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.