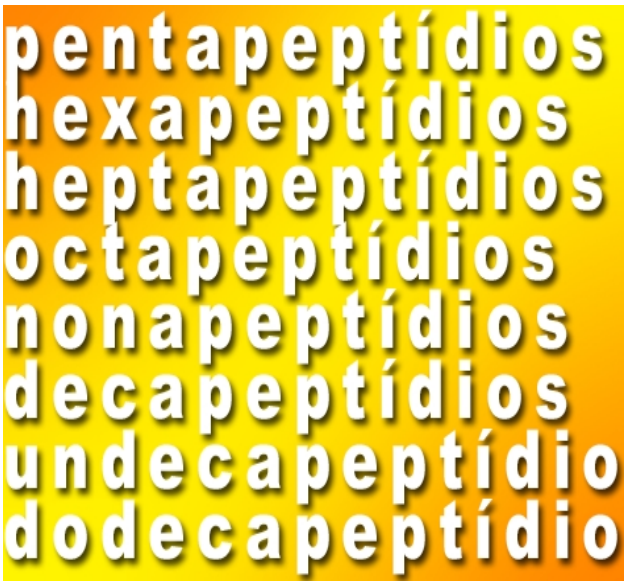
Ministério da Agricultura, Pecuária e Abastecimento

Comunicado Técnico 42

Dezembro, 2002
Campinas, SP

ISSN 1677-8464



Estudo da Influência de Seqüências Polipeptídicas e de Códons na Determinação da Estrutura Secundária de Proteínas

Goran Neshich[1]
Jair Lage de Siqueira Neto[2]

Acredita-se que a seqüência linear de aminoácidos que forma uma proteína é de alguma forma responsável pela determinação de sua estrutura espacial, ainda que não seja um processo completamente compreendido (Anfinsen & Taniuchi, 1966). As interações não covalentes são as forças mais importantes para estabilidade das estruturas protéicas (Stryer, 1995).

A estrutura secundária de proteínas é formada por conformações locais dos polipeptídios (Creighton, 1993). Pontes de hidrogênio entre resíduos linearmente próximos estabilizam estas conformações locais que formam padrões encontrados em todas as proteínas. Dois desses padrões de estrutura secundária mais comuns são as hélices alfa e as fitas beta (Chothia et al., 1997).

Uma hélice alfa surge a partir da formação de pontes de hidrogênio entre o átomo de hidrogênio ligado ao átomo de nitrogênio eletronegativo da ligação peptídica e ao átomo de oxigênio eletronegativo da carbonila do quarto aminoácido seguinte no sentido carboxi-terminal das ligações peptídicas. É uma estrutura consideravelmente estável tendo em vista que todos os resíduos participam das pontes de hidrogênio (com exceção dos resíduos das extremidades da hélice) (Lehninger et al., 2000).

Na conformação de fitas beta, a cadeia principal polipeptídica forma uma estrutura de *zig-zag* em que os resíduos se arranjam lado a lado. Neste caso, as pontes de hidrogênio são formadas entre segmentos adjacentes da cadeia de polipeptídios. As cadeias adjacentes podem ser tanto paralelas quanto antiparalelas no caso de fitas beta, dependendo da orientação amino-carboxila (Lehninger et al., 2000)

Com as informações já concedidas de como se estabilizam as estruturas secundárias, pode-se entender a importância da seqüência de aminoácidos na proteína para formação dessas conformações locais. No entanto, esta seqüência linear dos aminoácidos não é a única determinante da formação da estrutura secundária. Temos alguns casos observados a partir de proteínas cristalizadas e resolvidas por raios X que uma mesma seqüência de determinada extensão de aminoácidos pode dar origem a diferentes estruturas secundárias.

O propósito desse trabalho foi descrever e quantificar os casos de polipeptídicos idênticos em diferentes estruturas secundárias, encontrados em banco de dados de estruturas: *Protein Data Bank* (PDB) (Berman et al., 2000) ou bancos de dados derivados do PDB. Em etapa posterior, procurar-se-á entender melhor os mecanismos de formação de estruturas secundárias, a

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

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

partir da hipótese de que a seqüência nucleotídica que codifica a seqüência de aminoácidos, pode estar influenciando de alguma forma, a interação entre resíduos espacialmente próximos e conseqüentemente contribuindo para a determinação da estrutura secundária do polipeptídio.

Material e Métodos

Como primeiro passo em busca dos nossos objetivos, deveria ser selecionada uma base de dados não redundante contendo estruturas protéicas com baixo grau de similaridade. A base de dados escolhida foi obtida no *ASTRAL SCOP Genetic Domain Sequence* (Brenner et al., 2000) que é derivada da base de dados *PDB SEQRES*. Foram então escolhidos três bancos de dados com diferentes níveis de similaridade: ASTRAL SCOP 10, ASTRAL SCOP 40 e ASTRAL SCOP 70, com valores máximos de 10%, 40% e 70% de identidade respectivamente. Estas bases de dados foram obtidas em http://astral.stanford.edu/scopseq-1.61.html no mês de agosto de 2002.

Em seguida, foi desenvolvido um programa que executa uma busca exaustiva de todos os possíveis padrões de tetrapeptídios presentes em cada uma das bases escolhidas. O programa gera ainda um arquivo contendo o resíduo seguinte a cada padrão de tetrapeptídio encontrado. Com isso, para se obter os pentapeptídios, não seria mais necessária uma busca exaustiva, que se tornaria computacionalmente inviável para padrões um pouco maiores, exigindo um custo de processamento incompatível com nossa atual estrutura. Os pentapeptídios seriam obtidos a partir do arquivo gerado na busca dos tetrapeptídios. E ao se fazer a busca para os pentapeptídios, já se gera um novo arquivo que executa a busca dos hexapeptídios e assim por diante. Esse processo foi executado até encontrar polipeptídios de tamanho 12, nas três bases de dados.

A partir deste ponto, buscou-se as estruturas secundárias de cada um dos polipeptídios obtidos. A base de dados contendo as estruturas secundárias foi obtida em ftp://ftp.rcsb.org/pub/pdb/derived_data/ss.txt disponibilizada pelo PDB. Neste arquivo, a estrutura secundária foi determinada de acordo com a descrição e implementação do método DSSP de Kabsch & Sander (1983). A descrição das estruturas secundárias em 7 grupos (*helix*; *residue in isolated beta bridge*; *extended beta strand*; *3 10 helix*; *pi helix*; *hydrogen bonded turn*; e *bend*) não é a mais indicada para fazer o tipo de análise que se pretende. Por isso usou-se outra classificação em apenas 3 grupos: *Helix*, *Strand* e *Coil*. Para fazer a conversão, um script interpreta como *Helix*: *helix*, *3 10 helix* e *pi helix*; como *Strand*: *residue in isolated beta bridge* e *extended beta strand;* e como *Coil*: *hydrogen bonded turn*, *bend* e *"blank"*. Com esta conversão as análises foram significativamente facilitadas.

Um estudo paralelo, ainda em fase de desenvolvimento, é a análise de uma possível relação entre as trincas de ácido nucléico (códons) e a estrutura secundária do polipeptídios por elas codificadas. Para esse estudo, a base de dados de estrutura escolhida foi o próprio *Protein Data Bank*, o maior banco de dados de estruturas da atualidade. O *download* dos arquivos .pdb desta base de dados foi feito na data de 11/dez/2002 em ftp://ftp.rcsb.org/pub/pdb/data/structures/all/pdb/ .

A partir dos arquivos .pdb, selecionou-se somente àqueles que se referem a polipeptídios, excluindo-se seqüências nucleotídicas, pequenas moléculas e outras estruturas depositadas que não são proteínas. Deveria-se também ter um grau de confiabilidade na estrutura secundária descrita nos arquivos .pdb, e para isso selecionou-se apenas aqueles que apresentavam resolução da estrutura maior que 2.0Å, o que acarretaria em uma maior precisão nas análises posteriores.

A seqüência polipeptídica de cada cadeia dos arquivos .pdb selecionados foi então obtida no arquivo pdb_seqres.txt em formato pasta, disponível em ftp://ftp.rcsb.org/pub/pdb/derived_data/seqres.txt. A obtenção da seqüência nucleotídica codificadora de um determinado polipeptídio foi feita através do programa *Blast* (Altschul et al.,1990), utilizando mais especificamente o programa *TblasntN*, que compara a seqüência do polipeptídios contra uma tradução nos 6 *frames* de cada seqüência de um banco de nucleotídeos não redundante. O banco escolhido foi o nt, constituído por todas seqüências nucleotídicas não redundantes presentes em GenBank (NCBI Gene Data Bank) + EMBL (European Molecular Biology Laboratory - Nucleotide Sequence Database) + DDBJ (DNA Database of Japan) + PDB (Protein Data Bank), excluindo-se EST (Expressed Sequence Tags), STS (Sequence Tagged Sites), GSS (Genome Survey Sequence) e HTGS (High Throughput Genomic Sequences).

Foi desenvolvido um algoritmo capaz de executar o programa *TblastN* localmente de forma automática para todas as seqüências polipeptídicas obtidas a partir dos arquivos .pdb selecionados. Desta maneira reduziu-se drasticamente o tempo de execução da atividade, viabilizando este tipo de estudo em larga escala.

Em seguida, os formulários de saída do programa *TblastN* foram avaliados para verificar se o melhor *match* teve porcentagem de identidade maior que 95%. Para os casos que não superaram esse valor, a seqüência de nucleotídica foi excluída. Para os demais casos, considerou-se que a seqüência nucleotídica tem alta probabilidade de ter codificado o polipeptídio em questão, portanto sendo aceita.

A partir deste momento, deverá ser feito uma comparação entre as seqüências nucleotídicas e protéicas, pareando-se cada aminoácido com seu

códon respectivo e ainda anexar a informação de qual estrutura secundária o resíduo está fazendo parte. Desta forma poderão ser então feitos cálculos para se saber em que medida os códons estão influenciando a determinação da estrutura secundária.

Resultados

Os resultados demonstrando quantitativamente os polipeptídios idênticos e suas relações com estruturas secundárias são demonstrados nas Tabelas a seguir.

Tabela 1. Resultados obtidos na base de dados ASTRAL SCOP 10.

ASTRAL SCOP 10	*Número de padrões com mínimo de 2 ocorrências*	*Número de ocorrências do padrão nas proteínas*	*Média (prot/ padrão)*	*Padrões com somente 1 tipo de estr. sec.*	*Padrões com mais de 1 tipo de estr. sec.*	*% de polipep. que apresentam 1 único padrão de estr.sec.*
Tetrapeptídios	92.650	471.443	5,09	3.993	88.657	4,31
Pentapeptídios	65.487	146.713	2,24	6.168	59.319	9,42
Hexapeptídios	6.213	12.643	2,03	578	5.635	9,30
Heptapeptídios	452	939	2,08	52	400	11,50
Octapeptídios	50	128	2,56	10	40	20,00
Nonapeptídios	10	45	4,50	5	5	50,00
Decapeptídios	5	33	6,60	3	2	60,00
Undecapeptídios	3	27	9,00	2	1	66,67
Dodecapeptídios	1	21	21,00	0	1	0,00
Total	164.871	631.992	3,83	10.811	154.060	6,55

Tabela 2. Resultados obtidos na base de dados ASTRAL SCOP 40.

ASTRAL SCOP 10	*Número de padrões com mínimo de 2 ocorrências*	*Número de ocorrências do padrão nas proteínas*	*Média (prot/ padrão)*	*Padrões com somente 1 tipo de estr. sec.*	*Padrões com mais de 1 tipo de estr. sec.*	*% de polipep. que apresentam 1 único padrão de estr.sec.*
Tetrapeptídios	112.327	750.585	6,68	3.639	108.688	3,23
Pentapeptídios	128.951	304.529	2,36	11.371	117.580	8,81
Hexapeptídios	15.911	32.712	2,06	1.753	14.158	11,01
Heptapeptídios	1.974	4.092	2,07	322	1.652	16,31
Octapeptídios	642	1.354	2,11	133	509	20,71
Nonapeptídios	345	736	2,13	71	274	20,57
Decapeptídios	223	474	2,13	42	181	18,83
Undecapeptídios	145	311	2,14	24	121	16,55
Dodecapeptídios	92	203	2,21	12	80	13,04
Total	260.610	1.094.996	4,20	17.367	243.243	6,66

Tabela 3. Resultados obtidos na base de dados ASTRAL SCOP 70.

ASTRAL SCOP 10	Número de padrões com mínimo de 2 ocorrências	Número de ocorrências do padrão nas proteínas	Média (prot/ padrão)	Padrões com somente 1 tipo de estr. sec.	Padrões com mais de 1 tipo de estr. sec.	% de polipep. que apresentam 1 único padrão de estr.sec.
Tetrapeptídios	122.671	999131	8,14	3.682	118.989	3,00
Pentapeptídios	196.858	494002	2,51	20.815	176.043	10,57
Hexapeptídios	45.117	98106	2,17	9.814	35.303	21,75
Heptapeptídios	16.530	29460	2,18	4.023	9.507	29,73
Octapeptídios	9.929	21399	2,16	2.872	7.057	28,95
Nonapeptídios	8.005	17004	2,12	2.149	5.856	26,84
Decapeptídios	6.671	13980	2,10	1.635	5.036	24,50
Undecapeptídios	5.634	11664	2,07	1.253	4.381	22,23
Dodecapeptídios	4.850	9931	2,05	994	3.856	20,40
Total	416.265	1.694.677	4,07	47.237	366.028	11,35

Tabela 4. Descrição dos parâmetros apresentados nas Tabelas 1, 2 e 3.

Número de padrões com mínimo de 2 ocorrências	Se refere à quantidade de padrões polipeptídicos que se repetem pelo menos uma vez, ou seja, com no mínimo duas ocorrências na base de dados.
Número de ocorrências do padrão nas proteínas	Referente a quantas vezes os padrões polipeptídicos foram encontrados nas bases de dados, contando-se somente casos onde o padrão se repete pelo menos 1 vez
Média (prot/padrão)	É a média aritmética do número de vezes que os padrões polipeptídicos foram encontrados pela quantidade de padrões encontrados.
Padrões com somente 1 tipo de estr. sec.	Referente à quantidade de padrões polipeptídicos que se repetem pelo menos uma vez e apresentam sempre uma mesma estrutura secundária.
Padrões com mais de 1 tipo de estr. sec.	Se refere à quantidade de padrões polipeptídicos que se repetem pelo menos uma vez e apresentam mais de uma estrutura secundária
% de polipep. Que apresentam 1 único padrão de estr.sec.	É a porcentagem (em relação ao número total de padrões polipeptídicos) do número de padrões polipeptídicos que apresentam uma única estrutura secundária

Já na etapa posterior de obtenção de seqüências nucleotídicas, buscou-se a partir de 19.469 entradas da base de dados do PDB, somente as estruturas correspondentes a proteínas, descartando-se seqüências nucleotídicas e pequenas moléculas, reduzindo o universo dos dados.

O primeiro filtro aplicado se referiu à resolução da estrutura, fixado em 2,0Å. Somente as estruturas com resolução maior que este valor estipulado fariam parte da pesquisa. Totalizaram-se 1.759 entradas de arquivos .pdb a partir do primeiro filtro, correspondendo a um total de 2.959 cadeias polipeptídicas, já que cada entrada de pdb pode apresentar mais de uma cadeia protéica.

Foi então realizado o "*TblastN* de todas essas 2.959 cadeias. Destes, 244 formulários não apresentaram nenhuma seqüência nucleotídica correspondente, apresentando a mensagem de "*No hits found*".

O segundo filtro aplicado se refere a porcentagem de identidade do melhor *match*. Apenas 442 cadeias apresentaram identidade superior a 95%, sendo então consideradas as seqüências nucleotídicas correspondentes as seqüências protéicas.

Será então feito, em uma próxima etapa, o alinhamento entre as seqüências nucleotídicas e de proteínas para pareamento dos códons com os resíduos e suas estruturas secundárias. Será feita também uma análise

estatística na tentativa de concluir algo sobre a importância das seqüências de DNA na determinação da estrutura secundária das proteínas

Discussão e Conclusões

Inicialmente pode parecer estranho que haja mais pentapeptídios idênticos que tetrapeptídios idênticos nas bases de dados ASTRAL SCOP 40 e 70, já que cada pentapeptídio é também ao mesmo tempo 2 tetrapeptídios diferentes. É necessário então analisar estes dados obtidos antes de iniciar a discussão realmente importante do estudo.

A ocorrência de pentapeptídios idênticos mais numerosa que a de tetrapeptídios nas bases de dados ASTRAL SCOP 40 e ASTRAL SCOP 70 se deve a dois fatores: tamanho das bases de dados e possibilidades de combinação dos 20 aminoácidos para se formar tetra e pentapeptídios. A base de dados ASTRAL SCOP 10 tem 2.915 cadeias e 517.628 resíduos; ASTRAL SCOP 40 tem 4.383 cadeias e 796.111 resíduos; e ASTRAL SCOP 70 tem 5.809 cadeias e 1.046.255 resíduos. Com 20 aminoácidos são possíveis 160.000 tetrapeptídios e 3.200.000 pentapeptídios. Fazendo-se cálculos, é possível concluir que é necessário uma base de dados de tamanho mínimo de 320.010 resíduos para tornar possível uma maior ocorrência de pentapeptídios idênticos em relação a tetrapeptídios idênticos. E é fácil notar que quanto maior for a base de dados, menor será a possibilidade de ter-se um polipeptídio de tamanho n repetido mais vezes que um polipeptídio de tamanho n-1 repetido. Melhor explicitando: é necessária uma base de dados de tamanho mínimo de 6.400.012 para ter-se a possibilidade de encontrar mais hexapeptídios idênticos (ou seja, repetidos pelo menos uma vez) que pentapeptídios idênticos, já que pode-se formar 64.000.000 hexapeptídios, mas apenas 3.200.000 pentapeptídios. Por isso, é mais provável que o número maior de tetrapeptídios idênticos seja mais abundante que o de pentapeptídios idênticos na base de dados ASTRAL SCOP 10, do que na base ASTRAL SCOP 40, já que a primeira base é menor que a segunda. Usando este raciocínio, pode-se justificar a maior quantidade de pentapeptídios idênticos que de tetrapeptídios idênticos nas bases de dados ASTRAL SCOP 40 e ASTRAL SCOP 70, mas maior quantidade de tetrapeptídios idênticos que de pentapeptídios idênticos na base ASTRAL SCOP 10.

Em seguida tem-se para outra questão. Pode-se notar inicialmente que a porcentagem de padrões que apresentam uma única estrutura secundária e baixa, não superando os 30%, excetuando-se para polipeptídios com comprimento superior a 9 resíduos na base de dados ASTRAL SCOP 10. Estas porcentagens discrepantes acontecem por causa do pdb 1GKU, que apresenta uma seqüência de polialanina. Neste arquivo existe uma seqüência de 32 resíduos de alanina, sendo entendidos como 24 nonapeptídios idênticos, e apresentam diferentes conformações de estrutura secundária ao longo desta seqüência.

Como as porcentagens de ocorrências de uma única estrutura secundária por um padrão polipeptídico são baixas, pode-se suspeitar que as seqüências dos aminoácidos dentro de uma região contínua e limitada não determinam exclusivamente a formação da estrutura secundária. Nota-se ainda que polipeptídios muito pequenos, como tetra ou pentapeptídios apresentam as menores porcentagens de uma única estrutura secundária por padrão. Isto pode ser explicado pelo fato de que polipeptídios de tamanhos muito pequenos (menores que 7 resíduos) não são suficientes para caracterizar uma estrutura secundária determinada. As maiores porcentagens de uma única estrutura secundária por padrão polipeptídico estão entre os polipeptídios de tamanho 7, 8 e 9 resíduos. Este dado pode significar que as estruturas secundárias têm comprimento médio por volta de 8 resíduos, e por isso, polipeptídios com 7, 8 ou 9 aminoácidos apresentam estrutura secundária mais conservada. Mesmo assim, essas porcentagens, apesar de serem as maiores não superam 30%, como já citado anteriormente. Pode-se então sugerir que a seqüência dos aminoácidos dentro de uma região contínua e limitada está tendo aproximadamente 30% de influência na determinação da estrutura secundária. Resta então saber o que está determinando os outros 70% de influência para formação da estrutura secundária.

Possivelmente a resposta da dúvida supracitada está nas seqüências nucleotídicas que codificam as proteínas. Pelo fato de o código genético ser degenerado a grande maioria dos aminoácidos são codificados por mais de um códon nucleotídica. Por outro lado, cada códon é exclusivo de um aminoácido, e ainda há uma correlação recíproca de exclusividade entre os códons e os RNA transportadores, que carregam os aminoácidos. Esta pode ser a chave da questão da determinação da estrutura secundária.

No processo de tradução protéica, o ribossomo ao interagir com o RNA mensageiro na montagem da proteína, interage também com dois RNA transportadores simultaneamente, promovendo a ligação peptídica entre os dois aminoácidos carregados pelos seus respectivos RNA transportadores. E como os RNAt são exclusivos para os códons, cada RNAt apresenta uma dinâmica de interação com o ribossomo, podendo favorecer a formação de um anglo de ligação entre os aminoácidos, e assim sendo, pode também favorecer a formação de uma determinada estrutura secundária.

Para confirmar esta hipótese, será feito o alinhamento das 442 cadeias polipeptídicas com resolução de raios

X maior que 2,0Å, juntamente com sua estrutura secundária e com as suas seqüências nucleotídicas correspondentes. Desta maneira, será possível relacionar os códons e as estruturas secundárias. Uma análise estatística quantificando o quanto os códons estão relacionados à estrutura secundária será de fundamental importância para melhorar o nosso conhecimento nesta área ainda pouco compreendida.

Referências Bibliográficas

ALTSCHUL, S. F.; GISH, W.; MILLER, W.; MYERS, E. W.; LIPMAN, D. J. Basic local alignment search tool. **Journal of Molecular Biology**, v. 215, n. 3, p. 403-410, Oct. 1990.

ANFINSEN, C. B.; TANIUCHI, H. The Amino acid **sequence of an extracellular nuclease of *Staphylococcus aureus*. The** Journal of Biological Chemistry, v. 241, n. 19, p. 4366-4385, Oct. 1966.

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.

BRENNER, S. E.; KOEHL, P.; LEWITT, M. The ASTRAL compendium for sequence and structure analysis. **Nucleic Acids Research**, v. 28, p. 254-256, 2000.

CHOTHIA, C.; HUBBARD, T. J. P.; BRENNER, S. E.; BARNS, H.; MURZIN, A. Protein folds in the all-a and all-b classes. **Annual Reviews of Biophysics and Biomolecular Structure**, v. 26, p. 597-627, 1997.

CREIGHTON, T. E. (Ed.). **Proteins**: structures and molecular properties. 2nd ed. New York: Freeman, 1993.

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

KABSCH, W.; SANDER, C. On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations. **Proc. Natl. Acad. of Science USA**, v. 81, p. 1075-1078, Feb. 1984.

LEHNINGER, A. L.; NELSON, D. L.; COX, M. M. **Lehninger principles of biochemistry**. 3rd ed. New Yorik: Wohth, 2000. 1152 p.

STRYER, L. **Biochemistry**. 4. ed. New York: W.H. Freeman, 1995. 1064 p.

Comunicado Técnico, 42

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