PROGRAMAS COMPUTACIONAIS AUXILIARES NA BUSCA DE HOMOLOGIAS EM SEQUÊNCIA DE PROTEÍNAS

Proteínas são moléculas orgânicas encontradas nos organismos vivos, com funções específicas, principalmente, catálise (enzimas), transporte de íons e moléculas, formação das estruturas da parede e das membranas celulares, reserva e defesa (ex: anticorpos). A estrutura molecular determina, quase sempre, a função da proteína no organismo. A atividade enzimática e o mecanismo de transporte são exemplos clássicos, em que centros ativos da cadeia protéica estão envolvidos em processos metabólicos vitais. As moléculas protéicas são relativamente grandes, formadas de cadeias de aminoácidos (estrutura primária), que podem adotar conformações moleculares distintas (estruturas secundárias e terciárias), dependendo da ordem em que os aminoácidos aparecem na cadeia. A modificação dessa seqüência é verificada na escala evolucionária por troca, remoção ou inserção de um ou mais aminoácidos, na estrutura primária. Se estas alterações estruturais ocorrem em centros ativos, em decorrência de modificações no código genético, a atividade biológica da proteína pode mudar. O conhecimento da estrutura de uma proteína pode, então, ser a chave para o entendimento do seu modo de ação e abrir perspectivas enormes para a modificação da estrutura, visando obter maior efetividade funcional ou mesmo novas funções biológicas.

A ferramenta mais direta para se conhecer a estrutura de uma proteína é a difração de raios-X. Esta técnica oferece, no entanto, algumas dificuldades operacionais, tais como, complexidade, alto custo, análise bastante demorada para estruturas como as das macromoléculas e dificuldade na obtenção do monocristal.

Alternativamente, existem algoritmos e técnicas computacionais capazes de oferecer modelos satisfatórios das estruturas secundárias e terciárias de proteínas, com base nas estruturas primárias e mais um conjunto de informações adicionais, por exemplo, de diversas espectroscopias (infravermelho, ressonância magnética nuclear etc.) e de conhecimento das estruturas cristalinas de proteínas homólogas. A busca de seqüências homólogas constitui, assim, o primeiro importante passo da modelagem.

As estruturas primárias de proteínas (como as de DNA) estão disponíveis em alguns bancos de dados internacionais, acessáveis e transferíveis por redes remotas de computadores. Uma vez que uma base de dados esteja localmente disponível, o passo seguinte é fazer a busca e a extração do banco de dados daquelas seqüências consideradas parentes da proteína que se deseja modelar. O

International Centre for Genetic Engineering and Biotechnology, em Trieste, Itália, por exemplo, mantém bancos de dados e programas computacionais, aos quais o CNPMS tem acesso via rede remota. Uma primeira análise visual da estrutura primária permite identificar as regiões homólogas, em seqüências protéicas. Existem programas que usam algoritmos numéricos para a busca posterior de homologia em grandes números de estruturas primárias. Havendo, das seqüências homólogas selecionadas, uma ou mais de uma estrutura cristalina conhecida, a modelagem é enormemente dirigida e facilitada.

Apesar de existirem muitos programas de computador, com algoritmos de modelagem disponíveis em laboratórios de pesquisa especializados, alguns desses programas, sobretudo os de comparação de seqüência, devem ser modificados ou desenvolvidos localmente, seja para melhorar a eficiência da análise, seja para adequá-los aos recursos computacionais locais.

A tarefa de desenvolvimento de programas de análise preliminar levou à produção de dois programas de computador, em linguagem C, com o objetivo de fazer a busca e extração de seqüências de proteínas (programa *pullout*) em bancos de dados, no formato PIR (Protein Identification Resource) e apresentar na tela do computador, com opção de impressão, em equipamento padrão HP LaseJetII, as seqüências dispostas lado a lado e os códigos de uma letra dos aminoácidos homólogos marcados, por inversão de tela (programa *pe*), por definição de cores (programa *pec*) ou por variação do tom de cinza, na impressora. Primeiramente, os programas foram desenvolvidos para sistema Unix (originalmente, Xenix, em IBM PC). Posteriormente, foi desenvolvida uma versão do programa *pullout* para o sistema VMS/VAX. Um exemplo de saída do programa *pe* é apresentado na Figura 28a, com critério de identidade dos aminoácidos. O programa com formato de tela colorido permite a utilização de outros critérios, como características físico-químicas, preferência por determinadas estruturas secundárias etc.

O exemplo apresentado na Figura 28a vale-se de quatro proteínas. É, no entanto, apresentada somente uma parte da seqüência total. A primeira delas é uma proteína de ligação de D-ribose (*jgecr*), em *Escherichia coli*. As demais são proteínas periplasmáticas de transporte de arabinose (*plabp*), de leucina + isoleucina + valina (*p2liv*) e uma específica de transporte de leucina (*p2leu*). Conhece-se as estruturas cristalinas destas três últimas. O problema consiste, então, em modelar a estrutura de outras proteínas de transporte em bactérias, como a que serve de exemplo aqui (transporte de ribose, *jgecr*). A análise visual é seguida da busca de regiões homólogas por algoritmos numéricos e gráficos da estrutura tridimensional, até se chegar a um modelo consistente da estrutura que explique sua função. - José Domingos Fabris, Mark Johnson, Tom Blundell.