

Anotação Funcional Computacional de Proteínas

Fotos e ilustrações cedidas pelos autores

Novos métodos computacionais poderão preencher lacunas do sistema de anotação atual

Introdução

A característica mais importante de uma proteína é sua função. Pode até mesmo se dizer que a existência de uma proteína depende da sua função; enquanto que o DNA não-codificante de um organismo pode incluir cópias não-transcritas de genes. Sendo assim, o custo energético de sintetizar uma proteína assegura que somente proteínas com funções necessárias para um organismo sejam produzidas. A função de uma proteína pode ser descrita em vários níveis de detalhes, do fisiológico – proteína X está envolvida no processo de replicação de células –, até o químico – proteína X catalisa a hidrólise de um certo substrato. Para se determinar experimentalmente a função molecular de uma proteína, é necessário purificá-la (às vezes com a ajuda de técnicas modernas de biologia molecular), e, em seguida, testar sua atividade biológica. Os resultados podem ou não fornecer dados sobre as funções *in vivo* da proteína. Alternativamente, pode-se utilizar novas metodologias, como microarranjo (*microarrays*) ou análise proteômica, quando o objetivo é focalizar diretamente nos níveis de expressão de determinadas proteínas, ou na expressão dos genes que as codificam, sob diferentes condições ambientais, ou em diferentes etapas do desenvolvimento. Esses métodos fornecem indicações da função *in vivo* da proteína, mas, ao contrário dos ensaios, dizem pouco sobre a função em termos químicos e bioquímicos. Todas essas técnicas exigem um investimento significativo em equipamento e tempo, tanto que não podemos pensar em estudar diretamente mais do que uma minúscula fração de proteínas de interesse.

Ao contrário, seqüências biológicas são atualmente obtidas a um custo relativamente baixo. Isso reflete no crescimento exponencial do tamanho dos bancos de dados de seqüências. Porém, essa vasta quantidade de dados é de pouco valor científico ou

aplicado, sem a sua adequada anotação funcional. Como experimentos laboratoriais dificilmente vão ser capazes de tratar essa grande quantidade de dados, o caminho alternativo é através da análise computacional. Embora já existam sistemas computacionais capazes de anotar, até certo ponto, todas as novas seqüências que vêm sendo determinadas, estes ainda apresentam graves falhas. Além de produzir uma anotação significativamente incompleta, erros estão sendo introduzidos na anotação de algumas seqüências que, pela natureza do sistema, podem rapidamente ser propagados a outras seqüências a serem analisadas.

Essa revisão é dividida em três partes. Na primeira, descreve-se brevemente o modo atual de anotação funcional computacional, destacando suas falhas. Na segunda parte, são discutidas as novas possibilidades para a anotação funcional computacional, cujo desenvolvimento foi estimulado pelos projetos genoma. E finalmente, as novas idéias que buscam informações sobre função através de análises de estruturas são avaliadas. Um resumo do fluxo de dados durante o processo de anotação funcional está ilustrado na Figura 1.

O sistema atual de anotação funcional computacional

Atualmente, novas seqüências biológicas são anotadas funcionalmente simplesmente através da comparação com seqüências existentes, que são armazenadas em bancos de dados como, por exemplo, o GenBank (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein>). BLAST (Altschul *et al*, 1990) é o programa padrão para essa comparação, devido à sua extrema eficiência. Esse programa possibilita a comparação das milhares de novas seqüências geradas diariamente, com as depositadas em bancos de dados, que vêm crescendo expo-

Daniel John Rigden

Pesquisador na Área de Bioinformática
Embrapa Recursos Genéticos e
Biotecnologia – Cenargen/Embrapa
Brasília, DF
daniel@cenargen.embrapa.br

Luciane Vieira de Mello

Pesquisadora na Área de Bioinformática
Embrapa Recursos Genéticos e
Biotecnologia – Cenargen/Embrapa
Brasília, DF
mello@cenargen.embrapa.br

nencialmente. Assim, a nova seqüência é comparada com outra já existente e bem caracterizada, e que apresentou o maior grau de similaridade com a nova seqüência, sendo sua função transferida para esta. Dependendo do grau de similaridade, a anotação pode ser modificada de 'Proteína X' para 'Proteína X-provável' ou 'Semelhante à proteína X', refletindo assim uma incerteza na transferência de função, em casos onde a similaridade entre as duas seqüências seja considerada baixa.

A principal vantagem do sistema atual encontra-se na sua eficiência, que, mesmo em face da avalanche de seqüências novas, possibilita a anotação rápida de todas as novas seqüências. Porém, está ficando cada vez mais claro que o sistema atual tem sérias falhas. Uma falha não muito grave é a incapacidade do sistema em anotar novas seqüências que não apresentam similaridade significativa com seqüências existentes. Os resultados de projetos de genoma mostram que, em cerca de 40% dos casos, uma seqüência não mostra similaridade significativa com uma proteína já caracterizada (Gerlt and Babbitt, 2000). Nesses casos, o sistema atual é incapaz de fornecer uma anotação útil.

Uma falha mais grave, é uma série de problemas capazes de introduzir erros nas anotações funcionais dos bancos de dados. Uma vez que não existem dados experimentais sobre a grande maioria das proteínas, o sistema computacional transfere anotações de funções para novas seqüências com uma freqüência muito maior do que a transferência proveniente de dados laboratoriais. Assim, fica claro que qualquer erro que seja introduzido na anotação computacional, será rapidamente transmitido a múltiplas novas seqüências (Karp, 1998).

Uma fonte rica de erros de interpretação de seqüências encontra-se na interpretação errônea, ou superinterpretação dos resultados do BLAST (Pertsemilidis e Fondon, 2001). O BLAST mede similaridade local de duas seqüências. Entre as propriedades não medidas pelo programa estão a similaridade global, a similaridade funcional, a similaridade



Fig 1: Fluxo de dados durante o processo de anotação funcional. Linhas interrompidas indicam tradução (DNA → Proteína)

estrutural e a homologia (ancestral em comum). O mau entendimento do algoritmo do programa e, portanto, das limitações associadas aos seus resultados, pode levar usuários leigos a conclusões erradas (Pertsemilidis e Fondon, 2001). Um artigo publicado na revista Nature (Ichikawa *et al.*, 1997), e subseqüentemente retratado, é um exemplo importante (e famoso) de como erros de interpretação podem levar a conclusões errôneas do estudo. Problemas adicionais podem haver nos sistemas automatizados, nos quais a anotação é feita sem intervenção humana (Doerks *et al.*, 1998). Por exemplo, a maior similaridade local entre uma nova seqüência e seqüências existentes pode ficar fora das regiões responsáveis pela atividade da proteína. Assim, a anotação da nova proteína ficará, pelo menos, incompleta e, algumas vezes, incorreta. Também são comuns os casos nos quais a seqüência mais parecida com a nova proteína não possui uma função anotada, ou é anotada com uma função secundária da proteína. Dessa forma, a anotação mais

adequada é ignorada pelos sistemas automatizados, uma vez que o grau de similaridade da nova proteína é menor com tais proteínas. A comparação das anotações automatizadas realizadas por três diferentes grupos do genoma de *Mycoplasma genitalium* mostrou que as anotações possuíam, pelo menos, 8% de erro (Brenner, 1999).

Embora os erros de interpretação claramente contribuam para uma anotação errônea, um outro fator ainda mais problemático é a anotação por comparação, ou seja, a relação complicada entre o grau de similaridade existente entre duas seqüências, e a similaridade funcional entre elas. Resumindo, com alta identidade entre as seqüências (>80%), pode-se assumir que as suas funções sejam idênticas. Porém, na faixa de baixa identidade (<30%) é freqüentemente observado que existem diferenças nas suas funções, e pode haver proteínas claramente relacionadas evolucionariamente, mas com funções totalmente diferentes (Todd *et al.*, 2001). Uma das medidas que foi usada para analisar a relação

identidade de seqüência e similaridade de função foi o código EC. Esse número, que tem quatro campos na forma a.b.c.d, aloca às enzimas baseado nas suas atividades catalíticas – cada atividade diferente recebendo um determinado código. O primeiro dígito do código indica atividade geral, ex. hidrolase, com os dígitos seguintes referindo-se a detalhes da atividade. Assim, duas enzimas que catalisam o mesmo tipo de reação, mas que utilizam substratos diferentes, terão códigos compartilhados nos primeiros três dígitos, mas com o último dígito diferente. Foi observado que, acima de 50% de identidade de seqüência entre um par de enzimas, a variação no código EC é rara, porém presente (Figura 2a). Na faixa de 30% a 40%, a situação é diferente; só os três primeiros números podem ser previstos com uma precisão de 90%. Com menos de 30% de identidade entre duas enzimas, pares de seqüências apresentando diferenças até mesmo no primeiro dígito dos códigos EC são comuns (Figura 2a). Para a anotação computacional funcional, essas consi-

derações teriam pouca importância se, na maioria dos casos, houvesse um alto grau de identidade entre a nova proteína e a mais semelhante presente no banco de dados. Isso porque, dessa forma, poderíamos ter alta confiança na identidade de função entre as duas proteínas. No entanto, infelizmente, como mostra um estudo recente (Devos e Valencia, 2001; Figura 2b) isso está longe de ser verdade. Analisando o grau de identidade entre proteínas anotadas para três genomas e as seqüências mais parecidas disponíveis, foi observado que, num caso típico (50% dos casos), somente 25%-35% de identidade de seqüência (Figura 2b). Porém, como explicado acima, é justamente nessa faixa de identidade de seqüência que a relação entre identidade de seqüência e similaridade de função permitem a transferência confiável de função. Resumindo, na faixa de identidade de seqüência na qual uma anotação funcional é típica, uma fração significativa das anotações vai ser provavelmente realizada erroneamente. Grandes erros, por exemplo no primeiro dígito do código EC, vão ser menos comuns do que erros considerados menores, ou seja, no último dígito do código, por exemplo. Nas anotações dos três genomas analisados, foi estimado que o primeiro dígito estava errado em 2% dos casos, enquanto que, para o último dígito, mais de 30% das anotações estavam incorretas.

Assim, tendo-se conhecimento das limitações dos métodos de anotação atualmente disponíveis e utilizados, seja pela anotação equivocada, seja pela incapacidade de anotar cerca de 40% das proteínas, novos métodos computacionais para anotação funcional vêm sendo buscados. Hoje, após alguns anos de progresso notável, existem novas metodologias complementares ao sistema tradicional de comparação de seqüências. Na sua maioria, elas podem ser divididas em duas categorias. Primeiro,

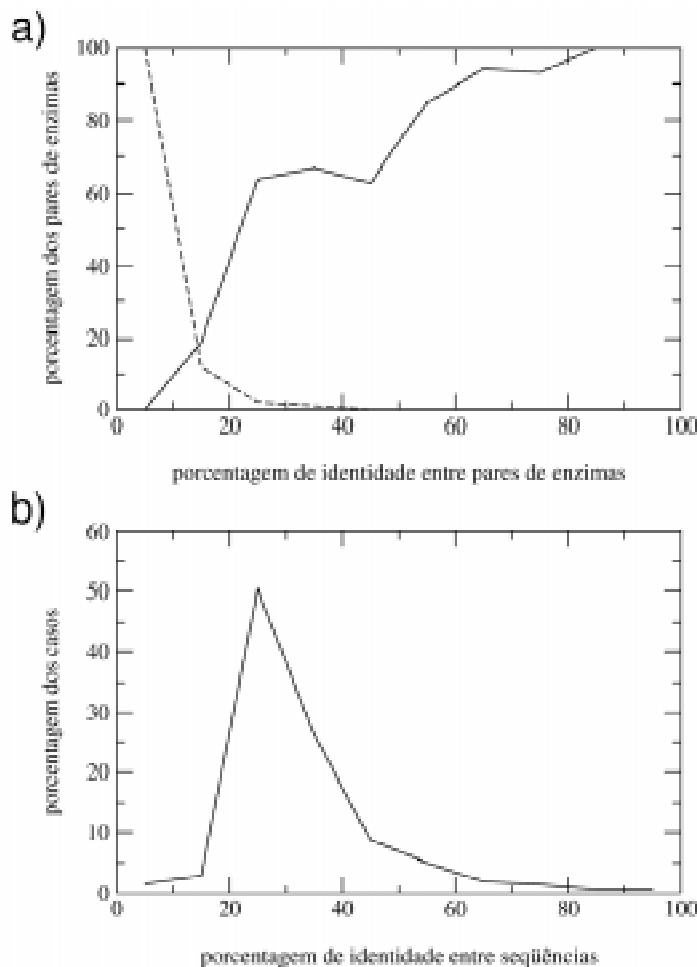


Fig 2: O problema fundamental da transferência de anotação funcional. a) Aos níveis mais baixos de identidade de seqüência, a porcentagem dos casos nos quais a função é idêntica (4 dígitos do código EC são iguais) é baixa (linhas contínuas); e os pares com funções não-relacionadas (nenhum dígito do código EC em comum) alta (linhas interrompidas). b) Tipicamente, durante o processo de anotação funcional computacional, a porcentagem de identidade entre a proteína a ser anotada e a seqüência encontrada no banco de dados é baixa - entre 20% e 40%. (Dados de Devos e Valencia, 2001)

as metodologias em decorrência dos projetos genoma (Marcotte, 2000). Isso porque esses projetos geraram informações, que são a base das novas técnicas, tais como a posição de determinados genes, ou, simplesmente, devido à grande quantidade de seqüências atualmente disponíveis. A segunda categoria contém metodologias que utilizam o aspecto estrutural (Thornton *et al.*, 2000). Esses aspectos estruturais são provenientes, tanto de modelos protéicos, como de estruturas tridimensionais determinadas experimentalmente. Ao contrário da situação atual, na qual a maioria das estruturas determinadas experimentalmente

são de proteínas de funções conhecidas, os projetos de genoma estrutural (Thornton, 2001) vão ter como resultado muitas proteínas com estruturas determinadas, porém com funções desconhecidas.

Genômica computacional

Entre os cinco diferentes métodos que podem ser agrupados sob esse título, três são estreitamente dependentes das seqüências provenientes dos projetos de genomas completos¹, assim não se aplicando às seqüências derivadas de outras fontes, como genomas expressos (funcionais) e proteomas. Esses métodos são denominados perfis filogenéticos (*filogenetic profile*), contexto genômico (*genome context*) e genoma diferencial (*subtraction of genome*).

O mais simples, porém o menos eficiente, desses métodos é o genoma diferencial (Huynen *et al.*, 1997). Esse método procura localizar genes envolvidos em aspectos fisiológicos importantes de um organismo pela comparação do seu genoma com o de um organismo parecido, mas com características diferentes. Por exemplo, pode-se comparar os genomas de duas bactérias, filogeneticamente próximas, sendo que uma possui patogenicidade e a outra não. Assim, espera-se que os genes associados com a patogenicidade estejam presentes somente no genoma da bactéria patogênica. Embora resulta-

dos interessantes venham sendo obtidos, a desvantagem do método é que os genes associados com a propriedade de interesse sempre farão parte de uma grande lista de genes, incluindo muitos que estão presentes no organismo patogênico, mas que não estão associados com a doença.

A técnica de perfil filogenético (Pelligrini *et al.*, 1999) é baseada numa proposta muito simples - que componentes de complexos macromoleculares ou enzimas de uma certa via metabólica vão ser herdados concomitantemente. Assim, os componentes isolados dos complexos ou vias, que, quando presentes

isoladamente nas células, são incapazes de exercer suas funções, não são encontrados separadamente. Na primeira etapa, um perfil de uma proteína é construído, composto de dados de presença ou ausência da proteína em vários genomas. Depois, faz-se uma busca por outras proteínas com o mesmo perfil de presença ou ausência, ou um perfil pouco diferente. Essas são indicadas como proteínas possivelmente relacionadas funcionalmente com a proteína utilizada para a construção do perfil. No trabalho original, perfis construídos para proteínas do ribossomo, do flagelo (complexos macromoleculares) e da via biosintética de histidina (via metabólica) produziram resultados que estavam de acordo com os dados experimentais, demonstrando a validade desse método (Pelligrini *et al.*, 1999). A dependência do método de perfis filogenéticos dos genomas completos é devida aos estudos de genes ou de proteínas expressas não fornecerem dados definitivos sobre a presença ou ausência de um particular gene no genoma relevante.

Métodos de contexto genômico usam a existência de agrupamento (*clusters*) de genes nos genomas de procaríotos (Overbeek *et al.*, 1999a). Embora as razões e os mecanismos responsáveis pela manutenção desses agrupamentos sejam desconhecidos, sua característica mais marcante é a composição de genes funcionalmente relacionados. Assim, podemos inferir uma relação funcional entre os genes presentes em novos agrupamentos descobertos. Dois aspectos distintos, mas complementares, dos agrupamentos, têm poder para preverem a relação de função – a conservação de uma distância pequena entre um par de genes (Overbeek *et al.*, 1999b) e a conservação da ordem dos genes no DNA (Overbeek *et al.*, 1999a). Assim, podemos comparar genomas (e não seqüências individuais, como é tradicionalmente feito) buscando agrupamentos de genes em genomas filogeneticamente distantes, e inferir uma relação funcional entre os genes componentes. Observa-se que proteínas que se interagem fisicamente apresentam uma tendência particular de serem codificadas por genes de ordem conservada. Dessa forma, há uma dependência entre os métodos de contexto genômico pelas seqüências oriundas dos projetos de genoma completo. Isso ocorre, uma vez que projetos de genoma expresso e

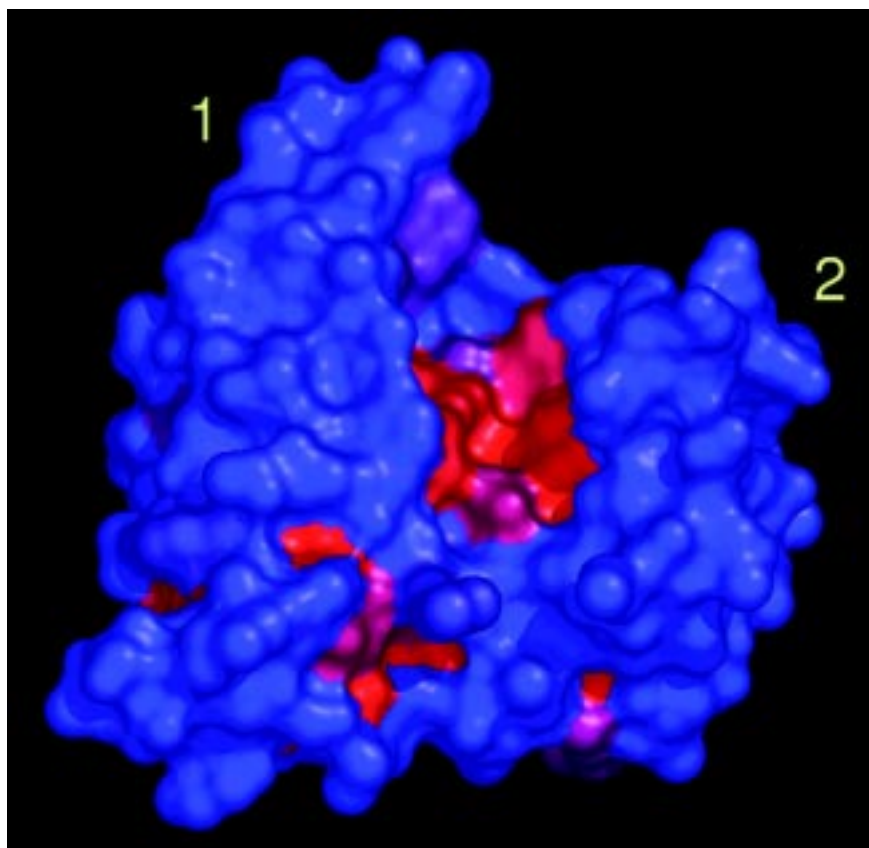


Fig 3: A grande cavidade entre os domínios 1 e 2 da estrutura experimental de uma proteína de proteção de plantas contém uma região com vários resíduos conservados (vermelho). Esse padrão está presente em toda a sua família protéica. Assim, é altamente indicada como um sítio de ligação

proteoma não fornecem informações sobre posicionamento dos genes no DNA do organismo.

Existem outros métodos, recentemente desenvolvidos, que podem ser aplicados a qualquer seqüência, independente da sua origem. Assim, são igualmente aplicáveis aos resultados de projetos de genoma completo, genoma expresso e proteoma, bem como às seqüências determinadas individualmente por experimentos tradicionais. Porém, vale notar que foi a quantidade de dados de seqüência provenientes, principalmente, dos projetos genoma que incentivaram o desenvolvimento dessas novas técnicas. A primeira dessas técnicas baseia-se nas consequências de eventos de fusão de genes (Marcotte *et al.*, 1999). Foi observado que proteínas presentes separadamente num genoma estão, às vezes, presentes como uma única proteína, do tamanho igual à soma dos dois componentes, em outros genomas. Essa observação necessariamente implica uma relação funcional entre os dois compo-

nentes, pois seria uma desvantagem para o organismo a expressão de duas proteínas não relacionadas funcionalmente, em conjunção. A observação de um caso dessa natureza é uma forte indicação de que as proteínas, quando presentes individualmente num organismo, podem interagir. Faz-se essa inferência porque o motivo mais forte que levaria a fusão de duas proteínas seria a proximidade das duas numa via metabólica. Assim, depois da fusão, a transferência do substrato de um componente ao outro seria facilitada. Porém, a fusão pode também ser tolerada, ou até favorecida, em termos evolucionários, em caso de duas proteínas com funções relacionadas. Outras análises adicionais mostraram-se capazes de apontar casos de interação entre dois componentes protéicos, quando existentes separadamente em um determinado organismo (Marcotte *et al.*, 1999).

Enquanto as análises de contexto genômico e fusão de genes, principalmente orientadas para a identificação de

¹ O termo genoma completo se refere-se aos projetos genoma que seqüenciam todo o conteúdo genético (DNA) de um organismo. O termo genoma estrutural foi utilizado como no Inglês, *structural genome*, que se refere à estrutura protéica.

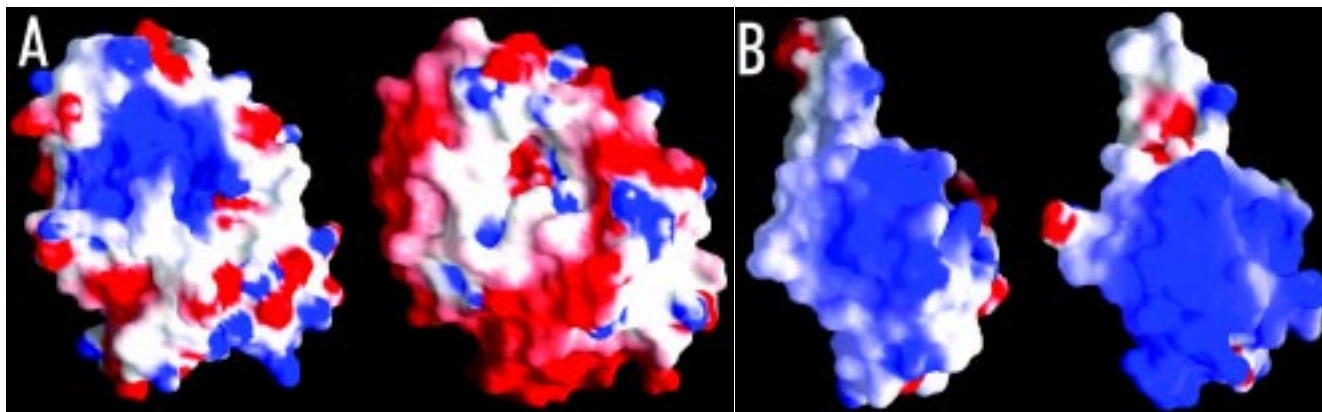


Fig 4: Similaridade em características eletrostáticas é correlacionada com similaridade de função. **a)** diferenças em características eletrostáticas entre *Phosphoglycerate mutase* (esquerda) e YhfR (direita) indicam funções diferentes, mesmo que as duas proteínas exibam cerca de 30% de identidade de seqüência (Rigden *et al.*, 2001). **b)** mesmo exibindo somente 16% de identidade de seqüência, similaridades eletrostáticas sugerem que as proteína *roLA* (direita) e papillomavirus (esquerda), são capazes de se ligarem ao DNA (Rigden e Carneiro, 1999)

proteínas funcionalmente relacionadas podem indicar pares de proteínas que interagem entre si, o último novo método dessa categoria – similaridade de árvores filogenéticas (Pazos e Valencia, 2001) – funciona no sentido contrário. Ou seja, este busca por pares de proteínas que interagem e que, portanto, têm funções relacionadas. Esse método fundamenta-se na observação de que a evolução coordenada de proteínas que se interagem leva as suas árvores filogenéticas a serem mais parecidas do que seria esperado. Assim, analisando a correlação entre árvores construídas por duas proteínas (ou mais precisamente, a correlação entre suas distâncias evolucionárias), quando achamos uma correlação significativa, há indicação de interação entre as proteínas. Dados experimentais já comprovaram a lógica utilizada nesse método, onde interações de proteínas já conhecidas foram destacadas pelos altos coeficientes de correlação entre suas árvores.

A questão que ainda existe é o quanto esse conjunto de novos métodos pode ajudar a preencher as lacunas no sistema atual de anotação funcional computacional. Uma resposta parcial encontra-se na avaliação quantitativa das técnicas descritas acima aplicadas ao genoma de *Mycoplasma genitalium* (Huynen *et al.*, 2000). Observou-se que a conservação de ordem de genes é a mais poderosa técnica, uma vez que pôde ser aplicada a 37% dos genes, seguida pela análise de perfil filogenético (11% dos genes), aparência de genes em agrupamentos sem ordem conservada (8%), e, finalmente, pela técnica de fusão de genes (6%). No total, foram obtidas informações sobre 50% do complemento

genético de *M. genitalium* através desses métodos. Essa figura é uma subestimativa da sua utilidade, uma vez que nem a técnica de genoma diferencial (não aplicável a somente um genoma), nem a de similaridade de árvores filogenéticas (recentemente publicada) foram aplicadas. Também é importante lembrar que o crescimento do uso dessas técnicas depende do crescimento dos bancos de dados de seqüências e, em particular, da disponibilidade de um número ainda maior de genomas completos. Em alguns casos, pode-se esperar que o poder da técnica cresça de acordo com o quadrado do número de genomas completos disponíveis. Para finalizar, é importante lembrar que essas técnicas, às vezes, podem produzir resultados vagos como, por exemplo, “proteínas A e B têm funções relacionadas”. No entanto, por apresentarem grande eficiência, está ficando claro que a combinação delas com os métodos tradicionais de buscas por homólogas nos bancos de dados levarão a um conhecimento bem mais profundo das novas seqüências.

Bioinformática estrutural

Embora métodos tradicionais de anotação funcional trabalhem somente com as seqüências protéicas, sabe-se que é a estrutura tridimensional de uma proteína, não simplesmente a sua seqüência, que determina a sua atividade. Quando a proteína se dobra, os resíduos importantes são orientados em suas corretas posições para a formação das regiões funcionais – proteínas desnaturadas, em geral, não exibem atividade. Essas regiões funcionais são, na sua maioria, interfaces para a ligação da proteína a outras

moléculas. Os métodos tradicionais funcionam devido às bem conhecidas relações entre seqüência, estrutura e função de proteínas. Em geral, proteínas de uma mesma família, embora não apresentando grande similaridade de seqüência, conservam a mesma estrutura tridimensional; estrutura esta que é mais conservada do que seqüência. Sabe-se também que mais importante do que a porcentagem total de identidade entre duas seqüências, é a identidade de resíduos chaves, responsáveis pela sua função. Assim, assumindo que a estrutura conservou a orientação tri-dimensional relativa desses resíduos, as proteínas possuirão a mesma função. Com essas relações estabelecidas, justifica-se, até um certo ponto (veja acima), a suposição da conservação de função quando se observa conservação de seqüência.

Mas o que acontece quando os resíduos importantes não são conservados, mesmo com grande conservação da seqüência em geral? Ou se outras mudanças na seqüência afetarem a região funcional, bloqueando o acesso ao sítio catalítico, por exemplo. Nesses casos, e em muitos outros (Gerlt e Babbitt, 2000), a análise pura de seqüência levará a conclusões erradas sobre a função, gerando os erros que, como vimos anteriormente, podem-se perpetuar rapidamente nos bancos de dados. Pode-se evitar alguns desses problemas através de uma extrapolação da seqüência em estrutura – a modelagem protéica. A grande conservação da estrutura tridimensional, mesmo após mutações em muitos resíduos, possibilita a construção de um modelo de uma proteína, em casos em que um molde adequado encontra-se disponível. Com o modelo

construído, fica disponível uma outra bateria de análises para determinar a probabilidade da conservação de função entre duas proteínas.

A seguir, serão descritas técnicas que podem ser utilizadas na busca da determinação da função de proteínas de estrutura conhecida. Como mencionado anteriormente, essas serão principalmente geradas pelos projetos de genoma estrutural (Thornton, 2001). Existem duas categorias de ferramentas de bioinformática estrutural disponíveis para a inferência de função a partir de estrutura protéica (um modelo ou uma estrutura experimental). A primeira busca por possíveis sítios de ligação (a presença dos quais pode-se esperar em quase todas as proteínas); e a segunda procura localizar possíveis sítios de catálise (só aplicáveis às enzimas).

Uma vez que a ligação de uma determinada molécula a uma proteína acontece na sua superfície, é nessa região que a busca por possíveis sítios de ligação ocorre. Uma análise bastante simples, mas surpreendentemente eficiente, é a da geometria (Laskowski *et al.*, 1996). A necessidade freqüente de uma proteína em se ligar com alta afinidade e alta especificidade, exige a formação de múltiplas interações entre a proteína e o ligante. Em particular, nos casos de ligantes pequenos, a alta afinidade e a especificidade são adquiridas pela acomodação do ligante numa região de depressão na superfície da estrutura protéica. Seguindo essa lógica, uma análise demonstrou que sítios de ligação são, muitas vezes, encontrados na maior depressão da superfície de uma proteína. Por exemplo, em casos de enzimas monoméricas, o sítio catalítico encontrou-se presente na maior depressão da superfície em 83% dos casos. Quando a maior depressão também contém uma alta concentração de resíduos conservados (ex. Figura 3), a probabilidade de o sítio de ligação encontrar-se nessa região aumenta ainda mais.

Uma outra característica importante da superfície da proteína é o seu campo eletrostático. Algumas proteínas empregam interações eletrostáticas para atração do ligante (ex., proteínas de ligação ao DNA) ou para localização subcelular à membrana (ex., citocroma C). Essas proteínas exploram a carga inata do ligante ou da membrana e a força, a longa distância, das interações eletrostáticas. Diferenças em campo eletrostático podem ser indicativas de diferenças em função, como foi visto para a enzima *phosphoglycerate mutase* (fosfoglicerato) e uma proteína homóloga que exibe atividade catalítica muito

diferente (Rigden *et al.*, 2001; Figura 4a). Pela mesma lógica, similaridades em características eletrostáticas podem reforçar a noção de similaridade funcional entre duas proteínas. Um outro exemplo é o modelo construído da proteína *rolA*, a base de uma proteína que liga a DNA, mas que compartilha somente 16% de identidade de seqüência. Enquanto 16% de identidade de seqüência não assegura similaridade em função (Figura 2a), o modelo também exibe uma região altamente positiva (Figura 4b), em acordo com dados experimentais mostrando a ligação entre *rolA* e DNA.

Recentemente, as propriedades eletrostáticas e hidrofóbicas de superfícies de proteínas foram sujeitas a um outro modo de análise – mapas de superfícies de proteínas (Pawlowski e Godzik, 2001). Aproximando as formas das proteínas como esferas, resíduos carregados e hidrofóbicos são marcados, construindo-se um mapa. Demonstrou-se que a similaridade dos mapas de proteínas possui maior relação com sua similaridade de função do que com a similaridade de seqüência. Dois outros métodos procuram possíveis sítios catalíticos e, portanto, só se aplicam às enzimas.

Dois outras técnicas relacionadas buscam sítios de ligação. A primeira, que se aplica somente às interfaces proteína-proteína, utiliza redes neurais em conjunção com o conhecimento sobre os resíduos mais comumente encontrados em tais interfaces (Zhou e Shan, 2001). Cerca de 70% dos resíduos localizados nas interfaces analisadas foram identificados corretamente. A segunda técnica trabalha com informações de conservação de seqüência junto com uma estrutura protéica, buscando agrupamentos ao nível tri-dimensional de resíduos altamente conservados em um alinhamento múltiplo de seqüências homólogas (Aloy *et al.*, 2001). Esses agrupamentos representam previsões de sítios de ligação ao substrato ou a outras proteínas. O papel fundamental de conservação de seqüência nesse método reflete-se na dependência do sucesso obtido da variação presente no alinhamento de seqüências; somente nos casos de alinhamentos contendo seqüências mais diversas foram obtidos bons resultados. Felizmente, com a alta e crescente produção de seqüências esta limitação vai pesar cada vez menos.

Dois outros métodos procuram possíveis sítios catalíticos e, portanto, são aplicáveis somente às enzimas. O primeiro baseia-se na observação de evolu-

ção convergente. Com o número crescente de estruturas protéicas determinadas, ficou claro que várias classes de enzimas, mesmo não tendo uma relação evolucionária, usam conjuntos estruturalmente semelhantes de resíduos catalíticos para efetuar as suas reações químicas. O mais bem conhecido desses exemplos é a tríade catalítica Asp-His-Ser, visto pela primeira vez em serino proteases e, desde então, em várias outras classes de proteinases e lipases (Wallace *et al.*, 1996). Através de uma análise das características geométricas dessas tríades de origens independentes, pode-se formular regras para a identificação de futuros novos casos de evolução convergente (ex. Aghajari *et al.*, 1998; Hakansson *et al.*, 2000). É claro que a obtenção do conhecimento do mecanismo químico de uma nova enzima, possivelmente obtido através desse método, representa um grande passo para o bom entendimento da sua função.

Um método que identifica resíduos possivelmente catalíticos através do cálculo de curvas de titulação teórica foi recentemente publicado (Ondrechen *et al.*, 2001). Esse método se fundamenta na observação de que resíduos catalíticos ácidos ou básicos estão freqüentemente situados em microambientes que perturbam os seus valores pK_a . Essas mudanças otimizam as características do sítio catalítico para o químico ácido-base envolvido na catálise, melhorando assim a eficiência da enzima. Através de cálculos teóricos com várias estruturas de enzimas, observou-se que resíduos com curvas de titulação perturbadas estavam situados principalmente nos seus sítios catalíticos respectivos.

A idéia de usar modelos derivados de seqüências a serem anotadas funcionalmente pressupõe que as estruturas resultantes são de qualidade adequada. Nesse aspecto, dois fatores positivos podem ser identificados. Primeiro, a modelagem em si é uma área muito ativa de pesquisa, na qual avanços significativos (fora do âmbito deste artigo) estão sendo realizados continuamente. Segundo, para vários desses métodos mencionados, já foi vista uma relativa insensibilidade a erros presente nas estruturas (Zhou e Shan, 2001; Pawlowski e Goszik, 2001; Aloy *et al.*, 2001).

Conclusão

A determinação da função de uma proteína é uma tarefa árdua, e deve ser realizada por especialistas. Como se mostrou ao longo deste artigo, a interpretação direta/simples de resultados, especialmente provenientes do BLAST (método mais

utilizado na anotação funcional de uma nova proteína), pode levar a resultados/conclusões errôneos. Para se afirmar com segurança a função de uma nova proteína, muitas vezes faz-se necessária a utilização de mais de uma das técnicas aqui descritas, visto que é a associação de vários resultados que indicará a função protéica tão procurada. A diminuição das falhas que levam a uma interpretação errada do genoma refletirá diretamente na diminuição da perda de todo um investimento nas primeiras etapas de um projeto genoma. Isso é, sabendo-se que a anotação é o processo de interpretação da seqüência crua, e que fornece informações biológicas, a melhoria das técnicas de anotação visa a um melhor aproveitamento prático/aplicado dos genomas que vêm sendo determinados em campos como a agricultura (ex: melhor entendimento de mecanismos de defesa das plantas), e na medicina (ex: produção de vacinas e desenvolvimento de novos fármacos).

É verdade que não possuímos (ainda) muitos especialistas nessa área, que ainda se encontra em fase de crescimento, e, como dito anteriormente, mesmo sendo a anotação de genoma um foco de intensa pesquisa, os sistemas atuais estão longe de ser infalíveis. Porém, esforços vêm sendo realizados por diferentes Instituições de Pesquisas e Órgãos Financiadores, que visam à formação de novos pesquisadores, e vêm financiando projetos de pesquisa em bioinformática. Os projetos genoma vêm crescendo exponencialmente em todo o mundo, e a bioinformática é uma área que deverá crescer para que a demanda gerada por esses projetos possa ser atendida. No entanto, cabe ressaltar que a anotação de genoma é simplesmente uma das diferentes frentes da bioinformática, que abrange aplicações de computação em biologia molecular, através de uma série de outras técnicas (Luscombe *et al.*, 2001).

Referência Bibliográfica

- Aghajari N, Feller G, Gerday C, Haser R. (1998) Crystal structures of the psychrophilic alpha-amylase from *Alteromonas haloplanctis* in its native form and complexed with an inhibitor. *Protein Sci.* 7:564-572.
- Aloy P, Querol E, Aviles FX, Sternberg MJ. (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J Mol Biol.* 311:395-408.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990) Basic local alignment search tool. *J Mol Biol.* 215:403-410.
- Brenner SE. (1999) Errors in genome annotation. *Trends Genet.* 15:132-133.
- Devos D, Valencia A. (2001) Intrinsic errors in genome annotation. *Trends Genet.* 17:429-431.
- Doerks T, Bairoch A, Bork P. (1998) Protein annotation: detective work for function prediction. *Trends Genet.* 14:248-250.
- Gerlt JA, Babbitt PC. (2000) Can sequence determine function? *Genome Biol.* 1:REVIEWS0005.
- Hakansson K, Wang AH, Miller CG. (2000) The structure of aspartyl dipeptidase reveals a unique fold with a Ser-His-Glu catalytic triad. *Proc Natl Acad Sci U S A.* 97:14097-14102.
- Huynen MA, Diaz-Lazcoz Y, Bork P. (1997) Differential genome display. *Trends Genet.* 13:389-390.
- Huynen M, Snel B, Lathe W 3rd, Bork P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.* 10:1204-1210.
- Ichikawa T, Suzuki Y, Czaja I, Schommer C, Lessnick A, Schell J, Walden R. (1997) Identification and role of adenyl cyclase in auxin signalling in higher plants. *Nature.* 390:698-701.
- Karp PD. (1998) What we do not know about sequence analysis and sequence databases. *Bioinformatics.* 14:753-754.
- Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. (1996) Protein clefts in molecular recognition and function. *Protein Sci.* 5:2438-2452.
- Luscombe NM, Greenbaum D, Gerstein M. (2001) What is bioinformatics? A proposed definition and overview of the field. *Methods Inf Med.* 40:346-358.
- Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science.* 285:751-753.
- Marcotte EM. (2000) Computational genetics: finding protein function by nonhomology methods. *Curr Opin Struct Biol.* 10:359-365.
- Ondrechen, MJ, Clifton, JG, Ringe, D. (2001) THEMATICS: a simple computational predictor of enzyme function from structure. *Proc Natl Acad Sci U S A.* 98:12473-12478.
- Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. (1999a) The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A.* 96:2896-2901.
- Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. (1999b) Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol.* 1:93-108.
- Pawlowski K, Godzik A. (2001) Surface map comparison: studying function diversity of homologous proteins. *J Mol Biol.* 309:793-806.
- Pazos F, Valencia A. (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng.* 14:609-614.
- Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A.* 96:4285-4288.
- Pertsemlidis A, Fondon JW 3rd. (2001) Having a BLAST with bioinformatics (and avoiding BLASTphemy). *Genome Biol.* 2:2002.
- Rigden, DJ, Carneiro, M (1999). A structural model for the *rolA* protein and its interaction with DNA. *Proteins,* 37 697-708.
- Rigden, DJ, Bagyan, I, Lamani, E, Setlow, P, Jedrzejak, MJ (2001) A cofactor-dependent phosphoglycerate mutase homologue from *Bacillus* species is actually a broad specificity acid phosphatase. *Protein Sci.,* 10, 1835-1846.
- Thornton JM, Todd AE, Milburn D, Borkakoti N, Orengo CA. (2000) From structure to function: approaches and limitations. *Nat Struct Biol.* 7 Suppl:991-994.
- Thornton J. (2001) Structural genomics takes off. *Trends Biochem Sci.* 26:88-89.
- Todd AE, Orengo CA, Thornton JM. (2001) Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol.* 307:1113-1143.
- Wallace AC, Laskowski RA, Thornton JM. (1996) Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci.* 5:1001-1013.
- Zhou HX, Shan Y. (2001) Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins.* 44:336-343.