Ministério da Agricultura, Pecuária e Abastecimento

Comunicado Técnico 37

Dezembro, 2002
Campinas, SP

ISSN 1677-8464

Análise do Grau de Conservação de Resíduos em Proteínas com Estrutura 3D Resolvida Utilizando o SMS

Roberto Hiroshi Higa[1]
Christian Baudet[2]
Adauto Luiz Mancini[3]
Paula Kuser Falcão[4]
Goran Neshich[5]

As proteínas são macromoléculas presentes em todo organismo vivo, sendo de importância fundamental para a realização de praticamente todas as suas funções. Ela determina a forma e a estrutura da célula e serve como instrumento para o reconhecimento molecular e de catálise (Alberts et al., 1994).

No curso da evolução, quando uma proteína com uma característica de superfície útil surge, sua estrutura básica é incorporada em muitas outras proteínas. Por isso, muitas proteínas encontradas nos organismos atuais possuem funções distintas, mas relacionadas, apresentando uma série de aminoácidos similares ou conservados quando alinhadas (Alberts et al., 1994).

Dessa forma, quando uma nova proteína é seqüenciada, é prática comum, atualmente, procurar por seqüências similares utilizando-se ferramentas de busca como Blast (Altschul et al., 1990, 1997), cuja função é encontrar seqüências similares em bancos de dados de seqüências. O objetivo é encontrar outras proteínas, cuja função já tenha sido determinada, e que possam fornecer algum subsídio sobre a função da nova proteína. Outra abordagem para análise da função de uma proteína, quando ela possui estrutura tridimensional resolvida, é o estudo de sua estrutura, procurando as partes que possam fornecer pistas sobre como ela realiza a sua função. Por exemplo, aminoácidos conservados durante a evolução podem indicar que ali existe um sítio de ligação — *binding site* (o lugar na superfície da proteína onde outras moléculas — proteínas, DNAs ou outras moléculas - com as quais ela interage se ligam) ou um sítio ativo — *active site* (o lugar na superfície proteína onde acontece a função propriamente dita, por exemplo uma reação química em que a proteína atua como enzima).

O Sting Millennium Suite (SMS) (Structural Bioinformatics Group, 2002) – possui duas ferramentas que auxiliam a análise do grau de conservação de resíduos para proteínas com estrutura depositada no Protein Data Bank (PDB) (Berman et al., 2000), ou seja, que tiveram sua estrutura tridimensional resolvida. Uma delas é o ***ConS****ensus* ***Seq****uence* (ConSSeq), cuja função é, exatamente, apresentar dados para análise do grau de conservação de aminoácidos em uma proteína. A outra é o Protein Dossier, que apresenta

[1] *M.Sc. em Engenharia Elétrica, Pesquisador da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: roberto@cnptia.embrapa.br)*

[2] *Estudante de Engenharia da Computação, Estagiário da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: christian@cnptia.embrapa.br)*

[3] *Bacharel em Ciência da Computação, Pesquisador da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: adauto@cnptia.embrapa.br)*

[4] *Ph.D. em Física Aplicada, Cristalografia de Proteínas, Pesquisadora da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: paula@cnptia.embrapa.br)*

[5] *Ph.D. em Biofísica, Pesquisador da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: neshich@cnptia.embrapa.br)*

Embrapa

uma figura descrevendo um conjunto de parâmetros físico-químicos da proteína, incluindo uma medida de grau de conservação de aminoácidos como um de seus parâmetros.

Este trabalho apresenta como as ferramentas ConSSeq e Protein Dossier do SMS são utilizadas para análise do grau de conservação de aminoácidos em proteínas com estrutura tridimensional resolvida[6].

A seguir, apresenta-se o *Homology Derived Secondary Structure of Proteins* (HSSP), que é a base de dados utilizada pelo SMS para avaliação do grau de conservação de aminoácidos e o conceito de Entropia Relativa, que é a medida de conservação utilizada pelo HSSP. Na seqüência, ilustra-se como utilizar o ConSSeq e o Protein Dossier para análise de conservação, finalizando-se com uma discussão sobre as perspectivas de trabalhos futuros relacionados à análise de conservação de aminoácidos e as ferramentas do SMS.

HSSP e Entropia Relativa

O HSSP é um banco de dados, cujo objetivo é expandir o universo de proteínas com informações sobre sua estrutura, através da utilização do conceito de homologia entre seqüências (Sander & Schneider, 1991; Schneider et al., 1997).

Inicialmente, Sander & Schneider (1991) avaliaram um conjunto de proteínas com estrutura tridimensional resolvida e formularam uma relação entre o grau de similaridade seqüencial, o grau de similaridade estrutural e o comprimento de um alinhamento. Em seguida, eles estabeleceram que uma homologia estrutural seria válida se a similaridade seqüencial entre as duas seqüências fosse superior a um limiar que dependia do comprimento do alinhamento. Este limiar é representado pela seguinte fórmula (Sander & Schneider, 1991):

$$t(L) = 290.15L^{-0.562}$$

onde:

L indica o comprimento do alinhamento e

t(L) indica o grau de identidade[7] mínimo que implica em homologia estrutural.

Os dados do HSSP são criados a partir do estabelecimento da relação de homologia estrutural entre proteínas com estrutura tridimensional resolvida – banco de dados PDB (Berman et al., 2000) - e proteínas com seqüências conhecidas – banco de dados Swiss-Prot (Bairoch & Apweiler, 2000). A Fig. 1 ilustra esse processo.

Para cada seqüência no banco de dados PDB é realizada uma busca, utilizando-se Blast, por seqüências similares no banco de dados Swiss-Prot. As seqüências

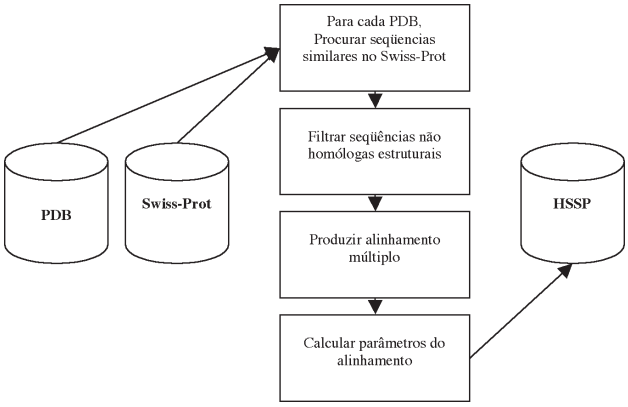

Fig. 1. Geração do banco de dados HSSP.

encontradas são filtradas, de acordo com o limiar definido para estabelecimento da homologia estrutural e, em seguida um alinhamento múltiplo é produzido de forma incremental, utilizando-se uma versão modificada do algoritmo de Smith-Waterman para alinhamento local entre duas seqüências (Setubal & Meidanis, 1997). Finalmente, parâmetros como entropia relativa e outros calculados pelo DSSP – Definition of Secondary Structure of Proteins (Kabsch & Sander, 1983) são acrescentados e armazenados no formato padrão do HSSP.

Em geral, existe um arquivo HSSP correspondente a cada arquivo PDB[8]. Cada arquivo HSSP contém (Schneider et al., 1997):

- a seqüência primária da proteína de estrutura conhecida com a indicação de estrutura secundária e área acessível por solvente calculados pelo DSSP (Kabsch & Sander, 1983);
- seqüências oriundas do Swiss-Prot consideradas homólogos estruturais da seqüência com estruturas conhecidas alinhadas;
- a variabilidade para cada posição do alinhamento segundo medidas diferentes, uma delas a entropia relativa; e
- o número de seqüências que geraram a posição ou ocupância.

A medida de entropia relativa, utilizada pelo SMS como medida de grau de conservação, é definida como:

[6] Proteínas com estrutura resolvida experimentalmente por difração de raios-X ou ressonância magnética nuclear, e com dados depositados no Protein Data Bank (PDB).

[7] Em um alinhamento entre duas seqüências, o grau de identidade significa o número de posições no alinhamento em que as duas seqüências possuem o mesmo tipo de aminoácido.

[8] Para arquivos PDB que contenham dados apenas sobre DNA, não existe o arquivo HSSP correspondente.

$$relent(i) = S(i) / \ln 20$$

onde:

$$S(i) = \sum_{R}^{20} f_{Ri} \ln f_{Ri}$$

f_{Ri} é a freqüência de ocorrência do tipo de resíduo R[9] na posição i do alinhamento.

Módulos do SMS para Análise de Conservação

O Protein Dossier é um módulo do SMS, cuja funcionalidade é produzir e apresentar uma figura, onde uma série de parâmetros físico-químicos utilizados para análise da estrutura tridimensional da proteína e sua relação com a função são sumarizadas (Fig. 2).

Um dos parâmetros apresentados é a entropia relativa —*Relative Entropy*, apresentada em tons de vermelho. Tons de vermelho mais forte indicam posições onde diversas mutações ocorreram durante a evolução enquanto tons de vermelho mais fraco indicam variação pequena. No caso extremo, preto significa que a posição é completamente conservada.

Para analisar com mais detalhes o grau de conservação dos aminoácidos da proteína em estudo, utilizando o SMS, a ferramenta indicada é o ConSSeq.

O ConSSeq apresenta ao mesmo tempo, além da entropia relativa, a seqüência de consenso[10] e um *Logo* que representa a proporção com que cada tipo de aminoácido ocorre em uma faixa de 10 posições. A ferramenta permite que um *Logo* seja elaborado para cada posição da seqüência bastando, para isto, utilizar a barra de rolagem para posicionar o mouse na posição desejada da seqüência. Apertando o botão esquerdo do mouse, o logo é elaborado para um intervalo de dez resíduos com o resíduo selecionado na posição central da figura.

A Fig. 3 ilustra a apresentação visual do ConSSeq para a cadeia "E" do arquivo com PDB ID igual a "1cho", uma proteína da família chymotrypsin retirada de tecido de pâncreas bovino. Na parte superior é apresentado, para cada posição da seqüência da proteína em estudo, um gráfico de barras que indica o valor da entropia relativa, a seqüência propriamente dita, e a seqüência de consenso. A cor vermelha indica alta entropia relativa enquanto a azul indica entropia relativa baixa.

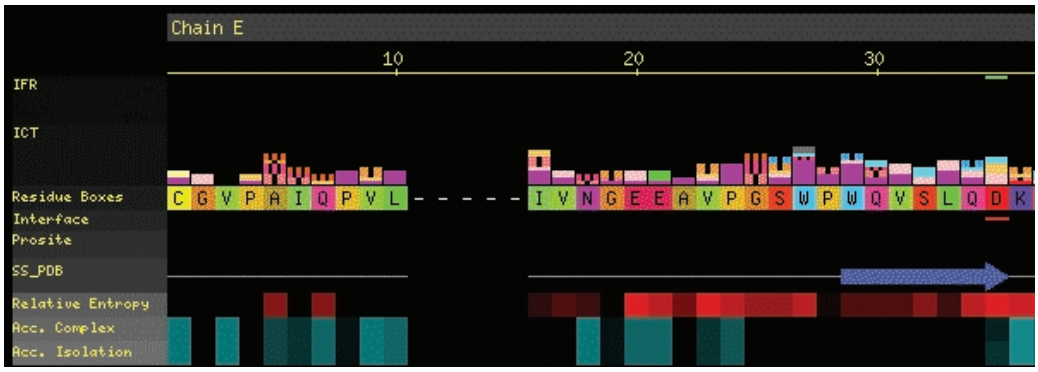

Fig. 2. Apresentação da entropia relativa no Protein Dossier.

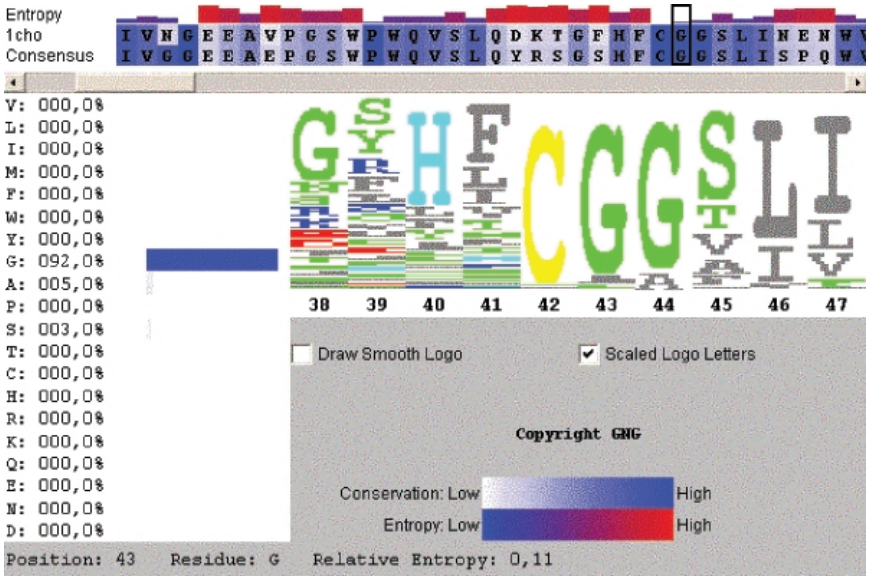

Fig. 3. Apresentação de informações sobre conservação pelo ConSSeq.

[9] Cada um dos 20 tipos de aminoácios podem ser associados aos números inteiros 1 a 20. Assim, pode-se utilizar R=1 para Alanina, R=2 para Valina, etc.

[10] Considerando um conjunto de seqüências alinhadas, a seqüência de consenso é formada atribuindo-se a cada posição o resíduo mais freqüente no alinhamento.

Na parte inferior, é possível verificar, para uma posição específica e na sua vizinhança, qual a freqüência de ocorrência de cada tipo de aminoácido, ou seja, de que forma a natureza está trocando o tipo de aminoácido naquela posição. À esquerda, é apresentado um histograma que indica a freqüência de ocorrência de cada tipo de aminoácido na posição especificada e à direita, considerando um intervalo de 10 posições, centrado na posição especificada, é apresentado um *Logo*[11] que indica a freqüência de ocorrência de cada tipo de aminoácido em cada uma das posições.

Discussão e Trabalhos Futuros

Os aminoácidos de uma proteína, conservados durante a evolução, podem revelar informações importantes sobre a sua função como a localização de sítios de ligação e sítios ativos. Por isso, a importância da análise destes parâmetros para compreensão da função da proteína e de sua relação com a estrutura.

O SMS permite que o grau de conservação de aminoácidos seja analisado através de duas ferramentas: o Protein Dossier que inclui a entropia relativa entre os parâmetros apresentados em formato de figura e o ConSSeq que além da entropia relativa, também apresenta a seqüência de consenso para comparação com a seqüência original e como cada posição da seqüência original está variando.

Entretanto, tanto o Protein Dossier quanto o ConSSeq baseiam-se na utilização de dados do HSSP para apresentação do grau de conservação de aminoácidos e o HSSP, por sua vez, apresenta as seguintes limitações:

- sua freqüência de atualização é menor que a atualização do PDB, que é semanal, tal que, para as estruturas mais novas depositadas no PDB, as entradas no HSSP são criadas e disponibilizadas com algum atraso;
- não é possível obter os dados do HSSP para estruturas não depositadas no PDB, o que inviabiliza a análise de conservação de aminoácidos para estruturas modeladas; e
- não há como adaptar o procedimento do HSSP para utilizar um alinhamento diferente (ex.: uma família protéica), que muitas vezes é mais interessante que o alinhamento fornecido pelo HSSP.

Para superar as limitações apresentadas acima, pretende-se construir um procedimento similar ao utilizado pelo HSSP, mas capaz de:

- produzir um banco de dados com as mesmas características do HSSP; e
- utilizar um alinhamento diferente para medir o grau de conservação, viabilizando a verificação do grau de conservação para estruturas não depositadas no PDB.

Assim, espera-se que o SMS possa oferecer ao usuário um conjunto de ferramentas mais flexível e interessante no que tange à análise de conservação de aminoácidos para proteínas.

[11] O logo é composto pelo código de uma letra para aminoácidos e deve ser interpretado da seguinte forma: quanto mais freqüente o aminoácido na posição determinada, maior o tamanho da letra correspondente ao seu código no logo.

Referências Bibliográficas

ALBERTS, B.; JOHNSON, A.; LEWIS, J.; RAFF, M.; ROBERTS, K.; WALTER, P. **Molecular biology of the cell**. 3rd ed. New York: Garland Pub., 1994. 1294 p.

ALTSCHUL, S. F.; GISH, W.; MILLER, W.; MYERS, E. W.; LIPMAN, D. J. Basic local alignment search tool. **J. Mol. Biol.**, v. 215, p. 403-410, 1990.

ALTSCHUL, S. F.; MADDEN, T. L.; SCHAFFER, A. A.; ZHANG, J.; ZHANG, Z.; MILLER, W.; LIPMAN, D. L. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acid Research**, v. 25, n. 17, p. 3389-3402, 1997.

BAIROCH, A.; APWEILER, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. **Nucleic Acids Research**, v. 28, n.1, p. 45-48, 2000.

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, n. 1, p. 235-242, 2000.

KABSCH, W.; SANDER, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. **Biopolymers**, v. 22, n. 12, p. 2577-2637, 1983.

SANDER, C.; SCHNEIDER, R. Database of homolgy-derived protein structures and the structural meaning of sequence alignment. **Proteins**: Struc. Function and Genetics, v. 9, p. 56-68, 1991.

SCHNEIDER, R.; DE DARUVAR, A.; SANDER, C. The HSSP database of protein structure sequence alignments. **Nucleic Acids Research**, v. 25, n. 1, p. 226-230, 1997.

SETUBAL, J.C.; MEIDANIS, J. **Introduction to computational molecular biology**. Boston: PWS Publ., 1997. 296 p.

STRUCTURAL BIOINFORMATICS GROUP. **Sting Millennium Suite**. Disponível em: <http://www.nbi.cnptia.embrapa.br/>. Acesso em: 21 dez. 2002.

Comunicado Técnico, 37

Embrapa Informática Agropecuária
Área de Comunicação e Negócios (ACN)
Av. André Tosello, 209
Cidade Universitária - "Zeferino Vaz"
Barão Geraldo - Caixa Postal 6041
13083-970 - Campinas, SP
Telefone (19) 3789-5743 - Fax (19) 3289-9594
e-mail: sac@cnptia.embrapa.br

1ª edição
2002 - on-line

Expediente

Supervisor editorial: *Ivanilde Dispato*
Normalização bibliográfica: *Marcia Izabel* Fugisawa Souza
Capa: *Intermídia Publicações Científicas*
Editoração Eletrônica: *Intermídia Publicações Científicas*