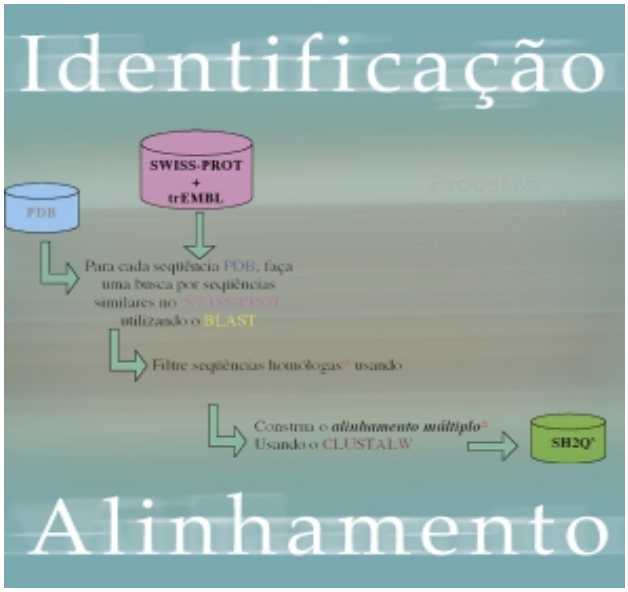
Comunicado Técnico 48
Setembro, 2003
Campinas, SP



ISSN 1677-8464

Análise Preliminar de um Processo para Identificação e Alinhamento de Seqüências Homólogas para Proteínas com Estrutura Resolvida

Roberto Hiroshi Higa[1]
Paula Regina Kuser[2]
Michel Eduardo Beleza Yamagishi[3]
Adauto Luiz Mancini[4]
Goran Neshich[5]

O Homology-derived Structures of Proteins (HSSP) é uma base de dados que agrega informações de proteínas com estrutura tridimensional resolvida a seqüências de proteínas sem estrutura (Sander & Schneider, 1991; Dodge et al., 1998). Para cada arquivo do Protein Data Bank - PDB (Berman et al., 2000), o HSSP possui um arquivo que relata um alinhamento formado por seqüências homólogas obtidas da base de dados de seqüências de proteínas SWISS-PROT (Bairoch & Apweiler, 2000). Este alinhamento é construído por um método iterativo baseado em programação dinâmica, que perfaz um processo de alinhamento de perfis (*profile*) de seqüências. Uma função de *threshold* para homologia estrutural (Sander & Schneider, 1991) é utilizada para decidir se uma seqüência é considerada homóloga ou não. O HSSP pode ser utilizado para definição de padrões de seqüência com significado estrutural e para estudo de evolução e dobramento de proteínas, entre outras aplicações.

No *Sting Millennium Suite* (SMS) (Neshich et al., 2003), ferramenta desenvolvida pelo Núcleo de Bioinformática Estrutural (NBI) da Embrapa Informática Agropecuária, disponível no site http://sms.cbi.cnptia.embrapa.br, o HSSP é utilizado para análise do grau de conservação de resíduos no contexto estrutural, através do valor de Entropia Relativa, relatado para cada posição de cada alinhamento. Entretanto, o fato da Entropia Relativa ser extraída do HSSP restringe a disponibilidade deste parâmetro àqueles arquivos PDBs considerados na atualização mais recente do HSSP. Assim, ele pode não estar disponível para arquivos PDBs recentes e ainda não processados pelo HSSP. Além disso, como os dados do HSSP são derivados dos dados do PDB, a Entropia Relativa nunca está disponível para arquivos locais correspondentes a estruturas não depositados no PDB (em processo de resolução ou modelados por homologia). Desta forma, para tornar o valor do parâmetro Entropia Relativa disponível, em todas as

[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. (email: 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] *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)*

[5] *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

situações de uso do SMS (arquivo PDB, ou arquivo local), pretende-se incorporar o cálculo deste parâmetro ao próprio SMS.

Os alinhamentos múltiplos relatados pelo HSSP são construídos utilizando-se um programa proprietário (não disponível publicamente) denominado MaxHom (Sander & Schneider, 1991). Assim, para que o parâmetro Entropia Relativa possa estar disponível em todas as situações de uso do SMS, torna-se necessário a elaboração de um processo alternativo para construção desses alinhamentos.

O objetivo deste trabalho é apresentar e fazer uma avaliação preliminar de um processo alternativo, denominado *Sequences Homologue to the Query [Structure-having] Sequence - SH2Q^s*, para elaboração de alinhamentos múltiplos semelhantes àqueles relatados no HSSP. O processo aqui apresentado baseia-se em programas de domínio público para busca em bases de dados de seqüências – Blast (Altschul et al., 1990, 1997) e para alinhamento múltiplo de seqüências – ClustalW (Thompson et al., 1994.). O critério para avaliação do mesmo é o grau de similaridade entre as medidas de Entropia Relativa, quando comparadas com os mesmos valores relatados pelo HSSP.

Material e Métodos

- Processo para construção de alinhamento múltiplo

A Fig. 1 ilustra o método utilizado para identificação das seqüências homólogas da seqüência *query*, construção do alinhamento múltiplo e cálculo da Entropia Relativa para cada posição do alinhamento.

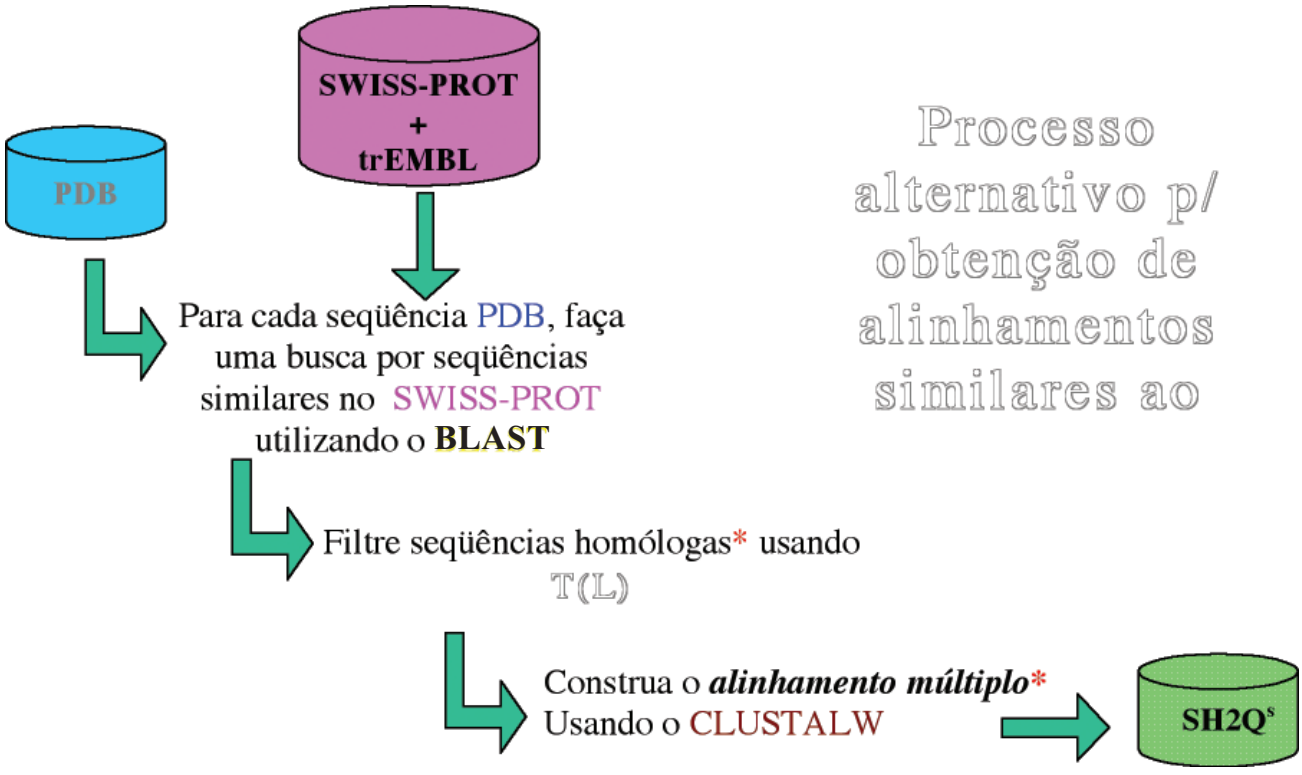


Fig.1. Processo de determinação do SH2Q^s.

Para cada seqüência de proteína contida em cada arquivo PDB é realizada uma busca Blast sobre o banco de dados de seqüências de proteínas SWISS-PROT e trEMBL. Para a busca Blast, os parâmetros utilizados foram utilizados a matriz de similaridade BLOSUM 60 para comparação de aminoácidos e o valor de limiar de significância estatística (E-value[6]) definido pelo modelo estocástico de Karlin & Altschul (1990), igual a 0.01. O resultado da busca é um conjunto de seqüências localmente alinhadas à seqüência *query*, segundo a heurística utilizada por Blast. Cada seqüência alinhada com a seqüência *query* é avaliada segundo a função de *threshold* T(L), proposta por Sander & Schneider (1991).

$$T(L) = 290.15L^{-0.562} \quad (1)$$

onde: *L* é o comprimento do alinhamento considerado.

O conjunto de seqüências resultante desse processo de filtragem é alinhado utilizando-se o programa de alinhamento múltiplo ClustalW. O algoritmo implementado por este programa utiliza uma série de heurísticas e uma estratégia de alinhamento progressivo para construção de um alinhamento múltiplo ótimo. Para comparação de aminoácidos, utilizou-se a série de matrizes de similaridade BLOSUM, o valor de penalidade para inserção de um gap no alinhamento múltiplo (*gap open penality*) igual a 3.0 e o valor para alongamento de um gap (*elongation penality)* igual a 0.1. Os demais parâmetros disponíveis para customização do alinhamento como definição de resíduos hidrofílicos e atribuição de valores de penalidades específicos para cada aminoácido, separação de gaps e adiamento da incorporação de seqüências muito divergentes ao alinhamento múltiplo foram todos desabilitados, ou seja, as heurísticas controladas por estes parâmetros não são utilizadas. Esta forma de utilização de parâmetros do ClustalW foi definida para tornar o processo de alinhamento múltiplo similar ao realizado pelo programa MaxHom, que utiliza apenas as penalidades para abertura e alongamento de gap como parâmetros. Os alinhamentos múltiplos foram construídos segundo duas estratégias diferentes:

- *HSSP-Like*: o alinhamento múltiplo é construído adicionando-se as seqüências em ordem de maior similaridade com a seqüência *query*;
- *Clustal-Like*: a ordem em que as seqüências são adicionadas ao alinhamento múltiplo é decidido segundo a estratégia padrão do ClustalW, ou seja, todos os pares de seqüências são comparados entre si e uma árvore guia, agrupando as seqüências mais similares, é construída. As seqüências são adicionadas ao alinhamento múltiplo obedecendo à estrutura da árvore guia, das folhas para a raíz. Na comparação de pares de seqüências, utiliza-se o parâmetro do ClustalW "lento e acurado" (programação dinâmica) e a matriz de similaridade BLOSUM 30.

Um detalhe importante na construção do alinhamento múltiplo é que apenas as subseqüências que formaram os alinhamentos locais encontrados pelo Blast são passadas para o ClustalW para construção do alinhamento múltiplo. Com isto, implementa-se um processo de alinhamento múltiplo baseado em programação dinâmica e alinhamentos locais.

• Cálculo de Entropia Relativa

Para cada posição do alinhamento múltiplo é calculada a freqüência relativa para cada um dos possíveis 20 aminoácidos. A Entropia para a variável x é dada por:

$$S(x) = -\sum_x p_x \log(p_x) \quad (2)$$

onde p_x é a freqüência relativa do aminoácido x para $1 \leq x \leq 20$.

A Entropia Relativa para a variável x relativa à variável y, onde y possui distribuição uniforme, é calculada da seguinte forma:

$$relent(x, y) = \frac{S(x)}{S(y)} \quad (3)$$

• Medidas de similaridade

Para medir a similaridade entre os valores de Entropia Relativa obtidos através do processo proposto e os valores relatados pelo HSSP, foram utilizadas duas medidas, o desvio médio quadrático (*rmsd)* e o cosseno (*cosine)*, dadas por:

$$rmsd = \sqrt{\frac{1}{N}\sum_{r=1}^{r=N}(r_x - r_y)^2} \quad (4)$$

onde: r denota um aminoácido da seqüência protéica considerada;

N denota o comprimento da seqüência protéica considerada;

r_x denota o valor de Entropia Relativa para o aminoácido r calculado através de SH2Q[2];

r_y denota o valor de Entropia Relativa para o aminoácido r relatado pelo HSSP.

[6] O valor de E-value pode ser relacionado com o valor de P-value de acordo com a seguinte fórmula:
P – value = 1 – e $^{-E\text{-value}}$.

$$\cos ine = \frac{\langle X, Y \rangle}{\|X\| \|Y\|} \quad (4)$$

onde: X denota um vetor, onde cada elemento X_i indica o valor de Entropia Relativa, calculado através de SH2Q[2], do aminoácido na posição i da seqüência protéica considerada;

Y denota um vetor, onde cada elemento Y_i indica o valor de Entropia Relativa, relatado pelo HSSP, do aminoácido na posição i da seqüência protéica considerada.

• Arquivos PDBs utilizados

Os seguintes arquivos PDBs foram utilizados neste experimento: 1muo, 1ps2, 1ijc. Todos eles descrevem uma única seqüência de proteína. Para cada uma dessas seqüências, foi construído um alinhamento múltiplo utilizando-se o processo proposto. Para cada posição do alinhamento obtido foi calculada a Entropia Relativa. Esses valores foram, então, comparados com os valores relatados pelo HSSP, utilizando-se o desvio médio quadrático e o cosseno como medidas de similaridade.

Resultados e Discussão

O número de seqüências obtido com a busca utilizando Blast foi sempre maior que o número de seqüências relatadas pelo HSSP. Uma possível explicação para este fato é que o processo de busca utilizado na construção do HSSP esteja sendo mais restritiva que o utilizado para construção do SH2Q[S] ou a sua atualização não reflita a última versão liberada do SWISS-PROT, pois o banco de dados do HSSP deveria ser completamente reconstruído a cada versão liberada do SWISS-PROT. De qualquer forma, a maior disponibilidade de seqüências para construção do alinhamento resultará em um valor mais confiável de Entropia Relativa relatada pelo SH2Q[S]. Para remover a influência destas seqüências adicionais e viabilizar a análise dos valores de Entropia Relativa baseado na utilização do ClustalW para construção do alinhamento múltiplo, apenas aquelas seqüências encontradas através do Blast e que também são relatadas pelo HSSP foram consideradas. Por isso, o número de seqüências alinhadas nos 3 experimentos foi sempre inferior ao relatado pelo HSSP.

As Fig. 2, 3 e 4 apresentam comparações dos valores de Entropia Relativa para cada resíduo de cada um dos arquivos PDB utilizados nos experimentos. A similaridade relativa aos valores relatados pelo HSSP

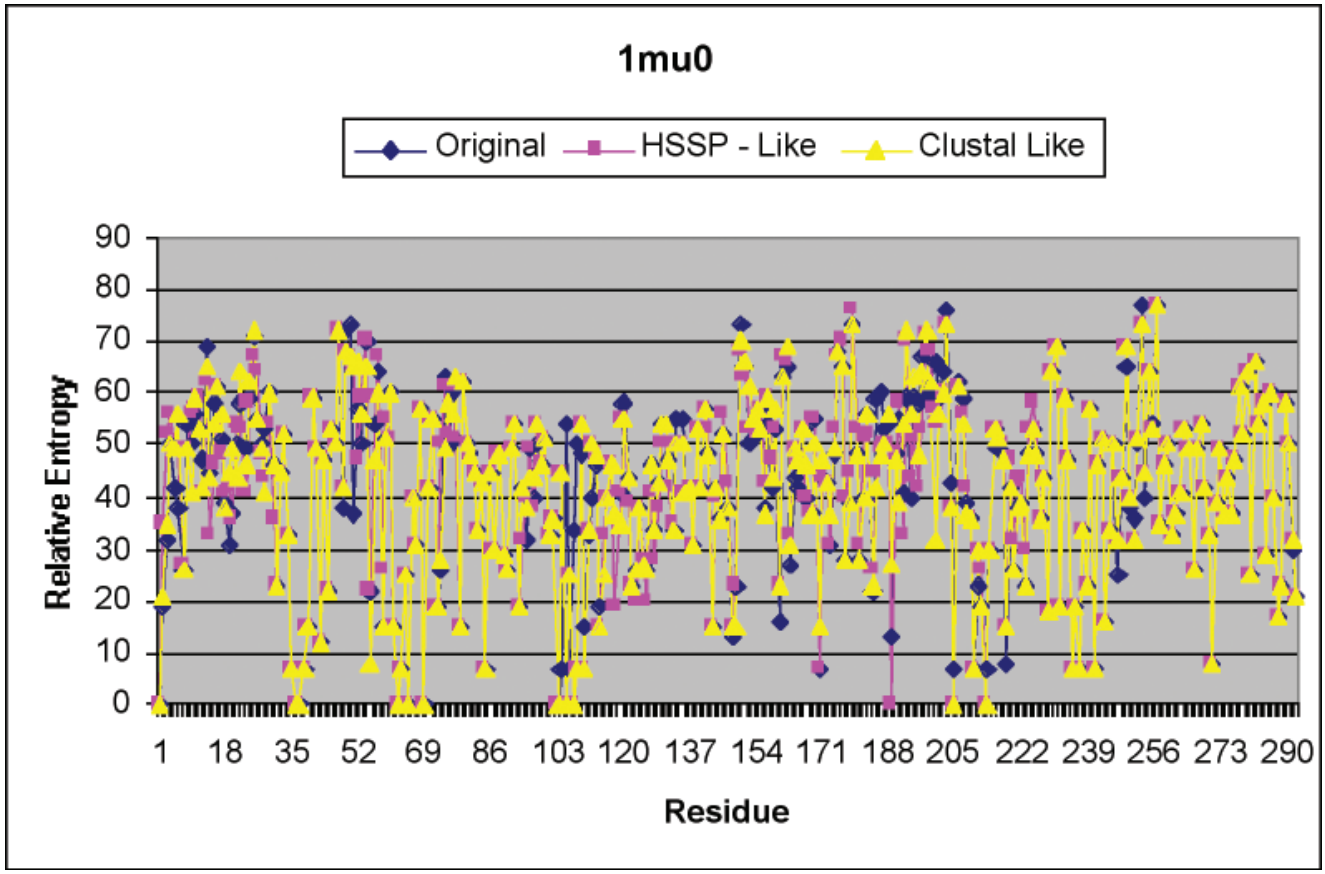


Fig. 2. Comparação dos valores de Entropia Relativa - pdb 1mu0.

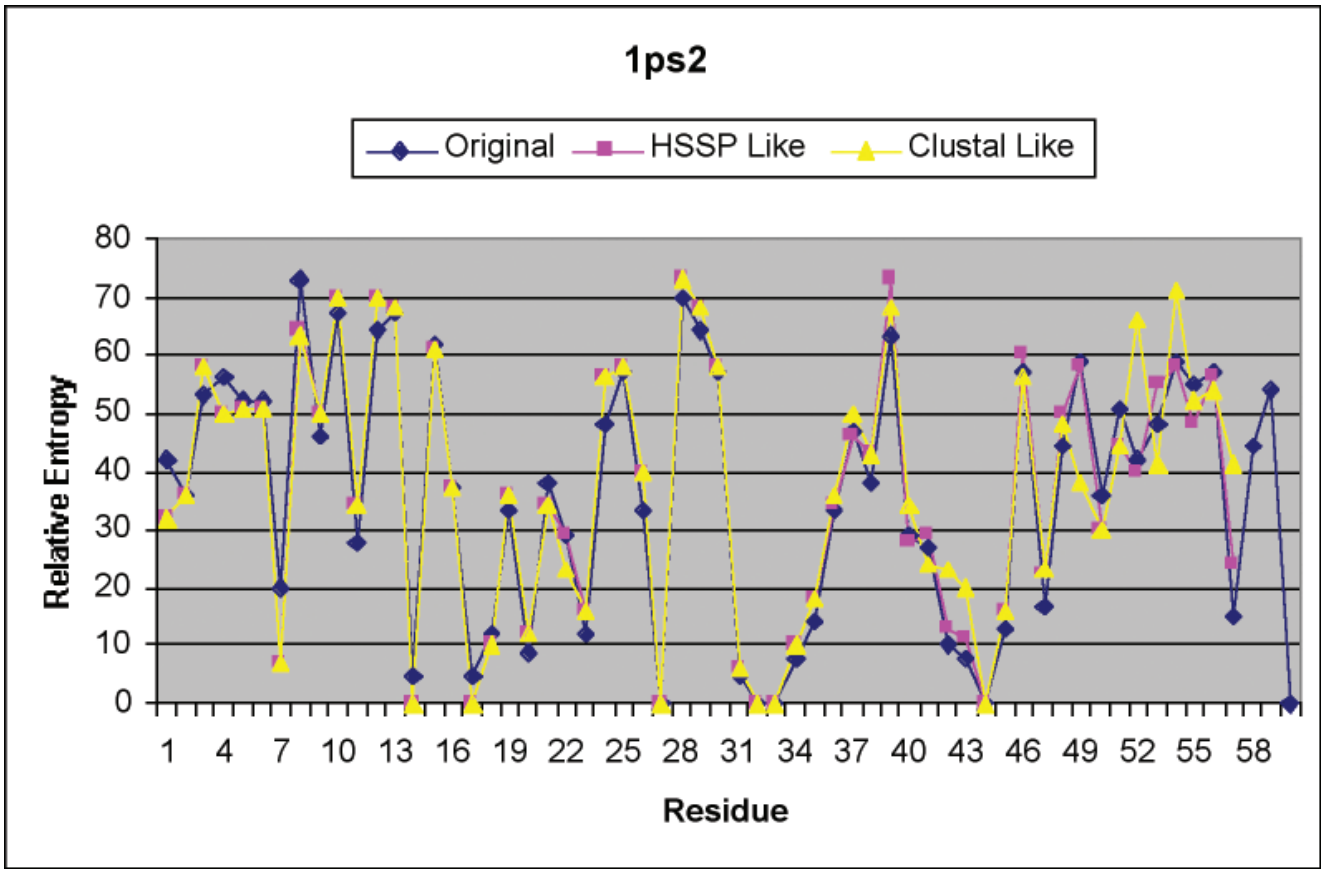


Fig. 3. Comparação dos valores de Entropia Relativa – pdb 1ps2.

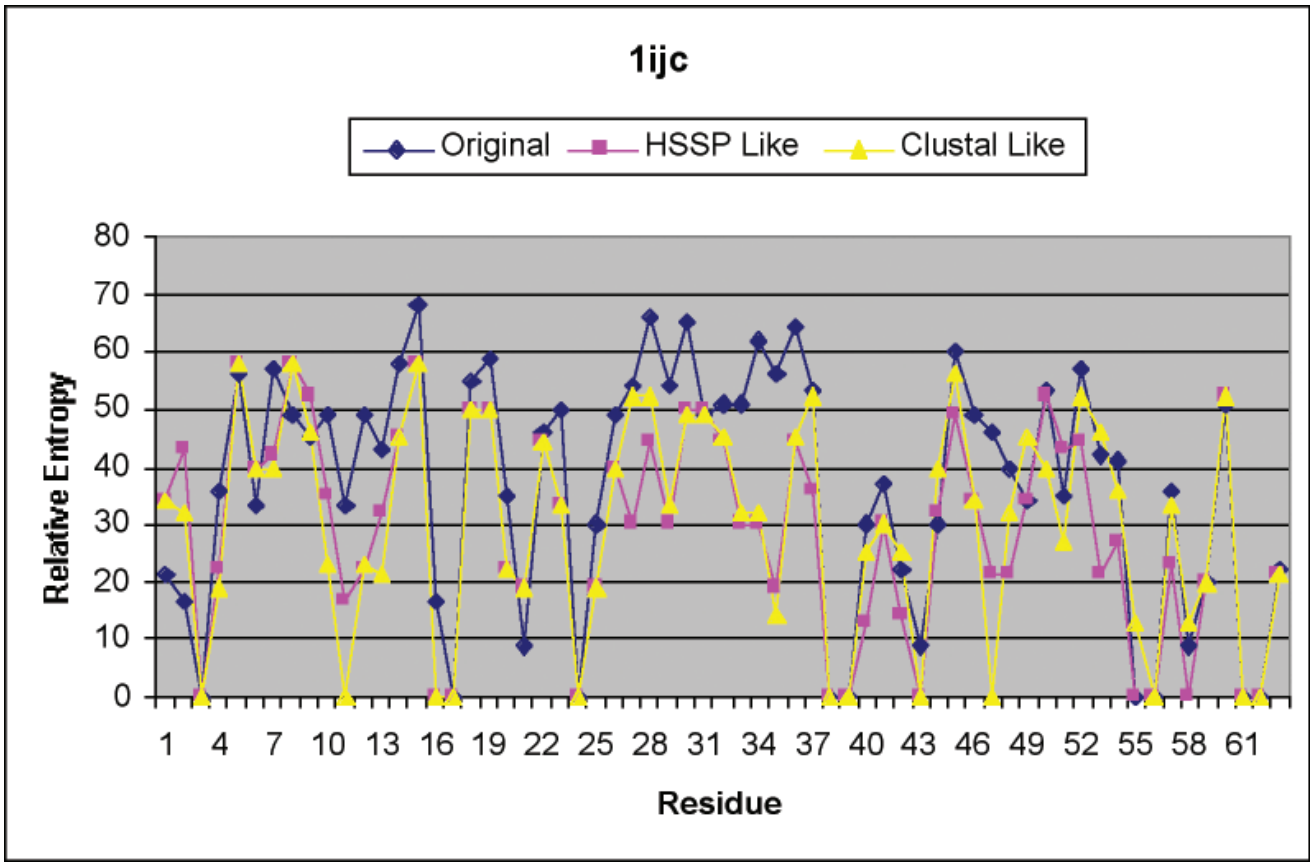


Fig. 4. Comparação dos valores de Entropia Relativa – pdb 1ijc.

variam de 4,733234 até 14,40954 para o *rmsd*, e de 0,946823 até 0,99403 para o *cosine*, sendo que os resultados mais similares foram obtidos para o PDB 1ps2 e os menos similares para o PDB 1ijc.

A Tabela 1 apresenta um sumário dos resultados obtidos para cada arquivo PDB. As duas medidas de similaridade indicam que o processo proposto produz um alinhamento similar ao relatado pelo HSSP, com valores de Entropia Relativa muito próximos. A maior variação obtida para o pdb 1ijc pode ser explicada pelo número de seqüências consideradas no alinhamento: dentre as 13 seqüências relatadas pelo HSSP, apenas 6 foram utilizadas no alinhamento. Cabe observar que para o PDB 1ijc, mesmo com um desvio maior em relação aos valores relatados pelo HSSP, os valores obtidos apresentaram uma tendência similar à obtida quando a Entropia Relativa é calculada para um subconjunto das seqüências em um alinhamento, ou seja, devido à menor quantidade de dados, a variabilidade e, por conseqüência, a Entropia Relativa tendem a apresentar um valor menor. Isto pode ser observado nas regiões entre os resíduos 9-17 e 25-35. Cabe ressaltar que a confiabilidade desse valor de Entropia Relativa é menor quando comparado ao valor obtido na situação em que se dispunha de mais seqüências alinhadas. Ainda referente ao pdb 1ijc, o valor do *cosine* muito próximo de 1 indica que, mesmo com um desvio maior, os trechos identificados como mais e menos conservados na seqüência são coincidentes com os relatados pelo HSSP. Em outras palavras, o perfil dos valores de Entropia Relativa é o mesmo que o relatado pelo HSSP (Fig.4).

Com respeito à ordem em que as seqüências são alinhadas ao alinhamento múltiplo – *Clustal Like* e *HSSP Like*, para os PDBs 1mu0 e 1ijc, os valores obtidos, tanto para o *rmsd* quanto para o *cosine*, foram muito similares nos dois casos, enquanto que para o PDB 1ps2, a estratégia *HSSP Like* apresentou um resultado um pouco superior. Dessa forma, os resultados obtidos são inconclusivos. Caso uma das estratégias tivesse se mostrado claramente superior à outra, esta informação poderia ser utilizada para guiar o esforço de implementação do processo automatizado de geração do SH2Q^{S}. A partir dos resultados obtidos, essa definição só poderá ser feita ou a partir de novos experimentos ou pela utilização de outros critérios como eficiência computacional da implementação resultante.

Tabela 1. Valores de similaridade para os pdbs 1mu0, 1ps2 e 1ijc.

	1mu0		*1ps2*		*1ijc*	
	HSSP Like	*Clustal Like*	*HSSP Like*	*Clustal Like*	*HSSP Like*	*Clustal Like*
Rmsd	7,251059	7,186589	4,733234	7,606553	14,11686	14,40954
Cosine	0,982856	0,98257	0,99403	0,984657	0,955148	0,946823

Considerações Finais

O experimento realizado mostrou que o processo proposto, com parâmetros:

- E-value igual a 0.01 e matriz de similaridade BLOSUM 60 para Blast; e inserção de gap igual a 3.0, alongamento de gap igual a 0.1, série de matrizes de similaridade BLOSUM e definição de resíduos hidrofílicos e atribuição de valores de penalidades específicos, separação de gaps e adiamento da incorporação de seqüências muito divergentes ao alinhamento progressivo desabilitados para ClustalW, pode ser utilizado para obter alinhamentos semelhantes aos relatados pelo HSSP (obtidos utilizando-se o programa MaxHom), com valores de Entropia Relativa para cada aminoácido também muito semelhantes. Embora, este seja um resultado preliminar, pois foram utilizados apenas 3 arquivos PDBs (1mu, 1ps2 e 1ijc) e cada um com não mais que 20 seqüências alinhadas, ele é suficientemente encorajador para que esforços sejam realizados no sentido de automatizar o processo proposto.

Uma vez que se tenha um processo automatizado, novos experimentos com um número mais representativo de PDBs deverão ser realizados para:

- avaliar a similaridade entre os valores de Entropia Relativa dos alinhamentos obtidos com relação aos relatados pelo HSSP, o que confirmaria os resultados parciais aqui apresentados;
- medir a performace do alinhamento para os casos em que o número de seqüências homólogas detectadas é muito grande, o que permitirá estabelecer os parâmetros para execução *on-line* do processo;

- reavaliar as estratégias de adição de seqüências ao alinhamento múltiplo, pois o resultado do experimento realizado foi inconclusivo;
- avaliar valores diversos para os parâmetros do Blast e, principalmente, do ClustalW, promovendo um ajuste fino ao processo proposto.

Finalmente, algumas melhorias e extensões serão avaliadas, podendo vir a ser ou não implementadas como, por exemplo, a utilização da nova função de *threshold* proposta por Rost (1999) e a construção de uma base de dados inversa ao HSSP, organizada por entradas do SWISS-PROT e não do PDB (Schafferhans et al., 2003).

Referências Bibliográficas

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 Acids Res., 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 Res., 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, p. 235-242, 2000.

DODGE, C.; SCHNEIDER, R.; SANDER, C. The HSSP database of protein structure-sequence alignments and family profiles. Nucleic Acids Res., v. 26, n. 1, p. 313-315, 1998.

KARLIN, S.; ALTSCHUL, S. F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. USA, v. 87, n. 6, p. 2264-2268, 1990.

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 Res., v. 31, n. 13, p. 3386-3392, 2003.

ROST, B. Twilight zone of protein sequence alignments. Protein Engineering, n. 12, p. 85-94, 1999.

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.

SCHAFFERHANS, A.; MEYER, J. E. W.; O'DONOGHUE, S. I. The PSSH database of alignments between sequences and tertiary structures. Nucleic Acids Res., v. 31, n. 1, p. 494-498, 2003.

THOMPSON, J. D.; HIGGINS, D. G.; GIBSON, T .J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res., v. 22, p. 4673-4680, 1994.

Comunicado Técnico, 48

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

Ministério da Agricultura, Pecuária e Abastecimento

Governo Federal

1ª edição
2003 - on-line