CARLOS HENRIQUE DA SILVEIRA

PROTEIN CUTOFF SCANNING: APLICAÇÃO DA VARREDURA EXAUSTIVA DE DISTÂNCIAS INTER-RESÍDUOS NA ANÁLISE DE CONTATOS INTRACADEIA EM PROTEÍNAS GLOBULARES.

Belo Horizonte

Fevereiro de 2008

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

CARLOS HENRIQUE DA SILVEIRA

PROTEIN CUTOFF SCANNING: APLICAÇÃO DA VARREDURA EXAUSTIVA DE DISTÂNCIAS INTER-RESÍDUOS NA ANÁLISE DE CONTATOS INTRACADEIA EM PROTEÍNAS GLOBULARES

Um estudo comparativo de técnicas de prospecção de contatos dependentes e independentes de distâncias delimitadoras (*cutoff*).

Projeto de tese apresentado ao Curso de Pós-graduação em Bioinformática da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Doutor em Bioinformática.

Orientador — Prof. Dr. Marcelo Matos Santoro
Laboratório Marcos Luiz dos Mares-Guia de Enzimologia e Físico-Química de Proteínas, Departamento de Bioquímica-Imunologia, Instituto de Ciências Biológicas – ICB, Universidade Federal de Minas Gerais – UFMG, Belo Horizonte – MG.

Co-Orientador — Prof. Dr. Carlos Henrique Inácio Ramos
Departamento de Química Orgânica, Instituto de Química - IQ, Universidade Estadual de Campinas – UNICAMP, Campinas – SP.

Co-Orientador — Prof. Dr. Wagner Meira Junior
Departamento de Ciência da Computação, Instituto de Ciências Exatas – ICEx, Universidade Federal de Minas Gerais – UFMG, Belo Horizonte – MG.

Co-Orientador — Dr. Goran Neshich
Núcleo de Bioinformática Estrutural, Centro Nacional de Pesquisa Tecnológica em Informática para a Agricultura – CNPTIA, Empresa Brasileira de Pesquisa Agropecuária – EMBRAPA, Campinas – SP.

"*Nadie rebaje a lágrima o reproche*
Esta declaración de la maestria
De Dios, que com magnifica ironia
Me dio a la vez los libros e la noche"

Jorge Luis Borges
Poema de los Dones.

"*Qualquer coisa, conforme se considera, é um assombro ou um estorvo, um tudo ou um nada, um caminho ou uma preocupação. Considerá-la cada vez de um modo diferente é renová-la, multiplicá-la por si mesma. É por isso que o espírito contemplativo que nunca saiu da sua aldeia tem contudo à sua ordem o universo inteiro. Numa cela ou num deserto está o infinito. Numa pedra dorme-se cosmicamente*."

Fernando Pessoa
Livro do Desassossego.

Agradecimentos

Já houve um sem número de momentos em que eu achei que jamais escreveria os agradecimentos de uma tese. Temia que essa escrita se tornasse um daqueles textos imaginários, que a gente compõe só no pensamento, nas noites tristes de insônia. Estar escrevendo esses agradecimentos agora me é tão estranho, tão surreal, tão improvável que fico pensando se não estou apenas cumprindo os angustiados desejos da imaginação de alguém. Talvez isso explique esse sentimento de atemporalidade que me invade, como se de repente uma parte do tempo curvasse sobre si mesma, e fizesse desse fugaz momento uma eternidade local.

Aproveito o delírio do meu devaneador para perpetuar aqui a minha gratidão a todos aqueles que contribuíram para o moto-contínuo dessa tese. Importante considerar que em textos idílicos como este, a cronologia e a ordem dos eventos não têm muita relevância. Todos têm o mesmo peso onírico. Já dizia João de Pessoas que o que faz um rio não é só a sua nascente, mas todos os seus afluentes.

Agradeço, pois:

À Deus ou ao Princípio Antrópico a existência, a dor e a alegria de ser;

Aos meus pais, Sô Eduardo e Dona Lurdinha, o dom da vida, a educação pelo exemplo, a infinita força da humildade, da luta, da alegria festeira não condicionada ao ter e ao poder, pelas orações, pelo amor sem distâncias, sem tempo, sem limites;

À Laila, pelo amor, pela cadência, pela vitalidade, pela ternura, pelo companheirismo, pela integridade de sua pessoa, pela sua alma que não tem tamanho. Só o verdadeiro amor suporta tantos sacrifícios! **MUITO OBRIGADO!** Essa tese eu dedico à você!

À minha filha, Luana, pelo sentido da vida, pela razão de viver, pela inocência sapeca, pela alegria sincera, descontaminada, pelo centro da falinha, pelos infinitos desenhos que adornam as cópias dos artigos citados nessa tese.

Aos meus irmãos, Márcio (Boizezim) e Elaine (Beeeeeeem). Eu não seria assim sem vocês. Obrigado pela força contínua e irrestrita. Obrigado por vocês existirem. Bruna e Valdei, sei que estavam torcendo também.

À Bita, minha "sogrinha", pela força imaterial e imprescindível ajuda material. Ao Sô Ricardo, pela força espiritual. Às minhas amigas do outro lado Joana D'arc, Dona Aparecida, Ir. Marta, Vovó São.

Ao meu orientador e amigo, Prof. Marcelo Santoro. Pela inspiração, pelo exemplo, pelo cuidado, pela sabedoria, pelo amor ao conhecimento. Sou muito grato por ter tido a oportunidade de compartilhar a alegria de me formar um cientista ao seu lado. Obrigado, mestre !

Aos meus co-orientadores, formais e informais: Carlos Ramos, Wagner Meira, Goran Neshich, Raul Habesch, Júlio Lopes. Nos momentos cruciais vocês estavam sempre presentes. Foi no ombro de vocês que eu pude ver mais longe.

Carlos Ramos: seu apôio às minhas viagens em Campinas foram fundamentais. Minha eterna gratidão. Goran Neshich: obrigado por abrir seu laboratório a essa tese. Meira, jamais vou esquecer as suas pertinentes observações. Júlio, obrigado pela presença constante.

Ao Raul, meu amigo estranho, que merece muito mais linhas e entrelinhas que esse espaço pode oferecer: obrigado do fundo da minha alma por tudo que você é e faz.

Aos coordenadores do curso durante minha passagem pela Bioinfo, professores Beirão, Sérgio Campos, Glória Franco.

Aos super-amigos da liga fantástica do Doutorado: Raquel, Cristina, Caio e Waisberg. Esta pós não teria a menor graça sem vocês. Raquel, eu lhe disse: você é um dos resultados dessa tese. Tenho muito orgulho de você. Cris, você é o lado mais alegre desse doutorado, e sem seu senso prático, tudo seria mais difícil. Caio, você é o mano mais velho, e como tal me salvou nestes momentos finais que foram tão difíceis. Meu muito obrigado (sem beijinhos)! Waisberg, você é uma raridade como intelectual e pessoa. Não sei como agradecer sua imensa ajuda, não só nas discussões temáticas, como também nos inúmeros *papers* que com toda boa vontade você me arrumou. Não posso deixar de registrar minha gratidão pelas discussões estatísticas com o Deive e Bráulio. À todos os demais amigos de doutorado, cujos nomes pela grande enumeração eu não saberia citar.

Ao Douglas! A Raquel está para início do meu doutorado assim como o Douglas está para o fim. Você foi fundamental na consolidação de tudo que está registrado aqui. Eu agradeço recursivamente ao destino por nos ter cruzado os caminhos. Aqui cabe um louvor ao Prof Meira, pelo seu grande talento em descobrir e apoiar novos talentos. À Kellen por sua efêmera mas não menos importante contribuição.

Aos amigos de todos os tempos e lugares: Leo, Rico, Phodão, Rogério e Carla, Marquinhos e Cida, Jamil, Wandeca, Zema, Jader, Dr. Doido, Prof. Letra, Dr. Omni, tia Poly, Jacque, Myrinha, Lu e Agenor e outros que minha danificada memória foi incapaz de lembrar. Ainda que em tempos e lugares diferentes, essa tese não seria a mesma sem vocês.

À Inspetoria Madre Mazzarello, em especial Ir. Eliane, Ir. Olga, Ir. Maria Helena e Ir. Divina, Ir.Arlete, Ir. Amélia. Num momento muito crítico, vocês me acolheram. Ir. Eliane, a senhora foi fundamental. Muito obrigado!

Aos amigos do extinto GREI – Grupo de Estudos Interdisciplinares da UFMG, muito especialmente, aos professores Rogério Parentoni, Romeu Guimarães, Hugo Mari, e Chico Muleta. Foi com vocês que eu comecei minha pós-graduação.

Aos meus mestres póstumos, Dr. Marcos Luiz dos Mares Guia e Saul Gdansky Jaccquieri. Não há palavras para dizer o quanto vocês influenciaram minha formação como cientísta.

Por fim, à CEMIG (Centrais Elétricas de Minas Gerais) e à COPASA.(Companhia de Saneamento de Minas Gerais). Sem elas, o vidro do *box* do meu banheiro não ficaria embaçado e eu não teria resolvido muitos dos inúmeros desafios que esta tese me impôs.

Sonho , 02 de fevereiro de 2008

Resumo

Neste trabalho foi feita uma análise comparativa entre duas metodologias clássicas no estudo de contatos em proteínas: a dependente de um delimitador de distância (CD - *Cutoff Dependent*) e outra que não é dependente de um delimitador, a decomposição de Delaunay (DT – Delaunay *Tessellation*). Essas técnicas foram avaliadas usando-se duas formas diferentes de representação de resíduos (centróides): pelo carbono alfa (CA) e pelo centro geométrico da cadeia lateral (GC). Um banco de dados foi montado, compreendendo dois conjuntos chamados ALPHA e BETA contendo cadeias das duas principais classes do sistema de classificação CATH: *all-alpha* e *all beta*, respectivamente. Um delimitador em 7.0 Å emergiu como um importante parâmetro de distância na análise dos contatos inter-resíduos em proteínas. Este valor marca o ponto de bifurcação no comportamento das curvas de contatos entre as técnicas CD e DT. Até 7,0 Å, as propriedades CD e DT são unificadas numa mais abrangente: nesta distância, todos os contatos (arestas) são totais e verdadeiro-positivos (completos e não-oclusos). A distância de 7,0 Å é o ponto também em que a primeira camada de vizinhos encontra-se otimamente separada das demais, constituindo-se principalmente de contatos de primeira-ordem. É demonstrado que 7,0 Å é um ponto de transição entre os comportamentos lineares e quadráticos da curva do número total de vizinhos por resíduo. Também é mostrado que a técnica DT tem uma conhecida anomalia em sua contagem de arestas que, em proteínas, pode produzir omissões indesejáveis e sistemáticas afetando principalmente a rede de contatos de proteínas betas com centróides em CA. Uma técnica auxiliar reconhecida por tratar essa anomalia é o quase-Delaunay (AD – *Almost* Delaunay). É observado que mesmo AD não se mostra uma técnica proveitosa em proteínas. É empiricamente demonstrado que DT+AD convergem para CD, na medida que o parâmetro de perturbação em AD cresce. Isto alerta que DT e técnicas correlatas devem ser usadas com precaução em proteínas. Como conseqüência, no estrito intervalo de 0,0 Å a 7,0 Å, CD revela-se uma metodologia mais simples, completa e confiável. Por fim, é evidenciado também que a redução na representação dos resíduos aos centróides CA e GC pode introduzir tendências estatísticas na análise de vizinhos em delimitadores até 6,8 Å, com CA em favor ALPHA e GC em favor de BETA. Para valores acima de 6,8 Å, este viés parece ser eliminado. Isto provê um argumento a mais em benefício do limite em 7,0 Å, como um parâmetro de referência, robusto e de carácter geral, a ser usado de forma segura como um confiável delimitador de distância nos estudos em massa de contatos de proteínas.

Abstract

In this study we carried out a comparative analysis between two classical methodologies used to prospect residue contacts in proteins: the traditional cutoff dependent (CD) approach and the cutoff free Delaunay tessellation (DT). Additionally, two alternative coarse-grained forms to represent protein residues were tested: using alpha carbon (CA) and using side chain geometric center (GC). A database was built, comprising two top classes according to CATH classification: all alpha and all beta. We found that the cutoff value at about 7.0 Å emerges as an important distance parameter in analysis of contacts in proteins. This value was not only independent of residue representation and of protein class but it was also the point where CD and DT methods diverged regarding their results. Up to 7.0 Å, CD and DT properties are unified, which implies that at this distance all identified contacts (edges) are fully true-positives (complete and not occluded). This unification may also imply that the edges distribution up to 7.0 Å is constituted mainly by contacts involving buried sites of the first coordination shell. We also have shown that DT techniques have a known anomaly, comprehending points near the degenerate condition, which in proteins may produce dangerous and systematic errors affecting mainly the contact network in beta chains with CA residue representation. The almost-Delaunay (AD) approach has been proposed to solve this DT anomaly. We found that even AD may not be an advantageous solution. We empirically demonstrated that the DT+AD results converge to CD, as the AD threshold perturbation parameter grows. This warns that DT and correlated techniques should be used with care in contacts analysis of proteins. As a consequence, in the strict range up to 7.0 Å, the CD approach revealed to be a simpler, more complete and reliable technique than DT (or DT +AD) to prospect protein contacts. Finally, we have shown that coarse-grained residue representation may introduce bias in the analysis of neighbors in cutoffs up to 6.8 Å, with CA in favor of all alpha proteins and GC in favor of all beta proteins. Beyond 6.8 Å, this bias is apparently eliminated. This provides an additional argument in benefice of the value 7.0 Å as an important lower bound cutoff to be used in contact analysis of proteins, for both CA and GC coarse-grained models.

Sumário

Lista de Figuras

Lista de Tabelas

***Abreviaturas**:*

CD: método tradicional de aferir contatos dependente de um delimitador (***C**utoff **D**ependent*); **DT**: método de prospeção de contatos independente de delimitador por tesselação de Delaunay (***D**elaunay **T**essellation*); **VD**: diagramas de Voronoi (*Voronoi **D**iagrams*); **AD**: método auxiliar de prospeção de contatos quase-Delaunay (***A**lmost **D**elaunay*); **CA**: **C**arbono **A**lfa; **GC**: centro geométrico (***G**eometric **C**enter*); **ALPHA**: subconjunto de cadeias toda alfa segundo CATH; **BETA**: subconjunto de cadeias toda beta segundo CATH; **BIC**: critério Bayesiano de Informação (***B**ayesian **I**nformation **C**riterion*).

1. Introdução

É possível rastrear os primeiros estudos com proteínas aos trabalhos pioneiros de composição química dos seres vivos conduzidos pelo químico holandês Gerhardus Johannes Mulder (1802-1880) e o famoso químico sueco Jöns Jakob Berzelius (1779–1848), no início do século XIX [1]. Mulder, em artigo publicado em 1839 no Journal für Praktische Chemie[2], escreve, conforme versão em inglês extraído de [3]:

> "*I have been occupied for some time with the study of the most essential substances of the animal kingdom, the fibrin, the albumin and the gelatin. Since the publication of this work I continued to study these substances. Berzelius communicated with me concerning the published results and gave me good advice for which I express my sincere thanks*".

A comunicação na qual Mulder referia-se é provavelmente uma carta enviada a ele por Berzelius, no ano anterior a essa publicação, datada de 10 de Julho de 1838, onde encontramos essa passagem histórica:

> "*Le nom protéine que je vous propose pour l'oxyde organique de la fibrine et de l'albumine, je voulais le dériver de πρωτειοσ (**protêios**)*[1] *parce qu'il paraît être la substance primitive ou principale de la nutrition animale.*"[5]

Mulder e Berzelius estavam corretos. Sabemos hoje que as proteínas são uma das mais importantes moléculas dos seres vivos. Elas estão envolvidas em uma ampla gama de processos bioquímicos: nos componentes estruturais; nas reações enzimáticas; na contração muscular, movimento ciliar, flagelar, deformação e divisão celular; na regulação gênica; nas respostas imunológicas; na auto-reconstituição e reparação de tecidos; no controle hormonal; no transporte de substâncias pelos fluidos corporais e transporte celular intermembrana; na constituição da membrana; no impulso nervoso; na reserva e armazenagem de nutrientes; e em outras funções não inclusas ou não delimitadas pelas categorizações acima, como nos venenos e toxinas, agentes antimicrobianos, substâncias anticongelantes etc. Não é surpresa que elas sejam a segunda substância mais encontrada nos seres vivos: sabemos que da massa

1 - O grifo e o termo entre parênteses são meus. Próton, protozoário e protótipo são exemplos de palavras que compartilham com proteína esse mesmo étimo com sentido de antecedência, o primeiro, o primordial [4]

de uma célula, 15% é proteína; a água é a primeira com 70%[6].

Embora toda essa intricada rede de funções estivesse sendo paulatinamente destrinchada desde Mulder e Berzelius, na virada do século XX pouco se sabia sobre a natureza química das proteínas, a despeito dos avanços feitos na identificação de seus blocos construtores, os aminoácidos, e na caracterização da ligação peptídica em 1902 por Emil Ficher (1852-1919) e Franz Hofmeister (1850-1922)[8]. Curioso notar que o primeiro aminoácido proteogênico (integrado ao código genético padrão) foi isolado antes de Berzelius pensar em "protêios": a Asparagina foi descoberta por Pierre Jean Robiquet (1780 - 1840) em 1806[9]. O último dos 20 aminoácidos codificados na síntese biológica[2], a Treonina, só foi descrito em 1938, por William Cumming Rose (1887 – 1985)[10].

Dentre muitas de suas intrigantes propriedades como longas cadeias poliméricas de aminoácidos, duas em especial pareciam contraditórias para a época: proteínas podiam formar cristais; e proteínas podiam desnaturar, perder sua função sob ação de certos agentes físicos ou químicos tais como calor, pressão, agitação mecânica, luz ultravioleta, pH e osmólitos[11]. Há registros de quem tenha conseguido desnaturar proteínas até por ondas supersônicas[12].

A capacidade das proteínas cristalizarem-se, formando diferentes cristais conforme a molécula, indicava que sua função dependia de uma organização espacial estrita[6]. De maneira geral, a formação de um cristal exige um arranjo coerente e repetitivo de unidades estruturalmente semelhantes. No final da década 1940, Linus Pauling e Robert Corey, com base em dados de cristalografia por raios X, anteveriam o papel das pontes de hidrogênio na formação de estruturas secundárias como as alfas hélices[14] e folhas betas[15]. Mais 10 anos seriam necessários para que Max Perutz (1914-2002) e John C. Kendrew (1917-1997) dominassem a técnica de difração de raios X para proteínas inteiras e confirmassem as previsões de Pauling e Corey, através da primeira resolução estrutural completa da hemoglobina humana e a mioglobina[16] de cachalote[17], respectivamente[3].

Se por um lado a formação de cristais direcionava o pensamento em proteínas para um mundo estruturado, o estranho fenômeno da desnaturação induzia ao caminho reverso. Alfred Mirsky (1900-1974) e Mortimer Louis Anson (1901–1968) foram os primeiros a levantar de forma convincente evidências contrárias ao consenso da época sobre desnaturação em proteínas, que postulava que este processo era um fenômeno irreversível. Seus históricos

2 - Além dos 20 aminoácidos clássicos tende-se a aceitar hoje também como o 21º e 22º, a selenocisteina e a pirrolisina, codificadas por algumas bactérias[7]

3 - Nas décadas por vir, outras técnicas, como o NMR[13] detalhariam (sem necessidade de cristais) não somente a estrutura, mas também o comportamento de certas proteínas em solução.

artigos[18][19] entre 1925 e 1930 demonstravam perturbadoramente que era possível recuperar hemoglobinas desnaturadas ou coaguladas (uma forma aglutinada geralmente precipitada da proteína desnaturada), preservando-lhes as mesmas propriedades de uma hemoglobina intacta[18]. Apesar de Mirsky e Anson levantarem algumas hipóteses sobre essa reversibilidade, eles não chegaram a dar-lhes o peso retórico de uma teoria.

A historiografia em proteínas irá creditar então ao chinês Hsien Wu (1893-1959)[12] em 1931, a primeira tentativa de uma teoria lúcida da desnaturação[20]. Assumir proteínas como longos polímeros de aminoácidos levava à constatação (um tanto evidente, agora) de que em proteínas globulares a cadeia deveria estar dobrada ou enovelada em si mesma. Wu sugeriu que a desnaturação poderia ser vista como um processo de desorganização estrutural em decorrência do desempacotamento da cadeia. Ele ponderava que os agentes desnaturantes desmantelariam uma proteína pelo enfraquecimento das interações polares e/ou Coulombicas[4]. Na ausência do desnaturante, essas forças reconduziriam a proteína ao seu estado nativo-funcional.

Juntamente com Alfred Mirsky (1900-1974), Linus Pauling, em 1936, proporia uma explicação semelhante a de Wu, mas conferindo às pontes de hidrogênio intracadeias o mote principal por trás desse processo [21]. O seu posterior sucesso da previsão das alfa-hélices, matematicamente confirmadas pelas coordenadas atômicas das globinas de Perutz e Kendrew na década de 1960, dogmatizaria o papel das pontes de hidrogênio não só em proteínas, mas em toda bioquímica [20].

Na década de 1950, Walter Kauzmann (?-) começaria a chamar atenção para os fatores entrópicos (em especial da água) na estabilidade da cadeia, criando um dos mais importantes conceitos da físico-química de proteínas: a interação hidrofóbica [22]. Mas, o pensamento bioquímico corrente encontrava-se rigidamente enjaulado pelas pontes de hidrogênio, e a hidrofobicidade de Kauzmann foi recebida com previsível negligência [20].

Ao longo das décadas de 1950 e 1960 o problema da desnaturação é vetorialmente invertido, em 180º, e vai se tornando o problema do enovelamento, como relembra Irving Klotz [23]. Certamente os avanços da nascente biologia molecular na caracterização do seu dogma central (DNA -> RNA -> PROTEÍNA) contribuíram para essa reversão [24] principalmente o processo de elucidação da síntese protéica *in-vivo* iniciado por Paul Zamecnik (1913-), em 1950 [6]. Afinal, se uma proteína era linearmente montada

4 - Na versão em inglês[12] no artigo de Wu essas interações são denominadas de "*secondary valence bonds*", onde são destacados apenas os grupos aminos e carboxílicos.

aminoácido a aminoácido a partir da informação codificada em ácidos nucléicos, a última etapa da decodificação genética teria de ser seu correto enovelamento numa molécula funcional.

É no contexto dessa transição que Christian B. Anfinsen (1916-1995), renaturando ribonucleases A (entre 1955-1962), reuniria uma massa hipercrítica de evidências para conceber o que muitas vezes é uma das coisas mais difíceis em ciência: o óbvio. O óbvio de que a reversibilidade da desnaturação *in-vitro*[5] implicava que toda informação que uma proteína globular de baixo peso precisava para se reenovelar estaria codificada na seqüência de seus resíduos[26]. Anfinsen unificaria teorias como as de Wu, Mirsky-Pauling e Kauzmann numa abrangente hipótese termodinâmica do enovelamento, ao considerar todos os tipos de interações como parte do processo. Conforme suas palavras no Nobel Lecture, em 1972 [27]:

> "*This hypothesis states that the three-dimensional structure of a native protein in its normal physiological milieu (...) is the one in which the Gibbs free energy of the whole system is lowest; that is, that the native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence, in a given environment*."

Mas essa hipótese criava um sério problema de otimização para a cinética do enovelamento. Sendo o estado desnaturado uma miríade de cadeias abertas em indefinidas conformações, como uma proteína consegue orientar seu próprio processo de enovelamento de modo que todas essas cadeias convirjam para um conjunto homogêneo nativamente estruturado?

Verbi gratia, para uma pequena proteína constituída por 100 aminoácidos, se considerarmos que cada resíduo adota uma de duas conformações de maior estabilidade, teremos um espaço conformacional representado por 10^{30} estruturas [29]. Cyrus Levinthal (1922-1990), em 1969, ponderava que o processo de renaturação não poderia envolver uma busca por força bruta das conformações com menor energia livre nesse imenso espaço de estruturas [30]. No nosso exemplo, se assumirmos que a mudança de uma conformação para outra demande até 10 picosegundos (10^{-11} segundos), aquela pequena proteína de 100 resíduos poderia levar da ordem de 100 bilhões[6] de anos para enovelar-se, a despeito de dados experimentais indicarem que isso em geral ocorre na faixa de milisegundos a segundos [31].

---

5 - Hoje sabemos que proteínas grandes e complexas podem necessitar de outras proteínas auxiliares para enovelar [25]

6 - O que é muito mais que os 14 bilhões de anos aceitos para a idade do Universo [28]. Mesmo assumindo paralelismo na mudança de conformação, o tempo ainda fica astronômico: da ordem de 1 bilhão de anos.

Essa alegoria com números, que acabou sendo interpretada como um paradoxo, foi uma forma criativa que Levinthal encontrou para semear no imaginário acadêmico a sua notável hipótese cinética do enovelamento [30]: que na seqüência dos resíduos de uma proteína deveriam estar codificados também os caminhos e as etapas que conduziriam à estrutura nativa, e não somente a informação para produzir (a qualquer custo) as estruturas mais termodinamicamente estáveis, como havia colocado Anfinsen [32].

E mais: o próprio Levinthal reconhecia que sua hipótese não cercava a possibilidade de que nem sempre o caminho mais eficiente (a melhor cinética) conduziria às melhores estruturas[7] (com menor energia livre) [33]. Logo, o relojoeiro cego por trás da evolução estaria encriptando na seqüência dos resíduos de uma proteína ambos aspectos do seu enovelamento: a cinética e a termodinâmica. Quebrar esse código, conforme já dito, significaria a elucidação do próximo passo na tradução da informação gênica. Implicaria em compor uma teoria que explicasse o que se convencionou chamar de o problema do enovelamento de proteínas ou PFP (*Protein Folding Problem*)[35]. Mas, por onde começar?

"*So there must be folding intermediates*", foi o que Robert L. Baldwin (1927-) diz ter ouvido do grande químico de polímeros Paul Flory (1910-1985) num dos antológicos seminários de Levinthal (*How to Fold Gracefully*), em Stanford, em 1968 [36]. A proposta de Flory era tentadora: se aquela mesma proteína hipotética de 100 resíduos que mencionamos tiver 4 intermediários obrigatórios no seu processo de enovelamento, então o espaço de busca reduziria de 10^{30} para 4 x (2^{25}) ou 10^{8} conformações; e a estrutura nativa poderia ser encontrada em milissegundos[38] . Era um bom começo. Estava deflagrada a corrida em busca de intermediários no processo de enovelamento[8].

De 1970 em diante, a comunidade envolvida no PFP cresce rapidamente tal qual os seqüenciamentos, a resolução de estruturas, as técnicas de controle e manipulação gênica, os aparatos físico-químicos de monitoramento estrutural, e o poder de cálculo e armazenamento computacionais. Crescimento que foi munindo os pesquisadores com um poder cada vez maior de fogo[41]. permitindo: análises comparativas de seqüências e estruturas[42]; re-engenharia de proteínas[43], seja pela síntese química[44] ou expressão de mutantes[45]; identificação de intermediários[46], estados de transição[47] e rotas de enovelamento[48];

7 - No jargão computacional de hoje podemos classificar a hipótese de Levinthal como uma estratégia de busca gulosa. De fato, já está demonstrado que o problema de encontrar a menor energia livre em modelos discretos de proteínas é NP-Hard [34]

8 - Mas, infelizmente, nem sempre os intermediários foram encontrados e surgiram dúvidas se seriam mesmo produtivos ou meras armadilhas cinéticas (desvios de rotas)[29]. Um dos principais críticos a essa hipótese é Thomas Creighton, que não acredita nem que eles possam acelerar o processo de enovelamento. Muito pelo contrário, julga demonstrar que a renaturação é mais lenta em proteínas com cinética de mais de dois-estados.[37]

simulações de dinâmica molecular[49]. Tudo conectado e confrontado por modelos teóricos cada vez mais arrojados[9][50-55]. Os avanços são muitos, mas parece que ainda não há uma teoria consenso[10] do enovelamento. Como a seqüência codifica a estrutura de uma proteína persevera como uma das maiores questões em aberto da ciência moderna, e o problema central da biologia estrutural.

Enquanto o PFP atrapalhava o sono de teóricos e experimentalistas engajados no desafio de elucidá-lo, outros obstinados pesquisadores seguiram adiante em outra frente, na tarefa de catalogar os seres vivos em seu nível molecular. Esse paulatino ganho em escala exigia a contrapartida da criação de repositórios referenciais para o armazenamento da crescente onda de dados que inundavam os periódicos. As décadas de 70 e 80 marcariam o início das operações dos que viriam a ser os grandes bancos de dados moleculares da atualidade. O primeiro deles, armazenando as coordenadas atômicas e informações correlatas, é o famoso PDB – *Protein Data Bank*, que entrou em produção para acesso internacional em 1977 [59], seguidos pelos bancos de seqüências nucleotídicas mantidos por europeus (EMBL-DL – *European Molecular Biology Laboratory - Data Library*)[60] e por americanos (Gen Bank do NCBI - *National Center for Biotechnology Information*)[61], ambos com raízes em 1980. Logo após vieram PIR (*Protein Information Resource*)[62] em 1984 e SwissProt[57] em 1986, concentrando dados bem anotados de seqüências protéicas. A partir daí, muitos outros foram criados[11], incluindo versões colaborativas ou unificadas deles, como INSDC (*International Nucleotide Sequence Databases Collaboration*)[60], Uniprot[63] e wwPDB[64].

Mas uma grande mudança de paradigma coroaria o apagar das luzes do milênio passado: a humanidade entraria na era "ômica", inaugurada em 1995 pelo primeiro seqüenciamento completo do DNA de um ser vivo de replicação autônoma[12], a bactéria *Haemophilus influenzae*[66]. Vieram no rastro: primeiro eucarioto em 1996, com a levedura *Saccharomyces cerevisiae*[67]; a bactéria *Escherichia coli*[68], em 1997; o primeiro ser vivo pluricelular em 1998, com o verme nematódeo *Caenorhabditis elegans*[69]; a *Drosophila melanogaster*[70], como primeiro inseto em 2000; a *Arabidopsis thaliana*[71], como primeira

9 - Destes modelos, dois destacam-se e confrontam-se: o do funil, proposto por Ken Dill[29]; e o enovelamento *in concert* defendido por Thomas Creighton[35]. No primeiro, um colapso hidrofóbico restringe o espaço conformacional, produzindo estruturas cada vez mais compactas e muitas rotas possíveis para se alcançar o estado nativo. No segundo, as conformações não-enoveladas estão em equilíbrio, e somente uma rota de máxima cooperatividade conduz ao estado funcional naturado.

10 - É crescente o sucesso de técnicas que se valem em maior ou menor grau do empirismo na predição de estruturas a partir de seqüências, como atestam os desafios lançados pelas reuniões bianuais do CASP (*Critical Assessment of techniques for protein Structure Prediction*) [39]. Porém, prever é uma condição necessária mas não suficiente para uma teoria científica[40].

11 - Como DDBJ[56], TrEMBL[57], AceDB[58] e outros.

12 - O sequenciamento de ácidos nucléicos virais já tinha sido inaugurado desde 1976[65].

planta também em 2000; o *Homo sapiens*[72], em 2001. No instante que essa tese é escrita, o site GOLD[13][73] registra perto de 700 seqüenciamentos concluídos, de um total de 3520 projetos em andamento.

Se antes o foco estava no estudo mais isolado de determinados genes e proteínas, agora queria-se não só o mapeamento de todos eles, mas também suas evoluções no tempo e correlações com as diversas instâncias do organismo. Abandonava-se uma visão compartimentalizada e hierárquica dos processos biológicos, para uma concepção holística e conexionista, representada por complexas redes de interações e relações. Tão forte é o paradigma que ele exigiu (com também o fizera Mulder e Berzelius) um novo vocabulário para propagar essa idéia de "coleção do tudo", estilizado pelos sufixos "ômica" para a designação da nova ciência e "oma" para o seu objeto em estudo[74]. Assim, nascia a genômica e seu genoma, a proteômica e seu proteoma, a transcriptômica e seu transcriptoma, a metabolômica e seu metaboloma, e tantas outras "ômicas" e "ômas"[14].

A virada do milênio foi também o momento em que a Internet capilarizaria o planeta, tornando a WEB um fenômeno ubíquo. Agora, diferentes bancos de dados podem ser sincronizados e integrados, e os lançamentos de novas entradas e consultas aos registros, feitos remotamente e instantaneamente a partir de qualquer computador conectado. Logo, a informação biológica crescia explosivamente não só em quantidade, mas também em disponibilidade, retroalimentando-se numa espiral virtuosa. Um simples teste pode demonstrar o quão impressionante é essa alegação. Uma consulta ao Google Scholar[15] retorna para o verbete "*proteins*" entre 2000 e 2008 cerca de 411000 registros de artigos. De 1900 a 1999, a mesma pesquisa retorna em torno de 302000 ocorrências. Só nos primeiros 8 anos do novo milênio há mais quantidade e disponibilidade de artigos do que em todo o último século do milênio passado! Como bem observou Frederic Richards[75]:

> "*During the past century the amount of detailed information about proteins has been increasing and today can only be described as a torrent. Our ability to assimilate this vast amount of data has not kept pace. Each group of closely related proteins is represented not only by a large number of individual papers but by its own set of specialist journals*."

13 - http://www.genomesonline.org/index.htm

14 - Nesse momento estão registrados em sites como http://omics.org cerca de 195 diferentes *ômicas.

15 - http://scholar.google.com

É nesse efervescente momento que emerge a Bioinformática como uma ferramenta auxiliar ao limitado cérebro humano, essencial na estruturação e manipulação dessa gigantesca profusão de dados. Como não podia deixar de ser, é neste contexto também que nasce essa tese. Seu objetivo inicial era estudar a relação seqüência e estrutura em globinas, buscando quem sabe alguma contribuição ao PFP. Mas durante o seu desenvolvimento ela acabou por esbarrar num problema metodológico envolvendo a forma como contatos inter-resíduos são calculados, que pareceu-nos ser mais promissor que a proposta original. Houve então um desvio de rota[16], que culminou no artigo submetido para a *Proteins: Structure, Function and Bioinformatics*. Nós julgamos que essa questão metodológica que se apresentava mexia com conceitos fundamentais da análise em massa de proteínas (um dos entraves da era "ômica"), e um estudo criterioso poderia repercutir nas diversas aplicações que utilizassem o padrão de contatos como base de seus algoritmos, inclusive aquelas que tentassem mapear a relação seqüência e estrutura em proteínas.

Esta introdução fez um rápido e panorâmico vôo sobre a história das proteínas. Agora em que os fatos estão mais contextualizados, é hora de vermos o que essa tese tem a contribuir para a evolução desse novo paradigma.

16 - Que espero não ter sido uma armadilha cinética.

2. Contatos em Proteínas

Contatos inter-resíduos e/ou interatômicos têm sido usados em uma ampla gama de estudos envolvendo proteínas. A sua precisa e correta determinação é de fundamental importância em muitos dos algoritmos de análise e comparação estrutural, tais como: densidade de empacotamento[76-79], similaridade funcional[80], relações evolucionárias[42], classificações topológicas[81-82], alinhamentos estruturais[83], avaliação estrutural[84], predição de estrutura terciária[85-86], análise de redes de contatos[87-89], potenciais empíricos,[90-92] previsão de estabilidade termodinâmica[75], inferências sobre o processo de enovelamento[93-94], interações proteína-proteína e proteína-ligante[95] etc.

Tão diverso quanto suas aplicações é a forma como os contatos vêm sendo definidos. O método clássico e mais simples é através do estabelecimento de delimitadores de distâncias. Dado dois pontos ou sites $\{i,j\}$ de um conjunto de átomos ou resíduos, i estará em contato com j se o último estiver dentro de uma esfera de raio r centrada no primeiro. O raio r é o delimitador de distância (ou *cutoff)*. Veja Figura 1a. Aqui já aparece um primeiro problema: que delimitador usar? A literatura é pródiga em oferecer uma ampla gama de opções. Em nível atômico: 3,8[96], 4,5[97], 5,0[98], 5,5[99], 6,0[94]; em nível de resíduos: 6,5[100], 7,0[88], 8,0[101], 9,0[80] Å. A despeito de importantes tentativas de racionalização na seleção do delimitador[100-104], na maioria das vezes percebe-se que o valor escolhido ou foi arbitrário ou atendeu a otimizações específicas de cada caso.

Há muitos outros tipos de definições de contatos correlatos ao método clássico descrito acima. Um deles pode ajustar o delimitador ao tamanho dos sites (geralmente compreendendo o raio de van der Waals dos átomos), somando aos raios de i e j uma distância fixa r [105]. Desde que r seja devidamente ajustado, essa forma tenta garantir que somente a primeira-ordem de contatos (a primeira camada não-oclusa de vizinhos) seja levada em conta. Novamente, o problema é a escolha de um valor adequado para r, que tem sido situado pela literatura entre 0,6 e 3,0 Å[42][105-88]. Os contatos também podem ter pesos atribuídos por uma função, representando uma área de contato[109], uma energia potencial[87], uma distância Euclideana[82] ou mesmo algum tipo de normalização[97], como aqueles feitos utilizando funções de distribuição radial (RDF - *Radial Distribution Function*)[110][91]. Adicionalmente, os contatos podem não ser visto apenas como uma

coleção de pares de sites. É possível estendê-los para outras dimensões formando *n-eplas*, sendo comum contatos de 3[111] e 4[112] eplas. Há ainda outras formas mais complexas de atribuir contatos, como os contatos tipo OSP (*Occluded Surface Packing*)[77], o método SPCD[84] (*Small-Probe Contact Dot*), e os contatos de ordem relativa ou RCO (*Relative Contact Order*)[94]. Mas, afora alguns detalhes divergentes, todas essa técnicas guardam em comum o uso explícito ou implícito (em geral através do raio da sonda) de um delimitador de distâncias.

Ademais, a transição no extremo do delimitador pode ser feita de uma forma discreta ou contínua. No tipo discreto, os contatos são aferidos na forma do tudo-ou-nada: ou ele está dentro da região delimitadora ou não está. Para esses casos, geralmente são utilizadas funções degraus tipo Heaviside[100]. O problema é que essa maneira discreta é suscetível a pequenas mudanças nas coordenadas dos pontos na região fronteiriça do delimitador. Uma alternativa é suavizar essa região de transição usando uma função sigmoidal[49]. Maiorov e Grippen[113] demonstraram que apesar dessa suavização ser útil na análise de estruturas homólogas, quando ela é aplicada a qualquer conjunto de proteínas, o modelo discreto se correlaciona linearmente muito bem com o modelo contínuo. (coeficiente de correlação de 0,997).

Uma outra forma de assinalar contatos em proteínas é através de diagramas de Voronoi[114] e decomposição de Delaunay[115], também conhecidos como tesselação de Voronoi e Delaunay. Seu uso em proteínas remonta ao pioneirismo de F. Richards em 1974 e J. Finney em 1975, nos cálculos de volumes e densidades de empacotamento, e vem crescendo ao longo dos últimos anos em inúmeras outras aplicações[116]. Nós podemos definir tesselação como uma forma de cobertura de um espaço *d* dimensional. Para um espaço Euclideano R^d, implica na possibilidade de usar uma coleção de politopos (uma generalização do conceito mais familiar "polígono" para qualquer dimensão) criados a partir dos sites para preencher de forma justa uma região d-dimensional sem sobreposições, falhas ou buracos. Essa cobertura estrita é capaz de capturar relações especiais entre o conjunto de pontos distribuídos no espaço. As tesselações de Voronoi e Delaunay são um caso especial desse tipo de cobertura que, através de regras geométricas exatas, produz um padrão de conectividade envolvendo sempre os vizinhos mais próximos de cada site (Figura 1b). Em proteínas, a tesselação de Delaunay irá resultar numa completa decomposição do volume ocupado por elas em tetraedros justapostos de tal forma que suas arestas representarão os contatos e os vértices (ou nós) os átomos ou resíduos (dependendo do nível escolhido).

É comum também classificar os métodos de geração de contatos em dependentes e independentes de um delimitador de distâncias. No primeiro caso, um parâmetro delimitador é um pré-requisito essencial na definição do contato. Um exemplo típico desse caso é o método clássico de aferir contatos descrito acima. Delimitadores também podem ser usados no cálculo da energia de contatos em certas simulações de mecânica e dinâmica molecular, especialmente no truncamento das forças de longa distâncias[117]. No segundo caso, inversamente, um parâmetro delimitador não é necessário para a definição do contato. A decomposição de Delaunay, enquanto uma abstração matemática[17], é um representante dessa categoria, já que a composição de contatos a partir dele é feita segundo critérios geométricos sem nenhuma necessidade de um delimitador.

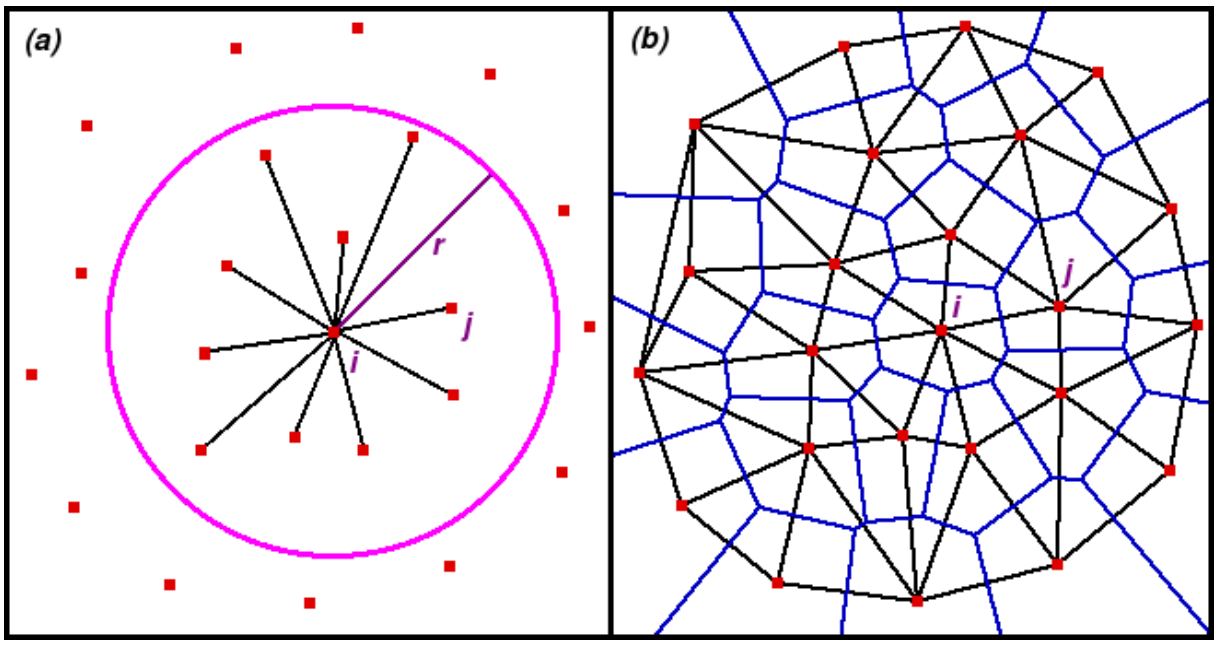


Figura 1: Exemplo de prospecção de contatos por métodos delimitador dependente e independente. (**a**) Método tradicional dependente de um delimitador de distância. Sites *i* e *j* (quadrados vermelhos) estão em contato porque *j* está dentro da esfera com centro em *i* e de raio *r*. (**b**) Tesselações de Voronoi (arestas em azul) e Delaunay (arestas em preto) como exemplo de um método delimitador independente. Para mais detalhes sobre ambos os métodos veja seção Materiais e Métodos. Figuras adaptadas deVoroGlide[119].

Outro exemplo de contatos gerados sem uso de um delimitador é aquele em que o limite de distância é levado ao infinito. Nesse caso, uma função de energia inversa à distância pode ser usada como peso das arestas para desvalorizar contatos entre sites separados por longas distâncias. O método de oclusão de C. Veloso *et al* é um exemplo dessa estratégia[87].

17 - Na prática, há quem [112] restrinja o tamanho da aresta Delaunay para evitar a criação de contatos de longa distância sem significado físico, principalmente envolvendo sites em superfícies de proteínas do PDB não artificialmente solvatadas.

Por ele, dois sites estarão em contato se não há nenhum outro interveniente entre eles, isto é, se não há um terceiro que os esteja ocluindo. No método Veloso, uma função de energia que decai com o inverso da distância é usada para depreciar contatos longos e não oclusos. Em simulações de mecânica e dinâmica molecular, as técnicas Ewald usadas principalmente na medição da energia de contatos eletrostáticos também podem ser classificadas como delimitadoras independentes[118].

Os contatos podem ainda ser classificados quanto à granulosidade dos pontos[120]. Nos modelos de granulação fina, os sites são concebidos em nível atômico, produzindo uma representação detalhada (porém mais complexa) da proteína. Essa granulosidade também pode ser usada para mapear contatos entre resíduos. O que geralmente se faz é assumir que dois resíduos estarão em contato se qualquer de seus átomos pesados[18] estiverem próximos o suficiente[98]. Em outras variações desse modelo, a escolha de quais átomos pesados considerar é mais restrita[121]. Outras ainda aplicam certas estatísticas sobre o conjunto de átomos dos resíduos em contato para atribuir-lhe um peso[97]. Modelos de granulação grossa, por outro lado, tentam reduzir a complexidade do sistema usando uma resolução baixa na representação dos sites[120]. Em geral, essa simplificação é feita reduzindo o resíduo a um ponto representativo chamado de centróide, a partir dos quais os cálculos de contatos são feitos. Usualmente são escolhidos como centróides: carbonos alfas (CA)[101][103], carbonos betas (CB)[122], centro geométrico (GC)[100] ou baricentro (BC)[123] da cadeia lateral (que também podem incluir ou não alguns átomos da cadeia principal).

Apesar de toda a diversidade de definições e classificações de contatos descritas acima, nós podemos inferir um ponto em comum para a maioria deles: mapear a presença ou localização dos sites num dado espaço objetivando extrair ou explorar preferências subjacentes em sua distribuição espacial. Mas, estariam os métodos relatados aqui enxergando esse objetivo comum da mesma maneira? Qual seria a interferência nas estatísticas de contatos se usássemos uma ou outra destas metodologias? E sobre os centróides, quais seriam as conseqüências de escolhermos um ou outro tipo de representação de resíduos? Seria possível estimar um delimitador ótimo ou deveríamos usar uma forma de gerar contatos que fosse independente do delimitador, como as tesselações de Voronoi e Delaunay? Essas são algumas das questões que esta tese pretende escrutinar.

Como essas metodologias compõem a base algorítmica da maior parte das aplicações

18 - Qualquer átomo senão hidrogênio (e seus isótopos).

usadas hoje em dia em bioinformática estrutural, passa a ser de fundamental importância conhecer a fundo suas idiossincrasias, seus limites, seus pontos de divergências e em que condições elas podem viciar ou tender os resultados. Por questões práticas nós focamos nossa atenção em alguns aspectos representativos das definições e classificações de contatos explicitadas aqui. Nós examinamos as relações entre o método clássico de aferir contatos definido por um delimitador (ou *cutoff dependent* – CD) e o delimitador independente Delaunay *tessellation* (DT), ambos no nível dos resíduos, utilizando como centróides o carbono alfa (CA), e o centro geométrico da cadeia lateral (GC). Essa matriz de 4 variáveis foi aplicada a dois conjuntos de dados, cada um com 91 cadeias de proteínas globulares não relacionadas em suas estruturas primárias: uma toda alfa (*all alpha*) e outra toda beta (*all beta*), conforme classificação do CATH[124]. Como um estudo de caso nós analisamos como essas metodologias reconhecem a vizinhança estruturada dos contatos de primeira-ordem, visando verificar quais suas implicações na avaliação do empacotamento em proteínas. Resultados interessantes emergiram destas comparações, tocando não somente em certas questões dos fundamentos do empacotamento de resíduos, como também da aplicabilidade da decomposição de Delaunay e técnicas correlatas na aferição de contatos em proteínas.

3. Objetivos

3.1 Objetivo Geral

Analisar a aplicação da varredura exaustiva de distâncias inter-resíduos na análise de contatos intracadeia em proteínas globulares ("*PROTEIN CUTOFF SCANNING*").

3.2 Objetivos Específicos

- Examinar comparativamente as relações entre duas técnicas clássicas de aferir contatos em proteínas: o tradicional que depende de um delimitador de distância (CD) e o delimitador independente Delaunay *tessellation* (DT).
- Examinar comparativamente a influência nas estatísticas de contatos para CD e DT quando os resíduos são representados por seus carbonos alfas (CA) e seus centros geométricos das cadeias laterais (GC).
- Examinar comparativamente como a matriz composta por CD, DT x CA, GC comporta-se quando aplicada a dois conjuntos disjuntos de cadeias de proteínas: toda alfa (*all alpha*) e toda beta (*all beta*), conforme sistema de classificação do CATH[124].

4. Materiais e Métodos

4.1 Base de dados

Doravante, chamaremos "alfa" e "beta" qualquer elemento de estrutura secundária em hélice[19] ou folha beta, respectivamente, independente de seus subtipos. Logo, estão incluídos em "alfa", por exemplo, os seguintes tipos de hélices: α-hélices, 3_{10}-hélices, π-hélices e a mais rara hélice "da mão esquerda" (*left hand helix*). Através do sistema avançado de busca do PDB[126] e do STING_DB[127], foram construídas duas bases denominadas ALPHA e BETA de igual tamanho, amostradas do sistema de classificação CATH. O conjunto ALPHA contém apenas proteínas da categoria *all-alpha* do CATH, e BETA da categoria *all-beta*. Todas as proteínas passaram pelos seguintes filtros: resolução menor que 2,0 Å, *R-Value Working*[20] menor que 0,2, identidade de seqüência menor que 30% e tamanho da cadeia entre 50 e 600 resíduos. Uma busca inicial feita em Novembro de 2007 retornou com esses critérios 248 proteínas para ALPHA e 314 para BETA. No intuito de reforçar o sinal das estruturas secundárias nos dois conjuntos, nós checamos o perfil destes conteúdos usando o sistema DSSP[21][81]. Para ALPHA, foram aceitas aquelas com mais de 35% de conteúdo "alfa" e menos de 12% de conteúdo "beta". Para BETA, em adição do conteúdo de "alfa" ser menor de 12%, foi imposto que o número relativo de resíduos em "beta" fosse ao menos duas vezes maior que os em "alfa", além de que a quantidade de conteúdo não assinalado fosse menor que 65%. Estes valores foram escolhidos após testes que nos permitiram ajustá-los a um nível que julgamos adequado à nossa análise. A cardinalidade dos dois conjuntos ao fim dessa rodada ficou reduzida a 158 para ALPHA e 148 para BETA.

4.2 Padronização dos arquivos PDB

A fim de uniformizar o conteúdo dos arquivos PDB retornados do estágio anterior de

19- A fita beta, apesar de poder ser classificada como uma hélice de passo 2, será tratada como "beta".
20 - Nós usamos *R-Value Working* ou *R-Value Working Test* porque verificamos empiricamente que esses campos estão mais freqüentemente anotados nos arquivos PDB que o *Free R-Value* .
21 - Como STING_DB usa uma anotação de estrutura secundária mais restringente (coincidência dos sistemas DSSP[81], STRIDE[125] e notação nativa do PDB, tanto em tamanho quanto em conteúdo de elementos de estrutura secundária) nós optamos por usar apenas o DSSP para que fosse possível lidar com um número maior de cadeias na filtragem.

checagem dos padrões DSSP, nós usamos o pacote PDBEST[128], uma ferramenta em desenvolvimento no nosso grupo, composta por uma coleção de scripts PERL que aplicam um conjunto de regras (definidas pelo usuário) sobre os arquivos PDBs originais, e retorna-os padronizados e filtrados (veja Figura 2).

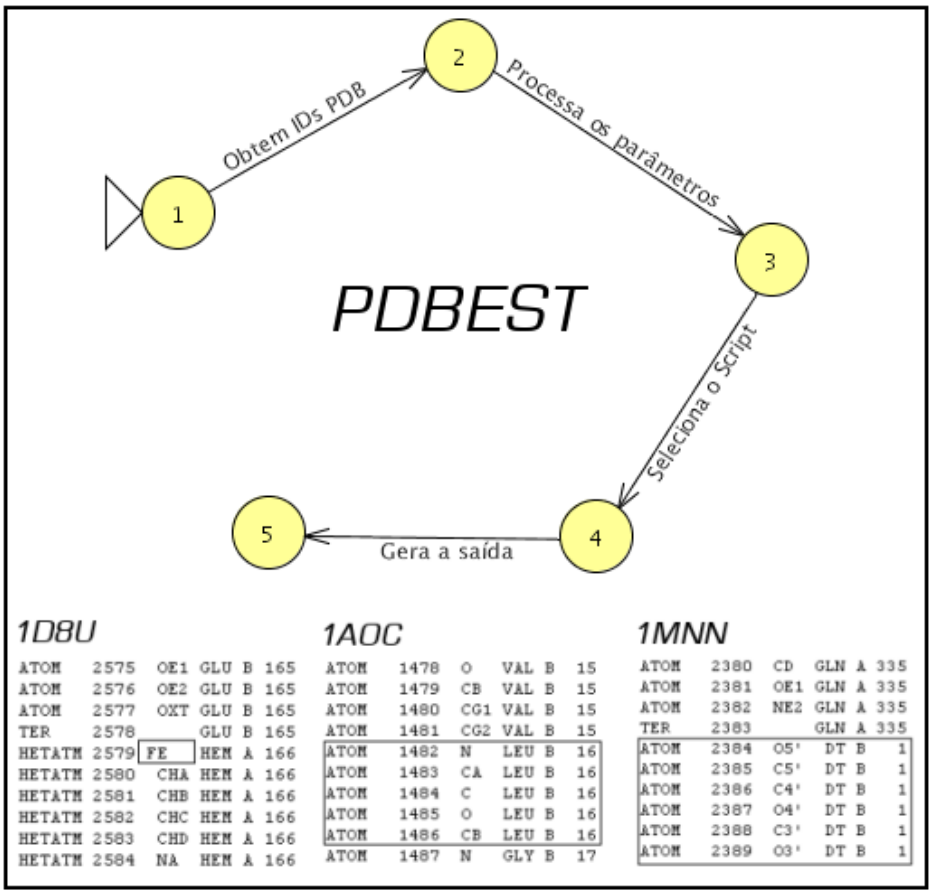


Figura 2: Máquina de estados e exemplos de questões tratadas pelo PDBEST. Na 1D8U[129] o nome do átomo FE está fora da posição esperada. Na 1AOC[130] o resíduo LEU-16 não está completo. Na 1MNN[131] na cadeia B os nomes dos resíduos não são de aminoácidos.

Nossos dados PDBESTs foram compostos após aplicação das seguintes regras (descritas em linguagem de alto nível):

- Re-enumere as cadeias, resíduos e átomos;
- Separe as cadeias em arquivos;
- Exclua cadeias com seqüências idênticas na TAG SEQRES, mantendo apenas a primeira;
- Desconsidere cadeias que tenham resíduos faltando átomos;

- Descarte cadeias cujo nome dos resíduos não seja um dos 20 aminoácidos proteogênicos padrões (exceção para seleno-metionina);
- Se houver átomos com mais de uma ocupação (*occupancy*), escolher a mais provável;
- Se houver modelos, pegar o primeiro;
- Se detectar falhas de notação nos campos cadeia, resíduo e átomo, corrija-as; se a correção não for possível, registre no arquivo de log a falha;

Ao final de todo esse processo, as bases ALPHA e BETA ficaram niveladas em 91 cadeias de proteínas cada (Tabela I). Um sumário estatístico dos dados pode ser visto na Figura 3. Sobre esses dados podemos fazer algumas considerações pertinentes. A primeira é sobre a assimetria entre o conteúdo "alfa" em ALPHA (61,6 ± 11,6 %) e "beta" em BETA (46,6 ± 7,9 %) na Figura 3a. Isso deve ser um resultado natural, dada as características topológicas dessas duas estruturas secundárias. Um segmento em "alfa" irá concentrar mais resíduos que um segmento de igual comprimento (em Å) em "beta". Isso faz com que o cálculo da quantidade relativa de conteúdo "alfa" por segmento fique maior que a quantidade relativa de conteúdo "beta" por segmento. Raciocínio inverso pode ser aplicado ao conteúdo assinalado como "outros" (não "alfa" e não "beta"), que será menor no conjunto ALPHA e maior no conjunto BETA. A segunda é sobre a assimetria positiva (*right skewed*) na distribuição de densidade para o tamanho das cadeias. Também isso é um resultado natural da amostragem, e deve refletir a distribuição da população dada pelo conjunto de cadeias armazenadas no PDB. O importante é que as estatísticas não-paramétricas de homogeneidade das médias, variâncias e de aderência (*goodness-of-fit*) indicam que nossos conjuntos ALPHA e BETA apresentam distribuições bem equivalentes. Isso assegura que do ponto de vista do tamanho da cadeia não há nenhuma tendência favorecendo ALPHA ou BETA.

TABELA I: Arquivos PDBs da Base de Dados

Base de Dados	Proteínas*
ALPHA (91 cadeias)	1LMB3, 1B0N1, 1M451, 1VRK1, 1A7W1, 1ALV1, 1AMZ1, 1BGF1, 1DK81, 1DNU2, 1EYV1, 1FC31, 1FT51, 1G331, 1G4I1, 1GPQ2, 1GV21, 1HBK1, 1HBN2, 1HE11, 1I2T1, 1I8O1, 1J7Y2, 1JFB1, 1K0M1, 1KG21, 1KQF3, 1L9L1, 1LJ81, 1LKP1, 1M1N2, 1M4R1, 1M8Z1, 1M9X2, 1MTY2, 1MTY3, 1MXR1, 1MZ41, 1N1J1, 1N1J2, 1N2A1, 1NOG1, 1O081, 1O831, 1OOH1, 1OR01, 1OW41, 1OWL1, 1PBW1, 1PPR1, 1Q081, 1QGI1, 1QMG1, 1QOY1, 1R8S2, 1RRM1, 1SQ21, 1T6U1, 1T7R1, 1TX41, 1TZV1, 1TZY2, 1TZY4, 1VDK1, 1VLG1, 1W531, 1WDC2, 1WKU1, 1WOL1, 1WPB1, 1WVE2, 1K961, 1YOY1, 1YYD1, 1Z101, 2ABK1, 2BAA1, 2CCH2, 2CIW1, 2CZ21, 2EUT1 ,2GC44 ,2GKM1, 2I5N1, 2I5N3, 2I5N4, 2INC1, 2INC2, 451C1, 5CSM1,1BZR1
BETA (91 cadeias)	1JIW2, 1F582, 1SBW1, 1TGS1, 1A121, 1BHE1, 1C9O1, 1CRU1, 1EAJ1, 1EUR1, 1EUW1, 1F8E1, 1FLT2, 1FNS1, 1FNS2, 1GQ81, 1GSK1, 1GUI1, 1HOE1, 1I0C1, 1IBY1, 1J831, 1K121, 1KV71, 1LK33, 1LR51, 1M9Z1, 1NSZ1, 1O5U1, 1O6S2, 1OFL1, 1OFZ1, 1OH41, 1PBY2, 1PMH1, 1PNF1, 1PQ71, 1PXV2, 1QHV1, 1RG81, 1RMG1, 1ROC1, 1RWI1, 1SFD1, 1SQ91, 1SR43, 1SVB1, 1SVP1, 1T2W1, 1T611, 1T612, 1TCZ1, 1TUD1, 1UAC2, 1UMH1, 1USR1, 1UV41, 1UWW1, 1UXZ1, 1V051, 1V6P1, 1VPS1, 1WD31, 1XQH1, 1Y0M1, 1Y7B1, 1ZE31, 1ZE32, 1ZGO1, 2A2Q3, 2ADF2, 2ADF3, 2AG41, 2AGY1, 2BCM1, 2DJF1, 2FCB1, 2FGQ1, 2FK91, 2GC42, 2GC43, 2H3L1, 2HS11, 2IAV1, 2IVZ1, 2J1N1, 2O8L1, 2POR1, 2SIL1, 3EZM1, 1K5C1

* Os PDB IDs foram concatenados com o ID da cadeia re-enumerada conforme sua ocorrência nos PDBs originais.

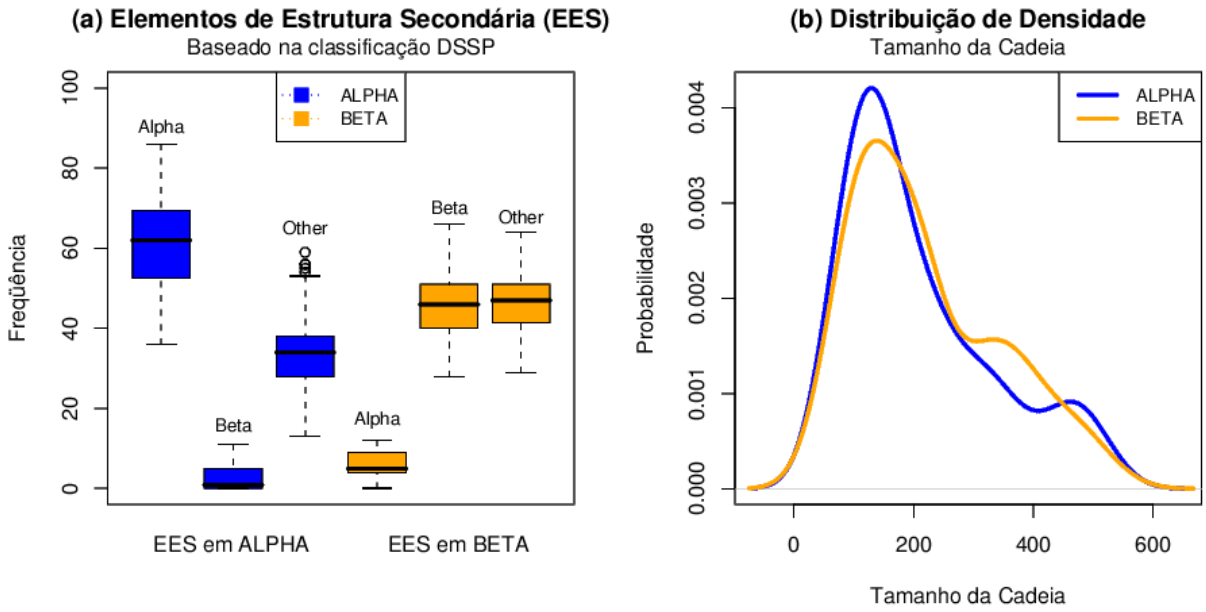


Figura 3: Perfil estatístico dos conjuntos de dados ALPHA e BETA. (**a**) A distribuição de elementos de estruturas secundárias segundo DSSP para ALPHA (em azul) e BETA (em laranja). O conteúdo "alfa" teve média e desvio padrão de 61,6 ± 11.6 %, com min/max de 36,0%/86,0%. O conteúdo "beta" teve média e desvio padrão de 46,6 ± 7,9 %, com min/max de 28,0%/66,0%. (**b**) Distribuição de densidade para o tamanho da cadeia nos conjuntos ALPHA and BETA, em azul e laranja respectivamente. A média e desvio padrão em ALPHA foi de 210 ± 125 resíduos, com min/max of 61/522 resíduos, totalizando 19163 resíduos. A média e desvio padrão em BETA foi de 221 ± 122 resíduos com min/max of 59/534 resíduos, totalizando 20127 resíduos. A homogeneidade das médias e variâncias foram aferidas pelos testes não-paramétricos Wilcoxon[132] e Fligner-Killeen[133], dando *p-values* de 0,44 e 0,23, respectivamente. O teste de Kolmogorov-Smirnov[134] assegurou a aderência (*goodness-of-fit*) das distribuições entre si, com *p-value* de 0,31.

4.3 Carácter globular dos conjuntos

Outra homogeneidade importante que devemos asseverar é que os conjuntos ALPHA e BETA tenham o mesmo perfil globular, o que pode ser inferido pela relação superfície/volume. Chothia e Janin[135] demonstraram, em uma aproximação à sólidos de contorno similar, que a relação entre a área acessível ao solvente (A_s) e massa molecular (M) me proteínas poderia ser dada por:

$$[1] \quad A_s = k_a M^d$$

onde k_a e d são constantes. Eles encontraram um $k_a \simeq 11{,}1$ e $d \simeq 0{,}70$. Este último foi assumido por eles como sendo suficientemente próximo de 2/3 (0,666...) conforme esperado para uma esfera perfeita. Nós podemos modificar[22] a equação [1] para produzir uma razão superfície (A_s) por volume (V) em função do número de resíduos (tamanho da cadeia) n:

$$[2] \quad \frac{A_s}{V} = k_b n^{-\frac{1}{3}}$$

que pode ser posta numa forma linear com uma transformação log-log:

$$[3] \quad \ln\left(\frac{A_s}{V}\right) = \ln(k_b) - \frac{1}{3}\ln(n)$$

onde K_b é uma constante. A Figura 4 mostra os gráficos para equações [2] e [3]. A regressão linear em [3] oferece as seguintes estatísticas, ao nível de confidência de 0,95: para ALPHA, intercepto igual a 1,00 ± 0,24, inclinação – 0,36 ± 0,04, coeficiente de determinação de 0,74; para BETA, intercepto igual a 0,95 ± 0,22, inclinação – 0,36 ± 0,04, coeficiente de determinação de 0,78. Podemos ver por esses parâmetros que ambos os conjuntos são

22 - Se $A_s/V = k_1 r^{-1}$, e se o volume é dado por $V = k_2 n$ e por $V = k_3 R^3$, então $r = k_4 n^{1/3}$ e $A_s/V = k_b n^{-1/3}$, sendo k_1, k_2, k_3, k_4 constantes.

razoavelmente homogêneos quanto ao carácter globular. A inclinação é virtualmente a mesma e o valor esperado de 1/3 (0,333...) está dentro do intervalo de confiança a 0,95.

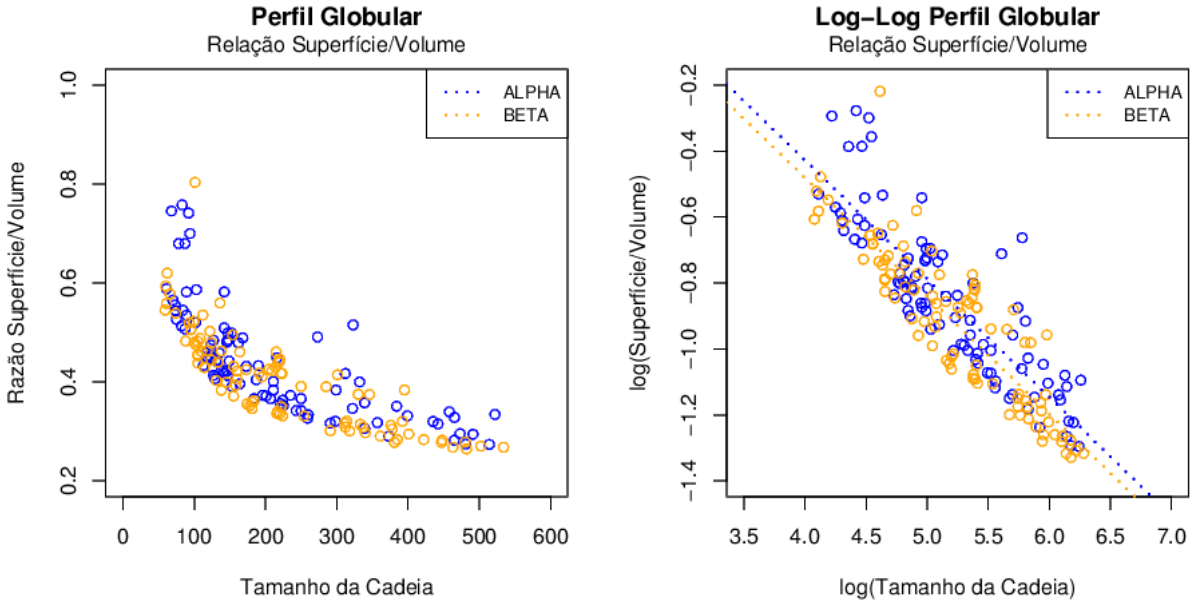


Figura 4: Perfil globular inferido pela relação superfície/volume para os conjuntos ALPHA e BETA.

Tanto volume quanto superfície foram calculados pelos programas de Gerstein *et. al.* [137][138] `calc-surface()` e `calc-volume()`, usando o tradicional método B[76] de Richards e também adotando os raios de van der Waals de Richards[23]. Importante frisar que as cadeias não foram solvatadas para esses cálculos. Testes preliminares para estimação de volume da mioglobina 1MBO[139] mostraram que a influência do solvente é pequena: 20521,9 $Å^3$ sem solvente e 20985,4 $Å^3$ com solvente[24] (diferença de aproximadamente 2%). Esse desvio pareceu-nos irrelevante para a precisão requerida nesta análise de superfície/volume. Além do que os resultados da regressão ficaram dentro do esperado teoricamente. Retornaremos à questão do solvente mais adiante.

4.4 Contatos

23 - C = 2.00 Å, O = 1,40 Å, N = 1,70 Å, S = 1,80 Å, P = 2,00 Å, demais = 1,80 Å

24 - A caixa de água foi adicionada usando os parâmetros *default* do GROMACS[136], através das rotinas `editconf()` e `genbox()`. Nenhuma otimização foi feita.

A definição de contatos que adotamos aqui é essencialmente geométrica e não energética. Envolve a determinação do conjunto de pontos vizinhos a um determinado centróides, tendo as distâncias Euclideanas como peso. Todos os contatos são definidos também no nível dos resíduos, com um centróides por resíduo: ou por seus carbonos alfas (CA) ou pelos centros geométricos das cadeias laterais (GC), reduzidas aos carbonos alfas para o caso das glicinas.

4.4.1 Contatos delimitadores dependentes

Nós implementamos a versão tradicional do método de aferir contatos (que abreviaremos por **CD** – ***C**utoff **D**ependent*) usando um delimitador: um contato é computado entre um par de resíduos $\{i,j\}$ se a distância Euclideana entre seus centróides for menor ou igual a um determinado valor. Para uma descrição mais matemática do método, nós vamos adotar aqui o elegante formalismo encontrado em Miyazawa & Jernigan[100]. Eles começam por definir uma função de contato:

$$[4] \quad C(i,j,r,d_r) = \begin{cases} 0 & if \ \ |i-j| \le d_r \\ H(r,d_{i,j}) & if \ \ |i-j| > d_r \end{cases}$$

onde i e j são posições dos resíduos na cadeia, r o delimitador de distância (*cutoff*), d_r é a distância em resíduos consecutivos ao longo da estrutura primária, H é uma função degrau tipo Heaviside expressa como:

$$[5] \quad H(r,d_{i,j}) = \begin{cases} 1 & if \ \ d_{i,j} \le r \\ 0 & if \ \ d_{i,j} > r \end{cases}$$

onde $d_{i,j}$ é a distância Euclideana entre resíduos $\{i,j\}$ em Å. Equação [5] verifica se os resíduos i e j estão dentro da esfera de raio r centrada em i. Equação [4] checa se o intervalo na seqüência entre i e j é maior ou igual a d_r. Se ambos forem verdadeiros, o contato é aceito. Isto é feito para todos os pares $\{i,j\}$ não contando pares previamente aceitos ($j > i$). Ou seja, somente as arestas do grafo de contatos são computadas. Logo, para uma proteína o número total de arestas seria computado por:

$$[6] \quad N(n,r,d_r) = \sum_{i=1}^{n} \sum_{j>i}^{n} C(i,j,r,d_r)$$

onde n é o número total de resíduos. A matriz de distâncias $M_{p,r}$ é construída aplicando [6] para todas as p proteínas das bases de dados, para a maioria dos casos variando r entre 0,0 Å à 28,0 Å, com passo de 0,2 Å:

$$[7] \quad M_{p,r} = \sum_{p=1}^{P} \sum_{r=0}^{R} N(n,r,d_r)$$

Nós analisamos o efeito do delimitador r sobre o perfil de contatos, numa técnica que nós denominamos varredura exaustiva de distâncias em proteínas (*protein cutoff scanning*). Isto forma uma distribuição acumulativa de distâncias em r. A distribuição de densidade foi estimada pela diferença em [6]:

$$[8] \quad Diff\ N(n,r,d_r) = N(n,r+s,d_r) - N(n,r,d_r)$$

onde s é o passo.

Detalhes importantes:

- Neste trabalho nós não diferenciamos contatos locais de não-locais na seqüência, de modo que $d_r = 0$. Chamamos de contatos locais aqueles vizinhos e próximos na estrutura primária; não-locais, os distantes.

- A equação [6] conta o número de arestas no grafo de contatos para a proteína *p*. O número total de vizinhos é obtido multiplicando por 2 o número total de arestas (*Handshaking* lema[140]). Se esse resultado é normalizado pelo número de resíduos, ele irá fornecer a coordenação média por resíduo de cada proteína.

4.4.2 Contatos delimitadores independentes

4.4.2.1 Diagramas de Voronoi

Os diagramas de Voronoi são um constructo geométrico assim denominado em honra ao matemático russo Georgy Voronoi (1868-1908), que foi quem em 1908 o generalizou para o caso *n* dimensional[114]. No entanto, as idéias básicas para baixas dimensões remontam aos trabalhos pioneiros de Johanm P. G. L. Dirichlet (1805-1859)[141], K. F. Gauss (1777-1855) [142] e R. Descartes (1596-1650)[143]. Há uma certa confusão histórica sobre quem primeiro conseguiu juntar esses conceitos num modelo aplicável ao mundo real. Há quem credite o caso bidimensional[144] ao meteorologista Alfred H. Thiessen (1872- ?) e a versão tridimensional[145] aos físicos Eugene P. Wigner (1902-1995) e Frederick Seitz (1911-).

Figura 5 ilustra como exemplo a construção de um diagrama de Voronoi no plano Euclideano 2d. Sua generalização para dimensões maiores segue-se naturalmente. Antes de prosseguir, estabeleceremos aqui algumas definições básicas. Chamaremos um *k-simplex* um "casco convexo" (*convex hull*)(veja figura 5g) de *k+1* sites afins independentes num espaço Euclideano $\mathbb{R}^k$. Estes *simplices*, no contexto das tesselações de Voronoi/Delaunay, vão identificar suas entidades básicas: 0-*simplex* para pontos, 1-*simplex*, para retas, 2-*simplex* para triângulos, 3-*simplex* para tetraedros e assim por diante.

Faremos uso aqui do formalismo adotado por Aurenhammer[146] na descrição matemática de um constructo Voronoi. Seja *S* o conjunto de *n* pontos (ou sites) especiais em um espaço Euclideano *2d*. Seja *p*, *q* dois destes distintos pontos de *S*. Seja $\delta(.)$ uma função de distâncias Euclideanas. Nós podemos definir uma região de influência de *p* sobre *q*, *R(p,q)*, como o semi-plano em que todos os seus pontos estão ao menos mais próximos de *p* que de *q*:

$$\forall\ p,q\ \in \mathbb{R}$$
$$[9] \quad R(p,q)\ =\ \{\ x\ \in\ \mathbb{R}^2\ \mid\ \delta(x,p) < \delta(x,q)\}$$

É fácil ver que todos os pontos que são eqüidistantes de *p* e *q* formam um bissetor que delimita a área de influência de ambos (Figura 5a). Agora, nós podemos estender essa bissecção para todos os *n-1* sites restantes de *S*, concebendo uma região de dominância de *p* que nós chamaremos de *VR(p, S)* como a intersecção:

$$[10] \quad VR(p,S) = \bigcap_{q \in S-\{p\}} R(p,q)$$

Aplicando em todos os sites de *S*, nós iremos criar um particionamento do espaço que contém *S*:

$$[11] \quad V(S) = \bigcup_{p,q \in S,\, p \neq q} VR(p,S) \cap VR(q,S)$$

V(S) será uma função Voronoi que irá decompor o espaço que contém *S* em polígonos (Figura 5a-f). Está função têm três propriedades importantes:

1. Todos os polígonos desenhados são convexos;

2. Todas as arestas são eqüidistantes de exatos dois sites;

3. Todos os vértices são eqüidistantes de ao menos três sites;

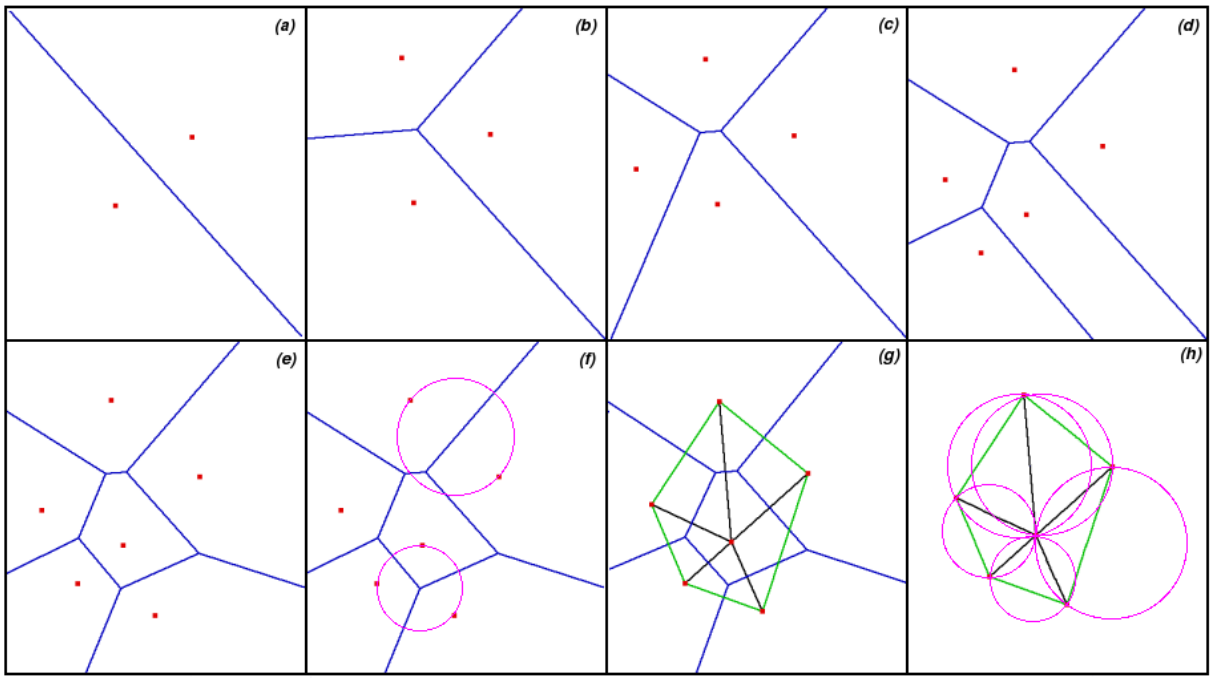


Figura 5: Ilustração do processo de construção de um diagrama de Voronoi em 2d e algumas de suas propriedades. (**a**) Um bissetor (azul) delimitando duas regiões e seus sites (quadrados vermelhos). (**b** até **e**) Processo de particionamento do espaço com 3, 4, 5 e 6 sites, respectivamente. Perceba que as arestas sempre estão eqüidistantes de dois pontos e que a vizinhança é composta por um critério de maior proximidade (sempre a maior possível). Note também que os polígonos (fechados) formados serão sempre convexos. (**f**) Círculos (em rosa) evidenciando duas importantes propriedades: toda aresta será eqüidistante de exatos dois sites; e todo vértice será eqüidistante de ao menos três sites. (**g**) A decomposição de Delaunay como um grafo dual dos diagramas de Voronoi (arestas em preto e verde). A aresta em verde destaca o casco convexo (*convex hull*), formando um "envelope" ao redor dos sites. Veja que o casco convexo é um subconjunto das arestas de Delaunay. (**h**) A propriedade do círculo vazio: três sites compor-se-ão um *simplex* se e somente se não há nenhum ponto no interior do círculo circunscrito a eles. Perceba que todos os vértices em Voronoi são também o centro dos círculos em Delaunay (compare com f). Figuras adaptadas de VoroGlide[119]

4.4.2.2 Tesselação de Delaunay

Um constructo geométrico relacionado aos diagramas de Voronoi (VD) é a tesselação de Delaunay (DT)[115]. Voronoi em seu celebrado artigo de 1908 já tinha percebido que o

grafo dual aos seus diagramas em redes regulares (*lattices*) tinham importantes características. Um grafo dual é obtido pela associação de um vértice a cada região do grafo alvo, compondo arestas no novo grafo se e somente se essas regiões do grafo alvo compartilham uma aresta[140]. Um outro russo, Boris Delaunay(1890-1980), estendeu a concepção original de Voronoi de redes regulares (*lattices*) para qualquer conjunto de pontos no espaço, através de um método muito engenhoso: três sites compor-se-iam em uma triangulação de Delaunay se e somente se o círculo circunscrito a eles não contivesse nenhum outro site. Aplicando esse método a todos os sites irá decompor o espaço ocupado por eles em triângulos justapostos, fazendo emergir a notável propriedade de que somente os vizinhos mais próximos e não oclusos por outro estarão conectados.

Este elegante resultado pode ser mais bem entendido se olharmos na dualidade dos grafos. É fácil perceber que há uma explícita correspondência entre os elementos dos grafos Voronoi *V(S)* e Delaunay *D(S)*: vértices em *V(S)* correspondem com triângulos em *D(S)*; regiões em *V(S)* correspondem com vértices em *D(S)*; e arestas em *V(S)* vão ser ortogonais às arestas em *D(S)*, mesmo que as primeiras não necessariamente interceptem as segundas. Se, como já vimos, *V(S)* mapeia as regiões de acordo com um critério de máxima proximidade, então as arestas em *D(S)* conectam os sites mais próximos possíveis.

Neste trabalho, nós usaremos o programa ADGCAL[147] na construção dos grafos de Delaunay (que abreviaremos por **DT** – ***D****elaunay* ***T****essellation*) em proteínas. Faremos uso também das mesmas equações [4] a [8] utilizadas na metodologia CD para compor a varredura de distâncias (*cutoff scanning*) em DT, mas obviamente apenas considerando as informações retornadas pelo ADGCAL.

4.5 Solvatação

Importante registrar que as cadeias de nossa base de dados não foram solvatadas, seja para os cálculos de superfície e volume (como já visto), seja para a determinação dos contatos Delaunay. Para esse primeiro trabalho nós queríamos ver como os perfis de distâncias comportavam-se acrescentando o mínimo possível de artificialidades aos dados experimentais

oriundos do PDB. A adição de águas virtuais a uma proteína por um programa de computador não deixa de ser uma interferência artificial nos dados em análise. E se ela tiver que ser feita, tem que ser muito criteriosa para causar o mínimo de interferência e ser o mais próxima possível da realidade. Sendo assim, por questões estratégicas, nós deixamos para um outro momento a análise do efeito da solvatação no perfil dos contatos (vide a seção Limitações e Perspectivas).

5. Resultados e Discussões

Primeiramente, iremos fazer uma análise comparativa ressaltando as principais propriedades e alguns problemas intrínsecos a essas duas metodologias: delimitador dependente (CD) e tesselação de Delaunay (DT).

5.1 Delimitador Dependente - CD

Uma das grandes vantagens deste método tradicional de aferir contatos (CD) é que ele faz uma completa varredura combinatória de todas as arestas possíveis no grafo de contatos, dadas as esferas de busca delimitadas por um raio *r,* a partir de cada site. Logo, se existem *n* diferentes vértices ou sites (resíduos) no interior destas esferas, haverá *C(n,2)* combinações de arestas enumeradas por CD, o que nos dá $O(n^2)$ contatos (Figura 7a). Como em um volume com pontos distribuídos de uma maneira aproximadamente uniforme é esperado que o número de sites cresça em $O(r^3)$, então o número de contatos em função das distâncias deverá variar em $O(r^6)$. Isto não é um crescimento exponencial, mas é um polinômio de elevado grau, e será esperado que o número de contatos cresça de forma vigorosa com a distância *r*, embora nós tenhamos evidências de que os coeficientes para os expoentes mais altos sejam bem pequenos (vide Anexo A). Em sistemas bem empacotados, como em proteínas, na medida que *r* cresce, as esferas de busca irão extrapolar todas as camadas de empacotamento ou coordenação (*coordination shell*) até alcançar o limite da proteína e seus resíduos mais externos. Sites mais internalizados irão sentir esse efeito de borda em delimitadores mais altos do que sites próximos à superfície, mas o efeito geral será que o crescimento cumulativo no número de contatos por distância será contido. O resultado será uma distribuição cumulativa sigmoidal assintótica a *C(n,2)*. Veja, como exemplo, o comportamento dos contatos para a mioglobina 1BZR na Figura 6a. Naturalmente, dada a simetria das curvas sigmoidais, sua distribuição de densidade (primeira derivada) pode ser confundida com uma Gaussiana, mas ela continuará polinomial, agora com $O(r^5)$ (Figura 6b).

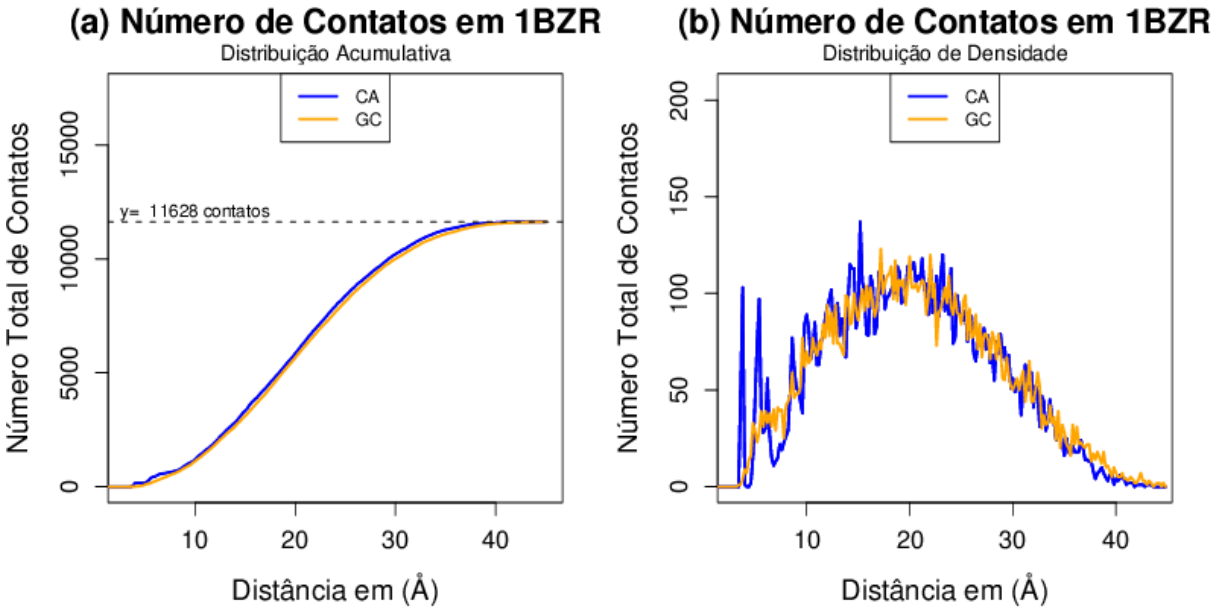


Figura 6: distribuição cumulativa e de densidade para o número total de contatos (arestas) na mioglobina 1BZR[148] com resíduos sendo representados por carbonos alfa (CA – em azul) e centro geométrico da cadeia lateral (GC – em laranja). (**a**) Distribuição cumulativa. Perceba que ela é uma sigmóide assintótica a C(153,2) = 11628 contatos. (**b**) Distribuição de densidade. A representação de resíduos por CA em distâncias menores reflete a maior organização da cadeia principal com acentuadas oscilações no perfil da curva. A representação de resíduos por GC é mais homogênea.

Enfatizando, o grande destaque da metodologia CD é que ela é total ou exaustiva, enumerando todos os contatos (arestas) que podem existir num determinado intervalo de distância. Essa propriedade global tem sido explorada de forma interessante, por exemplo, na análise do comportamento geral do empacotamento em função de distâncias[110]. Mas, se o objetivo é sondar apenas os contatos de primeira ordem envolvendo os vizinhos não oclusos da primeira camada de coordenação, a metodologia CD não oferece uma forma fácil e segura de evidenciá-la sem o risco de ter que confrontar com contatos falso-positivos (oclusos e contados) e falso-negativos (não oclusos e não contados). Usualmente o que se tem feito é tentar estimar um limite estatisticamente ótimo que minimize esses riscos, em geral observando o comportamento da densidade de contatos em função das distâncias.

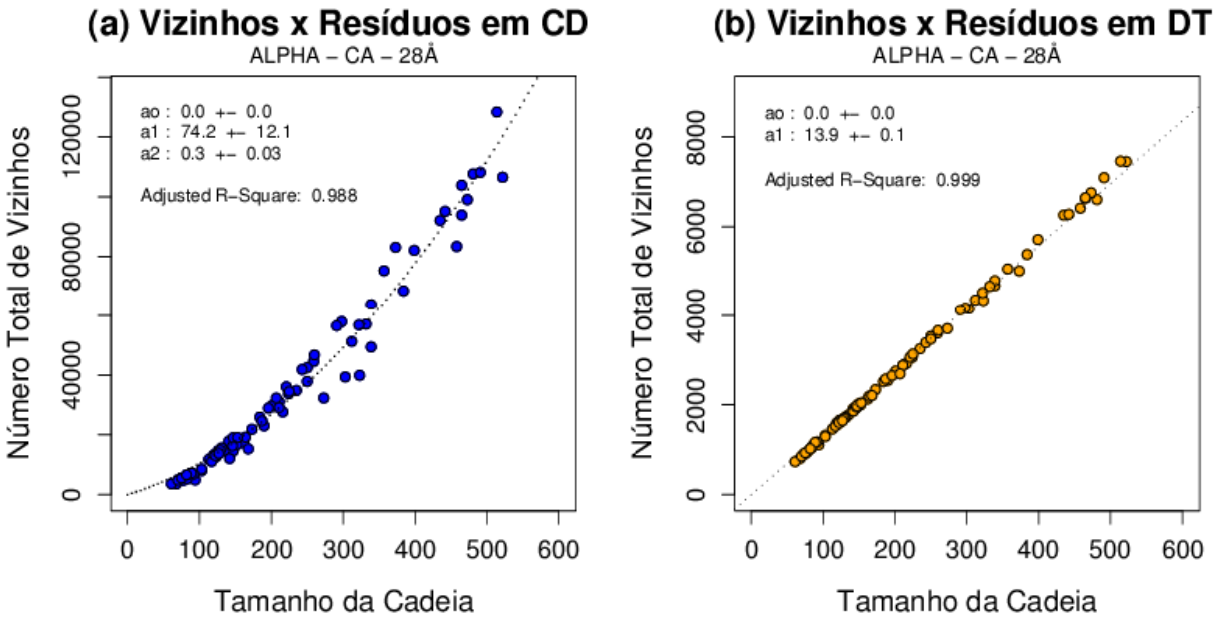


Figura 7: Total de vizinhos por resíduos com limite de 28,0 Å para a maior distância inter-resíduo, num comparativo entre as metodologias CD e DT para conjuntos ALPHA com representação de resíduos por CA. (**a**) Contatos por resíduos em CD. Os números no canto esquerdo indicam os valores dos coeficientes e erros padrões, ao nível de confidência de 0,95, para o ajuste quadrático, juntamente com o coeficiente de determinação. (**b**) Contatos por resíduos em DT. Os números no canto esquerdo mostram os valores dos coeficientes e erros padrões, ao nível de confidência de 0,95, para o ajuste linear, juntamente com o coeficiente de determinação.

É provável que os indianos Manavalan & Ponnuswamy[101], em 1977, tenham sido os primeiros a enfrentarem esse desafio. Examinando um conjunto de 14 proteínas, e representando seus resíduos por CA, eles encontraram que os hidrofóbicos estavam maximamente agrupados (*clustered*) entre 6 e 8 Å, e sugeriram o último como um valor ideal. A seguir, Miyazawa & Jerningan (1985)[100], analisando o padrão radial das densidades de contatos de resíduos internos à cadeia[25], evidenciaram um pico entre 5,0 e 5,5 Å, seguido por um vale em torno de 6,5 Å. Este foi o valor assumido como o delimitador ideal para a análise de potenciais empíricos que eles estavam fazendo envolvendo um conjunto especial de 42 proteínas. É importante ressaltar que eles usaram GC como centróide de seus resíduos (CA para Glicina). Seguindo uma metodologia similar, Zhang *et. al*. (1997)[102] conseguiram estimar seu melhor delimitador em 6,0 Å, mas aferindo seus contatos inter-resíduos a partir do grau de proximidade dos seus átomos pesados, numa seleta coleção de 89 proteínas. Furuichi & Koehl (1998) tentaram uma abordagem diferente, contrastando os efeitos do tamanho da cadeia no padrão de contatos. Usando uma representação de resíduos por CA, eles construíram dois subconjuntos com 68 proteínas ao todo, chamados de S e L, o primeiro

25 - Eles consideraram resíduos como internos aqueles cujos centróides distavam menos que 7.0 Å do centro geométrico da proteína.

designando aquelas com menos de 130 resíduos em suas cadeias, e o segundo com as demais. Dois resultados surpreendentes foram encontrados: primeiro, foi que as duas distribuições S e L revelaram-se idênticas até 10,0 Å, indicando que as interações de curta escala eram independentes do tamanho da cadeia; segundo, que o poder preditivo de seus potenciais empíricos referendados pelos *scores* de avaliação divergiam entre 7,0 e 8,0 Å. Em face disso, eles ficaram com o valor de 8,0 Å como o seu melhor delimitador. Mais recentemente, Kamagata & Kuwajima (2006) introduziram um viés experimental na determinação do delimitador ideal. Eles verificaram uma correlação surpreendentemente alta entre o número de contatos inter-resíduos densamente agrupados (*clusters*) *Nc* e o log das constantes cinéticas dos intermediários no enovelamento de 12 proteínas não-dois-estados. Na definição dos agrupamentos *Nc*, eles utilizaram uma granulação mais fina na contagem dos contatos inter-resíduos, verificando a proximidade entre seus átomos pesados. Variando o delimitador era possível interferir na quantidade de contatos em *Nc*. Eles então testaram a influência do delimitador no coeficiente correlação linear entre *Nc* e o log das constantes, e viram que não havia alterações significativas em distâncias maiores que 5,5 Å.

5.2 Decomposição de Delaunay - DT

Se CD tem que enfrentar o problema da ambigüidade na definição dos contatos de primeira ordem, técnicas delimitadores independentes como a decomposição de Delaunay (DT) sentem-se razoavelmente à vontade nessa questão. Isso porque a forma geometrica como os contados são definidos reduzem consideravelmente as chances de produzir contatos inadequados. Para a maioria dos casos há uma garantia matemática de que os contatos mais próximos estarão não oclusos porque, conforme visto, as bissecções que formam os separadores de regiões nos diagramas de Voronoi (VD) sempre são traçadas entre os sites de maior proximidade, o que praticamente elimina as chances de que uma aresta seja composta atravessando outros sites no meio do caminho[26].

26 - Porque se há um site intermediando outros dois, ele estará mais próximo de ambos, e a aresta será feita com ele.

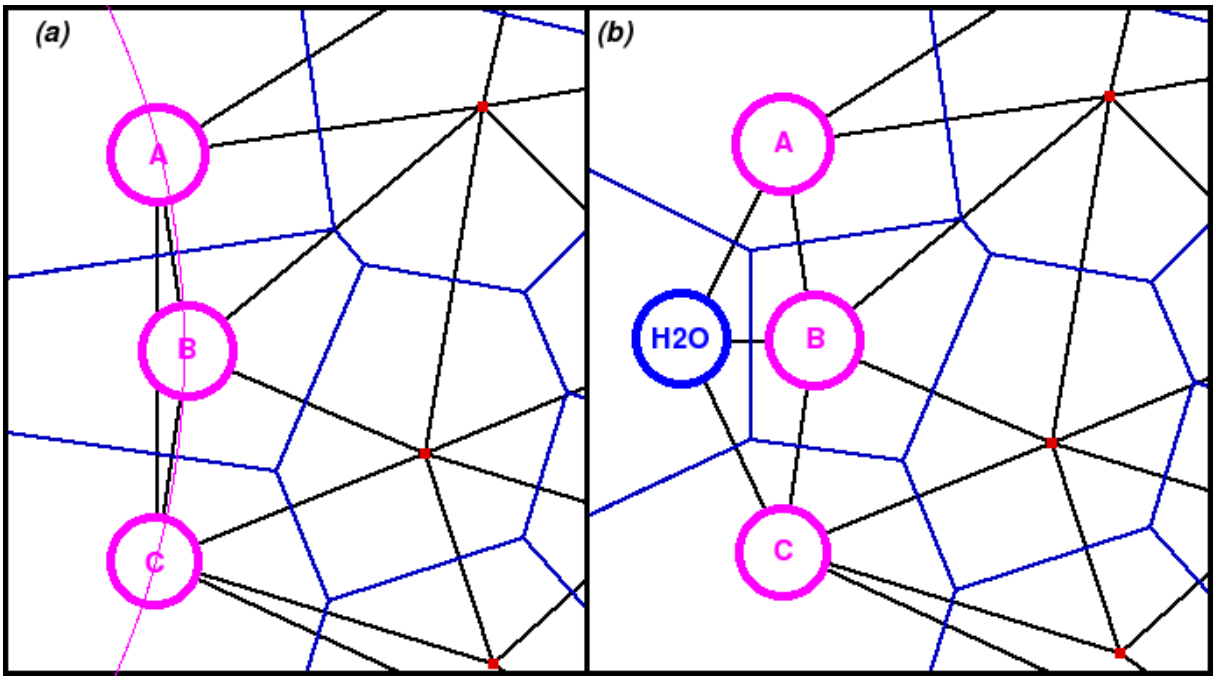


Figura 8: Caso raro de oclusão nas tesselações de Voronoi/Delaunay. (**a**) Os sites *A*, *B* e *C* estão numa situação quase-linear, com o centro da esfera comum a eles muito deslocado à esquerda (próximo do infinito). Nessa condição, *B* está parcialmente ocluindo *A* e *C*, mas ainda assim uma aresta {*A,C*} é feita (traços pretos). (**b**) A solvatação resolve o problema, já que a presença do solvente intervem na aresta {*A,C*}.

Exceções podem ocorrer em sites na superfície da proteína. Em certas configurações, arestas podem ser construídas envolvendo sites em uma situação de quase-linearidade, como exemplificado na Figura 8. Mas, esses casos tendem a ser raros e podem ser contornados com alguns procedimentos usuais em aplicações de Voronoi/Delaunay em proteínas, como a solvatação[76].

Seja como for, é justamente esse alto senso geométrico e a aparente desambigüidade alguns dos motivos pelos quais DT/VD vêm sendo largamente utilizados como venerados métodos de identificação de contatos em proteínas[116]. E eles saem-se melhor ainda quando o alvo é a mais correta e precisa identificação possível da vizinhança imediata a cada site, que já sabemos ser a maior dificuldade da metodologia CD.

Outra intrigante propriedade também contribui para esse culto ao método: independente do número de dimensões, o número de arestas por site em aglomerados aproximadamente uniformes tende a crescer em *O(n)* no caso médio[149]. Em proteínas nós podemos constatar que essa assertiva é de fato verdadeira. A Figura 7b claramente revela uma alta correlação linear entre o número total de vizinhos (com arestas limitadas no tamanho à 28

Å) e o tamanho da cadeia. Esta linearidade indica que DT consegue capturar um limitado número médio de vizinhos para cada site, o que faz com que o número total de contatos cresça em proporção direta ao tamanho da proteína. A Figura 9 tenta ilustrar de forma intuitiva essa visão.

Como conseqüência, a inclinação da reta conterá valiosa informação a cerca do número médio de vizinhos por site, na distância de corte escolhida. Para nosso conjunto ALPHA CA exibido na Figura 7b, com corte de 28 Å para o tamanho máximo de uma aresta, vemos que o valor estimado para inclinação da reta ali exposta é de 13,9 ± 0.1 vizinhos por site (a 0,95 de nível de confidência). Para os demais conjuntos esse número médio foi de: BETA CA: 14,0 ± 0.1, ALPHA GC: 13,4 ± 0.1, BETA GC: 13,6 ± 0.1, todos a 0,95 no nível de confidência. Estes resultados são bem próximos do valor 13,97 encontrado por Soyer et. al. (2000)[123] para o número médio de faces em poliedros Voronoi de uma coleção não redundante de 40 proteínas, com os resíduos estando representados pelo baricentro da cadeia lateral.

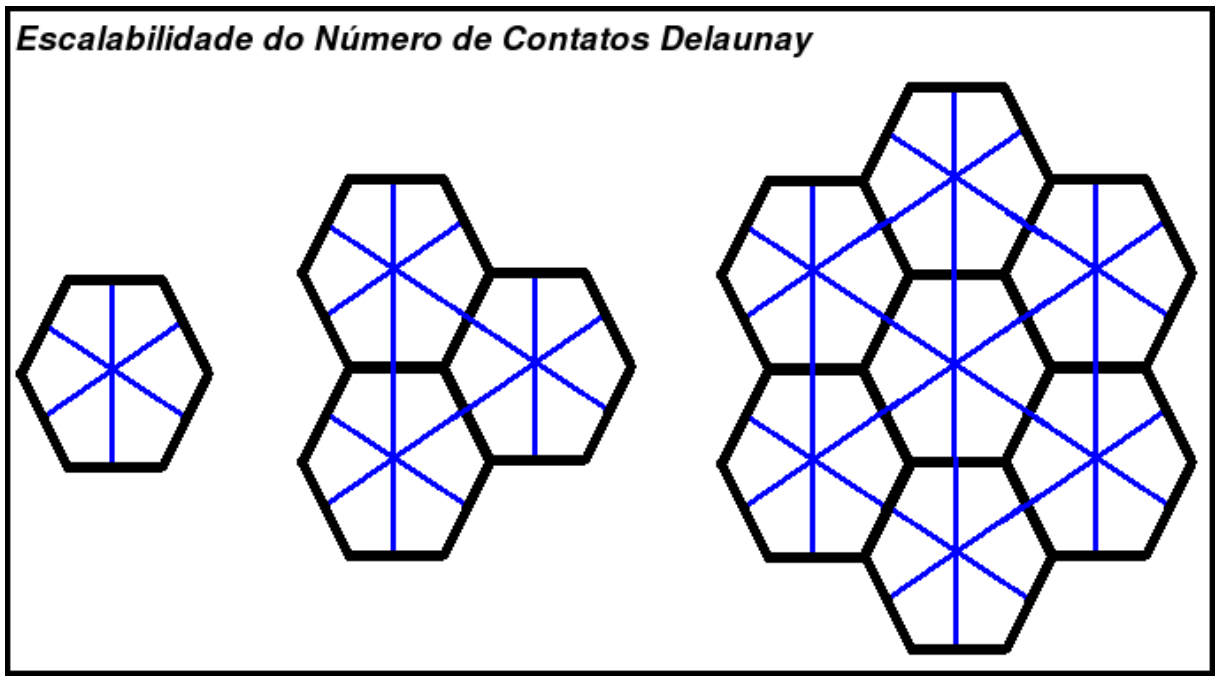


Figura 9: Um exemplo intuitivo da escalabilidade linear entre o número total de contatos com o tamanho da proteína. Perceba que o número de arestas Delaunay (em azul) irá crescer proporcionalmente com o número de células (ou favos) Voronoi, aqui representando no conjunto o tamanho de uma proteína.

Note que ambos os ajustes nas figuras 7a e 7b foram forçados a terem o intercepto em zero, dado que é muito razoável esperar que com zero resíduos o número total de vizinhos

será também zero. Mas, é possível verificar que as pequenas proteínas (com poucos resíduos) tendem a projetar seus pontos no gráfico abaixo da linha ajustada, tanto em CD quanto em DT. Provavelmente isto ocorre porque quando os delimitadores alcançam distâncias maiores, a contribuição para a contagem das arestas de proteínas pequenas é cada vez menor frente as proteínas grandes (e vice-versa. Vide Figura 10). Para CD, a varredura de distâncias irá alcançar os limites das pequenas proteínas mais rapidamente, exaurindo sua capacidade de contribuir com mais arestas (Figura 11a). Em DT, quando a varredura alcança valores de distâncias maiores, as contribuições para a contagem de arestas serão dadas principalmente por sites na superfície. E como proteínas pequenas tem relativamente menor área, a participação delas no processo geral de contagem será menor que em proteínas volumosas[27](Figura 11b).

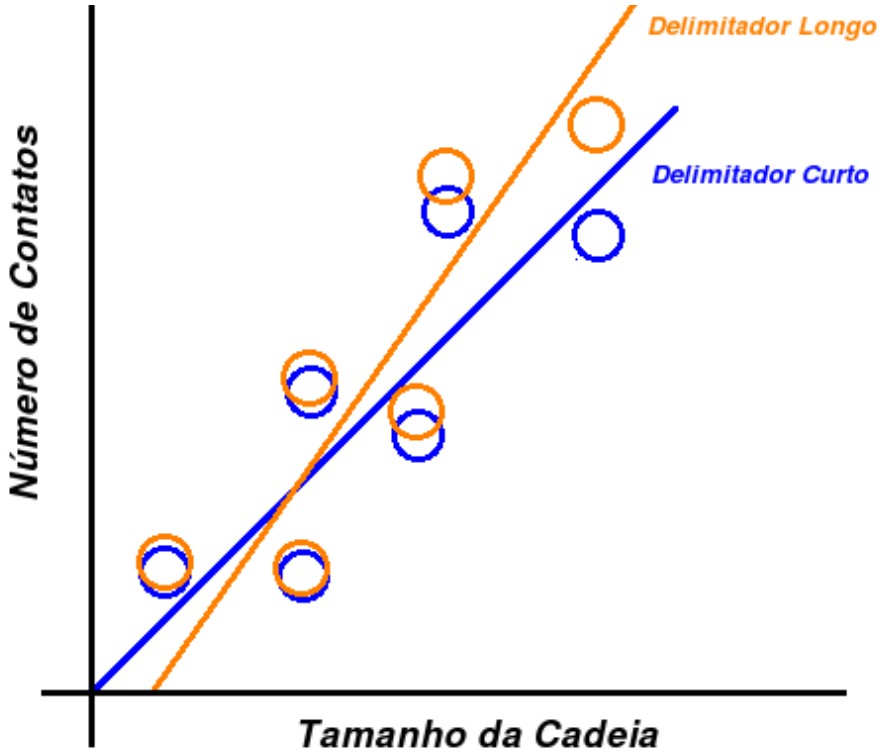


Figura 10: Ilustração da influência do tamanho das proteínas na regressão linear (em DT) conforme o valor do delimitador de distâncias. Se esse delimitador opera em curtas distâncias, todas as proteínas podem contribuir de forma igual na contagem dos contatos. Mas, se ele permite longas distâncias, a capacidade de participação das proteínas menores é exaurida mais rapidamente que das maiores. Isso produz uma taxa maior de crescimento no número de contatos para proteínas grandes, desviando do zero o intercepto da reta.

27 - Obviamente que, se as proteínas estiverem solvatadas, esse efeito será minimizado em DT. Vide seção Limitações e Perspectivas.

Normalizar o total de vizinhos pelo número de resíduos seria uma solução ? Vemos na Figura 12 que infelizmente isso não resolve. Como CD é quadrático por natureza na relação arestas por nós, seria esperado que a normalização pelo número de resíduos o linearizasse. Mas, não é isso que nós vemos na Figura 12a. O ajuste linear fica meio forçado. As cadeias com menos de 200 resíduos parecem seguir uma reta e as maiores outra. De fato, a linearização piorou o modelo, já que o coeficiente de determinação do ajuste que era de 0,968 caiu para 0,851. A explicação para esse resultado continua relacionada às diferenças no tamanho da cadeia. Em 28,0 Å, as proteínas pequenas já estão muito próximas de cobrir todas as combinações possíveis entre seus sites. Como sabemos que isso é seguramente $O(n^2)$, a normalização por n irá linearizá-lo com tranqüilidade. Por outro lado, nesta distância de 28,0 Å as grandes ainda não esgotaram todas as combinações possíveis de contatos entre seus sites. Logo, o seu total de vizinhos normalizados pela quantidade de resíduos fica aquém do esperado, colocando-os à direita da reta teórica.

Para DT a normalização também não é solução. Como a relação arestas x nós em DT é linear, a normalização pelo número de resíduos deveria redundar numa reta paralela ao eixo *X*. Mas, vemos na Figura 12b que ela de fato aproxima-se disso mas ainda mantém uma pequena inclinação. Também aqui a normalização piora o modelo, com queda no coeficiente de determinação de 0,999 para 0,696[28]. Há, como em CD, um conjunto de pequenas proteínas que não se encaixam bem na reta sem intercepto traçada. A explicação das diferenças nos tamanhos da cadeia continuam valendo aqui também, com a superfície das grandes proteínas superestimando o número médio de vizinhos por site.

Trata-se sem dúvida de dois exemplos alarmantes de que nem sempre a normalização de um conjunto heterogêneo de dados aplaina suas diferenças. Toda normalização tem que ser feita de forma criteriosa, ou pode-se, como visto aqui, piorar um modelo invés de incrementá-lo.

28 Devemos ressaltar que o R^2 aqui está contaminado pela perda de correlação na medida que a reta fica paralela à abscissa.

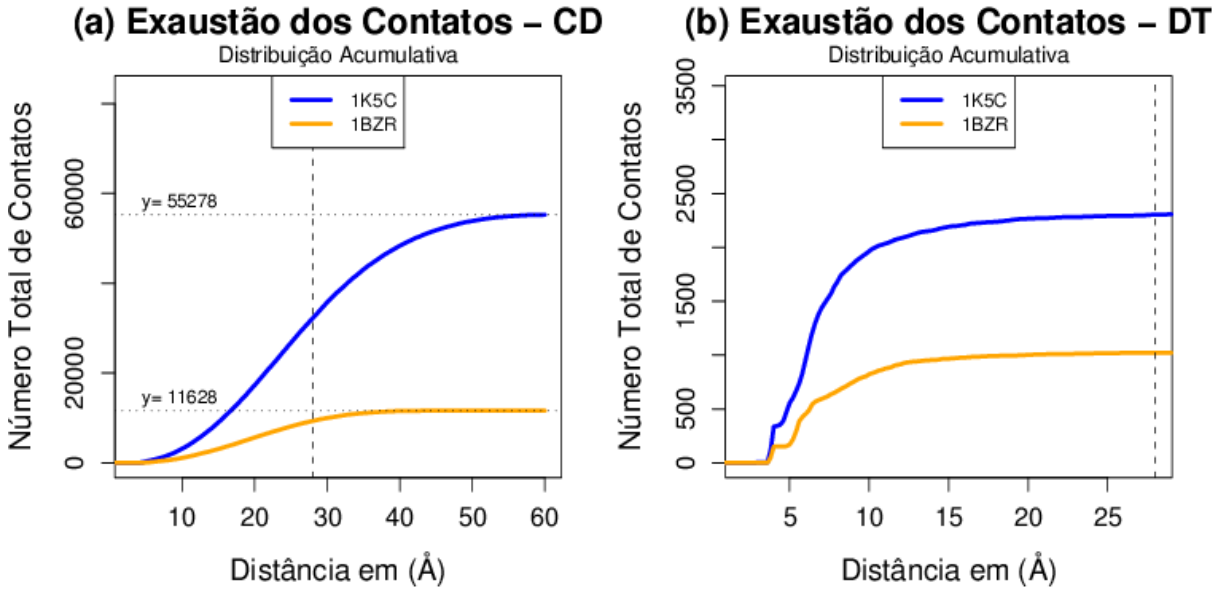


Figura 11: Efeito da exaustão na capacidade de contribuir com contatos conforme o tamanho da proteína (com representação de resíduos usando GC). A 1K5C [150] é uma endopolygalacturonase com 333 resíduos. A 1BZR[148] é uma mioglobina com 153 resíduos. Como a mioglobina é menor, ela passa a contribuir menos para a contagem dos contatos na medida que a varredura de distâncias aumenta. A linha tracejada destaca o delimitador a 28,0 Å. (**a**) O efeito diferencial da exaustão na metodologia CD. (**b**) O efeito diferencial da exaustão na metodologia DT.

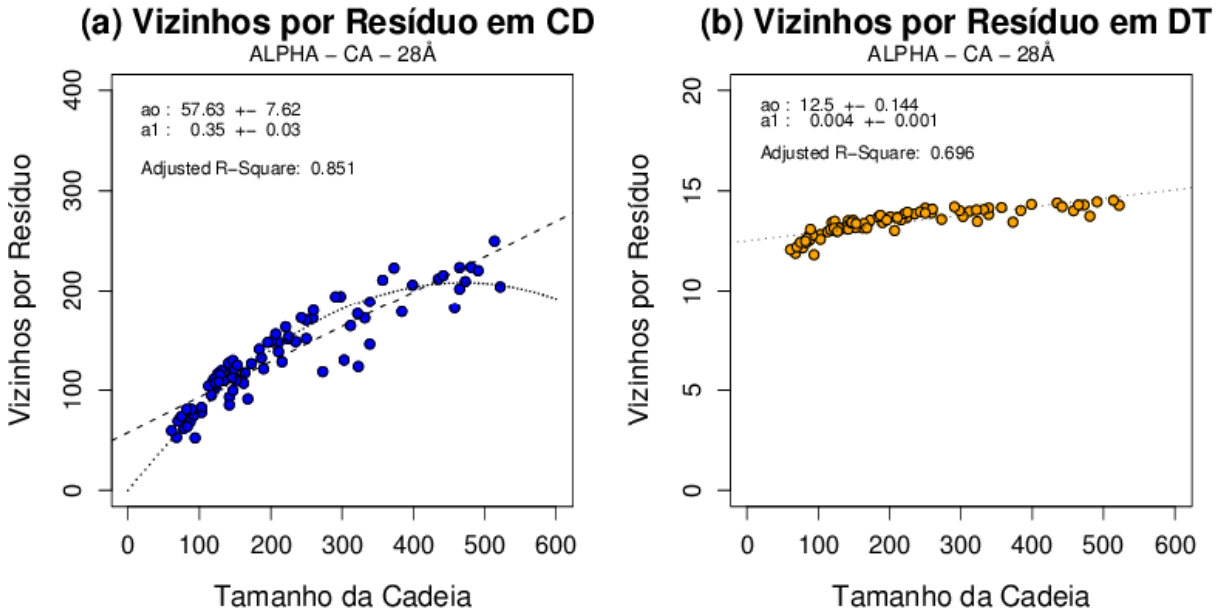


Figura 12: Total de vizinhos normalizados pelo tamanho da cadeia com limite de 28 Å para a maior distância inter-resíduo, num comparativo entre as metodologias CD e DT para conjuntos ALPHA com representação de resíduos por CA. (**a**) Vizinhos normalizados por resíduos em CD. Os números no canto esquerdo indicam os valores dos coeficientes e erros padrões, ao nível de confidência de 0,95, para o ajuste quadrático, juntamente com o coeficiente de determinação. O ajuste esperado a uma reta não é bom. (**b**) Vizinhos normalizados por resíduos em DT. Também o ajuste esperado a uma reta paralela às abscissas não é satisfatório.

Dito isto fica fácil de ver que o número médio de vizinhos por resíduos em distâncias maiores poderá sofrer um viés estatístico induzido pelo perfil do tamanho das proteínas na base de dados. Conforme bem observou Furuichi & Koehl[103], isto naturalmente impõe um limite superior ao valor ótimo do delimitador de distância a ser usado. Em outras palavras, um delimitador ideal não deve ultrapassar o raio médio das menores proteínas presentes no banco de dados.

Para termos uma idéia da dimensionalidade em Å de nossas proteínas, nós estimamos seu raio médio aproximando o seu volume à esferas perfeitas. Isto retornou as seguintes estatísticas gerais, juntando ALPHA e BETA num conjunto só: modal em 16 Å, média e mediana de 18 Å, min/max de 11 Å e 26 Å, respectivamente. Certamente que estes dados nos mostram que nosso delimitador inicial em 28 Å é grande demais para ser confiável. Logo, as estatísticas sobre nossa base de dados sugerem um limite superior para distâncias não maior que 11 Å, o raio estimado de nossa menor proteína.

Mas, o método DT não é de todo livre de problemas. É sabido que DT não é robusto a ruídos na localização dos pontos. Há situações em que pequenos movimentos nos centróides podem levar a diferenças substanciais no arranjo de seus *simplices* (Figura 13). DT, em 3D, requer para uma completa decomposição por poliedros que todos os pontos estejam numa condição geométrica geral, isto é, que não existam cinco sites co-esféricos[29]. Isto por quê cinco pontos numa esfera admitem cinco poliedros que concomitantemente satisfazem o critério de Delaunay da esfera vazia. Sites cujas coordenadas estejam próximos a esta condição anômala ao método DT são sensíveis à bruscas mudanças em seu padrão de arestas. Perceba que a técnica CD também compartilha esse problema, mas somente para sites que estejam na região limítrofe do delimitador. Para todos os demais, CD é robusto à ruídos na localização de seus pontos.

29 - O mesmo vale para os *simplices* em dimensões menores: quatro pontos co-circulares e três pontos co-lineares (onde o círculo comum pode ser visto como tendo raio infinito).

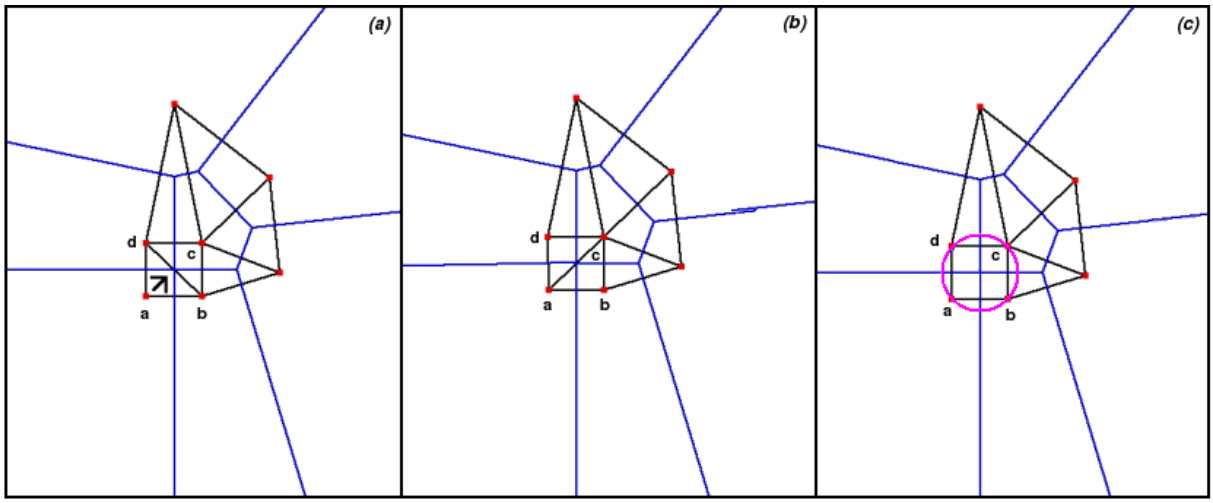


Figura 13: Ilustração de como ruídos nas posições dos sites podem mudar o perfil de arestas Delaunay. (**a-b**) Quando quatro pontos estão próximos de uma condição co-circular, mínimas mudanças em suas localizações podem alterar o padrão das arestas. Um pequeno movimento do site *a* em direção a *c*, por exemplo, mudou a aresta {*b,d*} para {*a,c*} (**c**) Se quatro pontos são **exatamente** co-circulares, a face Delaunay não pode ser um triângulo e o vértice Voronoi associado terá grau quatro. Esta é uma situação degenerada[151] e cada algoritmo lida com ela à sua maneira (se é que chega a ser tratado). O que se faz, usualmente, é submeter os pontos à pequenas pertubações aleatórias de modo a tentar recondicioná-lo na condição geral[152]. Figuras adaptadas a partir de VoroGlide[119]

Um real exemplo em proteínas pode ilustrar o quão sério pode ser essa anomalia DT. A Figura 14 mostra em amarelo todos os vizinhos encontrados por DT ao redor do resíduo ILE-167 (em azul claro) da *all-beta* endopolygalacturonase (1K5C)[150], resolvida com resolução de 0,96 Å. Todos os resíduos estão representados por seus CAs. DT corretamente identifica dez vizinhos imediatos à ILE-167, mas ignora solenemente outros dois legítimos contatos: ASN 188 e CYS 190, ambos em violeta. Isto aconteceu porque os quatro resíduos {CYS-166, ASN-188, GLN-189, ILE-167} estão numa condição próxima ao estado degenerado ao algoritmo, i. e., eles estão quase co-esféricos. E como o par {CYS-166, GLN-189} tem sites que estão mais próximos entre si que o par {ILE-167, ASN-188}, DT obedientemente traça uma aresta com sites do primeiro par e ignora os do segundo. Há uma clara simetria na vizinhança de ILE-167 que DT parece não ser capaz de processá-la corretamente. Pior, trata-se de um erro sistemático, com DT tendendo a omitir para muitos resíduos numa fita, dois ou mais contatos genuínos com as fitas vizinhas. Por exemplo, a aresta {THR-136, VAL-168} foi preterida por uma diferença de 0,14 Å com a {ILE-167, ILE-135}. Fica claro que essa técnica não é assim tão livre de ambigüidade na apuração dos contatos como vem sendo alardeado na literatura[123],[116].

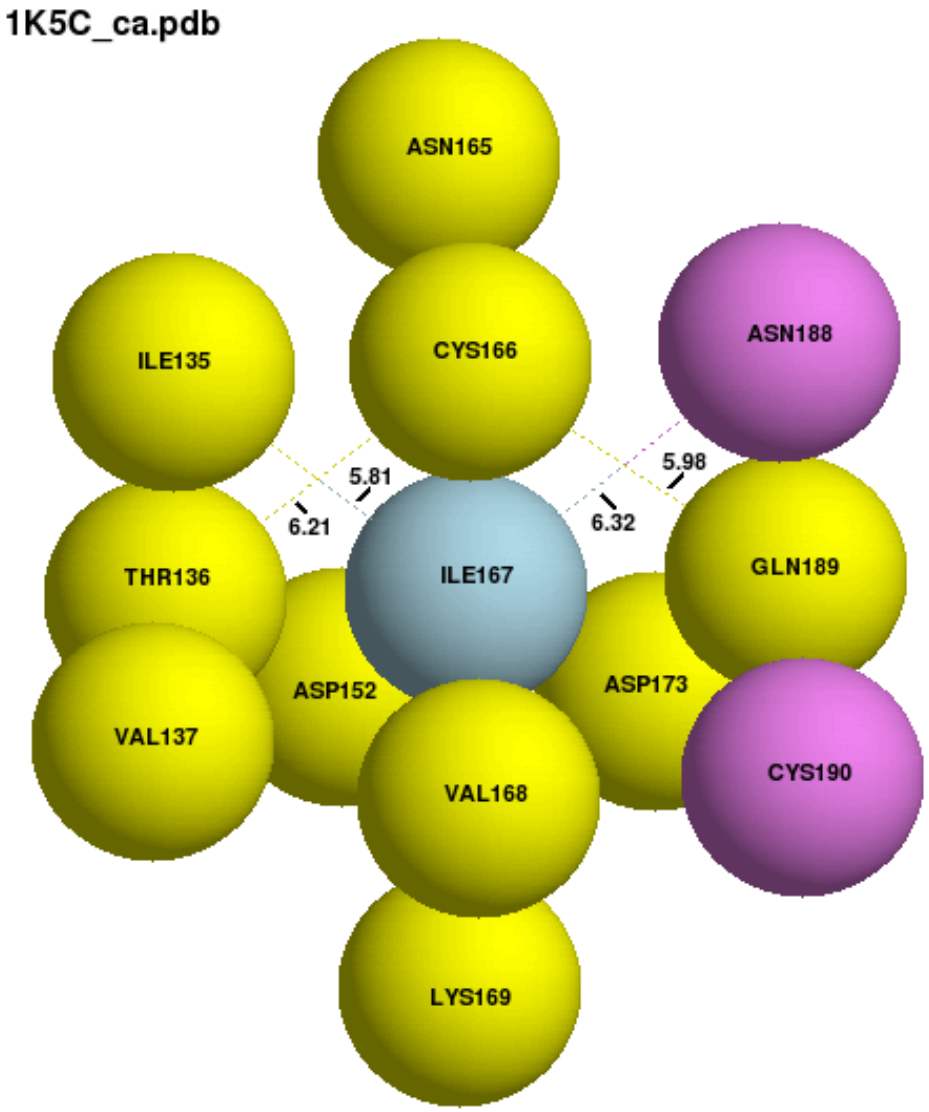


*Figura 14: E*xemplificação da anomalia DT em proteínas. Vemos acima a vizinhança imediata construída por DT sobre o resíduo ILE-167 (azul claro) da proteína *all-beta* Endopolygalacturonase 1K5C[150]. A representação dos resíduos usada foi CA, com destaque em modelos CPK. Em amarelo, vemos dez vizinhos corretamente caracterizados por DT e que têm uma aresta associada à ILE-167. Em violeta, estão dois legítimos vizinhos de ILE-167 que DT não foi capaz de reconhecer: ASN-188 e CYS-190. As linhas tracejadas e respectivos números indicam as distâncias em Å.

Bandyopadhyay & Snoeyink[153] vem tentando atacar esse problema com uma nova metodologia denominada de Decomposição quase-Delaunay (**AD** - ***A**lmost-**D**elaunay Tessellation*). Dado um conjunto de sites S em R^3, $Q \subset S$ pontos irá compreender um conjunto de quase-Delaunay simplices $AD(\epsilon)$ se e somente se perturbando cada site de S até um limite ϵ, o alterado Q atende ao critério Delaunay da esfera vazia. Veja Figura 15 para mais

detalhes. De fato, AD parece realmente ser capaz de identificar arestas omitidas por DT para pontos próximos da condição degenerada. Ele conseguiu detectar corretamente, por exemplo, o par de contatos {ILE-167, ASP-188} e {ILE-167-CYS-190} citados no caso da anomalia DT para a endopolygalacturonase (1K5C). Nós iremos retornar a essa questão em breve, mas antes nós temos que analisar como DT se relaciona com CD.

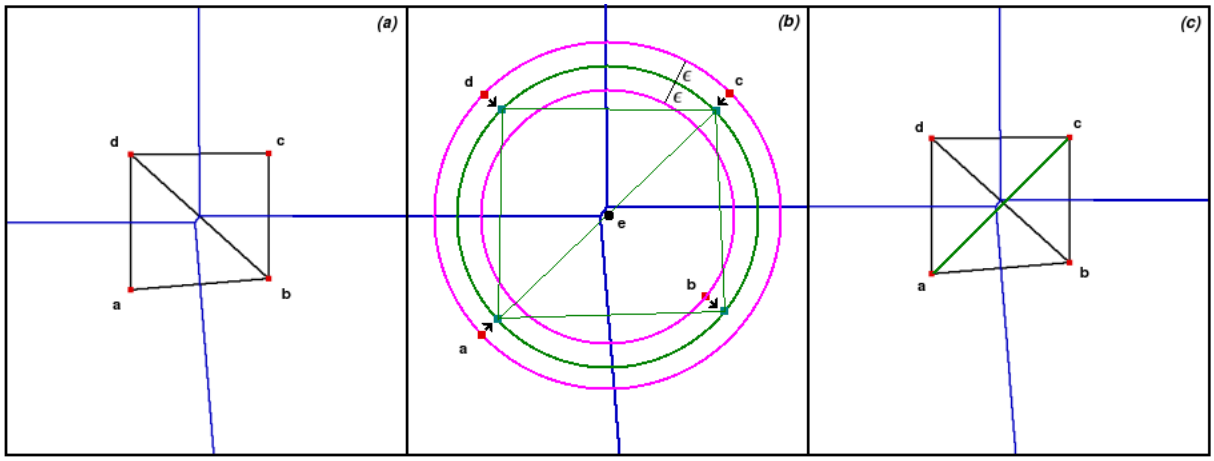


Figura 15: aplicação da técnica de decomposição quase-Delaunay (AD) a quatro pontos próximos do estado degenerado num espaço Euclideano 2d. A computação dos *simplices* AD envolve sua redução ao problema do anel de largura mínima (*minimum-width annulus*)[154] (**a**) Quatro sites próximos ao estado degenerado, montando um paralelo com a situação mostrada na figura anterior. Os simplices {*a,d,c*} e {*a,b,c*} são candidatos a AD. (**b**) Começa-se por traçar um anel de largura 2ε de tal forma que o círculo interno com centro em *e* esteja vazio e toque ao menos um site, e o círculo mais externo contenha os três sites candidatos. Se ao perturbar todos os sites com um deslocamento ε, o critério Delaunay do círculo vazio for atendido, então a largura 2ε é mínima e os *simplices* {*a,d,c*} e {*a,b,c*} são AD. (**c**) Com a inclusão das arestas validadas por AD(ε), nós podemos perceber agora que os sites *a,b,c,d* estão completamente conectados entre si. Obviamente que este constructo deixou de ser uma tesselação de Delaunay. Figuras geradas a partir de VoroGlide[119]

5.3 Confrontando CD e DT

Agora que conhecemos algumas propriedades e idiossincrasias de ambas as metodologias, é hora de compararmos como CD e DT apuram os contatos dos conjuntos ALPHA e BETA em função das distâncias, nas representações de resíduos CA e GC.

Figura 16 mostra a sobreposição das distribuições de densidade em CD e DT para o número médio de vizinhos no intervalo de distâncias entre 0,0 e 28,0 Å, para a base de dados ALPHA, com centróides CA e GC. Vemos que na distância de 7,0 Å, as curvas para CD e DT bifurcam-se. Até 7,0 Å as duas distribuições são em essência as mesmas, independente dos

centróides usados na representação dos resíduos. Acima de 7,0 Å, CD explode com grande variabilidade, certamente devido a diversidade do tamanho das cadeias presente nas bases de dados. Perceba que essa forte variância está ocorrendo mesmo com o número de vizinhos normalizados pelo tamanho da cadeia. Isto indica a presença de processos combinatórios na enumeração das arestas que parecem ser sensíveis ao número de sites. É importante lembrar que o número de contatos por distância cresce de acordo com uma polinomial de alto grau.

Dados equivalentes do conjunto BETA são mostrados na Figura 17. Embora aqui o ponto de bifurcação apareça também em torno de 7,0 Å, o início da divergência pode ter ocorrido antes, perto de 6,2 Å. Note que no intervalo entre 6,2 e 7,2 Å, a despeito das diferenças, ambas as curvas CD e DT ainda estão correlacionadas, com DT computando um número de vizinhos ligeiramente menor que CD. É possível que essa antecipação (de 7,0 Å para 6,2 Å) tenha ocorrido como conseqüência da anomalia Delaunay já descrita e exemplificada nas Figuras 13 e 14, que deixa de considerar contatos legítimos entre sites próximos da condição degenerada do algoritmo. Contribui para isso a topologia quase planar das fitas betas que tornam a posição dos CAs muito próximas de um plano. Os dados da Figura 17 nos permitem, porém, estimar o grau dessa anomalia pela diferença entre as áreas das curvas CD e DT até 7.0 Å. Os erros relativos foram os seguintes: 5,1 ± 0,3% para BETA CA, 0,6 ± 0,2% para BETA GC, 1,9 ± 0,3% para ALPHA CA e 0,3 ± 0,2% para ALPHA GC, no nível de confidência a 0,95. Veja que a falha é mais evidente na representação dos resíduos por CA. A despeito dos valores serem relativamente baixos, é importante reforçar que ao menos em estruturas “beta” temos indícios de que essa falha não é aleatória, mas ocorre de forma sistemática.

Na nossa opinião, o fato dessas duas distribuições serem equivalentes até 7,0 Å tem duas importantes conseqüências. A primeira é que ela unifica elegantemente as propriedades das duas metodologias CD e DT numa só: todas as arestas até 7,0 Å serão completas, dadas pela enumeração combinatória de todos os contatos que possam existir entre os sites inclusos na esfera de busca (propriedade CD); mas também, todas as arestas vão compor contatos legítimos, com uma garantia geométrica de conexão envolvendo apenas sites não oclusos (propriedade DT).

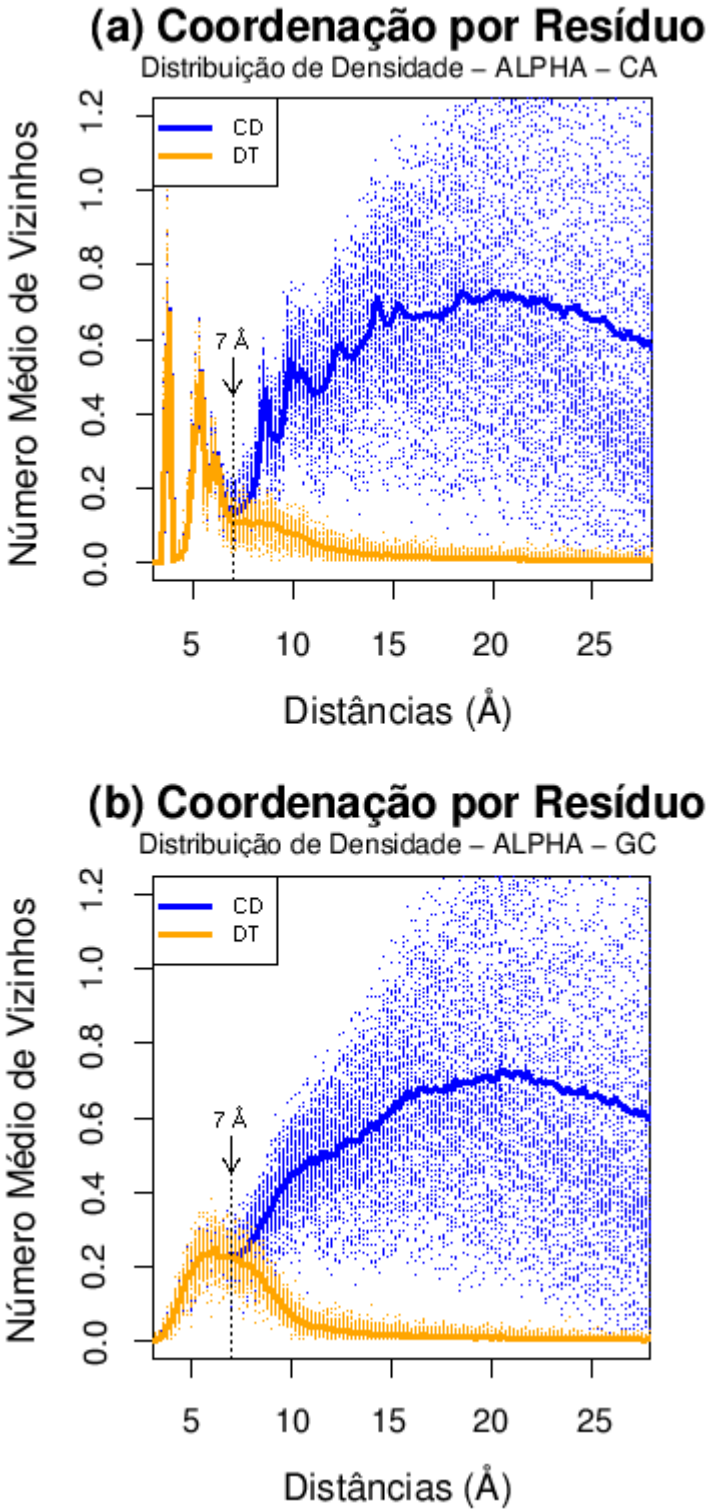


Figura 16: Comparação das técnicas CD e DT para as distribuições de densidade do número médio de vizinhos em função das distâncias para ALPHA. Linhas grossas indicam os pontos médios. 7,0 Å assinala o ponto de bifurcação. (**a**) Perfil para representação dos resíduos por CA. (**b**) Perfil para representação dos resíduos por GC.

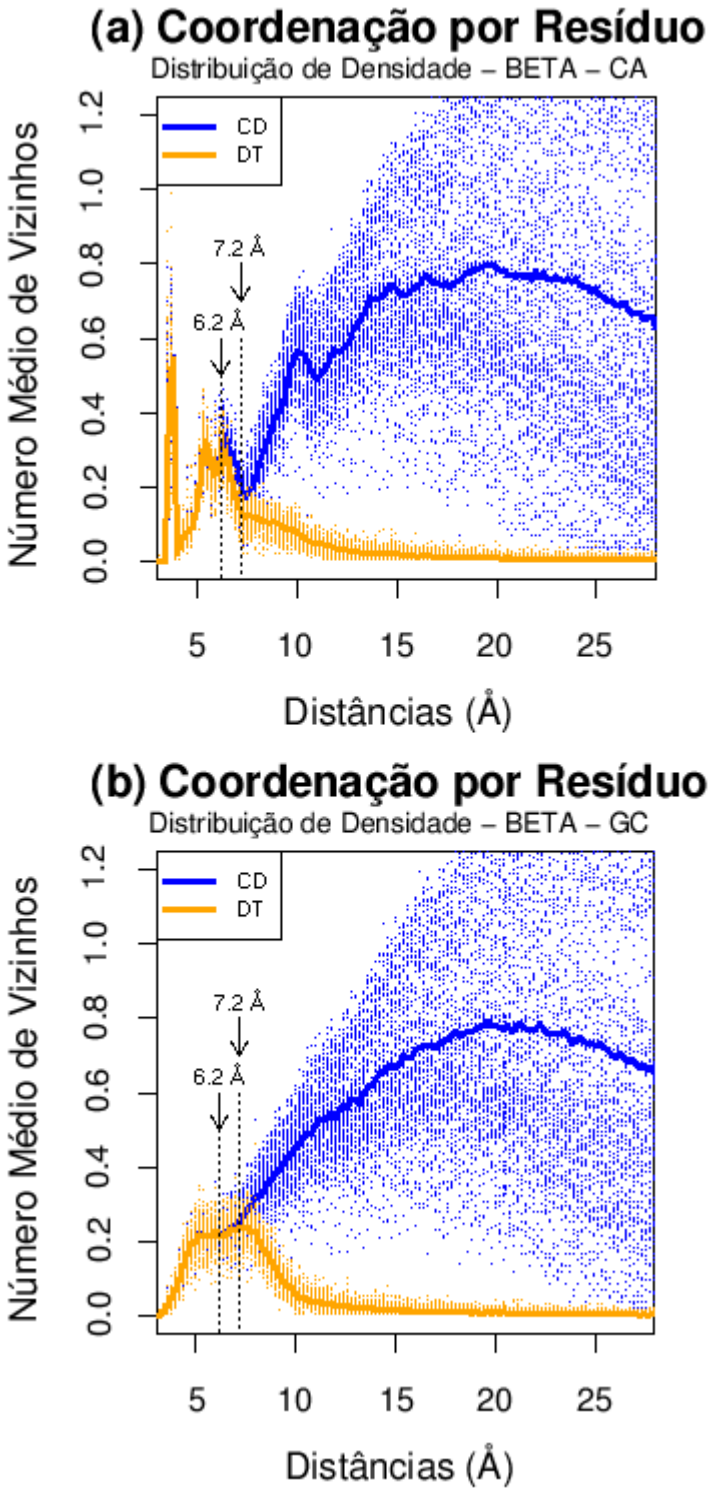


Figura 17: Comparação das técnicas CD e DT para as distribuições de densidade do número médio de vizinhos em função das distâncias para BETA. Linhas grossas indicam os pontos médios. O início da divergência é antecipado para 6,2 Å devido a anomalia em DT. A bifurcação acontece entre 6,2 e 7,2 Å (**a**) Perfil para representação dos resíduos por CA. (**b**) Perfil para representação dos resíduos por GC.

O fato dos resultados exibidos nas figuras 16 e 17 terem sido independentes tanto das classes (*all-alpha* e *all-beta*) quanto das representações (CA e GC) exploradas aqui, faz de 7.0 Å um potencial candidato a um limite de distância inferior de carácter geral, a ser usado como um delimitador referencial em estudos de contatos de proteínas. Enquanto esse limite inferior é virtualmente independente do tamanho das proteínas presentes no banco de dados, o limite superior (conforme nós vimos) o é. Logo, para as bases de dados montadas aqui, o delimitador ideal deveria estar entre 7,0 Å e 11,0 Å. No entanto, em 7,0 Å as garantias de não-oclusão entre os sites em contato serão maiores.

A segunda conseqüência é que, até o ponto de bifurcação, CD deveria herdar também o comportamento linear de DT. Como CD é quadrático por natureza, seria esperado que ele passasse por uma transição, de um modelo parabólico para linear próximo de 7,0 Å. A fim de checar essa possibilidade, nós aplicamos um teste de seleção de modelos usando o Critério de Informação Bayesiana ou **BIC (*B**ayesian **I**nformation **C**riterion*)[156], para avaliar quais dos comportamentos (linear ou quadrático) melhor se ajustam aos dados de CD para cada intervalo de distâncias. BIC retorna um número que mede a qualidade do ajuste dos dados ao modelo. Quanto menor for esse número, melhor é a avaliação do modelo. Mas, como os valores absolutos de BIC não têm significado em si, o usual é calcular a diferença entre os diversos modelos em avaliação. Para nosso teste, nós computamos valor D = BIC(linear) – BIC(quadrático). Um D positivo irá indicar que o modelo quadrático é superior; um D negativo, que o linear é superior. Burnham & Anderson[157], como uma máxima heurística (*rule of thumb*), sugerem $D \leq 10$ como um limite superior na avaliação do mérito dos modelos, mas recomendando $D \leq 2$ como um intervalo mais seguro e ideal. Em outras palavras, quanto mais próximo D é de zero mais indistintos são os modelos, e neste caso uma possibilidade é ficar com aquele que opera com menos parâmetros (princípio da navalha de Occam)[30]. Figura 18 apresenta os resultados da aplicação de BIC aos nossos dados. Nós podemos ver que nas distâncias mais longas o melhor modelo é indiscutivelmente quadrático. Mas, na medida que o delimitador regride, essa propriedade quadrática vai cedendo lugar a um modelo linear. A 7,0 Å, conforme previsto, todas as curvas (independente se ALPHA ou BETA, se CA ou GC) estão ou muito próximas ou abaixo de zero.

30 - Sei que esse princípio é polêmico. A Wikipedia inglesa tem um excelente artigo [155] com uma ampla discussão sobre o tema. E me fez até relembrar a máxima de Albert Einstein(1879-1955): "*Everything should be made as simple as possible, but not simpler.*". Seja como for, para os propósitos dessa tese o princípio parece justificável, já que não usei um modelo formal para ajudar na decisão da questão. Para ser sincero, não usei porque não encontrei nada convincente na busca (não muito profunda) que fiz sobre o problema.

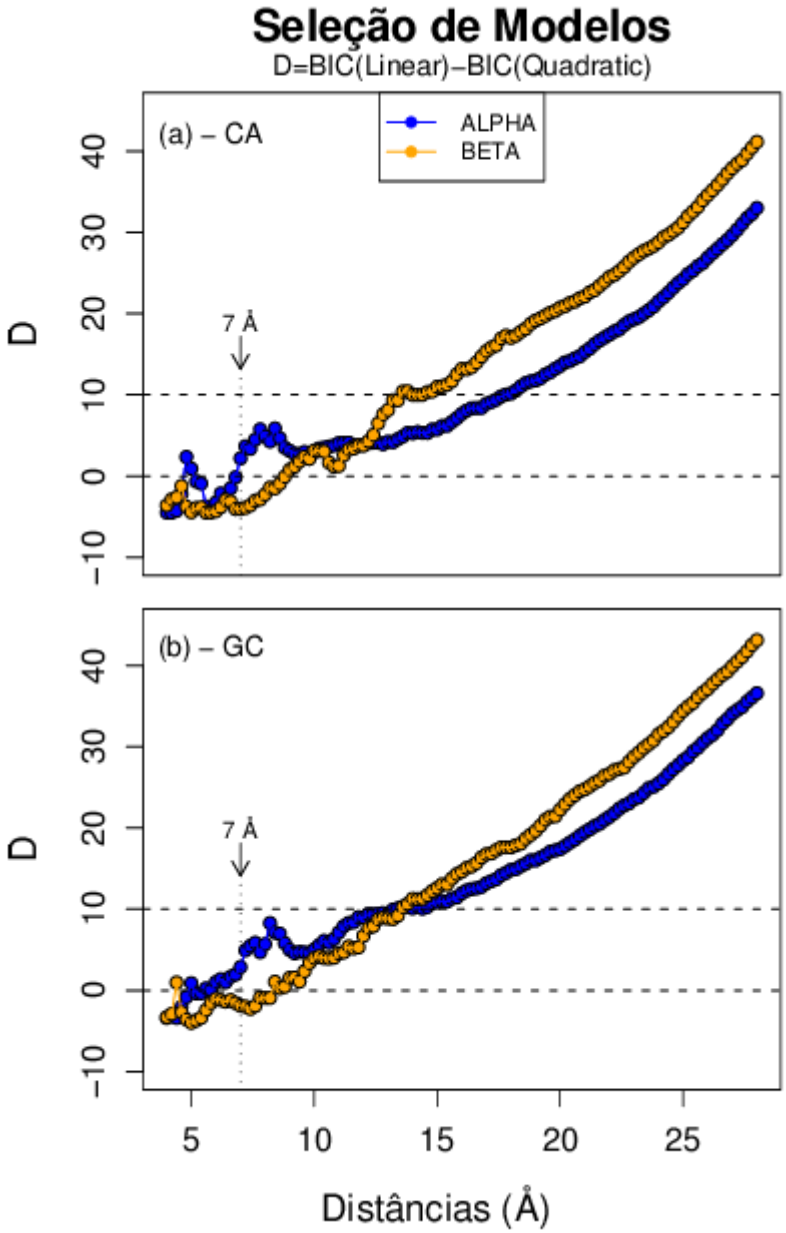


Figura 18: Teste de seleção de modelos lineares contra quadráticos usando *Bayesian Information Criterion* (BIC)[156] para avaliação das distribuições em cada intervalo de distâncias do delimitador. A ordenada contém a diferença entre os números BIC, *D* = BIC(Linear) – BIC(Quadrático). Quanto menor o valor retornado pela função BIC(.) mais adequado é o modelo. Logo, um *D* positivo favorece o modelo quadrático; negativo, o modelo linear. (**a**) Teste de seleção de modelo para representação dos resíduos por CA. Quando o delimitador aceita longas distâncias, a versão quadrática é favorecida. Na medida que essas distâncias decrescem, o sistema vai se tornando cada vez mais linear. Em 7,0 Å, *D* é próximo de zero em ambos conjuntos ALPHA e BETA. (**b**) O mesmo em (a) aplicado à representação por GC. Novamente o comportamento é similar à representação por CA.

Podemos ainda fazer algumas observações curiosas. Em distâncias mais longas, ambas as curvas parecem ter a mesma taxa de decaimento, com ALPHA estando ligeiramente mais próximo da linearidade que BETA. Em algum ponto em torno de 20,0 Å, ALPHA desacelera sua queda, enquanto BETA segue adiante sem bruscas alterações. O resultado é que BETA acaba atingindo D = 0 antes de ALPHA. Não nos parece trivial explicar esse estranho comportamento, que parece tocar diferenças topológicas sutis na distribuição de resíduos em ALPHA e BETA. Pode ser que seja um reflexo do padrão mais regular do empacotamento em estruturas "beta", se é que essas diferenças de comportamento são estatisticamente significantes. Essa é uma questão que merece ser estudada mais aprofundadamente.

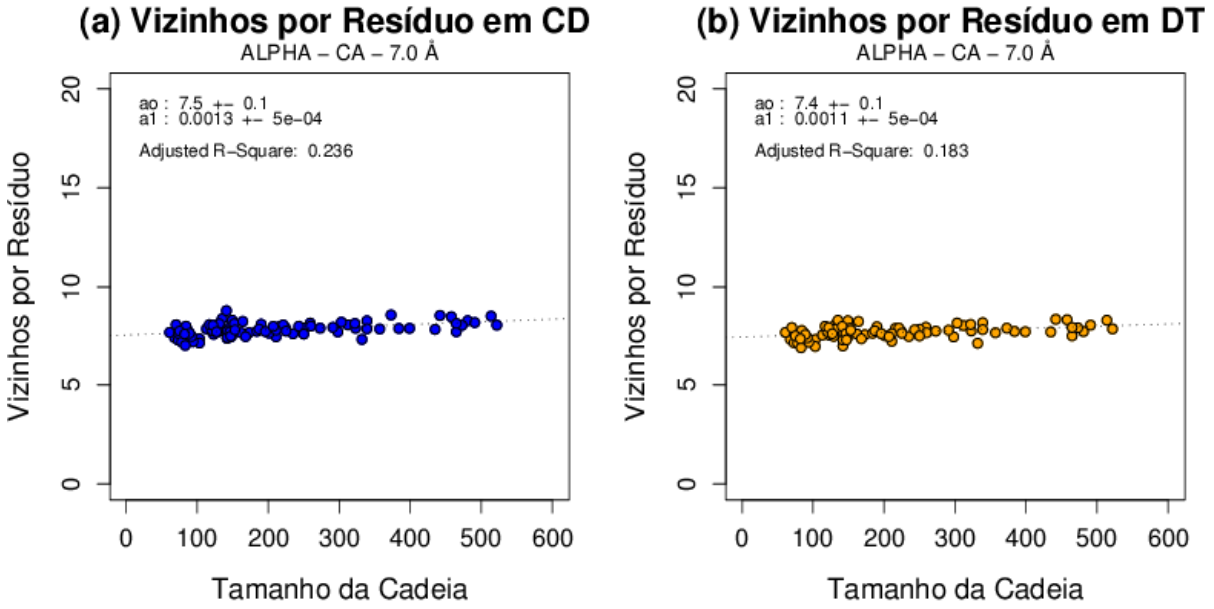


Figura 19: Total de vizinhos normalizados pelo tamanho da cadeia com limite de 7,0 Å para a maior distância inter-resíduo, num comparativo entre as metodologias CD e DT para conjuntos ALPHA com representação de resíduos por CA. (**a**) Vizinhos normalizados por resíduos em CD. Os números no canto esquerdo indicam os valores dos coeficientes e erros padrões, ao nível de confidência de 0,95, para o ajuste quadrático, juntamente com o coeficiente de determinação. Vemos agora que CD perdeu sua característica quadrática. (**b**) Vizinhos normalizados por resíduos em DT. As cadeias com menos de 100 resíduos parecem agora integradas aos demais pontos.

Conforme observado por Bandyopadhyay & Snoeyink (2004) através do lema 4.1[153], este comportamento linear pode ter raiz no empacotamento razoavelmente uniforme que se verifica em proteínas. Nós acreditamos que isto constitui, na verdade, uma assinatura topológica da primeira camada de coordenação, envolvendo contatos primários entre sites não oclusos. Dessa forma, o delimitador em 7,0 Å pode ser visto também como a distância ideal

onde a primeira-ordem de contatos em proteínas está otimamente separada das demais ordens. Como, neste momento, nós não podemos fornecer uma prova formal dessa proposição, nós iremos conjecturar sua verdade, com base em todas as evidências empíricas relatadas até aqui.

Na Figura 19 nós vemos o mesmo tipo de gráfico da Figura 12, mas agora com um delimitador de 7,0 Å. Conforme esperado, os gráficos de CD e DT ficam bem similares. Note que há uma perda de correlação entre o tamanho da cadeia e o número médio de vizinhos, ainda que a inclinação não seja exatamente zero (ver nota de rodapé 28). Também não temos mais problemas com as proteínas pequenas, que agora parecem integradas à regressão.

5.4 É quase-Delaunay (AD) uma solução?

Agora nós estamos em condições de avaliar a solução de Bandyopadhyay & Snoeyink (2004), proposta para contornar a anomalia DT que ocorre em sites próximos da condição degenerada. Na Figura 20 é mostrado um gráfico de AD (usando parâmetro de perturbação de 2,0 Å e delimitador de tamanho de aresta de 28,0 Å) contra CD e DT, para o conjunto BETA com CA. Muito curiosamente nós percebemos que AD comporta-se como se fosse a diferença entre CD e DT, até uma certa distância que é dependente da perturbação utilizada. Isso significa que, na medida que a perturbação cresce, AD+DT tende à CD. Se isso é verdade, não há benefício aparente em preferir a metodologia AD+DT em troca de CD. Logo, AD+DT não parece uma solução vantajosa frente a CD.

Entretanto, um fator positivo da metodologia AD é que com ela nós podemos inferir com mais precisão o ponto onde DT inicia sua divergência de CD. Se considerarmos esse ponto como aquele onde metade das proteínas tem ao menos uma aresta tipo AD, nós podemos estimá-lo em: 6,2 Å para ALPHA CA; 7,0 Å para ALPHA GC; 6,2 Å para BETA CA; 6,8 Å para BETA GC. Isso confirma um fato já observado antes: que a anomalia DT ocorre principalmente na representação de resíduos por CA.

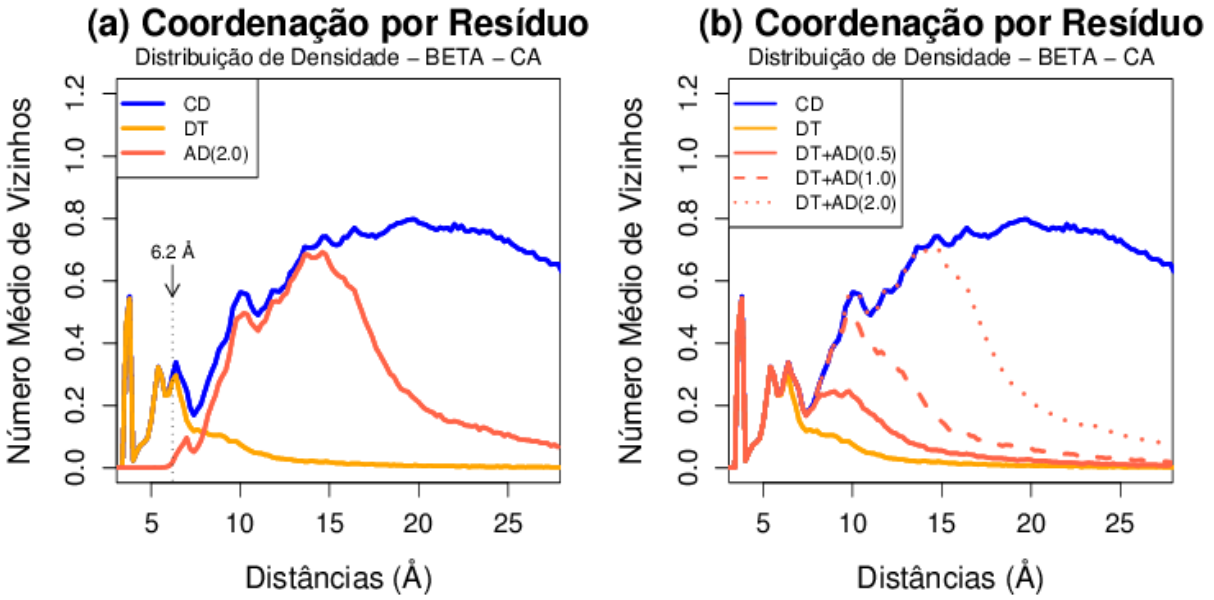


Figura 20: Comparação das curvas representando o número médio de resíduos em função da distância entre as metodologias quase-Delaunay (AD em vermelho), delimitador dependente (CD em azul) e decomposição de Delaunay (DT em laranja). Em todas as curvas AD o delimitador de tamanha da aresta (*prune parameter*) usado foi de 28,0 Å. Dados produzidos a partir do conjunto BETA com representação dos resíduos por CA. Os números entre parênteses para AD na legenda indicam o valor máximo do parâmetro de perturbação (ε) testados. (a) Comparativo de AD(ε = 2) contra CD e DT. Percebe-se que AD é um complemento de DT em relação a CD, pelo menos até 15,0 Å, e que ele é diferente de zero por volta de 6,5 Å. (b) Gráfico com CD, DT e a soma de DT + AD com ε de 0,5 Å (linha cheia em vermelho), 1,0 Å (linha tracejada em vermelho) e 2,0 Å (linha pontilhada em vermelho). Vê-se claramente que DT+AD tende a produzir resultados bem similares a CD na medida que o parâmetro de perturbação cresce.

5.5 Um Estudo de Caso

Por fim, nós implementamos um estudo de caso para analisar a influência da adoção de centróides CA e GC no padrão da vizinhança média dos resíduos. Agora que já demonstramos que CD e DT estão unificados até 7,0 Å, nós podemos usar um tanto tranqüilamente a primeira, por ser mais simples que a segunda.

A Figura 21 mostra a distribuição acumulativa para o número médio de vizinhos em função da distância, comparando os conjuntos ALPHA e BETA. Para centróides CA (Figura 21a) nós podemos distinguir ao menos 3 regiões, separadas pelos pontos de distâncias em 5,2 Å e 6,8 Å. Curioso notar que 6,8 Å parece ser um marco na mudança do comportamento da variância, visualmente claro pelo menor espalhamento dos pontos amostrais antes e depois desta distância referencial. Acreditamos que essa menor variabilidade até 6,8 reflete o

comportamento mais organizado (e restritivo) que a cadeia principal impõe à representação CA. Para a representação GC (Figura 21b), o padrão de variância é mais homogêneo, talvez porque a cadeia lateral espelhe seus maiores graus de liberdade no cálculo do seu centro geométrico. Apesar disso, também visualizamos um ponto de convergência em 6,8 Å.

Em ordem de termos uma idéia mais estatisticamente confiável das diferenças do padrão de vizinhos em ALPHA e BETA, nós decidimos por explorar a homogeneidade das médias/medianas ao longo das distâncias. Para isso, aplicamos testes paramétricos e não-paramétricos na avaliação das diferenças das médias/medianas, a saber: *Student-t* com correção Welsh para variâncias desiguais[158] e Wilcoxon *rank sum test*[132], respectivamente. Essa redundância com um teste não-paramétrico teve como objetivo assegurar os resultados retornados pelo teste-t, que é sabido requerer normalidade como pré-requisito à confiabilidade de suas estimativas, normalidade esta que há como nós garantirmos. Entretanto, é notável que os dois testes tenham produzido curvas correlatas para a maioria das regiões. Em ambos um *p-value* baixo indica que as diferenças no número médio de vizinhos em ALPHA e BETA são significativas, rejeitando ou colocando em dúvida a hipótese nula de homogeneidade. Para representação CA, na Figura 22a, vemos dois picos agudos sobre os valores de distância 5,2 Å e 6,8 Å, os mesmos pontos de interseção da Figura 21a, embora Wilcoxon pareça não concordar com o teste-t no pico em 5,2 Å. Com a representação GC, todos os *p-values* tenderam a ficar abaixo do limite arbitrário de significância, exceto numa região com um largo pico em torno de 6,8 Å (Figura 22b).

Olhando conjuntamente os dados das Figuras 21 e 22 nós podemos observar que, entre 5,2 Å e 6,8 Å, os centróides CA e GC parecem não concordar qual dos dois conjuntos, ALPHA ou BETA, tem um maior número médio de vizinhos por resíduo. CA parece introduzir um viés estatístico a favor de ALPHA, e GC a favor de BETA. Se dois diferentes pesquisadores tivessem cada um isoladamente escolhido um tipo de centróide, um CA e outro GC, e tivessem também utilizado delimitadores menores que 6,8 Å, eles poderiam chegar a resultados contraditórios a respeito se ALPHA ou BETA encerram em média mais vizinhos por resíduo.

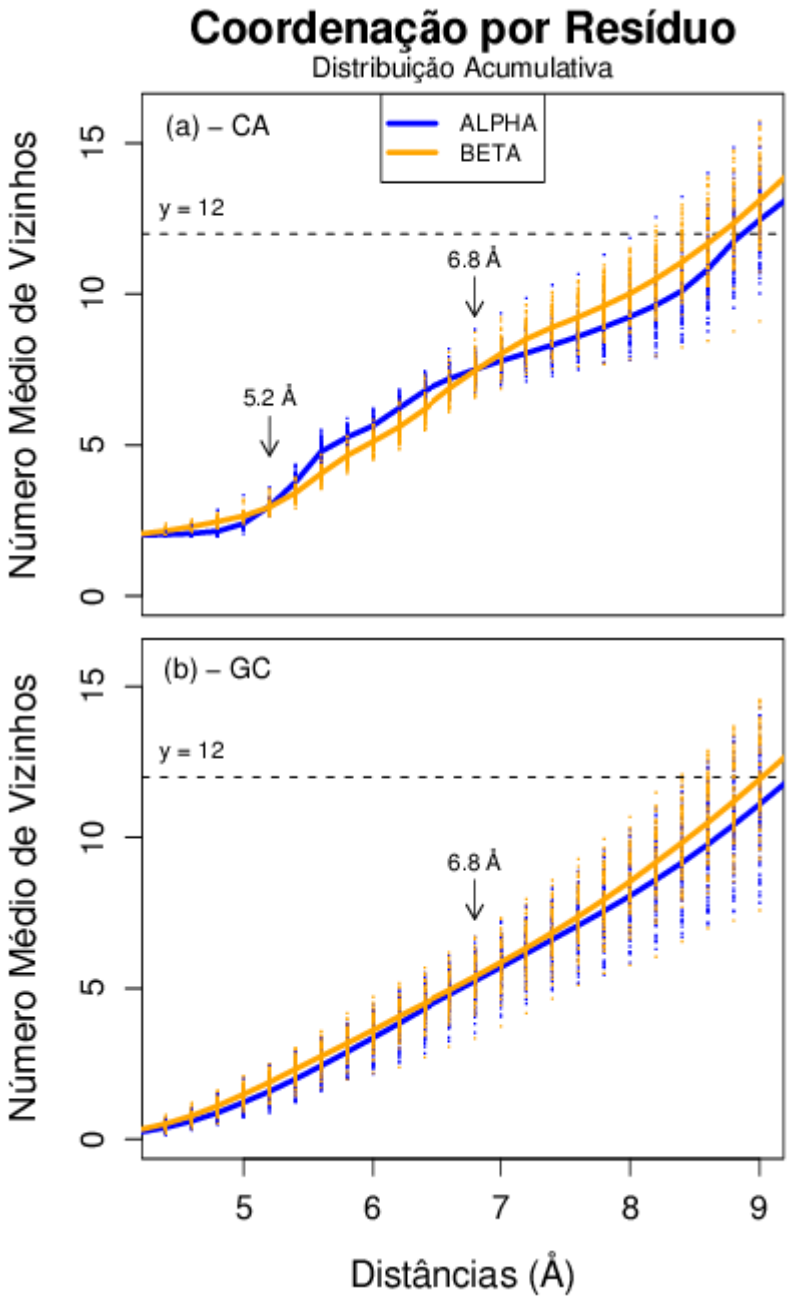


Figura 21: Distribuição acumulativa para o número médio de vizinhos em função das distâncias usando metodologia CD. As linhas grossas indicam a média dos pontos, sendo azul ALPHA e laranja BETA. A linha tracejada marca o ponto onde o número médio de vizinhos alcança o limite de 12, um número característico de vizinhos encontrado no empacotamento máximo de esferas perfeitas em 3d. (**a**) A coordenação média para a representação de resíduos CA. As setas destacam as regiões onde as curvas ALPHA e BETA parecem convergir entre si, compreendendo valores de 5,2 Å e 6,8 Å. (**b**) Os mesmos dados para representação GC, mostrando uma convergência em 6,8 Å, que coincide com um dos pontos de interseção em (a).

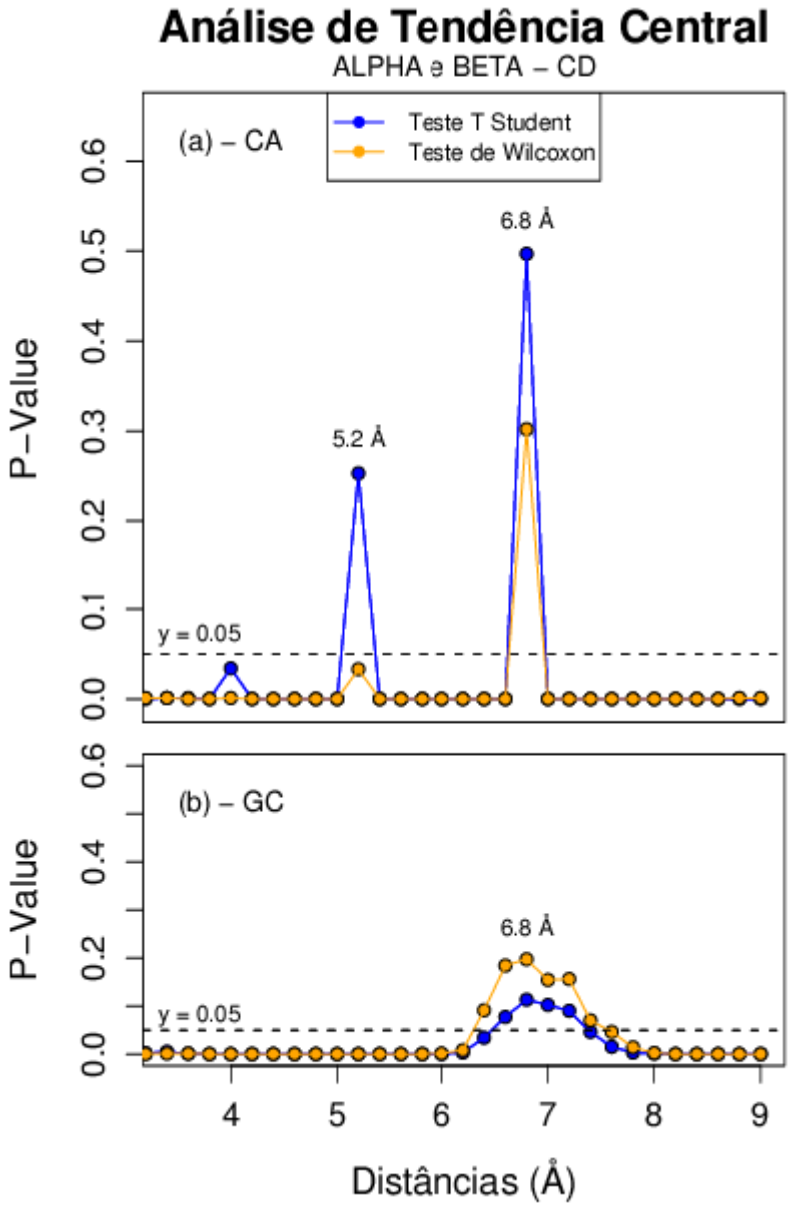


Figura 22: Análise de tendência central para a homogeneidade das médias/medianas entre ALPHA e BETA para o número médio de vizinhos por distância usando metodologia DT, conforme dados da figura anterior. Para essa avaliação foram aplicados os testes: paramétrico *Student t-test* com correção Welsh para diferenças nas variâncias[158], em azul; não-paramétrico Wilcoxon *rank sum test*[132], em laranja. A linha tracejada traça o limite arbitrário para o *p-value* de 0,05, ou nível de confidência a 0,95. Quanto maior o *p-value*, mais indistintos são as médias/medianas. (a) Perfil das homogeneidades para representação de resíduos por CA. Vê-se dois picos agudos em 5,2 Å e 6,8 Å, embora os dois testes não concordem a respeito do primeiro pico. Outras regiões apresentam p-value abaixo do limite de significância. (b) Mesmos dados para representação GC. Há somente um pico brando em torno de 6,8 Å.

O reflexo disso seria sentido nos algoritmos que mapeiam ou manipulam contatos. Por exemplo, se o objetivo fosse fazer inferências a respeito da densidade de empacotamento, aqueles dois pesquisadores estariam em desacordo sobre qual grupo seria mais compacto: se o contendo proteínas ricas em estruturas “alfa” ou em estruturas “beta”. Acima de 6,8 Å, no entanto, esta ambigüidade é minimizada, com ambos os centróides CA e GC ao menos concordando que estruturas “beta” tem um número médio estatisticamente mais elevado de vizinhos por resíduo. Isto constitui, sem dúvida, um argumento adicional em benefício do uso de um delimitador de distâncias em 7,0 Å.

Para concluir, gostaríamos de ressaltar dois pontos que julgamos ser importante no fechamento desse trabalho. Primeiramente, que nós estabelecemos aqui um valor candidato a um **limite inferior** para um delimitador ótimo a ser usado na apuração de contatos em proteínas. Vimos que esse limite inferior é virtualmente independente das dimensões das proteínas presentes no banco de dados. Mas, como o limite superior é dependente do tamanho das proteínas, um delimitador ideal também o será. O ajuste fino de seu valor dependerá de um estudo mais aprofundado sobre o perfil das oclusões, estudo esse já em andamento em nosso grupo de pesquisa[87]. Segundo, que o valor de 7,0 Å é um limite inferior se o objetivo da pesquisa é obter a mais completa enumeração possível dos contatos não oclusos da primeira camada de vizinhos. Se a meta, por outro lado, for mapear apenas as interações de curtas distâncias (como por exemplo, aquelas menores de 4,0 Å), um limite inferior obviamente deixa de ter sentido.

6. Limitações e Perspectivas

Nós iremos comentar aqui as limitações e perspectivas deste trabalho conjuntamente, uma vez que as restrições da primeira podem servir como motivações para a segunda. Acreditamos que os desdobramentos desse trabalho são muitos, oferecendo várias frentes de estudo.

Como o próprio nome sugere, a base de dados é a base de tudo. Montar um banco de dados equilibrado, não viciado e bem desenhado é um grande desafio a praticamente qualquer projeto em Bioinformática. Todas as conclusões feitas nesse trabalho estão circunscritas ao banco de dados que nós montamos. São conseqüências dos critérios de amostragem e filtragem operacionalizados sobre o conjunto universo dado pelo PDB (que em si é também uma pequena amostragem das estruturas presentes nos seres vivos). Outras composições, com diferentes proteínas, devem ser feitas para certificar o quão robusto são os resultados encontrados aqui, se o nosso processo de amostragem e filtragem não teria induzido de forma não intencional algum viés estatístico.

Por exemplo, nós decidimos por respeitar a distribuição assimétrica (*skewed*) do tamanho da cadeia que resultou no banco de dados final. Mas isso pode ter desequilibrado a representatividade das proteínas pequenas e grandes em detrimento daquelas que compõem a região modal. Em um trabalho futuro, seria interessante analisar de forma mais profunda essa influência, estratificando as cadeias por tamanho e dando o mesmo peso em número de proteínas (ou de resíduos) a cada estrato. Seja como for, nós estamos relativamente seguros de que nossa análise de contatos até 7,0 Å é essencialmente independente do tamanho da cadeia (Figura 19).

Neste trabalho, por razões práticas, nós limitamos nossa análise a uma matriz de 2x2x2 dimensões: dois modelos (CD, DT), dois centróides (CA, GC), duas classes de proteínas (*all alpha* e *all beta*). Certamente, há outras variáveis a explorar. Por exemplo, seria interessante ver como ficam os dados em outras classes, como a *alpha-beta* do CATH. Ou mesmo de outros sistemas de classificação, como o SCOP[159], que divide a *alpha-beta* em duas: *alpha/beta* e *alpha+beta*; e o STING[127], que classifica como *alfa* e *beta* aquelas que têm 100% de estruturas em hélices e fitas, respectivamente. Outras formas de representar os resíduos também precisam ser estudadas, como os centróides nos carbonos betas (CB), ou em

granulosidades mais finas, como os contatos inter-resíduos estabelecidos pelas distâncias entre seus átomos mais pesados. Uma análise numa granulosidade totalmente atômica seria um desafio computacional muito bem vindo.

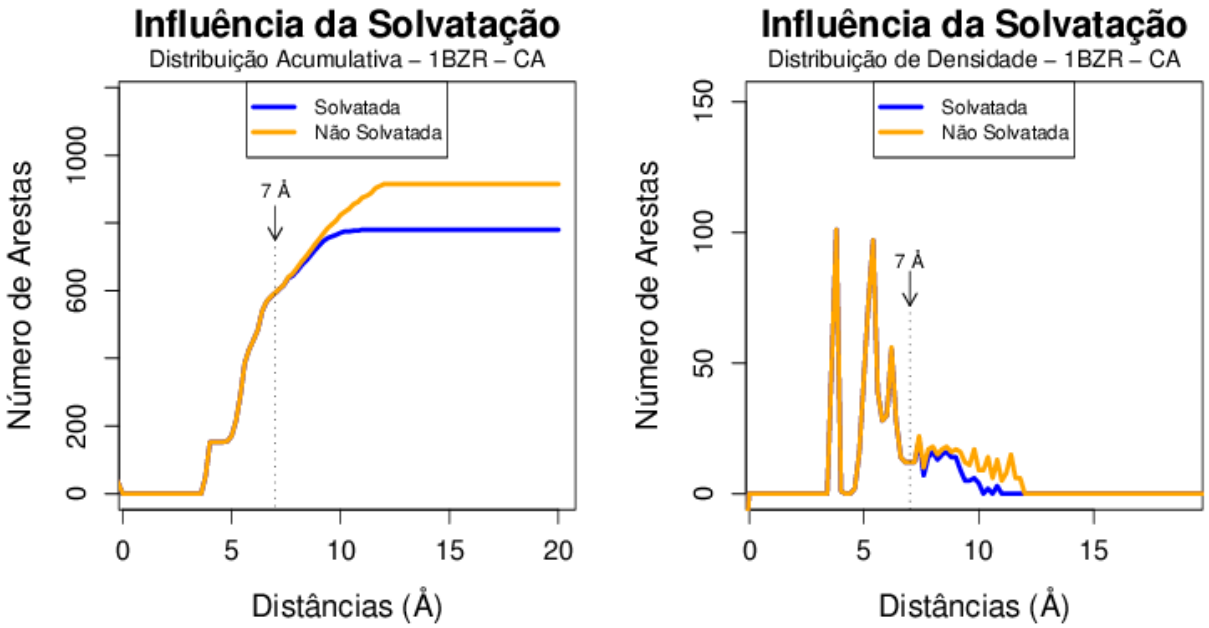


Figura 23: Influência da solvatação no perfil das arestas inter-resíduos em DT para 1BZR com representação CA. Vê-se que até 7.0 Å as distribuições com e sem solvente são idênticas.

Solvatação é outro grande problema, principalmente para a técnica DT. É sabido que sites de superfície podem produzir células de Voronoi abertas, de volume infinito. Isso certamente é um problema para alguns algoritmos, como os que calculam volume e áreas, mas não é um impedimento para a decomposição de Delaunay em poliedros (reveja Figura 5). Mas, em proteínas não solvatadas, os sites de superfície estão livres para compor arestas com os demais na mesma condição, criando um número irreal de vizinhos por resíduo. Nós vimos que isso pode, de fato, enviesar as estatísticas de contatos quando é escolhido um delimitador muito grande (reveja Figuras 11 e 12). Conforme já dito, nós conscientemente optamos neste presente trabalho por não hidratar artificialmente nossas proteínas, para que fosse possível ter uma visão dos contatos com seus dados originais. Aliado a isso, testes preliminares de solvatação feitos com a mioglobina 1BZR, usando o bem elaborado algoritmo de adição de águas do programa VORO3D[160], nos forneceram fortes evidências de que esse processo afeta apenas as arestas com comprimento maior que 7,0 Å (Figura 23). Certamente que tal

fato vem reforçar nossa hipótese de que a distribuição de arestas até 7,0 Å foi gerada principalmente por sites bem empacotados envolvidos na primeira-ordem de contatos, mais internos à proteína, não expostos à superfície. Um estudo mais criterioso sobre o papel do solvente no padrão de contatos já está em andamento em nosso grupo.

Outra questão que permanece aberta é se o número de vizinhos em função da distância pode ser usado como uma forma de inferir a densidade de empacotamento em proteínas. Este número irá depender de maneira complexa não somente da forma de empacotamento, mas também do tamanho e do formato dos sites. Fleming & Richards[77], usando uma métrica de contatos que leva em conta não somente distâncias mas também áreas (OSP - *Occluded Surface Packing*) em um conjunto de 152 proteínas não-homólogas, levantaram indícios de que proteínas ricas em hélices tendem a ter uma densidade de empacotamento maior que ricas em folhas betas. Também mostraram que resíduos aromáticos parecem empacotar-se melhor que alifáticos. Liang & Dill[79] já evidenciaram que proteínas grandes tendem a ser menos compactas que as pequenas. Angelov *et. al*.[161], analisando as propriedades Voronoi de uma coleção de 39 proteínas, demonstraram que há uma tendência de uma correlação positiva entre o número de faces por célula (que equivale ao número de vizinhos) e os volumes Voronoi dos resíduos, embora glicina, alanina, lisina e arginina sejam pontos fora da curva (*outliners*). Conforme esperado, glicina e triptofano foram os extremos, com o primeiro tendo em média 13.36 vizinhos, e o segundo 14.86 vizinhos, uma diferença de 1.50 vizinhos. Como os autores não fornecem o erro padrão, fica difícil julgar se essa diferença é estatisticamente significativa ou não, apesar de parecer em termos absolutos muito pequena. Enquanto Kuntz & Crippen[162] encontraram não-homogeneidades entre a densidade local das cadeias laterais hidrofóbicas e das cadeias principais, Tsai *et. al.*[78] já evidenciaram algo oposto: que se as águas estruturadas presentes nos registros PDBs são acrescidas ao cálculo, o empacotamento geral das proteínas torna-se surpreendentemente uniforme, não diferenciando núcleo de superfície. A despeito de toda essa complexidade, se nós assumirmos Bayesianamente a priori que as conclusões de Fleming & Richards são verdadeiras, que as diferenças de empacotamento entre os diferentes resíduos não sejam tão significantes assim, e se nós aceitarmos os centróides em CA e GC como assinaturas das contribuições da cadeia principal e cadeia lateral, respectivamente, talvez seja possível lançar um outro olhar sobre os dados da Figura 21: que a tendência de proteínas “alfas” serem mais bem empacotadas que “betas” esteja vindo do carácter helicoidal da cadeia principal. Trata-se de algo que precisa ser

estudado mais profundamente.

Nós estamos finalizando também uma análise multivariada envolvendo 43 parâmetros do STING_DB, a fim de ampliar o espectro das variáveis envolvidas e distinguir o que de fato é determinante nas características do empacotamento de hélices e fitas. Outras frentes ainda incluem: a influência das distâncias dos resíduos na seqüência ($d_r > 0$) sobre o perfil dos contatos; e a avaliação dos contatos envolvendo grandezas vetoriais (similaridade de grafos) e não somente grandezas escalares (distribuição de distâncias) como foi feito.

7. Conclusões

Ao final dessa jornada, nós sentimos ter tocado em diferentes questões. Primeiramente, houve a emergência do delimitador em 7,0 Å como um importante parâmetro de distância em análise de contatos de proteínas. Nós julgamos ter demonstrado que, até essa distância, as técnicas de prospecção de contatos delimitador dependente (CD) e decomposição de Delaunay (DT) convergem em seus resultados, o que nos permitiu unificar suas propriedades numa só: que até 7,0 Å todos os contatos são totais e verdadeiro-positivos. Defendemos também a hipótese de que essa nova propriedade é uma assinatura topológica da primeira camada de coordenação, envolvendo na sua maior parte vizinhos imediatos não-oclusos. A distância de 7,0 Å seria, portanto, aquela em que a primeira ordem de contatos estaria otimamente separada das demais ordens de vizinhos. Foi importante constatar também que nesta distância, os resultados eram indiferentes quanto ao tipo de centróide, se carbono alfa (CA) ou centro geométrico da cadeia lateral (GC), e tipo de classe, se *all-alpha* ou *all-beta*. A unificação dos modelos CD/DT também afirmou a condição linear da primeira ordem de contatos. Acreditamos que nós conseguimos mostrar que há uma mudança de modelo linear para quadrático quando a distância máxima dada pelo delimitador extrapola as primeiras camadas de vizinhos, ou seja, quando ela é maior que 7,0 Å.

Outra inesperada conclusão concerne à aplicabilidade da técnica DT na prospecção de contatos em proteínas. Nós vimos que DT carrega intrinsecamente uma inconveniente anomalia que o faz rejeitar arestas legítimas quando os sites encontram-se numa situação próxima da condição degenerada ao algoritmo. Embora a interferência dessa anomalia estimada por nós tenha sido pequena (da ordem de 5% para BETA CA), nós observamos que esse erro não é aleatório, mas sistemático, afetando de forma estruturada os contatos de proteínas "beta". Nós investigamos também a solução usualmente adotada para tratar essa anomalia: a metodologia quase-Delaunay (AD). Foi empiricamente demonstrado que AD tende a ser um complemento de DT e que, portanto, suas somas tendem à CD na medida que o parâmetro de perturbação cresce. Se com CD que é uma técnica muito mais simples nós temos a mesma garantia até 7,0 Å de contatos completos e verdadeiro-positivos, por que usar DT ou DT+AD ? É muito importante enfatizar que nós não estamos de modo algum condenando DT ou técnicas correlatas como não úteis ao estudo de contatos em proteínas.

Angelov *et. al*.[161] tem encontrado resultados curiosos explorando parâmetros topológicos das células de Voronoi, tais como o número de arestas por face, que pode ser avaliado como um peso ao contato, relacionado à simetria e ao grau de interação entre vizinhos. Nosso trabalho pode apenas afirmar, no estrito intervalo entre 0,0 até 7,0 Å, em grafos cujos contatos inter-resíduos tem como peso as distâncias Euclideanas, que DT ou DT+AD parecem não ser necessários, já que com CD nós obtemos de maneira mais simples, resultados mais completos e confiáveis.

Finalmente, nosso estudo de caso comparando os centróides CA e GC mostrou que o uso de delimitadores de distância menores que 6,8 Å podem conduzir a resultados contraditórios no que diz respeito a quem organiza em média mais vizinhos por resíduos, se cadeias "alfa" ou "beta". Vimos que nessa situação CA parece favorecer "alfa" e GC "beta". Mas felizmente, acima de 6,8 Å, os resultados concordam quanto ao fato de proteínas "beta" formarem em média mais vizinhos por resíduo. Investigações mais detalhadas serão necessárias para avaliar o real grau desse viés, principalmente em aplicações onde a precisão dos contatos é importante, com nos potenciais empíricos. Seja como for, esse resultado também reforça a credibilidade de 7,0 Å como um parâmetro de referência, robusto e de carácter geral, a ser usado de forma segura como um confiável delimitador de distância nos estudos em massa de contatos de proteínas.

Como bem disse Rubem Alves[40]:

> "*Da mesma forma como os anzóis pré-determinam os resultados da pescaria, os métodos pré-determinam o resultado da pesquisa. Porque os métodos são preparados de antemão para pegar aquilo que desejamos pegar*"

8. Bibliografia

[1] RAMOS, C. H. I. História - Proteínas I. *CBME Informação,* n. 2, p. 2, fev. 2004.

[2] MULDER, G. J. Ueber die Zusammensetzung einiger thierischen substanzen. *Journal für Praktische Chemie,* v. 16, p. 129-152, 1839.

[3] TEICH, M.; NEEDHAM, D. *Documentary History of Biochemistry, 1770-1940.* New York: Continuum International Publishing Group, 1991. p. 275-277.

[4] BUENO, F. S. *Grande dicionário etimológico-prosódico da língua portuguesa.* São Paulo: Lisa, vol. IV. 1988. p. 1826.

[5] PROTEIN. In: *WIKIPEDIA,* The free encyclopedia. Disponível em: <http://en.wikipedia.org/wiki/Protein/>. Acesso em janeiro de 2008.

[6] NELSON, D. L.; COX, M. M. *Lehninger Principles of Biochemistry.* 4. ed. New York: W. H. Freeman and Company, 2000. p. 75.

[7] ZHANG, Y.; BARANOV, P. V.; ATKINS, J. F.; GLADYSHEV, V. N. Pyrrolysine and selenocysteine use dissimilar decoding strategies. *Journal of Biological Chemistry,* v. 280, n. 21, p. 20740-20751, mar. 2005.

[8] RAMOS, C. H. I. História - Proteinas II. *CBME Informação,* n. 3, p. 3, abril de 2004.

[9] WAROLIN, C. Pierre Jean Robiquet (1780-1840). *Revue d'historie de la pharmacie.* v. 47, n. 321, p. 97-110, 1999.

[10] McCOY, R. H., MEYER, C. E., ROSE, W. C. Feeding experiments with mixtures of highly purified amino acides VIII: isolation and identification of a new essential amino acid. *Journal of Biological Chemistry,* v. 12, n. 1, p. 283-302, 1935.

[11] RAMOS, C. H. I. Protein folding & misfolding (Editorial). *Protein & Peptide Letters,* v. 12, n. 3, 2005

[12] WU, H. Studies on Denaturation of Proteins XIII. A Theory of Denaturation. *Advanced Protein Chemistry,* v. 46, p. 6-26, 1995. (Reprinted from the Chinese Journal of Physiology, v. 5, n. 4, p. 321-344, 1931)

[13] WÜTHRICH, K. Autobiography. In: *Les Prix Nobel. The Nobel Prizes 2002,* Stockholm: Editor Tore Frängsmyr, 2003.

[14] PAULING, L.; COREY, R. B. Atomic coordinates and structure factors for two helical configurations of polypeptide chains. *Proceedings of the National Academy of Sciences - PNAS,* v. 37, n. 5, p. 235-240, 1951.

[15] PAULING, L.; COREY, R. B. The pleated sheet, a new layer configuration of polypeptide chains. *Proceedings of the National Academy of Science - PNAS,* v. 37, n. 5, p. 251-256, 1951.

[16] MAX F. PERUTZ - BIOGRAPHY. In: *Nobel Lectures, Chemistry 1942-1962,* Amsterdam: Elsevier publishing company, 1964.

[17] JOHN C. KENDREW - BIOGRAPHY. In: *Nobel Lectures, Chemistry 1942-1962,* Amsterdam: Elsevier publishing company, 1964.

[18] ANSON, M. L.; MIRSKY, A. E. On some general properties of proteins. *Journal of General Physiology,* v. 9, p. 169-179, 1925.

[19] MIRSKY, A. E.; ANSON, M. L. Protein coagulation and its reversal: the reversal of the coagulation of hemoglobin. *Journal of General Physiology,* v. 13, p. 133-143, 1930.

[20] TANFORD, C. How protein chemists learned about the hydrophobic factor. *Protein Science,* v. 6, p. 1358-1366, 1997.

[21] MIRSKY, A. E.; PAULING, L. On the structure of native, denatured and coagulated proteins. *Proceeding of National Academy of Sciences - PNAS,* v. 22, n. 7, p. 439-447, 1936.

[22] KAUZMANN, W. Some factors in the interpretation of protein denaturation. *Advances in Protein Chemistry,* v. 14, p. 1-63, 1993.

[23] KLOTZ, I. Solvent water and protein behavior. *Protein Science,* v. 2, p. 1992-1999, 1993.

[24] HOROWITZ, N. H. One-gene-one-enzyme: remembering biochemical genetics. *Protein Science,* v. 4, p. 1017-1019, 1995.

[25] FELDMAN, D. E.; FRYDMAN, J. Protein folding in vivo: the importance of molecular chaperones. *Current Opinion in Structural Biology,* v. 10, n. 1, p. 26-33, 2000.

[26] MOUDRIANAKIS, E. N. From protein coagulation and reversible denaturation to the protein folding problem: Chris Anfisen defining the transition. *The FASEB Journal,* v. 10, p. 179-183.1996.

[27] ANFINSEN, C. B. Studies on the principles that govern the folding of protein chains. In: *Nobel Lectures, Chemistry 1971-1980,* Singapore: Editor-in-Charge Tore Frängsmyr, Editor Sture Forsén, World Scientific Publishing Co, 1993.

[28] LINEWEAVER, C. H.; DAVIS, T. M. Equívocos sobre o Big-Bang. *Scientific American Brasil,* n. 35, P. 32-35, abril de 2005.

[29] DILL, K. A. Polymer principles and protein folding. *Protein Science,* v. 8, p. 1166-1180, 1999

[30] LEVINTHAL, C. How to fold graciously?*Proceeding of Mössbauer spectroscopy in biological systems meeting,* Illinois: University of Illinois Press, 1969. p. 22-24.

[31] DOBSON, C. M. The nature and significance of protein folding. In: PAIN, R. H. (Org.). *Mechanisms of protein folding.* 2 ed. New York: Oxford University Press, 2000. p. 1-33.

[32] HONIG, B. Protein folding: from the Levinthal paradox to structure prediction. *Journal of Molecular Biology,* v. 293, p. 283-293, 1999.

[33] BALDWIN, R. L. Protein folding from 1961 to 1982. *Nature Structural Biology,* v. 6, n. 9, p. 814-817, 1999.

[34] UNGER, R.; MOULT, J. Finding the lowest free energy conformation of a protein is a NP-hard problem: proof and implication. *Bulletin of Mathematical Biology,* v. 55, n. 6, p. 1183-1198, 1993.

[35] CREIGHTON, T. E. The protein folding problem. *Science,* v. 240, p. 267-344, 1988.

[36] BALDWIN, R. L. The problem was to find the problem. *Protein Science,* v. 6, p. 2031-2034, 1997.

[37] CREIGHTON, T. E. How important is the molten globule for correct protein folding? *Trends in Biochemistry - TIBS,* v. 22, p. 6-10, 1997.

[38] FENG, H.; Zhou, Z.; Bai, Y. A protein folding pathway with multiple folding intermediates at atomic resolution. *Proceeding of the National Academy of Sciences - PNAS,* v. 102, n. 14, p. 5026-5031, 2005.

[39] VALENCIA, A. Protein refinement: a new challenge for CASP in its 10th anniversary. *Bioinformatics,* v. 21, n. 3, p. 277, 2005.

[40] ALVES, R.*Filosofia da Ciência:* introdução ao jogo e sua regras. 15a ed. São Paulo: Brasiliense, 1992.

[41] RAMOS, C. H. I. O segundo passo na tradução da informação gênica: o enovelamento de proteinas. *Bioscience Journal,* Uberlândia: Edição Especial, p. 39-52, 2004.

[42] LESK, A. M.; CHOTHIA, C. How different amino acids sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *Journal of Molecular Biology,* v. 136, p. 225-270, 1980.

[43] LEATHERBARROW, R. J.; FERSHT, A. R. Protein engineering. *Protein Engineering Design and Selection,* v. 1, p. 7-16, 1986.

[44] KOCHENDOERFER, G. G.; SALOM, D.; LEAR, J. D.; WILK-ORESCAN, R.; STEPHEN, B.; KENT, H.; DeGRADO, W. F. Total chemical synthesis of the integral membrane protein influenza a virus M2:role of its c-terminal domain in tetramer assembly.*Biochemistry,* v. 38, p. 11905-11913, 1999.

[45] RIBEIRO, E. A; RAMOS, C. H. I. Circular Permutation and Deletion Studies of Myoglobin Indicate that the Correct Position of Its N-Terminus Is Required for Native Stability and Solubility but Not for Native-like Heme Binding and Folding. *Biochemistry,* v. 44, p. 4699-4709, 2005.

[46] CREIGHTON, T. E.; DARBY, N. J.; KEMMINK, J. The roles of partly folded intermediates in protein folding. *FASEB Journal,* v. 10, p. 110-118, 1996.

[47] MATOUSCHEK, A.; KELLIS, J. T.; SERRANO, L.; FERSHT, A. L. Mapping the transition state and pathway of protein folding by protein engineering. *Nature,* v. 340, p. 122-126, 1989.

[48] TSUI, V.; GARCIA, C.; CAVAGNERO, S.; STUZDAK, G.; DYSON, H. J.; WRIGHT, P. E. Quench-flow experiments combined with mass spectometry show apomyoglobin folds through an obligatory intermediate. *Protein Science,* v. 8, p. 45-49, 1998.

[49] BROOKS, B. R.; BRUCCOLERI, R. E.; OLAFSON, B. D.; STATES, D. J.; SWAMINATHAN, S.; KARPLUS, M. CHARMM: a program for macromolecular energy minimization and dynamics calculations. *Journal of Computational Chemistry,* v. 4, p. 187-217, 1983.

[50] DAGGETT, V.; FERSHT, A. R. Is there a unifying mechanism for protein folding? *TRENDS in*

Biochemical Sciences, v. 28, n. 1, p. 18-26, 2003.

[51] CREIGHTON, T. E. Protein Folding. *Biochemical Journal,* v. 270, p. 1-16, 1990.

[52] SALI, A.; SHAKHNOVICH, E.; KARPLUS, M. How does a protein folding? *Nature,* v. 369, p. 248-251, 1994.

[53] BALDWIN, R. L.; ROSE, G. D. Is protein folding hierarchic? I. Local structure and peptide folding. *TIBS,* v. 24, p. 26-33, 1999.

[54] BALDWIN, R. L.; ROSE, G. D. Is protein folding hierarchic? II - Folding intermediates and transition states. *TIBS,* v. 24, p. 77-83, 1999

[55] DILL, K. A. Polymer principles and protein folding. *Protein Science,* v. 8, p. 1166-1180.1999.

[56] TATENO, Y.; GOJOBORI, T. DNA data bank of Japan in the age of information biology. *Nucleic Acids Research,* v. 25, p. 489-491, 2003.

[57] BOECKMANN, B.; BAIROCH, A.; APWEILER, R.; BLATTER, M.; ESTREICHER, A.; GASTEIGER, E.; MARTIN, M. J.; MICHOUD, K.; O'DONOVAN, C.; PHAN, I.; PILBOUT, S.; SCHNEIDER, M. The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research,* v. 31, p. 365-370, 2003.

[58] STEIN, L. D.; THIERRY-MIEG, J. Scriptable access to the Caenorhabditis elegans genome sequence and other AceDB databases. *Genome Research,* v. 8, n. 12, p. 1308-1315, 1998.

[59] BERNSTEIN, F. C.; KOETZLE, T. F; WILLIAMS, G. J.; MEYER JR, E. F.; BRICE, M. D.; RODGER, J. R.; KENNARD, O.; SHIMANOUCHI, T.; TASUMI, M. The Protein Data Bank: a computer-based archival file for macromolecular structures. *Journal of Molecular Biology,* v. 112, n. 3, p. 535–542, 1977.

[60] STOESSER, G.. BAKER, W.; VAN DEN BROEK, A.; CAMON, E.; GARCIA-PASTOR, M.; KANZ, C.; KULIKOVA, T.; LEINONEN, R.; LIN, Q.; LOMBARD, V.; LOPEZ, R.; REDASCHI, N.; STOEHR, P.; TULI, M. A.; TZOUVARA, K.; VAUGHAN, R. The EMBL Nucleotide Sequence Database. *Nucleic Acids Research,* v. 30, p. 21–26, 2002.

[61] BENSON, D. A.; KARSCH-MIZRACHI, I.; LIPMAN, D. J.; OSTELL, J.; WHEELER, D. L. GenBank. *Nucleic Acids Research,* v. 31, p. 23-27, 2003.

[62] WU, C. H.; YEH, L. S.; HUANG, H.; ARMINSKI, L.; CASTRO-ALVEAR, J.; CHEN, Y.; HU, Z.; KOURTESIS, P.; LEDLEY, R. S.; SUZEK, B. E.; VINAYAKA, C. R.; ZHANG, J.; BARKER, W. C. The protein information resource. *Nucleic Acids Research,* v. 31, p. 345–347, 2003.

[63] APWEILER, R.; BAIROCH, A.; WU, C.H.; BARKER, W. C.; BOECKMANN, B.; SERENELLA, F.; GASTEIGER, E.; HUANG, H.; LOPEZ, R.; MAGRANE, M.; MARTIN, M. J.; NATELE, D. A.; O'DONOVAN, C.; REDASCHI, N. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research,* v. 32, p. D115, D119, 2004.

[64] BERMAN, H. M.;HENRICK, K.; NAKAMURA, H. Announcing the worldwide Protein Data Bank. *Nature Structural Biology,* v. 10, n. 12, p. 980, 2003.

[65] FIERS, W.; CONTRERAS, R.; DUERINCK, F.; HAEGEMAN, G.; ISERENTANT, D.; MERREGAERT, J.; MIN JOU, W.; MOLEMANS, F.; RAEYMAEKERS, A.; VAN DEN BERGHE, A.; VOLCKAERT, G.; YSEBAERT, M.Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature,* v. 260, p. 500-507, 1976.

[66] FLEISCHMANN, R. D.; ADAMS, M. D.; WHITE, O.; CLAYTON, R. A.; KIRKNESS, E. F.; KERLAVAGE, A. R.; BULT, C. J.; TOMB, J. F.; DOUGHERTY, B.A.; MERRICK, J. M. et al.Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science,* v. 269, p. 496-512, 1995.

[67] GOFFEU, A.; BARRELL, B.G; BUSSEY, H.; DAVIS, R. W.; DUJON, B.; FELDMANN, H.; GALIBERT, F.; HOHEISEL, J. D.; JACQ, C.; JOHNSTON, M.; LOUIS, E. J.; MEWES, H. W.; MURAKAMI, Y.; PHILIPPEN, P.; TETTELIN, H.; OLIVER, S. G. Life with 6000 genes. *Science,* v. 274, p. 563-567, 1996.

[68] BLATTNER, F. R.; PLUNKETT, G.; BLOCH, C. A.; PERNA, N. T.; BURLAND, V.; RILEY, M.; COLLADOVIDES, J.; GLASNER, J. D.; RODE, C. K.; MAYHEW, G. F.; GREGOR, J.; DAVIS, N. W.; KIRKPATRICK, H. A.; GOEDEN, M. A.; ROSE, D. J.; MAU, B.; SHAO, Y. The complete genome sequence of *Escherichia coli K-12. Science,* v. 277, p. 1453- 1474, 1997.

[69] Genome sequence of the nematode *C. elegans, Science,* v. 282, p. 2012-2018, 1998.

[70] ADAMS, M.D.; CELNIKER, S.E.; HOLT, R.A.; EVANS, C.A.; GOCAYNE, J.D.; AMANATIDES, P.G.; SCHERER, S.E.; LI, P.W.; HOSKINS, R.A.; GALLE, R.F.; GEORGE, R.A.; LEWIS, S.E.; RICHARDS, S.; ASHBURNER, M.;HENDERSON, S.N.; SUTTON, G.G.; WORTMAN, J.R.; YANDELL, M.D.;

ZHANG, Q.; CHEN, L.X.; BRANDON, R.C.; ROGERS, Y.H.C.; BLAZEJ, R.G.; CHAMPE, M.; PFEIFFER, B.D.; WAN, K.H.; DOYLE, C.; BAXTER, E.G.; HELT, G.; NELSON, C.R.; MIKLOS, G.L.G.; ABRIL, J.F.; AGBAYANI, A.; AN, H.J.; ANDREWS- PFANNKOCH, C.; BALDWIN, D.; BALLEW, R.M.; BASU, A.; BAXENDALE, J.; BAYRAKTAROGLU, L.; BEASLEY, E.M.; BEESON, K.Y.; BENOS, P.V.; BERMAN, B.P.; BHANDARI, D.; BOLSHAKOV, S.; BORKOVA, D.; BOTCHAN, M.R.; BOUCK, J.; BROKSTEIN, P.; BROTTIER, P.; BURTIS, K.C.; BUSAM, D.A.; BUTLER, H.; CADIEU, E.; CENTER, A.; CHANDRA, I.; CHERRY, J.M.; CAWLEY, S.; DAHLKE, C.; DAVENPORT, L.B.; DAVIES, A.; DE PABLOS, B.; DELCHER, A.; DENG, Z.M.; MAYS, A.D.; DEW, I.; DIETZ, S.M.; DODSON, K.; DOUP, L.E.; DOWNES, M.; DUGAN-ROCHA, S.; DUNKOV, B.C.; DUNN, P.; DURBIN, K.J.; EVANGELISTA, C.C.; FERRAZ, C.; FERRIERA, S.; FLEISCHMANN, W.; FOSLER, C.; GABRIELIAN, A.E.; GARG, N.S.; GELBART, W.M.; GLASSER, K.; GLODEK, A.; GONG, F.C.; GORRELL, J.H.; GU, Z.P.; GUAN, P.; HARRIS, M.; HARRIS, N.L.; HARVEY, D.; HEIMAN, T.J.; HERNANDEZ, J.R.; HOUCK, J.; HOSTIN, D.; HOUSTON, D.A.; HOWLAND, T.J.; WEI, M.H.; IBEGWAM, C.; JALALI, M.; KALUSH, F.; KARPEN, G.H.; KE, Z.X.; KENNISON, J.A.; KETCHUM, K.A.; KIMMEL, B.E.; KODIRA, C.D.; KRAFT, C.; KRAVITZ, S.; KULP, D.; LAI, Z.W.; LASKO, P.; LEI, Y.D.; LEVITSKY, A.A.; LI, J.Y.; LI, Z.Y.; LIANG, Y.; LIN, X.Y.; LIU, X.J.; MATTEI, B.; MCINTOSH, T.C.; MCLEOD, M.P.; MCPHERSON, D.; MERKULOV, G.; MILSHINA, N.V.; MOBARRY, C.; MORRIS, J.; MOSHREFI, A.; MOUNT, S.M.; MOY, M.; MURPHY, B.; MURPHY, L.; MUZNY, D.M.; NELSON, D.L.; NELSON, D.R.; NELSON, K.A.; NIXON, K.; NUSSKERN, D.R.; PACLEB, J.M.; PALAZZOLO, M.; PITTMAN, G.S.; PAN, S.;POLLARD, J.; PURI, V.; REESE, M.G.; REINERT, K.; REMINGTON, K.; SAUNDERS, R.D.C.; SCHEELER, F.; SHEN,H.; SHUE, B.C.; SIDEN-KIAMOS, I.; SIMPSON, M.; SKUPSKI, M.P.; SMITH, T.; SPIER, E.; SPRADLING, A.C.; STAPLETON, M.; STRONG, R.; SUN, E.; SVIRSKAS, R.; TECTOR, C.; TURNER, R.; VENTER, E.; WANG, A.H.H.; WANG, X.; WANG, Z.Y.; WASSARMAN, D.A.; WEINSTOCK, G.M.; WEISSENBACH, J.; WILLIAMS, S.M.; WOODAGE, T.; WORLEY, K.C.; WU, D.; YANG, S.; YAO, Q.A.; YE, J.; YEH, R.F.; ZAVERI, J.S.; ZHAN, M.; ZHANG, G.G.; ZHAO, Q.; ZHENG, L.S.; ZHENG, X.Q.H.; ZHONG, F.N.; ZHONG, W.Y.; ZHOU, X.J.; ZHU, S.P.; ZHU, X.H.; SMITH, H.O.; GIBBS, R.A.; MYERS, E.W.; RUBIN, G.M. AND VENTER, J.C. The genome sequence of *Drosophila melanogaster. Science,* v. 287, p. 2185-2195, 2000.

[71] Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature,* v. 408, p. 796-815, 2000.

[72] VENTER, J.C.; ADAMS, M.D.; MYERS, E.W.; LI, P.W.; MURAL, R.J.; SUTTON, G.G.; SMITH, H.O.; YANDELL, M.; EVANS, C.A.; HOLT, R.A.; GOCAYNE, J.D.; AMANATIDES, P.; BALLEW, R.M.; HUSON, D.H.; WORTMAN, J.R.; ZHANG, Q.; KODIRA, C.D.; ZHENG, X.Q.H.; CHEN, L.; SKUPSKI, M.; SUBRAMANIAN, G.; THOMAS, P.D.; ZHANG, J.H.; MIKLOS, G.L.G.; NELSON, C.; BRODER, S.; CLARK, A.G.; NADEAU, C.; MCKUSICK, V.A.; ZINDER, N.; LEVINE, A.J.; ROBERTS, R.J.; SIMON, M.; SLAYMAN, C.; HUNKAPILLER, M.; BOLANOS, R.; DELCHER, A.; DEW, I.; FASULO, D.; FLANIGAN, M.; FLOREA, L.; HALPERN, A.; HANNENHALLI, S.;KRAVITZ,S.; LEVY, S.; MOBARRY, C.; REINERT, K.; REMINGTON, K.; ABU-THREIDEH, J.; BEASLEY, E.; BIDDICK, K.;BONAZZI, V.; BRANDON, R.; CARGILL, M.; CHANDRAMOULISWARAN, I.; CHARLAB, R.; CHATURVEDI, K.; DENG, Z.M.; DI FRANCESCO, V.; DUNN, P.; EILBECK, K.; EVANGELISTA, C.; GABRIELIAN, A.E.; GAN, W.; GE, W.M.; GONG, F.C.; GU, Z.P.; GUAN, P.; HEIMAN, T.J.; HIGGINS, M.E.; JI, R.R.; KE, Z.X.; KETCHUM, K.A.; LAI, Z.W.; LEI, Y.D.; LI, Z.Y.; LI, J.Y.; LIANG, Y.; LIN, X.Y.; LU, F.; MERKULOV, G.V.; MILSHINA, N.; MOORE, H.M.; NAIK, A.K.; NARAYAN, V.A.; NEELAM, B.; NUSSKERN, D.; RUSCH, D.B.; SALZBERG, S.; SHAO, W.; SHUE, B.X.; SUN, J.T.; WANG, Z.Y.; WANG, A.H.; WANG, X.; WANG, J.; WEI, M.H.; WIDES, R.; XIAO, C.L.; YAN, C.H.; YAO, A.; YE, J.; ZHAN, M.; ZHANG, W.Q.; ZHANG, H.Y.; ZHAO, Q.; ZHENG, L.S.; ZHONG, F.; ZHONG, W.Y.; ZHU, S.P.C.; ZHAO, S.Y.; GILBERT, D.; BAUMHUETER, S.; SPIER, G.; CARTER, C.; CRAVCHIK, A.; WOODAGE, T.; ALI, F.; AN, H.J.; AWE, A.; BALDWIN, D.; BADEN, H.; BARNSTEAD, M.; BARROW, I.; BEEYSON, K.; BUSAM, D.; CARVER, A.; CENTER, A.; CHENG, M.L.; CURRY, L.; DANAHER, S.; DAVENPORT, L.; DESILETS, R.; DIETZ, S.; DODSON, K.; DOUP, L.; FERRIERA, S.; GARG, N.; GLUECKSMANN, A.; HART, B.; HAYNES, J.; HAYNES, C.; HEINER, C.; HLADUN, S.; HOSTIN, D.; HOUCK, J.; HOWLAND, T.; IBEGWAM, C.; JOHNSON, J.; KALUSH, F.; KLINE, L.; KODURU, S.; LOVE, A.; MANN, F.; MAY, D.; MCCAWLEY, S.; MCINTOSH, T.; MCMULLEN, I.; MOY, M.; MOY, L.; MURPHY, B.; NELSON, K.; PFANNKOCH, C.; PRATTS, E.; PURI, V.; QURESHI, H.; REARDON, M.; RODRIGUEZ, R.; ROGERS, Y.H.; ROMBLAD, D.; RUHFEL, B.; SCOTT, R.; SITTER, C.; SMALLWOOD, M.; STEWART, E.; STRONG, R.; SUH, E.; THOMAS, R.;

TINT, N.N.; TSE, S.; VECH, C.; WANG, G.; WETTER, J.; WILLIAMS, S.; WILLIAMS, M.; WINDSOR, S.; WINN-DEEN, E.; WOLFE, K.; ZAVERI, J.; ZAVERI, K.; ABRIL, J.F.; GUIGO, R.; CAMPBELL, M.J.; SJOLANDER, K.V.; KARLAK, B.; KEJARIWAL, A.; MI, H.Y.; LAZAREVA, B.; HATTON, T.; NARECHANIA, A.; DIEMER, K.; MURUGANUJAN, A.; GUO, N.; SATO, S.; BAFNA, V.; ISTRAIL, S.; LIPPERT, R.; SCHWARTZ, R.; WALENZ, B.; YOOSEPH, S.; ALLEN, D.; BASU, A.; BAXENDALE, J.; BLICK, L.; CAMINHA, M.; CARNES-STINE, J.; CAULK, P.; CHIANG, Y.H.; COYNE, M.; DAHLKE, C.; MAYS, A.D.; DOMBROSKI, M.; DONNELLY, M.; ELY, D.; ESPARHAM, S.; FOSLER, C.; GIRE, H.; GLANOWSKI, S.; GLASSER, K.; GLODEK, A.; GOROKHOV, M.; GRAHAM, K.; GROPMAN, B.; HARRIS, M.; HEIL, J.; HENDERSON, S.; HOOVER, J.; JENNINGS, D.; JORDAN, C.; JORDAN, J.; KASHA, J.; KAGAN, L.; KRAFT, C.; LEVITSKY, A.; LEWIS, M.; LIU, X.J.; LOPEZ, J.; MA, D.; MAJOROS, W.; MCDANIEL, J.; MURPHY, S.; NEWMAN, M.; NGUYEN, T.; NGUYEN, N.; NODELL, M. The sequence of human genome. *Science,* v. 291, p. 1304-1351, 2001.

[73] BERNAL, A.; EAR, U.; KYRPIDES, N. Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Research,* v. 29, n. 1, p. 126-127, 2001.

[74] LEDERBERG, J.; McCRAY, A. T. 'Ome Sweet Omics' - A Genealogical Treasury of Words. *The Scientist,* v. 15, n. 7, p. 8, 2001.

[75] RICHARDS, F. M. Protein stability: still an unsolved problem. *Cell Molecular Life Science - CMLS,* v. 53, p. 790-802, 1997.

[76] RICHARDS, F. M. The interpretation of protein structures: total volumes, group volume distributions and packing density. *Journal of Molecular Biology.*v. 82, p. 1-14, 1974.

[77] FLEMING, P. J.; RICHARDS, F. M. Protein Packing: dependence on protein size, secondary structure and amino acid composition. *Journal of Molecular Biology,* v. 299, p. 487-498, 2000.

[78] TSAI, J.; TAYLOR, R.; CHOTHIA, C.; GERSTEIN, C. The packing density in proteins: standard radii and volumes. *Journal of Molecular Biology,* v. 290, p. 253-266, 1999.

[79] LIANG, J.; DILL, K. A. Are proteins well-packed? *Biophysical Journal,* v. 81, p. 751-766, 2001.

[80] LISEWSKI, A. M.; LICHTARGE, O. Rapid detection of similarity in protein structure and function through contact metric distances. *Nucleic Acids Research,* v. 34, p. 1-10, 2006.

[81] KABSCH, W.; SANDER, C. Dictionary of protein secondary structure: pattern recognition hydrogen-bonded and geometrical features. *Biopolymers,* v. 22, p. 2577-2637, 1983.

[82] MELO, R. C.; LOPES, C. E. R.; FERNANDES JR, F.; SILVEIRA, C. H.; SANTORO, M. M.; CARCERONI, R. L.; MEIRA JR, W.; ALBUQUERQUE, A. A. A contact map matching approach to protein structure similarity analysis. *Genetics and Molecular Research,* v. 5, p. 284-308, 2006.

[83] HOLM, L.; SANDER, C. Protein structure comparison by alignment of distance matrix. *Journal of Molecular Biology,* v. 233, p. 123-138, 1993.

[84] WORD, J. M.; LOVELL, S. C.; LABEAN, T. H.; TAYLOR, H. C.; ZALIS, M. E.; PRESLEY, B. K.; RICHARDSON, J. S.; RICHARDSON, D. C. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicity hydrogen atoms. *Journal of Molecular Biology,* v. 285, p. 1711-1733, 1999.

[85] SAMUDRALA, R.; MOULT, J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *Journal of Molecular Biology,* v. 275, p. 895-916, 1998.

[86] BOWIE, J. U.; LÜTHY, R.; EISENBERG, D. A method to identify protein sequences that fold into a know three-dimensional structure. *Science,* v. 253, p. 164-170, 1991.

[87] VELOSO, C.; SILVEIRA, C. H.; MELO, R. C; RIBEIRO, C.; LOPES, C. E. R., SANTORO, M. M.; MEIRA JR, W. On the characterization of energy network of proteins. *Genetics and Molecular Research,* v. 6, n. 4, p. 799-820, 2007.

[88] ALTIGAN, A. R.; AKAN, P.; BAYSAL, C. Small-World communication of residues and significance for protein dynamics. *Biophysical Journal,* v. 292, p. 85-91, 2004.

[89] KANNAN, N.; VISHVESHWARA, S. Identification of side-chain cluster in protein structures by a graph spectral method. *Journal of Molecular Biology,* v. 292, p. 441-464, 1999.

[90] JERNIGAN, R. L.; BAHAR, I. Structure-derived potentials and protein simulations. *Current Opinion in Structure Biology,* v. 6, p. 195-209, 1996.

[91] BAHAR, I.; JERNIGAN, R. L. Inter-residue potentials in globular protein and the dominance of highly specific hydrophilic interaction at close separation. *Journal of Molecular Biology,* v. 266, p. 195-214, 1997.

[92] SIPPL, M. J. Knowledge-based potentials for proteins. *Current Opinion in Structural Biology,* v. 5, p. 229-235, 1995.

[93] PTITSYN, O. B.; TING, K. H. Non-functional conserved residues in globins and their possible role as a folding nucleus. *Journal of Molecular Biology,* v. 291, p. 671-682, 1999.

[94] PLAXCO, K. W.; SIMONS, K. T.; BAKER, D. Contact order , transition state placement and the refolding rates of single domains proteins. *Journal of Molecular Biology,* v. 277, p. 985-994, 1998.

[95] MELO, R. C.; RIBEIRO, C.; MURRAY, C. S.; VELOSO, C. J. M.; SILVEIRA, C. H.; NESHICH, G.; MEIRA JR., W.; CARCERONI, R. L.; SANTORO, M. M. Finding protein-protein interactions patterns by contact-maps matching: BPTI-Serine protease complexes as a case study. *Genetic and Molecular Research,* v. 6, p. 946-963, 2007.

[96] MANCINI, A. L.; HIGA, R. H.; OLIVEIRA, A.; DOMINIQUINI, F.; KUSER, P. R.; YAMAGISHI, M. E.; TOGAWA, R. C.; NESHICH, G. STING Contacts: a web-based application for identification and analysis of amino acid contacts within protein structure and across protein interfaces. *Bioinformatics,* v. 20, p. 2145-2147, 2004.

[97] HERINGA, J.; ARGOS, P. Side-chain clusters in proteins structures and their role in protein folding. *Journal of Molecular Biology,* v. 220, p. 151-171, 1991.

[98] GODZIK, A.; KOLINSKI, A.; SKOLNICK, J. Topology fingerprint approach to the inverse protein folding problem. *Journal of Molecular Biology,* v. 227, p. 227-238, 1992.

[99] GREGORET, L. M.; COHEN, F. E. Protein folding: effect of packing density on chain conformation. *Journal of Molecular Biology,* v. 219, p. 109-122, 1991.

[100] MIYAZAWA, S.; JERNIGAN, R. L. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules,* v. 18, p. 534-552, 1985.

[101] MANAVALAN, P.; PONNUSWAMY, P. K. A study of the preferred environment of amino acid residues in globular proteins. *Achives in Biochemistry and Biophysics,* v. 184, p. 476-487, 1977.

[102] ZHANG, C.; VASMATZIS, G.; CORNETTE, J. L.; DELISI, C. Determination of atomic desolvation energies from the structures of crystallized proteins. *Journal of Molecular Biology,* v. 267, p. 707-726, 1997.

[103] FURUICHI, E.; KOEHL, P. Influence of protein structure databases on the predictive power of statistical pair potentials. *Proteins, Structure, Function and Genetics,* v. 31, p. 139-149, 1998.

[104] KAMAGATA, K.; KUWAJIMA, K. Surprisingly high correlation between early and late stages in non-two-stage protein folding. *Journal of Molecular Biology,* v. 357, p. 1647-1654, 2006.

[105] TANAKA, S.; SCHERAGA, H. A. Model of protein folding: inclusion of short-, medium-, and long-range interactions. *Proceedings of National Academy of Science - PNAS,* v. 72, p. 3802-3806, 1975.

[106] RODIONOV, M. A.; JOHNSON, M. S. Residue-residue contact substitution probabilities derived from aligned three-dimensional structures and the identification of common folds. *Protein Science,* v. 3, p. 2366-2377, 1994.

[107] HERNÁNDEZ, G.; LEMASTER, D. M. Hybrid native partitioning of interactions among nonconserved residues in chimeric proteins. *Proteins, Structure, Functions and Genetics,* v. 60, p. 723-731, 2005.

[108] BLADES, M. J.; ISON, J. C.; RANASINGHE, R.; FINDLAY, J. B. C. Automatic generation and evaluation of sparce protein signatures for families of protein structural domains. *Protein Science,* v. 14, p. 13-23, 2004.

[109] SHRAKE, A.; RUPLEY, J. A. Environment and exposure to solventof protein atoms: lyzozyme and insulim. *Journal of Molecular Biology,* v. 79, p. 351-371, 1973.

[110] KIRKWOOD, J. G. Molecular Distribution in Liquids. *Journal of Chemical Physics,* v. 7, p. 919-925, 1939.

[111] GODZIK, A.; SKOLNICK, J. Sequence structure matching in globular proteins: application to supersecondary and tertiary structure determination. *Proceeding of National Academy of Science - PNAS,* v. 89, p. 12098-12102, 1992.

[112] TROPSHA, A.; CARTER Jr, C. W.; CAMMER, S.; VAISMAN, I. Simplicial neighborhood analysis of protein packing (SNAPP): a computational geometry approach to studying proteins. *Methods in Enzimology,* v. 374, p. 509-544, 2003.

[113] MAIOROV, V. N.; CRIPPEN, G. M. Contact potential that recognizes the correct folding of globular proteins. *Journal of Molecular Biology,* v. 227, p. 876-888, 1992.

[114] VORONOI, G. M. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. deuxième Mémoire: Recherches sur les paralléllоèdres primitifs. *J. Reine Angew. Math.,* v. 134, p. 198-287, 1908.

[115] DELAUNAY, B. Sur la sphère vide. A la memoire de Georges Voronoi. *Izv. Akad. Nauk. SSSR.,* v. 7, p. 793-800, 1934.

[116] POUPON, A. Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. *Current Opinion in Structural Biology,* v. 14, p. 233-241, 2004.

[117] LONCHARICH, R. J.; BROOKS, B. R. The effects of truncating long-range forces on protein dynamics. *Proteins, Structure, Functions and Genetics,* v. 6, p. 32-45, 1989.

[118] DARDEN, T.; YORK, D.; PEDERSEN, L. Particle mesh Ewald: a N.LogN method for Ewald sums in large systems. *Journal of Chemical Physics,* v. 98, p. 10089-10092, 1993.

[119] ICKING, C. R.; KLEIN, P.; KÖLLNER, L. Java applets for the dynamic visualization of Voronoi diagrams. In: *Computer science in perpective.* New York: Springer-Verlag New York, Inc., 2003.

[120] MONGE, A.; LATHROP, E. J. P.; GUNN, J. R. SHENKIN, P. S.; FRIESNER, R. A. Computer modeling of protein folding: conformational and energetic analysis of reduced and detailed protein models. *Journal of Molecular Biology,* v. 247, p. 995-1012, 1995.

[121] BAHAR, I.; JERNIGAN, R. L. Coordination geometry of nonbonded residues in globular proteins. *Folding & Design,* v. 357, p. 357-370, 1996.

[122] YUAN, X.; BYSTROFF, C. Non-sequential structure-based alignments reveal topology-independent core packing arrangements in proteins. *Bioinformatics,* v. 21, p. 1010-1019, 2005.

[123] SOYER, A.; CHOMILIER, J.; MORNON, J. P.; JULLIEN, R.; SADOC, J. F. Voronoi tessellation reveals the condensed matter character of folded proteins. *Physical Review Letters,* v. 85, p. 3532-3535, 2005.

[124] PEARL, F. M.; BENNETT, C. F.; BRAY, J. E.; HARRISON, A. P.; MARTIN, N.; SHEPHERD, A.; SILLITOE, I.; THORTON, J.; ORENGO, C. A. The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Researchs,* v. 31, p. 452-455, 2003.

[125] FRISHMAN, D.; ARGOS, P. Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function and Genetics,* v. 23, p. 566-579, 1995.

[126] BERMAN, H. M.; WESTBROOK, J.; FENG, Z.; GILLILAND, G.; BHAT, T. N.; WEISSIG, H.; SHINDYALOV, I. N.; BOURNE, P. E. The protein data bank. *Nucleic Acids Research,* v. 28, p. 235-242, 2000.

[127] NESHICH, G.; MAZONI, I.; OLIVEIRA, S. R. M.; YAMAGISHI, M. E. B.; KUSER, P. R. F.; BORRO, L. C.; MORITA, D. U.; SOUZA, K. R. R.; ALMEIDA, G. V.; RODRIGUES, D. N.; JARDINE, J. G.; TOGAWA, R. C.; MANCINI, A. L.; HIGA, R. H.; CRUZ, S. A. B.; VIEIRA, F. D.; SANTOS, E. H.; MELO, R. C.; SANTORO, M. M. The star STING server: a multiplatform environment for protein structure analysis. *Genetics and Molecular Research,* v. 5, p. 717-722, 2006.

[128] PIRES, D. E. V.; SILVEIRA, C. H.; SANTORO, M. M.; MEIRA JR., W. PDBEST - PDB Enhanced Structures Toolkit. *Proceedings of 3rd International Conference of Brazilian Association for Bioinformatics and Computational Biology - X-Meeting 2007,* p. 39, 2007.

[129] HARGROVE, M. S.; BRUCKER, E. A.; STEC, B.; SARATH, G.; ARREDONDO-PETER, R.; KLUCAS, R. V.; OLSON, J. S.; PHILLIPS JR., G. N. Crystal structure of a nonsymbiotic plant hemoglobin. *Structure & Folding Designing,* v. 8, p. 1005-1014, 2000.

[130] BERGNER, A.; OGANESSYAN, V.; MUTA, T.; IWANAGA, S.; TYPKE, D.; HUBER, R.; BODE, W. Crystal structure of a coagulogen, the clotting protein from horseshoe crab: a structural homologue of nerve growth factor. *EMBO Journal,* v. 15, p. 6789-6797, 1996.

[131] LAMOUREUX, J. S.; STUART, D.; TSANG, R.; WU, C.; GLOVER, J. N. Structure of the sporulation-specific transcription factor Ndt80 bound to DNA. *EMBO Journal,* v. 21, p. 5721-5732, 2002.

[132] MANN, H. B.; WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics,* v. 18, p. 50-60, 1947.

[133] FLIGNER, M. A.; KILLEN, T. J. Distribution-free two sample tests for scale. *Journal of the American Statistical Association,* v. 71, p. 210-213, 1976.

[134] MASSEY JR., F. J. The Kolmogorov-Smirnov teste for goodness of fit. *Journal of the American Statiscal Association,* v. 46, p. 68-78, 1951.

[135] CHOTHIA, C.; JANIN, J. The nature of the accessible and buried surfaces in proteins. *Journal of Molecular Biology,* v. 105, p. 1-14, 1976.

[136] SPOEL, D. V. D.; LINDAHL, E.; HESS, B.; GROENHOF, G.; MARK, A. E.; BERENDSEN, H. J. C.

GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry,* v. 26, n. 16, p. 1701-1718, 2005.

[137] GERSTEIN, M.; TSAI, J.; LEVITT, M. The volume of atoms on the protein surface: calculated from simulation using Voronoi polyhedra. *Journal of Molecular Biology,* v. 249, p. 955-966, 1995.

[138] HARPAZ, Y.; GERSTEIN, M.; CHOTHIA, C. Volume changes on protein folding. *Structure,* v. 2, p. 641-649, 1994.

[139] PHILLIPS, S. E. V. Structure and refinement of oxymyoglobin at 1.6 A resolution. *Journal of Molecular Biology,* v. 142, p. 531-554, 1980.

[140] TOWNSEND, M. *Discrete mathematics:* applied combinatorics and graph theory. Menlo Park: the Benjamin/Cummings Publishing Company, Inc, 1987. p. 208-210.

[141] DIRICHLET, J. P. G. L. Über die reduction der positiven quadratischen formen mit drei unbestimmten ganzen zahlen. *J. Reme u. Angew. Math.,* v. 40, p. 209-227, 1850.

[142] GAUSS, C. F. Recursion der untersuchungen iiber die eigenschaften der positiven ternaren quadratischen formen von Ludwig August Seeber. *J. Reine Angew.,* v. 20, p. 312-320, 1840.

[143] DESCARTES, R. *Principia Philosophiae.* Amsterdan: Ludovicus Elzevirius, 1644.

[144] THIESSEN, H. Precipitation averages for large areas. *Montly Weather Review,* v. 39, p. 1082-1084, 1911.

[145] WIGNER, E.; SEITS, F. On the constitution of metallic sodium. *Physical Review,* v. 43, p. 804-810, 1933.

[146] AURENHAMMER, F.; KLEIN, R. Voronoi Diagrams. in: SACK, J.; URRUTIA, G.; *Handbook of Computational Geometry.* Amsterdam: Elsevier Science, 2000. p. 201-290.

[147] BANDYOPADHYAY, D.; SNOEYINK, J.; Almost-Delaunay simplices: robust neighbor relations for imprecise 3D points using CGAL. *Computer Geometry,* v. 38, p. 4-15, 2007.

[148] KACHALOVA, G. S.; POPOV, A. N.; BARTUNIK, H. D. A steric mechanism for inhibition of CO binding to heme proteins. *Science,* v. 284, p. 473-476, 1999.

[149] DWYER, R. A. Higher-dimensional Voronoi diagrams in linear expected time. *Journal of Discrete and Computational Geometry,* v. 6, p. 343-367, 1991.

[150] SHIMIZU, T.; NAKATSU, T.; MIYAIRI, K.; OKUNO, T.; KATO,H. Active-site architeture of endopolygalacturonase l from Stereum purpureum revelead by crystal structures in native and ligand-bound forms at atomic resolution. *Biochemistry,* v. 41, p. 6651-6659, 2002.

[151] EDELSBRUNNER, H.; MÜCKE, E. P.; Simulation of simplicity: a technique to cope with degenerate cases in geometric algorithms. *ACM Trans. Graph.*v. 9, p. 66-104, 1990.

[152] BARBER, C.B.; DOBKIN, D. P.; HUHDANPAA, H. The Quick Hull algorithm for convex hulls. *ACM Trans. Math. Soft.,* v. 22, p. 469-483, 1996.

[153] BANDYOPADHYAY, D.; SNOEYINK, J. Almost-Delaunay simplices: nearest neighbor relations for imprecise points. *ACM-SIAM Symp. on Disc. Algorith.,* session 5A, p. 410-419, 2004.

[154] DUNCAN, C. A.; GOODRICH, M. T.; RAMOS, E. A. Efficient approximation and optimization algorithms for computational metrology. *ACM-SIAM Symp. on Disc. Algorit.,* p. 121-130, 1997.

[155] OCCAM'S RAZOR. In: WIKIPEDIA, The Free Encyclopedia. Disponível em <http://en.wikipedia.org/wiki/Occam's_Razor>. Acesso em janeiro de 2008.

[156] SCHWARZ, G. Estimating the dimension of a model. *The Annals of Statistics,* v. 6, p. 461-464, 1978.

[157] BURNHAM, K. P.; ANDERSON, D. R. Multimodel inference: understanding AIC e BIC in model selection. *Sociological Methods Research,* v. 33, p. 261-304, 2004.

[158] WELCH, B. L. The generalization of "students" problem when several different population variances are involved. *Biometrika,* v. 34, p. 28-35, 1947.

[159] MURZIN, A. G.; BRENNER, S. E.; HUBBARD, T.; CHOTHIA, C. SCOP: a structural classification of proteins database for investigation of sequences and structures. *Journal of Molecular Biology,* v. 247, p. 536-540, 1995.

[160] DUPUIS, F.; SADOC, J. F.; JULIEN, R.; ANGELOV, B.; MORNON, J. P. Voro3D: 3D Voronoi tessellation applied to proteins structures. *Bioinformatics,* v. 21, p. 1715-1716, 2005.

[161] ANGELOV, B.; SADOC, J. F.; JULIEN, R.; SOYER, A.; MORNON, J. P. Nonatomic solvent-driven Voronoi tessellation of proteins: an open tool to analyze protein folds. *Protein: Structure, Function and Bioinformatics,* v. 49, p. 446-456, 2002.

[162] KUNTZ, L. D.; CRIPPEN, G. M. Protein Densities. *International Journal of Peptide Research,* v. 13, p.

223-228, 1978.

9. ANEXO A

9.1 Polinômios Habeschianos

Prólogo:

Esta tese passou por diversas mudanças de fases. Da proposta original apresentada como projeto para o ingresso no programa de PhD em Bioinformática na UFMG, pouca coisa restou. Durante o embate com o projeto e na medida que eu aprofundava na literatura, mudanças de rumo foram necessárias para afinar a tese ao objetivo formal de produzir algo inédito e com alguma relevância. Não é nada fácil fazer ciência na era "ômica". Já nas fases finais, para por um pouco de ordem no caos, eu resolvi registrar de forma livre num diário tudo que vinha pesquisando. O resultado disso foram 1107 páginas de anotações, 1214 figuras, totalizando 1.5 GB de dados. Devo dizer que uma boa parte do tempo do meu doutorado foi investido na busca de uma equação geral que descrevesse o comportamento das distribuições cumulativa e de densidade para o número de contatos em função das distâncias, dado um sistema de esferas distribuídas de forma aproximadamente uniforme num espaço Euclideano. Como um entusiasta da matemática, almejava uma abordagem que fosse mais dedutiva e que pudesse embasar os estudos indutivos que eu vinha fazendo desde o início do projeto. Cheguei a rascunhar folhas e mais folhas com várias tentativas de dedução, sem muito sucesso. Foi quando o físico Raul Habesch, amigo de longas eras, passou comigo uma semana histórica em Belo Horizonte. Trancafiados em meu escritório, depois de muitos cafés e garatujas alfanuméricas, conseguimos ao final chegar a um conjunto de equações candidatas, que vem se ajustando bem aos dados experimentais. Para nosso espanto, a dedução descambou em singelos polinômios. Como nós ainda não tivemos condições de verificar a originalidade desse achado, julgamos que seria uma melhor estratégia fazer uma publicação mais cuidadosa à parte, depois de uma rigorosa revisão da literatura. Por isso, ela não saiu no artigo submetido para a *Proteins: Structure, Function and Bioinformatics*. Mas, como foi um dos resultados dessa tese, fica aqui o seu registro para a posteridade. Se ela por acaso não trouxer originalidade alguma, valeu pela experiência de viver a indescritível emoção de deduzir em equação. É justo dizer que o Raul empreendeu, com o brilho lógico que lhe é peculiar, a maior parte do nosso esforço dedutivo, de forma que o mérito da descoberta certamente cabe a ele. Assim, em sua homenagem, eu estou chamando essas equações de **Polinômios Habeschianos**. Detalhes sórdidos dessa aventura estão registrados no meu diário.

Distribuição de distâncias em lattices finitos

Definição do problema

Queremos saber, inicialmente para uma área finita de um *lattice* 2D e posteriormente um volume finito de um *lattice* 3D, quais as distâncias possíveis entre pontos do *lattice*, e com que freqüências essas distâncias aparecem.

Deve ser evitada a seguinte redundância na contagem das distâncias: Dados dois pontos P_1 e P_2, se contabilizarmos a distância de P_1 a P_2 , não contabilizaremos de P_2 a P_1.

Representando o problema 2D

Representaremos o *lattice* 2D a partir de um ponto *P(0,0)*, uma base de vetores $\hat{e}_{0,1}$ e $\hat{e}_{1,0}$, e coordenadas (*a,b*) pertencentes a N^2 , de forma que qualquer ponto do *lattice* possa ser descrito como

$$P(a,b) = P(0,0) + a.\ \hat{e}_{1,0} + b.\ \hat{e}_{0,1}$$

A transição de um ponto para outro é feita por um vetor, chamado deslocamento:

$$P(a+i,b+j) = P(a,b) + i.\ \hat{e}_{1,0} + j.\ \hat{e}_{0,1}$$

Para simplificar a notação, $(i,j) \equiv i.\ \hat{e}_{1,0} + j.\ \hat{e}_{0,1}$.

Então o deslocamento fica:

$$P(a+i,b+j) = P(a,b) + (i\ ,\ j) \qquad [1]$$

Até aqui, pura álgebra linear.

Os limites do *lattice* são: $0 <= a <= m$, $0 <= b <= n$

Portanto, os deslocamentos são limitados a $-m < i < m$, $-n < i < n$

Primeira Regra da Contagem:

Se o deslocamento (i , j) for contado, então o deslocamento (-i , -j) **não** deve ser contado.

Justificativa: o vetor (-i , -j), aplicado em *P(a+i,b+j),* liga exatamente os mesmos pontos que o vetor (i, j) aplicado em *P(a,b)* , portanto essa distância só deverá ser contada em um dos casos.

Segunda Regra da Contagem:

A distância entre um ponto e o próprio ponto não deve ser contada.

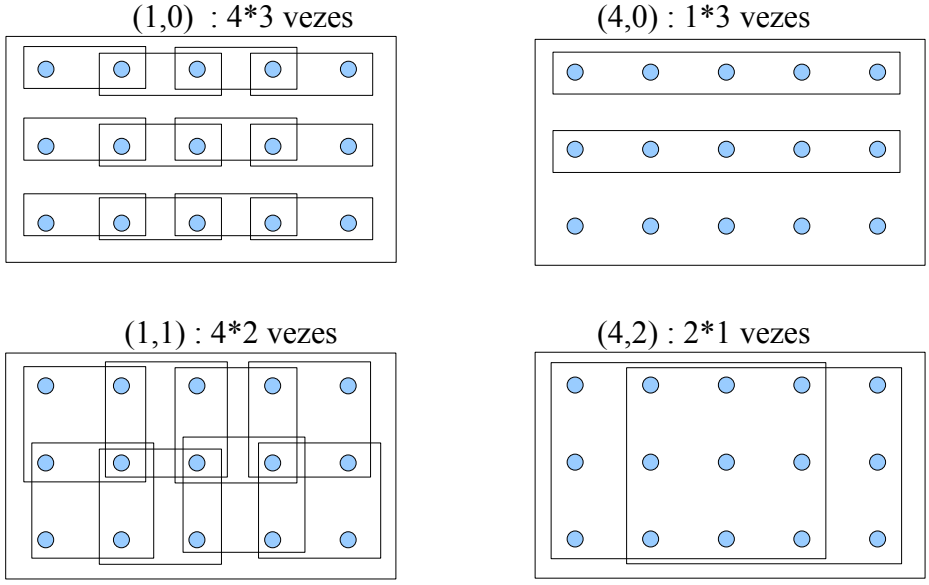
O *lattice* é um espaço discreto, assim como o espaço dos vetores de deslocamento. Portanto, deve haver um número finito de possibilidades de deslocamento, em um subespaço finito. Esse número é dado pela combinação de m.n elementos tomados 2 a 2 :

$$N_d = (m.n)! / ((m.n-2)!.2) = (m.n)(m.n-1)/2 \qquad [2]$$

Quantas vezes um determinado deslocamento pode ser aplicado no *lattice*?

Pensemos no caso limite: o deslocamento (0,0) pode ser aplicado uma vez em cada ponto, portanto *m* vezes em uma linha e *n* vezes em uma coluna , portanto *m.n* vezes. (obs: por enquanto, não foi pressuposto que $\hat{e}_{0,1}$ e $\hat{e}_{1,0}$ sejam ortogonais. Quando falamos em "linha" e "coluna", podem ser direções oblíquas). Porém, pela segunda regra da contagem, desconsideraremos a distância (0,0)

Já o deslocamento (1,0), que "ocupa" uma posição na linha, só pode ocorrer *m – 1* vezes em uma linha, mas continua *n* vezes nas colunas. O deslocamento (m-1, 0) só ocorre uma vez em cada linha (ligando o primeiro ao último ponto), nas *n* linhas. (fig 1)



Podemos concluir que o deslocamento (i,j) ocorre $F_{i,j} = (m-i).(n-j)$ vezes. Caso o deslocamento tenha sido no sentido de *i* ou *j* negativo, o que vale é o valor absoluto do deslocamento.

$$F_{i,j} = (m-|i|).(n-|j|)$$

Vamos trabalhar com o caso particular de um *lattice* quadrado: $m = n$.

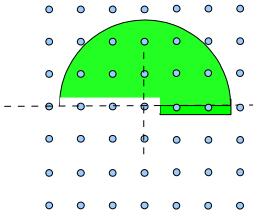
$$F_{i,j} = (n-|i|).(n-|j|) \qquad [3a]$$

E, pela Segunda Regra da Contagem:

$$F_{0,0} = 0 \qquad [3b]$$

Distribuição

Queremos obter a função *f(r)* que dá o número de distâncias menores ou iguais a *r*.



Ela deve corresponder a soma de todas as freqüências de distâncias compatíveis com a condição acima.

Considerando as freqüências dadas por [3a] e [3b],

$$f(r)=\sum_{i=-n+1}^{n-1}\sum_{j=1}^{n-1}F(i,j)\ +\sum_{i=1}^{n-1}F(i,0) \qquad \text{na condição } \sqrt{(i^2+j^2)}\leqslant r \qquad [4]$$

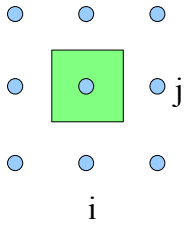
A região -n < *j* < *0* está excluída pela Primeira Regra da Contagem. Da mesma forma, a região {j = 0 , i < 0}, que é oposta a {j = 0 , i >= 0}

Função contínua

É mais simples trabalhar com o contorno curvo usando integrais do que somatórios. Procuremos então uma aproximação, por integrais, do somatório das freqüências. Consideremos uma "célula discreta", quadrada, ao redor de cada ponto (i,j), delimitada por

{ -r_0/2 <= x <= +r_0/2 , -r_0/2 <= y <= +r_0/2 }. Procuramos uma função contínua *Fc*(*x,y*) tal que o valor da integral sobre a célula seja igual a F(i,j) :

$$F(i,j)=\int_{(j-1/2)r_0}^{(j+1/2)r_0}\int_{(i-1/2)r_0}^{(i+1/2)r_0}Fc(x,y)\,dx\,dy$$



Prosseguiremos o cálculo no primeiro quadrante. O valor da integral será o mesmo para os outros quadrantes.

Pela expressão [3a]:

$$(n-i).(n-j)=\int_{(j-1/2)r_0}^{(j+1/2)r_0}\int_{(i-1/2)r_0}^{(i+1/2)r_0}Fc(x,y)\,dx\,dy$$

Suponhamos ainda que exista uma solução na forma *Fc(x,y) = Fx(x).Fy(y)* .

$$(n-i).(n-j)=\int_{(i-1/2)r_0}^{(i+1/2)r_0}Fx(x)\,dx.\int_{(j-1/2)r_0}^{(j+1/2)r_0}Fy(y)\,dy$$

Separando as variáveis,

$$(n-i)=\int_{(i-1/2)r_0}^{(i+1/2)r_0} Fx(x)\,dx \qquad (n-j)=\int_{(j-1/2)r_0}^{(j+1/2)r_0} Fy(y)\,dy$$

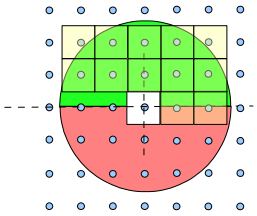
achamos como soluções:

$$Fx(x)=\left(n-\frac{x}{r_0}\right).\frac{1}{r_0} \qquad Fy(y)=\left(n-\frac{y}{r_0}\right).\frac{1}{r_0}$$

$$Fc(x,y)=\left(n-\frac{x}{r_0}\right)\left(n-\frac{y}{r_0}\right).\frac{1}{r_0^2} \quad \text{ou} \quad Fc(r,\theta)=\left(n-\frac{r}{r_0}\cos\theta\right)\left(n-\frac{r}{r_0}sen\,\theta\right).\frac{1}{r_0^2} \quad [5]$$

Para os outros quadrantes, $Fc(x,y) = Fc(|x|,|y|)$

A integral sobre a célula passa a representar, no espaço contínuo, o valor de *F(i,j)* no espaço discreto. O somatório das freqüências pode ser aproximado, então, por:



$$f(r)=\sum_{i=-n+1}^{n-1}\sum_{j=1}^{n-1} F(i,j) \;+\sum_{i=1}^{n-1} F(i,0) \qquad \text{na condição } \sqrt{(i^2+j^2)}\leqslant r$$

$$=2\int_0^r\int_0^{\pi/2} Fc(r,\theta)\,r\,d\theta\,dr \;-f_0\;+E_n(r)$$

Mais uma vez, reduzimos o problema ao primeiro quadrante, por simetria. A função $E_n(r)$ fornece o erro cometido na aproximação, que é devido ao "corte" irregular das células básicas pela fronteira circular. A constante f_0 aparece devido à Segunda Regra da Contagem, e corresponde à metade da contribuição do ponto (0,0) se ele tivesse freqüência F(0,0) = (n-i) (n-j) = n^2 . Esse valor deve ser descontado porque na definição de Fc(i,j) não foi levada em conta a segunda regra, e a integral inclui metade da célula básica do ponto (0,0).

Integrando:

$$f(r)=f_h(r)\;-f_0\;+E_n(r)$$

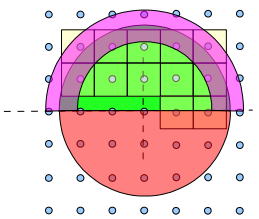
$$f_h(r)=\frac{1}{4}\left(\frac{r}{r_0}\right)^4-\frac{4}{3}n\left(\frac{r}{r_0}\right)^3+\frac{\pi}{2}n^2\left(\frac{r}{r_0}\right)^2 \qquad \text{para} \quad r_0 <= r < n.r_0 \quad [6]$$

$$f_0=1/2\,n^2$$

Na região $0 < r < r_0$, por definição $f(r) = 0$.

Fora da região $0 < r < n.r_0$, f(r) deixa de ser uma função polinomial.

Uma estimativa do erro *En(r)* pode ser feita considerando que, em média, metade da área das células básicas na fronteira ficarão de fora. Consideramos como área da fronteira a faixa entre
$r + r_0/2$ e $r - r_0/2$:



$$|E_n(r)| \;<\; 2\int_{r-r_0/2}^{r+r_0/2}\int_0^{\pi/2}\frac{1}{2}Fc(r,\theta)\,r\,d\theta\,dr$$

$$|E_n(r)| \;<\; \left(\frac{r}{r_0}\right)^3-4n\left(\frac{r}{r_0}\right)^2+\left(\frac{1}{4}+\pi n^2\right)\left(\frac{r}{r_0}\right)-\frac{1}{3}n$$

Interessa também achar a função *D(r)* , correspondente à **densidade** de valores, tal que

$$f(r)=\int_0^r D(R)\,dR \;+\; cte$$

Que é a derivada parcial de *f(r)* em relação a *r* .

$$D(r)=\frac{1}{r_0}\left[\left(\frac{r}{r_0}\right)^3-4n\left(\frac{r}{r_0}\right)^2+\pi n^2\left(\frac{r}{r_0}\right)\right]$$

Também nos interessa a função diferença finita $Diff_a(r) = f(r+a) - f(r)$:

$$Diff_a(r)=\frac{a}{r_0}.\left(\frac{r}{r_0}\right)^3+\left[-4n\frac{a}{r_0}+\frac{3}{2}\left(\frac{a}{r_0}\right)^2\right]\left(\frac{r}{r_0}\right)^2+\left[\left(\frac{a}{r_0}\right)^3-4n\left(\frac{a}{r_0}\right)^2+\pi n^2\left(\frac{a}{r_0}\right)\right]\left(\frac{r}{r_0}\right)+$$
$$+\left[\frac{\pi}{2}n^2\left(\frac{a}{r_0}\right)^2-\frac{4}{3}n\left(\frac{a}{r_0}\right)^3+\frac{1}{4}\left(\frac{a}{r_0}\right)^4\right]$$

No limite, $\lim_{a\to 0}\frac{Diff_a(r)}{a}=D(r)$

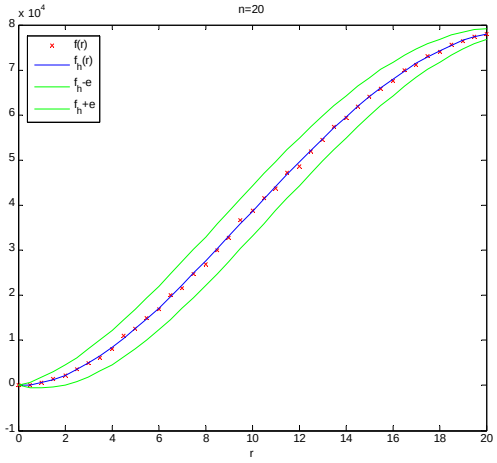


Figura A1: Ajuste dos polinômios Habeschianos à conjuntos de esferas aleatoriamente distribuídas num *lattice* 2D de lado n=20 Pelos gráficos, notamos que a estimativa de erro parece ser superestimada.

Lattice 3D

No caso de um cubo com *n* pontos na aresta, acrescentamos uma dimensão ao problema. Os pontos do *lattice* são:

$$P(a+i,b+j,c+k) = P(a,b,c) + i.\ \hat{e}_{1,0,0} + j.\ \hat{e}_{0,1,0} + k.\ \hat{e}_{0,0,1} = P(a,b,c) + (i, j, k)$$

As freqüências:

$$F_{i,j,k} = (n-|i|).(n-|j|).(n-|k|) \qquad [3c]$$

$$F_{0,0,0} = 0 \qquad [3d]$$

As distâncias:

$$d(i,j,k) = d_0.\mathrm{sqrt}(i^2 + j^2 + k^2)$$

Para montar a distribuição discreta, desconsideramos a região k<0 , e também {k=0 , j<0}, e {k=0 , j=0 , i<0 }, pela Segunda Regra da Contagem

$$f(r) = \sum_{i=-n+1}^{n-1} \sum_{j=-n+1}^{n-1} \sum_{k=1}^{n-1} F(i,j,k) + \sum_{i=-n+1}^{n-1} \sum_{j=1}^{n-1} F(i,j,0) + \sum_{i=+1}^{n-1} F(i,0,0)$$

na condição $d \leqslant r$

A aproximação para função contínua :

$$Fc(x,y,z) = \left(n - \frac{x}{r_0}\right)\left(n - \frac{y}{r_0}\right)\left(n - \frac{z}{r_0}\right).\frac{1}{r_0^3}$$

$$Fc(r,\theta,\varphi) = \left(n - \frac{r}{r_0}\cos\varphi\, sen\,\theta\right)\left(n - \frac{r}{r_0} sen\,\varphi\, sen\,\theta\right)\left(n - \frac{r}{r_0}\cos\theta\right).\frac{1}{r_0^3}$$

$$f(r) = \int_0^{nr_0} \int_{\substack{-nr_0 \\ d \leqslant r}}^{nr_0} \int_{-nr_0}^{nr_0} Fc(x,y,z)\, dx\, dy\, dz \ - f_0$$

Aqui, $f_0 = 1/2\ n^3$.

Em coordenadas esféricas, e reduzindo ao primeiro octante:

$$f(r) = 4 \int_0^r \int_0^{\pi/2} \int_0^{\pi/2} Fc(r,\theta,\varphi)\, r^2\, sen\,\theta\, d\varphi\, d\theta\, dr \ - f_0 \ + E_n(r)$$

Integrando:

$$f_h(r)=-\frac{1}{12}\left(\frac{r}{r_0}\right)^6+\frac{4}{5}n\left(\frac{r}{r_0}\right)^5-\frac{3}{4}\pi n^2\left(\frac{r}{r_0}\right)^4+\frac{2}{3}\pi n^3\left(\frac{r}{r_0}\right)^3 \quad \text{para} \quad r_0 <= r < n.r_0 \ [7]$$

$$f_0=\frac{1}{2}n^3$$

$f(r) = 0$, para $0 < r < r_0$

A estimativa de erro:

$$E_n(r) \ < \ 4\int_{r-r_0/2}^{r+r_0/2}\int_0^{\pi/2}\int_0^{\pi/2}\frac{1}{2}Fc(r,\theta,\varphi)r^2\,sen\theta\,d\varphi\,d\theta\,dr$$

$$E_n(r) \ < \ -\frac{1}{2}\left(\frac{r}{r_0}\right)^5+4n\left(\frac{r}{r_0}\right)^4+4\left(-\frac{5}{48}-\frac{3}{4}\pi n^2\right)\left(\frac{r}{r_0}\right)^3+$$
$$+4\left(\frac{\pi}{2}n^3+\frac{1}{2}n\right)\left(\frac{r}{r_0}\right)^2+4\left(-\frac{3}{16}\pi n^2-\frac{1}{128}\right)\left(\frac{r}{r_0}\right)+\frac{\pi}{6}n^3+\frac{1}{20}n$$

A densidade:

$$D(r)=\frac{1}{r_0}\left[-\frac{1}{2}\left(\frac{r}{r_0}\right)^5+4n\left(\frac{r}{r_0}\right)^4-3\pi n^2\left(\frac{r}{r_0}\right)^3+2\pi n^3\left(\frac{r}{r_0}\right)^2\right]$$

A Diff :

$$Diff_1(r)=-\frac{1}{2}\left(\frac{a}{r_0}\right)\left(\frac{r}{r_0}\right)^5+\left[4n\left(\frac{a}{r_0}\right)-\frac{5}{4}\left(\frac{a}{r_0}\right)^2\right]\left(\frac{r}{r_0}\right)^4+$$
$$+\left[-3n^2\pi\left(\frac{a}{r_0}\right)+8n\left(\frac{a}{r_0}\right)^2-\frac{5}{3}\left(\frac{a}{r_0}\right)^3\right]\left(\frac{r}{r_0}\right)^3+$$
$$+\left[2\pi n^3\left(\frac{a}{r_0}\right)-\frac{9}{2}\pi n^2\left(\frac{a}{r_0}\right)^2+8n\left(\frac{a}{r_0}\right)^3-\frac{5}{4}\left(\frac{a}{r_0}\right)^4\right]\left(\frac{r}{r_0}\right)^2+$$
$$+\left[2n^3\pi\left(\frac{a}{r_0}\right)^2-3n^2\pi\left(\frac{a}{r_0}\right)^3+4n\left(\frac{a}{r_0}\right)^4-\frac{1}{2}\left(\frac{a}{r_0}\right)^5\right]\left(\frac{r}{r_0}\right)+$$
$$-\frac{1}{12}\left(\frac{a}{r_0}\right)^6+\frac{4}{5}n\left(\frac{a}{r_0}\right)^5-\frac{3}{4}n^2\pi\left(\frac{a}{r_0}\right)^4+\frac{2}{3}n^3\pi\left(\frac{a}{r_0}\right)^3$$

Gráfico para solução 3D:

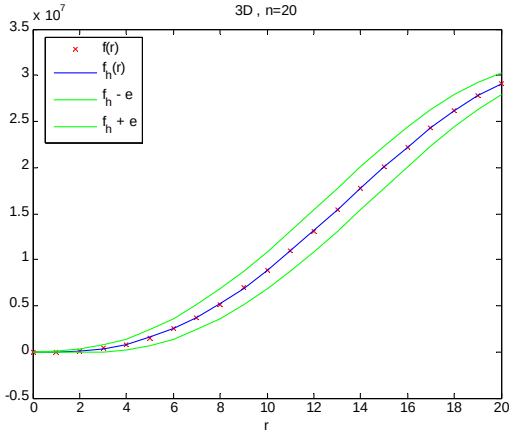


Figura A2: Ajuste dos polinômios Habeschianos à conjuntos de esferas aleatoriamente distribuídas num *lattice* 3D de lado n=20.

Ajustes em Outros Contextos

Mostramos a seguir o ajuste aos polinômios Habeschianos em outros agrupamentos de esferas:

Lattice Cúbico Randômico (PT11.pdb)

Conjunto de 1000 esferas, separadas inicialmente 1.0.unidade de distância num *lattice* cúbico de 10 esferas de lado. Sobre as coordenadas de cada ponto foi aplicado um ruído randômico de ± 1.0,de igual probabilidade (distribuição uniforme), efeito que estamos chamando de “termalização”.

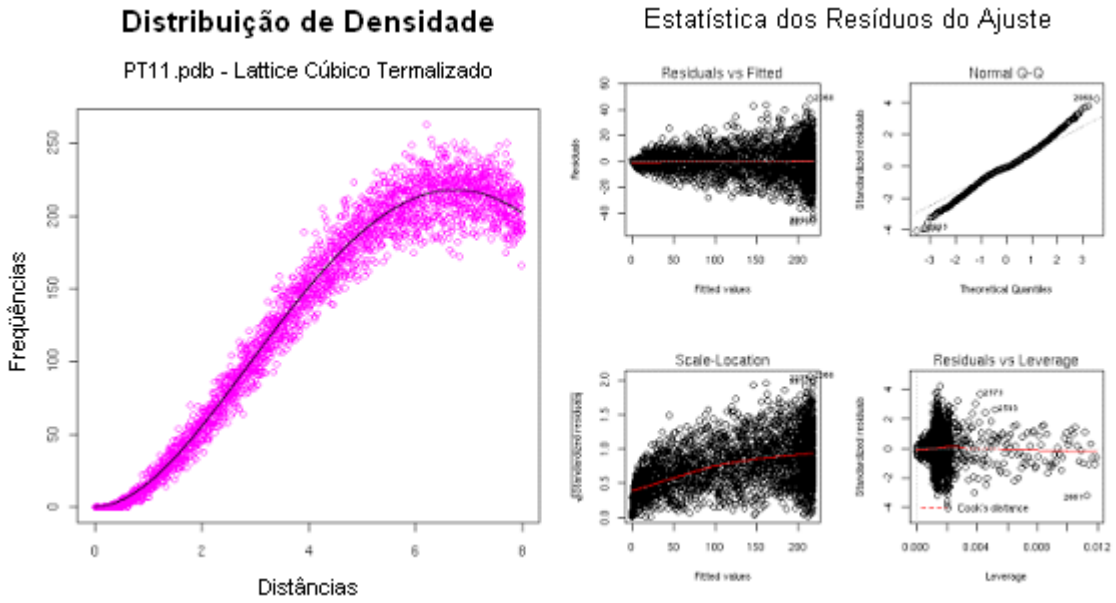


Figura A3: Esquerda: Gráficos do ajuste à distribuição de densidade para o aglomerado de esferas PT11.pdb. Direita: Gráfico com a estatística dos resíduos do ajuste.

Proteína: Mioglobina(1BZR.pdb)

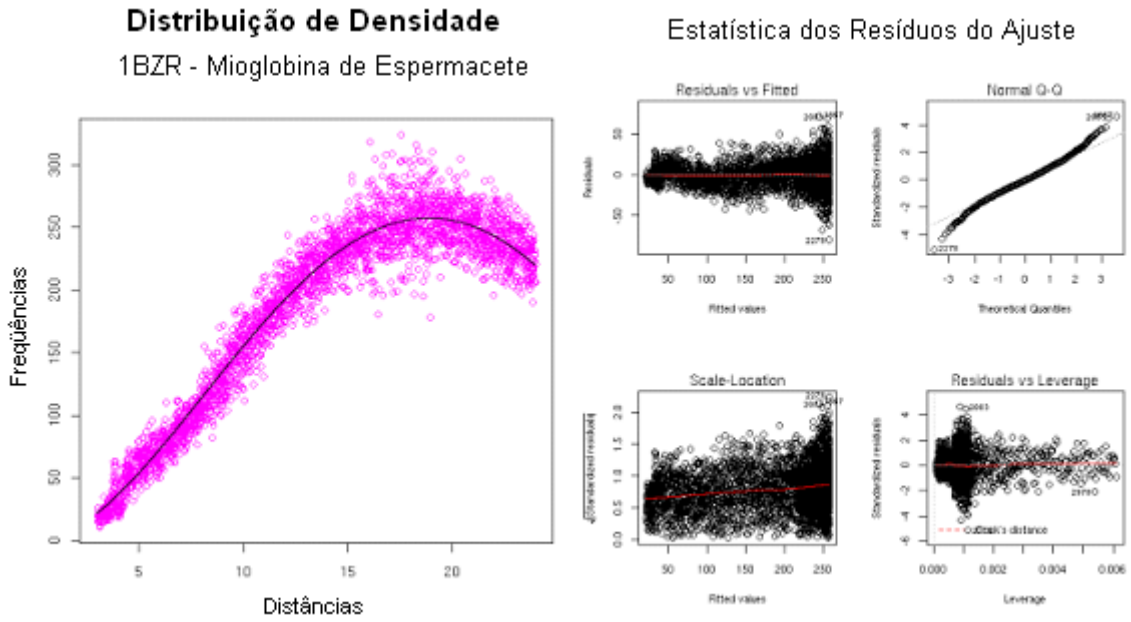


Figura A4: Esquerda: Gráficos do ajuste à distribuição de densidade para a mioglobina de Espermacete (Physeter catodon) Direita: Gráfico com a estatística dos resíduos do ajuste. Distância em Angstrons.

Proteína: Inibidor de Protease(1JIW-I.pdb)

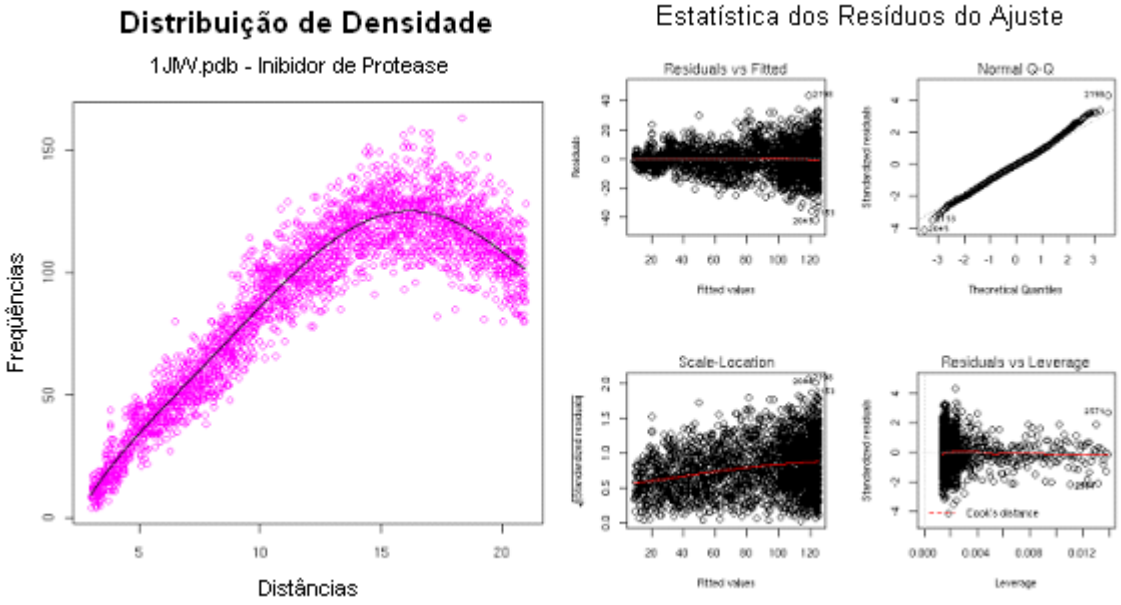


Figura A5: Esquerda: Gráficos do ajuste à distribuição de densidade para o inibidor de protease da bactéria (Pseudomonas aeruginosa) Direita: Gráfico com a estatística dos resíduos do ajuste. Distância em Angstrons.

Estrelas em Hipparcos (H002.pdb)

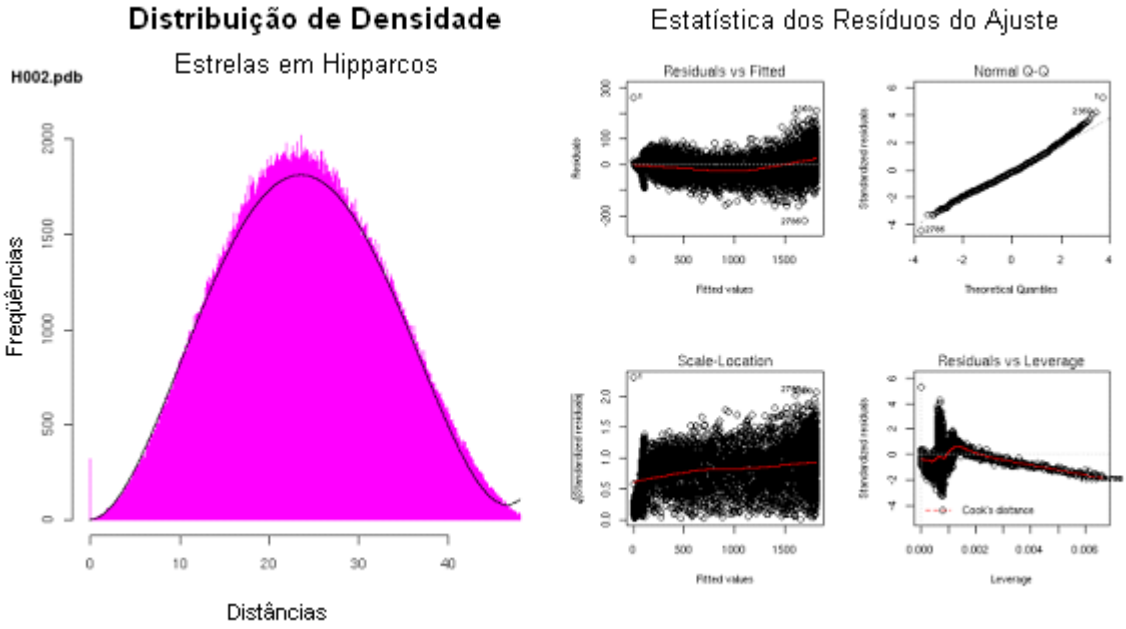


Figura A6: Subconjunto da base de dados de coordenadas estelares HYG(2.0) localizada em http://astronexus.com/node/34, com 3023 estrelas em Hipparcos. Distâncias em *parsecs*.

CONJECTURA SSH (Santoro, Silveira, Habesch):

Enunciando Provisório:

"*A distribuição de distâncias envolvendo um aglomerado de pontos suficientemente distribuídos num espaço Euclideano pode ser aproximada por polinômios Habeschianos*"

Figura A7: Representação por esferas do *lattice* cúbico "termalizado" (PT11.pdb), do aglomerado estelar em Hipparcos (H001.pdb), da proteína mioglobina de baleia (1BZR.pdb) e da proteína inibidor de protease da bactéria *Pseudomonas aeruginosa* (1JIW:I.pdb)