

WESLEY DIAS MACIEL

**Um Modelo para Descobertas Baseadas em  
Literatura Biológica**

**Uma Avaliação usando Patentes como Literatura de  
Referência**

Belo Horizonte  
14 de agosto de 2009



UNIVERSIDADE FEDERAL DE MINAS GERAIS  
INSTITUTO DE CIÊNCIAS BIOLÓGICAS  
INSTITUTO DE CIÊNCIAS EXATAS  
PROGRAMA DE DOUTORADO EM BIOINFORMÁTICA

**Um Modelo para Descobertas Baseadas em  
Literatura Biológica**  
**Uma Avaliação usando Patentes como Literatura de  
Referência**

Tese apresentada ao Programa de Doutorado em Bioinformática da Universidade Federal de Minas Gerais, como requisito parcial para a obtenção do grau de Doutor em Bioinformática.

WESLEY DIAS MACIEL

Belo Horizonte  
14 de agosto de 2009





UNIVERSIDADE FEDERAL DE MINAS GERAIS

FOLHA DE APROVAÇÃO

# Um Modelo para Descobertas Baseadas em Literatura Biológica

Uma Avaliação usando Patentes como Literatura de  
Referência

WESLEY DIAS MACIEL

Tese defendida e aprovada pela banca examinadora constituída por:

Prof. Ph. D. Sérgio Vale Aguiar Campos - Orientador  
Universidade Federal de Minas Gerais

Ph. D. Alessandra Conceição Faria Aguiar Campos - Co-orientadora  
Universidade Federal de Minas Gerais

Prof. Ph. D. José Palazzo Moreira de Oliveira  
Universidade Federal do Rio Grande do Sul

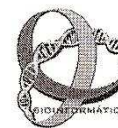
Prof. Ph. D. Mark Alan Junho Song  
Pontifícia Universidade Católica de Minas Gerais

Prof. Ph. D. Rubén Dario Sinisterra Millán  
Universidade Federal de Minas Gerais

Prof Ph. D. Marcelo Matos Santoro  
Universidade Federal de Minas Gerais

Belo Horizonte, 14 de agosto de 2009





**ATA DA DEFESA DA TESE DE DOUTORADO DE WESLEY DIAS MACIEL.**  
Aos quatorze dias do mês de agosto de 2009 às 14h00min, reuniu-se no Instituto de Ciências Exatas da Universidade Federal de Minas Gerais a Comissão Examinadora da tese de doutorado, indicada no três de julho de 2009, durante a 66ª reunião do Colegiado do Programa, para julgar, em exame final, o trabalho intitulado “Um Modelo para Descoberta Baseadas em Literatura Biológica”, requisito final para a obtenção do grau de Doutor em Ciências, Área de Concentração: Bioinformática. Abrindo a sessão o Presidente da Comissão, Prof. Dr. Sérgio Vale Aguiar Campos da Universidade Federal de Minas Gerais, 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 do resultado final. Foram atribuídas as seguintes indicações: Prof. Dr. José Palazzo Moreira de Oliveira, Porto Alegre, RS, aprovado; Prof. Dr. Mark Alan Junho Song da Pontifícia Universidade Católica de Minas Gerais, Belo Horizonte, MG, aprovado; Prof. Dr. Marcelo Matos Santoro da Universidade Federal de Minas Gerais, Belo Horizonte, MG, aprovado; Prof. Dr. Ruben Dario Sinisterra Millán da Universidade Federal de Minas Gerais, Belo Horizonte, MG, aprovado; Drª Alessandra Conceição Faria Aguiar Campos, co-orientadora da Universidade Federal de Minas Gerais, Belo Horizonte, MG, aprovado; Prof. Dr. Sérgio Vale Aguiar Campos, orientador, da Universidade Federal de Minas Gerais, Belo Horizonte, MG, 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 da Comissão encerrou a reunião e lavrou a presente ata que será assinada por todos os membros participantes da Comissão Examinadora. Belo Horizonte, aos quatorze dias de agosto de 2009.

  
Prof. Dr. José Palazzo Moreira de Oliveira – UFRGS


  
Prof. Dr. Mark Alan Junho Song – PUC-MG

  
Prof. Dr. Marcelo Matos Santoro – UFMG

  
Prof. Dr. Ruben Dario Sinisterra Millán – UFMG

  
Profª Drª Alessandra Conceição Faria Aguiar Campos – co-orientadora – UFMG

  
Prof. Dr. Sérgio Vale Aguiar Campos – orientador – UFMG

  
Prof. Dr. Wagner Mjira Junior  
Coordenador do Programa de Doutorado em  
Bioinformática - UFMG / Portaria DAP nº 2005





A DEUS, a meu pai, Miguel Alves Maciel, a minha mãe, Maria Fátima Dias Maciel, e a meus irmãos, Weik Dias Maciel e Walisson Dias Maciel.



# Agradecimentos

Início a escrita dos meus agradecimentos já pensando na enorme responsabilidade que me aguarda. Minha mente inicia longas viagens pelo passado. Penso imediatamente em meus professores e amigos de infância, de pessoas que ajudaram e possibilitaram a continuação de meus estudos, da moça que me vendia o lanche na cantina, dos pesquisadores ilustres nos congressos que participei, tanta gente... Todos com enorme contribuição na minha vida e formação, dando origem a algo grande demais, complexo demais para ser mencionado e agradecido aqui. Por isso, já começo meus agradecimentos pedindo a DEUS que retribua a todas essas pessoas a valiosa ajuda que um dia me prestaram e que também me torne capaz de sempre ajudar aqueles que buscam o crescimento, a evolução. Peço desculpas àqueles cujos nomes, por ventura, não apareçam nesse texto, mas dede já apresento a todos o meu sincero e cordial muito obrigado!

A conclusão deste trabalho não teria sido possível sem a ajuda e o comprometimento de duas pessoa muito especiais: o meu orientador, o professor Sérgio Vale Aguiar Campos, e minha co-orientadora, a Alessandra Conceição Faria Aguiar Campos. Entretanto, agradecê-los somente pela orientação e co-orientação é definitivamente muito pouco. Eu também preciso agradecê-los por diversas outras coisas que promoveram intensamente o meu crescimento e desenvolvimento durante nosso convívio nesses anos de doutorado. Foram tantos ensinamentos, tanta ajuda e tanto apoio que me tornei eternamente grato por conhecê-los. Hoje, além do enorme respeito e admiração que sinto por eles, também sinto-me privilegiado por ter recebido suas atenções nessa época de tão intenso trabalho.

No desenvolvimento do trabalho, o conhecimento e direcionamento do professor Marcos André Gonçalves foram determinantes na solução de vários problemas. Além disso, suas inúmeras sugestões ao revisar nossos artigos promoveram um forte amadurecimento de minhas idéias sobre o projeto. Então, é por essas tão preciosas ajudas, que eu o agradeço enormemente.

Eu também não poderia deixar de expressar meus sinceros agradecimentos a todos os professores que compuseram minha banca de defesa de tese: José Palazzo

Moreira de Oliveira, Mark Alan Junho Song, Rubén Dario Sinisterra Millán e Marcelo Matos Santoro. Inicialmente, agradeço-os extremadamente por aceitarem a participação em minha banca de defesa e por destinarem tempo à leitura e correções sobre a tese. Além disso, agradeço-os pelas opiniões e sugestões apresentadas no momento da defesa e também pelo apoio e cumprimentos pelo trabalho.

Agradeço também a todos os professores do programa de doutorado em Bioinformática da Universidade Federal de Minas Gerais pelos ensinamentos, apoio e direcionamentos durante minha passagem pelo curso. Em especial, gostaria de agradecer aos professores Sérgio Vale Aguiar Campos, Paulo Sérgio Lacerda Beirão, Glória Regina Franco, Marcelo Matos Santoro, Eduardo Martin Tarazona Santos, Frederico Ferreira Campos Filho e Gregorio Saravia Atuncar, pois, em muitos momentos, contribuíram de forma especial para o meu crescimento com suas idéias, aulas, palavras, exemplos e ações.

No início de meu curso, contei com a ajuda e apoio de três pesquisadores que foram muito importantes naquela fase do trabalho: o professor Claudionor José Nunes Coelho Júnior, François Artiguenave e o professor Júlio César Dias Lopes. Agradeço-os por me apresentarem a idéia e o caminho inicial do que seria pesquisado e por me incentivarem nos primeiros projetos do trabalho. Recordando-me ainda dessa fase inicial de meu curso de doutorado, também agradeço enormemente aos professores Nivio Ziviani, Antonio Alfredo Ferreira Loureiro, Mark Alan Junho Song e Manoel Palhares Moreira por minhas indicações ao doutorado em Bioinformática.

A vida acadêmica durante o doutorado teria sido excessivamente árdua se não fosse a amizade, convívio, ajuda, atenção e apoio de meus colegas da Bioinformática, do ICB-UFMG e do DCC-UFMG. Por essa razão é que registro aqui o meu sincero agradecimento a todos eles, especialmente a Michael Waisberg, Raquel Cardoso de Melo Minardi, Cristina Ribeiro, Adriano Barbosa da Silva, Ubirajara Fumega e Júlio César Torres Rodrigues. Sinto também a enorme importância de agradecer a atenção e ajuda dos funcionários da secretaria da Bioinformática, especialmente a Kátia Moraes Leite, Alberto Salazar Costa, Carlos Eduardo Fernandes dos Santos e Pollyanna Martins de Almeida. Também agradeço a todos os funcionários do DCC-UFMG e, em particular, a todo o pessoal da secretaria da pós-graduação em Ciência da Computação.

Também apresento meu enorme agradecimento à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa de estudos do programa especial BIOMICRO, que foi muito importante e me permitiu a realização do trabalho. Agradeço também à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) pelo apoio financeiro nos congressos que participei.

Agradeço ainda a meus amigos Manoel Palhares Moreira e Márcia Carolina Marra de Oliveira por terem dado o incentivo e apoio inicial aos meus estudos em bioinformática. Agradeço a meu amigo Christiano Coutinho de Moraes pelas várias vezes em que se dispôs a me ajudar na realização deste trabalho. Agradeço a meu amigo Mário Sérgio Alvim pelo debate, idéias e sugestões que ajudaram a direcionar o trabalho. Agradeço também a minhas queridas amigas Álica de Castro Silva e Jaqueline Martins Ferreira pelas orações e apoio nos momentos difíceis. Agradeço a Maria Madalena Magnabosco e a Maura Horta Simões pelos ensinamentos e pela ajuda no desvendar e compreender a vida. Agradeço a Hanna Safar, Daniela Melo Gomes, Leila Ribeiro Junqueira, Alba Valéria Souto Melo Moraes, Cristiane Amorim de Paula, Icek Majer Weireich, Stefânia Villela Moreira, João Carlos Guedes da Silva, Afonso Abreu Fernandes, Renato Teixeira Penna Mascarenhas, Silvestre Silva Melo, Cláudio Maciel de Sena, Carlos Alberto Rodrigues da Silva, Paulo Alves de Oliveira, Antônio Minelvino e Laura de Freitas Xavier pela atenção e ensinamentos, em particular aqueles sobre saúde.

Faço um agradecimento especial a Maria Cecília de Souza Silva por seu cuidado, atenção, sua enorme amizade e seus ensinamentos e orações. Faço outro agradecimento especial a Marcelo Antônio de Menezes pelo apoio, ajuda, atenção, revisão de textos, planejamento de figuras e pelos momentos de descontração que permitiram dar continuidade ao trabalho. À Maria Cecília e ao Marcelo também agradeço por terem me apresentado novos horizontes e uma nova realidade sobre as coisas grandiosas e importantes da vida.

Sem palavras para expressar minha enorme gratidão e amor, agradeço a meu pai, Miguel Alves Maciel, minha mãe, Maria Fátima Dias Maciel, e meus irmãos, Weik Dias Maciel e Walisson Dias Maciel, por serem meus pilares, por estarem incondicionalmente presentes em todos os momentos da minha vida e por serem meus motivos de real e profunda felicidade.

Por fim, e com amor eterno, agradeço a DEUS, por ser minha força, luz e inspiração, por me guiar nos momentos de dificuldade e assegurar o meu crescimento, amadurecimento e evolução.

Ah, agradeço também a você que agora inicia a leitura deste trabalho. Muito obrigado!



## Resumo

Entidades biológicas de diferentes categorias como alvo biológico, doença, fármaco e gene interagem entre si em um sistema biológico formando uma complexa rede. Frequentemente, essas entidades desempenham mais de uma atividade no sistema. Algumas dessas atividades são bem conhecidas e integram o conhecimento em biologia. Entretanto, outras atividades não são bem conhecidas ou permanecem desconhecidas por um longo tempo e, comumente, são descobertas ao acaso. Neste trabalho, desenvolvemos uma abordagem sistemática para inferir interações novas entre entidades biológicas, explorando coleções textuais que cobrem o conhecimento em biologia. Nosso modelo de inferência usa o modelo de espaço vetorial para construir uma rede biológica formada por sub-redes n-dimensionais de interações conhecidas entre entidades e que são extraídas da coleção de documentos. Cada dimensão de uma sub-rede representa uma categoria diferente do sistema biológico, como eco-sistemas, organismos, órgãos, tecidos, células, organelas, genes, proteínas, doenças e fármacos. A partir das interações conhecidas e estabelecidas em cada sub-rede, o modelo aplica uma relação de transitividade, para inferir novas interações entre entidades. Testamos e validamos nosso modelo através de uma coleção de documentos formada pela seção de reivindicação de patentes. Nós construímos a rede de acordo com os anos em que as patentes foram publicadas e observamos que novas interações encontradas em um ano foram confirmadas por patentes que não estavam na coleção de documentos e que foram publicadas em anos subsequentes. Além dessa validação baseada na data de publicação das patentes, também procuramos por artigos científicos publicamente disponíveis na *Web* com o objetivo de confirmar novas interações inferidas nas sub-redes. Por exemplo, o melhor resultado encontrado em nosso modelo indica a interação entre o neurotransmissor adrenalina e o gene receptor de androgênio. Encontramos um artigo que afirma que o efeito antiapoptótico da adrenalina parcialmente depende do receptor de androgênio. Nosso modelo é capaz de encontrar novas interações entre entidades através de conexões implícitas da literatura biológica. Além disso, e mais importante, nosso modelo é capaz de ordenar essas novas interações com base no valor da similaridade apresentada no modelo de espaço vetorial. Essa ordenação auxilia os pesquisadores

em seus trabalhos, indicando as interações entre entidades mais promissoras a serem consideradas.



## Abstract

Biological entities from different categories such as diseases, drugs, genes and targets interact among themselves in a biological system developing a complex network. Frequently, these entities perform more than one activity in the system. Some of these activities are well known and integrate the knowledge in life sciences. However, other activities are not well known or remain unknown for a long time and are discovered by chance. In this work we have developed a systematic approach to predict new interactions between biological entities by exploiting textual collections that cover the knowledge in life sciences. Our inference model use the vector space model to construct a biological network formed by n-dimensional subnetworks of known entity interactions extracted from the textual collection. Each subnetwork dimension represent a different category of a biological system such as ecosystems, organisms, organs, tissues, cells, organelles, genes, proteins, diseases and drugs. From the known interactions established in each subnetwork the model applies a transitive closure in order to predict the new entity interactions. We have tested and validated our model with a textual collection formed by patent claims. Iterating the model according to the years in which the patents were issued, we have observed that new interactions found in a year were confirmed by patents not in the collection and issued in a more recent year. In addition to our validation based on the patent issue dates, we have also looked for papers publicly available on the Web in order to confirm some of the new interactions found. For instance, the best result found in our model relates the interaction between the adrenaline neurotransmitter and the androgen receptor gene. We have found a paper reporting that the antiapoptotic effect of adrenaline partially depends on androgen receptor. Our model is able to find new entity interactions based on implicit connections in the biological literature. In addition and most importantly, our model has been able to rank these new interactions based on the similarity value stated in the vector space model. This ranking helps researchers to carry out their works pointing out promising entity interactions to consider.



# Conteúdo

<b>Lista de Figuras</b>	<b>xiii</b>
<b>Lista de Tabelas</b>	<b>xvii</b>
<b>Lista de Algoritmos</b>	<b>xix</b>
<b>1 Introdução</b>	<b>1</b>
1.1 Trabalhos Relacionados . . . . .	3
1.2 Contribuição . . . . .	9
<b>2 O Sistema BioSearch</b>	<b>11</b>
2.1 Os Módulos do Sistema . . . . .	12
2.1.1 A Tarefa de Coleta . . . . .	13
2.1.2 A Tarefa de Indexação da Coleção de Documentos . . . . .	16
2.1.3 A Tarefa de Construção da Rede Biológica . . . . .	19
2.1.4 A Tarefa de Consulta das Interações da Rede Biológica . . . . .	40
2.1.5 O Custo Total do Sistema . . . . .	43
2.1.6 O Desempenho do Sistema . . . . .	44
<b>3 O Modelo de Inferência</b>	<b>47</b>
3.1 A Rede Biológica e suas Sub-Redes . . . . .	48
3.2 A Formação dos Espaços Dimensionais . . . . .	49
3.2.1 As Interações entre Entidades Biológicas . . . . .	50
3.3 As Representações das Sub-Redes no Modelo . . . . .	51
3.4 A Construção das Sub-Redes . . . . .	53
3.4.1 A Identificação das Interações Conhecidas . . . . .	54
3.4.2 A Inferência das Novas Interações . . . . .	58
3.4.3 A Convergência em Sub-Redes . . . . .	60
3.5 O Conjunto Solução Retornado pelo Modelo . . . . .	62
3.5.1 A Descrição do Conjunto Solução das Novas Interações . . . . .	63

<b>4</b>	<b>Metodologia</b>	<b>65</b>
4.1	O Modelo . . . . .	66
4.1.1	As Interações Biológicas . . . . .	69
4.1.2	A Validação do Modelo . . . . .	71
4.2	O Sistema . . . . .	73
4.2.1	A Tarefa de Coleta . . . . .	73
4.2.2	A Tarefa de Indexação da Coleção de Documentos . . . . .	76
4.2.3	A Tarefa de Construção da Rede Biológica . . . . .	77
4.2.4	A Tarefa de Consulta das Interações da Rede Biológica . . . . .	78
4.2.5	A Formação da Base de Dados . . . . .	80
<b>5</b>	<b>Resultados</b>	<b>87</b>
5.1	A Construção da Rede . . . . .	87
5.1.1	O Espaço de Busca de Novas Interações . . . . .	89
5.1.2	O Histórico de Formação das Interações entre Entidades . . . . .	91
5.1.3	A Distribuição da Evidência de Interação . . . . .	94
5.2	A Validação . . . . .	95
<b>6</b>	<b>Discussão</b>	<b>101</b>
<b>7</b>	<b>Conclusão</b>	<b>109</b>
	<b>Apêndices</b>	<b>115</b>
	Apêndice A Os Nomes das Entidades Biológicas . . . . .	115
	Apêndice B O Diagrama de Entidades e Relacionamentos (DER) . . . . .	118
	Apêndice C A Distribuição da Evidência de Interação . . . . .	120
	Apêndice D A Distribuição das Patentes de Confirmação . . . . .	121
	Apêndice E A Confirmação do Sistema de Busca do USPTO . . . . .	124
	<b>Bibliografia</b>	<b>129</b>

# Lista de Figuras

2.1	Representação do sistema <i>BioSearch</i> . . . . .	11
2.2	Representação da arquitetura em 3 camadas do sistema <i>BioSearch</i> . . . . .	12
2.3	Tarefa 1: coleta dos documentos que descrevem o conhecimento em biologia. . . . .	13
2.4	Tarefa 2: indexação da coleção de documentos que descreve o conhecimento em biologia. . . . .	16
2.5	Índice invertido. . . . .	17
2.6	Cálculo do peso das entidades na coleção de documentos. . . . .	18
2.7	Tarefa 3: construção da rede biológica. . . . .	19
2.8	Representação da formação de espaços dimensionais. . . . .	21
2.9	Representação das iterações para formação de espaços dimensionais. . . . .	22
2.10	Representação da formação de interações através do produto cartesiano entre entidades. . . . .	23
2.11	Representação do processo de leitura das entidades que formam uma rede. . . . .	26
2.12	Representação da formação do produto cartesiano. . . . .	27
2.13	Interseção de listas invertidas. . . . .	29
2.14	Relação de transitividade em uma matriz representando as conexões de uma sub-rede bidimensional genérica. . . . .	34
2.15	Tarefa 4: consulta das interações da rede biológica. . . . .	41
3.1	Representação da formação dos espaços dimensionais. . . . .	49
3.2	Representação de interações entre entidades no espaço dimensional <i>fármaco</i> $\times$ <i>doença</i> . . . . .	52
3.3	Grafo ponderado representando interações da sub-rede <i>fármaco</i> $\times$ <i>doença</i> . . . . .	53
3.4	Representação de uma sub-rede quadridimensional em um espaço bidimensional. . . . .	54
3.5	Relação de transitividade em uma matriz representando as interações em uma sub-rede bidimensional genérica. . . . .	54

3.6	Construção das sub-redes. . . . .	55
3.7	Representação dos vetores de pesos de um documento e de uma consulta. . . . .	56
3.8	Convergência. . . . .	61
3.9	Representação das condições impostas pelo conjunto solução. . . . .	64
4.1	Grafo de interação entre entidades biológicas construído a partir de patentes biotecnológicas. . . . .	70
4.2	Tarefa 1: coleta de patentes que descrevem o conhecimento em biologia. . . . .	73
4.3	<i>Ranking</i> gerado pelo sistema de busca do USPTO . . . . .	74
4.4	Interface <i>Web</i> para administradores do sistema <i>BioSearch</i> . . . . .	75
4.5	Interface <i>Web</i> para usuários do sistema <i>BioSearch</i> . . . . .	76
4.6	Pesquisa através do tipo de interação estabelecida na rede. . . . .	77
4.7	Pesquisa na sub-rede <i>doença</i> $\times$ <i>fármaco</i> . . . . .	78
4.8	<i>Ranking</i> de resposta a consulta na sub-rede <i>doença</i> $\times$ <i>fármaco</i> . . . . .	79
4.9	Patente que comprova uma interação conhecida e escolhida pelo usuário. . . . .	80
4.10	Passos do processo de inferência que levaram à indicação de uma nova interação. . . . .	81
5.1	Representação do espaço de busca para algumas das interações possíveis envolvendo o fármaco aspirina. . . . .	90
5.2	Inferência da melhor interação encontrada pelo modelo. . . . .	92
5.3	Distribuição das evidências de interação (I). . . . .	93
5.4	Número de interações com patentes de confirmação por ano. . . . .	96
5.5	Distribuição das interações inferidas em 2004 e que possuem confirmações patenteadas em 2005 (I). . . . .	97
5.6	Distribuição das interações inferidas em 2004 e que possuem confirmações patenteadas em 2005 (II). . . . .	97
5.7	Distribuição das interações inferidas em 2004 e que possuem confirmações patenteadas em 2005 (III). . . . .	98
5.8	Página do USPTO confirmando a inexistência de patentes em que co-ocorram as entidades adrenalina e receptor de androgênio (I). . . . .	100
1	Diagrama de entidades e relacionamentos da base de dados do sistema <i>BioSearch</i> . . . . .	119
2	Distribuição das evidências de interação (II). . . . .	120
3	Distribuição das evidências de interação (III). . . . .	121
4	Distribuição das interações inferidas em 2004 e que possuem confirmações patenteadas em 2005 (IV). . . . .	122

5	Distribuição das interações inferidas em 2004 e que possuem confirmações patenteadas em 2005 (V). . . . .	123
6	Página do USPTO confirmando a inexistência de patentes em que co-ocorram as entidades adrenalina e receptor de androgênio (II). . .	124
7	Página do USPTO confirmando a inexistência de patentes em que co-ocorram as entidades adrenalina e receptor de androgênio (III). . .	125
8	Página do USPTO confirmando a inexistência de patentes em que co-ocorram as entidades adrenalina e receptor de androgênio (IV). . .	126
9	Página do USPTO confirmando a inexistência de patentes em que co-ocorram as entidades adrenalina e receptor de androgênio (V). . .	127





# Lista de Tabelas

4.1	Número de entidades em cada categoria biológica. . . . .	66
4.2	Sub-redes do sistema biológico. . . . .	67
4.3	Exemplo de interações entre entidades biológicas construído a partir de patentes biotecnológicas. . . . .	69
4.4	Entidades biológicas consideradas nos experimentos. . . . .	71
5.1	Descrição da rede de interações biológicas. . . . .	87
5.2	Ordenação das sub-redes de acordo com as melhores interações infe- ridas pelo modelo. . . . .	88
5.3	As 5 melhores novas interações de toda a rede de interação. . . . .	94
5.4	As 5 melhores novas interações encontradas na sub-rede <i>alvo</i> $\times$ <i>gene</i> . . . . .	95
5.5	As 5 interações conhecidas com maior evidência de interação em 2005 e que se tornaram novas em 2004. . . . .	99
5.6	Distribuição das patentes de confirmação por nível do <i>ranking</i> de res- posta de cada sub-rede. . . . .	100



# Lista de Algoritmos

2.1	Módulo 1: extração de entidades e categorias biológicas. . . . .	14
2.2	Módulo 2: coleta de documentos. . . . .	15
2.3	Módulo 3: pré-processamento dos documentos. . . . .	15
2.4	Módulo 4: construção do índice invertido. . . . .	18
2.5	Módulo 5: formação dos espaços n-dimensionais. . . . .	20
2.6	Módulo 6: formação das interações possíveis entre entidades. . . . .	24
2.7	Módulo 6: produto cartesiano das entidades em cada sub-rede. . . . .	25
2.8	Módulo 7: construção das sub-redes. . . . .	28
2.9	Módulo 7: encontrar as interações conhecidas de uma sub-rede - versão: média aritmética das similaridades. . . . .	29
2.10	Módulo 7: encontrar as interações conhecidas de uma sub-rede - versão: soma das similaridades. . . . .	30
2.11	Módulo 7: encontrar as interações conhecidas de uma sub-rede - versão: máxima similaridade. . . . .	31
2.12	Módulo 8: modelo de espaço vetorial. . . . .	32
2.13	Módulo 9: convergência de novas interações em uma sub-rede. . . . .	35
2.14	Módulo 9: inferência de novas interações. . . . .	37
2.15	Módulo 10: processamento de consulta. . . . .	41



# Capítulo 1

## Introdução

Em um sistema biológico, encontramos entidades de diferentes categorias desempenhando atividades importantes no sistema. Algumas entidades constituem os blocos construtores do sistema, como genes e proteínas. Outras entidades são responsáveis por indicar ou mudar o estado do sistema, como doenças e fármacos. A ação de uma entidade pode mediar ou interferir na ação de outras entidades, desenvolvendo uma rede complexa de interações. Frequentemente, as entidades desempenham mais de uma atividade no sistema. Algumas dessas atividades são conhecidas e integram o conhecimento em biologia. No entanto, algumas atividades não são bem conhecidas ou permanecem desconhecidas por um longo tempo e, comumente, são descobertas ao acaso. Os fármacos, por exemplo, têm uma atividade farmacológica principal e atividades secundárias responsáveis por efeitos colaterais no sistema biológico. Contudo, os efeitos colaterais de um fármaco podem ser explorados como novos usos no tratamento de muitas doenças. Um exemplo importante é o citrato de sildenafil (Viagra) que foi originalmente desenvolvido para o tratamento de angina e hipertensão. Entretanto, os testes clínicos do Viagra revelaram seu efeito colateral de aumentar a função erétil (Silverman, 2004). Em nosso trabalho, desenvolvemos uma abordagem sistemática para construir uma rede de interação entre entidades biológicas com base em uma coleção de documentos que descreve o conhecimento em biologia. Dessa forma, conseguimos inferir novas interações entre as entidades a partir das interações já conhecidas e estabelecidas na rede.

Após o seqüenciamento do genoma humano, diversas pesquisas passaram a proporcionar um grande aumento do conhecimento em biologia. Desde então, muito desse conhecimento passou a ser publicado na *World Wide Web*, ou simplesmente *Web*, com o intuito de aumentar e acelerar o número de descobertas. Um fenômeno similar aconteceu quando a *Web* foi criada, dando origem às bibliotecas digitais. Diversas páginas eletrônicas foram publicadas num ritmo acelerado, gerando um

enorme emaranhado de informação interconectada. Várias pesquisas foram, então, desenvolvidas com o objetivo de extrair e analisar a informação publicada nesse emaranhado. Dessas pesquisas, o modelo de espaço vetorial (Salton and McGill, 1983; Baeza-Yates and Ribeiro-Neto, 1999; Witten *et al.*, 1999) vem sendo considerado um importante método para recuperar informação na *Web*. Nesse trabalho, desenvolvemos um modelo de inferência baseado no modelo de espaço vetorial, para descobrir novos relacionamentos entre entidades biológicas de categorias distintas como ecossistemas, organismos, órgãos, tecidos, células, organelas, genes, proteínas, doenças e fármacos. Nosso modelo constrói uma rede de interações conhecidas entre as entidades de categorias distintas, usando uma coleção textual formada por documentos publicados na *Web* que descrevem o conhecimento em biologia. O modelo percorre e analisa essa rede para inferir novas interações entre as entidades biológicas através de uma relação de transitividade que implementamos e avaliamos nesse trabalho. O processo de inferência abordado na relação de transitividade explora as atividades principais e secundárias que as entidades desempenham no sistema biológico, com base no princípio de que "**IF** uma entidade  $x$  interage com as entidades  $y$  e  $w$  **AND** uma outra entidade  $z$  também interage com a entidade  $y$ , **THEN**  $z$  possivelmente também interage com a entidade  $w$ ".

Em nossa rede biológica, as interações conhecidas entre entidades representam a co-ocorrência dessas entidades em pelo menos um dos documentos que compõem a literatura biológica e que estão na coleção textual. Nosso objetivo é usar essas interações já conhecidas entre as entidades para inferir e ordenar interações novas. Assim, nós podemos apresentar as interações mais promissoras que devem ser analisadas para promover avanços no conhecimento sobre biologia.

Implementamos nosso modelo em um sistema chamado *BioSearch* (BioSearch, 2009) e construímos uma prova de conceito para testá-lo, usando entidades de 4 categorias diferentes: alvo biológico, doença, fármaco e gene. Nossa coleção de documentos é formada pela seção de reivindicação de 17.830 patentes coletadas no sítio *Web* de patentes dos Estados Unidos da América, *United States Patent and Trademark Office* (USPTO), (USPTO, 2009). Utilizamos a seção de reivindicação, porque este é o campo mais importante na especificação de uma patente, uma vez que apresenta a invenção, definindo o escopo de proteção (Shinmori *et al.*, 2003; USPTO, 2009). A partir de 266.528 interações possíveis entre as entidades de nossa rede, o modelo identificou 1.027 interações conhecidas nas seções de reivindicação das patentes pertencentes a nossa coleção de documentos e inferiu 3.195 interações novas. Dessa forma, o modelo construiu uma rede com um total de 4.222 interações que podem ser consultadas por pesquisadores, para elaboração de estratégias que

promovam novos avanços científicos e tecnológicos.

Nós validamos nossos experimentos usando a data de publicação das patentes que formam a coleção de documentos, reconstruindo nossa rede de interações biológicas num intervalo de 30 anos. Observamos que as novas interações encontradas em um ano foram confirmadas por patentes publicadas em anos subsequentes. Por exemplo, a interação entre a doença ataque do coração e o gene *ppar-gama* é relatada por 1 patente publicada em 2005. Quando essa patente é removida da coleção, 61 patentes publicadas até o ano de 2004 passam a indicar essa interação como uma interação nova para a rede. Nós também pesquisamos artigos científicos publicamente disponíveis na *Web*, para confirmar algumas das novas interações encontradas. Por exemplo, o melhor resultado encontrado em nosso modelo indica a interação entre o neurotransmissor adrenalina e o gene receptor de androgênio na sub-rede bidimensional *alvo*  $\times$  *gene*. Nenhuma seção de reivindicação das patentes que formam nossa coleção de documentos relata essa interação. No entanto, Sastry *et al.* (2007) afirmam em 2007 que o efeito antiapoptótico da adrenalina parcialmente depende do receptor de androgênio.

## 1.1 Trabalhos Relacionados

Estudos em biologia sistêmica analisam a operação de sistemas biológicos complexos e as perturbações que afetam esses sistemas, como as perturbações provocadas pela administração de fármacos (Butcher *et al.*, 2004; Naylor, 2004). Os avanços em biologia sistêmica têm um grande valor para diversas áreas como medicina, farmácia, agricultura e química. No descobrimento de novos fármacos, por exemplo, a biologia sistêmica fornece importantes meios para identificação racional de novos alvos biológicos (Kitano, 2002; Hood and Perlmutter, 2004). Os estudos nessa área também têm um importante papel no desenvolvimento de modelos preditivos de doenças humanas, uma vez que muitas dessas doenças ocorrem a partir de interações moleculares complexas ao invés de uma única alteração molecular (Butcher *et al.*, 2004; Naylor, 2004). Além disso, a biologia sistêmica também promove uma avaliação de predisposições a certas doenças, diagnóstico de doenças e identificação do progresso de doenças (Naylor, 2004; Hood and Perlmutter, 2004). Em nosso trabalho, desenvolvemos um modelo para construção de uma rede que representa sistemas biológicos a partir da literatura que descreve o conhecimento em biologia.

Avanços em métodos de alto desempenho resultaram na descrição de várias redes de interações moleculares em sistemas biológicos. A reconstrução e análise dessas redes biológicas *in-silico* têm recebido importante atenção em biologia sistêmica

(Friedman, 2004; Alon, 2003; Ambesi-Impiombato and di Bernardo, 2006; Alm and Arkin, 2003; Hopkins, 2007; Csermely *et al.*, 2005) e muitos trabalhos têm reconstruído essas redes através de perfis de expressão obtidos através de experimentos com microarranjos (Brazma and Vilo, 2000). O trabalho desenvolvido por Lamb *et al.* (2006) tem uma grande importância nessa área de pesquisa. Os autores usam o termo assinatura para designar o conjunto de características de entidades das categorias gene, doença e fármaco. Eles descreveram estados biológicos em termos dessas assinaturas, para mapear relações entre entidades dessas 3 categorias biológicas. Além disso, eles criaram uma base de dados para armazenar assinaturas de fármacos e genes e desenvolveram uma ferramenta para casamento de padrão que detecta as similaridades entre essas assinaturas. Ideker *et al.* (2001) também contribuíram significativamente para as pesquisas na área de biologia sistêmica. Nesse trabalho, eles aplicaram uma estratégia para integrar dados sobre interações celulares. Assim, eles conseguiram assimilar esses dados em um modelo biológico capaz de prever o comportamento celular.

Além desses trabalhos, diversos algoritmos têm sido apresentados para construção de redes biológicas. Basso *et al.* (2005) desenvolveram um algoritmo para reconstruir redes de interações celulares a partir de perfis de expressão obtidos em experimentos com microarranjos e, então, inferir novas interações nessa rede. O algoritmo identifica co-regulações gênicas estatisticamente significantes na rede através de uma métrica para relacionamentos usada na teoria da informação. Grabe and Neuber (2005) apresentaram um algoritmo para um modelo de biologia sistêmica capaz de simular a homeostase da epiderme humana e de prever propriedades dermatológicas básicas. Gardner *et al.* (2003) construíram um modelo de interações reguladoras que se baseia na análise de múltiplas regressões lineares de perfis transcricionais. A abordagem é usada para explicar as propriedades funcionais de redes genéticas e identificar alvos moleculares de compostos farmacológicos. Yamanishi *et al.* (2008) desenvolveram um método supervisionado para prever novas interações entre fármacos e alvos biológicos. Eles integraram em um espaço bidimensional a similaridade estrutural química dos fármacos e a similaridade da seqüência de aminoácidos dos alvos. Guimerà *et al.* (2007) propuseram uma abordagem que se baseia no mapeamento de interações entre agentes bioquímicos, como uma ferramenta para identificação de alvos biológicos para fármacos. Embora muito importantes e interessantes, esses trabalhos restringem o número de categorias biológicas usadas para construir a rede biológica. Além disso, eles não aplicam o modelo de espaço vetorial como arcabouço algébrico capaz de extrair as interações entre entidades biológicas a partir de uma coleção de documentos que descreve o conhecimento em biologia.



Em nosso trabalho, usamos as categorias biológicas para reduzir o espaço de pesquisa de interações biológicas e promover resultados mais apurados. Além disso, o modelo de espaço vetorial é uma ferramenta para recuperação de informação cuja similaridade nos propiciou um importante meio para ordenação das interações que são estabelecidas na rede biológica.

Nós também podemos construir redes biológicas a partir de outras fontes de informação diferentes daquelas diretamente relacionadas a perfis de expressão obtidos em experimentos com microarranjos. Uma forma alternativa de construir essas redes é explorar o conhecimento relatado na literatura biológica. Nesse sentido, algoritmos para descobertas baseadas em literatura são uma importante fonte de investigação, porque permitem identificar conexões explícitas e implícitas na literatura biológica que podem levar a novas interações entre entidades (Bruza and Weeber, 2008). Swanson (1986, 1990) é o pioneiro nas pesquisas sobre descobertas baseadas em literatura. Ele usou o silogismo  $A \rightarrow B$  AND  $B \rightarrow C$  THEN  $A \rightarrow C$ , para descobrir conexões implícitas entre documentos de uma literatura. Smalheiser and Swanson (1998) e Swanson *et al.* (2006) implementaram esse silogismo em um sistema chamado *ARROWSMITH*, para facilitar a descoberta baseada em literatura. Weeber *et al.* (2001) contribuíram para essa área de pesquisa, apresentando um modelo que usa técnicas de processamento natural de linguagem (PNL) para encontrar conceitos na literatura biomédica e reduzir o espaço de busca. Hristovski and B. Peterlin (2005) implementaram o silogismo proposto por Swanson em um sistema interativo chamado *BITOLA*. O objetivo desse sistema é dar suporte à descoberta de genes candidatos em relacionamento etiológico com doenças. Na implementação desse sistema, eles consideraram a localização dos genes nos cromossomos, para restringir o número de interações encontradas na literatura. Hristovski *et al.* (2006) apresentaram um método para melhorar a descoberta baseada em literatura, usando predicados semânticos. Esses predicados semânticos correspondem a relações semânticas que são extraídas de textos biomédicos através de sistemas de processamento natural de linguagem. Os predicados semânticos são empregados no método com o objetivo de facilitar a avaliação de novas interações. No entanto, esses trabalhos não se dedicam à construção de redes de interações biológicas.

Considerando a construção de redes e mapas de interações através de algoritmos para descobertas baseadas em literatura, Wren *et al.* (2004) descreveram um método baseado no silogismo proposto por Swanson com o objetivo de identificar relacionamentos potenciais em uma literatura biomédica. Os autores definiram categorias de interesse (e. g. genes, doenças, fenótipos e compostos químicos) e modelaram os relacionamentos em uma rede usando a teoria de lógica *fuzzy*. Por outro lado,

Campillos *et al.* (2008) construíram uma rede conectando fármacos comercializados e proteínas alvo, extraíndo termos relevantes a partir das bulas desses fármacos. Eles desenvolveram uma métrica de similaridade para efeitos colaterais e analisaram a probabilidade de dois fármacos compartilharem um mesmo alvo. Jung and Gudivada (1995) implementaram uma ferramenta baseada em redes neuronais artificiais, para construir um mapa de relacionamentos entre doenças pediátricas e seus sintomas a partir de uma coleção de documentos. Nessa ferramenta, eles implementaram uma teoria da psicologia chamada *teoria da construção pessoal, personal construct theory*. Li *et al.* (2009) propuseram um método que usa dados sobre genes, doenças e fármacos extraídos de resumos publicados no PubMed (PubMed, 2009), para construção de um mapa de interações moleculares. Entretanto, esses trabalhos não exploram a capacidade do modelo vetorial de recuperar informação a partir de coleções textuais. Dessa forma, eles não o empregam no processo de inferência com o objetivo de encontrar novas interações entre entidades biológicas a partir da literatura.

Quanto à avaliação e validação de resultados apresentados por algoritmos para descobertas baseadas em literatura, Kostoff (2008c,a) apresenta a importância de distinguirmos descoberta de inovação. O autor explica que as descobertas ocorrem quando encontramos algo previamente desconhecido ou não reconhecido. Por outro lado, a inovação reflete a mudança de uma prática usada correntemente para uma nova, supostamente melhor. Além disso, o autor discute sobre a quantidade e qualidade de descobertas potenciais. Nessa discussão, ele demonstra a relevância de processos rigorosos de avaliação, para assegurar a inexistência de trabalhos anteriores que relatem os resultados apontados pelos algoritmos para descobertas baseadas em literatura. Entretanto, o autor não esclarece que esse problema relaciona-se com a proporção de documentos da literatura que é coberta pela base de dados do sistema e não com a estratégia de descoberta e ordenação das interações.

Kostoff *et al.* (Kostoff *et al.*, 2008b) também apresentaram uma metodologia genérica para descobertas baseadas em literatura. Nesse trabalho, eles ainda defendem o uso de patentes como uma boa fonte de informação para se identificar interações já conhecidas. Eles usam essa metodologia para identificar interações relacionadas ao fenômeno de Raynaud (Kostoff *et al.*, 2008d), à doença dos olhos chamada catarata (Kostoff, 2008b), à doença de Parkinson (Kostoff and Briggs, 2008), à esclerose múltipla (Kostoff *et al.*, 2008c) e à purificação de água (Kostoff *et al.*, 2008e). Além disso, Kostoff *et al.* (Kostoff *et al.*, 2008a) compilaram as lições aprendidas nesses experimentos e apresentaram diretrizes para pesquisas futuras. Nesses trabalhos, os autores não usaram filtros numéricos para reduzir o número de novas interações e ordená-las. Portanto, um enorme esforço humano é necessário para examinar as

interações entre entidades.

Em nosso modelo, ordenamos as interações com base na similaridade especificada no modelo de espaço vetorial, com o propósito de reduzir o esforço humano ao examiná-las. Além disso, nessa fase do trabalho, o objetivo básico não é assegurar uma cobertura completa da literatura biológica. Ao invés disso, o objetivo é fornecer uma prova de conceito capaz de mostrar a capacidade de nosso modelo de descobrir e ordenar as interações. Por isso, usamos uma pequena coleção textual para avaliar o modelo. Estamos cientes de que muitas das novas interações inferidas através dessa coleção de documentos já foram apresentadas em trabalhos anteriores. Entretanto, usamos essas interações já apresentadas na literatura em nosso favor, empregando-as no processo de validação dos resultados.

Algoritmos de mineração de dados (Fan *et al.*, 2006; Tan *et al.*, 2002; Berendt *et al.*, 2002) e recuperação de informação (Salton and McGill, 1983; Baeza-Yates and Ribeiro-Neto, 1999; Witten *et al.*, 1999) fornecem meios importantes para extração de relacionamentos entre entidades biológicas a partir de coleções de documentos. Dessa forma, GuoDong and Min (2007) investigaram a incorporação de diversos conhecimentos léxicos, sintáticos e semânticos em uma estratégia para extração de relacionamentos usando *support vector machines* (SVM). Os autores trataram a extração de relacionamentos como um problema de classificação que emprega técnicas de *machine learning*. Além disso, eles também mostraram que informação semântica, como a disponibilizada na WordNet (WordNet, 2008), pode ser aplicada na tarefa de extração de relacionamentos. Analogamente, Wendlandt and Driscoll (1991) propuseram uma estratégia para melhorar o desempenho de recuperação de um sistema de recuperação de informação através de uma modelagem semântica. A estratégia considera a ocorrência de entidades do mundo real e seus relacionamentos em um coleção de documentos e emprega um conceito lingüístico conhecido como *papeis temáticos*, para reconhecer propriedades desses relacionamentos. As estratégias descritas nesses trabalhos não são aplicadas na construção de redes de interações. Além disso, eles não usam o modelo de espaço vetorial para extrair os relacionamentos da coleção textual. De modo contrário, Glenisson *et al.* (2003) usaram o modelo de espaço vetorial para criar agrupamentos de genes a partir de uma coleção de documentos. Eles também demonstraram o efeito da utilização de várias estratégias de pesos e fontes de informação na configuração funcional dos agrupamentos. Entretanto, esse trabalho também não é dedicado à construção de redes de interação.

Também é importante mostrar que a grande quantidade de dados biológicos publicados na *Web* tem motivado o desenvolvimento de algoritmos, para integração desses dados de forma a promover novos avanços em biologia. Tiffin *et al.* (2005),

por exemplo, desenvolveram um algoritmo que usa ontologia para relacionar os tecidos associados a doenças e os genes expressos nesses tecidos. As relações entre genes, tecidos e doenças são extraídas de resumos de artigos e documentos descrevendo perfis de expressão gênica. Gopalacharyulu *et al.* (2005) desenvolveram uma abordagem para mineração e integração de dados biológicos baseada em contexto. A abordagem usa a premissa de que o relacionamento entre entidades biológicas, como genes e organelas celulares, pode ser representado como uma rede complexa. Os autores adotaram uma métrica para atribuir pesos às conexões da rede que descreve a distância entre as entidades no espaço original. Karthikeyan *et al.* (2006) apresentaram um algoritmo distribuído para minerar automaticamente informação química a partir da Internet. Esse algoritmo procura por informação química em máquinas de busca e integra os resultados em um formato estruturado. Cheung *et al.* (2005) desenvolveram uma estratégia para resolver o problema da falta de padrão que dificulta a tarefa de integração de dados biológicos. Eles exploraram tecnologias como *resource description framework* (RDF) para representar, armazenar e pesquisar dados e meta-dados biológicos. Neshich *et al.* (2004), Wishart *et al.* (2006), Chen *et al.* (2002), Ihlenfeldt *et al.* (2002) e Hewett *et al.* (2002) integraram dados biológicos como fármacos, genes e proteínas minerados a partir de diversas fontes na *Web* e geraram bases de dados com importantes aplicações em biologia. Entretanto, esses trabalhos não consideram a integração dos dados biológicos, para construção de uma rede de interações. Além disso, eles restringem o número de categorias biológicas dos dados integrados e não exploram o modelo de espaço vetorial em suas implementações.

Existem várias fontes de informação biológica publicamente disponíveis na *Web* que podem ser usadas em estratégias de descoberta baseada em literatura e em integração de dados, como sítios de patentes, sítios de bulas de medicamentos e sítios de resumos de artigos científicos. Os pesquisadores comumente recorrem a patentes em seus trabalhos, porque elas possuem um grande valor como fonte de informação estratégica, técnica e relacionada a negócios (Lechter *et al.*, 1990; Tseng *et al.*, 2007). Trippe (2003) até descreve a *patinformática* como a ciência que analisa a informação contida em patentes com o objetivo de descobrir relacionamentos e tendências. Além disso, Shinmori *et al.* (2003) propuseram um *framework* para representar a estrutura da seção de reivindicação de patentes e um método para analisar automaticamente essa seção, uma vez que esta é a seção mais importante na especificação de uma patente. Mukherjea and Bamba (2004) desenvolveram um sistema para recuperar informação a partir de patentes biomédicas. O sistema identifica e classifica termos biológicos que ocorrem nas patentes e os integra através de conceitos encontrados

em dicionários biomédicos. Larkey (1999) descreveu o sistema de recuperação e classificação de patentes desenvolvido para o USPTO. Fall *et al.* (2003) verificou os melhores meios de lidar com a classificação de patentes e apresentou uma comparação entre o desempenho de classificação de vários algoritmos. Contudo, esses trabalhos não se dedicam à construção de redes de interação e não implementam processos de inferência. Tseng *et al.* (2007) descreveu e avaliou diversas técnicas de mineração de dados, para automatizar o processo de criação de mapas de patentes. Esses mapas são usados para melhorar tarefas de análise das patentes, como classificação, organização, compartilhamento de conhecimento e consultas sobre invenções já patenteadas. No entanto, esse trabalho não emprega o modelo de espaço vetorial na construção de seus mapas de patentes e não implementa processos de inferência.

## 1.2 Contribuição

Neste trabalho, criamos um modelo para construir uma rede complexa de interações entre entidades biológicas a partir de coleções de documentos que representam o conhecimento em biologia, como patentes, artigos científicos e bulas de medicamentos. Nosso modelo combina o modelo de espaço vetorial com uma relação de transitividade, dando origem a um processo de inferência que indica novas interações entre entidades biológicas. A rede é formada por sub-redes de interações entre entidades de categorias biológicas distintas. A vantagem de usar as categorias é aplicá-las na criação de espaços  $n$ -dimensionais. Esses espaços dimensionais restringem o espaço de pesquisa das interações entre entidades e promovem resultados mais apurados.

As interações entre entidades são inicialmente estabelecidas na rede quando procuramos pela ocorrência dessas entidades na coleção textual. Essas interações representam as interações conhecidas que são descritas na coleção de documentos. As interações conhecidas recebem um valor que representa a evidência de interação entre as entidades, conforme indicado pela coleção de documentos. Esse valor é determinado com base no modelo de espaço vetorial. A vantagem de usar o modelo de espaço vetorial é explorar seu poder de recuperar informação a partir de coleções textuais, com o objetivo de encontrar as ocorrências das entidades nos documentos e calcular a evidência de interação entre elas.

Nosso modelo usa as interações estabelecidas na rede para indicar interações novas através da relação de transitividade que avaliamos no processo de inferência. Nossa relação de transitividade explora as atividades principais e secundárias que as entidades exercem no sistema biológico, assumindo que "IF uma entidade  $x$  in-

terage com as entidades  $y$  e  $w$  **AND** uma outra entidade  $z$  também interage com a entidade  $y$ , **THEN**  $z$  possivelmente também interage com a entidade  $w$ ". A nova interação também recebe um valor que mensura a evidência de interação entre as entidades que ela relaciona. O valor da evidência de interação de uma interação nova é determinado com base na evidência de interação das interações que satisfazem a condição imposta pela relação de transitividade. Dessa forma, podemos ordenar as interações estabelecida na rede e observar os resultados mais promissores indicados pelo modelo.

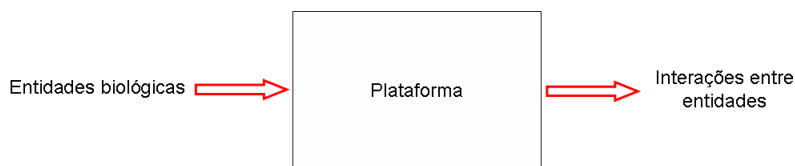
Nós implementamos nosso modelo em um sistema chamado *BioSearch* (BioSearch, 2009). O sistema é um repositório de interações entre entidades biológicas em que usuários podem pesquisar interações conhecidas e novas entre essas entidades. Consultando as interações conhecidas, os usuários podem estudar o que tem sido publicado na literatura biológica e guiar suas pesquisas. Por outro lado, consultando as interações novas, os usuários podem analisar as interações mais promissoras para suas pesquisas.

Nós realizamos nossos experimentos considerando uma coleção textual formada pela seção de reivindicações de patentes coletadas no sítio *Web* do USPTO. Esses experimentos demonstram que o modelo é capaz de restringir os melhores resultados, indicando as interações biológicas mais relevantes. Os experimentos mostram ainda que a literatura de patentes é uma boa fonte de informação para o descobrimento de novas interações entre entidades biológicas. Os testes de validação evidenciam que muitas interações inferidas pelo modelo em um ano foram confirmadas por patentes publicadas em anos subseqüentes. Esses testes também mostram que muitas patentes de confirmação foram encontradas no topo dos *rankings* de resposta de cada sub-rede. Além disso, encontramos artigos científicos publicamente disponíveis na *Web* que confirmam interações novas apontadas pelo modelo.

## Capítulo 2

# O Sistema BioSearch

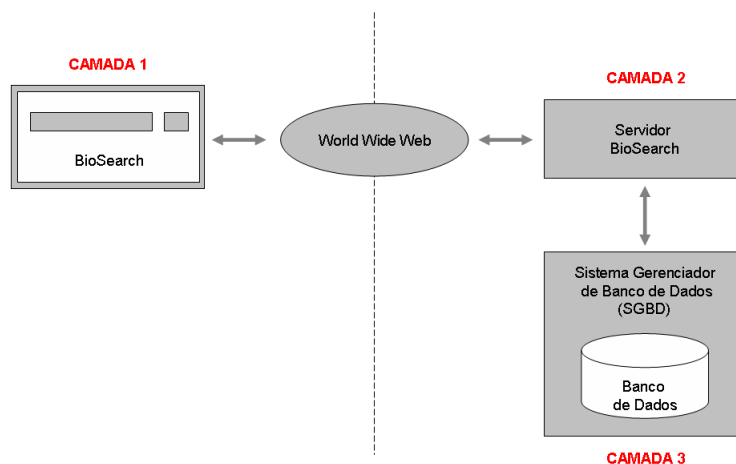
Em nosso trabalho desenvolvemos um modelo de inferência, para construir uma rede de interações entre entidades de um sistema biológico. Nós implementamos esse modelo de inferência em um sistema chamado *BioSearch* cuja entrada é um conjunto de entidades biológicas agrupadas em categorias (Figura 2.1). O sistema usa essa entrada para coletar na *Web* uma coleção de documentos em que essas entidades ocorrem. O sistema utiliza a coleção de documentos para estabelecer associações já conhecidas entre entidades de categorias biológicas distintas. Essas associações são empregadas na construção de uma rede de interações que representa as interações entre entidades no sistema biológico. As interações conhecidas recebem um valor que mede a evidência de interação entre as entidades com base na coleção de documentos. Além disso, as interações entre entidades de categorias distintas permitem dividir a rede em sub-redes caracterizadas pelas categorias que as constituem. A partir das interações em cada sub-rede, o sistema infere novas interações entre as entidades e atribui uma evidência de interação a essas novas interações. A informação de saída do sistema é uma lista ordenada de interações entre entidades biológicas. A lista é ordenada em ordem decrescente da evidência de interação que quantifica o relacionamento entre as entidades nas sub-redes.



**Figura 2.1:** Representação da entrada e saída do sistema *BioSearch*.

Além da construção da rede biológica, outra função importante do sistema é permitir que usuários consultem as interações conhecidas e novas entre entidades a partir da *Web*. Para isso, o sistema é desenvolvido em uma arquitetura cliente-

servidor *multi-thread* de 3 camadas (Coulouris *et al.*, 2005), com o objetivo de separar o módulo que implementa a interface de interação com usuários, os módulos que implementam a lógica do modelo de inferência e o sistema gerenciador de banco de dados (Figura 2.2). Na primeira camada da arquitetura está a interface cliente para *Web* que permite a interação dos usuários com o sistema. Na segunda camada está o servidor do sistema que implementa a lógica do modelo de inferência. Na terceira camada está o sistema gerenciador de banco de dados onde é armazenada toda a rede biológica criada pelo sistema.



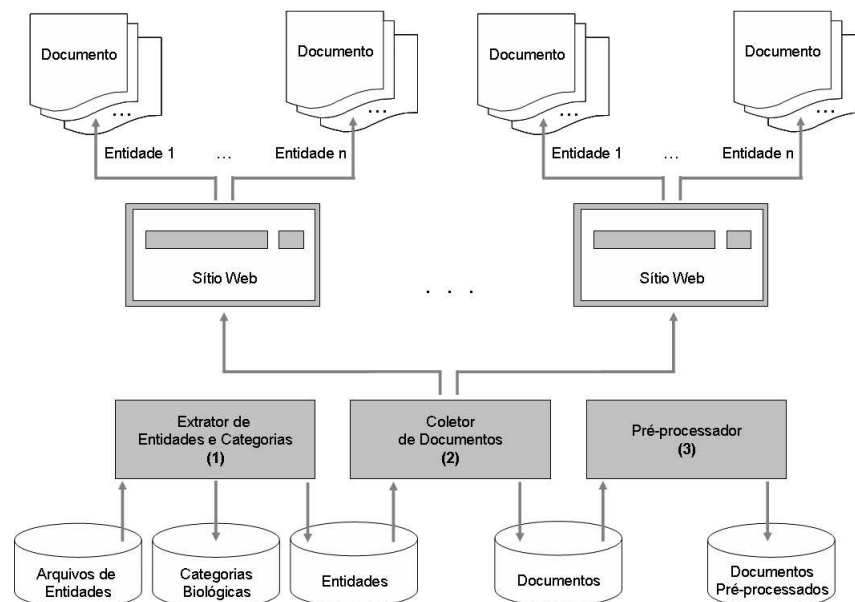
**Figura 2.2:** Representação da arquitetura cliente-servidor em 3 camadas do sistema *BioSearch*.

## 2.1 Os Módulos do Sistema

Os módulos de nosso sistema são agrupados de acordo com 4 tarefas distintas. A primeira tarefa do sistema é a coleta da coleção de documentos que descreve o conhecimento em biologia. A segunda tarefa é a indexação dessa coleção de documentos. A terceira tarefa é a construção da rede biológica, para inferência de novas interações entre entidades. Por fim, a quarta tarefa é a consulta das interações da rede biológica pelos usuários através da *Web*.

Para o desenvolvimento e análise dos algoritmos empregados em cada tarefa de nosso sistema, consideremos o conjunto  $C = \{c_1, c_2, \dots, c_{n_C}\}$  de categorias do sistema biológico, o conjunto  $E = \{e_1, e_2, \dots, e_{n_E}\}$  das entidades do sistema biológico pertencentes a cada categoria  $c_i \in C$ , o conjunto  $Q = \{q_1, q_2, \dots, q_{n_Q}\}$  de consultas formadas pela conjunção de entidades em  $E$  e o conjunto de documentos  $D = \{d_1, d_2, \dots, d_{n_D}\}$  que constitui a coleção de documentos de nosso modelo de inferência e descreve o conhecimento em biologia. Nesses conjuntos,  $n_C$ ,  $n_E$ ,  $n_Q$  e  $n_D$  são as cardinalidades de  $C$ ,  $E$ ,  $Q$  e  $D$  respectivamente.





**Figura 2.3:** Tarefa 1: coleta dos documentos que descrevem o conhecimento em biologia. O sistema forma uma coleção de documentos a partir de entidades biológicas e suas categorias.

### 2.1.1 A Tarefa de Coleta

O objetivo da tarefa de coleta é formar uma coleção de documentos que descreva o conhecimento em biologia na base de dados do sistema. Essa coleção de documentos deve apresentar interações já conhecidas entre as entidades biológicas fornecidas como entrada para o sistema, para que a rede de interações possa ser construída nas tarefas subsequentes do sistema. A tarefa de coleta possui 3 módulos (Figura 2.3) e inicia a partir de arquivos em modo texto contendo nomes de entidades biológicas, o sítio na *Web* em que essas entidades foram encontradas e suas respectivas categorias, como alvo biológico, doença, fármaco e gene.

#### A Identificação de Entidades e Categorias Biológicas

O módulo *extrator de entidade e categorias* (módulo 1) é o responsável por ler os arquivos de entidades fornecidos como entrada para o sistema e extrair as entidades que neles ocorrem. Em seguida, o módulo armazena as entidades na base de dados, agrupando-as por categoria biológica e pelo sítio *Web* onde foram encontradas. Dessa forma, esse módulo é o responsável por formar o conjunto de categorias  $C$  e o conjunto de entidades  $E$ .

Para analisarmos o custo de formação dos conjuntos  $C$  e  $E$ , consideremos que, na entrada do sistema, temos um arquivo de entidades para cada categoria biológica separadamente. Assim, temos  $n_C$  arquivos de entrada. Consideremos também que o número de entidades em cada categoria é  $n_E/n_C$ . Dessa forma, temos que o custo

de formar os conjuntos  $C$  e  $E$  é da ordem  $O(n_C \times n_E/n_C) = O(n_E)$ , pois para cada categoria em  $C$  o sistema precisa ler  $n_E/n_C$  entidades dessa categoria (Algoritmo 2.1).

---

**Algorithm 2.1** Módulo 1: extração de entidades e categorias biológicas.

---

**procedure** ExtrairEntidadesCategorias ()

**begin procedure**

- 1 **for each** arquivo de categoria  $c_i \in C$  contendo entidades biológicas tal que  $1 \leq i \leq n_C$  **do**
- 2     armazene na base de dados a categoria  $c_i \in C$ , formando o conjunto  $C$
- 3     extraia as entidades presentes no arquivo de entidades da categoria  $c_i$
- 4     armazene na base de dados as entidades encontradas, formando o conjunto  $E$
- 5     armazene na base de dados os sítios  $Web$  em que a categoria  $c_i$  foi encontrada
- 6 **end for**

**end procedure**

---

## A Formação da Coleção de Documentos

As entidades na base de dados formam um *log* de consulta que é lido pelo módulo *coletor de documentos* (módulo 2). Esse módulo coletor é responsável por submeter cada entidade do *log* de consultas em sítios *Web* públicos que contenham documentos descrevendo o conhecimento em biologia e que relatem interações das entidades pesquisadas, como sítios de patentes, sítios de bulas de medicamentos e sítios contendo resumos de artigos científicos. O módulo coletor engloba módulos específicos para fazer a interface com cada um dos sítios *Web* considerados na formação da coleção de documentos. Os sítios *Web* retornam uma lista de documentos relevantes para cada entidade pesquisada. Os documentos dessa lista são armazenados na base de dados pelo módulo coletor, formando a coleção de documentos  $D$ .

Para analisarmos o custo de coletar a coleção de documentos, consideremos que o número  $k_1$  de sítios *Web* pesquisados é pequeno e sempre constante. Consideremos ainda que o número de documentos retornados para cada entidade é determinado por  $k_2 \times n_D$ , sendo  $k_2$  um número no intervalo  $[0, 1]$  que indica a proporção média de documentos na base de dados retornados para cada entidade nos sítios *Web* de pesquisa. Assim, temos que o custo de coletar o conjunto  $D$  é da ordem  $O(n_E \times k_1 \times k_2 \times n_D) = O(n_E \times n_D)$ , pois, para cada entidade em  $E$ , o sistema coleta  $k_2 \times n_D$  documentos em cada um dos  $k_1$  sítios *Web* de pesquisa (Algoritmo 2.2).

## O Pré-processamento da Coleção de Documentos

Todos os documentos da coleção são pré-processados, gerando visões dos documentos originais (Baeza-Yates and Ribeiro-Neto, 1999). Essas visões são textos

---

**Algorithm 2.2** Módulo 2: coleta de documentos.

---

```

procedure ColetarDocumentos ()
begin procedure
1  for each entidade  $e_i \in E$  tal que  $1 \leq i \leq n_E$  do
2    for each sítio Web considerado para a formação da coleção de documentos do
3      submeta  $e_i$  no sítio Web
4      colete os documentos retornados pelo sítio Web
5      armazene os documentos retornados pelo sítio Web na base de dados, formando  $D$ 
6    end for
7  end for
end procedure

```

---

padronizados que serão usados nos módulos das demais tarefas do sistema. O módulo *pré-processador* (módulo 3) lê cada documento da coleção e executa operações comuns a documentos de todos os sítios *Web* de pesquisa, como eliminação de TAGs HTML e conversão de todos os caracteres para minúsculos, e também operações específicas, como extrair a seção de reivindicação de patentes. Os documentos pré-processados são armazenados na base de dados e constituem a versão final da coleção de documentos que será usada pelos outros módulos do sistema. O custo de obter a coleção de documentos pré-processada é da ordem  $O(n_D)$  (Algoritmo 2.3).

---

**Algorithm 2.3** Módulo 3: pré-processamento dos documentos.

---

```

procedure PreProcessarDocumentos ()
begin procedure
1  for each documento  $d_i \in D$  tal que  $1 \leq i \leq n_D$  do
2    Extraia de  $d_i$  a seção  $s$  de interesse
3    Elimine TAGs HTML de  $s$ 
4    Elimine de  $s$  caracteres especiais diferentes de letras e números
5    Converta os caracteres de  $s$  para caracteres minúsculos
6    armazene o texto pré-processado de  $s$  na base de dados
7  end for
end procedure

```

---

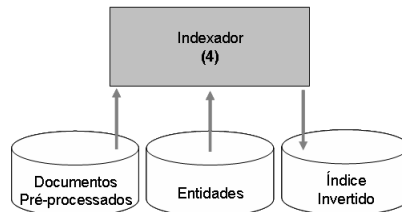
### O Custo da Tarefa de Coleta

O custo da tarefa de coleta é determinado pela soma dos custos de seus 3 módulos. O custo de formar os conjuntos  $C$  e  $E$  é da ordem  $O(n_E)$ . O custo de coletar os documentos do conjunto  $D$  é da ordem  $O(n_E \times n_D)$ . O custo de obter a coleção de documentos pré-processada é da ordem  $O(n_D)$ . Assim, o custo da tarefa de coleta é da ordem  $O(n_E \times n_D)$  (Equação 2.1).

$$\begin{aligned}
O(n_E + (n_E \times n_D) + n_D) &= O(n_E \times (1 + n_D) + n_D) \\
&= O((n_E \times n_D) + n_D) \\
&= O((n_E + 1) \times n_D) \\
&= O(n_E \times n_D)
\end{aligned} \tag{2.1}$$

### 2.1.2 A Tarefa de Indexação da Coleção de Documentos

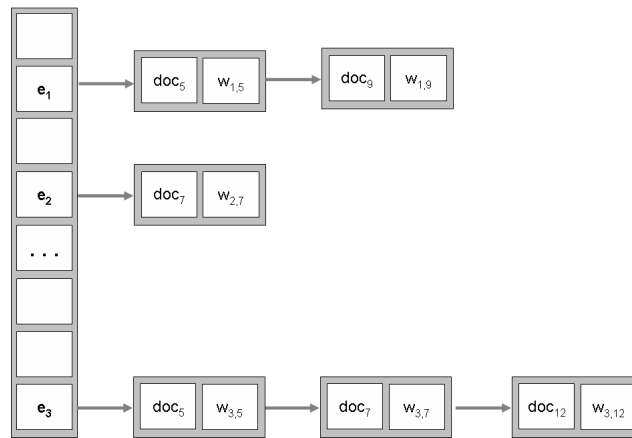
A tarefa de indexação possui apenas um módulo chamado *indexador* (Figura 2.4) cuja função é construir um *índice invertido* (Witten *et al.*, 1999) que descreve a localização das entidades biológicas na coleção de documentos (Figura 2.5). O objetivo de criar o índice invertido é sumarizar a ocorrência das entidades, tornando as buscas necessárias no modelo de espaço vetorial mais eficientes (Salton and McGill, 1983; Baeza-Yates and Ribeiro-Neto, 1999; Witten *et al.*, 1999).



**Figura 2.4:** Tarefa 2: indexação da coleção de documentos que descreve o conhecimento em biologia.

O índice invertido é uma tabela *hash* (Cormen *et al.*, 2001; Ziviani, 2007) em que cada entrada é usada pelo indexador para armazenar uma entidade biológica. Esta entidade biológica é também chamada de chave da entrada que a armazena na tabela *hash*. A entrada da tabela *hash* que recebe uma entidade aponta para uma lista encadeada (Cormen *et al.*, 2001; Ziviani, 2007) com a localização das ocorrências da entidade na coleção de documentos. Cada nodo da lista encadeada armazena um documento em que a entidade ocorre e o peso da entidade para esse documento. Assim, os nodos da lista são gerados para armazenar um par ordenado do tipo  $\langle \text{identificador}, \text{peso} \rangle$ . O identificador é um número que identifica cada documento na coleção. O peso é calculado através da estratégia de peso TFIDF do modelo de espaço vetorial e corresponde ao peso da entidade biológica no documento.

O módulo indexador é responsável por encontrar e contar as ocorrências das entidades em cada documento e na coleção de documentos. A partir da contabilização das ocorrências das entidades é que esse módulo calcula os pesos das entidades para cada documento. O módulo armazena as entidades, os documentos em que as entidades ocorrem e os pesos das entidades para cada documento no índice invertido.



**Figura 2.5:** Índice invertido. Nesse índice,  $w_{3,7}$ , por exemplo, é o peso da entidade  $e_3$  no documento  $d_7$ .

Assim, consideremos uma coleção formada pelos documentos  $d_1$ ,  $d_2$ ,  $d_3$ ,  $d_4$ ,  $d_5$  e  $d_6$  (Figura 2.6). Consideremos ainda que as entidades que ocorrem nessa coleção de documentos são  $e_1$ ,  $e_2$ ,  $e_3$ ,  $e_4$ ,  $e_5$ ,  $e_6$  e  $e_7$ . Para calcularmos o peso de  $e_3$  no documento  $d_1$ , por exemplo, precisamos **(a)** identificar a entidade mais freqüente de  $d_1$  e **(b)** calcular a freqüência dessa entidade. Em seguida, **(c)** calculamos a freqüência de  $e_3$  em  $d_1$  e **(d)** determinamos o valor do fator TF, dividindo a freqüência de  $e_3$  pela freqüência da entidade que mais ocorre em  $d_1$ . O passo seguinte é **(e)** determinar o número de documentos da coleção em que  $e_3$  ocorre e o **(f)** número total de documentos da coleção, para **(g)** computarmos o valor do fator IDF. Multiplicando os fatores TF e IDF, **(h)** determinamos o peso de  $e_3$  em  $d_1$ . Por fim, **(i)** inserimos o nodo  $\langle d_1, w_{3,1} \rangle$  no índice invertido, usando  $e_3$  como chave. Depois de criar o índice invertido, o sistema pode pesquisá-lo para encontrar os documentos relacionados a uma dada consulta que representa a interação entre entidades em uma sub-rede.

### O Custo da Tarefa de Indexação da Coleção de Documentos

O custo da tarefa de indexação da coleção de documentos corresponde ao custo de criar o índice invertido. O custo de criação do índice invertido é calculado com base no número  $n_D$  de documentos da coleção e no número  $n_E$  de entidades fornecidas como entrada para o sistema. Em uma passada pela coleção de documentos, determinamos a freqüência das entidades em cada documento, a entidade mais freqüente de cada documento e também o número de documentos em que cada entidade ocorre. Essa passada pela coleção de documentos gera um custo da ordem  $O(n_D \times n_E)$ . Em seguida, para cada documento da coleção, calculamos o peso de cada entidade e geramos o nodo de cada entidade no índice invertido, o que gera outro custo da ordem  $O(n_D \times n_E)$  (Algoritmo 2.4). Assim, o custo de gerar o índice invertido é da ordem

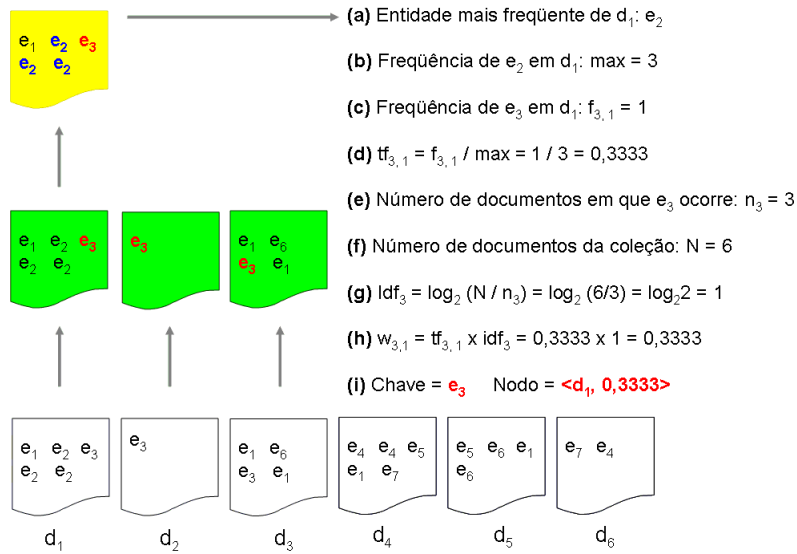


Figura 2.6: Cálculo do peso das entidades na coleção de documentos através da estratégia TFIDF.

$O(n_D \times n_E)$  (Equação 2.2).

$$\begin{aligned}
 O((n_D \times n_E) + (n_D \times n_E)) &= O(2 \times n_D \times n_E) \\
 &= O(n_D \times n_E)
 \end{aligned} \tag{2.2}$$

---

**Algorithm 2.4** Módulo 4: construção do índice invertido.

---

**procedure** ConstruirIndiceInvertido ()

**begin procedure**

```

1   $N \leftarrow n_D$ 
2  for each documento  $d_i \in D$  tal que  $1 \leq i \leq n_D$  do
3     $max \leftarrow$  número de ocorrências da entidade mais freqüente de  $d_i$ 
4    for each entidade  $e_j \in E$  tal que  $1 \leq j \leq n_E$  do
5       $f \leftarrow$  número de ocorrências de  $e_j$  em  $d_i$ 
6       $tf \leftarrow f / max$ 
7       $n \leftarrow$  número de documentos de  $D$  em que  $e_j$  ocorre
8       $idf \leftarrow \log(N/n)$ 
9       $w \leftarrow tf \times idf$ 
10      $nodo \leftarrow$  par ordenado  $\langle d_i, w \rangle$ 
11     insira  $nodo$  no índice invertido usando  $e_j$  como chave na tabela hash
12   end for
13 end for

```

**end procedure**

---

### 2.1.3 A Tarefa de Construção da Rede Biológica

Em nosso trabalho, procuramos por interações ainda não conhecidas entre entidades de um sistema biológico e que podem ser reveladas pela análise de documentos que compõem a literatura que descreve o conhecimento em biologia (Swanson, 1990; Smalheiser and Swanson, 1998; Swanson *et al.*, 2006; Weeber *et al.*, 2001; Hristovski *et al.*, 2006; Wren *et al.*, 2004; Bruza and Weeber, 2008; Campillos *et al.*, 2008). Para tanto, na tarefa de construção da rede biológica os objetivos são formar os espaços dimensionais das sub-redes, gerar interações possíveis entre as entidades biológicas, encontrar interações conhecidas entre essas entidades a partir da coleção de documentos, construir as sub-redes com as interações conhecidas e, então, inferir interações novas a partir das interações já estabelecidas nas sub-redes.

#### A Formação dos Espaços Dimensionais

A tarefa de construção da rede engloba 5 módulos (Figura 2.7) e inicia com a formação dos espaços n-dimensionais que caracterizam cada sub-rede. Essa função fica a cargo do módulo *combinador de categorias* (módulo 5) que lê as categorias biológicas a partir da base de dados e as combina, para formar os espaços n-dimensionais que então são armazenados na base de dados.

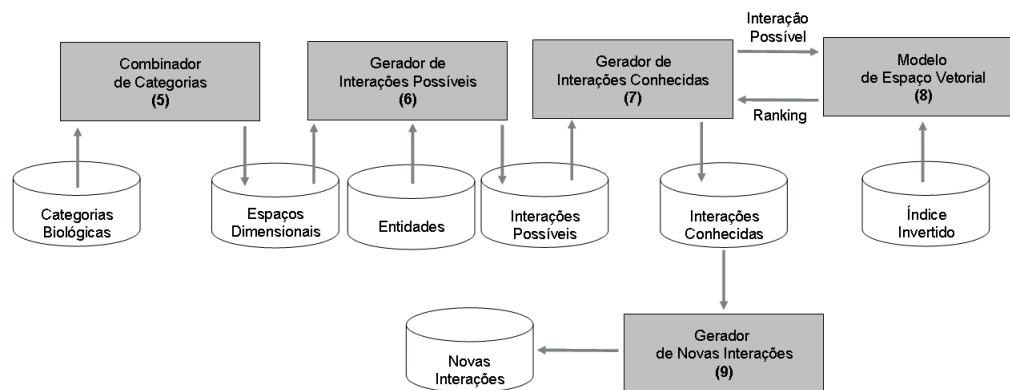


Figura 2.7: Tarefa 3: construção da rede biológica.

Nossa implementação do módulo combinador de categorias fundamenta-se no paradigma da programação dinâmica (Cormen *et al.*, 2001; Ziviani, 2007), com o objetivo de gerar espaços dimensionais a partir de espaços dimensionais previamente gerados (Algoritmo 2.5). Nessa implementação nós usamos um vetor  $v$  para armazenamento das categorias de  $C$  e 4 filas FIFO (*First In First Out*): *temp*, *nodo*, *subrede* e *rede*. A fila *temp* é usada para armazenar temporariamente uma categoria do vetor  $v$  ou um espaço dimensional da rede. A fila *nodo* é usada para

---

**Algorithm 2.5** Módulo 5: formação dos espaços n-dimensionais.

---

**function** GerarEspaçosDimensionais ()

**begin function**

```

1  seja  $v$  um vetor com as  $n_C$  categorias biológicas consideradas no sistema
2  sejam  $temp$ ,  $nodo$ ,  $subrede$  e  $rede$  4 filas (FIFO) inicialmente vazias
3
4  for  $i = 1$  to  $n_C$  do
5    enfileirar ( $temp$ ,  $v[i]$ ) {Enfileira o elemento da posição  $i$  de  $v$  em  $temp$ .}
6  end for
7  while  $temp$  não estiver vazia do
8     $nodo \leftarrow$  desenfileirar ( $temp$ ) {Retira o primeiro elemento de  $temp$  e o atribui a  $nodo$ .}
9     $i \leftarrow$  índice da última categoria enfileirada em  $nodo$ 
10   for  $j = i + 1$  to  $n_C$  do
11      $subrede \leftarrow$   $nodo$  {Substitui o conteúdo de  $subrede$  pelo conteúdo de  $nodo$ .}
12     enfileirar ( $subrede$ ,  $v[j]$ ) {Enfileira o elemento da posição  $j$  de  $v$  em  $subrede$ .}
13     enfileirar ( $temp$ ,  $subrede$ ) {Enfileira  $subrede$  em  $temp$ .}
14     enfileirar ( $rede$ ,  $subrede$ ) {Enfileira  $subrede$  em  $rede$ .}
15   end for
16 end while
17 return  $rede$ 

```

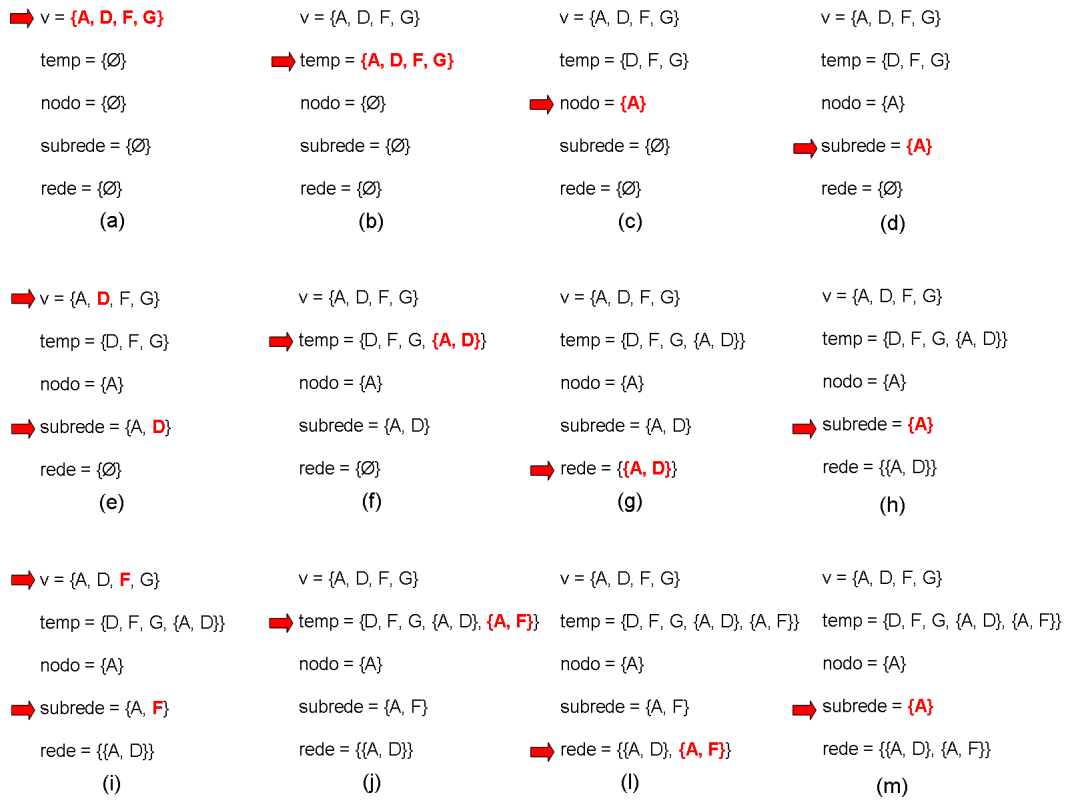
**end function**

---

armazenar elementos desenfileirados da fila  $temp$  e que darão origem a novos espaços dimensionais. Cada elemento armazenado na fila  $subrede$  representa um novo espaço dimensional da rede. Os elementos armazenados na fila  $subrede$  são enfileirados na fila  $temp$  para dar origem à novos espaços dimensionais e são enfileirados também na fila  $rede$ . A fila  $rede$  é a saída do combinador de categorias e armazena todos os espaços dimensionais possíveis da rede.

Para exemplificarmos o funcionamento do módulo combinador de categorias, consideremos a formação dos espaços dimensionais de uma rede a partir das categorias alvo biológico (**A**), doença (**D**), fármaco (**F**) e gene (**G**) (Figura 2.8). O combinador de categorias inicia a formação dos espaços dimensionais lendo as categorias de  $C$  a partir da base de dados e armazenando-as em  $v$  (Figura 2.8 (a)). Os elementos de  $v$  são enfileirados em  $temp$  (Figura 2.8 (b)). Então, o primeiro elemento de  $temp$  é desenfileirado e atribuído à  $nodo$  (Figura 2.8 (c)). A fila  $subrede$  recebe o conteúdo de  $nodo$  (Figura 2.8 (d)). A partir do vetor  $v$ , cada elemento subsequente à última categoria em  $subrede$  é também enfileirada em  $subrede$  (Figura 2.8 (e)). Dessa forma, a cada elemento de  $v$  enfileirado em  $subrede$ , o módulo combinador de categoria determina um novo espaço dimensional. Após determinar um espaço dimensional, o módulo enfileira o conteúdo de  $subrede$  em  $temp$  e em  $rede$  (Figura

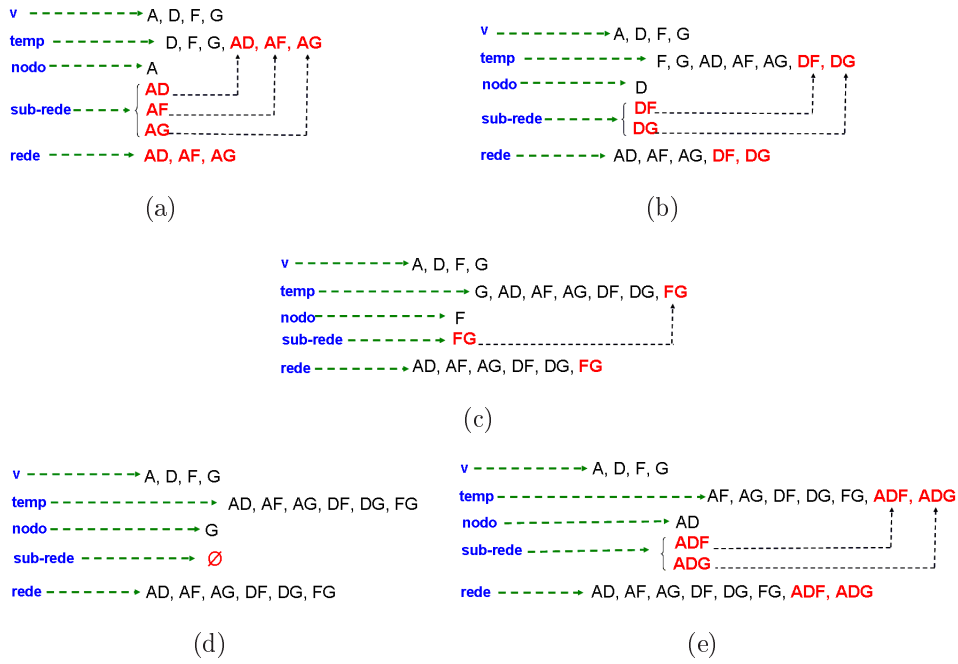




**Figura 2.8:** Representação da formação de espaços dimensionais a partir das categorias alvo biológico (**A**), doença (**D**), fármaco (**F**) e gene (**G**).

2.8 (**f e g**)). Então, *subrede* volta a receber o conteúdo de *nodo* e o processo continua até que todos os elementos de  $v$  tenham sido enfileirados em *subrede* (Figura 2.8 (**h-m**)). Em seguida, um novo elemento é desenfileirado de *temp* e atribuído a *nodo*. O combinador de categorias termina de gerar os espaços dimensionais quando a fila *temp* se torna vazia.

O combinador de categorias inicia a formação dos espaços dimensionais lendo as categorias de  $C$  a partir da base de dados e armazenando-as no vetor  $v$ . O custo de gerar o vetor  $v$  com as categorias de  $C$  é da ordem  $O(n_c)$ . Em seguida, os elementos do vetor  $v$  são enfileirados na fila *temp* com um custo também da ordem  $O(n_c)$ . Então, o primeiro elemento da fila *temp* é desenfileirado e atribuído à fila *nodo*. A fila *subrede* recebe o conteúdo de *nodo* e, a partir desse ponto, inicia a formação das sub-redes bidimensionais. Através de iterações sucessivas, os elementos do vetor  $v$  são enfileirados um a um em *subrede* (Figura 2.9). Assim, o custo de gerar as sub-redes bidimensionais é dado por  $O(\frac{n_c!}{2! \times (n_c - 2)!})$ . Quando todas as sub-redes bidimensionais são formadas, inicia a formação das sub-redes tridimensionais com um custo da ordem  $O(\frac{n_c!}{3! \times (n_c - 3)!})$ . Esse processo continua até que sejam formadas as sub-redes  $n_C$ -dimensionais com custo da ordem  $O(\frac{n_c!}{n_c! \times (n_c - n_c)!})$ . Então, o custo de gerar todas as sub-redes é dado por  $O(\sum_{i=2}^{n_C} \frac{n_c!}{i! \times (n_c - i)!})$ . Dessa forma, o custo do



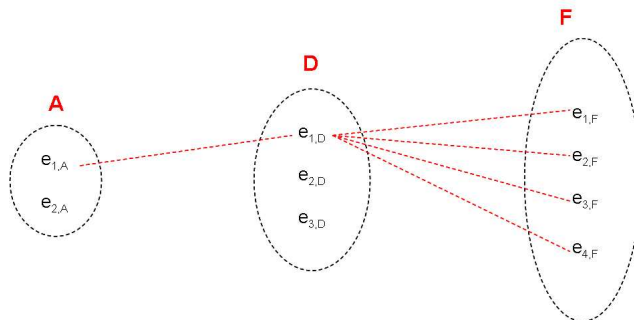
**Figura 2.9:** Representação das iterações para formação de espaços bidimensionais e tridimensionais a partir das categorias alvo biológico (**A**), doença (**D**), fármaco (**F**) e gene (**G**).

módulo combinador de categorias é da ordem  $O(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!})$  (Equação 2.3).

$$\begin{aligned}
O(n_C + n_C + \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!}) &= O(2 \times n_C + \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!}) \\
&= O(n_C + \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!}) \\
&= O(n_C + \sum_{i=2}^{n_C} \frac{n_C \times (n_C - 1)!}{i! \times (n_C - i)!}) \\
&= O(n_C + n_C \times \sum_{i=2}^{n_C} \frac{(n_C - 1)!}{i! \times (n_C - i)!}) \\
&= O(n_C \times (1 + \sum_{i=2}^{n_C} \frac{(n_C - 1)!}{i! \times (n_C - i)!})) \\
&= O(n_C \times \sum_{i=2}^{n_C} \frac{(n_C - 1)!}{i! \times (n_C - i)!}) \\
&= O(\sum_{i=2}^{n_C} \frac{n_C \times (n_C - 1)!}{i! \times (n_C - i)!}) \\
&= O(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!}) \tag{2.3}
\end{aligned}$$

## As Interações Entre Entidades Biológicas

A rede criada em nosso sistema é formada por sub-redes de interações entre entidades biológicas. Essas interações podem ser de três tipos: *possíveis*, *conhecidas* e *novas*. As interações possíveis de uma sub-rede correspondem a todas as tuplas do espaço n-dimensional dessa sub-rede. Essas tuplas são geradas pelo produto cartesiano entre as entidades das categorias que formam o espaço n-dimensional (Figura 2.10). Dessa forma, o número de tuplas de um espaço n-dimensional indica o número de interações possíveis entre as entidades das categorias que constituem a sub-rede. Esse número é o produto das cardinalidades das categorias que formam o espaço n-dimensional. Por outro lado, as interações conhecidas são aquelas cujas entidades co-ocorrem em documentos da coleção. Por fim, as interações novas são aquelas inferidas a partir das interações conhecidas ou de interações previamente inferidas.



**Figura 2.10:** Representação do produto cartesiano para formação das interações possíveis entre as entidades de uma sub-rede cujo espaço dimensional é formado pelas categorias alvo biológico (**A**), doença (**D**) e fármaco (**F**).

## As Interações Possíveis

Para gerar as interações possíveis, o módulo *gerador de interações possíveis* (módulo 6) lê os espaços dimensionais e suas entidades a partir da base de dados e realiza o produto cartesiano entre as entidades. Nossa implementação do módulo gerador de interações possíveis também fundamenta-se no paradigma da programação dinâmica. Assim, usamos interações possíveis geradas num passo do algoritmo para dar origem a outras interações possíveis da rede biológica. Nessa implementação, usamos um procedimento para iniciar a formação das interações possíveis que é o responsável por ler as entidades de cada sub-rede (Algoritmo 2.6). Usamos também um procedimento recursivo para realizar o produto cartesiano entre as entidades (Algoritmo 2.7).

Para fazer a leitura das entidades de cada sub-rede, consideramos uma pilha FILO (*First In Last Out*) produto para armazenamento de cada resultado interme-

---

**Algorithm 2.6** Módulo 6: formação das interações possíveis entre entidades.

---

```

function GerarInteraçõesPossíveis ()
begin function
  1 seja produto uma pilha (FILO) inicialmente vazia
  2 sejam rede, subrede, categoria, entidades e resultado quatro filas (FIFO) inicialmente vazias
  3 seja conjunto um vetor inicialmente vazio cujas posições armazenam filas (FIFO)
  4
  5 rede ← todos os espaços n-dimensionais formados na rede biológica
  6 while rede não estiver vazia do
  7   subrede ← desenfileirar (rede) {Retira o primeiro elemento de rede e o atribui a subrede.}
  8   tamanho ← 1
  9   while subrede não estiver vazia do
 10    categoria ← desenfileirar (subrede) {Retira o primeiro elemento de subrede e o atribui
      a categoria.}
 11    entidades ← todas as entidades de categoria
 12    conjunto[tamanho] ← entidades
 13    tamanho ← tamanho + 1
 14  end while
 15  ProdutoCartesiano (conjunto, resultado, produto, tamanho, 1)
 16  for i = 1 to tamanho do
 17    conjunto[i] ← ∅
 18  end for
 19 end while
 20 return resultado
end function

```

---

diário do produto cartesiano e 5 filas FIFO: *rede*, *subrede*, *categoria*, *entidades* e *resultado*. Nós usamos a fila *rede* para armazenar os espaços dimensionais que integram a rede biológica (Figura 2.11). Por outro lado, usamos a fila *subrede* para receber os elementos desenfileirados da fila *rede*. A fila *categoria* armazena cada uma das categorias armazenadas em *subrede* e a fila *entidades* armazena as entidades biológicas dessas categorias. O conteúdo da fila *entidades* é armazenado no vetor *conjunto*, formando sub-conjuntos que serão usados na geração do produto cartesiano. A fila *resultado* é usada para armazenar o resultado obtido no produto cartesiano entre as entidades. Esse resultado do produto cartesiano corresponde às interações possíveis da rede.

O vetor *conjunto*, a fila *resultado* e a pilha *produto* são passadas como parâmetro para o procedimento recursivo. Nesse procedimento, consideramos ainda uma fila FIFO *subconjunto* para armazenar cada sub-conjunto de entidades do vetor *conjunto* e um vetor *entidades*. A cada iteração recursiva do procedimento, as interações entre entidades vão sendo formadas, enfileirando-se cada entidade do vetor

---

**Algorithm 2.7** Módulo 6: produto cartesiano das entidades em cada sub-rede.

---

**procedure** ProdutoCartesiano (*conjunto*, *resultado*, *produto*, *tamconjunto*, *n*)

**parâmetros de entrada** passados por referência:

*conjunto* um vetor cujas posições armazenam filas (FIFO) de entidades

*resultado* uma fila (FIFO)

*produto* uma pilha (FILO)

**parâmetros de entrada** passados por valor:

*tamconjunto* um inteiro

*n* um inteiro

**begin procedure**

1 seja *subconjunto* uma fila (FIFO) inicialmente vazia

2 seja *entidades* um vetor inicialmente vazio

3

4  $subconjunto \leftarrow conjunto[n]$

5  $tamsubconjunto \leftarrow$  tamanho de *subconjunto*

6 **for**  $i = 1$  to  $tamsubconjunto$  **do**

7      $entidades[i] \leftarrow$  **desenfileirar** (*subconjunto*) {Retira o primeiro elemento de *subconjunto* e atribui à posição *i* de *entidades*.}

8 **end for**

9 **for**  $i = 1$  to  $tamsubconjunto$  **do**

10      $entidade \leftarrow entidades[i]$

11     **empilhar** (*produto*, *entidade*) {Insere *entidade* no topo de *produto*.}

12     **if**  $n < tamconjunto$  **then**

13         **ProdutoCartesiano** (*conjunto*, *resultado*, *produto*, *tamconjunto*,  $n + 1$ )

14     **else**

15         **enfileirar** (*resultado*, *produto*) {Insere *produto* no final de *resultado*.}

16     **end if**

17     **desempilhar** (*produto*) {Retira o topo de *produto*.}

18 **end for**

**end procedure**

---

*entidades* na fila *produto* (Figura 2.12). Quando um resultado parcial do produto cartesiano é obtido na pilha *produto*, ele é enfileirado na fila *resultado*.

Iniciamos a determinação do custo do módulo gerador de interações possíveis com base no número de espaços dimensionais da rede que é  $O(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!})$ . Para cada espaço dimensional, o gerador de interações possíveis deve ler as  $n_E/n_C$  entidades de cada categoria biológica que forma o espaço dimensional (Algoritmo 2.6). Assim, se temos *i* categorias em um dado espaço dimensional, o gerador de interações possíveis terá um custo  $O(i \times n_E/n_C)$  ao ler as entidades para esse espaço dimensional.

O custo de gerar o produto cartesiano entre as entidades de cada espaço dimensional é determinado elevando-se o número  $n_E/n_C$  de entidades de cada categoria ao número de categorias de cada espaço dimensional (Algoritmo 2.7). Dessa forma, se



**Figura 2.11:** Representação do processo de leitura das entidades que formam uma rede cujas categorias biológicas são alvo (**A**), doença (**D**), fármaco (**F**) e gene (**G**).

temos  $i$  categorias em um espaço dimensional, esse custo é da ordem  $O((n_E/n_C)^i)$ . Logo, o custo do gerador de interações possíveis é da ordem  $O(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i))$ .

## As Interações Conhecidas

No sistema, cada sub-rede é representada por um grafo ponderado (Ziviani, 2007). Nesse grafo, os nodos são as entidades das categorias que formam o espaço  $n$ -dimensional da sub-rede. As arestas são as interações entre entidades de categorias distintas e correspondem a elementos do produto cartesiano entre as entidades da sub-rede. Além disso, os pesos das arestas medem a evidência de interação entre as entidades. A evidência de interação é determinada com base no modelo de espaço vetorial, quando procuramos pela ocorrência das entidades na coleção de documentos.

No *módulo gerador de interações conhecidas* (módulo 7), o grafo de uma sub-rede  $n$ -dimensional é representado por uma matriz. Cada célula dessa matriz representa uma aresta da sub-rede e, portanto, uma interação entre entidades. Para criar essa matriz, o módulo gerador de interações conhecidas relaciona cada elemento do produto cartesiano de uma sub-rede a uma célula da matriz que representa essa sub-rede (Algoritmo 2.8). Assim, a matriz contém todas as interações possíveis de uma sub-rede. Todas essas interações possíveis serão pesquisadas na coleção de documentos. As interações que forem encontradas na coleção de documentos serão as interações

$\Rightarrow$ conjunto = $\{\{e_{1,A}, e_{2,A}\}, \{e_{1,D}, e_{2,D}, e_{3,D}\}\}$ subconjunto = $\{\emptyset\}$ entidades = $\{\emptyset\}$ entidade = $\emptyset$ produto = $\{\emptyset\}$ resultado = $\{\emptyset\}$ (a)	$\Rightarrow$ conjunto = $\{\{e_{1,A}, e_{2,A}\}, \{e_{1,D}, e_{2,D}, e_{3,D}\}\}$ $\Rightarrow$ subconjunto = $\{e_{1,A}, e_{2,A}\}$ entidades = $\{\emptyset\}$ entidade = $\emptyset$ produto = $\{\emptyset\}$ resultado = $\{\emptyset\}$ (b)	conjunto = $\{\{e_{1,A}, e_{2,A}\}, \{e_{1,D}, e_{2,D}, e_{3,D}\}\}$ subconjunto = $\{e_{2,A}\}$ $\Rightarrow$ entidades = $\{e_{1,A}\}$ entidade = $\emptyset$ produto = $\{\emptyset\}$ resultado = $\{\emptyset\}$ (c)	conjunto = $\{\{e_{1,A}, e_{2,A}\}, \{e_{1,D}, e_{2,D}, e_{3,D}\}\}$ subconjunto = $\{\emptyset\}$ $\Rightarrow$ entidades = $\{e_{1,A}, e_{2,A}\}$ entidade = $\emptyset$ produto = $\{\emptyset\}$ resultado = $\{\emptyset\}$ (d)
$\Rightarrow$ conjunto = $\{\{e_{1,A}, e_{2,A}\}, \{e_{1,D}, e_{2,D}, e_{3,D}\}\}$ subconjunto = $\{\emptyset\}$ entidades = $\{e_{1,A}, e_{2,A}\}$ $\Rightarrow$ entidade = $e_{1,A}$ produto = $\{\emptyset\}$ resultado = $\{\emptyset\}$ (e)	conjunto = $\{\{e_{1,A}, e_{2,A}\}, \{e_{1,D}, e_{2,D}, e_{3,D}\}\}$ subconjunto = $\{\emptyset\}$ entidades = $\{e_{1,A}, e_{2,A}\}$ entidade = $e_{1,A}$ $\Rightarrow$ produto = $\{e_{1,A}\}$ resultado = $\{\emptyset\}$ (f)	conjunto = $\{\{e_{1,A}, e_{2,A}\}, \{e_{1,D}, e_{2,D}, e_{3,D}\}\}$ subconjunto = $\{\emptyset\}$ entidades' = $\{\emptyset\}$ entidade' = $\emptyset$ $\Rightarrow$ produto = $\{e_{1,A}\}$ resultado = $\{\emptyset\}$ (g)	conjunto = $\{\{e_{1,A}, e_{2,A}\}, \{e_{1,D}, e_{2,D}, e_{3,D}\}\}$ subconjunto = $\{\emptyset\}$ $\Rightarrow$ subconjunto' = $\{e_{1,D}, e_{2,D}, e_{3,D}\}$ entidades' = $\{\emptyset\}$ entidade' = $\emptyset$ produto = $\{e_{1,A}\}$ resultado = $\{\emptyset\}$ (h)
conjunto = $\{\{e_{1,A}, e_{2,A}\}, \{e_{1,D}, e_{2,D}, e_{3,D}\}\}$ subconjunto' = $\{e_{2,D}, e_{3,D}\}$ $\Rightarrow$ entidades' = $\{e_{1,D}\}$ entidade' = $\emptyset$ produto = $\{e_{1,A}\}$ resultado = $\{\emptyset\}$ (i)	conjunto = $\{\{e_{1,A}, e_{2,A}\}, \{e_{1,D}, e_{2,D}, e_{3,D}\}\}$ subconjunto' = $\{e_{3,D}\}$ $\Rightarrow$ entidades' = $\{e_{1,D}, e_{2,D}\}$ entidade' = $\emptyset$ produto = $\{e_{1,A}\}$ resultado = $\{\emptyset\}$ (j)	conjunto = $\{\{e_{1,A}, e_{2,A}\}, \{e_{1,D}, e_{2,D}, e_{3,D}\}\}$ subconjunto' = $\{\emptyset\}$ $\Rightarrow$ entidades' = $\{e_{1,D}, e_{2,D}, e_{3,D}\}$ entidade' = $\emptyset$ produto = $\{e_{1,A}\}$ resultado = $\{\emptyset\}$ (l)	conjunto = $\{\{e_{1,A}, e_{2,A}\}, \{e_{1,D}, e_{2,D}, e_{3,D}\}\}$ subconjunto' = $\{\emptyset\}$ entidades' = $\{e_{1,D}, e_{2,D}, e_{3,D}\}$ $\Rightarrow$ entidade' = $e_{1,D}$ produto = $\{e_{1,A}\}$ resultado = $\{\emptyset\}$ (m)
conjunto = $\{\{e_{1,A}, e_{2,A}\}, \{e_{1,D}, e_{2,D}, e_{3,D}\}\}$ subconjunto' = $\{\emptyset\}$ entidades' = $\{e_{1,D}, e_{2,D}, e_{3,D}\}$ entidade' = $e_{1,D}$ $\Rightarrow$ produto = $\{e_{1,A}, e_{1,D}\}$ resultado = $\{\emptyset\}$ (n)	conjunto = $\{\{e_{1,A}, e_{2,A}\}, \{e_{1,D}, e_{2,D}, e_{3,D}\}\}$ subconjunto' = $\{\emptyset\}$ entidades' = $\{e_{1,D}, e_{2,D}, e_{3,D}\}$ entidade' = $e_{1,D}$ produto = $\{e_{1,A}, e_{1,D}\}$ $\Rightarrow$ resultado = $\{\{e_{1,A}, e_{1,D}\}\}$ (o)	conjunto = $\{\{e_{1,A}, e_{2,A}\}, \{e_{1,D}, e_{2,D}, e_{3,D}\}\}$ subconjunto' = $\{\emptyset\}$ entidades' = $\{e_{1,D}, e_{2,D}, e_{3,D}\}$ entidade' = $e_{1,D}$ $\Rightarrow$ produto = $\{e_{1,A}\}$ resultado = $\{\{e_{1,A}, e_{1,D}\}\}$ (p)	conjunto = $\{\{e_{1,A}, e_{2,A}\}, \{e_{1,D}, e_{2,D}, e_{3,D}\}\}$ subconjunto' = $\{\emptyset\}$ entidades' = $\{e_{1,D}, e_{2,D}, e_{3,D}\}$ $\Rightarrow$ entidade' = $e_{2,D}$ produto = $\{e_{1,A}\}$ resultado = $\{\{e_{1,A}, e_{1,D}\}\}$ (q)

**Figura 2.12:** Representação da formação do produto cartesiano entre as entidades que formam uma rede cujas categorias biológicas são alvo (A), doença (D), fármaco (F) e gene (G).

conhecidas da sub-rede. Aquelas interações que não forem encontradas na coleção de documentos passarão pelo processo de inferência e poderão se tornar as interações novas indicadas pelo sistema.

O módulo gerador de interações conhecidas inicia todas as células da matriz com o valor 0, indicando a ausência de interação entre entidades. Então, ele dá início à busca das interações conhecidas na coleção de documentos. O valor da evidência de interação das interações conhecidas pode ser determinado por diversas estratégias, como a média aritmética das similaridades retornadas pelo modelo de espaço vetorial (Algoritmo 2.9), a soma das similaridades retornadas (Algoritmo 2.10) ou ainda a máxima similaridade retornada (Algoritmo 2.11).

O elemento do produto cartesiano armazenado em cada célula da matriz é processado pelo módulo do *modelo de espaço vetorial* (módulo 8) como uma consulta que expressa uma conjunção entre entidades. A conjunção entre as entidades é im-

---

**Algorithm 2.8** Módulo 7: construção das sub-redes.

---

```

function ConstruirSubRede ()
begin procedure
  1 for each sub-rede que compõe a rede biológica do
  2   sejam  $NW$  e  $S$  duas matrizes
  3   relacione cada célula de  $NW$  e de  $S$  a um elemento do produto cartesiano da sub-rede
  4   for  $i = 1$  to numerolinhas do
  5     for  $j = 1$  to numerocolunas do
  6        $NW[i, j] \leftarrow 0$ 
  7        $S[i, j] \leftarrow 0$ 
  8     end for
  9   end for
 10   EncontrarInteraçõesConhecidas ( $NW$ )
 11   Convergência ( $NW, S$ )
 12   apresente o conjunto solução  $S$ 
 13 end for
end procedure

```

---

portante para assegurar que os documentos em que as entidades ocorrem não sejam ortogonais. Dessa forma, a consulta indica que o módulo do modelo de espaço vetorial deve encontrar documentos na coleção que possuem ocorrências de todas as entidades presentes na consulta.

O módulo do modelo de espaço vetorial processa a conjunção entre entidades expressa na consulta, localizando no índice invertido as listas encadeadas dessas entidades. Em seguida, o módulo procura os documentos comuns dessas listas, determinado a interseção entre elas. Após determinar a interseção entre as listas encadeadas, o módulo gera um índice invertido reduzido, contendo apenas as listas referentes às entidades da consulta. As listas do índice invertido reduzido contêm apenas os nodos que armazenam os documentos comuns entre as entidades que formam a consulta. Por exemplo, consideremos uma consulta enviada ao módulo do modelo de espaço vetorial formada pela conjunção das entidades  $e_1$  e  $e_3$  (Figura 2.13). Consideremos também que, no índice invertido, o único documento em que essas duas entidades estão presentes é  $d_5$ . Então, o módulo gera o índice invertido reduzido de  $e_1$  e  $e_3$ , cujas listas encadeadas possuem apenas um nodo que armazena o identificador de  $d_5$  e o peso dessas entidades em  $d_5$ .

Assim que a interseção entre as listas encadeadas é processada, o módulo do modelo de espaço vetorial determina a similaridade entre a consulta e cada documento na interseção das listas encadeadas (Algoritmo 2.12). O módulo do modelo de espaço vetorial gera uma lista encadeada cujos nodos armazenam esses documen-



**Algorithm 2.9** Módulo 7: encontrar as interações conhecidas de uma sub-rede. Nesta versão do algoritmo, a estratégia para determinar a evidência de interação das interações conhecidas corresponde à média aritmética das similaridades retornadas pelo modelo de espaço vetorial.

**procedure** EncontrarInteraçõesConhecidas ( $NW$ )

**parâmetros de entrada** passados por referência:

$NW$  uma matriz

**begin procedure**

```

1  sejam ranking e q duas listas encadeadas inicialmente vazias
2  for  $i = 1$  to numerolinhas do
3    for  $j = 1$  to numerocolunas do
4      contador  $\leftarrow 0$ 
5       $q \leftarrow$  entidades de  $NW[i, j]$ 
6      ranking  $\leftarrow$  ModeloEspacoVetorial ( $q$ )
7      while ranking não estiver vazia do
8        nodo  $\leftarrow$  desalistar (ranking) {Retira o primeiro elemento de ranking e o atribui a nodo.}
9        sim  $\leftarrow$  similaridade armazenada no par ordenado  $\langle d, sim \rangle$  de nodo
10        $NW[i, j] \leftarrow NW[i, j] + sim$ 
11       contador  $\leftarrow$  contador + 1
12     end while
13     if contador > 0 then
14        $NW[i, j] \leftarrow NW[i, j] / contador$ 
15     end if
16   end for
17 end for
end procedure

```

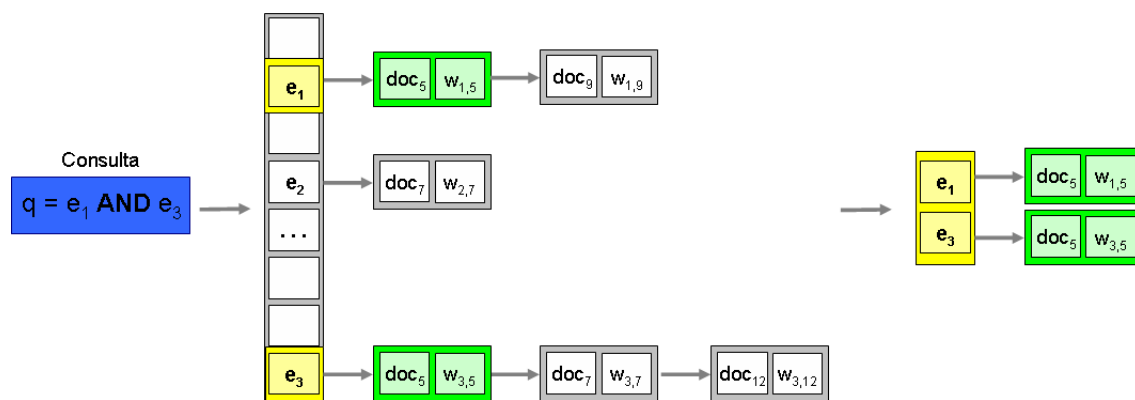


Figura 2.13: Interseção de listas invertidas.

tos e suas respectivas similaridades e a retorna para o módulo gerador de interações conhecidas. Essa lista é chamada de *ranking* e seus nodos são ordenados pelo valor

---

**Algorithm 2.10** Módulo 7: encontrar as interações conhecidas de uma sub-rede. Nesta versão do algoritmo, a estratégia para determinar a evidência de interação das interações conhecidas corresponde à soma das similaridades retornadas pelo modelo de espaço vetorial.

---

**procedure** *EncontrarInteraçõesConhecidas* (*NW*)

**parâmetros de entrada** passados por **referência**:

*NW* uma matriz

**begin procedure**

```

1  sejam ranking e q duas listas encadeadas inicialmente vazias
2  for i = 1 to numerolinhas do
3    for j = 1 to numerocolunas do
4      q ← entidades de NW[i, j]
5      ranking ← ModeloEspacoVetorial (q)
6      while ranking não estiver vazia do
7        nodo ← desalistar (ranking) {Retira o primeiro elemento de ranking e o atribui a nodo.}
8        sim ← similaridade armazenada no par ordenado < d, sim > de nodo
9        NW[i, j] ← NW[i, j] + sim
10     end while
11   end for
12 end for
end procedure

```

---

da similaridade. O módulo gerador de interações conhecidas calcula a evidência de interação entre as entidades da célula que geraram a consulta, com base em todas as similaridades presentes no *ranking*. O valor dessa evidência de interação indica que o relacionamento entre essas entidades já é conhecido e que foi encontrado na coleção de documentos.

Depois de pesquisar todas as consultas de uma matriz, nós temos todas as interações conhecidas da sub-rede. Contudo, algumas células da matriz permanecem com valor 0, indicando que algumas interações entre entidades não são mencionadas na coleção de documentos. Essas células com valor 0 representam as potenciais interações novas entre as entidades biológicas que elas relacionam, porque a coleção de documentos não apresenta evidências de que essas interações foram previamente relatadas por pesquisadores.

Após identificar todas as interações conhecidas de uma sub-rede, o módulo gerador de interações conhecidas pode encontrar entidades que não interagem com as demais entidades dessa sub-rede. Na matriz da sub-rede, todas as células da linha e coluna relacionadas a essas entidades isoladas são iguais a zero. Os zeros indicam a ausência de relacionamento dessas entidades com as demais, o que impossibilita

---

**Algorithm 2.11** Módulo 7: encontrar as interações conhecidas de uma sub-rede. Nesta versão do algoritmo, a estratégia para determinar a evidência de interação das interações conhecidas corresponde à identificação da maior similaridade retornada pelo modelo de espaço vetorial.

---

**procedure** EncontrarInteraçõesConhecidas (*NW*)

**parâmetros de entrada** passados por referência:

*NW* uma matriz

**begin procedure**

```

1  sejam ranking e q duas listas encadeadas inicialmente vazias
2  for i = 1 to numerolinhas do
3    for j = 1 to numerocolunas do
4      q ← entidades de NW[i, j]
5      ranking ← ModeloEspacoVetorial (q)
6      while ranking não estiver vazia do
7        nodo ← desalistar (ranking) {Retira o primeiro elemento de ranking e o atribui a nodo.}
8        sim ← similaridade armazenada no par ordenado < d, sim > de nodo
9        if sim > NW[i, j] then
10         NW[i, j] ← sim
11        end if
12      end while
13    end for
14  end for
end procedure

```

---

a descoberta de novas interações que envolva essas entidades na sub-rede. Então, o módulo gerador de interações elimina da matriz as linhas e colunas de entidades isoladas da sub-rede. Essa eliminação permite reduzir espaço de armazenamento em memória e tempo de processamento do processador durante a análise das sub-redes.

Iniciamos a avaliação do custo de encontrar as interações conhecidas da rede através do cálculo do custo de gerar as matrizes das sub-redes. Esse custo equivale a ler as interações possíveis da rede a partir da base de dados e que é da ordem  $O(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i))$ . Em seguida, cada célula das matrizes é pesquisada no módulo do modelo de espaço vetorial com um custo também da ordem  $O(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i))$ . Assim, o custo de encontrar as interações conhecidas se torna  $O(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i))$  (Equação 2.4).

$$O\left(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right) + \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right)\right) =$$

---

**Algorithm 2.12** Módulo 8: modelo de espaço vetorial.

---

**function** ModeloEspacoVetorial (*consulta*)

**parâmetro de entrada** passado por **valor**:

*consulta* uma lista encadeada de entidades

**begin function**

```

1  seja indice uma tabela hash inicialmente vazia
2  sejam ranking e docs duas listas encadeadas inicialmente vazias
3
4  indice ← interseção entre as entidades presentes em consulta no índice invertido
5  docs ← documentos distintos de indice
6  while docs não estiver vazia do
7    d ← desalistar (docs) {Retira o primeiro elemento de docs e o atribui a d.}
8    soma ← 0
9    tamanho ← tamanho de consulta
10   for each entidade  $e_i \in \textit{consulta}$  tal que  $1 \leq i \leq \textit{tamanho}$  do
11      $w_1 \leftarrow$  peso de  $e_i$  no documento  $d$  armazenado em indice
12      $w_2 \leftarrow$  peso de  $e_i$  em consulta
13     soma ← soma +  $w_1 \times w_2$ 
14   end for
15   a ← soma dos quadrados dos pesos de todas as entidades que ocorrem em d
16   b ← soma dos quadrados dos pesos de todas as entidades que ocorrem em consulta
17   sim ← soma / ( $\sqrt{a} \times \sqrt{b}$ )
18   nodo ← par ordenado  $\langle d, \textit{sim} \rangle$ 
19   alistar (ranking, nodo) {Insere nodo no final de ranking.}
20 end while
21 return ranking

```

**end function**

---

$$\begin{aligned}
O\left(2 \times \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right)\right) &= \\
O\left(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right)\right) & \quad (2.4)
\end{aligned}$$

Durante o processo de pesquisa das interações conhecidas, o custo de encontrar a lista invertida de cada entidade das consultas no índice invertido é da ordem  $O(1)$ . Cada consulta possui uma entidade para cada categoria que compõe o espaço dimensional da sub-rede. Assim, o número de entidades de uma consulta equivale ao número de dimensões da sub-rede. Dessa forma, temos que o custo de encontrar as interações conhecidas se torna  $O\left(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right)\right)$  (Equação 2.5).

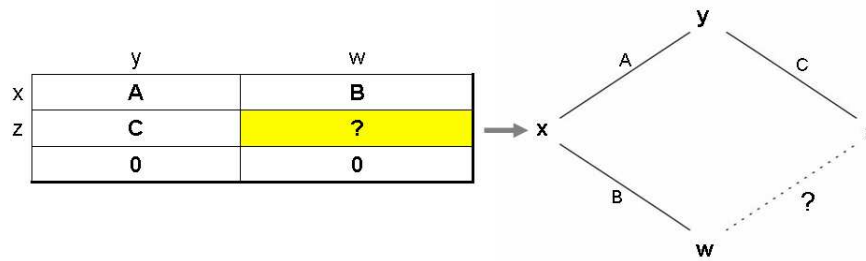
$$O\left(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right) \times i \times 1\right) =$$

$$\begin{aligned}
O\left(\sum_{i=2}^{n_C} \frac{n_C!}{i \times (i-1)! \times (n_C-i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right) \times i\right) &= \\
O\left(\sum_{i=2}^{n_C} \frac{n_C!}{(i-1)! \times (n_C-i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right)\right) &= \\
O\left(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right)\right) & \quad (2.5)
\end{aligned}$$

Consideremos agora que o tamanho das listas encadeadas no índice invertido é  $n_L$ . Assim, para um espaço dimensional formado por  $i$  categorias biológicas, a determinação da interseção entre essas listas gerará um custo da ordem  $O((n_L)^i)$ . Então, o processo de determinar a interseção das listas encadeadas no índice invertido torna o custo de encontrar as interações conhecidas da ordem  $O\left(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right) \times (n_L)^i\right)$ .

Para determinarmos as similaridades entre os documentos nas interseções das listas encadeadas e as consultas, precisamos das somas dos quadrados dos pesos das entidades em cada documento e em cada consulta. A soma do quadrado dos pesos das entidades de cada documento é calculada durante a tarefa de indexação da coleção de documentos. Dessa forma, durante o cálculo da similaridade, esses valores são lidos da base de dados com um custo da ordem  $O(n_D)$ . Por outro lado, em nosso sistema, o peso das entidades nas consultas é sempre 1. Então, a soma dos quadrados dos pesos das entidades de cada consulta equivale ao número de entidades da consulta. Assim, numa sub-rede com  $i$  dimensões a soma dos quadrados dos pesos das entidades em cada consulta gera um custo  $O(i)$ , porque há uma entidade para cada dimensão da sub-rede por consulta. Dessa forma, o custo de encontrar as interações conhecidas é da ordem  $O\left(n_D \times \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right) \times (n_L)^i\right)$  (Equação 2.6).

$$\begin{aligned}
O\left(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right) \times (n_L)^i \times n_D \times i\right) &= \\
O\left(\sum_{i=2}^{n_C} \frac{n_C!}{i \times (i-1)! \times (n_C-i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right) \times (n_L)^i \times n_D \times i\right) &= \\
O\left(\sum_{i=2}^{n_C} \frac{n_C!}{(i-1)! \times (n_C-i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right) \times (n_L)^i \times n_D\right) &= \\
O\left(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right) \times (n_L)^i \times n_D\right) &= \\
O\left(n_D \times \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right) \times (n_L)^i\right) & \quad (2.6)
\end{aligned}$$



**Figura 2.14:** Relação de transitividade em uma matriz representando as conexões de uma sub-rede bidimensional genérica.

## A Inferência de Novas Interações

Depois que todas as interações já conhecidas são encontradas através das buscas na coleção de documentos e são estabelecidas na rede, o módulo *gerador de novas interações* (módulo 9) inicia o processo de inferência das novas interações. O módulo gerador de novas interações infere uma nova interação em uma sub-rede quando encontra uma *relação de transitividade* entre 3 interações já estabelecidas nessa sub-rede. Nós definimos que 3 interações estabelecidas na rede estão em uma relação de transitividade quando satisfazem a condição  $(x, y)$  and  $(x, w)$  and  $(z, y) \rightarrow (z, w)$  (Figura 2.14). Essa relação de transitividade explora as atividades principais e secundárias das entidades no sistema biológico, indicando que "IF a entidade  $x$  interage com as entidades  $y$  e  $w$  AND a entidade  $z$  interage com a entidade  $y$ , THEN  $z$  possivelmente também interage com a entidade  $w$ ". Após a inferência de uma nova interação, o módulo registra a evidência de interação dessa nova interações na matriz que representa a sub-rede.

Quando as novas interações são registradas na matriz de uma sub-rede, elas podem formar triplas com as interações já conhecidas, tornando-se capazes de contribuir para a inferência de outras interações novas entre as entidades. Como resultado disso, mais interações novas podem ser descobertas, aplicando-se uma nova iteração do processo de inferência. Isso é possível, porque interações descobertas no processo de inferência podem formar triplas com interações já conhecidas da sub-rede ou com outras interações novas, satisfazendo a condição imposta pela relação de transitividade. De forma geral, as novas interações inferidas pelo módulo em uma dada iteração podem ser usadas em iterações subseqüentes, para ajudar na inferência de outras novas interações. Dessa maneira, o módulo gerador de novas interações infere novas interações entre entidades biológicas utilizando as interações conhecidas de uma sub-rede ou também as interações previamente inferidas. Então, novas descobertas levam ao surgimento de outras, desencadeando um processo evolutivo de descobrimento de novas interações. Esse processo pode ser repetido até que, através

da aplicação de várias iterações do processo de inferência, não sejam formadas novas triplas que satisfaçam a condição imposta pela relação de transitividade ou que todas as interações possíveis da sub-rede sejam convertidas em interações conhecidas e em interações novas. Assim, é possível que o módulo gerador de novas interações consiga convergir a matriz de uma sub-rede para outra em que todas as interações possíveis entre entidades são apresentadas como interações já conhecidas ou como interações novas (Algoritmo 2.13).

---

**Algorithm 2.13** Módulo 9: convergência de novas interações em uma sub-rede.

---

**procedure** *Convergência* (*NW*, *S*)

**parâmetros de entrada** passados por **referência**:

*NW* e *S* duas matrizes

**begin procedure**

```

1  zeroscorrentes ← 0
2  iteracao ← 1
3  for i ← 1 to numerolinhas do
4    for j ← 1 to numerocolunas do
5      if NW[i, j] = 0 then
6        zeroscorrentes ← zeroscorrentes + 1
7      end if
8    end for
9  end for
10 zerosprevios ← zeroscorrentes + 1
11 while zeroscorrentes < zerosprevios and zeroscorrentes > 0 do
12   InferirNovasInterações (NW, S, iteracao)
13   zerosprevios ← zeroscorrentes
14   for i ← 1 to numerolinhas do
15     for j ← 1 to numerocolunas do
16       if NW[i, j] = 0 and S[i, j] ≠ 0 then
17         NW[i, j] ← S[i, j]
18         zeroscorrentes ← zeroscorrentes - 1
19       end if
20     end for
21   end for
22   iteracao ← iteracao + 1
23 end while
end procedure

```

---

O módulo gerador de novas interações usa um fator para penalizar a evidência de interação das novas interação a cada aplicação do processo de inferência. O fator de penalidade é usado para dividir a evidência de interação atribuída às novas interações e equivale ao número de iterações do processo de inferência. Dessa forma,

na primeira iteração do processo de inferência, esse fator é igual a 1 e, por isso, não altera a evidência de interação calculada para as interações descobertas a partir das interações já conhecidas. Na iteração seguinte do processo de inferência, o valor do fator de penalidade é incrementado e passa a ser igual a 2. Isso significa que a evidência de interação das novas interações inferidas com base nas interações conhecidas e com base nas interações novas da iteração 1 serão divididas por 2. Assim, a cada iteração subsequente do processo de inferência, o módulo gerador de novas interações aumenta o valor desse fator de penalidade. Com isso, o módulo gerador de novas interações privilegia as descobertas das primeiras iterações do processo de inferência em detrimento das últimas, visto que as descobertas das últimas iterações só são possíveis com a identificação das descobertas de iterações anteriores.

No processo de inferência, o módulo gerador de novas interações percorre a matriz, procurando pelas células que permanecem com valor igual a 0. Assim que ele encontra uma dessas células, ele procura por todas as triplas de interações estabelecidas na sub-rede que satisfaçam a condição imposta pela relação de transitividade. A evidência de interação de uma nova interação inferida pelo módulo é calculada com base nas evidências de interação dessas 3 interações que satisfazem a condição de transitividade. Assim como para as interações conhecidas, podemos usar diversas estratégias para esse cálculo, como a média aritmética das evidências de interação da tripla de interações, a soma dessas evidências de interação ou também podemos escolher a maior evidência de interação da tripla. Na implementação corrente do sistema, avaliamos a estratégia da média aritmética.

Se muitas triplas de interações satisfazem a condição imposta pela relação de transitividade, o módulo gerador de novas interações aplica uma estratégia para determinar o valor da nova interação. Várias estratégias podem ser empregadas, por exemplo: somar o valor das evidências de interação calculado a partir de cada tripla, calcular a média aritmética dessas evidências de interação ou escolher a maior evidência de interação obtida a partir dessas triplas. Em nossa implementação corrente, o módulo seleciona a tripla que resulta na maior evidência de interação. A evidência de interação escolhida é registrada na célula que permanecia com o valor igual a 0 e representa uma nova interação inferida pelo módulo. Esse processo é conduzido até que todas as células da matriz recebam um valor diferente de 0 ou até que não haja mais triplas de interações na sub-rede que satisfaçam a condição imposta pela relação de transitividade (Algoritmo 2.14). Depois que todas as novas interações são identificadas, o módulo gerador de novas interações armazena todas as interações das matrizes que representam as sub-redes na base de dados, para que possam ser consultadas pelos usuários através da *Web*.



Entretanto, quando há mais de uma tripla de interações que satisfazem a condição imposta pela relação de transitividade, podemos novamente aplicar várias estratégias para determinar o valor da evidência de interação da nova interação. Por exemplo, (Equação 3.19).

---

**Algorithm 2.14** Módulo 9: inferência de novas interações.

---

```

procedure InferirNovasInterações (NW, S, penalidade)
parâmetros de entrada passados por referência:
NW e S duas matrizes
parâmetros de entrada passados por valor:
penalidade um inteiro
begin procedure
1  media ← 0
2  divisor ← 3 × penalidade
3  for i ← 1 to numerolinhas do
4    for j ← 1 to numerocolunas do
5      if NW[i, j] = 0 then
6        for k ← 1 to numerolinhas do
7          for l ← 1 to numerocolunas do
8            if k ≠ i and l ≠ j then
9              if NW[k, l] ≠ 0 and NW[i, l] ≠ 0 and NW[k, j] ≠ 0 then
10               average ← (NW[k, l] + NW[i, l] + NW[k, j])/divisor
11               if average > S[i, j] then
12                 S[i, j] ← average
13               end if
14             end if
15           end if
16         end for
17       end for
18     end if
19   end for
20 end for
end procedure

```

---

Para determinarmos o custo de inferir as novas interações, observamos que o módulo tem um custo inicial da ordem  $O(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i))$ , para contar o número de células iguais a zero na rede (Algoritmo 2.13). Esse custo equivale ao número de interações possíveis da rede. Além disso, observamos que, em cada iteração do processo de inferência, o módulo tem um outro custo da ordem  $O(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i))$ , para contar as células que permanecem com valor igual a 0 e registrar na matriz o valor das novas interações inferidas. Então, consideremos que o módulo gerador de novas interações aplique  $n$  iterações do pro-

cesso de inferência, para inferir todas as interações novas da rede. Assim, o custo do módulo torna-se  $O(n \times \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i))$  (Equação 2.7).

$$\begin{aligned}
& O\left(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i) + \right. \\
& \quad \left. n \times \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i) \right) = \\
& O((n + 1) \times \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i)) = \\
& O(n \times \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i)) \tag{2.7}
\end{aligned}$$

Durante o processo de inferência, o módulo percorre a matriz de cada sub-rede procurando as células com valor igual a 0 (Algoritmo 2.14). Isso gera um custo da ordem  $O(\frac{n_C!}{i! \times (n_C - i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i))$  para cada uma das sub-redes com  $i$  dimensões. Sempre que o módulo encontra uma célula com valor igual a 0, ele percorre toda a matriz novamente para encontrar as interações que satisfazem a relação de transitividade. Isso gera novamente outro custo da ordem  $O(\frac{n_C!}{i! \times (n_C - i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i))$  para cada uma das sub-redes com  $i$  dimensões. Dessa forma, o custo de encontrar as novas interações da rede em cada uma das iterações do processo de inferência é da ordem  $O(\sum_{i=2}^{n_C} (\frac{n_C!}{i! \times (n_C - i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i))^2)$  (Equação 2.8).

$$\begin{aligned}
O\left(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i) \times \frac{n_C!}{i! \times (n_C - i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i) \right) = \\
O\left(\sum_{i=2}^{n_C} (\frac{n_C!}{i! \times (n_C - i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i))^2 \right) \tag{2.8}
\end{aligned}$$

Por fim, para determinarmos o custo total do módulo gerador de novas iterações, somamos o custo de contabilizar as células que permanecem com valor igual a 0 e o custo de encontrar as novas interações da rede nas  $n$  iterações do processo de inferência. Logo, o custo total de encontrar todas as interações novas da rede é da ordem  $O(n \times \sum_{i=2}^{n_C} (\frac{n_C!}{i! \times (n_C - i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i))^2)$  (Equação 2.9).

$$\begin{aligned}
O\left(n \times \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i) + \right. \\
\left. n \times \sum_{i=2}^{n_C} (\frac{n_C!}{i! \times (n_C - i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i))^2 \right) =
\end{aligned}$$

$$\begin{aligned}
& O\left(n \times \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right) + \right. \\
& \quad \left. \left(\frac{n_C!}{i! \times (n_C - i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right)\right)^2\right) = \\
& O\left(n \times \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right) \times \right. \\
& \quad \left. \left(1 + \frac{n_C!}{i! \times (n_C - i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right)\right)\right) = \\
& O\left(n \times \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right) \times \right. \\
& \quad \left. \frac{n_C!}{i! \times (n_C - i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right)\right) = \\
& O\left(n \times \sum_{i=2}^{n_C} \left(\frac{n_C!}{i! \times (n_C - i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right)\right)^2\right) \tag{2.9}
\end{aligned}$$

### O Custo da Tarefa de Construção da Rede Biológica

O custo da tarefa de construção da rede biológica corresponde à soma dos custos de formar os espaços dimensionais e de encontrar as interações possíveis, conhecidas e novas. Temos que o custo de:

- Formar os espaços dimensionais é da ordem  $O\left(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!}\right)$  (Página 22, Equação 2.3).
- Determinar as interações possíveis é da ordem  $O\left(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right)\right)$  (Página 26).
- Identificar as interações conhecidas é da ordem  $O\left(n_D \times \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right) \times (n_L)^i\right)$  (Página 33, Equação 2.5).
- Inferir as novas interações é da ordem  $O\left(n \times \sum_{i=2}^{n_C} \left(\frac{n_C!}{i! \times (n_C - i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right)\right)^2\right)$  (Página 39, Equação 2.9).

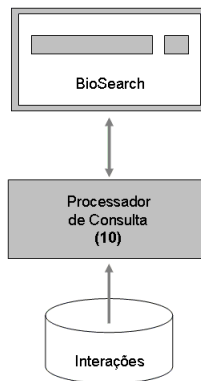
Logo, o custo da tarefa de construção da rede biológica é da ordem  $O\left(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right) + n_D \times \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right) \times (n_L)^i + n \times \sum_{i=2}^{n_C} \left(\frac{n_C!}{i! \times (n_C - i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right)\right)^2\right)$  (Equação 2.10).

$$\begin{aligned}
& O\left(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!}\right) + \left(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right)\right) + \\
& \quad \left(n_D \times \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times \left(i \times \frac{n_E}{n_C} + \left(\frac{n_E}{n_C}\right)^i\right) \times (n_L)^i\right) +
\end{aligned}$$

$$\begin{aligned}
& \left( n \times \sum_{i=2}^{n_C} \left( \frac{n_C!}{i! \times (n_C - i)!} \times \left( i \times \frac{n_E}{n_C} + \left( \frac{n_E}{n_C} \right)^i \right)^2 \right) \right) = \\
& O \left( \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} + \frac{n_C!}{i! \times (n_C - i)!} \times \left( i \times \frac{n_E}{n_C} + \left( \frac{n_E}{n_C} \right)^i \right) + \right. \\
& \quad \left. n_D \times \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times \left( i \times \frac{n_E}{n_C} + \left( \frac{n_E}{n_C} \right)^i \right) \times (n_L)^i + \right. \\
& \quad \left. n \times \sum_{i=2}^{n_C} \left( \frac{n_C!}{i! \times (n_C - i)!} \times \left( i \times \frac{n_E}{n_C} + \left( \frac{n_E}{n_C} \right)^i \right)^2 \right) \right) = \\
& O \left( \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times \left( 1 + i \times \frac{n_E}{n_C} + \left( \frac{n_E}{n_C} \right)^i \right) + \right. \\
& \quad \left. n_D \times \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times \left( i \times \frac{n_E}{n_C} + \left( \frac{n_E}{n_C} \right)^i \right) \times (n_L)^i + \right. \\
& \quad \left. n \times \sum_{i=2}^{n_C} \left( \frac{n_C!}{i! \times (n_C - i)!} \times \left( i \times \frac{n_E}{n_C} + \left( \frac{n_E}{n_C} \right)^i \right)^2 \right) \right) = \\
& O \left( \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times \left( i \times \frac{n_E}{n_C} + \left( \frac{n_E}{n_C} \right)^i \right) + \right. \\
& \quad \left. n_D \times \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times \left( i \times \frac{n_E}{n_C} + \left( \frac{n_E}{n_C} \right)^i \right) \times (n_L)^i + \right. \\
& \quad \left. n \times \sum_{i=2}^{n_C} \left( \frac{n_C!}{i! \times (n_C - i)!} \times \left( i \times \frac{n_E}{n_C} + \left( \frac{n_E}{n_C} \right)^i \right)^2 \right) \right) \quad (2.10)
\end{aligned}$$

### 2.1.4 A Tarefa de Consulta das Interações da Rede Biológica

Depois que todas as interações são estabelecidas na rede biológica, elas são disponibilizadas para pesquisa através da *Web* (Figura 2.15). As pesquisas realizadas na interface cliente para *Web* são analisadas pelo módulo *processador de consulta* (módulo 10). Nessa interface cliente para pesquisa, os usuários expressam suas necessidades de informação através de uma consulta (Baeza-Yates and Ribeiro-Neto, 1999). Então, o processador de consulta procura na base de dados as interações da rede que satisfazem a consulta especificada. O processador de consulta apresenta como resposta um conjunto de interações ordenadas em ordem decrescente de evidência de interação.



**Figura 2.15:** Tarefa 4: consulta das interações da rede biológica.

Na interface cliente, o usuário pode especificar quais entidades ele deseja consultar. No entanto, quando nenhuma entidade é especificada, o sistema assume que todas as entidades da base devem ser consideradas para satisfazer a consulta. O usuário também pode informar se deseja consultar a rede completa ou alguma sub-rede específica. Dessa forma, o sistema permite consultas globais em toda a rede e também consultas locais, para analisar as interações em cada sub-rede separadamente. Além disso, o usuário pode especificar também se deseja consultar apenas as interações conhecidas, apenas as interações novas, ou ambas (Algoritmo 2.15).

---

**Algorithm 2.15** Módulo 10: processamento de consulta.

---

**function** `ProcessarConsulta` (*consulta*)

**parâmetro de entrada** passado por **valor**:

*consulta* uma lista encadeada

**begin function**

```

1  sejam ranking e temp duas filas (FIFO) inicialmente vazias
2
3  temp ← interações da rede
4  while temp não estiver vazia do
5    t ← desenfileirar (temp) {Retira o primeiro elemento de temp e o atribui a t.}
6    if t relaciona as entidades específicas em consulta then
7      if t possui o espaço dimensional especificado em consulta then
8        if t possui a iteração especificada em consulta then
9          alistar (ranking, t) {Insere t no final de ranking.}
10         end if
11       end if
12     end if
13 end while
14 ordene (ranking)
15 return ranking
  
```

**end function**

---

## O Custo da Tarefa de Consulta das Interações da Rede Biológica

Para analisarmos o custo do módulo processador de consulta, consideremos que  $k_3$  é uma constante no intervalo  $[0, 1]$  que indica a proporção de interações possíveis que se tornaram interações conhecidas ou novas na rede. Inicialmente o módulo lê as interações da rede a partir da base de dados o que gera um custo da ordem  $O(k_3 \times \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i)) = O(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i))$ . Em seguida, ele seleciona entre as interações lidas aquelas que possuem as entidades, espaço dimensional e iteração especificados na consulta, o que gera outro custo da ordem  $O(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i))$ . Por fim, o módulo ordena o *ranking* de resposta, gerando um custo da ordem  $O(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i) \times \log(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i)))$ . Logo o custo de consultar as interações da rede é da ordem  $O(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i) \times \log(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i)))$  (Equação 2.11).

$$\begin{aligned}
& O\left(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i) + \right. \\
& \quad \left. \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i) + \right. \\
& \quad \left. \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i) \times \right. \\
& \quad \left. \log\left(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i)\right)\right) = \\
& O\left(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i) \times \right. \\
& \quad \left. (1 + 1 + \log\left(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i)\right))\right) = \\
& O\left(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i) \times \right. \\
& \quad \left. \log\left(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i)\right)\right) \tag{2.11}
\end{aligned}$$

### 2.1.5 O Custo Total do Sistema

O custo total do sistema corresponde à soma dos custos de suas 4 tarefas que são:

- Tarefa de coleta:  $O(n_E \times n_D)$  (Página 16, Equação 2.1).
- Tarefa de indexação:  $O(n_D \times n_E)$  (Página 18, Equação 2.2).
- Tarefa de construção da rede:  $O(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i) + n_D \times \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i) \times (n_L)^i + n \times \sum_{i=2}^{n_C} (\frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i))^2)$  (Página 40, Equação 2.10).
- Tarefa de consulta:  $O(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i) \times \log(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i)))$  (Página 42, Equação 2.11).

Logo, o custo total do sistema é da ordem  $O(n_E \times n_D + n_D \times \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i) \times (n_L)^i + n \times \sum_{i=2}^{n_C} (\frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i))^2 + \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i) \times \log(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i)))$  (Equação 2.12).

$$\begin{aligned}
& O(n_E \times n_D + n_D \times n_E + \\
& \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i) + \\
& n_D \times \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i) \times (n_L)^i + \\
& n \times \sum_{i=2}^{n_C} (\frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i))^2 + \\
& \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i) \times \\
& \log(\sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i)) =
\end{aligned}$$

$$\begin{aligned}
& O(2 \times n_E \times n_D + \\
& n_D \times \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i) \times (n_L)^i + \\
& n \times \sum_{i=2}^{n_C} (\frac{n_C!}{i! \times (n_C-i)!} \times (i \times \frac{n_E}{n_C} + (\frac{n_E}{n_C})^i))^2 +
\end{aligned}$$

$$\begin{aligned}
& \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times \left( i \times \frac{n_E}{n_C} + \left( \frac{n_E}{n_C} \right)^i \right) \times \\
& \left( 1 + \log \left( \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times \left( i \times \frac{n_E}{n_C} + \left( \frac{n_E}{n_C} \right)^i \right) \right) \right) = \\
& \qquad \qquad \qquad O(n_E \times n_D + \\
& n_D \times \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times \left( i \times \frac{n_E}{n_C} + \left( \frac{n_E}{n_C} \right)^i \right) \times (n_L)^i + \\
& n \times \sum_{i=2}^{n_C} \left( \frac{n_C!}{i! \times (n_C - i)!} \times \left( i \times \frac{n_E}{n_C} + \left( \frac{n_E}{n_C} \right)^i \right) \right)^2 + \\
& \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times \left( i \times \frac{n_E}{n_C} + \left( \frac{n_E}{n_C} \right)^i \right) \times \\
& \log \left( \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \times \left( i \times \frac{n_E}{n_C} + \left( \frac{n_E}{n_C} \right)^i \right) \right) \tag{2.12}
\end{aligned}$$

### 2.1.6 O Desempenho do Sistema

Como podemos observar, o número de entidades e de categorias consideradas na construção da rede são os principais responsáveis por grandes implicações no desempenho do sistema. O aumento no número de entidades e de categorias promove um grande aumento no número de interações possíveis na rede, o que demanda grande tempo de processamento para realizar o produto cartesiano entre as entidades e maior capacidade de memória para armazenamento dessas interações. Conseqüentemente, esse grande número de interações possíveis torna as matrizes de cada sub-rede maiores, também resultando em aumento no tempo de processamento e maior consumo de espaço em memória. Além disso, o grande número de interações possíveis nas matrizes também aumenta consideravelmente o número de buscas para encontrar as interações conhecidas na coleção de documento. Por fim, o aumento no tamanho das matrizes das sub-rede também tem reflexo na inferência das novas interações, aumentando o tempo de processamento para determinar as triplas que satisfazem a condição de transitividade.

Na implementação corrente do sistema, empregamos processamento paralelo em todos os módulos da arquitetura, para obtermos ganhos de desempenho. Além disso, os dados são mantidos pré-processados em todos os módulos para melhorar o desempenho de tarefas subseqüentes. Em versões futuras, implementaremos ainda uma arquitetura distribuída para o sistema, com o objetivo de aproveitarmos as



vantagem do processamento e armazenamento em rede.



## Capítulo 3

# O Modelo de Inferência

Em um sistema biológico encontramos várias entidades que o constituem, que indicam seu estado ou que alteram seu estado. Essas entidades podem ser classificadas em categorias como eco-sistemas, organismos, órgãos, tecidos, células, organelas, genes, proteínas, alvos biológicos, doenças, fármacos, etc. Cada uma dessas entidades possui funções importantes para o sistema. Além disso, a ação de uma entidade pode mediar ou interferir na ação de entidades de categorias diferentes, formando uma rede complexa de interações.

Uma rede biológica é formada pelas entidades de um sistema biológico e pelas interações entre essas entidades. As entidades correspondem aos nodos da rede e as interações entre as entidades formam as conexões da rede. Através da rede, nós podemos evidenciar interações já conhecidas do sistema e avaliar possíveis interações novas ainda não observadas entre as entidades. Muitas interações são identificadas, por exemplo, quando o resultado de um experimento com microarranjos é analisado, ou quando é realizado um estudo da co-ocorrência dessas entidades em uma coleção de documentos que descreve o conhecimento em biologia, como artigos científicos, bulas de medicamentos, anotações sobre resultados de bancada, patentes, etc (Swanson, 1990; Smalheiser and Swanson, 1998; Swanson *et al.*, 2006; Weeber *et al.*, 2001; Hristovski *et al.*, 2006; Campillos *et al.*, 2008). Além disso, também podemos empregar métodos de inferência na rede que use as interações já estabelecidas para indicar interações novas.

Neste trabalho, desenvolvemos um modelo de inferência que relaciona entidades de categorias biológicas diferentes, para formar uma rede composta de sub-redes  $n$ -dimensionais. Cada sub-rede representa as interações entre essas entidades em sistemas biológicos. Além disso, todas as interações entre entidades das sub-redes são estabelecidas com base em uma coleção de documentos que descreve o conhecimento em biologia, usando-se o modelo de espaço vetorial. As interações entre entidades

nas sub-redes são usadas para inferir novas interações a partir de uma relação de transitividade que empregamos no processo de inferência do modelo. As interações de uma sub-rede recebem um valor que mede a evidência de interação entre as entidades e que também é calculado com base no modelo de espaço vetorial.

Desenvolvemos também uma estratégia que permite o modelo inferir novas interações entre entidades a partir de interações previamente inferidas. Dessa forma, o modelo desencadeia um processo evolutivo de descobrimento de novas interações em que novas descobertas levam ao surgimento de outras. Como resposta, o modelo apresenta um conjunto de interações ordenadas em ordem decrescente da evidência de interação que liga as entidades biológicas, permitindo analisar as interações com maior ou menor evidência de interação na rede ou em cada sub-rede separadamente.

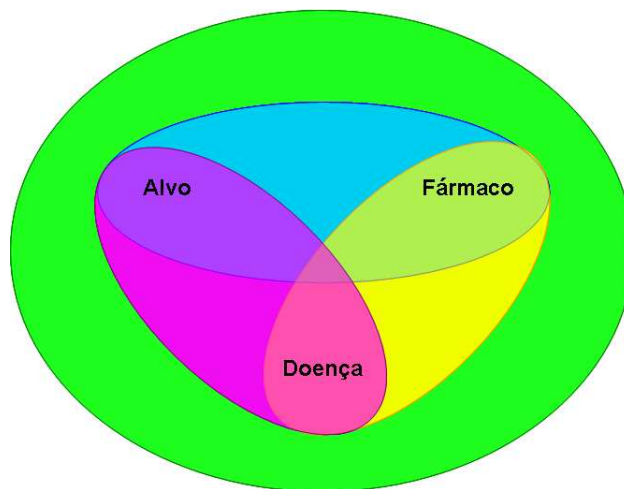
### 3.1 A Rede Biológica e suas Sub-Redes

Em nosso modelo a construção da rede biológica é baseada numa coleção de documentos que descreve o conhecimento em biologia. Todas as interações entre entidades que compõem a rede são inicialmente estabelecidas com base na ocorrência dessas entidades na coleção de documentos. As interações entre entidades são modeladas usando a teoria dos grafos (Cormen *et al.*, 2001; Ziviani, 2007). Um grafo é definido por um par  $G = (V, A)$  onde  $V = \{v_1, v_2, \dots, v_{n_V}\}$  é um conjunto não vazio de vértices com cardinalidade  $n_V$  e  $A = \{(v_i, v_j) \mid v_i, v_j \in V, 1 \leq i \leq n_V, 1 \leq j \leq n_V\}$  é um conjunto de arestas que ligam os vértices do grafo.

Em nossa abordagem, o conjunto de vértices  $V$  é formado pelas diversas entidades que compõem o sistema biológico. O conjunto de arestas  $A$  é formado pelas interações entre entidades de categorias biológicas distintas. As interações de  $A$  são identificadas quando as entidades são mencionadas simultaneamente em algum dos documentos da coleção textual. Entidades de uma mesma categoria não são conectadas no modelo. Dessa forma, as arestas criam partições no grafo com o objetivo de relacionar entidades de categorias distintas. Assim, as interações entre entidades de duas categorias distintas formam um grafo bipartido (Ziviani, 2007; Yamanishi *et al.*, 2008), já as interações entre entidades de  $n$  categorias diferentes formam um grafo multipartido, ou n-partido.

As interações entre entidades de categorias distintas também são responsáveis pela formação de sub-grafos. Um grafo  $G' = (V', A')$  é chamado de sub-grafo de um grafo  $G = (V, A)$  se  $V' \subseteq V$  e  $A' \subseteq A$  (Ziviani, 2007). Em nosso modelo, chamamos de dimensão cada uma das categorias biológicas envolvidas na formação de um sub-grafo. Portanto, o menor número de dimensões encontradas nos sub-grafos

do modelo é 2 e o maior número de dimensões encontradas equivale ao número de categorias que compõem o sistema biológico. Além disso, as interações entre entidades de categorias distintas recebem um peso. Esse peso mensura a evidência de interação  $\epsilon$  que liga essas entidades na rede que representa o sistema biológico, conforme indicado pela coleção de documentos. Um grafo  $G = (V, A)$  que possui um peso  $\epsilon$  associado a cada uma de suas arestas  $A = \{(v_i, v_j, \epsilon) \mid v_i, v_j \in V, \epsilon \in \mathbb{R}\}$  é chamado de grafo ponderado (Ziviani, 2007). Nesse trabalho nos referimos a um grafo ponderado como uma rede e a um sub-grafo ponderado como sub-rede. Por essa razão, dizemos que em nosso modelo temos uma rede biológica formada por sub-redes n-dimensionais.



**Figura 3.1:** Representação da combinação das categorias *alvo*, *doença* e *fármaco* para formação dos espaços dimensionais  $alvo \times doença$ ,  $alvo \times fármaco$ ,  $doença \times fármaco$  e  $alvo \times doença \times fármaco$  que integram uma rede biológica.

## 3.2 A Formação dos Espaços Dimensionais

Para a construção e pesquisa na rede biológica de nosso modelo, consideremos o conjunto  $C = \{c_1, c_2, \dots, c_{n_C}\}$  de categorias do sistema biológico, o conjunto  $E = \{e_1, e_2, \dots, e_{n_E}\}$  de entidades do sistema biológico pertencentes a categorias de  $C$ , o conjunto  $Q = \{q_1, q_2, \dots, q_{n_Q}\}$  de consultas formadas pela conjunção de entidades em  $E$  e que também é o conjunto de interações possíveis entre essas entidades e o conjunto  $D = \{d_1, d_2, \dots, d_{n_D}\}$  de documentos que descrevem o conhecimento em biologia e que representa a coleção textual de nosso modelo de inferência. Nesses conjuntos,  $n_C$ ,  $n_E$ ,  $n_Q$  e  $n_D$  são as cardinalidades de  $C$ ,  $E$ ,  $Q$  e  $D$  respectivamente.

A construção da rede inicia com a formação dos espaços dimensionais que caracterizam as sub-redes. Cada espaço dimensional é uma combinação das categorias

do sistema biológico que compõem o conjunto  $C$ . Por exemplo, se o conjunto  $C$  é formado pelas categorias *alvo*, *doença* e *fármaco*, a combinação dessas categorias dará origem 4 espaços-dimensionais (Figura 3.1). Cada um desses espaços dimensionais corresponde a uma sub-rede da rede completa. Nessa rede, 3 sub-redes têm 2 dimensões (*alvo*  $\times$  *doença*, *alvo*  $\times$  *fármaco* e *doença*  $\times$  *fármaco*) e 1 sub-rede tem 3 dimensões (*alvo*  $\times$  *doença*  $\times$  *fármaco*). Consideremos então que  $C_{n_C, m}$  é o conjunto resultante da combinação das  $n_C$  categorias de  $C$  tomadas  $m$  a  $m$  e que  $C_m^{n_C}$  é a cardinalidade de  $C_{n_C, m}$  (Equação 3.1). Consideremos ainda o conjunto de espaços dimensionais  $S = \{s_1, s_2, \dots, s_{n_S}\}$  em que cada espaço dimensional  $s_i \in S$  representa uma sub-rede específica do sistema biológico. Assim, a cardinalidade  $n_S$  de  $S$  equivale ao número de sub-redes que formam a rede biológica e é determinado pela soma do número de combinações possíveis a partir de  $C$  (Equação 3.2).

$$C_m^{n_C} = \text{card}(C_{n_C, m}) = \frac{n_C!}{m! \times (n_C - m)!} \quad (3.1)$$

$$n_S = \text{card}(S) = \sum_{i=2}^{n_C} \frac{n_C!}{i! \times (n_C - i)!} \quad (3.2)$$

Iniciamos a geração de  $S$  pela formação dos espaços dimensionais das sub-redes bidimensionais. Esses espaços dimensionais correspondem às combinações das  $n_C$  categorias biológicas de  $C$  tomadas duas a duas ( $C_{n_C, 2}$ ). Por exemplo, *alvo*  $\times$  *doença*, *alvo*  $\times$  *fármaco* e *doença*  $\times$  *fármaco*. Em seguida, geramos os espaços tridimensionais que correspondem às combinações das  $n_C$  categorias de  $C$  tomadas três a três ( $C_{n_C, 3}$ ). Por exemplo, *alvo*  $\times$  *doença*  $\times$  *fármaco*. Continuamos as combinações até que tenhamos o espaço  $n_C$ -dimensional que corresponde à combinação das  $n_C$  categorias biológicas de  $C$  tomadas  $n_C$  a  $n_C$  ( $C_{n_C, n_C}$ ). Dessa forma, o conjunto  $S$  pode ser descrito como a união dos espaços dimensionais de todas as sub-redes que constituem a rede de interações biológicas (Equação 3.3).

$$S = \bigcup_{i=2}^{n_C} C_{n_C, i} = \{C_{n_C, 2} \cup C_{n_C, 3} \cup \dots \cup C_{n_C, n_C}\} \quad (3.3)$$

### 3.2.1 As Interações entre Entidades Biológicas

As interações possíveis entre entidades em uma sub-rede correspondem às tuplas do espaço  $n$ -dimensional  $s_i \in S$  dessa sub-rede. Essas tuplas são obtidas pelo produto cartesiano entre as entidades biológicas  $e_i \in E$  das categorias que formam  $s_i$ . Então, consideremos o conjunto das  $n$  categorias que formam  $s_i$  como  $C_{s_i} = \{c_{1, s_i}, c_{2, s_i}, \dots, c_{n, s_i}\}$ , sendo  $C_{s_i} \subseteq C$ ,  $n$  a cardinalidade de  $C_{s_i}$  e  $n_{c_{1, s_i}}, n_{c_{2, s_i}}, \dots,$

$n_{c_{n,s_i}}$  as cardinalidades de  $c_{1,s_i}, c_{2,s_i}, \dots, c_{n,s_i}$ , respectivamente. O produto cartesiano que gera as tuplas possíveis de  $s_i$  é dado por  $s_i = \{c_{1,s_i} \times c_{2,s_i} \times \dots \times c_{n,s_i} \mid e_1 \in c_{1,s_i}, e_2 \in c_{2,s_i}, \dots, e_n \in c_{n,s_i}\}$ . Assim, a cardinalidade  $t_{s_i}$  de  $s_i$  equivale ao número de tuplas de  $s_i$  e indica o número de interações possíveis entre as entidades  $e_i \in E$  das categorias de  $s_i$  (Equação 3.4).

$$t_{s_i} = \text{card}(s_i) = \prod_{x=1}^n \text{card}(c_{x,s_i}) \quad (3.4)$$

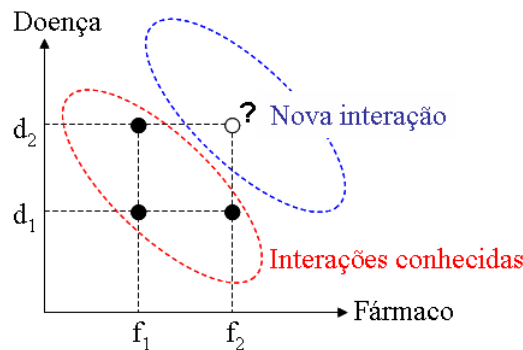
onde  $c_{x,s_i} \in C_{s_i}$  é a categoria biológica  $x$  do espaço dimensional  $s_i$  e  $\text{card}(c_{x,s_i})$  é a cardinalidade de  $c_{x,s_i}$ , número de entidades biológicas de  $c_{x,s_i}$ . Dessa forma, o total de interações possíveis na rede corresponde ao número total de tuplas em toda a rede e é calculado pela soma das tuplas em cada sub-rede (Equação 3.5).

$$t_{rede} = \sum_{i=1}^{n_s} t_{s_i} \quad (3.5)$$

Em nosso modelo, as interações já conhecidas entre entidades biológicas correspondem a interações possíveis da rede que são encontradas na coleção de documentos que descreve o conhecimento em biologia. Por outro lado, as novas interações correspondem a interações possíveis da rede que não são encontradas na coleção de documentos e que são inferidas pelo modelo a partir das interações já estabelecidas na rede. Por exemplo, consideremos as interações entre os fármacos  $f_1$  e  $f_2$  e as doenças  $d_1$  e  $d_2$  em uma sub-rede com espaço dimensional *fármaco*  $\times$  *doença* (Figura 3.2). Nessa sub-rede, o modelo encontra interações conhecidas em documentos da coleção textual que indicam o uso do fármaco  $f_1$  para o tratamento das doenças  $d_1$  e  $d_2$ . O modelo também encontra documentos relatando uma interação conhecida que indica o uso do fármaco  $f_2$  no tratamento da doença  $d_1$ . Então, os fármacos  $f_1$  e  $f_2$  provavelmente compartilham alguma característica comum responsável pela eficácia desses dois compostos no tratamento das doenças  $d_1$  e  $d_2$ . Assim, o modelo infere uma nova interação na sub-rede *fármaco*  $\times$  *doença* ligando o fármaco  $f_2$  e a doença  $d_2$ . A nova interação representa um novo uso para o fármaco  $f_2$ .

### 3.3 As Representações das Sub-Redes no Modelo

Cada sub-rede é representada por um grafo ponderado cujos pesos medem a evidência de interação entre entidades. Nesse grafo, os nodos são as entidades das categorias que formam o espaço dimensional  $s_i$  da sub-rede, as arestas representam as interações entre entidades de categorias distintas e a evidência de interação é



**Figura 3.2:** Representação de interações entre entidades em uma sub-rede com espaço dimensional *fármaco*  $\times$  *doença*.

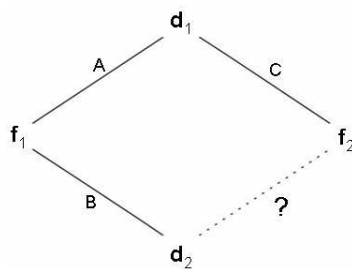
determinada pelo modelo de espaço vetorial, quando procuramos pela ocorrência das entidades na coleção de documentos. Assim, consideremos, por exemplo, um grafo ponderado que representa a sub-rede *fármaco*  $\times$  *doença* e que mostra as interações entre os fármacos  $f_1$  e  $f_2$  e as doenças  $d_1$  e  $d_2$  (Figura 3.3). Nesse grafo, o fármaco  $f_1$  trata a doença  $d_1$  com evidência de interação  $A$  e  $d_2$ , com evidência de interação  $B$ . Além disso, o fármaco  $f_2$  trata a doença  $d_1$  com evidência de interação  $C$ . Logo, o modelo atribui uma evidência de interação à nova interação que liga o fármaco  $f_2$  e a doença  $d_2$  cujo valor é determinado com base em  $A$ ,  $B$  e  $C$ .

O grafo de uma sub-rede  $n$ -dimensional é representado por uma matriz  $NW_n$ , ou simplesmente  $NW$ , que recebe as entidades biológicas das dimensões da sub-rede em suas linhas e colunas. Na matriz  $NW_n$ ,  $n$  é o número de dimensões usadas para construir a sub-rede  $n$ -dimensional. No trabalho, usamos indistintamente  $NW_n$ , ou  $NW$ , para nos referirmos tanto a uma sub-rede  $n$ -dimensional quanto à matriz que a representa.

Cada célula da matriz  $NW$  representa uma interação entre entidades da sub-rede e é referida por  $NW_{i,j}$ , onde  $i$  é uma linha qualquer da matriz e  $j$ , uma coluna. Além disso, o número de células  $NW_{i,j}$  na matriz  $NW$  de espaço dimensional  $s_i$  é dado pelo número de tuplas  $t_{s_i}$  de  $s_i$  (Equação 3.4). Na implementação do modelo, todas as sub-redes  $n$ -dimensionais são representadas através de uma matriz bidimensional. Portanto, o modelo transforma todos os espaços  $n$ -dimensionais em espaços bidimensionais.

O modelo transforma os espaços  $n$ -dimensionais em espaços bidimensionais, determinando o número de linhas e colunas que a matriz bidimensional  $NW$  deve possuir para armazenar todas as  $t_{s_i}$  interações possíveis da sub-rede  $n$ -dimensional (Equações 3.6, 3.7 e 3.8). Por exemplo, consideremos a representação de uma sub-rede quadridimensional em um espaço bidimensional (Figura 3.4). O espaço dimen-





**Figura 3.3:** Grafo ponderado representando interações da sub-rede *fármaco*  $\times$  *doença*.

sional dessa sub-rede é formado pelas categorias  $c_1$ ,  $c_2$ ,  $c_3$  e  $c_4$ . As linhas da matriz  $NW$  recebem o produto cartesiano das entidades de  $c_1$  e  $c_2$ . Por essa razão, o número de linhas de  $NW$  corresponde ao produto das cardinalidades de  $c_1$  e  $c_2$ . Por outro lado, as colunas da matriz recebem o produto cartesiano das entidades de  $c_3$  e  $c_4$ . Dessa forma, o número de colunas de  $NW$  é igual ao produto das cardinalidades de  $c_3$  e  $c_4$ .

$$linhas_{NW} = \prod_{x=1}^{\lfloor \frac{n}{2} \rfloor} card(c_{x,s_i}) \quad (3.6)$$

e

$$colunas_{NW} = \prod_{x=\lfloor \frac{n}{2} \rfloor + 1}^n card(c_{x,s_i}) \quad (3.7)$$

com

$$t_{s_i} = linhas_{NW} \times colunas_{NW} \quad (3.8)$$

onde  $n$  é o número de categorias que formam  $s_i$ .

### 3.4 A Construção das Sub-Redes

A construção das sub-redes de interações inicia com a identificação das interações conhecidas entre as entidades biológicas. As interações conhecidas são estabelecidas nas sub-redes consultando-se a coleção de documentos. Depois que todas as interações conhecidas são identificadas na coleção de documentos e são estabelecidas nas sub-redes, o modelo inicia o processo de inferência das novas interações. O modelo infere uma nova interação em uma sub-rede quando encontra uma *relação de transitividade* entre 3 interações já estabelecidas nessa sub-rede. Nós definimos que 3 interações estabelecidas em uma sub-rede estão em uma relação de transitividade quando satisfazem a condição  $(x, y)$  and  $(x, w)$  and  $(z, y) \rightarrow (z, w)$  o que significa que

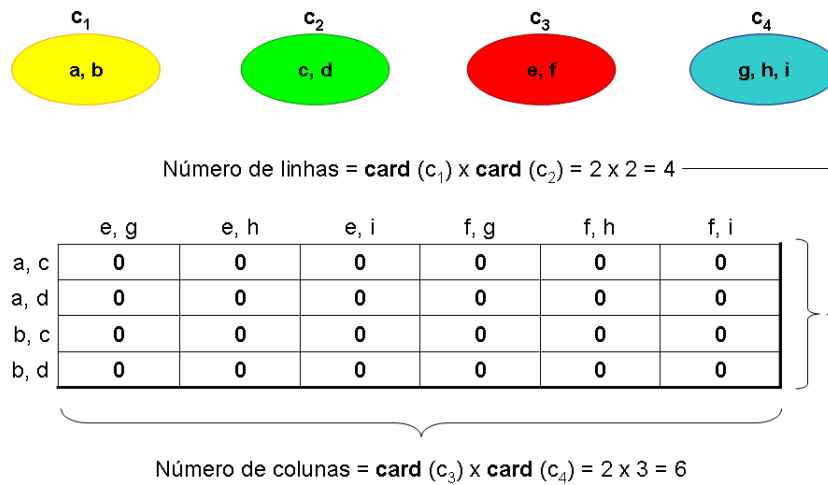


Figura 3.4: Representação de uma sub-rede quadridimensional em um espaço bidimensional.

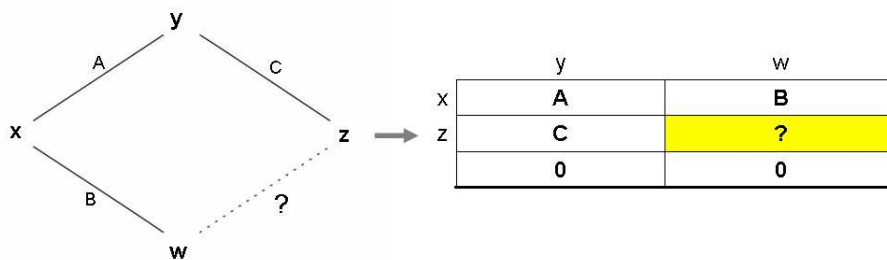


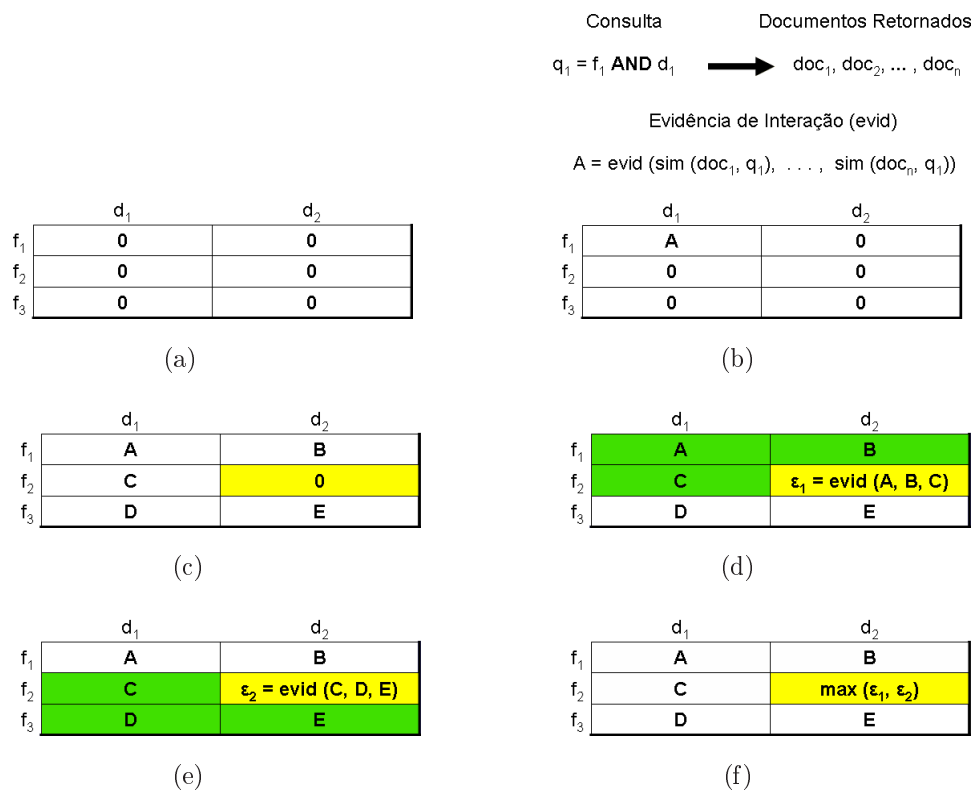
Figura 3.5: Relação de transitividade em uma matriz representando as interações de uma sub-rede bidimensional genérica.

"**IF** a entidade  $x$  interage com as entidades  $y$  e  $w$  **AND** a entidade  $z$  interage com a entidade  $y$  **THEN**  $z$  possivelmente também interage com a entidade  $w$ " (Figura 3.5). Assim, o modelo estabelece uma nova interação  $(z, w)$  em uma sub-rede sempre que encontra uma tripla de interações satisfazendo a condição imposta pela relação de transitividade. Dessa forma, o processo de inferência explora as atividades principais e secundárias que as entidades exercem no sistema biológico, para indicar as novas interações.

### 3.4.1 A Identificação das Interações Conhecidas

Na construção das sub-redes, nosso modelo inicia todas as células  $NW_{i,j}$  da matriz  $NW_n$  que representa cada sub-rede com o valor 0, indicando a ausência de interação entre entidades (Figura 3.6(a)). Nós usamos as entidades relacionadas em cada célula para formar uma consulta para o modelo de espaço vetorial. As entidades relacionadas em cada célula correspondem a uma interação possível na sub-rede e são unidas por um conectivo de conjunção  $E$  ( $AND$ ) para formar uma

consulta  $q_i \in Q$  (Figura 3.6(b)). Em nosso modelo, esta consulta representa a conjunção entre entidades de categorias distintas. A conjunção entre as entidades é muito importante, porque ela assegura que os documentos em que as entidades ocorrem não são ortogonais (Salton and McGill, 1983; Baeza-Yates and Ribeiro-Neto, 1999; Witten *et al.*, 1999). Esses documentos não podem ser ortogonais, porque eles necessariamente têm que possuir ocorrências de todas as entidades relacionadas na consulta. Em seguida, cada consulta é então submetida ao modelo de espaço vetorial. Assim, o modelo de espaço vetorial pesquisa a coleção de documentos, para identificar cada documento  $d_i \in D$  que contém todas as entidades relacionadas em cada célula da matriz  $NW$ .



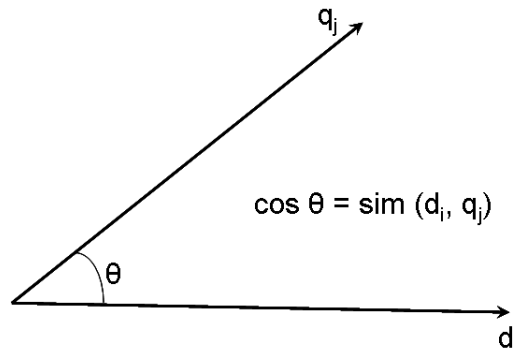
**Figura 3.6:** Construção das sub-redes.

O modelo de espaço vetorial atribui um peso para cada entidade da rede biológica que ocorre na coleção de documentos com base na estratégia de peso TFIDF (Equação 3.9) (Salton and McGill, 1983; Baeza-Yates and Ribeiro-Neto, 1999; Witten *et al.*, 1999). Esses pesos medem a importância da entidade para uma consulta da matriz e para um documento da coleção.

$$w_{x,i} = tf_{x,i} \times idf_x = \frac{f_{x,i}}{\max_{j,i} f_{j,i}} \times \log\left(\frac{N}{n_x}\right) \quad (3.9)$$

Na estratégia de peso TFIDF,  $w_{x,i}$  é o peso da entidade  $e_x \in E$  em um documento  $d_i$  da coleção  $D$ ,  $tf_{x,i}$  é a frequência normalizada da entidade  $e_x$  no documento  $d_i$  e indica a importância de  $e_x$  em  $d_i$ . O fator  $tf_{x,i}$  é calculado dividindo-se a frequência de  $e_x$  pela frequência da entidade que mais ocorre em  $d_i$ . Além disso,  $idf_x = \log(N/n_x)$  é a frequência inversa de  $e_x$  e indica a importância de  $e_x$  na coleção de documentos,  $f_{x,i}$  é a frequência da entidade  $e_x$  no documento  $d_i$ ,  $max_{j,i}$  é o número de vezes que a entidade mais frequente  $e_j \in E$  ocorre no documento  $d_i$ ,  $N$  é o número de documentos na coleção e  $n_x$  é o número de documentos na coleção em que a entidade  $e_x$  ocorre.

O fator  $tf_{x,i}$  da estratégia de peso TFIDF representa informação local da coleção de documentos, uma vez que ele interfere no valor do peso  $w_{x,i}$  de uma entidade biológica apenas levando em consideração a frequência local dessa entidade em um documento. O fator  $tf_{x,i}$  aumenta o peso de entidades que são frequentes em um documento e diminui o peso das entidades com menor frequência em um documento. Por outro lado, o fator  $idf_x$  representa informação global da coleção de documentos, uma vez que ele interfere no valor do peso  $w_{x,i}$  de uma entidade biológica conforme a importância global dessa entidade na coleção de documentos. O fator  $idf_x$  aumenta o peso de entidades que são pouco frequentes na coleção, entidades com baixo valor de  $n$ , e diminui o peso de entidades que são muito frequentes na coleção, entidades com alto valor de  $n$ . Dessa forma, o fator  $idf_x$  penaliza as entidades corriqueiras da coleção e promove as entidades raras.



**Figura 3.7:** Representação dos vetores de pesos de um documento  $d_i$  e de uma consulta  $q_j$ .

Cada consulta da matriz  $NW$  recebe um valor de similaridade para cada documento da coleção, com base no modelo de espaço vetorial. O valor dessa similaridade representa a relevância de um documento para uma consulta. Para calcular o valor dessa similaridade, as consultas e os documentos são descritos através de vetores de pesos das entidades que possuem (Figura 3.7). O valor desses pesos é determinado pela estratégia de peso TFIDF. Então, temos um vetor de pesos para cada

documento  $d_i$  e para cada consulta  $q_j$ . No modelo de espaço vetorial, a similaridade  $sim(d_i, q_j)$  entre um documento  $d_i$  e uma consulta  $q_j$  é definida pelo cosseno do ângulo formado entre esses dois vetores (Equação 3.10).

$$sim(d_i, q_j) = \frac{\sum_{x=1}^t (w_{x,i} \times w_{x,j})}{\sqrt{\sum_{x=1}^t (w_{x,i})^2} \times \sqrt{\sum_{x=1}^t (w_{x,j})^2}} \quad (3.10)$$

Na equação de similaridade definida no modelo vetorial,  $d_i$  é o documento  $i$  da coleção,  $q_j$  é a consulta  $j$  da matriz que representa uma sub-rede,  $t$  é o número de entidades biológicas da rede,  $w_{x,i}$  é o peso da entidade  $e_x$  no documento  $d_i$ ,  $w_{x,j}$  é o peso da entidade  $e_x$  na consulta  $q_j$ . Em nosso modelo, os pesos das entidades em uma consulta são sempre iguais a 1 ( $w_{x,j} = 1$ ).

A similaridade  $sim(d_i, q_j)$  corresponde à relevância do documento  $d_i$  para a conjunção de entidades biológicas que formam a consulta  $q_j$ . Portanto, a similaridade  $sim(d_i, q_j)$  quantifica o relacionamento das entidades que formam a consulta  $q_j$  e que foi encontrado no documento  $d_i$ .

A célula da matriz  $NW$  que liga as entidades de uma consulta  $q_j$  recebe um valor com base nas similaridades retornadas pelo modelo de espaço vetorial para essa consulta (Figura 3.6(b)). Dessa forma, esse valor corresponde ao peso da aresta que conecta essas entidades na sub-rede. Além disso, esse valor representa a evidência de interação  $\epsilon$  da interação conhecida entre essas entidades. A evidência de interação entre duas entidades é dada por uma função  $evid()$ . Várias estratégias podem ser empregadas na função  $evid()$ , para determinar o valor da evidência de interação a partir das similaridades retornadas pelo modelo de espaço vetorial. Por exemplo, o valor da evidência de interação  $\epsilon$  de uma interação conhecida entre a entidade  $e_a \in E$  da categoria  $a \in C$  e a entidade  $e_b \in E$  da categoria  $b \in C$  pode ser determinado pela soma das similaridades retornadas (Equação 3.11), pela média aritmética das similaridades retornadas (Equação 3.12) ou ainda equivaler à máxima similaridade retornada (Equação 3.13).

$$\epsilon = \sum_{i=1}^n sim(d_i, q_j) = \sum_{i=1}^n sim(d_i, e_a \text{ AND } e_b) \quad (3.11)$$

$$\epsilon = \frac{\sum_{i=1}^n sim(d_i, q_j)}{n} = \frac{\sum_{i=1}^n sim(d_i, e_a \text{ AND } e_b)}{n} \quad (3.12)$$

$$\epsilon = \max_{i=1}^n sim(d_i, q_j) = \max_{i=1}^n sim(d_i, e_a \text{ AND } e_b) \quad (3.13)$$

Nessas estratégias para determinar a evidência de interação das interações conhecidas,  $n$  é o número de documentos em que  $e_a$  e  $e_b$  co-ocorrem. Além disso,  $\text{sim}(d_i, q_j)$  é a similaridade do documento  $d_i$  para a consulta  $q_j = e_a \text{ AND } e_b$ , segundo o modelo de espaço vetorial.

Após pesquisar todas as consultas da matriz  $NW$ , temos todas as interações conhecidas da sub-rede. Contudo, algumas células da matriz permanecem com valor 0, indicando que algumas interações entre entidades não são mencionadas na coleção de documentos (Figura 3.6(c)). Essas células com valor 0 representam as potenciais novas interações entre as entidades biológicas que elas relacionam, porque a coleção de documentos não apresenta evidências de que essas interações entre entidades biológicas foram previamente mencionados por pesquisadores.

### 3.4.2 A Inferência das Novas Interações

No processo de inferência de novas interações em cada sub-rede, o modelo procura pelas células  $NW_{i,j}$  da matriz  $NW$  que permanecem com o valor inicial 0, após o processo de construção da sub-rede. Assim que o modelo encontra uma dessas células, ele procura por todas as triplas de interações já conhecidas da sub-rede que satisfazem a condição imposta pela relação de transitividade (Figuras 3.6(d) e 3.6(e)). A partir das triplas de interações conhecidas que satisfazem a relação de transitividade, o modelo determina um novo valor para a célula  $NW_{i,j}$ . Esse novo valor é registrado em  $NW_{i,j}$  e representa a evidência de interação de uma nova interação na sub-rede.

A evidência de interação de uma nova interação descoberta pelo modelo é calculada com base nas evidências de interação das 3 interações que satisfazem a condição imposta pela relação de transitividade. Assim como no cálculo do valor da evidência de interação das interações conhecidas, várias estratégias também podem ser empregadas na função  $\text{evid}()$  para determinar o valor da evidência de interação de uma interação nova. Por exemplo, o valor da evidência de interação  $\epsilon$  de uma nova interação entre a entidade  $e_a \in E$  da categoria  $a \in C$  e a entidade  $e_b \in E$  da categoria  $b \in C$  pode ser determinado pela soma das evidências de interação das interações que satisfazem a condição imposta pela relação de transitividade (Equação 3.14), pela média aritmética dessas evidências de interação (Equação 3.15) ou, ainda, equivaler ao maior valor dessas evidências de interação (Equação 3.16).

$$\epsilon = \epsilon_1 + \epsilon_2 + \epsilon_3 \tag{3.14}$$

$$\epsilon = (\epsilon_1 + \epsilon_2 + \epsilon_3)/3 \quad (3.15)$$

$$\epsilon = \max(\epsilon_1, \epsilon_2, \epsilon_3) \quad (3.16)$$

Nessas estratégias para determinar a evidência de interação de uma nova interação,  $\epsilon_1$ ,  $\epsilon_2$  e  $\epsilon_3$  são os valores das evidências de interação da primeira, segunda e terceira interação que formam a tripla que satisfaz a condição imposta pela relação de transitividade, respectivamente. Em nossa implementação corrente do modelo, empregamos a média aritmética dessas evidências de interação. Entretanto, quando há mais de uma tripla de interações que satisfazem a condição imposta pela relação de transitividade, podemos novamente aplicar várias estratégias para determinar o valor da evidência de interação da nova interação. Por exemplo, podemos somar o valor das evidências de interação calculado a partir das triplas (Equação 3.17), calcular a média aritmética dessas evidências de interação (Equação 3.18) ou escolher a maior evidência de interação obtida a partir dessas triplas (Equação 3.19).

$$\epsilon = \sum_{i=1}^n \epsilon_i \quad (3.17)$$

$$\epsilon = \frac{\sum_{i=1}^n \epsilon_i}{n} \quad (3.18)$$

$$\epsilon = \max_{i=1}^n \epsilon_i \quad (3.19)$$

Nessas estratégias,  $\epsilon$  é a evidência de interação escolhida para a nova interação. Além disso,  $\epsilon_i$  é a evidência de interação calculada a partir da tripla de interações  $i$  que satisfaz a condição imposta pela relação de transitividade. Por fim,  $n$  é o número de triplas de interações que satisfazem a condição imposta pela relação de transitividade.

Em nossa implementação corrente, ao encontrar mais de uma tripla de interações que satisfazem a condição imposta pela relação de transitividade, o modelo escolhe aquela que resulta no maior valor de evidência de interação para a nova interação (Figura 3.6(f)). A evidência de interação escolhida é registrada na célula  $NW_{i,j}$  que permanecia com o valor inicial 0 e representa uma nova interação inferida pelo modelo. Esse processo é conduzido até que todas as células  $NW_{i,j}$  da matriz  $NW$  recebam um valor diferente de 0 ou até que não haja mais interações conhecidas na sub-rede que satisfaçam a condição imposta pela relação de transitividade.

### 3.4.3 A Convergência em Sub-Redes

Em nosso modelo, podemos inferir novas interações entre entidades biológicas a partir de interações previamente descobertas. Dessa forma, novas descobertas levam ao surgimento de outras, desencadeando um processo evolutivo de descobrimento de novas interações. Isso é possível, porque as interações descobertas no processo de inferência podem formar triplas com interações já conhecidas da sub-rede ou com outras interações novas, satisfazendo a condição imposta pela relação de transitividade. Dessa forma, através da aplicação de várias iterações do processo de inferência, o modelo pode convergir a matriz  $NW$  de uma sub-rede  $n$ -dimensional para outra em que todas as interações possíveis entre entidades são apresentadas. Na iteração 0, o modelo descobre todas as interações conhecidas descritas na coleção de documentos. Na iteração 1, o modelo descobre todas as novas interações que podem ser inferidas com base nas interações conhecidas que foram encontradas na coleção de documentos durante a iteração 0. Na iteração 2, o modelo descobre as novas interações que podem ser inferidas com base no resultado das iterações 0 e 1. O modelo interrompe as iterações quando todas as células da matriz que representa uma sub-rede recebem valores diferentes de 0 ou quando não é mais possível encontrar interações que satisfaçam a relação de transitividade.

Durante o processo de inferência, nosso modelo pode encontrar entidades numa sub-rede que não interagem com as demais entidades dessa sub-rede. Essas entidades isoladas impedem que o modelo infira novas interações que as envolva na sub-rede. Na matriz  $NW$  de uma sub-rede, todas as células  $NW_{i,j}$  das linhas ou colunas relacionadas a essas entidades são iguais a zero, indicando a ausência de relacionamento dessas entidades e também impossibilitando a descoberta de novas interações relacionadas a essas entidades na sub-rede. Imaginemos, por exemplo, uma sub-rede bidimensional formada pelas categorias  $c_1 = \{e_1, e_2, e_3, e_4, e_5\}$  e  $c_2 = \{e_6, e_7, e_8, e_9, e_{10}\}$  (Figura 3.8(a)). Na iteração 0 do modelo ( $I_0$ ), as linhas da matriz  $NW$  dessa sub-rede recebem as entidades de  $c_1$  e as colunas, as entidades de  $c_2$  (Figura 3.8(b)). Depois de pesquisar a coleção de documentos e encontrar as interações conhecidas, o modelo descobre que nessa sub-rede as entidades  $e_5$  e  $e_9$  não interagem com as demais entidades. Então, a linha referente a  $e_5$  e a coluna referente a  $e_9$  são eliminadas da matriz  $NW$ , gerando uma versão compacta dessa matriz (Figura 3.8(c)). Na implementação de nosso modelo, as linhas e colunas de entidades isoladas das sub-redes são eliminadas da matriz  $NW$  para reduzir espaço de armazenamento em memória e tempo de processamento do processador durante a análise das sub-redes.

Em seguida, o modelo aplica várias iterações do processo de inferência na matriz  $NW$  da sub-rede  $n$ -dimensional, para inferir novas interações a partir de interações



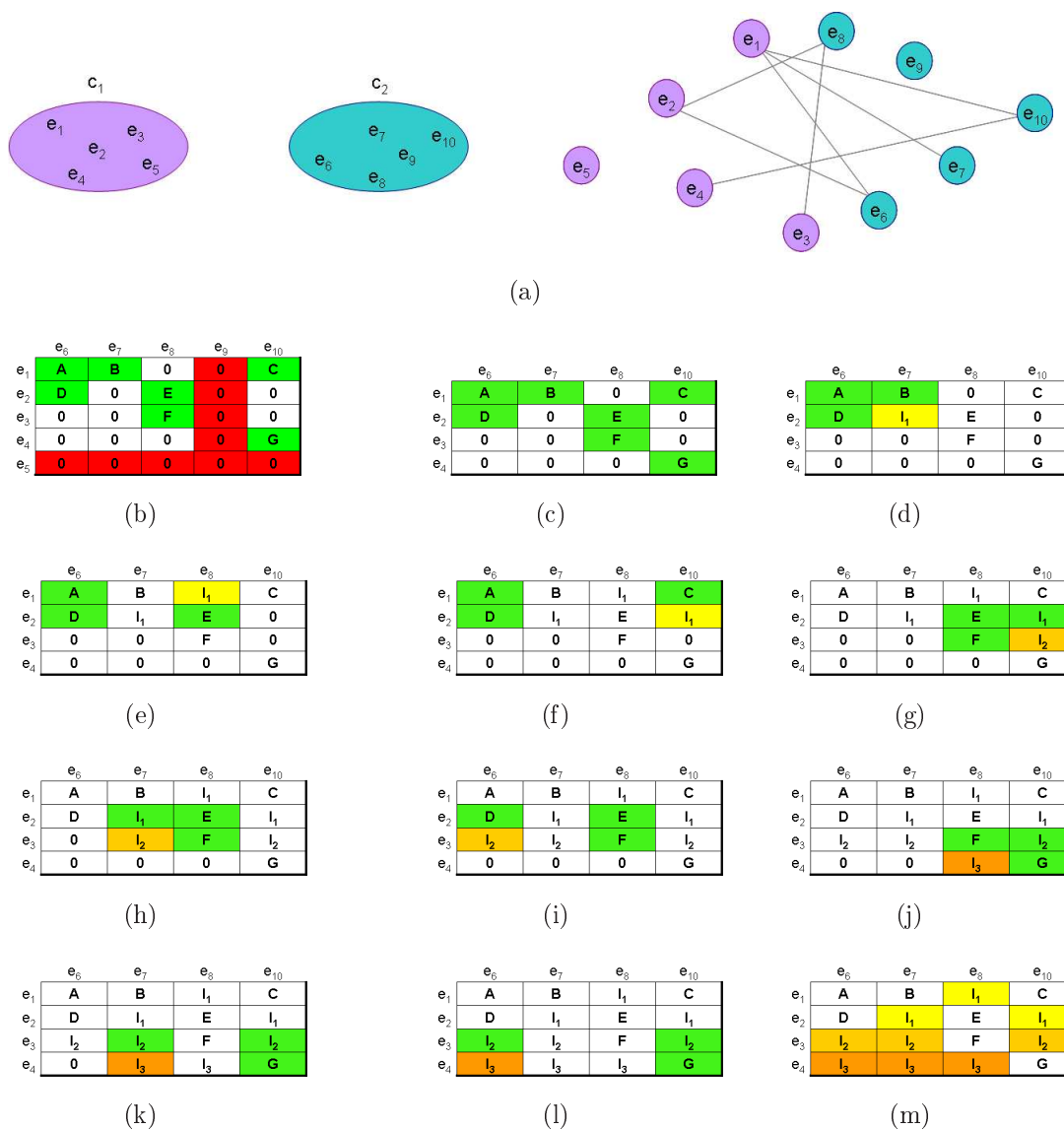


Figura 3.8: Convergência em uma sub-rede bidimensional.

previamente descobertas. Assim, o modelo desencadeia o processo evolutivo nas sub-redes que leva à inferência de novas interações biológicas à medida que novas descobertas são acrescentadas à rede e, conseqüentemente, ao conhecimento agregado à coleção de documentos. Na iteração 1 do modelo ( $I_1$ ), novas interações são descobertas no processo de inferência. Essas descobertas baseiam-se nas interações já conhecidas e encontradas em  $I_0$  (Figuras 3.8(d)-3.8(f)). Na segunda iteração do processo de inferência ( $I_2$ ), o modelo descobre novas interações, baseando-se nas interações conhecidas encontradas em  $I_0$  e nas inferidas em  $I_1$  (Figuras 3.8(g)-3.8(i)). Por fim, na terceira iteração do processo de inferência ( $I_3$ ), o modelo infere novas interações com base nas interações conhecidas encontradas em  $I_0$  e nas inferidas em

$I_1$  e  $I_2$  (Figuras 3.8(j)-3.8(l)). Kostoff (2008a) afirma que as novas interações inferidas a partir de interações conhecidas correspondem a inovações. Por outro lado, ele afirma que as novas interações inferidas em iterações subsequentes a partir de interações também novas correspondem a descobertas.

O modelo usa um fator para penalizar a evidência de interação das novas interações a cada aplicação do processo de inferência. O fator de penalidade é usado para dividir a evidência de interação atribuída às novas interações encontradas na iteração. Na primeira iteração do processo de inferência ( $I_1$ ), esse fator é igual a 1 e, por isso, não altera a evidência de interação calculada para as novas interações descobertas nessa iteração. A cada iteração subsequente do processo de inferência, o modelo aumenta o valor desse fator. Na implementação corrente do modelo, a evidência de interação é dividida pelo número da iteração. Assim, o modelo privilegia as descobertas das primeiras iterações do processo de inferência em detrimento das últimas, visto que as descobertas das últimas iterações só são possíveis com a identificação e comprovação das descobertas obtidas em iterações anteriores. O modelo finaliza a aplicação do processo de inferência quando todas as células  $NW_{i,j}$  da matriz  $NW$  recebem um valor diferente de 0 ou quando não é mais possível encontrar novas interações a partir das descobertas de iterações prévias (Figura 3.8(m)).

### 3.5 O Conjunto Solução Retornado pelo Modelo

Nosso modelo apresenta como resposta um conjunto de interações ordenadas em ordem decrescente de evidência de interação. O conjunto pode conter todas as interações da rede ou ser dividido em conjuntos menores para cada sub-rede. Dessa forma, podemos realizar consultas globais no modelo, para analisarmos as interações na rede completa ou podemos realizar consultas locais, para analisarmos as interações em cada sub-rede separadamente.

Para a formação do conjunto solução, cada célula  $NW_{i,j}$  de uma sub-rede  $n$ -dimensional  $NW$  com espaço dimensional  $s_p \in S$  é definida como

$$NW_{i,j} = \{e_{1,x}, \dots, e_{n,y}, \delta, \epsilon, \iota\} \quad (3.20)$$

onde  $1 \leq i \leq \text{linhas}_{NW}$ ,  $1 \leq j \leq \text{colunas}_{NW}$ ,  $n$  é o número de dimensões que constituem o espaço dimensional  $s_p$  de  $NW$ ,  $e_{k,w}$  é a  $k$ -ésima entidade biológica  $e_k \in E$  da categoria biológica  $c_w \in C_{s_p}$ , sendo  $C_{s_p} \subseteq C$  o conjunto de categorias que formam  $s_p$ . Além disso,  $\delta \subseteq D$  é um conjunto de documentos em que cada documento  $d_r \in \delta$  possui todas as entidades biológicas  $e_{k,w}$  de  $NW_{i,j}$ ,  $\epsilon$  é a evidência de interação entre as entidades biológicas de  $NW_{i,j}$  e  $\iota$  é a iteração do processo

de inferência em que a interação entre as entidades de  $NW_{i,j}$  foi estabelecida na sub-rede. Quanto temos  $\iota = 0$ , a célula  $NW_{i,j}$  representa uma interação conhecida da sub-rede que foi encontrada na coleção de documentos. Por outro lado, quando  $\iota > 0$ , a célula  $NW_{i,j}$  representa uma nova interação inferida na sub-rede.

Com base na definição de cada célula  $NW_{i,j}$  de uma sub-rede n-dimensional  $NW$ , podemos estabelecer uma condição de existência para as interações dessa sub-rede. Podemos afirmar que existe uma interação em uma sub-rede n-dimensional  $NW$  se, e somente se, a célula  $NW_{i,j}$  que representa essa interação na matriz de  $NW$  possui  $\epsilon > 0$ .

### 3.5.1 A Descrição do Conjunto Solução das Novas Interações

A partir da condição de existência estabelecida para as interações de uma sub-rede n-dimensional  $NW$ , podemos definir uma descrição para o conjunto solução contendo apenas as novas interações inferidas pelo modelo. Assim, o conjunto solução  $\sigma$  das novas interações de uma sub-rede  $NW$  de nosso modelo de inferência é definido como

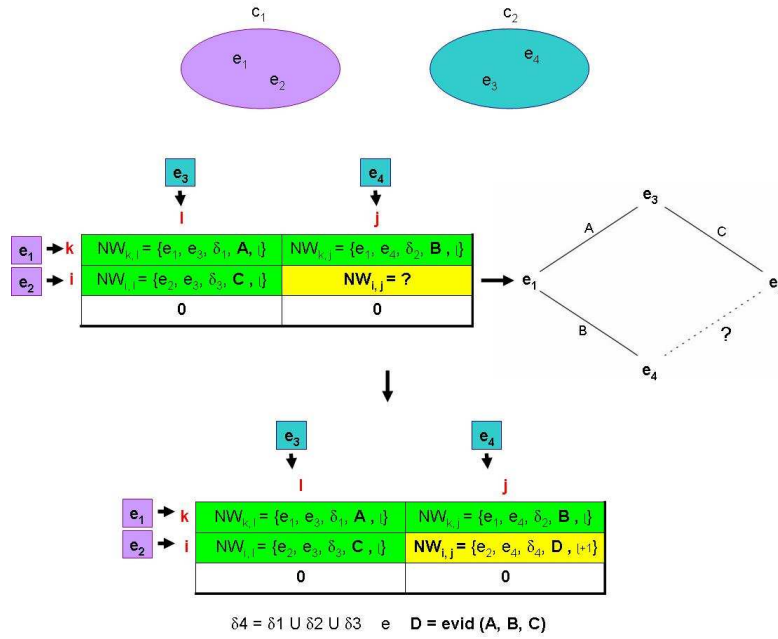
$$\sigma = \{NW_{i,j} | \exists NW_{k,l} \wedge \exists NW_{k,j} \wedge \exists NW_{i,l} \wedge \nexists NW_{i,j}\} \quad (3.21)$$

onde  $1 \leq i \leq linhas_{NW}$ ,  $1 \leq k \leq linhas_{NW}$  e  $i \neq k$  e  $1 \leq j \leq colunas_{NW}$ ,  $1 \leq l \leq colunas_{NW}$  e  $j \neq l$ .

Para exemplificarmos a formação do conjunto solução das novas interações com base nessa definição, consideremos uma sub-rede bidimensional formada pelas categorias  $c_1 = \{e_1, e_2\}$  e  $c_2 = \{e_3, e_4\}$  (Figura 3.9). Na matriz  $NW$  que representa essa sub-rede, as linhas recebem as entidades de  $c_1$  e as colunas, as entidades de  $c_2$ .

As interações já estabelecidas na sub-rede que satisfazem a relação de transitividade são apresentadas nas células  $NW_{k,l}$ ,  $NW_{k,j}$  e  $NW_{i,l}$ . Essas interações foram encontradas na coleção de documentos ou inferidas previamente pelo modelo, com base nos documentos presentes em  $\delta_1 \in D$ ,  $\delta_2 \in D$  e  $\delta_3 \in D$ , respectivamente. A evidência de interação de  $NW_{k,l}$ ,  $NW_{k,j}$  e  $NW_{i,l}$  é  $A$ ,  $B$  e  $C$ , respectivamente. Além disso, essas interações foram estabelecidas na sub-rede na iteração  $\iota$  do modelo. Uma nova interação entre as entidades  $e_2$  e  $e_4$  é representada na célula  $NW_{i,j}$  e é inferida com base na tripla de interações representadas nas células  $NW_{k,l}$ ,  $NW_{k,j}$  e  $NW_{i,l}$ .

O conjunto de documentos  $\delta_4 \in D$  que levam à descoberta da relação entre  $e_2$  e  $e_4$  em  $NW_{i,j}$  corresponde à união dos conjuntos de documentos de  $NW_{k,l}$ ,  $NW_{k,j}$  e  $NW_{i,l}$ ,  $\delta_4 = \delta_1 \cup \delta_2 \cup \delta_3$ . A evidência de interação  $\epsilon$  de  $NW_{i,j}$  é calculada com base na evidência de interação de  $NW_{k,l}$ ,  $NW_{k,j}$  e  $NW_{i,l}$ ,  $D = \epsilon = evid(A, B, C)$ ,



**Figura 3.9:** Representação das condições impostas pelo conjunto solução contendo apenas as novas interações.

sendo  $\text{evid}()$  a função escolhida para determinar a evidência de interação das novas interações. Além disso, a iteração em que o modelo infere  $NW_{i,j}$  é  $\iota + 1$ .

No conjunto solução, as novas interações são ordenadas pelo valor da evidência de interação. Dessa forma, o conjunto solução  $\sigma$ , que é a saída de nosso modelo de inferência, é uma lista de células  $NW_{i,j} = \{e_{1,x}, \dots, e_{n,y}, \delta, \epsilon, \iota\}$  ordenada em ordem decrescente da evidência de interação  $\epsilon$ .

As condições impostas pelo conjunto solução asseguram que, ao encontrar uma célula  $NW_{i,j} = \{e_{1,x}, \dots, e_{n,y}, \emptyset, 0, 0\}$  da matriz  $NW$ , o modelo procura a tripla de interações já estabelecidas na sub-rede  $NW_{k,l}$ ,  $NW_{k,j}$  e  $NW_{i,l}$  que satisfazem a condição imposta pela relação de transitividade na matriz  $NW$ .  $NW_{i,j}$  recebe uma evidência de interação calculada com base nas evidências de interação das interações em  $NW_{k,l}$ ,  $NW_{k,j}$  e  $NW_{i,l}$  que satisfazem a condição imposta pela relação de transitividade. Então, o modelo insere a nova interação  $NW_{i,j} = \{e_{1,x}, \dots, e_{n,y}, \delta, \epsilon, \iota\}$  no conjunto solução e ordena esse conjunto novamente em ordem decrescente de evidência de interação.

# Capítulo 4

## Metodologia

Nós podemos modelar um sistema biológico através da rede de interações entre as entidades que o constitui. Além disso, podemos classificar essas entidades biológicas em categorias, como eco-sistemas, organismos, órgãos, tecidos, células, organelas, genes, proteínas, alvos biológicos, doenças, fármacos, etc. Nessa representação, interações entre entidades do sistema correspondem às conexões da rede. Essas interações são conseguidas, por exemplo, quando o resultado de um experimento com microarranjos é analisado, ou quando é realizado um estudo da co-ocorrência dessas entidades biológicas em uma coleção de documentos que descreve o conhecimento em biologia, como artigos científicos, bulas de medicamentos, anotações sobre resultados de bancada, patentes, etc. A vantagem da representação das interações do sistema em uma rede é que podemos obter diversas informações sobre essas entidades, sobre as interações dessas entidades e ainda empregá-las em um processo de inferência. Através do processo de inferência, conseguimos prever novas interações entre as entidades do sistema com base em interações já estabelecidas na rede.

Nós criamos um modelo para construção de uma rede que representa as interações entre entidades de sistemas biológicos. Além disso, nós também desenvolvemos um processo de inferência que permite prever novas interações a partir das interações já estabelecidas na rede. Esse processo de inferência explora as atividades principais e secundárias que as entidades desempenham nos sistemas biológicos. No modelo, categorias biológicas distintas são combinadas para formar sub-redes de interações possíveis entre entidades. Cada categoria de uma sub-rede é considerada como uma dimensão dessa sub-rede. Todas as interações possíveis entre as entidades de uma sub-rede são pesquisadas em uma coleção de documentos que descreve o conhecimento em biologia, para determinar quais interações já são conhecidas na literatura. No modelo, apenas interações entre entidades de categorias distintas são consideradas na construção das sub-redes n-dimensionais.

## 4.1 O Modelo

Em nossos experimentos, as sub-redes que integram nossa rede de interações são formadas pelas combinações entre 4 categorias biológicas: alvo, doença, fármaco e gene (Tabela 4.1). Nós escolhemos essas categorias com base em duas razões. A primeira razão é a grande importância que as entidades dessas categorias possuem para as pesquisas na área de saúde. A segunda razão é que podemos alcançar diversas aplicações práticas para a sociedade com o estudo e análise das interações entre entidades dessas categorias.

**Tabela 4.1:** Número de entidades em cada categoria biológica.

Categoria	Entidades	Sítios <i>Web</i>
Alvo	23	The Free Dictionary (TFD, 2009), Therapeutic Target Database (TTD, 2007), Drug Bank (DrugBank, 2009)
Doença	22	Karolinska Institute (KI, 2009), Mayo Clinic (MC, 2009), Therapeutic Target Database, Drug Bank, Medline Plus (MedlinePlus, 2009)
Fármaco	22	Drugs.com (drugs.com, 2009), Patient.uk (patient.uk, 2009), Therapeutic Target Database, Drug Bank
Gene	20	Kyoto Encyclopedia of Genes and Genomes (KEGG, 2009), HUGO Gene Nomenclature Committee (HGNC, 2009), NCBI Entrez Gene (NCBI, 2009)
Total	87	

A categoria doença corresponde a um conjunto de estados possíveis de um sistema biológico (e.g. câncer, diabetes, AIDS, etc). A categoria fármaco corresponde a um conjunto de moléculas capazes de alterar o estado de um sistema biológico (e.g. tamoxifeno, galantamina, citrato de sildenafil, etc). As categorias alvo e gene correspondem a conjuntos de blocos construtores do sistema biológico. A categoria gene é o conjunto de blocos construtores responsável por gerar outros blocos construtores (e.g. complexo de histocompatibilidade principal classe 1, peptidilproil isomerase, etc). A categoria alvo é o conjunto de blocos construtores sobre os quais atuam os fármacos (e.g. epinefrina, cicloxigenase 1, acetilcolina, etc). Assim, um alvo pode ser, por exemplo, uma proteína, uma enzima ou mesmo um gene. Contudo, em nossos experimentos todas as categorias de uma sub-rede são conjuntos

disjuntos. Por exemplo, as categorias alvo e gene não possuem entidades em comum quando formam a sub-rede cujo espaço dimensional é  $alvo \times gene$ .

Combinando as 4 categorias biológicas tratadas nos experimentos, temos uma rede composta de 11 sub-redes, sendo 6 sub-redes de 2 dimensões, 4 sub-redes de 3 dimensões e 1 sub-rede de 4 dimensões (Tabela 4.2). Nessas sub-redes, as interações conhecidas são estabelecidas quando encontramos co-ocorrências das entidades que compõem as interações possíveis dessas sub-redes nos documentos da coleção. Por outro lado, as interações novas são aquelas inferidas a partir das interações já estabelecidas na rede.

**Tabela 4.2:** Sub-redes do sistema biológico.

Dimensões	Sub-rede	Espaço Dimensional
2	1	$alvo \times doen\c{c}a$
	2	$alvo \times f\acute{a}rmaco$
	3	$alvo \times gene$
	4	$doen\c{c}a \times f\acute{a}rmaco$
	5	$doen\c{c}a \times gene$
	6	$f\acute{a}rmaco \times gene$
3	7	$alvo \times doen\c{c}a \times f\acute{a}rmaco$
	8	$alvo \times doen\c{c}a \times gene$
	9	$alvo \times f\acute{a}rmaco \times gene$
	10	$doen\c{c}a \times f\acute{a}rmaco \times gene$
4	11	$alvo \times doen\c{c}a \times f\acute{a}rmaco \times gene$

Para formarmos o conjunto de entidades, inicialmente procuramos por doenças relacionadas a vários órgãos diferentes do corpo humano, como o coração, pulmão e cérebro (Apêndice A). Em seguida, procuramos por alguns dos fármacos comumente usados no tratamento dessas doenças. Depois, procuramos pelos genes e alvos biológicos relacionados a essas doenças ou sobre os quais agem os fármacos escolhidos. Nosso objetivo ao empregar essa estratégia era fornecer relações para o modelo que evidenciassem interações já conhecidas entre entidades, quando a coleta dos documentos e as buscas na coleção textual fossem realizadas. A partir do conjunto de entidades formado até então, o modelo poderia inferir novas interações, por exemplo, que indicassem o uso de um fármaco no tratamento de uma doença em um órgão para a qual o fármaco não tivesse sido originalmente planejado. Por fim, procuramos alguns alvos, doenças, fármacos e genes que não tivessem alguma relação aparente com as entidades já selecionadas. Nosso objetivo dessa vez era fornecer entidades para o modelo que pudessem ampliar a possibilidade de inferência de novas interações. De uma forma mais importante que a estratégia de seleção descrita, na escolha

das entidades mantivemos os mesmos dois princípios que nós levamos a selecionar as categorias biológicas. Dessa forma, a escolha das entidades também esteve sempre fundada na importância que elas têm para as pesquisas na área de saúde e nas aplicações práticas para a sociedade que podem ser alcançadas através do estudo e análise das interações entre elas.

A coleção de documentos que usamos em nossos experimentos é formada pela seção de reivindicação de 17.830 patentes coletadas no sitio *Web* de patentes dos Estados Unidos da América (USPTO) (USPTO, 2009). Essas patentes foram publicadas entre 01 de janeiro de 1976 e 31 de dezembro de 2005. Elas foram retornadas pelo sistema de busca do USPTO quando as entidades biológicas selecionadas para a construção de nossa rede foram nele pesquisadas. No sistema de busca do USPTO, essas pesquisas são representadas como *aclm/”entidade” and isd/1/1/1976 → 31/12/2005*. Nessa representação, *aclm* especifica que as entidades biológicas devem ser pesquisadas na seção de reivindicação das patentes, *entidade* é o nome de uma entidade biológica e *isd* especifica a data de publicação das patentes que se deseja pesquisar, respectivamente. Os nomes das entidades são colocados entre aspas duplas para indicar o modo de busca por frase no sistema de busca do USPTO. No sistema de busca do USPTO, a busca por frase é implementada como um casamento de cadeias de caracteres permitindo erros, ou casamento aproximado (Ziviani, 2007).

Em nossos experimentos, a data 1 de janeiro de 1976 é considerada como data inicial para a coleta das patentes, pois é a data de publicação a partir da qual o sistema de busca do USPTO disponibiliza consultas pela seção de reivindicação. Por outro lado, a data 31 de dezembro de 2005 é considerada como data final para coleta, pois podemos usar as patentes publicadas posteriormente a essa data, para validar as interações novas inferidas pelo modelo.

As relações entre as entidades são pesquisadas nessa coleção de documentos através do modelo de espaço vetorial que é usado como arcabouço algébrico para recuperação de informação. Um problema ao se extrair as relações entre entidades na coleção de documentos é a falta de informação semântica sobre essas entidades biológicas no texto das patentes. Dessa forma, corremos o risco de estabelecer associações falsas quando procuramos pela co-ocorrência dessas entidades na coleção de documentos (Fan *et al.*, 2006; GuoDong and Min, 2007; Berendt *et al.*, 2002; Jung and Gudivada, 1995; Schuffenhauer *et al.*, 2002; Cheung *et al.*, 2005). Esse risco existe, porque a simples co-ocorrência das entidades no texto das patentes não indica necessariamente que haja uma interação entre elas. Em consequência dessas interações incorretamente estabelecidas na rede, o modelo pode inferir novas interações que não são verdadeiras para o sistema biológico. Assim, o ruído causado pelas intera-



ções falsas extraídas da coleção de documentos resulta na propagação de interações espúrias pela rede.

O modelo estabelece associações falsas entre entidades biológicas na rede quando, por exemplo, encontra no texto das patentes sentenças escritas pelos redatores para relatar que as entidades não interagem entre si. Dessa forma, seria necessário criar estratégias para analisar essas sentenças e identificar as "não relações" entre entidades. Como um outro exemplo, as associações falsas também ocorrem quando os redatores das patentes constroem sentenças para diferenciar a invenção que está sendo patenteada de invenções prévias, ou também quando escrevem sobre uma entidade biológica diferente daquela usada na invenção apenas para exemplificar uma situação ou condição. No entanto, uma patente é um texto semi-estruturado com campos específicos e com propósitos bem definidos que podem auxiliar a recuperação de informação. A seção de reivindicação das patentes, por exemplo, é o campo que especifica o assunto que o aplicante considera como invenção, delimitando o escopo de proteção da patente (USPTO, 2009). Por essa razão, em nossos experimentos optamos por usar apenas a seção de reivindicação de cada patente durante a construção das sub-redes n-dimensionais. Assim, conseguimos reduzir as falsas associações entre entidades.

**Tabela 4.3:** Exemplo de interações entre entidades biológicas das categorias *fármaco* e *doença*. O exemplo foi construído a partir de uma coleção de documentos formada pela seção de reivindicação de 3 patentes biotecnológicas do USPTO.

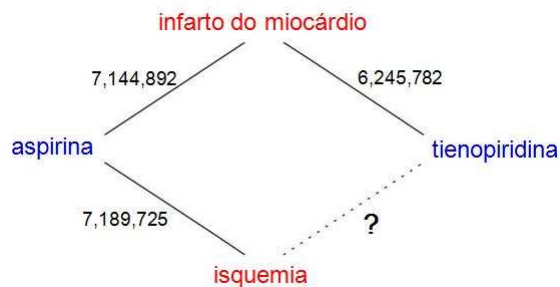
Fármaco	Patente	Doença	Disponibilidade
aspirina	7,144,892	infarto do miocárdio	✓
aspirina	7,189,725	isquemia	✓
tienopiridina	6,245,782	infarto do miocárdio	✓
tienopiridina	?	isquemia	?

### 4.1.1 As Interações Biológicas

Em nosso trabalho, assumimos que existem interações ainda não conhecidas entre entidades de um sistema biológico. Além disso, assumimos também que essas interações podem ser descobertas pela análise de coleções de documentos biológicos que descrevem interações já conhecidas entre essas entidades (Swanson, 1986, 1990; Smalheiser and Swanson, 1998; Swanson *et al.*, 2006; Weeber *et al.*, 2001; Wren *et al.*, 2004; Bruza and Weeber, 2008; Campillos *et al.*, 2008). Assim, desenvolvemos um modelo para construção e inferência em uma rede biológica com o objetivo de identificar essas novas interações. O modelo cria a rede biológica procurando in-

terações conhecidas na coleção de documentos. Em seguida, ele identifica relações de transitividade entre as interações estabelecidas na rede para inferir as interações novas.

Consideremos, por exemplo, uma coleção de documentos biológicos formada pelas seções de reivindicação das patentes 7,144,892, 7,189,725 e 6,245,782 do USPTO (Tabela 4.3). Na seção de reivindicação da patente 7,144,892 encontramos ocorrências do fármaco *aspirina* e da doença *infarto do miocárdio*. Na seção de reivindicação da patente 7,189,725 encontramos ocorrências do fármaco *aspirina* e doença *isquemia*. Por outro lado, na seção de reivindicação da patente 6,245,782 encontramos ocorrências do fármaco *tienopiridina* e doença *infarto do miocárdio*. Então, a partir dessas interações conhecidas e encontradas na coleção de documento, nosso modelo cria um grafo que mapeia a ocorrência dessas entidades biológicas na coleção de documentos (Figura 4.1). Com base nas conexões do grafo, o modelo encontra uma relação de transitividade entre as entidades biológicas. Segundo essa relação de transitividade o modelo infere que *aspirina* e *tienopiridina* podem compartilhar propriedades comuns que levam à indicação de *tienopiridina* no tratamento de *isquemia*. Então, o modelo estabelece uma nova aresta no grafo conectando o fármaco tienopiridina à doença isquemia.



**Figura 4.1:** Grafo de interação entre entidades biológicas das categorias *fármaco* e *doença*. Exemplo construído a partir de uma coleção de documentos formada pela seção de reivindicação de 3 patentes biotecnológicas do USPTO.

Nesse nosso exemplo, entretanto, não consideramos os sinônimos nem os termos relacionados das entidades tratadas, o que diminui a precisão do processo de inferência do modelo. Essa limitação pode ser exemplificada pela seção de reivindicação da patente 6,245,782 onde encontramos ocorrências da entidade *ataque isquêmico transitório* que é um nome relacionado à doença *isquemia*. Dessa forma, se esse nome relacionado fosse considerado na construção da rede, a interação entre *tienopiridina* e *isquemia* seria apontada pelo modelo como uma interação também já conhecida e não como uma nova interação. Por essa razão, em nosso modelo tratamos esse problema através da criação de grupos de nomes para cada entidade biológica.

Nós consideramos um total de 189 nomes de entidades em nossos experimentos

(Tabela 4.4). Elas estão dispostas em 87 grupos de nomes. O sistema de busca do USPTO não retornou patentes para as entidades de 2 grupos, o que resultou em 85 grupos indexados pelo modelo de espaço vetorial. Além disso, não foi possível encontrar interações para as entidades de 2 dos 85 grupos indexados, o que resultou em 83 grupos presentes na rede de interações. Cada grupo contém o nome da entidade biológica e seus nomes relacionados, como sinônimos. Por exemplo, consideramos diabetes mellitus tipo 2, diabetes não-insulino-dependente e diabetes tipo 2 como uma mesma entidade biológica da categoria *doença*. Quando procuramos pela ocorrência de cada entidade na coleção de documentos, consideramos todos os nomes do grupo dessa entidade. Assim, a frequência de cada um dos nomes do grupo é somada para obter a frequência do grupo. No entanto, apenas uma entidade de cada grupo é usada para representar o grupo na construção da rede biológica. Assim, para o exemplo anterior, apenas a doença diabetes tipo 2 é usada para representar seu grupo na rede. Algumas variações sintáticas dos nomes também são consideradas nos grupos. Por exemplo, "Alzheimer's disease" e "Alzheimer disease".

**Tabela 4.4:** Entidades biológicas consideradas nos experimentos e seus grupos de nomes relacionados.

	Number
Nomes de entidades	189
Grupos de nomes	87
Grupos indexados	85
Grupos na rede	83

Quando uma nova interação entre entidades é inferida pelo modelo, ele atribui uma evidência de interação a essa nova interação na rede. Em nossos experimentos, observamos a evidência de interação das interações conhecidas e também das novas interações. Com isso, conseguimos analisar o processo de inferência de nosso modelo. Além disso, ordenamos as interações da rede pelo valor da evidência de interação. Assim, foi possível verificar quais interações o modelo indica como os relacionamentos mais promissores a serem estudados. Nós conduzimos esse experimento em toda a rede e em cada sub-rede.

### 4.1.2 A Validação do Modelo

Validamos os resultados de nosso modelo considerando a data de publicação das patentes de nossa coleção de documentos. Nós criamos a rede de interação com base em todos os documentos da coleção. Então, observamos as interações conhecidas e novas que foram estabelecidas na rede. Em seguida, removemos da coleção de

documentos todas as patentes publicadas no ano mais recente, no caso o ano de 2005, e construímos a rede novamente. Assim, pudemos observar as interações conhecidas e novas que foram estabelecidas na rede através das patentes publicadas até o ano de 2004. Nós executamos esses passos, reconstruindo a rede para cada um dos 30 anos que são abrangidos pelas datas de publicação das patentes de nossa coleção desde 01/01/1976 até 31/12/2005.

Esse experimento foi realizado com o objetivo de verificar se interações novas encontradas até um dado ano de publicação seriam confirmadas por patentes publicadas em um ano de publicação mais recente. Assim, conseguimos demonstrar que nosso modelo é capaz de inferir interações que são confirmadas por patentes publicadas no USPTO. Além disso, nós também observamos a distribuição das patentes de confirmação ao longo dos *rankings* de resposta gerados por nosso modelo. Para esse experimento, dividimos os *rankings* gerados anualmente para cada sub-rede em seções de até 100 interações por seção. Nós chamamos essas seções de níveis dos *rankings*. Depois, observamos a distribuição das patentes de confirmação em cada um desses níveis. O objetivo desse experimento é verificar a capacidade de nosso modelo indicar as interações que possuem patentes de confirmação no topo do *ranking* de resposta. Dessa forma, pudemos demonstrar que o topo dos *rankings* de resposta retornados por nosso modelo possuem as melhores interações novas a serem estudadas.

Nós também realizamos pesquisas no sistema de busca do USPTO, procurando patentes que confirmassem as interações novas inferidas através das patentes publicadas até o ano de 2005. Esse experimento foi realizado submetendo as 100 primeiras interações novas apontadas pelo modelo em cada sub-rede no sistema de busca do USPTO. Nesse experimento, os *rankings* de resposta de cada sub-rede foram divididos em 10 seções de até 10 interações por seção. Como no experimento anterior, também chamamos essas seções de níveis. Em seguida, observamos a distribuição das patentes de confirmação nesses níveis. Para esse experimento, nós pesquisamos patentes publicadas de 01/01/1976 até 25/04/2009. A data 01/01/1976 foi escolhida com o objetivo de validar a tarefa de coleta. Assim, conseguimos verificar se alguma patente que relata as novas interações pesquisadas no sistema de busca do USPTO não tinha sido adquirida durante a tarefa de coleta. Por outro lado, a data 25/04/2009 foi escolhida por ser o dia de realização do experimento.

Além da validação baseada nas datas de publicação das patentes, também procuramos por artigos científicos publicamente disponíveis na *Web* que confirmassem as interações novas inferidas pelo modelo. Esse experimento foi realizado para a primeira nova interação indicada no topo do *ranking* de resposta de cada sub-rede.

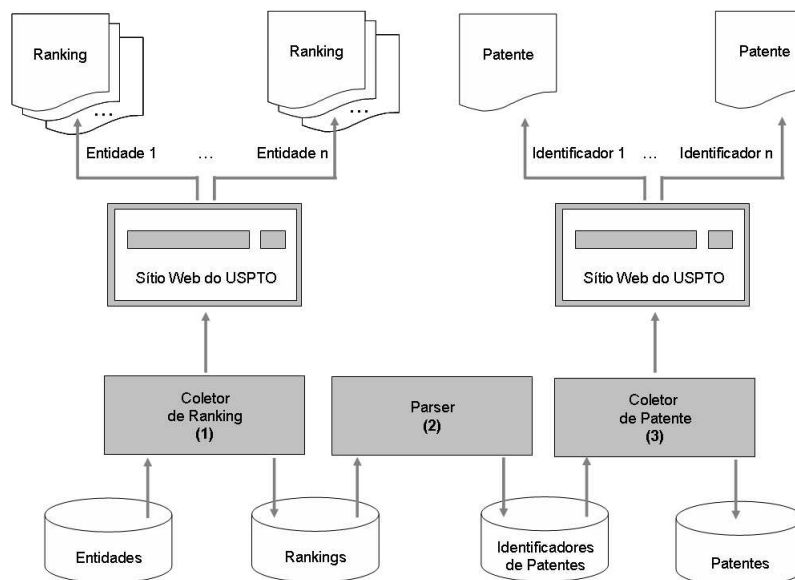


Figura 4.2: Tarefa 1: coleta de patentes do USPTO que descrevem o conhecimento em biologia.

## 4.2 O Sistema

Nós implementamos nosso modelo de inferência em um sistema chamado *BioSearch* (BioSearch, 2009). O sistema *BioSearch* foi completamente escrito na linguagem de programação Java, *Java Standard Edition programming language* (Java, 2009). Todos os algoritmos foram escritos usando o *Java Standard Edition Development Kit* 5.0 (JDK, 2009) no ambiente integrado de desenvolvimento NetBeans 5.5.1 (NetBeans, 2009). Para o desenvolvimento da aplicação *Web*, nós usamos a *Java Servlet platform technology* (Servlet, 2009). Para extrair a seção de reivindicação das patentes, nós usamos o analisador léxico e gerador de código para a linguagem Java Jflex (Jflex, 2007).

Nós usamos a *Java Database Connectivity Application Programming Interface* (JDBC, 2009), para realizar a comunicação do sistema *BioSearch* com o sistema gerenciador de banco de dados. Nós modelamos a base de dados usando o DBDesigner 4 (DBDesigner, 2009) e a criamos no sistema gerenciador de banco de dados PostgreSQL 8.3.7-1 (PostgreSQL, 2009) (Apêndice B - Figura 1). Além disso, nós usamos o PostgreSQL JDBC driver 8.2-504 JDBC 3 (Driver, 2009) para estabelecer as conexões da aplicação com o sistema gerenciador de banco de dados.

### 4.2.1 A Tarefa de Coleta

No sistema *BioSearch*, o módulo coletor engloba módulos específicos para fazer a interface com cada sítio *Web* considerado na formação da coleção de documentos.

Em nossos experimentos, desenvolvemos esses módulos para a coleta de patentes no sítio *Web* de patentes do USPTO (Figura 4.2).



**Figura 4.3:** Primeira página do *ranking* de patentes gerado pelo sistema de busca do USPTO. Esse *ranking* foi gerado ao pesquisarmos a entidade "type 2 diabetes" e restringirmos a data de publicação das patentes ao período de 1 de janeiro de 1976 a 31 de dezembro de 2005.

Em nossa implementação do módulo coletor, um módulo *coletor de ranking* (módulo 1) lê os nomes das entidades biológicas armazenados na base de dados e os submete para pesquisa no sistema de busca do USPTO. O coletor procura pelas patentes do USPTO publicadas entre 1 de janeiro de 1976 e 31 de dezembro de 2005 em cuja seção de reivindicação encontramos os nomes das entidades pesquisadas. Por exemplo, a patente 6,974,826 é coletada por ter sido publicada em 13 de dezembro de 2005 e por possuir ocorrências da entidade "diabetes tipo 2" em sua seção de reivindicação.

O casamento do nome das entidade na seção de reivindicação das patentes é realizado através de busca por frase. Esse método de busca foi escolhido para assegurar a ocorrências das entidades na seção de reivindicação das patentes adquiridas durante a tarefa de coleta. Também avaliamos buscas através dos conectivos OR e AND entre os termos que compõem o nome das entidades. Entretanto, esses conectivos geraram um grande número de patentes como resposta e nem todas essas patentes possuíam apropriadamente as ocorrências das entidades pesquisadas. Por outro lado, a busca por frase nos permitiu restringir o conjunto de patentes que satisfazem às consultas do coletor, eliminar patentes que não promovem a identifi-

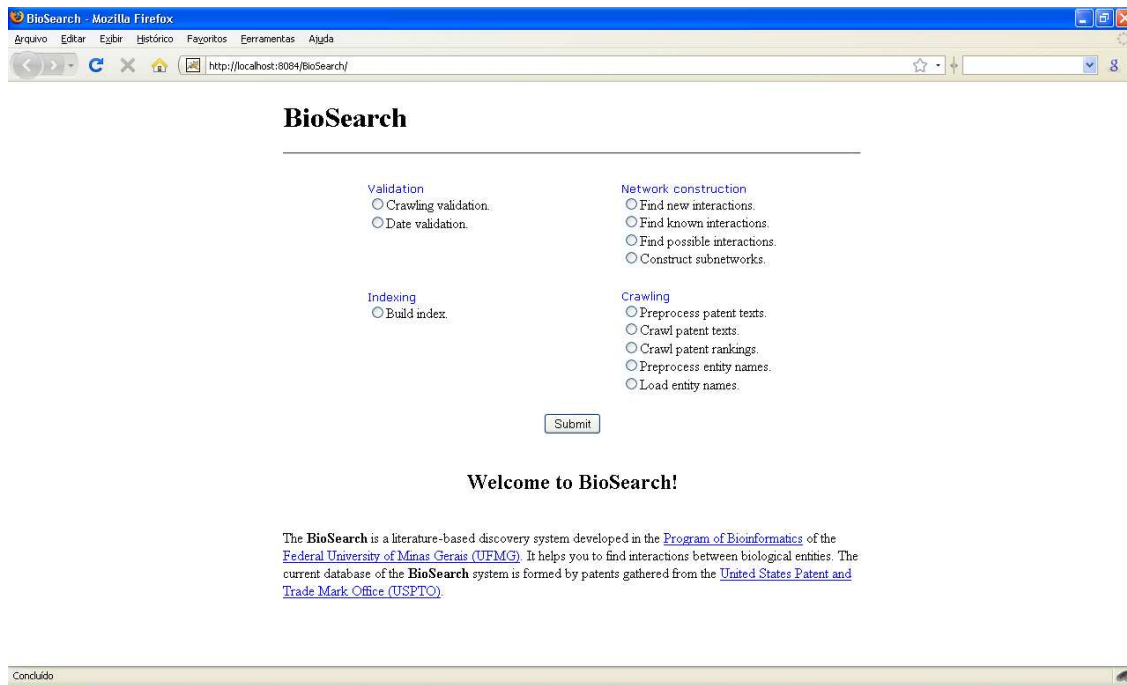


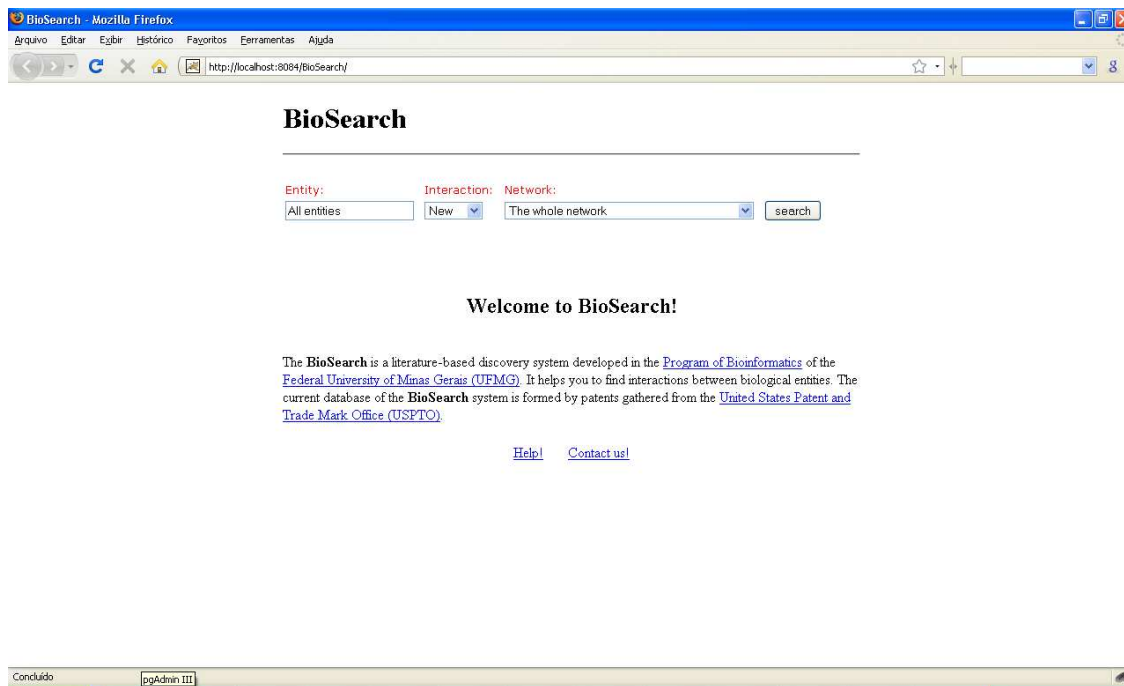
Figura 4.4: Interface *Web* para administradores do sistema *BioSearch*.

cação de interações biológicas e, conseqüentemente, diminuir o tamanho da coleção textual do sistema. Além disso, através da busca por frase, conseguimos mostrar que o modelo de inferência implementado é capaz de predizer novas interações entre entidades biológicas mesmo nesse cenário que é o mais restritivo, se comparado com aquele das buscas através dos conectivos OR e AND. Em conseqüência da redução no número de patentes retornadas pela busca por frase, conseguimos também melhorar o desempenho do sistema.

O sistema de busca do USPTO retorna um *ranking* de patentes para cada entidade pesquisada (Figura 4.3). Cada linha de um *ranking* possui o número identificador e título de uma patente relacionada à entidade pesquisada. Essas linhas são ordenadas pela data de publicação das patentes. O *ranking* é dividido em páginas de no máximo 50 linhas.

O módulo coletor de ranking de nosso sistema coleta todas as páginas dos *rankings* retornados pelo sistema de busca do USPTO e as armazena na base de dados. Um módulo de *parsing* (módulo 2) lê esses *rankings* na base de dados e extrai deles os números identificadores e títulos das patentes. Os números identificadores das patentes são passados por um filtro que remove números repetidos e números de patentes já coletadas anteriormente. Os números identificadores que restam após a tarefa de filtragem são armazenados na base dados. Em seguida, o módulo *coletor de patente* (módulo 3) lê os números identificadores das patentes que ainda não foram

coletadas no sistema de busca do USPTO a partir da base de dados. Ele pesquisa esses números identificadores no sistema de busca do USPTO, para que o texto das patentes sejam recuperados. Assim que o sistema de busca do USPTO retorna o texto de uma patente, o módulo coletor de patente armazena esse texto na base de dados.



**Figura 4.5:** Página inicial do sistema *BioSearch* para consultar as interações da rede biológica.

## 4.2.2 A Tarefa de Indexação da Coleção de Documentos

Na tarefa de indexação, o módulo indexador é responsável por construir um índice invertido. O índice invertido descreve a localização dos grupos de nomes das entidades biológicas na coleção de documentos. Esse índice sumariza a ocorrência dos grupos, tornando as busca necessárias no modelo de espaço vetorial mais eficientes (Salton and McGill, 1983; Baeza-Yates and Ribeiro-Neto, 1999; Witten *et al.*, 1999).

Em nossos experimentos, apenas as entidades biológicas são usadas para a formação do índice invertido. Os demais termos das patentes não são inseridos no índice invertido, porque não são usados na construção da rede de interações. Assim, conseguimos economizar espaço de armazenamento e tempo de processamento, construindo uma versão reduzida do índice invertido.



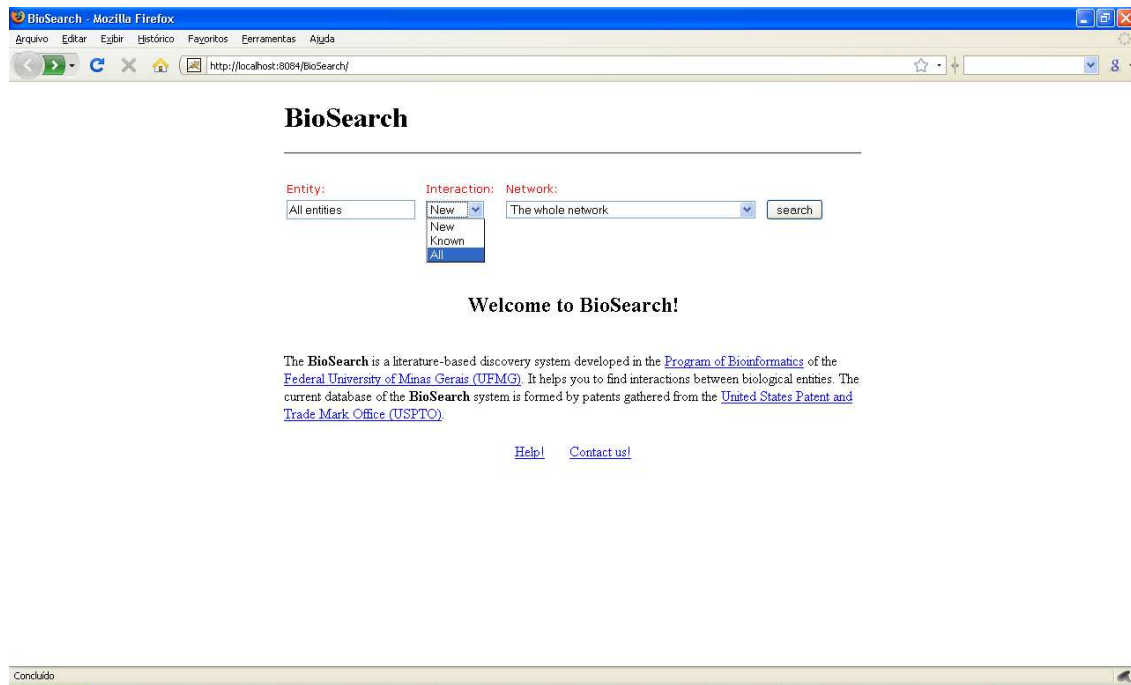


Figura 4.6: Pesquisa através do tipo de interação estabelecida na rede.

### 4.2.3 A Tarefa de Construção da Rede Biológica

Em nosso sistema, cada sub-rede  $n$ -dimensional é representada por um grafo ponderado (Ziviani, 2007). Nesse grafo, os nodos são as entidades das categorias que formam o espaço  $n$ -dimensional da sub-rede. Por outro lado, as arestas são as interações entre entidades de categorias distintas. Além disso, os pesos das arestas representam a evidência de interação entre as entidades ligadas pela aresta. Essa evidência de interação é determinada com base no modelo de espaço vetorial.

Inicialmente, as arestas das sub-redes são estabelecidas procurando-se as co-ocorrências das entidades na coleção de documentos. Denominamos essas arestas de interações conhecidas entre as entidades das sub-redes, pois são interações relatadas na base de patentes. Os grafos de todas as sub-redes são representados por matrizes que recebem as entidades biológicas das dimensões das sub-redes em suas linhas e colunas.

Após identificar todas as interações conhecidas de uma sub-rede, o sistema remove dessa sub-rede as entidades que não interagem com as demais. Essas entidades isoladas impedem que o sistema infira novas interações que as envolva na sub-rede. Na matriz da sub-rede, todas as células das linhas ou colunas relacionadas a essas entidades são iguais a zero, indicando a ausência de relacionamentos. Essa remoção permite reduzir espaço de armazenamento na memória e tempo de processamento do processador, durante a tarefa de inferência em cada sub-rede. Em seguida, o sistema

usa as interações que satisfazem a relação de transitividade em cada sub-rede, para inferir as interações novas. O sistema armazena todas as interações das sub-redes na base de dados para que possam ser consultadas e analisadas através da *Web*.

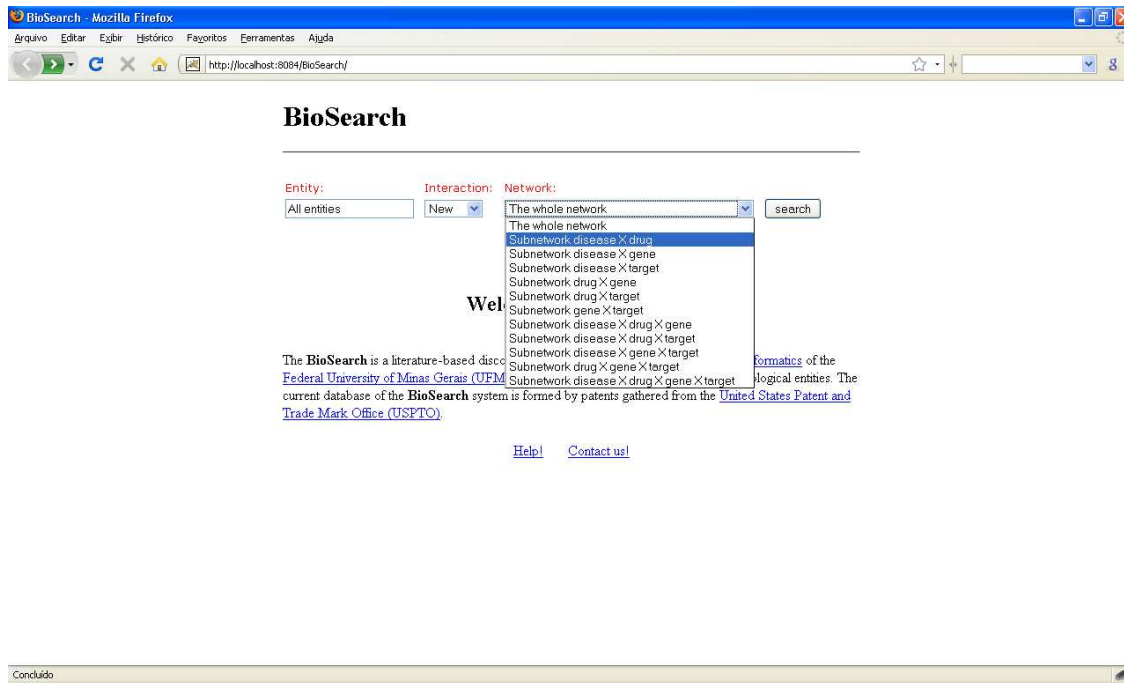


Figura 4.7: Pesquisa local na sub-rede *doença × fármaco*.

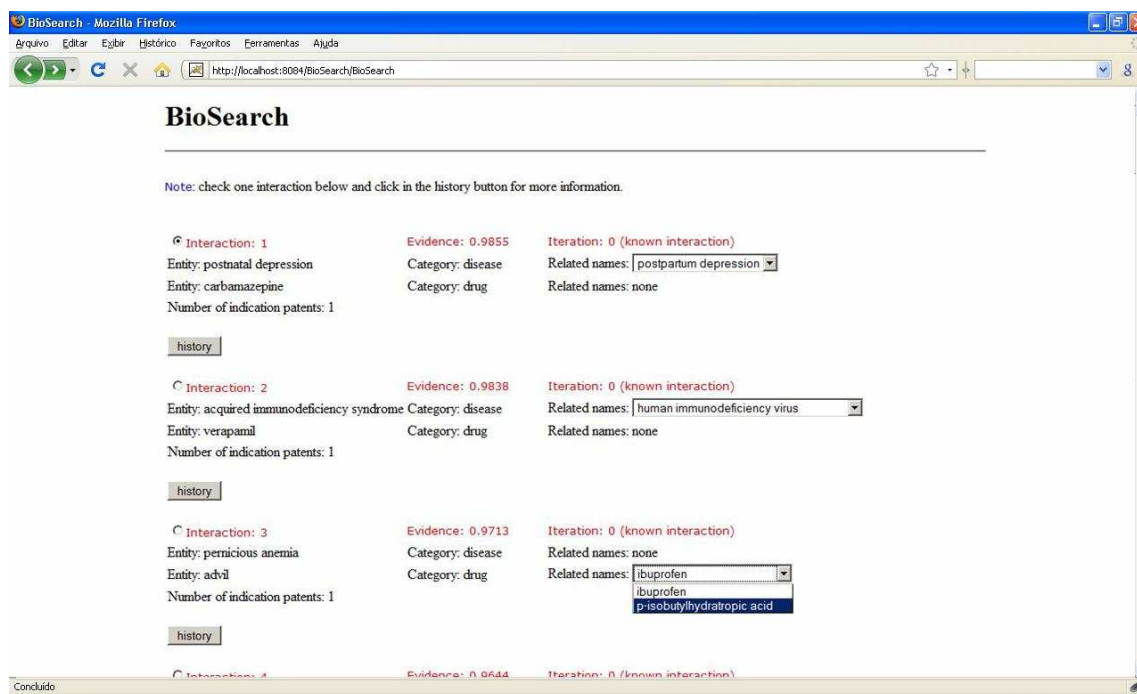
#### 4.2.4 A Tarefa de Consulta das Interações da Rede Biológica

Depois que todas as interações são estabelecidas na rede biológica, nós as disponibilizamos para consulta. Para isso, implementamos duas interfaces cliente que permitem o acesso ao sistema através da *Web*: uma destinada aos administradores do sistema e a outra, aos usuários comuns, como pesquisadores interessados em biotecnologia.

Na interface cliente para administradores, disponibilizamos os serviços relacionados a 3 das tarefas do sistema *BioSearch*: coleta, indexação e construção da rede biológica (Figura 4.4). Além disso, disponibilizamos também serviços relacionados à validação do modelo: validação através de coletas no *sítio Web* de patentes do USPTO e a validação baseada na data de publicação das patentes.

Implementamos a tarefa de consulta às interações da rede biológica em uma outra interface cliente para os usuários do sistema (Figura 4.5). Nessa interface, o usuário pode especificar se deseja pesquisar todas as entidades biológicas consideradas na construção da rede ou apenas algumas delas que lhe são de interesse. No entanto,

quando nenhuma entidade é especificada na interface, o sistema assume que todas as entidades da rede devem ser consideradas para satisfazer a consulta.



**Figura 4.8:** *Ranking* de resposta a uma consulta local por interações conhecidas na sub-rede *doença* × *fármaco*.

Na interface cliente para consulta das interações, os usuários podem especificar integralmente os nomes das entidade ou apenas parte deles. Por exemplo, consideremos que o usuário informe a seqüência de caracteres "epi" como nome da entidade biológica a ser pesquisada. Nesse caso, o sistema considera todas as entidades da rede que possuem essa seqüência de caracteres no nome, para satisfazer a consulta (e.g. *epilepsy*, *carbamazepine*, *epidermal growth factor receptor 2* e *epinephrine*).

Implementamos também na interface para usuários um serviço que permite a formulação de consultas com base no tipo das interações estabelecidas na rede. Assim, a interface permite que os usuários consultem as interações conhecidas da rede, as interações novas ou ambas (Figura 4.6). Além disso, criamos ainda duas formas de restringir o resultado das consultas nessa interface, através da especificação de consultas globais ou consultas locais. Nas consultas globais, os usuários pesquisam interações em toda a rede biológica. Por outro lado, nas consultas locais, os usuários pesquisam as interações das sub-redes separadamente (Figura 4.7).

Assim que o usuário submete sua consulta, o sistema procura na base de dados as interações que a satisfazem. O sistema retorna como resposta um conjunto de interações ordenadas em ordem decrescente de evidência de iteração (Figura 4.8). Chamamos esse conjunto de interações de *ranking* de resposta do sistema. Na página



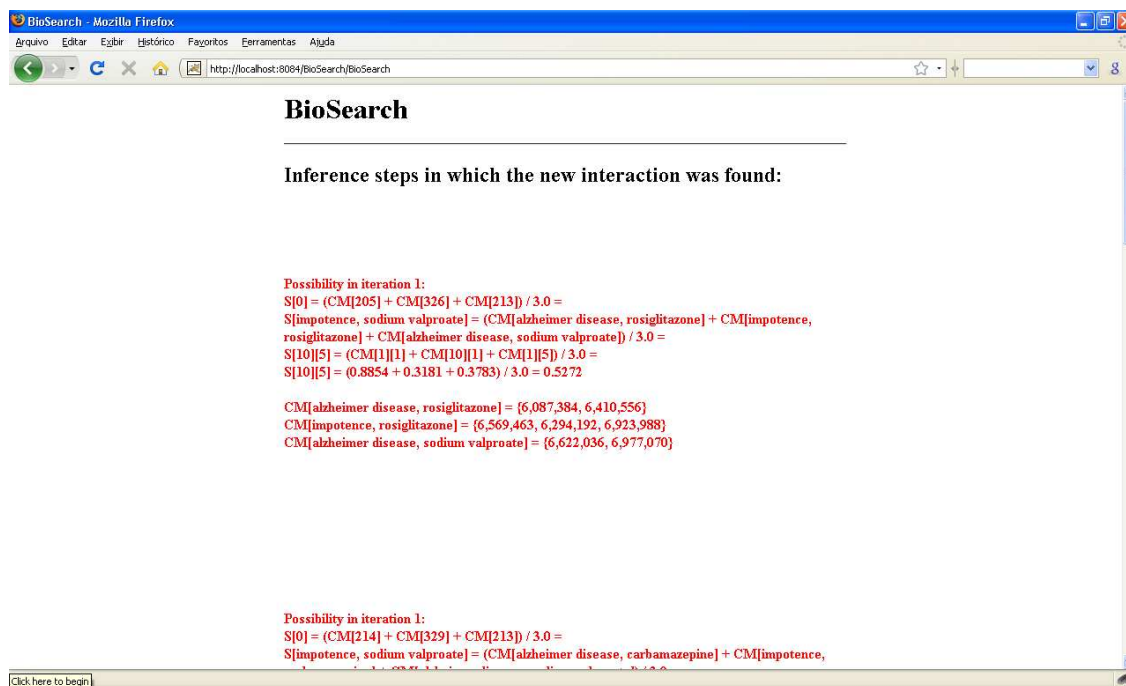
**Figura 4.9:** Patente que comprova uma interação conhecida e escolhida pelo usuário.

que exibe esse *ranking* de resposta, o sistema apresenta o nome das entidades e também os nomes relacionados a essas entidades. Nessa página, o sistema apresenta ainda as categorias das entidades, a iteração do processo de inferência em que cada interação foi estabelecida na rede, o valor da evidência de interação das interações, a posição das interações no *ranking* e o número de patentes que indicam as interações.

A página do *ranking* de resposta possui um botão que permite observar o histórico de formação das interações da rede biológica. Quando o usuário do sistema pressiona esse botão para uma interação conhecida, o sistema apresenta o número identificador das patentes que evidenciaram essa interação na coleção de documentos (Figura 4.9). Clicando no número identificador de qualquer uma dessas patentes, o usuário recebe a página da patente no sítio *Web* do USPTO. Por outro lado, quando o usuário do sistema pressiona esse botão para uma interação nova, o sistema apresenta todos os passos do processo de inferência que levaram à indicação dessa nova interação na rede (Figura 4.10).

#### 4.2.5 A Formação da Base de Dados

Durante a formação da base de dados que utilizamos na construção da rede biológica, fizemos vários refinamentos sucessivos que levaram à formação da base descrita em nossos experimentos e também a vários aperfeiçoamentos na implementação do sistema. Esses refinamentos foram realizados considerando 6 variáveis principais:



**Figura 4.10:** Passos do processo de inferência que levaram à indicação de uma nova interação.

1. As categorias biológicas.
2. O grupo de nomes das entidades.
3. O sítio *Web* para coleta das patentes.
4. A forma de utilização do nome das entidades, para gerar as consultas empregadas na coleta das patentes:
  - (a) A aplicação do conectivo OR entre os termos que compõem o nome das entidades.
  - (b) A aplicação do conectivo AND entre os termos que compõem o nome das entidades.
  - (c) A aplicação da busca por frase.
5. A seção das patentes considerada na tarefa de coleta: a coleta realizada através da busca das ocorrências das entidades em todo o corpo das patentes ou em alguma seção específica.
6. A data de publicação das patentes coletadas.

Dentre essas variáveis, entretanto, duas foram mantidas sempre constantes:

1. O sítio *Web* para a coleta das patentes, no caso o sítio *Web* do USPTO.
2. A data de publicação das patentes, fixada no período que vai de 01/01/1976 até 31/12/2005.

## As Categorias Biológicas

Para a formação de nossa base de dados, inicialmente consideramos entidades das categorias atividade biológica, doença, enzima, fármaco, gene e proteína. A categoria atividade biológica possuía entidades cujos nomes eram sentenças que descreviam a ação de enzimas e proteínas no organismo, por exemplo, *peptidase inhibitor*. Entretanto, a categoria atividade biológica foi eliminada dos experimentos por dois motivos. O primeiro motivo é que ela possuía entidades que contribuíam para a coleta de muitas patentes que não se relacionavam com biologia, visto que muitos termos no nome dessas entidades se relacionavam com patentes de diversas áreas de pesquisa. O segundo motivo é que as entidades dessa categoria também promoviam um grande número de patentes repetidas entre os *rankings* de resposta coletados. Então trabalhamos com as categorias doença, enzima, fármaco, gene e proteína. Posteriormente, unimos as categorias enzima e proteína, para dar origem a uma única categoria chamada alvo biológico. Assim, são consideradas 4 categorias na construção da rede: alvo biológico, doença, fármaco e gene.

## O Grupo de Nomes das Entidades

Inicialmente, não tínhamos realizado nenhum estudo sobre os nomes das entidades. Dessa forma, as entidades consideradas na base de dados não possuíam grupos de nomes. Por exemplo, os sinônimos de uma entidade eram todos tratados como entidades distintas. A consequência da falta desses grupos de nomes foi uma perda de precisão durante a coleta e indexação da coleção textual e também durante a construção da rede. Durante a coleta, muitas patentes referentes a uma entidade não eram coletadas, porque o nome usado para designar a entidade nas patentes não era o mesmo usado pelo sistema durante a tarefa de coleta. Durante a indexação, a contabilização da frequência das entidades na coleção textual também era afetada. Cada nome era tratado como relacionado a uma entidade distinta, desconsiderando-se, por exemplo, os sinônimos e acrônimos de cada entidade. Logo, a inexistência dos grupos de nomes influenciava o valor do peso das entidades e, conseqüentemente, o valor da evidência de interação. Durante a construção da rede, a falta dos grupos de nomes influenciou na identificação das interações conhecidas e na inferência das novas interações. Na identificação das interações conhecidas, muitas interações eram tratadas como distintas, quando na verdade relacionavam as mesmas entidades. Isso era muito visível principalmente nos *rankings* de resposta das sub-redes que eram muito longos, por apresentarem várias interações que se referiam às mesmas entidades. Um problema semelhante acontecia na inferência das novas interações. Muitas

interações novas eram indicadas na rede com base em interações conhecidas que na verdade não eram interações distintas. Assim, fizemos a identificação dos sinônimos, acrônimos e das variações sintáticas dos nomes das entidades e cada entidade passou a possuir seu grupo de nomes.

### A Forma de Utilização dos Nomes das Entidades na Coleta de Patentes

Na primeira tarefa de coleta das patentes, conectamos os termos que compunham o nome das entidades nas consultas através do operador lógico OR. O conectivo OR indica ao sistema de busca do USPTO que as patentes retornadas devem possuir pelo menos um dos termos que formam o nome das entidades. Assim, um número maior de patentes é considerado relevante para responder cada consulta. Dessa forma, o conectivo OR foi responsável pela recuperação de um grande número de patentes. Além disso, esse conectivo também foi responsável pela ocorrência de várias patentes repetidas entre os *rankings* de resposta coletados para entidades distintas.

A vantagem na utilização do conectivo OR foi que o grande número de patentes coletadas através dele nos permitiu uma maior cobertura dos temas relacionados a biologia. Submetendo a consulta *peptidase OR inhibitor*, por exemplo, conseguimos recuperar patentes em que ocorriam outros tipos de inibidores, como o *Acetyl-CoA carboxylase inhibitor* na patente 6,979,741.

A utilização do conectivo OR, por outro lado, apresentou dois inconvenientes. O primeiro inconveniente é que muitas das patentes retornadas através dele não se relacionavam ao assunto especificado nas consultas. Por exemplo, no caso anterior da consulta *peptidase OR inhibitor*, o sistema de busca do USPTO retorna a patente 6,980,897 que nem mesmo é uma patente relacionada à biologia. Entretanto, essa patente possui uma ocorrência do termo *inhibitor* na sentença "... according to a select position signal inputted from an inhibitor switch...". Como podemos observar, nessa patente o termo *inhibitor* não é empregado no mesmo sentido daquele expresso na consulta. Logo, esse problema semântico torna essa patente inadequada para o propósito de identificação de interações conhecidas que envolvam a entidade *peptidase inhibitor*. O segundo inconveniente na utilização do conectivo OR é o aumento do gasto com tempo de rede, com tempo de processamento e com espaço de armazenamento. O conectivo OR torna maiores os *rankings* de resposta retornados para as entidades, uma vez que sua utilização resulta em um número maior de patentes relevantes para cada consulta. A implicação disso é um gasto maior com tempo de rede no módulo de coleta do sistema, para coletar os *rankings* do USPTO e, depois, para recuperar as patentes listadas nesses *rankings*. Além disso, as várias patentes repetidas entre *rankings* de resposta do USPTO que são retornados para entidades

distintas geram um gasto maior com tempo de processamento. O gasto com tempo de processamento aumenta, porque aumenta o esforço necessário para identificar e eliminar as repetições. Por fim, as patentes que não se relacionam com o assunto especificado nas consultas resultam em um desperdício de espaço de armazenamento em memória e de tempo de processamento durante a construção da rede.

Posteriormente, substituímos o conectivo OR entre os termos que compunham o nome das entidades nas consultas pelo conectivo AND. A vantagem na utilização do conectivo AND é assegurar que o sistema de busca do USPTO retorne apenas as patentes que possuem todos os termos das consultas. No entanto, a utilização desse conectivo apresentou uma desvantagem: o conectivo AND não garante que a simples ocorrência de todos os termos da consulta no texto de uma patente realmente seja suficiente para indicar a existência do nome da entidade consultada nessa patente. Por exemplo, o sistema de busca do USPTO retorna a patente 6,979,691 quando realizamos a coleta de patentes relacionadas à entidade *cardiac ischemia*, através da submissão da consulta *cardiac AND ischemia*. Essa patente possui ocorrências de *ischemia* e de *cardiac glycosides*, mas não possui ocorrências de *cardiac ischemia*. Logo, a patente 6,979,691 não é adequada para a identificação de interações conhecidas que envolvam a entidade *cardiac ischemia*.

Por fim, substituímos o conectivo AND entre os termos que compunham o nome das entidades nas consultas pela busca por frase. Dessa maneira, o sistema de busca do USPTO só retorna patentes em que encontramos casamentos dos nomes das entidades consultadas. Nessa fase da construção da base de dados, observamos que a busca por frase é a mais adequada para a coleta das patentes e identificação das interações conhecidas na coleção de documentos. A busca por frase restringe melhor o grupo de patentes que satisfazem às consultas. Com isso, conseguimos uma diminuição do tamanho dos *rankings* de resposta do USPTO, do número de patentes repetidas entre *rankings* de resposta do USPTO que são retornados para entidades distintas, do número de patentes coletadas e, conseqüentemente, uma redução do custo com tempo de rede, com tempo de processamento e com armazenamento em memória.

No sistema de busca do USPTO, a busca por frase é implementada como um casamento de cadeias de caracteres permitindo erros. Por essa razão, durante a tarefa de coleta essa busca ainda acaba por retornar algumas patentes não desejadas. Entretanto, a quantidade dessas patentes indesejadas é bem menor que aquela conseguida através da utilização dos conectivos OR e AND. Nas tarefas de indexação e de construção da rede, utilizamos a busca por frase como um casamento de cadeias de caracteres não permitindo erros, ou casamento exato. O casamento exato é uma



forma mais restritiva, porém mais apropriada para identificar as interações conhecidas entre as entidades. Essa forma de busca é mais apropriada em nosso sistema, porque nos permite filtrar as patentes indesejadas, impedindo-as de contribuírem para a identificação de interações conhecidas.

### **A Seção das Patentes Considerada na Tarefa de Coleta**

Inicialmente, realizamos a tarefa de coleta considerando a ocorrência das entidades que formavam as consultas em todo o corpo das patentes. Entretanto, essa estratégia retornou um grande número de patentes inadequadas para a criação da rede. Essas patentes eram inadequadas, porque nelas as entidades utilizadas na coleta eram mencionadas em muitas seções sem, contudo, serem os objetos de proteção. Por essa razão, nas coletas seguintes, pesquisamos as entidades que formavam as consultas apenas na seção de reivindicação das patentes. A seção de reivindicação foi escolhida, porque é esta a seção mais relevante das patentes. O objetivo da seção de reivindicação é evidenciar claramente as particularidades da invenção ou criação, apresentando o escopo de proteção das patentes e especificando o objeto de proteção (USPTO, 2009).



# Capítulo 5

## Resultados

### 5.1 A Construção da Rede

Em nossos experimentos a rede biológica possui 266.528 interações possíveis. Pesquisando a coleção de documentos, nosso modelo identificou 1.027 interações conhecidas (Tabela 5.1). Com base nessas interações conhecidas, o modelo inferiu 3.195 novas interações, totalizando 4.222 interações em toda a rede. A sub-rede tridimensional *alvo*  $\times$  *doença*  $\times$  *fármaco* possui o maior número de interações conhecidas (199) e também o maior número de novas interações (958). A sub-rede bidimensional *doença*  $\times$  *fármaco* foi a que apresentou o segundo maior número de interações conhecidas (192). No entanto, a sub-rede bidimensional *alvo*  $\times$  *doença* foi a que apresentou o segundo maior número de interações novas (346).

**Tabela 5.1:** Descrição da rede de interações biológicas.

Dimensões	Sub-rede	Espaço Dimensional	Interação		Total
			Conhecida	Nova	
2	1	<i>alvo</i> $\times$ <i>doença</i>	138	346	484
	2	<i>alvo</i> $\times$ <i>fármaco</i>	105	294	399
	3	<i>alvo</i> $\times$ <i>gene</i>	50	175	225
	4	<i>doença</i> $\times$ <i>fármaco</i>	192	270	462
	5	<i>doença</i> $\times$ <i>gene</i>	76	184	260
	6	<i>fármaco</i> $\times$ <i>gene</i>	38	130	168
3	7	<i>alvo</i> $\times$ <i>doença</i> $\times$ <i>fármaco</i>	199	958	1.157
	8	<i>alvo</i> $\times$ <i>doença</i> $\times$ <i>gene</i>	55	269	324
	9	<i>alvo</i> $\times$ <i>fármaco</i> $\times$ <i>gene</i>	34	76	110
	10	<i>doença</i> $\times$ <i>fármaco</i> $\times$ <i>gene</i>	71	304	375
4	11	<i>alvo</i> $\times$ <i>doença</i> $\times$ <i>fármaco</i> $\times$ <i>gene</i>	69	189	258
Total			1.027	3.195	4.222

Nós observamos que em todas as sub-redes o número de interações inferidas pelo modelo é bem maior que o número das interações já conhecidas. Na sub-rede tridimensional *alvo*  $\times$  *fármaco*  $\times$  *gene*, por exemplo, o número de interações conhecidas é o menor de toda a rede (34). No entanto, o número de interações novas é maior que o dobro das conhecidas (76). Esses resultados indicam que, através de descobertas já realizadas e patenteadas sobre as entidades biológicas consideradas em nossos experimentos, muito conhecimento ainda pode ser explorado analisando-se conexões implícitas entre os documentos dessa literatura.

**Tabela 5.2:** Ordenação das sub-redes de acordo com as melhores interações inferidas pelo modelo.

Dimensões	Sub-rede	Espaço Dimensional	Nova Interação	Evidência de Interação	
2	3	alvo gene	adrenaline androgen receptor	0.9757	
	5	doença gene	acquired immunodeficiency syndrome transforming growth factor, beta 1	0.9738	
	2	alvo fármaco	cyclooxygenase 2 verapamil	0.9597	
	4	doença fármaco	erectile dysfunction divalproex	0.9470	
	1	alvo doença	cyclic-gmp phosphodiesterase arrhythmia	0.9272	
	6	fármaco gene	ciclosporin androgen receptor	0.8211	
	3	7	alvo doença fármaco	adrenaline alzheimer dementia acetylsalicylic acid	0.8807
9		alvo fármaco gene	lymphotoxin acarbose apolipoprotein a 1	0.8723	
8		alvo doença gene	choline acetylase parkinson disease apolipoprotein e	0.8695	
10		doença fármaco gene	gout hydrochlorothiazide endothelin 1	0.8357	
4		11	alvo doença fármaco gene	hmg-coa reductase breast adenocarcinoma tamoxifen ppar-gamma	0.7826

Nós ordenamos as novas interações de cada sub-rede pelo valor da evidência de interação. Exceto quando expresso de modo contrário, todos os resultados apresen-

tados neste trabalho que retratam a evidência de interação entre as entidades foram obtidos considerando a média aritmética das similaridades retornadas pelo modelo de espaço vetorial, para calcularmos a evidência de interação das interações conhecidas. Nesse experimento, observamos que a sub-rede bidimensional *alvo*  $\times$  *gene* foi a que apresentou a interação com maior evidência de interação em toda a rede (Tabela 5.2). Essa interação indica um relacionamento entre o hormônio adrenalina e o gene receptor de androgênio com evidência de interação de 0,9757. Entre as sub-redes tridimensionais, a sub-rede *alvo*  $\times$  *doença*  $\times$  *fármaco* foi a que apresentou a interação com maior evidência de interação. Essa interação indica um relacionamento entre o hormônio adrenalina, o mal de Alzheimer e o ácido acetilsalicílico com evidência de interação de 0,8807. O melhor resultado da rede quadridimensional indica um relacionamento entre a enzima HMG-CoA redutase, o cancer de mama, o fármaco tamoxifeno e o gene ppar-gamma com evidência de interação de 0,7826.

### 5.1.1 O Espaço de Busca de Novas Interações

Em nosso modelo, os resultados de uma sub-rede com maior número de dimensões são mais apurados, porque restringem melhor o espaço de busca de interações novas (Figura 5.1). Por exemplo, consideremos uma pessoa interessada em realizar uma pesquisa sobre possíveis novas interações envolvendo o ácido acetilsalicílico, a aspirina. No início da pesquisa, não é possível prever as interações mais promissoras, sem antes realizar experimentos que as revelem. O problema é que o total de interações possíveis entre as entidades torna o espaço de busca das novas interações muito grande, dificultando a realização desses experimentos, principalmente se consideramos os custos e o tempo necessário para realizá-los. Entretanto, os resultados mostram que nosso modelo é capaz de prever as melhores interações entre entidades e que um maior número de dimensões permite reduzir o espaço de busca das novas interações.

Imaginemos que numa busca inicial no sistema *BioSearch* esse pesquisador decida analisar as interações da aspirina com os alvos HMG-CoA redutase, *cachectin* e acetilcolinesterase (linhas pontilhadas na figura 5.1). Para essas consultas, o sistema informa que a interação entre a aspirina e o alvo HMG-CoA redutase já é conhecida, possui evidência de interação  $\epsilon_2 = 0.7607$  ( $\epsilon_n$  é o valor da evidência de interação na sub-rede n-dimensional,  $n = 2, 3, 4, \dots$ ) e foi encontrada em 26 patentes (e. g. patentes 6,967,212 e 6,875,782). Essa pessoa é informada também de que a interação entre a aspirina e o alvo *cachectin* também já é conhecida, possui evidência de interação  $\epsilon_2 = 0.9743$  e foi encontrada na patente 6,391,832. O sistema informa ainda que a interação entre a aspirina e o alvo acetilcolinesterase é nova e possui evidência de

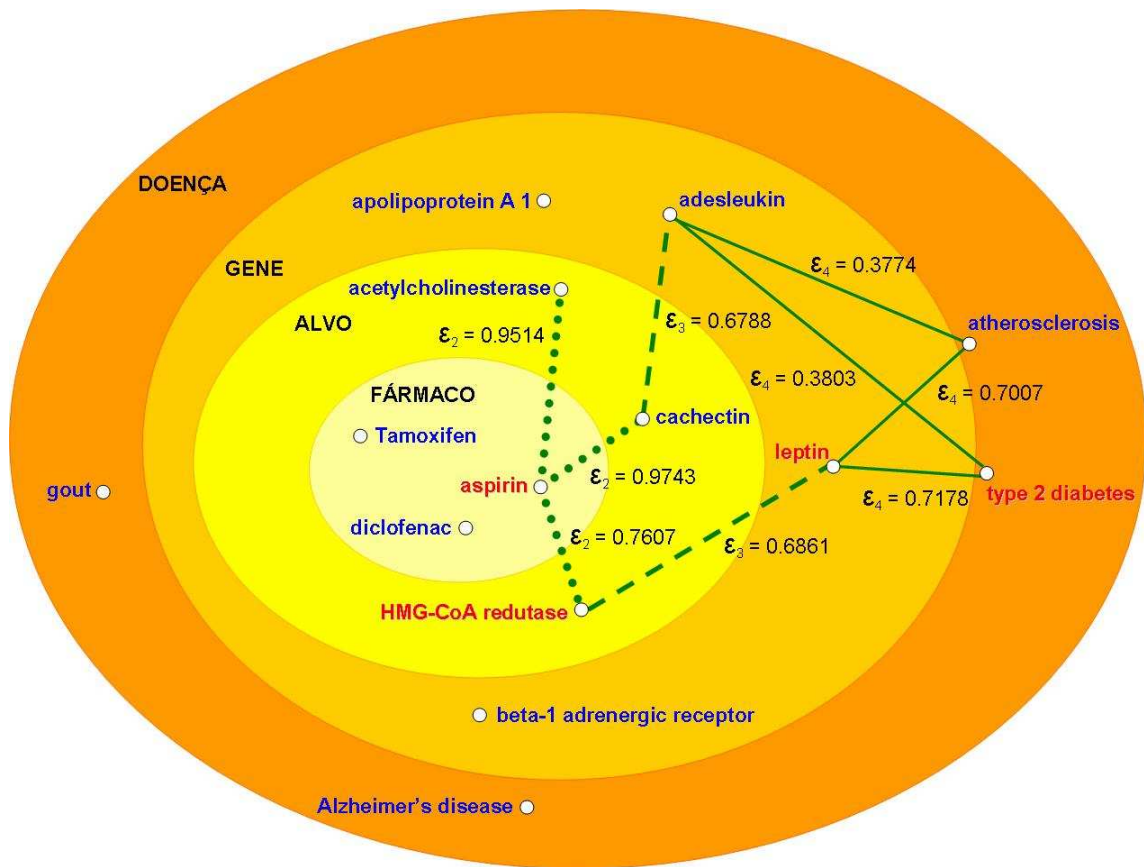


Figura 5.1: Representação do espaço de busca para algumas das interações possíveis envolvendo o fármaco aspirina.

interação  $\epsilon_2 = 0.9514$ .

Nesse ponto da análise, a melhor alternativa para essa pessoa seria realizar sua pesquisa sobre a interação entre aspirina e o alvo acetilcolinesterase, visto que essa interação tem evidência de interação alta e que as duas outras interações já são conhecidas. No entanto, o espaço de busca de novas interações para essas entidades ainda é muito grande, pois elas podem interagir com diversas outras entidades. Assim, somente analisando as interações dessas entidades em sub-redes com mais dimensões, é possível selecionar as interações mais relevantes.

Continuando sua análise com o objetivo de refinar os resultados apresentados pelo modelo, o pesquisador decide incluir a dimensão gene às suas consultas. Desse modo, o modelo promove uma redução do espaço de busca e aumenta as chances de sucesso da pesquisa. O sistema informa que 2 interações novas são conseguidas nesse espaço tridimensional: a interação entre aspirina, o alvo HMG-CoA redutase e o gene leptina é nova e possui evidência de interação  $\epsilon_3 = 0.6861$  e a interação entre aspirina, o alvo *cachectin* e o gene *aldesleukin* é também nova e possui evidência de interação  $\epsilon_3 = 0.6788$  (linhas tracejadas na figura 5.1). O pesquisador descobre

nesse momento que a interação entre apirina e o alvo acetilcolinesterase revela-se menos interessante, uma vez que nenhuma interação entre essas entidades é formada na sub-rede tridimensional.

Buscando conexões na sub-rede quadridimensional, o pesquisador descobre 4 novas interações possíveis, todas elas indicando a interação das entidades consideradas até o momento com as doenças diabetes tipo 2 e aterosclerose. O sistema aponta que as interações entre aspirina, *cachectin*, *aldesleukin* e as doenças diabetes tipo 2 ou aterosclerose são menos evidentes, porque são descobertas na iteração 2 do modelo. Com base nas afirmações de Kostoff (2008a), essas interações podem corresponder a novas descobertas em biotecnologia, pois foram inferidas através de outras interações novas estabelecidas na rede durante a iteração 1 do modelo. No entanto, o sistema revela que as interações entre apirina, HMG-CoA redutase, leptina e as doenças diabetes tipo 2 ou aterosclerose possuem evidência de interação alta, porque são descobertas na iteração 1 do modelo. Novamente com base nas afirmações de Kostoff (2008a), essas interações podem corresponder a inovações em biotecnologia, pois são inferidas unicamente através de interações já conhecidas. O sistema mostra ainda que a conexão entre apirina, HMG-CoA redutase, leptina e diabetes tipo 2, com evidência de interação  $\epsilon_4 = 0.7178$ , revela-se a mais promissora para a pesquisa (linhas contínuas na figura 5.1).

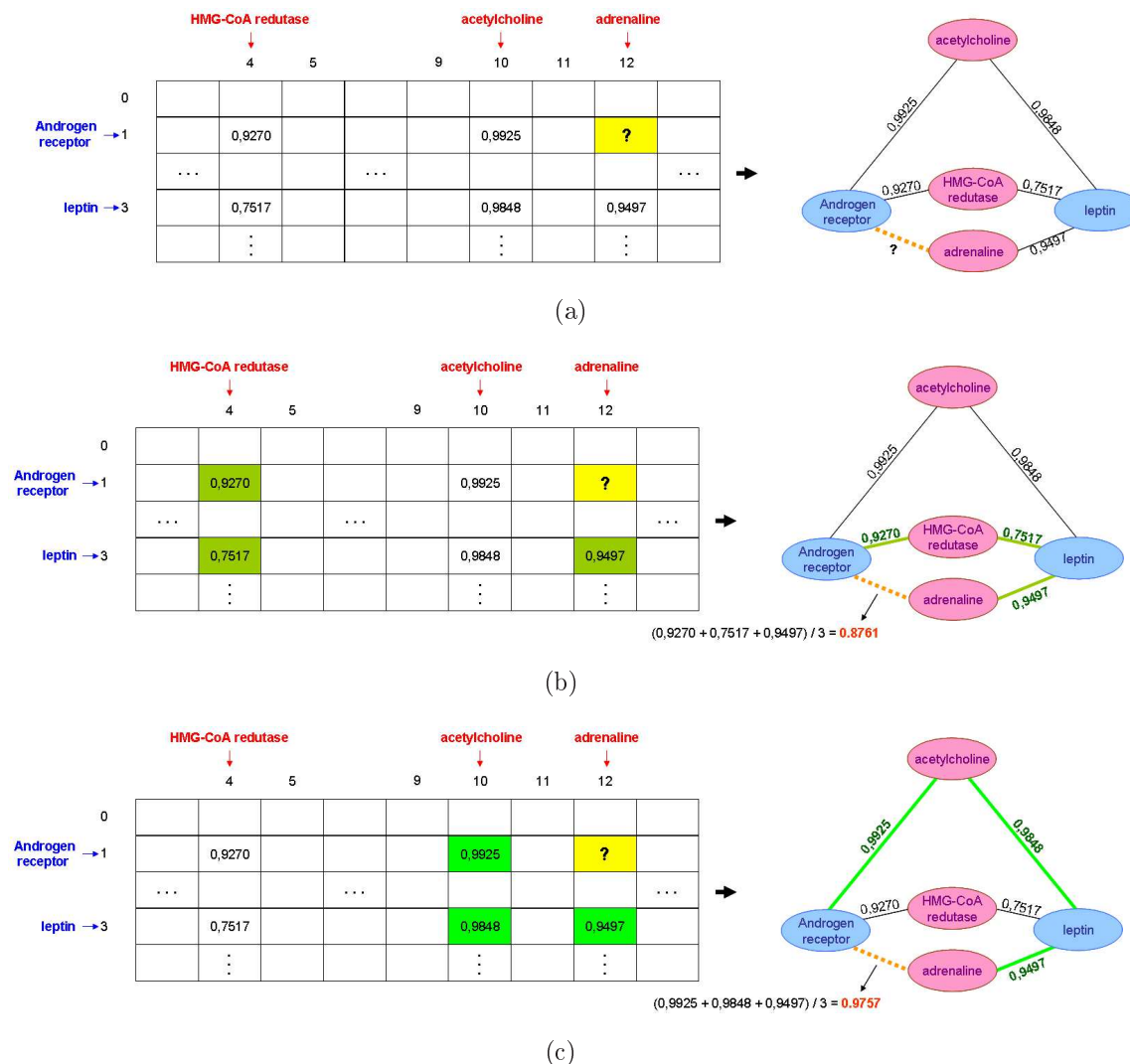
### 5.1.2 O Histórico de Formação das Interações entre Entidades

Na interface do sistema *BioSearch* podemos observar o histórico que mostra a formação de todas as interações estabelecidas na rede pelo modelo. A melhor interação inferida pelo modelo, por exemplo, indica uma nova interação possível entre a adrenalina e o receptor de androgênio e é descoberta em três passos do processo de inferência na matriz que representa as interações da sub-rede *alvo*  $\times$  *gene* (Figura 5.2).

No primeiro passo, o modelo identifica a possibilidade da nova interação entre o alvo adrenalina e o gene receptor de androgênio (Figura 5.2(a)). No segundo passo do processo de inferência, o modelo encontra 3 interações já conhecidas que satisfazem a relação de transitividade e levam à inferência da nova interação entre o alvo adrenalina e o gene receptor de androgênio. A primeira interação conhecida<sup>1</sup> é entre o gene receptor de androgênio e o alvo HMG-CoA redutase com evidência

---

<sup>1</sup>Interação estabelecida na rede com base nas patentes 6,306,874, 6,313,138, 6,358,970, 6,420,382, 6,472,403, 6,479,512, 6,645,974, 6,838,584, 6,872,724 e 6,958,340.



**Figura 5.2:** Inferência da melhor interação encontrada pelo modelo. Entidades representadas em vermelho pertencem à categoria alvo e as representadas em azul, à categoria gene. (a) Possível nova interação entre o alvo adrenalina e o gene receptor de androgênio. (b) Primeira tripla de interações que indica a nova interação. (c) Segunda tripla de interações que indica a nova interação.

de interação 0,9270. A segunda interação conhecida<sup>2</sup> é entre o gene leptina e o alvo HMG-CoA redutase com evidência de interação 0,7517. A terceira interação conhecida<sup>3</sup> é entre o gene leptina e o alvo adrenalina com evidência de interação 0,9497. Essas interações conhecidas evidenciam a nova interação com o valor 0,8761 (Figura 5.2(b)).

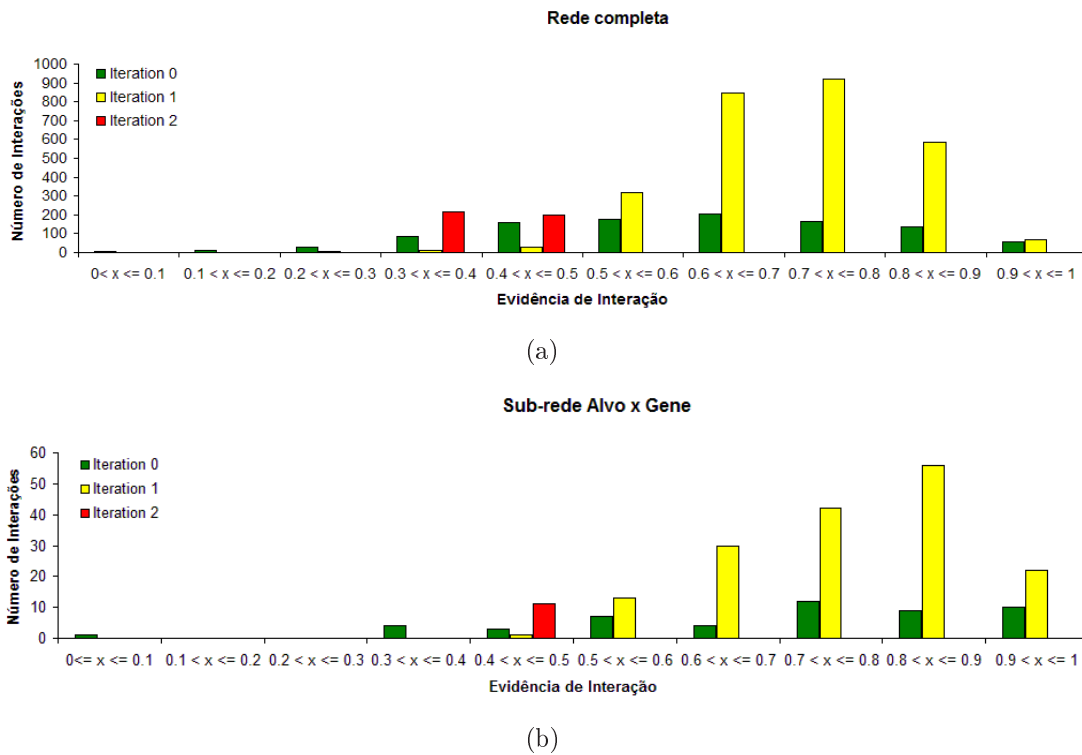
No terceiro passo do processo de inferência, o modelo encontra uma outra tripla de interações já conhecidas que também satisfazem a relação de transitividade e

<sup>2</sup>Interação estabelecida na rede com base nas patentes 6,130,214, 6,369,067, 6,440,961, 6,444,816, 6,465,444, 6,486,141 e 6,486,142.

<sup>3</sup>Interação estabelecida na rede com base nas patentes 6,603,058 e 6,867,005.



levam à inferência da nova interação entre o gene receptor de androgênio e o alvo adrenalina. Nessa nova possibilidade, a primeira interação conhecida<sup>4</sup> é entre o gene receptor de androgênio e o alvo acetilcolina com evidência de interação 0,9925. A segunda interação conhecida<sup>5</sup> é entre o gene leptina e o alvo acetilcolina com evidência de interação 0,9848. A terceira interação conhecida<sup>6</sup> é entre o gene leptina e o alvo adrenalina com evidência de interação 0,9497. Nessa possibilidade, as interações conhecidas evidenciam a nova interação com valor 0,9757 (Figura 5.2(c)).



**Figura 5.3:** Distribuição das evidências de interação. As interações da iteração 0 do modelo correspondem às interações já conhecidas. As interações da iteração 1 são inferidas a partir das interações conhecidas. As interações da iteração 2 são inferidas com base nas interações das iterações 0 e 1. (a) Distribuição das evidências de interação em toda a rede. (b) Distribuição das evidências de interação na sub-rede *alvo × gene*.

Como a evidência de interação encontrada no terceiro passo do processo de inferência é maior que a encontrada no segundo passo, ela se torna o valor da evidência de interação da nova interação entre o alvo adrenalina e o gene receptor de androgênio. Depois do terceiro passo do processo de inferência, o modelo não encontra outras triplas de interações que evidenciem o novo relacionamento entre o alvo adrenalina e o gene receptor de androgênio. Então, a evidência de interação encontrada no terceiro passo do processo de inferência (0,9757) é armazenada na matriz indicando

<sup>4</sup>Interação estabelecida na rede com base nas patentes 6,139,735 e 6,387,268.

<sup>5</sup>Interação estabelecida na rede com base nas patentes 6,503,713, 6,528,315 e 6,832,114.

<sup>6</sup>Interação estabelecida na rede com base nas patentes 6,603,058 e 6,867,005.

a nova interação entre as duas entidades.

Neste trabalho, nosso principal objetivo é apresentar o modelo desenvolvido e mostrar sua capacidade de inferir novas interações entre entidades. Por essa razão, não nos preocupamos em capturar o aspecto semântico em que as entidades biológicas foram empregadas nas coleção de documentos. Por exemplo, no histórico que demonstra a formação da nova interação entre o alvo adrenalina e o gene receptor de androgênio não encontramos ocorrências do gene leptina na patente 6,130,214, mas sim da proteína leptina que é codificada por esse gene. Já na patente 6,503,713, a ocorrência dessa entidade refere-se ao RNA mensageiro que codifica a proteína leptina.

**Tabela 5.3:** As 5 melhores novas interações indicadas pelo modelo em toda a rede de interação.

Dimensões	Sub-rede	Espaço Dimensional	Nova Interação	Evidência de Interação
2	11	alvo gene	adrenaline androgen receptor	0.9757
2	5	doença gene	acquired immunodeficiency syndrome transforming growth factor, beta 1	0.9738
2	5	doença gene	acquired immunodeficiency syndrome plasminogen activator, urokinase	0.9719
2	3	alvo gene	hmg-coa reductase von willebrand factor	0.9717
2	5	doença gene	alzheimer dementia plasminogen activator, urokinase	0.9639

### 5.1.3 A Distribuição da Evidência de Interação

Em nossos experimentos, poucas interações possuem alta evidencia de interação. Apenas 2,91% das interações na rede completa estão no intervalo  $]0,9, 1]$  e 14,32% das interações na sub-rede  $alvo \times gene$ , por exemplo, possuem evidência de interação nesse intervalo (Figura 5.3).

Na rede completa, a maioria das interações novas encontradas na iteração 1 possuem evidência de interação no intervalo  $]0,6, 0,8]$  (63,61%) e apenas 2,41% dessas interações possuem evidência de interação no intervalo  $]0,9, 1]$ . Na sub-rede  $alvo \times gene$ , a maioria das novas interações encontradas na iteração 1 possui evidência de interação no intervalo  $]0,8, 0,9]$  (34,15%) e 13,41% dessas interações possuem evidência de interação no intervalo  $]0,9, 1]$ . Em todas as sub-redes, o maior número de interações concentra-se no intervalo  $]0,6, 0,8]$ . Todas as interações novas da iteração 2 tiveram evidência de interação menor ou igual a 0,5. Interessantemente,

a sub-rede tridimensional *doença*  $\times$  *fármaco*  $\times$  *gene* e a sub-rede quadridimensional *alvo*  $\times$  *doença*  $\times$  *fármaco*  $\times$  *gene* são as que possuem o maior número de interações novas na iteração 2. Segundo Kostoff (2008a), essas seriam, então, as sub-redes com maior possibilidade de encontramos novas descobertas em biologia. Além disso, observamos que a iteração 2 é a última iteração do modelo (Apêndice C). Esse resultado é importante, pois mostra que as iterações do modelo não são executadas indefinidamente até que cada nodo seja conectado com todos os demais nodos de uma sub-rede.

Nós verificamos quais eram as 5 melhores novas interações apontadas pelo modelo em toda a rede (Tabela 5.3). Essas melhores interações da rede incluem as dimensões alvo, doença e gene, sendo três interações encontradas na sub-rede *doença*  $\times$  *gene* e duas encontradas na sub-rede *alvo*  $\times$  *gene*. Além disso, examinamos a sub-rede *alvo*  $\times$  *gene*, que é a sub-rede com a nova interação mais promissora da rede. Observamos que as cinco melhores interações inferidas nessa sub-rede pelo modelo incluem os alvos adrenalina, HMG-CoA redutase, receptor de fibrinogênio e receptor de progesterona e os genes receptor de androgênio, fator de von Willebrand e o gene leptina (Tabela 5.5).

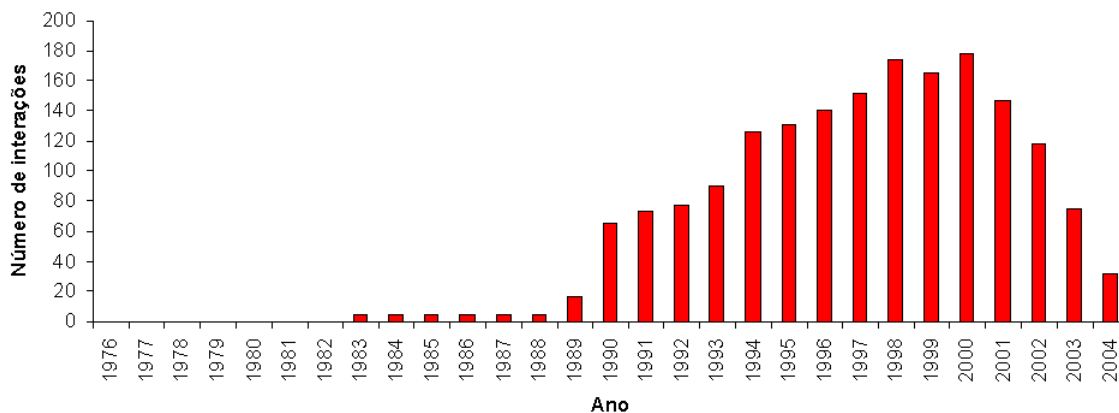
**Tabela 5.4:** As 5 melhores novas interações encontradas na sub-rede *alvo*  $\times$  *gene*.

Dimensões	Sub-rede	Espaço Dimensional	Nova Interação	Evidência de Interação
2	11	alvo	adrenaline	0.9757
		gene	androgen receptor	
2	11	alvo	hmg-coa reductase	0.9717
		gene	von willebrand factor	
2	11	alvo	fibrinogen receptor	0.9639
		gene	androgen receptor	
2	11	alvo	progesterone receptor	0.9632
		gene	von willebrand factor	
2	11	alvo	fibrinogen receptor	0.9614
		gene	leptin	

## 5.2 A Validação

Validamos nossos experimentos usando a data de publicação das patentes que formam a coleção de documentos. As patentes da coleção foram publicadas entre 01/01/1976 e 31/12/2005. Nesse experimento, criamos a rede usando todos os documentos da coleção e observamos as interações conhecidas e novas que foram

estabelecidas. Em seguida, removemos da coleção todas as patentes publicadas em 2005 e construímos a rede novamente. Assim, pudemos observar as interações conhecidas e novas que foram estabelecidas na rede através das patentes publicadas até o ano de 2004. Comparando a rede formada até o ano de 2004 com aquela formada até o ano de 2005, pudemos observar, por exemplo, interações que eram novas em 2004 e que tornaram-se conhecidas em 2005. Isso nos permitiu concluir que essas interações novas indicadas na rede de 2004 possuíam confirmações patenteadas em 2005. Nós executamos esses passos para cada um dos 30 anos abrangidos pelas datas de publicação das patentes de nossa coleção.

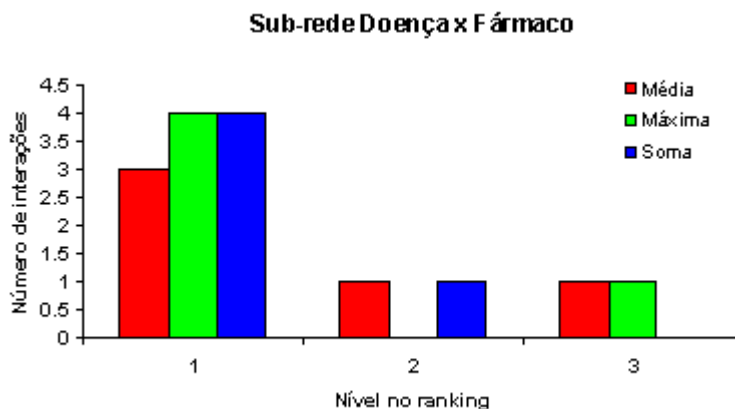


**Figura 5.4:** Número de interações com patentes de confirmação por ano.

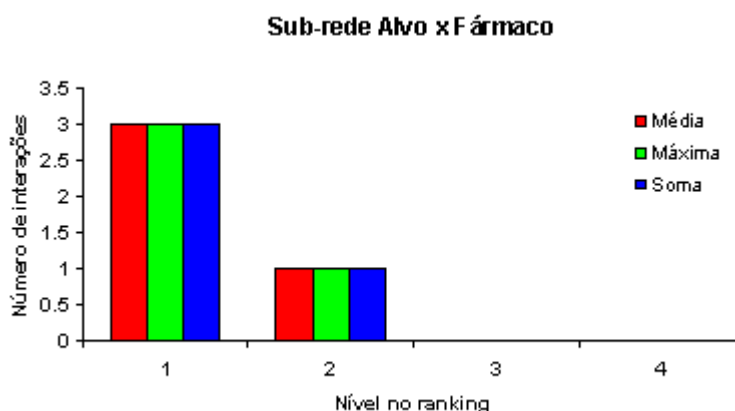
Examinando os resultados desse experimento anualmente, observamos que as interações novas inferidas em um ano foram confirmadas por patentes publicadas em um ano mais recente (Figura 5.4). Esses resultados mostram que nosso modelo infere interações que são confirmadas por patentes já publicadas no USPTO. Por exemplo, removendo todas as patentes publicadas em 2005, nosso modelo inferiu 2.930 interações novas através das patentes publicadas até 2004. Entre essas interações novas, 32 possuem patentes de confirmação publicadas em 2005. Esse experimento também nos permitiu observar que há um aumento e concentração de patentes de confirmação entre 1990 e 2000 e uma diminuição desse número a partir de 2001 até 2004. Esses resultados refletem a sazonalidade do grande número de patentes de biotecnologia requeridas nos anos 90 e sua redução a partir de 2001 (Horn and Lipsey, 2004).

Analisamos a distribuição das 32 interações indicadas em 2004 e com patentes de confirmação em 2005 nos *rankings* de novas interações de cada sub-rede. Para realizar esse estudo, dividimos os *rankings* das interações novas inferidas em 2004 de cada sub-rede em níveis de até 100 interações por nível. Então, observamos a distribuição das 32 interações novas com patentes de confirmação nesses níveis.

Para essa análise, apresentamos os resultados das 3 estratégias que usamos para determinar o valor das interações conhecidas: a média aritmética, máxima e soma das similaridades retornadas pelo modelo de espaço vetorial.



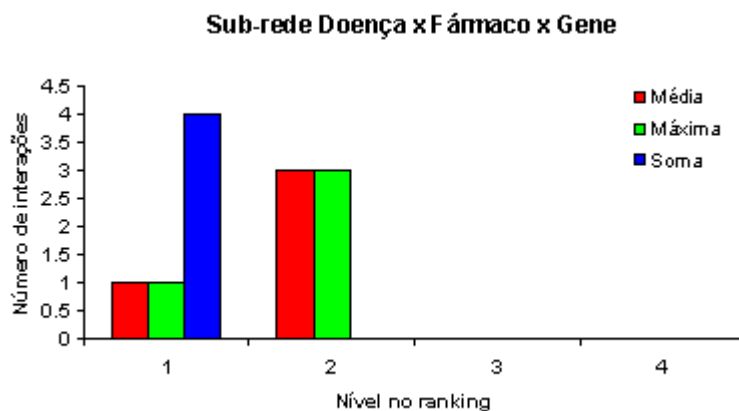
**Figura 5.5:** Distribuição das interações inferidas em 2004 e que possuem confirmações patenteadas em 2005 no *ranking* da sub-rede *doença*  $\times$  *fármaco*.



**Figura 5.6:** Distribuição das interações inferidas em 2004 e que possuem confirmações patenteadas em 2005 no *ranking* da sub-rede *alvo*  $\times$  *fármaco*.

Na sub-rede bidimensional *doença*  $\times$  *fármaco*, por exemplo, observamos que o modelo inferiu 275 novas interações em 2004. Dessas interações, 5 foram confirmadas através de patentes publicadas em 2005 (Figura 5.5). Na estratégia da média aritmética das similaridades para determinar o valor das interações conhecidas, 3 interações foram colocadas no primeiro nível do *ranking*, 1 no segundo nível e 1 no terceiro nível. Na estratégia da máxima similaridade, 4 interações foram colocadas no primeiro nível e 1 no terceiro nível. Na estratégia da soma das similaridades, 4 interações foram colocadas no primeiro nível e 1 no segundo nível.

Verificamos também que na sub-rede bidimensional *alvo*  $\times$  *fármaco*, o modelo inferiu 282 novas interações em 2004. Dessas interações, 4 foram confirmadas através de patentes publicadas em 2005 (Figura 5.6). Nessa sub-rede, as 3 estratégias usadas



**Figura 5.7:** Distribuição das interações inferidas em 2004 e que possuem confirmações patenteadas em 2005 no ranking da sub-rede *doença × fármaco × gene*.

para determinar o valor das interações conhecidas colocaram 3 interações no primeiro nível e 1 no segundo nível.

Num último exemplo, observamos que o modelo inferiu 308 novas interações na sub-rede *doença × fármaco × gene* em 2004. Dessas interações, 4 foram confirmadas através de patentes publicadas em 2005 (Figura 5.7). As estratégias da média aritmética e máxima similaridade para determinar o valor das interações conhecidas colocaram 1 interação no primeiro nível e 3 no segundo nível. Por outro lado, a estratégia da soma das similaridades colocou todas as interações no primeiro nível (Apêndice D).

Verificamos também quais foram as interações conhecidas que, estabelecidas na rede em 2005, tornaram-se interações novas em 2004, quando as patentes publicadas em 2005 foram removidas da coleção de documentos (Tabela 5.5). Muitas dessas interações conhecidas são evidenciadas por apenas 1 patente em 2005. Depois que as patentes de 2005 são removidas da coleção, observamos que muitas patentes sugerem essas interações em 2004 como sendo novas interações para a rede. Por exemplo, a interação entre a doença ataque do coração e o gene *ppar-gama* é relatada por 1 patente publicada em 2005. Quando essa patente é removida da coleção, 61 patentes publicadas até o ano de 2004 passam a indicar essa interação como uma nova interação para a rede. Esse resultado mostra que a literatura de patentes possui muitas relações implícitas que podem ser exploradas para promover novos avanços científicos e tecnológicos. Além disso, esse resultado mostra ainda que em geral essas relações são evidenciadas por um grande número de patentes. O conhecimento dessas relações implícitas é importante, porque podemos usá-lo para acelerar o processo de pesquisa, de descoberta e de inovação científica e tecnológica em biotecnologia.

Também pesquisamos o sítio *Web* de patentes do USPTO, procurando patentes de confirmação para as novas interações inferidas através das patentes publicadas

**Tabela 5.5:** As 5 interações conhecidas com maior evidência de interação em 2005 e que se tornaram novas em 2004.

Sub-rede	Espaço Dimensional	Interação	Evidência em 2005	Evidência em 2004	Patentes em 2005	Indicações em 2004
5	doença gene	heart attack ppar-gamma	0.9999	0.8324	1	61
1	alvo doença	adrenaline cardiac ischemia	0.9866	0.8676	1	36
11	alvo doença fármaco gene	hmg coa reductase breast cancer tamoxifen kennedy disease	0.9190	0.6383	1	5
2	alvo fármaco	gp iib/iiia neoral	0.9137	0.8354	1	103
4	doença fármaco	HIV bonyl	0.9041	0.8825	1	30

até o ano de 2005. Nós procuramos por patentes publicadas entre 01/01/1976 e 25/04/2009. Esse experimento foi realizado para as 100 primeiras interações novas de cada sub-rede (a sub-rede tridimensional  $alvo \times fármaco \times gene$  possui apenas 76 interações novas em 2005).

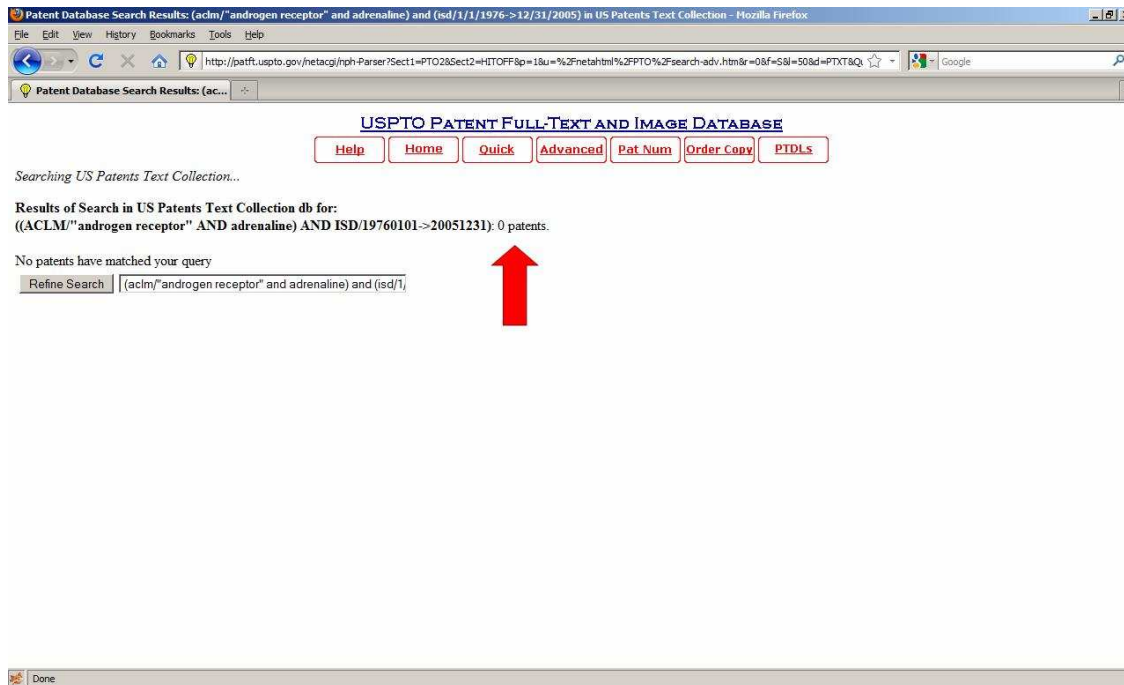
Nesse segundo experimento de validação, nós descobrimos 86 patentes de confirmação para as 1.076 interações novas pesquisadas no sítio *Web* do USPTO. Como esperado, não encontramos nenhuma patente de confirmação publicada antes de 01/01/2006 para as novas interações inferidas pelo modelo. Em seguida, analisamos a distribuição dessas patentes de confirmação no *ranking* de cada sub-rede (Tabela 5.6). Os *rankings* foram divididos em 10 níveis de até 10 interações por nível. A maioria das patentes de confirmação (56%) foram encontradas para interações novas situadas até o nível 5 dos *ranking* de cada sub-rede.

Além das nossas análises baseadas na data de publicação das patentes, nós realizamos uma terceira validação de nossos resultados. Nesse terceiro processo de validação, procuramos por artigos científicos publicamente disponíveis na *Web* para confirmar as novas interações inferidas pelo sistema. Por exemplo, o melhor resultado encontrado em nosso modelo indica a interação entre o neurotransmissor adrenalina e o gene receptor de androgênio na sub-rede bidimensional  $alvo \times gene$ . No USPTO, nenhuma seção de reivindicação das patentes publicadas no período que vai de 01/01/1976 até 31/12/2005 possui a co-ocorrência das entidades adrenalina e receptor de androgênio (Figura 5.8 e Apêndice E). Por isso, nenhuma seção de reivindicação das patentes que formam nossa coleção de documentos relata interação entre essas entidades. Entretanto, Sastry *et al.* (2007) afirmam em 2007 que o

**Tabela 5.6:** Distribuição das patentes de confirmação por nível do *ranking* de resposta de cada sub-rede.

Nível no	Sub-rede											Total
<i>Ranking</i>	1	2	3	4	5	6	7	8	9	10	11	
1	0	1	1	0	1	2	2	0	0	0	0	7
2	0	1	0	7	2	1	0	0	0	0	0	11
3	0	0	2	3	4	0	0	0	0	0	0	9
4	2	1	0	0	4	3	0	0	0	0	2	12
5	3	0	0	1	0	0	0	0	3	2	0	9
6	0	2	0	1	1	1	0	2	0	0	0	7
7	0	1	2	1	3	0	0	0	0	0	0	7
8	0	0	0	1	1	3	0	0	0	0	0	5
9	1	1	2	4	0	0	0	0	0	0	0	8
10	2	0	0	4	1	2	0	2	0	0	0	11
Total	8	7	7	22	17	12	2	4	3	2	2	86

efeito antiapoptótico da epinefrina, ou adrenalina, parcialmente depende do receptor de androgênio. A ativação do receptor de androgênio por dihidrotestosterona ou pelo análogo do androgênio R1881 é conhecida por proteger células da próstata de entrarem em apoptose. Além disso, uma diminuição modesta no efeito antiapoptótico da epinefrina em células onde a expressão do receptor de androgênio foi reduzida usando-se uma abordagem shRNA apresentou evidências de que a epinefrina reduz a sensibilidade de células cancerígenas de entrarem em apoptose através de interações com receptores beta2-adrenérgicos.

**Figura 5.8:** Página do USPTO confirmando a inexistência de patentes publicadas no período de 01/01/1976 até 31/12/2005 em que haja co-ocorrência das entidades adrenalina e receptor de androgênio na seção de reivindicação.



# Capítulo 6

## Discussão

Alcançamos resultados muito interessantes em nossa estratégia de combinar o modelo de espaço vetorial em um processo de inferência. Nós usamos essa estratégia para modelar sistemas biológicos a partir de coleções textuais e prever novas interações entre entidades desses sistemas. Modelar sistemas biológicos é uma tarefa complexa, porque devemos considerar um grande número de parâmetros biológicos que afetam a interação entre as entidades, como a concentração das entidades envolvidas na interação e o papel que elas desempenham em várias reações diferentes. Além disso, sistemas biológicos não são sistemas lineares. Por essa razão, perturbações em sistemas biológicos comumente levam a resultados inesperados. Assim, temos que estudar sistemas biológicos em diferentes níveis de abstração. Somente dessa forma, conseguimos construir simplificações que permitam diminuir a complexidade desses sistemas, para que possamos realizar nossas análises (Holme, 2008).

Em nosso modelo, escolhemos trabalhar em um nível de abstração que descreve sistemas biológicos com base em coleções textuais. Para alcançarmos essa descrição dos sistemas, focamos na recuperação de informação sobre entidades biológicas a partir das coleções textuais através do modelo de espaço vetorial. Em seguida, expressamos essa informação em nossa relação de transitividade que explora as atividades principais e secundárias que as entidades exercem nos sistemas biológicos. Com isso, conseguimos criar nossa representação do sistema biológico, formando uma rede de interações. Nessa representação, temos uma perda de informação sobre o sistema real, mas a abordagem nos permitiu realizar várias análises com importantes descobertas. Uma dessas descobertas importantes foi a indicação de como o conhecimento biológico está interconectado na literatura que o descreve. Descobertas como essa nos encorajam a realizar investigações mais profundas de parâmetros biológicos que devemos estudar, para alcançarmos melhores resultados no processo de inferência e na estratégia de *ranking* das interações (Hristovski and B. Peterlin,

2005). Além de melhorar o processo de inferência e de *ranking* das interações, esses parâmetros têm uma outra importante função: impedir a propagação de ruído através da rede. Interações espúrias incorretamente estabelecidas durante a construção da rede propagam interações indevidas no processo de inferência. Assim, o conjunto de parâmetros nos ajuda impor restrições ao estabelecimento de interações na rede. A extração desses parâmetros a partir de diversas fontes de informação e a integração deles em nosso modelo também é um importante desafio a ser tratado.

Neste trabalho, nosso objetivo básico não é assegurar uma cobertura completa da literatura biológica e nem criar uma grande rede de interações. Ao invés disso, apresentamos uma prova de conceito que mostra a aplicação de nosso modelo e sua utilidade em descobrir e ordenar interações com base em conexões implícitas da literatura biológica. Por esse motivo, usamos uma pequena coleção textual, restrita a patentes de biotecnologia, com o único objetivo de avaliar nosso modelo. Dessa forma, estamos cientes de que muitas novas interações inferidas pelo modelo em nossos experimentos atuais já foram reportadas em trabalhos anteriores, como artigos científicos. Entretanto, usamos essas interações já comprovadas e publicadas em trabalhos anteriores para validar nossos resultados, uma vez que não existem no momento coleções textuais disponíveis para validação de sistemas para descobertas baseadas em literatura (Kostoff, 2008a; Smalheiser and Torvik, 2008; Kostoff, 2008c).

Os resultados experimentais que obtivemos demonstram que a seção de reivindicação nos documentos da literatura de patentes biotecnológicas possui conexões implícitas que podemos explorar para alcançarmos maiores avanços em biologia. De forma semelhante, trabalhos relacionados (Swanson, 1986, 1990; Smalheiser and Swanson, 1998; Swanson *et al.*, 2006; Weeber *et al.*, 2001; Wren *et al.*, 2004; Bruza and Weeber, 2008; Campillos *et al.*, 2008; Kostoff, 2008a; Kostoff *et al.*, 2008b,d; Kostoff, 2008b; Kostoff and Briggs, 2008; Kostoff *et al.*, 2008c,e,a; Li *et al.*, 2009) têm apresentado resultados semelhantes em experimentos realizados com títulos e resumos de artigos científicos. Além disso, a literatura de patentes biotecnológicas tem proporcionado bons meios de avaliarmos nosso modelo. Iterando o modelo com base nos anos em que as patentes foram publicadas, conseguimos mostrar que novas interações indicadas a partir de patentes mais antigas foram confirmadas por patentes publicadas em anos mais recentes. Esses testes também mostraram que muitas confirmações foram encontradas para interações no topo de nossos *rankings* de resposta. Considerando, por exemplo, a estratégia de *ranking* que atribui às interações conhecidas a soma das similaridades retornadas pelo modelo de espaço vetorial, nós tivemos 69% de confirmação entre as primeiras 100 novas interações.

Em nossos experimentos também verificamos que, com exceção da sub-rede *fár-*

*maco*  $\times$  *gene*, todas as demais sub-redes bidimensionais tiveram suas melhores interações com evidência de interação maior que aquelas das sub-redes tridimensionais. Além disso, todas as sub-redes tridimensionais tiveram suas melhores interações com evidência de interação maior que aquela da sub-rede quadridimensional. Isso acontece, porque torna-se mais difícil encontrar documentos na coleção em que ocorram simultaneamente as entidades de todas as dimensões de uma sub-rede que possui muitas dimensões, o que leva à diminuição do valor da evidência de interação. Entretanto, os resultados de sub-redes com mais dimensões fornecem indicações mais apuradas de novos relacionamentos, embora possam ter evidência de interação menor que a de sub-redes com menos dimensões. Esses resultados são mais apurados porque sub-redes com mais dimensões restringem melhor o espaço de busca de novas interações. Desse modo, interações da sub-rede quadridimensional têm maior chance de levarem a novas descobertas e inovações que as interações das sub-redes tridimensionais. Da mesma forma, as interações das sub-redes tridimensionais têm maior chance de levarem a novas descobertas e inovações que as interações das sub-redes bidimensionais. Assim, resultados encontrados em uma sub-rede não devem ser comparados com resultados de sub-redes que possuem um número diferente de dimensões, apenas considerando-se o valor da evidência de interação.

A análise da evidência de interação das interações conhecidas e encontradas na coleção de patentes nos permitiu identificar três informações principais e importantes: (i) os temas mais estudados em biologia, (ii) as pessoas e instituições que estudaram esses temas e (iii) quando esses temas foram estudados. Por outro lado, a análise da evidência de interação das novas interações nos ajudou a entender as melhores maneiras como o conhecimento em biologia pode ser ampliado a partir do conhecimento já adquirido na área. Isso é possível, porque novas interações com alta evidência de interação correspondem a novos avanços com fortes evidências a partir da observação do conhecimento corrente. Essas interações com alta evidência de interação representam os melhores caminhos a serem seguidos por pesquisadores de forma a terem sucesso em seus experimentos.

Outra forma de analisar as novas interações é levando em consideração o número de documentos na coleção textual que as indicam. Interações indicadas por um grande número de documentos representam os assuntos que estão em voga nas pesquisas científicas na data considerada para análise. Por outro lado, as interações indicadas por poucos documentos representam os assuntos que estão à margem nas pesquisas científicas na data considerada para análise. Assim, as interações com alta evidência de interação e com muitas patentes de indicação correspondem a interações com maior probabilidade de estarem sendo estudadas por um maior número de

pesquisadores concorrentes. Em contraste, novas interações com evidência de interação não tão alta, mas com poucas patentes de indicação representam possibilidades de extensão do conhecimento biológico com poucas evidências a partir do conhecimento corrente. Embora essas interações possam não ser as mais promissoras em princípio, elas têm menos chances de estarem sendo desenvolvidas por pesquisadores concorrentes, dadas as poucas evidências que elas têm a partir do conhecimento corrente, ou, pelo menos, têm grandes chances de estarem sendo estudadas por poucos pesquisadores.

As 5 melhores novas interações indicadas pelo modelo em toda a rede de interação mostram que as sub-redes cujos espaços dimensionais são formados pelas categorias alvo biológico, doença e gene possuem os melhores resultados. Isso revela uma grande quantidade de estudos envolvendo entidades dessas categorias e, conseqüentemente, a grande importância que esses estudos têm tido para a pesquisa em biologia. Além disso, o modelo também apresenta, através das novas interações, as indicações do que ainda pode ser obtido como progresso nas áreas da biologia que são indicadas por essas categorias. Por exemplo, na área da biologia que estuda alvos biológicos e genes, o sistema mostra que o estudo da interação entre adrenalina e o receptor de androgênio pode trazer resultados muito promissores, pois possui evidência de interação 0.9757.

Em nossa implementação corrente do modelo, não usamos processamento natural de linguagem (PNL) (Weeber *et al.*, 2001) e nem heurísticas para capturar o contexto em que os nomes das entidades são empregados na coleção de documentos. Entretanto, os nomes de entidades que selecionamos foram satisfatórios para o nosso propósito de testar e avaliar o modelo. Nessa implementação, criamos grupos de nomes para cada entidade biológica. Contudo, acreditamos que podemos alcançar resultados ainda melhores e interessantes se considerarmos ontologias (Fensel, 2002; Alani *et al.*, 2003; Schuffenhauer *et al.*, 2002; Ontology, 2008) e tesouros (Silveira, 2003) como ferramentas para identificação dos nomes das entidades e das categorias que devemos usar na criação da rede. Além disso, também podemos empregar ontologias e tesouros para reconhecer a semântica em que as entidades são empregadas nos documentos. Com isso, conseguimos uma maneira mais adequada de controlar e normatizar o processo de seleção de entidades para indexação e para construção da rede.

Os resultados que encontramos através da literatura biológica nos estimula a empregar o modelo também em outras áreas do conhecimento. O modelo pode ter aplicações importantes nas áreas em que a análise e inferência de relacionamentos entre entidades contribui para o avanço da ciência, especialmente em ciências experi-

mentais, a exemplo da física, química e engenharias, como a engenharia de materiais. Outra aplicação importante para o modelo é na construção de redes sociais, uma vez que o estudo das interações entre indivíduos e instituições pode revelar características e propriedades importantes dessas redes. Uma dessas redes com inúmeras contribuições e benefícios para a ciência, para a tecnologia e, conseqüentemente, para a sociedade é a de interações entre inventores, proprietários de patentes e instituições. Determinando, por exemplo, as interações conhecidas entre instituições e os proprietários das patentes de uma certa tecnologia, é muito importante inferir quais instituições teriam maior interesse em firmar contratos para transferência dessa tecnologia.

Uma outra consideração importante sobre nosso modelo é o número de entidades e categorias para a construção da rede. Um aumento no número de entidades e categorias tem implicações significativas no desempenho do sistema. Para aumentarmos o ganho de desempenho em nossa implementação corrente, priorizamos o desenvolvimento de uma arquitetura com processamento paralelo para o sistema. Além disso, nessa arquitetura os dados são pré-processados e armazenados na base de dados com o objetivo de melhorar o tempo de processamento em cada módulo do sistema. O desenvolvimento de uma estratégia distribuída para essa arquitetura e para o armazenamento da rede biológica também é uma outra contribuição que consideramos importante para o trabalho.

Com base em nossos experimentos, podemos apresentar um resumo das vantagens que encontramos em nosso modelo e sistema e também dos trabalhos futuros para aperfeiçoá-los. As vantagens que destacamos são:

1. Desenvolvemos um sistema para dar suporte à atividade cotidiana de pesquisadores em biologia. Quando um pesquisador inicia a investigação das interações possíveis em que uma entidade pode ter participação, ele encontra um número muito grande de interações. Esse grande número de interações dificulta o trabalho do pesquisador, pois teoricamente ele precisa analisar todas essas interações possíveis para conseguir chegar à interação mais promissora da entidade. Dessa forma, essa busca pela interação mais promissora é custosa, principalmente se consideramos os recursos financeiros e o tempo necessários para realizá-la. Entretanto, nosso sistema auxilia e facilita esse trabalho de busca das interações, pois permite que esse pesquisador tenha uma visão do estado da arte em biologia, apresentando os assuntos já estudados na área. Além disso, o sistema também apresenta os estudos que ainda podem ser realizados para ampliar o conhecimento adquirido, a partir dos estudos já realizados.
2. O sistema é capaz de distinguir as melhores interações entre as entidades. To-

das as interações da rede criada pelo sistema podem ser ordenadas em ordem de relevância, segundo a coleção de documentos. O sistema realiza essa ordenação através de um peso que é atribuído a todas as interações. Esse peso é a evidência de interação entre as entidades. A evidência de interação é atribuída às interações com base no modelo de espaço vetorial. Assim, a evidência de interação reflete o número de documentos que tratam essas interações na coleção textual.

3. O sistema também permite observar o número de documentos da coleção textual que promoveram a formação de cada interação da rede. O número de documento que relatam a interação revela o interesse que essa interação desperta na comunidade científica. Assim, alguns pesquisadores podem se interessar por aquelas interações com maior número de documentos, porque elas indicam as pesquisas mais realizadas e que despertam o maior interesse da comunidade científica em uma dada época. Por outro lado, alguns pesquisadores podem optar por aquelas interações com menor número de documentos por serem menos pesquisadas e, conseqüentemente, representarem nichos menos explorados nas pesquisas científicas.
4. O sistema permite identificar através das interações da rede quais são as entidades mais pesquisadas e quais são as menos pesquisadas em biologia. O mesmo também pode ser conseguido sobre as categorias usadas na construção da rede, o que resulta na identificação das áreas mais pesquisadas ou menos pesquisadas em biologia. Essas informações são importantes, porque permitem que os pesquisadores criem estratégias para dar um melhor direcionamento a suas pesquisas.
5. A criação dos espaços dimensionais na rede é uma boa forma de restringir e organizar o espaço de busca das interações. Assim, o sistema consegue identificar os melhores resultados da rede por área de pesquisa. Essas áreas são demarcadas pelas categorias consideradas nos espaços dimensionais das sub-redes. Além disso, um espaço dimensional com maior número de dimensões permite reduzir o espaço de busca das novas interações, tornando as busca mais fáceis de serem realizadas e os resultados de cada busca mais apurados.
6. O sistema permite que o usuário observe como todas as interações da rede são formadas. Isso é possível, porque o sistema apresenta um histórico que demonstra como todas as interações foram estabelecidas na rede. Ele mostra as patentes que levaram à formação das interações conhecidas e os passos do processo de inferência que promoveram a indicação de todas as interações novas.

7. O sistema permite que pesquisadores consultem a rede com base no tipo das interações. Pesquisando interações inferidas na iteração 0 do sistema, o pesquisador encontra as interações já conhecidas da rede. Pesquisando interações da iteração 1, o pesquisador encontra as interações novas que podem corresponder a inovações para o estado da arte em biologia. Pesquisando interações inferidas a partir da iteração 2 do sistema, o pesquisador encontra interações novas que podem corresponder a descobertas para o estado da arte em biologia. Assim, o sistema possibilita a identificação do que já é conhecido em biologia, de inovações e também de novas descobertas na área.

Por outro lado, destacamos os seguintes pontos a serem abordados nos trabalhos futuros:

1. Incluir informação auxiliar que permita ou inviabilize o estabelecimento das interações entre entidades na rede. Nossa modelagem de um sistema biológico baseia-se apenas na co-ocorrência das entidades em uma coleção de documentos. No entanto, a co-ocorrência das entidades nos documentos é insuficiente para assegurar a existência das interações no sistema real. Portanto, é preciso criar regras para estabelecer as interações na rede, com o objetivo de eliminar interações indevidas ou inviáveis biologicamente. A dificuldade de identificar a semântica em que as entidades são empregadas nos documentos é um fator que leva à criação dessas interações espúrias na rede.
2. Assegurar que todos os nomes e termos relacionados às entidades sejam considerados na construção da rede, por exemplo, sinônimos, acrônimos e variações sintáticas. Os nomes das entidades têm implicações importantes na precisão do processo de inferência. A falta de algum nome de uma entidade interfere na identificação das interações conhecidas em que essa entidade participa. Assim, o modelo infere uma nova interação entre entidades que, entretanto, já é conhecida em um documento. Porém, os nomes que foram usados para referenciar as entidades nesse documento não foram os mesmos usados durante a construção da rede.
3. Aprimorar a tarefa de coleta. A cobertura da coleção de documentos também tem implicações na precisão do processo de inferência. O sistema pode inferir uma nova interação apenas porque os documentos em que ela é descrita não foram coletados. Esse problema de cobertura exige o desenvolvimento de estratégias eficazes e eficientes de coleta que assegurem a recuperação dos documentos necessários para a construção da rede. Essas estratégias de coleta devem ser implementadas com base em princípios que minimizem os custos de armazenamento, de processamento e de tempo de rede.

4. Criar a rede a partir de outros campos das patentes, como o título, o resumo e a seção de descrição. Além disso, criar a rede também a partir de outras literaturas publicamente disponíveis na *Web*, como artigos científicos e bulas de medicamentos.
5. Analisar e implementar o processo de inferência em sub-redes unidimensionais.
6. Criar estratégias que promovam o aprimoramento do processo de geração das interações possíveis. Em nossa implementação corrente, a geração das interações possíveis da rede é realizada através do produto cartesiano entre as entidades. O aumento no número de entidades e categorias da rede afeta significativamente o custo de processamento e armazenamento do produto cartesiano.
7. Empregar o modelo e o sistema desenvolvidos em outros campos da ciência e da tecnologia. O trabalho desenvolvido pode ter importantes aplicações em outras áreas de estudo, promovendo avanços importantes para a sociedade. Dessa forma, devemos empregar o sistema em áreas como a física, a química e as engenharias.



# Capítulo 7

## Conclusão

Neste trabalho apresentamos um modelo para inferir novas interações entre entidades biológicas a partir de coleções textuais. Nós usamos esse modelo para construir uma rede de interações com base em uma relação de transitividade que explora as atividades principais e secundárias das entidades em um sistema biológico. O modelo emprega a capacidade do modelo de espaço vetorial de recuperar informação, para estabelecer as interações entre entidades a partir da coleção textual. Consideramos diferentes categorias das entidades biológicas ao construirmos a rede. Cada combinação de um conjunto de categorias dá origem à um espaço dimensional que caracteriza uma sub-rede na rede completa. Esses espaços dimensionais dividem o espaço de pesquisa de novas interações, fornecendo resultados mais apurados. Além disso, e ao contrário de muitos trabalhos relacionados, nosso modelo é capaz de relacionar múltiplas entidades biológicas simultaneamente e resultados mais apurados são encontrados em sub-redes com maior número dimensões.

Em nossos experimentos, usamos uma coleção de documentos formada pela seção de reivindicação de patentes. A literatura de patentes é uma fonte de informação muito adequada para nosso trabalho, porque possui um grande valor estratégico, técnico e relacionado a negócios. Além disso, a literatura de patentes propiciou um ótimo meio para testarmos o modelo e mostrarmos sua utilidade, porque criamos nossa rede de interações de acordo com os anos em que as patentes foram publicadas. Assim, patentes publicadas em anos mais recentes confirmaram as indicações de interações encontradas em anos anteriores.

Durante a formação da base de dados para construção da rede biológica, observamos a importância do contexto semântico em torno do nome das entidades. A preservação dessa semântica durante a tarefa de coleta é responsável por reduzir o tempo de rede, economizar espaço de armazenamento em memória e melhorar o desempenho do sistema. Durante as tarefas de indexação e de construção da rede, a

semântica do nome das entidades é responsável principalmente por manter a precisão do processo de identificação das interações conhecidas, evitando o estabelecimento de interações inadequadas. Dessa forma, constatamos que a utilização dos conectivos OR e AND entre os termos que formam o nome das entidades não é apropriada no sistema.

Uma estratégia simples para preservar a semântica dos nomes das entidades é utilizar a busca por frase. A busca por frase é uma forma mais segura e apropriada para identificação de casamentos desses nomes no texto das patentes, evitando a maioria dos problemas causados pelos conectivos OR e AND, como a coleta de patentes inadequadas para identificação de interações conhecidas. Uma segunda estratégia é procurar a ocorrência do nome das entidades em seções das patentes que ofereçam condições de minimizar a formação de interações inadequadas, como a seção de reivindicação.

Durante os experimentos, o objetivo básico não foi assegurar uma cobertura completa da literatura biológica em nossa base de dados. Ao invés disso, o objetivo foi fornecer uma prova de conceito que comprovasse a capacidade de nosso modelo descobrir novas interações com base em conexões implícitas entre os documentos da literatura biológica. Além disso, comprovamos também que nosso modelo é capaz de ordenar as interações da rede através de um valor que mede a evidência de interação das entidades biológicas a partir da coleção de documentos. Para calcular a evidência de interação, o modelo usa o valor da similaridade que o modelo de espaço vetorial atribui às entidades. Dessa forma, nosso modelo é capaz de restringir os melhores resultados e indicar as interações biológicas mais promissoras.

Em nossos resultados, observamos uma diminuição da evidência de interação em sub-redes com maior número de dimensões. Isso acontece em decorrência da dificuldade de se encontrar documentos em que co-ocorram as entidades de todas as dimensões de uma sub-rede com muitas dimensões. Dessa maneira, verificamos que não podemos comparar resultados de sub-redes que possuem um número diferente de dimensões com base apenas no valor da evidência de interação. Essa comparação é equivocada, porque os resultados das sub-redes com maior número de dimensões fornecem indicações mais apuradas de novos relacionamentos. Esses resultados são mais apurados, porque o aumento no número de dimensões contribui para restringir o espaço de busca de novas interações.

Nossos testes de validação demonstram que muitas interações inferidas por nosso modelo em um ano foram confirmadas por patentes publicadas em anos subsequentes. Esses testes também mostram que muitas patentes de confirmação foram encontradas no topo dos *rankings* de resposta de nosso modelo. Considerando as novas

---

interações inferidas através de patentes publicadas até o ano de 2005, por exemplo, nosso modelo confirmou 8% dessas interações. Além disso, 56% dessas confirmações estão entre as 550 primeiras interações novas inferidas em cada sub-rede. Esse resultado é especialmente alto, quando consideramos o decrescente número de patentes biotecnológicas publicadas a partir de 2001.

Como trabalho futuro, construiremos sub-redes a partir de outros campos das patentes (e. g. título, resumo e seção de descrição) e a partir de outras literaturas publicamente disponíveis na *Web* (e. g. artigos científicos e bulas de medicamentos). Também analisaremos e implementaremos o processo de inferência em sub-redes unidimensionais. Para melhorar o processo de inferência de nosso modelo, avaliaremos técnicas de processamento natural de linguagem (PNL), heurísticas, ontologias e tesouros que contribuam para a identificação das entidades na coleção de documentos e também do contexto em que elas são empregadas nessa coleção. Uma das primeiras heurísticas que empregaremos para tratar o contexto semântico das entidades será a avaliação de critérios de proximidade entre as ocorrências das entidades nos documentos. Realizaremos novas análises de nossa estratégia de *ranking*, com o objetivo de aperfeiçoá-la e de propor estratégias novas. Verificaremos o impacto de certificados de adição (artigo 76 da lei nº 9.279, de 14 de maio 1996) na tarefa de construção da rede. Além disso, analisaremos se as patentes de confirmação encontradas no USPTO são dos inventores ou instituições que possuíam as indicações das novas interações nos anos anteriores à publicação dessas patentes. Também verificaremos novas aplicações para nosso modelo, principalmente em outras ciências, como a física, a química e a engenharia.



# Apêndices



## Apêndice A Os Nomes das Entidades Biológicas

Apresentamos abaixo os nomes das entidades biológicas consideradas na construção de nossa rede de interações. As entidades são apresentadas em suas respectivas categorias. Cada linha nas categorias apresenta um grupo de nomes das entidades consideradas em nossos experimentos. Os nomes em cada grupo são separados pelo caracter ";".

### Categoria Alvo Biológico

1. *cgmp specific phosphodiesterase type 5; 3',5'-cyclic-nucleotide phosphodiesterase; cyclic amp phosphodiesterase; phosphodiesterase 5.*
2. *cyclic-gmp phosphodiesterase.*
3. *cyclooxygenase 1.*
4. *cyclooxygenase 2; prostaglandin-endoperoxide synthase 2.*
5. *er-beta; estrogen receptor beta.*
6. *collagen receptor.*
7. *glycoprotein iib/iiia receptor; gpiib-iiia receptor; glycoprotein iib/iiia; gp iib/iiia; gpiib/iiia receptor; fibrinogen receptor; platelet gpiib-iiia; platelet gpiib/iiia receptor; platelet glycoprotein iib-iiia.*
8. *er-alpha; estrogen receptor alpha; oestrogen receptor.*
9. *3-hydroxy-3-methylglutaryl coenzyme a reductase; hmg coa reductase; hmg-coa reductase.*
10. *c-reactive protein.*
11. *progesterone receptor.*
12. *epidermal growth factor receptor 2.*
13. *p2y purinoceptor; adp receptor; p2y receptor.*
14. *serotonin receptor.*
15. *cyclophilin a; peptidyl-prolyl cis-trans isomerase a; rotamase.*
16. *acetylcholinesterase; cholinesterase.*
17. *intrinsic factor.*
18. *choline acetyltransferase; choline o-acetyltransferase; choline acetylase.*
19. *acetylcholine.*
20. *tryptophan hydroxylase.*
21. *epinephrine; adrenaline.*
22. *tumor necrosis factor-alpha; cachectin.*
23. *tumor necrosis factor-beta; lymphotoxin.*

## Categoria Doença

1. *diabetes mellitus type 2; noninsulin-dependent diabetes mellitus; type 2 diabetes; non-insulin-dependent diabetes mellitus.*
2. *alzheimer disease; alzheimer dementia; alzheimer's dementia; alzheimer's disease; alzheimers disease.*
3. *atherosclerosis.*
4. *acquired immunodeficiency syndrome; human immunodeficiency virus; human immunodeficiency virus disease.*
5. *breast cancer; cancer of the breast; cancer of breast; breast adenocarcinoma.*
6. *gouty arthritis; gout.*
7. *parkinson disease; parkinson's disease; parkinsonism; parkinsons disease.*
8. *rheumatoid arthritis.*
9. *lung cancer.*
10. *cardiac ischemia.*
11. *impotence; erectile dysfunction.*
12. *thrombosis.*
13. *arrhythmia; cardiac arrhythmia; dysrhythmia.*
14. *epilepsy.*
15. *ovarian cancer; cancer of the ovary.*
16. *prostate cancer; cancer of the prostate; prostate carcinoma.*
17. *pulmonary arterial hypertension; pulmonary hypertension.*
18. *myocardial infarction; acute myocardial infarction; heart attack.*
19. *postpartum depression; postnatal depression.*
20. *dementia with lewy bodies; lewy body dementia; diffuse lewy body disease; cortical lewy body disease; senile dementia of lewy type.*
21. *pernicious anemia.*
22. *basal cell carcinoma.*

## Categoria Fármaco

1. *sildenafil citrate; viagra.*
2. *rosiglitazone.*
3. *acarbose.*
4. *rivastigmine.*
5. *donepezil; donepezil hydrochloride; aricept.*
6. *sodium valproate; divalproex sodium; divalproex.*
7. *carbamazepine.*



8. *trazadone; desyrel.*
9. *atorvastatin; lipitor.*
10. *ibuprofen; p-isobutylhydratropic acid; advil.*
11. *naproxen; naprosyn; anaprox; bonyl.*
12. *tamoxifen.*
13. *aminoglutethimide.*
14. *hydrochlorothiazide; triamterene.*
15. *acetylsalicylic acid; aspirin.*
16. *cyclosporine; ciclosporin; neoral.*
17. *diclofenac; voltaren; cataflam.*
18. *flecainide.*
19. *propranolol; propranalol; propanolol; propanalol.*
20. *verapamil.*
21. *olanzapine.*
22. *galanthamine; galantamine.*

## Categoria Gene

1. *d2 dopamine receptor; dopamine receptor d2; dopamine d2 receptor.*
2. *androgen receptor; dihydrotestosterone receptor; testosterone receptor; kennedy disease.*
3. *hydroxymethylglutaryl-coa reductase.*
4. *apolipoprotein e.*
5. *leptin.*
6. *peptidylprolyl isomerase a.*
7. *apolipoprotein a-1; apolipoprotein a1; apolipoprotein a 1; apolipoprotein a-i; apolipoprotein ai; apolipoprotein a i.*
8. *interleukin 2; t-cell growth factor; aldesleukin.*
9. *plasminogen activator; urokinase.*
10. *tumor necrosis factor precursor.*
11. *tumor protein p53; tumor suppressor p53; phosphoprotein p53; tumor suppressor is p53.*
12. *major histocompatibility complex class i; major histocompatibility complex class 1.*
13. *major histocompatibility complex class ii.*
14. *transforming growth factor; beta 1.*
15. *cgmp-binding cgmp-specific phosphodiesterase.*

16. *von willebrand factor; von willebrand's factor; factor von willebrand; von willebrands factor.*
17. *endothelin 1.*
18. *hemoglobin alpha 2; hemoglobin alpha chain; alpha-globin.*
19. *peroxisome proliferator-activated receptor gamma; ppar-gamma.*
20. *beta-1 adrenergic receptor; beta-1 adrenoceptor.*

## Apêndice B O Diagrama de Entidades e Relacionamentos (DER)

Diagrama de Entidades e Relacionamentos (DER) da base de dados desenvolvida para o sistema *BioSearch*. A base de dados foi modelada no DBDesigner 4 (DBDesigner, 2009) e criada no sistema gerenciador de banco de dados PostgreSQL 8.3.7-1 (PostgreSQL, 2009).

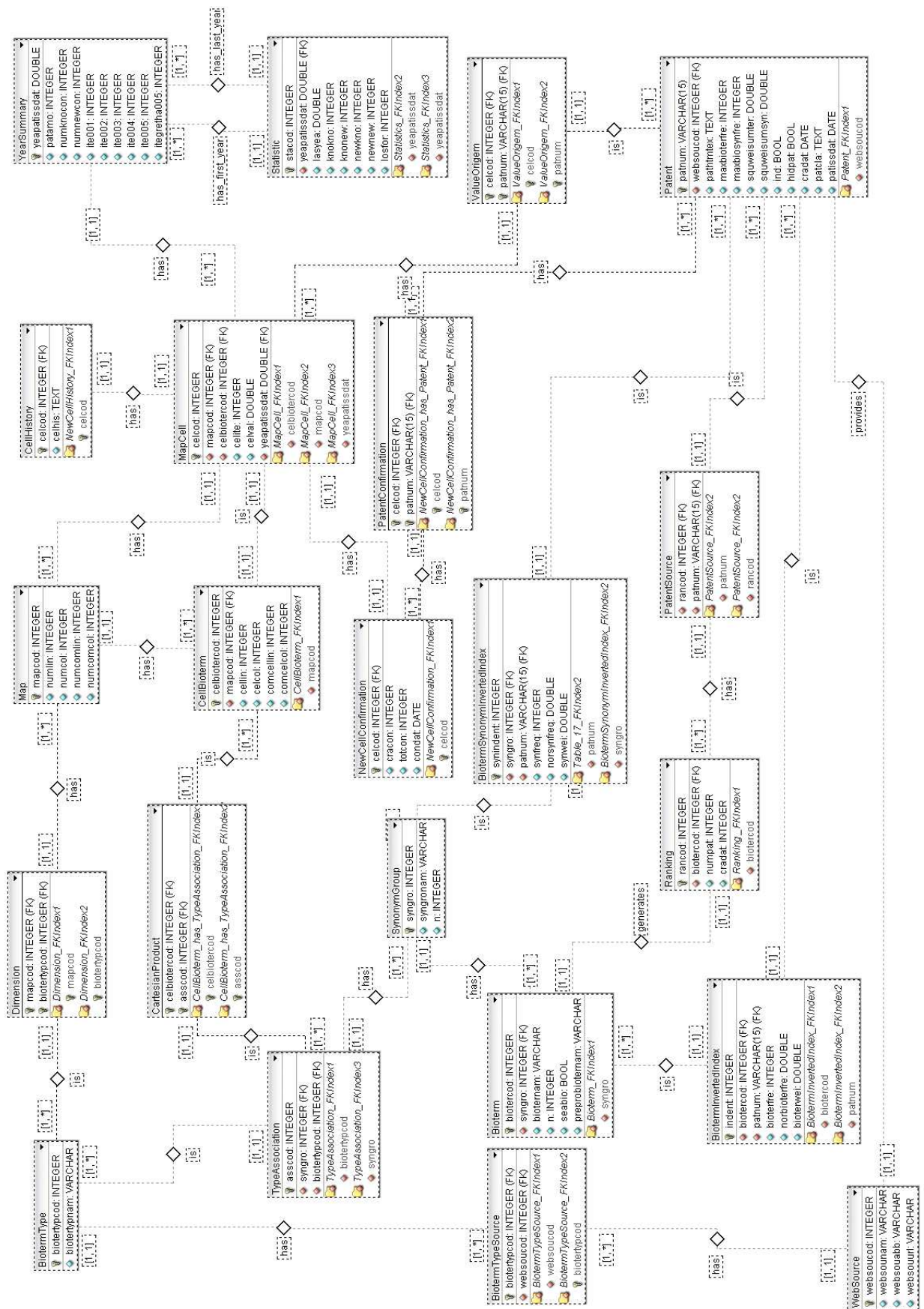
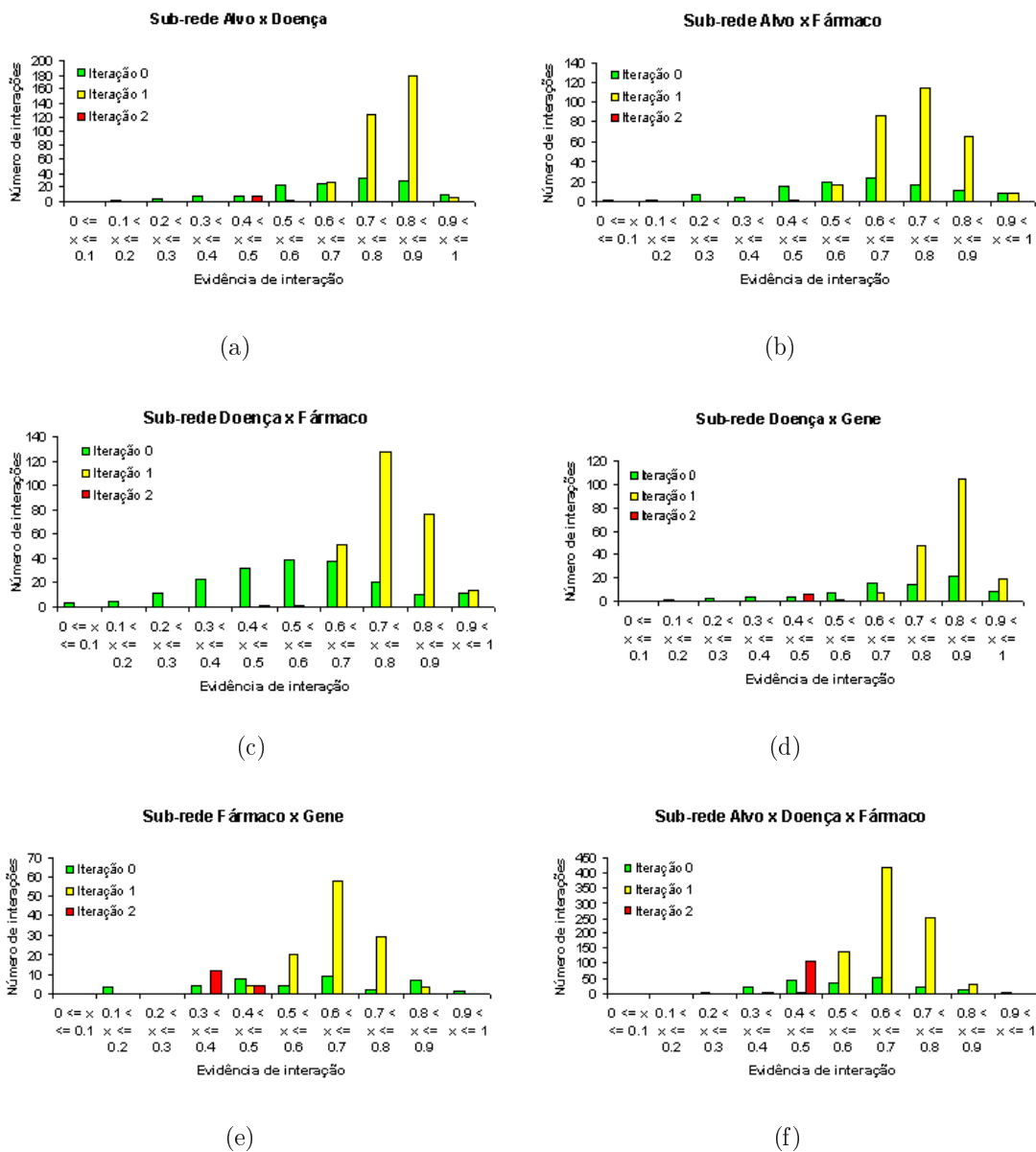


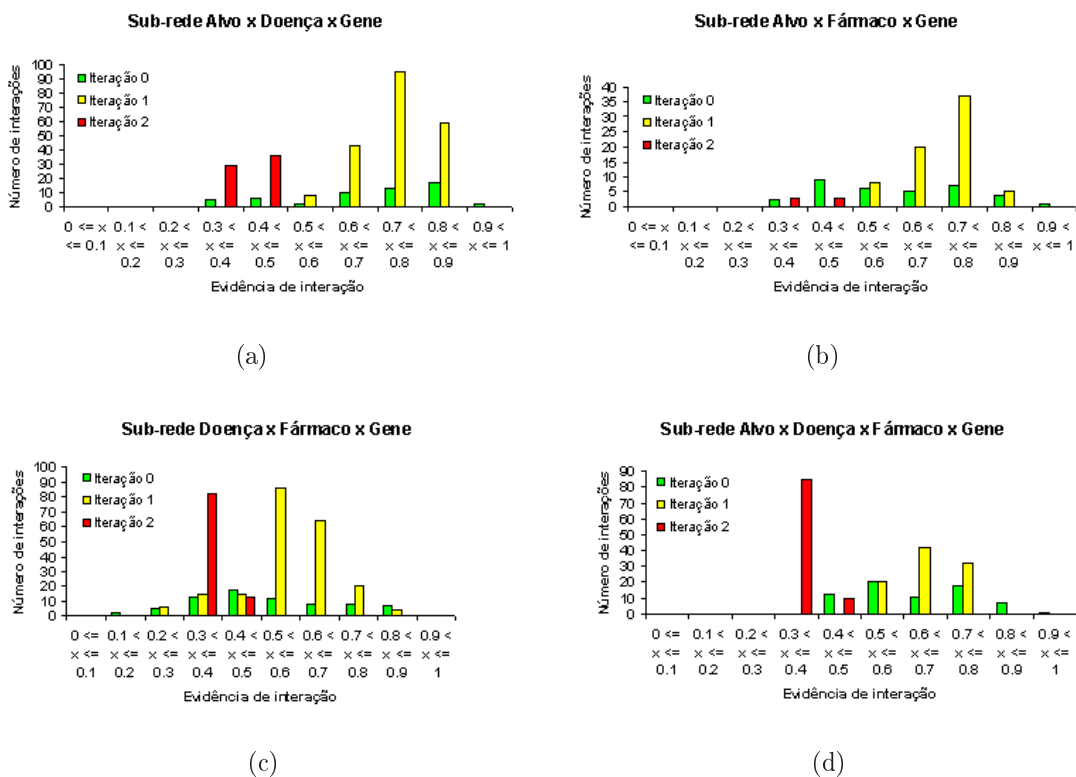
Figura 1: Diagrama de entidades e relacionamentos (DER) da base de dados do sistema *BioSearch*.

## Apêndice C A Distribuição da Evidência de Interação

Distribuição do valor da evidência de interação por sub-rede.



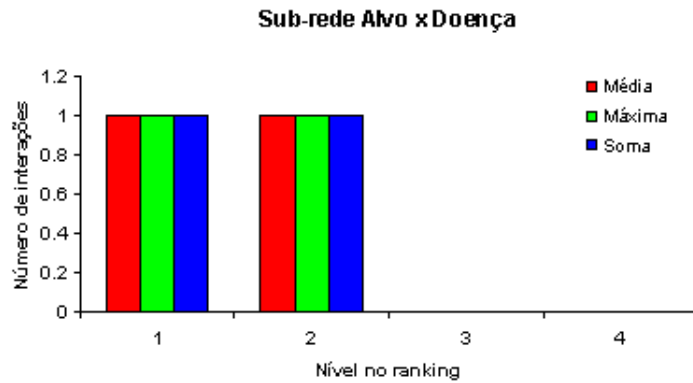
**Figura 2:** Distribuição das evidências de interação (I). (a) Sub-rede *alvo* × *doença*. (b) Sub-rede *alvo* × *fármaco*. (c) Sub-rede *doença* × *fármaco*. (d) Sub-rede *doença* × *gene*. (e) Sub-rede *fármaco* × *gene*. (f) Sub-rede *alvo* × *doença* × *fármaco*.



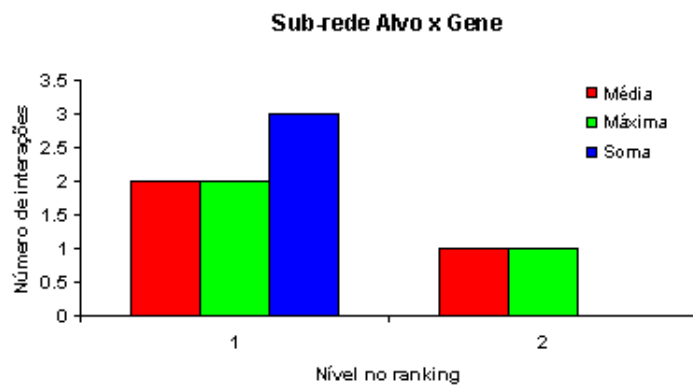
**Figura 3:** Distribuição das evidências de interação (II). (a) Sub-rede *alvo* × *doença* × *gene*. (b) Sub-rede *alvo* × *fármaco* × *gene*. (c) Sub-rede *doença* × *fármaco* × *gene*. (d) Sub-rede *alvo* × *doença* × *fármaco* × *gene*.

## Apêndice D A Distribuição das Patentes de Confirmação

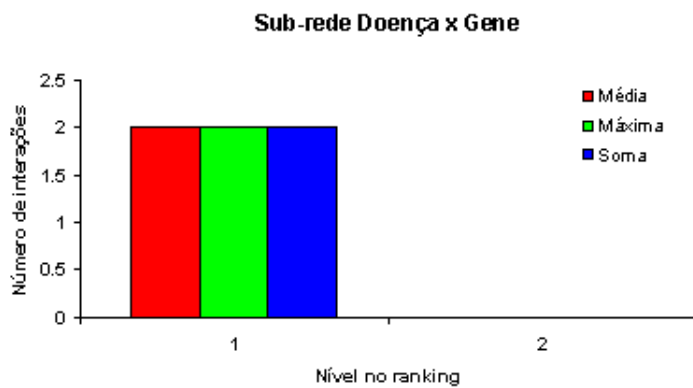
Distribuição das 32 interações inferidas em 2004, que possuem confirmações patenteadas em 2005, no *ranking* das sub-redes. Não foram encontradas patentes de confirmação para as interações da sub-rede *fármaco* × *gene* e da sub-rede *alvo* × *fármaco* × *gene*. Na sub-rede *fármaco* × *gene* foram inferidas 119 novas interações e na sub-rede *alvo* × *fármaco* × *gene*, 76 novas interações. Cada nível dos *rankings* apresentados nos grafos possuem até 100 novas interações.



(a)

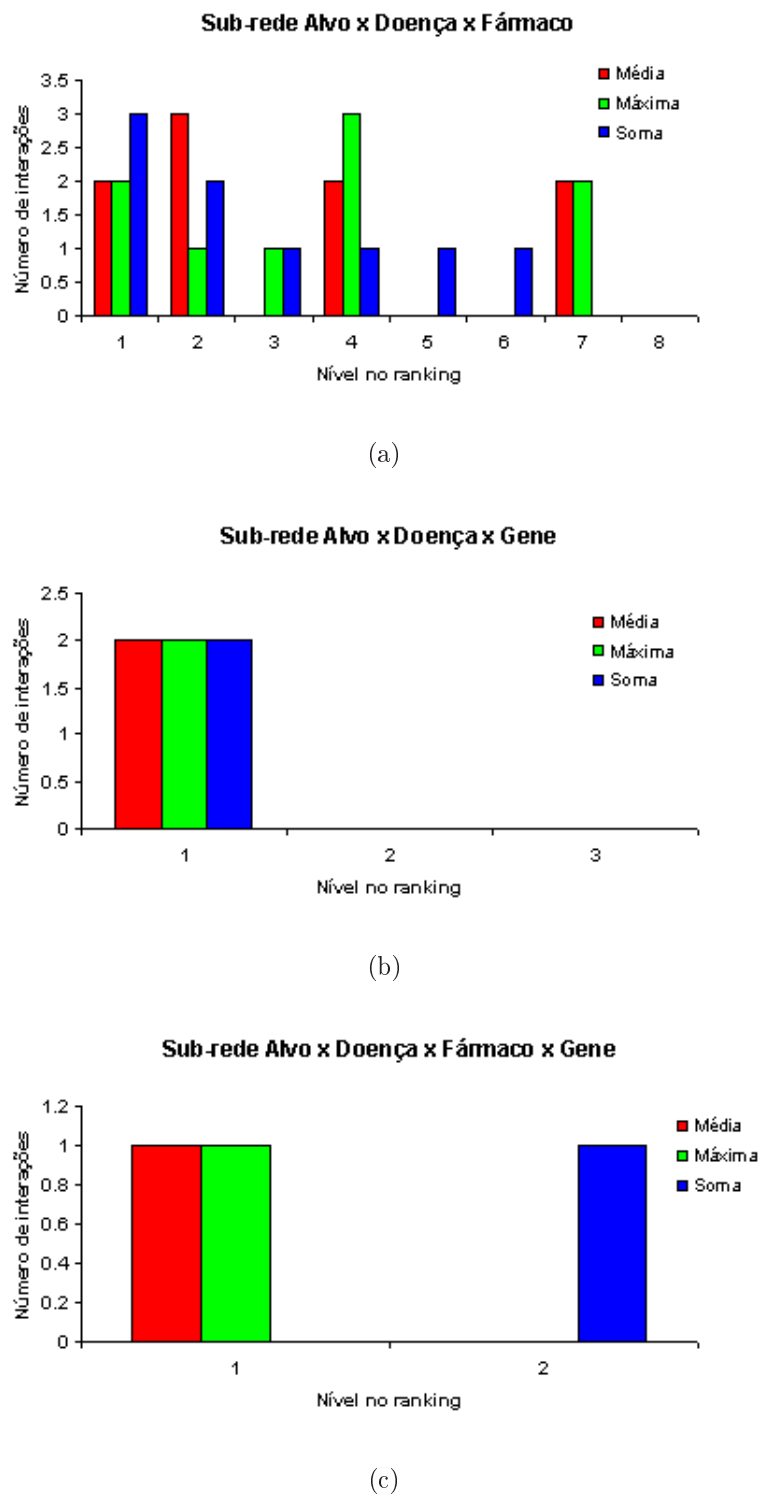


(b)



(c)

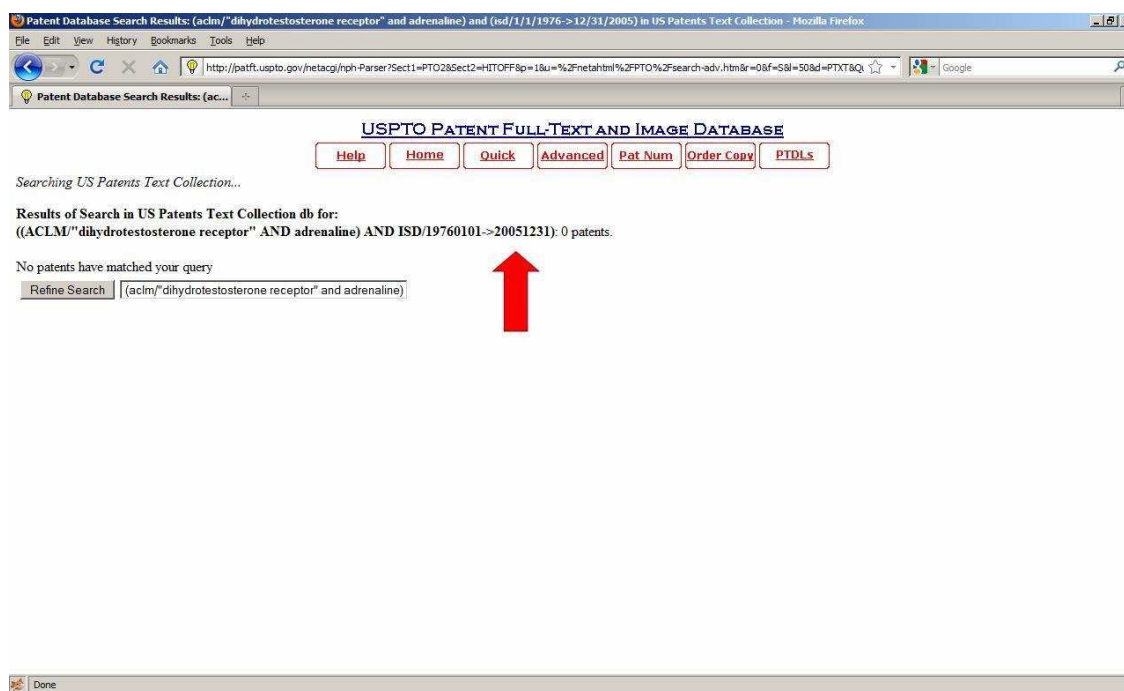
**Figura 4:** Distribuição das interações inferidas em 2004 e que possuem confirmações patenteadas em 2005 (I). (a) Sub-rede *alvo* × *doença*: 348 novas interações, 2 patentes de confirmação. (b) Sub-rede *alvo* × *gene*: 152 novas interações, 3 patentes de confirmação. (c) Sub-rede *doença* × *gene*: 167 novas interações, 2 patentes de confirmação.



**Figura 5:** Distribuição das interações inferidas em 2004 e que possuem confirmações patenteadas em 2005 (II). (a) Sub-rede *alvo* × *doença* × *fármaco*: 786 novas interações, 9 patentes de confirmação. (b) Sub-rede *alvo* × *doença* × *gene*: 242 novas interações, 2 patentes de confirmação. (c) Sub-rede *alvo* × *doença* × *fármaco* × *gene*: 175 novas interações, 1 patente de confirmação.

## Apêndice E A Confirmação do Sistema de Busca do USPTO

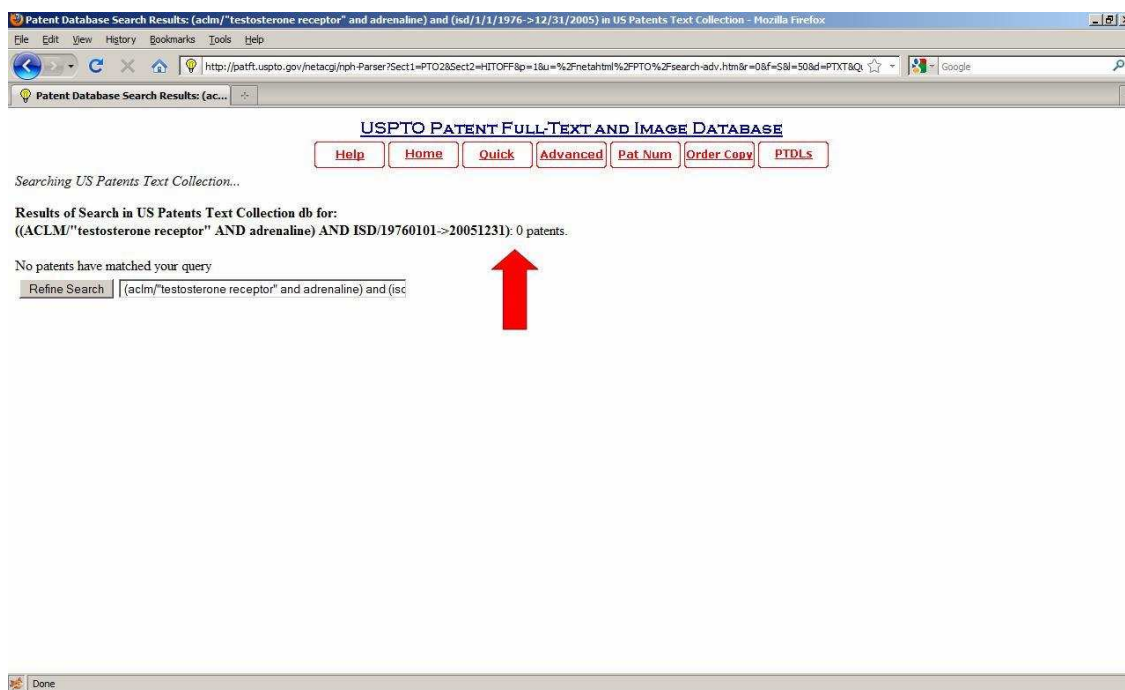
Páginas do USPTO que confirmam a inexistência de patentes publicadas no período que vai de 01/01/1976 até 31/12/2005 em que haja a co-ocorrência das entidades adrenalina e receptor de androgênio na seção de reivindicação. Cada página mostra esta confirmação para as combinações dos grupos de nomes que consideramos para as entidade adrenalina e receptor de androgênio em nossos experimentos.



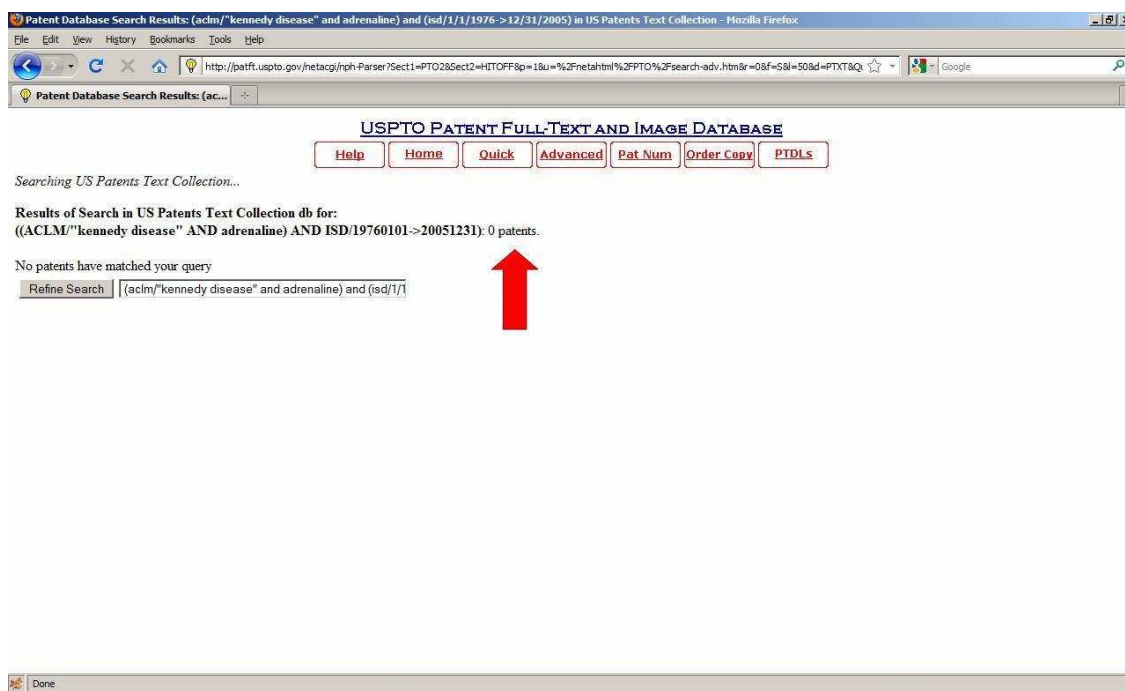
(a)

**Figura 6:** Página do USPTO confirmando a inexistência de patentes publicadas no período de 01/01/1976 até 31/12/2005 em que haja co-ocorrência das entidades adrenalina e receptor de androgênio na seção de reivindicação (I). (a) Consulta formada pelos nomes *dihydrotestosterone receptor* e *adrenaline*.



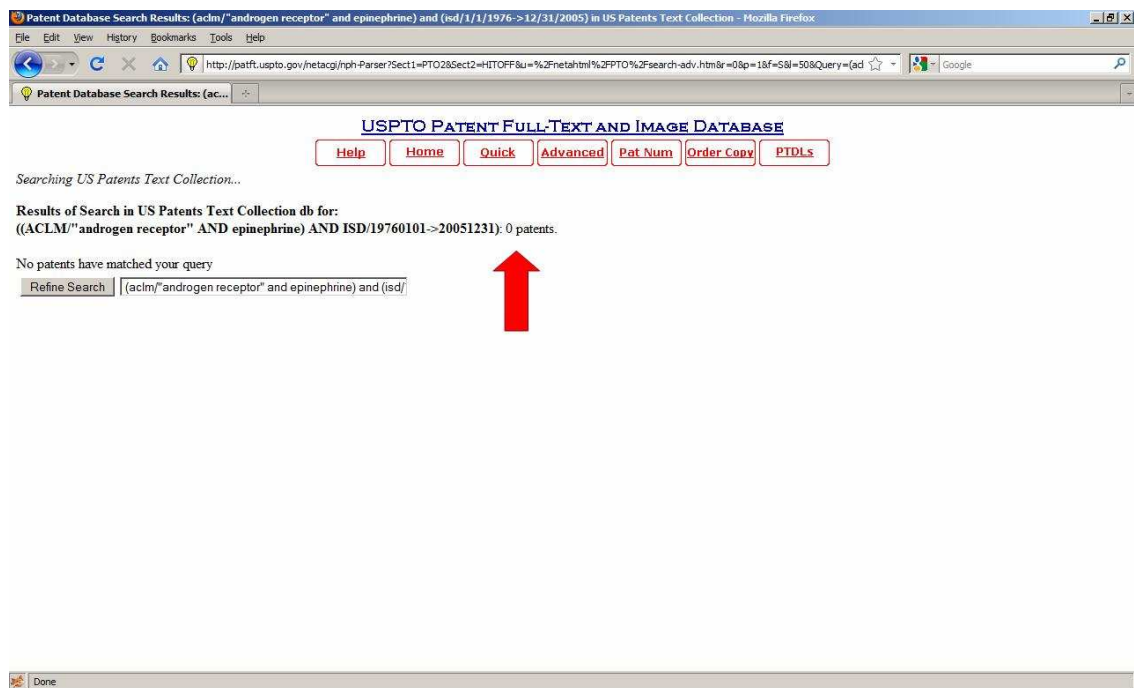


(a)

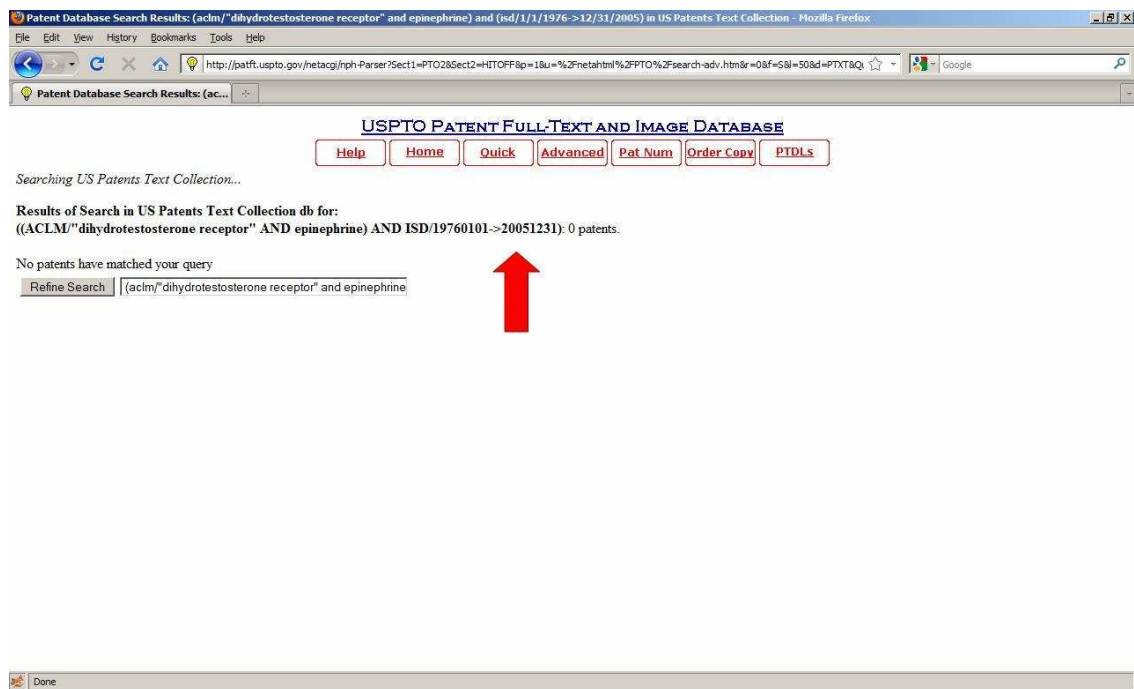


(b)

**Figura 7:** Página do USPTO confirmando a inexistência de patentes publicadas no período de 01/01/1976 até 31/12/2005 em que haja co-ocorrência das entidades adrenalina e receptor de androgênio na seção de reivindicação (II). (a) Consulta formada pelos nomes *testosterone receptor* e *adrenaline*. (b) Consulta formada pelos nomes *kennedy disease* e *adrenaline*.

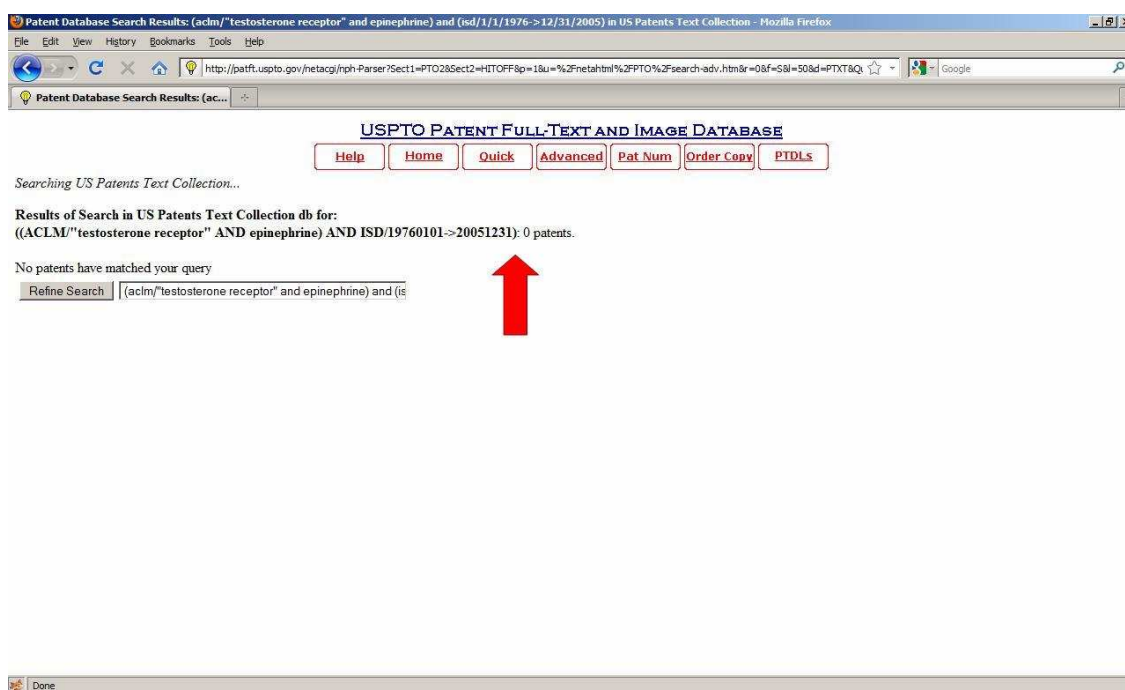


(a)

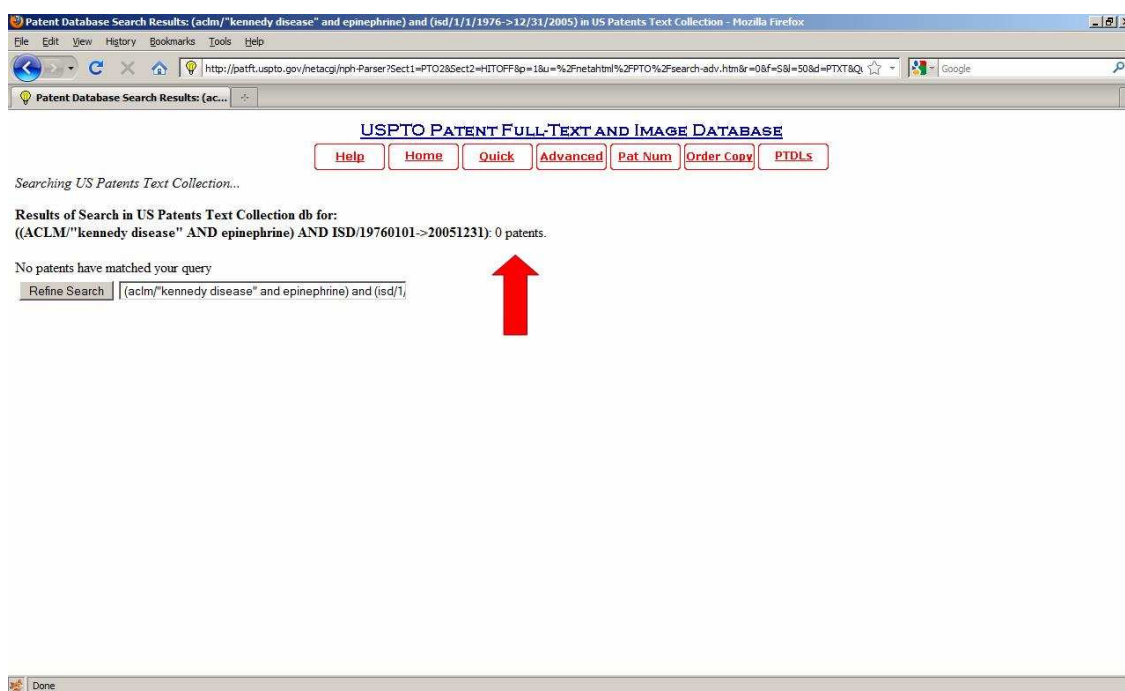


(b)

**Figura 8:** Página do USPTO confirmando a inexistência de patentes publicadas no período de 01/01/1976 até 31/12/2005 em que haja co-ocorrência das entidades adrenalina e receptor de androgênio na seção de reivindicação (III). (a) Consulta formada pelos nomes *androgen receptor* e *epinephrine*. (b) Consulta formada pelos nomes *dihydrotestosterone receptor* e *epinephrine*.



(a)



(b)

**Figura 9:** Página do USPTO confirmando a inexistência de patentes publicadas no período de 01/01/1976 até 31/12/2005 em que haja co-ocorrência das entidades adrenalina e receptor de androgênio na seção de reivindicação (IV). (a) Consulta formada pelos nomes *testosterone receptor* e *epinephrine*. (b) Consulta formada pelos nomes *kennedy disease* e *epinephrine*.



# Bibliografia

- Alani, H., Kim, S., Millard, D. E., Weal, M. J., Hall, W., Lewis, P. H., and Shadbolt, N. R. (2003). Automatic ontology-based knowledge extraction from web documents. *IEEE Intelligent Systems: Natural Language Processing*, **18**, 14–21.
- Alm, E. and Arkin, A. P. (2003). Biological networks. *Current Opinion in Structural Biology*, **13**, 193–202.
- Alon, U. (2003). Biological networks: The tinkerer as an engineer. *Science*, **301**, 1866–1867.
- Ambesi-Impiombato, A. and di Bernardo, D. (2006). Computational biology and drug discovery: from single-target to network drugs. *Current Bioinformatics*, **1**, 3–13.
- Baeza-Yates, R. A. and Ribeiro-Neto, B. A. (1999). *Modern Information Retrieval*. ACM Press / Addison-Wesley.
- Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005). Reverse engineering of regulatory networks in human b cells. *Nature Genetics*, **37**(4), 382–390.
- Berendt, B., Hotho, A., and Stumme, G. (2002). Towards semantic web mining. In *Proceedings of the International Semantic Web Conference (ISWC02)*, pages 264–278.
- BioSearch (2009). <http://luar.dcc.ufmg.br/BioSearch>.
- Brazma, A. and Vilo, J. (2000). Gene expression data analysis. *Federation of European Biochemical Societies*, **480**, 17–24.
- Bruza, P. and Weeber, M. (2008). *Literature-Based Discovery*. Springer.
- Butcher, E. C., Berg, E. L., and Kunkel, E. J. (2004). Systems biology in drug discovery. *Nature Biotechnology*, **22**(10), 1253–1259.

- Campillos, M., Kuhn, M., Gavin, A., Jensen, L. J., and Boork, P. (2008). Drug target identification using side-effect similarity. *Science*, **321**, 263–266.
- Chen, X., Ji, Z. L., and Chen, Y. Z. (2002). Ttd: Therapeutic target database. *Nucleic Acid Research*, **30**(1), 412–415.
- Cheung, K., k. Y. Yip, Smith, A., deKnikker, R., Masiar, A., and Gerstein, M. (2005). Yeasthub: A semantic web use case for integrating data in the life sciences domain. *Bioinformatics*, **21 Suppl. 1**, i85–i96.
- Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001). *Introduction to Algorithms*. MIT Press and McGraw-Hill, second edition.
- Coulouris, G., Dollimore, J., and kindberg, T. (2005). *Distributed Systems: Concepts and Design*. Addison Wesley/Pearson Education, fourth edition.
- Csermely, P., Ágoston, V., and Pongor, S. (2005). The efficiency of multi-target drugs: The network approach might help drug design. *Trends in Pharmacological Sciences*, **26**, 178–182.
- DBDesigner (2009). <http://fabforce.net/dbdesigner4/>.
- Driver, P. J. (2009). <http://jdbc.postgresql.org/index.html>.
- DrugBank (2009). <http://www.drugbank.ca/>.
- drugs.com (2009). <http://www.drugs.com/>.
- Fall, C. J., Törösvári, A., Benzineb, K., and Karetka, G. (2003). Automated categorization in the international patent classification. In *Proceedings of the ACM SIGIR Forum 37(1)*, pages 10–25.
- Fan, W., Wallace, L., Rich, S., and Zhang, Z. (2006). Tapping the power of text mining. *Communications of the ACM*, **49**(9), 76–82.
- Fensel, D. (2002). Ontology-based knowledge management. *IEEE Computer: Cover Feature*, pages 56–59.
- Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, **303**, 799–805.
- Gardner, T. S., di Bernardo, D., Lorenz, D., and Collins, J. J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, **301**, 102–105.

- Glenisson, P., Antal, P., Mathys, J., Moreau, Y., and Moor, B. D. (2003). Evaluation of the vector space representation in text-based gene clustering. In *In Proceedings of the Eighth Annual Pacific Symposium on Biocomputing*, pages 391–402, Lihue, Hawaii, USA.
- Gopalacharyulu, P. V., Lindfors, E., Bounsaythip, C., Kivioja, T., Yetukuri, L., Hollmén, J., and Orešič, M. (2005). Data integration and visualisation system for enabling conceptual biology. *Bioinformatics*, **21 Suppl. 1**, i177–i185.
- Grabe, N. and Neuber, K. (2005). A multicellular systems biology model predicts epidermal morphology, kinetics and  $ca^{2+}$  flow. *Bioinformatics*, **21**(17), 3541–3547.
- Guimerà, R., Sales-Pardo, M., and Amaral, L. A. N. (2007). A network-based method for target selection in metabolic networks. *Bioinformatics*, **23**(13), 1616–1622.
- GuoDong, Z. and Min, Z. (2007). Extracting relation information from text documents by exploring various types of knowledge. *Information Processing and Management*, **43**(4), 969–982.
- Hewett, M., Oliver, D. E., Rubin, D. L., Easton, K. L., Stuart, J. M., Altman, R. B., and Klein, T. E. (2002). Pharmgkb: the pharmacogenetics knowledge base. *Nucleic Acid Research*, **30**(1), 163–165.
- HGNC (2009). <http://www.genenames.org/>.
- Holme, P. (2008). Model validation of simple-graph representations of metabolism. *Journal of the Royal Society Interface*.
- Hood, L. and Perlmutter, R. M. (2004). The impact of systems approaches on biological problems in drug discovery. *Nature Biotechnology*, **22**(10), 1215–1217.
- Hopkins, A. (2007). Network pharmacology: Network biology illuminates our understanding of drug action. *Nature Biotechnology*, **25**(10), 1110–1111.
- Horn, C. E. V. and Lipsey, C. E. (2004). *Biotechnology Innovation Report 2004 - Benchmarks*. Finnegan, Henderson, Farabow, Garrett and Dunner, LLP.
- Hristovski, D. and B. Peterlin, J. A. Mitchell, S. M. H. (2005). Using literature-based discovery to identify disease candidates genes. In *International Journal of Medical Informatics*, volume 74, pages 289–298.
- Hristovski, D., Friedman, C., Rindfleisch, T. C., and Peterlin, B. (2006). Exploiting semantic relations for literature-based discovery. In *American Medical Informatics Association Symposium Proceedings*, pages 349–353.

- Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929–934.
- Ihlenfeldt, W., Voigt, J. H., Bienfait, B., Oellien, F., and Nicklaus, M. C. (2002). Enhanced cactus browser of the open nci database. *American Chemical Society: Journal of Chemical Information and Modeling*, **42**, 46–57.
- Java (2009). <http://java.sun.com/>.
- JDBC (2009). <http://java.sun.com/products/jdbc/overview.html>.
- JDK (2009). <http://java.sun.com/javase/6/docs/technotes/guides/index.html#jre-jdk>.
- Jflex (2007). <http://jflex.de/>.
- Jung, G. S. and Gudivada, V. N. (1995). Automatic determination and visualization of relationships among symptoms for building medical knowledge bases. In *Proceedings of the 1995 ACM Symposium on Applied Computing*, pages 101–107, Nashville, Tennessee, United States.
- Karthikeyan, M., Krishnam, S., and Pandey, A. K. (2006). Harvesting chemical information from the internet using a distributed approach: Chemxtreme. *American Chemical Society: Journal of Chemical Information and Modeling*, **46**, 452–461.
- KEGG (2009). <http://www.genome.jp/kegg/>.
- KI (2009). <http://www.mic.ki.se/diseases/alphalist.html>.
- Kitano, H. (2002). Computational systems biology. *Nature*, **420**, 206–210.
- Kostoff, R. N. (2008a). Literature-related discovery (lrd): Introduction and background. *Technological Forecasting and Social Change*, **75**, 165–185.
- Kostoff, R. N. (2008b). Literature-related discovery (lrd): Potential treatments for cataracts. *Technological Forecasting and Social Change*, **75**, 215–225.
- Kostoff, R. N. (2008c). *Where is the Discovery in Literature-Based Discovery?*, chapter 5, pages 57–72. Springer.
- Kostoff, R. N. and Briggs, M. B. (2008). Literature-related discovery (lrd): Potential treatments for parkinson’s disease. *Technological Forecasting and Social Change*, **75**, 226–238.



- Kostoff, R. N., Block, J. A., Solka, J. L., Briggs, M. B., Rushenberg, R. L., Stump, J. A., Johnson, D., Lyons, T. J., and Wyatt, J. R. (2008a). Literature-related discovery (lrd): Lessons learned, and future research directions. *Technological Forecasting and Social Change*, **75**, 276–299.
- Kostoff, R. N., Briggs, M. B., Solka, J. L., and Rushenberg, R. L. (2008b). Literature-related discovery (lrd): Methodology. *Technological Forecasting and Social Change*, **75**, 186–202.
- Kostoff, R. N., Briggs, M. B., and Lyons, T. J. (2008c). Literature-related discovery (lrd): Potential treatments for multiple sclerosis. *Technological Forecasting and Social Change*, **75**, 239–255.
- Kostoff, R. N., Block, J. A., Stump, J. A., and Johnson, D. (2008d). Literature-related discovery (lrd): Potential treatments for raynaud’s phenomenon. *Technological Forecasting and Social Change*, **75**, 203–214.
- Kostoff, R. N., Solka, J. L., Rushenberg, R. L., and Wyatt, J. A. (2008e). Literature-related discovery (lrd): Water purification. *Technological Forecasting and Social Change*, **75**, 256–275.
- Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J., Subramanian, A., Ross, K. N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S. A., Haggarty, S. J., Clemons, P. A., Wei, R., Carr, S. A., Lander, E. S., and Golub, T. R. (2006). The connectivity map: using gene-expression signatures to connect small molecules, genes and disease. *Science*, **313**, 1929–1935.
- Larkey, L. S. (1999). A patent search and classification system. In *Proceedings of the Fourth ACM conference on Digital libraries*, pages 179–187, Berkeley, California, United States.
- Lechter, M. A., Clifford, E. C., Famiglio, R. B., and Joenk, R. J. (1990). *Successful Patents and Patenting for Engineers and Scientists*. The Institute of Electrical and Electronics Engineers, Inc., New York.
- Li, J., Zhu, X., and Chen, J. Y. (2009). Building disease-specific drug-protein connectivity maps from molecular interaction networks and pubmed abstracts. *PLoS Computational Biology*, **5**(7).
- MC (2009). <http://www.mayoclinic.com/>.
- MedlinePlus (2009). <http://medlineplus.gov/>.

- Mukherjea, S. and Bamba, B. (2004). Biopatentminer: An information retrieval system for biomedical patents. In *Proceedings of 30th Very Large Database (VLDB) Conference*, pages 1066–1077.
- Naylor, S. (2004). Systems biology, information, disease and drug discovery. *Drug Discovery World*, **5**, 23–33.
- NCBI (2009). <http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene>.
- Neshich, G., Rocchia, W., Mancini, A. L., Yamagishi, M. E. B., Kuser, P. R., Fileto, R., Baudet, C., Pinto, I. P., Montagner, A. J., Palandrani, J. F., Krauchenco, J. N., Torres, R. C., Souza, S., Togawa, R. C., and Higa, R. H. (2004). Java protein dossier: a novel web-based data visualization tool for comprehensive analysis of protein structure. *Nucleic Acid Research*, **32**, **Web Server issue**, W595–W601.
- NetBeans (2009). <http://www.netbeans.org/>.
- Ontology, G. (2008). <http://www.geneontology.org/>.
- patient.uk (2009). <http://www.patient.co.uk/>.
- PostgreSQL (2009). <http://www.postgresql.org/>.
- PubMed (2009). <http://www.ncbi.nlm.nih.gov/pubmed>.
- Salton, G. and McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill Book Co., New York.
- Sastry, K. S. R., Karpova, Y., Prokopovich, S., Smith, A. J., Essau, B., Gersappe, A., Carson, J. P., Weber, M. J., Register, T. C., Chen, Y. Q., Penn, R. B., and Kulik, G. (2007). Epinephrine protects cancer cells from apoptosis via activation of camp-dependent protein kinase and bad phosphorylation. *Journal of Biological Chemistry*, **282**(19), 14094–14100.
- Schuffenhauer, A., Zimmermann, J., Stoop, R., van der Vyver, J., Lecchini, S., and Jacoby, E. (2002). An ontology for pharmaceutical ligands and its application for in silico screening and library design. *American Chemical Society: Journal of Chemical Information and Modeling*, **42**, 947–955.
- Servlet, J. (2009). <http://java.sun.com/products/servlet/>.
- Shinmori, A., Okumura, M., Marukawa, Y., and Iwayama, M. (2003). Patent claim processing for readability: Structure analysis and term explanation. In *In Proceedings of the Workshop on Patent Corpus Processing*, volume 20, pages 56–65, Sapporo, Japan.

- Silveira, M. L. (2003). *Recuperação Vertical de Informação: Um Estudo de Caso na Área Jurídica*. Master's thesis, Universidade Federal de Minas Gerais.
- Silverman, R. (2004). *The Organic Chemistry of Drug Design and Drug Action*, chapter 2. Elsevier, second edition.
- Smalheiser, N. R. and Swanson, D. R. (1998). Using arrowsmith: a computer-assited approach to formulating and assessing scientific hypotheses. *Computer Methods and Programs in Biomedicine*, **57**, 149–153.
- Smalheiser, N. R. and Torvik, V. I. (2008). *The Place of Literature-Based Discovery in Contemporary Scientific Practice*, chapter 2, pages 13–22. Springer.
- Swanson, D. R. (1986). Fish-oil, raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, **30**(1), 7–18.
- Swanson, D. R. (1990). Medical literature as a potencial source of new knowledge. *Bulletin of the Medical Library Association*, **78**(1), 29–37.
- Swanson, D. R., Smalheiser, N. R., and Torvik, V. L. (2006). Ranking indirect connections in literature-based discovery: The role of medical subject headings. *Journal of the American Society for Information Science and Technology*, **57**(11), 1427–1439.
- Tan, P., Kumar, V., and Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 32–41, Edmonton, Alberta, Canada.
- TFD (2009). <http://www.thefreedictionary.com/>.
- Tiffin, N., Kelso, J. F., Powell, A. R., Pan, H., Bajic, V. B., and Hide, W. A. (2005). Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Research*, **33**(5), 1544–1552.
- Trippe, A. J. (2003). Patinformatics: tasks to tools. *Elsevier: World Patent Information*, **25**, 211–221, doi: 10.1016/S0172-2190(03)00079-6.
- Tseng, Y., Lin, C., and Lin, Y. (2007). Text mining techniques for patent analysis. *Information Processing & Management*, **In Press, Corrected Proof**.
- TTD (2007). <http://xin.cz3.nus.edu.sg/group/ttd/ttd.asp>.
- USPTO (2009). <http://www.uspto.gov/>.

- Weeber, M., Klein, H., de Jong-van den Berg, L. T. W., and Vos, R. (2001). Using concepts in literature-based discovery: Simulating swanson's raynaud-fish oil and migraine-magnesium discoveries. *Journal of the American Society for Information Science and Technology*, **52**(7), 548–557.
- Wendlandt, E. B. and Driscoll, J. R. (1991). Incorporating a semantic analysis into a document retrieval strategy. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 270–279, Chicago, Illinois, United States.
- Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006). Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acid Research*, **34**, Database issue, D668–D672.
- Witten, I. H., Moffat, A., and Bell, T. C. (1999). *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann Publishing, second edition.
- WordNet (2008). <http://wordnet.princeton.edu/>.
- Wren, J. D., Bekeradjian, R., Stewart, J. A., Shohet, R. V., and Garner, H. R. (2004). Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics*, **20**(3), 389–398.
- Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., and Kanehisa, M. (2008). Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, **24**, i232–i240.
- Ziviani, N. (2007). *Projeto de Algoritmos com Implementações em Java e C++*. Thomson Learning.