Quim. Nova, Vol. 26, No. 2, 253-259, 2003

Divulgação

MODELAGEM DE PROTEÍNAS POR HOMOLOGIA

Osvaldo Andrade Santos Filho[#] e Ricardo Bicca de Alencastro*
Departamento de Química Orgânica, Instituto de Química, Universidade Federal do Rio de Janeiro, Centro de Tecnologia , Bloco A, Cidade Universitária, Ilha do Fundão, 21949-900 Rio de Janeiro - RJ

Recebido em 29/1/02; aceito em 23/9/02

PROTEIN HOMOLOGY MODELING. Modeling methods to derive 3D-structure of proteins have been recently developed. Protein homology-modeling, also known as comparative protein modeling, is nowadays the most accurate protein modeling method. This technique can produce useful models for about an order of magnitude more protein sequences than there have been structures determined by experiment in the same amount of time. All current protein homology-modeling methods consist of four sequential steps: fold assignment and template selection, template-target alignment, model building, and model evaluation. In this paper we discuss in some detail the protein-homology paradigm, its predictive power and its limitations.

Keywords: protein structure prediction; homology modeling; structural biology.

INTRODUÇÃO

As maiores esperanças no campo das ciências médicas nas próximas décadas estão, sem dúvida, centradas no Projeto Genoma. Os pares de bases do genoma humano que foram seqüenciados recentemente constituem um "rascunho" importante para a elucidação do genoma completo[1,2]. Desafio maior, entretanto, será a elucidação estrutural de novos alvos moleculares, principalmente proteínas, enzimas e receptores, e novas drogas que certamente emergirão dos dados provenientes do genoma humano e de outros genomas em estudo. Estas estruturas serão a base da revolução da "medicina do futuro" [3], em que a compreensão dos fenômenos biológicos a nível molecular terá papel cada vez mais relevante. Esta nova medicina molecular certamente implicará em avanços significativos das técnicas de diagnóstico e tratamento. Com isso, espera-se que seja possível iniciar os tratamentos clínicos no feto ou na primeira infância, muito antes do surgimento dos primeiros sintomas.

Neste contexto, esforços têm sido feitos em todo o mundo, por instituições governamentais e privadas, no sentido de elucidar o maior número possível de estruturas tridimensionais (estruturas terciárias e quaternárias) de proteínas[4-7]. Apesar das consideráveis inovações técnicas, sobretudo nas áreas de cristalografia de raios-X e difração de nêutrons e de ressonância magnética nuclear (RMN), muitos problemas básicos persistem. A obtenção de amostras em quantidade suficiente para os ensaios necessários é, em muitos casos, difícil e os cristais obtidos nem sempre têm a qualidade necessária para o trabalho experimental (somente uma em cada vinte proteínas, aproximadamente, produz cristais adequados)[3]. Além disso, em certas classes de proteínas, como por exemplo as proteínas de membrana celular, a determinação estrutural é um desafio. Essas proteínas raramente cristalizam e dificilmente podem ser tratadas de modo satisfatório por RMN.

Por outro lado, a elucidação das seqüências de aminoácidos (estruturas primárias) é uma tarefa relativamente mais simples. Por isto, nota-se hoje um grande hiato entre o número de estruturas primárias e secundárias disponíveis. Para se ter uma idéia do problema, em outubro de 2001 o SWISS-PROT[8-10], o mais importante banco de dados de estruturas primárias, incluía 101.602 seqüências de resíduos de aminoácidos e, no mesmo período, somente 14.301 estruturas protéicas estavam disponíveis no PDB[11], o principal banco de dados de estruturas terciárias de proteínas. Em conseqüência, foram desenvolvidos outros métodos de elucidação de estruturas tridimensionais de proteínas. É possível, em princípio, predizer a estrutura tridimensional de proteínas a partir de sua estrutura primária. Este método é conhecido como modelagem de proteínas *ab initio*[12-17]. A determinação da estrutura secundária de proteínas por este método é um sério desafio, sendo hoje o maior problema não resolvido da biologia molecular estrutural[18]. Ainda não dispomos de algoritmos capazes de simular, com precisão, a ação das leis que regem o processo de enovelamento ou empacotamento ("folding"). A principal razão é que estas leis ainda não são perfeitamente conhecidas, o que torna muito difícil obter conformações simultaneamente estáveis e funcionais a custo computacional razoável. Apesar dessas dificuldades, progressos na predição da estrutura tridimensional de peptídeos e pequenas proteínas por métodos *ab initio* têm sido relatados recentemente[12-17].

A ferramenta mais bem sucedida de predição de estruturas tridimensionais de proteínas é a modelagem por homologia, também conhecida como modelagem comparativa ("comparative protein modeling")[19-26]. Esta abordagem baseia-se em alguns padrões gerais que têm sido observados, em nível molecular, no processo de evolução biológica[27]:

(a) homologia entre seqüências de aminoácidos implica em semelhança estrutural e funcional;
(b) proteínas homólogas apresentam regiões internas conservadas (principalmente constituídas de elementos de estrutura secundária: hélices-α e fitas-β);
(c) as principais diferenças estruturais entre proteínas homólogas ocorrem nas regiões externas, constituídas principalmente por alças ("loops"), que ligam os elementos de estruturas secundárias.

Outro fato importante é que as proteínas agrupam-se em um número limitado de famílias tridimensionais[28]. Estima-se que existam cerca de 5.000 famílias protéicas[28]. Conseqüentemente, quando se conhece a estrutura de pelo menos um representante de uma família,

*e-mail: bicca@iq.ufrj.br
[#]Endereço atual: Laboratory of Molecular Modeling and Design (M/C-781), The University of Illinois at Chicago, College of Pharmacy, 833 South Wood Street, Chicago, IL 606012-7231, USA

é geralmente possível modelar, por homologia, os demais membros da família.

Apresentamos neste artigo os princípios básicos da metodologia de modelagem de proteínas por homologia, suas limitações e perspectivas futuras de aplicação.

PROTEÍNAS HOMÓLOGAS

O mecanismo evolucionário de duplicação de genes, associado às mutações, leva a divergências moleculares e, conseqüentemente, à formação de famílias de proteínas estruturalmente relacionadas. Proteínas derivadas de um ancestral comum são ditas **homólogas**[29]. Em função do número de mutações envolvidas, as seqüências de aminoácidos de proteínas homólogas podem ser, idênticas, semelhantes ou dissemelhantes. Conseqüentemente, a semelhança entre as seqüências de aminoácidos em proteínas homólogas, expressa pelo grau (percentual) de identidade, é menos preservada do que a semelhança de estruturas tridimensionais. Em outras palavras, as estruturas tridimensionais de proteínas homólogas tendem a se conservar porque a estrutura ancestral comum é crucial para a manutenção da função das proteínas[29].

A conservação de resíduos em proteínas homólogas é notável. Seqüências de resíduos de aminoácidos de proteínas com cerca de 30% de identidade, apenas, podem ter excelente sobreposição das cadeias principais ("protein backbone"), com desvios de mínimos quadrados ("rmsd") da ordem de 2 Å[30], comparáveis aos valores de "rmsd" da ordem de 0,7 Å encontrados em proteínas idênticas em diferentes formas cristalinas[31] e da ordem da resolução da estrutura cristalográfica de muitas das proteínas disponíveis no PDB.

É importante enfatizar que a conservação da estrutura terciária de proteínas homólogas não é uma propriedade intrínseca das proteínas, mas uma conseqüência da evolução, regida por restrições funcionais. Assim por exemplo, a mudança randômica de 70% dos aminoácidos constituintes de uma proteína levaria certamente a uma grande mudança conformacional e possível perda da função[30].

MODELAGEM DE PROTEÍNAS POR HOMOLOGIA

A modelagem de uma proteína (proteína-problema) pelo método da homologia baseia-se no conceito de evolução molecular. Isto é, parte-se do princípio de que a semelhança entre as estruturas primárias desta proteína e de proteínas homólogas de estruturas tridimensionais conhecidas (proteínas-molde) implica em similaridade estrutural entre elas[31].

Os métodos correntes de modelagem de proteínas por homologia implicam basicamente em quatro passos sucessivos:

- identificação e seleção de proteínas-molde;
- alinhamento das seqüências de resíduos;
- construção das coordenadas do modelo;
- validação.

A Figura 1 mostra um esquema geral do processo de modelagem de proteínas por homologia. Para cada um dos passos existe um grande número de métodos, programas e servidores específicos da rede mundial de computadores (Internet) (Tabela 1). Servidores são computadores que fornecem dados e serviços para a Internet, compartilhando seus recursos.

Identificação e seleção de proteínas-molde

A primeira etapa do método é a identificação de pelo menos uma proteína de estrutura tridimensional conhecida, que servirá de molde para a determinação da estrutura da proteína-problema[23]. Duas situações são, em geral, possíveis:

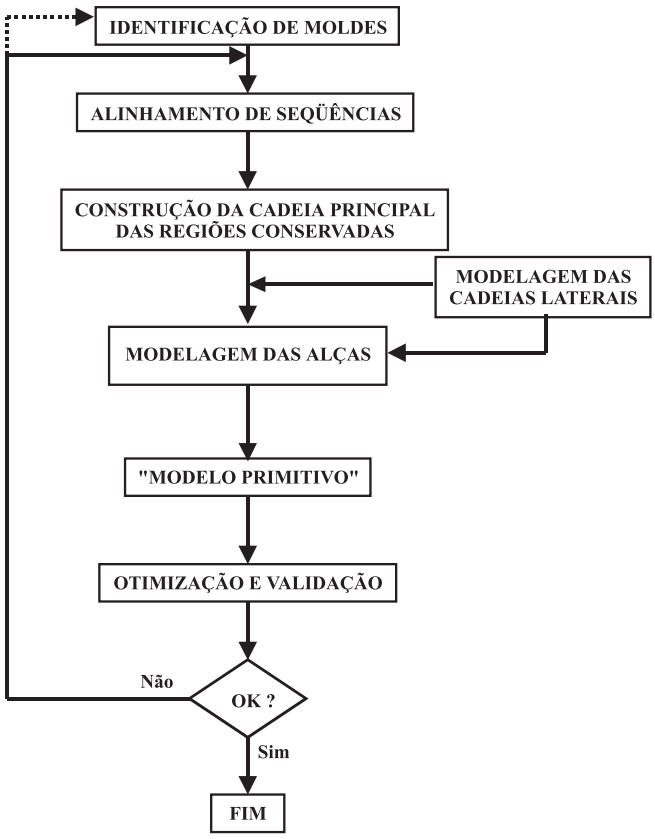

Figura 1. *Esquema geral da modelagem de proteínas por homologia*

(a) conhece-se a família protéica a que pertence a proteína-problema;
(b) não se sabe a que família a proteína-problema pertence.

No primeiro caso, é suficiente selecionar a proteína-molde diretamente do PDB. Como um exemplo deste tipo de seleção de proteínas-molde, veja a referência 20.

No segundo caso, o método mais simples é procurar de forma sistemática um ou mais moldes adequados em um banco de dados de estruturas primárias derivadas de proteínas armazenadas no PDB. Podem ser utilizadas ferramentas de busca, como o BLAST[36] ou o FastA[37], bancos de dados de seqüências de pares de bases de DNA ou de aminoácidos, como por exemplo, o GenBank[34] ou o SWISS-PROT[8-10] ou, ainda, outras ferramentas, como o "Protein Identification Resource" (PIR)[51].

Se o grau de identidade entre as estruturas primárias das proteínas-molde e da proteína–problema for igual ou superior a cerca de 25%, quando o número de resíduos é superior a 80, existe grande probabilidade de que estas proteínas tenham estruturas tridimensionais semelhantes[52] e pode-se construir um modelo para a proteína-problema.

Nos casos em que a identidade entre as seqüências é inferior a cerca de 25%, a melhor maneira de atacar o problema é utilizar métodos de compatibilidade seqüência-estrutura ("threading" ou "3D template matching methods")[53-55]. Estes métodos baseiam-se na avaliação da compatibilidade entre as estruturas terciárias da proteína-problema e de potenciais proteínas-molde. Isto é, escolhe-se a proteína-molde cuja estrutura terciária é "mais adequada" em relação à seqüência da proteína-problema. A avaliação seqüência/estrutura é feita por meio de um potencial empírico derivado de uma tabela de contatos de resíduos observados em proteínas de estrutura conhecida[24]. Esses métodos de seleção de estruturas-molde são menos preci-

Tabela 1. Alguns programas e servidores da Internet úteis na modelagem por homologia (maio de 2002)

Nome	Tipo*	Endereço na *Internet*	Ref.
Bancos de dados			
SRS	s	srs.ebi.ac.uk/	32
CATH	s	www.biochem.ucl.ac.uk/bsm/cath/	33
GenBank	s	www.ncbi.nlm.nih.gov/GenBank	34
MODBASE	s	guitar.rockefeller.edu/modbase/	35
PDB	s	www.rcsb.org/pdb/	11
SWISS-PROT	s	www.ebi.ac.uk/swissprot	8-10
Bioinformática geral			
Expasy	s	www.expasy.org/	
Fontes de estruturas-molde			
BLAST	s	www.ncbi.nlm.nih.gov/BLAST/	36
FastA	s	www2.ebi.ac.uk/fasta3	37
UCLA-DOE FRSVR	s	fold.doe-mbi.ucla.edu/	38
PROFIT	p	www.bioinf.org.uk/software/	39
Alinhamento de seqüências			
BLAST	s	www.ncbi.nlm.nih.gov/BLAST/	40
CLUSTAL	s	www.ebi.ac.uk/clustalw/	41
FastA	s	www2.ebi.ac.uk/fasta3	37
MULTALIN	s	prodes.toulouse.inra.fr/multalin/multalin.html	42
Modelagem de proteínas			
MODELLER	p	guitar.rockefeller.edu/modeller/modeller.html	43
SWISS-MOD	s	www.expasy.ch/swissmod	44
SWISS-MODEL	s	www.expasy.org/swissmod/SWISS-MODEL.html	45-47
SwissPdbViewer	p	ca.expasy.org/spdbv/	47
Validação de modelos			
BIOTECH	s	biotech.embl-heidelberg.de:8400/	48
PROCHECK	p	www.biochem.ucl.ac.uk/~roman/procheck/procheck.html	49
WHATCHECK	p	www.cmbi.kun.nl/swift/whatcheck/	48
PROVE	s	www.ucmb.ulb.ac.be/UCMB/PROVE	50

* s: servidor; p: programa

sos que os métodos precedentes, mas poderão vir a se tornar ferramentas muito úteis para a seleção dos moldes[24].

Alinhamento das seqüências de resíduos de aminoácidos

Definido o molde, passa-se ao alinhamento da seqüência-problema com a seqüência-molde. O objetivo do alinhamento é alinhar resíduos estruturalmente equivalentes levando em conta características estruturais comuns, tais como, elementos de estrutura secundária e resíduos catalíticos[19]. No processo, ocorrem espaços vazios ("gaps"), representados no alinhamento por linhas tracejadas (Figura 2), que correspondem, principalmente, às regiões de alças.

Se, na etapa de identificação de moldes, mais de uma estrutura for selecionada, procura-se alinhar a seqüência-problema com as seqüências de todos os moldes disponíveis. Neste caso, obtém-se um alinhamento múltiplo. Este tipo de alinhamento é considerado

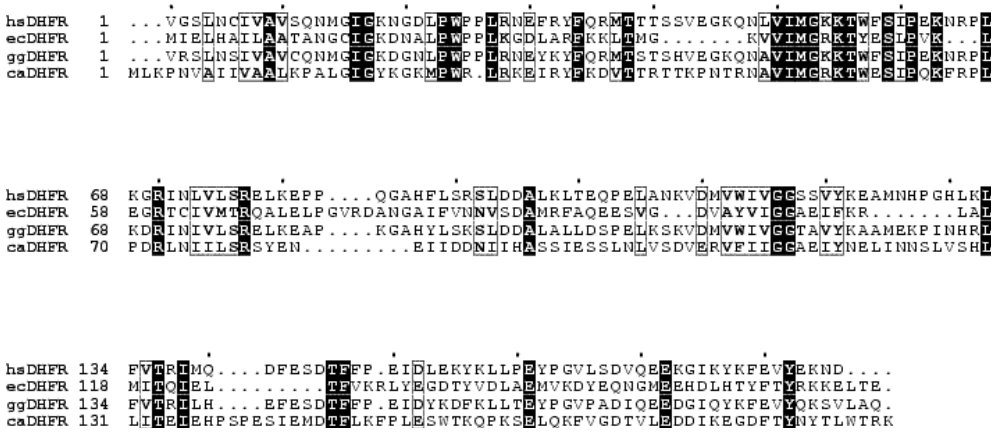

Figura 2. *Alinhamento múltiplo das seqüências de di-hidrofolato redutase (DHFR) humana, de* Escherichia coli*, de fígado de galinha e de* Candida albicans*. As áreas escuras do alinhamento indicam resíduos conservados idênticos e os retângulos indicam regiões em que mais de 80% dos resíduos são semelhantes, de acordo com suas propriedades físico-químicas. O programa CLUSTAL[41] foi utilizado para o alinhamento*

como sendo mais confiável que o alinhamento simples que envolve apenas duas seqüências. A razão disto está em que o alinhamento múltiplo permite detectar mais facilmente as características estruturais comuns de proteínas homólogas.

Como exemplo de alinhamento múltiplo, a Figura 2 mostra um alinhamento de seqüências de di-hidrofolato redutases (DHFRs) de quatro fontes: humana (hsDHFR)[56], de *Escherichia coli* (ecDHFR)[57], de fígado de galinha (ggDHFR)[58] e de *Candida albicans* (caDHFR)[59].

A maior parte dos métodos de alinhamento baseia-se em técnicas de programação dinâmica[60,61]. Os programas mais utilizados nos processos de alinhamentos são o BLAST[40], o FastA[37], o CLUSTAL[41], e o MULTALIN[42], todos disponíveis como servidores na Internet (Tabela 1).

Após o alinhamento, é possível reconhecer regiões estruturalmente conservadas e regiões variáveis. As primeiras correspondem às regiões de máxima similaridade, isto é, em que as conformações são muito semelhantes. Nas regiões variáveis não há, em geral, correspondência estrutural, encontram-se principalmente alças.

O melhor alinhamento de várias seqüências de estrutura conhecida é obtido por sobreposição das moléculas ou a partir de restrições espaciais. Pode-se melhorar a qualidade do alinhamento das seqüências com o auxílio de outras informações. Assim, por exemplo, quando se tem várias seqüências homólogas, pode-se construir um perfil estrutural da família protéica[62]. Além disso, pode-se usar informações provenientes das estruturas-molde como guia para o alinhamento[63]. As informações permitem modificar iterativamente o alinhamento e, portanto, a estrutura tridimensional da proteína-problema. Cautela é importante neste tipo de procedimento para evitar modificações drásticas na conformação do arcabouço estrutural do modelo.

Como a avaliação de um modelo tridimensional é mais fácil que a de um alinhamento de seqüências, um bom procedimento é gerar vários modelos alternativos e escolher o mais satisfatório quando há dúvida sobre o alinhamento[64,65].

Construção do modelo (geração das coordenadas cartesianas)

Modelagem das regiões estruturalmente conservadas

A localização no espaço tridimensional do maior número possível de átomos da proteína-problema depende do alinhamento[23]. A construção da parte interna da proteína-problema baseia-se na idéia de que a conformação da cadeia principal da estrutura-molde pode ser transferida para ela. Existem muitos métodos de modelagem dessas regiões conservadas. Os mais importantes são a modelagem pela união de corpos rígidos ("modeling by rigid-body assembly")[66-68], a modelagem pela combinação de segmentos ("modeling by segment matching")[69,70] e a modelagem pela satisfação de restrições espaciais ("modeling by satisfaction of spatial restraints")[43, 71-73]. Por razões de espaço, não descreveremos estes métodos.

Modelagem das regiões de alças

As diferenças funcionais entre membros de uma mesma família protéica são, em geral, conseqüência de diferenças estruturais na superfície externa das proteínas. Estas diferenças provêm de substituições, eliminações e inserções de resíduos nas cadeias de proteínas homólogas, principalmente nas alças, as regiões mais expostas da proteína. Por serem regiões estruturalmente variáveis, as alças geralmente determinam a especificidade das proteínas, contribuindo para a estrutura dos sítios de ligação (no caso de enzimas, sítios ativos). Conseqüentemente, a modelagem das alças é uma etapa essencial na geração de modelos protéicos por homologia[74].

A modelagem das alças pode ser vista como um problema de enovelamento de proteínas. A conformação de um segmento de cadeia polipeptídica deve ser obtida a partir de sua seqüência. As alças, entretanto, são geralmente pequenas e isto provoca distorções de enovelamento difíceis de avaliar *a priori*[74]. Além disto, a conformação de um dado segmento (de alça) é influenciada pela estrutura do resto da proteína[74]. Em outras palavras, as conformações das alças devem ter baixa energia interna e não interagir desfavoravelmente com o restante da proteína[65]. Devido a estas complicações, a modelagem de alças é, sem dúvida, a etapa mais difícil da predição de modelos de proteínas. Conseqüentemente, é nestas regiões que a probabilidade de ocorrerem erros de modelagem é maior.

Os métodos de predição de alças classificam-se em: métodos *ab initio*[75-77] e métodos comparativos, baseados na semelhança com fragmentos de estruturas conhecidas[78-80]. A predição de alças por métodos *ab initio* baseia-se em uma análise conformacional que depende do ambiente estrutural e é guiada por uma função de energia. Os métodos comparativos utilizam bancos de dados de estruturas de alças de proteínas armazenadas no PDB, nos quais pode-se procurar, para cada alça da proteína-problema, estruturas que tenham as conformações energética e geometricamente mais favoráveis. Uma vez obtidos os modelos de alça, o sistema é otimizado por mecânica molecular. Esta metodologia é mais eficiente quando o número de aminoácidos das alças é pequeno, geralmente menor do que oito[23].

Pode-se reduzir, em alguns casos, as limitações inerentes aos métodos comparativos pela aplicação de métodos híbridos, que combinam a abordagem por bancos de dados e os métodos *ab initio*[81-84]. As alças modeladas são selecionadas de acordo com critérios como a exposição de grupos hidrofóbicos na superfície e a necessidade de baixas energias conformacionais relativas.

Modelagem das cadeias laterais

Como no caso das alças, as conformações das cadeias laterais também são modeladas a partir de dados estéricos e energéticos de estruturas semelhantes obtidos experimentalmente[85,86]. A modelagem é feita com o auxílio de bancos de dados de rotâmeros permitidos de cadeias laterais (χ_n), derivados de estruturas terciárias do PDB de alta resolução[23]. A Figura 3 define esquematicamente os ângulos torcionais de proteínas envolvidos na análise.

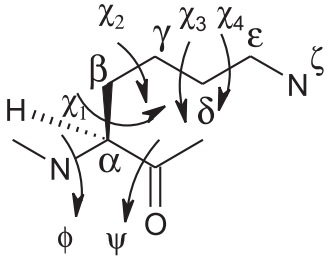

Figura 3. *Definição dos ângulos torcionais da cadeia principal (ϕ e ψ) e das cadeias laterais (χ_n)*

A utilização de "bibliotecas" de cadeias laterais (ou "coleções", termo menos utilizado na área) como, por exemplo, a desenvolvida por Dunbrack Jr. e colaboradores[87], é um dos métodos mais eficientes de predição da estrutura das cadeias laterais de proteínas. Elas permitem estimar a probabilidade de cada cadeia lateral assumir uma determinada conformação (χ_n) em função dos ângulos conformacionais da cadeia principal (ψ e ϕ) e do tipo de resíduo. A precisão aproximada do método é de 82% para o ângulo diedro χ_1 e de

72% para os diedros χ_1 e χ_2, tomados simultaneamente. Esta precisão é alcançada para pares de estruturas que compartilham de 30 a 90% de identidade.

Quando as cadeias laterais estão localizadas na superfície da proteína, deve-se incluir termos de solvatação no processo de modelagem[88,89]. Além disso, a inclusão de termos que descrevem as ligações hidrogênio melhora significativamente a predição da estrutura das cadeias[90].

Pontes de dissulfeto são um problema especial e as cadeias laterais deste tipo são modeladas com o auxílio de informações estruturais derivadas de pontes de dissulfeto equivalentes em estruturas semelhantes[91,92].

Otimização do modelo gerado

Após a construção do modelo da proteína-problema é necessário otimizá-lo. As interações desfavoráveis entre átomos não-ligados ("nonbonded contacts"), bem como as energias de ângulos torcionais e de ligação são otimizadas por mecânica molecular[65]. Como regra geral, esta otimização não pode ser extensiva para não desviar o sistema de seu molde original[23]. Conseqüentemente, as etapas de otimização e validação do modelo devem ser efetuadas simultaneamente.

Validação do modelo

A validação é uma etapa essencial que pode ser executada em diferentes níveis de organização estrutural. Deve-se avaliar a qualidade do empacotamento global da proteína, os possíveis erros estruturais em regiões localizadas e os parâmetros estereoquímicos[25]. A Tabela 1 lista alguns métodos de validação.

Como a qualidade do modelo depende muito da estrutura da proteína-molde utilizada, é preciso que esta última tenha sido obtida em boa resolução e com um fator-R satisfatório. É preferível usar como modelos estruturas cuja resolução seja igual ou superior a 2 Å e fator-R inferior a 20%[19].

Um primeiro requisito a que uma proteína adequadamente modelada deve atender é ter uma estrutura terciária satisfatória[93]. Sua qualidade depende da proteína escolhida como molde e do alinhamento calculado. É importante verificar se existem grandes diferenças conformacionais não explicadas entre os elementos de estrutura secundária (regiões conservadas) das estruturas-molde e da estrutura–modelada[19].

A qualidade estereoquímica do modelo é de importância fundamental. O programa mais utilizado na avaliação dos parâmetros estereoquímicos, o PROCHECK[49], avalia os comprimentos de ligação, os ângulos planos, a planaridade dos anéis de cadeias laterais, a quiralidade, as conformações das cadeias laterais, a planaridade das ligações peptídicas, os ângulos torcionais da cadeia principal e das cadeias laterais, o impedimento estérico entre pares de átomos não-ligados e a qualidade do gráfico de Ramachandran[94]. O gráfico de Ramachandran é particularmente útil porque ele define os resíduos que se encontram nas regiões energicamente mais favoráveis e desfavoráveis e orienta a avaliação da qualidade de modelos teóricos ou experimentais de proteínas.

É necessário avaliar também as interações entre a estrutura modelada e o meio, essencialmente água. Neste tipo de análise, pode-se usar o programa WHATCHECK[48], que dá informações sobre a formação de regiões centrais hidrofóbicas, a acessibilidade de resíduos e átomos a moléculas de solvente (água), a distribuição espacial de grupos iônicos, a distribuição das distâncias atômicas e das ligações hidrogênio da cadeia principal. No mesmo contexto, o programa PROSAII[95] avalia o ambiente de cada aminoácido da proteína modelada (validação localizada), tendo como referência o ambiente esperado em proteínas análogas de alta resolução.

Outras metodologias baseadas em mecânica molecular, cálculos de energia livre de solvatação ou métodos estatísticos têm sido testadas[96]. Estas metodologias são, em princípio, capazes de estimar a qualidade da estrutura terciária dos modelos de proteínas.

LIMITAÇÕES DO MÉTODO E POSSÍVEIS ERROS

Em muitos casos, a probabilidade de ocorrência de erros em um modelo de proteína construído pelo método da homologia é inversamente proporcional ao grau de identidade entre as seqüências de resíduos utilizadas. No entanto, Chothia e Lesk demostraram que relações mais complexas entre a qualidade de modelos gerados e o grau de identidade com o molde são possíveis[31]. Em conseqüência, modelar uma proteína com baixo grau de identidade com o molde utilizado é geralmente muito mais difícil.

Em geral, os erros não se distribuem igualmente por toda a estrutura[21]. Como um primeiro critério, os rmsd das distâncias entre os átomos da cadeia principal de estruturas conservadas de pares de proteínas dependem do grau de identidade das seqüências utilizadas[31]. Pares de proteínas com grau de identidade superior a 50%, aproximadamente, apresentam, em geral, rmsd inferior a cerca de 1,0 Å para a proteína como um todo.

As regiões cuja estrutura é conservada são mais confiáveis que as regiões de alças. Quando o grau de identidade é superior a cerca de 30%, os resíduos formados por hélices-α e fitas-β têm, em geral, bom alinhamento com os resíduos correspondentes do molde. Abaixo deste grau de identidade, as fitas aparentemente se alinham melhor que as hélices[21].

As etapas mais problemáticas da modelagem de proteínas por homologia são o alinhamento das seqüências e a modelagem das regiões de alças. Muitos avanços têm sido feitos nos métodos de alinhamento de seqüências[96]. Já a modelagem de alças ainda é um fator limitante, principalmente quando o número de resíduos é grande[21].

Em alguns casos, cerca de 27% de identidade entre as seqüências é suficiente para um bom alinhamento e, conseqüentemente, para um bom modelo. Há casos, porém, em que apesar do alto grau de identidade e do bom alinhamento não se obtém modelos satisfatórios[21]. Cada modelagem é um problema particular e deve-se ter sempre em mente que a construção de um modelo de proteína é uma tarefa que deve ser fundamentada em princípios de evolução molecular para cada grupo específico de proteínas homólogas. Quando as proteínas têm vários domínios, a construção do modelo é um processo ainda mais complexo.

Em resumo, os erros normalmente encontrados em modelos de proteínas gerados por homologia podem ser divididos em cinco categorias[64,97]:

1. *Erros de modelagem das cadeias laterais.* Quando o grau de identidade entre a estrutura-problema e a estrutura-molde é baixo podem surgir problemas de encaixe de algumas cadeias laterais no corpo principal do modelo, formando regiões interrompidas.
2. *Deslizamento em regiões do alinhamento.* Divergências entre seqüências podem provocar distorções durante a localização dos resíduos da cadeia principal, mesmo que o enovelamento global do modelo seja satisfatório. A utilização de várias proteínas-molde pode reduzir este tipo de erro.
3. *Erros em regiões sem molde*. Este tipo de erro pode ocorrer em segmentos da estrutura-problema que não têm equivalentes na estrutura-molde, principalmente em regiões de alças grandes. A avaliação crítica do alinhamento proposto e do ambiente em que a alça está localizada pode levar à superação deste tipo de erro.

4. Erros de alinhamento. Ocorrem principalmente quando o grau de identidade entre as seqüências envolvidas é inferior a 30%. Para reduzi-los, deve-se dar preferência a alinhamentos múltiplos e, se possível, modificar iterativamente as regiões do alinhamento tidas como problemáticas.

5. Moldes incorretos. Este tipo de erro é mais provável quando o grau de identidade é inferior a 25%. Nestes casos, deve-se buscar parâmetros experimentais e modificar com cautela o alinhamento.

COMENTÁRIOS FINAIS

A modelagem por homologia pode ser útil quando as estruturas de proteínas baseadas em cristalografia de raios-X ou RMN não estão disponíveis. A modelagem de proteínas por homologia tem sido importante nas áreas de biologia estrutural, bioquímica e biofísica, particularmente em estudos relacionados com os genomas. Aqui, seu potencial é imenso, pois a técnica é capaz de acelerar o processo de elucidação de estruturas protéicas em curto espaço de tempo e a custos reduzidos.

Apesar das limitações inerentes ao método, a modelagem de proteínas por homologia é uma ferramenta adequada para a predição teórica da estrutura de proteínas. A técnica tem sido aplicada com sucesso na indústria farmacêutica e na pesquisa fundamental, sendo uma poderosa alternativa para a superação dos problemas envolvidos na elucidação de estruturas protéicas por técnicas experimentais.

Para uma revisão do progresso na área de modelagem de proteínas, bem como a validação e limitações das técnicas atualmente utilizadas, recomenda-se ao leitor uma consulta ao "Critical Assessement of Techniques for Protein Structure Prediction" (CASP)[99]. Trata-se de uma conferência bianual, onde se avaliam os métodos teóricos disponíveis para a predição de estrutura de proteínas.

Durante a conferência, especialistas em cristalografia e em RMN de proteínas submetem aos participantes seqüências de proteínas cujas estruturas terciárias foram recentemente elucidadas e ainda não divulgadas. Vários grupos, das áreas de modelagem de proteínas por homologia, por métodos *ab initio* e enovelamento de proteínas ("protein folding"), tentam elucidar as estruturas obtidas experimentalmente. Deste modo, os métodos teóricos para predição de proteínas podem ser validados e uma visão geral do estado-da-arte é obtida.

O CASP5 aconteceu entre os dias 1 e 5 de dezembro de 2002 no Asilomar Conference Center em Pacific Grove, próximo a Monterey, Califórnia, EUA[99].

AGRADECIMENTOS

Os autores agradecem o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da Fundação de Amparo à Ciência do Estado do Rio de Janeiro (FAPERJ; processo E-26/151898/2000), sem o qual seria impossível a realização deste projeto. O. A. Santos Filho agradece ao CNPq a bolsa pós-doutoral (processo 150089/00-7 RN).

REFERÊNCIAS

1. Venter, J. C.; Adams, M. D.; Myers, E. W.; *et al.*; *Science* **2001**, *291*, 1304.
2. Lander, E. S.; Linton, L. M.; Birren, B.; *et al.*; *Nature* **2001**, *409*, 860.
3. Maggio, E. T.; Ramnarayan, K.; *Trends Biotechnol.* **2001**, *19*, 266.
4. Norwell, J. C.; Machalek, A. Z.; *Nat. Struct. Biol.* **2000**, *7*, 931.
5. Terwilliger, T.C.; *Nat. Struct. Biol.* **2000**, *7*, 935.
6. Yokoyama, S.; Hirota, H.; Kigawa, T.; Yabuki, T.; Shirouzu, M.; Terada, T.; Ito, Y.; Matsuo, Y.; Kuroda, Y.; Nishimura, Y.; Kiogoku, Y.; Miki, K.; Hasui, R.; Kuramitsu, S.; *Nat. Struct. Biol.* **2000**, *7*, 943.
7. Heinemann, U.; *Nat. Struct. Biol.* **2000**, *7*, 940.
8. Gasteiger, E.; Jung, E.; Bairoch, A.; *Curr. Issues Mol. Biol*. **2001**, *3*, 47.
9. Bairoch, A.; Apweiler, R.; *Nucleic Acids Res*. **2000**, *28*, 45.
10. Bairoch, A.; Apweiler, R.; *J. Mol. Med*. **1997**, *75*, 312.
11. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E.; *Nucleic Acids Res.* **2000**, *28*, 235.
12. Ortiz, A. R.; Kolinski, A.; Skolnick, J.; *J. Mol. Biol.* **1998**, *277*, 419.
13. Lee, J.; Scheraga, H. A.; Rackovsky, S.; *J. Comput. Chem.* **1997**, *18*, 1222.
14. Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Oldziej, S.; Scheraga, H. A.; *J. Comput. Chem.* **1997**, *18*, 874.
15. Osguthorpe, D. J.; *Proteins Suppl.* **1997**, *1*, 172.
16. Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D.; *J. Mol. Biol.* **1997**, *268*, 209.
17. Simons, K. T.; Ruczinski, I.; Kooperberg, C.; Fox, B. A.; Bystroff, C.; Baker, D.; *Proteins* **1999**, *34*, 82.
18. Sternberg, M. J.; Bates, P. A.; Kelly, L. A.; MacCallum, R. M.; *Curr. Opin. Struct. Biol.* **1999**, *9*, 368.
19. Santos Filho, O. A.; *Tese de Doutorado*, Instituto Militar de Engenharia, Rio de Janeiro, Brasil, 2000.
20. Santos Filho, O. A.; Bicca de Alencastro, R.; Figueroa Villar, J. D.; *Biophys. Chem.* **2001**, *91*, 305.
21. D'Alfonso, G.; Tramontano, A.; Lahm, A.; *J. Struct. Biol.* **2001**, *134*, 246.
22. Sánchez, R.; Šali, A.; *J. Comput. Phys.* **1999**, *151*, 388.
23. Peitsch, M. C. Em *Practical Application of Computer-Aided Drug Design*; Charifson, P. S., ed.; Marcel Dekker: New York, 1997.
24. Rost, B.; Sander, C.; *Annu. Rev. Biophys. Biomol. Struct.* **1996**, *25*, 113.
25. Johnson, M. S.; Srinivasan, N.; Sowdhamini, R.; Blundell, T. L.; *Crit. Rev. Biochem. Mol. Biol.* **1994**, *29*, 1.
26. Thornton, J. M.; Swindells, M. B. Em *Molecular Structures in Biology*; Diamond, R.; Koetzle, T. F.; Prout, K.; Richardson, J. S., eds.; Oxford University Press: Oxford, 1993.
27. Branden, C.; Tooze, J.; *Introduction to Protein Structure*, Gerland: New York, 1991.
28. Wolf; Y. I.; Grishin, N. V.; Koonin, E. V.; *J. Mol. Biol.* **2000**, *299*, 897.
29. Höltje, H.-D.; Folkers, G. Em *Molecular Modeling: Basic Principles and Applications*; Mannhold, R.; Kubinyi, H.; Timmerman, H., eds.; VCH: Weinheim, 1997.
30. Benner, S. A.; Cannarozzi. G.; Gerloff, D.; Turcotte, M.; Chelvanayagam, G.; *Chem. Rev.* **1997**, *97*, 2725.
31. Chothia, C.; Lesk, A. M.; *EMBO J.* **1986**, *5*, 823.
32. Kreil, D. P.; Etzold, T.; *Trends Biochem. Sci.* **1999**, *24*, 155.
33. Orengo, C. A.; Pearlm, F. M. G.; Bray, J. E.; Todd, A. E.; Martin, A. C.; Conte, L.; Thornton, J. M.; *Nucleic Acids Res.* **1999**, *27*, 275.
34. Benson, D. A.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Raoo, B. A.; Wheeler, D. L.; *Nucleic Acids Res.* **2000**, *28*, 15.
35. Sánchez, R.; Šali, A.; *Bioinformatics* **1999**, *15*, 1060.
36. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J.; *J. Mol. Biol.* **1990**, *215*, 403.
37. Pearson, W. R.; *Methods Enzymol.* **1990**, *183*, 63.
38. Fischer, D.; Eisenberg, D.; *Protein Sci.* **1996**, *5*, 947.
39. Flockner, H.; Braxenthaler, M.; Lackner, P.; Jaritz, M.; Ortner, M.; Sippl, M. J.; *Proteins* **1995**, *23*, 376.
40. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Miller, W.; Lipman, D. J.; *Nucleic Acids Res.* **1997**, *25*, 3389.
41. Jeanmougin, F.; Thompson, J. D.; Gouy, M.; Gibson, D. G.; Higgins, T. J.; *Trends Biochem. Sci.* **1998**, *23*, 403.
42. Corpet, F.; *Nucleic Acids Res.* **1988**, *16*, 10881.
43. Šali, A.; Blundell, T. L.; *J. Mol. Biol.* **1993**, *234*, 779.
44. Peitsch, M. C.; Jongeneel, C. V.; *Int. Immunol.* **1993**, *5*, 233.
45. Peitsch, M. C.; *Bio-technol.* **1995**, *13*, 658.
46. Peitsch, M. C.; *Biochem. Soc. Trans.* **1996**, *24*, 274.
47. Guex, N.; Peitsch, M. C.; *Electrophoresis* **1997**, *18*, 2714.
48. Hooft, R. W. W.; Vriend, G.; Sander, C.; Abola, E. E.; *Nature* **1996**, *381*, 272.
49. Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M.; *J. Appl. Crystallogr.* **1993**, *26*, 283.
50. Pontius, J.; Richelle, J.; Wodak, S. J.; *J. Mol. Biol.* **1996**, *264*, 121.
51. Barker, W. C.; Garavelli, J. S.; Huang, H.; McGarvey, P. B.; Orcutt, B. C.; Srinivasarao, G. Y.; Xiao, C.; Yeh, L. -S. L.; Ledley, R. S.; Janda, J. F.; Pfeiffer, F.; Mewes, H. -W.; Tsugita, A.; Wu, C.; *Nucleic Acids Res.* **2000**, *28*, 41.
52. Sander, C.; Schneider, R.; *Proteins* **1991**, *9*, 56.
53. Bowie, J. U.; Lüthy, R.; Eisenberg, D.; *Science* **1991**, *253*, 164.
54. Godzik, A.; Kolinski, A.; Skolnick, J.; *J. Mol. Biol.* **1992**, *227*, 227.
55. Jones, D. T.; Taylor, W. R.; Thornton, J. M.; *Nature* **1992**, *358*, 86.
56. Davies, J. F.; Delcamp, T. J.; Prendergast, N. J.; Ashford, V. A.; Freisheim, J. H.; Kraut, J.; *Biochemistry* **1990**, *29*, 9467.

57. Sawaya, M. R.; Kraut, J.; *Biochemistry* **1997**, *36*, 586.
58. McTigue, M. A.; Davies, J. F.; Kaufman, B. T.; Kraut, J.; *Biochemistry* **1992**, *31*, 7264.
59. Whitlow, M.; Howard, A. J.; Stewart, D.; Hardman, K. D.; Kuyper, L. F.; Baccanari, D. P.; Fling, M. E.; Tansik, R. L.; *J. Biol. Chem.* **1997**, *272*, 30289.
60. Needleman, S. B.; Wunsch, C. D.; *J. Mol. Biol.* **1970**, *48*, 443.
61. Smith, T. F.; Waterman, M. S.; *J. Mol. Biol.* **1981**, *147*, 195.
62. Rychlewski, L.; Zhang, B.; Godzik, A.; *Fold. Des.* **1998**, *3*, 229.
63. Sánchez, R.; Šali, A.; *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 13597.
64. Sánchez, R.; Šali, A.; *Proteins (Suppl.)* **1997**, *1*, 50.
65. Leach, A. R.; *Molecular Modelling: Principles and Applications*, Longman: Essex, 1996.
66. Blundell, T. L.; Sibanda, B . L.; Sternberg, M. J. E.; Thornton, J. M.; *Nature* **1987**, *326*, 347.
67. Browne, W. J.; North, A. C. T.; Phillips, D. C.; Brew, K.; Hill, T. C.; *J. Mol. Biol.* **1969**, *42*, 65.
68. Greer, J.; *Proteins* **1990**, *7*, 317.
69. Claessens, M.; Custem, E. V.; Lasters, I.; Wodak, S.; *Protein Eng.* **1989**, *4*, 335.
70. Levitt, M.; *J. Mol. Biol.* **1992**, *226*, 507.
71. Aszódi, A.; Taylor, W. R.; *Fold. Des.* **1996**, *1*, 325.
72. Havel, T. F.; Snow, M. E.; *J. Mol. Biol.* **1991**, *217*, 1
73. Srinivasan, S.; March, C. J.; Sudarsanam, S.; *Protein Sci.* **1993**, *2*, 227.
74. Fiser, A.; Do, R. K. G.; Šali, A.; *Protein Sci.* **2000**, *9*, 1753.
75. Fine, R. M.; Wang, H.; Shenkin, P. S.; Yarmush, D. L.; Levinthal, C.; *Proteins* **1986**, *1*, 342.
76. Moult, J.; James, M. N. G.; *Proteins* **1986**, *1*, 146.
77. Bruccoleri, R. E.; Karplus, M.; *Biopolymers* **1987**, *26*, 137.
78. Greer, J.; *Proc. Natl. Acad. Sci. U.S.A.* **1980**, *77*, 3393.
79. Jones, T. H.; Thirup, S.; *EMBO J.* **1986**, *5*, 819.
80. Chothia, C.; Lesk, A. M.; *J. Mol. Biol.* **1987**, *196*, 901.
81. Chothia, C.; Lesk, A. M.; Levitt, M.; Amit, A. G.; Mariuzza, R. A.; Philips, S. E. V.; Poljak, R. J.; *Science* **1986**, *233*, 755.
82. Martin, A. C. R.; Cheetham, J. C.; Rees, A. R.; *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 9268.
83. Mas, M. T.; Smith, K. C.; Yarmush, D. L.; Aisaka, K.; Fine, R. M.; *Proteins* **1992**, *14*, 483.
84. van Vlijmen, H. W. T.; Karplus, M.; *J. Mol. Biol.* **1997**, *267*, 975.
85. Vásquez, M.; *Curr. Opin. Struct. Biol.* **1996**, *6*, 217.
86. Sánchez, R.; Šali, A.; *Curr. Opin. Struct. Biol.* **1997**, *7*, 206.
87. Bower, M. J.; Cohen, F. E.; Dunbrack Jr., R. L.; *J. Mol. Biol.* **1997**, *267*, 1268.
88. Cregut, D.; Liautard, J. P.; Chiche, L.; *Protein Eng.* **1994**, *7*, 1333.
89. Wilson, C.; Gregoret, L. M.; Agard, D. A.; *J. Mol. Biol.* **1993**, *229*, 996.
90. Dunbrack, R. L.; Karplus, M.; *J. Mol. Biol.* **1993**, *230*, 543.
91. Jung, S. H.; Pastan, I.; Lee, B.; *Proteins* **1994**, *19*, 35.
92. Šali, A.; Overington, J. P.; *Protein Sci.* **1994**, *3*, 1582.
93. Laskowski, R. A.; MacArthur, M. W.; Thornton, J. M.; *Curr. Opin. Struct. Biol.* **1998**, *5*, 631.
94. Ramachandran, G. N.; Sasisekharan, V.; *Adv. Prot. Chem.* **1968**, *23*, 283.
95. Sippl, M. J.; *Proteins* **1993**, *17*, 355.
96. Sauder, J. M.; Arthur, J. W.; Dunbrack, R. L.; *Proteins* **2000**, *40*, 6.
97. Šali, A.; Potterton, L.; Yuan, F.; Vlijmen, H.; Karplus, M.; *Proteins* **1995**, *23*, 318.
98. Xu, L. Z.; Sánchez, R.; Šali, A.; Heintz, N.; *J. Mol. Biol.* **1996**, *271*, 24711.
99. http:// predictioncenter.llnl.gov/casp5/Casp5.html, acessada em Maio 2002.