máquina a extrapolação do conhecimento disponível de interações conhecidas às desconhecidas. A abordagem proposta foi avaliada com a combinação das bases ATC/OMS, KEGG, EXPASY e ENZYME, sendo drugs.com o padrão ouro. Foram contemplados 1.390 fármacos e 18.340 interações medicamentosas conhecidas, melhor classificadas pelo modelo conjugado com o algoritmo *RandomCommittee*. Obteve-se kappa=0,871, precisão=0,959 e área sob a curva ROC=0,985. Dentre 947.015 pares desconhecidos, 12.482 foram classificados como interação (26,0% com frequência de citações MEDLINE). A relevância das interações medicamentosas foi verificada com a frequência de citações MEDLINE e pela incidência nas bases populacionais ELSA, Estudo Longitudinal da Saúde do Adulto, e SIGAF, Sistema de Gerenciamento de Assistência Farmacêutica (SES-MG). O metamodelo proposto consiste em uma relevante forma de construir conhecimento preditivo de interações medicamentosas ao adotar técnicas de mineração de dados e grandes bases de dados biológicas e populacionais.

**Palavras-chave:** Saúde Pública, Mineração de Dados, Farmacoepidemiologia, Interações de Medicamentos.

# Abstract

Several drug interactions, differential effect of the combination of two or more drugs compared to the isolated use, are documented only after broad usage by populations. Due to the complexity of determining clinical and epidemiological computational methods arise as a complement or alternative to the discovery of new interactions from data warehouses and traditional experience of information expressed in natural language. This study presents a deductive, holistic and heuristic metamodel entitled DataMInt, which combines techniques of extraction, engineering, processing and analysis to generate predictive models of drug interactions powered by integrating biological databases and population data with the hypothesis space of drug combinations. From the vectorization of data as text, number or ontology; metrics of distance between the instances are combined under different treatments, filters and methods of selection of data in order to generate models that delineate the latent knowledge that characterizes a drug interaction. The metamodel applies the concept 'entity-attribute', since the entities are best characterized as the number of attributes grows and as more entities described, increases the informative and discriminative power of the attribute. A large space of hypotheses enables machine learning techniques to extrapolate the available knowledge from known to unknown interactions. The proposed approach was evaluated with a combination of bases ATC / WHO, KEGG, and ExPASy ENZYME, drugs.com as gold standard. 1,390 and 18,340 known drugs and drug interactions were included respectively, and classified the best model in conjunction with RandomCommittee algorithm, yielding kappa = 0.871, accuracy = 0.959 and the area under the ROC curve = 0.985. Among 947 015 unknown pairs, 12,482 were classified as interaction (26.0% with citations MEDLINE). The relevance of drug interactions was verified with the frequency citations in the MEDLINE database and the incidence of ELSA, Longitudinal Study of Adult Health and SIGAF, the Pharmaceutical Care Management (SES-MG) system data. The proposed metamodel consists in a relevant way to build predictive knowledge of drug interactions by adopting Data Mining techniques in large data bases of biological and population data.

**Keywords:** Computational Biology, Data Mining, Pharmacoepidemiology, Drug Interactions.

# Resumo Estendido

O número de fármacos existentes e a crescente demanda por novas tecnologias farmacêuticas inviabiliza a avaliação exaustiva destinada ao conhecimento pleno dos efeitos isolados e das combinações terapêuticas ou casuais. Frequentemente, a interação medicamentosa, efeito diferenciado da combinação de dois ou mais fármacos em relação ao uso isolado, é documentada apenas após a manifestação em populações. Constitui um fenômeno complexo, cuja determinação bioquímica e farmacológica demanda corroboração clínica e epidemiológica. A caracterização da interação requer a avaliação de aspectos químicos, biológicos, psicológicos, comportamentais e sociais. Contudo, a previsão canônica de interações medicamentosas está restrita aos ensaios laboratoriais ou clínicos que elaboram modelos farmacocinéticos, relativos à absorção, metabolismo e eliminação; ou farmacodinâmicos, associados ao mecanismo de ação. Seja *in vitro*, *in vivo* ou *in populo*, as abordagens tradicionais estão limitadas a avaliação indutiva de uma quantidade restrita de informações destinadas a responder a uma hipótese específica, distante da avaliação do fenômeno enquanto categoria. Devido à complexidade, torna-se dispendiosa a definição acurada da interação medicamentosa, por demandar novos ciclos de hipóteses e análises para atingir o limiar de informação que subsista a prática clínica. Métodos computacionais se colocam como complemento ou alternativa a diversas demandas com elevado custo humano. A partir de quantidades massivas de dados e da experiência tradicional expressa em linguagem natural, modelos preditivos *in silico* vem estabelecendo novo conhecimento, na temática proposta, ao integrar dados biológicos e populacionais.

O presente trabalho apresenta um metamodelo dedutivo, holístico e heurístico, intitulado DataMInt, para descoberta de conhecimento em bancos de dados[3]. O metamodelo conjuga técnicas de extração, engenharia, processamento e análise para gerar modelos preditivos alimentados pela integração de bases de dados com o espaço de hipóteses de combinações de fármacos. A partir da vetorização de dados estruturados na forma de texto, número ou ontologia, métricas de distância entre as instâncias são combinadas sob diversos tratamentos, filtros e métodos de seleção de dados, de modo a delinear o conhecimento latente que caracteriza uma interação medicamentosa. O metamodelo abriga o conceito "entidade-atributo", visto que as entidades são

[3] A Descoberta de Conhecimento em Banco de Dados elenca um conjunto de técnicas preditivas que incluem armazenamento de dados, inteligência artificial ou aprendizado e máquina, análises estatísticas, formas de validação, dentre outras. É conhecida como KDD (*Knowledge Discovery in Databases*) ou Mineração de Dados (*Data Mining*).

melhor caracterizadas conforme cresce o número de atributos e quanto mais entidades descritas, aumenta o poder informativo e discriminativo do atributo. Um espaço de hipóteses amplo possibilita às técnicas de aprendizado de máquina a extrapolação do conhecimento disponível de interações conhecidas às desconhecidas.

A abordagem proposta foi avaliada com a combinação das bases ATC/OMS, KEGG, EXPASY e ENZYME, sendo o padrão ouro coletado a partir do drugs.com. Foram aplicadas técnicas de seleção de atributos e remoção de ruído como a avaliação da entropia e Decomposição em Valores Singulares, SVD. Foi realizada validação cruzada entre quatro classes de acordo com a respectiva gravidade ou caráter inerte/sinérgico da interação medicamentosa.

Foram contemplados 1.390 fármacos e 18.340 interações medicamentosas conhecidas, melhor classificadas pelo modelo conjugado com o algoritmo *RandomCommittee*. Obteve-se kappa=0,871, precisão=0,959 e área sob a curva ROC=0,985. Dentre 947.015 pares desconhecidos, 12.482 foram classificados como interação (26,0% com frequência de citações MEDLINE). A relevância das interações medicamentosas foi verificada com a frequência de citações MEDLINE e pela incidência nas bases populacionais ELSA, Estudo Longitudinal da Saúde do Adulto, e SIGAF, Sistema de Gerenciamento de Assistência Farmacêutica, fornecida pela Secretaria de Saúde do Estado de Minas Gerais.

O metamodelo proposto consiste em uma relevante forma de construir conhecimento preditivo de interações medicamentosas ao adotar técnicas de mineração de dados e grandes bases de dados biológicas e populacionais.

**Palavras-chave:** Saúde Pública, Mineração de Dados, Farmacoepidemiologia, Interações de Medicamentos.

# Lista de Figuras

# Lista de Tabelas

# Lista de expressões latinas

- *ad hoc*: para isto, para o caso específico
- *a posteriori*: pelo que se segue, em consequência de uma hipótese
- *a priori*: admitido como evidente, independe da experiência
- e.g., *exempli gratia*: por exemplo
- i.e., *id est*: isto é, ou seja
- *in memoriam*: em memória de
- *in populo*: estudos em populações
- *in vitro*: ensaios laboratoriais
- *in vivo*: estudos em seres vivos, incluindo estudos clínicos
- *in silico*: ensaios computacionais

# Lista de Símbolos e Notações

Tabela 1: **Símbolos e notações matemáticas.**

| Símbolo | Descrição |
|---|---|
| $x \leftarrow y$ | a variável $x$ recebe o valor de $y$ |
| $x \Leftarrow y$ | a variável $x$ é concatenada a $y$, p.ex. se $x = 5$ e $y = 3$, $x \leftarrow 8$; se $x =$"aba" e $y =$"ccc", $x \leftarrow$"abaccc" |
| $\|X\|$ | cardinalidade ou número de elementos distintos do conjunto |
| $\|x\|$ | cardinalidade ou número de elementos distintos do vetor |
| $\|[X]\|$ | cardinalidade ou número de elementos distintos do conjunto tratado como vetor, desta forma são conservadas as repetições |
| $\|\|x\|\|$ | produto interno do vetor $x$ |
| $[x, y, ..., z]$ ou $\vec{x}$ | vetor, sem a seta, indica vetor tradado como ponto no espaço n-dimensional |
| $(x, y)$ | par ordenado |
| $X = \{x_1, x_2, ..., x_n\}$ | elementos do conjunto $X$ de cardinalidade $n$. |
| $\wedge$ | operador lógico "E" |
| $\vee$ | operador lógico "OU" |
| $\forall$ | para todos |
| $\exists$ | existe |
| $\nexists$ | não existe |
| $\subset$ | subconjunto |
| $\subseteq$ | subconjunto ou igual |
| $\in$ | pertence |
| $\notin$ | não pertence |
| $\ll$ | muito menor que |
| $\gg$ | muito maior que |
| $\approx$ | aproximação |
| $\|$ | tal que |

continua na próxima página...

Tabela 1: **Símbolos e notações matemáticas** ...continuação

| Símbolo | Descrição |
|---|---|
| $\sum x$ | somatório de $x$, p.ex., $\sum_{i=1}^{4} i = 1+2+3+4 = 10$ |
| $\prod x$ | produtório de $x$, p.ex., $\prod_{i=1}^{4} i = 1 \times 2 \times 3 \times 4 = 4! = 24$ |
| $Y \bigcup X$ | união de todos os elementos $Y$ com os de $X$ |
| $Y \bigcap X$ | intersecção, ou seja, o subconjunto resultante dos elementos comuns aos conjuntos $X$ e $Y$ |
| $\lceil x \rceil$ | Próximo número inteiro |
| $\lfloor x \rfloor$ | Número inteiro anterior |
| $\prec$ | precede |
| $\succ$ | sucede |

Tabela 2: **Entidades do modelo.** Estas entidades podem ser entendidas como domínios para o espaço de hipóteses.

| Símbolo | Descrição |
|---|---|
| $U$ | um conjunto de pacientes ou usuários de serviço de saúde $u_1$, $u_2$, ..., $u_n$ |
| $F$ | um conjunto de fármacos $f_1$, $f_2$, ..., $f_n$ |
| $A$ | um conjunto de associações de fármacos $a_1$, $a_2$, ..., $a_n$, dado $A = \{a \mid a = F_x \wedge F_x \subseteq F \forall \mid F_x \mid \geq 2\}$ |
| $S$ | um conjunto usuários de fármacos $s_1$, $s_2$, ..., $s_n$, dado $S = \{s \mid s \in U(F)\} \therefore S = U(F) \subseteq U$ |
| $T$ | um conjunto de usuários de polifarmácia $t_1$, $t_2$, ..., $t_n$, dado $T = \{t \mid t \in U(A)\} \therefore T = U(A) \subseteq U(F) \subseteq U$ |
| $G$ | um conjunto fármacos utilizados $g_1$, $g_2$, ..., $g_n$, tal que $G = \{g \mid g \in F(U)\} \therefore G = F(U) \subseteq F$ |
| $H$ | um conjunto de fármacos associados $h_1$, $h_2$, ..., $h_n$, dado $H = \{h \mid h \in F(A)\} \therefore H = F(A) \subseteq F$ |
| $B$ | um conjunto de associações de fármacos utilizadas por pacientes $b_1$, $b_2$, ..., $b_n$, dado $B = \{b \mid b \in A(U)\} \therefore B = A(U) \subseteq A$ |
| $V$ | um conjunto de associações projetadas a partir de fármacos $G$ utilizados $v_1$, $v_2$, ..., $v_n$, dado $V = \{v \mid v \in F(U) \bigcap F(A)\} \therefore V = F(U) \bigcap F(A) \subseteq F$ |

Tabela 3: **Índices das entidades.**

| Símbolo | Descrição |
|---|---|
| **índices superiores, classificação de associação** | |
| $+$ ou $\oplus$ | associação sinérgica |
| 0 ou $\odot$ | associação inerte |
| $-$ ou $\ominus$ | associação adversa |
| 1 | interação menor |
| 2 | interação moderada |
| 3 | interação maior |
| **índices inferiores, fontes de dados** | |
| $k$ | associação conhecida |
| $p$ | associação prevista |
| $c$ | Drugs.com |
| $b$ | DrugBank |
| $a$ | ATC/OMS |
| $e$ | Drugs.com $\wedge$ DrugBank |
| $u$ | Drugs.com $\vee$ DrugBank |
| $q$ | associações eleitas para o espaço de busca |
| $r$ | treino |
| $t$ | teste |
| $\breve{X}$ | elemento contendo citações em textos científicos |

Exemplos de entidades combinadas com os índices:

- $F^{\ominus}$ um conjunto de fármacos com associações adversas conhecidas ou previstas.
- $F_c^{\ominus}$ um conjunto de fármacos com associações adversas classificadas segundo o sítio Drugs.com.
- $A_p^{\oplus}$ associações previstas como sinérgicas.
- $A_c^{3\oplus}$ associações classificadas pelo Drug.com como maiores.
- $\breve{A}_p^{+}$ citações das associações previstas como sinérgicas.
- $B_e^{\ominus}$ associações adversas classificadas segundo o Drugs.com e DrugBank.
- $S_c^{\ominus}$ usuários de fármacos com associações adversas classificadas segundo o sítio Drugs.com.

- $T_c^-$ usuários de associações adversas classificadas segundo o sítio Drugs.com.

Apenas o conjunto $U$ não admite índices. Os conjuntos $S$ e $T$ não admitem citações.

Tabela 4: **Funções.**

| Símbolo | Descrição |
|---|---|
| $\Pi$ | conjunto de funções de coleta de dados $\pi_1$, $\pi_2$, ..., $\pi_n$ |
| $\Xi$ | conjunto de funções de incorporação dos dados $\xi_1$, $\xi_2$, ..., $\xi_n$ |
| $\Psi$ | conjunto de funções de transformação dos dados $\psi_1$, $\psi_2$, ..., $\psi_n$ |
| $\Phi$ | conjunto de funções de transformação de matrizes $\phi_1$, $\phi_2$, ..., $\phi_n$ |
| $E$ | conjunto de funções de formação do espaço de hipóteses $\varepsilon_1$, $\varepsilon_2$, ..., $\varepsilon_n$ |
| $\Delta$ | conjunto de métricas de distância $\delta_1$, $\delta_2$, ..., $\delta_n$ |
| $\Theta$ | conjunto de funções de seleção de atributos $\theta_1$, $\theta_2$, ..., $\theta_n$ |
| $\Gamma$ | conjunto de funções de aproximação (modelo de aprendizado de máquina) $\gamma_1$,$\gamma_2$, ..., $\gamma_n$ |
| $\Omega$ | conjunto de medidas de desempenho $\omega_1$, $\omega_2$, ..., $\omega_n$ |
| $\Sigma$ | conjunto de medidas de incidência e prevalência $\sigma_1$, $\sigma_2$, ..., $\sigma_n$ |

Tabela 5: **Representação de dados.**

| Símbolo | Descrição |
|---|---|
| $D$ | conjunto potência de dados de fármaco $D_1$, $D_2$, ..., $D_n$ *in natura*, logo $D = 2^D$ |
| $M$ | conjunto potência de matrizes binárias ou matrizes de frequência $M_1$, $M_2$, ..., $M_n \therefore D = 2^D$, tomado a partir de funções $\Psi$ em dados $D$. |
| $W$ | conjunto potência de matrizes binárias ou de frequência decompostas $W_1$, $W_2$, ..., $W_n \therefore W = 2^W$, tomado a partir de funções $\Phi$ em matrizes $M$ |
| $Q$ | conjunto potência de lista ou matriz de adjacência contendo o espaço de hipóteses de associação de fármaco $Q_1$, $Q_2$, ..., $Q_n \therefore Q = 2^Q$, tomado a partir de funções $E$ dos domínios $F$ e $A$ |
| $N$ | conjunto potência de matrizes de distância $N_1$, $N_2$, ..., $N_n \therefore N = 2^N$, tomado a partir de funções $\Delta$ em matrizes $M$ ou $W$ |
| $Y$ | conjunto potência de matrizes de distância $Y_1$, $Y_2$, ..., $Y_n \therefore Y = 2^Y$ com atributos selecionados, tomado a partir de funções $\Theta$ em matrizes $N$ |

continua na próxima página...

Tabela 5: **Representação de dados** ...continuação

| Símbolo | Descrição |
|---|---|
| $C$ | conjunto potência de dados de associação de fármaco $C_1$, $C_2$, ..., $C_n$ *in natura*, logo $C = 2^C$ |
| $R$ | conjunto potência de dados de previsão $R_1$, $R_2$, ..., $R_n$ $\therefore$ $R = 2^R$, tomado a partir de funções $\Gamma$ em matrizes $N$ ou $Y$ e $C$ |
| $P$ | conjunto potência de dados de desempenho $P_1$, $P_2$, ..., $P_n$ $\therefore$ $P = 2^P$, tomado a partir de funções $\Omega$ em dados $R$ |
| $P$ | conjunto potência de dados de incidência ou prevalência comparativa das previsões em populações $P_1$, $P_2$, ..., $P_n$ $\therefore$ $P = 2^P$, tomado a partir de funções $\Sigma$ em dados $R$ |

# Lista de Algoritmos

# Prólogo

## Como definir saúde?

Sair do ponto de vista exclusivamente biológico para abordar as esferas sociais e psíquica certamente é o ato ético a ser perseguido. Ao medicar (e medicalizar) a sociedade, não basta avaliarmos restritamente parâmetros dinâmicos ou cinéticos desdenhando questões como mudanças na qualidade de vida, comportamentais ou em políticas públicas. Um indivíduo doente pode ser fruto de uma sociedade doente.

O entendimento de como um medicamento age não pode ser reduzido ao mecanismo fisiológico. Um medicamento apenas pode ser considerado eficaz, seguro e efetivo se agregar qualidade de vida de forma ética, ou seja, com equidade, onde o exercício pleno da cidadania não ocorra apenas com a igualdade dos deveres, mas com o nivelamento das dimensões físicas, psíquicas e sociais e das tantas outras que definem o gênero humano.

Ao conceituar saúde, tentamos modelar formas de explicá-la e reduzi-la a fenômenos predizíveis. Trespassando a rudimentalidade da técnica disponível, os primeiros anatomistas usaram a matemática para gerar modelos e superar o que apenas olhos nus poderiam perscrutar. Com o advento da ciência moderna, modelos mecanicísticos foram capazes de traduzir as informações fisiopatológicas em fenômenos aproximadamente previsíveis. No entanto, o acúmulo de informações convocou os homens a migrarem de modelos absolutamente observacionais ou racionalistas para avaliações que envolvessem processamento massivo de dados com técnicas computacionais, correlacionando o humanamente impensável.

Ainda hoje, muitos fisicalistas acreditam poder explicar os fenômenos biológicos apenas com dados e linguagem biológica. Neste ponto de vista, determinado comportamento biológico pode ser previsto com equações como a de de Michaelis Menten. Acredita-se que esses mecanismos de expressão estão reduzidos apenas a fatores bioquímicos e à transcrição. No entanto, a verdade obtida com a metodologia científica experimental é apenas parcial, sendo insuficiente para o entendimento do papel enzimático apenas considerar um conjunto de aspectos que remetem apenas ao objeto de estudo, pois esta enzima participa de um sistema, o qual a produz em determinada quantidade e a expressa em locais específicos do organismo sob estímulos de retroalimentação. Ao sairmos de um modelo estritamente celular para a construção de uma ontologia dos fenômenos e categorias, vem sendo observado aspectos que podem superar o âm-

bito bioquímico devido a fatores externos como interação entre organismos e o meio. Em se tratando de nossa espécie, agregamos complexidade social e psíquica, ampliando as fronteiras para avaliação das conexões com o sistema imunoneurológico. Se pretendemos decifrar a natureza em mecanismos, não podemos subestimá-la quanto a capacidade de promover impactos moleculares, incluindo a expressão em função das ações do que circunda o indivíduo em sua coletividade e ambiente. A avaliação isolada não é plenamente capaz de explicar um objeto complexo e está fadada a estagnar em si, colocando em xeque a prática de tomar o todo como a simples soma de partes. O conhecimento das interações, conexões, torna mais complexa a avaliação de sistemas.

Em nosso âmbito, não basta avaliar o medicamento como mero agente metabólico. Uma ação não esperada de um fármaco pode ser descoberta com estudos de utilização de medicamentos que avaliam hábitos dos usuários ou prescritores. A Assistência Farmacêutica deve lidar com o desafio da disponibilidade e qualidade do uso. A Atenção Farmacêutica deve investigar de perto caso a caso ao levantar subjetividades que possam levar a não adesão ao tratamento. Substâncias estigmatizadas, falta de informação, comportamentos de profissionais da saúde e dos pacientes devem ser observados e orientados. Diante da saúde institucionalizada, um pacto entre a sociedade, gestores e pesquisadores deve inaugurar um ciclo que respeite subjetividades e universalize boas práticas.

Longe de intentar a solução definitiva para modelar algo da concepção da saúde, entende-se que este trabalho tange, ou ao menos almeja, tais questões levantadas nesta dita pós-modernidade que tenta irromper com paradigmas de causa e efeito. Assim, amplas e diversificadas evidências do ponto de vista biológico com as técnicas consolidadas de avaliação epidemiológica podem gerar um modelo amplo que permita a compreensão individual do caráter da saúde e da interação medicamentosa, objeto do presente estudo.

Neste trabalho, a semântica de um dado conhecimento disponível é modelada matematicamente para ser analisada por métodos computacionais, relacionando-a ao perfil terapêutico de populações. Ao integrar grande número de informações, traçou-se os veios para uma nova abordagem dedutiva na formação da pergunta e na condução do método para estudos de utilização de medicamentos.

Este processo esboça uma aprendizagem que pretende aproximar da realidade o modelo de manutenção da saúde apregoado neste início do século XXI. Se conceitos psíquicos ou de qualidade de vida não foram abordados neste texto, ao menos pavimentou-se o caminho na prospecção de múltiplas variáveis. Esta capacidade de processamento pode ser utilizada para consolidar a visão biopsicossocial e possivelmente abrigar um holístico estudo de medicamentos.

# O que esperar com a leitura desta tese?

O veio do estudo é realizar previsões de interações medicamentosas por métodos computacionais. Elencou-se como interações medicamentosas às potenciais do ponto de vista farmacológico por se tratarem de informações de relativa ampla disponibilidade e cobertura dos fármacos existentes. Quanto ao conjunto de técnicas, optou-se por aplicar métodos de mineração de dados, por serem capazes de lidar com grandes massas de dados e pelo poder de obtenção de informações não triviais, latentes, ou seja, não deduzíveis diretamente.

Os resultados adquiridos advém de estudos em saúde pública e ciências da computação, com insumos para a farmacologia clínica e estudos de biologia sistêmica. Muito há de ser feito para estabelecer a previsão de interações medicamentosas com técnicas de mineração de dados como área do conhecimento que paute decisões clínicas e governamentais. Este trabalho inaugura um contexto diferenciado, heterodoxo, intento válido, de caminhos e descaminhos da busca pelo pela invenção diferente, mesmo diante de alguma inerência apontada por Lavoisier.

Grato pelo interesse!

# Sumário

# Capítulo 1

# Introdução

A interação medicamentosa ocorre quando o efeito de um fármaco é modificado pela presença de outro, caracterizada por manifestações terapêuticas ou adversas diferenciadas do uso isolado. Embora o uso corrente do termo posiciona a interação medicamentosa como um evento adverso, é uma prática comum a combinação de fármacos objetivando-se potenciação dos efeitos terapêuticos. O desfecho negativo é caracterizado pelo aumento da toxicidade de pelo menos um dos fármacos ou pela redução do efeito terapêutico, podendo ser ainda mais prejudicial [Vonbach, 2007].

A seguir, são destacados aspectos do fenômeno estudado em relação à formação de evidências experimentais (*in vitro*, *in vivo*), clínicas e epidemiológicas (*in populo*) e computacionais (*in silico*).

## 1.1 A dualidade farmacológica entre a atividade terapêutica e a tóxica

Os fármacos são substâncias benéficas, contudo podem causar doenças e morte [Vonbach, 2007]. Em 2010, os medicamentos foram responsáveis por 27,7% dos casos de intoxicação no Brasil [SINITOX, 2013]. Morbidades induzidas por fármacos se tornaram um problema frequente com elevação de gastos, sendo responsável por 6,5% das admissões hospitalares com 2,3% de óbitos dentre estes casos [Vonbach, 2007].

Com o deslumbramento despertado pelas tecnologias farmacêuticas em face do ganho de longevidade e qualidade de vida, aliado à pressão aos profissionais e ao sistema de saúde exercida com estratégias de marketing da indústria farmacêutica cada vez mais agressivas para penetração no mercado [Campos Neto et al., 2012], incrementa-se o uso indiscriminado e a polifarmácia[1] com consequente aumento do risco de eventos adversos relativos ao número de casos de combinação de medicamentos.

[1] Polifarmácia é definida como o uso simultâneo de dois ou mais medicamentos. No contexto do presente trabalho, é sinônimo de "combinação de medicamentos" e "associação de medicamentos".

"Deus ajude o paciente quando o cardiologista prescrever claritromicina." Esta afirmação de Walton-Shirley [2013] ilustra o perigo da combinação deste fármaco com sinvastativa, a qual pode ter a concentração aumentada em dez vezes no organismo; ou com a digoxina, fármaco em que a dosagem terapêutica é próxima da dosagem tóxica.

### 1.1.1 Interação medicamentosa e as práticas da saúde baseada em evidência

A detecção de interações medicamentosas ocorre desde o desenvolvimento dos fármacos ao monitoramento pós-venda. Um obstáculo para o crescimento do uso clínico de novas tecnologias farmacêuticas, mais eficazes e capazes de erradicar doenças especificas, é a falha no entendimento sistêmico da dinâmica celular. Em contraste, a indústria farmacêutica frequentemente depara-se com a falta de informação para seleção de alvos terapêuticos específicos e seguros, praticando investimentos estratosféricos em pesquisa e desenvolvimento [Kriete & Eils, 2006]. Adicionalmente à indústria, grupos de proteção ao consumidor, usuários de medicamentos e agências governamentais estão fortemente interessados em identificar reações adversas a fármacos incluindo interações medicamentosas [Page et al., 2012].

Antes e após o lançamento do fármaco, a identificação da melhor evidência é um aspecto fundamental para o uso seguro de medicamentos, sobretudo para os profissionais de saúde diretamente envolvidos no processo farmacoterápico. A relevância desse conhecimento cresce juntamente com o arsenal terapêutico disponível nos serviços de saúde, cuja incorporação de novas classes terapêuticas, novas formas farmacêuticas e sistemas de liberação de fármacos, gera um fator de risco para erros de medicação [Carvalho et al., 2013] o que demanda geração de novas evidências.

A saúde baseada em evidência é o consensual, explícito e diligente uso da melhor evidência atualizada na tomada de decisão clínica. A obtenção da melhor evidência envolve buscas sistemáticas de uma questão clínica restrita a uma população alvo com intervenção e desfechos bem definidos. Neste intuito, as bases de busca que mais se destacam são a MEDLINE e a EMBASE [Tanjong-Ghogomu et al., 2009].

Diversos tipos estudos intuem o grau de evidência que deve pautar a decisão clínica, sendo usualmente os ensaios experimentais/laboratoriais àqueles com menor evidência e revisões sistemáticas com metanálise considerados o de maior evidência. Centros colaboradores para saúde baseada em evidência, como o Cochrane ou Oxford, hierarquizaram como melhor nível de evidência as revisões sistemáticas de ensaios clínicos controlados e randomizados, seguidas respectivamente de resultados de ensaios clínicos controlados e randomizados de elevada qualidade, ensaios clínicos não randomizados e estudo observacional, estudos experimentais, e, em última instância, opinião de especialistas (anexo A.4.1).

### 1.1.2 A sociedade vulnerável à falha na segurança de fármacos

Mesmo diante do elevado investimento com ensaios clínicos, novos fármacos ainda chegam ao mercado com falhas não detectadas [Strandell et al., 2013]. O número de fármacos cujo licenciamento foi afetado devido a reações adversas foi 34 nas décadas de 50 e 60; 137 entre 70 e 80; e 113 entre 90 e 2010 com permanência de 5 anos no mercado para grande parte dos casos [Aronson, 2011]. Em diversos países, falhas devido a interações medicamentosas motivaram a remoção de produtos, tais como Fenoxipropazina (1966), mebanazina (1975), tranilcipromina (1987), sorivudina (1993), nialamida (1995), mibefradil (1997), bromocriptina (1998), astemizol (2001) [Stephens, 2005].

Estes fatos devem-se, sobretudo, à limitações dos estudos clínicos. Antes de ganhar o mercado, os fármacos são testados em apenas alguns milhares de pacientes, sendo posteriormente usados por milhões. Como resultado, muitos casos de eventos adversos não identificados nos ensaios clínicos são observados em populações maiores [Higgins & Green, 2011; Page et al., 2012].

## 1.2 Identificação de interações medicamentosas

A farmacoepidemiologia avalia a interação medicamentosa enquanto objeto de estudo do desfecho de saúde relacionado à utilização de medicamentos em populações. Em particular destacam-se estudos de utilização de medicamentos em coortes de pacientes acompanhados por grande quantidade de tempo [Ceccato et al., 2013] e verificação de padrões em bases de dados de notificação de eventos adversos, como a do departamento estadunidense de alimentos e medicamentos, FDA. Apesar da existência dessas técnicas, ainda à beira do XXI, constatou-se que poucos estudos foram desenvolvidos para categorizar a prevalência de interações medicamentosas potenciais e sua gravidade em populações [Peng et al., 2003], restando inúmeras combinações cujos efeitos são desconhecidos ou pouco relatados.

Logo, urge a necessidade de estudos farmacoepidemiológicos pós-marketing de utilização de medicamentos que sejam capazes de detectar eventos raros de segurança em função de populações expostas e não expostas ao tratamento simultâneo com outros fármacos. Alguns estudos são mostrados a seguir.

### 1.2.1 Prevalência de interações medicamentosas

Um estudo britânico mostrou que 16% dentre 18.820 admissões hospitalares mostraram interações medicamentosas com aumento de 2% a 3% na mortalidade [Walton-Shirley, 2013]. Em um hospital suíço, 21% das admissões causadas por medicamentos foram relacionadas à interação medicamentosa, correspondendo a 13% do total [Vonbach, 2007].

Pasina et al. [2013] observaram 2.712 pacientes hospitalizados com idade superior a 65 anos durante três meses. Praticamente 19% foram expostos a pelo menos uma interação me-

dicamentosa de severidade considerada maior. A mortalidade foi significativamente maior em relação a pacientes expostos a pelo menos duas interações medicamentosas consideradas graves. Os autores sugeriram monitoramento cuidadoso para a minimização dos riscos.

Dentre os fatores de risco para interação medicamentosa adversa, destacam-se a polifarmácia, número de fármacos administrados, idade avançada e a prorrogação de internação hospitalar, com consequente elevação nos custos e prevalência de comorbidades [Linnarsson, 1993; Moura et al., 2011; Pinto et al., 2013]. Ressalta-se o papel da interação medicamentosa enquanto uma das principiais causas preveníveis de reações adversas [Snyder et al., 2012].

### 1.2.2 Dificuldades nas buscas por evidências

Possivelmente associado à falta de estudos que indiquem de forma completa quais medicamentos interagem, a elevada prevalência de risco às interações medicamentosas potenciais foi contradita por estudos clínicos que indicaram valores inferiores na prática. Becker et al. [2007] realizaram uma revisão sistemática que recuperou vinte e três trabalhos no MEDLINE e EMBASE entre 1990 e 2006 sobre interações medicamentosas em pacientes hospitalizados. Foi demonstrado que as interações medicamentosas em estudos com grande número de pacientes causaram 0,054% das incursões de emergência, 0,57% das admissões hospitalares e 0,12% das re-hospitalizações. Em idosos, as interações medicamentosas foram responsáveis por 4,8% das admissões. As morbidades mais comuns foram sangramento gastrointestinal, níveis irregulares de pressão arterial e arritmia cardíaca. Com estes dados, os autores concluíram que as interações medicamentosas estão limitadas a um número reduzido de fármacos e mitigaram sua importância ao salientar a incerteza sobre os impactos clínicos sob a baixa prevalência observada.

A controversa oscilação entre achados potenciais e clínicos mostra a dificuldade em se detectar ou atribuir fatos clínicos a interações medicamentosas. Quando não envolve mecanismos tradicionais farmacocinéticos[2] a interação torna-se um fenômeno de difícil detecção. Seja em estudos controlados ou em dados históricos, uma resposta apontada para esta divergência é a possibilidade de interações sub-notificadas.

Hazell & Shakir [2006] verificaram que a média de sub-notificação em doze países pode atingir 94%. As principais dificuldades envolvem pacientes e profissionais da saúde e foram apontadas por Aronson [2011] como o desconhecimento sobre a importância da notificação, subestimação dos efeitos suspeitos, letargia ou indiferença sobre a contribuição da notificação e complacência por acreditar-se que apenas são licenciados fármacos seguros.

A incerteza sobre o limiar de relevância da reação contribui para a sub-notificação, sendo fator de divergência para a classificação de interações medicamentosas. A constatação da interação enquanto parte da natureza dos fármacos envolvidos traz informação relevante, sobretudo

[2] Definição na página 143.

em condições de saúde específicas [Aronson, 2011], contribuindo para o estabelecimento da relação causal de sua ocorrência na prática clínica.

Outra deficiência na avaliação de interações medicamentosas é a existência de comorbidades. Pacientes idosos frequentemente apresentam 2 ou 3 morbidades. Embora o seguimento de protocolos específicos para cada morbidade seja comum, não existe a preocupação da reavaliação quando outros protocolos estão envolvidos [Huang et al., 2013].

Desta forma, a prática clínica deve ser norteada por estudos clínicos e epidemiológicos constantemente atualizados quanto a qualidade da evidência. O volume de dados crescentemente gerados demanda formação de repositórios propensos a recuperação da informação que permita estabelecer a associação de eventos sinérgicos ou tóxicos ao uso concomitante de medicamentos, dado que, ainda hoje, pouco é conhecido diante das possibilidades de combinações.

## 1.3 A dualidade farmacológica se reflete na terminologia das interações

A dualidade, ou mesmo, dubiedade do caráter benéfico ou adverso da combinação entre substâncias pode causar divergências nos estudos e na terminologia adotada. Ainda, uma substância pode apresentar atividade farmacológica apenas na presença de outra.

Um exemplo abordado no presente trabalho é a combinação entre insulina e losartana. Embora diversos estudos apontem para o efeito sinérgico do aumento da sensibilidade à insulina[3] com a presença de losartana [Jin & Pan, 2007], existem relatos de reações adversas [DRUG INFORMER, 2013]. Takagi & Umemoto [2012] realizaram uma revisão sistemática com metanálise, a qual reúne evidências para a verificação de uma tendência global das farmacoterapias nesta linha de combinação. Os autores recomendaram como opção mais segura a combinação de insulina com telmisartana em detrimento dos demais fármacos desta classe que trata problemas circulatórios.

Ao buscar nomenclatura correlata a "interação medicamentosa", verificou-se na base MeSH [Lipscomb, 2000] que este termo é uma ramificação de "toxicidade farmacológica", descrita como "manifestação de efeitos adversos de fármacos administrados terapeuticamente ou para fins diagnósticos, não incluindo envenenamento acidental ou intencional".

Embora a definição não seja completa, o fármaco que afeta outro de modo benéfico pode ser descrito como "adjuvante farmacológico", definido pela base MeSH como "agente que melhora a ação do princípio ativo (sinergismo) podendo afetar a absorção, mecanismo de ação, metabolismo[4] ou excreção (farmacocinética[5])". Um exemplo que se adequaria a esta definição é a combinação de clavulanato a amoxicilina, onde o primeiro reduz o metabolismo do

[3] Antidiabético.

[4] Informações sobre metabolismo de fármacos são dadas na seção A.3.

[5] Definição na página 143.

segundo aumentando a capacidade antimicrobiana, sem, contudo, apresentar ação terapêutica isoladamente.

A definição não é completa pois, tradicionalmente, a substância para ser considerada fármaco deve ter uma ação terapêutica própria, caso contrário, a substância é considerada apenas adjuvante terapêutico. Desta forma, permanece a confusão entre substâncias ativas e inertes (como os excipientes) ou substâncias com potencial farmacológico indireto (como o clavulanato).

Provavelmente o termo mais adequado seja "sinergismo farmacológico", definido como "ação de um fármaco na melhora da efetividade de outro fármaco". Em uma busca realizada pelo presente autor em setembro de 2013, recuperou-se 136.025 citações MEDLINE com o termo "*drug interaction*[MeSH Terms]" e 56.057 com "*drug synergism*[MeSH Terms]". Uma evidência de que "sinergismo terapêutico" é correlato de "interação medicamentosa", a despeito da ausência na hieraquia de termos MeSH, é a recuperação das 56.057 citações ao associar-se os termos com o operador "*AND*" e nenhuma citação com o operador "*NOT*".

Apesar da busca com palavras-chave recuperar um número elevado de citações em relação à busca por pares específicos, observou-se neste trabalho, conforme mostrado na figura 5.2 que a busca MEDLINE por combinações específicas não recuperou metade das interações medicamentosas conhecidas. Grande parte das informações recuperadas rementem a estudos de combinações com fins terapêuticos, permanecendo a controvérsia se existe alguma tendência em orientar esforços para a compreensão das interações sinérgicas ou adversas.

A pesquisa e definição da interação medicamentosa enquanto terapêutica ou adversa não é direta ou isenta de confusão. Outro fato é a não detecção de compêndios de combinações recomendáveis, ou, ao menos, inertes, provavelmente devido à praxe de contra-indicar a polifarmácia.

### 1.3.1 A dualidade terminológica se reflete no estabelecimento da relação causal de fenômenos clínicos

Além do estabelecimento de novos alvos terapêuticos e da terminologia, outra dificuldade em se determinar interações medicamentosas está na correlação dos eventos adversos ou terapêuticos à combinação de fármacos em cada condição clínica, sobretudo nos eventos de baixa prevalência. Embora o cenário ideal seja a detecção dos eventos em estudos clínicos randomizados com determinação das rotas metabólicas[6] e mecanismos de ação, nem todas as interações são descritas desta forma.

Dentre as falhas na cobertura dos eventos adversos nos ensaios clínicos, destacam-se o pequeno número de pacientes em termos epidemiológicos; a duração do tratamento que pode chegar a apenas uma dose e a oscilação na dosagem devido à dinâmica do desenvolvimento das

[6]Rota metabólica é a ocorrência de redes de moléculas e proteínas capazes anabolizar (sintetizar) ou catabolizar (degradar, quebrar) moléculas. A rota é caracterizada pela participação das moléculas formadas em etapas subsequentes de metabolismo.

formulações, sendo frequente o uso de baixas dosagens e exclusão de populações específicas como grávidas ou indivíduos com histórico clínico desfavorável. Devido a estes fatores, a generalização dos resultados torna-se limitada [Strom & Kimmel, 2007].

Verificou-se a tendência das interações medicamentosas sinérgicas serem estudadas clinicamente, enquanto as adversas são observadas predominantemente em populações. Porém, dados epidemiológicos relatando desfechos clínicos negativos de interações medicamentosas são raros, por esta razão, os estudos avaliam interação medicamentosa potencial [Vonbach, 2007] a qual é registrada na literatura, contudo, não pode ser confirmada devido a não coleta de desfechos clínicos associados ao conjunto de fármacos utilizados.

## 1.4 A informação como ferramenta de promoção à saúde

A tomada da decisão em associar ou não determinados fármacos deve ser ponderada quanto a qualidade da evidência e possível impacto na prática clínica. Uma estratégia para mitigar os impactos das interações medicamentosas é a promoção do acesso a informações previamente avaliadas quanto à qualidade da evidência [Walton-Shirley, 2013]. A constituição da evidência de efetividade deve vir acompanhada da segurança. Desta forma, enfocando os vários níveis de evidência, os estudos devem contemplar a avaliação de eventos terapêuticos e adversos, incluindo interações medicamentosas.

Os aspectos técnicos e regulatórios acerca da efetividade e segurança dos medicamentos são dinâmicos. Diante da crescente riqueza de informação, técnicas inteligentes e holísticas de interpretação devem fornecer subsídios para que os profissionais de saúde se pautem na melhor evidência disponível [Kriete & Eils, 2006].

A demanda do pronto acesso a informações aumentou os investimentos em sistemas de apoio à decisão, os quais contribuem no ato da prescrição, dispensação, administração e monitoramento dos medicamentos. Em geral, são compostos por bancos de dados e sofisticados sistemas de recuperação de informação [Hemens et al., 2011].

Módulos de detecção de interações medicamentosas em sistemas de auxílio a prescrição são úteis na prática clínica [Vonbach, 2007; Walton-Shirley, 2013]. Acredita-se que sistemas automatizados ofereçam benefícios ao cuidado de pacientes com alertas em tempo real quando contém informações acuradas. No entanto, a qualidade dos alertas pode variar conforme a base adotada, sobretudo na cobertura de casos, estratégia de busca e classificação da gravidade. A escolha da ferramenta deve envolver aspectos de sensibilidade, especificidade [Vonbach, 2007], avaliando-se os casos de sinergismo apontado como interação adversa (falso positivo) e casos de interação adversa potencial apontados como inertes ou sinérgicos (falso negativo). Porém, as informações acerca da segurança não estão amplamente disponíveis refletindo-se na terminologia adotada para a definição do caráter da combinação estudada.

### 1.4.1 Sistemas computadorizados

Os sistemas de auxílio a prescrição ou dispensação se apresentam como solução para a avaliação de interações medicamentosas por realizarem alertas a partir de dadoes e informações integradas de pacientes e medicamentos [Snyder et al., 2012]. Os sistemas computadorizados de auxílio a tomada de decisão são desenvolvidos com intuito de reduzir variabilidade, padronizar e validar intervenções [Sucher et al., 2008] e aumentar a qualidade no serviço de saúde [Kawamoto et al., 2005], incluindo desfechos de segurança como interação medicamentosa [Wong et al., 2010].

Métodos sofisticados de recuperação de informação foram estabelecidos diante da crescente disponibilidade de dados científicos e informações acerca de medicamentos. Diversos repositórios são disponibilizados por entidades como a Organização Mundial de Saúde (ATC[7], Drug dictionary, WHO-ART[8], CID-10[9]), Agência Europeia de Medicamentos (EVMPD[10]), Agência Estadunidense de Alimentos e Medicamentos - FDA (COSTART[11]) e Conferência Internacional de Harmonização (MedDRA[12], MedLEE) [Mann & Andrews, 2007]. Neste ínterim, muitos estabelecimentos de saúde investem em sistemas computadorizados de auxílio à tomada de decisão para contínua atualização de interações medicamentosas conhecidas [Wong et al., 2010].

Apesar desses esforços, alguns estudos demonstraram a falta de uma evidência definitiva acerca da contribuição dos softwares de auxilio à tomada de decisão de cunho clínico incluindo verificação de interações medicamentosas [Sim et al., 2001; Whiting et al., 2004; Wong et al., 2010; Hemens et al., 2011; Jaspers et al., 2011]. A utilidade destes sistemas perpassa pela redução do tempo, esforço ou inciativa requerida dos clínicos para acatar as recomendações [Kawamoto et al., 2005].

Os sistemas não configurados para apresentar informações clínicas relevantes e alertas oportunos levam à "fadiga aos alertas". Os usuários frequentemente ignoram as informações por considerarem excessivas ou irrelevantes, reduzindo o impacto clínico das ferramentas [Snyder et al., 2012; Troiano et al., 2013].

Diante destes fatores, os sistemas de apoio à decisão ainda não contribuem de forma significativa em desfechos de saúde [Sim et al., 2001; Hemens et al., 2011], ou mesmo, interações medicamentosas [Wong et al., 2010]. Além dos alertas desnecessários, outra limitação é a restrição ao conhecimento armazenado [Snyder et al., 2012]. As ferramentas apenas respondem a um conjunto limitado de fármacos, interações e regras manualmente estipuladas e frequentemente são baseadas em poucas fontes sem a avaliação e atualização devida.

Além das dificuldades citadas, como a subnotificação, dubiedade na nomenclatura e a existência de comorbidades que podem modificar o curso do tratamento; as lacunas de infor-

[7] Anatomical-Therapeutic-Chemical.
[8] Dicionário hierárquico de reações adversas suspeitas usado pelo Centro de Monitoramento Uppsala.
[9] Classificação Internacional de Doenças.
[10] EudraVigilance Medicinal Product
[11] *Coding Symbol for a Thesaurus of Adverse Reaction Terms*.
[12] *Medical Dictionary for Regulatory Activities*.

mações farmacológicas acerca da dose-dependência de muitas interações medicamentosas [Villacorta Linaza et al., 2010], a natureza do processo regulatório de aprovação, variações genéticas e demográficas podem trazer obstáculos ao reconhecimento de interações medicamentosas [Percha & Altman, 2013]. Ainda, o complicado desenvolvimento de softwares para descoberta inteligente de interações medicamentosas ajustadas a modelos de verossimilhança e impacto clínico potencial, requer a avaliação de grande quantidade de casos que assegurem exatidão e interpretação apropriada da relevância das informações extraídas acerca de morbidades [Wong et al., 2010].

### 1.4.2 Técnicas de aprendizado de máquina para problemas biológicos complexos

Bancos de dados relacionais, modelos de processamento de linguagem natural e aprendizado de máquina vem sendo desenvolvidos para disponibilizar alertas e informações preditivas. As técnicas de aprendizado de máquina representam uma alternativa para superar as limitações que envolvem avaliação simultânea de diversas entidades implicando em respostas complexas. Modelos preditivos que aplicam técnicas de aprendizado de máquina obtiveram consideráveis avanços no contexto biológico como identificação de epidemias [Gomide et al., 2011], termos biomédicos [Krauthammer & Nenadic, 2004; Torii et al., 2004], previsão de função enzimática [da Silveira et al., 2012], interação fármaco-gene [Tari et al., 2010], inibição de sítio ativo de enzimas [Gonçalves-Almeida et al., 2012], função de proteína [Pires et al., 2011] ou função terapêutica de fármacos [Wang et al., 2013].

A previsão de fenômenos biológicos, incluindo interação medicamentosa, não é trivial dada a complexidade e o número de elementos envolvidos. Usualmente, lidar com a complexidade da linguagem farmacológica tradicional envolve transposição em linguagem computacional por sofisticadas modelagens estruturais ou descritivas na forma de entidades biológicas e relacionamentos ou ações, ou, ainda, de forma hierárquica. Contudo, a modelagem pode consumir esforços e recursos humanos cuja especialização requer substancial treinamento da acuidade que estabeleça uma visão objetiva e abrangente para modelar o minimundo a ser explorado.

O poder de expressão da modelagem de informações massivas e problemas biológicos complexos deve ser elaborado juntamente com técnicas sofisticadas que permitam explorar ferramentas estado da arte da computação. O processo conhecido como KDD, descoberta de conhecimentos em bancos de dados, é um aliado que combina métodos tradicionais de análise estatística com técnicas sofisticadas para processar grandes volumes de dados. Este conjunto de técnicas extraem padrões úteis em dados de alta dimensionalidade (com centenas ou milhares de atributos), complexos e heterogêneos (texto, números, datas ou hierarquias) [Zaki & Meira Jr, 2014; Tan et al., 2005]. Os modelos viabilizam o aprendizado de máquina com a geração de respostas por meio de observações cuja retroalimentação tende a elevar a performance ao longo da experiência adquirida [Russel & Norvig, 2003].

A combinação de técnicas do processo KDD viabiliza diversas tarefas como a seleção de atributos em dados de elevada dimensionalidade; visualização para auxílio a descoberta de conhecimento a partir de estruturas globais e complexas rotas biológicas; classificação e taxonomia, isto é, assinalar um conjunto de entidades a uma determinada classe de acordo com instâncias previamente conhecidas, armazéns e mineração de dados como *corpus* de textos científicos ou bancos de dados de farmacovigilância e análise de redes biológicas [Peng et al., 2010].

Três elementos são necessários para o estabelecimento de modelos preditivos computacionais. O primeiro é o espaço de busca, o qual corresponde ao conjunto de previsões possíveis, interações e eventos clínicos que são as variáveis independentes da função de aprendizagem. Em outras palavras, o espaço de busca é a consulta (*query*) que define aquilo que se deseja conhecer. O segundo elemento é a fonte de dados e informações destinadas ao modelo preditivo, ou seja, as variáveis alocadas no eixo das ordenadas no espaço multidimensional, como informações de proteínas e elementos biológicos, descrição farmacológica, resumos científicos ou notificações espontâneas de eventos adversos. Nesta etapa é definido o modelo de dados na forma de matrizes ou grafos que possibilitam a correlação entre as entidades e os eventos avaliados. A terceira etapa é a validação das previsões. A partir da comparação das previsões frente a um padrão ouro é evidenciada a capacidade de apreensão das características que regem a relação fármaco-evento usadas para novas atribuições à instâncias conhecidas [Sojda, 2007]. Outra forma de validação é o acompanhamento dos eventos por especialistas que possam julgar a correspondência da saída do modelo.

### 1.4.3 Descoberta de conhecimento relacionado a eventos adversos e interações medicamentosas

Alguns trabalhos vem demonstrando êxito na previsão de eventos adversos e interações medicamentosas.

Gurulingappa et al. [2013] usaram processamento de linguagem natural para detectar automaticamente sinais de eventos adversos a partir de texto e fontes abertas com base em modificações na utilização dos fármacos para finalidades terapêuticas não regulamentadas. Page et al. [2012] demonstraram a importância na busca em bases contendo anos de pesquisa epidemiológica na seleção de exemplos positivos e negativos para o aprendizado de máquina de eventos clínicos.

Wilk et al. [2013] propuseram um método para identificar e encaminhar reações adversas, incluindo interações medicamentosas, em pacientes com mais de uma morbidade de acordo com os respectivos protocolos clínicos de manejo da doença. Os operadores do domínio do conhecimento de interações e revisões foram combinados com programação de restrições lógicas [Gelfond & Lifschitz, 1988, 1991]. Os operadores caracterizaram reações adversas e descreveram revisões aos modelos lógicos requeridos para encaminha-los após a resolução do caso

clínico.

A obtenção do conhecimento sobre interações farmacocinéticas[13] é a primeira escolha para a verificação de combinações de fármacos, dado que são intuitivas por envolverem variáveis frequentemente mensuráveis (relações da concentração do fármaco) ou possuírem relação direta, como a verificação de enzimas metabólicas compartilhadas ou outras biomoléculas.

Embora a exploração farmacocinética seja mais comum, existem modelos farmacodinâmicos[14] como o proposto por Huang et al. [2013]. Os autores elaboraram uma métrica para mensurar a afinidade das interações entre fármacos e alvos terapêuticos. Os autores adotaram interações medicamentosas farmacodinâmicas como padrão ouro positivo em um modelo Bayesiano probabilístico. Dentre 9.626 interações farmacodinâmicas potenciais conhecidas, foi obtido com este mapeamento um acerto de 82% das instâncias.

Frequentemente, os modelos de dados partiram do conhecimento direto estabelecido, tal como o farmacocinético ou combinação de eventos bem conhecidos. Dentre estes modelos para previsão de interações medicamentosas destacam-se ferramentas baseadas em programação em lógica matemática[15] [Segura-Bedmar et al., 2011b], mineração de textos[16] [Duke et al., 2012] científicos, mineração de rotas metabólicas na forma de grafos [Lin et al., 2010], detecção de padrões estruturais dos fármacos com biomoléculas [Vilar et al., 2012] ou mineração de lócus gênicos análogos [Lin et al., 2007]. Notoriamente, o aprendizado de máquina, quando aplicada à linguagem natural agrega a capacidade descritiva do homem com a de processamento de informações pelo computador.

O desafio defrontado por estas técnicas é ultrapassar o reconhecimento das interações para abrigar mecanismos e, em última instância, realizar previsões de interações desconhecidas, auxiliando a prática clínica e tomada de decisão em saúde pública [Percha & Altman, 2013].

Em uma revisão sistemática (capítulo 3) verificou-se que os modelos abordam uma criteriosa escolha dos atributos que reconhecidamente estão relacionados a interações medicamentosas. Destacam-se estudos em bases de uso de medicamentos [Kinney, 1986; Estacio-Moreno et al., 2008; Harpaz et al., 2010a; Lin et al., 2010; Duke et al., 2012], relações de fármacos e biomoléculas de metabolismo como citocromos [Duke et al., 2012; Gottlieb et al., 2012], alvos terapêuticos compartilhados [Gottlieb et al., 2012], rotas metabólicas [Tari et al., 2010], indicações terapêuticas [Gottlieb et al., 2012], interação fármaco-proteína ou biomolécula de modo geral [Lin et al., 2010; Percha et al., 2012; Tari et al., 2010; Gottlieb et al., 2012] e combinações de fármacos e efeitos adversos [Estacio-Moreno et al., 2008; Gottlieb et al., 2012; Harpaz et al., 2010a].

A estratégia da verificação do conhecimento consolidado reproduz de modo sofisticado o verificado nas áreas laboratoriais, clínicas e epidemiológicas e pode estar limitada às informa-

[13] Definição na página 143.

[14] Definição na página 145.

[15] Este paradigma de programação pode ser visto nos trabalhos de Gelfond & Lifschitz [1988, 1991], sendo bastante conhecida a linguagem Prolog e derivações.

[16] Definição na página 33

ções explicitamente convencionadas à interação medicamentosa. A linguagem farmacológica tradicional é tratada por complexas modelagens que formam estruturas hierárquicas ou difusas de entidades biológicas e relacionamentos de fenômenos ou ações. A transposição da linguagem farmacológica é um recurso intuitivo baseado no conhecimento tradicional explícito. Conforme citado, boa parte das estruturas elaboradas viabilizaram a previsão de interações medicamentosas sob modelagem farmacocinética. As variáveis deste domínio são frequentemente mensuráveis (relações da concentração do fármaco) ou são diretamente relacionáveis por compartilhar a ação de enzimas metabólicas.

### 1.4.4 Limitações das técnicas existentes

Considerando a tendência em modelar os dados farmacológicos a partir do significado explícito dos elementos envolvidos, não verificou-se na literatura um modelo geral para descoberta de conhecimento em bancos de dados de combinações medicamentosas que não demande artifícios de seleção manual criteriosa dos atributos que descrevem as entidades abordadas e modelem computacionalmente a linguagem farmacológica com o compromisso de consistência explícita, ou seja, necessariamente geram-se modelos restritos à compreensão, racionalidade, humana.

Os modelos limitados à acuidade e forma de expressão humanas, ou que devem obrigatoriamente remontar ao conhecimento atual, restringem a capacidade de verificação de novos conhecimentos. Um exemplo paradigmático da tendência em aproximar a concepção da verdade a estruturas arraigadas na consciência humana é a insistência observada na década de 90 em se enxergar o código genético[17] como uma cadeia de letras (bases nitrogenadas) que formam palavras (códons), as quais são interpretadas (transcrição) por enzimas e constituem posteriormente frases (proteínas). Embora a codificação em linha seja frequentemente observada, um modelo mais apropriado para a interpretação do código genético é na forma de grafos[18], dado que, contrariamente ao senso comum de então, as estruturas não correspondem a uma linearidade tácita, mas dinâmica, ora degenerada, ora polissêmica, cuja informação pode ser interpretada de várias formas conforme o processo.

Diante do exposto, verificou-se que a regra geral implementada por estes modelos assume *a priori*, que existem características farmacológicas que participam do fenômeno da interação baseada em fatores biológicos que envolvem rotas metabólicas ou dinâmicas. A explicação almejada, em última instância, é a determinação de cada interação a partir do compartilhamento de biomoléculas como receptores ou enzimas ou ações fisiológicas.

### 1.4.5 Uma proposta holística

Os estudos que determinam especificidades farmacodinâmica ou farmacocinética procuram explicar diretamente os fenômenos abordados ao formar uma estrutura causal com a hipótese

[17] Estruturas de DNA e RNA, conforme descrito na página 139.
[18] Definição na página 152.

proposta. Mais especificamente, desejam saber qual atributo do fármaco ou da interação está relacionado com a interação.

O ponto de vista adotado pelo presente modelo é de que, conhecer o fármaco como um todo, ou seja, de forma holística, é pré-requisito para a verificação de padrões indiretos que cerceiam as explicações tipicamente fornecidas. Com isso, pretende-se ampliar a capacidade explicativa dos fenômenos e identificar padrões desconhecidos de modo a ampliar as explicações dos modelos farmacológicos e inovar com modelos preditivos em conhecimento latente. Assim, toda e qualquer informação acerca dos fármacos, incluindo características farmacotécnicas, clínicas e epidemiológicas, caracteriza o fármaco e, consequentemente, podem fornecer subsídio que estabeleça relações diretas e indiretas com o sistema em que está inserido (no caso nos referimos ao sistema biológico do nível molecular ao indivíduo e social quando remete-se a populações).

Na presente abordagem, o fármaco é representado como um vetor de características empíricas (e.g., absorção, biodisponibilidade) ou de elementos coletados em diversas bases de dados (e.g., mapas metabólicos, informações acerca de enzimas). A exploração do conhecimento implícito acerca dos fármacos requer a avaliação sem escolha *a priori* de atributos, ou seja, sem pré-determinação de quais atributos estão relacionados com o fenômeno estudado. Desta forma, não são utilizados somente os atributos que reconhecidamente explicam o fenômeno de interações medicamentosas. Ao tomar cada fármaco como um conjunto de características inicialmente independentes, o modelo proposto estabelece modelos dedutivos de exploração do universo completo de possibilidades de interação entre todos os fármacos descritos ao estabelecer uma estrutura comparativa global de cada atributo para a extrapolação local do conhecimento acerca das interações medicamentosas previamente conhecidas.

Diante das limitações clínicas, experimentais e computacionais, conjectura-se que interações medicamentosas devem ser avaliadas com abordagem diferente do paradigma predominante. Conjectura-se que a exploração exaustiva de atributos não usuais com técnicas estado da arte de mineração de dados fornece uma alternativa genuinamente capaz de descoberta de novo conhecimento.

## 1.5 Especificidades da solução

As modelagens usuais para o problema de previsão de interações medicamentosas avaliam de forma especializada o conhecimento biológico, o que implica na formação de métricas de distâncias específicas para interpretar a relação entre substâncias e restrição do espaço de hipóteses (número de combinações avaliadas[19]) devido, em parte, à limitação ao contexto do conhecimento em vigor. No entanto, as descobertas se afastam do caráter especulativo quando observadas, ao menos, quanto a utilização por populações. Este pontos são relatados a seguir.

[19] Definição na página 172.

**Exploração de atributos *a priori vs a posteriori*** Conforme exposto, a capacidade de extração do conhecimento pelas relações comparativas entre os fármacos é reduzida ao avaliar apenas os atributos diretamente relacionados ao fenômeno de interação medicamentosa. A caracterização dos fármacos em sua essência não pode ocorrer apenas na correlação direta da farmacologia da interação. Acredita-se que o conhecimento latente emerja ao estabelecer relações entre as características do objeto estudado e o mundo, o sistema envolvido.

**Métrica de distância para comparação fármacos representados como "entidade-atributo"** As comparações não são realizadas com os objetos em si, mas com um comparador intermediário usado para todos os elementos do espaço de hipóteses. As métricas podem ser aprimoradas, no entanto, o uso de diversas métricas de distância permite a exploração sob múltiplos prismas de modo a estabelecer diferentes visões do mesmo conceito. A multiplicidade de visões exploradas sistematicamente pode criar modelos com maior caráter informativo do que o uso de apenas uma visão dada como certa. Tal abordagem pode evitar o viés, a paralaxe da incerteza inerente a uma observação.

**Exploração seletiva *versus* exploração completa do espaço de hipóteses** Os modelos preditivos baseados em um reduzido espaço de hipóteses possuem menor número de fármacos, o que mitiga a definição do escopo de cada atributo utilizado para o modelo preditivo e a capacidade de generalização. A incerteza da observação decresce com o número de observações. Conforme abordado na seção B.2.1, a avaliação indutiva parte de um modelo restritivo a cada contexto específico para estabelecer generalizações. No entanto, a verificação de relações complexas demandam a avaliação da rede como um todo, de modo que as características de cada elemento sejam avaliadas simultaneamente entre os objetos próximos e os semelhantes e, ainda, de forma sistêmica. Deste modo, os atributos serão melhor caracterizados conforme contemplarem mais pontos de vista (formas de apreensão) e mais instâncias (objetos e relações ontológicas[20]).

**O modelo é limitado ao contexto abordado, à caracterização das dimensões (entidades e atributos)** Se o domínio do conhecimento é limitado ao contexto farmacocinético, apenas interações deste tipo serão identificadas. Se gerado restritamente a partir de coleta de dados populacionais, apenas os fármacos envolvidos em polifarmácias serão usados para determinar relações. Interações sinérgicas não serão identificadas se a base do conhecimento apenas contemplar interações adversas. Um padrão-ouro com poucas instâncias ou documentação restringirá a cobertura preditiva e, possivelmente, a especificidade, dada a reduzida caracterização dos atributos periféricos que subsidiem a demanda da explicação causal com aporte canônico a ser utilizada ao transpor previsões ao teste empírico.

[20] A ontologia é a categorização do ser enquanto objeto de estudo, ou seja, avalia a realidade e a condição existencial de entes. Na computação representa a identificação e a determinação de papéis que contribuem para sua definição em um dado sistema. Definição na página 153.

**A utilidade das interações deve ser observada na prática clínica** Os autores que usam modelos em fontes clínico-populacionais, embora sejam limitadas aos fármacos e combinações, possuem maior especificidade nos achados. Os demais trabalhos, sobretudo ao abrigarem grande número de fármacos, devem estabelecer correspondências em usuários de medicamentos, e, inclusive, caracterizar o uso e amadurecer observações que caracterizam desfechos clínicos.

A seguir, será delineado o cenário de aplicação dos fenômenos farmacológicos aos casos possíveis de interação medicamentosa diante da consistência esperada para a prática da saúde baseada em evidências.

## 1.6 Objetivo

O presente trabalho visa estabelecer um modelo para descoberta de conhecimento em bancos de dados massivos para detecção de interação medicamentosa potencial existente e previsão de novas interações, ou seja, ainda não comprovadas pelos métodos científicos canônicos. O modelo deve caracterizar cada elemento de um amplo conjunto de fármacos pela extração direta ou indireta do conhecimento atribuído e categorizado na forma de atributos. Esta extração deve ocorrer diante da comparação de todos os atributos disponíveis e combinações possíveis par a par.

Objetiva-se viabilizar de forma automática a geração de modelos preditivos que sejam capazes de selecionar e correlacionar fatos farmacológicos catalogados para a previsão de novas interações medicamentosas. A extrapolação do conhecimento farmacológico é realizada mediante modelagem dos dados e uso de técnicas estado da arte de aprendizado de máquina. A relevância das previsões é avaliada quanto ao uso por populações e por revisão da literatura científica.

A premissa maior do modelo fundamenta-se na existência de características intrínsecas do fármaco relacionadas diretamente ou indiretamente a capacidade de interação.

A descoberta de novas interações envolve o uso inicial de toda e qualquer informação. Nenhuma variável é descartada *a priori*, visto que o modelo deve extrair a semântica implícita a partir da definição do escopo das características que definem o fármaco frente a exploração completa da combinação dos fármacos selecionados. Desta forma, a correlação de atributos tratados de forma independente possibilita o posicionamento de cada fármaco frente aos demais e de cada combinação contida no espaço de hipóteses. Este posicionamento faz com que atributos diretamente relacionados a interações possam ser complementados pelos demais atributos que descrevem o fármaco. Acredita-se que esta forma de modelar dados e informações possibilita a descoberta de novo conhecimento sem demandar especialistas para escolha de atributos, liberando-os ao âmbito da *praxis*.

Propõe-se um modelo que gera modelos com acuidade preditiva, ou seja, um metamodelo. O metamodelo abriga o conceito “entidade-atributo”, visto que as entidades são melhor carac-

terizadas conforme cresce o número de atributos e quanto mais entidades descritas, aumenta o poder informativo e discriminativo do atributo.

A integração do conhecimento gerado acerca de interações medicamentosas previstas estabelece novos veios condutores para fomentar estudos clínicos, populacionais e revisões sistemáticas.

### 1.6.1 Objetivos específicos

- Identificar propostas de previsão de interações medicamentosas com métodos de inteligência artificial por meio de uma revisão sistemática.
- Estabelecer a temática da previsão de interações medicamentosas por métodos de inteligência artificial e situar o modelo proposto.
- Coletar e harmonizar dados farmacológicos, padrão-ouro de interações conhecidas e dados farmacoepidemiológicos.
- Construir um metamodelo de engenharia de dados, processamento de dados e validação a partir de padrão ouro para conjugar técnicas algébricas que favoreçam a construção do modelo preditivo supervisionado de interações medicamentosas.
- Verificar sistematicamente as previsões na literatura científica.
- Avaliar a relevância das previsões em populações de usuários de medicamentos.
- Disponibilizar o código-fonte do modelo com licença GNU *General Public License* (Licença Pública Geral).

O trabalho proposto introduziu inovações na exploração computacional de interações medicamentosas, cujos aportes são sumarizados a seguir.

## 1.7 Contribuições

O metamodelo proposto e implementado intitula-se DataMInt, *Data Mining of Interaction* e a logo incialmente proposta é mostrada na figura 1.1. Os pilares trazidos neste texto possibilitaram descobertas, cujas contribuições estão pontuadas a seguir.

**Exploração completa do espaço de hipóteses de fármacos aos pares** O algoritmo gerado a partir do modelo proposto abrangeu uma quantidade inédita de fármacos e combinações, tangendo o valor de um milhão de combinações, considerando todos os pares possíveis do conjunto de fármacos avaliados (capítulo 5).


Figura 1.1: Logo proposto para o metamodelo implementado de Mineração de Interações Medicamentosas: DataMInt - *Data Mining of Interaction*. Simboliza uma árvore, linha dos algoritmos de aprendizado de máquina mais bem sucedidos nesta primeira abordagem.

**Extração de conhecimento farmacológico latente para a previsão de interações** A avaliação do fármaco enquanto entidade expressa em um espaço n-dimensional promoveu a previsão de interações medicamentosas enquanto características intrínsecas dos fármacos observados em conjunto, sem a adoção direta do fenômeno da interação medicamentosa na entrada dos modelos (capítulo 4).

**Manejo da complexidade biológica** O modelo proposto representa uma solução viável para a extração de conhecimento. A abordagem não demanda laborioso pré-tratamento técnico (*ad hoc*) das informações, se adequando a diversas formas de expressá-la, seja em linguagem natural (texto), numérica, categórica ou hierárquica (taxonomia[21] ou ontologia).

**Desempenho e performance** A validação do modelo atingiu os melhores niveis de desempenho observados na literatura (capítulo 3) perante um amplo padrão ouro. As previsões ocorrem com bom desempenho (elevada acurácia e precisão) e performance (reduzido tempo de processamento), relevante ao crescente acúmulo de dados científicos (capítulo 4). O modelo estabelece vínculo das interações conhecidas com as previstas por extrapolação das funções de aprendizagem baseadas em fontes disjuntas ao padrão ouro (capítulo 5) o que afasta efeito de sobreposição (*overffiting*).

**Verificação da utilidade dos achados com base farmacológica em usuários de medicamentos** Poucos trabalhos verificaram populacionalmente os resultados previstos com base farmacológica, questionando-se a relevância das previsões, dado que podem jamais virem a ser utilizadas. Embora o espaço de hipóteses tenha sido na ordem de um milhão de pares

[21] Definição hierárquica de grupos de objetos com base em características comuns.

de fármacos, verificou-se, em duas populações, a ampla utilização das combinações previstas como interações (capítulo 6).

**Previsão de combinações com potencial sinérgico** Não verificou-se na literatura modelo que possibilite em caráter amplo e sistemático, a detecção de farmacoterapias com potencial sinérgico, ou, ao menos, seguro. Embora esta área deva ser mais explorada, o modelo proposto apresenta com ineditismo esta questão (capítulo 5).

**Sistematização do conhecimento enquanto disciplina da previsão computacional de interações medicamentosas** Ao realizar a primeira revisão sistemática do tema (capítulo 3), o presente trabalho lançou a pedra fundamental para elencar o que se produziu de conhecimento, propiciando estabelecer enquanto disciplina esta modalidade de avaliar interações medicamentosas a partir de atributos farmacológicos e epidemiológicos com base em estruturas de dados complexas e processamento de grandes quantidades de informação.

## 1.8 Organização do texto

Este texto foi segmentado em sete capítulos, contudo, o material complementar de apoio pode ser relevante ao leitor conforme mostrado na figura 1.2.

A introdução emerge a importância do tema à luz dos fatos históricos e científicos.

Além do referencial teórico contido no capítulo 2, o aspecto transdisciplinar de técnicas de bioinformática demanda a condução de leitores a domínios infensos à sua área de atuação. Como artifício, ofertou-se uma possível horizontalização da temática com um referencial teórico distribuído no apêndice A entre as seções A.1 ao A.5. O percurso traçado para a descoberta de interações medicamentosas é descrito no apêndice A.1. O apêndice A.2 relata fundamentos das disciplinas *in vitro*, frequentemente chamadas de "Laboratório Molhado" (do inglês *wet lab*). Estudos *in vivo* ou ensaios clínicos são descritos no apêndice A.3. Questões epidemiológicas, ou *in populo*, são abordadas no apêndice A.4. Finalmente, a metologia computacional que constitui o domínio *in silico* foi complementada no apêndice A.5.

No capítulo 3 são relatados dez trabalhos identificados por uma inédita revisão sistemática da literatura. São artigos completos, análogos ao modelo proposto, de abordagens de inteligência artificial voltadas à previsão de interações medicamentosas com base farmacológica ou populacional.

O modelo sugerido é conceituado no capítulo 4. Uma extensão deste capítulo é abordada no apêndice B, o qual discorre sobre as implicações em investigar interações sobre o ponto de vista da teoria do conhecimento, algébrico, computacional e de mineração de dados.

Interações medicamentosas potenciais sob o ponto de vista farmacológico são exploradas no capítulo 5, e sob o ponto de vista populacional no capítulo 6.

Finalmente, contribuições, colaborações e desdobramentos são apontados no capítulo 7.

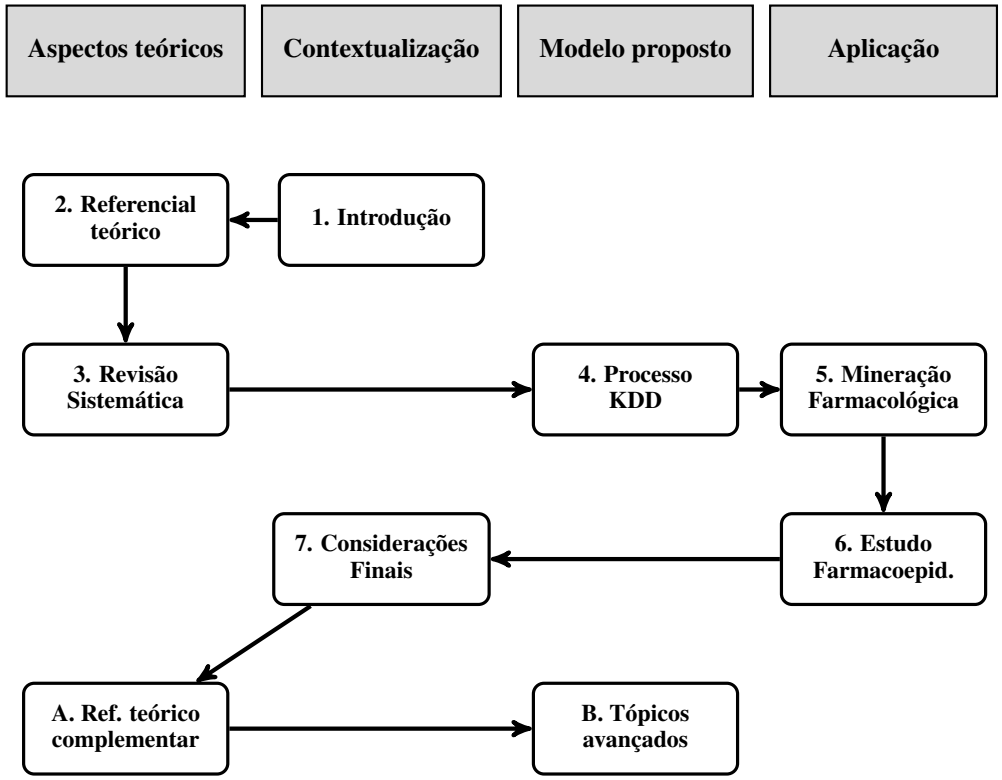



Figura 1.2: **Dinâmica de leitura da tese.** Os capítulos são indicados com número e os apêndices com letras. A tese pode ser lida ao longo das setas ou colunas de modo que o leitor se familiarize com o tema proposto.

Referências, sobretudo ao referencial teórico, são disponibilizadas na forma de *link*, motivo pelo qual recomenda-se a leitura em formato eletrônico.

# Capítulo 2

# Referencial teórico

A seguir são colocadas informações básicas para o entendimento do metamodelo DataMInt. Aspectos complementares podem ser vistos nos apêndices A.1. experimentação científica, A.2. *in vitro* A.3. *in vivo*, A.4 *in populo* e A.5 *in silico*.

## 2.1 Fármaco

Os fármacos são moléculas com atividade biológica obtidas a partir de bactérias, fungos, animais, vegetais ou síntese química. Possuem potencial de prevenir ou curar doenças com melhora do estado físico ou mental [Katzung, 2003].

O fármaco é uma tentativa de mimetizar algum papel biológico de biomoléculas, na tentativa de corrigir a homeostasia afetada por alguma causa interna ou externa ao organismo e altera processos fisiológicos de tecidos ou organismos [WHO, 1994].

Além das finalidades paliativas, profiláticas ou curativas, o fármaco pode ser utilizado com fins diagnósticos [Katzung, 2003].

### 2.1.1 Propriedades físico-químicas

O fármaco pode ocorrer nos três estados da matéria, na forma sólida, líquida ou gasosa. Este fator influencia a via de administração, e.g., oral, injetável ou inalação.

As propriedades de carga/polaridade e permeabilidade em membrana também influem na formulação de um fármaco. Caso o sítio de ação, em geral um receptor, não for acionado na medida necessária devido a baixa permeação ou degradação do fármaco pelo sistema gastro-intestinal ou hepático, as características de permeação são moduladas pela modificação de grupamentos químicos, o que o torna estável e eficaz. A molécula concebida para ativar após metabolização é chamada **pró-fármaco**.

Pureza e produtos de degradação, viscosidade, densidade, solubilidade, pH ou pka (escala logarítmica correspondente à acidez), ponto de fusão ou ebulição, granulometria, cor, absorti-

vidade molar, hidratação, massa molecular, quiralidade, polimorfismo e teores de grupamentos químicos são exemplos de atributos físicos ou químicos que não são comumente atrelados à explicação de interações medicamentosas, porém constituem parte da identidade de um fármaco. Estas características combinadas com as demais podem ser usadas pelo modelo integrativo, introduzido no capítulo 4, para a detecção de padrões não triviais de modo a captar possíveis interações diante da definição da essência de cada fármaco comparativamente.

## 2.2 Medicamento

O medicamento abriga substâncias ativas e adjuvantes farmacotécnicos ou terapêuticos em uma forma farmacêutica[1] para viabilizar a qualidade do uso e o objetivo farmacológico pretendido [BRASIL, 2010a].

Os ensaios preliminares *in vitro* e *in vivo*, em geral, empregam o fármaco na forma de substância química purificada e solubilizada em algum solvente ou mistura e acondicionada em condições brandas, i.e., sob temperatura reduzida e ao abrigo da luz. O desenvolvimento farmacotécnico acompanha a fase subsequente à definição da atividade terapêutica, de modo a veicular o fármaco com a melhor performance farmacológica e estabilidade química.

Após a verificação aproximada das dosagens terapêuticas, tóxicas e letais e alguns mecanismos bioquímicos, em geral, a atuação em enzimas metabólicas importantes como as do citocromo P450, a forma farmacêutica é escolhida em função das características do fármaco e dos pacientes alvo. Fármacos administrados com a mesma finalidade terapêutica, porém por vias diferentes devem atingir a mesma biodisponibilidade, isto é, a concentração sanguínea capaz de promover a ação terapêutica.

A escolha da forma farmacêutica é crítica para a **adesão** ao tratamento. O sucesso em tratamentos longos será limitado caso houver desconforto como aplicações frequentes ou dolorosas, ou ainda, características referentes a organolépticas[2] desagradáveis. Um medicamento na forma de comprimidos pode não ser aceito para o tratamento de crianças, pacientes com dificuldade de deglutição ou concomitante a dada condição de saúde que impele ao vômito. A forma farmacêutica, desinformação, acesso, preço, tratamentos estigmatizados como o uso de neurolépticos ou medicamentos para hanseníase, podem causar o uso inadequado ou insuficiente do tratamento, culminando na falha terapêutica por falta de adesão.

O fármaco pode ser considerado **eficaz** e obter efeitos farmacológicos favoráveis sem que o medicamento seja **eficiente**, ou seja, não manifeste resultados terapêuticos esperados. O medicamento condensa as características técnicas e sociais, ambas igualmente importantes para o estudo do efeito de medicamentos em uso concomitante.

[1] Forma física de veiculação do fármaco como comprimidos, drágeas, cápsulas, pomada, gotas, injetável, entre outras.

[2] Cor, odor, sabor, entre outras características percebidas pelos sentidos humanos.

## 2.3 Eventos Adversos

**Reação Adversa a Medicamento**, RAM, é qualquer resposta prejudicial e não intencional nas doses normalmente usadas de medicamentos. A ANVISA, Agência Nacional de Vigilância Sanitária, classifica **efeito colateral** como reação adversa inesperada e séria. O evento inesperado é desconhecido por pesquisadores e, portanto, não catalogado. O evento sério pode resultar em morte, hospitalização prolongada ou morbidades com prognóstico desfavorável como câncer [BRASIL, 2009].

### 2.3.1 Classificação de reações adversas a medicamentos

Tradicionalmente, eventos adversos são classificados como reações tipo A (exageradas) e tipo B. O primeiro grupo envolve respostas exageradas, sendo geralmente dose-dependente e previsíveis. O segundo grupo é relacionado às ações farmacológicas desconhecidas frequentemente causadas por mecanismos imunológicos ou farmacogenéticos, sendo comum a não dose-dependência [Lee, 2009].

Os aspectos mecanicísticos de reações adversas a medicamentos são geralmente concentrados em agentes biológicos sob fatores de susceptibilidade, mecanismos farmacológicos ou imunológicos ou ações de metabólitos. A sigla EIDOS condensa uma classificação, em que $E$ corresponde às espécimes extrínsecas que iniciam o efeito, $I$ a espécime intrínseca afetada, $D$ indica o fator de distribuição do agente, $O$ indica o desfecho fisiológico ou patológico e $S$ corresponde à sequela, ao evento adverso propriamente dito [Aronson, 2011].

Os fatores clínicos de eventos adversos podem ser descritos pelo sistema DoTS, o qual inclui planejamento de farmacovigilância, aspectos de prevenção e recomendações de procedimentos regulatórios para novos fármacos [Calderón-Ospina & Bustamante-Rojas, 2010; Aronson, 2011]

## 2.4 Interação Medicamentosa

Em 1972, a Organização Mundial da Saúde descreveu interação medicamentosa como efeito "nocivo e não compreendido, o qual pode ocorrer em doses normalmente empregadas pelo homem para profilaxia, diagnóstico ou tratamento de doenças, ou para modificação de função fisiológica" [Lin et al., 2010].

A interação medicamentosa ocorre quando um ou mais fármacos afetam a atividade, metabolismo ou toxicidade de outro fármaco.

O delineamento das causas da interação ocorre na esfera epidemiológica, diante da avaliação dos fatores que relacionam o uso de medicamentos à eventos de saúde; e na esfera farmacológica, a qual mapeia relações entre fármacos, biomoléculas e entes biológicos. A integração deste conhecimento contribui para o manejo da farmacoterapia mitigando danos quando a combinação é inevitável.

### 2.4.1 Classificação de interações medicamentosas

Interações medicamentosas são classificadas dicotomicamente, conforme sua existência [Wishart et al., 2008] ou de modo mais detalhado segundo a **gravidade** em "maior", "moderada" ou "menor" [DRUGS.COM, 2011; Tatro, 2012], **desdobramento** (rápido em até 24h, ou retardado em dias ou semanas) e **documentação** ("estabelecida" em estudos controlados; "provável", sem prova clínica; "suspeita", com evidências que precisam de maiores estudos; "possível", pode ocorrer, mas os dados são limitados; e "improvável", duvidoso, não há boa evidência de efeito clínico) [Tatro, 2012].

Mecanismo, gerenciamento (recomendações para redução ou prevenção dos efeitos), efeitos, acompanhamento (parâmetros clínicos ou laboratoriais) e ajuste de dosagem frequentemente complementam a informação sobre a interação classificada.

A classificação pode ser adequada diante das especificidades metodológicas. Harpaz et al. [2010b] realizou a avaliação de efeitos adversos e interações medicamentosas adversas com base em regras de combinações conforme mostrado na tabela 2.1. Estas regras indicam a presença de combinações espúrias em que não ocorre a interação devido ao evento adverso ser atribuído a um dos fármacos, ou quando um fármaco trata o evento adverso de outro.

Tabela 2.1: **Taxonomia para reações adversas e interações medicamentosas.** Harpaz et al. [2010b] quantificou uma amostra de 6.725 medicamentos contidos em 163.944 notificações de eventos adversos suspeitos do FDA, agência estadunidense de fármacos e alimentos.

| Nível | Descrição | Casos |
|---|---|---|
| **Medicamento** ($n \approx 30.000$ entradas) | | |
| A1 | Medicamentos conhecidamente associados/ tratam a mesma indicação | 57% |
| A2 | Medicamentos com o mesmo ingrediente ativo | 2% |
| A3 | Fármacos supostamente não relacionados | 41% |
| **Efeito adverso** ($n = 3.402$) | | |
| B1 | Um dos fármacos conhecidamente causam o efeito | 22% |
| B2 | Todos os medicamentos causam o efeito | 21% |
| B3 | Nenhum dos medicamentos causam o efeito | 27% |
| B4 | Associações confusas, medicamentos usados para tratar efeitos adversos | 30% |
| **Interação** ($n = 1.868$) | | |
| C1 | Interação medicamentosa conhecida | 35% |
| C2 | Interação medicamentosa desconhecida | 65% |

### 2.4.2 Interação farmacodinâmica

A interação farmacodinâmica direta ocorre quando os fármacos atuam no mesmo sítio, como agonistas ou antagonistas, ou quando atuam em vias distintas culminando no mesmo efeito. Fármacos psicoativos combinados, como opioides e sedativos, frequentemente acionam os receptores, os propagadores da ação, com consequente modificação da dinâmica molecular (e.g.,

potenciação ou competição) e, consequentemente, do efeito farmacológico [Aronson, 2011]. O efeito anticoagulante da varfarina é aumentado com o uso de esteroides (hormônio) ou tetraciclinas (antibiótico).

A interação farmacodinâmica indireta ocorre quando um fármaco interfere no efeito farmacológico, terapêutico ou tóxico de outro de forma independente do efeitos de ambos. Por exemplo, medicamentos usados para o tratamento de arritmia cardíaca podem ser afetados por modificações no balanceamento eletrolítico causado por diuréticos [Byrne, 2003].

### 2.4.3 Interação farmacocinética

Uma interação farmacocinética[3] ocorre quando um medicamento afeta as taxas de absorção, distribuição, metabolismo ou excreção de outro fármaco. Este tipo de alteração é monitorado com parâmetros clínico-laboratoriais como a concentração sérica máxima, tempo de meia vida, entre outros. A absorção de antibióticos como fluoroquinolonas ou tetraciclinas é prejudicada na presença de alimentos ou antiácidos que contenham ferro ou cálcio [Byrne, 2003].

### 2.4.4 Mecanismos de interações medicamentosas

A interação medicamentosa pode ser descrita biologicamente como eventos moleculares encadeados, dado que o produto de determinada reação torna-se o substrato da seguinte.

Nesta abordagem, uma reação química é segmentada em elementos conceituais, tais como reagentes, produtos, reações, estequiometrias, taxas e parâmetros cinéticos. O posicionamento dos componentes em biocompartimentos pode especializar os papéis biológicos na análise ou simulação da rede de reações [Hucka et al., 2003].

A reconstrução de redes metabólicas despertou o desenvolvimento de ferramentas que automatizam grande parte do esforço. Estas ferramentas localizam genes associados a enzimas, recuperam informações em bancos de dados específicos para descoberta de funções conforme a classificação de enzimas EC e realizam o encadeamento dos eventos e das biomoléculas.

A DIO [Yoshikawa et al., 2004] é uma ontologia específica de interações medicamentosas, a qual permite a descrição encadeada de cada interação fármaco-biomolécula perfazendo o mecanismo da interação medicamentosa com enzimas e biomoléculas e consequências biológicas como a inibição ou indução. Outras ontologias correlacionam termos médicos como a UMLS [Bodenreider, 2004]), celulares ou rotas bioquímicas (GO [Ashburner et al., 2000]), relação fármaco-doenças (KEGG [Kanehisa, 2013]) ou redes semânticas que englobam estes aspectos de modo geral [Chen et al., 2009].

[3] "Farmacocinética" é definida na página 143.

## 2.5 Mineração de dados

Segundo Han & Kamber [2001] e Wang et al. [2005] a Mineração de Dados (*Data Mining*), também conhecida como **Processo de Descoberta de Conhecimento em grandes Bases de Dados**, KDD (*Knowledge Discovery in Databases - KDD*), em sua forma mais fundamental, é a extração de informações interessantes, não triviais, implícitas, previamente desconhecidas e potencialmente úteis a partir bases de dados massivas. É também conhecida como um conjunto de procedimentos que transforma dados em conhecimento a partir da extração de fontes originais para a análise dos modelos e padrões encontrados [Zaki & Meira Jr, 2014].

As bases de dados armazenam informações separadas em atributos com semântica implícita e em formatos diversificados, tais como números, datas, textos ou lista de valores. A descrição padronizada dos atributos e utilizada pelos Sistemas de Gerenciamento de Banco de Dados, SGBD, é armazenada na forma de **metadados**.

Técnicas de pré-processamento lidam com metadados e conjuntos de atributos de alta dimensionalidade para a redução no custo do processamento e melhora do desempenho. A consequente elevação das taxas de acerto evita perda das relações semânticas dos metadados. Ferramentas estatísticas com implementação de algoritmos eficientes vem se tornando um alvo importante para os mineradores de dados, a fim de resolver os problemas citados [Kriegel et al., 2007; Zaki & Meira Jr, 2014].

### 2.5.1 Aplicações

A mineração de dados (*Data Mining*) despontou na década de 90 a partir da viabilização do acúmulo de dados em armazéns conhecidos como *data warehouse* e consequente necessidade de extrair informações úteis.

Destacam-se dois ramos na mineração de dados em biologia sistêmica. O primeiro é a descoberta de conhecimento para extração de padrões ocultos de grandes massas de dados experimentais, resultando em hipóteses. O segundo constitui a análise baseada em simulação, a qual testa hipóteses com experimentos *in silico*, disponibilizando previsões para serem testadas por estudos *in vitro* ou *in vivo* [Wang et al., 2005].

A serviço da descoberta de interações medicamentosas o processo KDD é utilizado para a extração de dados farmacológico ou populacional, tratamento dos dados acerca de medicamentos para a formação da estrutura adequada aos algoritmos de mineração de dados, culminando na análise dos modelos e previsões de interações medicamentosas. Este processo vem sendo adotado de forma bem sucedida em diversos âmbitos, destacando-se os esforços do centro UPPSALA, colaborador da OMS, para monitoramento de sinais, isto é, eventos clínicos potencialmente correlacionados com medicamentos [Mann & Andrews, 2007].

### 2.5.2 Tarefas da Mineração de Dados

A aplicação do conjunto de técnicas do processo KDD para a geração do modelo preditivo com aprendizado de máquina inicia-se com a coleta, seguida da engenharia dos dados (limpeza, discretização, normalização, redução de dimensionalidade) para a definição dos atributos que alimentarão os modelos preditivos. Os modelos são formados a partir de métricas de similaridade entre entidades e funções de aproximação que delineiam o comportamento dos dados a partir de tarefas de *a*) previsão (classificação ou regressão), *b*) agrupamento ou *c*) regras de combinações (verificação de padrões frequentes). Na última etapa, a análise dos resultados permite a avaliação quanto a capacidade informativa do conhecimento minerado, ou seja, se o conhecimento é útil e não trivial [Zaki & Meira Jr, 2014].

### 2.5.3 Tarefas descritivas

As tarefas descritivas são realizadas na análise por agrupamento e padrões frequentes. A primeira aplica métricas no espaço n-dimensional que são capazes de distinguir os dados com funções de similaridade. A segunda avalias regras baseadas na incidência no conjunto de dados.

#### 2.5.3.1 Análise por Agrupamento

Segundo MacCuish & MacCuish [2011] *cluster analysis* é o estudo de métodos para agrupamento de dados quantitativamente, também conhecido como taxonomia numérica, a qual segue uma tendência humana natural em agrupar coisas, criar classes com ou sem profundidade no significado.

Esta análise procura separar os dados em grupos, os quais devem ter correspondência de significado e utilidade, capturando sua estrutura essencial.

Em muitos casos, as técnicas por agrupamento representam apenas um ponto de partida para outros propósitos, como o da sumarização de dados [Tan et al., 2005].

O objetivo das técnicas de agrupamento é salientar as similaridades dos elementos com o conjunto, bem como as diferenças em relação aos outros grupos. Consequentemente, uma boa métrica de separação prima pela homogeneidade dos agrupamentos. Dentre as técnicas destacam-se *K-means*, agrupamento hierárquico aglomerativo e DBSCAN [Tan et al., 2005; Zaki & Meira Jr, 2014].

### 2.5.4 Tarefas preditivas

As tarefas preditivas estabelecem valores contidos em um atributo a partir de outros atributos do conjunto. São contempladas a *a*) classificação, designação de uma classe a um objeto dentre um conjunto de classes pré-estabelecidas e *b*) regressão, definição que ocorre a partir de preditor

contínuo. Em outras palavras, é a tarefa de aprendizado de uma função alvo a qual mapeia cada conjunto de atributos em relação a um atributo pré-definido [Tan et al., 2005].

Assim como a segmentação ontológica das combinações ou eventos, isto é, caracterização descritiva e hierárquica de objetos e classes de objetos, as tarefas descritivas são capazes de agrupar entidades. Caso seja eleita uma entidade como exemplo do grupo ou seja formado um arquétipo a partir das entidades que compõe o agrupamento definido, o espaço de busca pode ser reduzido a esta representação e os efeitos de interações extrapolados para as entidades do subconjunto. As técnicas por agrupamento podem ser usadas para geração automática de taxonomia. Informação complementar acerca da exploração ontológica de entidades e atributos encontra-se no anexo B.3.2.3.

#### 2.5.4.1 Classificação

Este conjunto de técnicas atribuem classes pré-determinadas a partir do **treino** em instâncias conhecidas, adotando-se um conjunto de registros multi-atributos definidos por uma variável discreta chamada **classe**. O *teste* relaciona variáveis dos atributos (ou valores dos atributos) às categorias pré-definidas no treino [Veloso et al., 2006] em instâncias tomadas como desconhecidas contidas no padrão ouro. O modelo gerado é usado para realizar previsões nas instâncias desconhecidas, porém, sendo desejável a mesma estrutura de dados e teor de informação do treino para evitar falsas extrapolações.

Existem diversos modelos de classificação consolidados, como redes neurais, modelos estatísticos com discriminantes linear/quadráticos, árvores de decisão e algoritmos genéticos. Dentre esses métodos, árvores de decisão são particularmente apropriadas para mineração de dados. Árvores de decisão podem construídas relativamente rápido quando comparadas com outros métodos, além de serem de fácil compreensão [Veloso et al., 2006].

A seguir são destacadas técnicas consolidadas na tarefa de classificação.

**Árvore de decisão** **Árvores** são estruturas formadas por atributos distribuídos em **nodos** e **folhas**. Um **nodo raiz** é ramificado a outros nodos que podem se ramificar novamente. Os nodos terminais são chamados **folhas**. As folhas recebem as classes e os nodos intermediários recebem valores que são capazes de distinguir as instâncias. Os algoritmos de árvore de decisão são capazes de distinguir qual posição no vetor de medidas pode discriminar hierarquicamente as interações medicamentosas.

Considerando valores entre 0 e 1, uma ramificação da árvore hipotética com duas classes (“interação” e “não interação”) seria “absorção”>0,9, “toxicidade”<0,2 e “classificação anatômica”>0,75 $\rightarrow$ ”interação”. Nesta ramificação de decisão, os fármacos com perfis semelhantes de absorção e classificação anatômica e com toxicidades diferentes tendem a ser classificados como “interação”.

As árvores de decisão tem custo relativamente baixo, possibilitando rápida construção de modelos. Porém, encontrar uma árvore de decisão ótima é um problema $np-$completo[4] devido ao número elevado de hipóteses. Muitos algoritmos adotam soluções heurísticas[5].

Dentre os algoritmos destacam-se J48 e florestas randômicas.

**Classificadores Bayesianos** Frequentemente, instâncias que compartilham os mesmos atributos podem não ser designadas à mesma classe, ou seja, o padrão pode não ser determinístico.

O teorema de Bayes avalia a probabilidade conjunta de um evento $x \in X$ e um outro evento $y \in Y$ conforme ilustrado na equação 2.1[6].

$$P(Y|X) = \frac{P(X|Y)}{P(X)} \tag{2.1}$$

O classificador de Bayes adota o produto das probabilidades condicionais dos atributos $X$ com $d$ dimensões em função da classe $y$, supondo que os atributos sejam independentes:

$$P(X|Y=y) = \prod_{i=1}^{d} P(X_i|Y=y) \tag{2.2}$$

O método Bayesiano é particularmente útil na presença de pontos de ruído[7] isolados, visto que ao balanceá-los com a probabilidade condicional e diante de atributos irrelevantes, a probabilidade $P(X_i|Y)$ tende a se tornar uniforme [Tan et al., 2005] aproximando-se dos demais pontos.

**Classificação baseada em Regras** Esta técnica é usada em tarefas descritivas ou quando as regras tem como consequente (último termo) a classe a ser determinada. Nesta técnica ocorre a extração de subsequências ou subestruturas que aparecem no conjunto de dados com frequência não inferior a um limiar especificado pelo usuário. Adotam-se diferentes formas estruturais como subgrafos ou subárvores as quais podem ser combinadas com itens frequentes ou subsequências. Em um grafo[8], o padrão estrutural é caracterizado quando se identifica uma subestrutura frequente [Han et al., 2007]. Encontrar padrões frequentes desempenha um papel essencial nas combinações iminentes, correlações e em muitos outros relacionamentos interessantes entre os dados. Sendo assim, contribui para indexação de dados, classificação, agrupamento e outras tarefas de mineração de dados [Han et al., 2007].

Esta técnica abriga a classificação de registros sob as regras "se... então". Os algoritmos de classificação baseada em regras avaliam a relação entre atributos **antecedentes** que impli-

[4] Isto é, complexo, não tratável em tempo computacional.

[5] As computação heurística contempla métodos que oferecem uma solução ótima, não determinística, com foco em problemas não praticáveis com as técnicas estatísticas e computacionais fora de seu domínio.

[6] Lê-se $P(A|B)$ como probabilidade condicional de A dado B.

[7] "Ruído" são dados inadequados que não representam informação relativa à instância.

[8] Informações acerca da teoria dos grafos encontram-se na página 152.

cam nas classes **consequentes**, computando-se a frequência com que características associadas aparecem no banco. As frequências são ponderadas em relação ao total de instâncias abrangido e em relação ao número de instâncias como um todo.

As regras são avaliadas quanto a confiança relativa ao número de instâncias cobertas. Ou seja, para uma interação ser considerada "não interação" pondera-se quantas instâncias de "absorção"<0.9 devem ser consideradas. Se o limiar for inferior a algo pré-definido, esta regra é descartada.

As regras são agrupadas em relação às classes e ordenadas crescentemente conforme o número de termos. Desta forma, espera-se que as regras mais simples tenham melhor capacidade de expressão (uma aplicação para o conceito de navalha de Occam).

A expressividade do conjunto de regras é semelhante à da árvore de decisão, pois são expressas de forma completa e mutuamente excludentes. A diferenciação ocorre quando o classificador emite um número maior de regras diante de um conjunto delimitado de registros, restringindo e complexificando o perímetro de decisão.

A interpretação do conjunto de regras é frequentemente mais fácil do que para a estrutura gerada pela árvore de decisão, porém, o desempenho pode não ser o mesmo.

Frequências desbalanceadas das instâncias alocadas nas classes, como é o caso do presente estudo, podem influir na capacidade discriminativa de classificadores que ordenam as classes de forma balanceada.

**Classificador do vizinho mais próximo** A técnica de $k$ vizinhos mais próximos (KNN) toma a característica das instâncias como representações no espaço $d$-dimensional. A distância do par de fármacos é avaliada em relação aos demais, tomando-se os mais próximos como vizinhos. O centro dos pares considerados vizinhos é estabelecido e adquire um rótulo respectivo à classe. Caso os vizinhos tenham mais de um rótulo, a classe majoritária é atribuída.

A escolha de um valor de $k$ pequeno influi na sobreposição com casos considerados como ruído, ou seja, que não contribuem para a classificação. Se $k$ for abrangente, o classificador pode incluir pontos de dados longe da vizinhança e valorizar a classe mais incidente.

A técnica baseada no vizinho mais próximo atribui cada instância a uma classe baseando-se em medidas de similaridade entre instâncias posicionadas no espaço n-dimensional.

Classificadores como árvores de decisão ou baseados em regras são conhecidos como **aprendizes gulosos** pois descobrem um modelo mapeando cada atributo de entrada conforme a classe mais próxima. **Aprendizes sob demanda** memorizam a instância inteira e avaliam as correspondências exatas das instâncias desconhecidas com as conhecidas, descartando elementos sem esta característica.

Os classificadores por vizinhança são susceptíveis a ruído, pois fazem previsões baseadas em informações locais, ao contrário de árvores de decisão e regras de combinação que tentam contemplar globalmente o espaço de entrada.

Estes classificadores demandam ponderação das dimensões, pois todos atributos são tomados como vetores no mesmo espaço.

#### 2.5.4.2 Regressão

A regressão é a previsão do valor de um atributo numérico para determinadas instâncias baseada em funções preditivas estabelecidas a partir dos demais atributos. Função preditiva é a função que descreve instâncias conhecidas com base na projeção numérica dos dados, para, destarte, ser utilizada na previsão de instâncias desconhecidas.

Nesta categoria estão os métodos que podem ser descritos como tradicionais equações matemáticas, porém não enquadram-se em outras classificações, a exemplo do Naive Bayes (equação 2.2).

São exemplos de classificadores a regressão linear; regressão logística; SMO[9], baseado em núcleos polinomiais ou Gaussianos; *VotedPerceptron* e *RBFNetwork*, implementação de uma base radial Gaussiana, a qual deriva os centros e distâncias de unidades ocultas usando *k-means*[10] e combinando-as com regressão logística [Witten & Frank, 2005].

### 2.5.5 Avaliação da previsão

Métodos supervisionados de aprendizado de máquina que realizam a classificação constroem modelos baseados em instâncias conhecidas. Desta forma, a classificação é ponderada tomando-se casos de treino (instâncias conhecidas) como se fossem desconhecidos (teste), verificando-se os compromissos entre os acertos e erros dos valores positivos e negativos relativos a cada classe.

#### 2.5.5.1 Validação cruzada

O processo de validação cruzada consiste em distribuir randomicamente, sem reposição, instâncias conhecidas em partições contendo a mesma proporção das classes em relação ao todo. A cada iteração, uma partição é usada como teste para a avaliação do treino realizado nas demais partições. O processo é repetido até que todas as partições sejam testadas. O desempenho é avaliado tomando-se alguma relação de média pré-determinada.

#### 2.5.5.2 Desempenho

Cada métrica de desempenho é calculada conforme a incidência de acertos e erros das previsões que o classificador faz em instâncias conhecidas.

A escolha do modelo com base no desempenho ocorre com a ponderação dos valores mais próximos de zero segundo a taxa de erro e mais próximos de 1 segundo as demais métricas,

[9] *Sequencial Minimal Optimization*
[10] Técnica de análise por agrupamento.

destacando-se o coeficiente kappa e à área sob a curva ROC. Estas métricas são derivadas da matriz de confusão.

**Matriz de confusão** Na matriz de confusão são alocadas as frequências dos valores atribuídos às classes em relação à classe correta. Desta forma, se dentre 100 instâncias, 10 forem conhecidamente interação medicamentosa e o classificador atribuir 8 corretamente, os valores serão distribuídos conforme mostrado na tabela 2.2. Nesta tabela de elementos $f_{ij}$, o elemento $i$ representa os casos reais e o $j$ representa os previstos. Os valores corretos assumem que $i = j$, sendo $i = 1$ em classificações dicotômicas para verdadeiro positivos e $i = 0$ para verdadeiros negativos. Os incorretos são alocados nas posições em que $i \neq j$, sendo falsos positivos, $i = 1$ e falsos negativos $i = 0$.

Tabela 2.2: **Matriz de confusão hipotética**

| **↓real\|previsto→** | **interação** | **não interação** |
|---|---|---|
| **interação** | $f_{11} = 8$ | $f_{10} = 2$ |
| **não interação** | $f_{01} = 0$ | $f_{00} = 90$ |

A relação destes valores é adotada para a criação das métricas de desempenho. As equações abaixo podem ser aplicadas na classificação que envolva mais de duas classes.

A métrica mais intuitiva é a **precisão** conforme define Tan et al. [2005].

$$\text{Precisão} = \frac{\text{previsões corretas}}{\text{total de previsões}} = \frac{\sum_{i=1}^{n} f_{ii}}{\sum_{i=1,j=1}^{n} f_{ij}} \tag{2.3}$$

O cálculo da equação 2.3 intui o conceito de acurácia segundo Zaki & Meira Jr [2014].

A taxa de erro é a razão dos casos opostos à precisão.

$$\text{Taxa de erro} = \frac{\text{previsões incorretas}}{\text{total de previsões}} = 1 - \text{Precisão} = \frac{\sum_{i \neq j} f_{ij}}{\sum_{i=1,j=1}^{n} f_{ij}} \tag{2.4}$$

A medida-$F$ é a média harmônica entre precisão e revocação.

$$\text{medida-}F = 2 \times \frac{\text{precisão} \times \text{revocação}}{\text{precisão } + \text{revocação}} = \frac{2\sum_{i=1}^{n} f_{ii}}{2\sum_{i=1}^{n} f_{ii} + \sum_{i \neq j} f_{ij}} \tag{2.5}$$

A cobertura, ou revocação é obtida pela fração de previsões corretas para uma classe em relação aos número de elementos da classe $k$.

$$\text{cobertura} = \text{revocação} = \frac{\text{previsões corretas da classe}}{\text{total de previsões da classe}} = \frac{\sum_{i=1}^{n} f_{ik}}{\sum_{i=1}^{n} f_{k}} \tag{2.6}$$

A taxa de verdadeiros positivos da classe, conhecida como sensibilidade, é a fração de previsões corretas em relação à soma dos valores positivos.

$$\text{sensibilidade} = \frac{f_{kk}}{\sum_{k=1}^{n} f_k} \tag{2.7}$$

A taxa de verdadeiros negativos, conhecida como especificidade, é a revocação da negativa da classe.

$$\text{especificidade} = \frac{(\sum_{k=1}^{n} f_k) - f_{kk}}{\sum_{k=1}^{n} f_k} \tag{2.8}$$

**Coeficiente kappa de Cohen** A principal medida adotada pelo modelo foi o coeficiente kappa, o qual avalia a concordância entre as previsões conforme as classes. A acurácia da distribuição esperada ao acaso é avaliada em função da razão entre os valores corretos esperados ($\forall i = j$) em relação ao total, conforme observado nas equações 2.9 e 2.10.

$$\text{acurácia randômica} = \frac{\sum_{i=j}^{n}(\sum_{k=1}^{n} f_{ik} \times \sum_{k=1}^{n} f_{kj})}{(\sum_{i=1,j=1}^{n} f_{ij})^2} \tag{2.9}$$

$$\text{kappa} = \frac{\text{acurácia total} - \text{acurácia randômica}}{1 - \text{acurácia randômica}} \tag{2.10}$$

Segundo Landis & Koch [1977] um coeficiente superior a 0,81 indica uma concordância fidedigna. Logo, o modelo que obtém valor inferior a este limiar deve ser descartado.

**Área sob a curva ROC** A curva da Característica de Operação do Receptor, cuja sigla do inglês é conhecida por ROC, é uma representação gráfica para cada classe em que o eixo das abscissas recebe a taxa de verdadeiros positivos e o eixo das ordenadas recebe a taxa de verdadeiros negativos.

A curva é gerada a partir da ordenação das probabilidades emitidas pelo classificador para cada instância conhecida. A cada $k$ previsões, por exemplo, $k = 0,1 \times n$, a matriz de confusão é gerada e são calculadas a sensibilidade e a especificidade, cada qual constituirão as coordenadas $(x, y)$ respectivamente. Deste modo, um gráfico semelhante à figura 2.1 será gerado.

A área sob a curva ROC (AUC) é obtida usando-se a somatória das áreas dos trapézios de cada $k$ intervalo entre as coordenadas $x$ e $y$.

$$\text{AUC} = \sum_{i=1}^{k} \frac{y_i - y_{i-1}}{2(x_i - x_{i-1})} \tag{2.11}$$

### 2.5.6 Mineração de texto

Mineração de texto (*text mining*) pode ser caracterizada como o processo de análise de texto para extrair informação útil para fins específicos.

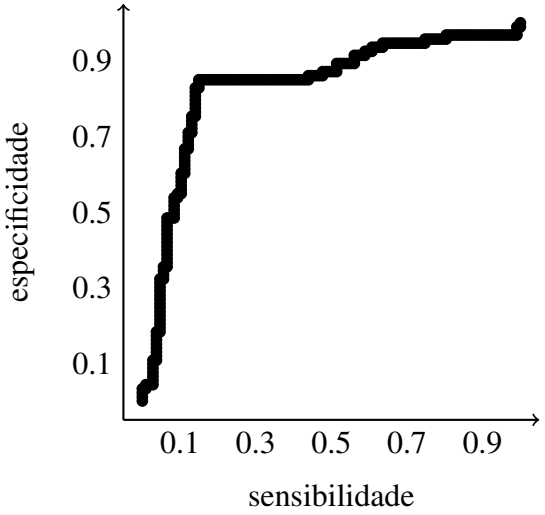



Figura 2.1: **Curva ROC hipotética.**

Ao contrário de sistemas de banco de dados, o texto é desestruturado, amorfo e difícil de lidar com algoritmos. No entanto, na cultura atual, o texto é o veículo mais comum para o intercâmbio formal de informações. O campo da mineração de texto geralmente lida com material cuja função é a comunicação de fatos, informações ou opiniões não classificadas desta forma. A motivação para tentar extrair informações de texto automaticamente é convincente, mesmo que o sucesso seja apenas parcial. Constitui-se, portanto, um desafio empregar-se esta técnica, sobretudo no contexto de extração de informação não trivial a partir da literatura farmacológica relacionada a medicamentos [Witten et al., 1999].

As informações conhecidas de um fármaco são frequentemente estruturadas em texto para avaliação dos profissionais de saúde em detrimento do formato computacionalmente extraível, como em bancos de dados relacionais [Duke & Bolchini, 2011].

A informação textual descreve a substância quanto a farmacocinética, farmacodinâmica, toxicologia, administração, posologia entre outros. Esta informação carrega valor semântico individual, a qual, no ato da prescrição, dispensação ou administração deve ser avaliada conjuntamente aos demais fármacos concomitantes do paciente seja sob a ótica terapêutica ou quanto a possibilidade de eventos não esperados.

Devido a terminologia frequentemente difusa, a avaliação pode ser prejudicada, sobretudo quando disponibilizada na forma de alertas por ferramentas de apoio a decisão.

# Capítulo 3

# Inteligência artificial para previsão de interações medicamentosas

Neste capítulo é apresentada uma revisão sistemática que objetivou recuperar estudos completos semelhantes ao presente trabalho. Identificou-se trabalhos de inteligência artificial que foram implementados e validados com base clínica ou a partir da literatura para descoberta de interações medicamentosas.

## 3.1 Métodos

A revisão sistemática foi realizada de acordo com as recomendações do centro colaborador Cochrane [Higgins & Green, 2011] e com o método PRISMA[1] [Liberati et al., 2009].

### 3.1.1 Elegibilidade

Apenas foi selecionado trabalho completo de conteúdo original publicado em periódico veiculado por fontes reconhecidas no meio científico.

Selecionou-se modelos implementados que adotaram técnicas de inteligência artificial ou aprendizado de máquina com ou sem abordagem de mineração de dados capazes de realizar detecção e que realizaram previsão de interação medicamentosa em humanos com base em dados farmacológicos ou clínico-epidemiológicos.

Foram incluídos trabalhos validados clinicamente ou experimentalmente, ou ainda, verificado na literatura quanto à capacidade de detectar, simular, prever ou identificar interação medicamentosa.

[1] *Preferred Reporting Items for Systematic Reviews and Meta-analysis*

### 3.1.2 Estratégia de busca

As buscas foram realizadas em fevereiro de 2013 nas bases EMBASE, MEDLINE, Base Cochrane para registros de ensaios clínicos controlados e LILACS.

Os termos e análogos, segundo a base MeSH[2] do Instituto Nacional de Informação Biotecnológica dos Estados Unidos relativos a "*artificial intelligence*" e "*drug interaction*" foram combinados usando o operador "*AND*", com sinônimos associados por "*OR*". A estratégia de busca completa é mostrada no apêndice C.

Foi realizada busca manual nas referências citadas pelos artigos incluídos, bem como as listadas em revisões sistemáticas afins. Preteriu-se "*machine learning*" como termo principal por ser considerado uma ramificação de inteligência artificial segundo a base MeSH.

### 3.1.3 Seleção

Desenvolveu-se uma plataforma *webservice* chamada Revis, implementada em php e mysql, para listar os trabalhos coletados. Esta ferramenta aloca os trabalhos para a equipe de revisores de modo que duas opiniões concordantes incluam ou excluam o estudo em três etapas.

Na primeira etapa é realizada a leitura de títulos e informações de fundo do periódico, tais como veículo de publicação, ano, idioma e autores, com o objetivo de realizar uma primeira poda com base em diferenças grosseiras com os objetivos do estudo. A segunda etapa inclui a leitura do resumo e, por fim, é realizada a leitura completa dos trabalhos selecionados na etapa antecedente.

O cegamento é assegurado pela restrição ao acesso com senha individual e pela alocação randômica dos trabalhos.

Esta ferramenta contribuiu para diversas revisões sistemáticas submetidas a eventos e periódicos [Lemos et al., 2013; Machado et al., 2013]. A concordância inter-examinador foi altamente satisfatória segundo critério estabelecido por Landis & Koch [1977], com kappa 0,88 (0,86 a 0,90 para intervalo de confiança de 95%).

### 3.1.4 QUADAS

Não foi encontrada ferramenta específica para avaliação da qualidade clínica de estudos que adotam técnicas computacionais para previsão de eventos em saúde. Embora não esteja contido no escopo da revisão, considerou-se o propósito de investigar interação medicamentosa como correlato ao processo de diagnóstico. Desta forma, adaptou-se o método QUADAS[3] [Whiting et al., 2004] usado para revisão sistemática de estudos com fins diagnósticos.

A ferramenta original consiste em quatorze questões respondidas como "sim", "não" ou "incerto". Foram aplicadas nove questões em trabalhos com dados de pacientes. O QUADAS

[2] *Medical Subject Headings*
[3] *Quality Assessment of Diagnostic Accuracy Studies*

foi aplicado por dois revisores independentes e discordâncias foram resolvidas posteriormente por consenso.

### 3.1.5 Síntese de dados e análise

Os aspectos reportados foram a abordagem computacional, a fonte de dados, o método de validação e desfechos de saúde. Não foi realizada metanálise devido a diferença metodológica dos estudos, participantes e medidas de desfecho.

## 3.2 Resultados

591 registros foram identificados a partir das quatro bases adotadas, um destes a partir de busca manual. Após a exclusão de registros duplicados, restaram 574. Dentre os estudos excluídos, 263 (46,6%) não envolveram interações medicamentosas, frequentemente restritos à interação fármaco-biomolécula, como citocromo P450. 179 (31,75) foram excluídos devido ao tipo de estudo, ou seja, por não serem métodos implementados e validados. 122 (21,6%) foram excluídos devido a intervenção, visto que não são métodos de inteligência artificial ou aprendizado de máquina.

51 artigos foram eleitos para a leitura completa. Dentre os 41 excluídos, 19 foram por tipo de estudo[4], 12 por tipo de participante[5] e 10 por tipo de intervenção[6]. Finalmente, 10 estudos corresponderam aos critérios de elegibilidade conforme mostrado na figura 3.1.

Dentre os estudos incluídos, cinco foram conduzidos nos Estados Unidos, dois na Europa e três na Ásia, abrangendo um período de 26 anos (tabela 3.1). Os objetivos explicitados foram farmacovigilância [Estacio-Moreno et al., 2008], mineração de texto [Duke et al., 2012; Harpaz et al., 2010a; Tari et al., 2010; Segura-Bedmar et al., 2011b; Zhang et al., 2012a], mineração de dados [Harpaz et al., 2010a], padronização ou taxonomia de fármacos [Duke et al., 2012; Harpaz et al., 2010a], geração de um sistema computadorizado de apoio a decisão clínica [Kinney, 1986; Gottlieb et al., 2012], estabelecimento de alvos terapêuticos [Lin et al., 2010], elucidação de mecanismos de fármacos [Lin et al., 2010; Gottlieb et al., 2012; Percha et al., 2012; Zhang et al., 2012a], análises do vocabulário específico para recuperação de informações [Duke et al., 2012; Percha et al., 2012; Segura-Bedmar et al., 2011b; Zhang et al., 2012a] e sugestão de novas interações medicamentosas [Segura-Bedmar et al., 2011b; Gottlieb et al., 2012; Zhang et al., 2012a].

[4] [Ardizzone et al., 1988; Boyce et al., 2009; Cerrito, 2001; Del Fiol et al., 2000; Duda et al., 2005; Eyers & Reamtong, 2008; Fuhr, 2008; Gardner & Rizack, 1990; Gebhart, 2011; Gordon, 2008; Grime et al., 2010; Hampton, 2011; Han et al., 2012; Hartge et al., 2006; Hripcsak et al., 1996; Leone et al., 2010; Preferansky, 1992a,b; Yoon et al., 2011]

[5] [Broccatelli et al., 2012; Burton et al., 2009; Cheng et al., 2011; Harpaz et al., 2010b; Krejsa et al., 2003; Kuperman et al., 1994; Michielan et al., 2009; Segura-Bedmar et al., 2010; Speedie et al., 1992; van Puijenbroek et al., 2002; Villier et al., 2012; Yap et al., 2006]

[6] [Boyce et al., 2009; Del Fiol & Haug, 2009; Duke & Bolchini, 2011; Ebrahiminia et al., 2006; Escousse et al., 1987; Gray et al., 1991; Kam et al., 2011; Lin et al., 2011; Takarabe et al., 2011; Takigawa et al., 2011]

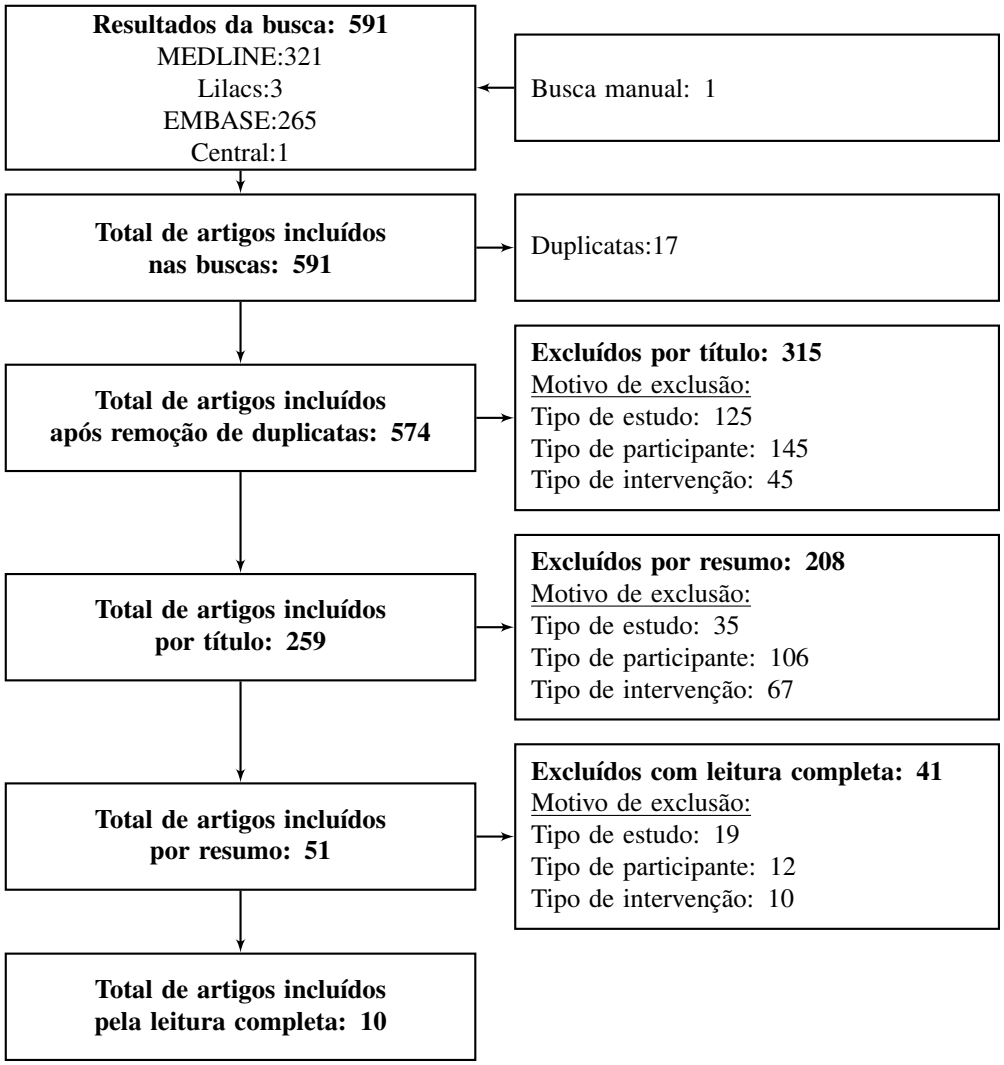



Figura 3.1: **Diagrama de inclusão de estudos de inteligência artificial aplicados a previsão de interações medicamentosas em humanos.** A busca foi realizada em fevereiro de 2013 nas bases EMBASE, MEDLINE, Cochrane e LILACS. O tipo de participante é "interação medicamentosa", o tipo de intervenção é "inteligência artificial" e o tipo de estudo é "modelo implementado e validado".

Apenas Estacio-Moreno et al. [2008], Lin et al. [2010] e Duke et al. [2012] abordaram desfechos de saúde, sendo eventos adversos, taxa de óbitos e miopatia, respectivamente. Nenhum estudo apresentou informações sobre tempo de processamento ou complexidade dos algoritmos.

### 3.2.1 Validação

Segura-Bedmar et al. [2011b], Tari et al. [2010] e Zhang et al. [2012a] adotaram precisão (proporção de verdadeiros positivos em relação ao total de positivos), revocação (proporção de verdadeiros positivos em relação a soma dos verdadeiros positivos e falsos negativos) e métrica-F (média harmônica entre precisão e revocação). Gottlieb et al. [2012] enfocou seus resultados na forma de área sob a curva ROC com validação cruzada. Percha et al. [2012] adotou uma

Tabela 3.1: **Características dos Estudos incluídos.**

| Estudo | Local | Modelo | Base de dados | Padrão Ouro | Validação |
|---|---|---|---|---|---|
| Kinney [1986] | EUA | PLM | Prontuários | Manual | média, desvio-padrão, qui-quadrado |
| Estacio-Moreno et al. [2008] | França | RA | Notificações | limiar da *British Medicines and Healthcare products Regulatory Agency* | Qui-quadrado |
| Harpaz et al. [2010a] | EUA | RA | Notificações | Manual (Micromedex, Epocrates) | suporte, *relative reporting ratio* |
| Lin et al. [2010] | Taiwan | Grafos, Agrupamento | Notificações | DrugBank, TTD, DART, and ADME-APs | Coeficiente de agrupamento |
| Tari et al. [2010] | EUA | PLN, PLM | Texto científico | DrugBank | precisão, revocação, métrica-F, acurácia |
| Segura-Bedmar et al. [2011a] | Espanha | PLN, PLM | Textos científicos | DrugBank | precião, revocação, métrica-F |
| Duke et al. [2012] | EUA | PLN, RA | Prontuários | Textos científicos | Revocação, acurácia |
| Gottlieb et al. [2012] | Israel | Classificação por regressão logística | BDF | Drugs.com, DrugBank | Validação cruzada, área curva ROC |
| Percha et al. [2012] | EUA | PLN, classificação | Textos científicos | DrugBank, drug lexicon | deixe-um-fora |
| Zhang et al. [2012a] | China | PLN; grafos; classificação por árvores, regressão logística e SVM | BDF | DDI Extraction Challenge, 2011 corpora | precisão, revocação, métrica-F, curva ROC, MCC |

PLN: Processamento de linguagem natural. PLM: Programação em Lógica Matemática. RA: Regras de Associação BDF: Bancos de dados com informação de fármacos, biomoléculas e medicamentos.

variação de deixa-um-fora aplicada a florestas randômicas. Duke et al. [2012] usou revocação e acurácia (proporção de verdadeiros positivos em relação ao total de instâncias). Estacio-Moreno et al. [2008] e Kinney [1986] usaram qui-quadrado.

### 3.2.2 Qualidade clínica

Apesar da amostra de tamanho reduzido, Kinney [1986] foi o único a estabelecer a eficácia do método diretamente em pacientes. Visando estabelecer comparações de qualidade pelo critério QUADAS, adicionalmente, foi considerado que os estudos de Estacio-Moreno et al. [2008], Harpaz et al. [2010a], Lin et al. [2010], Duke et al. [2012] e Gottlieb et al. [2012] proporcionaram diagnósticos válidos, apesar das previsões não terem sido avaliadas na prática clínica. A adaptação ao contexto preditivo não prejudicou a avaliação QUADAS, visto que a ferramenta

Tabela 3.2: **QUADAS - avaliação da qualidade dos estudos incluídos** e realizados em bases clínico-populacionais.

| **Estudo** | Kinney [1986] | Estacio-Moreno et al. [2008] | Lin et al. [2010] | Harpaz et al. [2010a] | Duke et al. [2012] | Gottlieb et al. [2012] |
|---|---|---|---|---|---|---|
| O espectro de pacientes foi representativo dentre os pacientes que recebem o teste na prática? | incerto | sim | sim | sim | sim | sim |
| O critério de seleção foi claramente descrito? | sim | não | sim | sim | sim | sim |
| O padrão de referência foi ligado corretamente à classificação da condição alvo? | sim | incerto | sim | sim | sim | sim |
| A amostra inteira ou uma seleção randômica da amostra receberam verificação usando um padrão diagnóstico de referência? | incerto | incerto | sim | sim | sim | sim |
| Os pacientes receberam o mesmo padrão de referência independentemente dos resultados do teste? | incerto | incerto | sim | sim | sim | sim |
| O padrão de referência foi independente do teste, isto é, o teste não incorporou o parte do padrão de referência? | sim | incerto | sim | sim | incerto | não |
| A execução do teste foi descrita detalhadamente, de modo a permitir sua replicação? | não | não | não | não | não | não |
| A execução do padrão de referência foi descrita detalhadamente, de modo a permitir sua replicação? | não | não | não | não | incerto | sim |
| Os resultados não interpretados ou intermediários dos testes foram reportados? | incerto | não | sim | não | sim | sim |

não incorpora um escore final de qualidade, devido a impossibilidade de determinar objetivamente o peso de cada aspecto abordado (tabela 3.2).

Estacio-Moreno et al. [2008], Harpaz et al. [2010a], Lin et al. [2010], Duke et al. [2012] e Gottlieb et al. [2012] usaram elevado número de registros, sendo potencialmente representativos dentre os casos possíveis. Estacio-Moreno et al. [2008] não explicitou o critério de seleção dos casos de interação e se o mesmo padrão foi empregado em todos os casos para verificar interações medicamentosas. Não ficou claro se a base de conhecimento usada nas previsões foram independentes do padrão ouro usado para validação em [Estacio-Moreno et al., 2008; Duke et al., 2012].

Tabela 3.3: **Precisão dos trabalhos incluídos**. Avaliação de interações medicamentosas em relação ao padrão ouro de interações conhecidas.

| Estudo | Registros e fonte de interações medicamentosas | Fármacos | associações | precisão |
|---|---|---|---|---|
| Kinney [1986] | 120 prontuários | 342* | 27 | 37,0% |
| Estacio-Moreno et al. [2008] | 3.249 notificações | 527 | 593 | 18,5% |
| Harpaz et al. [2010a] | 169.040 notificações | 6.725 | 100 | 35,0% |
| Lin et al. [2010] | 1.952 notificações | 527 | 110 | 17,2% |
| Segura-Bedmar et al. [2011b] | 3775 sentenças | 3.313 | 3.160 | 52,1% |
| Tari et al. [2010] | 17 milhões de resumos | 579 | 315 | 77,7% |
| Duke et al. [2012] | 817.059 prontuários | 232 | 196 | 62,8% |
| Gottlieb et al. [2012] | 5.039 CRD, 20.452 não CRD e 1.227 DrugBank | 671 | 37.212 | 93,0% |
| Percha et al. [2012] | 354.805 sentenças | 2.910 | 5.000 | 79,3% |
| Zhang et al. [2012a] | 579 documentos biomédicos | 625 | 805 | 63,1% |

CRD: Interações medicamentosas relatadas a citocromos. * Considerou-se 3,8 medicamentos por paciente.

Lin et al. [2010] e Gottlieb et al. [2012] foram os únicos a discutir resultados intermediários ao comparar os falso negativos e verdadeiros negativos, demonstrando uma possível fonte de erro sobre as interações medicamentosas desconhecidas, essencial para guiar novos estudos.

Kinney [1986], Lin et al. [2010], Percha et al. [2012], Segura-Bedmar et al. [2011b] e Gottlieb et al. [2012] implementaram técnicas validadas usando padrão ouro e base de conhecimento independentes, essencial para evitar *overffiting*[7], ou seja, assegura a generalização do modelo a casos desconhecidos. Somente Segura-Bedmar et al. [2011b] e Zhang et al. [2012a] disponibilizaram o código-fonte, o que viabiliza a reprodução dos resultados. Contudo, Gottlieb et al. [2012] disponibilizaram as previsões, possibilitando comparação dos achados.

A precisão variou de 18,5% [Estacio-Moreno et al., 2008] a 93,0% [Gottlieb et al., 2012] baseada no número de casos falso-positivos em um conjunto de interações medicamentosas conhecidas (tabela 3.3).

### 3.2.3 Síntese dos estudos

A partir da análise dos trabalhos elencados, verificou-se que as etapas presentes em todos os modelos foram (I) coleta de dados, (II) seleção de atributos, (III) processamento dos dados,

[7] O *overffiting* ocorre quando os modelos reproduzem os dados ao invés de representá-los, indicando uma pobre estratégia de amostragem ou validação.

(IV) definição do espaço de hipóteses (V) técnica de aprendizado de máquina, (VI) respostas preditivas e (VII) validação.

#### 3.2.3.1 Kinney [1986]

Interações medicamentosas descritas em livros-texto tiveram as probabilidades de ocorrência ponderadas com 1.330 regras. A partir da resposta de algumas questões clínicas relacionadas ao medicamento ingerido, o sistema EXSYS[8] correlaciona as informações e identifica os princípios ativos diretamente ou com auxílio do *soundex*. Diante das interações previstas com base nas regras inseridas, o sistema busca na literatura armazenada sugestão de tratamentos alternativos.

Seis médicos residentes usaram o sistema interativo por um mês em 90 pacientes com média de 3,8 medicamentos (entre 0 e 16). A partir de 27 interações medicamentosas potenciais, 37,0% (n=10) foram confirmadas com a análise dos dados clínicos. Dentre os pacientes com interações medicamentosas, 20,0% não possuíam histórico médico (n=55) e 45,7% (n=35) detinham histórico médico completo. O autor concluiu que a falta de informação clínica afeta o desempenho do algoritmo.

#### 3.2.3.2 Estacio-Moreno et al. [2008]

Foi introduzida a técnica FCA (*Formal Concept Analysis*) a qual consiste em selecionar casos prévios de sinais de interações medicamentosas potenciais, ou seja, captar a correlação entre fármacos e eventos adversos, e síndromes potenciais, onde dois ou mais fármacos ou eventos adversos são relatados conjuntamente.

A técnica FCA constrói uma estrutura hierárquica como uma malha de objetos e atributos dotada de pacientes e características sócio-demográficas, fármacos e eventos adversos, e estabelece a razão da proporção de relatos segundo Evans et al. [2001], qui-quadrado e um limiar escolhido para uma quantidade mínima de casos que devem ser observados para definir a relevância do padrão ou relação.

A partir do critério da *British Medicines and Healthcare products Regulatory Agency*, 3.249 casos de notificação de farmacovigilância foram testados. Identificou-se 527 fármacos, 639 eventos adversos, 110 pares classificados como interações medicamentosas a partir de 593 relações significativas. Concluiu-se que apesar desta técnica requerer uma busca exaustiva, é útil para evitar combinações espúrias.

#### 3.2.3.3 Harpaz et al. [2010a]

Uma abordagem do algoritmo *a priori* foi otimizada e paralelizada constituindo uma implementação adaptável a ampla gama de casos segundo os autores. Esta técnica identifica as regras de combinação para um conjunto de fármacos em relação a um conjunto de efeitos adversos,

[8] *rule-based backward-chaining system*

por considerar a ocorrência das regras diante de um limiar mínimo, o qual reduz o impacto da atribuição do evento a causas aleatórias.

Para mapear os nomes dos fármacos, os pesquisadores usaram MedDRA e MedLEE, um sistema de processamento de linguagem natural, para atrelar o efeito adverso e o fármaco a um código UMLS. Cerca de 24 mil medicamentos foram reduzidos a 6.725 substâncias com o uso do RXNorm para desambiguação. Notificações duplicadas (4.094 dentre 169.040) foram tratadas semi-manualmente quando as notificações apresentaram pelo menos 8 fármacos por evento adverso, minimizando chance de duplicação pelo acaso.

Os pesquisadores encontraram 1.704 e 164 combinações entre fármacos e eventos adversos com 2 e 3 fármacos, respectivamente, dentre aproximadamente 30 mil casos sem o filtro. Os pesquisadores enfatizaram a presença de vários casos espúrios, como fármacos associados com elevada frequência, porém sem interação; ou casos em que um fármaco trata o evento adverso de outro.

As cem combinações ranqueadas com suporte superior a 20 e Risco Relativo de pelo menos 2, conforme Szarfman et al. [2002] e outras cem selecionadas randomicamente, foram manualmente acuradas por especialistas identificando-se 35 interações medicamentosas conhecidas no grupo ranqueado segundo os padrões de referência Micromedex e Epocrates.

#### 3.2.3.4 Lin et al. [2010]

Adotou-se um modelo baseado em grafos para demonstrar a elevada complexidade da relação entre fármacos e biomoléculas alvo, como enzimas ou receptores.

Os autores integraram as bases de dados TTD [Chen et al., 2002], DrugBank [Wishart et al., 2006], DART [Ji et al., 2003] e ADME-AP [Sun et al., 2002] para analizar 1.952 eventos adversos suspeitos na base de notificações espontâneas do FDA. Adotou-se como desfecho primário as taxas de óbito. Uma ontologia de alvos terapêuticos foi elaborada a partir das bases ENZYME [Enz, 2007], GPCRDB [Horn et al., 1998], NRDB [Vroling et al., 2012] e LGIC [Novere & Changeux, 1999], correlacionada aos termos ATC usando entradas UniProt/Swiss-Prot.

A conexão dos fármacos e alvos, possibilitou aos pesquisadores calcular o coeficiente de agrupamento, ponderado entre 0 e 1, o qual proporcionou a avaliação do número de alvos partilhados dos fármacos, os quais caracterizariam a interação. A partir de 198 (10,1%) óbitos, o coeficiente de agrupamento médio e o número de alvos comuns foi praticamente duas vezes maior em relação ao caso de sobreviventes, evidenciando-se a relevância das 19 interações medicamentosas identificadas.

#### 3.2.3.5 Tari et al. [2010]

Foi realizada mineração de texto com técnica de processamento de linguagem natural para extração de interações medicamentosas baseadas em regras lógicas a partir do metabolismo de

fármacos.

A integração do conhecimento biológico, em geral, rotas metabólicas com as restrições estequiométricas e sinalização, ou rotas farmacocinéticas, foi realizada com *parse trees*[9] a partir da ferramenta *Link Grammar*. A ferramenta BANNER foi usada para o reconhecimento dos nomes de genes e proteínas. MetaMap foi usado para os nomes dos fármacos.

O método GNAT foi aplicado para desambiguar cada menção de gene pela identificação dos símbolos oficiais pelo BANNER. Consultas PTQL [Tari et al., 2009] seguidas do uso de regras lógicas construídas em AnsProlog [Gelfond & Lifschitz, 1988, 1991], extraíram 132 resultados explícitos e 5.133 implícitos de aproximadamente 17 milhões de resumos MEDLINE. Dentre eles, 128 corresponderam ao DrugBank [Wishart et al., 2006]. Dentre 315 resultados adicionais escolhidos para a avaliação, 256 estavam corretos (81.3%), sendo 171 (54,3%) relacionados ao citocromo CYP3A4.

#### 3.2.3.6 Segura-Bedmar et al. [2011b]

Foi introduzida uma abordagem de processamento de linguagem natural para a extração em 579 documentos coletados pelo robô *Kapow's free RoboMaker screen-scraper1* e analizados pela ferramenta MMTx[10] da UMLS, a qual realizou diversas etapas de separação das sentenças, *tokenization*, *POS-tagging* e decompôs sintaticamente para a ligação das frases com os conceitos da base UMLS Metathesaurus.

3.775 (65,0% do total) sentenças com duas ou mais menções de fármacos foram manualmente acuradas por um farmacêutico, resultado em 3.160 (10,3%) interações medicamentosas, dentre 30.757 pares de fármacos. Foi realizada classificação com SVM obtendo-se entre 51,0% e 73,8% de desempenho nas métricas adotadas.

#### 3.2.3.7 Duke et al. [2012]

Processamento de linguagem natural foi utilizado para mineração de resumos MEDLINE sobre informações relacionadas ao complexo enzimático do citocromo P450 com consequente inibição ou indução de pares de fármacos *in vitro* e, a partir destas, indutivamente em textos com teor *in vivo*.

A informação extraída foi manualmente acurada por três revisores independentes. A relevância da interação medicamentosa foi avaliada de acordo com a participação de cada enzima relativa ao par de fármacos em "maior", "menor" e "não envolvida". Parâmetros farmacocinéticos como a constante de inibição foram classificados em "forte", "moderado" e "não envolvido".

A partir de 817.059 registros médicos, os pesquisadores realizaram um estudo farmacoepidemiológico em três coortes e avaliaram a presença do substrato e da enzima inibidora em

[9] São árvores utilizadas para decomposição sintática de frases determinando-se os agentes para auxílio a interpretação do texto. A quebra das frases em subestruturas é realizada pela técnica de *tokenization*.

[10] *MetaMap Transfer*

pacientes que apresentaram exposição prévia à medicação e experienciaram um mês de miopatia. Dentre 1,492 fármacos, 232 foram extraídos dos resumos com respectivos substratos e inibidores. A partir da informação *in vitro* 13.197 interações medicamentosas foram previstas, sendo 3.670 prescritas. A partir dos resumos *in vivo*, 196 interações prescritas foram previstas, sendo que 123 (62,7%) mostraram significância clínica.

#### 3.2.3.8 Gottlieb et al. [2012]

A partir de sete medidas de similaridades os autores exploraram o espaço completo de pares de fármacos com (320.182) ou sem (304.769) relação conhecida com citocromos.

(I) A primeira medida de similaridade foi baseada quimicamente no escore bidimensional de Tanimoto, o qual estabelece uma relação da "impressão digital química" de substâncias. (II) A similaridade baseada nos ligantes relaciona propriedades bidimensional de receptores proteicos e fármacos. (III) Os efeitos colaterais foram avaliados usando métodos de mineração em texto. (IV) Os autores utilizaram a base ATC/OMS para estabelecer a similaridade dos fármacos associando suas probabilidades pelo compartilhamento de arestas e ancestrais comuns. (V) O sequenciamento de alvos terapêuticos foi avaliado a partir de escores de alinhamento pré-definidos. (VI) Foi mensurada a distância de cada par de fármacos baseado em uma rede proteína-proteína utilizando-se o caminho mais curto. (VII) Foi calculado o escore de similaridade semântica baseada em três ontologias fornecidas pelo *Gene Ontology*.

As medidas foram combinadas resultando em 49 atributos. Os pares foram relacionados usando média geométrica. Os escores foram calculados para as interações conhecidas e posteriormente as desconhecidas. A partir de validação cruzada, o melhor modelo de regressão logística foi utilizado para realizar as previsões.

Os autores avaliaram a correlação de classes de fármacos quanto a classificação ATC nível 3, cuja combinação não é recomenda pelos achados. Foram avaliadas correlações entre interações medicamentosas e efeitos adversos. Houve um a sobreposição de 39% com os efeitos adversos reportados pelo FDA. A prevalência das previsões foi 19% considerando 9.413 pacientes hospitalizados com uso crônico de dois ou mais medicamentos.

Devido à separação das interações em dois grupos, o modelo sugere o mecanismo da interação medicamentosa como farmacocinético ou farmacodinâmico. Os autores disponibilizaram as previsões em um sítio web.

#### 3.2.3.9 Percha et al. [2012]

A partir de um *corpusBase de análise formada por textos em um determinado idioma* estabelecido por 17,5 milhões de resumos MEDLINE, foi estabelecida uma rede semântica de genes e pares de fármacos conectados. Foi realizada inferência do relacionamento mecanicístico usando processamento de linguagem natural seguida de classificação por *Ramdom Forest*

As dependências foram representadas por grafos, cujas sentenças foram extraídas do *corpus* a partir da ferramenta PharmGKB [Klein et al., 2001]. O léxico abrigou 731 genes farmacológicos conhecidos e 2.910 fármacos distintos ou classes farmacológicas. Os termos normalizados das sentenças corresponderam aos vetores de frequência, constituindo uma matriz em que as linhas são os caminhos mais curtos entre os pares de fármacos calculado com o algoritmo *breadth-first*.

Foi realizada validação cruzada em um conjunto treino com 5.000 interações DrugBank [Wishart et al., 2006] usadas como exemplos positivos e outras 5.000 amostras randômicas de fármacos usados como exemplos negativos. Cada árvore de termos foi classificada como um ponto de treinamento de interagentes de acordo com o número de votos positivos. O fármacos corretamente assinalados foram 79,3% dentre 354.805 sentenças (48,5% com pares de fármacos conhecidos). 36.429 pares de fármacos classificados como positivos não foram identificados no DrugBank, sendo tratados como interações potencialmente desconhecidas.

Segundo os autores, as sentenças com maior escore disponibilizaram informação de elevada confiança sobre o mecanismo fármaco-gene dentre as relações determinadas a partir do léxico extraído.

#### 3.2.3.10 Zhang et al. [2012a]

Este trabalho venceu o desafio para extração de interações medicamentosas a partir do *corpus* de 579 documentos biomédicos [Segura-Bedmar et al., 2011b] com 30.853 pares de fármacos e 3.158 interações conhecidas construídas a partir de processamento de linguagem natural.

O autor construiu uma abordagem *hash* de subgrafos emparelhados em núcleo único. A linguagem foi extraída usando estrutura de dependência e representação em grafos em ordem linear para as sentenças candidatas. A seguir, a operação *hash* computou o valor das identificações hierárquicas de cada nodo, mapeando os grafos em pares de subgrafos no espaço de atributos.

Dentre 7.026 pares de fármacos (10,8% ou 756 interações medicamentosas conhecidas) usados no conjunto de testes, 508 (67,2%) foram corretamente assinalados a interações medicamentosas e 297 foram falsos positivos.

## 3.3 Discussão

Os trabalhos identificados com a revisão sistemática foram avaliados quanto a qualidade experimental relativas e ao contexto de aplicação e a validação.

As fontes de treino/teste e as previsões devem ser explicitamente diferentes para evitar sobreposição, ou seja, acomodação do modelo em relação às instâncias fornecidas com possível perda da capacidade de generalização. O estudo de Kinney [1986] construiu as regras a partir da mesma literatura de validação. Estacio-Moreno et al. [2008] aplicou o domínio do

conhecimento como fonte de validação, mas usou critérios da *British Medicines and Healthcare products Regulatory Agency* para validar as afirmações acerca dos resultados de interações medicamentosas apontados com elevada probabilidade.

O desempenho dos algoritmos de inteligência artificial pode ser avaliado por métricas variadas, dificultando a comparação dos estudos, sobretudo sob diferentes desenhos experimentais [Catal, 2012].

As abordagens de validação como padrão ouro, comparação em tempo real com saída simulada, opinião de especialistas e análise de sensibilidade em relação à valores de entrada e saída são as métricas mais populares de comparação [Sojda, 2007], as quais culminam na verificação dos valores verdadeiros ou falsos usando casos conhecidos positivos e negativos conforme observado nos estudos de Tari et al. [2010], Segura-Bedmar et al. [2011b], Duke et al. [2012], Gottlieb et al. [2012], Percha et al. [2012] e Zhang et al. [2012a]. No entanto, somente Zhang et al. [2012a] e Gottlieb et al. [2012] mostraram análise da área da curva ROC[11] e apenas Tari et al. [2010], Segura-Bedmar et al. [2011b] e Zhang et al. [2012a] mostraram resultados relativos à métrica-F. Estas métricas, juntamente com o cálculo kappa de concordância entre classes, são mais expressivas do que as demais métricas de validação isoladas e corroboram a robustez dos métodos, devendo ser primeira escolha.

Não houve métrica unívoca para análise comparativa do desempenho. Calculou-se a precisão não reportada como artifício de comparação, a partir da razão do número de acertos em relação aos casos avaliados. Embora a comparação direta não seja possível, a precisão apresentou-se como uma forma de mensurar a variabilidade do escopo das abordagens. O DrugBank foi empregado por Percha et al. [2012], Segura-Bedmar et al. [2011b], Tari et al. [2010] e Zhang et al. [2012a] dentre as cinco maiores precisões calculadas, sendo que Gottlieb et al. [2012] também usou Drugs.com. Conjectura-se que o desempenho proporcionado pelo DrugBank deve-se à disponibilidade de um grande número de substâncias que aumenta a quantidade de informações para o treino com consequente ampliação da capacidade de generalização dos modelos. Outro fator para esta predominância é a dificuldade em identificar interações medicamentosas a partir de bases populacionais devido a restrição às combinações observadas, vigorando o trabalho de Duke et al. [2012] com os melhores resultados.

O critérios QUADAS não foi aplicado integralmente devido às diferenças inerentes entre observações clínicas e estudos *in silico*. O período coberto, o cegamento e as perdas clínicas não são aplicáveis aos estudos selecionados nesta revisão. Dentre os estudos que adotaram notificações de farmacovigilância Estacio-Moreno et al. [2008], Harpaz et al. [2010a], Lin et al. [2010], Duke et al. [2012] e Gottlieb et al. [2012] demonstraram a utilidade desta fonte de informação para a previsão de interações medicamentosas, em virtude da avaliação dos padrões das notificações de eventos adversos acompanhados dos fármacos ingeridos. Embora Kinney [1986] tenha usado um número reduzido de pacientes, seus resultados mostram a utilidade

[11] área obtida a partir da probabilidade dos casos falso positivos nas abssissas e a probabilidade dos verdadeiros positivos nas ordenadas

prática em se utilizar inteligência artificial no contexto clínico.

Os trabalhos de Percha et al. [2012], Segura-Bedmar et al. [2011b] e Zhang et al. [2012a] adotaram resumos MEDLINE entre outros documentos científicos. Os trabalhos evidenciaram esta rica fonte de interações medicamentosas potenciais, sobretudo diante da exploração de resultados *in vitro* ou *in vivo*, dada a descoberta do contexto biológico compartilhado com o homem como genes e proteínas. Observou-se importante contribuição dos estudos quando avaliaram a relevância das previsões ao informar a prevalência em pacientes ou populações.

As ferramentas de processamento de linguagem natural foram frequentemente usadas apenas para evidenciar interações conhecidas, motivo de exclusão de muitos trabalhos. Porém, quando usadas como meio para a construção de uma estrutura de dados juntamente com algum algoritmo de aprendizado de máquina, mostrou-se útil para a previsão de interações desconhecidas.

Como verificado por Wong et al. [2010], taxas menos expressivas de previsão de interações medicamentosas podem ocorrer devido a busca não automatizada ou avaliação especializada de interações e da baixa qualidade em delinear os dados, especialmente a partir da literatura acumulada manualmente. Os modelos podem falhar em distinguir os casos espúrios do variado repertório de notificações que não descrevem os fatores e contaminam os casos usados para estimar a incidência ou prevalência [Sim et al., 2001].

O monitoramento de eventos adversos foi desenvolvido em diversos trabalhos excluídos, porém estes estudos não os correlaciona a combinação de fármacos. Outros estudos, embora tenham usado métodos estatísticos avançadas não foram considerados como inteligência artificial por não possibilitarem a descoberta de novo conhecimento a partir da estrutura de dados definida, apenas realizando detecção de padrões explícitos com pressupostos restritivos.

### 3.3.1 Limitações e qualidades da revisão

Outras abordagens, como relatos de bases de conhecimento usadas pelas ferramentas de inteligência artificial para explorar interações medicamentosas, foram deliberadamente excluídas por não serem ferramentas implementadas e validadas. Foram excluídos textos incompletos ou resumos, bem como não realizou-se busca em outras fontes como literatura cinzenta, pois o objetivo do trabalho é identificar modelos academicamente fundamentados de qualidade chancelada por veículos científicos reconhecidos. Não realizou-se exclusão por motivo de idioma.

Demonstrou-se com os 573 artigos encontrados que este número foi apropriado devido à inclusão de diversos sinônimos na busca que abrangeu as principais bases de dados científicas. O uso de dois revisores independentes na seleção dos estudos trouxe confiabilidade aos resultados, cumprindo-se os objetivos pretendidos.

## 3.4 Sumário

A variabilidade dos esforços verificados nesta revisão sistemática demonstrou que a previsão de interações medicamentosas e interação fármaco-biomolécula não é trivial. Apesar de ainda não ter sido evidenciada a eficácia clínica de métodos de auxílio a tomada de decisão [Sim et al., 2001; Hemens et al., 2011], evidenciou-se que a inteligência artificial é uma técnica promissora para a promoção dos cuidados com a saúde [Jaspers et al., 2011].

Recomenda-se aos estudos futuros a exposição da matriz de confusão ou das previsões para possibilitar a comparação dos estudos com metanálise em uma tentativa de mostrar tendência global em prever interações medicamentosas com métodos de aprendizado de máquina. Recomenda-se, ainda, a disponibilização do código fonte para replicação com diferentes estruturas de dados, previsões e bases de conhecimento, demonstrando a robustez do modelo.

A análise sistemática dos trabalhos inaugura a conceituação da previsão computacional de interações medicamentosas por métodos de aprendizado de máquina enquanto disciplina.

# Capítulo 4

# Descoberta de conhecimento em bancos de dados

Neste capítulo são oferecidos aspectos teóricos da abordagem proposta de descoberta de conhecimento acerca de interações medicamentosas, situando o modelo proposto em relação aos demais modelos descritos no capítulo 3. Os aspectos inovadores da aplicação do modelo são enfatizados ao longo dos capítulos 4 e 5.

Os modelos recuperados com a revisão sistemática descrita no capítulo 3 apresentaram soluções para a previsão de interações medicamentosas com ênfase em processamento de linguagem natural, formação de regras lógicas e exploração de dados de notificação espontânea.

Contudo, o universo de fármacos e combinações foi restrito, dado o número de fármacos explorado. Especula-se que a redução do escopo do domínio do conhecimento se deveu à falta de informações farmacológicas diretamente correlacionada a interações medicamentosas ou a capacidade de processamento de algumas implementações que escolheram a solução *a priori*, ou seja, com base em juízos *ad hoc* de como o problema seria resolvido.

O DataMInt foi concebido sob o pressuposto de que a **natureza dos fármacos** e a **relação causal destes com o fenômeno** estudado não podem ser conhecidas em si[1] ou restritas a apenas um ou a um conjunto finito de atributos. Logo, a técnica escolhida para relacionar objeto e fenômeno não pode ser posta *a priori*[2], deve ser submetida à experimentação. No entanto, os métodos tradicionais restringem a capacidade de explicação do fenômeno por partirem de

[1] Embora pretenda-se conhecer o fármaco em si, concebe-se a impossibilidade.Talvez, a melhor característica da ciência é admitir seu caráter de refutabilidade, o que impede determinismos.

[2] A definição *a priori* da solução advém do dedutivismo apregoado desde Descartes, ou seja, devemos racionalizar o problema e colocá-lo na forma de uma hipótese, uma explicação geral que valida a realidade com a experimentação. Partindo-se de uma base empírica (concepção baconiana-humiana), acredita-se que não é possível criar uma hipótese sem que antes haja a experimentação do fenômeno, a sua captura, em última (ou primeira) instância, deve advir dos sentidos. O que se propõe com o DataMInt não é uma solução direta, um modelo de apropriação inicial dos sentidos ou da razão, mas um metamodelo que, através da experiência (dados de entrada), possa gerar simultaneamente um modelo (racionalização, concepção de algo já dado) capaz de apreender o fenômeno em questão de modo a predizê-lo. Inaugura-se uma concepção dedutivista-indutivista ampla, ou holística, como queira; por não se ater apenas ao domínio da hipótese, mas por gerar múltiplas hipóteses, bem acima da capacidade humana em processá-las.

um vocabulário limitado a técnica a qual cerceia o escopo dos dados de entrada. O DataMInt visa abranger um conjunto de técnicas de extração e engenharia que contemple toda forma de caracterizar os objetos envolvidos com fenômeno e o mundo para a formação dos dados de entrada, bem como adotar diversas formas de processamento e análise de modo que cada nuance dos dados possa ser captada em função da determinação prática almejada.

Logo, demanda-se um modelo que seja capaz de abranger a larga quantidade de dados disponíveis de modo a lidar com o universo completo de fármacos conhecidos e combinações decorrentes para prever interações que, em última instância, não possuam informações clínicas diante do uso concomitante. Seja no processo de detecção clínica de interações consolidadas, seja na fomentação de novos estudos, os modelos de previsão devem ser capazes de correlacionar o maior número de informações possível para assegurar a generalidade, contudo, sem perda da coerência com o conhecimento disponível.

A abordagem proposta caracteriza cada fármaco pela extração direta ou indireta de conhecimento a partir da comparação de todos os atributos entre si e entre as instâncias disponíveis do conjunto de fármacos. O modelo objetiva associar as técnicas disponíveis para cada etapa do processo de descoberta de conhecimento em uma estrutura de dados que correlacione conhecimento de fármacos categórico, numérico ou em linguagem natural mediante a transformação e comparação na forma de matrizes (comparação global) e vetores (comparações locais) numéricos ou binários adotando-se diversas métricas de distância.

A modelagem pode envolver o uso de qualquer técnica de classificação, estabelecendo a escolha do processo diante da capacidade de representar com elevado desempenho o conhecimento disponível em um dado padrão ouro. Desta forma, acredita-se que a extrapolação do conhecimento existente indicia a qualidade da previsão de novas interações medicamentosas.

## 4.1 O processo KDD

As etapas do *benchmark* do Processo de Descoberta de Conhecimento em Bancos de Dados são mostradas na figura 4.1. A partir da definição do problema[3] é iniciado o processo KDD para a extração, engenharia, mineração e análise dos dados.

**Extração de dados** é o conjunto de procedimentos de coleta e integração dos dados para os propósitos da mineração e análise. Durante a extração são definidas entidades e atributos relevantes para a tarefa de mineração de dados descrita na seção 4.5, bem como a forma de integração dos dados em um repositório que atenda a demanda do acesso.

**Engenharia de dados** é o conjunto de procedimentos que preparam os dados para a mineração. A primeira etapa é a limpeza dos dados para tornar a base mais próxima da realidade.

[3] A definição do problema não implica em formulação de uma hipótese.

Posteriormente, é realizada a transformação dos dados para torná-los compatíveis com os algoritmos de processamento em termos da natureza e do volume das operações.

**Mineração de dados** é a aplicação de uma ou mais técnicas que extraem informação a partir dos dados. Nesta etapa são adotados, e.g., algoritmos de agrupamento ou classificação.

**Análise de dados** é a verificação dos modelos e das informações extraídas, avaliando-se o desempenho da estratégia de mineração de dados. Nesta etapa são definidos ajustes que possam demandar nova coleta e extração dos dados.

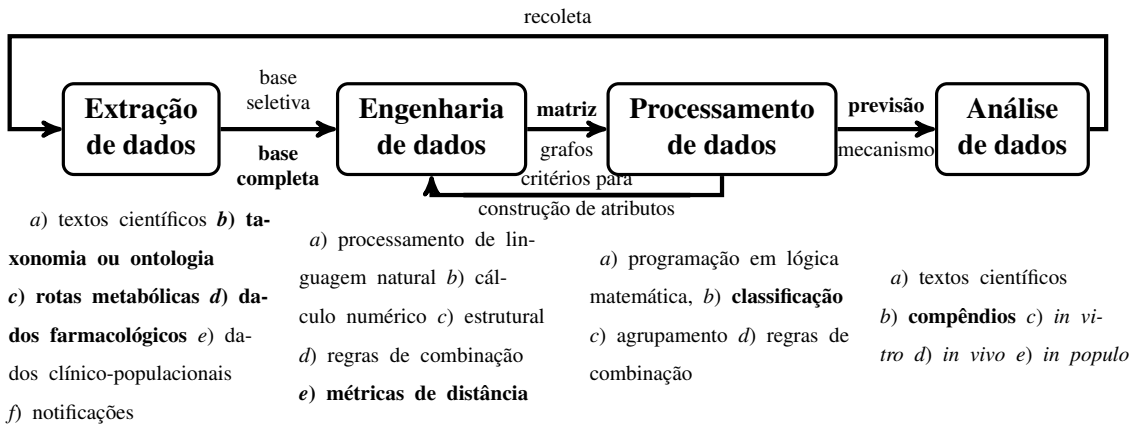



Figura 4.1: **Processos para descoberta de conhecimento em Bancos de Dados.** As etapas percorridas pelo modelo proposto estão em destaque.

O modelo proposto extrai as características latentes que tornam subconjuntos de fármacos, chamados de combinação de fármacos, propensos a interagir entre si, assinalando a existência da interação.

A modelagem para extração dos padrões de interação medicamentosa é realizada a partir de um padrão-ouro de interações conhecidas, o qual define o nível de evidência em saúde pretendido diante dos compromissos entre cobertura e especificidade. Desta forma, pretende-se elaborar um modelo robusto para diferentes contextos da avaliação de interação entre entidades biológicas a partir da descrição dos elementos envolvidos.

O modelo adota técnicas do processo KDD para realizar automaticamente o pré-tratamento dos atributos disponíveis, de modo a viabilizar a extração de caraterísticas preditivas pelos algoritmos de aprendizado de máquina pertinentes à matriz de entrada gerada sem a necessidade de sofisticados modelos farmacodinâmicos ou farmacocinéticos constituídos manualmente por especialistas. A extração ocorre com base na ampla informação disponível.

O tratamento dos dados relaciona características locais de cada fármaco do conjunto às características globais do atributo que o descreve. A matriz de entrada nos algoritmos de aprendizado de máquina é estruturada de modo que as instâncias de subconjuntos de fármacos sejam representadas nas linhas e cada atributo em uma coluna. A abordagem de classificação que

satisfaz os critérios de desempenho conforme validação cruzada com instâncias conhecidas, é posteriormente aplicada no espaço de hipóteses das combinações desconhecidas de fármacos.

## 4.2 Definição do problema

A definição do problema é pré-requisito para a descoberta de conhecimento, sendo a etapa em que os requisitos são definidos a partir da perspectiva do cenário da aplicação.

O problema é definido a partir de quatro elementos sintetizados no acrônimo PICO [Liberati et al., 2009], o qual foi utilizado para a definição da resposta almejada com a realização da revisão sistemática do capítulo 3. O PICO é constituído pelo o objeto do estudo (**p**articipante), a **i**ntervenção, a **c**omparação e o desfecho ou **o**bjetivo. Objeto de estudo é a avaliação das combinações sob o ponto de vista dos fármacos ou dos pacientes. A intervenção é a estratégia adotada para a resolução do problema, ou seja, o conjunto de bases, técnicas e algoritmos adotados. A forma de comparação dos resultados é estabelecida a partir da estratégia e pode adotar técnicas ou critérios bem estabelecidos na literatura, comparação com padrão ouro[4] ou avaliação clínica. Os desfechos ou objetivos do estudo incluem a elucidação de mecanismos, previsão de novas interações ou avaliação do risco de interação medicamentosa em grupos de pacientes.

Conforme relacionado na seção 3.2, a previsão de interações medicamentosas envolve diversos problemas, tais como farmacovigilância, padronização ou taxonomia de fármacos, elaboração de sistema de apoio à decisão, descoberta de alvos terapêuticos, elucidação de mecanismos de fármacos, análise de vocabulário para recuperação de informações e sugestão de novas interações.

Os requisitos para a descoberta de interações medicamentosas devem ser direcionados ao conjunto de fármacos estudados e ao tipo de informação almejada. Desta forma, deve ser definido o nível de especificidade ou generalidade dos dados utilizados e da informação demandada para a resolução dos problemas citados no parágrafo anterior. No nível mais específico, deseja-se avaliar a interação entre apenas dois fármacos, cuja modelagem pode abranger a afinidade por uma molécula ou simulação numérica de equações farmacodinâmicas para definir os níveis de concentração plasmática dos fármacos. Em um nível intermediário de generalidade, interações entre classes farmacológicas ou biomoléculas específicas podem ser requisitadas, como proposto no trabalho de Tari et al. [2010], focado na avaliação de interações medicamentosas via metabolismo de fármacos. O limite superior da generalidade é determinado pela avaliação de uma ampla gama de fármacos e formas de combinações.

A estratégia para a resolução é definida a partir do problema estabelecido. Isto envolve estipular as fontes de conhecimento (i.e., bases de dados farmacológicas e/ou epidemiológicas) bem como as técnicas de mineração de dados pertinentes. Duas fontes de conhecimento dividem as abordagens dada a relação entre fármacos, paciente e evento. A primeira fonte lida com

[4]Interações conhecidas e catalogadas na literatura em quantidade suficiente para assegurar a capacidade de generalização do modelo.

informação de usuários de medicamentos e procura detectar correlações entre os fármacos e eventos clínicos. A segunda fonte é obtida a partir de informações relativas às características bioquímicas e farmacológicas dos fármacos. As técnicas de mineração são determinadas a partir da quantidade de dados (i.e., em volumes grandes de dados um algoritmo pode ser mais eficiente que outro) e da informação pretendida, seja em uma análise exploratória, quando não se conhece a natureza dos dados, ou em uma tarefa específica como a determinação da combinação de fármacos enquanto interação ou combinação inerte.

O problema defrontado pelo presente trabalho é gerar um metamodelo que combina ferramentas bem estabelecidas no KDD em bancos de dados de informações farmacológicas (intervenção) capaz de prever interações medicamentosas (desfecho) a partir de um amplo espectro de fármacos (participante) e combinações, sendo a validação das previsões realizada com padrão ouro de interações catalogadas em grande número (comparação).

## 4.3 Extração de dados

A extração de dados determina quais entidades e atributos constituirão a estrutura de dados para a mineração das informações pretendidas. Os atributos abordados podem ser numéricos ou categóricos ou expressos em linguagem natural.

Atributos numéricos são quantidades que expressam medições físicas ou escalas numéricas.

Os atributos categóricos são “nomes” ou outras identificações que permitem operações que avaliam a presença ou a ausência da característica ou propriedade ou, ainda, intensidade ou ordem de precedência. Ainda, existem dados categóricos estruturados em que as entidades são interligadas por representações que definem comportamento, fenômeno ou sequência de eventos, i.e., mecanismos metabólicos em que fármacos são relacionados a enzimas e seus produtos de degradação.

Os atributos expressos em linguagem natural são textos, cujas características foram abordadas nas seções 2.5.6 e B.3.3.1.

A partir da definição da estratégia para a resolução do problema são estabelecidos os domínios que contém os dados a serem coletados e as entidades participantes da estrutura dos dados a ser constituída.

### 4.3.1 Definição do domínio do conhecimento

O domínio do conhecimento pode ser constituído a partir de fontes primárias, secundárias ou terciárias e abriga a área do conhecimento que se pretende explorar conforme o problema elaborado. O domínio do conhecimento é definido a partir da combinação de características exploradas por várias disciplinas e variações como química, bioquímica, farmacologia, farmacoepidemiologia ou farmacotécnica.

As fontes primárias advém da literatura com ampla aceitação no meio científico como revisões sistemáticas, ensaios clínicos randomizados ou estudos de coorte (e.g.: MEDLINE, EMBASE, LILACS)[5].

As bases secundárias são manualmente compiladas a partir da primeira fonte de dados, porém na forma de livros-texto (e.g., o livro de Tatro [2012]) ou compêndios (e.g.: micromedex, martindale, drugs.com, DrugBank, ATC/OMS ), permitindo a recuperação sistemática das informações.

As fontes terciárias são obtidas a partir da interrelação entre as entidades presentes nos bancos secundários, sendo estruturadas na forma de matrizes, grafos, bancos de dados hierárquicos ou bancos de dados relacionais tornando-as, em geral, utilizáveis apenas por algoritmos (e.g., KEGG). linhas em

O presente trabalho possibilita a exploração das fontes secundárias e terciárias, em que são extraídas informações acerca de fármacos advindos do DrugBank, KEEG e ATC/OMS e informações contidas no banco ExPASy e Enzyme sobre as enzimas associadas. O padrão ouro foi coletado no sítio drugs.com.

#### 4.3.1.1 Dados farmacológicos

Entende-se como dado farmacológico àquele relativo ao fármaco. Desta forma, são abrangidas as disciplinas que estudam a atuação farmacotécnica, fisiológica ou relativa ao uso dos fármacos e medicamentos. Pretende-se utilizar todo dado farmacológico disponível na forma de texto, caracterização química ou técnica.

**Descrição textual.** Os dados farmacológicos são usualmente descritos na forma de textos curtos com dezenas ou centenas de palavras, com ou sem números. Esta forma descritiva contém informação a ser consultada por profissionais de saúde na tomada de decisão quanto a interações farmacocinéticas (absorção, distribuição, metabolismo e excreção) ou farmacodinâmica (indicação, modo de ação, efeitos adversos, contra-indicação). Porém, algumas características numéricas são disponibilizadas isoladamente como tempo de meia vida, fração da ligação à proteínas plasmáticas ou massa molecular.

**Caracterização química.** O fármaco, na acepção mais básica, é uma substância química. Este contexto exclui uma pequena parcela de medicamentos à base de seres vivos sem isolamento da substância ativa. Enquanto substância química, os fármacos podem ser descritos quanto a massa molecular, solubilidade, índice de acidez, coeficiente de partição (fração que permanece em uma mistura de solventes aquosos e oleosos), ponto isoelétrico, ponto de fusão, ponto de ebulição, área de superfície polar, refratividade, isomeria, massa de hidratação, rota de síntese, produtos de degradação, entre outros.

[5] No apêndice encontram-se maiores explicações, sobretudo na seção A.4.

As características moleculares influenciam na escolha dos excipientes e da forma farmacêutica do medicamento que conterá o fármaco. Desintegração, dissolução, estabilidade, posologia, volume de distribuição, biodisponibilidade e aspectos organolépticos são fatores influenciados pela constituição molecular.

**Caracterização técnica.** As definições técnicas são estabelecidas para o controle de qualidade e influenciam na apresentação final do medicamento. O repasse da informação aos profissionais de saúde e ao consumidor (por vias acadêmicas ou estratégia de marketing), bem como o preço, forma farmacêutica e posologia podem influenciar na quantidade ou na qualidade da utilização do medicamento propiciando ou não ambiente para o uso concomitante e possíveis interações, bem como constituem características latentes que podem agrupar fármacos por similaridades.

Os fatores mencionados podem estar relacionados diretamente a interações medicamentosas ou indiretamente, afetando características do fármaco que venham a propiciar o cenário para sua interação. Logo, não são assumidos *a priori* quais atributos devem contribuir para a previsão de interações entre fármacos.

#### 4.3.1.2 Dados taxonômicos ou ontológicos

Ontologia é a representação das relações dos fármacos, biomoléculas e fenômenos em hierarquias com informações contextualizadas em níveis anatômicos, terapêuticos ou químicos. A taxonomia se atém a relação hierárquica com definições implícitas acerca das relações.

Os fármacos são distinguidos quanto ao nível anatômico (i.e., respiratório, cardiovascular, gastrointestinal), organismo afetado (i.e., humanos, bactérias ou fungos), terapêutico (i.e., anti-ácido, anti-inflamatório, vitamínico, anti-arrítmico) ou ação/família química (benzodiazepínico, aminoglicosídico, inibidor da monoaminoxidase, betabloqueador).

#### 4.3.1.3 Mecanismos farmacológicos

O mecanismo é uma estrutura de dados que contém informações de encadeamento de elementos em que uma etapa gera algum substrato ou causa algum efeito na etapa seguinte. De modo geral, expressam relações de entidades como fármacos, eventos e/ou biomoléculas como enzimas.

Os mecanismos farmacocinéticos são os mais explorados para a detecção laboratorial e *in silico* de interações medicamentosas por serem os mais intuitivos, dado que dois fármacos metabolizados pela mesma enzima possuem grande chance de competir e ter a concentração plasmática aumentada de pelo menos um deles. Proteínas transportadoras plasmáticas sofrem competição de boa parte dos fármacos, modificando a concentração dispersa no sangue. Os mecanismos farmacocinéticos mais abordados envolvem o complexo de enzimas que metabolizam os fármacos com destaque para o sistema de citocromos hepáticos.

Os mecanismos farmacodinâmicos relacionam sequências de manifestações químicas que culminam em efeitos clínicos. Por exemplo, a levodopa é convertida a dopamina, a qual aciona os receptores dopaminérgicos e mitiga os sintomas da doença de Parkinson.

### 4.3.2 Identificação do objeto farmacológico de estudo

Entende-se como objeto farmacológico o agente que modifica funções fisiológicas no organismo humano cujos efeitos são avaliados em função da combinação. A identificação deste agente pode ser realizada em sua forma química completa ou a partir da porção com atividade farmacológica. Salienta-se que o objetivo deste trabalho não é avaliar apenas fármacos individualmente, contudo, deve-se conhecer o fármaco para identificar a mesma entidade ao longo das fontes de dados adotadas. Em último caso, a identificação pode ser realizada pelo nome do fármaco ou nome químico, porém, com grande possibilidade de perda devido à divergência de nomenclatura. Devido a essa dificuldade, a verificação manual por especialistas tornou-se uma prática comum em diversos estudos. Exemplos de identificação são mostrados da tabela 4.1.

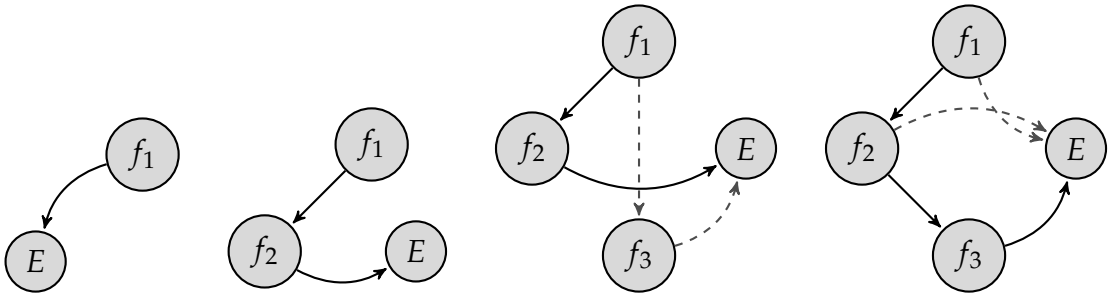
Tabela 4.1: **Exemplos de identificação do fármaco diclofenaco.**

| **Fonte** | **Códigos** |
|---|---|
| CAS | 15307-86-5 |
| ATC | D11AX18, M01AB05, M02AA15 ou S01BC03 |
| PubChem | CID 3033 |
| IUPHAR ligand | 2714 |
| DrugBank | DB00586 |
| ChemSpider | 2925 |
| UNII | 144O8QL0L1 |
| KEGG | D07816 |
| ChEBI | CHEBI:47381 |

O estudo deve escolher a base de identificação que agregar maior caráter informativo à combinação de fármacos. Caso a base de identificação não contemplar o conjunto de fármacos estudados, uma estratégia para evitar perdas é estabelecer um identificador próprio auto-soma e criar uma tabela relacional $n$ para $n$ constando o índice criado, a base vinculada e o respectivo identificador. Deve-se estabelecer claramente o critério para a identificação (e.g., nome químico, atividade terapêutica, nome genérico ou porção ativa da molécula) para estabelecer o mérito das combinações.

Uma alternativa à construção de uma tabela relacional é a escolha do índice contido na base que contempla o maior número de instâncias do objeto de estudo.

Uma confusão a ser evitada é a verificação da interação entre fármacos ou medicamentos. O estudo pode definir que um medicamento com vários ativos pode interagir com outro, independente se apenas um destes ativos for responsável pela interação com o outro medicamento.




(a) $n = 1$ (b) $n = 2, k = 2$ (c) $n = 3, k = 2$ (d) $n = 3, k = 3$

Figura 4.2: **Mecanismos de eventos da interação de objetos.** Em $n = 1$ o evento $e$ pode ser associado diretamente ao objeto (figura a). A elucidação do mecanismo que estabelece a relação objeto-evento deve considerar cada possibilidade de combinação. Assim, na figura (b), tanto $(f_1, f_2) \rightarrow e$, como $(f_2, f_1) \rightarrow e$ são explicações do universo de hipóteses para o evento $e$. Na figura (d) $f_1$ interage com $f_2$ o qual interage com $f_3$ gerando o evento $e$. As linhas tracejadas na figura (c) demonstra rota alternativa partindo de $f_1$, quando o tamanho das associações estudadas $k$ é igual a 2. Não constitui objeto de estudo de interações os casos em que um (figura a) ou mais (figura d, tracejado) objetos são vinculados a um evento sem interagir entre si.

Embora este ponto de vista seja válido se forem consideradas as vias de administração, formas farmacêuticas ou combinações farmacológicas, a discriminação pelas substâncias ativas torna o entendimento da interação mais intuitivo e direto, mesmo quando mais de duas substâncias estão envolvidas ao mesmo tempo na interação, conforme ilustrado na figura 4.2. Nesta figura são mostradas formas de atrelar os fármacos ao evento estudado. O caso (c) pode ser diferenciado do caso (d) se o fármaco $f_3$ não estiver envolvido no mecanismo da interação, tornando um erro atrelar o evento aos três fármacos simultaneamente.

No presente estudo, adotou-se o drugcard oferecido no DrugBank, visto que esta base apresenta o maior número de substâncias farmacologicamente ativas, sendo a principal fonte para a engenharia de dados utilizada no modelo preditivo mostrado no capítulo 5.

## 4.4 Engenharia de dados

A etapa de engenharia consiste na limpeza e transformação dos dados no formato de entrada para o processamento. Uma vez coletados os dados, é realizada identificação dos fármacos e o tratamento para geração do formato de entrada adequado à técnica de aprendizado de máquina escolhida.

A relação entre os atributos é frequentemente definida manualmente por especialistas. No entanto, o presente modelo trata minimamente cada atributo, deixando para a técnica de aprendizado de máquina o papel de selecioná-los conforme a respectiva contribuição preditiva. Desta forma, embora muitos atributos possam ser concatenados ou agrupados segundo uma avaliação especializada, o modelo os trata separadamente para não haver perda do escopo individual pre-

servando as características da informação diante da fonte adotada. Acredita-se que um escopo melhor definido amplie a capacidade discriminativa do modelo e contribua para a coerência das previsões.

A seguir são descritas as etapas da transformação dos dados. Inicialmente cada atributo acerca de fármacos[6] é decomposto em matrizes binárias, tratando-se cada fármaco como um ponto, representado nas linhas, disposto em um espaço n-dimensional. Nesta etapa, pode ser realizada tentativa de remoção de ruídos como a aplicação da decomposição por valores singulares, SVD. Em seguida, para cada atributo, o conjunto de combinações é formado a partir da tomada de distância entre todos os pontos, constituindo um vetor de distâncias com cardinalidade igual ao número de combinações. Os vetores são concatenados de modo que cada linha da matriz resultante represente uma combinação (par de fármacos) e cada coluna um atributo de combinação. Esta matriz é a entrada para a mineração de dados. Para cada tratamento ou tomada de distância é gerado um novo atributo de combinação.

### 4.4.1 Limpeza dos dados

A limpeza agrega qualidade aos dados ao torná-los mais fidedignos à realidade que expressam. Dados de baixa qualidade são usualmente consequência do processo de coleta, na ocorrência de armazenamento de valores espúrios. Contudo, a falta ou incompletude de dados farmacológicos é uma característica intrínseca aos bancos de dados, sobretudo quando contemplam novas tecnologias farmacêuticas, dado que o conhecimento acumulado é frequentemente proporcional ao tempo de lançamento do fármaco. A limpeza dos dados trata três problemas. O primeiro é a duplicação de dados, o segundo é a ocorrência de valores faltantes e o terceiro é a ocorrência de ruído.

A deduplicação ocorre quando dois ou mais registros estão presentes para a mesma entidade. Este fato é decorrente de uma identificação incorreta de fármacos em que a mesma entidade é representada mais de uma vez no banco de dados. Existem ferramentas de relacionamento de registros que avaliam probabilisticamente conjuntos de instâncias com características aproximadamente comuns, de modo a agrupar elementos que atendam a um limiar de similaridade. A aplicação desta técnica é comum para o tratamento de bases populacionais. No presente estudo, a estratégia adotada para evitar fármacos duplicados foi definir um critério para identificação unívoca descrito na seção 5.2.2.

A ocorrência de valores faltantes ou perdidos, ou seja, valores não fornecidos, é geralmente consequência do aspecto dinâmico da coleta dos dados. A melhor estratégia para lidar com este problema é a recoleta periódica (automatizada com um "robô" ou acionada manualmente) objetivando completar (e atualizar) os dados para cada fármaco. Outra estratégia adotada foi a abordagem de diferentes fontes de dados. Desta forma, a combinação das bases ATC/OMS, DrugBank e KEGG possibilita ampla descrição, mitigando o impacto de ausências

[6] Seção 4.3.1.1.

locais. Uma terceira via é estimar os valores faltantes com avaliação comparativa da distribuição dos atributos e das instâncias. Embora nesta abordagem não tenha sido realizada diretamente a estimação dos valores faltantes, a tomada de cada fármaco como um vetor de características permite que cada combinação, também um ponto no espaço n-dimensional, tenha o impacto de faltantes diluído na avaliação entre combinações de fármacos, possibilitando o estabelecimento de relações comutativas de entidades semelhantes.

O ruído é um problema dos dados de entrada e, frequentemente, representa uma contaminação sem diferenças salientes em relação aos dados verdadeiros. O tratamento é o mesmo para exceções, visto que estas são tratadas a partir da avaliação do conjunto dos dados. Devido ao tratamento dos dados na forma numérica, o presente modelo adotou a técnica de Decomposição em Valores Singulares, um tratamento matemático que avalia as características latentes das instâncias em relação ao conjunto de dados. Esta técnica é explicada com maiores detalhes neste capítulo após a definição de como as matrizes são geradas (seção 4.4.5).

### 4.4.2 Transformação dos dados farmacológicos

A transformação dos dados farmacológicos é a transposição ao formato de entrada dos algoritmos de processamento.

O formato de entrada dos algoritmos de aprendizado de máquina que processarão e formarão os modelos preditivos variam. Podem ocorrer diretamente na forma de texto; representação da ligação entre fármacos, moléculas e eventos (i.e., listas ou matrizes de adjacência) ou numérica. Porém, nesta abordagem, adotou-se apenas o formato de matriz matrizes binárias ou de frequência.

Cada atributo é reduzido a uma matriz numérica binária ou inteira, em que os fármacos estão dispostos nas linhas e os termos do atributo nas colunas. Os termos são extraídos distintamente para cada tipo de atributo, seja numérico, texto ou taxonômico/mecanicístico.

#### 4.4.2.1 Atributo numérico

Um atributo numérico é expresso na forma de número inteiro ou decimal obtido a partir de uma aferição ou escala comparativa.

Durante o tratamento, os atributos numéricos, e.g., biodisponibilidade, massa molecular e hidrofobicidade, são submetidos a diversas operações matemáticas previamente escolhidas como inversão, logaritmação, exponenciação, radiciação, cálculo do cosseno ou ponderação (distribuição entre os valores 0 e 1)[7]. Cada transformação gera um novo atributo. Em outras palavras, se foram usadas cinco operações, o mesmo número de novas colunas de atributos são adicionadas na matriz de combinação. Um tratamento adicional é segmentar cada coluna em

[7]Exemplos em notação matemática: $log(x)$, $log_2(x)$, $ln(x)$, $cos(x)$, $x^2$, $\sqrt{x}$, $x^{-1}$ e $|x|$

um número pré-determinado de categorias (e.g., igual ao número de classes do padrão-ouro[8] ou em uma segregação que favoreça essa determinação) formando-se um vetor binário com uma posição para cada categoria. O valor indicado para o fármaco é alocado na respectiva posição do vetor com o valor 1. Por exemplo, três fármacos contendo $d_1 = 5$, $d_2 = 6$ e $d_3 = 11$, em um atributo segmentado em $[0;5[$, $[5;10[$, $[10;15[$ e $[15;20[$, são descritos como vetores $\overrightarrow{m_1} = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix}$, $\overrightarrow{m_2} = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix}$ e $\overrightarrow{m_3} = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}$.

#### 4.4.2.2 Atributo em formato categórico e em texto

O atributo em formato categórico é aquele que contém um ou mais valores (termos) discretos que indicam presença ou ausência de características, podendo ou não ser ordinais (i.e., posologia "manhã", "tarde", "noite"), expressar intensidade (i.e., interação "menor", "moderada", "maior" ou "rápido" e "tardio") ou estímulo (i.e., ação de "agonismo" ou "antagonismo"). Termo é definido como sequência de caracteres iniciada e/ou terminada por um caractere delimitador (i.e., espaço), sem que o delimitador seja incluído. O dicionário de termos é formado pelo conjunto de termos distintos contidos no atributo.

Atributos como "mecanismo de ação" ou "metabolismo" ocorrem na forma de texto, ou seja, sequência de termos que intuem uma frase, e são convertidos a matrizes de termos conforme explicado a seguir.

A formação da matriz de termos envolve a geração do dicionário de termos e alocação da posição ou frequência análoga ao tratamento dos atributos numéricos, entretanto, mais de uma posição no vetor pode ser modificada diante da presença de vários termos para o mesmo fármaco. As descrições $1 : n$ dos fármacos são concatenadas em um bloco de texto. Ocorre a redução da caixa para letras minúsculas e remoção de caracteres não alfanuméricos[9] (exceto espaço)

O dicionário é elaborado a partir da coleção distinta e ordenada dos termos. Um exemplo hipotético, sem maiores tratamentos, é dado a seguir.

> "Não forces o poema a desprender-se do limbo.
> Não colhas no chão o poema que se perdeu.
> Não aludes o poema. Aceita-o,
> como ele aceitará sua forma definitiva e concentrada no espaço."
>
> O dicionário formado a partir do texto será
>
> dicionário=[a; aceitao; aceitará; aludes; chão; colhas; como; concentrada; definitiva; desprenderse; do; e; ele; espaço; forces; forma; limbo; não; no; o; perdeu; poema; que; se; sua].

[8] O padrão ouro na forma de classes corresponde à resposta preditiva almejada para combinação de fármacos desconhecidas. Mais informações na seção 4.5.1.

[9] 0 a 9, *a* a *z*

Supondo cada linha do texto acima como pertencente aos fármacos $f_1$ a $f_4$, respectivamente, a matriz $M^{m\times n}$ correspondente ao atributo, terá cada posição formada formada por

$$m_{ij} = \begin{cases} 1 & \text{se } (i,j) \in \text{dicionário} \\ 0 & \text{em caso contrário.} \end{cases}$$

Adotando-se esta regra, cada linha constitui um vetor de termos correspondente a um fármaco: $\vec{m}_1 = [100000000110001011010100 0]$, $\vec{m}_2 = [000011000000000000011111110]$, $\vec{m}_3 = [000100000000000000010101000]$ e $\vec{m}_4 = [001000111001110100100000 1]$.

As palavras muito ou pouco frequentes, neste caso, com frequência igual a 1 ou 4, poderão ser removidas do dicionário por não diferenciarem as instâncias, restando [não; no; o; poema]. A matriz resultante será:

$$\text{atributo} = \begin{array}{c} \\ f_1 \\ f_2 \\ f_3 \\ f_4 \end{array} \begin{array}{c} \begin{array}{cccc} \text{não} & \text{no} & \text{o} & \text{poema} \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

Logo, cada coluna $j$ referente ao termo pode ser selecionada se contemplar um limiar de expressividade $l \leq \sum_{i=1}^{m} m_{ij} \leq 1 - l$ que permita a distinção dos dados. Neste caso, $l$ deve situar-se no intervalo $]0;0,5[$, sendo comum valores de 1% a 20%.

Adicionalmente, outras técnicas de mineração de texto podem ser usadas. Destacam-se a extração de radicais linguísticos chamada *stemming* e a remoção de palavras comuns de baixa expressividade para os modelos, em geral, alcunhadas *stop words*. Consequentemente, palavras como "enzima" e "enzimático" são consideradas como uma. Artigos, preposições ou pronomes usualmente são eliminados por constarem na lista de *stop words*.

A primeira frase abaixo não foi submetida ao processo de remoção de palavras comuns e redução ao tronco linguístico, ao contrário da segunda:

1. "*following the 5 mg once daily dose the median time to maximum concentration is 2 hours*"

2. "*follow 5 mg daili dose median time maximum concentr 2 hour*"

#### 4.4.2.3 Atributo taxonômico, ontológico ou mecanicístico

Estes atributos são tratados como categóricos, conforme introduzido na seção 4.3.1.2. Taxonomias ou ontologias são classificações hierárquicas que definem o fármaco e papéis biológicos em níveis com significação estabelecida, tais como compartimentos biológicos, atividade química, atividade farmacológica (i.e., agonismo, antagonismo). Uma taxonomia bastante utilizada no contexto de medicamentos é a classificação ATC/OMS exemplificada na tabela B.1, a

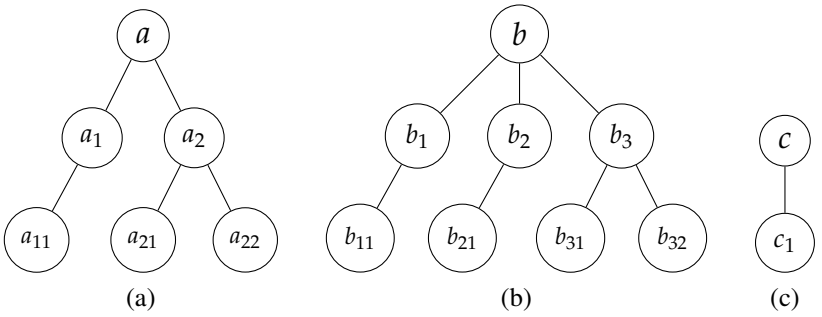



Figura 4.3: **Exemplo de classificaÃ§Ã£o hierÃ¡rquica**

qual distingue o fármaco em cinco níveis. Outras taxonomias incluem a classificação EC para enzimas ou *Gene Ontology* para modelos biológicos em geral.

Da mesma forma que são tratados os demais atributos categóricos, são formados vetores de frequência para os dados taxonômicos. Porém, cada nível é tratado como um atributo distinto. Quando não há níveis definidos, como em rotas metabólicas ou mecanismos de ação farmacodinâmica, o dado é tratado na forma de bloco único, gerando apenas um atributo.

Exemplificando, seja a hierarquia $D$ mostrada na figura 4.3 contendo três níveis, considerando, ainda, cinco fármacos $f_1$ a $f_5$ classificados por descritores "d", cujos índices correspondem respectivamente a um descritor de cada fármaco $d_1 = \{a_{11}, b_{11}\}$, $d_2 = \{a_{21}, b_{31}, c_1\}$, $d_3 = \{a_{22}, b_{21}\}$, $d_4 = \{b_{11}\}$ e $d_5 = \{b_{11}\}$[10]. Cada nível pode expressar um atributo conforme ilustrado a seguir.

$$\text{classificação} = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{array} \begin{array}{c} \text{nível 1} \\ \left[\begin{array}{cc} a_{11} & b_{11} \\ a_{21} & b_{31} \\ a_{22} & b_{21} \\ b_{11} & \\ b_{11} & \end{array}\right] \end{array} \begin{array}{c} \text{nível 2} \\ \left[\begin{array}{ccc} a_1 & b_1 & \\ a_2 & b_3 & c_1 \\ a_2 & b_2 & \\ b_1 & & \\ b_1 & & \end{array}\right] \end{array} \begin{array}{c} \text{nível 3} \\ \left[\begin{array}{ccc} a & b & \\ a & b & c \\ a & b & \\ b & & \\ b & & \end{array}\right] \end{array}$$

Os termos $a_{ij}$, $a_i$ e $a$, relativos a cada nível, formam o dicionário, bastando indicar a presença da classificação com o dígito 1 para cada fármaco. A seguir são mostradas as matrizes correspondentes aos atributos $M_1$, $M_2$ e $M_3$ advindos da classificação.

[10] Aqui é diferenciado $f$ e $d$ pois o fármaco, em si não é tomado por sua descrição, analogamente, "y" expressa informação diferente do que "f(x)" na expressão $y = f(x)$.

$$M_1 = \begin{array}{c} \\ m_1 \\ m_2 \\ m_3 \\ m_4 \\ m_5 \end{array} \begin{array}{c} \begin{array}{ccccccc} a_{11} & a_{21} & a_{22} & b_{11} & b_{21} & b_{31} & b_{32} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

$$M_2 = \begin{array}{c} \\ m_1 \\ m_2 \\ m_3 \\ m_4 \\ m_5 \end{array} \begin{array}{c} \begin{array}{cccccc} a_1 & a_2 & b_1 & b_2 & b_3 & c_1 \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array}, \quad M_3 = \begin{array}{c} \\ m_1 \\ m_2 \\ m_3 \\ m_4 \\ m_5 \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix} \end{array}$$

### 4.4.3 Espaço de hipóteses

O espaço de hipóteses é o conjunto de combinações de fármacos cuja interação deseja-se classificar, chamado conjunto consulta $Q$ (*query*). Esta definição pode ocorrer *a priori* ou *a posteriori*. Os estudos que definem *a priori* são aqueles que possuem um domínio de fármacos definido, bem como a descrição de cada elemento deste conjunto, e são baseados em bancos de dados farmacológicos. Os *a posteriori*, caso do presente estudo, definem o número de fármacos e combinações decorrentes conforme a base é avaliada, sendo comum na exploração de textos científicos e populações de usuários de medicamentos.

O espaço de hipóteses é influenciado pelo processo de Descoberta de Conhecimento em Banco de Dados aplicado. Os modelos identificados na revisão sistemática (capítulo 3) extraem variáveis diretamente relacionadas ao contexto de fármacos e, por isso, são chamados de indutivos, ou seja, partem de conhecimentos específicos para a posterior generalização das respostas. O modelo proposto realiza a extração de dados a partir de informações difusas, ou seja, que podem não estar atreladas diretamente ao contexto de interações medicamentosas, objetivando caracterizar de modo amplo cada instância de fármaco, para, posteriormente, constituir os atributos específicos da combinação. Esta linha de pensamento, que parte de um conceito amplo para uma conclusão específica, é chamada de método dedutivo[11].

[11]Com esse texto pode-se concluir que propõe-se duas formas de avaliar o método científico fugindo das tradicionais formas de ver (*a priori*-dedutiva e *a posteriori*-indutiva. O sistema indicado aqui de forma controversa coloca que sem a experiência se estabelecem métodos *a priori* indutivos no escopo de aplicação, ou seja, pequenas formulações que tendem a se ampliar. É uma interpretação da falha do método cartesiano, pois é de bom tom na ciência contemporânea não fazer generalizações amplas demais. Esse cuidado no discurso não foi suficiente para que a ciência, em quase sua totalidade, fosse praticada de forma fragmentária, com pequenas perguntas que não respondem as atividades práticas, que se tornam úteis apenas quando unidas discursivamente de forma subjetiva. O método proposto é uma abordagem "dedutiva" e *a posteriori* por não tecer uma hipótese inicial e explorar todo

O entendimento de ambos os métodos é complementar, por esta razão são explicados abaixo.

#### 4.4.3.1 Método indutivo

O método indutivo limita o espaço de hipóteses ao contexto das variáveis assumidas como relacionadas à previsão de interações medicamentosas.

Estacio-Moreno et al. [2008], Harpaz et al. [2010b] e Duke et al. [2012] avaliaram regras de combinação em bases populacionais e extraíram correlações em casos reais de polifarmácia e eventos de saúde (i.e., morbidades, parâmetros clínicos).

Lin et al. [2010] e Gottlieb et al. [2012] elaboraram redes de biomoléculas e fármacos, limitando o conjunto de fármacos às combinações existentes.

Os estudos de mineração de textos científicos realizados por Tari et al. [2010], Segura-Bedmar et al. [2011a] e Percha et al. [2012] também estão limitados às combinações descritas na literatura consultada, bem como aos fármacos identificados pelos indexadores adotados e construíram modelos focados em relações de biomoléculas.

Esta abordagem é caracterizada pela maior proximidade com a realidade consolidada cientificamente. Porém, a restrição do contexto aos atributos assumidos *a priori* como relacionados à interação medicamentosa e, consequentemente, do número de fármacos cobertos pelo conhecimento disponível; torna limitada a capacidade de descoberta de novo conhecimento.

#### 4.4.3.2 Método dedutivo

O método proposto estabelece uma estrutura global de fármacos e combinações de fármacos para definir a propriedade local da interação entre cada fármaco. Esta estrutura é tão mais completa quanto mais informações acerca do maior número de instâncias de fármacos for fornecida. O modelo não avalia a interação entre fármacos em si, ou atributos de interações, mas gera novos atributos de combinações baseados nas características dos fármacos posicionadas frente aos demais.

A essência do modelo é tornar possível a avaliação do universo de fármacos e combinações de fármacos com o maior número de informações e instâncias. As razões para isto são apresentadas a seguir.

O fármaco é representado por um conjunto de atributos. Ao mesmo tempo, o escopo do atributo é delimitado pelo número de fármacos descritos. A capacidade preditiva que o atributo pode oferecer ao modelo em conjunto com demais atributos depende da sua capacidade de expressar a informação. Ou seja, somente há extração de semântica quando um número suficiente de observações refletem a amplitude do conceito do atributo.

---

o universo de informações para, por intermédio de centenas de "hipóteses" (talvez seja tido como um método "ostensivo"), chamadas aqui de "modelos", responder a um conjunto de milhares de "perguntas" do tipo "se um fármaco $x$ pode interagir com um fármaco $y$". Certamente esta confusão deve desenvolvida com maior rigor para tratar como a abordagem proposta se enquadra na velha discussão Descartes/Kant *vs* Bacon/Hume.

Por exemplo, o escopo do atributo “massa” não será definido em toda sua amplitude ao tomarmos apenas os fármacos elementares como lítio, potássio, magnésio, ferro, zinco. Este subconjunto contempla valores de uma a duas ordens de grandeza, os quais não correspondem a maioria dos fármacos e não representa a amplitude do atributo. Logo, as massas moleculares dos demais fármacos devem ser incluídas para não ocorrer a caracterização do atributo de forma enviesada.

Os atributos das combinações de fármacos são obtidos com o processamento dos atributos dos fármacos. Da mesma forma que a caracterização de um dado atributo demanda o maior número de instâncias possível para assegurar sua representatividade e capacidade de generalização, quanto mais completo for o conjunto de fármacos e atributos de fármacos, melhor será a representação das combinações e atributos de combinações, visto que os atributos criados para caracterizar as combinações remetem a atributos dos fármacos.

Cada atributo de fármaco melhor representa o conjunto quanto maior for o número de instâncias. Por extensão, os atributos de combinações derivados dos atributos de fármacos são melhor representados quanto maior for o número de combinações envolvidas, em outras palavras, quanto maior for o espaço de hipóteses. Tratando-se a combinação como subconjuntos de fármacos, o cenário ideal é construído com a comparação de $n$ fármacos do mesmo domínio, $k$ a $k$, sem repetição. Usualmente o valor de $k$ é igual a 2, sendo o abordado neste texto. Porém, outros domínios devem ser explorados, dado que o consumo médio de medicamentos em determinadas populações pode chegar a 5, sendo frequente o uso de dez substâncias ativas ou mais.

Evidentemente, o valor de $k = 2$ representa a redução ao padrão mínimo de combinação. Logo, para um paciente que estiver associando dez fármacos, é válido reduzir a combinação ao universo de $\binom{10}{2} = 45$ pares de combinações possíveis diante da inviabilidade em avaliar todas as possibilidades $\sum_{i=2}^{n=10} \binom{10}{i}$. Esta informação corrobora a necessidade em explorar o conjunto completo de combinações de fármacos par-a-par antes de estabelecer domínios com valores de $k > 2$, visto que o efeito observado no paciente pode estar relacionado a apenas um desses pares, sendo pouco provável, ou passível de observação clínica, a relação do evento aos dez fármacos simultaneamente.

Além da capacidade de generalização necessária para a construção da rede de atributos de fármaco e atributos de combinação, a avaliação do espaço completo de hipóteses é necessária para a cobertura de todos os casos possíveis. Esta cobertura somente é viabilizada devido a exploração das características que tornam os fármacos propensos a interagir, e não diretamente da interação em si, dado que o número de instâncias conhecidas, ou mesmo, de combinações entre fármacos-eventos ou fármacos-biomoléculas detém, pelo menos, uma a duas ordens de grandeza a menos que o universo de possibilidades usualmente abordado. Em geral os outros modelos estabelecem generalizações comutativas. Por exemplo, se o fármaco $f_1$ interage com $f_2$ e $f_2$, também deve interagir com $f_3$. Neste caso, são avaliados os aspectos correlatos da combinação entre $f_1$ e $f_3$ em relação às interações relacionadas. O mesmo conceito se aplica

caso $f_2$ for uma biomolécula.

Embora Segura-Bedmar et al. [2011a] e Gottlieb et al. [2012], dentre outros autores, tenham relatado o conjunto de hipóteses como a exploração de todos os pares de fármacos contemplados, apenas Gottlieb et al. [2012] explorou o espaço completo do número combinações, porém, em um conjunto restrito de fármacos.

A restrição do número de fármacos ou combinações do espaço de hipóteses deve-se à premissa que impõe a necessidade do vínculo corroborável pelo domínio do conhecimento científico estabelecido *a priori*, seja farmacológico ou farmacoepidemiológico. Desta forma, os autores optam por tratar o conhecimento explícito previamente de modo a gerar uma estrutura coerente com o domínio do conhecimento assumido.

A abordagem proposta estabelece o vínculo fármaco-atributo e combinação-atributo extemporaneamente[12] ou *a posteriori* com a extração da semântica implícita, latente. O papel da seleção do conhecimento relevante é realizado pelos algoritmos de aprendizado de máquina. Estes algoritmos extraem informações preditivas a partir dos dados modelados como medidas de diferença ou similaridade entre pares de fármacos com interações conhecidas para extrapolar às combinações desconhecidas.

A modelagem dos dados proposta assume que qualquer dado farmacológico disponível pode influenciar na capacidade do fármaco interagir, logo, nenhum atributo é descartado *a priori*. Não é realizado nenhum pressuposto taxonômico ou farmacológico para a seleção dos atributos, evitando que pares de fármacos sem as informações eleitas no escopo metodológico não sejam comparados e suas combinações sejam impossibilitadas de serem previstas. Salienta-se que o espaço de hipóteses deve ser formado por todas as combinações de $k$ a $k$, a partir do conjunto de fármacos $F$.

### 4.4.4 Construção dos dados de combinações de fármacos

A construção dos dados $N$ de combinações $A$ de fármacos $F$ ocorre a partir da formação da matriz $N$, a qual contém um atributo em cada coluna e uma combinação em cada linha a partir das descrições $D$ coletadas dos fármacos $F$.

Conforme exposto, cada atributo $D_x \subseteq D$ relativo ao conjunto $F$ constituirá uma matriz $M_x \subseteq M$ ou matriz decomposta $W_x \subseteq W$, em que cada linha representa um fármaco $f \in F$. Cada representação $m \in M_x$ ou $w \in W_x$ do fármaco $f$ na forma de vetor de um atributo será tratada como um ponto no espaço n-dimensional. A matriz de entrada $N$ para o modelo de aprendizado relativa ao espaço de hipóteses $Q \subseteq A$ é formada a partir do cálculo da distância $n_{ij} = \delta(m_i, m_j)$ ou $n_{ij} = \delta(w_i, w_j)$ dos respectivos vetores de cada par $f_i$ e $f_j$, $\forall i \neq j$ relativos a cada atributo $M_x$, $M_y$,..., $M$ ou $W_x$, $W_y$,..., $W$.

As tomadas de distância $\delta \in \Delta$ entre os vetores de $M$ ou $W$ que representam fármacos são avaliadas quanto a pelo menos uma métrica de diferença e uma de proximidade. Logo, são

[12] Se pudermos cunhar um intermediário entre *a prior* e *a posteriori* para não incorrer adequadamente no segundo termpo, eu sugeriria *extemporaneus*.

gerados, no mínimo, dois novos atributos que formarão a matriz de entrada para os modelos de aprendizado de máquina.

O cálculo mais difundido para o grau de diferença é obtido com a distância euclidiana (equação 4.1). Esta métrica visa satisfazer as propriedades de positividade, em que elementos iguais tem distância igual a zero; simetria, onde a distância $(x, y)$ é igual a $(y, x)$ e diferença triangular, em que a distância entre um terceiro elemento é proporcional em relação aos outros dois.

$$d(X, Y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \tag{4.1}$$

Estas propriedades refletem padrões que não são quantificáveis na comparação de conjuntos, dado que as repetições são tratadas como um elemento, ou diferenças não métricas, como o tempo. Neste último caso, o valor não intui uma distância a não ser que seja tratado, i.e., ao categorizar o horário de ingestão de um medicamento para "manhã", "tarde" ou "noite", dado que 14h30min e 14h49min intuem a mesma informação.

A medida de proximidade mais usual para lidar com informações desse tipo é a semelhança de cosseno. Nesta medida, a comparação de dois vetores binários equivalentes de fármacos resulta em valor igual a 1. Os vetores totalmente distintos resultam em valor igual a 0, dada a posição ortogonal[13] que assumem.

Esta abordagem é útil para avaliação de atributos na forma de texto, onde grande parte das matrizes são esparsas[14] As normalizações mantém os valores iguais a zero, não sendo um recurso para este problema. O produto obtido para cada posição entre dois vetores faz com que os valores dessemelhantes sejam desconsiderados. A consequência direta é a comparação de fármacos com magnitudes diferentes conforme o atributo, ou seja, dois fármacos com teores diferentes de termos assinalados nos vetores podem ser comparados, desde que algumas posições sejam comuns a ambos.

O $\cdot$ na equação 4.2 indica o produto interno do vetor $x \cdot y = \sum_{k=1}^{n} x_k y_k$, e $\|x\|$ é o comprimento do vetor $\|x\| = \sqrt{\sum_{k=1}^{n} x_k^2} = \sqrt{x \cdot x}$.

$$cos(X, Y) = \frac{x \cdot y}{\|x\| \, \|y\|} \tag{4.2}$$

### 4.4.5 Decomposição em Valores Singulares

A matriz binária de cada atributo de fármaco pode ser decomposta por tratamentos matemáticos para a remoção de ruído ou redução da dimensionalidade. Uma técnica utilizada pelo presente trabalho é a SVD, Decomposição em Valores Singulares.

[13] Em ângulo reto

[14] Uma matriz constituída em grande parte por zeros é chamada matriz esparsa.

A técnica conhecida como SVD (*Singular Value Decomposition*) é a função $\phi(M) = W$, cuja transformação $W \in \mathbb{R}^{m \times n}$ gera outras três matrizes, $USV^T$, sendo duas ortogonais, $U^m$ e $(V^T)^n$, e uma diagonal $S^{m \times n}$, a qual contém o vetor dos valores singulares, cuja utilidade é explicada abaixo.

Os valores singulares contidos na matriz diagonal são ordenados, isto é, $s_i \geq s_{i+1}$ sendo $1 \leq i \leq n$, e são usados para calcular a aproximação capaz de recuperar a matriz original que satisfaça o produto da equação $W' = USV^T$.

A operação $W = (S \times V^T)^T$ retorna uma matriz $m \times n$ com as dimensões originais, porém com os valores ordenados de forma decrescente conforme os valores singulares. A vantagem desta técnica é a evidenciação das características latentes das instâncias em relação ao conjunto de dados, particularmente útil em matrizes esparsas [Elden, 2006].

A matriz gerada pode ser reduzida ao se identificar o valor $k \ll min(m,n)$, o qual remete à dimensão de aproximação $W \approx (\Sigma \times \Psi^T)^T$ para $\Sigma \in \mathbb{R}^{m \times k}$, e $\Psi \in \mathbb{R}^{n \times k}$, cuja matriz resultante deve prover a capacidade informativa da original, porém com dimensão reduzida. Por esta razão, SVD é frequentemente empregada para redução de dimensionalidade.

O SVD é aplicado na matriz binária $M_x$ relativa de cada atributo $D_x$ de fármaco, gerando um novo atributo (ou coluna) em $N$. No entanto, o SVD pode ser usado na matriz de combinações de fármacos, se $N$ for constituída de colunas semanticamente correspondentes, i.e., quando a grandeza de todas as colunas intuírem o mesmo conceito e amplitude dos dados sob a mesma medida de distância. Por mesmo conceito entende-se escolha de variáveis análogas, i.e., variáveis discursivas tipo texto, ou valores numéricos com um fator de ponderação comum (i.e., normalização).

### 4.4.6 Treino e teste

O treino corresponde à fração de combinações conhecidas conforme um padrão ouro $C$ e o teste remete ao conjunto de combinações desconhecidas ou tratadas como desconhecidas enquanto artifício de validação. Na exploração do espaço de hipóteses $Q$ de combinações de medicamentos a serem consultadas é comum a relação $|C| \ll |Q|$, sendo inferior em duas ou mais ordens de grandeza.

A tomada de cada atributo como matriz de termos ocorre para o conjunto completo de fármacos. Porém, a tomada completa das distâncias para a formação da matriz de atributos de combinações ocorre somente quando as previsões forem realizadas com o modelo que apresentar melhor desempenho. Durante a fase de validação, somente as distâncias dos atributos dos fármacos das combinações do treino são tomadas.

## 4.5 Processamento de dados

O precessamento dos dados é a aplicação de uma ou mais técnicas que extraem padrões a partir dos dados, os quais são usados para estabelecer modelos preditivos.

A tarefa de mineração de dados adotada nesta abordagem foi a classificação (vide referencial teórico na seção 2.5.4.1). Em outras palavras, pretende-se assinalar os pares de fármacos em categorias pré-estabelecidas que culminem na identificação da interação medicamentosa.

A escolha do classificador deve ocorrer conforme a natureza dos dados. No entanto, assumiu-se que a natureza dos dados é desconhecida *a priori*. Logo, foi desenvolvido um método automático, capaz de aplicar diversos algoritmos de classificação e avaliação do desempenho conforme os resultados preditivos esperados.

A estruturação dos dados enquanto uma matriz numérica viabilizou a aplicação de diversos algoritmos.

### 4.5.1 Respostas preditivas

As respostas preditivas são os ponderadores da função-alvo. As respostas preditivas são àquelas modeladas para a alocação das instâncias às classes. O modelo em si abrange a resposta preditiva na medida que contém a explicação para a alocação das instâncias à classe abrigando informações quanto às variáveis mais importantes e aos padrões detectados.

Devido a escolha da tarefa de classificação, a resposta preditiva é categórica. Os pares de fármacos são avaliados, em última instância, como “interação” e ”não interação”.

É importante salientar que este aspecto dicotômico não é determinístico. O mesmo par de fármacos pode ser considerado interação ou não segundo a fonte do padrão-ouro avaliada. Ainda, um par de fármacos amplamente conhecido como interagente, pode ou não interagir na prática conforme as condições de saúde do paciente ou ser sinérgico ou terapêutico.

A resposta preditiva possui um compromisso com a classificação das instâncias conhecidas a partir do padrão-ouro (treino). O padrão-ouro adotado classifica a interação como “menor”, “moderada” ou ”maior”. Contudo, esta categorização representa caráter secundário para os objetivos pretendidos nesta abordagem.

A exploração completa do espaço de hipóteses deve ser comparada com as instâncias de treino. Mesmo que todos os pares do espaço de hipóteses sejam atrelados a estas três categorias, a ponderação mais relevante é a probabilidade deste enquadramento. Pares de fármacos do teste classificados com valores superiores a um limite definido, i.e., 95%, são considerados como similares às instâncias de treino, e, portanto, são rotulados como “interação prevista”. Os pares descartados podem ser considerados como inertes enquanto não houver evidência contrária. Desta forma, todo o espaço de hipóteses é rotulado.

### 4.5.2 Seleção de atributos

A seleção de atributos é a escolha daqueles que melhor representam a realidade a ser expressa e contribuem para a construção do modelo preditivo.

A seleção pode ocorrer *a priori*, ou seja, adotando-se atributos que reconhecidamente estão atrelados ao contexto de interações medicamentosas. Um exemplo frequente de tratamento *a priori* é a formação de uma estrutura de dados em que as enzimas do citocromo são arestas e os fármacos vértices. Esta estrutura fornece intuitivamente a determinação de interações farmacocinéticas pela verificação dos padrões comutativos citados anteriormente. A seleção de atributos *a posteriori*, é realizada pela estratégia de aprendizado de máquina, ou ainda, manualmente, sob critérios do domínio do conhecimento, e, ao final do modelo, a partir da avaliação individual de sua contribuição preditiva. A seleção de atributos pode ocorrer internamente no processo de aprendizado de máquina por meio de técnicas supervisionadas, ou seja, baseadas nas instâncias conhecidas; ou por meio de técnicas não supervisionadas.

**Seleção supervisionada** A seleção supervisionada é aquela baseada nos atributos das instâncias conhecidas, ou seja, das interações classificadas. Logo, a seleção é realizada em função da classe e estendida para os demais atributos no treino, quando apenas as classes escolhidas serão adotadas.

Um exemplo de seleção de atributos é a tomada dos subconjuntos diante da avaliação de todas as possibilidades no espaço $2^n$. Porém, este número é frequentemente proibitivo de ser explorado. Tipicamente, o espaço de busca explorado por métodos gulosos toma a direção a partir de um dos extremos da matriz de entrada. Em cada etapa, uma alteração local é feita para o atual subconjunto de atributos diante de qualquer adição ou exclusão de um atributo.

## 4.6 Análise de dados

Após a definição do problema, escolha do modelo de dados e estabelecimento do modelo preditivo, o processo KDD culmina com a análise dos dados é o uso dos modelos e das informações mineradas, verificando-se a efetividade da estratégia de mineração de dados. Nesta etapa são definidos ajustes que possam demandar nova coleta e extração dos dados.

### 4.6.1 Avaliação *ad hoc* da previsão de instâncias desconhecidas

A interação prevista e desconhecida (ausente do padrão ouro) deve ser avaliada em função da probabilidade conjunta em se identificar uma interação ao acaso dentre as interações conhecidas.

A avaliação da chance de um fármaco interagir com os demais pode ter como ponto de partida as interações conhecidas. Por exemplo, o fármaco lepirudina possui 14 interações conhecidas segundo as bases DrugBank e Drugs.com dentre os demais 1.388 fármacos que

formarão os pares adotados para classificação. Logo, a probabilidade de acertar ao acaso será de $p(Lepirudina|F) = 14 \div 1388 = 1,01\%$.

Considerando que a mediana e os quartis da frequência de interações para o conjunto de interações conhecidas são $\tilde{X} = 22$, $Q_1 = 6$ e $Q_3 = 52$; a chance de prever ao acaso a interação de um fármaco qualquer em relação aos demais é cerca de 1,59% (0,43% a 3,74%).

Em outra abordagem, tomam-se os casos conhecidos em relação ao tamanho do universo de pares de fármacos. Seja o $A$ o conjunto de pares de fármaco e $A_k$ o conjunto de instâncias conhecidas, as cardinalidades $|A| = 965.355$ e $|A_k| = 41.654$, a proporção de casos conhecidos em relação ao total é $1 : 23$ ou $p = 4,31\%$.

Se o valor de interações existentes estiver estagnado ao número de conhecidas, a identificação de novos casos torna-se um fenômeno raro e a proporção estimada é $\hat{p} \ll p$. Porém, se existem combinações ainda sem avaliação, presumindo-se que existe, ao menos, o dobro de interações, o valor esperado é assumido como $\hat{p} \gtrsim 2p$.

A grosso modo, esta conjectura deve-se à correspondência das citações MEDLINE em 51,6% nas buscas pelos nomes genéricos combinados em relação aos 29,6% obtidos para os pares desconhecidos, projetando o impacto desta importante fonte em relação às demais (seção 5.5.1.1).

#### 4.6.1.1 Amostragem

Seja $\hat{p} = p$, supondo igual proporção de casos desconhecidos; $\hat{q} = 1 - p$; o total de casos $N = |A| - |A_k|$; $\hat{d} = 5\%$, a precisão absoluta desejada e $Z^2_{1-\alpha \div 2} = 1,96$; o cálculo da amostragem, baseado na equação 4.3 [Scheaffer et al., 2011], sugere a observação de 64 casos para 95% de confiança ou 78 para 97% de confiança.

$$n = \frac{N \times \hat{p} \times \hat{q}}{\frac{\hat{d}}{1,96^2}(N-1) + \hat{p} \times \hat{q}} \tag{4.3}$$

### 4.6.2 Comparação com outros estudos

A comparabilidade entre estudos requer analogia metodológica viabilizada pela adoção de fontes similares do domínio do conhecimento e objetivos aproximadamente comuns.

A restrição inerente à diversidade dos estudos faz com que valores de acurácia e precisão devam ser vistos parcimoniosamente devido ao grau de correspondência com a realidade abordada, sobretudo diante das ferramentas de validação.

Diversas formas de comparação foram apresentadas em detalhe no capítulo 3 com critérios relativos as bases de dados, validação e achados diante dos objetivos propostos. Verificou-se que a objetividade das métricas de validação não é suficiente para pontuar a relevância dos estudos, tornando imprescindível a comparação discursiva e crítica por pesquisadores experimentados.

## 4.7 Sumário do modelo

A seguir são condensadas as etapas percorridas pelo modelo, cujo panorama é dado na figura **??**.

(I) Os dados $D_1$ coletados e tratados por funções $\pi$ a partir de diversas bases do conhecimento são atrelados ao fármacos por métodos de extração $\xi$ (frequentemente de linguagem natural) que atendam aos compromissos de univocidade respectiva à fração ativa do fármaco, constituindo os dados armazenados $D_3$. (II) Características $1 : n$ são agrupadas em um atributo quando intuem o mesmo conceito ou alocadas em atributos distintos, como o caso de ontologias ou taxonomias hierárquicas. (III) Após a definição do conjunto de fármacos a ser explorado com base nos dados disponíveis, cada atributo $D_x \subseteq D_3$ de fármaco é decomposto a uma matriz de frequência $M$ por uma função $\psi$, em que cada linha representa um fármaco e cada coluna um termo da decomposição do atributo. (IV) O dígito 1 ou a frequência, conforme pré-estabelecido, é assinalado na posição correspondente ao termo e ao fármaco, sendo as posições os restantes preenchidas com zero. (V) A matriz pode sofrer redução da dimensionalidade por filtros $\psi$ com remoção de atributos com filtros pré-estabelecidos. (VI) Cada matriz $M$ correspondente ao atributo de fármaco pode sofrer Decomposição por Valores Singulares pela função $\phi$ gerando uma nova matriz $W$. (VII) Para cada matriz de atributo de fármaco $M$ ou $W$, as distâncias entre as linhas correspondentes aos fármacos das combinações $C$ de interação conhecida segundo padrão-ouro integradas ao espaço de hipóteses $Q$ pela função $\varepsilon$ são calculadas por diversas métricas $\delta$. Cada métrica gera um atributo para a matriz de combinação de fármacos. A matriz de combinações de fármacos terá o número de atributos de fármacos sem SVD $M$, mais o mesmo número com SVD $W$, sendo estes números multiplicados pelo número de métricas $\delta$ usadas para o cálculo da distância entre os vetores dos fármacos. (VIII) Os vetores $n_{ij}$ das distâncias $\delta(m_i, m_j)$ e $\delta(w_i, w_j)$ respectivas a cada atributo de fármaco, em que cada célula é uma combinação $a = (f_i, f_j)$, são concatenados de modo a formarem a matriz de combinações de fármacos, em que cada linha corresponde a mesma combinação $a_i \in A$ de fármacos e cada coluna um tratamento de atributo de fármaco elaborado nas etapas anteriores. (IX) Funções $\theta$ de seleção de atributos pode reduzir verticalmente a matriz $N$. (X) As funções $\gamma$ constroem modelos na matriz de combinações $N$, fragmentando-a horizontalmente em $k$ partes de forma aleatória estratificada, mantendo a proporcionalidade das classes assinaladas pelo padrão ouro $C$ em cada parte $k$. (XI) O treino é realizado em $k$ iterações tomando-se $k-1$ partes, por diversas pré-configurações de modelos de aprendizado de máquina, incluindo ou não seleção de atributos, sendo o desempenho avaliado na parte $k$ não utilizada para treino com base em métricas derivadas da matriz de confusão. (XII) Após avaliação nas $k$ partes, o desempenho médio é calculado por funções $\omega$ a partir das previsões $R$ e armazenado em $P$, sendo escolhido o melhor modelo $\gamma$ conforme métrica pré-estabelecida (i.e., $kappa \rightarrow 1$). (XIII) A configuração do melhor modelo ($\Psi_x$, $\Phi_x$, $E_x$, $\Delta_x$, $\Theta_x$, $\Gamma_x$) quanto aos atributos selecionados e métricas de distâncias é reaplicada no cálculo das distâncias entre todas as linhas correspondentes

tes a cada fármaco da matriz pré-processada e cada atributo de fármaco selecionado, gerando uma nova matriz de distâncias $N$ relativa ao universo completo de combinações $Q$. As previsões são realizadas no espaço completo de hipóteses adotando-se o algoritmo de aprendizado de máquina com maior desempenho, incluindo as instâncias conhecidas, agora tratadas como desconhecidas. (XIV) O desempenho final $P$ é calculado com base na previsão das instâncias conhecidas. (XV) As previsões são avaliadas por especialistas com modelos $\sigma$, verificando-se em $V$ a relevância com dados $K$ relativos ao uso por populações, compêndios e periódicos.

# Capítulo 5

# Mineração farmacológica de interações

O modelos de previsão de interações medicamentosas elencados no capítulo 3 e o proposto são elaborados em função de dois tipos de fontes de dados. A primeira fonte é populacional e a segunda é farmacológica. Fontes populacionais com registros na ordem de dezenas de milhares detém potencialmente padrões de perfis de utilização de medicamentos que podem ser atrelados a eventos clínicos, contudo, frequentemente demandam um número mínimo de observações para averiguação de interações, o que reduz o espectro de fármacos abrangido. A fonte farmacológica verifica padrões nas redes bioquímicas ou farmacológicas que possam atar eventos e combinações de medicamentos e explicar sua interação, desde que se conheça estas rotas.

A modelagem aplicada neste capítulo lida com dados farmacológicos, porém, objetiva avaliar as características dos fármacos em si, como pré-requisito para a verificação das características potencialmente diretas que explicam sua interação. Admite-se que a propriedade de interagir está relacionada, em alguma instância, às propriedades que constituem o fármaco, cuja natureza é apreendida a partir da integração das características, e podem agregar padrões preditivos comparativamente. Neste nível da abordagem, as características que descrevem rotas envolvidas nas interações não precisam ser manualmente determinadas, tão pouco os padrões populacionais de uso. Isto se deve ao posicionamento dos fármacos frente a múltiplas características avaliadas independentemente, de modo que a sobreposição destas características aproxima fármacos e interações, identificando a característica desejada por extrapolação das similaridades com instâncias conhecias.

A técnica consiste em processar cada atributo de fármaco, projetando-o como um ponto no espaço n-dimensional. Cada ponto é representado por um vetor binário gerado a partir da decomposição do atributo em características dicotômicas. O espaço de hipóteses é constituído a partir da combinação de todos os fármacos elencados aos pares. Medidas de diferença e proximidade entre os vetores de cada fármaco e atributo são tomadas constituindo vetores de

distâncias, em que cada posição representa a mesma combinação de fármacos. Os vetores combinados constituem matrizes de entrada para os modelos de aprendizado de máquina, sendo os atributos alocados nas colunas e cada combinação representada por uma linha. A função de aproximação com melhor desempenho, a partir da validação cruzada, é eleita para as previsões farmacológicas. O impacto das previsões é avaliado quanto à frequência de citações MEDLINE.

## 5.1 Definição do problema

A interação medicamentosa é consequência da modificação da ação de um fármaco por outro. A descoberta de interações ocorre a partir da verificação individual da natureza da combinação ou com a busca em padrões que associem o uso de medicamentos a condições de saúde a partir de grandes bases de dados.

Não é possível adotar-se bases populacionais para a avaliação exploratória de cada combinação potencial em um conjunto de fármacos dada a ausência das combinações entre todos os fármacos, o que pode chegar a cerca de vinte e cinco milhões de possibilidades, baseando-se no número de substâncias farmacológicas atualmente conhecidas. A abordagem proposta supera esta limitação por basear-se no conhecimento estabelecido para cada fármaco, não da avaliação do fenômeno da interação em si. A comparação global de fármacos possibilita a verificação local da potencialidade de interação ou o caráter inerte de elementos do conjunto.

A informação da potencialidade da interação é útil no caso em que um profissional de saúde estiver monitorando um paciente com uso de polifarmácia. A existência de uma interação medicamentosa é constatada clinicamente, sendo desejável conhecer a potencialidade dada a quantidade de combinações possíveis. Desta forma, um paciente que estiver usando seis medicamentos terá $\binom{6}{2} = 15$ combinações teóricas de pares de medicamentos. Caso houver relato de que algum desses pares possua potencialidade de interagir, a atenção dada a combinação deve ser proporcionalmente elevada diante dos riscos, ou ainda, a respectiva medicação deve ser preventivamente suspensa ou substituída.

Por esta razão, o problema foi inicialmente definido como a apreensão das características intrínsecas dos fármacos, comparados par-a-par, que possibilitam a verificação de sua interação. A solução constitui uma primeira linha de informação para combinações desconhecidas, dicotomizando-as entre “interações potenciais”, quando suas características se aproximam das instâncias conhecidas, ou “combinações inertes”, quando não há fatores preditivos o bastante para elencá-las como interação. Para que isso ocorra deve-se estabelecer uma estratégia que seja capaz de abrigar o maior número de informações possível, de modo que os fármacos menos conhecidos sejam caracterizados em dimensões que permitam a comparação com fármacos dotados de interações conhecidas, justificando o uso de variáveis que não contenham informações diretamente relacionadas com o fenômeno da interação.

A completeza das informações requer a habilidade da coleta em diferentes fontes de dados. Ainda, deve ser mensurada a relação da qualidade da informação em termos de cobertura

e especifidade dada a característica dialética do funcionamento do fármaco e das interações. Os fármacos podem ser agentes terapêuticos ou tóxicos e, da mesma forma, as combinações podem apresentar sinergismo ou serem interações adversas. O compromisso entre a acuidade da informação ou seu caráter de alerta geral, onde não há nenhum conhecimento, deve ser explicitado e avaliado em função do padrão ouro utilizado. A definição do padrão ouro deve deixar claro o nível de corroboração das evidências utilizadas pois será refletido nos padrões apreendidos pelo modelo, logo, repercute no posicionamento das informações geradas frente ao preconizado pelas práticas de saúde baseadas em evidência.

Seja para a previsão de interações totalmente novas ou para a observação de padrões que superem a inerente obsolescência dos bancos alimentados com interações conhecidas, o modelo de descoberta de interações medicamentosas com base em técnicas *in silico* mostra sua importância frente às constantes atualizações do conhecimento e da demanda por manipulação de dados em vertiginosa acumulação.

O método proposto abrange os requisitos abordados por ser capaz de oferecer alertas dicotômicos de interações medicamentosas potenciais e ainda estabelecer a possível classe relativa ao padrão ouro, com a ressalva do compromisso entre a cobertura e a especificidade do padrão-ouro abordado.

A seguir é mostrada uma aplicação do processo em que foi utilizado um padrão-ouro de dados do sítio Drugs.com disjunto da base de conhecimento construída a partir da integração entre DrugBank, KEEG, ATC/OMS, ExPASy e ENZYME. O Drugs.com apresenta informações de interações medicamentosas que versam pela generalidade, porém, sem serem corroboradas em grande parte pelo DrugBank, o qual também fornece informações de interações medicamentosas. Logo, o modelo implementado é orientado a oferecer um alerta para interações completamente desconhecidas.

## 5.2 Extração de dados

### 5.2.1 Definição do domínio do conhecimento

Algoritmos em linguagem bash foram implementados para coleta das bases a partir do DrugBank, compilação e exportação para o SGBD (Sistema Gerenciados de Banco de Dados) MySQL 5.5.310 ubuntu 0.12.04.2 [Widenius et al., 2002].

A coleta (figura 5.1) foi realizada diretamente a partir das respectivas páginas web, sendo as marcas de codificação do formato hiper-texto utilizadas como referência para a incorporação do Drugs.com, DrugBank, KEGG e ExPASy.

O ATC/OMS e os números EC (ENZYME) foram obtidos na forma de planilha e convertidos em formato mysql para incorporação no SGBD.

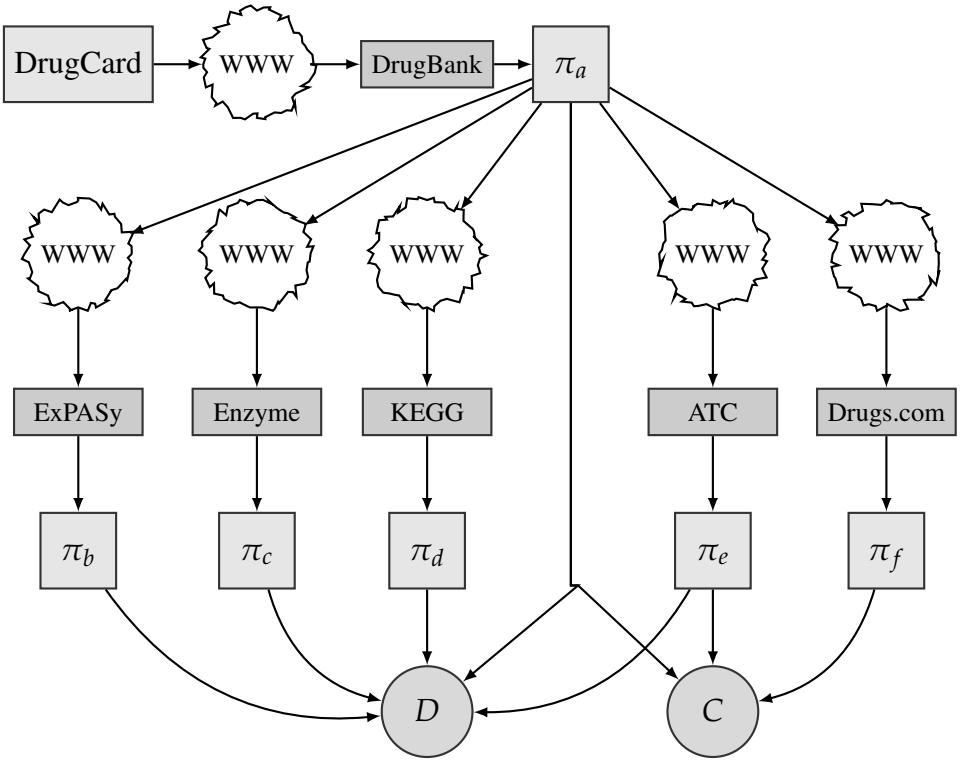



Figura 5.1: **Coleta de dados farmacológicos.** A coleta das descrições relativas a fármacos presentes no DrugBank realizada com funções $\Pi$ de extração e conversão para cada base de dados de fármacos $D_x \subseteq D$ e combinações $C$ de interações conhecidas.

#### 5.2.1.1 Dados farmacológicos

Variou-se o drugcard de DB00001 a DB08914 para a coleta de todo o DrugBank. Os campos respectivos a cada fármaco foram associados pelo drugcard na coleta de informações relativas às outras bases. Por exemplo, se uma enzima foi relacionada pelo DrugBank ou KEGG, os registros coletados da base ENZYME foram atrelados ao drugcard que originou a busca. Fármacos sem interseção com o DrugBank não foram empregados.

Ao final, uma tabela com 6.937 registros e 329 campos, sendo 164 originais e 165 derivados ou associativos, foi construída para o objeto "fármaco". A tabela 5.1 mostra a distribuição do atributos originalmente coletados. A lista completa de atributos encontra-se no apêndice D.

#### 5.2.1.2 Mecanismos bioquímicos farmacológicos

Os mecanismos (p.ex, *ExPASy_reaction*, *kegg_pathway_orthology*) foram tratados como atributo categórico (seção 4.4.2.2).

#### 5.2.1.3 Dados taxonômicos ou ontológicos

**Interação farmacológica** A base Drugs.com foi a fonte de interações medicamentosas potenciais adversas. Os nomes genéricos correspondentes aos fármacos DrugBank ou ATC/OMS

Tabela 5.1: **Atributos originais coletados** para caracterização do objeto "fármaco".

| Base | Campos texto | Campos numéricos | Exemplo |
|---|---|---|---|
| DrugBank | 62 | 16 | *absorption*, *mass*, *description* |
| KEGG | 53 | 9 | *disease*, *orthology*, *activity* |
| ATC | 10 | 0 | $atc1$, *atc1 name* |
| ENZYME | 8 | 0 | $eclevel1$, *ec level1 name* |
| ExPASy | 6 | 0 | *reaction*, *name*, *general comments* |

foram utilizados como referência para a coleta das informações de interação relativas a cada fármaco. Foram coletados 351.164 registros de interações entre medicamentos, incluindo nomes comerciais. Estes registros foram associados ao respectivos drugcards resultando em 18.123 combinações distintas.

Recuperou-se para cada combinação as respectivas gravidades classificadas pelo Drugs.com ("menor", "moderada" e "maior").

**Combinação segura** Embora a polifarmácia seja amplamente utilizada, não foi localizada base de dados que contenha informações de interações benéficas, tão pouco, combinações seguras ou inertes.

A classificação ATC/OMS possui algumas substâncias que compartilham o mesmo código químico-terapêutico (nível 5). Por exemplo, "J01CR02" define amoxicilinina e ácido clavulânico para infecção de uso sistêmico. O DrugBank possui informação de fármacos associados na mesma apresentação. Os fármacos contidos no mesmo medicamento ou sob o mesmo ATC foram relacionados dois-a-dois como combinações seguras. A ausência do dado referente ao benefício terapêutico fez com que estas combinações fossem tratadas como casos neutros.

Foram adotados como exemplos negativos as combinações de fármacos que compartilham o mesmo ATC nível químico, assumindo *a priori* sua segurança. Admite-se que alguns casos possam manifestar interação medicamentosa diante de regimes terapêuticos em condições de saúde desfavoráveis, porém, este grupo constitui um conhecimento diferente do que as interações explicitamente adversas descritas pelo DrugBank.

As características das interações medicamentosas referentes as duas bases são sumarizadas na tabela 5.2.

### 5.2.2 Identificação do objeto farmacológico de estudo

O objeto farmacológico de estudo é o conjunto $A$ de combinações $a$ de fármacos $f \in F$ aos pares, sendo $a = \{f_x, f_y | f \in F \wedge x \neq y\}$. A identificação unívoca dos fármacos $f$ constitui a identificação do objeto de estudo.

Tabela 5.2: **Interações medicamentosas coletadas**. As combinações seguras padrão-ouro advém de fármacos que compartilham o mesmo ATC/OMS. As combinações seguras do DrugBank advém das apresentações comerciais que contém mais de um ativo. Coleta realizada em julho de 2013.

| **Classificação** | **Padrão-ouro**[†] | **DrugBank** | **Padrão-ouro⋃DrugBank** |
|---|---|---|---|
| Seguras/Não | 195 | 687 | 802 (90,9%) |
| Menor | 1.218 | | |
| Moderada | 14.659 | | |
| Maior | 2.246 | | |
| Sim | 18.123[‡] | 12.786 | 27.077 (87,6%) |

[†] Drugs.com e ATC/OMS. [‡] Menor⋃Moderada⋃Maior.

Foi realizada indexação a partir da identificação da porção ativa de cada fármaco fornecida pelo DrugBank. A escolha do drugcard ocorreu devido à indexação disponível para diversas fontes passíveis de serem sistematicamente recuperáveis. Ainda, o DrugBank prioriza estudos relacionados a medicamentos, contendo grande número de fármacos. São 6.937 substâncias os quais geram 24.057.516 de pares de combinação, sendo este o universo de hipóteses máximo a ser explorado.

Embora o DrugBank ofereça integridade referencial para o KEGG e ATC, foi realizada indexação manual de forma complementar para aumentar a quantidade dos dados de fármacos e favorecer a sobreposição das características para a detecção de padrões.

## 5.3 Engenharia de dados

Foi implementado um algoritmo em bash para integração do dados armazenados no servidor MySQL, constituição dos vetores de termos, cálculo de distâncias, SVD e classificação usando ferramentas do scilab [Scilab Enterprises, 2012] e weka [Witten & Frank, 2005]. Os experimentos foram realizados em um computador AMD Phenon II X6 1075T, com 16GB RAM 1,333MHz, ubuntu 12.04 64-bit, kernel linux 3.2.0-48-generic Gnome 3.4.2.

### 5.3.1 Limpeza dos dados

A limpeza dos dados é um pré-tratamento necessário para assegurar a correspondência dos dados com a realidade. Porém, devido ao aspecto exploratório, o mínimo de pré-tratamento foi realizado, resumindo-se à remoção de ruídos com SVD a partir da extração das características latentes dos dados.

Não foi realizada reposição de dados faltantes. Se o fármaco não apresentou a informação para determinado atributo, gerou-se vetor nulo. As tomadas de distância cujo denominador fosse considerado como zero foi dada como desconhecida, não sendo considerada pelo classi-

ficador. A estratégia para mitigar o impacto foi utilizar diferentes métricas de distância e um grande número de atributos para reduzir o impacto global de uma ausência local.

A remoção de ruído foi realizada adotando-se Decomposição em Valores Singulares (seção 4.4.5) após a formação das matrizes para cada atributo de fármaco.

## 5.3.2 Transformação dos dados farmacológicos

A transformação dos dados é uma etapa crítica para o processo de mineração e consiste na geração das matrizes de entrada no formato requerido pelos algoritmos de aprendizado de máquina.

Dentre os seis mil fármacos coletados, utilizou-se neste experimento os que continham algum tipo de informação dada pelo DrugBank e ATC simultaneamente. O espaço de hipóteses abrigou 965.355 pares de 1.390 fármacos. O espaço de hipóteses possui 18.340 interações conhecidas de 971 fármacos de acordo com Drugs.com, correspondendo a 1,9% do total de combinações (tabela 5.3).

Este experimento abrangeu atributos texto como “indicação”, “farmacologia”, “toxicologia”, “subgrupo químico” e 20 atributos numéricos discretizados como “solubilidade em água”, “ponto de fusão” e ”peso molecular médio”. Não adotou-se *steming* e *stop words* neste experimento devido a minimização do pré-tratamento citada anteriormente.

### 5.3.2.1 Atributo numérico

Os atributos numéricos foram adotados conforme o valor original e também foram convertidos a bases $log(x)$, $x^2$, $\sqrt{x}$, $x^{-1}$ e $|x|$. Os atributos foram discretizados em seções com o mesmo número de classes do padrão ouro e foram convertidos em vetor binário, cuja presença foi assinalada para cada instância de fármaco na posição relativa à seção.

### 5.3.2.2 Atributo em formato categórico

Cada variável $1 : n$, como forma farmacêutica ou código ATC, foi concatenada para cada fármaco utilizando-se caractere espaçador. Os caracteres não alfanuméricos foram removidos e os restantes reduzidos a letras minúsculas, mantendo-se o um espaço entre cada termo (seção 4.4.2.2).

### 5.3.2.3 Atributo taxonômico, ontológico ou mecanicístico

Cada nível da classificação ATC e das enzimas relacionadas foi alocado em um atributo distinto, bem como os nomes de cada nível. Cada termo foi tratado como uma posição no vetor de cada fármaco, sendo tratado analogamente como os atributos na forma de texto.

### 5.3.3 Espaço de hipóteses

O espaço de hipóteses é conjunto de objetos de estudo que devem ser atrelados ao evento levantado pela definição do problema.

O espaço de hipóteses foi elaborado de modo a conter a combinação de $n$ fármacos para-par. A exploração completa do espaço de hipóteses aos pares é necessária para a formação de matrizes de atributos de fármacos que contemplem um grande número de instâncias e assegurem a generalidade da definição do atributo. Espera-se que atributos bem definidos elevem a correspondência com a realidade comparada.

### 5.3.4 Construção dos dados de combinações de fármacos

A construção dos dados de combinações de fármacos é a etapa final na elaboração das matrizes de entrada em que se relaciona os dados dos fármacos com a tomada de distâncias dos vetores correspondentes.

Os vetores de frequência de termos para cada atributo foram avaliados com e sem SVD e sob as distâncias euclidiana (equação 4.1) e cosseno (equação 4.2) normalizada a $-log((1 + cos(x,y)) \div 2)^2$ conforme ilustrado na equação 5.1, sendo $x$ e $y$ linhas das matrizes $M$ ou $W$ correspondentes aos fármacos $f_i$ e $f_j$, respectivamente.

$$\delta(x,y) = \left( -\log \frac{1 + \frac{x \cdot y}{\|x\| \cdot \|y\|}}{2} \right)^2 \tag{5.1}$$

As distâncias euclidiana e de cosseno constituíram 6 matrizes, sendo duas para cada distânca sem SVD, duas com SVD e duas concatenadas com e sem SVD. Estas matrizes foram submetidas ao algoritmo *CfsSubetEval* LinearForwardSelection do weka para seleção supervisionada de atributos seguido de reamostragem não supervisionada, o qual gerou mais seis matrizes, perfazendo um conjunto experimental de 12 matrizes.

### 5.3.5 Decomposição em Valores Singulares

A Decomposição em Valores Singulares foi realizada como parte do processo de limpeza dos dados por se tratar de uma operação de remoção de ruído. Não realizou-se redução de dimensionalidade devido ao custo combinatorial em se identificar o valor de $k$ para a redução de cada matriz em relação à formação da matriz final de combinações (seção 4.4.5).

Seja um conjunto de atributos de fármacos $D_1$, $D_2$, ..., $D_w(F) \in D$ convertidos a matrizes de frequência $M$. A decomposição destas matrizes e sua recuperação com a operação $(S \times V^T)^T$ resulta nas matrizes $W_x \subseteq W$, sendo $1 \leq x \leq w$ e $w$ equivalente ao número de atributos de fármacos. Tanto o valor de $q$ quanto os respectivos valores de $k$ relativos à redução de dimensionalidade são variáveis para cada $M_x^{n \times q}$ e $W_x^{n \times q}$. Não obstante, a definição do valor de $k$ somente pode ocorrer com a formação da matriz de combinações e avaliação do desem-

penho conforme cada algoritmo de aprendizado de máquina. Logo, a otimização da matriz de combinações requer a avaliação local de cada valor de $k$ quanto à contribuição no desempenho das previsões, correspondendo a uma iteração do modelo. O número máximo de iterações para cada classificador que determina a otimização do valor de $k$ para cada atributo decomposto em $W$ é calculado com a equação 5.2.

$$\text{Número máximo de iterações} = w\prod_{i=1}^{w} q_i \tag{5.2}$$

O problema torna-se $np$-completo [1], se considerarmos que cada valor de $k$ pode variar conforme o classificador $\gamma$ adotados, ou ainda, um valor de $k$ pode demandar otimização em relação aos demais valores de $k$ escolhidos, com um número máximo de iterações resultando na equação 5.3.

$$\text{Número máximo de iterações} = w! \times |M| \times \prod_{i=1}^{w} q_i \tag{5.3}$$

Não foi realizada redução de dimensionalidade na matriz de atributo de fármaco devido ao custo computacional, dada a quantidade de matrizes geradas para cada atributo respectivas aos tratamentos (operações matemáticas e filtros). Acredita-se que os valores reduzidos ou zerados ao final dos vetores causam pouco impacto na tomada de distâncias, evidenciando-se as características latentes com a remoção de ruído, tornando natural a poda do valor de $k$ conforme exposto na seção 4.4.5.

### 5.3.6 Treino e teste

O treino é a construção das funções de aproximação pelos algoritmos de aprendizado de máquina com base nas instâncias conhecidas (base de treino). O teste é realizado nas instâncias desconhecidas, constituindo as previsões (base de teste).

As matrizes de atributos de fármacos foram geradas para todos os fármacos, não apenas àqueles que participam da base adotada como padrão-ouro. No entanto, as combinações foram separadas em treino e teste conforme o conhecimento das instâncias. Logo, foram calculadas as distâncias entre os fármacos cujas combinações são conhecidas e verificado o melhor desempenho relativo ao tratamento dos dados, à seleção e ao classificador. Posteriormente, todas as distâncias foram tomadas perfazendo o conjunto final de teste contendo o espaço completo de hipóteses.

[1] Complexidade não tratável computacionalmente de forma exaustiva

Tabela 5.3: Classificadores adotados no modelo geral para previsão de interações farmacológicas.

| Modelo | Classificador |
|---|---|
| Bayes | BayesNet, NaiveBayesUpdateable |
| Functions | Logistic, SMO, SimpleLogistic, MultilayerPerceptron, RBFNetwork |
| Lazy | IB1, IBk, KStar, LWL |
| Meta | AdaBoostM1, AttributeSelectedClassifier, Bagging, ClassificationViaRegression, CVParameterSelection, Dagging, Decorate, END, FilteredClassifier, Grading, LogitBoost, MultiBoostAB, MultiClassClassifier, MultiScheme, nestedDichotomies.ClassBalancedND, nestedDichotomies.DataNearBalancedND, nestedDichotomies.ND, OrdinalClassClassifier, RacedIncrementalLogitBoost, RandomCommittee, RandomSubSpace, RotationForest, Stacking, Vote |
| Misc | HyperPipes, VFI |
| Rules | ConjunctiveRule, JRip, OneR, NNge, Ridor, ZeroR, PART |
| Trees | DecisionStump, FT, J48, J48graft, LADTree, LMT, NBTree, RandomForest |

## 5.4 Mineração de dados

### 5.4.1 Respostas preditivas

A resposta preditiva almejada é a caracterização da interação como inerte ou não inerte a partir da elevada correspondência com um grupo de características adotada pelo modelo para a designação das classes definidas a partir do padrão ouro (“segura”, “menor”, “moderada” e “maior”). Logo, a classificação com elevada probabilidade em alguma dessas classes reporta sua proximidade com os elementos cuja interação possui algum tipo de evidência.

### 5.4.2 Seleção supervisionada

O modelo é aplicado inicialmente ao conjunto completo de atributos. Porém, adotou-se a seleção de subconjunto de características para avaliar atributos redundantes que podem reduzir a precisão.

Conforme exposto na seção 4.4.4, foi adotado o método *CfsSubsetEval* de seleção para frente. O conjunto inicia vazio, sendo os atributos adicionados um a um. O algoritmo considera o valor preditivo de subconjuntos de atributos e avalia a redundância entre eles. O *CfsSubsetEval* indica conjuntos de atributos elevada correlação com a classe, porém de baixa intercorrelação [Witten & Frank, 2005].

As doze matrizes com 18.340 instâncias conhecidas foram submetidas aos 52 classificadores mostrados na tabela 5.3 sob validação cruzada em 10 partições (seção 2.5.5.1).

Os três algoritmos e a matriz com o maior coeficiente *Kappa* (seção 2.5.5.2) tiveram

os parâmetros manualmente variados verificando-se os resultados de curva ROC, acurácia e precisão.

O algoritmo cujo pré-tratamento e classificação atingiu melhores resultados foi escolhido para realizar a previsão no espaço completo de pares hipotéticos. As previsões inferiores a 0,95 de probabilidade foram descartadas.

As evidências científicas das previsões foram avaliadas como frequência de citações MEDLINE dos nomes e sinônimos de cada par de fármacos. A busca se deu na estrutura análoga a (((insulin AND aspart) OR (insulin AND detemir)) AND (budesonide OR desonide)).

O modelo implementado é mostrado no algoritmo 5.1.

## 5.5 Análise de dados

A decomposição SVD e a combinação das medidas de seno e cosseno foram fatores preponderantes nos experimentos com acurácia superior a 0,9. Logo, a extração de características latentes demanda diferentes abordagens para viabilizar aos classificadores e estabelecer modelos preditivos mais acurados.

A combinação das distâncias agregou poder preditivo aos classificadores, seguido do uso isolado da distância de cosseno e da distância euclidiana. A distância de cosseno mostrou-se superior devido ao predomínio de atributos na forma de texto [Tan et al., 2005].

A seleção de atributos para o melhor classificador "RandomCommittee" (tabela 5.5) resultou em 17 atributos, sendo advindos das distâncias de (I) cosseno ("*description*", "*drug reference*", "*generic name*", "*atc level 1*", "*atc level 3*", "*name atc level 1*", "*name atc level 3*"); (II) euclidiana ("*brand mixtures*", "*chemical structure*", "*atc level 2*", "*name atc level 5*") e (III) ambas ("*absorption*", "*organisms affected*", "*name atc level 2*").

Observou-se a participação dos atributos ATC/OMS em aproximadamente 50% dos atributos selecionados pelos modelos mais bem sucedidos. Absorção e organismos afetados estiveram presentes em todas as seleções de atributos. Nenhum atributo numérico foi selecionado pelo algoritmo *CfsSubsetEval* neste experimento.

Com validação cruzada, as técnicas de metaprendizado conquistaram melhor desempenho dentre os 53 classificadores em 946 experimentos, seguidos por árvores e *lazy*. Os desempenhos com acurácia superior a 0,9 são mostradas na tabela 5.4

### 5.5.1 Previsão de instâncias desconhecidas

As instâncias desconhecidas foram obtidas a partir da exploração do conjunto completo de hipóteses para $k = 2$. Foram previstas 54.816 interações (5,79% de $|Q|$). 51 combinações foram classificadas como interação grave, 12.369 como interação moderada, 62 como interação leve e 42.334 como combinações seguras. Apesar das classes $C$ atribuídas ao conjunto $A_k$ serem

**Algoritmo 5.1 Processos do modelo exaustivo de mineração de interações medicamentosas.** $D$ representa um conjunto de descritores de um conjunto de fármaco $f \in F$. $A$ é o conjunto completo de combinações possíveis de fármacos aos pares. $C$ é o conjunto de classes de acordo com o padrão ouro assinalado para as instâncias conhecidas $A_k$, dado $A_k = \{f_i, f_j | i \neq j\}$ e $A_k \subset A$. $N$ é a matriz de distâncias das combinações de fármacos $a \in A$, sendo $Q = A$. $Y$ é o conjunto de treino contendo as distâncias das combinações conhecidas. $\Delta$ é o conjunto de métricas de distâncias. $\Theta$ é o conjunto de variações usadas dos parâmetros do algoritmo *CfsSubsetEval* de seleção de atributos. $\Gamma$ são as técnicas de aprendizado de máquina utilizadas para a classificação com validação cruzada. $R$ é o resultado do desempenho calculado. $P$ é o conjunto de previsões de interações medicamentosas de acordo com o melhor modelo de mineração de dados.

**para** $i \leftarrow 1$ **até** $|D|$ **faça** ▷ Calcula as distâncias para cada atributo $D_x \subset D$.
  $aux \leftarrow D_i$;
  **se** $D_i$ é contínuo **então**
    $aux \leftarrow$ DISCRETIZA$(D_x)$;
  **fim se**
  $M_i \leftarrow$ MATRIZFREQUÊNCIA$(aux)$; ▷ Cada elemento $m \in M$ e $w \in W$ representa um fármaco no espaço n-dimensional.
  $W_i \leftarrow \text{SVD}(M_i)$;
  **para cada** $\delta \in \Delta$ **faça**
    **para** $x \leftarrow 1$ **até** $|F|$ **faça**
      **para** $y \leftarrow x + 1$ **até** $|F|$ **faça** ▷ Calcula a distância entre todos os fármacos. Cada $n \in N$ representa uma combinação $a \in A$.
        $N \Leftarrow \delta(m_x, m_y)$;
        $N \Leftarrow \delta(w_x, w_y)$;
      **fim para**
    **fim para**
  **fim para**
**fim para**
**para cada** $n \in N$ **faça** ▷ Obtém as instâncias de treino $A_k$.
  **se** $n(f_i, f_j)$ é uma interação conhecida $\in C$ **então**
    $Y_0 \Leftarrow (n, c)$
  **fim se**
**fim para**
**para** $i \leftarrow 1$ **até** $|\Theta|$ **faça** ▷ Realiza seleção de atributos.
  $Y_i \leftarrow \theta(Y_0)$;
**fim para**
**para cada** $\gamma \in \Gamma$ **faça** ▷ Treino.
  **para** $i \leftarrow 0$ **até** $|\Theta|$ **faça**
    $R \Leftarrow \gamma(Y_i)$;
  **fim para**
**fim para**
**retorna** $P \Leftarrow$ MELHORMODELO$(R, N, \Theta, \Delta)$

Tabela 5.4: **Desempenho dos classificadores adotados no modelo geral para previsão de interações farmacológicas.** Em todos os casos foi foi realizado SVD.

| Distância | Modelo | Classificador | Precisão | Kappa | EMA |
|---|---|---|---|---|---|
| ambas | meta | RandomCommittee | 0,9585 | 0,8707 | 0,0354 |
| ambas | trees | RandomForest | 0,9579 | 0,8683 | 0,0466 |
| cosseno | meta | RandomCommittee | 0,9561 | 0,8635 | 0,0368 |
| cosseno | trees | RandomForest | 0,9553 | 0,8598 | 0,0481 |
| euclidiana | meta | RandomCommittee | 0,9519 | 0,8511 | 0,0382 |
| euclidiana | trees | RandomForest | 0,9501 | 0,8443 | 0,0497 |
| ambas | meta | RotationForest | 0,9471 | 0,8307 | 0,0529 |
| ambas | lazy | IB1 | 0,9402 | 0,8243 | 0,0299 |
| ambas | lazy | IBk | 0,9401 | 0,8239 | 0,0300 |
| cosseno | meta | RotationForest | 0,9386 | 0,8018 | 0,0580 |
| cosseno | lazy | IB1 | 0,9299 | 0,7939 | 0,0350 |
| cosseno | lazy | IBk | 0,9298 | 0,7934 | 0,0351 |
| ambas | trees | RandomTree | 0,9291 | 0,7924 | 0,0354 |
| cosseno | trees | RandomTree | 0,9282 | 0,7908 | 0,0358 |
| ambas | trees | J48graft | 0,9197 | 0,7483 | 0,0498 |
| ambas | meta | OrdinalClassClassifier | 0,9143 | 0,7344 | 0,0555 |
| ambas | meta | ND | 0,9130 | 0,7331 | 0,0548 |
| ambas | meta | ND,DataNearBalancedND | 0,9122 | 0,7329 | 0,0539 |
| ambas | meta | ND,ClassBalancedND | 0,9113 | 0,7289 | 0,0551 |

ND: NestedDichotomies. EMA: Erro médio absoluto

Tabela 5.5: **Desempenho do classificador RandomCommittee** com 50 iterações, 2 sementes, seleção de atributos e SVD em matriz de combinações de fármacos com distância euclidiana e cosseno.

| Conhecidos\Previstos | Maior | Moderada | Menor | Segura | conhecidos | Representatividade |
|---|---|---|---|---|---|---|
| **Maior** | 1890 | 382 | 5 | 0 | 2277 | 12,42% |
| **Moderada** | 56 | 14567 | 23 | 3 | 14649 | 79,87% |
| **Menor** | 7 | 249 | 960 | 3 | 1219 | 6,65% |
| **Segura** | 0 | 26 | 1 | 168 | 195 | 1,06% |
| **Previstos** | 1953 | 15224 | 989 | 174 | 18340 | 100,00% |
| | | | | | **Média simples** | **Média ponderada** |
| **Taxa de FP** | 0,0039 | 0,1771 | 0,0017 | 0,0003 | 0,0458 | 0,1475 |
| **Taxa de FN** | 0,1700 | 0,0056 | 0,2114 | 0,1378 | 0,1312 | 0,0355 |
| **Acurácia** | 0,9755 | 0,9597 | 0,9843 | 0,9982 | 0,9794 | 0,9631 |
| **Sensibilidade** | 0,8300 | 0,9944 | 0,7886 | 0,8622 | 0,8688 | 0,9645 |
| **Especificidade** | 0,9764 | 0,9738 | 0,9851 | 0,9985 | 0,9835 | 0,9750 |
| **Precisão** | 0,9677 | 0,9568 | 0,9709 | 0,9657 | 0,9653 | 0,9588 |
| **métrica F** | 0,9759 | 0,9709 | 0,9759 | 0,9694 | 0,9730 | 0,9717 |
| **curva ROC** | 0,9840 | 0,9850 | 0,9810 | 0,9730 | 0,9808 | 0,9846 |

FP: Falso Positivo. FN: Falso Negativo.

desbalanceadas, a proporção das previsões seguiram o comportamento esperado, apontando massivamente (77,4%) para combinações seguras.

A coleta da base de dados DrugBank resultou em 9.324 interações, porém somente 1.574 corresponderam a interações sugeridas pelo Drugs.com (439 maiores, 1.116 moderadas e 19 menores) e 4 ao grupo de combinações seguras. A disparidade confirma a ausência reportada por Coloma et al. [2013] de uma lista definitiva de interações medicamentosas.

Embora a conciliação com a base DrugBank apenas tenha confirmado 12,4% das interações medicamentosas potenciais adversas do Drugs.com, adotou-se esta base devido à maior cobertura, necessária para a verificação de interações totalmente desconhecidas. Desta forma, optou-se por abranger grande número de instâncias em detrimento da determinação definitiva de dada interação. A isenção em relação às bases do domínio do conhecimento, ou seja, o não uso da base DrugBank tanto na construção das matrizes $N$ de entrada nos modelos quanto no padrão ouro, remete a um benefício adicional ao evitar-se *overffiting*, ou seja, alguma sobreposição dos dados originais que enviesem a formação dos dados de treino de modo a prejudicar previsões na base de testes.

As interações medicamentosas previstas corresponderam a 148 reportadas no DrugBank (146 classificadas como moderada e 2 como leve). No entanto 766 combinações previstas como seguras são reportadas pelo DrugBank como interações.

Embora tenha sido usado o termo "combinação segura", os pares envolvidos podem consistir em "interação terapêutica". Por exemplo, a inibição que clavulanato causada na enzima betalactamase produzida por algumas bactérias, reduz a degradação do antibiótico, elevando sua efetividade. Trata-se de uma interação farmacocinética sinérgica. Desta forma, pode-se esperar atuações não inertes de pares de fármacos entre si, o que foi captado pelo algoritmo.

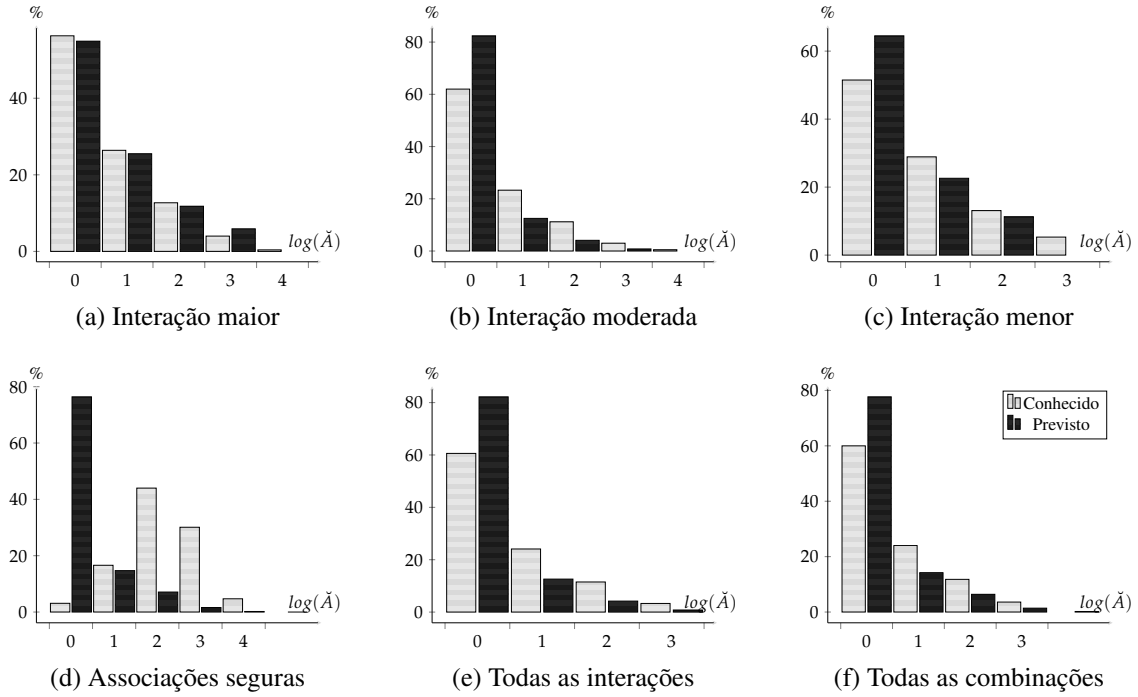
Embora a representatividade dos falso negativos tenha sido baixa (1,82%), sua existência demonstra a necessidade de mais informações preditivas acerca de combinações seguras ou terapêuticas. Uma interpretação alternativa é de que o uso de fármacos com o mesmo código ATC fornece insumo para detecção de interações medicamentosas superior à premissa da segurança. Esta afirmação é baseada na força preditiva das variáveis ATC para o modelo, as quais corresponderam até 50% das variáveis selecionadas. Somando as previões de interação medicamentosa adversa com as realizadas com base em pares ATC/OMS, o modelo foi capaz de detectar 11,8% das interações exclusivas do DrugBank.

53 fármacos participaram das combinações previstas como graves, sendo 24 para o sistema nervoso, 11 para o sistema cardiovascular e 4 para o sistema músculoesquelético. Cloreto de potássio esteve presente em 86,56% das previsões graves. Ulobetasol (corticosteroide), fulvestrant (agente antineoplásico), metilaminolevulinato (tratamento de queratoses), lubiprostona (agente antineoplásico), e terlipressina (hormônio hipofisário) não possuem qualquer menção de interação medicamentosa no DrugBank e no Drugs.com, indicando que a ferramenta pode ser usada para explorar fármacos com interação medicamentosa desconhecida.

#### 5.5.1.1 Amostragem

A comparação dos casos previstos e conhecidos em função de citações MEDLINE é mostrada na figura 5.2.

Pelo menos uma citação foi encontrada em 29,6% do conjunto $A_p$ e em apenas 51,6% das combinações conhecidas $A_k$. Os valores encontrados para interações previstas são justificáveis devido à possível falta de estudos científicos. No entanto, o volume de interações conhecidas




(a) Interação maior (b) Interação moderada (c) Interação menor

(d) Associações seguras (e) Todas as interações (f) Todas as combinações

Figura 5.2: **Densidade de citações MEDLINE para 18.340 pares de interações conhecidas e 54.816 previstas, incluindo combinações seguras.** O número de citações em escala logarítmica expresso nas abscissas foram obtidos com busca (($name_{a1}$ OR $name_{a2}$ OR ... OR $name_{an}$) AND ($name_{b1}$ OR $name_{b2}$ OR ... OR $name_{bm}$)). As coordenadas expressam o percentual de pares de fármacos.

que não retornaram citações sugere que outras bases como EMBASE e LILACS devem ser usadas neste tipo de avaliação e em estudos preditivos.

Realizou-se busca manual aleatória de casos sem citação no MEDLINE. A interação entre fluconazol e clozapina retornou 83 citações na busca EMBASE. Os compêndios de Baxter [2008], Jacomini & da Silva [2011] e Tatro [2012], não relatam esta interação. Porém o serviço web micromex [Micromedex, 2013] a relata como interação grave. A identificação sugere que os estudos não podem limitar-se a apenas uma fonte interações medicamentosas. Os achados a partir de estudos de mineração de textos científicos tornam-se bastante restritos ao ater-se apenas ao MEDLINE.

Especula-se que este viés de publicação, para determinados fármacos, provavelmente se deve a orientação econômica, regulatória e científica ou da ocorrência da informação apenas no dossiê técnico que culmina na bula. Maiores investigações são demandadas para delinear o viés de publicação das interações medicamentosas.

As interações entre galantamina ou seleginina associadas a oxibato de sódio foram previstas e estão ausentes dos compêndios mencionados. O primeiro fármaco é usado no tratamento de Alzheimer moderado, enquanto o segundo é usado para tratamento inicial de doença de Parkinson. O oxibato é um anestésico com uso *off label*[2] para depressão, insônia, narco-

[2]Sem indicação autorizada.

lepsia e alcoolismo, cujo uso foi mencionado dentre os 38 resumos encontrados para as duas combinações. A interação é passível de ocorrer devido a modificação no sistema de inibição da monoaminoxidase com consequente redução de dopamina, visto que o oxibato promove a ativação dos receptores do ácido gamaminobutírico.

Outros exemplos de interações graves e moderadas previstas são entre clorpromazina e propranolol, insulina e folinato de cálcio, insulina e norgestimato, cloranfenicol e amoxicilina, imunoglobulina e tolbutamida, insulina e levotiroxina, bleomicina e metotrexato, propranolol e salbutamol e clonidina e metildopa.

### 5.5.2 Comparação com outros estudos

Dada a heterogeneidade dos estudos, a comparação direta muitas vezes é dificultada devido aos objetivos que implicam na escolha de diferentes parâmetros de desempenho.

Segura-Bedmar et al. [2011b], Duke et al. [2012], Percha et al. [2012] e Zhang et al. [2012b] usaram processamento de linguagem natural para extração de interações medicamentosas em bases biomédicas, incluindo a base de resumos MEDLINE.

O desempenho está relacionado à dificuldade na extração dos termos e na especifidade da informação desejada. Os trabalhos de Duke et al. [2012] e Gottlieb et al. [2012] são exemplos que focaram nas interações relacionadas aos processos de indução ou inibição enzimática mediados pelo sistema citocromo. Neste sentido, a delimitação do escopo aumenta a especifidade da informação e reduz a cobertura dos casos. As fontes de dados (populacional ou farmacológica) influem na correspondência do acerto. Logo, níveis inferiores em bases populacionais podem apresentar caráter menos especulativo do que abordagens farmacológicas. Cada comparação entre estudos pode exigir diferentes abordagens para definir-se qual conquistou melhor correspondência com os objetivos pretendidos. Diante destas ressalvas, as características dos estudos correlatos encontram-se sumarizadas na Tabela 5.6.

A representação semântica na forma de grafos, realizada por Percha et al. [2012] e Zhang et al. [2012b], define novas arestas entre pares de fármacos com a extração de regras ou padrões específicos entre os fármacos e outra entidade biológica. O método proposto não define intermediários biológicos, sendo difuso neste aspecto, porém, informações são extraídas quanto a representatividade de determinado atributo para a extração de interações. A derivação destes atributos enquanto entidade possibilita a verificação da contribuição preditiva e mecanicística.

A extração de interações enquanto exploração de todas as possibilidades dos pares de fármacos foi descrita por Segura-Bedmar et al. [2011b] como um problema combinatorial. No entanto, Gottlieb et al. [2012] relatou explorar todo o conjunto de possibilidades, ainda assim, restrito a 687 fármacos. Possivelmente, esta restrição dos demais estudos se deva ao custo das tradicionais técnicas de extração de semântica que demandam um conjunto substancial de sentenças manualmente acuradas ou métricas para cada atributo e, de modo geral, estão limitadas à semântica explícita em textos com ampla variação temática. A abordagem proposta

Tabela 5.6: **Comparação entre estudos** de previsão computacional de interações farmacológicas.

| **Estudo** | **Modelo** | $\|\mathbf{F}\|$ | $\|\mathbf{A}\|$ | $\|\mathbf{A_k}\|$ | **Fonte de $\|\mathbf{A_k}\|$** | **Cobertura** |
|---|---|---|---|---|---|---|
| Segura-Bedmar et al. [2011a] | SVM | 3.313 | 30.757 | 3.160 | DrugBank | 52.1% |
| Duke et al. [2012] | Regras de combinação | 232 | 13.197 | 196 | MEDLINE | 62.8% |
| Gottlieb et al. [2012] | Regressão logística | 671 | 37,212 | | DrugBank, Drugs.com | 93.0% |
| Percha et al. [2012] | RandomForest, regressão logística, SVM | 2.910 | 10.000 | 5.000 | DrugBank, drug lexicon | 79.3% |
| Zhang et al. [2012a] | *Graph pairwise sigle kernel* | 625 | 30.583 | 756 | *DDI Extraction challenge* | 67.2% |
| Modelo proposto | RamdomCommitee | 1.390 | 965.355 | 18.300 | Drugs.com | 95.0% |

$F$: fármacos. $A$: combinações. $A_k$: combinações conhecidas.

extrai a semântica implícita em atributos com elevada densidade semântica, isto é, com escopo bem delimitado.

Esta abordagem demonstrou que é possível identificar interações medicamentosas desconhecidas mediante as verificação de padrões de distância entre fármacos com consequente extração das características intrínsecas das combinações.

## 5.6 Sumário

O modelo proposto foi capaz de representar uma realidade de forma fidedigna ao padrão ouro. O número inferior de citações MEDLINE nas interações previstas em relação às conhecidas demonstra o caráter de descoberta de conhecimento em explorar combinações ainda não abordadas pela literatura científica. Devido aos elevados valores de concordância com o padrão-ouro, o modelo possui potencial para representar interações altamente corroboradas, podendo ser utilizado como alerta juntamente a uma ferramenta de auxílio a prescrição ou dispensação. Ainda, os resultados sugerem uma exploração especulativa do caráter de reações subnotificadas, ou seja, sem corroboração por populações, demonstrando o potencial destas informações para aplicação na área de saúde pública e destinação de novos estudos acadêmicos.

# Capítulo 6

# A utilização de previsões farmacológicas em estudos farmacoepidemiológicos

A abordagem da mineração farmacológica, mostrada no capítulo 5, oferece um recurso particularmente útil para o desenvolvimento e o conhecimento de fármacos. Estudos *in vitro* ou *in vivo* podem alimentar-se da exploração comparativa dos fármacos com interações desconhecidas em relação aos demais. Trata-se de uma mineração predominantemente prospectiva.

A interação medicamentosa que não envolve mecanismos tradicionais farmacocinéticos, torna-se um fenômeno de difícil detecção. Seja em estudos controlados ou focado em dados históricos, a verificação de combinações previstas como interação traz um alerta, sobretudo quanto a possibilidade de sub-notificação das ocorrências desse evento. Desta forma, deve ser demonstrada a utilidade das previsões quanto ao perfil de utilização por populações.

Os objetivos deste capítulo são (I) caracterizar a relevância das previsões de interação na avaliação de bases populacionais com representatividade estabelecida a partir da abordagem de mineração farmacológica proposta e (II) avaliar o compromisso das previsões em relação ao padrão ouro e outras fontes consultadas.

Para isso, o capítulo está organizado em três seções. Na primeira, são descritos o desenho do estudo e os procedimentos metodológicos adotados. As características das bases populacionais são sumarizadas, assim como os critérios de inclusão de pacientes e o cálculo da prevalência do uso de medicamentos e a classificação do padrão-ouro de interações medicamentosas e das previsões. Ademais, é caracterizada a representatividade da classificação adotada por duas bases de interações medicamentosas conhecidas, sendo Drugs.com o padrão-ouro usado para o aprendizado de máquina e a segunda a base DrugBank. A interseção e a união dessas bases foram utilizadas para estabelecer a evidência do ponto de vista de especificidade e cobertura de interações medicamentosas. Desta forma são comparadas interações duplamente qualificadas e aquelas catalogadas em pelo menos uma fonte.

Na segunda seção os resultados obtidos com a abordagem proposta são apresentados. As interações medicamentosas previstas são posicionadas frente a estas quatro fontes de comparação. Na sequência é realizada comparação das interações medicamentosas quanto aos grupos anatômicos. A classificação anatômica sugere quais sistemas são demandados e quais podem sofrer com os impactos das interações medicamentosas potenciais. Concluindo a caracterização do uso, o estudo das combinações mais prevalentes do padrão-ouro, interseção e previsão é aprofundado ao nível químico. Desta forma é estabelecida uma perspectiva do impacto da prescrição ou dispensação em função da prevalência observada.

Ainda na seção de resultados são avaliados pares amostrados de previsões das combinações utilizadas em relação a três compêndios e à literatura científica. Finalmente, é realizada a verificação desses pares com previsões realizadas por outra ferramenta desenvolvida por Gottlieb et al. [2012].

A terceira seção encerra o capítulo com a discussão dos resultados obtidos.

## 6.1 Métodos

### 6.1.1 Desenho do estudo

Para identificar e caracterizar as combinações medicamentosas previstas duas populações foram avaliadas. A população descrita na base ELSA [Aquino et al., 2012] foi observada em profundidade por um curto período com a intenção de coletar dados sobre todos os medicamentos utilizados. A segunda população registrada na base SIGAF [Guerra Júnior et al., 2008] apresenta um conjunto mais restrito de fármacos padronizados por se tratar de uma base administrativa. Contudo, a base SIGAF apresenta um grande número de observações e amplo intervalo de tempo.

**Base ELSA** O Estudo Longitudinal de Saúde do Adulto representa o maior estudo epidemiológico da América Latina [ELSA, 2009]. Objetiva contribuir com informação relevante concernente ao desenvolvimento e progressão clínica e subclínica de doenças crônicas, em particular, doenças cardiovasculares e diabetes. Foram coletadas variáveis socioeconômicas, 17 variáveis relativas à prescrição, 19 sobre o uso de medicamentos, 799 variáveis de fármacos derivadas do uso recente de medicamentos de uso contínuo e 51 variáveis laboratoriais. Foram disponibilizados 15.105 registros contendo 49.713 indicações de uso de medicamento.

**Base SIGAF** O Sistema Integrado de Gerenciamento da Assistência Farmacêutica foi desenvolvido pela Superintendência de Assistência Farmacêutica da Secretaria de Saúde do Estado de Minas Gerais. Este sistema integra as unidades de saúde e abriga a gestão de insumos farmacêuticos incluindo a dispensação. A base fornecida consistiu em 7.103.636 registros de dispensações para 544.120 pacientes entre abril de 2010 a fevereiro de 2013.

#### 6.1.1.1 Critério de combinação e simultaneidade do tratamento

A base ELSA foi disponibilizada em tabela única contendo um paciente em cada tupla e uma coluna para cada atributo. Tanto os dados clínicos, quanto os registros dicotômicos relativos ao consumo do medicamento antecedem no máximo duas semanas à entrevista. Logo, os medicamentos assinalados ao paciente foram considerados como concomitantemente usados, extraindo-se combinações aos pares.

A simultaneidade de tratamentos medicamentosos da base SIGAF foi traçada a partir do código fornecido como identificador do paciente e combinações de medicamentos cujo intervalo entre as datas de dispensação foi de até quinze dias, adotando-se o mesmo critério de continuidade da base ELSA.

Os fármacos foram identificados quanto ao drugcard do DrugBank, desmembrando-se a apresentação na forma de combinação conforme o número de substâncias ativas presentes.

#### 6.1.1.2 Critérios de elegibilidade

Foram selecionados pacientes em que pelo menos um fármaco foi identificado quanto ao drugcard.

### 6.1.2 Taxonomia das combinações

A notação completa está presente na lista de símbolos e notações na página xxv.

Foram utilizadas as interações conhecidas $A_k$ caracterizadas na tabela 5.2 para estabelecer a relevância do padrão ouro, bem como para caracterizar as populações.

Os domínios de $A_k$ estabelecidos para combinações inertes (0) e interações menores, moderadas ou maiores ($-1$, $-2$ e $-3$, respectivamente) foram $A_c = \{a_c^{1-}, a_c^{2-}, a_c^{3-}\}$ para interações coletadas a partir do sítio Drugs.com, $A_b = \{a_b^0, a_b^{1-}\}$ para o sítio DrugBank, $A_a = \{a_a^0\}$ para ATC/OMS.

A evidenciação da interação medicamentosa foi graduada em dois níveis, derivando-se um grupo advindo da interseção entre DrugBank e Drugs.com e outro a partir da união. Os conjuntos derivados da intersecção e união foram denotados como $A_e = \{a_e^0, a_e^{1-}, a_e^{3-}\}$ e $A_u = \{a_u^0, a_u^{1-}, a_u^{3-}\}$, respectivamente. A definição da interação classificada como "maior" ($-3$) pelo Drugs.com foi mantida nestes grupos, sendo as demais interações indicadas com $-1$.

As combinações seguras e interações previstas farmacologicamente $A_p$ foram denotadas como $A_p = \{a_p^0, a_p^{1-}, a_p^{2-}, a_p^{3-}\}$.

Embora o grupo inicialmente caracterizado como inerte possa apresentar potencial sinérgico, ou ao menos, capacidade de interação, manteve-se a nomenclatura 0 ao invés de usar-se o sinal $+$ destinado ao sinergismo. As informações narradas a cerca deste grupo são postas parcimoniosamente devido ao caráter especulativo do argumento em se usar combinações comumente usadas no mercado ou sob a mesma classificação terapêutica no nível químico.

### 6.1.3 Prevalência das combinações

A prevalência de cada classificação foi calculada com a razão dos pacientes que foram expostos a pelo menos uma combinação pertencente ao grupo em relação ao total de pacientes que fizeram uso de um ou mais medicamentos (equação 6.1).

As prevalência para cada combinação foram calculada da mesma forma.

$$\text{Prevalência} = \frac{\text{pacientes expostos}}{\text{total de pacientes}} = \frac{|T|}{|S|} \tag{6.1}$$

$S$ corresponde a um conjunto de usuários de fármacos e $T$ a usuários de polifarmácia.

### 6.1.4 Citações

As citações MEDLINE intuem um grau de evidência peculiar por acrescentar resultados de fronteira do domínio *in vitro* aos demais que corroboram evidência de interações.

Os nomes genéricos de cada par da amostra foram associados com sinônimos e termos correlatos à "interação medicamentosa" conforme exemplificado na estratégia de busca abaixo.

> ***Prednisolone****[Title/Abstract] AND* ***Salbutamol****[Title/Abstract] AND ( "Drug Interactions" OR "Drug Interaction" OR "Interaction, Drug" OR "Interactions, Drug" OR "Previous Indexing" OR "Drug Antagonism" OR "Drug Synergism" OR "Drug Agonism" OR "Drug Partial Agonism" OR "Agonism, Drug Partial" OR "Partial Agonism, Drug" OR "Drug Agonism, Partial" OR "Agonism, Partial Drug" OR "Partial Drug Agonism" OR "Drug Antagonism" OR "Antagonism, Drug" OR "Antagonisms, Drug" OR "Drug Antagonisms"OR "Drug Inverse Agonism" OR "Agonism, Drug Inverse" OR "Inverse Agonism, Drug"OR "Drug Synergism" OR "Drug Synergisms" OR "Synergism, Drug" OR "Synergisms, Drug" OR "Drug Potentiation" OR "Drug Potentiations" OR "Potentiation, Drug" OR "Potentiations, Drug")*

### 6.1.5 Análise de dados

Análise descritiva dos dados foi conduzida com a apresentação em tabelas da distribuição de frequências (relativas e absolutas) das variáveis selecionadas, utilizando-se MySQL 5.5.310 ubuntu 0.12.04.2 [Widenius et al., 2002] e R versão 3.0.2 [R Core Team, 2013] como ferramentas de manipulação dos dados e análises estatísticas. Os intervalos de confiança das médias foram estimados com a distribuição $t-$ student com $\alpha = 0,05$.

confidencialidade

## 6.2 Resultados

### 6.2.1 Perfil de utilização de medicamentos e combinações

#### 6.2.1.1 Aspectos gerais

Dentre $U = 15.005$ pacientes da base ELSA, $S = 8.890$ (59,2%) foram associados aos fármacos indexados segundo o DrugBank. Foram coletados 5,5 relatos de uso de fármacos por paciente, sendo a mediana igual a 4 fármacos, $Q_1 = 2$ e $Q_3 = 8$.

A base SIGAF apresentou $U = 544.120$ pacientes, dentre os quais $S = 542.415$ (99,67%) foram associados aos medicamentos indexados, sendo 13,4 substâncias distintas por paciente no período estudado (1.142 dias entre a primeira e a última dispensação), com mediana igual a 5, $Q_1 = 2$ e $Q_3 = 16$.

A polifarmácia, uso de dois ou mais medicamentos, foi verificada em 67,1% e 69,2% dos usuários de medicamentos das bases ELSA e SIGAF, respectivamente, sob o mesmo critério de simultaneidade de quinze dias.

A partir do conjunto de fármacos $F = 1.660$ abrangidos pelas combinações drugs.com, DrugBank, ATC e previsas; foram relatados 502 na base ELSA e 409 foram dispensados aos usuários contemplados na base SIGAF. Esta diferença reflete a diversidade de medicamentos oriundos da base ELSA, visto que as informações foram coletadas diretamente com os pacientes. Desta forma, são contemplados medicamentos vendidos comercialmente não incluídos em listas padronizadas.

Os pacientes da base ELSA associaram $H = 494$ fármacos (29,8% dos fármacos de combinações classificadas) e $H = 402$ foram associados (24,6%) pelos pacientes SIGAF.

Na base ELSA identificou-se $B = 11.014$ combinações distintas, sendo $B_k = 1.314$ (11,9%) classificadas. Na base SIGAF foram classificadas $B_k = 3.091$ (14,6%) combinações dentre $B = 21.108$ observadas. As proporções indicam que o número de combinações tende a aumentar com a sequência de observações. Possivelmente a disparidade seria ainda maior se o espectro de fármacos da base SIGAF fosse o mesmo da base ELSA.

#### 6.2.1.2 Classificação de combinações

O perfil da classificação de combinações é apresentado nas tabelas 6.1 e 6.2 para as bases ELSA e SIGAF respectivamente.

As tabelas do perfil contemplam quatro secções verticais cinco e horizontais. Cada seção horizontal indica respectivamente a classificação drugs.com, DrugBank, a interseção entre Drugs.com e DrugBank, a união e, finalmente, as previsões baseadas no modelo farmacológico do capítulo 5. A primeira seção vertical, indica o montante de fármacos $F_k$ e combinações classificadas (conhecidas e previstas) $A_k$, bem como a relação $\breve{A} \div A$ de citações MEDLINE para as combinações do conjunto. A prevalência é assinalada em seguida pela quantidade de usuários das combinações classificadas por cem usuários de medicamentos. No terceiro extrato

consta o universo de combinações projetado com base nos fármacos usados, ou seja, a quantidade de combinações que participariam deste grupo caso todos os fármacos fossem associados. O último nível vertical indica os fármacos derivados das combinações utilizadas, bem como as combinações e citações MEDLINE.

Conforme verificado nas tabelas 6.1 e 6.2, relativo a $|G_k| \div |F_k|$, cerca de 49% e 44% dos fármacos com interações maiores duplamente conhecidas $A_e^{3-}$ foram utilizados pelas populações ELSA e SIGAF, respectivamente. Esta tendência refletiu-se nas demais classificações, variando de 26,8% a 51,8%. As previsões abrangeram 4,7% dos pacientes da base ELSA e 1,5% dos pacientes da base SIGAF. Contudo, 6,5% e 25,5% apresentaram utilização para as interações do grupo previsto como inerte ou sinérgico para as bases ELSA e SIGAF respectivamente.

Conforme observado na relação $|B_k| \div |A_k|$, as combinações consideradas maiores e duplamente documentadas $B_e^{3-}$ foram 3,4% (ELSA) e 12,1% (SIGAF) dentre as conhecidas. Observando-se as interações previstas, nenhuma interação grave foi verificada na base ELSA, porém 1,6% das combinações foram assinaladas como moderadas (194) e 5,0% como leves. Na base SIGAF observou-se 11,8% das interações maiores, 3,2% das moderadas e 5,0% das leves previstas.

Nas tabelas 6.1 e 6.2, os estratos horizontais $A_e$ e $A_u$ relativos a interseção e união respectivamente mostram os extremos da variação que a quantificação de interações medicamentosas baseadas em compêndios pode assumir. A prevalência de interações maiores $\{3^-\}$ na base ELSA variou de 0,3% a 3,3% considerando-se o primeiro valor como evidência corroborada ou o segundo com alguma documentação. Esta variação foi atenuada para 1,1% a 2,1% com o maior número de observações vistas na base SIGAF. Considerando todas as interações, i.e., $A^-$, a proporção da prevalência das interações medicamentosas com algum indício de risco $A_u^-$ em relação a interações com maior nível de evidência foi de $11:1$ na base ELSA, porém, foi de cerca de $3:1$ na base SIGAF.

Ao comparar o número de citações por interações medicamentosas, todas as classes da base Drugs.com apresentaram, mais que o dobro das interações verificadas no DrugBank, mostrando um maior alinhamento com o veículo MEDLINE. Esta tendência refletiu-se em ambas as base ELSA e SIGAF.

Em ambas as populações quase a totalidade dos fármacos usados foram associados. Porém, as combinações usadas $B_k$ concentraram-se em menor número de fármacos quando comparadas às combinações conhecidas dentre os fármacos utilizados $V_k$ em todas as classificações avaliadas, conforme visto nas respectivas colunas das tabelas 6.1 e 6.2. Verificou-se na base ELSA a relação $V \div B$ variou de $5:1$ para os casos leves ou inertes a $13:1$ para os casos inertes previstos, enquanto a maior quantidade de observações da base SIGAF nivelou esse número para $2:1$ a $4:1$. Dentre as proporções das interações potenciais adversas, observou-se que as conhecidas variaram de $7:1$ a $9:1$ enquanto as previstas variaram de $6:1$ a $9:1$, demonstrando que a proporção da presença das interações distintas previstas é semelhante àquelas

observadas para as conhecidas. A razão média entre as 30 relações $B_k \div A_k$ (15 para cada base, sendo 3 para Drugs.com, interseção e união; 2 para o DrugBank e 4 para as previsões) $B$ foi de 13,4% com intervalo de confiança $IC_{(95\%)} = [12,2;14,6]$ para a base ELSA e 39,8% com $IC_{(95\%)} = [33,8;45,8]$ para a base SIGAF.

A concentração das combinações observadas em relação às possíveis reflete a dificuldade em se determinar interações medicamentosas adotando-se apenas fontes populacionais. Previsões *in silico* ou a avaliação das combinações previstas com base em uso por populações, estão limitadas aos números de fármacos e de observações. Ressalta-se que para obter as interações previstas, foram avaliados quase um milhão de pares de fármacos.

Avaliando-se ambas as bases conjuntamente, foram classificados $H = 464$ fármacos associados (28,0%, $n = 1660$). Este número variou em 198 a 362 considerando combinações duplamente documentadas ou advindas da união entre Drugs.com e DrugBank. Nesta mesma ordem, identificaram-se 127 a 333 fármacos com interações medicamentosas adversas. Dentre as combinações previstas, identificou-se 394 fármacos, sendo 294 relativos a interações medicamentosas adversas.

Em 4.061 combinações verificadas nas populações, 223 (5,5%) a 2.317 (57,1%) podem ser consideradas como interação medicamentosa adversa segundo a interseção e a união das fontes consultadas. Esta disparidade reflete a não convergência dos compêndios e a limitada aplicabilidade no contexto clínico. O uso de uma base ampla como o Drugs.com ou mais restritiva como o DrugBank deve ser atrelado às informações de grupos específicos de pacientes. As previsões abrangeram $A_p^- = 561$ (13,8%) interações medicamentosas adversas observadas nas populações.

De modo geral, a razão entre citações MEDLINE e combinações apresentou valores superiores entre combinações consideradas seguras $\breve{A}^0$ em relação às interações medicamentosas potenciais adversas $\breve{A}^-$. O intervalo de confiança do primeiro grupo (n=10) foi $IC_{(95\%)} = [649,31;1728,49]$ enquanto o segundo (n=20) apresentou $IC_{(95\%)} = [85,3731;178,227]$. Esta diferença reflete a preponderância de combinações seguras ou menores em relação às interações medicamentosas, refletindo a o padrão da busca geral observada na figura 5.2.

Nas populações, verificou-se que as interações potenciais consideradas maiores, $B^{3-}$, variaram de 15 a 149 considerando interseção e união das fontes consultadas. As razões de citações MEDLINE para as quatro intersecções de interações medicamentosas adversas tiveram $IC_{(95\%)} = [47,0924;174,708]$, a união apresentou variação $IC_{(95\%)} = [134,446;274,154]$ ($p = 0,0690$ em teste t pareado bicaudado). Estes resultados frustram a expectativa de que interações duplamente documentadas recuperariam mais citações MEDLINE.

A razão de citações MEDLINE em relação as combinações usadas por populações acompanharam, de modo geral, as tendências observadas em relação ao total de combinações conhecidas. Conforme evidenciado na figura 5.2 as combinações seguras e interações menores possuíram os menores índices e nenhuma citação recuperada para os pares consultados.

A tendência mais marcante foi a verificação de poucas citações recuperadas quanto aos

Tabela 6.1: **Representatividade e prevalência de combinações conhecidas e previstas de fármacos na base ELSA**, Estudo Longitudinal de Saúde do Adulto, 2013.

| | Classificação | | | Prevalência | Fármacos usados e combinações derivadas | | | | | Associações usadas e fármacos derivados | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A** | $\lvert\mathbf{F_k}\rvert$ | $\lvert\mathbf{A_k}\rvert$ | $\frac{\lvert\check{\mathbf{A}}_\mathbf{k}\rvert}{\lvert\mathbf{A_k}\rvert}$ | $\frac{\lvert\mathbf{T_k}\rvert}{\lvert\mathbf{S}\rvert}$ | $\lvert\mathbf{G_k}\rvert$ | $\frac{\lvert\mathbf{G_k}\rvert}{\lvert\mathbf{F_k}\rvert}$ | $\lvert\mathbf{V_k}\rvert$ | $\frac{\lvert\mathbf{B_k}\rvert}{\lvert\mathbf{A_k}\rvert}$ | $\frac{\lvert\check{\mathbf{V}}_\mathbf{k}\rvert}{\lvert\mathbf{V_k}\rvert}$ | $\lvert\mathbf{H_k}\rvert$ | $\frac{\lvert\mathbf{H_k}\rvert}{\lvert\mathbf{F_k}\rvert}$ | $\lvert\mathbf{B_k}\rvert$ | $\frac{\lvert\mathbf{B_k}\rvert}{\lvert\mathbf{A_k}\rvert}$ | $\frac{\lvert\check{\mathbf{B}}_\mathbf{k}\rvert}{\lvert\mathbf{B_k}\rvert}$ |
| $A_c^{1-}$ | 407 | 1.218 | 196,5 | 5,6% | 210 | 51,6% | 494 | 40,6% | 389,3 | 210 | 51,6% | 91 | 7,5% | 976,0 |
| $A_c^{2-}$ | 819 | 14.659 | 75,6 | 17,4% | 316 | 38,6% | 3.641 | 24,8% | 113,4 | 316 | 38,6% | 456 | 3,1% | 229,5 |
| $A_c^{3-}$ | 561 | 2.246 | 150,6 | 3,3% | 237 | 42,2% | 433 | 19,3% | 93,0 | 237 | 42,2% | 58 | 2,6% | 179,7 |
| $A_b^{0}$ | 342 | 687 | 497,6 | 3,5% | 176 | 51,5% | 227 | 33,0% | 1.236,8 | 174 | 50,9% | 45 | 6,6% | 2.093,8 |
| $A_b^{1-}$ | 1.220 | 12.786 | 17,0 | 8,6% | 377 | 30,9% | 2.158 | 16,9% | 41,9 | 377 | 30,9% | 245 | 1,9% | 42,1 |
| $A_e^{0}$ | 342 | 687 | 497,6 | 3,5% | 176 | 51,5% | 227 | 33,0% | 1.236,8 | 174 | 50,9% | 45 | 6,6% | 2.093,8 |
| $A_e^{1-}$ | 412 | 1.144 | 100,2 | 3,3% | 196 | 47,6% | 354 | 30,9% | 173,1 | 196 | 47,6% | 53 | 4,6% | 145,5 |
| $A_e^{3-}$ | 250 | 439 | 68,9 | 0,3% | 122 | 48,8% | 142 | 32,3% | 60,3 | 122 | 48,8% | 15 | 3,4% | 55,8 |
| $A_u^{0}$ | 396 | 802 | 641,4 | 3,6% | 205 | 51,8% | 286 | 35,7% | 1.325,2 | 203 | 51,3% | 52 | 6,5% | 2.248,0 |
| $A_u^{1-}$ | 1.319 | 27.077 | 52,5 | 22,3% | 406 | 30,8% | 5.797 | 21,4% | 108,0 | 406 | 30,8% | 724 | 2,7% | 269,7 |
| $A_u^{3-}$ | 561 | 2.246 | 150,6 | 3,3% | 237 | 42,2% | 433 | 19,3% | 93,0 | 237 | 42,2% | 58 | 2,6% | 179,7 |
| $A_p^{0}$ | 1.244 | 41.549 | 103,7 | 6,5% | 373 | 30,0% | 3.710 | 8,9% | 118,3 | 372 | 29,9% | 283 | 0,7% | 81,1 |
| $A_p^{1-}$ | 72 | 60 | 17,5 | 0,1% | 33 | 45,8% | 17 | 28,3% | 2,1 | 33 | 45,8% | 3 | 5,0% | 2,3 |
| $A_p^{2-}$ | 1.344 | 12.217 | 23,4 | 4,7% | 432 | 32,1% | 1.834 | 15,0% | 21,5 | 432 | 32,1% | 194 | 1,6% | 25,2 |
| $A_p^{3-}$ | 53 | 51 | 126,7 | 0,0% | 22 | 41,5% | 19 | 37,3% | 300,0 | 22 | 41,5% | 0 | 0,0% | 0,0 |

**Linhas**. $\mathbf{A_c}$ Drugs.com, 0 são combinações seguras sob o mesmo código da classificação ATC/OMS e $\{1-, 2-, 3-\}$ são interações medicamentosas adversas menores, moderadas e maiores, respectivamente. $\mathbf{A_b}$ DrugBank, 0 indica combinações seguras e $1-$ interações medicamentosas adversas. $\mathbf{A_e}$, $\mathbf{A_u}$ interseção e união entre Drugs.com, DrugBank e combinações seguras ATC/OMS. As combinações moderadas e menores foram agrupadas em $A_e^{-1}$ e $A_u^{1-}$. $\mathbf{A_p}$ combinações previstas computacionalmente análogas à $\mathbf{A_c}$. **Colunas**. $\mathbf{A}$ Universo de combinações. $\mathbf{F_k}$ Fármacos das combinações classificadas. $\mathbf{A_k}$ Associações classificadas. $\check{\mathbf{X}}$ Soma das citações MEDLINE coletadas em agosto de 2013 com a combinação dos nomes genéricos de cada par de fármacos dos grupos $A_k$, $V_k$ e $B_k$. $\mathbf{V_k}$ Universo de combinações conhecidas projetadas dentre os fármacos utilizados. $\mathbf{G_k}$ Fármacos utilizados e classificados. $\mathbf{H_k}$ Fármacos associados e classificados. $\mathbf{B_k}$ Associações utilizadas e classificadas. $\mathbf{T_k}$ Usuários das combinações classificadas. $\mathbf{S}$ Usuários de medicamentos. $T_k \div S$ Prevalência. † O critério de simultaneidade e o cálculo da prevalência foi realizado com base no relato de uso de medicamentos por $|S| = 8.890$ pacientes em até 15 dias antecedentes à entrevista.

Tabela 6.2: **Representatividade e prevalência de combinações conhecidas e previstas de fármacos na base SIGAF/SES-MG**, Sistema Integrado de Gerenciamento da Assistência Farmacêutica, 2010 a 2013.

| | Classificação | | | Prevalência | Fármacos usados e combinações derivadas | | | | | Associações usadas e fármacos derivados | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A** | $\lvert\mathbf{F_k}\rvert$ | $\lvert\mathbf{A_k}\rvert$ | $\frac{\lvert\breve{\mathbf{A}}_\mathbf{k}\rvert}{\lvert\mathbf{A_k}\rvert}$ | $\frac{\lvert\mathbf{T_k}\rvert}{\lvert\mathbf{S}\rvert}$ | $\lvert\mathbf{G_k}\rvert$ | $\frac{\lvert\mathbf{G_k}\rvert}{\lvert\mathbf{F_k}\rvert}$ | $\lvert\mathbf{V_k}\rvert$ | $\frac{\lvert\mathbf{B_k}\rvert}{\lvert\mathbf{A_k}\rvert}$ | $\frac{\lvert\breve{\mathbf{V}}_\mathbf{k}\rvert}{\lvert\mathbf{V_k}\rvert}$ | $\lvert\mathbf{H_k}\rvert$ | $\frac{\lvert\mathbf{H_k}\rvert}{\lvert\mathbf{F_k}\rvert}$ | $\lvert\mathbf{B_k}\rvert$ | $\frac{\lvert\mathbf{B_k}\rvert}{\lvert\mathbf{A_k}\rvert}$ | $\frac{\lvert\breve{\mathbf{B}}_\mathbf{k}\rvert}{\lvert\mathbf{B_k}\rvert}$ |
| $A_c^{1-}$ | 407 | 1.218 | 196,5 | 6,3% | 187 | 45,9% | 336 | 27,6% | 376,4 | 185 | 45,5% | 131 | 10,8% | 707,8 |
| $A_c^{2-}$ | 819 | 14.659 | 75,6 | 22,0% | 272 | 33,2% | 2.772 | 18,9% | 155,9 | 266 | 32,5% | 1.033 | 7,0% | 180,0 |
| $A_c^{3-}$ | 561 | 2.246 | 150,6 | 2,1% | 210 | 37,4% | 416 | 18,5% | 179,8 | 205 | 36,5% | 149 | 6,6% | 189,4 |
| $A_b^{0}$ | 342 | 687 | 497,6 | 25,3% | 150 | 43,9% | 203 | 29,5% | 480,1 | 149 | 43,6% | 118 | 17,2% | 744,1 |
| $A_b^{1-}$ | 1.220 | 12.786 | 17,0 | 16,8% | 323 | 26,5% | 1.858 | 14,5% | 50,0 | 317 | 26,0% | 635 | 5,0% | 51,9 |
| $A_e^{0}$ | 342 | 687 | 497,6 | 25,3% | 150 | 43,9% | 203 | 29,5% | 480,1 | 149 | 43,6% | 118 | 17,2% | 744,1 |
| $A_e^{1-}$ | 412 | 1.144 | 100,2 | 5,6% | 174 | 42,2% | 274 | 24,0% | 186,2 | 171 | 41,5% | 126 | 11,0% | 135,0 |
| $A_e^{3-}$ | 250 | 439 | 68,9 | 1,1% | 109 | 43,6% | 136 | 31,0% | 144,3 | 108 | 43,2% | 53 | 12,1% | 107,2 |
| $A_u^{0}$ | 396 | 802 | 641,4 | 36,2% | 170 | 42,9% | 245 | 30,5% | 637,9 | 169 | 42,7% | 147 | 18,3% | 929,2 |
| $A_u^{1-}$ | 1.319 | 27.077 | 52,5 | 28,8% | 345 | 26,2% | 4.556 | 16,8% | 127,5 | 338 | 25,6% | 1.620 | 6,0% | 178,3 |
| $A_u^{3-}$ | 561 | 2.246 | 150,6 | 2,1% | 210 | 37,4% | 416 | 18,5% | 179,8 | 205 | 36,5% | 149 | 6,6% | 189,4 |
| $A_p^{0}$ | 1.244 | 41.549 | 103,7 | 25,5% | 305 | 24,5% | 2.453 | 5,9% | 135,2 | 298 | 24,0% | 774 | 1,9% | 163,1 |
| $A_p^{1-}$ | 72 | 60 | 17,5 | 0,0% | 29 | 40,3% | 11 | 18,3% | 42,9 | 29 | 40,3% | 3 | 5,0% | 3,0 |
| $A_p^{2-}$ | 1.344 | 12.217 | 23,4 | 12,5% | 360 | 26,8% | 1.272 | 10,4% | 48,7 | 353 | 26,3% | 392 | 3,2% | 97,5 |
| $A_p^{3-}$ | 53 | 51 | 126,7 | 0,0% | 23 | 43,4% | 21 | 41,2% | 133,6 | 22 | 41,5% | 6 | 11,8% | 373,8 |

**Linhas**. $\mathbf{A_c}$ Drugs.com, 0 são combinações seguras sob o mesmo código da classificação ATC/OMS e $\{1-, 2-, 3-\}$ são interações medicamentosas adversas menores, moderadas e maiores, respectivamente. $\mathbf{A_b}$ DrugBank, 0 indica combinações seguras e $1-$ interações medicamentosas adversas. $\mathbf{A_e}$, $\mathbf{A_u}$ interseção e união entre Drugs.com, DrugBank e combinações seguras ATC/OMS. As combinações moderadas e menores foram agrupadas em $A_e^{-1}$ e $A_u^{1-}$. $\mathbf{A_p}$ combinações previstas computacionalmente análogas à $\mathbf{A_c}$. **Colunas**. $\mathbf{A}$ Universo de combinações. $\mathbf{F_k}$ Fármacos das combinações classificadas. $\mathbf{A_k}$ Associações classificadas. $\breve{\mathbf{X}}$ Soma das citações MEDLINE coletadas em agosto de 2013 com a combinação dos nomes genéricos de cada par de fármacos dos grupos $A_k$, $V_k$ e $B_k$. $\mathbf{V_k}$ Universo de combinações conhecidas projetadas dentre os fármacos utilizados. $\mathbf{G_k}$ Fármacos utilizados e classificados. $\mathbf{H_k}$ Fármacos associados e classificados. $\mathbf{B_k}$ Associações utilizadas e classificadas. $\mathbf{T_k}$ Usuários das combinações classificadas. $\mathbf{S}$ Usuários de medicamentos. $T_k \div S$ Prevalência. $^{\dagger}$ O critério de simultaneidade e o cálculo da prevalência foi realizado com base em dispensações com intervalo de até 15 dias para $|S| = 542.415$ durante o intervalo de 1.142 dias.

fármacos previstos. Contraditoriamente ao senso comum, não observou-se de forma unânime a correspondência direta entre o número de citações e a relevância para todas as interações medicamentosas. O fato da correspondência ser ainda menor para as previstas, sugere o ineditismo das previsões, sendo admissível devido ao número de interações medicamentosas conhecidas que não recuperaram citações. Conjectura-se que poucos estudos publicados foram suficientes para motivar a inclusão de grande parte das combinações presentes em determinado compêndio, provavelmente devido à sua relevância quanto ao nível de evidência.

Desta forma, cabe um refinamento na estratégia de busca para recuperar as citações específicas à interação presente no compêndio e mensurar a frequência. Salienta-se o viés desta informação por não contemplar citações de outras fontes igualmente relevantes como o EMBASE, LILACS, CENTRAL, entre outras. Contudo, embora a busca não tenha sido sistemática, é um artefato inédito de comparação da evidência científica de pares de fármacos a partir de um veículo que condensa uma parcela importante das publicações.

#### 6.2.1.3 Avaliação por grupo anatômico-terapêutico

Um medicamento pode apresentar várias classificações ATC/OMS, cuja escolha é realizada mediante o diagnóstico. Devido a avaliação dos fármacos enquanto entidades químicas, construiu-se a tabela 6.3 considerando todas as classificações assinaladas para o mesmo fármaco no nível anatômico. Nesta tabela é relatada a quantidade de combinações no total e os respectivos teores de interações medicamentosas potenciais adversas conforme classificação Drugs.com, interseção e união com DrugBank e conforme previsões *in silico*.

A combinação de fármacos indicados para os mesmo sistema pode sugerir interação medicamentosa por duplicidade terapêutica. Conforme relação $B_a \div B$, descrita na tabela 6.3, dentre as combinações classificadas (conhecidas e previstas) para o mesmo sistema lideraram o cardiovascular (15,1% dentre 4.061 combinações) e nervoso (13,7%), com a maior diversidade de uso. Destacam-se combinações entre anti-infecciosos (3,9%). A prevalência de antibióticos associados em ambiente não hospitalar foi 1,4% e 13,3% nas bases ELSA e SIGAF, sendo que metade das combinações representa algum perigo e, além destas, 4,4% foram previstas como interação medicamentosa.

Dentre as classificações com maior número de combinações de fármacos que tratam sistemas anatômicos diferentes, liderou aquelas com o sistema cardiovascular, presente em sete dentre os quinze sistemas distintos com maior diversidade de combinações (tabela 6.3). Dentre estes sete grupos, a prevalência de combinações com o sistema cardiovascular variou de 5,6 a 28,8% na base ELSA e 11,8% a 28,7% na base SIGAF, liderando combinações com aparelho digestivo e metabolismo em ambas as bases.

A relação de citações e combinações entre todas as combinações classificadas $A_k$ foi superior em todas as instâncias do Drugs.com e inferior em todas as previsões, exceto na combinação entre fármacos para o aparelho digestivo e anti-infecciosos, em que foram observadas

332,1 citações MEDLINE por combinação prevista, sugerindo algum padrão de correspondência.

Observando a prevalência das 91 combinações entre sistemas anatômicos, verificou-se que a maior proporção de interações medicamentosas potenciais adversas foi para o grupo de medicamentos indicados para o sistema músculo-esquelético associado com medicamentos para o sangue e órgãos hematopoiéticos ($n = 26$) ou medicamentos para o sistema cardiovascular ($n = 140$), ultrapassando 90% de casos indicados em pelo menos uma das bases. As interações duplamente documentadas foram lideradas por medicamentos anti-infecciosos associados com medicamentos para o sangue e órgãos hematopoiéticos ($n = 20$; 29,0%) ou para o sistema nervoso ($n = 44$; 17,7%).

As combinações documentadas em pelo menos uma das bases que apresentaram mais de 1500 citações MEDLINE por combinação utilizada foram entre medicamentos para o sangue e órgãos hematopoiéticos com hormônios ($n = 35$; 77,8%) ou aparelho respiratório ($n = 38$; 70,4%), entre medicamentos para aparelho digestivo e hormônios ($n = 60$; 52,6%) e entre dermatológicos e hormônios ($n = 36$; 54,6%).

As combinações mais prevalentes foram entre medicamentos para o aparelho digestivo, circulatório e nervoso ultrapassando 25% na base ELSA e 28% na base SIGAF. A evidência na combinação entre estes três grupos foi de 3,5% para interações adversas duplamente qualificadas $A_e^-$ a 74,9% àquelas identificadas em pelo menos uma das bases.

#### 6.2.1.4 Associações mais prevalentes

A seguir são descritas as combinações mais prevalentes conforme classificação ATC/OMS nível 5 (químico).

**Drugs.com** As trinta combinações mais prevalentes sugeridas pelo Drugs.com como interação medicamentosa potencial adversa são mostradas na tabela 6.4.

Ibuprofeno liderou as combinações mais utilizadas. Segundo o Drugs.com. o ibuprofeno possui 65 combinações com algum grau de risco (25 confirmadas pelo DrugBank), sendo 57 observadas mais de uma vez nas populações estudadas. Este medicamento é considerado seguro por ser dotado de elevada biodisponibilidade e solubilidade, sendo vendido sem a obrigatoriedade da apresentação de prescrição médica[1]. A aparente segurança e a indução ao consumo devem ser reavaliadas quando outros medicamentos são usados concomitantemente.

**Interseção** Dentre as evidências duplamente documentadas, uma combinação maior envolvendo medicamento de venda livre ocorreu entre ibuprofeno e varfarina. Outra combinação com medicamento de venda livre foi entre paracetamol e ciprofloxacino, usado por cerca de 60 pacientes da base SIGAF.

[1] Medicamentos de venda livre são também chamados OTC, *out-the-counter*.

Tabela 6.3: **Associações medicamentosas mais diversificadas segundo classificação ATC/OMS por nível anatômico utilizadas pelas populações ELSA e SIGAF** e interações medicamentosas adversas potenciais. Os percentuais são relacionados à coluna $B_a$.

| **ATC** nível 1 | | $\frac{\lvert\mathbf{B_c}\rvert}{\lvert\mathbf{B_a}\rvert}$ | $\frac{\lvert\check{\mathbf{B}}_\mathbf{c}\rvert}{\lvert\mathbf{B_c}\rvert}$ | $\frac{\lvert\mathbf{B_e}\rvert}{\lvert\mathbf{B_a}\rvert}$ | $\frac{\lvert\check{\mathbf{B}}_\mathbf{e}\rvert}{\lvert\mathbf{B_e}\rvert}$ | $\frac{\lvert\mathbf{B_u}\rvert}{\lvert\mathbf{B_a}\rvert}$ | $\frac{\lvert\check{\mathbf{B}}_\mathbf{u}\rvert}{\lvert\mathbf{B_u}\rvert}$ | $\frac{\lvert\mathbf{B_p}\rvert}{\lvert\mathbf{B_a}\rvert}$ | $\frac{\lvert\check{\mathbf{B}}_\mathbf{p}\rvert}{\lvert\mathbf{B_p}\rvert}$ | $\lvert\mathbf{B_a}\rvert$ | $\frac{\lvert\mathbf{B_a}\rvert}{\lvert\mathbf{B}\rvert}$ | $\frac{\lvert\check{\mathbf{B}}_\mathbf{a}\rvert}{\lvert\mathbf{B_a}\rvert}$ | $\frac{\lvert\mathbf{T}\rvert}{\lvert\mathbf{S}\rvert}^{\dagger}$ | $\frac{\lvert\mathbf{T}\rvert}{\lvert\mathbf{S}\rvert}^{\ddagger}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C | C | 31,1% | 643,1 | 4,7% | 159,9 | 40,3% | 511,3 | 5,9% | 51,2 | 615 | 15,1% | 339,9 | 22,7% | 30,3% |
| N | N | 24,2% | 421,8 | 2,2% | 275,3 | 40,0% | 279,7 | 2,5% | 29,4 | 557 | 13,7% | 150,0 | 11,3% | 18,2% |
| A | C | 69,1% | 465,3 | 3,5% | 156,5 | 74,9% | 429,2 | 18,5% | 71,1 | 482 | 11,9% | 404,1 | 28,8% | 28,7% |
| C | D | 78,1% | 346,2 | 8,2% | 100,2 | 81,9% | 329,9 | 10,5% | 15,0 | 465 | 11,5% | 342,2 | 21,3% | 13,0% |
| C | N | 56,8% | 231,0 | 11,3% | 168,2 | 70,8% | 186,1 | 26,3% | 35,2 | 407 | 10,0% | 151,7 | 25,6% | 28,1% |
| D | N | 70,2% | 268,9 | 13,1% | 92,3 | 78,9% | 239,1 | 16,2% | 14,1 | 389 | 9,6% | 192,2 | 19,0% | 11,9% |
| C | R | 68,5% | 380,0 | 4,2% | 179,3 | 76,2% | 344,0 | 15,2% | 37,6 | 336 | 8,3% | 282,5 | 20,5% | 14,5% |
| A | N | 53,6% | 278,8 | 6,9% | 165,8 | 66,9% | 223,5 | 24,1% | 16,6 | 332 | 8,2% | 287,1 | 25,5% | 22,9% |
| A | J | 50,0% | 283,1 | 6,2% | 353,1 | 58,0% | 244,2 | 10,1% | 332,1 | 276 | 6,8% | 219,8 | 5,8% | 19,1% |
| N | R | 65,2% | 312,4 | 6,6% | 224,7 | 73,6% | 276,6 | 15,8% | 17,5 | 273 | 6,7% | 221,5 | 15,3% | 13,8% |
| C | J | 59,9% | 173,3 | 9,0% | 179,1 | 74,9% | 138,7 | 9,4% | 34,5 | 267 | 6,6% | 113,5 | 5,6% | 21,5% |
| A | D | 53,8% | 978,5 | 5,6% | 145,5 | 61,4% | 857,0 | 16,5% | 30,9 | 249 | 6,1% | 681,9 | 18,2% | 12,3% |
| J | N | 66,7% | 67,4 | 17,7% | 22,0 | 84,7% | 53,0 | 8,4% | 2,2 | 249 | 6,1% | 45,8 | 5,1% | 19,8% |
| A | A | 29,1% | 1859,7 | 1,3% | 57,7 | 33,8% | 1604,0 | 14,3% | 31,5 | 237 | 5,8% | 967,4 | 18,6% | 16,2% |
| D | J | 62,2% | 451,9 | 5,6% | 531,6 | 73,0% | 385,5 | 4,7% | 85,1 | 233 | 5,7% | 312,8 | 4,7% | 13,2% |
| A | R | 52,0% | 909,5 | 0,5% | 85,0 | 55,5% | 852,2 | 23,0% | 66,4 | 200 | 4,9% | 525,9 | 19,8% | 13,8% |
| C | H | 82,3% | 618,7 | 2,1% | 87,3 | 83,9% | 607,2 | 12,5% | 16,2 | 192 | 4,7% | 667,1 | 19,3% | 11,8% |
| C | G | 65,8% | 346,0 | 9,1% | 116,7 | 72,2% | 316,0 | 22,5% | 35,2 | 187 | 4,6% | 258,6 | 13,1% | 14,6% |
| J | J | 44,0% | 786,4 | 3,8% | 623,0 | 53,5% | 654,2 | 4,4% | 128,4 | 159 | 3,9% | 620,2 | 1,4% | 13,3% |
| geral | $Q_1$ | 50,0% | 66,5 | 0,0% | 0,0 | 56,7% | 61,7 | 5,9% | 10,0 | 23,5 | 0,6% | 101,1 | 1,5% | 2,6% |
| (91) | $\tilde{X}$ | 57,9% | 248,4 | 5,4% | 77,3 | 70,6% | 214,6 | 12,9% | 30,2 | 51,5 | 1,3% | 273,7 | 3,4% | 6,3% |
| | $Q_3$ | 68,5% | 602,3 | 10,3% | 188,3 | 78,6% | 505,7 | 20,0% | 85,1 | 141,5 | 3,5% | 623,9 | 10,0% | 12,9% |

*ATC* Classificação ATC no nível anatômico para cada fármaco $f_i$ e $f_j$. **A** Aparelho digestivo e metabolismo, **B** Sangue e órgãos hematopoiéticos, **C** Aparelho cardiovascular, **D** Dermatológicos, **G** Aparelho geniturinário e hormônios sexuais, **H** Preparações hormonais sistêmicas, excluindo hormônios sexuais e insulinas, **J** Anti-infecciosos para uso sistêmico, **L** Antineoplásicos e imunomoduladores, **M** Sistema músculo-esquelético, **N** Sistema nervoso, **P** Produtos antiparasitários, inseticidas e repelentes, **R** Aparelho respiratório, **S** Órgãos sensoriais, **V** Vários. $\mathbf{B_x}$ Associações medicamentosas potenciais adversas classificadas segundo *p* previsão, *c* Drusg.com, *d* DrugBank, *e* interseção e *u* união das bases Drugs.com e DrugBank. $\mathbf{B_a}$ Total de combinações classificadas pela *ATC* nível 1 e usadas pelas populações ELSA ou SIGAF. **B** Total de combinações utilizadas ($n = 4061$). $\check{\mathbf{X}}$ Soma das citações MEDLINE coletadas em agosto de 2013 com a combinação dos nomes genéricos de cada par de fármacos. $\tilde{X}$ mediana. $Q_1$ e $Q_3$ primeiro e terceiro quartis. $\frac{\lvert\mathbf{T}\rvert}{\lvert\mathbf{S}\rvert}$ prevalência de expostos ao grupo de combinações em relação ao total de usuários de medicamentos $^{\dagger}\lvert S\rvert = 8.890$ (ELSA) e $^{\ddagger}\lvert S\rvert = 542.415$ (SIGAF). Houve redundância na frequência do fármaco que possui mais de uma ATC.

Tabela 6.4: **Associações mais prevalentes conforme classificação Drugs.com.**

| Associação | | ATC $f_i$ | ATC $f_j$ | Classificação | B̆ | ELSA† | SIGAF‡ |
|---|---|---|---|---|---|---|---|
| carbamazepina | hidrocortisona | N | A C D H S | moderada | 64 | 4,477% | |
| somatropina | hidrocortisona | H | A C D H S | moderada | 5.400 | 3,645% | |
| clorotiazida | ibuprofeno | C | C G M | moderada | 34 | | 3,219% |
| hidrocortisona | amiodarona | A C D H S | C | maior | 10 | 2,520% | |
| somatropina | lidocaína | H | A C D N R S | menor | 12 | 2,284% | |
| loratadina | ibuprofeno | C R | C G M | moderada | 24 | 0,023% | 2,279% |
| losartana | ibuprofeno | C | C G M | moderada | 5 | | 1,984% |
| ibuprofeno | captopril | C G M | C | moderada | 22 | | 1,727% |
| fluoxetina | ibuprofeno | N | C G M | moderada | 43 | | 1,700% |
| azitromicina | amoxicilina | J S | J | menor | 527 | | 1,326% |
| losartana | hidrocortisona | C | A C D H S | moderada | 22 | 1,260% | |
| prednisona | ibuprofeno | A H | C G M | moderada | 65 | | 1,253% |
| ibuprofeno | dexametasona | C G M | A C D H R S | moderada | 150 | | 1,186% |
| loratadina | prednisona | C R | A H | moderada | 29 | | 1,174% |
| insulina | clorotiazida | A | C | moderada | 191 | | 1,164% |
| omeprazol | ciprofloxacino | A | J S | menor | 32 | 0,023% | 1,140% |
| ciprofloxacino | ibuprofeno | J S | C G M | moderada | 38 | 0,011% | 1,115% |
| enalapril | ibuprofeno | C | C G M | moderada | 16 | | 1,108% |
| fluconazol | miconazol | D J | A D G J S | moderada | 409 | | 1,051% |
| mebendazol | metronidazol | P | A D G J P | moderada | 61 | | 1,005% |
| insulina | metformina | A | A | moderada | 3958 | | 0,990% |
| cetoconazol | dexametasona | D G J | A C D H R S | moderada | 167 | | 0,984% |
| loratadina | dexametasona | C R | A C D H R S | moderada | 24 | | 0,980% |
| prednisona | clorotiazida | A H | C | moderada | 72 | | 0,940% |
| clorotiazida | dexametasona | C | A C D H R S | moderada | 39 | | 0,928% |
| anlodipino | ibuprofeno | C | C G M | moderada | 14 | | 0,850% |
| omeprazol | ferro | A | B C N V | moderada | 22 | | 0,819% |
| diclofenaco | clorotiazida | D M S | C | moderada | 23 | | 0,812% |
| insulina | captopril | A | C | moderada | 262 | | 0,737% |
| insulina | somatropina | A | H | moderada | 19.467 | 0,709% | |

*ATC* Classificação ATC no nível anatômico para cada fáramco $f_i$ e $f_j$. **A** Aparelho digestivo e metabolismo, **B** Sangue e órgãos hematopoiéticos, **C** Aparelho cardiovascular, **D** Dermatológicos, **G** Aparelho geniturinário e hormônios sexuais, **H** Preparações hormonais sistêmicas, excluindo hormônios sexuais e insulinas, **J** Anti-infecciosos para uso sistêmico, **L** Antineoplásicos e imunomoduladores, **M** Sistema músculo-esquelético, **N** Sistema nervoso, **P** Produtos antiparasitários, inseticidas e repelentes, **R** Aparelho respiratório, **S** Órgãos sensoriais, **V** Vários. **B̆** Citações MEDLINE coletadas em agosto de 2013 com a combinação dos nomes genéricos. † Prevalência da combinação em relação ao relato de uso de medicamento(s) por 8.890 pacientes por até 15 dias antecedentes à entrevista ($t = 1$). ‡ Prevalência da combinação em relação ao total de 542.415 pacientes que tiveram medicamentos dispensados em um intervalo de 1.142 dias, considerando intervalo máximo de 15 dias entre as dispensações como critério para combinação.

Outras combinações consideradas maiores foram entre fluconazol e sinvastatina, haloperidol e lítio, fluoxetina e lítio, hidroclorotiazida e lítio; sendo nenhuma delas observada na base ELSA.

A tabela 6.5 relata as trinta combinações classificadas mais prevalentes nas populações estudadas.

**Avaliação de casos previstos** O FDA alertou para o uso de diuréticos que geram perda de magnésio, o qual inclui a hidroclorotiazida em concomitância com medicamentos inibidores da bomba de prótons [FDA, 2011]. A interação prevista entre omeprazol e hidroclorotiazida não foi confirmada por nenhuma fonte consultada, contudo, a interação considerada moderada entre omeprazol e furosemida pode causar hipomagnesemia de acordo com o Drugs.com. A redução dos níveis de magnésio pode causar arritmia, palpitações, espasmo muscular, tremor ou convulsões.

Benzodiazepinas como o clonazepam associadas com tiazídicos como a hidroclorotiazida

Tabela 6.5: **Associações mais prevalentes conforme interseção entre Drugs.com e DrugBank.**

| **Associação** | | **ATC** $f_i$ | **ATC** $f_j$ | **Classificação** | **B̆** | ELSA† | SIGAF‡ |
|---|---|---|---|---|---|---|---|
| cafeína | ciprofloxacino | N | J S | moderada | 95 | | 1,156% |
| torasemide | ibuprofeno | C | C G M | moderada | 66 | | 1,024% |
| estradiol | prednisolona | G L | A C D H R S | moderada | 513 | 0,990% | 0,021% |
| varfarina | hidrocortisona | B | A C D H S | moderada | 28 | 0,855% | |
| propranolol | ibuprofeno | C | C G M | moderada | 111 | | 0,845% |
| atenolol | ibuprofeno | C | C G M | moderada | 54 | | 0,830% |
| omeprazol | cetoconazol | A | D G J | moderada | 78 | | 0,589% |
| lidocaína | timolol | A C D N R S | C S | moderada | 23 | 0,360% | |
| fluconazol | sinvastatina | D J | C | maior | 20 | | 0,331% |
| ibuprofeno | carvedilol | C G M | C | moderada | 2 | | 0,303% |
| sinvastatina | cetoconazol | C | D G J | maior | 46 | | 0,248% |
| ciprofloxacino | ferro | J S | B C N V | moderada | 18 | | 0,210% |
| prednisona | cetoconazol | A H | D G J | moderada | 74 | | 0,178% |
| losartana | lítio | C | D N | moderada | 18 | 0,158% | 0,067% |
| fenobarbital | dexametasona | N | A C D H R S | moderada | 616 | | 0,142% |
| haloperidol | lítio | N | D N | maior | 574 | | 0,142% |
| prednisona | fenobarbital | A H | N | moderada | 66 | | 0,136% |
| fluconazol | amitriptilina | D J | N | moderada | 7 | | 0,129% |
| fluoxetina | lítio | N | D N | maior | 267 | | 0,127% |
| lidocaína | carvedilol | A C D N R S | C | moderada | 2 | 0,124% | |
| carbamazepina | metronidazol | N | A D G J P | moderada | 24 | | 0,114% |
| fluconazol | carbamazepina | D J | N | moderada | 19 | | 0,113% |
| cafeína | norfloxacino | N | J | moderada | 41 | | 0,111% |
| cafeína | lítio | N | D N | moderada | 137 | | 0,107% |
| metronidazol | fenobarbital | A D G J P | N | moderada | 25 | | 0,103% |
| clorotiazida | lítio | C | D N | maior | 55 | | 0,096% |
| amitriptilina | cetoconazol | N | D G J | moderada | 17 | | 0,095% |
| varfarina | ibuprofeno | B | C G M | maior | 136 | | 0,092% |
| prednisolona | cetoconazol | A C D H R S | D G J | moderada | 123 | 0,011% | 0,091% |
| carbamazepina | cetoconazol | N | D G J | moderada | 40 | | 0,090% |

*ATC* Classificação ATC no nível anatômico para cada fáramco $f_i$ e $f_j$. **A** Aparelho digestivo e metabolismo, **B** Sangue e órgãos hematopoiéticos, **C** Aparelho cardiovascular, **D** Dermatológicos, **G** Aparelho geniturinário e hormônios sexuais, **H** Preparações hormonais sistêmicas, excluindo hormônios sexuais e insulinas, **J** Anti-infecciosos para uso sistêmico, **L** Antineoplásicos e imunomoduladores, **M** Sistema músculo-esquelético, **N** Sistema nervoso, **P** Produtos antiparasitários, inseticidas e repelentes, **R** Aparelho respiratório, **S** Órgãos sensoriais, **V** Vários. **B̆** Citações MEDLINE coletadas em agosto de 2013 com a combinação dos nomes genéricos. † Prevalência da combinação em relação ao relato de uso de medicamento(s) por 8.890 pacientes por até 15 dias antecedentes à entrevista ($t = 1$). ‡ Prevalência da combinação em relação ao total de 542.415 pacientes que tiveram medicamentos dispensados em um intervalo de 1.142 dias, considerando intervalo máximo de 15 dias entre as dispensações como critério para combinação.

apresentam mais episódios de hiponatremia severa do que o uso dos fármacos isoladamente [Liamis et al., 2013]. Confusão, convulsão, fatiga, cefaleia, irritabilidade, perda de apetite, espasmos muscular são sintomas causados pela redução dos níveis de sódio.

A angiotensina II está relacionada ao desenvolvimento de problemas vasculares, cardíacos e renais. Especula-se que bloqueadores de angiotensina I afetam a sensibilidade à insulina. No entanto, embora estudos apontem para a a melhora da sensibilidade [Jin & Pan, 2007], permanece a controvérsia da melhora com losartana devido a relatos de eventos adversos relacionados a resistência à insulina [DRUG INFORMER, 2013]. Em metanálise verificou-se que a telmisartana vem se mostrando mais específica para a angiotensina I e eficaz do que outros antagonistas de angiotensina como a losartana [Takagi & Umemoto, 2012]. Embora a previsão possa ser tomada como falso-positiva, a combinação entre estes dois fármacos deve ser realizada com cautela.

Tabela 6.6: **Associações mais prevalentes conforme previsão farmacológica**

| Associação | | ATC $f_i$ | ATC $f_j$ | Classificação | B̆ | ELSA† | SIGAF‡ |
|---|---|---|---|---|---|---|---|
| omeprazol | clorotiazida | A | C | moderada | 11 | | 5,166% |
| clorotiazida | clonazepam | C | N | moderada | 0 | | 1,950% |
| insulina | losartana | A | C | moderada | 148 | 0,158% | 1,045% |
| diazepam | captopril | N | C | moderada | 16 | | 0,992% |
| insulina | sinvastatina | A | C | moderada | 195 | | 0,983% |
| insulina | metilfenidato | A | N | moderada | 8 | 0,855% | 0,002% |
| amitriptilina | clorotiazida | N | C | moderada | 9 | | 0,831% |
| metformina | propranolol | A | C | moderada | 11 | | 0,589% |
| loratadina | diazepam | C R | N | moderada | 13 | 0,011% | 0,484% |
| fluoxetina | prednisona | N | A H | moderada | 5 | | 0,461% |
| insulina | peniramina | A | R | moderada | 8 | 0,416% | |
| cafeína | escopolamina | N | A N S | moderada | 95 | | 0,342% |
| metronidazol | clonazepam | A D G J P | N | moderada | 2 | | 0,334% |
| estradiol | dexametasona | G L | A C D H R S | moderada | 136 | | 0,329% |
| insulina | levotiroxina | A | H | moderada | 1534 | 0,011% | 0,283% |
| prednisolona | salbutamol | A C D H R S | R | moderada | 352 | | 0,271% |
| diazepam | metronidazol | N | A D G J P | moderada | 24 | | 0,249% |
| tenofovir | hidrocortisona | J | A C D H S | moderada | 5 | 0,248% | |
| hidrocortisona | lopinavir | A C D H S | ? | moderada | 2 | 0,248% | |
| levonorgestrel | ferro | G | B C N V | moderada | 3 | | 0,232% |
| insulina | atorvastatina | A | C | moderada | 162 | 0,214% | |
| hidrocortisona | famotidine | A C D H S | A | moderada | 0 | 0,214% | |
| miconazol | captopril | A D G J S | C | moderada | 2 | | 0,209% |
| phenytoin | captopril | N | C | moderada | 17 | | 0,195% |
| diazepam | ranitidine | N | A | moderada | 40 | | 0,186% |
| insulina | methyldopa | A | C | moderada | 51 | | 0,183% |
| sinvastatina | prednisolona | C | A C D H R S | moderada | 44 | | 0,169% |
| atenolol | ferro | C | B C N V | moderada | 8 | | 0,165% |
| fluconazol | propranolol | D J | C | moderada | 4 | | 0,162% |
| insulina | nortriptilina | A | N | moderada | 6 | | 0,155% |

*ATC* Classificação ATC no nível anatômico para cada fáramco $f_i$ e $f_j$. **A** Aparelho digestivo e metabolismo, **B** Sangue e órgãos hematopoiéticos, **C** Aparelho cardiovascular, **D** Dermatológicos, **G** Aparelho geniturinário e hormônios sexuais, **H** Preparações hormonais sistêmicas, excluindo hormônios sexuais e insulinas, **J** Anti-infecciosos para uso sistêmico, **L** Antineoplásicos e imunomoduladores, **M** Sistema músculo-esquelético, **N** Sistema nervoso, **P** Produtos antiparasitários, inseticidas e repelentes, **R** Aparelho respiratório, **S** Órgãos sensoriais, **V** Vários. **B̆** Citações MEDLINE coletadas em agosto de 2013 com a combinação dos nomes genéricos. † Prevalência da combinação em relação ao relato de uso de medicamento(s) por 8.890 pacientes por até 15 dias antecedentes à entrevista ($t = 1$). ‡ Prevalência da combinação em relação ao total de 542.415 pacientes que tiveram medicamentos dispensados em um intervalo de 1.142 dias, considerando intervalo máximo de 15 dias entre as dispensações como critério para combinação.

## 6.2.2 Verificação das previsões

Foram avaliadas 78 previsões ($\alpha = 0,97$) por amostragem conforme descrito na seção 4.6.1.1 caculada segundo Scheaffer et al. [2011].

### 6.2.2.1 Compêndios

Foram identificadas 10 (12,82% da amostragem) combinações citadas nos compêndios avaliados [Jacomini & da Silva, 2011; Tatro, 2012; Micromedex, 2013].

A interação entre o beta-bloqueador **carvedilol** e **cimetidina** foi apontada por Tatro [2012] como moderada, de efeito rápido e documentação satisfatória ("provável"). Segundo o autor, a cimetidina pode reduzir os efeitos hepáticos de primeira passagem ao reduzir o fluxo sanguíneo e inibir o metabolismo via CYP2D6. Jacomini & da Silva [2011] consideraram como risco a ser avaliado e apontaram efeitos adversos como insônia, tontura, sintomas gastrointestinais e hipotensão postural.

O corticosteroide **prednisona** juntamente com o anti-fúngico **fluconazol**, foi considerado por Tatro [2012] como interação moderada, de documentação intermediária ("suspeita") e efeito retardado. A inibição do metabolismo da prednisona via CYP3A4 pode reduzir a eliminação e aumentar sua toxicidade.

O uso concomitante de **metformina** e **propranolol** foi relatado por Jacomini & da Silva [2011] como interação medicamentosa de risco, devido à consequente piora no controle dos níveis glicêmicos. Segundo as autoras o uso deve ser evitado.

A concentração plasmática da **loratadina** pode ser aumentada diante do uso concomitante de **cimetidina**, recomendando-se avaliação do risco diante da necessidade do uso [Jacomini & da Silva, 2011]. Este fato está bem documentado e é classificado como interação menor [Micromedex, 2013].

O uso concomitante de **losartana** e **meloxicam** é bem documentado e apresenta risco moderado de reduzir os efeitos anti-hipertensivos e causar nefrotoxicidade [Micromedex, 2013].

**Diazepam** e **ranitidina** foi uma interação considerada como improvável por Tatro [2012], com gravidade menor e desfecho rápido devido à possível alteração da biodisponibilidade observada em voluntários.

O uso de **insulina** com **levotiroxina** ou ***Ginkgo biloba*** são razoavelmente documentados e podem resultar em moderado decréscimo da efetividade do agente anti-diabético [Micromedex, 2013].

Tatro [2012] sugeriu interação de ação rápida, porém de severidade menor com evidência classificada como "possível" para a combinação de **paracetamol** e **escopolamina**. O início da ação do paracetamol pode ter o efeito retardado levemente reduzido, devido à queda na motilidade gastrointestinal dos anticolinérgicos[2]. Jacomini & da Silva [2011] indicou que esta interação não possui significação clínica.

**Levodopa** e **clonidina** é descrita por Tatro [2012] como possível, de efeito moderado e retardado, contudo, sem mecanismo descrito.

#### 6.2.2.2 Citações MEDLINE

Foram apresentadas 246 citações para 37 pares (47,4%, $n = 78$, mínimo=1, $Q_1 = 2$, mediana=4, $Q_3 = 6,65$, máximo=31). Associações terapêuticas ou tóxicas foram atribuídas a 13 (16,6%) combinações em 18 publicações.

Mallik et al. [2008] identificaram *in vitro* uma competição direta entre **verapamil** e **varfarina** para ligação no mesmo sítio de uma isoforma de albumina sérica, o que pode comprometer a liberação plasmática de um dos fármacos.

Houve inibição competitiva do **diclofenaco** por **fenitoína** via citocromo CYP2C9 observada a partir de ensaios enzimáticos *in vitro*Leemann et al. [1993].

[2]Substâncias anticolinérgicas inibem, de modo geral, receptores estimulados pela acetilcolina, principal mediadora da inervação parassimpática, responsável pela redução de batimentos cardíacos ou estimulação da contração da musculatura lisa intestinal.

**Haloperidol** foi capaz de inibir *in vitro* os efeitos da liberação de cálcio pela pregnenolona, um precursor da progesterona, reduzindo os efeitos da **pentazocina**[3], sugerindo que estes fármacos atuam em um mecanismo comum [Hong et al., 2004]. Bergeron et al. [1999] haviam relatado o antagonismo da progesterona no receptor $\sigma$ o qual é bloqueado pelo haloperidol.

Em um ensaio duplo-cego cruzado com nove homens, observou-se redução significativa da concentração plasmática de **prednisona** pela **cimetidina** ou **ranitidina**, contudo, sem alteração clinicamente significativa [Sirgo et al., 1985].

O uso de **prednisolona** e **salbutamol** fez com que 15 pacientes possuindo obstrução crônica das vias aéreas não obtivessem resultados clínicos com o tratamento [Curzon et al., 1983].

Hiperfagia estimulada pela administração crônica de um neuroesteroide precursor da **progesterona** em camundongos foi reduzida a hipofagia e analgesia com o uso de **fluoxetina**. Os autores sugeriram o envolvimento de receptores $5 - HT(2)$ relacionados a serotonina neste mecanismo [Kaur & Kulkarni, 2002].

Combinações de opioides, como a **morfina**, e anti-inflamatórios não esteroide, como a **nimesulida**, podem ter efeito direto espinhal sobre o processamento da informação nociceptora, o que pode ser alcançado por mecanismos adicionais, independentes da inibição da síntese de prostaglandinas ou ativação de receptores opioides. A combinação pode reduzir as doses necessárias para a analgesia da morfina [Pinardi et al., 2005; Miranda & Pinardi, 2009].

Reações extrapiramidais de elevadas dosagens de **metoclopramida** para combate a êmese durante o tratamento de câncer, pode ser mitigada com o uso de **lorazepam** [Seynaeve et al., 1991].

A liberação de prolactina diante do estresse cirúrgico foi reduzida com a combinação de **dexametasona** e **prometazina**, não havendo qualquer modificação com o uso isolado [Chapler et al., 1978].

Oransay et al. [2011] sugeriram **teofilina** como antídoto-terapia para a cardiotoxicidade em ratos induzida por **amitriptilina**. Especula-se que o efeito se deve ao antagonismo não seletivo da adenosina[4] pela teofilina.

Foi observado um efeito sinérgico de **metronidazol** e **rifampicina** *in vitro* contra o microrganismo *Bacteroides fragilis*, com aumento de 50% da atividade antimicrobiana em relação ao uso isolado [Ralph & Amatnieks, 1980].

Estudos em camundongos e sugeriram que agonistas de receptores imidazóis como a **clonidina** potencializaram os efeitos antidepressivos da **fluoxetina** devido a elevação do níveis cerebrais de agmatina, um neurotransmissor putativo [Rénéric et al., 2002; Taksande et al., 2009]. Ratos com agressividade induzida por apomorfina apresentaram efeitos anti-agressivos com a combinação, porém não manifestaram estes efeitos quando os fármacos foram usados

[3] O haloperidol reverte o efeito antagônico da pentazocina na analgesia da morfina.

[4] A amitriptilina induz vasodilatação parcial e, consequentemente, hipotensão ao acionar a adenosina, um receptor $\alpha 2$.

Tabela 6.7: **Associações previstas e corroboradas por outro modelo.** Gottlieb et al. [2012] desenvolveram uma ferramenta de previsão e sugestão do manejo de interações medicamentosas. O mecanismo é sugerido como farmacocinético ou farmacodinâmico. 78 combinações previstas e utilizadas por populações foram avaliadas.

| Associação | | ATC nível 1 | | Escore | Pr | Farmacocinética |
|---|---|---|---|---|---|---|
| alprazolam | diclofenaco | N | D M S | 0,723 | m | CYP3A4 |
| clonidina | bromazepam | C N S | N | 0,563 | m | - |
| clonidina | clorotiazida | C N S | C | 0,612 | m | - |
| clonidina | hidralazina | C N S | C | 0,687 | - | - |
| clonidina | indapamida | C N S | C | 0,798 | m | - |
| clonidina | metildopa | C N S | C | 0,931 | m | - |
| clonidina | paroxetina | C N S | N | 0,715 | m | - |
| diclofenaco | nalbufina | D M S | N | 0,875 | m | - |
| diclofenaco | risperidona | D M S | N | 0,661 | m | - |
| diclofenaco | trazodona | D M S | N | 0,604 | m | CYP2D6 |
| loratadina | cimetidina | C R | A | 0,428 | m | CYP3A4, CYP2C8 |
| loratadina | paroxetina | C R | N | 0,770 | m | CYP3A4 |
| loratadina | sertralina | C R | N | 0,791 | m | CYP3A4 |
| lorazepam | verapamil | N | C | 0,773 | m | CYP3A4 |
| losartan | oxcarbazepina | C | N | 0,636 | m | CYP3A4, CYP2C9 |
| risperidona | prednisolona | N | A C D H R S | 0,672 | m | - |
| verapamil | varfarina | C | B | 0,450 | m | CYP2C9 |

*ATC* Classificação ATC no nível anatômico para cada fáramco $f_i$ e $f_j$. **A** Aparelho digestivo e metabolismo, **B** Sangue e órgãos hematopoiéticos, **C** Aparelho cardiovascular, **D** Dermatológicos, **H** Preparações hormonais sistêmicas, excluindo hormônios sexuais e insulinas, **M** Sistema músculo-esquelético, **N** Sistema nervoso, **R** Aparelho respiratório, **S** Órgãos sensoriais, **Pr** Procedimento, onde *m* assinala a necessidade de monitoramento dos efeitos, de modo geral, equivalente a gravidade moderada. *CYP* são modalidades de enzimas do complexo do citocromo, responsável pelo metabolismo concomitante destes pares.

isoladamente [Skrebuhhova-Malmros et al., 2001].

Riedel et al. [1995] observou o efeito neuroprotetor da **cafeína** diante da redução da memória de curto e longo prazo causada pela **escopolamina** em 16 voluntários sadios. Posteriormente, este fenômeno foi avaliado *in vivo* por Botton et al. [2010], cuja prevenção foi observada em camundongos a partir de testes envolvendo reconhecimento de objetos.

#### 6.2.2.3 Comparação com outro modelo

Gottlieb et al. [2012] realizaram tratamento de relações farmacodinâmicas e farmacocinéticas e posterior classificação baseado em medidas de distância. Os 17 (21,80%) resultados correspondentes às previsões são mostrados na tabela 6.7.

Esta ferramente é baseada em atributos selecionados *a priori*. A diferença da ferramenta de comparação em relação à proposta é a escolha e tratamento de variáveis farmacológicas diretamente relacionadas às interações ao invés de selecionar uma ampla quantidade de variáveis de modo automático.

## 6.3 Discussão

A observação da prevalência em dados históricos de uso de fármacos é importante para caracterizar a proximidade das previsões com a realidade [Duke et al., 2012], sobretudo em face do tamanho do universo explorado e da inexistência de uma variável direta concernente ao uso por populações na base do conhecimento adotada.

A avaliação das previsões em bases populacionais foi observada em seis dentre os dez estudos recuperados pela revisão sistemática mostrada no capítulo 3. Bases que relatam utilização de medicamentos propiciaram diretamente a extração de informação preditiva por Estacio-Moreno et al. [2008] e Harpaz et al. [2010a]. Prontuários e notificações associadas ao conhecimento farmacológico possibilitaram a extração de interações medicamentosas por Kinney [1986], Lin et al. [2010] e Gottlieb et al. [2012]. Duke et al. [2012] mostrou que os textos científicos associados a bases populacionais provém informação preditiva para interações medicamentosas. Logo, torna-se adequado o posicionamento das previsões diante destas fontes.

As populações observadas mostraram representatividade diante da cobertura de fármacos e combinações. Verificou-se que praticamente metade dos usuários de medicamentos apresentaram algum grau de polifarmácia. A prevalência observada nas populações correspondeu à expectativa geral, a qual pode ultrapassar 80% em populações específicas como de pacientes idosos. A observação de interações medicamentosas potenciais podem variar de 17 a 33% conforme a base populacional avaliada [Loyola Filho et al., 2008; Rozenfeld et al., 2008; Pinto et al., 2013]. Agregando as interações medicamentosas potenciais adversas, a bases bases ELSA ($B_p^{1-} = 22,3\%$, $B_p^{3-} = 3,3\%$) e SIGAF ($B_p^{1-} = 28,8\%$, $B_u^{3-} = 2,1\%$) apresentaram tendências similares.

As combinações utilizadas para o treino do modelo cobriram 47,3% das 4.016 observadas em populações. Ressalta-se que 61,6% das combinações usadas foram documentadas em pelo menos uma fonte. A partir da similaridade com as interações do treino, foram previstas outras 1.560 como potencialmente não inertes, correspondendo a 38,4% das combinações utilizadas. A amplitude das previsões refletiu a amplitude da base de treino.

O trabalho de Duke et al. [2012] utilizou uma abordagem a partir da previsão de interações medicamentosas adversas com base em literatura de experimentos laboratoriais farmacocinéticos. Duke et al. [2012] observaram, dentre 13.197 interações medicamentosas previstas *in vitro*, 3.670 (29,7%) com uso verificado em prontuários médicos de 800 mil pacientes. Este valor é inferior ao conquistado pelo presente trabalho se considerarmos a totalidade das previsões. Porém, dentre as 12.369 previsões de interações medicamentosas potenciais adversas, 561 (4,5%) foram utilizadas pelas populações ELSA ou SIGAF. A comparação entre os trabalhos é limitada devido à orientação geral do presente estudo em contraposição as modelagem de busca por interações específicas envolvendo citocromos realizada por Duke et al. [2012]. Os autores identificaram, 196 interações avaliadas em ensaios clínicos farmacocinéticos (1,5% dentre as previsões), sendo 123 confirmadas como interações medicamentosas adversas e 73 como não

interações (precisão $= 62,8\%$). O valor baixo de casos confirmados possivelmente se deve à incidência do evento observado (miopatia).

O fato de muitas interações afetarem apenas subconjuntos de indivíduos [Aronson, 2011] pode ser apontada causa da divergência entre a potencialidade sinérgica ou adversa em relação ao observado na literatura e nas previsões. Conforme ilustrado, a informação do potencial terapêutico ou a dúvida em relação à manifestação nociva da combinação entre paracetamol e escopolamina pode ser relevante para a prescrição a um paciente com problemas hepáticos ou de mobilidade intestinal.

As previsões amostradas foram confirmadas acima da chance ao acaso. A incidência de quase 13% das previsões em compêndios e 17% em trabalhos preliminares mostra uma cobertura de 29,5% de previsões de combinações não inertes. Embora as técnicas de levantamento adotadas neste texto não tenham sido exaustivas, esta cobertura é comparável aos trabalhos que realizaram verificação manual, variando de 17,2% a 37,0% [Kinney, 1986; Estacio-Moreno et al., 2008; Harpaz et al., 2010a; Lin et al., 2010].

O maior ganho do modelo foi a capacidade de detectar combinações não inertes, ou seja, combinação de fármacos com potencial sinérgico ou adverso. A divergência da classificação da gravidade refletiu a natureza especulativa da base adotada, a qual não obteve substancial confirmação por outras bases. No entanto, a motivação pelo alerta de interações não inertes possibilita o monitoramento de inúmeros tratamentos que habitualmente são prescritos sem que se conheça plenamente os efeitos da combinação dos fármacos.

# Capítulo 7

# Considerações finais

Elaborou-se um modelo capaz de reconhecer interações medicamentosas potenciais diante das similaridades com interações conhecidas.

A abordagem implementada extrai a semântica implícita de cada atributo diante da amplitude tomada em um espaço n-dimensional de fármacos. Ao invés de coletar informação de cada par para estabelecer um conhecimento global, foi estabelecido um modelo comparativo global de todas as entidades para deduzir a natureza de casos específicos. A simplificação dos atributos enquanto medidas de distância viabilizou um modelo que permitiu a integração de diversas características de forma computacionalmente viável e expressiva.

O grande desafio defrontado pelos demais métodos citados foi a extração de conhecimento a partir de amplas bases de dados sem contexto específico, ou a formação de contexto manualmente estabelecido a partir de elaboradas técnicas com potencial restrição da capacidade de generalização. As técnicas de mineração de dados aplicadas neste modelo permitiram ao modelo a extração de padrões preditivos para grande parte do espaço de busca ao adotar atributos com contextos restritos como “absorção” ou “organismo afetado” de modo a reduzir o vocabulário abordado e elevar sua expressividade.

A existência de informações sugestivas na literatura ou a partir de previsões acerca de combinações medicamentosas, a despeito da ausência do consenso quanto a natureza sinérgica ou adversa, sugere alguma tendência a atividade biológica diferenciada do uso separado.

Antes da definição da intensidade da interação medicamentosa, deve-se definir se a combinação é inerte ou não. Logo, a maioria de combinações cuja atividade não seja conhecida consensualmente devem ser evitadas ou usadas parcimoniosamente até que maiores estudos sejam realizados, mesmo diante do potencial sinérgico. O modelo contribui para que este tipo de informação possa ser prestada na forma de alerta elevando-se o grau de monitoramento.

Em face das previsões falso-negativas documentadas na literatura o modelo possui relevante acuidade na identificação de combinações medicamentosas não inertes, ou seja, com potencial sinérgico ou adverso. Logo, delineou-se um modelo capaz de apreender características intrínsecas dos fármacos as quais possibilitam a previsão de sua interação com os demais

sem necessidade de avaliação utilização e desfechos clínicos diretos da combinação.

A contra-indicação de uma combinação apenas deve ser realizada a partir da corroboração laboratorial, clínica e epidemiológica em diversos subgrupos populacionais, ou seja, segundo o infundido pelas práticas de saúde baseada em evidência. As entidades com poder de decisão em saúde pública devem estabelecer conjuntamente com a academia quais combinações e interações medicamentosas são consensuais para instituir protocolos clínicos que restrinjam seu uso.

A continuidade da pesquisa requer estudos em profundidade com base em padrões corroborados para a descoberta de interações proibitivas e bases maiores, para verificação especulativa. As previsões relacionadas a cada fármaco devem ser investigadas com técnicas de revisão sistemática e estudos populacionais retrospectivos ou prospectivos. As regras emitidas pelo modelo em função dos atributos relevantes devem nortear estudos que traçarão os possíveis mecanismos baseados em relatos de morbidade e estudos laboratoriais ou clínicos.

A quantidade de combinações previstas se mostrará relevante conforme sejam corroboradas. Sendo a corroboração dos pares interagentes diminuta, o advento do modelo será a identificação de um fenômeno raro. Ao contrário, se forem numerosas as confirmações, indica que existe um padrão intrínseco que determinará interações que modifiquem o destino de um fármaco ainda na fase de desenvolvimento.

A sinergia de grande parte das combinações medicamentosas sem qualquer estudo quanto a interações permanece oculta. A exploração sistemática de todo o espaço de busca possibilita a emergência de novas atividades terapêuticas advindas de combinações desconhecidas ainda não observadas em populações.

Uma potencial aplicação para este modelo é a verificação de excipientes quanto à capacidade de serem inertes. A utilização de moléculas cujo potencial farmacológico é desconhecido ou tido como inexistente, quando comparadas quimicamente, pode ampliar a lista dos adjuvantes terapêuticos, visto que, ao que tudo indica, sua descoberta é acidental, dado que estudos sistemáticos em humanos de substâncias associadas ocorre em menor número devido à complexidade que cresce exponencialmente com o número de substâncias e devido a aspectos de segurança. Muitos fármacos com elevada toxicidade podem ter seus efeitos mitigados quando associados a outros fármacos, ou potencializados quando elevadas dosagens são necessárias. Logo, o modelo foi capaz de representar uma realidade seguindo o padrão ouro com a extração das características latentes de fármacos, sendo promissor no contexto clínico ou para pautar decisões em saúde pública, sobretudo quando pouco se conhece a respeito de um dado medicamento.

A investigação das características dos fármacos e consequente descoberta de polifarmácias que mitiguem os efeitos adversos ou promovam a redução de dosagens via potencialização da farmacoterapia pode introduzir novos regimes terapêuticos com critérios objetivos de eficácia e segurança.

# Referências Bibliográficas


(1994). *Lexicon of alcohol and drug terms*. World Health Organization.

(2007). Enzyme nomenclature database.

(2011). FDA Drug Safety Communication: Low magnesium levels can be associated with long-term use of Proton Pump Inhibitor drugs (PPIs) . `http://www.fda.gov/drugs/drugsafety/ucm245011.htm`. Accessed: 2013-09-20.

Abbagnano, N. (2007). *Dicionário de filosofia*. Martins Fontes.

Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K. & Walter, P. (2002). *Molecular biology of the cell*. Garland, 4 edição.

Aquino, E. M. L.; Barreto, S. M.; Bensenor, I. M.; Carvalho, M. S.; Chor, D.; Duncan, B. B.; Lotufo, P. A.; Mill, J. G.; Molina, M. D. C.; Mota, E. L. A.; Passos, V. M. A.; Schmidt, M. I. & Szklo, M. (2012). Brazilian Longitudinal Study of Adult Health (ELSA-Brasil): objectives and design. *Am. J. Epidemiol.*, 175(4):315--24.

Ardizzone, E.; Bonadonna, F.; Gaglio, S.; Marceno, R.; Nicolini, C.; Ruggiero, C. & Sorbello, F. (1988). Artificial intelligence techniques for cancer treatment planning. *Med Inform (Lond)*, 13(3):199--210.

Aronson, J. K. (2011). *Adverse Drug Reactions: History, Terminology, Classification, Causality, Frequency, Preventability*, pp. 1--119. John Wiley & Sons, Ltd.

Ashburner, M.; Ball, C.; Blake, J.; Botstein, D.; Butler, H.; Cherry, M.; Davis, A.; Dolinski, K.; Dwight, S. & Eppig, J. (2000). Gene Ontology: Tool for the Unification of Biology. *Nature Genetics*, 25(1):25--29.

Baxter, K. (2008). *Stockley's Drug Interactions, 8th Edition: A Source Book of Interactions, Their Mechanisms, Clinical Importance and Management*. Drug Interactions (Stockley) Series. Pharmaceutical Press.

Becker, M. L.; Kallewaard, M.; Caspers, P. W.; Visser, L. E.; Leufkens, H. G. & Stricker, B. H. (2007). Hospitalisations and emergency department visits due to drug–drug interactions: a literature review. *Pharmacoepidemiology and Drug Safety*, 16(6):641--651.

Berger, M. L.; Bingefors, K.; Hedblom, E. C.; Pashos, C. L. & Torrance, G. W. (2009). *Custo em saúde, qualidade e desfechos: o livro de termos da ISPOR*. Associação Brasileira de Farmacoeconomia e Pesquisa de Desfechos - ISPOR Brasil.

Bergeron, R.; De Montigny, C. & Debonnel, G. (1999). Pregnancy reduces brain sigma receptor function. *British Journal of Pharmacology*, 127(8):1769--1776.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl 1):D267--D270.

Botton, P. H.; Costa, M. S.; Ardais, A. P.; Mioranzza, S.; Souza, D. O.; da Rocha, J. B. T. & Porciúncula, L. O. (2010). Caffeine prevents disruption of memory consolidation in the inhibitory avoidance and novel object recognition tasks by scopolamine in adult mice. *Behavioural Brain Research*, 214(2):254--259.

Boyce, R.; Collins, C.; Horn, J. & Kalet, I. (2009). Computing with evidence Part I: A drug-mechanism evidence taxonomy oriented toward confidence assignment. *J Biomed Inform*, 42(6):979--989.

Brandao, C. M. R.; Felipe, F.; da Matta, M. G. P.; Afonso, G. J. A.; Gurgel, A. E. I.; Leal, C. M. & de Assis, A. F. (2013). Gastos público com medicamentos para o tratamento da osteoporose na pós-menopausa. *Revista de Saúde Pública*, 47:390--402.

BRASIL, A. A. N. d. V. S. (2009). Resolução nÂº 4, de 10 de fevereiro de 2009. Relatório técnico.

BRASIL, A. A. N. d. V. S. (2010a). Farmacopeia Brasileira. 1.

BRASIL, M. d. S. (2010b). Relacao nacional de medicamentos essenciais: RENAME.

BRASIL, M. d. S. (2011). DATASUS - CID 10.

Broccatelli, F.; Cruciani, G.; Benet, L. Z. & Oprea, T. I. (2012). BDDCS class prediction for new molecular entities. *Mol. Pharm.*, 9(3):570--580.

Brunton, L.; Lazo, J. & Parker, K. (2005). *Goodman & Gilman's The Pharmacological Basis of Therapeutics, Eleventh Edition*. McGraw Hill professional. Mcgraw-hill.

Burton, J.; Ijjaali, I.; Petitet, F.; Michel, A. & Vercauteren, D. P. (2009). Virtual screening for cytochromes p450: successes of machine learning filters. *Comb. Chem. High Throughput Screen.*, 12(4):369--382.

Byrne, B. (2003). Drug interactions: a review and update. *Endodontic Topics*, 4(1):9--21.

Calderón-Ospina, C. & Bustamante-Rojas, C. (2010). The DoTS classification is a useful way to classify adverse drug reactions: a preliminary study in hospitalized patients. *International Journal of Pharmacy Practice*, 18(4):230--235.

Campos Neto, O. H.; de Assis, A. F.; de Ávila, M. M. A.; Felipe, F.; Vasconcelos, B. F. L.; Leal, C. M. & Gurgel, A. E. I. (2012). Médicos, advogados e indústria farmacêuticatica na judicialização da saúde em Minas Gerais, Brasil. *Revista de Saúde Pública*, 46:784--790.

Carvalho, W. d. S.; Magalhães, S. M. S. & Reis, A. M. M. (2013). *Eventos adversos a medicamentos*, volume 1, pp. 145--84. COOPMED, 1 edição.

Caspi, R.; Foerster, H.; Fulcher, C. A.; Kaipa, P.; Krummenacker, M.; Latendresse, M.; Paley, S.; Rhee, S. Y.; Shearer, A. G.; Tissier, C.; Walk, T. C.; Zhang, P. & Karp, P. D. (2008). The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res*, 36(Database issue):D623--31.

Catal, C. (2012). Performance Evaluation Metrics for Software Fault Prediction Studies. *Acta Polytechnica Hungarica*, 9:4.

Ceccato, M. d. G. B.; Saturnino, L. T. M.; Almeida, C. C.; Oliveira, G. L. & Araújo, S. M. R. (2013). *Farmacoepidemiologia: o estado da arte no Brasil*, pp. 104--44. COOPMED, 1 edição.

Cerrito, P. (2001). Application of data mining for examining polypharmacy and adverse effects in cardiology patients. *Cardiovasc. Toxicol.*, 1(3):177--179.

Chapler, F.; Sherman, B. & Swanson, J. (1978). The effects of an antihistamine and/or a glucocorticoid on the prolactin response to surgical procedures. *Am J Obstet Gynecol*, 132(4):367--72.

Chen, H.; Ding, L.; Wu, Z.; Yu, T.; Dhanapalan, L. & Chen, J. Y. (2009). Semantic web for integrated network analysis in biomedicine. *Briefings in Bioinformatics*, 10(2):177--192.

Chen, X.; Ji, Z. L. & Chen, Y. Z. (2002). TTD: Therapeutic Target Database. *Nucleic Acids Research*, 30(1):412--415.

Cheng, F.; Yu, Y.; Zhou, Y.; Shen, Z.; Xiao, W.; Liu, G.; Li, W.; Lee, P. W. & Tang, Y. (2011). Insights into molecular basis of cytochrome p450 inhibitory promiscuity of compounds. *journal of chemical information and modeling*, 51(10):2482--2495.

Coloma, P.; Avillach, P.; Salvo, F.; Schuemie, M.; Ferrajolo, C.; Pariente, A.; Fourrier-Reglat, A.; Molokhia, M.; Patadia, V.; Lei, J. v. d.; Sturkenboom, M. & Trifirò, G. (2013). A reference standard for evaluation of methods for drug safety signal detection using electronic healthcare record databases. *Drug Safety*, 36(1):13--23.

Curzon, P. G.; Martin, M. A.; Cooke, N. J. & Muers, M. F. (1983). Effect of oral prednisolone on response to salbutamol and ipratropium bromide aerosols in patients with chronic airflow obstruction. *Thorax*, 38(8):601--4.

da Silveira, C. H.; Meira, W.; Silveira, S. A.; Rodrigues, A. O. & de Melo-Minardi, R. C. (2012). ADVISe: Visualizing the dynamics of enzyme annotations in UniProt/Swiss-Prot. *2012 IEEE Symposium on Biological Data Visualization (BioVis)*, 0:49--56.

Del Fiol, G. & Haug, P. J. (2009). Classification models for the prediction of clinicians' information needs. *J Biomed Inform*, 42(1):82--89.

Del Fiol, G.; Rocha, B. H.; Kuperman, G. J.; Bates, D. W. & Nohama, P. (2000). Comparison of two knowledge bases on the detection of drug-drug interactions. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pp. 171--175.

DRUG INFORMER (2013). Losartan Potassium Related Insulin Resistance. `http://www.druginformer.com/search/side_effect_details/cozaar/insulin%20resistance.html`. Accessed: 2013-09-20.

DRUGS.COM (2011). Prescription drug information, interacions and side effects. `http://www.drugs.com/zyrtec.html`.

Duda, S.; Aliferis, C.; Miller, R.; Statnikov, A. & Johnson, K. (2005). Extracting drug-drug interaction articles from MEDLINE to improve the content of drug databases. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pp. 216--220.

Duke, J. D. & Bolchini, D. (2011). A successful model and visual design for creating context-aware drug-drug interaction alerts. *AMIA Annu Symp Proc*, 2011:339--48.

Duke, J. D.; Han, X.; Wang, Z.; Subhadarshini, A.; Karnik, S. D.; Li, X.; Hall, S. D.; Jin, Y.; Callaghan, J. T.; Overhage, M. J.; Flockhart, D. A.; Strother, R. M.; Quinney, S. K. & Li, L. (2012). Literature Based Drug Interaction Prediction with Clinical Assessment Using Electronic Medical Records: Novel Myopathy Associated Drug Interactions. *PLoS Computational Biology*, 8(8).

Ebrahiminia, V.; Riou, C.; Seroussi, B.; Bouaud, J.; Dubois, S.; Falcoff, H. & Venot, A. (2006). Design of a decision support system for chronic diseases coupling generic therapeutic algorithms with guideline-based specific rules. *Stud Health Technol Inform*, 124:483--488.

Elden, L. (2006). Numerical linear algebra in data mining. *Acta Numerica*, pp. 327--384.

ELSA (2009). ELSA Brasil: the greatest epidemiological study in Latin America. *Rev Saude Publica*, 43(1).

Escousse, A.; Bianchetti, D. & Sgro, C. (1987). Database for practitioners with the Minitel system on side effects and drug interactions. *Therapie*, 42(1):57.

Estacio-Moreno, A.; Toussaint, Y. & Bousquet, C. (2008). Mining for adverse drug events with formal concept analysis. *Stud Health Technol Inform*, 136:803--808.

Evans, S. J.; Waller, P. C. & Davis, S. (2001). Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and drug safety*, 10(6):483--486.

Eyers, C. E. & Reamtong, O. (2008). All systems are go. *Genome Biology*, 9(5).

Fall, C. P.; Marland, E. S.; Wagner, J. M. & Tyson, J. J. (2002). Computational Cell Biology.

Ferreira, A. (2009). *Novo dicionário Aurélio da língua portuguesa*. Positivo, 4 edição.

Fuhr, U. (2008). Improvement in the handling of drug-drug interactions. *Eur. J. Clin. Pharmacol.*, 64(2):167--171.

Gardner, D. & Rizack, M. (1990). A Prolog knowledge base for drug interactions. *Comput. Biomed. Res.*, 23(2):139--152.

Gebhart, F. (2011). Data-mining uncovers hyperglycemic drug-drug interaction between paroxetine and pravastatin. *Drug Topics*, 155(8):25.

Gelfond, M. & Lifschitz, V. (1988). The Stable Model Semantics For Logic Programming. pp. 1070--1080. MIT Press.

Gelfond, M. & Lifschitz, V. (1991). Classical Negation in Logic Programs and Disjunctive Databases. *New Generation Computing*, 9:365--385.

Gomide, J.; Veloso, A.; Jr., W. M.; Almeida, V.; Benevenuto, F.; Ferraz, F. & Teixeira, M. (2011). Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. Em *ACM Web Science Conference (WebSci)*, pp. 1--8.

Gonçalves-Almeida, V. M.; Pires, D. E. V.; Minardi, R. C. d. M.; da Silveira, C. H.; Jr., W. M. & Santoro, M. M. (2012). HydroPaCe: understanding and predicting cross-inhibition in serine proteases through hydrophobic patch centroids. *Bioinformatics*, 28(3):342--349.

Gordon, E. J. (2008). Banking on DrugBank. *ACS Chemical Biology*, 3(1):6.

Gottlieb, A.; Stein, G. Y.; Oron, Y.; Ruppin, E. & Sharan, R. (2012). INDI: a computational framework for inferring drug interactions and their associated recommendations. *Molecular Systems Biology*, 8(1).

Gray, D. L.; Ash, S. R.; Jacobi, J. & Michel, A. N. (1991). The training and use of an artificial neural network to monitor use of medication in treatment of complex patients. *J Clin Eng*, 16(4):331--336.

Grime, K.; Ferguson, D. D. & Riley, R. J. (2010). The use of HepaRG and human hepatocyte data in predicting CYP induction drug-drug interactions via static equation and dynamic mechanistic modelling approaches. *Curr. Drug Metab.*, 11(10):870--885.

Guerra Júnior, A. A.; Pereira, L. A. M.; Silva, G. D. d.; Faleiros, D. R.; Bontempo, V.; Macedo, R. C. R.; Andrade, W. W.; Souza Filho, H. C. R.; Figueiredo, F. A. S.; Almeida, R. N. d. & Almeida, A. F. S. (2008). *Rede Farmacia de Minas - Plano Estadual de Estruturação da Rede de Assistência Farmaceutica: uma estratégia para ampliar o acesso e o uso racional de medicamentos no SUS*. Autêntica.

Gurulingappa, H.; Toldo, L.; Rajput, A. M.; Kors, J. A.; Taweel, A. & Tayrouz, Y. (2013). Automatic detection of adverse events to predict drug label changes using text and data mining techniques. *Pharmacoepidemiology and Drug Safety*, pp. n/a--n/a.

Hampton, T. (2011). Data mining approach shows promise in detecting unexpected drug interactions. *JAMA : the journal of the American Medical Association*, 306(2):144.

Han, J.; Cheng, H.; Xin, D. & Yan, X. (2007). Frequent pattern mining: current status and future directions. *Data Mining Knowledge Discovery*.

Han, J. & Kamber, M. (2001). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

Han, X.; Wang, Z.; Subhadarshini, A.; Karnik, S.; Strother, R. M.; Hall, S. D.; Jin, Y.; Flockhart, D. A.; Quinney, S. K.; Duke, J. D. & Li, L. (2012). Novel translational paradigm for drug-drug interaction research: A combination of literature-based discovery, electronic medical records and in vitro DDI screening assays. *Clinical Pharmacology and Therapeutics*, 91:S2.

Harpaz, R.; Chase, H. S. & Friedman, C. (2010a). Mining multi-item drug adverse effect associations in spontaneous reporting systems. *BMC Bioinformatics*, 11 Suppl 9:S7.

Harpaz, R.; Haerian, K.; Chase, H. S. & Friedman, C. (2010b). Statistical Mining of Potential Drug Interaction Adverse Effects in FDA's Spontaneous Reporting System. *AMIA Annu Symp Proc*, 2010:281--285.

Hartge, F.; Wetter, T. & Haefeli, W. E. (2006). A similarity measure for case based reasoning modeling with temporal abstraction based on cross-correlation. *Computer Methods and Programs in Biomedicine*, 81(1):41--48.

Hazell, L. & Shakir, S. (2006). Under-Reporting of Adverse Drug Reactions. *Drug Safety*, 29(5):385--396.

Hemens, B. J.; Holbrook, A.; Tonkin, M.; Mackay, J. A.; Weise-Kelly, L.; Navarro, T.; Wilczynski, N. L. & and, R. B. H. (2011). Computerized clinical decision support systems for drug prescribing and management: a decision-maker-researcher partnership systematic review. *Implement Sci*, 6:89.

Higgins, J. P. & Green, S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions*. The Cochrane Collaboration, Oxford, version 5.1.0 edição.

Hong, W.; Nuwayhid, S. J. & Werling, L. L. (2004). Modulation of bradykinin-induced calcium changes in SH-SY5Y cells by neurosteroids and sigma receptor ligands via a shared mechanism. *Synapse*, 54(2):102--110.

Horn, F.; Weare, J.; Beukers, M. W.; Horsch, S.; Bairoch, A.; Chen, W.; Edvardsen, O.; Campagne, F. & Vriend, G. (1998). GPCRDB: An Information system for G protein-coupled receptors. *Nucleic Acids Res*, 26:294--297.

Hornby, A. & Wehmeier, S. (2007). *Oxford Advanced Learner's Dictionary of Current English*. Oxford University Press.

Hripcsak, G.; Clayton, P. D.; Jenders, R. A.; Cimino, J. J. & Johnson, S. B. (1996). Design of a clinical event monitor. *Comput. Biomed. Res.*, 29(3):194--221.

Huang, J.; Niu, C.; Green, C. D.; Yang, L.; Hongkang, M. & J., H. J.-D. (2013). Systematic Prediction of Pharmacodynamic Drug-Drug Interactions through Protein-Protein-Interaction Network. *PLoS Comput Biol*, 9(3):e1002998.

Hucka, M.; Finney, A.; Sauro, H. M.; Bolouri, H.; Doyle, J. C.; Kitano, H.; Arkin, A. P.; Bornstein, B. J.; Bray, D.; Cornish-Bowden, A.; Cuellar, A. A.; Dronov, S.; Gilles, E. D.; Ginkel, M.; Gor, V.; Goryanin, I.; Hedley, W. J.; Hodgman, T. C.; Hofmeyr, J. H.; Hunter, P. J.; Juty, N. S.; Kasberger, J. L.; Kremling, A.; Kummer, U.; Novare, N. L.; Loew, L. M.; Lucio, D.; Mendes, P.; Minch, E.; Mjolsness, E.; Nakayama, Y.; Nelson, M. R.; Nielsen, P. M. F.; Sakurada, T.; Schaff, J. C.; Shapiro, B. E.; Shimizu, T. S.; Spence, H. D.; Stelling, J.; Takahashi, K.; Tomita, M.; Wagner, J. & Wang, J. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524--531.

Jacomini, L. C. L. & da Silva, T. M. (2011). *Interação Medicamentosa - Celmo Celeno Porto*. Guanabara Koogan.

Jaspers, M. W. M.; Smeulers, M.; Vermeulen, H. & Peute, L. W. P. (2011). Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. *JAMIA*, 18(3):327--334.

Ji, Z. L.; Han, L. Y.; Yap, C. W.; Sun, L. Z.; Chen, X. & Chen, Y. Z. (2003). Drug Adverse Reaction Target Database (DART) : proteins related to adverse drug reactions. *Drug safety*, 10:685--90.

Jin, H.-M. & Pan, Y. (2007). Angiotensin type-1 receptor blockade with losartan increases insulin sensitivity and improves glucose homeostasis in subjects with type 2 diabetes and nephropathy. *Nephrology Dialysis Transplantation*, 22(7):1943--1949.

Kam, H. J.; Kim, J. A.; Cho, I.; Kim, Y. & Park, R. W. (2011). Integration of heterogeneous clinical decision support systems and their knowledge sets: feasibility study with Drug-Drug Interaction alerts. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2011:664--673.

Kanehisa, M. (2013). Molecular network analysis of diseases and drugs in KEGG. *Methods Mol. Biol.*, 939:263--275.

Kanehisa, M.; Goto, S.; Furumichi, M.; Tanabe, M. & Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*, 38(Database issue):D355--60.

Katzung, B. G., editor (2003). *Farmacologia Básica e Clínica*. Guanabara Koogan S.A., 8 edição.

Kaur, G. & Kulkarni, S. K. (2002). Evidence for serotonergic modulation of progesterone-induced hyperphagia, depression and algesia in female mice. *Brain Research*, 943(2):206--215.

Kawamoto, K.; Houlihan, C. A.; Balas, E. A. & Lobach, D. F. (2005). Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ (Clinical research ed.)*, 330(7494):765+.

Kinney, E. L. (1986). Expert system detection of drug interactions: Results in consecutive inpatients. *Computers and Biomedical Research*, 19(5):462--467.

Kitano, H. (2002a). Computational systems biology. *Nature*, 420(6912):206--210.

Kitano, H. (2002b). Systems Biology: A Brief Overview. *Science*, 295(5560):1662--1664.

Klein, T. E.; Chang, J. T.; Cho, M. K.; Easton, K. L.; Fergerson, R.; Hewett, M.; Lin, Z.; Liu, Y.; Liu, S.; Oliver, D. E.; Rubin, D. L.; Shafa, F.; Stuart, J. M. & Altman, R. B. (2001). Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. *The pharmacogenomics journal*, 1(3):167--170.

Krauthammer, M. & Nenadic, G. (2004). Term identification in the biomedical literature. *J. of Biomedical Informatics*, 37(6):512--526.

Krejsa, C. M.; Horvath, D.; Rogalski, S. L.; Penzotti, J. E.; Mao, B.; Barbosa, F. & Migeon, J. C. (2003). Predicting ADME properties and side effects: the BioPrint approach. *Curr Opin Drug Discov Devel*, 6(4):470--480.

Kriegel, H.-P.; Borgwardt, K. M.; Kroger, P.; Pryakhin, A.; Schubert, M. & Zimek, A. (2007). Future trends in data mining. *Data Mining Knowledge Discovery*.

Kriete, A. & Eils, R. (2006). Chapter 1 - Introducing Computational Systems Biology. Em Kriete, A. & Eils, R., editores, *Computational Systems Biology*, pp. 1--14. Academic Press, Burlington.

Kuperman, G. J.; Bates, D. W.; Teich, J. M.; Schneider, J. R. & Cheiman, D. (1994). A new knowledge structure for drug-drug interactions. *Proc Annu Symp Comput Appl Med Care*, pp. 836--840.

Landis, J. R. & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159--174.

Laporte, J. & G.Tognoni (2007). *Principios de epidemiología del medicamento*. Masson-Salvat, 2 edição.

Lee, A. (2009). *Reações adversas a medicamentos*. ArtMed, 2 edição.

Leemann, T.; Transon, C. & Dayer, P. (1993). Cytochrome P450TB (CYP2C): A major monooxygenase catalyzing diclofenac 4'-hydroxylation in human liver. *Life Sciences*, 52(1):29--34.

Lemos, L. L. P.; Acurcio, F. D. A.; Almeida, A. M.; Araújo, V. E.; Barbosa, M. M.; Machado, M. A. A.; Costa, J. D. O. & Kakehasi, A. M. (2013). Rituximabe para o tratamento da artrite reumatoide: revisão sistemática. *Revista Brasileira de Reumatologi*.

Leone, R.; Magro, L.; Moretti, U.; Cutroneo, P.; Moschini, M.; Motola, D.; Tuccori, M. & Conforti, A. (2010). Identifying adverse drug reactions associated with drug-drug interactions: Data mining of a spontaneous reporting database in Italy. *Drug Safety*, 33(8):667--675.

Liamis, G.; Rodenburg, E. M.; Hofman, A.; Zietse, R.; Stricker, B. H. & Hoorn, E. J. (2013). Electrolyte Disorders in Community Subjects: Prevalence and Risk Factors. *The American Journal of Medicine*, 126(3):256--263.

Liberati, A.; Altman, D. G.; Tetzlaff, J.; Mulrow, C.; Gøtzsche, P. C.; Ioannidis, J. P. A.; Clarke, M.; Devereaux, P. J.; Kleijnen, J. & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ*, 339.

Lin, F. P.; Anthony, S.; Polasek, T. M.; Tsafnat, G. & Doogue, M. P. (2011). BICEPP: an example-based statistical text mining method for predicting the binary characteristics of drugs. *BMC Bioinformatics*, 12:112.

Lin, M.; Li, H.; Hou, W.; Johnson, J. A. & Wu, R. (2007). Modeling sequence-sequence interactions for drug response. *Bioinformatics*, 23(10):1251--7.

Lin, S.-F.; Xiao, K.-T.; Huang, Y.-T.; Chiu, C.-C. & Soo, V.-W. (2010). Analysis of adverse drug reactions using drug and drug target interactions and graph-based methods. *Artificial Intelligence in Medicine*, 48(2-3):161--166.

Linnarsson, R. (1993). Drug interactions in primary health care: A retrospective database study and its implications for the design of a computerized decision support system. *Scandinavian journal of primary health care*, 11(3):181--186.

Lipscomb, C. E. (2000). Medical Subject Headings (MeSH). *Bull Med Libr Assoc.* 88(3): 265–266.

Loyola Filho, A. I. d.; Elizabeth, U.; Firmo Josélia, O. A. & Lima-Costa, M. F. (2008). Influência da renda na associação entre disfunção cognitiva e polifarmácia: Projeto Bambuí. *Revista de Saúde Pública*, 42:89--99.

MacCuish, J. D. & MacCuish, N. E. (2011). Clustering in Bioinformatics and Drug Discovery.

Machado, M. A. A.; Maciel, A. A.; Pires, L. L. L.; Oliveira, C. J. D.; Maria, K. A.; Gurgel, A. E. I.; Leal, C. M. & Assis, A. F. D. (2013). Adalimumabe no tratamento da artrite reumatoide: uma revisão sistemática e metanálise de ensaios clínicos randomizados. *Revista Brasileira de Reumatologia*.

Mallik, R.; Yoo, M. J.; Chen, S. & Hage, D. S. (2008). Studies of verapamil binding to human serum albumin by high-performance affinity chromatography . *Journal of Chromatography B*, 876(1):69--75.

Mann, R. & Andrews, E. (2007). *Pharmacovigilance*. Wiley.

McGuinness, D. L. & van Harmelen, F. (2004). OWL Web Ontology Language Overview. W3C recommendation, W3C. http://www.w3.org/TR/2004/REC-owl-features-20040210/.

Michielan, L.; Stephanie, F.; Terfloth, L.; Hristozov, D.; Cacciari, B.; Klotz, K. N.; Spalluto, G.; Gasteiger, J. & Moro, S. (2009). Exploring potency and selectivity receptor antagonist profiles using a multilabel classification approach: the human adenosine receptors as a key study. *journal of chemical information and modeling*, 49(12):2820--2836.

Micromedex (2013). Healthcare Series [Internet database].

Milreu, P. V. (2008). *Análise de nutrientes utilizando redes metabólicas*. Tese de doutorado, Universidade Federal de Mato Grosso do Sul.

Miranda, H. F. & Pinardi, G. (2009). Lack of effect of naltrindole on the spinal synergism of morphine and non-steroidal anti-inflammatory drugs (NSAIDS). *J Physiol Pharmacol*, 60(2):71--6.

Mithani, A.; Preston, G. M. & Hein, J. (2009). Rahnuma: hypergraph-based tool for metabolic pathway prediction and network comparison. *Bioinformatics*, 25(14):1831--1832.

Moura, C.; Prado, N. & Acurcio, F. (2011). Potential drug-drug interactions associated with prolonged stays in the intensive care unit: a retrospective cohort study. *Clin Drug Investig*, 31(5):309--316.

Nelson, D.; Lehninger, A. & Cox, M. (2013). *Lehninger Principles of Biochemistry*. W.H. Freeman.

Novere, N. L. & Changeux, J.-P. (1999). The Ligand Gated Ion Channel Database. *Nucleic Acids Research*, 27(1):340--342.

Oransay, K.; Kalkan, S.; Hocaoglu, N.; Arici, A. & Tuncok, Y. (2011). An alternative antidote therapy in amitriptyline-induced rat toxicity model: theophylline. *Drug and Chemical Toxicology*, 34(1):53--60. PMID: 20954804.

Page, D. C.; Costa, V. S.; Natarajan, S.; Barnard, A.; Peissig, P. & Caldwell, M. (2012). Identifying Adverse Drug Events by Relational Learning. Em *Twenty-Sixth Conference on Artificial Intelligence (AAAI-12)*, Toronto, Canada.

Pasina, L.; Djade, C. D.; Nobili, A.; Tettamanti, M.; Franchi, C.; Salerno, F.; Corrao, S.; Marengoni, A.; Iorio, A.; Marcucci, M. & Mannucci, P. M. (2013). Drug-drug interactions in a cohort of hospitalized elderly patients. *Pharmacoepidemiology and Drug Safety*, pp. n/a--n/a.

Pavlopoulos, G. A.; Secrier, M.; Moschopoulos, C. N.; Soldatos, T. G.; Kossida, S.; Aerts, J.; Schneider, R. & Bagos, P. G. (2011). Using graph theory to analyze biological networks. *BioData Min*, 4:10.

Peng, C. C.; Glassman, P. A.; Marks, I. R.; Fowler, C.; Castiglione, B. & Good, C. B. (2003). Retrospective drug utilization review: incidence of clinically relevant potential drug-drug interactions in a large ambulatory population. *J Manag Care Pharm*, 9(6):513--22.

Peng, Y.; Zhang, Y. & Wang, L. (2010). Artificial intelligence in biomedical engineering and informatics: An introduction and review. *Artificial Intelligence in Medicine*, 48(2-3):71--73.

Percha, B. & Altman, R. B. (2013). Informatics confronts drug-drug interactions. *Trends in Pharmacological Sciences*, 34(3):178--184.

Percha, B.; Garten, Y. & Altman, R. B. (2012). Discovery and explanation of drug-drug interactions via text mining. *Pac Symp Biocomput*, pp. 410--421.

Pinardi, G.; Prieto, J. C. & Miranda, H. F. (2005). Analgesic synergism between intrathecal morphine and cyclooxygenase-2 inhibitors in mice. *Pharmacology Biochemistry and Behavior*, 82(1):120--124.

Pinto, M. C. X.; Felipe, F. & Pimenta, P. M. L. (2012). Potentially inappropriate medication use in a city of Southeast Brazil. *Brazilian journal of Pharmaceutical Sciences*, 48:79--86.

Pinto, M. C. X.; Malaquias, D. P.; Ferré, F. & Pinheiro, M. L. P. (2013). Potentially inappropriate medication use among institutionalized elderly individuals in southeastern Brazil.

Pires, D. E. V.; Melo-Minardi, R. C.; Santos, M. A.; da Silveira, C. H.; Santoro, M. M. & Meira Junior, W. (2011). Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. *BMC Genomics*, 12 Suppl 4:S12.

Preferansky, N. G. (1992a). From information-retrieval to expert drug interaction system. *Farmatsiya*, 41(4):9--14.

Preferansky, N. G. (1992b). Knowledge presentation by using rules in the expert system 'Drug Interaction'. *Farmatsiya*, 41(3):8--12.

R Core Team (2013). The R Reference Index.

Ralph, E. D. & Amatnieks, Y. E. (1980). Potentially synergistic antimicrobial combinations with metronidazole against Bacteroides fragilis. *Antimicrob. Agents Chemother*, 13(3).

Rénéric, J.-P.; Bouvard, M. & Stinus, L. (2002). In the rat forced swimming test, chronic but not subacute administration of dual 5-HT/NA antidepressant treatments may produce greater effects than selective drugs. *Behavioural Brain Research*, 136(2):521--532.

Riedel, W.; Hogervorst, E.; Leboux, R.; Verhey, F.; van Praag, H. & Jolles, J. (1995). Caffeine attenuates scopolamine-induced memory impairment in humans. *Psychopharmacology*, 122:158--168.

Rozenfeld, S.; M., F. M. J. & A., A. F. (2008). Drug utilization and polypharmacy among the elderly: a survey in Rio de Janeiro City, Brazil. *Revista Panamericana de Salud Publica*, 23:34--43.

Rumbaugh, J.; Jacobson, I. & Booch, G. (2005). *The Unified Modeling Language Reference Manual*. Addison-Wesley, Boston, MA, 2. edição.

Russel, S. & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*. Pearson Education Inc.

Scheaffer, R.; Mendenhall, W.; Ott, R. & Gerow, K. (2011). *Elementary Survey Sampling*. Advanced series. Brooks/Cole.

Scheer, M.; Grote, A.; Chang, A.; Schomburg, I.; Munaretto, C.; Rother, M.; Söhngen, C.; Stelzer, M.; Thiele, J. & Schomburg, D. (2011). BRENDA, the enzyme information system in 2011. *Nucleic Acids Research*, 39:670--676.

Scilab Enterprises (2012). *Scilab: Free and Open Source software for numerical computation*. Scilab Enterprises, Orsay, France.

Segura-Bedmar, I.; Crespo, M.; de Pablo-Sanchez, C. & Martinez, P. (2010). Resolving anaphoras for the extraction of drug-drug interactions in pharmacological documents. *BMC bioinformatics*, 11 Suppl 2:S1.

Segura-Bedmar, I.; Martinez, P. & de Pablo-Sanchez, C. (2011a). A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents. *BMC bioinformatics*, 12 Suppl 2:S1.

Segura-Bedmar, I.; Martinez, P. & de Pablo-Sanchez, C. (2011b). Using a shallow linguistic kernel for drug-drug interaction extraction. *journal of Biomedical Informatics*, 44(5):789--804.

Seynaeve, C.; Mulder, P.; Verweij, J. & Gralla, R. (1991). Controlling cancer chemotherapy-induced emesis. *Pharmaceutisch Weekblad*, 13(5):189--197.

Sim, I.; Gorman, P.; Greenes, R.; Haynes, R.; Kaplan, B.; Lehmann, H. & Tang, P. (2001). Clinical Decision Support Systems for the Practice of Evidence-based Medicine. *Journal of the American Medical Informatics Association*, 8(6).

SINITOX (2013). Casos, Óbitos e Letalidade de Intoxicação Humana por Agente e por Região. Brasil, 2010. Acessado em 22/02/2013.

Sirgo, M.; Rocci Jr, M.; Ferguson, R.; Eshelman, F. & Vlasses, P. (1985). Effects of cimetidine and ranitidine on the conversion of prednisone to prednisolone. *Clin Pharmacol Ther*, 37(5):534--8.

Skrebuhhova-Malmros, T.; Allikmets, L. & Matto, V. (2001). Additive Effect of Clonidine and Fluoxetine on Apomorphine-Induced Aggressive Behavior in Adult Male Wistar Rats. *Archives of Medical Research*, 32(3):193--196.

Snyder, B. D.; Polasek, T. M. & Doogue, M. P. (2012). Drug interactions: principles and practice. *Autralian Prescriber*, 35(3).

Sojda, R. (2007). Empirical evaluation of decision support systems: Needs, definitions, potential methods, and an example pertaining to waterfowl management. *Environmental Modelling & Software*, 22(2):269--277.

Speedie, S. M.; McNally, D.; Skarupa, S.; Michocki, R.; Rudo, C.; Metge, C.; Palumbo, F. & Knapp, D. (1992). Evaluating drug prescribing in a large, ambulatory population: application of an embedded expert system. *Proc Annu Symp Comput Appl Med Care*, pp. 621--625.

Stephens, M. (2005). *Appendix I: Drug Products Withdrawn from the Market for Safety Reasons*, pp. 667--702. John Wiley & Sons, Ltd.

Strandell, J.; Noren, N. G. & Hägg, S. (2013). Key Elements in Adverse Drug Interaction Safety Signals An Assessment of Individual Case Safety Reports. *Drug Safety*, 36(1):63--70.

Strom, B. & Kimmel, S. (2007). *Textbook of Pharmacoepidemiology*. Wiley.

Sucher, J. F.; Moore, F. A.; Todd, S. R.; Sailors, R. M. & McKinley, B. A. (2008). Computerized clinical decision support: a technology to implement and validate evidence based guidelines. *J Trauma*, 64(2):520--37.

Sun, L. Z.; Ji, Z. L.; Chen, X.; Wang, J. F. & Chen, Y. Z. (2002). ADME-AP: a database of ADME associated proteins. *Bioinformatics*, 18(12):1699--1700.

Szarfman, A.; Machado, S. G. & O'Neill, R. T. (2002). Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. *Drug Saf*, 25(6):381--392.

Takagi, H. & Umemoto, T. (2012). Telmisartan improves insulin sensitivity: A meta-analysis of randomized head-to-head trials. *International Journal of Cardiology*, 156(1):92--96.

Takarabe, M.; Shigemizu, D.; Kotera, M.; Goto, S. & Kanehisa, M. (2011). Network-based analysis and characterization of adverse drug-drug interactions. *journal of chemical information and modeling*, 51(11):2977--2985.

Takigawa, I.; Tsuda, K. & Mamitsuka, H. (2011). Mining significant substructure pairs for interpreting polypharmacology in drug-target network. *PloS one*, 6(2):e16999.

Taksande, B. G.; Kotagale, N. R.; Tripathi, S. J.; Ugale, R. R. & Chopde, C. T. (2009). Antidepressant like effect of selective serotonin reuptake inhibitors involve modulation of imidazoline receptors by agmatine. *Neuropharmacology*, 57(4):415--424.

Tan, P.-N.; Steinbach, M. & Kumar, V. (2005). *Introduction to Data Mining*. Addison-Wesley.

Tanjong-Ghogomu, E.; Tugwell, P. & Welch, V. (2009). Evidence-based medicine and the Cochrane Collaboration. 67:198--205+.

Tari, L.; Anwar, S.; Liang, S.; Cai, J. & Baral, C. (2010). Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. *Bioinformatics (Oxford, England)*, 26(18):i547--553.

Tari, L.; Hakenberg, J.; Gonzalez, G. & Baral, C. (2009). Querying parse tree database of Medline text to synthesize user-specific biomolecular networks. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 87--98.

Tatro, D. (2012). *Drug Interaction Facts 2013: The Authority on Drug Interactions*. Drug Interaction Facts. Lippincott Williams & Wilkins.

Torii, M.; Kamboj, S. & Vijay-Shanker, K. (2004). Using name-internal and contextual features to classify biological terms. *Journal of Biomedical Informatics*, 37(6):498--511.

Troiano, D.; Jones, M. A.; Smith, A. H.; Chan, R. C.; Laegeler, A. P.; Le, T.; Flynn, A. & Chaffee, B. W. (2013). The need for collaborative engagement in creating clinical decision-support alerts. *Am J Health Syst Pharm*, 70(2):150--3.

van Puijenbroek, E. P.; Bate, A.; Leufkens, H. G.; Lindquist, M.; Orre, R. & Egberts, A. C. (2002). A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiology and drug safety*, 11(1):3--10.

Veloso, A.; Jr., W. M. & Zaki, M. J. (2006). Lazy Associative Classification. Em *ICDM*, pp. 645--654.

Vilar, S.; Harpaz, R.; Uriarte, E.; Santana, L.; Rabadan, R. & Friedman, C. (2012). Drug–drug interaction through molecular structure similarity analysis. *journal of the American Medical Informatics Association*, 19(6):1066--1074.

Villacorta Linaza, P.; Ruano Camps, R.; Gallego Fernández, C.; Santos Ramos, B.; Rodríguez Terol, A. & Camacho, C. (2010). Calidad de las bases de datos sobre interacciones de antirretrovirales. *Medicina Clínica*, 134(15):678--683.

Villier, C.; Schir, E.; Logerot, S. & Mallaret, M. (2012). Drug interactions with colchicine: Results from a local data mining. *Fundamental and Clinical Pharmacology*, 26:74.

Vonbach, P. (2007). *Drug-Drug Interactions in the Hospital*. Tese de doutorado, Fakultät der Universität Basel.

Vroling, B.; Thorne, D.; McDermott, P.; Joosten, H.-J.; Attwood, T. K.; Pettifer, S. & Vriend, G. (2012). NucleaRDB: information system for nuclear receptors. *Nucleic Acids Research*, 40:377--380.

Walton-Shirley, M. (2013). Drug-Drug Interactions: Why There Was Standing Room Only.

Wang, Y.-C.; Chen, S.-L.; Deng, N.-Y. & Wang, Y. (2013). Network predicting drug's anatomical therapeutic chemical code. *Bioinformatics*, 29(10):1317--1324.

Wang, Y. H.; Li, Y.; Yang, S. L. & Yang, L. (2005). Classification of substrates and inhibitors of P-glycoprotein using unsupervised machine learning approach. *journal of chemical information and modeling*, 45(3):750--757.

Whiting, P.; Rutjes, A. W. S.; Dinnes, J.; Reitsma, J.; Bossuyt, P. M. M. & Kleijnen, J. (2004). Development and validation of methods for assessing the quality of diagnostic accuracy studies. *Health technology assessment*, 8:iii, 1--234.

WHO (2011). *Guidelines for ATC classification and DDD assignment*. WHO Collaborating Centre for Drug Statistics Methodology.

Widenius, M.; Axmark, D. & Mysql, A. B. (2002). *MySQL Reference Manual*. O'Reilly Media, Inc., 1 edição.

Wilk, S.; Michalowski, W.; Michalowski, M.; Farion, K.; Hing, M. M. & Mohapatra, S. (2013). Mitigation of adverse interactions in pairs of clinical practice guidelines using constraint logic programming. *Journal of Biomedical Informatics*, 46(2):341--353.

Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B. & Hassanali, M. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, 36(Database issue):D901--906.

Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z. & Woolsey, J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res*, 34(Database issue).

Witten, I. & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science.

Witten, I. H.; Bray, Z.; Mahoui, M. & Teahan, B. (1999). Text Mining: A new frontier for lossless compression. Em *In Data Compression Conference*, pp. 198--207. IEEE Press.

Wong, K. K. K.; Ngo, J. C. K.; Liu, S.; Lin, H.-Q.; Hu, C.; Shaw, P. & Wan, D. C. C. (2010). Interaction study of two diterpenes, cryptotanshinone and dihydrotanshinone, to human acetylcholinesterase and butyrylcholinesterase by molecular docking and kinetic analysis. *Chemico-Biological Interactions*, 187(1-3):335--339.

Yap, C. W.; Xue, Y.; Li, Z. R. & Chen, Y. Z. (2006). Application of support vector machines to in silico prediction of cytochrome p450 enzyme substrates and inhibitors. *Curr Top Med Chem*, 6(15):1593--1607.

Yoon, D.; Park, M. Y. & Park, R. W. (2011). Detection of drug-drug interactions from spontaneous reporting system data by multifactor dimensionality reduction. *Pharmacoepidemiology and Drug Safety*, 20:S350.

Yoshikawa, S.; Satou, K. & Konagaya, A. (2004). Drug interaction ontology (DIO) for inferences of possible drug-drug interactions. *Stud Health Technol Inform*, 107(Pt 1):454--8.

Zaki, M. J. & Meira Jr, W. (2014). *Fundamentals of Data Mining Algorithms*, volume 1. Cambridge University Press.

Zhang, J.; Jia, J.; Zhu, F.; Ma, X.; Han, B.; Wei, X.; Tan, C.; Jiang, Y. & Chen, Y. (2012a). Analysis of bypass signaling in EGFR pathway and profiling of bypass genes for predicting response to anticancer EGFR tyrosine kinase inhibitors. *Molecular bioSystems*, 8(10):2645--2656.

Zhang, Y.; Lin, H.; Yang, Z.; Wang, J. & Li, Y. (2012b). A single kernel-based approach to extract drug-drug interactions from biomedical literature. *PLoS ONE*, 7(11):e48901.

# Apêndice A

# Referencial teórico complementar

## A.1 Experimentação Científica na Saúde

Desde o iluminismo, vigora na ciência contemporânea o paradigma cartesiano de causa e efeito. Tradicionalmente, qualquer estudo científico possui três elementos fundamentais: o objeto, o agente e o ato. Por exemplo, diante de informações coletada a partir do **objeto** "paciente", deseja-se determinar o *agente* "associação de fármacos" responsável por causar o *ato* evento tóxico ou terapêutico. A lei que descreve esta correlação deve ser passível de reprodução sob determinadas condições para cada um dos três elementos, as quais contemplam a explicação necessária para afirmativas sob o rigor científico.

Existem três domínios empíricos que fomentam a prática da saúde baseada em evidência: *in vitro*, *in vivo* e *in populo*. Estas áreas são insumo para um quarto domínio, o *in silico*, o qual vem se afigurando como provedor de conhecimento, embora ainda não paute decisões sem corroboração dos demais domínios.

A figura A.1 exemplifica disciplinas em cada domínio empírico na área de interações medicamentosas. Nesta figura o agente e o objeto se alternam entre o paciente e o medicamento conforme o ponto de vista adotado.

### A.1.1 Pesquisa e desenvolvimento de fármacos

#### A.1.1.1 Pesquisa básica

Neste domínio são avaliados aspectos farmacológicos, fisiológicos, estados patológicos e produzidas novas tecnologias farmacêuticas.

A etiologia dos estados patológicos deve ser compreendida para a avaliação da resposta do corpo à doença e da doença à intervenção. A pesquisa básica é o principal ímpeto para a compreensão do corpo humano e o desenvolvimento da medicina moderna[Berger et al., 2009].

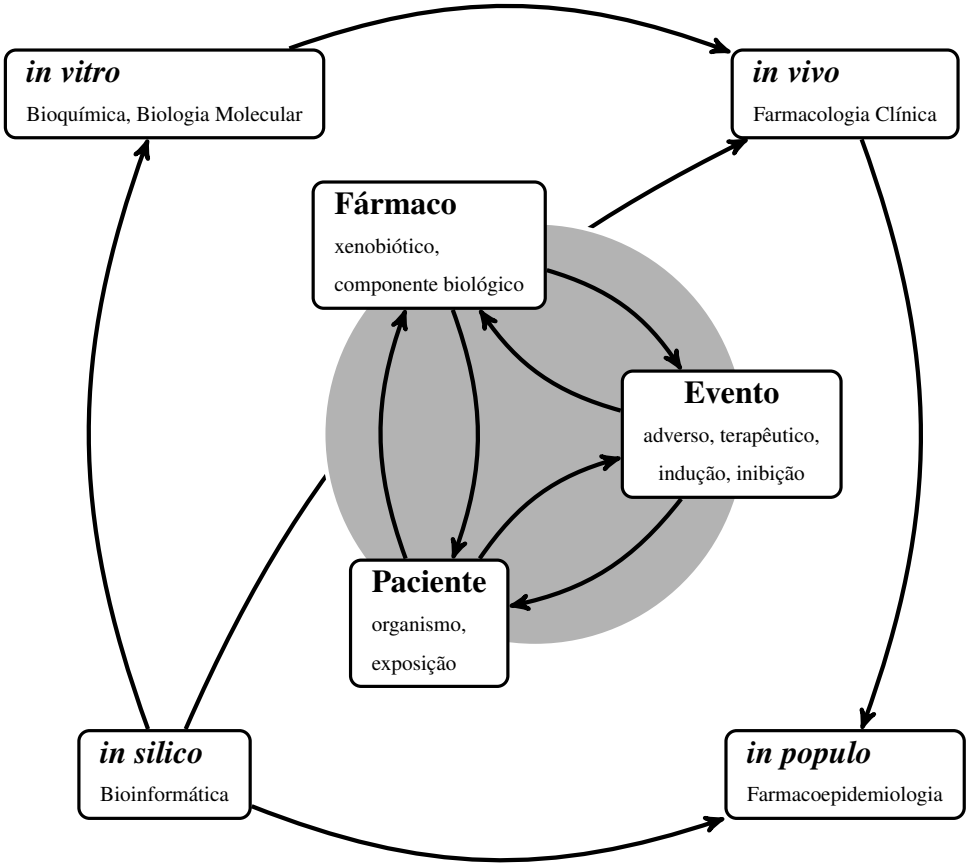



Figura A.1: **Domínios da ciência empírica.** O trajeto na determinação de eventos como efeito da exposição a um agente farmacológico, tradicionalmente inicia com abordagens *in vitro* ou *in vivo* cuja caraterística principal é a observação controlada de fenômenos isolados. Devido à presença dos fatores de variabilidade que prejudicam amostras, a generalização do efeito previamente observado culmina com a prevalência verificada em estudos populacionais. As etapas são potencializadas com o advento computacional, sobretudo diante da elevada processividade de casos e variáveis. Não obstante, a experimentação *in silico* pode substituir etapas, ou mesmo, ser fomentadora de novos conhecimentos a serem corroborados pelas demais áreas.

### A.1.1.2 Pesquisa pré-clínica

Os ensaios pré-clínicos contemplam o âmbito laboratorial e estudos em pelo menos duas espécies animais. É uma investigação focada na segurança. Os **modelos animais** procuram reproduzir as condições de saúde manifestadas em humanos e são usados para desenhar os primeiros ensaios clínicos.

Embora controverso, a importância dos testes em animais se deve ao uso de substâncias desconhecidas em humanos ser considerado antiético. No entanto, a extensão dos modelos animais é limitada devido às diferenças fisiológicas.

#### A.1.1.3 Pesquisa clínica

A partir das formas farmacêuticas eleitas (seção 2.2) quatro fases conduzem os experimento em humanos.

**Fase I** Testes realizados em voluntários sadios (por exemplo, entre 20 e 80). Objetiva-se identificar aspectos de segurança não evidenciados anteriormente, rotas metabólicas, vias de administração e efeitos biológicos.

**Fase II** Estudos prova de conceito [1] são realizados em cerca de 75 a 100 pacientes das populações-alvo. Adquirem-se as primeiras evidências de eficácia em humanos e dosagens terapêuticas.

**Fase III** Estudos conduzidos em alguns milhares de pacientes da população-alvo. São evidenciadas as características necessárias à aprovação regulatória ou novas indicações terapêuticas. Usualmente são realizados ensaios clínicos randomizados e controlados contra placebo ou terapia alternativa. São monitorados desfechos clínicos como acidente vascular cerebral, biomarcadores como colesterol e pressão arterial, qualidade de vida, entre outros.

**Fase IV** São os estudos pós-comercialização. São observados usos ***off label***, ou seja, não indicados para as codições-alvo; interações medicamentosas, dado que a polifarmácia não é mais controlada; eventos raros e diferenças no perfil de eficácia e segurança em subpopulações; variações na dosagem não contempladas anteriormente; entre outros.

### A.1.2 Evidência e relação causal

#### A.1.2.1 Estabelecimento da força da evidência

A **evidência direta** advém de estudos, randomizados ou não, em que a associação probabilística entre a intervenção e o desfecho é causal e não espúria.

A **evidência mecanicística** alega que o processo causal conecta a intervenção com o desfecho, e a ausência de plausibilidade química, biológica ou mecânica sugere a interação.

A **evidência paralela** estabelece a relação causal de uma hipótese sugerida em um estudo confrontada com estudos correlatos, verificando-se a consistência pela replicabilidade e a analogia dos efeitos e intervenções.

[1] Testes de prova de conceito são a implementação breve e/ou incompleta de um certo método ou ideia para demonstrar sua exequibilidade, ou uma demonstração em princípio, cujo propósito é verificar que algum conceito ou teoria é, provavelmente, capaz de ser explorado de uma forma útil. A prova de conceito é, usualmente, considerada um marco no caminho de um protótipo que ilustre plenamente o funcionamento do conceito ou mecanismo *sub judice*[Berger et al., 2009].

### A.1.2.2 Tipos de erros no estabelecimento de causação na associação de fatores

São apontados dois tipos de erros para no estabelecimento de uma associação de fatores. O **artefatual** ocorre pela chance (associação espúria ou falsa) ou com viés (variação sistemática). O erro **indireto** ocorre através da confusão.

A atribuição correta estabelece como **independente** os elementos associados (nenhuma associação) ou **causal** (direta, verdadeira).

O domínio *in vitro* constitui o cerne da fase pré-clínica, porém as técnicas e resultados estendem-se para os modelos *in vivo* inicialmente em animais e posteriormente em humanos.

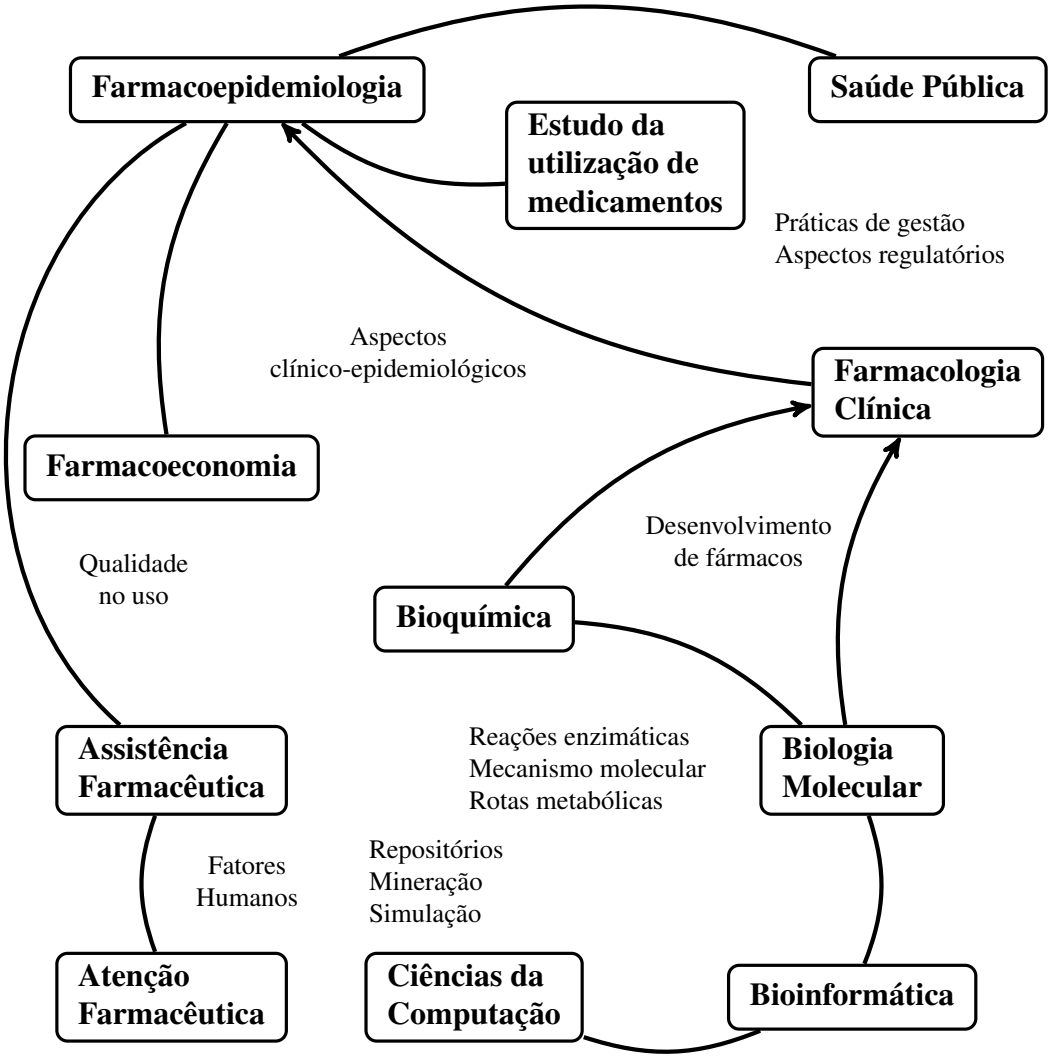



Figura A.2: O estudo de eventos, adversos ou terapêuticos relacionados à medicação e, consequentemente, o da associação de fármaocs envolve as áreas de conhecimento e técnicas citadas, cujo cerne é a atribuição de eventos ao uso de fármacos em humanos.

## A.2 O domínio *in vitro*

A introdução de uma substância não sintetizada pelo organismo, chamada **xenobiótico**, pode desencadear uma série de reações que, em última instância, tangem o domínio molecular. Desta forma, caracterizar a unidade fundamental dos seres vivos em função de sua estrutura contribui para o entendimento dos fenômenos que o uso de duas ou mais substâncias estranhas podem causar entre si, tomando como meio a célula e os tecidos.

### A.2.1 Biologia celular

As informações a seguir podem ser encontradas em livros-texto de Alberts et al. [2002] e Nelson et al. [2013].

A célula é delimitada por uma membrana permeável a solventes, íons e moléculas estruturada como uma **miscela**, cujo arcabouço é sustentado por um **citoesqueleto**.

A comunicação intra e intercelular é fundamental para a definição do papel no organismo e manutenção do ciclo celular que se encarrega das tarefas regulares durante a meia vida da célula, chamada **intérfase**, replicação e morte celular programada, a **apoptose**[2].

O solvente é o principal intermediário na comunicação intra e intercelular. O **sinal** frequentemente é uma biomolécula ou potencial de ação por gradientes de íons. A estrutura capacitada para captar o sinal é chamada de **receptor**.

A semântica dos sinais trocados dentro e fora da célula, basicamente expressa a produção ou **anabólise** de biomoléculas. Outra consequência da sinalização é ordem para a quebra ou **catálise** de moléculas, sobretudo para a geração de energia ou eliminação quando a molécula não for mais demandada.

Funções celulares se tornam mais especializadas ao observar-se organismos mais complexos. Desta forma, as células eucarióticas[3] incorporaram as **mitocôndrias** para auxiliar na geração de grande parte da energia pelo processo de oxidação ou **aeróbico**.

O núcleo abriga as informações de replicação das moléculas estruturais que migram pelo solvente intracelular, o **citoplasma**, para os **ribossomos** que são organelas responsáveis por decodificar estas mensagens e construir novas biomoléculas. Se estas biomoléculas são produzidas para exportação, o **complexo de Golgi** pode se encarregar de empacotá-las na forma de **vesículas**. Se demanda-se destruição de moléculas, o **lisossomo** é a estrutura que contém diversas enzimas para este fim. A grande esteira de transporte é o **retículo endoplasmático**, que pode ou não abrigar enzimas e ribossomos em suas paredes como um *pass-through*[4], contribuindo para construção ou degradação de biomoléculas. O **núcleo** abriga o repositório de

[2] A necrose é um processo patológico de morte celular em que a taxa de renovação do tecido tende a zero, ao contrário da apoptose que promove a renovação do tecido com a substituição por células novas.

[3] Plantas, animais e fungos.

[4] São aberturas nas paredes de linhas produtivas por onde os produtos são transferidos de um setor para outro por esteiras ou rolamentos.

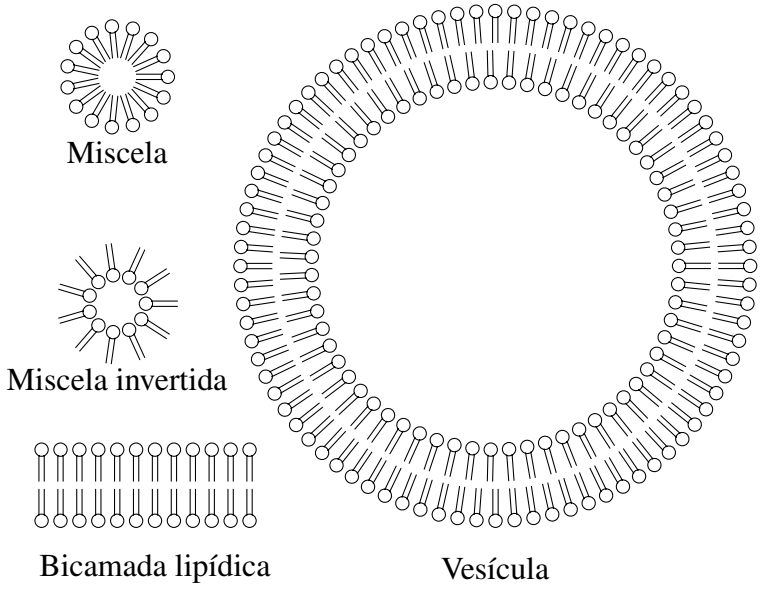



Figura A.3: **Miscela.** As células e vesículas são delimitadas por bicamadas lipídicas em que a porção apolar, dotada de uma cadeia de carbonos, encontra-se internalizada a esta camada, e a porção polar é voltada para o meio intracelular e extracelular. As miscelas e miscelas invertidas podem transportar em seu interior substâncias apolares e polares, respectivamente.

informações acerca da síntese da maior parte das biomoléculas do organismo em estruturas de DNA altamente enoveladas chamadas cromossomos.

#### A.2.1.1 Solventes, íons, moléculas e biomoléculas

As funções elementares das biomoléculas são **estrutural**, **energética** ou **sinal de comunicação**. No organismo as moléculas transitam predominantemente pelo **plasma**, que embora constituído em grande parte por água, possui em sua constituição outras biomoléculas dissolvidas.

**O meio** Desta forma, os elementos se dissolvem ou se dispersam conforme sua afinidade pelo solvente ou tamanho. Devido a base aquosa do plasma, as moléculas nele dissolvidas são de natureza **polar**, ou seja, dotadas de carga. As moléculas apolares tendem às propriedades lipídicas (por exemplo a gordura e o azeite) e são repelidas pelo meio aquoso, porém, de modo a não se dissolverem. No entanto, existem estruturas híbridas, como os fosfolipídeos da membrana, que são capazes de organizarem-se de modo a formar miscelas (figura A.3). Estas estruturas possuem uma propriedade chamada **anfótera** por abrigar características polares e apolares. As diferentes camadas formadas podem abrigar estruturas apolares de outros lipídeos ou proteínas, ao mesmo tempo que emergem as estruturas polares para o meio ou para o interior. Invertendo-se a polaridade, o núcleo da miscela pode abrigar moléculas apolares.

**Função estrutural** As proteínas são as principais moléculas que estruturam o organismo. Constituídas de aminoácidos, estruturam o citoesqueleto, formam a actina e miosina dos músculos; a queratina da pele, unha e cabelos, e as organelas. As enzimas são proteínas que realizam

**catálise** das reações, sem as quais o tempo e a energia demandadas tornaria inviável grande parte dos processos celulares de quebra ou formação de biomoléculas. **Carreadores**, como a albumina, são proteínas presentes no sangue capazes de transportar moléculas. A hemoglobina é uma **metaloproteína alostérica**[5].

**Função energética** Muita energia é gasta pela célula para manutenção dos gradientes iônicos e metabólitos através das membranas que viabilizam a atividade elétrica de células excitáveis[Fall et al., 2002]. Embora qualquer biomolécula possa ser quebrada para gerar energia, protagonizam os carboidrados ou açúcares como energia imediatamente disponível e os lipídeos como energia de armazenamento. Intermediárias em reações energéticas, as coenzimas são estrutas híbridas como os ATP[6], NAD[7] ou FAD[8], formadas por íons, carboidratos e **ácidos nucleicos**.

**Sinal de comunicação** Os ácidos nucléicos **adenina** (A), **citosina** (T), **guanina** (G) e **timina** (T) integram o DNA[9] e possuem uma correspondência direta com a decodificação de proteínas, pois o agrupamento de três aminoácidos formam um **códon** que expressa um aminoácido. Esta expressão, embora degenerada, isto é, com combinações que não expressam aminoácidos ou que expressam mais de um, é altamente eficiente havendo diversos pontos de controle que culmina na formação de uma proteína conforme as sequências de códons transcritas a RNA[10] são traduzidas pelos ribossomos. Hormônios, como a insulina, podem ser **peptídeos** que são constituídos por dois ou mais aminoácidos. Outra comunicação é a modificação da diferença de potencial da membrana pela intrusão ou extrusão de íons como sódio, potássio ou cálcio. Esta diferença de potencial é propagada segundo oscilação das cargas iônicas no citoplasma e no ambiente extracelular e aciona estruturas proteicas que atuam como canais eliminando ou incorporando **ativamente**, com gasto de anergia, ou **passivamente**, sem gasto de energia, moléculas para a manutenção da **homeostase**, ou equilíbrio, da célula e do organismo.

#### A.2.1.2 O fenômeno dinâmico

Em termos gerais refere-se a qualquer processo observado ao longo do tempo. As células são dinâmicas. Os **ciclos celulares** de crescimento, divisão, comunicação intra e intercelular, **movimentação e contração celular** exigem constante regulação de processos termodinâmicos para a homeostasia da célula.

[5] Metaloproteína é uma proteína ligada a um ou mais metais. Uma proteína alostérica possui diversos sítios de ligação cuja atividade é influenciada conforme a presença de ligantes.

[6] trifosfato de adenosina

[7] dinucleotídeo de nicotinamida-adenina

[8] dinucleotídeo de flavina-adenina

[9] Ácido desoxirribonucleico. As fitas duplas ligam-se respectivamente entre as purinas A e G e as pirimidinas T e C (A com T e G com C). A fita de nucleotídeos é estruturada pela ligação de desoxirriboses (açúcar de cinco carbonos) nas posições 5' e 3' por meio de um fosfato que confere a carga ácida do DNA.

[10] Ácido ribonucleico. Estrutura semelhante ao DNA, porém em fita simples tendo a timina substituída pela uracila

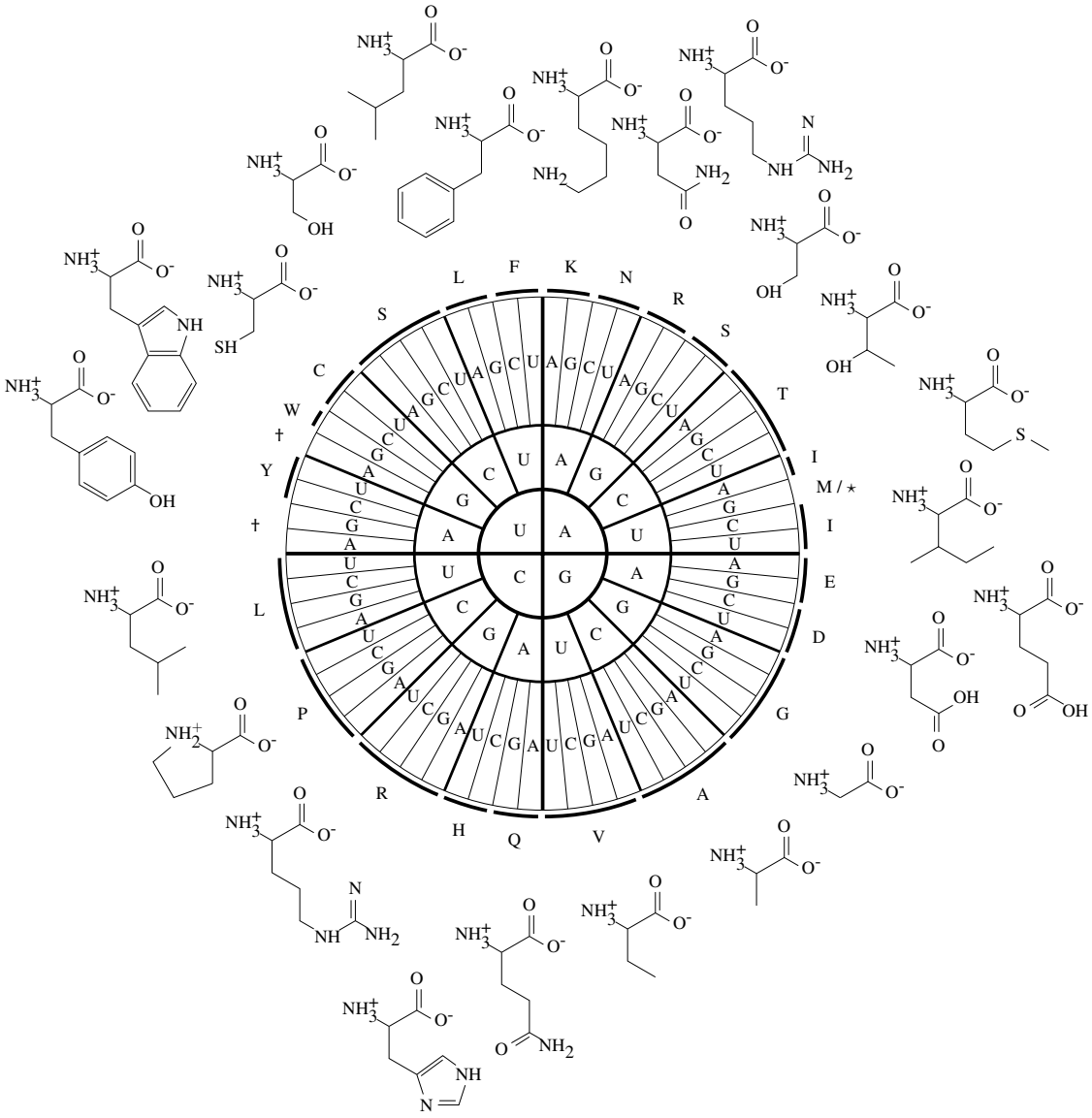


Figura A.4: **Bases nitrogenadas, códons e aminoácidos.** A combinação de três bases nitrogenadas, ou códon, transcrita a RNA por polimerases, são posteriormente traduzidas como um aminoácido. Os três círculos concêntricos devem ser lidos de modo centrífugo, por exemplo, o códon $UAA$ não possui significação, porém $UAC$ e $UAU$ são traduzidos ao aminoácido Y (tirosina). Esta imagem permite a observação da degeneração do código, dado que um código pode não expressar informação, e a redundância, ou seja, o códigos diferentes que expressam a mesma informação.

Ritmos circadianos são mudanças regulares nos processos celulares (do Latin *circa*, sobre, e *dies*, dia) em período de 24h. A modelagem deste relógio interno permite avaliar as adaptações da célula ao longo do dia, como, por exemplo, os fatores de emissão de neurotransmissores que induzem ao sono.

### A.2.2 Biologia de Sistemas

É a identificação dos elementos em um sistema e a análise de suas interrelações bem como a explicação das propriedades emergentes do sistema. Biologia de sistemas é a progressão natural da biologia molecular às ciências descritiva e qualitativa de redes biológicas de respostas dinâmicas[Kriete & Eils, 2006].

A biologia de sistemas culmina na integração de modelos de modo a gerar sequências de fenômenos conhecidas como **rotas**, as quais elucidam ciclos celulares, processos evolutivos, metabolismo, reações químicas em cadeia, doenças e síndromes de modo que possam ser visualizados e compreendidos.

Da associação com a bioinformática, surge o desafio em gerar estruturas de dados capazes de expressar o conhecimento sistêmico de modo a superar os limites da cognição humana, fator fundamental para entendimento da complexidade biológica.

## A.3 O domínio *in vivo*

Embora muitas conclusões possam ser tiradas ao se avaliar enzimas, células e tecidos isolados; no domínio *in vivo* ocorre a integração dos processos bioquímicos das células com a finalidade da manutenção das funções básicas do organismo. Esta integração é conhecida como **metabolismo**.

Esta seção introduz aspectos abordados em livros-texto de farmacologia básica como os escritor por Brunton et al. [2005] e Katzung [2003]. A **farmacologia básica** é segmentada em farmacocinética e farmacodinâmica.

### A.3.1 Farmacocinética

De modo geral a sigla ADME condensa os elementos básicos da farmacocinética: absorção, distribuição, metabolismo e excreção.

Cada fármaco realiza um trajeto característico para a entrada e saída do organismo. A partir da ingestão, a molécula ativa deve chegar ao sítio de ação. O organismo possui diversos mecanismos de defesa para substâncias estranhas constituindo compartimentos ou barreiras que devem ser vencidas pelo fármaco no trajeto ao sítio de ação de modo a torná-lo biodisponível. A biodisponibilidade é medida pela fração do fármaco que atinge a circulação sistêmica na forma

quimicamente inalterada[11]. O processo de transformação que o organismo impõe ao fármaco é estudado pela **farmacocinética** desde a **absorção** à **eliminação**.

#### A.3.1.1 Absorção ou Permeação

A passagem do fármaco ao longo das barreiras impostas pelo organismo pode ser realizada por difusão, transporte ou endocitose/exocitose. A **difusão** ocorre diante do gradiente de concentração em que solutos e solventes tendem a um determinado equilíbrio em um meio permeável.

Fármacos com peso molecular de 20 a 30 mil devem possuir alguma capacidade de difusão aquosa e lipídica para serem capazes de atravessar o plasma e as membranas celulares, conferida, em geral, segundo a capacidade de ionização enquanto ácidos ou bases fracas[12]. Moléculas maiores ou insolúveis são transportadas por carreadores ou transportadores de membrana conforme mencionado na seção A.2.1.1. Finalmente, algumas substâncias como o ferro e a vitamina $B_{12}$ são incorporadas por movimento de invaginação da membrana celular a qual engloba a molécula incorporando-a envolta por uma miscela.

#### A.3.1.2 Metabolismo e Eliminação

A excreção renal de um fármaco sem biotransformação é infrequente, visto que os fármacos geralmente apresentam peso molecular elevado ou grupos funcionais não ionizados ou parcialmente ionizados para facilitar a travessia por membranas.

Sem a eliminação, o tempo de circulação do fármaco no organismo, chamado meia-vida, poderia ser demasiado causando efeitos tóxicos. Desta forma, a biotransformação deve ser ponderada para a manutenção dos efeitos farmacológicos.

Todos os tecidos possuem alguma capacidade metabólica. A biotransformação ocorre na ingestão pelo trato gastrointestinal, inalação pelos pulmões ou, ainda, na passagem pelos rins. Porém no fígado ocorre a maior parte do metabolismo, devido ao complexo enzimático presente. Didaticamente, as reações são classificadas como de primeira passagem ou fase I, como o metabolismo da morfina que é absorvida inalterada e chegam diretamente ao fígado pelo sistema porta; e segunda passagem ou fase II (por exemplo, conjugação da acetilcisteína com glutationa), embora esta ordem não ocorra em todos os casos.

**Reações de fase I** Em geral, as reações desta fase convertem fármacos predominantemente lipofílicos (apolares) em substâncias mais polares ou hidrossolúveis por introdução de grupamentos $-OH$, $-NH_2$, $-SH$ que frequentemente os inativa.

[11] Se 100mg de um fármaco são administrados e 70mg atingirem a circulação, a biodisponibilidade será de 70%.

[12] Esta propriedade foi explorada inicialmente por Henderson-Hasselbalch em equações regidas por constantes de equilíbrio iônico como, p.ex. o ácido acetilsalicílico: $C_6H_7O_2COOH \overset{pk_a}{\rightleftharpoons} C_6H_7O_2COO^- + H^+ \therefore log\frac{protonado}{deprotonado} = pk_a - pH$

**Reações de fase II** Os fármacos que ainda não possuem polaridade suficiente para serem eliminados sofrem uma reação subsequente de conjugação do grupamento químico recém estabelecido com algum substrato endógeno, como ácido glicurônico, ácido sulfúrico, ácido acético ou aminoácido.

**Sistema microssomal, indução e inibição enzimática** Enzimas localizadas nas paredes dos retículos endoplasmáticos, quando extraídas a partir da lise de tecidos hepáticos são reagrupadas pelas membranas lamelares em estruturas chamadas microssomos. Os microssomos lisos (sem ribossomos) são responsáveis pela biotransformação de diversos fármacos, sobretudo em reações de oxidação que incluem o agente redutor NADPH. O uso continuado de fármacos que demanda metabolismo de proteínas microssomais, por exemplo, as associadas ao complexo do citocromo P450, podem induzir o sistema a aumentar a síntese ou reduzir a degradação das enzimas com consequente elevação da capacidade deste metabolismo, fenômeno conhecido como **indução enzimática**. A **inibição enzimática** ocorre quando substratos ligam-se fortemente às enzimas de modo a impedir sua atividade em outras moléculas.

## A.3.2 Farmacodinâmica

A farmacodinâmica avalia os efeitos bioquímicos e fisiológicos e seus mecanismos de ação. Uma análise completa da ação do fármaco possibilita as bases de um uso terapêutico racional e o desenho de novas e superiores tecnologias farmacêuticas[Brunton et al., 2005].

Os efeitos da maior parte dos fármacos advém da interação com macromoléculas. Esta interação inicia modificações fisiológicas e bioquímicas características. O receptor é o componente que interage quimicamente com o fármaco para iniciar uma dada reação.

### A.3.2.1 Receptores de fármacos

A maior parte dos receptores é formada por proteínas. Exemplos incluem fatores do crescimento, fatores de transcrição, enzimas, canais iônicos, ou mesmo, atuam em proteínas estruturais como a tubulina. Outros alvos relevantes são os ácidos nucleicos, particularmente para fármacos quimioterápicos.

Receptores normalmente são acionados em processos de regulação a partir de substâncias endógenas[13], como hormônios ou neurotransmissores. Fármacos cuja ação no receptor é superior à causada pelas substâncias endógenas atuam como **agonistas**. Fármacos que se ligam ao receptor sem realizar ação farmacológica são **antagonistas**, pois impedem ou prejudicam a ação da molécula endógena ou de outro fármaco agonista.

[13] Substâncias produzidas pelo próprio organismo, antônimo de xenobiótico.

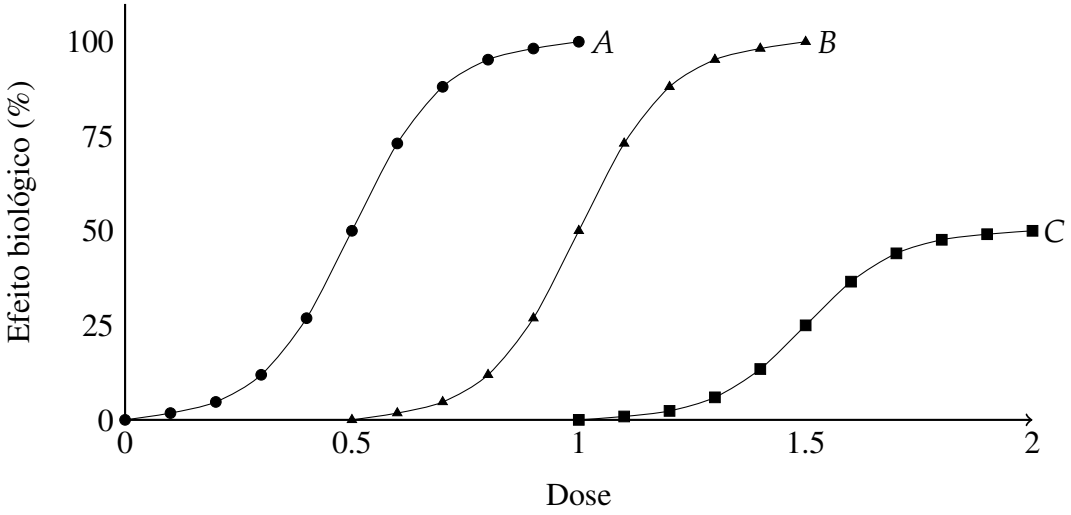



Figura A.5: **Relações de dose-efeito.** O fármaco *A* é mais potente que o *B*, porém ambos tem a mesma eficácia. O fármaco *C* é menos potente e menos eficaz que os fármacos *A* e *B*.

### A.3.2.2 Relação dose-efeito

A ocupação do receptor pelo fármaco frequentemente não é definitiva. A afinidade molecular entre fármaco e receptor, análoga a enzima-substrato ou antígeno-anticorpo, rege a relação dose-efeito. Fármacos com menor afinidade pelo receptor tendem a demandar maiores dosagens para a obtenção do efeito biológico. No entanto, o efeito pode não ser equiparado ao de outro fármaco, mesmo com a elevação da dose.

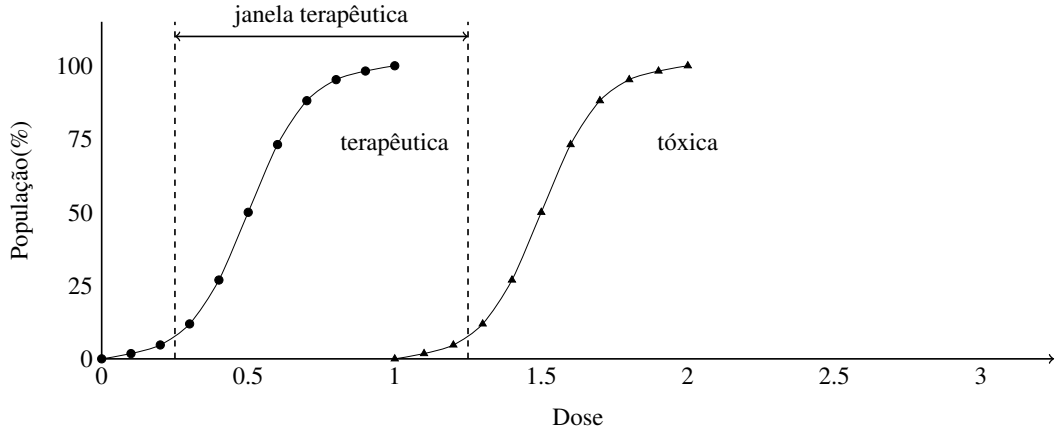
A instabilidade ou inespecificidade pode tornar um fármaco menos eficaz, conforme observado na figura A.5. Outra observação importante é estagnação do efeito em determinado limiar de dosagem, não observando-se diferenças com o aumento.

**Janela terapêutica** A diferença entre a atividade terapêutica e a tóxica de um fármaco pode estar na dosagem, conforme demonstrado na figura A.6. Fármacos como a digoxina[14] ou a varfarina[15] possuem baixo **índice terapêutico**, que corresponde a fração entre a dose tóxica e a dose efetiva. A penicilina[16] é um exemplo de elevado índice terapêutico, sendo comum o uso 10 vezes superior à dose mínima necessária para obtenção de resposta. Em casos como este, a elevação da biodisponibilidade não afeta de modo crítico os efeitos terapêuticos.

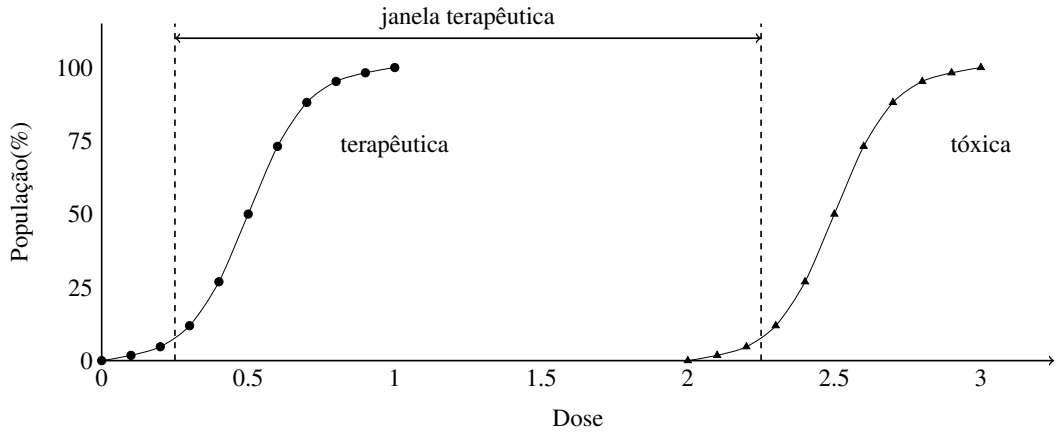
## A.4 O domínio *in populo*

Os estudos realizados nos domínios *in vitro* e *in vivo* devem partir do pressuposto de que as amostras analisadas são representativas. Porém, as amostras são quase nunca verdadeiramente

[14] Usado para o tratamento cardíaco.
[15] Usado como agente antitrombótico, ou seja, para controle da coagulação sanguínea.
[16] Usado via parenteral no combate de infecções como gonorreia, meningite, sífilis ou artrite séptica.




(a) Fármaco A




(b) Fármaco B

Figura A.6: **Janela terapêutica.** O fármaco *A* é menos seguro que o *B*, pois a distância entre a dose terapêutica e a dose tóxica na população observada é menor.

randômicas, pois é logisticamente impossível identificar todos os indivíduos que pertencem à população-alvo, e, em seguida, escolher aleatoriamente entre eles[Strom & Kimmel, 2007].

Uma **associação** de objetos de estudo poderá ser generalizada à **causação** diante de progressivos estudos populacionais que contemplem todos os casos possíveis.

## A.4.1 Níveis de evidência

### A.4.1.1 Revisão sistemática

Uma revisão sistemática deve ser focada em uma questão enquadrada em critérios pré-definidos. Devem ser especificado os tipos de particiantes, ou seja, quem ou o que se enquadra como objeto de estudo, o tipo de interveção e a comparação para a resolução do problema e, finalmente, os desfechos desejados com a abordagem estabelecida. Este fluxo conhecido como PICO [17]

[17] *Participants*, *Interventions*, *Comparisons* e *Outcomes*

Tabela A.1: **Nível de evidência para decisões clínicas** de terapias, prevenção, etiologia ou dano, segundo *Centre for Evidence-Based Medicine, Oxford.*

| Nível | Descrição |
|---|---|
| 1a | Revisão sistemática (com homogeneidade) de ensaios clínicos controlados e randomizados. |
| 1b | Ensaios clínicos controlados e randomizados (com estreito intervalo de confiança). |
| 1c | Ensaios clínicos controlados randomizados ou não. |
| 2a | Revisão sistemática (com homogeneidade) de estudos de coorte. |
| 2b | Estudo individual de coorte ou ensaios controlados randomizados de menor qualidade (por exemplo, <80% acompanhamento). |
| 2c | Pesquisa de desfechos; estudos ecológicos. |
| 3a | Revisão sistemática (com homogeneidade) de estudos caso-controle. |
| 3b | Estudo individual de caso-controle. |
| 4 | Série de casos; estudos de coorte ou caso controle de baixa qualidade. |
| 5 | Opinião de especialista sem apreciação crítica explícita ou com base em fisiologia, pesquisa de bancada ou "primeiros princípios". |

é amplamente preconizado por centros de excelência em revisão sistemática como o Centro Colaborador Cochrane[Higgins & Green, 2011].

**Metanálise** A metanálise é uma comparação estatística indireta de diversos estudos que abrangem as mesmas condições diante de critérios pré-estabelecidos. Uma crítica comum a metanálise é a "mistura de maçãs com laranjas", dado que estudos clinicamente diversos podem ser combinados distorcendo o significado diante de efeitos obscuros. No entanto, as diversas precauções e ponderações realizadas para estabelecer o viés dos estudos estabelecem a qualidade metodológica dos achados, tornando a ferramenta mais poderosa para avaliação da tendência de efeitos terapêuticos ou diagnósticos[Higgins & Green, 2011].

A metanálise culmina com a geração de um gráfico de floresta que em que os estudos são posicionados no eixo das ordenadas e o indicativo do efeito observado nas abscissas. Uma linha vertical segmenta os tratamentos favoráveis à intervenção e o grupo que sugere que o tratamento alternativo é superior. Tratamentos estatísticos indicam na forma de um prisma qual a tendência geral em relação ao eixo vertical a qual sugere a recomendação derivada do estudo de metanálise.

#### A.4.1.2 Ensaio clínico

É um estudo prospectivo de comparação da segurança e eficácia e/ou efetividade, de dois grupos, o da intervenção terapêutica e o do controle (placebo ou terapia existente). Quando **controlado**, a designação dos indivíduos ao grupo deve ser randomizada. Um estudo **cego** ocorre

quando o paciente desconhece o grupo a qual pertence. Um estudo **duplo-cego** ocorre quanto o médico ou o enfermeiro também desconhece qual tratamento foi procedido.

#### A.4.1.3 Estudo observacional

Método de pesquisa prospectiva para documentação de resultados clínicos, econômicos e/ou humanísticos da prática real na saúde, sob a ausência de restrições de um desenho experimental formal. Frequentemente são avaliadas bases de dados amplas em que a observação do evento desejado torna-se mais provável. Porém, muitas vezes são empregadas bases que não foram desenvolvidas para este fim, limitando as conclusões e a generalidade.

**Estudo de coorte** Avaliação do risco relativo de incidência de determinado evento em grupos de indivíduos expostos ($e$) e não expostos ($n$).

$$\frac{\frac{e}{e+\not{e}}}{\frac{n}{n+\not{n}}} \tag{A.1}$$

**Estudo de caso-controle** Avaliação da incidência do evento em função de grupo de indivíduos caso (com a doença) em relação a um grupo de indivíduos controle (sem a doença).

$$\frac{\frac{e}{e+n}}{\frac{\not{e}}{\not{e}+\not{n}}} \tag{A.2}$$

Nas equações A.1 e A.2 os casos em que o evento não foi observado são representados por $\not{e}$ e $\not{n}$ para expostos e não expostos, respectivamente.

#### A.4.1.4 Pesquisa de desfechos

Avalia o efeito das intervenções de cuidados à saúde sob aspectos relacionados ao paciente, abordando frequentemente alternativas de tratamento e avaliação de múltiplos tipos de resultados relacionados à doença.

#### A.4.1.5 Estudo ecológico

A **análise de tendência secular** examina a coincidência de tendências de uma causa presumida a uma exposição e da causa presumida de uma doença. Estas tendências podem ser avaliadas ao longo do tempo ou através de fronteiras geográficas. Esta análise é útil por oferecer uma rápida evidência a uma hipótese. No entanto são empregados apenas dados agregados dos indivíduos, não controlando-se as variáveis de confusão.

#### A.4.1.6 Desconcertamento

Confusão se refere ao efeito da exposição sob o estudo sendo misturado ao efeito de um terceiro fator. O terceiro fator deve ser um fator de risco para a doença, bem como associado à exposição. Fatores comuns de confusão incluem idade e sexo. A confusão pode ser controlada no desenho do estudo através da randomização, restrição (critérios de inclusão) ou equiparação. A confusão pode, também, ser avaliada e controlada na análise, através da análise estratificada ou métodos multivariáveis[Berger et al., 2009].

### A.4.2 Estudo de utilização de medicamentos

Segundo Laporte & G.Tognoni [2007], a utilização de medicamentos deve ser avaliada enquanto interação com o processo global de atenção à saúde, em face do diagnóstico e tratamento com decorrente modificação do curso natural da doença, culturalmente como é assumida na sociedade.

O estudo avalia a qualidade do consumo, a qual está ligada a detecção, avaliação, compreensão e prevenção de RAM (Reações Adversas a Medicamentos), incluindo interações medicamentosas nocivas.

Inclui a análise da oferta e informação de medicamentos, estudos quantitativos de consumo, estudos sobre a qualidade do consumo em outros fatores além da ocorrência de RAM, estudos de hábitos de prescrição e estudos de cumprimento da prescrição.

### A.4.3 Farmacovigilância

A vigilância de medicamentos abriga a detecção, avaliação, compreensão e prevenção de reações adversas e problemas relacionados em populações. Ocorre na fase IV após a comercialização, conforme visto na seção A.1.1.3.

A farmacovigilância realiza a identificação e valoração dos efeitos do uso, agudo ou crônico, dos tratamentos farmacológicos no conjunto da população ou subgrupos de pacientes expostos a tratamentos específicos. A atuação inclui estudos que valoram e quantificam a eficácia e eficiência dos fármacos, análise de estatísticas vitais, supervisão intensiva de pacientes hospitalizados, vigilância orientada a problemas e promoção de sistemas de notificação.

Dentre os aspectos abrigados na fase IV, encontra-se a observação dos medicamentos de fabricantes diferentes com o mesmo fármaco, forma farmacêutica, mesma via de administração e potência que devem ser **equivalentes**, apresentando o mesmo desempenho nos aspectos físico-químicos como liberação, pureza e uniformidade[BRASIL, 2010a].

Sobretudo em formulações pouco hidrossolúveis ou biodisponíveis, medicamentos equivalentes podem não ser bioequivalentes. Oscilações na fabricação de matérias primas, como

o polimorfismo[18] podem causar variações na produção e consequente perca do desempenho como queda do teor ou elevação dos produtos de degradação, trazendo impactos terapêuticos ou tóxicos.

### A.4.4 Saúde Pública

#### A.4.4.1 Assistência Farmacêutica

A Assistência Farmacêutica é um conjunto de ações voltadas à promoção, à proteção e à recuperação da saúde, tanto individual como coletiva, tem o medicamento como insumo essencial, visando o acesso e uso racional. Esse conjunto envolve a pesquisa, o desenvolvimento e a produção de medicamentos e insumos, bem como a seleção, programação, aquisição, distribuição, prescrição, dispensação, garantia da qualidade dos produtos e serviços, acompanhamento e avaliação da sua utilização, na perspectiva da obtenção de resultados concretos e da melhoria da qualidade de vida da população

Neste contexto, emergem modelos de gestão, promovendo o acesso ao medicamento, e a atuação clínica do farmacêutico com o monitoramento farmacoterapêutico, conhecido como atenção farmacêutica.

A avaliação de interações medicamentosas em populações fornece subsídios epidemiológicos para tomadas de decisão, bem como a avaliação no contexto clínico de resultados negativos associados à medicação - RNM. A integração de algoritmos para detecção de interações medicamentosas com informações dos pacientes, seja oriunda de prontuários médicos ou documentação farmacêutica, apresenta-se como uma importante ferramenta para o manejo da polifarmácia em populações.

## A.5 O domínio *in silico*

### A.5.1 Modelagem Computacional de Sistemas Biológicos

A modelagem computacional contribui com a descrição dos complexos sistemas biológicos, incluindo observações ao longo do tempo. Nas ciências físicas, métodos teoréticos em combinação com medições experimentais vem contribuindo para a neurobiologia e fisiologia[Fall et al., 2002].

Fall et al. [2002] descreveram cinco etapas para o fluxo experimental, teorético e computacional de modelos dinâmicos. 1) A partir do trabalho experimental, devem ser selecionados possíveis **mecanismos moleculares** com base na plausibilidade. Em muitos casos experimentalistas devem ser consultados. 2) A representação esquemática dos mecanismos deve primar pela generalidade contendo os passos elementares. 3) As leis fundamentais da física e da quí-

[18] Capacidade de formação cristalina diferenciada no agente químico, com possível modificação em características físico-químicas como solubilidade, compressibilidade, adsorção, entre outras.

mica podem ser usadas para traduzir os passos elementares em expressões matemáticas. 4) Estas expressões são combinadas com equações diferenciais tempo-dependentes para quantificar as mudanças descritas para todo o modelo. 5) As equações diferenciais devem ser avaliadas quanto ao sucesso da representatividade do modelo do sistema biológico.

Análogamente, o uso de técnicas de mineração de dados é uma alternativa ou complemento às equações diferenciais citadas na quarta e quinta etapa.

## A.5.2 Complexidade e custo computacional

A complexidade computacional define a viabilidade do processamento das informações. Alguns problemas não são computáveis, outros são computáveis mas são impraticáveis (chamados de problemas intratáveis) independente do processamento da máquina, alguns são computáveis, mas o algoritmo desenvolvido pode não ser o mais eficiente, ou equivalente ao processamento manual das informações[MacCuish & MacCuish, 2011].

## A.5.3 Teoria dos grafos

### A.5.3.1 Grafos simples não-direcionado

Seja um grafo $G$, definido pelo par $(V, E)$, onde $V$ é o conjunto de vértices ou nodos e $E$ é conjunto de arestas representando as conexões entre os nodos. Define-se como $E = \{(i, j) \,|i, j \in V\}$ como a única conexão entre os nodos $i$ e $j$. Nesse caso dizemos que $i$ e $j$ são vizinhos. Uma conexão multi-aresta consiste em duas ou mais arestas que tem os mesmos terminais. As conexões multi-aresta são especialmente importantes para as redes em que dois elementos podem ser ligados por mais de uma conexão. Nesses casos, cada conexão indica diferentes tipos de informação. [Pavlopoulos et al., 2011]

### A.5.3.2 Grafos direcionados

Definido por um tripleto $G = (V, E, f)$, onde f é uma função que mapeia a ordem dos vértices $V$ para cada elemento de $E$. Os pares ordenados dos vértices são chamados de arestas direcionadas, arcos ou setas. Este tipo de grafo é comumente utilizado para procedimentos onde deseja-se recuperar o fluxo da rede de interação sequencial dos elementos em um ou múltiplos pontos. Comum em redes metabólicas, transdução de sinais ou redes regulatórias. [Pavlopoulos et al., 2011]

### A.5.3.3 Grafos ponderados

Definido como um grafo $G = (V, E)$ onde $V$ é um conjunto de vértices e $E$ é um conjunto de arestas entre os vértices $E = \{(u, v) \,/ u, v \in V\}$ associadas, cada, a uma função de peso $w : E \to R$, onde $R$ denota o conjunto dos números Reais. Muitas das vezes, o peso $w_{ij}$ da aresta entre os nodos $i$ e $j$ representa a relevância da conexão. Frequentemente é empregado para

a captura da relevância de co-ocorrências de text-mining, similaridades estruturais de sequência entre proteínas, ou co-expressão de genes. [Pavlopoulos et al., 2011]

#### A.5.3.4 Grafos bipartidos

O conjunto $V$ pode ser particionado em dois subconjuntos $V_1$ e $V_2$, onde cada elemento $(u, v) \in E$ implica que cada $u \in V_1$ e cada $v \in V_2$ ou $v \in V_1$ e cada $u \in V_2$. Em outras palavras, não existe arestas entre elementos do mesmo subconjunto. Comumente usado para representação de reações enzimáticas em rotas metabólicas. [Pavlopoulos et al., 2011]

#### A.5.3.5 Hipergafos

Um hipergrafo é a generalização de um grafo ordinário, onde uma aresta, chamada hiperaresta, pode conectar mais que dois vértices. Mithani et al. [2009] geraram um hipergrafo, a partir do KEGG, onde os compostos são vértices e as hiperarestas são as conexões entre os compostos. A reação é tratada como um única entidade, possibilitando a captura do relacionamento entre o número de metabólitos envolvidos.

### A.5.4 Bioinformática

Segundo Wang et al. [2005] a bioinformática é a ciência do manejo, mineração de dados biológicos nos níveis genômico, metabolômico, proteômico, filogenético, celular ou do organismo como um todo.

#### A.5.4.1 Ontologia

Uma descrição ontológica deve adotar arcabouço descritivo que discrimine entidades, relações e papéis, estabelecendo uma linguagem compreendida por humanos e máquinas. A OWL, web of ontology é exemplo amplamente difundido de linguagem, recomendada pela W3C [McGuinness & van Harmelen, 2004], a qual padroniza a descrição de *a*) classes, *b*) propriedades, *c*) cardinalidade e *d*) relações de igualde de forma análoga à UML, linguagem unificada de modelagem[Rumbaugh et al., 2005].

Dentre as ferramentas destacam-se a KAAS - KEGG *Automatic Annotation Server* e o *pathologic*, ferramenta de previsão que acompanha o pacote *biocyc pathway tools*, sendo que a última contém também o módulo *pathway hunter tool*, o qual implementa técnicas que procuram localizar vias metabólicas alternativas. [Milreu, 2008]

A DIO [Yoshikawa et al., 2004], é uma ontologia específica de interações medicamentosas, a qual permite a descrição encadeada de cada interação fármaco-biomolécula perfazendo o mecanismo da interação medicamentosa sob a distinção de componentes biológicos, tais como enzimas e biomoléculas e consequências biológicas como a inibição ou indução. Outras on-

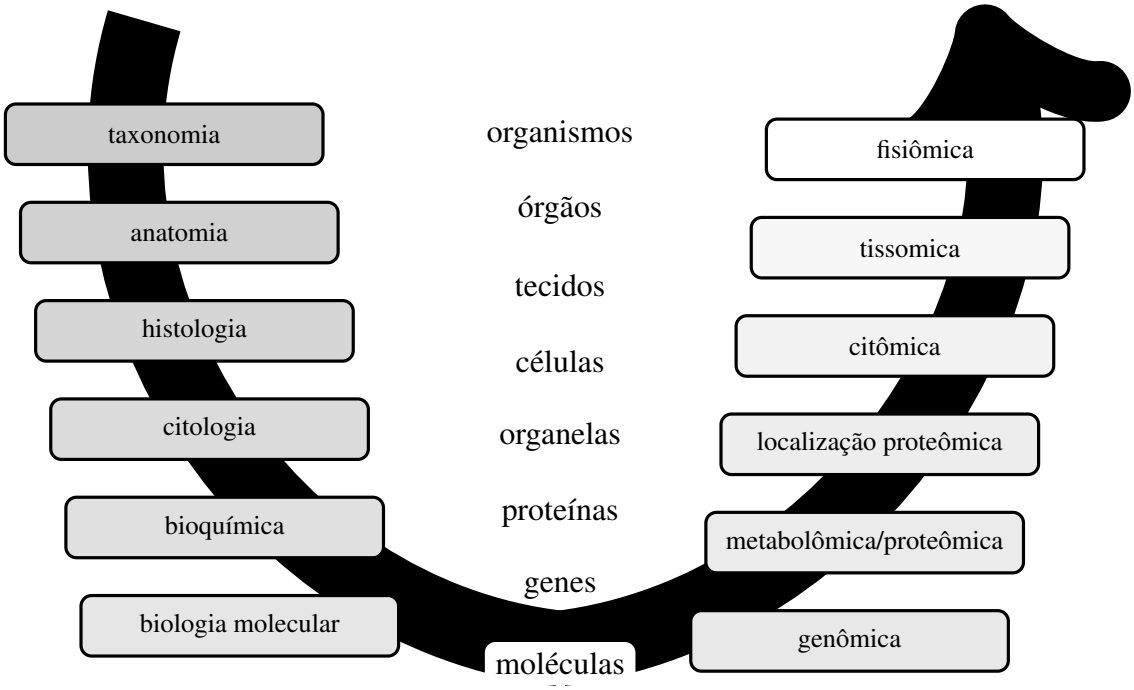



Figura A.7: **Biologia computacional de sistemas.** A evolução das sub-disciplinas na biologia através dos tempos ocorreu sempre focada em estruturas cada vez menores e questões mais detalhadas. A disponibilidade de técnicas de sequenciamento genético de elevado desempenho representou uma inflexão nas pesquisas relacionadas as bases da vida. As investigações estendidas a dados livre de hipóteses sobre entidades biológicas despontaram a genômica como a primeira dentre a crescente área das disciplinas "omicas". Apesar da genômica e proteômica funcional estarem longe de se fazerem completas, novos nichos do conhecimento lidam com fenótipos celulares, tecidos e níveis fisiológicos, constituindo disciplinas especializadas que preenchem os esparsos níveis de informação demandadas. A biologia de sistemas disponibiliza metodologias para combinar, modelar e simular entidades sobre diversos níveis de organização biológica (horizontal), como regulação de genes e redes de proteínas bem como abordagens multi-escala (vertical). Figura reproduzida de Kriete & Eils [2006].

tologias apenas correlacionam termos médicos (p.ex, UMLS), celulares (por exemplo, GO) ou rotas bioquímicas limitadas ao contexto biológico (por exemplo, KEEG).

### A.5.4.2 Biologia Computacional de Sistemas

O termo foi cunhado por Kitano em 2002[Kitano, 2002a,b] como o campo que possibilita o entendimento no nível sistêmico pela análise por técnicas computacionais de dados biológicos. Os domínios da Biologia de Sistemas Computacional se estendem da taxonomia à Biologia Molecular, da genômica à fisiômica, ou seja, do nível molecular ao de organismos[Kriete & Eils, 2006].

Citômica é o estudo sistemático da organização biológica e comportamento ao nivel celular, o qual tem sido desenvolvido com imagens computacionais ou citometria de fluxo.

Kriete & Eils [2006] denominam as áreas da biologia sistêmica como tecnologias facilita-

doras (**Enabling technologies**). Os autores discutem bases de dados para biologia de sistemas que agregam informação sobre as respostas de sistemas biológicos a perturbações genéticas ou ambientais, requerendo a integração dos mais divergentes tipos de dados para modelagem, simulação e previsão. O conceito de bases de dados integrativos perpassa por três grandes áreas dos dados gerados em biologia sistêmica: dados experimentais, elementos de biologia sistêmica e modelos matemáticos com suas simulações derivadas.

**Descoberta biológica pela análise e modelagem de redes bioquímicas** A identificação sistêmica e análise comportamental são as duas classes de ferramentas demandadas. Uma vez que os sistemas sejam identificados e o modelo construído, o comportamento pode ser estudado, seja por integração ou análise de responsividade a pertubações externas.

O processo de modelagem inicia com com uma abordagem reducionista, criando um modelo simplificador. Posteriormente, a geração do entendimento das estruturas delineadoras e componentes são representados com conceitos matemáticos e estatísticos. O modelo mínimo então cresce em complexidade, direcionada por novas hipóteses que podem não ser aparentemente a descrição fenomenológica. Então, um experimento é desenvolvido usando biologia sistêmica para testar como o modelo preditivo concorda com as observações experimentais.

Os parâmetros construtivos do modelo podem ser mensurados diretamente ou podem ser inferidos durante o processo de validação. No entanto, a propagação do erro através desses parâmetros representa um significativo desafio para o pesquisador. Se os dados e as previsões concordarem, um novo experimento pode ser desenhado e realizado. Caso contrário, aspectos metodológicos devem ser revistos. Esse processo continua até que seja coletada evidência experimental suficiente a favor do modelo.

A modelagem pode ser abordada sob a perspectiva *bottom-up* ou *top-down*. Na modelagem *bottom-up* é usada a abordagem reducionista no estudo dos componentes básicos para posteriormente integra-los e encontrar padrões relevantes e funções, com rotas metabólicas. No entanto, esta estratégia em geral tem capacidade limitada de traduzir o efeito das perturbações nessas rotas a células e seu papel celular. Esta abordagem não é efetiva em modelar entidades multicelulares (tecidos) ou organismos. A abordagem *top-down* é iniciada com o sistema intacto para depois decompô-lo em partes e interações. Aqui é estabelecido o conhecimento sobre o sistema adquirindo a capacidade de predicá-lo em módulos funcionais.

A decomposição de múltiplos componentes celulares em grupos permite a modelagem e simulação para resolução de instâncias no tempo e pode mimetizar uma propriedade biológica envolvida. A diferença crítica entre essas abordagens ocorre quando componentes e interações não são totalmente conhecidas.

**Seleção do modelo e simulação de processos celulares dinâmicos** Sistemas dinâmicos com tempo discreto tem sido longamente utilizados na biologia. Simulações computacionais requerem cuidadosas considerações como o nível de detalhe necessário para um modelo

representativo, visto que o detalhamento desnecessário irá tornar os modelos tão complexos tornando inviável o estudo numérico detalhado. Exemplos desta área são a estimação, modelagem e simulação de redes genéticas a partir da expressão gênica de dados de microarranjos, ou da abordagem discreta de modelagem *top-down* de redes bioquímicas a partir de bases experimentais *high-throughput*[19] em termos de engenharia reversa para construção de modelos dinâmicos discretos.

**Representações multiescala de células e fenótipos emergentes** O termo complexidade é frequentemente associado com impredizível. No entanto, sistemas biológicos complexos como as células são robustos e funcionalmente estáveis. A complexidade na biologia é atribuída a larga diversidade de elementos (p.ex. genes, proteínas e células). A caracterização desses elementos podem revelar uma variedade no espaço de estados, como a ativação de proteínas ou ciclo celular. Ainda, a diversas interações, alinearidades e retroalimentações em níveis hierárquicos biológicos contribuem na intrincada rede que aparenta ser um complexo de termos, contudo, com passível detecção de padrões generalizáveis.

A previsão do comportamento emergente do sistema ocorre em um limiar que admite uma diversidade de entidades que não inviabilize o tratamento das informações e impliquem um retorno que não seja simplista de modo a descaracterizar a utilidade da informação prevista.

É comumente reconhecida que a complexidade biológica está de acordo com a progressiva evolução trazida ao longo do acréscimo de complexidade das células e organismos através dos tempos. Esse julgamento coincide com a noção de que a maior complexidade é melhor em termos de sistemas adaptativos complexos e da capacidade de auto-organização.

Análises baseadas em computação e representações das propriedades emergentes são recentes, mas são campos essenciais na área de biologia sistêmica. O objetivo é conceitualizar e abstrair os princípios e o modelo das estruturas biológicas, incluindo níveis superiores de organização como células, tecidos e órgãos.

Esforços de modelagem são amplamente focados em um nível isolado ou escala, como a genômica ou proteômica, celular, tecido, órgão, sistema orgânico, corpo inteiro, comportamento ou população. Poucas pesquisas são devotadas ao desenvolvimento de ferramentas, técnicas, algoritmos e teoria matemática para integrar a continuidade desde a microescala até a macroescala.

Modelagem multiescala entrelaça conceitos no espaço estacionário e cruza escalas de tempo. Diferentes níveis organizacionais como redes genéticas regulatórias, módulos e rotas podem ser aninhados hierarquicamente. Modelos computacionais representando relações espaço-temporais não são limitados a uma resolução específica mas podem integrar multiescalas, incluindo abstrações flexíveis a simulações fisiológicas funcionais.

[19] Trabalhos em larga escala de tempo e distribuição de processos.

### A.5.4.3 Redes de reações bioquímicas

Segundo Kriete & Eils [2006] as redes metabólicas versam basicamente pelos modelos estruturais, regulatórios ou cinéticos.

Os **modelos estruturais** encontram-se no mais baixo nível de detalhamento, onde são distinguidas estruturas estequiométricas de uma rede de reações bioquímicas, por exemplo, catálises, transporte e ligação. Isto representa a topologia do fluxo de massa através da rede e identifica os substratos e produtos de todos os processos, mas não incorpora efeitos inibitórios ou ativatórios de efetores alostéricos. Frequentemente são expressos na forma de matrizes estequiométricas.

Os **modelos regulatórios** consideram as interações dinâmicas dos substratos e produtos de suas enzimas, dos efetores alostéricos de enzimas, fatores de transcrição, e influências regulatórias que regem retroalimentação e influencia o fluxo de massa incorrendo na descrição de todas as interações na rede bioquímica.

Os **modelos cinéticos** incorporam propriedades cinéticas aos processos e a concentração total dos motivos presentes na rede diante da descrição cinética. Disto decorre a parametrização de todas as equações de taxas de todos os processos na rede, o que determina o tipo de função e os parâmetros das taxas para toda reação da rede como funções de todas as concentrações intermediárias, não somente aquelas que são estequiométricas. As equações enzimáticas de Michaelis-Menten podem ser reversíveis ou irreversíveis ou complexas, com mecanismos ordenados, sequenciais ou randômicos ou enzimas multi-subunidade de mecanismos cooperativos, inclusive, com cadeias de retroalimentação.

Integrados com os modelos estequiométricos, os modelos cinéticos podem ultimamente ter a descrição em que todos os parâmetros (propriedades cinéticas totais, motivo-conservado e condições de ligação) é dado como um valor determinado experimentalmente.

# Apêndice B

# Tópicos avançados do modelo

## B.1 Aspectos epistemológicos e metafísicos da interação entre objetos

O modelo proposto deriva do **paradigma sistêmico** e versa pela integração das características do ente, ou ser, ao avaliarmos o sistema que o abriga enquanto entidade única. Nesta abordagem o **ser**, sujeito ou objeto é representado na forma de **entidade** definida como um conjunto de características que compõe sua **identidade**. Em outras palavras considera-se que qualquer coisa existente é dotada de atributos que a define. A percepção e a cognição limita a concepção da entidade que perpassa pelo poder de expressão da forma simbólica adotada [1].

Entende-se por percepção não somente o apreendido sensorialmente ou intelectivamente, mas a extensão da percepção que as máquinas ou ferramentas oferecem. Embora os registros apreendidos pelos diferentes instrumentos sejam frequentemente traduzidos para a cognição humana, possuem linguagem própria, cuja utilização direta nos modelos preditivos pode corrigir distorções de interpretação, visto que a tradução, embora traga aporte semântico, pode também determinar uma simplificação, ou seja, perca de informação verificável apenas em outras estruturas de pensamento que não às limitadas à cognição humana.

Uma vez o objeto de estudo definido enquanto um conjunto de percepções, a **interação** entre entes constitui o **fenômeno** ou **evento**, tratados como sinônimos. Da mesma forma, **dimensão**, **característica**, **atributo** e **descritor** intuem o mesmo conceito e abrigam a caracterização do objeto.

**Instancia** é um conjunto de categorias que cria um determinado domínio. **Categoria** é uma noção que agrupa uma classe de elementos da realidade. Se esses elementos constituem uma classe é devido a características comuns com a classe, ou seja, relações metonímicas.

O paradigma integrativo da determinação holística dos objetos enquanto parte de um sistema integrado e interagente se contrapõe ao reducionismo científico, cuja análise se limita a

[1] Este tipo de conhecimento é explorado pela semiótica.

reduzir a complexidade do objeto de estudo pela decomposição de suas partes, constituindo a realidade como a soma dos fragmentos.

### B.1.1 Interação entre objetos

O modelo integrativo sugere que a caracterização de um objeto ocorre apenas na observação de sua capacidade de interagir.

O conceito de **interação** converge para a **ação** ou **efeito** de uma dada entidade em outra[Hornby & Wehmeier, 2007] ou de forma recíproca[Ferreira, 2009]. A interação não é considerada entidade distinta, dado que existe apenas enquanto **transação**[Abbagnano, 2007]. No entanto, a associação de entidades pode trazer consequências diferenciadas da soma das ações individuais, diferenciando-as pela definição da capacidade de interagir.

Neste modelo, a interação é tratada como entidade por agregar as propriedades da conjunção de entidades e possuir comportamento próprio, diferente da associação sem interação, em que apenas ocorre a concomitância de entes que atuam sem modificação das ações individuais.

No contexto de medicamentos a definição mais frequente é a modificação do efeito esperado de um fármaco em função da associação a outro[Berger et al., 2009; Tatro, 2012], sendo a ação mútua evidenciada em menor proporção.

Sob a ótica da modelagem computacional descrita por Rumbaugh et al. [2005], a interação é parte dos três elementos que regem o comportamento básico de um objeto ou coleção de objetos, a qual envolve mensagens, ações e ligações (ou conexões).

O significante mais difundido para uma interação é representado por uma reta que liga dois objetos ou conceitos. Em outras palavras, é a aresta entre dois vértices, nodos, figuras ou pontos. Quando a ação ocorre de A para B, em geral, emprega-se uma seta na extremidade do elemento que sofre a ação.

Desta forma, os domínios do signo **interação** não restringem-se às relações recíprocas nem comutativas, ou seja, A pode interagir com B sem que B interaja com A. A interação não possui comportamento por si só, mas deriva do comportamento de um elemento em função de outro, não sendo possível a descrição sem os elementos que a geram. A interação carrega as propriedades circunscritas em tudo que se reconhece como participante da interação.

O entendimento da interação enquanto relação de dois entes torna-se necessário para avaliar a predizibilidade enquanto artefato ou enquanto caracterização do objeto de interesse. A descrição de uma interação pode ser reduzida aos aspectos básicos que a define de forma equipotente, de modo a reproduzir, inclusive, as propriedades não abordadas. A forma reducionista de avaliar a interação, em geral, isola os objetos interagentes e busca eliminar os fatores que não corroboram a explicação do evento. No entanto, a descrição reduzida restringe a comparação entre interações distintas devido à possível diferença de escopo em dada descrição. Ao invés de estudar independentemente cada interação em busca da generalidade, o modelo integrativo adota a descrição ampla dos objetos envolvidos sem o viés imediato da redução.

Possivelmente, o modelo proposto representa uma interpretação do mundo de aporte dedutivo que traz para a metodologia o apenas abrigado discursivamente nas seções de discussão dos textos científicos dos tempos atuais.

### B.1.2 Premissas do paradigma integrativo

São mostradas a seguir sete premissas que constituem dedutivamente ou *a priori* o pensamento proposto.

**A capacidade de interagir com outra entidade é uma característica imanente, e constitui parte da identidade que a define** Em outras palavras o elemento A não depende de B para ser dotado da característica de interagir com B, podendo ter essa capacidade distinguida, em última instância, isoladamente. Da mesma forma, B possui características intrínsecas, distintas ou não das características de A, para interagir com A. Somente haverá interação mútua quando A e B forem individual e simultaneamente dotados da característica de interagir.

**A interação é inerentemente relacionada com o meio** Esta interação apenas pode existir com um elemento intermediário C, o qual viabiliza e ao mesmo tempo modula as características relacionadas à interação entre A e B. Tanto A como B se abrigam em C, caso contrário, não podem constituir a abstração de uma entidade real. Pensar isoladamente A ou B torna a identidade da interação inerte no plano imaginário.

**A interação é previsível a partir da avaliação das características determinantes das entidades individualmente e conjugadas às características do meio** Ou seja, as características de A em relação ao meio, em alguma instância, são suficientes para verificar-se a potencialidade de interação com B, independendo das características de B.

**O meio, por sua vez, é entidade e, ao mesmo tempo, constituído por um conjunto de entidades** O meio, sendo identificado por características, constitui por si só uma entidade. Um conjunto de características do meio pode ser tomado como entidade conforme o contexto de apreensão do mundo.

**A característica que rege a propensão a interagir com outro ente perpassa pela interação com o meio** O meio, enquanto viabilizador da interação, ao mesmo tempo modula a característica que compõe a propensão de uma entidade interagir, dado que permeia tanto a interação de A com C, como B com C, as quais não são inertes, dado que não existe entidade isolada.

**A interação com outra entidade é uma interação com o meio, e por consequência, com as demais entidades** Tomar o meio como uma composição de entidades faz com que a avaliação da interação com uma entidade seja a interação com o meio que a abriga. Da mesma forma, como as entidades participam do meio, ao haver interação com o meio, ocorre interação do meio com as demais entidades.

**Logo, a interação entre entidades específicas pode ser prevista ao avaliar-se as interações entre as entidades que compõe o meio** A determinação da interação entre a entidade A e B necessariamente deve ocorrer com a verificação da interação entre A e C e B e C. Ou seja, sendo C composto por A e B, devem ser verificadas as características de A em si, B em si, A com B, B com A, A com (A com B), A com (B com A), B com (A com B) e B com (B com A), e assim sucessivamente. Desta forma, a interação será rastreada a partir das possibilidades que descrevem as entidades, e por consequência, o meio.

### B.1.3 O paradigma reducionista

A posição oposta ao paradigma integrativo é de que apenas existem características conjuntas e específicas nas entidades envolvidas que as fazem interagir, ou seja, não avalia-se a natureza da entidade enquanto constituinte de um sistema que a abriga. Uma ampliação é considerar, ainda, as características do meio em que elas interagem, embora frequentemente não seja imprescindível. Ou seja, os objetos devem ser avaliados no ato da interação ou rastreadas as suas consequências no objeto ou no meio que o caracteriza. Em geral, é desejável tomar de forma cartesiana a interação como a variável controlada (comumente alocada no eixo das abscissas) e as demais características isoladas como as variáveis apreendidas nas condições experimentais delimitantes do modelo (eixo das ordenadas). Os limites da variável controlada devem ser avaliados para observar a extensão do modelo, ou seja, em quais condições o fenômeno se comporta conforme o esperado.

Este pressuposto viabiliza a experimentação por se aproximar objetivamente às características intrínsecas da interação. Porém esta limitação impõe constantes mudanças discretas e sistemáticas de foco quando deseja-se ampliar o contexto. Assim, a restrição da delimitação imposta para explicar o fenômeno, frequentemente não permite conclusões no limiar prático necessário, demandando numerosas pesquisas com perguntas crescentemente específicas. Como uma lupa ao sol, desvia-se o olhar para iluminar outros fragmentos do fenômeno com o intento potencialmente inatingível de verificá-lo por completo ao unir as observações.

A única solução apresentada até o momento para integrar as observações pixalizadas do conhecimento científico é o treinamento do cérebro capacitando-o a estabelecer discursivamente o conhecimento apreendido. Logo, com o passar do tempo, a experiência potencialmente faz o pesquisador retomar a visão do todo para, destarte, vislumbrar a realidade composta pelos

múltiplos contextos do problema escolhido. Os pesquisadores que atingem esse patamar são frequentemente alcunhados de *expert* ou *ad hoc*.

### B.1.4 Propriedades do paradigma integrativo

Definir um objeto perpassa pela limitação cognitiva e pelo objetivo ao faze-lo. Desta forma, ao se perguntar sobre um triângulo qualquer, a definição mínima é a de um polígono dotado de três retas que se conectam por três vértices, dado o entendimento (significado) e da forma de expressar (significante) de reta e vértice. Nesta representação as quatro dimensões que definem o objeto são os tamanhos das retas do polígono e a propriedade, ou restrição, de que se conectem pelas extremidades.

Embora a definição mínima possa parecer suficiente, esta lei reducionista traduz apenas uma identidade circunscrita pelo contexto e não define completamente o triângulo, pois a essência é composta por todo e qualquer tipo de representação que se possa fazer do objeto e das características que o define. Assim, para dissecar a verdade sobre um objeto devemos inerentemente avaliar suas propriedades em si e para si e de si para o meio. No primeiro caso, existem propriedades tomadas como imanentes ao triângulo, como "a soma de todos os ângulos perfazem 180°". Porém esta condição imanente, somente ocorre em um espaço plano, constituindo um exemplo de caracterização que depende da apreensão baseada na representação semiótica que descreve a interação com o meio, cujas propriedades averiguadas definem a capacidade de interação com os demais elementos do conjunto de objetos geométricos. Desta forma, outas dimensões como o ângulo e tamanho das arestas e a propriedade da soma dos ângulos tornam mais exata ou próxima da verdade a definição.

Qualquer propriedade advém da estrutura de pensamento que a gera, limitada a uma forma de percepção da realidade que compõe a imagem ou o signo (significante atrelado ao significado) na ótica do observador. O triângulo também possui relações com outros triângulos ou figuras geométricas que contribuem na sua definição. Por exemplo, ao indagarmos "quantos triângulos podem ser derivados ao se ligar internamente o centro das arestas" ou "qual o epicentro de um triângulo se ligarmos seus vértices ao aro de um círculo ou um quadrado", estamos falando da relação deste triângulo com outros elementos que compõe o meio geométrico. Desta forma o completo entendimento do triângulo e dos demais polígonos passa pela avaliação de suas relações intrínsecas e das relações com o meio.

#### B.1.4.1 Dimensões discretas ou contínuas

A forma de cognição disponibilizada pela matemática permite estruturar as dimensões que definem um objeto como discretas ou contínuas.

Se considerarmos a descrição como discretas, ou seja, intuída como pertencente ao conjunto dos números $\mathbb{N}$, podemos dizer que os naturais positivos representam dimensões do que o objeto **é**, e os $\mathbb{N}^-$ do que o objeto **não é**. A característica "cor" de uma entidade é expressa na

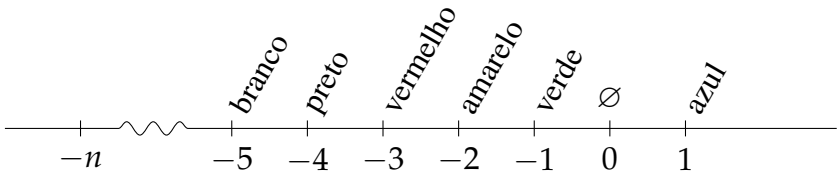



Figura B.1: **Dimensões discretas de um descritor.** Representação matemática da natureza discreta ou contínua de uma variável como redução a padrões binários da característica que compõe a identidade de uma entidade. A ausência de informação é representada pelo valor nulo.

figura B.1 como fatores mutuamente excludentes. Ser "azul" implica em não ser "verde" nem "amarelo".

Porém, a realidade tomada como intervalos pertencentes ao conjunto dos números $\mathbb{R}$ expressa a incerteza que aproxima de ter ou a não determinada característica. Por exemplo cada propriedade discreta pode adotar valores entre $]n, n+1]$ para $\mathbb{R}^+$ ou $[n, n-1[$ para $\mathbb{R}^-$. Logo, um valor posicionado no intervalo $]0, 1]$ referente a cor "azul" na figura B.1, indica a intensidade do azul conforme o padrão de apreensão da cor.

Talvez, a contrapercepção de um objeto, ou seja, sua intuição não emergida ao consciente, ou *ego*, possa ser expressa por uma matemática igual e opostamente não intuitiva, a dos números complexos. Assim, para cada dimensão de certeza ou graus de certeza, há um contradimensão de intuição. Embora a intuição não seja considerada escopo da ciência, frequentemente a fomenta, constiuinto a massa negra paradoxalmente marginalizada pela demanda de concretude da ciência, contudo, é mais uma forma de abstração com potencial igualmente passível de ser explorado quando houver método que o torne aceitável.

A massa negra é tudo aquilo que não é apreendido sensorialmente ou intelectivamente, mas pode trazer alguma coerência para os modelos propostos, análogos ao auto-intitulado maior erro de Einstein. Na ocasião, o famoso físico previu décadas antes a constante cosmológica que sustenta o movimento das galáxias, posteriormente atrelada à substância negra. Adimensionalidade não existe neste sistema. Ou se **é** ou se **não é**, não existindo objetos sem qualquer característica. No entanto, ao tomar a realidade de um objeto enquanto a sua percepção, ou seja, a imagem que fazemos dele, torna-se cabível o artefato de considerar como "nula" ($(\varnothing)$) a característica desconhecida, no lugar de tomar esta observação como adimensional ($\nexists$), sendo substituída quando apreendida alguma percepção, relativa a característica em questão.

### B.1.4.2 Identidade e imagem

A identidade de um objeto é obtida pela apreensão de todas as dimensões que o carateriza. O número de dimensões tende ao infinito, no entanto a percepção é finita. Logo, o apreendido é uma imagem ou aproximação da identidade.

Cada dimensão abriga um conjunto de observações possíveis de valores ou formas de percepção atribuídas à imagem e pode ser tomada em um plano cartesiano como um eixo

$x, y, z, ..., \infty = \varnothing$.

Assim, uma entidade é dotada de infinitas dimensões e cada dimensão abriga valores finitos ou infinitos segundo a capacidade de apreensão. A imagem tende ao mesmo limite, porém, no caso em que foi apreendida apenas uma dimensão verifica-se que $x \neq \varnothing$, $y = \varnothing$, $z = \varnothing$, $w = \varnothing, ..., \infty = \varnothing$.

A acuidade da percepção e consequente definição da realidade de identidade ocorre com um número de termos suficiente para apreender o conjunto de características determinantes para distinguir a capacidade de interação, dada a impossibilidade de apreender o objeto na totalidade.

#### B.1.4.3 As dimensões do objeto são as dimensões do meio

Isto reforça ainda mais a ideia do modelo integrativo. A existência do objeto perfaz a existência do meio. Atributos do meio também são atributos do objeto. Sendo assim, as dimensões do meio também são dimensões do objeto, bem como as dimensões do objeto são dimensões do meio, a medida que o objeto também constitui o meio. Desta forma, a acuidade da a imagem demanda sua representação contendo dimensões que caracterizam o meio ou pelos meios as entidade que o compõe. Em uma percepção dotada de múltiplas representações dos objetos com dimensões relativas ao meio, o meio já encontra-se representado nas dimensões dos seus objetos, desta forma, não é necessária a representação do meio e de suas dimensões disjuntamente à percepção, ou constituição da imagem do objeto.

#### B.1.4.4 Discreto, contínuo e formas de percepção

Tomar as características como discretas ou contínuas são apenas exercícios de cognição. Usualmente, não é por uma ou duas características que o objeto será distinguido. Desta forma, a maneira de expressar com acuidade é compensada pela tomada de um número suficiente de dimensões que sejam capazes de caracterizar o objeto tomando-se a finalidade do mundo sensível ou inteligível.

No entanto, o conhecimento da verdade sobre um objeto perpassa pela avaliação das dimensões do meio, logo, não se pode fazer afirmações sobre a verdade de um objeto sem tomar todas as percepções disponíveis ou sem um certo esforço em obter as dimensões possíveis a exaustão da capacidade cognitiva.

#### B.1.4.5 Realidade é percepção

Uma vez que não podemos apreender a verdade de um objeto por não tomarmos todas as características que o define, tomamos apenas algumas que constituem sua imagem, logo, ao falarmos da realidade sobre um objeto inerentemente estamos nos posicionando como observadores, derivando nossa percepção para falar sobre a realidade do objeto. A realidade é fruto da percepção e está relacionada ao observador. A realidade é aquilo que é para cada um.

Tomando **realidade** diferentemente do conceito de **verdade**, assume-se que a **verdade** existe, porém é intangível dada nossa limitação cognoscente. Nos é apenas cognoscível a realidade, e neste caso, cada observador, seja homem ou máquina, torna-se a medida de todas as coisas. A verdade consiste na paradoxal apreensão das infinitas dimensões de uma entidade, conforme exposto na seção B.1.4.10.

### B.1.4.6 A dialética do ser ou não ser

Tomando cada observação como uma representação cognitiva, ou imagem da entidade, a derivação das observações perfazem inerentemente a escolha ou interpretação da percepção em face de dotar ou não o objeto daquela caraterística em questão. Assim, independente da forma de cognição, por exemplo, se "cor" é exressa como "azul" ou "azul-marinho" ou 790nm[2], em última instância a apreensão da característica "cor" perpassa por ser ou não "cor". Neste caso é igualmente válido dizer "não preto", "não branco", "não amarelo", "azul", "azul-marinho", "790nm". E "não preto" caracteriza igualmente a entidade como "azul" o faz, mesmo diante da variação no poder de expressão, pois dizer "não preto" caracteriza "cor" como fator atrelado ao sistema e enquanto entidade que é. Desta forma, um conjunto de observações da entidade quanto a "cor" também constrói a imagem da "cor" em função das entidades, tornando-a igualmente como entidade.

Em um sistema em que **ser um** implica em **não ser um outro**, não é necessário caracterizar o **não ser**. Quando uma dada entidade tem uma característica que contradiga este princípio com a de outra entidade, esta característica é excluída mutuamente. Consequentemente, caracterizar o que esta entidade não é torna-se necessário para a constituição da imagem das entidades e da imagem das características.

### B.1.4.7 Propriedade fractal-comutativa

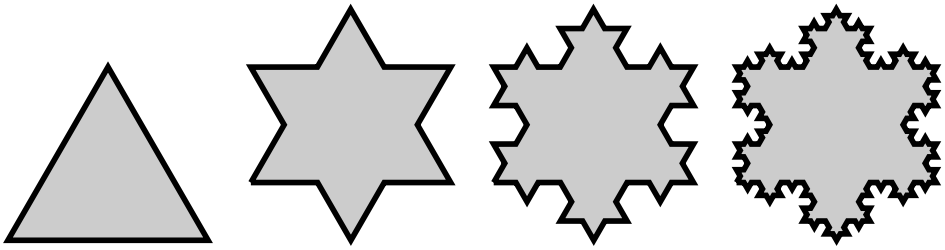
Conforme enunciado na seção B.1.2, a entidade se torna característica ao concebê-la como parte de uma entidade ou o meio. Por exemplo, o triângulo é a entidade que constitui o fractal da figura B.2. Cada iteração herda elementos da anterior, porém possui características próprias.

A propriedade do triângulo de gerar uma estrutura maior, faz com que as características das entidades derivadas também participem da natureza do triângulo. Logo, independente do nível ontológico observado, as características das parte remontam às do todo e vice-versa. O específico e o universal são apenas níveis de abstração conforme o meio observado.

### B.1.4.8 O ser e o tempo

Tomando-se a essência de um objeto como um conjunto das infinitas características que o definem, o tempo torna-se apenas mais uma dimensão destas características na medida que cada dimensão é apreendida ao longo da vida da entidade e do observador.

[2]Comprimento de onda da luz em escala $10^{-9}m$



(a) Iteração 1 (b) Iteração 2 (c) Iteração 3 (d) Iteração 4

Figura B.2: **Fractal** da formação sucessiva de três triângulos equiláteros a partir da divisão de cada aresta do triângulo equilátero original em três partes iguais de modo a formar um triângulo no centro de cada aresta. Esta estrutura converge para o polígono conhecido como floco de neve de Koch.

Tradicionalmente objetivou-se definir como essência as imanências imutáveis de um objeto. No entanto, existem características mutáveis que constituem a essência do objeto. Desta forma, cada dimensão possui um **escopo** que, em parte, é autônomo em relação às demais dimensões, cuja continuidade não é representada com apenas uma obervação em um momento específico, sem prejuízo a sua caracterização.

Ao averiguar um objeto movido pelo senso comum, o observador inerentemente confunde o tempo em que se posiciona com o tempo do próprio objeto. Com apenas uma observação não é possível apreender a verdade do objeto. O objeto deve ser apreendido durante todo o tempo de existência para obter-se compreensão fidedigna. Desta forma, constitui uma imagem mais próxima da verdade do objeto um conjunto de observações, na impossibilidade de observar pelo período completo da existência do objeto. Ou ainda, diversas observações de objetos similares em momento distintos da vida.

Conforme evidenciado na figura B.3, cinco apreensões foram realizadas do objeto de estudo. O tempo 4 apresenta um conjunto de condições que podem indiciar o fenômeno em estudo. Porém, o ente avaliado possui um histórico o qual faz parte da sua essência não podendo ser ignorado na avaliação, sobretudo em modelos preditivos. Deve-se priorizar uma observação contínua dos eventos para que cada dimensão seja corretamente traçada em sua amplitude de valores. Ressalta-se que o entendimento da amplitude de uma dimensão também requer a observação em outas instâncias.

Neste contexto, as previsões não tratam de encontrar descrições localizadas no futuro, mas detectar uma imanência, algo da essência do ente que o torna propenso a manifestar a característica em observação.

Porém ao considerar a dinâmica da dimensão avaliada, as observações devem ser realizadas simultaneamente para a entidade e o meio. Ao considerar o tempo tal qual nossa acuidade o define, isto implica em observar todo o tempo de todas as entidades, ou na impossibilidade, diversas observações de todas as entidades.

Desta forma, a imagem não mais representa fidedignamente a essência do objeto, mas o conjunto de suas imagens. O objeto deve ser avaliado como a junção das características

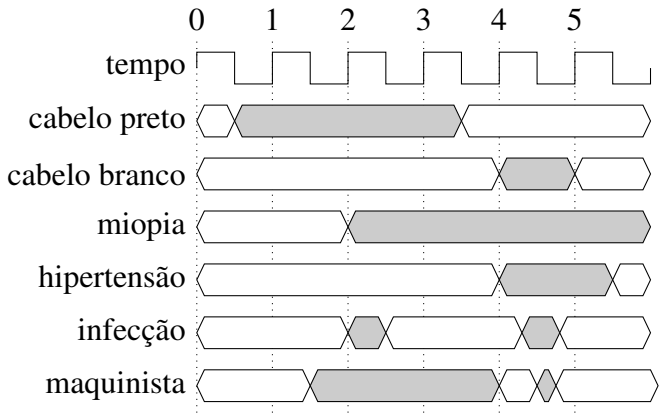



Figura B.3: **A essência de um objeto e o tempo.** A presença do atributo, representada em cinza, indica a continuidade ao longo da vida do objeto avaliado. A essência de um objeto é apreendida sob a conjunção de diversas imagens ao longo da vida deste. Desta forma, o tempo torna-se apenas uma das dimensões que descrevem o objeto, exceto no estudo de fenômenos dinâmicos, em que o tempo torna-se o próprio objeto de estudo. Em estudos dinâmicos o objeto se torna uma dimensão da entidade "tempo".

observadas em todos os momentos.

Se o tempo for o fator preditivo, restringindo a observação ao que se entende por dinâmica, cada imagem deve ser tratada distintamente da entidade, visto que não se trata mais de obter a verdade sobre uma entidade, mas apenas uma imagem resultante em função de outra caraterística, no caso o tempo.

Da mesma forma, devem ser observadas as entidades do meio para obter-se a realidade sobre uma determinada entidade. Deve-se faze-lo entre todas as imagens entre todas as observações na unidade do tempo adotada.

#### B.1.4.9 Dimensão enquanto entidade

O atributo de apenas uma entidade não constitui por si só uma entidade. A caracterização dos domínios de um descritor se faz com a composição de todas as entidades que compõe o meio. Ou seja, o perfil das entidades compõe o descritor e o conjunto de descritores compõe a entidade. A caracterização do atributo é modificada conforme o conjunto de entidades.

Por extensão, não é possível avaliar as características de um objeto isoladamente, mas sempre em interação com as demais entidades e com o meio.

#### B.1.4.10 Paradoxo da proporcionalidade inversa da razão entre o conhecido e o desconhecido

O conhecimento de um objeto emerge características que presumem novas formas de observar, ampliando a quantidade de itens desconhecidos sobre o objeto conforme ilustrado na figura B.4. A medida que se conhece um objeto, ampliam-se as fronteiras sobre o que se desconhece, em outras palavras, a medida que cresce o conhecido, cresce ainda mais o desconhecido.

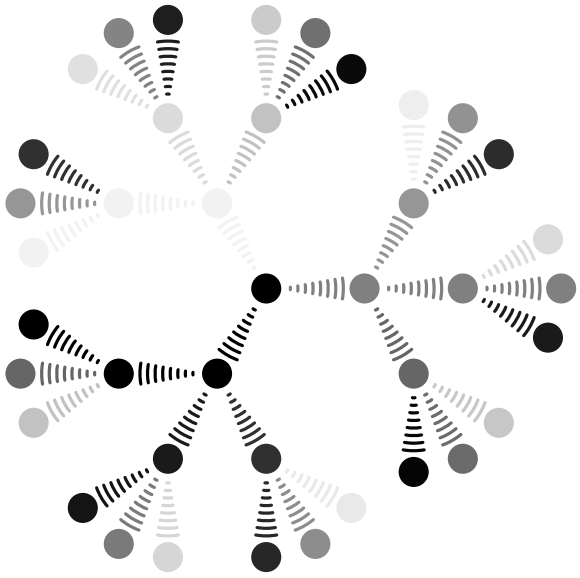


Figura B.4: **Paradoxo das dimensões do desconhecido.** A medida que uma propriedade torna-se mais conhecida, dela derivam outras, pela consequente ampliação da percepção, cuja natureza amplia os níveis do que se desconhece.

### B.1.5 Previsão de semelhantes

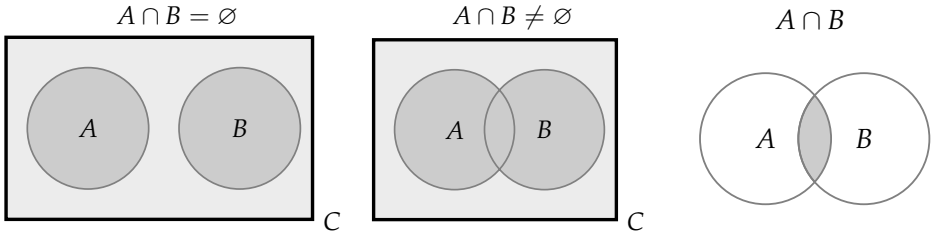
O ser apenas pode ser expresso na sua completude. No entanto, o fenômeno da interação pode ser previsto a partir da observação das características que traduzem a propensão à interação. Desta forma, na impossibilidade de avaliar todas as características das entidades que compõe o meio, seja pela restrição inerente da intangibilidade da definição completa do ser, seja por restrições computacionais, intelectivas ou devido ao desconhecimento, similaridades devem ser agrupadas a fim de reduzir o número de possibilidades, porém, sem perca da identidade que define a entidade.

Ainda que regidas por homens, as máquinas podem ser modeladas para possuírem uma forma própria de pensar e apreender. As similaridades podem ser verificadas por métodos gulosos que sejam capazes de processamento em tempo hábil. Logo se $A$ for semelhante a $B$ espera-se que ambos interajam com $C$ de maneira semelhante, sendo necessário apenas avaliar-se $A$ ou $B$ ou uma entidade que sintetiza as características apreendidas de $A$ ou $B$.

Na verdade, esta restrição é inseparável da decomposição de uma entidade em características que definem sua identidade. Inerentemente, o insumo para qualquer análise de uma entidade corresponde a análise de uma imagem desta entidade ao abrir mão de ao menos uma característica que a compõe. Sob este aspecto, as imagens de entidades diferentes podem equivaler, sendo as entidades tratadas como uma.

### B.1.6 Escopo dos paradigmas reducionista e integrativo

Os diagramas de Venn mostrados na figura B.5 ilustram o escopo dos paradigmas proposto e instituído. Na interação 3 é observado o reducionismo aplicado pelo modo vigente de estu-




(a) Interação 1 (b) Interação 2 (c) Interação 3

Figura B.5: **Escopo dos paradigmas do conhecimento.** As figuras (a) e (b) ilustram a demanda por todas as características que compõe as entidades em questão para a avaliação da interação conforme o paradigma integrativo, independente da ocorrência de características comuns. As características do meio devem ser incluídas na avaliação da interação. A figura (c) ilustra que o interesse do paradigma reducionista é apenas a intersecção das características de A e B, mantendo-se controladas ou isoladas as demais características, incluindo as do meio.

dar os fenômenos apenas em suas intersecções. Frequentemente, as variáveis ambientais são desconsideradas pela verificação aparente da não interferência o objeto de estudo, ou ainda, as condições do meio são isoladas e controladas.

O paradigma integrativo agrega múltiplas características disjuntas e conjuntas em prol da formação de uma visão global do objeto ou fenômeno estudado, trazendo consigo a desejável abordagem que tende a completeza, hoje, só adquirida e armazenada em cérebros com décadas de prostração em um tema sem que se envergue demais para erguer-se novamente e olhar ao redor.

Embora antagônicas existe uma relação entre essas duas abordagens na medida em que a insurgente pode usar como insumo os conhecimentos adquiridos com a tradicional, tomando-a como forma de cognição.

## B.1.7 Sumário dos paradigmas

Ambos os paradigmas podem responder a pergunta "se $A$ interage com $B$". No entanto, o paradigma integrativo requer que todas as entidades conhecidas semelhantes a $A$ sejam avaliadas em função da interação com entidades conhecidas e semelhantes a $B$.

### B.1.7.1 O paradigma integrativo

Não é possível analisar a interação de uma entidade sem avaliar a interação com o meio e, consequentemente, com as demais entidades que compõe o meio e da interação das demais entidades entre si.

Dada a impossibilidade de analisar a entidade em sua completude, apenas é possível analisar a imagem da entidade composta pelo conjunto de dimensões observadas comparativamente em relação às imagens de outras entidades.

Agrupar imagens similares viabiliza a análise e consequentemente a previsão da interação. Uma questão menor, a interação, não pode ser avaliada sem outras questões de igual grau que compõe uma questão maior, o meio. O meio, em si, pode constituir uma entidade quando comparado a outros meios ou grupos de entidades.

#### B.1.7.2 O modelo reducionista

As entidades estudadas devem ser isoladas e controladas as características que definem o meio, restringindo-se à característica ou fenômeno analisado. Uma vez evidenciado em quais condições o fenômeno é capaz de repetir, altera-se a característica observada controlando-se as demais, vislumbrando o dia em que todas as características tenham sido observadas.

Este modelo reconhece a impossibilidade de avaliar todas as características que definem uma entidade, pois a medida que um fenômeno é descrito, torna-se categorizado como característica, sendo objeto de novo estudo.

Uma questão maior necessariamente deve ser avaliada como o conjunto de estudos de questões mais específicas. A integração frequentemente ocorre apenas na forma discursiva nas discussões dos textos científicos ou intelectiva, não constituindo o cerne da técnica científica.

### B.1.8 Analogia computacional

A denominação adotada na abordagem metafísica possui termos correlatos na linguagem computacional. Assim, a **entidade** é tida como o **objeto** e as **características** como **atributos** estruturados com **metadados**. O **meio** ou **sistema** é composto pela união das objetos conhecidos. O sistema quando avaliado entre outros sistemas passa a constituir um objeto e tomar recursivamente como atributos os objetos que o compõe.

Objetos que possuem atributos de mesma natureza tem uma imagem comum tratada como **classe**. As relações entre classes são análogas às relações entre objetos. Uma classe pode ser tomada como **arquétipo** no contexto aglomerativo de atributos, visto que não se trata da imagem de um objeto existente que representa outros com as mesmas características, mas de uma imagem resultante de vários objetos semelhantes, porém com características distintas agrupadas com algum grau de similitude. **Instância** é um objeto da classe, por esta razão pode não representar um objeto existente, mas um nível de abstração da classe que acolhe propriedades de entidades existentes. Este conceito é importante pois a **classe** é o objeto de estudo na ciência de modo geral, não mais o objeto em si.

Assim como na **herança**, descrita pela modelagem orientação a objetos, os atributos de um objeto em um nível inferior também o são em um nível superior. Assim “unha” é atributo de “mão”, bem como de “pessoa” e de “população”. O tratamento dos atributos seguem primariamente a forma de registrar baseada na linguagem humana expressa cognitivamente e intermediada pelos sentidos como texto, números, imagens e sons. Secundariamente, os siste-

mas podem armazenar informação derivada que perfaça a própria visão dos algoritmos sobre o objeto na estrutura de linguagem própria da máquina.

Os objetos de interesse para avaliação da interação devem ser expressos, tanto quanto possível, com os atributos disponíveis em níveis superiores e inferiores. Assim, ao estudar "unha" ou "pessoa" devem ser agregados atributos que os diferenciem ao longo dos níveis ou modificações de classe como modelado segundo o **polimorfismo**. A análise será inútil se "unha" for um atributo descrito igualmente em todas as instâncias de "pessoa". Neste caso, devem ser acrescentadas instâncias diferentes de "unha" que agreguem a diferenciação de "pessoas".

Respostas aparentemente espúrias não devem ser descartadas, como a correlação de uma espécie de unha com o perfil de consumo de produtos de limpeza, sobretudo se a interação em questão for a aquisição deste tipo de produto. Em última instância nada é espúrio, pois a propensão de determinadas instâncias de "pessoa" em comprar o produto pode estar relacionada, por exemplo a fatores de expressão genômica polimórficos que interagem olfato com a formação da unha. O associação espúria pode ocorrer quando o meio ou a quantidade de apreensões podem ter sido insuficientemente observados ou quando a pergunta é restritiva, ou seja, não objetiva explicar o fenômeno enquanto ente.

O modelo proposto integra aprendizado de máquina ao conhecimento numérico, categórico ou na forma de texto e estabelece relações de dependência com o evento estudado.

## B.2 Aspectos algébricos da interação entre objetos

### B.2.1 Espaço de hipóteses

O atributo de um conjunto de entidades tomado como motivação do estudo é tratado como evento de estudo cuja presença ou incidência deseja-se determinar.

O **evento**, descrito a seguir, é a capacidade de interação entre dois objetos. A propensão do evento $e$ dentre um conjunto de eventos $E$ ocorrer diante de dada associação $a \in A$ entre entidades $f \in F$, expressa como $a \rightarrow e$ ($a$ implica em $e$), envolve a análise do conjunto de descritores (atributos) com valor semântico detonados em matrizes de frequência $M$ ou de distâncias $N$.

As características partilhadas pelos objetos com os desencadeadores diretos ou indiretos do evento permitem discriminar o papel a partir da avaliação sistêmica das entidades envolvidas.

Sendo assim, um modelo capaz de predizer o comportamento de associações, deve minerar características intrínsecas do conhecimento disponível sobre as entidades mediante a correspondência com os demais elementos. Em outras palavras, $a \rightarrow e$ é definido mediante a determinação do subconjunto de descritores $n_x$ que sejam vinculados ao evento ou conjunto de eventos. As técnicas de mineração de dados são empregadas para estabelecer o vínculo no espaço $N \times E$.

#### B.2.1.1 Espaço de hipóteses não mecanicístico para associações

Determinar interações requer a delimitação do universo de pesquisa das possíveis associações entre si. Logo, a combinação dos subconjuntos de objetos $F = \{f_1, f_2, ..., f_m | f_i \in F \wedge i = 1, 2, ..., m\}$ evidenciada na equação B.1 perfaz o universo de busca das possíveis associações.

$$A = F_1 \times F_2 \times ... \times F_m \forall F_i \subseteq F \wedge |F_i| \geq 2 \tag{B.1}$$

Cada associação é subconjunto de $F$. Porém a representação $a_i = F_i \subseteq F$ é verdadeira do ponto de vista matemático, mas não do semântico, visto que nem toda a associação se tornará entidade por não possuir características próprias. Constata-se que $|A| \leq \left|2^{|F|}\right| - |F|$ pois não é objetivo do presente modelo avaliar o subconjunto dos objetos isolados.

As associações que compõe o conjunto $A = \bigcup_{i=1}^{|A|} F_i$ obtido com a equação B.1 tem a amplitude calculada com as equações B.2 e B.3, onde $k$ representa a cardinalidade de cada associação.

$$min\,|A| = \sum_{k=2}^{|F|} \left\lceil \frac{|F|}{k} \right\rceil \tag{B.2}$$

$$max\,|A| = \sum_{k=2}^{|F|} \frac{|F|!}{k!(|F|-k)!} \tag{B.3}$$

Adotou-se $A^k$ como notação para a cardinalidade das associações. Por exemplo, o conjunto das associações binárias é representado por $A^2$, as ternárias por $A^3$.

Considerou-se $max|A| - min|A| \approx max|A|$ devido à diferença esperada para os casos práticos em que valores superiores a três ordens de grandeza para $|F|$ torna pouco representativo o valor mínimo.

#### B.2.1.2 Espaço de hipóteses para mecanismos de associações

A definição do mecanismo ou rota determina a ordem de precedência da ação de um objeto sobre outro como determinação da propensão ao evento. Neste caso, cada associação é tratada como vértice de um grafo completo, cujo espaço de hipóteses é constituído por todos os subgrafos direcionados possíveis sem repetição do mesmo fármaco.

Seja o grafo da associação $G_a = (F_a, E)$, onde $F_a$ é o conjunto de fármacos que compõe uma associação e $E$ são as arestas, deseja-se determinar o subgrafo $G'_a = (F'_a, E') \rightarrow e$, em que $F'_a \subseteq F_a$ e $E' \subseteq E$.

O número de subgrafos ou mecanismos possíveis é obtido com a equação B.4, considerando o número de arestas $|E| = |F_a|\,(|F_a| - 1)$.

$$|G_a| = 2^{|E|} \tag{B.4}$$

## B.2.2 Elementos do modelo preditivo

Os **objetos** variam na forma e conteúdo quanto à **descrição** disponibilizada pela fonte de informação. Um descritor pode ser empregado como **preditor**, o qual serve de parâmetro para avaliação sistemática dos **resultados** cuja verossimilhança é verificada conforme novos dados são coletados.

### B.2.2.1 Objeto

Os objetos $\{f_1, f_2, ..., f_m\} \in F$ são elementos da realidade cuja propensão a interação deseja-se conhecer em relação aos demais elementos do conjunto.

### B.2.2.2 Associação

A associação é dada pelo subconjunto de entidades $a = F_i \subseteq F$ de cardinalidade mínima igual a dois, definida na equação B.3.

A constituição do meio segundo o modelo integrativo requer a avaliação das associações entre todos os elementos que o constitui. No entanto, existe um domínio em que deseja-se avaliar associações conhecidas ou usuais. Desta forma constitui-se o conjunto $B$, o qual não necessariamente contempla todos os elementos em $F$, sendo o subconjunto de objetos que o constitui chamado $F_u$.

O conjunto de associações que conhecidamente culminam numa interação é denominado $A_k$. Em estudos populacionais, a relação esperada entre os conjuntos é mostrada na equação B.5.

$$|A_k| < |B| \ll |A| \tag{B.5}$$

### B.2.2.3 Atributo

A descrição que define uma entidade carrega valor intrínseco cujo potencial preditivo está relacionado à fonte de informação, completude, formato e alinhamento com o conhecimento existente. Distintas fontes de dados podem ser combinadas perfazendo imagens dos objetos cujo poder preditivo é denotado pela avaliação da semântica implícita.

Descritores podem ser categóricos (cor, classe terapêutica, doença, sexo), numéricos (tempo de meia vida, solubilidade), texto (posologia, mecanismo de ação), cadeia de caracteres (sequência proteica, lócus gênico), vetores (intervenções medicamentosas ao longo de unidades de tempo), grafos (rotas metabólicas) ou tempo (data de nascimento, intervalo da posologia).

O atributo é o subconjunto de descritores em função de um conjunto de entidades, sendo denotado como $M_x \subseteq M$. O número de atributos corresponde a quantidade de subconjuntos $D_x$.

**Atributo do objeto** Os objetos são definidos diferentemente conforme a forma de apreensão. A descrição de um conjunto de objetos ocorre pela união de distintos subconjuntos de descritores mostrados na equação B.6.

$$M = M_1 \bigcup M_2 \bigcup ... \bigcup M_n \tag{B.6}$$

**Atributo de associação e interação** A descrição da associação é a junção dos descritores dos respectivos objetos. Embora a associação não seja entidade do mundo real, é assim tratada para verificar-se a potencialidade de ser uma interação. O conjunto de atributos da associação é definido na equação B.7, sendo $A_k$ o conjunto de descritores para interações, ou seja, para associações conhecidas.

$$N = C \bigcup M \tag{B.7}$$

Em geral, $|A_k| \ll |A|$. Seja $A_p$ o conjunto de descritores de associações previstas, objetiva-se determinar quais descritores de cada objeto comporão a descrição $A_p \approx A_k$. A previsão da interação ocorre quando um elemento $a_p \neq \varnothing$, ou seja, quando exitem dimensões comuns de pelo menos um objeto à interação. O resgate dos elementos comuns podem constituir previsões comutativas entre associações que compartilhem os mesmos objetos. Além disso, relações indiretas com o sistema podem compor informações preditivas a qual culmina na extração semântica não trivial demonstrada na seção B.2.2.6.

Caso algum objeto não for descrito no conjunto de dados, a cardinalidade de $A$ pode ser inferior. Se o modelo preditivo não trabalhar com casos nulos, uma associação $a$ em que $\exists d = \varnothing \forall f \in F_x$ não será eleita para análise. Assim, o número de associações avaliadas está limitado às entidades descritas, dado que $|D| \leq |F| \, \forall d \neq \varnothing \wedge f \in F_x$.

No entanto, dado que a capacidade de interagir pode não estar relacionada a todos os objetos da associação, torna-se recomendável adotar modelos que contemplem a previsão da associação com pelo menos um elemento descrito.

**Preditor discreto** As interações são frequentemente classificadas de forma dicotômica ou ordinal. Um exemplo ordinal de domínio para classificação conforme significância clínica de interação medicamentosa é dado por $C = \{leve, moderada, grave, \varnothing\}$.

As associações a serem previstas possuem $a_p = \varnothing$. Nas interações que possuem $a_k \neq \varnothing$, possivelmente $\exists N \neq \varnothing$.

Neste caso, a tarefa de classificação pode ser assumida na forma supervisionada empregando-se as instâncias conhecidas para treinamento do modelo.

No caso de mineração de condições de saúde, sobretudo em base de pacientes, pode-se agrupar as interações medicamentosas e verificar a métrica que melhor discrimina cada doença ou detectar padrões frequentes verificando-se o envolvimento nas regras de associação com suporte razoável.

**Preditor contínuo** Dados numéricos pertencentes ao subconjunto de fármacos podem constituir informações preditivas. Constantes enzimáticas, duração do tratamento, pressão arterial ou glicemia são exemplos de preditores contínuos.

#### B.2.2.4 Imagem e espectro

A identidade de um objeto é apreendida na forma de uma imagem dada a impossibilidade de avaliá-lo por completo conforme abordado na seção B.1.4.2. Desta forma, a imagem do objeto $f$ é a apreensão do conhecimento disponível $d$ em relação ao escopo que cada atributo $D_x \subseteq D$ oferece em relação aos demais elementos do conjunto $F$. Ou seja, não é possível compreender uma dada característica $d \in D_x$ sem a avaliação inserida no âmbito de um dado número de instâncias $D_x \subseteq D$ para $\bigcup_{i=1}^{d \to |D|} D_i(F)$.

A disposição dos atributos para o observador, seja humano ou o computador, faz com que um subconjunto de $D$ constitua a imagem depurada para a análise. Logo, a imagem é denotada como $M_x \subseteq M$, aproximando-se da completude conforme a capacidade do observador em contemplar mais elementos $d$.

Além do modelo $\gamma \in \Gamma$ de **tratamento** e **seleção dos atributos** que construirão a imagem, outro recurso é a transformação da imagem em um espectro $W$ o qual deve herdar a expressão preditiva da imagem. Este espectro pode ser, por exemplo, uma interpretação descritiva dos elementos do conjunto $D$ ou algum modelo $\phi \in \Phi$ de decomposição matemática.

Salienta-se que a decomposição espectral pode modificar as dimensões em $D$, porém $\sum_{i=1}^{|F|} |D(f_i)| \forall f \in F$ deve permanecer inalterado, assegurando a correspondência $W \succ M$. Uma forma comum de representar o conjunto de descritores é a inserção dos descritores dos objetos ao longo das linhas, sendo cada atributo alocado em uma coluna, tal qual em uma matriz ou banco de dados. Nesta representação, o espectro resultante da transformação da imagem não pode sofrer modificação na distinção longitudinal em relação ao objeto, somente transversal em relação aos atributos.

#### B.2.2.5 Distância e amplitude

Atributos categóricos no formato de texto ou numérico discretizado decompostos vetorialmente resultam em matrizes de frequência. Desta forma, os vetores que representam a imagem $m$ ou espectro $w$ de cada objeto ou a variável numérica $d$ empregada diretamente, podem ser usados para calcular-se a distância entre as imagens dos objeto da associação, resultando em um valor numérico chamado **amplitude da associação**.

As métricas de distância são agrupadas no conjunto $\Delta$ e abordadas na seção 4.4.4, tendo exemplos enumerados no anexo C. Desta forma, cada métrica pode oferecer uma dimensão para a avaliação preditiva da associação, possibilitando ao modelo diferentes imagens com variáveis poderes de expressão conforme a estrutura original dos dados.

A distância $\delta \in \Delta$ entre dois elementos $f$ é denotada por $\delta(f_i, f_j)$. Sendo $a$ composto por $\{f_i, f_j\}$ esta distância representa a amplitude $\delta(a)$. O conjunto de amplitudes é representado por $\Delta(A)$.

Se objetos na condição $\exists d_i = \varnothing$ forem adotados e ainda assim o a contribuição preditiva do atributo ou conjunto de atributos em questão se mantiver, torna-se evidenciado que o conhecimento das características de apenas um dos fármacos pode ser suficiente para a classificação da interação. A força preditiva estará em apenas um dos elementos descritivos do par, com base no panorama de fármacos.

#### B.2.2.6 Modelos e resultados

Modelo é a conjunção das funções de aprendizagem $\gamma \in \Gamma$ com os descritores e preditores adaptados ao formato de entrada em função dos resultados esperados. A partir do elemento de desempenho que decide as ações a executar, o elemento de aprendizagem modifica o de desempenho para que ele tome decisões melhores[Russel & Norvig, 2003].

As análise de cada imagem $N$, espectro ou amplitude pelo modelo $m \in M$ gera um conjunto de previsões $R$ que devem ser avaliadas por métricas $P$ de desempenho $\forall c \neq \varnothing$ ou métodos de comparação de resultados ainda que manualmente. Métricas de desempenho são abordadas na seção 2.5.5.2.

A figura **??** é um exemplo que esquematiza a combinação da descrição dos objetos e associações em função do preditor, gerando respostas que retroalimentam o modelo, constituindo a aprendizagem.

## B.3 Mineração de interações entre objetos

Os atributos $D$ extraídos a partir de aferições numéricas, linguagem natural (p.ex. texto) ou ontológica (p.ex. estruturas hierárquicas) alimentam os modelos preditivos que devem elaborar funções de aprendizagem capazes de discriminar os termos mais relevantes para a detecção de interações.

Uma vez definido o espaço de busca deve-se compreender como os dados foram apreendidos. A interpretação humana ou computacional decorrente de um ou mais descritores deve gerar imagens cujos padrões sejam úteis para que os modelos possam apreender as características preditivas. A forma de apreender atributos é vista na seção B.3.3 e a decomposição é vista na seção 4.4.5.

Dentre os modelos existentes focou-se nos métodos supervisionados (introduzidos na seção 2.5.4.1), os quais geram funções de aproximação a partir de instâncias conhecidas. Porém abordagens como computação natural ou aprendizagem por reforço podem ser adotadas segundo as premissas do modelo integrativo.

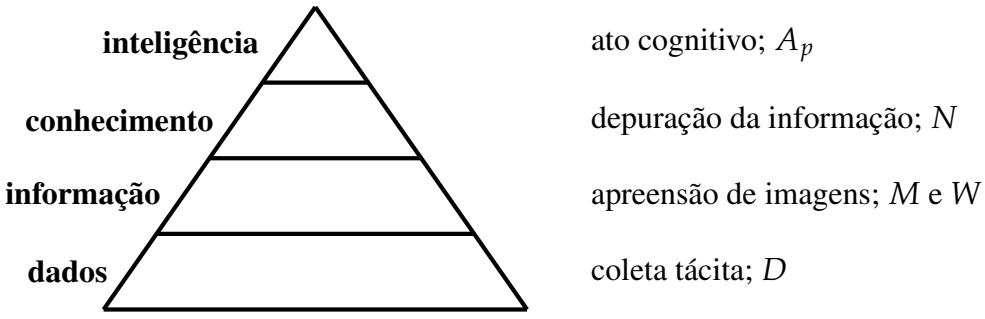



Figura B.6: **Pirâmide do conhecimento.** Os dados são coletados em grande quantidade, porém somente constituem informação quando compilados de forma relevante. As informações empenhadas mediante a incorporação e análise gera o conhecimento. A inteligência advém da superação do conhecimento pela criação de novas informações puramente intelectivas.

## B.3.1 Modelo de aprendizagem

A semântica fundamenta a significação e consequentemente possui aspecto nivelador da forma de apreensão e do conhecimento. Algum nivelamento semântico é uma característica necessária para a comunicação a fim de que emissor e receptor compartilhem a mesma concepção do sinal.

Na prática, as variáveis distorções do sinal fazem com que a apreensão da realidade sofra interpretações conforme o indivíduo. Estas variações repercutem do armazenamento das informações ao processo de aprendizagem, seja humana ou computacional.

Sendo a realidade percepção conforme enunciado na seção B.1.4.5, a concepção sobre um objeto é tanto fidedigna quanto mais dimensões forem contempladas. Logo, a fim de tornar as variações de apreensão da realidade frutos do acaso, diversas percepções de fontes diferentes devem ser consideradas pelo modelo preditivo, o qual deve caminhar pela pirâmide do conhecimento ilustrada na figura B.6 com dados em diversas e distintas dimensões.

### B.3.1.1 Preditor ou classe

O evento ou fator preditivo é o ponderador da função alvo de aprendizagem. Em outras palavras, é o atributo $c \in C$ que direciona a avaliação do modelo para obtenção da resposta almejada. O preditor herda as propriedades dos descritores narradas na seção B.2.2.3.

O principal fator da escolha do preditor é a capacidade resolutiva na distinção das características de interesse. Porém, qualquer atributo ou conjunto de atributos que descrevam uma associação distinta pode ser empregado em modelos preditivos. Na prática opta-se por descritores categóricos ou contínuos, tornando mais intuitiva a compreensão dos resultados minerados.

Em uma tarefa de classificação é a classe que deseja-se conhecer para cada associação. Na divisão das associações em agrupamentos, o preditor é o agente discriminador dos grupos, sendo, neste caso, possível adotar mais de um descritor, ou mesmo, entender quais atributos são diferenciadores dos elementos. Os preditores são os agentes consequentes na análise por regras de associação.

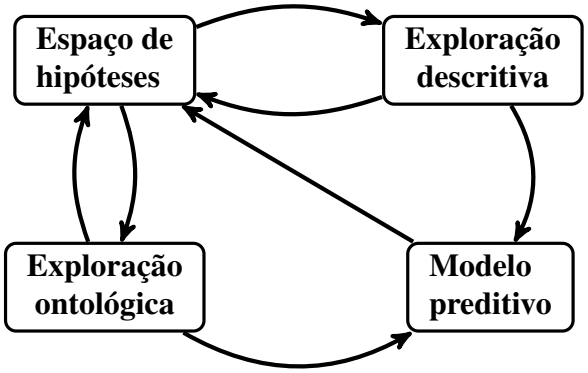



Figura B.7: **Processos para previsão de associações medicamentosas.** O espaço de hipóteses, constituído pelas possibilidades de associações vinculáveis a causa de um determinado evento, é explorado ontologicamente ou minerado com técnicas descritivas para o estabelecimento do modelo preditivo, o qual retroalimenta o espaço de associações-evento nas iterações subsequentes.

O preditor ou conjunto de preditores é um fator importante na performance do aprendizado de máquina, sendo desejável o balanceamento ao longo do conjunto de associações estudadas e objetividade na resposta preditiva almejada. No entanto, em tarefas supervisionadas constitui o principal limitante da capacidade de generalização do modelo.

## B.3.2 Exploração do espaço de hipóteses

Conforme abordado na seção B.2.1, a explosão combinatorial de associações possíveis, vista na figura B.8, decorre do acréscimo do número de entidades e pode demandar alternativa à exploração completa do universo de hipóteses, discorrido na seção B.3.2. A alternativa mais profícua é reduzir o universo de hipóteses à análise de associações mais simples como as binárias. Permanecendo inviável a análise completa do universo reduzido, outros recursos são a tomada de distâncias, considerando cada atributo como independente, a formação de arquétipos pelo uso de bases ontológicas relacionadas às entidades (seção A.5.4.1) ou a mineração de dados descritiva (seção 2.5.3). As previsões podem retroalimentar a exploração do espaço de hipóteses aperfeiçoando o espaço de busca ou o modelo da iteração anterior (figura B.7).

O espaço de hipóteses cresce exponencialmente com o número de entidades. Uma solução é agrupar imagens semelhantes conforme classificações ou ontologias disponíveis, ou ainda, empregar técnicas descritivas de mineração de dados.

Caso $|F|$ for na ordem de centenas ou milhares ainda em $A^2$, será impeditivo o processamento de todo o conjunto de hipóteses (equação B.3). Desta forma, novos subconjuntos $A_i \subset A$ serão constituídos a partir de arquétipos de grupos de objetos ou associações, conforme visto na seção B.3.2.3.

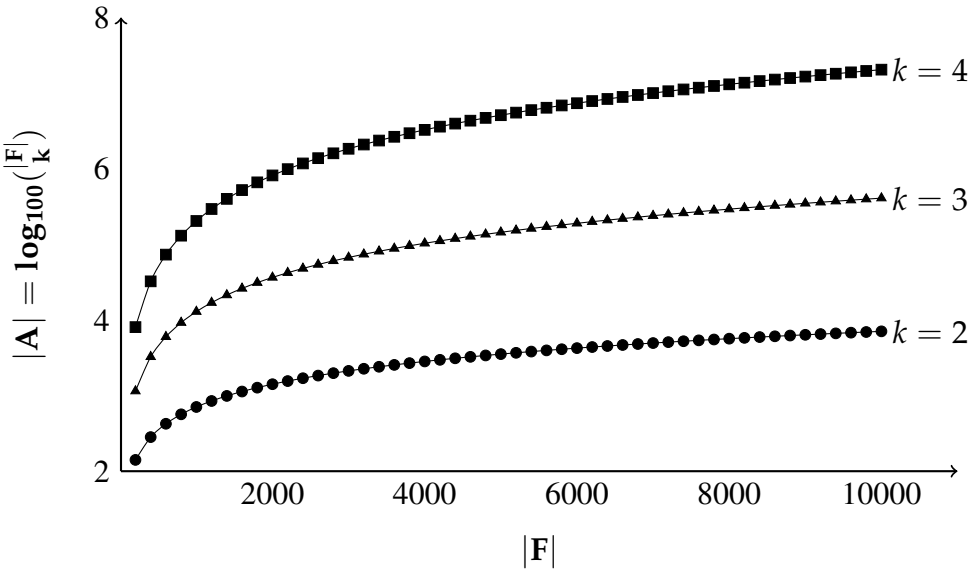



Figura B.8: **Espaço de associações.** O acréscimo de objetos $f \in F$ aumenta exponencialmente o espaço binomial de associações $A$.

### B.3.2.1 Associações binárias

Conforme abordado na seção 4.4.3.2, o recurso explorado neste texto é a avaliação de associações binárias. Não constitui um escopo tão limitado avaliar a relação entre objetos aos pares, dado que esta estrutura pode ser tomada como unidade fundamental para a avaliação de associações de outras ordens. A detecção de uma dada interação pode nortear comutativamente interações ternárias, quaternárias, e assim por diante. Por exemplo, se a interação de $A$ com $B$ necessita de $C$ para ocorrer, as descrições das associações de $A$ com $B$, $A$ com $C$ e $B$ com $C$ podem ser compiladas previamente para a avaliação da interação em questão com decorrente redução do espaço de busca.

### B.3.2.2 Exploração ontológica

Ao desenvolver conhecimento sobre determinado objeto, a descrição na forma ontológica ou hierárquica estabelece um norteador para a avaliação de interações. Esta avaliação é proporcional à similaridade esperada do comportamento de objetos pertencentes ao mesmo grupo (figura B.10).

A concepção ontológica remonta que objetos da mesma classe compartilham as mesmas características de determinada dimensão. Sob esta definição, o preditor também é uma dimensão ontológica, pois deseja-se saber o que o objeto é ou não a partir das imagens depuradas a partir dos dados.

Diante da ótica de uma ontologia, os descritores dos respectivos objetos, por exemplo $F_x = \{f_1, f_2, f_3\}$, resultam no arquétipo da classe $D_x = d_1 \bigcup d_2 \bigcup d_3$. Se o conjunto universo em uma dimensão for $D_x = \{a, b, c, d, e, f\}$, sendo $d_1 = \{a, b\}$, $d_2 = \{b\}$ e $d_3 = \{a, d\}$, uma descrição para o arquétipo será $d_x = \{a, b, d\}$. Ao agrupar por similaridade inúmeras

instâncias das entidades cuja interação deseja-se avaliar a consequente redução do conjunto de possibilidades atribuirá uma análise aproximada para a formação dos padrões, os quais, serão expandidos para os demais elementos do conjunto. Se $F_x$ interagem com $F_y$, será considerado que todos os elementos contidos em $F_x$ interagem com os elementos $F_y$. Espera-se que a concepção de grupo seja capaz de fornecer insumo para a avaliação dos atributos compartilhados entre arquétipos diferentes.

A conclusão em respostas dicotômicas será que "todos os elementos de um grupo interagem ou não com os de outro grupo". Se respostas valoradas em termos de probabilidade forem possibilitadas pelo modelo de extração ontológica, o teor reportado dirá que "alguns elementos de um grupo interagem com alguns elementos de outro grupo". Os grupos cuja presença de interações inexistir ou for desprezível serão descartados pelo modelo.

**Ontologia manualmente acurada** Um exemplo é apresentado na tabela B.1, a qual ilustra a redução do espaço de hipóteses ao agrupar os fármacos por classificação ATC da OMS[WHO, 2011]. Empregando-se combinações duas a duas, o quarto nível da ATC possibilita a comparação de cerca de 30 mil grupos farmacológicos ao invés de 9,4 milhões na comparação entre todos os pares de 4.342 fármacos.

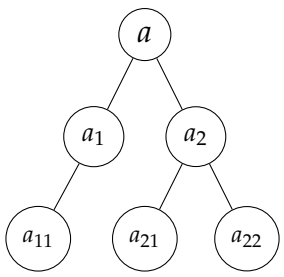
Tabela B.1: Espaço de hipóteses para avaliação de associações de acordo com o nível da classificação ATC.*

| nível | n | $\bar{x}$ | CV% | min | $\tilde{x}$ | max | $k = 2^{**}$ | $k = 3^{**}$ |
|---|---|---|---|---|---|---|---|---|
| anatômico | 14 | 345,2 | 52,8 | 93 | 273,0 | 684 | 91 | 455 |
| terapêutico | 90 | 53,7 | 91,5 | 1 | 39,5 | 277 | 4.005 | 121.485 |
| farmacológico | 243 | 19,9 | 93,4 | 1 | 14,0 | 106 | 29.646 | $2,4 \cdot 10^6$ |
| químico | 777 | 6,2 | 91,3 | 1 | 4,0 | 29 | $3,0 \cdot 10^5$ | $7,8 \cdot 10^7$ |

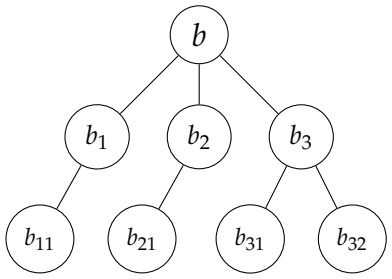
$n$ = total de elementos.
Descritores dos fármacos no grupo (4.342 subtâncias): $\bar{x}$ = média ponderada, $CV\%$ = coeficiente de variação = $\sigma \div \bar{x} \times 100$ ($\sigma$ =desvio padrão), min=mínimo, $\tilde{x}$ =mediana, max=máximo.
* Classificação segundo WHO [2011].
** Espaço de hipóteses $\left|A^k\right| = \frac{|F|!}{k!(|F|-k)!}$.

**Descoberta de ontologia** O poder preditivo decai inversamente à disseminação de uma dada característica nos elementos do conjunto de arquétipos. Desta forma, a ontologia manualmente acurada pode não ser útil na discriminação do modelo para a previsão, demandando atributos que melhor a caracterizem. Uma forma de contornar este problema é a detecção de uma ontologia implícita nos atributos, considerando a definição de ontologia enquanto forma estruturada de avaliação de um conjunto de objetos sob o mesmo domínio.

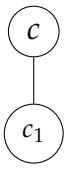
O algoritmo B.1 pode ser empregado na seleção de uma ontologia ou para remoção de atributos. Ele desconsidera atributos com elevado teor de casos nulos e na mesma medida inclui




(a)




(b)




(c)

$$F = \{f_1, f_2, f_3, f_4, f_5\}$$

$$A^2 = \begin{Bmatrix} \{f_1, f_2\} & \{f_1, f_3\} \\ \{f_1, f_4\} & \{f_1, f_5\} \\ \{f_2, f_3\} & \{f_2, f_4\} \\ \{f_2, f_5\} & \{f_3, f_4\} \\ \{f_3, f_5\} & \{f_4, f_5\} \end{Bmatrix}$$

(d)

| $d_h(f)$ | nível 3 | nível 2 | nível 3 |
|---|---|---|---|
| $d_h(f_1) =$ | $\{a_{11}, b_{11}\}$ | $\{a_1, b_1\}$ | $\{a, b\}$ |
| $d_h(f_2) =$ | $\{a_{21}, b_{31}\}$ | $\{a_2, b_3, c_1\}$ | $\{a, b, c\}$ |
| $d_h(f_3) =$ | $\{a_{22}, b_{21}\}$ | $\{a_2, b_2\}$ | $\{a, b\}$ |
| $d_h(f_4) =$ | $\{b_{14}\}$ | $\{b_1\}$ | $\{b\}$ |
| $d_h(f_5) =$ | $\{b_{14}\}$ | $\{b_1\}$ | $\{b\}$ |

(e)

nível 1

$$f_{ab} = \{f_1, f_3\}$$
$$f_{abc} = \{f_2\}$$
$$f_b = \{f_4, f_5\}$$

| nível 2 | | nível 3 | |
|---|---|---|---|
| $f_{a1b1}$ | $= \{f_1\}$ | $f_{a11b11}$ | $= \{f_1\}$ |
| $f_{a2b3c1}$ | $= \{f_2\}$ | $f_{a21b31}$ | $= \{f_2\}$ |
| $f_{a2b2}$ | $= \{f_3\}$ | $f_{a22b21}$ | $= \{f_3\}$ |
| $f_{b1}$ | $= \{f_4, f_5\}$ | $f_{b14}$ | $= \{f_4, f_5\}$ |

(f)

$$A^2_{n1} = \{\{f_{ab}, f_{abc}\} \quad \{f_{ab}, f_b\} \quad \{f_{abc}, f_b\}\}$$

$$A^2_{n2} = \begin{Bmatrix} \{f_{a1b1}, f_{a2b3c1}\} & \{f_{a1b1}, f_{a2b2}\} \\ \{f_{a1b1}, f_{b1}\} & \{f_{a2b3c1}, f_{a2b2}\} \\ \{f_{a2b3c1}, f_{b1}\} & \{f_{a2b2}, f_{b1}\} \end{Bmatrix}$$

$$A^2_{n3} = \begin{Bmatrix} \{f_{a11b11}, f_{a21b31}\} & \{f_{a11b11}, f_{a22b21}\} \\ \{f_{a11b11}, f_{b14}\} & \{f_{a21b31}, f_{a22b21}\} \\ \{f_{a21b31}, f_{b14}\} & \{f_{a22b21}, f_{b14}\} \end{Bmatrix}$$

(g)

$$p(A^2_{n1}) = \{\cancel{\{f_{ab}, f_{abc}\} = 0,3} \quad \cancel{\{f_{ab}, f_b\} = 0,7} \quad \{f_{abc}, f_b\} = 0,9\}$$

$$A^2_{n1} = \begin{Bmatrix} \cancel{\{f_1, f_2\}} & \cancel{\{f_1, f_3\}} & \cancel{\{f_1, f_4\}} & \cancel{\{f_1, f_5\}} & \cancel{\{f_2, f_3\}} \\ \{f_2, f_4\} & \{f_2, f_5\} & \cancel{\{f_3, f_4\}} & \cancel{\{f_3, f_5\}} & \cancel{\{f_4, f_5\}} \end{Bmatrix}$$

(h)

Figura B.9: **Redução ontológica do espaço de busca.** *a*) *b*) *c*) Ontologia $H$ a qual classifica os elementos de $F$ em três níveis. *d*) Conjunto universo $A$ dos elementos de $F$ combinados dois a dois. *e*) Classificação ontológica hipotética $H$ de cada elemento $f$. *f*) Arquétipos $f \in F_H$, os quais herdam as características dos elementos em cada nível ontológico. *g*) Espaço de busca em $F_H$. *h*) Resultado hipotético da previsão para o nível 1. *i*) Redução do espaço de busca considerando a remoção das associações com menor probabilidade (dois traços) ou considerando apenas a maior probabilidade (nenhum traço).

àqueles que possuem menos níveis distintos de classificação, permitindo um número controlado de grupos para a geração de arquétipos. O limiar deve situar-se entre 0 e 1. Se $l = 0,9$ significa que atributos com 10% de casos nulos em com 90% de casos distintos serão descartados. Este último fator é importante pois não é possível a verificação de padrões frequentes quando os casos distintos tendem a 100%.

Alternativamente, o algoritmo B.1 pode ser aplicado como filtro para a remoção das colunas em que houver uma quantidade de valores nulos ou altamente incidentes por não contribuírem para a distinção das instâncias. Este algoritmo demanda um limiar previamente fornecido.

O filtro remove colunas com elevado teor de casos nulos e, na mesma medida, inclui àqueles que possuem menos níveis distintos de classificação. O limiar fornecido deve situar-se entre 0 e 1. Se $l = 0,9$ significa que colunas com mais de 10% de casos nulos em com mais de 90% de casos distintos serão descartados. Este último fator é importante pois não é possível a verificação de padrões frequentes quando os casos distintos tendem a 100%.

**Algoritmo B.1** Filtro de atributos com base em um limiar de expressividade entre casos nulos e casos únicos.

```
1:  função FILTRAATRIBUTO(atributo M^{m×n}, limiar l)
2:      q ← m;                                          ▷ total de fármacos
3:      para cada coluna m_{mj} faça
4:          qnn = Σ_{i=1}^{m} m_{ij};                   ▷ casos não nulos
5:          se qnn/q > l então                          ▷ avalia a completude
6:              qu ← m;                                 ▷ casos únicos
7:              se qu/q < l então                       ▷ avalia a disseminação
8:                  R ⇐ m_{mj};                         ▷ concatena o atributo à saída
9:              fim se
10:         fim se
11:     fim para
12:     retorna R
13: fim função
```

**Ontologia enquanto descritor** A ontologia é uma característica do objeto e seu conjunto de informações podem ser usadas, por exemplo, na forma de vetor binário. Conforme visto na figura B.9 o primeiro nível da árvore pode gerar um vetor com três posições, o segundo com seis e o terceiro com sete. Uma associação cujos objetos pertencem ao mesmo domínio pode ser descrita com a frequência dos termos como no caso $\{f_2, f_4\}$ em que o vetor referente ao nível 1 será $d_{n1}(\{f_2, f_4\}) = \begin{bmatrix} 1 & 2 & 1 \end{bmatrix}$, dado que os dois objetos são classificados como "$b$". Esta abordagem pode oferecer aporte para árvores de decisão e técnicas de agrupamento.

### B.3.2.3 Exploração descritiva

A definição de cada objeto como vetor de características no espaço n-dimensional viabiliza a geração de um arquétipo o qual sintetiza as características em um único elemento mantendo a

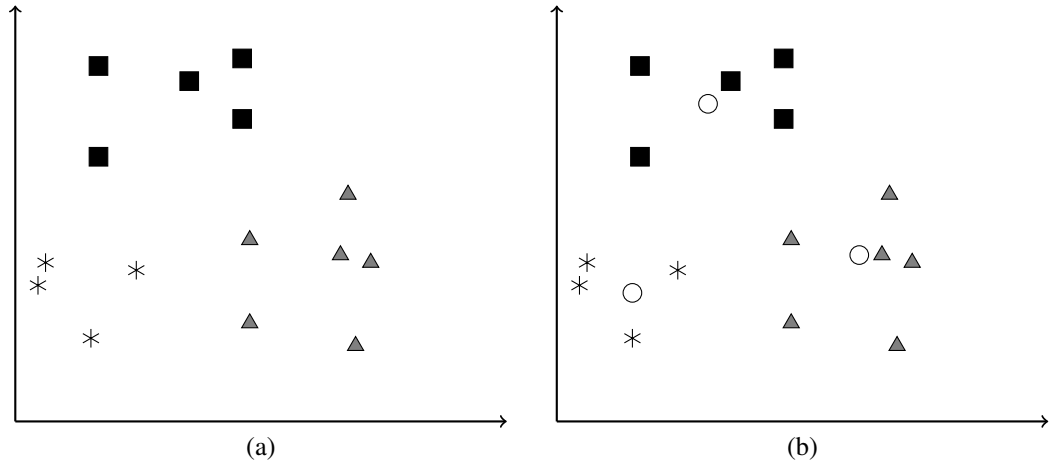



Figura B.10: **Arquétipos.** O espaço de hipóteses pode ser reduzido ao se apreender uma única imagem para cada o grupo. Assim, os grupos serão comparados por um elemento resultante chamado arquétipo, simbolizado pelo círculo.

relação de identidade do grupo conforme ilustrado na figura B.10.

As comparações entre os arquétipos reduzem o custo computacional alimentando inicialmente o modelo preditivo. Nas iterações posteriores, eleva-se o grau de refinamento com a avaliação entre os elementos dos grupos em que houve previsão em alto nível do evento em estudo.

A sumarização de características pode ser realizada por relações de frequência do aparecimento de termos descritivos, como a soma, $k$ atributos mais frequentes ou média. Outra definição ocorre na tomada de faixas de distâncias entre todas as entidades para cada atributo. Estas técnicas oferecem um modo intuitivo e rápido de sumarização. Porém, técnicas de de agrupamento como k-médias, agrupamento hierárquico aglomerativo ou DBSCAN, após a definição da métrica de distância e do critério de formação dos agrupamentos pode ser usados para distinção dos dados.

Nos casos em que houver ruído, ou seja, descritores que reduzem a capacidade discriminativa do modelo, empregam-se técnicas matemáticas como análise dos componentes principais (PCA) ou decomposição por valores singulares (SVD). Além da remoção dos ruídos, possibilitam redução de dimensionalidade ao extrair as relações que distinguem as instâncias.

Uma forma intermediária de reduzir o espaço de busca é a comparação dos elementos do conjunto de objetos com os arquétipos derivados.

#### B.3.2.4 Silogismo da Generalização do evento ao grupo de associações

Nos casos em que o modelo preditivo for supervisionado, ou seja, baseado em associações de efeito conhecido, o agrupamento deve ser formado em função do evento estudado. Caso contrário, as generalizações perderão a capacidade discriminativa.

Ao determinar a forma de agrupar associações, observam-se cinco possíveis desdobra-

mentos para cada subconjunto $A_i \subset A$, no espaço $A \times E$ [3]:

1. todas as interações do grupo implicam na mesma classificação,

$$a \rightarrow e \forall a \in A_i \tag{B.8}$$

2. cada interação implica em alguma classificação,

$$a \rightarrow e_1 \vee a \rightarrow e_2 \vee ... \vee a \rightarrow e_n \forall a \in A_i \tag{B.9}$$

3. associações desconhecidas em meio a interações com a mesma classificação,

$$a \rightarrow e \vee a \rightarrow ? \forall a \in A_i \tag{B.10}$$

4. associações desconhecidas em meio a interações com diferentes classificações,

$$a \rightarrow e_1 \vee a \rightarrow e_2 \vee ... \vee \rightarrow e_n \vee a \rightarrow ? \forall a \in A_i \tag{B.11}$$

5. nenhuma associação é conhecida.

$$a \rightarrow ? \forall a \in A_i \tag{B.12}$$

O arquétipo no primeiro caso é um bom artifício de comparação para o método preditivo e representa melhor os elementos do grupo por ser uma generalização consolidada.

No segundo caso devem ser conjugadas técnicas *multi-label*, as quais tratam o problema da mesma instância admitir mais de uma classificação. Cada grupo-evento deve ser tratado como um arquétipo.

Nos casos 3 a 5, não devem ser estabelecidos arquétipos para as instâncias desconhecidas, adotando-se os dados originais.

Desta forma, espera-se que generalizações associadas ao evento reduza o espaço de hipóteses. Caso os arquétipos não forem estabelecidos de acordo com o preditor, pode-se perder a especificidade se não houver correlação entre o evento e o método adotado para discriminar os grupos. Mesmo assim, este recurso pode ser adotado para identificar-se os grupos com menor tendência ao evento e descartar seus elementos nas iterações mais específicas.

[3] Considerou-se ? como desconhecido, diferente de nulo $\varnothing$.

### B.3.3 Manipulação de atributos

#### B.3.3.1 Mineração em texto

A extração de termos que compõe a descrição de um medicamento apresenta-se como a etapa crítica na mineração em texto.

Técnicas de processamento de linguagem natural decompõe frases em estruturas linguísticas como ”sujeito“ e ”predicado“, ou ainda classes gramaticais com ”substantivo“, ”adjetivo“, ”advérbio“, ”pronome“ e ”verbo“ para identificar as entidades envolvidas e coletar informações. Estas técnicas são destinadas, sobretudo, à extração de semântica humana em textos facultativamente tipados.

Contudo, o uso de campos adotados como definição de um aspecto do objeto provê alto valor semântico computacional cuja simples presença da palavra ou termo carrega padrões importantes para a leitura por modelos preditivos, mesmo sob a perda da verificação de sua afirmação ou negação no contexto. A definição humana de um objeto, mesmo sendo determinística, torna-se integrativa quando avaliada em conjunto com as demais definições dos elementos do conjunto conforme exposto na seção B.2.2.4.

Adotando-se a linguagem de mineração em texto, cada atributo $D_x \in D$ carrega um dicionário de termos, ou seja, um vetor de termos distintos. A presença ou frequência destes termos para cada objeto alimenta a matriz de cardinalidade $|F| \times |D|$.

Os objetos respectivos à associação podem ter a frequência conjunta de termos alocadas na matriz $|A| \times |D|$. Ou ainda, sob uma métrica de distância $\Delta$, vetores respectivos aos elementos de $F$ podem ter as distâncias avaliadas, constituindo a amplitude de cada associação em um vetor de cardinalidade $|A|$. Uma terceira forma é justapor os dicionários de modo que cada objeto ocupe sua respectiva posição, gerando-se uma matriz de dimensões $\max(k) \cdot |A| \times |D|$, onde $max(k)$ é a cardinalidade máxima observada no conjunto de associações. Os tratamentos dos descritores geram imagens $M$ que podem ser usadas no modelo integrativo, pois não partem somente da observação única de $f$, mas ao colocá-la como um vetor que indica o que $f$ expressa e o que não expressa, o posiciona em relação à amplitude do atributo obtida a partir da observação dos demais objetos.

A presença ou a ausência de uma palavra em uma definição não constitui sozinha o que o objeto é ou não é. No entanto, a conjunção de termos que indiquem o que participa ou não da identidade do objeto pode ser suficiente para caracterizá-lo, não sob a semântica humana pois esta se perde com o tratamento, mas sob a semântica matemática e computacional. Ainda, a avaliação de diversos atributos permite que os modelos verifiquem padrões e captem a essência do objeto, distinguindo-os segundo o preditor desejado.

### B.3.4 Decomposição de atributos

A imagem é uma interpretação direta dos dados coletados. O espectro é a sobreposição de múltiplas imagens considerando todo o conjunto de dados. O espectro não possui as características originais das imagens abrigadas, mas avaliado sobre um determinado filtro, traduz em si algo que o conjunto de imagens expressa. Um exemplo de decomposição espectral é a SVD[Elden, 2006] (seção 5.3.5).

Em diversas etapas a decomposição pode ser usada conforme ilustrado na figura **??**. A decomposição proximal ocorre diretamente em uma coleção de atributos numéricos dos descritores. A intermediária é realizada após a formação de uma imagem numérica como, por exemplo, vetores de termos. Por fim, a decomposição distal ocorre nas distâncias do conjunto de associações com os valores de cada atributo concatenados. A decomposição pode ser realizada de forma proximal após a computação da frequência de termos ou medial, em uma etapa antes do retorno de $M$.

Em qualquer etapa a decomposição somente pode ser realizada quando as posições referentes aos atributos contiverem aporte semântico de igual impacto na discriminação das entidades. Caso contrário, a decomposição não remontará ao significante da matriz original.

### B.3.5 Sumário do modelo

A essência de um objeto pertencente a um conjunto é captada a partir da comparação com os demais. Esta comparação pode ser realizada diretamente ou com uma métrica de distância. A definição do atributo ocorre a partir da observação entre todos os objetos do conjunto ou de todas as associações. A coleção de observações e decorrente ponderação das distâncias entre objetos pode apresentar padrões correlatos à característica em estudo. Os padrões de um número suficiente de instâncias conhecidas podem ser estendidos à instâncias desconhecidas.

# Apêndice C

# Estratégias de busca

## C.1 Medline

(

"Artificial Intelligence" OR "Intelligence, Artificial" OR "Computer Reasoning" OR "Reasoning, Computer" OR "Machine Intelligence" OR "Intelligence, Machine" OR "AI (Artificial Intelligence)" OR "AIs (Artificial Intelligence)" OR "Machine Learning" OR "Learning, Machine" OR "Knowledge Representation (Computer)" OR "Knowledge Representations (Computer)" OR "Representation, Knowledge (Computer)" OR "Representations, Knowledge (Computer)" OR "Computer Vision Systems" OR "Computer Vision System" OR "System, Computer Vision" OR "Systems, Computer Vision" OR "Vision System, Computer" OR "Vision Systems, Computer" OR "Knowledge Acquisition (Computer)" OR "Acquisition, Knowledge (Computer)" OR "Acquisitions, Knowledge (Computer)" OR "Knowledge Acquisitions (Computer)" OR "Expert Systems" OR "Expert System" OR "System, Expert" OR "Systems, Expert" OR "Fuzzy Logic" OR "Logic, Fuzzy" OR "Knowledge Bases" OR "Base, Knowledge" OR "Bases, Knowledge" OR "Knowledge Base" OR "Knowledgebases" OR "Knowledgebase" OR "Knowledge Bases (Computer)" OR "Base, Knowledge (Computer)" OR "Bases, Knowledge (Computer)" OR "Knowledge Base (Computer)" OR "Neural Networks (Computer)" OR "Network, Neural (Computer)" OR "Networks, Neural (Computer)" OR "Neural Network (Computer)" OR "Models, Neural Network" OR "Model, Neural Network" OR "Network Model, Neural" OR "Network Models, Neural" OR "Neural Network Model" OR "Perceptrons" OR "Perceptron" OR "Connectionist Models" OR "Connectionist Model" OR "Model, Connectionist" OR "Models, Connectionist" OR "Neural Network Models" OR "Robotics" OR "Support Vector Machines" OR "Support Vector Machine" OR "Vector Machine, Support" OR "Data Mining" OR "Mining, Data" OR "Text Mining" OR "Mining, Text" OR "Multifactor Dimensionality Reduction" OR "Multifactor Dimensionality Reductions" OR ?Natural Computing?

)

AND

(

“Drug Interactions” OR “Drug Interaction” OR “Interaction, Drug” OR “Interactions, Drug” OR “Previous Indexing” OR “Drug Antagonism” OR “Drug Synergism” OR “Drug Agonism” OR “Drug Partial Agonism” OR “Agonism, Drug Partial” OR “Partial Agonism, Drug” OR “Drug Agonism, Partial” OR “Agonism, Partial Drug” OR “Partial Drug Agonism” OR “Drug Antagonism” OR “Antagonism, Drug” OR “Antagonisms, Drug” OR “Drug Antagonisms” OR “Drug Inverse Agonism” OR “Agonism, Drug Inverse” OR “Inverse Agonism, Drug” OR “Drug Synergism” OR “Drug Synergisms” OR “Synergism, Drug” OR “Synergisms, Drug” OR “Drug Potentiation” OR “Drug Potentiations” OR “Potentiation, Drug” OR “Potentiations, Drug”

)

## C.2 Embase

“drug interaction”/exp OR “drug interactions” OR “interaction, drug”

AND

“artificial intelligence”/exp OR “Artificial Intelligence” OR “Intelligence Artificial” OR “Computer Reasoning” OR “Reasoning Computer” OR “Machine Intelligence” OR “Intelligence Machine” OR “AI Artificial Intelligence” OR “AIs Artificial Intelligence” OR “Machine Learning” OR “Learning Machine” OR “Knowledge Representation Computer” OR “Knowledge Representations Computer” OR “Representation Knowledge Computer” OR “Representations Knowledge Computer” OR “Computer Vision Systems” OR “Computer Vision System” OR “System Computer Vision” OR “Systems Computer Vision” OR “Vision System Computer” OR “Vision Systems Computer” OR “Knowledge Acquisition Computer” OR “Acquisition Knowledge Computer” OR “Acquisitions Knowledge Computer” OR “Knowledge Acquisitions Computer” OR “Expert Systems” OR “Expert System” OR “System Expert” OR “Systems Expert” OR “Fuzzy Logic” OR “Logic, Fuzzy” OR “Knowledge Base” OR “Neural Networks ComputER” OR “Network Neural Computer” OR “Model Neural Network” OR Perceptron OR “Connectionist Model” OR “Robotic” OR “Support Vector Machine” OR “Data Mining” OR “Text Mining” OR “Multifactor Dimensionality Reduction”

## C.3 Lilacs

(

mh:“Interações de medicamentos” OR “Drug Antagonism” OR “Drug Synergism” OR “Interacciones de Drogas” OR “toxicidade de drogas” OR mh:G07.690.812.240$ OR mh:G07.700.680.240$

)

AND

(

mh:“Inteligência Artificial” OR mh:L01.224.065$ OR mh:L01.725.500$ OR mh:G17.485$ OR mh:L01.224.065.605$ OR mh:“Mineração de Dados” OR “Minería de Datos” OR “Data Mining” OR mh:L01.470.625$ OR mh:L01.700.508.208.199$

)

# Apêndice D

# Atributos coletados

Os atributos coletados de fontes farmacológicas estão descritos abaixo.

## D.1 Atributos DrugBank em formato numérico

dbk experimental properties Caco2permeability, dbk experimental properties hydrogenacceptorcount, dbk experimental properties hydrogendonorcount, dbk experimental properties hydrophobicity, dbk experimental properties isoelectricpoint, dbk experimental properties logP, dbk experimental properties logS, dbk experimental properties meltingpoint, dbk experimental properties physiologicalcharge, dbk experimental properties pKaStrongestAcidic, dbk experimental properties pKaStrongestBasic, dbk experimental properties polarizability, dbk experimental properties polarsurfacearea, dbk experimental properties refractivity, dbk experimental properties rotatablebondcount, dbk prices, Weight average, Weight monoisotopic.

## D.2 Variável KEGG em formato numérico

drug mol weight.

## D.3 Atributos ATC em formato texto

atc1, atc1 name, atc2, atc2 name, atc3, atc3 name, atc4, atc4 name, atc5, atc5 name.

## D.4 Atributos DrugBank em formato texto

dbk absorption, dbk affected organisms, dbk ahfs codes, dbk atc codes, dbk brand mixtures, dbk brand names, dbk carriers, dbk categories, dbk chemical formula, dbk classes, dbk description, dbk dosage forms, dbk drug interactions, dbk enzymes, dbk experimental properties, dbk experimental properties boilingpoint, dbk experimental properties Caco2permeability, dbk

experimental properties hydrogenacceptorcount, dbk experimental properties hydrogendonorcount, dbk experimental properties hydrophobicity, dbk experimental properties isoelectricpoint, dbk experimental properties logP, dbk experimental properties logS, dbk experimental properties meltingpoint, dbk experimental properties physiologicalcharge, dbk experimental properties pKa, dbk experimental properties pKaStrongestAcidic, dbk experimental properties pKaStrongestBasic, dbk experimental properties polarizability, dbk experimental properties polarsurfacearea, dbk experimental properties refractivity, dbk experimental properties rotatablebondcount, dbk experimental properties watersolubility, dbk external links, dbk food interactions, dbk groups, dbk half life, dbk inchi, dbk inchi key, dbk indication, dbk iupac name, dbk kingdom, dbk manufacturers, dbk mechanism of action, dbk metabolism, dbk metabolism2, dbk name, dbk packagers, dbk patents, dbk pdb entries, dbk pharmacodynamics, dbk prices, dbk protein binding, dbk route of elimination, dbk safe associations, dbk safe associations cleaned, dbk smiles, dbk state, dbk substructures, dbk synonyms, dbk type, dbk volume of distribution, drugcardClearance, drugcardDrug Interactions, drugcardGeneral Reference, drugcardPathways, drugcardSynonyms, drugcardSynthesis Reference, drugcardTargets actions, drugcardTargets description, drugcardTargets gene, drugcardTargets name, drugcardTargets organism class, drugcardTargets pharmacological action, drugcardTargets uniprot, drugcardVolume of distribution, drugcardWeight average, drugcardWeight monoisotopic.

## D.5 Atributos ENZYME em formato texto

ec1, ec1 name, ec2, ec2 name, ec3, ec3 name, ec4, ec4 name.

## D.6 Atributos EXPASY em formato texto

expasy general comments, expasy name accepted, expasy name alternative, expasy reaction, expasy uniprot.

## D.7 Atributos KEGG em formato texto

kegg disease carcinogen, kegg disease category, kegg disease comment, kegg disease description, kegg disease drug, kegg disease drug entrie, kegg disease entry, kegg disease gene, kegg disease gene hsa, kegg disease gene ko, kegg disease icd10, kegg disease marker, kegg disease marker hsa, kegg disease medlineplus, kegg disease name, kegg disease omim, kegg disease pathway, kegg drug activity, kegg drug brite, kegg drug comment, kegg drug disease, kegg drug entry, kegg drug exact mass, kegg drug formula, kegg drug mol weight, kegg drug name, kegg drug other dbs cas, kegg drug other dbs drugbank, kegg drug other dbs nikkaji, kegg drug other dbs pubchem, kegg drug remark atc, kegg drug remark same as, kegg drug remark therapeutic category, kegg drug structure map, kegg drug target, kegg drug target hsa, kegg drug target

ko, kegg orthology brite, kegg orthology definition, kegg orthology definition ec, kegg orthology entry, kegg orthology gene, kegg orthology name, kegg orthology pathway, kegg pathway compound, kegg pathway description, kegg pathway description ec, kegg pathway disease, kegg pathway drug, kegg pathway entry, kegg pathway name, kegg pathway orthology, kegg pathway orthology ec.

# Anexo A

# Currículo do autor

Doutorado em bioinformática pela Universidade Federal de Minas Gerais (2013) nas temáticas mineração de dados, revisão sistemática de eficácia de medicamentos, estudo de utilização de medicamentos, interações medicamentosas e judicialização da saúde. Graduado em farmácia com habilitação em fármacos e medicamentos pela Universidade Federal de Alfenas (2008). Técnico em informática com ênfase em programação pelo CEFET-SP Uned Cubatão (2002). Durante a graduação realizou iniciação científica nas áreas de química analítica, extratos vegetais e atenção farmacêutica. Experiência em indústria farmacêutica nos setores de pesquisa e desenvolvimento, controle e garantia da qualidade (2008-2012).

## A.1 Formação acadêmica/titulação

**1999-2000** Curso técnico/profissionalizante em informática com ênfase em programação. Centro Federal de Educação Tecnológica (SP) Uned Cubatão.

**2003-2008** Graduação em Farmácia com habilitação em fármacos e medicamentos (indústria farmacêutica). Universidade Federal de Alfenas.

**2009-2013** Doutorado em Bioinformática (Conceito CAPES 6). Universidade Federal de Minas Gerais, UFMG, Brasil.

## A.2 Contribuições

- Brandao et al. [2013]
- Pinto et al. [2013]
- Campos Neto et al. [2012]
- Pinto et al. [2012]

## A.3 Prêmio

**2010** I Prêmio Estadual de Assistência Farmacêutica Aluísio Pimenta , Categoria Assistência Farmacêutica no âmbito do SUS - Sistemas de Gerenciamento de Dados - primeiro lugar, Secretaria de Estado de Saúde de Minas Gerais - SES/MG.

## A.4 Programas de computador sem registro

**2012** Ferré, Felipe ; SALES, M. H. ; NEVES, T. H. ; Acurcio, F. A. . Revis - Sistema de Revisão Sistemática.

**2011** Ferré, Felipe ; Silva, L ; Machado, MAA . Sistema Integrado de Gerenciamento da Assistência Farmacêutica de Minas Gerais - Módulo Cuidado Farmacêutico (SiGAF-MG).

**2009** Ferré, Felipe ; Marques, L. A. M. ; Miguel, E . FarClinic - Sistema para farmácia clínica.

**2008** Ferré, Felipe. BPFtotal. 2008.

## A.5 Contato

**Email** ferrebioinfo@gmail.com

**Sítio** http://dcc.ufmg.br/~ferre. Nesta página serão disponibilizadas revisões do texto, código-fonte, previsões e outras publicações.

# Anexo B

# Fontes de dados

Além das bases em saúde, uma fonte comum para verificação de interações medicamentosas é construida a partir de bancos de dados comerciais, como a Thomson Micromedex ou a DrugBank, disponibilizada gratuitamente[Wong et al., 2010; Tari et al., 2010].

São destacados três grupos de fontes de dados. O primeiro advém de repositórios públicos contendo bases secundárias, utilizadas para construção de bases terciárias de interações medicamentosas, redes metabólicas, informações sobre as enzimas, entre outras. A segunda fonte advém de informações clínicas relacionadas ao uso de medicamentos, coletadas de prontuários ou formulários ou bases administrativas, as quais são inseridas em banco de dados normalizados ou não para posteriormente serem empregados para a caracterização das interações medicamentosas potenciais verificadas a partir dos modelos desenvolvidos. O terceiro grupo é constituído de listas de referência empregadas para classificação e padronização da nomenclatura e formação de redes ontológicas.

## B.1 Repositórios públicos de dados

### B.1.1 BRENDA

Iniciada em 1987 pelo German National Research GBF (atual Helmholtz Centre for Infection Research), BRENDA (BRaunschweig ENzyme DAtabase) é um dos principais repositórios de dados sobre enzimas manualmente anotadas. Os dados são primariamente obtidos da literatura e incluem classificação e nomenclatura, reação e especificidade, parâmetros funcionais, ocorrência, estrutura enzimática, aplicação, informações sobre mutações, estabilidade, doenças, preparação e isolamento. Desde 2007, BRENDA vem sendo mantida pela Technische Universitat Braunschweig, Institute for Bioinformatics e Systems Biology [Scheer et al., 2011].

### B.1.2 DIO

Drug Interaction Ontology (DIO) foi desenvovido como representação formal do conhecimento farmacológico. Disponibiliza um modelo para acúmulo reutilizável de conhecimento sobre os componentes moleculares farmacológicos. Sua ontologia foi empregada para implementar um modelo relacional o qual inclui representação simbólica das possíveis interações fármaco-biomolécula. Esta modelagem permite a realização de consultas que distinguam as interações medicamentosas dentre as demais interações moleculares [Yoshikawa et al., 2004].

### B.1.3 DrugBank

A base de dados disponibilizada pelo DrugBank combina detalhadas informações químicas e farmacológicas sobre as substâncias, incluindo alvos terapêuticos (sequencia genômica, estrutura, rota metabólica). Possui entrada para 6.796 fármacos descritos por até 150 campos, incluindo absorção, biotransformação, peso molecular, indiciação de uso, entre outros. Esta base é coordenada por David Wishart, do Departamento de Ciências da Computação e Ciências Biológicas da Universidade de Alberta [Wishart et al., 2008].

### B.1.4 Drugs.com

O sitio drugs.com disponibiliza de forma independente informações sobre 24 mil medicamentos de venda livre ou de prescrição e produtos naturais. As informações incluem dados sobre interações entre pares de medicamentos as quais são classificadas conforme a gravidade (alta, moderada, leve). Como fontes de dados primárias do drug.com são empregadas Micromedex (atualizado em 5/6/2011), Cerner Multum (atualizado em 21/06/2011), Wolters Kluwer (atualizado em 1/06/2011) entre outras [DRUGS.COM, 2011].

### B.1.5 Gene Ontology

O projeto Gene Ontology disponibiliza ontologia sobre componentes celulares, partes de células ou ambiente extracelular, função molecular, processos biológicos, operações ou conjuntos de eventos moleculares pertinentes ao funcionamento ou integração de unidades como células, tecidos, órgãos ou organismos. Por exemplo, o citocromo c pode ser descrito pela função molecular com o termo *atividade de oxiredutase*, processo biológico “fosforilação oxidativa”e “indução de morte celular”e como componente celular pelos termos “matriz mitocondrial”e “mitocôndria interna de membrana”. O GO é estruturado como um grafo acíclico direcionado e cada termo é definido com relacionamentos de um ou mais termos no mesmo domínio ou vários outros domínios [Ashburner et al., 2000].

### B.1.6 Kegg

KEGG é uma base de dados iniciada em 1995, originalmente como parte do programa Genoma Humano japones. Desde então, vem disponibilizando informações sobre funções celulares em formas computávies, especialmente em redes moleculares (KEGG pathay maps) e listas hieráquicas (BRITE functional hierarchies) com recente foco em fármacos e doenças humanas (KEGG medicus). KEEG é mantido por pesquisadores da Universidade de Kioto, Japão [Kanehisa et al., 2010].

### B.1.7 MetaCyc

MetaCyc é uma banco de dados confiável, não redundante sobre o metabolismo de pequenas moléculas. Contém rotas metabólicas, reações e dados sobre enzimas exclusivamente demonstrados via experimental. MetaCyc é usado como referência para composição de componentes patológicos em ferramentas de previsão computacional de redes metabólicas disponibilizadas na base de dados de rotas e genoma (PGDB - Pathway/Genome Database ) [Caspi et al., 2008].

### B.1.8 Patika

A base de dados PATIKA integra dados de diversas fontes, incluindo UniProt, PubChem, GO, IntAct, HPRD e Reactome. A bases de dados é focada somente em rotas metabólicas humanas, disponibilizando grande diversidade de estados de diferentes entidades biológicas na ordem de milhares de reações.

### B.1.9 PubChem

PubChem é uma base de dados desenvolvida pela NCBI (National Center for Biotechnology Information) para disponibilizar acesso a comunidade científica às mais atualizadas e abrangentes fontes de estruturas químicas de pequenas moléculas orgânicas e sua atividade biológica. Abriga informações sobre compostos advindas de fontes confiáveis da literatura bem como de dados de programas de respositórios moleculares.

### B.1.10 SBML

É um formato livre e aberto para intercâmbio de modelos computacionais de processos biológicos. É particularmente útil para modelos de metabolismo, e sinalização celular. Tem sido desenvolvido internationalmente pela comunidade científica desde 2000.[Hucka et al., 2003]

## B.2 Listas de referência

### B.2.1 ATC

A classificação anatômica terapêutica química (ATC) e o sistema de dose diária definida (DDD) são padrões recomendados pela Organização Mundial da Saúde para estudos de utilização de medicamentos. O sistema é difundido internacionalmente com número crescente de usuários [WHO, 2011]. Cada medicamento é classificado conforme cinco níveis, ilustrado na tabela B.1

Produtos medicinais são classificados de acordo com o uso terapêutico e ativo principais, sob o princípio básico de apenas um código ATC para cada rota de administração (isto é, formas farmacêuticas com ingredientes similares e potência terão o mesmo código ATC)[WHO, 2011].

O código ATC é degenerado, visto que 15,6% das substâncias químicas possuem mais de uma classificação. Por exemplo, a betametasona possui onze diferentes códigos ATC de acordo com a utilização terapêutica.

Tabela B.1: Exemplo da classificação ATC. Descrição do fármaco metformina conforme a os cinco níveis da classificação anatômica terapêutica química (ATC) da OMS

| nível | descrição | código | descrição |
|---|---|---|---|
| 1 | grupo anatômico principal | A | trato alimentar e metabolismo |
| 2 | subgrupo terapêutico | A10 | fármacos usados na diabetes |
| 3 | subgrupo farmacológico | A10B | redutores de glicose sanguínea, excl. insulinas |
| 4 | subgrupo químico | A10BA | biguaninas |
| 5 | substância química | A10BA02 | metformina |

### B.2.2 RENAME

O Ministério da Saúde estabeleceu mecanismos que permitem a contínua atualização da Relação Nacional de Medicamentos Essenciais - Rename, sua implementação e ampla divulgação. Adotada em nível nacional, a Rename serve de instrumento básico para a elaboração das listas estaduais e municipais segundo sua situação epidemiológica, para a orientação da prescrição médica, para o direcionamento da produção farmacêutica e para o desenvolvimento científico e tecnológico.As Políticas de Medicamentos e de Assistência Farmacêutica estabelecem a atualização e a implementação da Rename como instrumento racionalizador das ações no âmbito da assistência farmacêutica e medida indispensável para o uso racional de medicamentos no contexto do SUS. A seleção dos medicamentos da Rename baseia-se nas prioridades nacionais de saúde, bem como na segurança, na eficácia terapêutica comprovada, na qualidade e na disponibilidade dos produtos [BRASIL, 2010b].

### B.2.3 CID-10

A CID-10 foi conceituada para padronizar e catalogar as doenças e problemas relacionados à saúde, tendo como referência a Nomenclatura Internacional de Doenças, estabelecida pela Organização Mundial de Saúde. Com base no compromisso assumido pelo Governo Brasileiro, a organização dos arquivos em meio magnético e sua implementação para disseminação eletrônica foi efetuada pelo DATASUS, possibilitando, assim, a implantação em todo o território nacional, nos registros de Morbidade Hospitalar e Ambulatorial, compatibilizando estes registros entre todos os sistemas que lidam com morbidade [BRASIL, 2011].

# Anexo C

# Métricas de distância contidas no algoritmo implementado

Medidas $\Delta$ de diferença ou similaridade implementadas na presente versão do modelo.

**Distância euclidiana**

$$d(X,Y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \tag{C.1}$$

**Distância de cosseno**

$$cos(X,Y) = \frac{x \cdot y}{\|x\| \, \|y\|} \tag{C.2}$$

O $\cdot$ indica o produto interno do vetor $x \cdot y = \sum_{k=1}^{n} x_k y_k$, e $\|x\|$ é o comprimento do vetor $\|x\| = \sqrt{\sum_{k=1}^{n} x_k^2} = \sqrt{x \cdot x}$.

**Minkowski**

$$d(x,y) = \left( \sum_{k=1}^{n} |x_k - y_k|^r \right)^{\frac{1}{r}} \tag{C.3}$$

**Chebychev**

$$d(x,y) = \max_{i=1}^{l} |x_i - y_i| \tag{C.4}$$

**Manhattan**

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i| \tag{C.5}$$

**Camberra**

$$d(x,y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i + y_i|} \tag{C.6}$$

### Jaccard estendido

$$d(x,y) = \frac{x \cdot y}{\|x\|^2 + \|y\|^2 - x \cdot y} \tag{C.7}$$

### Correlação

$$\text{correlação}(x,y) = \frac{\text{covariância}(x,y)}{\text{desvio padrão}(x) \times \text{desvio padrão}(y)} = \frac{s_{xy}}{s_x \times s_y} \tag{C.8}$$

Sendo:

$$\text{média}(a) = \bar{a} = \frac{\sum_{k=1}^{n} a_k}{n} \tag{C.9}$$

$$\text{desvio padrão}(a) = s_a = \sqrt{\frac{\sum_{k=1}^{n} (a_k - \bar{a})}{n-1}} \tag{C.10}$$

$$\text{covariância}(x,y) = s_{xy} = \frac{\sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})}{n-1} \tag{C.11}$$

# Anexo D

# Código-fonte

Implementação em R[1] do SHAMAM[2], análise semântico-heurístico para mineração de associações medicamentosas.

São mostrados algoritmos para geração das amplitudes de associações a partir das entidades descritas, classificação, performance e coleta de previsões.

## D.1 Funções primárias ou distais

Funções que não chamam outras funções SHAMAM, mas são chamadas[3].

Funções R de mineração em texto a partir de tabela mysql de entidades com descrição na estrutura "id|código|nome|...|$n$|".

### D.1.1 Dependências

```
1 library("DBI");
2 library("SDMTools");
3 library("RMySQL");
4 library("multicore");
5 library("tm");
6 library("R.utils")
7 library("tcltk")
```

Mantem o numero de decimais sobre controle devido a restrições do weka

### D.1.2 clean.matrix

```
1 clean.matrix=function(x,na.replace=0,inf.replace=1,decimals=12){
2   try(x[is.na(x)] <-na.replace);   try(x[is.nan(x)] <-na.replace);
```

[1] R versão 3.0.1 (16/05/16) *"Good Sport" Copyright (C) 2013 The R Foundation for Statistical Computing Platform*: x86-64-pc-linux-gnu (64-bit)

[2] *Semantic-Heuristic Analysis for Mining Association of Medicines*

[3] Arquivo shamam_funcoes13091501_distal.r

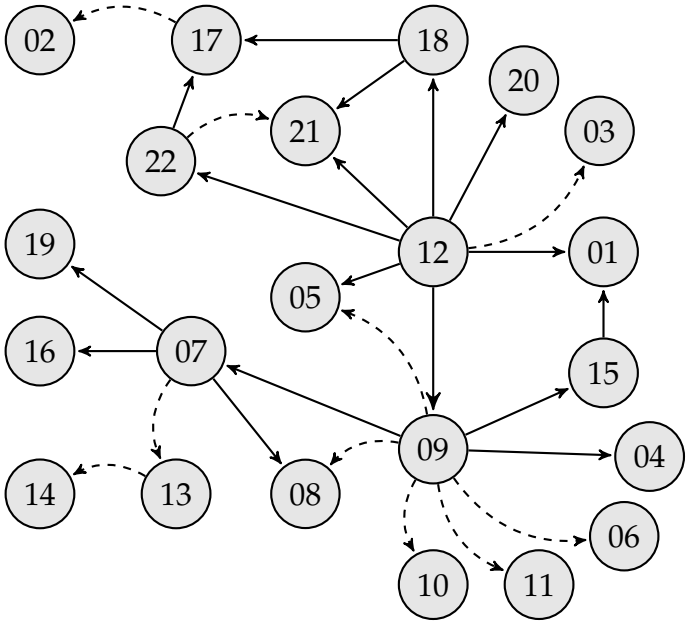



Figura D.1: **Funções implementadas em R para classificação geral.** Algoritmos: 01 clean.matrix, 02 csv2arff, 03 feature.clustering, 04 get.matrix.distances, 05 mysql.classification, 06 mysql.connection, 07 mysql.desc2matrix, 08 mysql.descriptor, 09 mysql.distances, 10 mysql.numeric.fields, 11 mysql.text.fields, 12 shamam, 13 split.desc, 14 split.str, 15 svd.filter, 16 tm.corpus2matrix, 17 weka.classification, 18 weka.classification.optimization, 19 weka.desc2matrix, 20 weka.feature.selection, 21 weka.performance, 22 weka.train.storming. Setas indicam chamadas, p.ex., 09 $\rightarrow$ 10, significa 09 chama 10. Linhas tracejadas representam chamadas opcionais.

```
  try(x[x==Inf]   <-inf.replace);  try(x[x==-Inf]   <-inf.replace);
  return(x)
}
```

### D.1.3 csv2arff

```
csv2arff=function(data=NULL,wekajar,csv=NULL,arff=NULL,train=FALSE){
  if (missing(wekajar)) stop("SHAMAM: Insert weka's java file directory.")
  if (is.null(arff))
    arff=paste("/tmp/shamam",gsub("[^[:alnum:]]","",Sys.time()),
               ".arff",sep="")
  if(is.null(data) & is.null(csv))
    stop("SHAMAM: Insert data or csv file name.")
  if (!is.null(data)){
    write.csv(data,file=paste(arff,".tmp",sep=""))
    csv=paste(arff,".tmp",sep="")
  }
  try(system(paste("java -Xmx50g -cp", wekajar,
                   "weka.core.converters.CSVLoader",
                   csv,">", arff, sep=" ")))
  if (train){
    try(system(paste("grep -v \'?$\' ",arff," > ",arff,
                     ".train.arff", sep="")))
    arff=paste(arff,".train.arff", sep="")
```

```
19    }
20    if (!is.null(data)) system(paste("rm ", arff,".tmp",sep=""))
21    else arff=NULL
22    return(arff)
23  }
```

### D.1.4 feature.clustering

```
1   feature.clustering=function(data,range=c(1,10),technique="kmeans",clusters=4){
2   #Divide cada atributo em n=clusters agrupamentos
3     if (missing(data)) stop("SHAMAM: Insert a matrix.")
4     if (!is.matrix(data)) stop("SHAMAM: Data class must be \"matrix\".")
5     col=ncol(data); if (range[2]>col) range[2]=col;
6     for(i in range[1]:range[2])
7       if (!is.na(sum(as.numeric(data[,i]))))
8         if (technique=="kmeans")
9           if(clusters>length(levels(as.factor(data[,i])))){
10            r=try(as.vector(kmeans(as.numeric(data[,i]),clusters)$cluster))
11            if (length(r)==length(data[,i])) data[,i]=r
12          }
13    return(data)
14  }
```

### D.1.5 get.matrix.distances

```
1   get.matrix.distances=function(f,var,my.con=NULL,a_c,full=FALSE,
2                                 distance=list("euclidean","cosine2")){
3     aux=list("euclidean","cosine","cosine2","jaccard","pearson","spearman",
4              "manhattam","camberra","minkowsky","chebychev")
5     #---------testa argumentos-----------------------------------------
6     if (missing(f)) stop("SHAMAM: Insert a matrix.")
7     if (missing(var)) stop("SHAMAM: Insert a variable or table name.")
8     d=aux%in%distance;
9     sum_d=sum(d)
10    if (!sum_d) stop("SHAMAM: Insert a valid distance metric.")
11    #Calcula a distancia entre todos os elementos em a_c
12    f=as.matrix(f);  f.nc=ncol(f); f.nr=nrow(f); if(is.null(f.nc)) f.nc=0;
13    if(f.nc>0){
14      names(d)=aux;
15      c=matrix(numeric(0), f.nr,2)
16      for (i in 1:f.nr){#Calculos preliminares
17        c[i,1]=crossprod(f[i,])
18        c[i,2]=var(f[i,])
19      }
20      #gera o conjunto completo de possibilidades se a_c nao for fornecido
21      if (is.null(a_c)){a_c.k=a_c; full=TRUE;} else a_c.k=NULL;
22      if (full){
23        #gera matriz vazia para conter a combinacao dos resultados
24        a_c=matrix(numeric(0),(choose(f.nr,2)),3)
25        names(a_c)=c("idA","idB","class")
26        i=0;
27        for (j in 1:(f.nr-1)){
```

```
28          for (k in (j+1):f.nr){
29            i=i+1; a_c[i,1]=j; a_c[i,2]=k; a_c[i,3]="?";
30            aux=a_c.k[which(a_c.k[,1]==j&a_c.k[,2]==k),3]
31            if (length(aux)==1) a_c[i,3]=aux
32          }
33        }
34      }
35      r=matrix(numeric(0), nrow(a_c),sum_d+3)
36      #avalia as distancias
37      i=0
38      for (ii in 1:nrow(a_c)){
39        j=as.numeric(a_c[ii,1]); k=as.numeric(a_c[ii,2]);
40        l=2;i=i+1;
41        aa=f[j,]; bb=f[k,]
42        a1=(aa - bb); a2=crossprod(aa, bb)
43        r[i,1]=j; r[i,2]=k;
44        if (d["euclidean"]) {l=l+1;r[i,l]=sqrt(sum(a1 ^ 2));}
45        if (d["cosine"])    {l=l+1;r[i,l]=a2/sqrt(c[j,1] * c[k,1])}
46        if (d["cosine2"])   {l=l+1;
47                             r[i,l]=(-log(1+(a2/sqrt(c[j,1] * c[k,1]))/2))^2}
48        if (d["jaccard"])   {l=l+1;r[i,l]=a2/(c[j,1]^2 + c[k,1]^2 - a2)}
49        if (d["pearson"])   {l=l+1;
50                             r[i,l]=cov(aa, bb,method=c("pearson"))/
51                             (c[j,2]*c[k,2])}
52  #       if (d["kendall"])   {l=l+1;
53  #                            r[i,l]=cov(aa, bb,method=c("kendall"))/
54  #                            (c[j,2]*c[k,2])}
55        if (d["spearman"])  {l=l+1;
56                             r[i,l]=cov(aa, bb,method=c("spearman"))/
57                             (c[j,2]*c[k,2])}
58        if (d["manhattam"]) {l=l+1;r[i,l]=sum(a1)}
59        if (d["camberra"])  {l=l+1;r[i,l]=sum(a1)/sum(aa+bb)}
60        if (d["minkowsky"]) {l=l+1;r[i,l]=(sum(a1 ^ 3))^1/3}
61        if (d["chebychev"]) {l=l+1;r[i,l]=max(a1)}
62      }#metrics  -------------------
63      r=as.matrix(r)
64      try(r[r=="NaN"] <-"?"); try(r[r=="NA"] <-"?");
65      try(r[is.na(r)] <-"?"); try(r[is.nan(r)] <-"?");
66      try(r[r==Inf] <- "?"); try(r[r==-Inf] <- "?")
67      #seta a classe na ultima coluna
68      for (i in 1:nrow(a_c)) r[i,sum_d+3]=a_c[i,3];
69      colnames(r)=c(paste(var,"_id1", sep=""),paste(var,"_id2",sep=""),
70                    paste(var,"_", names(c(which(d))),sep=""),"class")
71    }else {r=0}
72    return(r)
73  }
```

## D.1.6 mysql.classification

```
1  mysql.classification=function(my.con,my.tbl.ass,ida,idb,class,
2                                 my.tbl=NULL,id="id"){
3    if(!is.null(my.tbl)){
4      dbGetQuery(my.con, "DROP TABLE IF EXISTS tmp")
5      query=paste("CREATE TABLE tmp ENGINE=MyISAM",
```

```
                "SELECT @curRow := @curRow +1 AS row_number, A.*",
                "FROM ",my.tbl,"A JOIN (SELECT @curRow :=0)r")
    dbGetQuery(my.con, query)
    query=paste("SELECT B.row_number as idA, C.row_number as idB,",class,
                "FROM", my.tbl.ass," A, tmp B, tmp C",
                "WHERE A.idA=B.id AND A.idB=C.id AND",class,"<>\"?\" AND",
                class,"IS NOT NULL ORDER BY B.row_number, C.row_number")
  }
  else
    query=paste("SELECT",ida,",",idb,",", class, "FROM",my.tbl.ass,
                "WHERE",class,"<>\"?\" AND",class,"IS NOT NULL")
  r=dbGetQuery(my.con, query)
  dbGetQuery(my.con, "DROP TABLE IF EXISTS tmp")
  return(r)
}
```

### D.1.7 mysql.connection

```
mysql.connection=function(user,base){
  if(missing(user))user=readline(prompt="SHAMAM: Insert mysql user name")
  if(missing(base))base=readline(prompt="SHAMAM: Insert mysql base name")
  cat("Password: ")
  system("stty -echo")
  pass=readline()
  system("stty echo")
  cat("\n")
  return(dbConnect(MySQL(), user=user, password=pass, dbname=base))
}
```

### D.1.8 mysql.descriptor

```
mysql.descriptor = function(my.con,my.tbl,variable){
  if (missing(my.con) |missing(my.tbl) | missing(variable))
    stop("SHAMAM: Missing parameters on description collection.")
  #Obtem a descricao das entidades da variavel correspondente
  query=paste("SELECT ",variable, " as varx FROM", my.tbl)
  #Obtem a descricao das entidades da variavel correspondente
  o.my=dbGetQuery(my.con,query)
  return(o.my$var)
}
```

### D.1.9 mysql.numeric.fields

```
mysql.numeric.fields = function(my.con,my.tbl){
#Retorna a lista de campos em formato texto
  g=dbGetQuery(my.con, paste("DESC", my.tbl));
  nvars=nrow(g)
  var <- data.frame(var = rep(0));i=0;
  for (ii in 1:nvars){
    var.type=gsub("[^[:alpha:] ]", "",g[ii,2]);
```

```
8     if(var.type=="float" | var.type=="bigint" | var.type=="int" |
9        var.type=="double"){
10      i=i+1;
11       var=rbind(var,g[ii,1])
12    }
13  }
14  return(var[-1,])
15 }
```

### D.1.10 mysql.text.fields

```
1  mysql.text.fields = function(my.con,my.tbl){
2  #Retorna a lista de campos em formato texto
3    g=dbGetQuery(my.con, paste("DESC", my.tbl));
4    nvars=nrow(g)
5    var <- data.frame(var = rep(0));i=0;
6    for (ii in 1:nvars){
7      var.type=gsub("[^[:alpha:] ]", "",g[ii,2]);
8      if(var.type=="text" | var.type=="varchar" | var.type=="char" |
9        var.type=="enumsimnao"){
10       i=i+1;
11        var=rbind(var,g[ii,1])
12    }
13  }
14  return(var[-1,])
15 }
```

### D.1.11 split.str

```
1  split.str <- function(x, n) {
2    sst <- strsplit(x, '')[[1]]
3    m <- matrix('', nrow=n, ncol=(length(sst)+n-1)%/%n)
4    m[seq_along(sst)] <- sst
5    apply(m, 2, paste, collapse='')
6  }
```

### D.1.12 tm.corpus2matrix

```
1  tm.corpus2matrix = function(corpus,p.threshold=0,tfidf=TRUE){
2    if (p.threshold <0 | p.threshold >1)
3      stop("The threshold must be between 0 and 1. Put 0.02 for 2%.")
4    #coleta o dicionario e avalia a frequencia dos termos
5    d=Dictionary(DocumentTermMatrix(corpus))
6    x=(DocumentTermMatrix(corpus, list(dictionary = d)))
7    #normaliza  term frequency?inverse document frequency
8    if(tfidf) x=try(weightTfIdf(x, normalize = TRUE))
9    #Procede se houver palavras frequentes
10   if(length(x)){
11     x.nc=ncol(x)
12     #obtem a matriz de frequencia original
```

```
13      f=as.matrix(x)
14      #Remove as palavras 2% menos frequentes
15      ff = try(f[,colSums(f) > (p.threshold*x.nc)])
16      if (length(ff)>(10*nrow(f)))
17        f=ff#Usa a poda se restarem mais que 10 colunas
18      else print("SHAMAM Warning: Huge pruning threshold. Choosing original
19  matrix.")
20      return(f)
21    }else return(NULL)
22  }
```

### D.1.13 tm.get.corpus

```
1  tm.get.corpus = function(textmatrix,stemming=TRUE,stopwords=TRUE){
2    #importa somente alfa-numericos para o formato da cran library tm
3    o.df=DataframeSource(as.data.frame(gsub("[^[:alnum:] ]","",textmatrix)))
4    #Gera um corpus volatil e faz transformacoes------------------------
5    z.tm=Corpus(o.df)
6    z.tm=tm_map(z.tm, stripWhitespace)#Remove espacos brancos extras
7    z.tm=tm_map(z.tm, tolower)#Reduz para minusculas
8    z.tm=tm_map(z.tm, removePunctuation)#Remove pontuação
9    #Remove palavras comuns
10   if (stemming)  z.tm=tm_map(z.tm, removeWords, stopwords('english'))
11   if (stopwords) z.tm=tm_map(z.tm, stemDocument)#Reduz ao tronco ling.
12   z.tm=tm_map(z.tm, stripWhitespace)#Remove espaco branco extra
13   return(z.tm)
14 }
```

### D.1.14 weka.desc2matrix

```
1  weka.desc2matrix=function(wekajar, my.tbl.desc,timeout=300){
2      tmp=paste("/tmp/shamam",gsub("[^[:alnum:]]","",Sys.time()),"",sep="")
3      if(length(my.tbl.desc)==0) stop("SHAMAM: Insert a valid descriptor.")
4      write.csv(my.tbl.desc,file=paste(tmp,".csv",sep=""))
5      r=NULL
6      if(file.exists(paste(tmp,".csv",sep="")))
7        try(system(paste("timeout ",timeout," java -Xmx50g -cp ", wekajar,
8                " weka.filters.unsupervised.attribute.NominalToString -i ",
9                tmp,".csv -o ", tmp, ".arff", sep="")))
10
11     if(file.exists(paste(tmp,".arff",sep="")))
12     try(system(paste("timeout ",timeout," java -Xmx50g -cp ", wekajar,
13             " weka.filters.unsupervised.attribute.StringToWordVector -i ",
14             tmp,".arff -o ", tmp, "2.arff", sep="")))
15
16     if(file.exists(paste(tmp,"2.arff",sep="")))
17     try(system(paste("timeout ",timeout," java -Xmx50g -cp ", wekajar,
18             " weka.core.converters.CSVSaver -i ",
19             tmp,"2.arff -o ", tmp, "2.csv", sep="")))
20
21     if(file.exists(paste(tmp,"2.csv",sep="")))
22       r=try(read.csv(paste(tmp,"2.csv",sep=""),sep=","))
```

```
23
24      if (length(r)>1){
25        r=as.matrix(r); r=subset(r,select=-1)#remove primeira coluna
26      }else{
27        print(paste("SHAMAM: Weka error or step timeout",
28                    timeout,"seconds exceeded."))
29        r=NULL
30      }
31      return(r)
32  }
```

### D.1.15 weka.feature.selection

```
1   weka.feature.selection=function(data=NULL, wekajar, method, featselec,
2                                   return.data=TRUE,timeout=300,infile=NULL,
3                                   resample=FALSE,train.filter=FALSE){
4     m=paste("weka.attributeSelection",
5       c("ChiSquaredAttributeEval","FilteredAttributeEval",
6         "GainRatioAttributeEval","InfoGainAttributeEval",
7         "LatentSemanticAnalysis","SymmetricalUncertAttributeEval"),sep=".")
8     if (missing(wekajar)) stop("SHAMAM: Insert weka's java file directory.")
9     if (missing(method)){
10      method=1; print("SHAMAM Warning: Selecting intersection method.")
11    }else if(is.numeric(method)){
12        if(method<=0 | method>1)
13        stop("SHAMAM: Method must be a value in ]0,1] interval.")
14      }else{
15        if(method!="default") {
16          method=1; print("SHAMAM Warning: Selecting intersection method.")
17          }
18        }
19    if (missing(infile) & missing(data))
20      stop("SHAMAM: Insert a matrix, arff or csv file.")
21    if (!is.null(data)) infile=csv2arff(data,wekajar,arff=infile,
22                                        train=train.filter)
23    if (is.null(infile))
24      stop("SHAMAM: Arff conversion error. Verify weka files.")
25
26    if(method=="intersection" | is.numeric(method)){
27      if (missing(featselec)){
28        featselec=m
29        print("SHAMAM Warning: Missing ranker feature selection.")
30      }
31      print(paste("SHAMAM: Performing top ",method*100,"%",sep=""))
32      lm=length(m); r <- vector("list", lm); min=1/0# Seta minimo como infinito
33      for(i in 1:lm){#ranqueia as variaveis para cada metodo "m"
34        print(paste("SHAMAM: Performing",m[i],"ranker."))
35        aux=matrix(c(try(system(paste("timeout",timeout,"java -Xmx50g -cp",
36                        wekajar, m[i],"-i", infile,
37                        "| grep -v '^   0' | grep _ |
38                        awk '{print $1\",\"$2\",\"$3}' | grep -v ^0,|
39                        grep -v @| sed 's/,$//g'|
40                        awk -F',' '{print $NF\",\"$1}'"),
41              intern = TRUE))),ncol=2)
```

```
      r[[i]]=do.call(rbind,(strsplit(aux,",")))
      if(min>nrow(aux)) min=nrow(aux)
    }
    #Faz o topXX a partir dos atributos com relevancia maior que zero
    if (is.numeric(method)){
      p=ceiling(method*min); aux=r[[m[1]]][1:p,1]#coleta os top p%
      for(i in 2:lm) aux=intersect(aux,r[[m[i]]][,1])
    }
    r=as.list(aux)
    if (return.data)
      data=data[ , which(names(as.data.frame(data)) %in% rbind(r,"class"))]
      r=csv2arff(data,wekajar,arff=infile,train=train.filter)
  }
  if(method=="default"){
    method="weka.filters.supervised.attribute.AttributeSelection"
    print("SHAMAM: Performing AttributeSelection")
    arff=paste("/tmp/shamam",gsub("[^[:alnum:]]","",Sys.time()),
               ".arff",sep="")
    try(system(paste("timeout",timeout,"java -Xmx50g -cp", wekajar,
                     method,"-i", infile,"-o",arff)))
    try(system(paste("mv",arff,infile)))
    r=infile
  }
  if(resample){
    method="weka.filters.supervised.instance.Resample";
    option="-c last"
    try(system(paste("timeout",timeout,"java -Xmx50g -cp", wekajar,
                     method,"-i", infile,"-o","/tmp/shamam.arff",option)))
    try(system(paste("mv /tmp/shamam.arff",infile)))
    r=infile
  }
  return(r)
}
```

### D.1.16 weka.performance

```
weka.performance=function(weka.res,teste=0,as.vector=FALSE){
  #coleta apenas instancias do teste
  if(teste){f=which(weka.res$actual =="?");  f=weka.res[-f,];}
  else{f=weka.res}
  f$actual=factor(f$actual);           l=levels(f$actual);
  r=matrix(numeric(0),11,length(l));
  for(i in 1:length(l)){
    #Seta cada nivel como 0 ou 1 e avalia a performance
    q=f;
    q$actual=as.character(q$actual);  q$predic=as.character(q$predic);
    q$actual[q$actual==l[i]]="1";     q$actual[q$actual!="1"]="0";
    q$predic[q$predic==l[i]]="1";     q$predic[q$predic!="1"]="0";
    #Ajusta p para tornar-se a medida da predicao
    w=which(q$predic =="0");          q=as.matrix(q);
    1-as.numeric(q[w,3]);             q[w,3]=1-as.numeric(q[w,3]);
    #Obtem a performance
    a=accuracy(q[,1],q[,3]);
    r[1,i]=a$omission.rate;           r[2,i]=a$sensitivity;
```

```
    r[3,i]=a$specificity;              r[4,i]=a$prop.correct;
    r[5,i]=a$Kappa;                    r[6,i]=a$AUC;
    a=confusion.matrix(q[,1],q[,3]);
    r[7,i]=auc(q[,1],q[,3]);           r[8,i]=a[1,1];
    r[9,i]=a[1,2];                     r[10,i]=a[2,1];
    r[11,i]=a[2,2];
  }
  colnames(r)=l
  rownames(r)=c("omission_rate","sensitivity","specificity",
                "prop_correct","Kappa","AUC","AUC2","nn","sp_no",
                "so_np","ss")
  all_mean=rowMeans(r); r=cbind(r,all_mean)
  if (as.vector){
    y=merge(rownames(r),colnames(r))
    w=within(y, C <- paste(x, y, sep='__'))
    r=as.vector(r)
    names(r)=w$C
  }
  return(round(r,4))
}
```

## D.2 Funções secundárias ou mediais

Funções que chamam outras funções SHAMAM e são chamadas [4].

### D.2.1 mysql.desc2matrix

```
mysql.desc2matrix=function(my.con,my.tbl,variable=NULL,svd=FALSE,
                      stemming=TRUE,stopwords=TRUE,p.threshold=0,
                      classes=4,tfidf=TRUE,timeout=300,only.weka=FALSE){
  my.tbl.desc=mysql.descriptor(my.con,my.tbl,variable)
  if (is.numeric(my.tbl.desc)){
    r=kmeans(my.tbl.desc,classes)$cluster
  }
  else{
    if(!only.weka){
      my.tbl.corpus =
      tm.get.corpus(my.tbl.desc,stemming,stopwords)
      r=tm.corpus2matrix(my.tbl.corpus,p.threshold,tfidf=tfidf)
    } else r=NULL
    if (length(r)<1) {
      print(paste("SHAMAM Warning: Using weka to vectorize",variable))
      r=try(weka.desc2matrix(wekajar=wekajar,my.tbl=my.tbl.desc,
                             timeout=timeout))
      if (length(r)<1){
#         try(system("rm /tmp/*.arff /tmp/*.csv"))
        my.tbl.desc=split.desc(my.tbl.desc)#quebra de strings longas
        r=try(weka.desc2matrix(wekajar=wekajar,my.tbl=my.tbl.desc,
                               timeout=timeout))
      }
```

[4] Arquivo shamam_funcoes13091502_medial.r

```
24        if (svd & !is.null(r)) r=svd.filter(r)
25      }
26    }
27    return(r)
28  }
```

## D.2.2 mysql.distances

```
1  mysql.distances=function(my.con,my.tbl,variable=NULL,ignore="id",
2                           stemming=TRUE,stopwords=TRUE,p.threshold=0,
3                           svd.proximal=FALSE, classes=4,
4                           distance=list("euclidean","cosine"),
5                           wekajar=NULL,tfidf=TRUE,svd.distal=FALSE,
6                           full=FALSE,only.weka=FALSE,a_c=NULL,
7                           a_c.ida="idA",a_c.idb="idB",a_c.class=NULL,
8                           id="id"){
9  print("flag0")
10   if(svd.distal & length(distance)>1)
11     stop("SHAMAM: You can not choose svd.distal and more than one distance
12 metric.")
13   if(is.null(variable)){
14     strfield=mysql.text.fields(my.con,my.tbl)
15     numfield=mysql.numeric.fields(my.con,my.tbl)
16     variable=rbind(as.matrix(strfield),as.matrix(numfield))
17   }
18   variable=as.list(setdiff(variable,ignore)) #remove variaveis ignoradas
19   if (missing(my.con)) my.con=mysql.connection()
20   #gera arquivo de saida
21   if (is.null(a_c)) full==TRUE
22   else{print("flag1")
23     if(is.character(a_c))
24       a_c=try(mysql.classification(my.con=my.con, my.tbl.ass=a_c,
25                                    ida=a_c.ida, idb=a_c.idb, id=id,
26                                    class=a_c.class,my.tbl=my.tbl))
27
28   }
29 print("flag1b")
30   if (full)
31     r=matrix(numeric(0),
32              choose(length(
33                    mysql.descriptor(my.con,my.tbl,variable[1])),2),1)
34   else r=matrix(numeric(0), nrow(a_c),1)
35
36 print("flag2")
37
38   #obtem a matriz binaria para cada variavel a calcula as distancias
39   my.tbl.bin=NULL;
40   for (i in 1:length(variable)){
41     my.tbl.bin=try(mysql.desc2matrix(my.con,my.tbl,variable[i],
42                              svd.proximal,stemming,stopwords,p.threshold,
43                              tfidf=tfidf,only.weka=only.weka))
44
45 print("flag4")
46
```

```
    if (length(my.tbl.bin)>1){
      a_c.dist=get.matrix.distances(my.tbl.bin,var=variable[i],
                                    distance=distance,a_c=a_c,
                                    my.con=my.con, full=full)
      aux=as.matrix(a_c.dist[,3:(ncol(a_c.dist)-1)])
      if (length(distance)==1)
        colnames(aux)=paste(variable[i],"_",distance,sep="")
      r=cbind(r,aux)
      print(paste("Descriptor:",variable[i],"Length:",length(my.tbl.bin)))
    }else print(paste("SHAMAM Warning:",variable[i],
                      "descriptor not performed."))
  }
print("flag5")
  r=subset(r,select=-1)#remove primeira coluna
  if (svd.distal) {
    r=try(svd.filter(r))
    if (is.numeric(r))
      colnames(r)=c(paste("svd",c(1:ncol(r)),sep=""))
  }
print("flag6")
  class=a_c.dist[,length(distance)+3]
  r=cbind(as.matrix(r),class)# insere a classe
  return(r)
}
```

### D.2.3 split.desc

```
split.desc=function(descriptor,maxchar=40){
  if (missing(descriptor)) stop("SHAMAM: Insert a string.")
  r=descriptor
  for (i in 1:length(descriptor)){
#     try(system("rm /tmp/rsystem.csv /tmp/rsystem.dat"))
    x=unique(strsplit(gsub("[^[:alnum:] ]", "", descriptor[i]), " +")[[1]])
    for (j in 1:length(x)) if(length(x[j]))
      if(nchar(x[j])>=60) x[j]=paste(split.str(x[j],10),collapse=" ")
    x=sort(unique(strsplit(gsub("[^[:alnum:] ]", "",
                                paste(x,collapse=" ")), " +")[[1]]))
    r[i]=as.character(paste(x,collapse=" "))
  }
  return(r)
}
```

### D.2.4 svd.filter

```
svd.filter=function(matriz){
  if (ncol(matriz)>1){
    x=clean.matrix(matriz)
    x <- matrix(as.numeric(matriz), nrow=nrow(matriz))
    s=svd(t(x))
    r=t(diag(s$d)%*%t(s$v))
    if (ncol(r)!=ncol(x) | nrow(r)!=nrow(x)) r=matriz
  }else r=matriz
```

```
  if (setequal(r,matriz)) print("SHAMAM Error: SVD not performed.")
  return(r)
}
```

## D.2.5 weka.classification

```
weka.classification=function(data=NULL,wekajar,classifier,infile=NULL,
                             train.filter=FALSE,timeout=300,option=""){
  if (missing(wekajar)) stop("SHAMAM: Insert weka's java file directory.")
  if (missing(classifier)){
    classifier="weka.classifiers.meta.RandomCommittee"
    print("SHAMAM Warning: Missing classifier. Using RandomCommittee.")
  }
  if (missing(infile) & missing(data))
    stop("SHAMAM: Insert a matrix, arff or csv file.")
  if (!is.null(data)) infile=csv2arff(data,wekajar,arff=infile,
                                      train=train.filter)
  if (is.null(infile))
    stop("SHAMAM: Arff conversion error. Verify weka files.")
  r=(try(system(paste("timeout",timeout,"java -Xmx60g -cp", wekajar,
                      classifier,option,"-t", infile,
                      "-p 1 | sed \"s/[+()']//g\" | awk '{$1=$1}1' |
                      sed 's/[ :]/,/g'| sed 1d | sed 1d |sed 1d |sed 1d |
                      sed 1d | sort -n -k7 -t',' | sed 1d |
                      awk -F',' '{print $3\",\"$5\",\"$6}'"),
         intern = TRUE)))
   if (length(r)>1){
     r=data.frame(matrix(unlist(strsplit(r,split=",")), ncol=3, byrow=T))
     colnames(r)=c("actual","predic","p")
   }else{
     r=NULL
     print(paste("SHAMAM: Error or timeout",
                 timeout,"seconds exceeded. Classifier: ",classifier))
   }
  return(r)
}
```

## D.2.6 weka.classification.optimization

```
weka.classification.optimization=function(data=NULL, wekajar, classifier,
                                          infile=NULL, train.filter=FALSE,
                                          timeout=300){
  if (missing(wekajar)) stop("SHAMAM: Insert weka's java file directory.")
  if (missing(classifier))
    stop("SHAMAM - Error: Insert a classifier.")
  if (missing(infile) & missing(data))
    stop("SHAMAM: Insert a matrix, arff or csv file.")
  if (!is.null(data)) infile=csv2arff(data,wekajar,arff=infile,
                                      train=train.filter)
  if (is.null(infile))
    stop("SHAMAM: Arff conversion error. Verify weka files.")

```

```
14  best.kappa=0; r=NULL;
15  if (classifier=="weka.classifiers.meta.RandomCommittee" |
16      classifier=="weka.classifiers.trees.RandomForest"|
17      classifier=="weka.classifiers.meta.RotationForest"){
18    for (s in 1:4)
19      for (i in seq(5, 50, by = 5)){
20        option=paste("-S",s,"-I",i); e=NULL;
21        e=try(weka.classification(infile=infile, wekajar=wekajar,
22                                  classifier=classifier,
23                                  timeout=timeout, option=option))
24        if(length(e)>1){
25          aux=try(t(as.matrix(weka.performance(e,as.vector=TRUE))))
26          if(length(aux)>1){
27            aux=cbind(aux,option);
28            colnames(aux)[ncol(aux)]="classifier_option"
29            rownames(aux)=classifier
30            if(is.null(r)) r=aux else r=rbind(r,aux)
31          }
32        }
33      }
34  }else
35  #------------------------------------------------
36  if (classifier=="weka.classifiers.meta.nestedDichotomies.ClassBalancedND"){
37    for (s in 2:5){
38      option=paste("-S",s); e=NULL;
39      e=try(weka.classification(infile=infile, wekajar=wekajar,
40                                classifier=classifier,
41                                timeout=timeout, option=option))
42      if(length(e)>1){
43        aux=try(t(as.matrix(weka.performance(e,as.vector=TRUE))))
44        if(length(aux)>1){
45          aux=cbind(aux,option);
46          colnames(aux)[ncol(aux)]="classifier_option"
47          rownames(aux)=classifier
48          if(is.null(r)) r=aux else r=rbind(r,aux)
49        }
50      }
51    }
52  }else #------------------------------------------------
53  if (classifier=="weka.classifiers.trees.J48graft" |
54      classifier=="weka.classifiers.meta.AttributeSelectedClassifier" |
55      classifier=="weka.classifiers.meta.nestedDichotomies.ND" |
56      classifier=="weka.classifiers.meta.FilteredClassifier"|
57      classifier=="weka.classifiers.trees.J48"|
58      classifier=="weka.classifiers.meta.OrdinalClassClassifier"){
59    for (m in 2:5)
60      for (c in seq(0.1, 0.5, by = 0.1)){
61        option=paste("-C",c,"-M",m); e=NULL;
62        e=try(weka.classification(infile=infile, wekajar=wekajar,
63                                  classifier=classifier,
64                                  timeout=timeout, option=option))
65        if(length(e)>1){
66          aux=try(t(as.matrix(weka.performance(e,as.vector=TRUE))))
67          if(length(aux)>1){
68            aux=cbind(aux,option);
69            colnames(aux)[ncol(aux)]="classifier_option"
```

```
            rownames(aux)=classifier
            if(is.null(r)) r=aux else r=rbind(r,aux)
          }
        }
      }
  }else
  #-------------------------------------------------
  if (classifier=="weka.classifiers.trees.RandomTree"){
    for (s in 1:5)
      for (m in 0:4){
        option=paste("-S",s,"-M",m); e=NULL;
        e=try(weka.classification(infile=infile, wekajar=wekajar,
                                  classifier=classifier,
                                  timeout=timeout, option=option))
        if(length(e)>1){
          aux=try(t(as.matrix(weka.performance(e,as.vector=TRUE))))
          if(length(aux)>1){
            aux=cbind(aux,option);
            colnames(aux)[ncol(aux)]="classifier_option"
            rownames(aux)=classifier
            if(is.null(r)) r=aux else r=rbind(r,aux)
          }
        }
      }
  }#-------------------------------------------------
  else{
    e=NULL;
    e=try(weka.classification(infile=infile, wekajar=wekajar,
                              classifier=classifier,
                              timeout=timeout))
    if(length(e)>1){
      r=try(t(as.matrix(weka.performance(e,as.vector=TRUE))))
      if(length(r)>1){
        r=cbind(r,"");
        rownames(r)=classifier
        colnames(r)[ncol(r)]="classifier_option"
      }
    }
  }
  return(r)
}
```

## D.2.7 weka.train.storming

```
weka.train.storming=function(data=NULL,wekajar,infile=NULL,method=NULL,
                             train.filter=FALSE,timeout=300){
  if (is.null(method))
  m=paste("weka.classifiers",
    c("bayes.BayesNet","bayes.NaiveBayesUpdateable","functions.Logistic",
    "functions.SMO", "functions.SimpleLogistic",
    "functions.MultilayerPerceptron", "functions.RBFNetwork",
    "lazy.IB1", "lazy.IBk", "lazy.KStar", "lazy.LWL", "meta.AdaBoostM1",
    "meta.AttributeSelectedClassifier", "meta.Bagging",
    "meta.ClassificationViaRegression", "meta.CVParameterSelection",
```

```
    "meta.Dagging", "meta.Decorate", "meta.END",
    "meta.FilteredClassifier", "meta.Grading", "meta.LogitBoost",
    "meta.MultiBoostAB", "meta.MultiClassClassifier", "meta.MultiScheme",
    "meta.nestedDichotomies.ClassBalancedND",
    "meta.nestedDichotomies.DataNearBalancedND",
    "meta.nestedDichotomies.ND", "meta.OrdinalClassClassifier",
    "meta.RacedIncrementalLogitBoost", "meta.RandomCommittee",
    "meta.RandomSubSpace", "meta.RotationForest","meta.Stacking",
    "meta.Vote", "misc.HyperPipes", "misc.VFI", "rules.ConjunctiveRule",
    "rules.JRip",  "rules.OneR", "rules.NNge", "rules.Ridor",
    "rules.ZeroR", "rules.PART", "trees.DecisionStump", "trees.FT",
    "trees.J48", "trees.J48graft", "trees.LADTree","trees.LMT",
    "trees.NBTree", "trees.RandomForest", "trees.RandomTree",
    "trees.REPTree"),sep=".")
  else m=method
  if (missing(wekajar)) stop("SHAMAM: Insert weka's java directory.")
  if (missing(infile) & missing(data))
    stop("SHAMAM: Insert a classifing matrix, arff or csv file.")
  if (!is.null(data)) infile=csv2arff(data,wekajar,arff=infile,
                                      train=train.filter)
  lm=length(m)
  print("SHAMAM: Starting classification storming.")
  print(paste("...................End-predicted maximum:",
        (Sys.time()+timeout*lm)))
  r=NULL
  for(i in 1:lm){
    e=try(weka.classification(wekajar=wekajar,classifier=m[i],
                              infile=infile, train.filter=train.filter,
                              timeout=timeout))
    if(!is.null(e)){
      aux=try(as.matrix(weka.performance(e,as.vector=TRUE)))
      colnames(aux)=m[i]
      if (is.null(r)) r=aux else r=cbind(r,aux)
      print(paste(round(aux["Kappa__all_mean",1],4),m[i],"total:",
                                                          i,"/",lm))
    }
  }
  print(paste("SHAMAM: Best Kappa ",
              sort(r["Kappa__all_mean",])[length(r["Kappa__all_mean",])]))
  return(r)
}
```

## D.3 Função terciária ou proximal

Função que chama outras funções shamam, mas não é chamada [5].

### D.3.1 shamam

```
shamam=function(my.con,my.tbl,variable=NULL,ignore="id",
                stemming=TRUE,stopwords=TRUE,p.threshold=0.01,
```

[5] Arquivo shamam_funcoes13091503_proximal.r

```
3                 svd.proximal=TRUE, classes=4, a_c,
4                 distance=list("euclidean","cosine"), resample=FALSE,
5                 wekajar=NULL,tfidf=TRUE,svd.distal=FALSE,
6                 train.filter=FALSE,timeout=300,prediction=FALSE,
7                 cluster.distal=FALSE,vectorize.with.weka=FALSE,
8                 a_c.ida="idA",a_c.idb="idB",a_c.class=NULL,
9                 my.tbl.id="id",experiment=NULL){
10  #Verificacao dos parametros
11  if (missing(wekajar)) stop("SHAMAM: Insert weka's java directory.")
12  if (missing(a_c))
13    stop("SHAMAM: Insert known instances (idA,idB,class).")
14  #Coleta distancias da tabela mysql
15  data=mysql.distances(my.con=my.con, my.tbl=my.tbl,
16                           variable=variable, ignore=ignore,
17                           stemming=stemming, stopwords=stopwords,
18                           p.threshold=p.threshold,
19                           svd.proximal=svd.proximal,distance=distance,
20                           classes=classes, a_c=a_c, a_c.ida=a_c.ida,
21                           a_c.idb=a_c.idb, a_c.class=a_c.class,
22                           wekajar=wekajar, tfidf=tfidf,id=my.tbl.id,
23                           svd.distal=svd.distal,full=FALSE,
24                           only.weka=vectorize.with.weka)
25                           print("flag")
26  #outros clustering serao implementados
27  if (cluster.distal==TRUE | cluster.distal=="kmeans")
28    data=feature.clustering(data,range=c(1,(ncol(data)-1)),
29                            clusters=classes)
30  #adapta missing ao formato do weka
31  data=clean.matrix(data,na.replace="?",inf.replace="?",decimals=12)
32
33  #Primeira corrida. Identifica os top5 classificadores
34  infile=weka.feature.selection(wekajar=wekajar,data=data,
35                                method="default",resample=resample)
36  aux=t(weka.train.storming(infile=infile,wekajar=wekajar,timeout=timeout))
37  r=cbind(aux,""); colnames(r)[ncol(r)]="classifier_option"
38  r=cbind(r,"default"); colnames(r)[ncol(r)]="feature_selection"
39
40  #Segunda corrida. Otimiza a selecao de atributos dos top5
41  best5=as.list(names(sort(r[,"Kappa__all_mean"])[(nrow(r)-4):nrow(r)]))
42
43  for(i in 1:5){
44    aux=weka.classification.optimization(infile=infile,
45                                         classifier=best5[i],
46                                         timeout=timeout,wekajar=wekajar)
47    aux=cbind(aux,i); colnames(aux)[ncol(aux)]="feature_selection"
48    if(!is.null(aux)) if(ncol(aux)==ncol(r)) r=rbind(r,aux)
49
50    for(j in seq(0.1, 0.9, by = 0.2)){
51      infile2=weka.feature.selection(wekajar=wekajar,data=data,method=j)
52      aux=weka.classification.optimization(infile=infile2,
53                                           classifier=best5[i],
54                                           timeout=timeout,
55                                           wekajar=wekajar)
56      aux=cbind(aux,j); colnames(aux)[ncol(aux)]="feature_selection"
57      if(!is.null(aux)) if(ncol(aux)==ncol(r)) r=rbind(r,aux)
58    }
```

```
  }
  r=r[order(r[,"Kappa__all_mean"],r[,"feature_selection"],decreasing=TRUE),]
  print(experiment)
  print(infile)
  print(infile2)
  print(rownames(r)[1])
  try(write.csv(r,file=paste("~/Dropbox/experimento_",
                             experiment,".csv",sep="")))
  #fim treino
  #inicio teste-------------------------------------------------------------
  e3=NULL
  if (prediction){
    print("SHAMAM: Getting full predictions")
    data3=mysql.distances(my.con=my.con, my.tbl=my.tbl,
                          ignore=ignore,
                          stemming=stemming, stopwords=stopwords,
                          p.threshold=p.threshold,
                          svd.proximal=svd.proximal,
                          classes=classes, distance=distance,
                          wekajar=wekajar, tfidf=tfidf,
                          svd.distal=svd.distal,full=TRUE)
    best5=r[1:5,(ncol(r)-1):ncol(r)]
#     best5[,2][best5[,2]=="default"] <-10#substitui para nao ter corte
#     for(j in 1:5){}
    method=best5[1,2];
    if (method!="default") method=(as.numeric(best5[1,2])/10)
    data3=weka.feature.selection(data=data3, wekajar, method=method,
                                 featselec, return.data=TRUE,
                                 timeout=300,infile=NULL,
                                 resample=FALSE,train.filter=FALSE)
    e3=try(weka.classification(wekajar=wekajar,rownames(best5)[1],
                          data=data3, train.filter=train.filter,
                          timeout=timeout,full=TRUE))
#     r2.perf=try(weka.performance(r2))
#     r=list(r,r2.perf,r2,data2,data3,r.o)
#     names(r)=c("all_performances","best_Kappa_classifier",
#                "best_feature_selection","prediction_performance",
#                "prediction","data_train","data_test","optimization")
    if (!is.null(e3)){
      r=e3;  print("SHAMAM Error: Predictions not performed.");
  }

  }
  return(r)
}
```

# Índice Remissivo