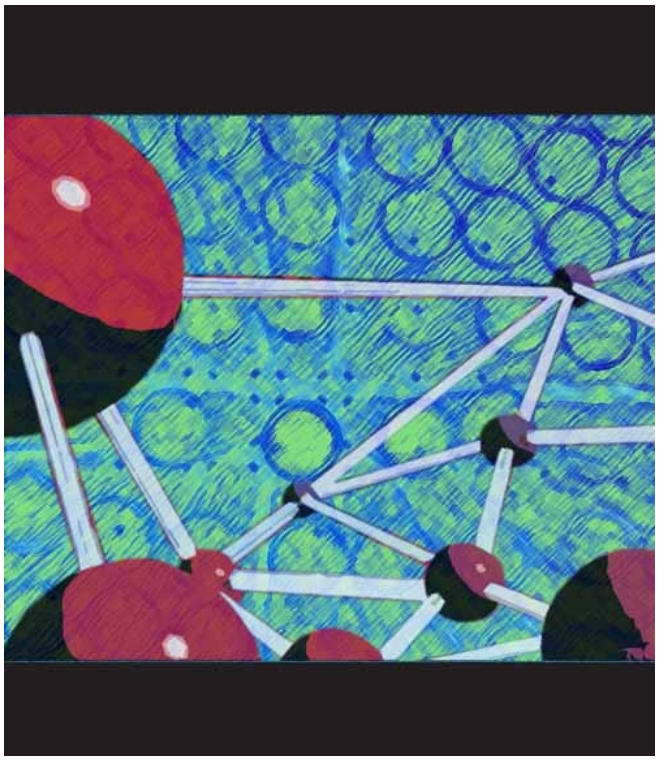
Comunicado Técnico 61

Outubro, 2004
Campinas, SP



ISSN 1677-8464

Grau de Conservação de Aminoácidos de Proteínas: Comparação entre Entropia Relativa e Pressão Evolucionária

Roberto Hiroshi Higa
Paula Regina Kuser Falcão
Michel Eduardo Beleza Yamagishi
Maria Fernanda Moura
Renato Fileto
Adauto Luiz Mancini
Goran Neshich

O Java Protein Dossier - JPD (Neshich et al., 2004) é um programa integrado à ferramenta STING (Neshich et al., 2003) desenvolvida pelo Núcleo de Bioinformática Estrutural, da Embrapa Informática Agropecuária, para apoiar a análise de proteínas e sua complexa relação estrutura/função. Ele realiza esta função apresentando ao biólogo molecular (usuário) um grande conjunto de parâmetros tanto biológicos quanto físico-químicos, organizados por aminoácido. Dessa forma, o usuário pode realizar a análise da proteína em estudo utilizando diversas perspectivas, por exemplo analisando resíduos conservados ao longo da evolução, se existem padrões na seqüência de resíduos associados a alguma função biológica, de acordo com o Prosite, analisando o potencial eletrostático na superfície da proteína (um potencial bastante negativo pode indicar um local preferencial para ligação com partes positivas de outras moléculas), etc.

Um dos parâmetros apresentados pelo JPD é o grau de conservação de aminoácidos. Este parâmetro é importante na análise de proteínas, pois permite identificar aminoácidos preservados pela natureza ao longo da evolução, seja pela importância funcional, seja pela importância na estabilidade da estrutura tridimensional da proteína. O JPD apresenta este parâmetro baseado em dois alinhamentos distintos, um extraído da base de dados HSSP - Homology-derived Structures of Proteins (Sander & Schneider, 1991) e o outro extraído de uma base de dados construída a partir de um processo alternativo desenvolvido no próprio Núcleo de Bioinformática Estrutural, denominado SH2Q[S] - Sequences Homologue to the Query *[Structure-having] Sequence* (Higa et al., 2003). Para cada um desses alinhamentos, dois métodos de aferição do grau de conservação são utilizados: entropia relativa e pressão evolucionária.

[1] *M.Sc. em Engenharia Elétrica, Pesquisador da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: roberto@cbi.cnptia.embrapa.br)*
[2] *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@cbi.cnptia.embrapa.br)*
[3] *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)*
[4] *M.Sc. em Engenharia Elétrica, Pesquisadora da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: fernanda@cbi.cnptia.embrapa.br)*
[5] *Doutor em Ciência da Computação, Pesquisador da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: fileto@cbi.cnptia.embrapa.br)*
[6] *Bacharel em Ciência da Computação, Pesquisador da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: adauto@cbi.cnptia.embrapa.br)*
[7] *Ph.D. em Biofísica, Pesquisador da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: neshich@cbi.cnptia.embrapa.br)*

Embrapa

A entropia relativa é calculada para cada posição do alinhamento de forma independente. Para cada posição do alinhamento é calculada a freqüência relativa para cada um dos possíveis 20 aminoácidos. A entropia para a posição *i* de um alinhamento é dada por:

$$S_i \qquad {}_x p_x \log(p_x) \qquad (1)$$

onde p_x é a freqüência relativa do aminoácido x na posição *i* do alinhamento.

A entropia relativa (*relent*) para a posição i de um alinhamento é calculada com relação à distribuição uniforme da seguinte forma:

$$relent(i) \quad \frac{S_i}{S_{unif}} \qquad (2)$$

onde S_i é a entropia na posição *i* do alinhamento e S_{unif} é entropia para a distribuição uniforme.

A pressão evolucionária, por sua vez, baseia-se em um método estatístico que utiliza como base a relação filogenética entre as seqüências alinhadas e o processo estocástico subjacente ao processo evolutivo (modelo de substituição de aminoácidos). Do ponto de vista computacional, este método é muito mais custoso do que o método da entropia relativa. Para calcular a pressão evolucionária, o JPD utiliza o software Rate4Site (Pupko et al., 2002).

Este trabalho apresenta uma comparação entre essas duas medidas utilizadas para medir o grau de conservação de aminoácidos, entropia relativa e pressão evolucionária. A motivação para a sua realização é apresentar um indicativo para os usuários do JPD sobre como interpretar os dados relativos a conservação de aminoácidos, uma vez que nem sempre as medidas apresentadas estão em acordo. Esta informação é especialmente importante na utilização da função "Select" do JPD, na qual o usuário fornece uma faixa de valores para cada um dos diversos parâmetros para seleção de aminoácido, incluindo entropia relativa e pressão evolucionária.

Material e Métodos

Conjunto de dados

Os dados utilizados foram extraídos de um subconjunto dos alinhamentos contidos na base de dados HSSP (*release* maio/2004). 200 cadeias protéicas foram selecionadas aleatoriamente, sendo considerados para comparação os valores de entropia relativa e pressão evolucionária para cada um dos aminoácidos de cada uma das cadeias amostradas.

Método de comparação

Sabe-se que a distribuição dos valores de pressão evolucionária possui distribuição normal com média igual a 0 e desvio padrão igual 1 (Pupko et al., 2002). Já para os valores de entropia relativa, não há referência a essa Informação. Por isso, foram realizadas cinco amostragens aleatórias, conforme descrito, e calculadas as estatísticas descritivas básicas (máximo, mínimo, moda, média, mediana, desvio padrão, curtose e assimetria) e traçado o histograma das distribuições. Rapidamente, verificou-se que os valores de entropia relativa não possuem distribuição normal.

Dessa forma, procedeu-se a comparação da seguinte maneira:

1. para verificar a equivalência entre as duas medidas, foram calculados os percentis (Martins & Donaire, 1979) 10, 20, 30, 40, 50, 60, 70, 80, 90 e 100 para as duas medidas, entropia relativa e pressão evolucionária. Em seguida contabilizou-se, para cada percentil, o número de resíduos selecionados por ambos os métodos, e por cada método individualmente; e
2. para comparar a sensibilidade em detectar resíduos conservados, calcularam-se os valores de entropia relativa e pressão evolucionária para que um conjunto de aminoácidos previamente conhecidos sejam evolucionariamente conservados - os aminoácidos envolvidos na reação catalítica do DNA Polimerase I e os aminoácidos que formam o núcleo catalítico da proteína quinase cAMP-dependente. Foram considerados conservados os aminoácidos que apresentaram valores de entropia relativa ou pressão evolucionária inferiores a um valor de limiar estabelecido. Os valores de limiar utilizados foram aqueles correspondentes aos percentis 10, 20, 30, 40 e 50.

Resultados e Discussão

As 200 cadeias protéicas selecionadas resultaram em um conjunto de dados contendo 49.469 aminoácidos. A Tabela 1 apresenta os percentis para a entropia relativa, calculada sobre esse conjunto de dados.

Tabela 1. Percentis para entropia relativa e pressão evolucionária.

Percentil	*Entropia Relativa*	*Pressão Evolucionária*
10	4	-0,885
20	14	-0,686
30	24	-0,544
40	32	-0,424
50	40	-0,316
60	47	-0,104
70	55	0,190
80	63	0,594
90	72	1,246
100	95	12,264

A Fig. 1 apresenta as proporções de aminoácidos em cada percentil (10, 20, 30, 40, 50, 60, 70, 80, 90 e 100) relatados exclusivamente por cada método, bem como a proporção comum a ambos os métodos. O baixo percentual de aminoácidos classificados no mesmo percentil por ambos os métodos é um indicativo de que não existe equivalência

entre eles. As discordâncias menores nos extremos indicam uma ligeira convergência dos métodos sobre quais são os resíduos extremamente conservados e extremamente não conservados.

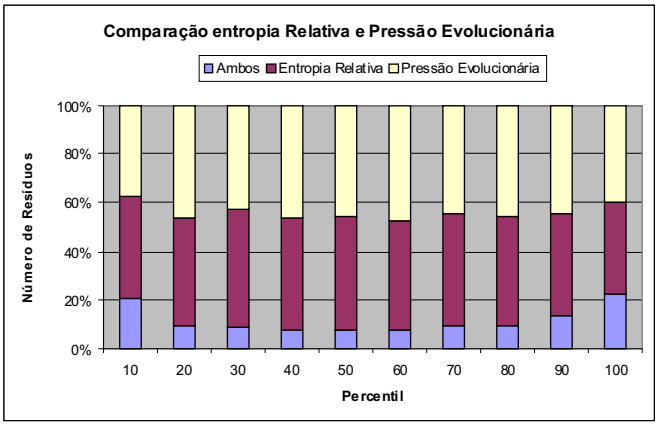


Fig. 1. Relação entre percentagem de aminoácidos pertencentes a cada percentil, segundo cada método, exclusivamente e por ambos os métodos.

Para comparar a sensibilidade dos métodos em detectar aminoácidos conservados, foram utilizados dois estudos de caso: proteína quinase cAMP-dependente e DNA polimerase I.

Proteína quinase cAMP-dependente

A proteína quinase cAMP-dependente é uma enzima com um papel fundamental no processo de transdução de sinais em vertebrados. Ela participa dos processos de regulação de síntese e quebra de glicogênio, síntese de ácidos gordurosos, oxidação de piruvato para acetil-CoA, movimentação de triacilglicerol e a regulação recíproca de glicose e gluconeogênesis. Como seu nome indica, ela é uma enzima dependente da presença de cAMP.

Esta enzima ocorre naturalmente como uma estrutura quaternária com quatro subunidades - duas regulatórias (R) e duas catalíticas (C). Na forma de holoenzima, as subunidades reguladoras estão ligadas ao sítio ativo das subunidades catalíticas, inativando-as. Quando uma molécula de cAMP está presente, ela se liga à subunidade reguladora, causando uma mudança conformacional que libera e ativa as duas subunidades C. Uma vez separada da subunidade R, a subunidade C pode realizar a sua função que é a adição de fosfato a uma Serina ou Treonina do substrato (Madhusudan et al., 1994).

Tabela 2. Número de resíduos identificados como conservados por cada método para os resíduos funcionais conservados da proteína quinase cAMP-dependente (pdb ID: 2cpk:E). Os limiares utilizados correspondem aos percentis 10, 20, 30, 40 e 50.

	10%	*20%*	*30%*	*40%*	*50%*
Entropia Relativa	14	18	21	28	30
Pressão Evolucionária	30	33	37	38	38

De 43 aminoácidos envolvidos na reação catalítica da enzima proteína quinase cAMP-dependente, através da pressão evolucionária foram detectados 38 aminoácidos conservados enquanto através da entropia relativa foram detectados 30 para um limiar correspondente ao percentil 50. Além disso, todos os 30 aminoácidos detectados pelo método de entropia relativa também foram detectados pelo método de pressão evolucionária. Utilizando-se um limiar correspondente aos percentis 10 e 20 e utilizando-se a entropia relativa, foram detectados 14 e 18 aminoácidos conservados, respectivamente, enquanto que pela pressão evolucionária foram detectados 30 e 33 aminoácidos (Tabela 2). Em todos os limiares testados, o método da pressão evolucionária detectou um número maior de aminoácidos relatados como conservados e funcionalmente importantes quando comparado ao método da entropia relativa.

DNA polimerase I

A família de proteínas DNA polimerases contém sítios que podem ser sobrepostos estruturalmente e uma seqüência altamente conservada. Ela tem um papel central na transmissão da informação genética de geração para geração. Seu padrão de dobramento global se assemelha à mão direita humana e contém três domínios distintos - a palma, os dedos e o dedão. Eles compartilham duas regiões conservadas: motivo A e motivo C, ambos localizados no interior do domínio palma. A seqüência primária de vários sítios ativos de DNA polimerases são excepcionalmente conservados, o que sugere que o motivo A evoluiu lentamente. Em particular, a seqüência DYSQIELR (aminoácidos 605 a 617) no motivo A é conservada em polimerases de organismos separados por milhões de anos de evolução. Recentemente, Patel & Loeb (2000) conduziram um experimento de mutagênese dirigida que indicou todos os aminoácidos do motivo A, a menos do aminoácido Asp:610, como mutáveis enquanto a atividade da enzima nativa era preservada.

Tabela 3. Número de resíduos identificados como conservados por cada método para os resíduos catalíticos do DNA polimerase I (pdb ID: 1ktq:A). Os limiares utilizados correspondem aos percentis 10, 20, 30, 40 e 50.

	10%	*20%*	*30%*	*40%*	*50%*
Entropia Relativa	5	7	8	8	10
Pressão Evolucionária	8	10	10	11	11

De 13 resíduos envolvidos na reação catalítica da enzima DNA polimerase I, o número máximo de resíduos detectados através da pressão evolucionária foi 11, enquanto através da entropia relativa foram detectados 10, quando um limiar correspondia ao percentil 50. Além disso, todos os 10 aminoácidos detectados pelo método de entropia relativa também foram detectados pelo método de pressão evolucionária. Utilizando-se um limiar correspondente aos percentis 10 e 20, através da entropia relativa, foram detectados 5 e 7 aminoácidos conservados, respectivamente, enquanto que pela pressão evolucionária

foram detectados 8 e 10 aminoácidos, respectivamente (Tabela 3). Em todos os limiares testados, o método da pressão evolucionária detectou um número maior de aminoácidos relatados como conservados e funcionalmente importantes quando comparado ao método da entropia relativa.

Considerações Finais

A avaliação dos métodos de entropia relativa e pressão evolucionária para medida do grau de conservação de aminoácidos mostrou que nenhum dos métodos é 100% eficaz. Em ambos os casos de teste, nenhum dos métodos conseguiu detectar todos os aminoácidos conservados, o que justifica a presença de ambos no JPD. Comparativamente, o método da pressão evolucionária mostrou-se superior ao método da entropia relativa em todas as situações testadas, principalmente quando limiares mais restritivos foram utilizados. Para limiares de conservação moderados, os resultados tenderam a se aproximar, ainda que os resultados obtidos pelo método da pressão evolucionária continuassem prevalentes.

Finalmente, é importante ressaltar a importância desta informação para os usuários do JPD. Além de auxiliar na interpretação de valores discrepantes entre os dois métodos, ela é de fundamental importância na utilização da função "Select" em que o usuário pode fornecer uma faixa de valores para os diversos parâmetros, inclusive para o grau de conservação (entropia relativa e pressão evolucionária).

Referências Bibliográficas

HIGA, R. H.; KUSER, P. R.; YAMAGISHI, M. E. B.; MANCINI, A. L.; NESHICH, G. **Análise preliminar de um processo para identificação e alinhamento de seqüências homólogas para proteínas com estrutura resolvida**. Campinas: Embrapa Informática Agropecuária, 2003. 7 p. (Embrapa Informática Agropecuária. Comunicado Técnico, 48). Disponível em: <http://www.cnptia.embrapa.br/modules/tinycontent3/content/2003/comtec48.pdf>. Acesso em: 20 dez. 2004.

MADHUSUDAN; TRAFNY, E. A.; XUONG, N. H.; ADAMS, J. A.; EYCK, L.; TAYLOR, S. S.; SOWADSKI, J. M. cAMP-dependent protein kinase: crystallographic insights into substrate recognition and phosphotransfer. **Protein Science**, v. 3, p. 176-87, 1994.

NESHICH, G.; ROCCHIA, W.; MANCINI, A. L.; YAMAGISHI, M. E. B.; KUSER, P. R.; FILETO, R.; BAUDET, C.; PINTO, I. P.; MONTAGNER, A. J.; PALANDRANI, J. F.; KRAUCHENCO, J. N.; TORRES, R. C.; SOUZA, S.; TOGAWA, R. C.; HIGA, R. H. [Java]Protein Dossier: a novel web-based data visualization tool for comprehensive analysis of protein structure. **Nucleic Acids Research**, v. 32 (Web Server issue), W595-W601, 2004. (DOI: 10.1093/nar/gkh480). Disponível em: <http://nar.oupjournals.org/cgi/content/full/32/suppl_2/W595>. Acesso em: 20 dez. 2004.

NESHICH, G.; TOGAWA, R. C.; MANCINI, A. L.; KUSER, P. R.; YAMAGISHI, M. E. B.; PAPPAS JUNIOR, G.; TORRES, W. V.; CAMPOS, T. F. e; FERREIRA, L. L.; LUNA, F. M.; OLIVEIRA, A. G.; MIURA, R. T.; INOUE, M. K.; HORITA, L. G.; SOUZA, D. F. de; DOMINIQUINI, F.; ÁLVARO, A.; LIMA, C. S.; OGAWA, F. O.; GOMES, G. B.; PALANDRANI, J. F.; SANTOS, G. F. dos; FREITAS, E. M. de; MATTIUZ, A. R.; COSTA, I. C.; ALMEIDA, C. L. de; SOUZA, S.; BAUDET, C.; HIGA, R. H. STING Millennium: a Web based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence. **Nucleic Acids Research**, v. 31, n. 13, p. 3386-3392, 2003.

MARTINS, G. de A.; DONAIRE, D. **Princípios de estatística**. São Paulo: Atlas, 1979. 200 p.

PATEL, P. H.; LOEB, L. DNA polymerase active site is highly mutable: evolutionary consequences. **Proceedings of the National Academy of Science of the United States of America**, v. 97, n. 10, p. 5095-5100, May, 2000.

PUPKO, T.; BELL, R. E.; MAYROSE, I.; GLASER, F.; BENTAL, N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. **Bioinformatics**, v. 18, p. 71-77, 2002.

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

Comunicado Técnico, 61

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 - 2004
Todos os direitos reservados.

Comitê de Publicações

Presidente: *Marcos Lordello Chaim (presidente)*
Membros Efetivos: *Carla Geovana Macário, Ivanilde Dispato, José Ruy porto de Carvalho, Luciana Alvim Santos Romani, Marcia Isabel Fugisawa Souza, Suzilei Almeida Carneiro (secretária)*
Suplentes: *Carlos Alberto Alves Meira, Eduardo Delgado Assad, Maria Angelica Andrade Leite, Maria Fernanda Moura, Maria Goretti Gurgel Praxedis*

Expediente

Supervisor editorial: *Ivanilde Dispato*
Normalização bibliográfica: *Marcia Izabel Fugisawa Souza*
Editoração eletrônica: *Área de Comunicação e Negócios*