Universidade de São Paulo

Pró-Reitoria de Graduação

Curso de Ciências Moleculares

Ciclo Avançado

Projeto de Iniciação Científica

Uma abordagem computacional para a análise de propriedades estruturais e termodinâmicas da especificidade pelo substrato de metaloproteases M2

1º semestre de 2011

Carlos Eduardo Oliveira Vido
Turma 18
no. USP: 6433797
carlos.vido@usp.br

Prof. Dr. Paulo Sérgio Lopes de Oliveira
Centro Nacional de Pesquisa em Energia e Materiais
Laboratório Nacional de Biocências
(019) 3512 1032
paulo.oliveira@lnbio.org.br

Introdução

Proteases são enzimas capazes de clivar ligações peptídicas, em um processo chamado proteólise. Pode-se dizer que a proteólise começou a ser estudada em 1836, quando o fisiólogo alemão Theodor Schwann descobriu e passou a estudar a pepsina[1]. Após o advento da bioinformática, foram identificadas mais de 600 proteases humanas, que compõem cerca de 2% do genoma da espécie. Além da sua atividade na degradação de proteínas, descobriu-se também que essas enzimas têm participação em cascatas de sinalização celular[2,3].

Devido à sua possível interferência em vias de sinalização, deficiências proteolíticas (tanto por excesso quanto por falta de atividade) estão envolvidas em doenças cardiovasculares, inflamatórias, neurodegenerativas, bacterianas, virais e parasíticas[3]. Recentemente descobriu-se ainda que metaloproteases da matriz celular (MMPs) têm participação no crescimento de tumores e na metástase[4]. Portanto é muito ampla a utilização de inibidores de proteases como fármacos contra diversas doenças. O exemplo mais trivial é o dos inibidores da ECA (enzima conversora de angiotensina), utilizados no combate de pressão alta, como o Captopril (desenvolvido nos anos 70).

Já se determinou como as principais classes de proteases de mamíferos agem[2]. O banco de dados MEROPS divide essas enzimas em 7 super-clãs, de acordo com aspectos estruturais e funcionais. Nas cisteíno- e treonino-proteases, o resíduo que batiza o grupo é o nucleófilo responsável pela clivagem da ligação peptídica. As asparagino-proteases clivam a si próprias em uma ligação asparaginil, e a asparagina é o nucleófilo.[5] As demais (aspartato-, glutamato-, serino-, treonino- e metaloproteases) são hidrolases nas quais o composto que dá nome à enzima é responsável por coordenar a molécula de água que hidrolisa a ligação.

Metaloproteases

As metaloproteases são caracterizadas pela presença de um grupo funcional metálico. O íon metálico é responsável por formar um complexo com uma molécula de água no sítio ativo, responsável pela clivagem. Ele pode ser coordenado por quaisquer três aminoácidos com cadeias laterais carregadas (*i.e.*: arginina, histidina, lisina, aspartato ou glutamato).

As metaloproteases são ubíquas à vida terrestre e, na sua maioria, são zinco-dependentes. Segundo a classificação do banco de dados MEROPS, elas constituem um super-clã que engloba 14 clãs e 61 famílias de proteases, dentre as quais está a família M2 das metalo-exopeptidases, que contém a ECA[5].

A família M2 pertence ao clã MA, caracterizado pela presença de um padrão Xaa-Xbb-Xcc-His-Glu-Xbb-Xbb-His-Xbb-Xdd (onde Xaa é treonina ou hidrofóbico, Xbb é sem carga, Xcc é qualquer aminoácido exceto prolina e Xdd é hidrofóbico), em que os resíduos de histidina ligados a um átomo de zinco e o glutamato tem função catalítica. As enzimas da família M2 são exopeptidases que agem nas proximidades do C-terminal de oligopeptídeos, e possuem um glutamato próximo do seu próprio C-terminal que participa na coordenação do íon zinco[5].

A ECA por muito tempo foi a única enzima nessa família. Ela hidrolisa dipeptídeos (e ocasionalmente tripeptídeos) da região C-terminal de seus substratos, e tem em sua estrutura dois átomos de cloro essenciais para seu funcionamento[5]. Sua estrutura também se caracteriza pela inclusão do sítio ativo em um sulco, protegido do acesso de ligantes volumosos por três alfa-hélices[6].

Bases de dados de proteases e seus substratos

Sabe-se que cada protease se liga a uma sequência específica de aminoácidos, de forma que conhecer a sequência exata ajudaria a criar inibidores altamente específicos. Como occorre com outras moléculas de interesse clínico e biológico, há na internet bancos de dados como o MEROPS e o CutDB, que podem ser úteis para esse propósito.

O MEROPS é um banco de dados de peptidases curado manualmente, em que substratos e inibidores são organizados hierarquicamente em espécies, famílias e clãs. Seu foco é distinguir peptidases através de sua especificidade, de acordo com a sequência que ela cliva[7]. O repositório tem ferramentas de busca por peptidases ou inibidores (nome, número de acesso ou número MEROPS), por seus genes (nome ou dentro de um genoma), por estruturas dentro de um clã ou família específico, por genoma comparativo (organismo ou peptidases em comum entre genomas), por especificidade do sítio de clivagem (quais peptidases clivam uma determinada ligação ou substrato, que sítios de clivagem há dentro de uma proteína, que ligações uma peptidase pode clivar ou quais peptidases podem ser afetadas por um dado inibidor) ou busca cruzada com outros bancos de dados.

O CutDB é um depositário de eventos proteolíticos, reunindo conteúdo submetido diretamente por usuários e informações publicadas em outros bancos de dados ou em artigos científicos. Os eventos proteolíticos são caracterizados por três parâmetros (protease, substrato e sítio de clivagem) e acompanhados de um contexto biológico e comentários. Buscas no banco de dados podem ser conduzidas através de seis menus: número MEROPS, nome da protease ou do substrato, organismo de origem da protease ou do substrato e doença relacionada. Há ainda um recurso de busca textual, que procura nesses campos e também por sítios de clivagem, ID no PubMed e autor ou último editor da entrada no banco de dados[8].

Nenhum dos repositórios apresentam programas que, dada uma proteína qualquer, permita analisar qual protease seria capaz de clivá-la[†]. Fazendo uma analogia com outras ferramentas de bioinformática, seria desejável um programa que executasse para proteases algum tipo de reconhecimento de padrões para sequências-alvo tais quais aqueles existentes para os fatores de transcrição.

Fatores de transcrição são biomoléculas que se ligam a sequências específicas de bases nitrogenadas nos ácidos nucléicos, regulando a transcrição de uma determinada região codificante[9]. Bancos de dados como o TRANSFAC concentram informações sobre fatores de transcrição, seus sítios de ligação, matrizes de distribuição de nucleotídeos e os genes regulados por cada fator. No *site* do banco de dados,

[†]O MEROPS tem uma função semelhante, mas que apenas aceita como entrada moléculas escolhidas em uma lista.

http://www.gene-regulation.com, diversos programas de busca permitem encontrar em uma sequência de DNA possíveis sítios de ligação de fatores de transcrição[10].

Porém a pequena variedade de bases permite a esses programas de predição uma abordagem de força-bruta que é inviável no estudo de proteases; *e.g.*: um sítio de transcrição composto por 5 bases nitrogenadas tem $4^5 = 1024$ sequências possíveis, enquanto um peptídeo de 5 aminoácidos pode ter $20^5 = 3,2 * 10^6$ composições diferentes.

Uma abordagem mais eficiente seria inicialmente determinar o esqueleto molecular do substrato de cada protease em particular. Analisando o seu sítio ativo, seria possível determinar quais são as exigências termodinâmicas para cada resíduo do peptídeo-alvo. No exemplo anterior, se soubéssemos que o primeiro aminoácido deve ser, digamos, carregado negativamente, dos 3,2 milhões de possibilidades seriam excluídas 90%, restando 320 mil. Aplicando restrições semelhantes aos demais aminoácidos, chegaríamos a um número ainda mais diminuto de possibilidades, permitindo então a abordagem de força-bruta.

Objetivos

Este projeto visa implementar uma abordagem computacional para o estudo de propriedades termodinâmicas e estruturais dos sub-sítios de ligação de substratos de metaloproteases. O resultado deste estudo permitirá um melhor entendimento dos mecanismos envolvidos na seletividade e afinidade destas enzimas por seus substratos.

Objetivos específicos:

1. Buscar no PDB estruturas de proteínas homólogas à ECA contendo em seus sítios ativos algum inibidor ou fragmentos de peptídeos.
2. Sobrepor as estruturas obtidas de modo a posicionar todos os ligantes das diferentes proteases no mesmo sistema de coordenadas da estrutura de ECA.
3. Criar um algoritmo capaz de analisar a estrutura dos inibidores/peptídeos e determinar os espaços não redundantes ocupados pelos mesmos dentro da estrutura de ECA, permitindo a identificação de sub-sítios putativos.
4. Mapear as propriedades físico-químicas dos sub-sitios encontrados em 3, tais como: volume, área e potenciais eletrostáticos.
5. Predizer potenciais aminoácidos capazes de ocupar os sub-sítios encontrados com base em cálculos de energia de ligação entre peptídeos obtidos do banco de dados MEROPS.

Métodos

Busca e sobreposição de estruturas de proteínas homólogas à ACE

As estruturas tridimensionais utilizadas para análise serão escolhidas dentre as disponíveis no Protein DataBank (PDB). Para analisar a estrutura do sítio ativo, deve-se levar em conta a mudança de conformação induzida pela acoplagem de um substrato no mesmo. Dessa forma, é necessário construir complexos protease-peptídeo para análise.

Portanto foi escolhida a enzima conversora de angiotensina complexada com o inibidor Lisinopril (número de acesso 1O86 no PDB) como estrutura modelo para o estudo. Outras estruturas homólogas serão utilizadas para produzir uma superposição, com o objetivo de tornar mais evidentes a geometria e as propriedades eletrostáticas do sítio ativo. Para isso, será utilizado o software FATCAT, que alinha estruturas de proteínas levando em consideração sua flexibilidade e os rearranjos estruturais que podem ocorrer em decorrência de sua função. Como resultado, esse software obtém resultados mais precisos do que programas que comparam proteínas como corpos rígidos[11].A partir do arquivo de alinhamento, pode-se fazer com que um software de exibição de estruturas tridimensionais mostre simultaneamente tantas proteínas quanto se deseje, e pode-se limitar a região de exibição para uma caixa ao redor das regiões de interesse, ignorando o restante.

Identificação de sub-sítios putativos

Após a sobreposição de uma quantidade razoável de complexos protease-substrato, espera-se a evidenciação de regiões que não sejam comuns a todos os substratos. Essas regiões revelam a presença de sub-sítios que ajudam a elucidar o mecanismo que rege diferenças de afinidade de uma dada enzima por substratos diversos. Resumidamente, esse processo será conduzido através de um programa capaz de varrer o espaço tridimensional da proteína na região dos sub-sítios, procurando por espaços contendo ligantes nas diferentes estruturas sobrepostas. O programa analisará os ligantes procurando por um conjunto único de átomos que preencha o maior número de espaços vazios das metaloproteases. Estas regiões serão definidas como os sub-sítios putativos de ligação do substrato.

Mapeamento das propriedades físico-químicas dos sub-sitios

Para tal análise, será necessário mapear os sub-sítios de ACE, analisando fatores volumétricos (*i.e.*, se o volume de uma determinada cavidade restringe a entrada de certos resíduos) e de caráter eletrostático, manifestas na forma de interações de atração/repulsão. A determinação volumétrica será realizada por um programa desenvolvido pelo estudante de mestrado Saulo Henrique Pires de Oliveira do grupo de Bioinformática Estrutural do LNBio. O cálculo do potencial eletrostático dos sub-sitios será realizado pelo programa YASARA (http://www.yasara.org).

Predição de potenciais aminoácidos capazes de ocupar os sub-sítios de ACE

Para medida de especificidade e afinidade por aminoácido em cada sub-sítio estabelecido acima, escolheremos cinco substratos conhecidos para ACE e construiremos a estrutura tridimensional de peptídeos derivados destes substratos na região de clivagem obtida da base de dados MEROPS. Estes petídeos serão ancorados manualmente à estrutura de ACE, utilizando como base os sub-sítios putativo encontrados nas etapas anteriores. Serão medidas então as energias de ligação de cada peptídeo à protease.O cálculo da energia de ligação entre a protease e seus substratos será realizado através da técnica de dinâmica molecular acoplada ao método de Poisson-Boltzman[12] (MM-PBSA) e ao método generalizado de Born[13] (MM-GBSA). Ambos calculam uma aproximação para a diferença entre a energia livre do complexo enzima-substrato e a de cada macromolécula sozinha.

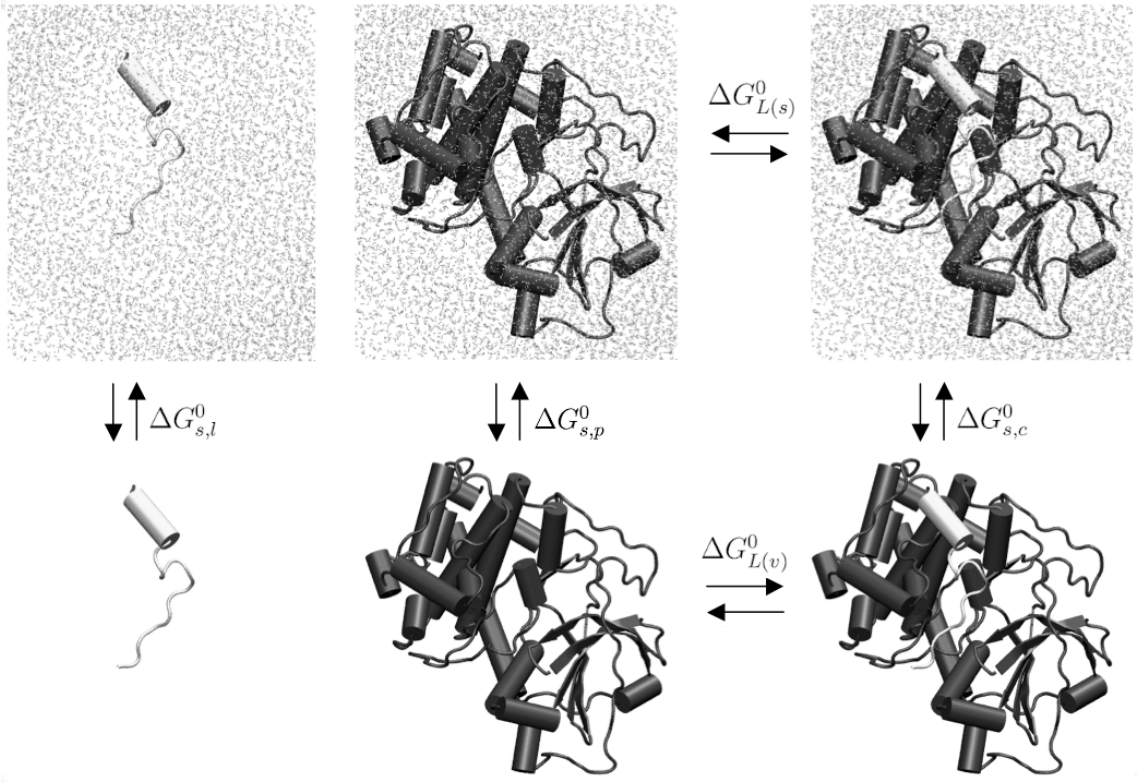

Figura 1: Esquema do ciclo termodinâmico a ser utilizado para calcular a energia de ligação entre a protease, mais escura, e o ligante. O método implementado visa calcular a diferença de energia livre (ΔG^0) entre duas moléculas ligadas e as mesmas moléculas interagindo diretamente com o solvente. Para estabelecer a influência do solvente na componente eletrostática, a aproximação de solvente implícito é utilizada. *s,l = solvente, ligante; s,p = solvente, protease; s,c = solvente, complexo; L(s) = ligação em solvente; L(v) = ligação no vácuo.*

O cálculo da energia de ligação pode ser feito de acordo com a equação:

$$\Delta G^0_{L(s)} = \Delta G^0_{L(v)} + \Delta G^0_{s,c} - (\Delta G^0_{s,p} + \Delta G^0_{s,l})$$

No método MM-PBSA, o $\Delta G^0_{L(v)}$ é obtido calculando-se a energia média de interação entre duas moléculas e, se assim desejado, a variação de entropia causada pela ligação. As energias de solvatação são calculadas pela equação linearizada de Poisson-Boltzmann, que considera a contribuição da energia eletrostática para a energia livre de solvatação e adiciona um termo empírico correspondente à contribuição hidrofóbica. Portanto, os termos à direita da igualdade acima podem ser expressos pelas equações:

$$\Delta G^0_{L(v)} = \Delta E_{mec.molecular} - T * \Delta S$$
$$\Delta G^0_{s,l} = G^0_{hidro} + (G^0_{E,\epsilon=80} - G^0_{E,\epsilon=1})$$
$$\Delta G^0_{s,p} = G^0_{hidro} + (G^0_{E,\epsilon=80} - G^0_{E,\epsilon=1})$$
$$\Delta G^0_{s,c} = G^0_{hidro} + (G^0_{E,\epsilon=80} - G^0_{E,\epsilon=1})$$

em que $\Delta E_{mec.molecular}$ é a variação de energia calculada por mecânica molecular, T é a temperatura, ΔS é a entropia, G^0_{hidro} é a contribuição hidrofóbica para a energia livre e $G^0_{E,\epsilon=x}$ é a parcela eletrostática da energia livre de solvatação com constante dielétrica $\epsilon = x$.

A contribuição da entropia pode ser estimada através de uma análise de modo normal nas três estruturas (ligante, protease e complexo), mas na prática o resultado pode ser desprezado, pois os valores são muito pequenos e essa análise, além de custosa computacionalmente, possui uma margem de erro muito grande.

As energias médias de interação entre a protease e seu substrato são obtidas através de cálculos feitos a partir de *snapshots* de uma simulação de dinâmica molecular em equilíbrio. O ideal seria realizar esse tipo de simulação para as três estruturas, mas o processo é otimizado analisando-se apenas o complexo e admitindo que as mudanças nas outras estruturas são irrelevantes.

Preparação para dinâmica molecular

Os complexos terão suas estruturas preparadas para minimização de energia pelo software YASARA, que tem uma função capaz de criar todos os átomos que não estão no modelo (*e.g.*, os átomos de hidrogênio). O programa Amber99 é capaz de calcular os parâmetros de força necessários. Serão criados então quatro modelos: a protease por si só, o peptídeo por si só, o complexo e o complexo imerso em soluto.

Minimização de energia e dinâmica molecular

O software Sander, do pacote AMBER, será responsável pela minimização de energia e dinâmica molecular. Será utilizada uma distância de corte de 8Å.

As estruturas dos complexos terão suas energias minimizadas, com o objetivo de eliminar contatos de van der Waals inadequados, utilizando 500 passos de gradiente descendente e até 500 passos de gradiente conjugado. Sobre todos os átomos de hidrogênio será utilizado o método Shake[14,15], com passo de 2fs, em todas as etapas da dinâmica molecular. Todos os átomos serão submetidos a restrição de movimento, para evitar movimentos muitos amplos causados pelo posicionamento de uma molécula de solvente demasiado próxima do complexo.

Por dinâmica molecular, se realizará a elevação da temperatura da solução de 0K a 300K e a eliminação de contatos de van der Waals inadequados e de distorções estereoquímicas, *e.g.* ligações entre átomos com ângulos ou comprimentos fora dos padrões observados. A simulação terá 25 mil passos.

Uma segunda simulação será realizada, a 300K e com as mesmas restrições, para atingir o equilíbrio de densidade no sistema. Novamente, serão utilizados 25 mil passos. Em sequência, 250 mil passos a 300K serão realizados para equilibrar o sistema.

Finalmente, será feita uma simulação com passo de 2ns a 300K para explorar a conformação do complexo protease-peptídeo. A cada cinco mil passos será salvo um *snapshot* do sistema para análise posterior.

Referências

1. Florkin, M. "Discovery of pepsin by Theodor Schwann". *Revue médicale de Liège* **12** (5), 139–44 [1957]

2. Turk, B. "Targeting proteases: successes, failures and future prospects". *Nature Rev. Drug Discov.* **5**, 785–799 [2006]

3. Drag, M, Salvesen, GS. "Emerging principles in protease-based drug discovery". *Nature Rev. Drug Discov.* **9**, 690–701 [2010]

4. Gialeli C, Theocharis AD, Karamanos NK. "Roles of matrix metalloproteinases in cancer progression and their pharmacological targeting". *FEBS J.* **278**, 16-37 [2010]

5. *MEROPS - the Peptidase Database.* http://merops.sanger.ac.uk

6. Hodis E, *et al.* "Proteopedia - a scientific 'wiki' bridging the rift between three-dimensional structure and function of biomacromolecules". *Genome Biol.* **9**(8), R121 [2008]

7. Rawlings, ND *et al.* "MEROPS: the peptidase database". *Nucleic Acids Res.* **38**, D227–D233 [2010]

8. Igarashi, Y *et al.* "CutDB: a proteolytic event database". *Nucleic Acids Res.* **35**, D546–D549 [2007]

9. Latchman, DS. "Transcriptional Factors: An Overview". *Int. J. Biochem. Cell Biol.* **29**, 1305–1312 [1997]

10. Matys, V *et al.* "TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes". *Nucleic Acids Res.* **34**, D108–D110 [2006]

11. Ye, Y, Godzik, A. "Flexible structure alignment by chaining aligned fragment pairs allowing twists". *Bioinformatics* **19**(2), ii246-55. [2003]

12. Luo, R, David, L, Gilson, MK. "Accelerated Poisson-Boltzmann Calculations for Static and Dynamic Systems". *J. Comput. Chem.*, **23**(13), 1244-53. [2002]

13. Case DA, Luo, R et al. AMBER 9 User Manual. [2006]

14. Ryckaert, J-P, Ciccotti G, Berendsen HJC. "Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes". *J. Comput. Phys.* **23**, 327–341[1977]

15. Miyamoto, S, Kollman PA. "SETTLE: An Analytical Version of the SHAKE and RATTLE Algorithm for Rigid Water Models". *J. Comput. Chem.*, **13**, 952–962 [1992]