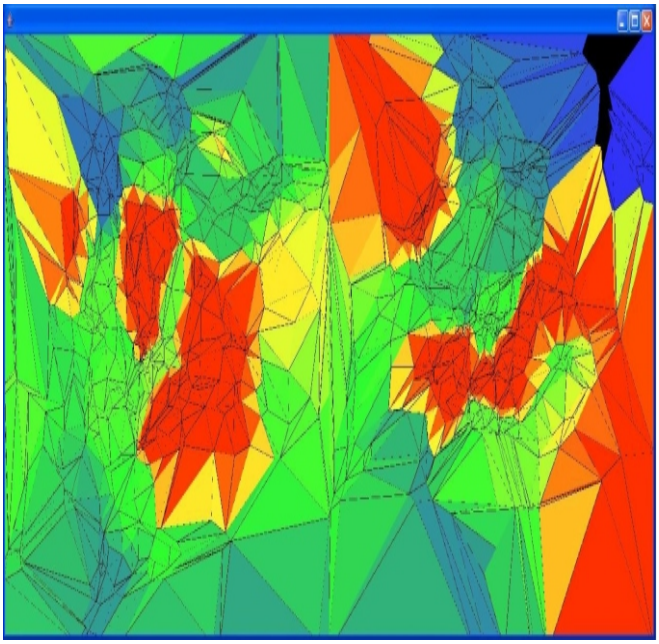
Comunicado Técnico 78

Dezembro, 2006
Campinas, SP



ISSN 1677-8464

Projeções de Superfície 3D no Plano para Análise de Interfaces Proteicas através do Sting

Marcelo Gonçalves Narciso[1]
Michel Eduardo Beleza Yamagishi[2]
Thiago Quinaglia[3]
Edgard Henrique dos Santos[4]
Fábio Danilo Vieira[5]
José Gilberto Jardine[6]
Ivan Mazoni[7]
Paula Regina Kuser Falcão[8]
Goran Neshich[9]

Uma área que tem crescido muito nos últimos anos é a Bioinformática. De acordo com definição encontrada na Wikimedia Foundation (2006), a Bioinformática combina conhecimentos de Química, Física, Biologia e Informática para processar dados biológicos. Assim, para tratar os dados existentes é importante desenvolver software para identificar genes, prever a configuração tridimensional de proteínas, identificar inibidores de enzimas, organizar e relacionar informação biológica ou simular células biológicas. O software Sting (Embrapa Informática Agropecuária, 2006b) é um software voltado para a bioinformática estrutural e, de forma geral, faz análise estrutural de proteínas e mostra de forma especializada qualquer proteína que está catalogada no Protein Data Bank (Research Collaboratory for Structural Bioinformatics, 2006). O PDB, como é conhecido, constitui-se em um repositório de arquivos com descrição de vários parâmetros relacionados a proteínas. Diversos tipos de análises podem ser feitos com este software.

Um dos recursos do Sting é a análise, de forma visual, de como algumas grandezas físico-químicas estão presentes na interface proteica. A proteína é mostrada em 3 dimensões (3D). Cada região da cadeia proteica é colorida com uma cor diferente, representando a variação de características físico-químicas (potencial eletrostático, eletrofobicidade, etc.) na cadeia.

Para visualizar todas as características de uma cadeia proteica é necessário ficar girando a superfície, visto que tem o fundo da superfície e também os lados. Para facilitar a análise de grandezas na cadeia proteica, faz-se necessário mapear a superfície em um plano (2D) sem que esta venha perder as características de seu formato original. A superfície

[1] Doutor em Computação, Pesquisador da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: narciso@cnptia.embrapa.br)
[2] Doutor em Matemática Aplicada, Pesquisador da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: michel@cbi.cnptia.embrapa.br)
[3] Bacharel em Ciência da Computação, Bolsita da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: quinaglia@cbi.@cnptia.embrapa.br)
[4] Bacharel em Ciência da Computação, Analista da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: edgard@cbi.@cnptia.embrapa.br)
[5] Bacharel em Tecnologia de Informação, Analista da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: fabio@cbi.@cnptia.embrapa.br)
[6] Ph.D. em Engenharia de Alimentos, Pesquisador da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (jardine@cnptia.embrapa.br)
[7] Bacharel em Ciência da Computação, Analista da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: ivan@cbi.cnptia.embrapa.br)
[8] Doutora em Cristalografia de Proteínas, Pesquisadora da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: paula@cbi.cnptia.embrapa.br)
[9] Ph.D. em Biofísica, Pesquisador da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (Neshich@cbi.cnptia.embrapa.br)

deverá estar mapeada em um retângulo para facilitar a visualização de qualquer grandeza desejada e também para se comparar com outras cadeias proteicas similares.

Para se fazer esta transformação, 3D para 2D, parte-se da triangularização da superfície 3D da interface. Esta triangularização nada mais é do que a coleção de todos os triângulos que compõem a superfície 3D. Cada triângulo é identificado pelos seus três vértices, com coordenadas espaciais. Um algoritmo que transforma as coordenadas destes vértices em 3D e faz o mapeamento destes para 2D foi feito de tal maneira que, localmente, as distâncias triangulares sejam preservadas (minimiza a distorção). Este trabalho descreve este algoritmo e também alguns resultados, que podem ser vistos no Sting (Embrapa Informática Agropecuária, 2006b).

Material e Métodos

O problema de projetar uma superfície 3D no plano em si não é novo, e há na literatura diversos algoritmos desenvolvidos com tal finalidade (Haker et al., 2000). Entretanto, a maioria desses algoritmos requer condições extremamente restritivas, como, por exemplo, que as superfícies sejam topologicamente equivalentes a esferas, isto é, que exista um ponto de onde qualquer segmento de reta partindo dele intercepte a superfície em um único ponto. Em outras palavras, que a superfície seja uma função. Obviamente, superfícies muito intrincadas dificilmente satisfazem esse tipo de condição. Em particular, para a interface de contato entre cadeias proteicas, essa condição não é realística. Se por um lado deseja-se preservar o máximo possível as características da superfície, por outro lado, as deformações advindas da projeção sobre o plano são inevitáveis. É necessário, nesses casos, obter um compromisso entre a preservação das características da superfície e as deformações. No caso da análise das interfaces proteicas, as características locais são mais importantes que as globais. Dessa forma, buscou-se um algoritmo que, embora permitisse deformações globais, preservasse localmente as características da superfície. É fácil entender porque se deseja preservar as características locais. Tome um exemplo onde a interface de contato entre duas cadeias proteicas seja caracterizada por um modelo chave-fechadura, ou seja, uma cadeia agindo como chave (com uma protuberância) e a outra como fechadura (com uma abertura). Para que uma análise dessa interface seja realizada satisfatoriamente, tanto a protuberância quanto a abertura, devem ser facilmente identificadas na projeção no plano.

Um algoritmo que atende a essas peculiaridades dessa aplicação é denominado *Shape Preserving Algorithm* (Floater, 1997). Este algoritmo gera as projeções da superfície no plano, de forma a manter ao formato inicial dos triângulos. Este algoritmo precisa que a superfície tridimensional seja aberta e só pode existir uma abertura. Como nem todas as superfícies das interfaces proteicas têm apenas uma abertura (podem ser fechadas ou ter mais de uma abertura) foi necessário fazer adaptações ao algoritmo, o qual foi feito em linguagem C e está disponível em Embrapa Informática Agropecuária (2006a). Sem estas adaptações, poucas proteínas poderiam ser mapeadas no plano, visto a diversidade de interfaces proteicas que ocorrem na natureza.

O algoritmo de Floater (1997) busca obter coordenadas (x,y) tal que estas descrevam a projeção de uma superfície em três dimensões (x1, y1, z1). A Fig. 1 a seguir mostra como é uma representação de uma superfície e a projeção desta em um plano.

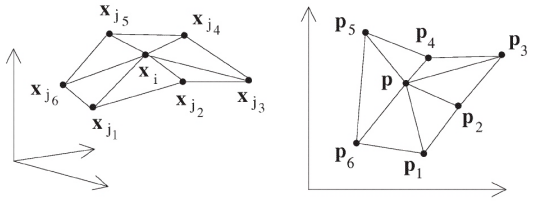


Fig. 1. Representação de uma superfície em 3D e em 2D.

Na Fig. 1, Xj_1, Xj_2, ..., Xj_6 e X_i representam pontos no espaço tridimensional (eixos x, y e z). P_1, P_2, ..., P_6 e P representam a projeção da superfície no plano. O algoritmo a seguir busca uma forma de transformar os pontos da superfície tridimensional em pontos do plano. Resumidamente, o algoritmo irá montar uma matriz A tal que

$$AX = B1 \text{ e } AY = B2,$$

onde X e Y são vetores cujos elementos, após determinados, formarão os pontos (x_i,y_i) pertencentes ao plano. B1 e B2 são vetores com elementos determinado a partir dos pontos da borda da superfície. A matriz **A** contém os coeficientes numéricos para se obter X e Y. O algoritmo que está descrito a seguir é responsável por obter os elementos da matriz A. Para melhor entendimento, levar em consideração a Fig. 1.

1. Seja S o número de pontos que compõe a superfície. Determinar quais pontos são exteriores (fazem parte da borda da superfície, isto é, da abertura) e quais são os interiores (não fazem parte da borda). Seja M o número de pontos exteriores e N o número de pontos interiores. Assim, $N + M = S$.

Os pontos exteriores são aqueles que não são envolvidos por pontos tais que estes formem um ciclo. Para exemplificar, na Fig. 1, P_1, P_2, ..., P_6 representam os pontos exteriores (observe que fazem um caminho fechado ou ciclo) e P representa o ponto interior.

2. Para todos os pontos interiores, que pertencem ao conjunto T de pontos da superfície, e o conjunto de pontos J que se conectam a algum elemento i do conjunto de pontos T, determinar uma matriz $_{\lambda ik}$, $i \in T$, $k \in J$, tal que a multiplicação dos elementos desta matriz pelo valor das coordenadas dos pontos que pertencem a J venham a gerar as coordenadas do ponto i.

Para ilustrar o que foi dito, seja a Fig. 1 e seja um ponto interior P do plano tal que este tenha ligações com os pontos

P_1, P_2, .. P_6. Estes pontos têm coordenadas no eixo XY. Para se determinar as coordenadas, supõe-se que o ponto interior P esteja na origem e o ponto P_1 na reta $y = 0$ e $x > 0$. Como são conhecidas as correspondentes coordenadas de P e P_1 no espaço, X_i e Xj_1, basta saber a distância **d** entre estes pontos que se terá as coordenadas de P_1 no plano, isto é, (**d**,0). Visto que tem-se as coordenadas de P e P_1 no plano, (0,0) e (d,0) respectivamente, pode-se, a partir destes pontos, determinar os demais pontos do plano.

Assim, sabendo-se a distância de P a P_2, isto é, a distância de X_i a Xj_2, e sendo **d1** esta distância, as coordenadas de P_2 seriam: $(d1\cos\theta, d1\,\mathrm{sen}\theta)$, θ é o ângulo entre as retas que passam por X_i e Xj_1, e por X_i e Xj_2. Assim, todos os demais pontos são determinados, bastando saber o ângulo entre as retas no espaço e o valor da distância de X_i a um ponto qualquer Xj_k, $k \in J$. Com os valores de P_1, P_2, .. P_6, calculados, e considerando-se as suas coordenadas como Px_k,e Py_k, $k \in J$, matematicamente, λ_{ik}, para um dado i, seria a solução das seguintes equações:

$$Pxi = \Sigma\, \lambda ik * Px_k,\ k \in J.$$

$$Pyi = \Sigma\, \lambda ik * Py_k,\ k \in J.$$

$$\Sigma\, \lambda is = 1,\ s \in T \cup J\ .$$

Mais detalhes podem ser vistos em Floater (1997).

3. Após conseguir todos os elementos da matriz λ, obter as coordenadas de todos os pontos no plano através da resolução dos sistemas $AX = B1$ e $AY = B2$, B1 e B2 são gerados a partir dos pontos exteriores. Seja E o conjunto de pontos exteriores.

$$B1i = \Sigma\, \lambda ir * Pyr,\ r \in E.$$

$$B2i = \Sigma\, \lambda ir * Pyr,\ r \in E.$$

Os elementos da matriz A, A_{ij}, são iguais a $-\lambda_{ij}$ ($A_{ij} = -\lambda_{ij}$) se $i \neq j$, e iguais a 1 se $i = j$. O número de elementos de uma linha ou coluna da matriz A é igual ao número de pontos interiores. O método para se resolver os sistemas foi o da triangularização, proposto por Gauss (Ruggiero & Vera, 2004).

Assim, com os pontos x_i e y_i ($i = 1,2,\ldots,n$) obtidos, tem-se o mapeamento dos pontos da superfície em 2D, conforme descrito em Floater (1997).

Este método, conforme mencionado anteriormente, só se aplica a superfícies que tenham somente uma abertura e sejam conexas e contínuas. Porém, existem muitos casos em que estas condições não são satisfeitas, como, por exemplo, uma superfície que tenha duas ou mais aberturas, ou ainda, não tenha abertura alguma. Para estes casos, o algoritmo foi adaptado para ser mais genérico. Foram adicionadas as seguintes considerações:

1. Caso a superfície seja totalmente fechada, basta escolher um ponto da mesma e retirá-lo da superfície. Assim, é feita uma abertura nesta superfície de forma a poder ser usado o algoritmo de Floater (1997).

2. Caso a superfície contenha mais de uma abertura, escolhe-se a abertura com maior número de pontos. Para as demais aberturas, insere-se um ponto central de tal forma que este se ligue com os demais pontos na da abertura. Este procedimento é feito para as demais aberturas que existirem. Assim ficará apenas uma abertura. As demais aberturas, na prática, serão planarizadas, porém sem o ponto fictício.

3. Caso a superfície total seja contida por duas ou mais superfícies ligadas por apenas um ponto, são escolhidos todos os pontos externos exceto os pontos de ligação entre as superfícies.

O algoritmo citado foi implementado em linguagem C devido à velocidade de processamento de que esta linguagem proporciona, e também por causa do sistema linear que pode conter muitas variáveis. O principal gargalo deste algoritmo é a resolução dos sistemas $AX = B1$ e $AY = B2$. A resolução dos dois sistemas é feita de forma simultânea, visto que a cada triangularização superior que é feita em A, tem-se o mesmo procedimento para B1 e B2 quanto a multiplicação por um escalar e subtração ou adição de valores. Ao se resolver os dois sistemas simultaneamente, $AX = B1$ e $AY = B2$, o tempo de execução do algoritmo reduziu em torno de 40% em média, conforme os testes feitos.

Assim, o algoritmo gera as coordenadas (x,y) em função das coordenadas tridimensionais (x1, y1, z1) da proteína. O software Sting lê estas coordenadas e mostra a figura 2D. Uma interface feita em Java produz as imagens de uma proteína em 3D e a respectiva imagem mapeada no plano 2D.

Em termos de complexidade, após a inserção da adaptação, tem-se que é necessário analisar se existem aberturas ou não. Esta é a dificuldade desta adaptação. A análise deve considerar todas as aberturas e deixar apenas a maior sem a inserção de ponto fictício. Por outro lado, caso não exista abertura, basta eleger o ponto com menos vizinhos (pontos que se ligam a ele) para ser retirado. Em princípio a idéia é simples, mas sem esta, o algoritmo de Floater só serviria para alguns casos.

Resultados

Conforme mencionado anteriormente, um dos recursos do Sting é a análise, de forma visual, de como algumas grandezas físico-químicas estão presentes na interface proteica. Usando o Java Protein Interface Viewer ou JPIV, um dos módulos do Sting (Yamagishi et al., 2006), pode-se visualizar a interface proteica tridimensionalmente, tal como mostrado na Fig. 2.

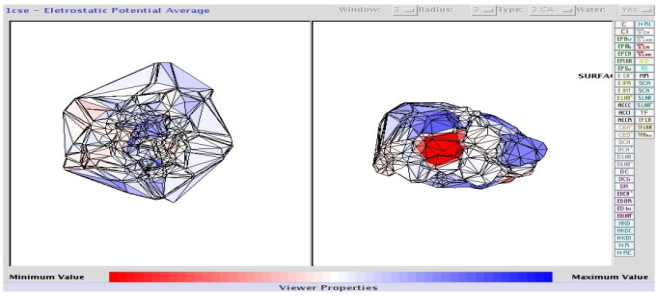


Fig. 2. Potencial eletrostático da proteína Serine Proteinase, cadeias E e I, respectivamente.

As imagens da Fig. 2 ao serem projetadas no plano 3D para 2D ficam da seguinte maneira (Fig. 3).

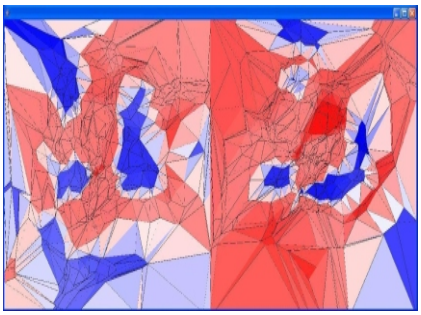


Fig. 3. Potencial eletrostático em 2D da proteína Serine Proteinase, cadeias E e I, respectivamente.

Na Fig. 3, são mostradas as cadeias E e I, com as cores representativas do potencial eletrostático (azul-positivo, vermelho-negativo), conforme visualizado no software Sting pelo JPIV.

Conforme visto nas figuras anteriores, fica mais fácil visualizar em que região o potencial eletrostático tem maior ou menor influência e a intensidade desta.

Para se obter dados sobre qualquer proteína, pode-se consultar o *site* do Protein Data Bank (Research Collaboratory for Structural Bioinformatics, 2006) que contém uma série de arquivos texto sobre diversas proteínas, que hoje está em torno de 40.000. Cada arquivo contém uma série de parâmetros sobre proteínas, incluindo os componentes, as cadeias, a localização de cada componente na cadeia, etc. Estes arquivos possuem as extensões .pdb e pdb, que significam protein data bank.

Existe um banco de dados com todos estes arquivos no site do Laboratório de Bioinformática (Embrapa Informática Agropecuária, 2006a), que são atualizados semanalmente. Assim, estes dados são usados para gerar as coordenadas das superfícies tridimensionais, após um pré-processamento, e estas coordenadas são a entrada para o algoritmo de Floater (1997) modificado, conforme descrito anteriormente, para então gerar as projeções no plano.

Alguns dos resultados podem ser mostrados a seguir, graficamente. Seja a proteína *serine proteinase*, que se encontra em um arquivo 1cho.pdb, descrita anteriormente. Dentre os vários parâmetros que podem ser visualizados são: potencial eletrostático (Fig. 3), hidrofobicidade (Fig. 4), conservação (Fig. 5), distribuição de energia (Fig. 6), etc. As figuras a seguir descrevem algumas destas características.

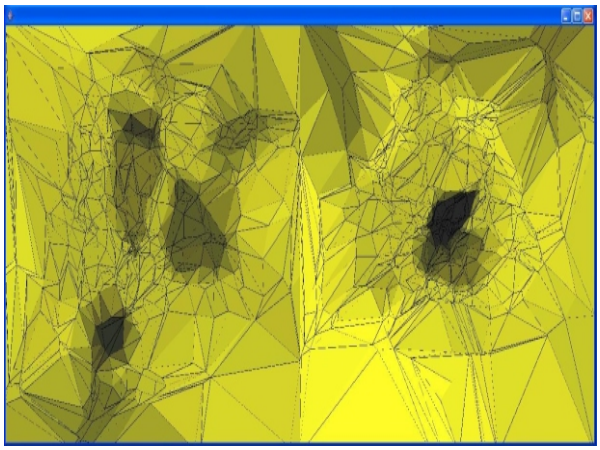


Fig. 4. Representação de características hidrofóbicas.

Na Fig. 4, a cor amarela significa "hidrófila" e escura significa "hidrofóbica" da proteína Serine Proteinase, cadeias E e I.

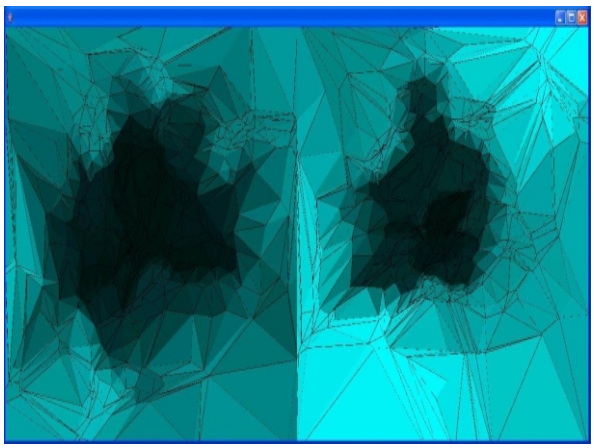


Fig. 5. Conservação (escuro = mais conservado) da proteína Serine Proteinase, cadeias E e I.

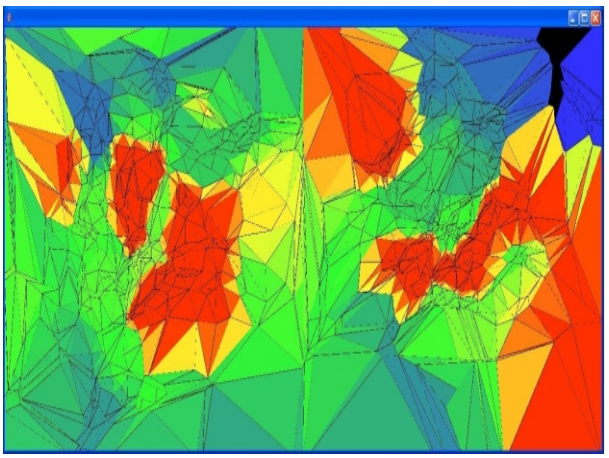


Fig. 6. Distribuição de energia (vermelha = menor energia) da proteína Serine Proteinase, cadeias E e I.

Na Fig. 5 são mostrados os resíduos mais conservados, os quais estão mais ao centro. Na Fig. 6, estão mostradas as concentrações de energia na superfície da proteína.

Para que o algoritmo executasse rapidamente, foi codificado em C, e foi necessária uma série de otimizações de cálculo. Não foi um algoritmo de fácil construção, pois exigiu muito conhecimento de geometria analítica, trigonometria, cálculo numérico, etc. para se determinar a matriz λ e conseqüentemente a matriz A, a qual é usada para o cálculo das coordenadas (x,y). O número de linhas de código chega a quase 4.000.

Devido ao grande número de arquivos a serem processados (em torno de 70.000 arquivos de entrada, com as coordenadas das superfícies), usou-se um cluster linux Mosix (Barak, 2006). Com este cluster, o qual contém 25 nós (processadores), foram divididas as tarefas de processamento em 10 nós. O cluster vai distribuindo o processamento nos 10 processadores livres que encontrar e monitora todo o processamento. Assim, não foi necessário executar este programa em outras máquinas da rede e o processamento foi muito tranqüilo. O tempo gasto para se fazer todo o processamento foi 1 dia.

Conclusões

O Algoritmo de Floater (1997), ao ser modificado para ser genérico, isto é aceitar figuras fechadas ou com mais de uma abertura, possibilitou a visualização das propriedades fisico-químicas de cada proteína, o que poucos softwares no mercado podem fazer. Assim, com a ajuda do Sting, pode-se ver qualquer proteína cadastrada no banco de dados Protein Data Bank (Research Collaboratory for Structural Bioinformatics, 2006) em 3D e 2D, juntamente com várias características físico-químicas, conforme a necessidade.

O Sting pode ser acessado em seu *site* (Embrapa Informática Agropecuária, 2006b), assim como uma série de funcionalidades para análise de proteínas, que são explicadas no próprio software.

Referências Bibliográficas

BARAK, A. *MOSIX*: grid and cluster management. Disponível em: <http://www.mosix.org>. Acesso em: 14 ago. 2006.

EMBRAPA INFORMÁTICA AGROPECUÁRIA. *Bioinformatics Laboratory*. Disponível em: <http://www.cbi.cnptia.embrapa.br/>. Acesso em: 14 ago. 2006a.

EMBRAPA INFORMÁTICA AGROPECUÁRIA. *Blue Star Sting - Structural Bioinformatics Group*. Disponível em: <http://www.cbi.cnptia.embrapa.br/SMS/>. Acesso em: 14 ago. 2006b.

FLOATER, M. S. Parametrization and smooth approximation of surface triangulations. *Computer Aided Geometry Design*, 14, n. 3, p. 231-250, Apr. 1997. Disponível em: <http://www.multires.caltech.edu/teaching/courses/cs101.3.spring02/cs101_files/resources/Parameterization/Floater.pdf>. Acesso em: 14 ago. 2006.

HAKER, S.; ANGENENT, S.; TANNENBAUM, A.; KIKINIS, R.; SAPIRO, G.; HALLE, M. Conformal surface parametrization for texture mapping. *IEEE Transactions on Visualization and Computer Graphics*, v. 6, n. 2, p. 1-9, Apr./June, 2000. Disponível em: <http://www.spl.harvard.edu:8000/pages/papers/tannenba/texture/texture.pdf>. Acesso em: 14 ago. 2006.

RESEARCH COLLABORATORY FOR STRUCTURAL BIOINFORMATICS. Protein Data Bank. *RCSB PDB - latest released structures*. Disponível em: <http://www.pdb.org/>. Acesso em: 14 ago. 2006.

RUGGIERO, M. A. G.; VERA, L. da R. *Cálculo numérico:* aspectos teóricos e computacionais. São Paulo :McGraw-Hill, 2004.

WIKIMEDIA FOUNDATION. Bioinformática. Disponível em: <http://pt.wikipedia.org/wiki/Bioinformática>. Acesso em: 14 ago. 2006.

YAMAGISHI, M. E. B.; FALCÃO, P. R. K.; BORRO, L. C.; OLIVEIRA, S. R. M.; SANTOS, E. H.; JARDINE, J. G.; VIEIRA, F. D.; MAZONI, I.; NARCISO, M. G.; NESHICH, G. Analysing protein-protein using JPIV. Pôster publicado no ISMB, 2006, Fortaleza. Disponível em: <http://ismb2006.cbi.cnptia.embrapa.br/poster_abstract.php?id=I-14>. Acesso em: 14 ago. 2006.

Comunicado Técnico, 78

Ministério da Agricultura, Pecuária e Abastecimento

Embrapa Informática Agropecuária
Área de Comunicação e Negócios (ACN)
Endereço: Caixa Postal 6041 - Barão Geraldo
13083-970 - Campinas, SP
Fone: (19) 3789-5743
Fax: (19) 3289-9594
e-mail: sac@cnptia.embrapa.com.br

1ª edição on-line - 2006