UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
CENTRO DE BIOTECNOLOGIA

MODELAGEM DE SERINO-PROTEASES E INIBIDORES COM EMPREGO DE FERRAMENTAS DE BIOINFORMATICA ESTRUTURAL

CRISTINA RUSSO

Porto Alegre, Março de 2006

CRISTINA RUSSO

MODELAGEM DE SERINO-PROTEASES E INIBIDORES COM EMPREGO DE FERRAMENTAS DE BIOINFORMATICA ESTRUTURAL

Dissertação submetida ao Programa de Pós-graduação em Biologia Celular e Molecular da UFRGS como requisito parcial para obtenção de grau de Mestre em Biologia Celular e Molecular

Orientador: Dr. Jorge Almeida Guimarães

Porto Alegre, Março de 2006

Instituições e fontes financiadoras

Este trabalho foi desenvolvido na Universidade Federal do Rio Grande do Sul, no Laboratório de Bioinformática Estrutural e de Bioquímica Farmacológica – Centro de Biotecnologia, UFRGS. Teve auxílio do Centro de Biotecnologia, financiado pela CAPES, CNPq e FAPERGS.

Banca Examinadora

Dr. Paulo A. Netz (Instituto de Química – UFRGS)

Dra. Paula Regina Kuser Falcão (Embrapa – Campinas)

Dr. Henrique Bunselmeyer Ferreira (Centro de Biotecnologia – UFRGS)

Dr. Carlos Termignoni (Centro de Biotecnologia – UFRGS)

Dr. Tarso Benigno Ledur Kist (Suplente) (Instituto de Biociências – UFRGS)

À minha família de ambos hemisférios

Agradecimentos

. À minha família: mãe, pai, Silvana, Andréa e Marcelo. Ao Rory (e também Ford), núcleo da minha família americana. Às adições Sam, Érico, Raquel, Ricardo, Bruno, Nisia, Wagner e demais amigos pelotenses.

Aos mentores, que sempre me incentivaram e fizeram minhas idéias crescerem. Conselheiros, não apenas no meio acadêmico, mas com a sabedoria para uma vida. Professores Hermes de Amorim e Jorge Guimarães, que me aceitaram em seu laboratório e me guiaram nesta jornada científica que está em andamento.

Aos meus colegas do laboratório de Bioinformática Estrutural, assim como os colegas dos laboratórios vizinhos do Centro de Biotecnologia. Aos muitos professores, como o Prof. Carlos Termignoni, Prof. Tarso Kist e a Prof.ª Ana Bazzan, com quem tive contato e que cultivaram de alguma forma minhas idéias.

RESUMO

A família das serino proteases inclui diversas proteínas envolvidas em uma variedade de processos fisiológicos, como digestão protéica, regulação da pressão arterial e da coagulação sanguínea, entre outros. Quando disfuncionais, relacionam-se à condições patológicas graves como trombose vascular, embolismo pulmonar, infarto agudo do miocárdio, isquemia cerebral, bem como com o quadro da hemofilia B, ocasionada pela deficiência de fator IX. A investigação dos mecanismos de ação e a compreensão do funcionamento de serino proteases e de seus inibidores, as serpinas, é essencial para desvendar mecanismos de doenças e para a modelagem de fármacos mais específicos e eficientes.

A pesquisa de fármacos anti-trombóticos busca desenvolver produtos capazes de interferir no processo hemostático que pode resultar no grave quadro trombo-embólico. A nitroforina-2 (NP-2), produzida na secreção salivar do inseto barbeiro *Rhodnius prolixus*, é uma proteína dotada de potente ação anti-hemostática, atuando como um inibidor específico do complexo tenase intrínseco (TF/FVIIa/FIX). Em trabalho anterior, nosso grupo identificou que o fragmento correspondente ao segmento 90-110 de NP-2 possui atividade anticoagulante. Um modelo do complexo fator IXa (fIXa)-NP-2 foi gerado a partir do "docking" do peptídeo de NP-2 na provável região de ligação com fIXa. Este complexo foi usado como ponto de partida para a simulação por dinâmica molecular e refinamento da estrutura. Novos peptídeos foram propostos como o objetivo de gerar estruturas com maior afinidade por fIXa. Identificamos o hexapeptídeo LKEADE como apresentando melhor afinidade por fIXa, o qual sugerimos como "template" para o planejamento racional de fármacos com ação anticoagulante e anti-trombótica.

Numa outra vertente de nosso trabalho, estudamos a superfamília de proteínas denominada serpinas (serpins, ou "SERine Protease INhibitors") que engloba um grande conjunto de proteínas dotadas de estruturas similares e provenientes de vários e distintos organismos. A função inicialmente identificada para as serpinas foi de inibição de serino proteases da coagulação sangüínea; entretanto, muitas serpinas perderam esta função.

Nestes estudos foram usados modelos ocultos de Markov para criar modelos e descrições possibilitando classificar as serpinas em relação à sua função biológica. Foram criadas assinaturas para cada grupo: seqüências consenso e padrões de aminoácidos distintos para cada modelo/função correspondente. Um modelo específico para serpinas da coagulação sangüínea foi criado. Para este modelo, foi gerada a expressão regular [IVTLM]-[FLVA]-F-S-P-[VLWYF]-[SG]-[IV] que descreve tal função. Além disso, ela codifica a seqüência de uma importante região envolvida na mudança conformacional e no funcionamento da serpina. Tanto esta seqüência quanto a estrutura secundária correspondente podem ser melhor investigadas, por sugerirem um alvo para inibição ou ativação.

ABSTRACT

The protein family of serine proteases comprises molecules involved in a wide range of physiological processes, such as regulation of blood pressure and coagulation. Dysfunctional serine proteases are related to pathological conditions such as embolism, heart failure, cerebral ischemia and also hemophilia B (the latter being a consequence of factor IX deficiency). The understanding of serine proteases mechanism of action, as well as of their inhibitors is a key step to the discover and develop new drugs.

In the study of homeostasis, the search for anti-thrombotic drugs seek to avoid thrombosis and related conditions. The protein nitrophorin-2 (NP-2), found in the salivary glands of the hematophagous insect *Rodnius prolixus*, is a specific inhibitor of the intrinsic tenase complex (TF/fVIIa/fIX). In previous studies, we identified the NP-2 fragment (amino acid sequence 90-110) showing anticoagulant activity. Models of the complex factor IXa (fIXa) – NP-2 were created. This complex was used as a starting point for the molecular dynamic simulations and structure refinement. New NP-2 peptides were suggested, in order to achieve higher affinity complexes. We identified the hexapeptide LKEADE as showing higher affinity for fIXa, which we suggest to be used on rational drug design studies for anticoagulant drugs.

The superfamily of proteins known as serpins (“Serine Protease Inhibitors”) involve a number of similar structures, found in a wide variety of organisms. Its function was initially identified as an inhibitor of blood clotting serine proteases. However, many serpins have lost that function in the course of natural evolution. Hidden Markov models were then used to create profiles classifying those proteins due to their biological function. We created signatures for each group of functional serpins in the form of consensus sequences and distinct amino acid patterns. A blood clotting-specific model was generated, which has yielded the regular expression [IVTLM]-[FLVA]-F-S-P-[VLWYF]-[SG]-[IV]. Such pattern also codes for a region of the structure involved in an extensive conformational change. The change is related to the activation of the serpin, and, therefore, both pattern and sequence should be investigated. Combined, these information suggest a powerful target for blood clotting inhibition.

Índice

Lista de figuras

Lista de abreviaturas

Serpins – inibidores de serino-proteases “Serine Protease Inhibitors” (pg. 9)

DM – Dinâmica Molecular (pg. 16)

PDB – Protein Data Bank (pg. 15)

EGF – fator de crescimento epidérmico “Epidermal Growth Factor” (pg. 17)

Gla – domínio rico em resíduos de ácido γ-carboxiglutâmico (pg. 17)

TF – fato tissular “Tissue Factor” (pg. 19)

fIX – fator IX (pg. 17)

fIXa – fator IX ativado (pg. 17)

fX – fator X (pg. 17)

fVIIa – fator VII ativado (pg. 17)

NP-2 – nitroforina-2 (pg. 20)

RMN – ressonância magnética nuclear (pg. 26)

HMM – modelos ocultos de Markov “Hidden Markov Models” (pg. 41)

SD – “Steepest Descent” (pg. 36)

TIP3 – “Transferable Interaction Potential Three Point”(pg. 39)

SPC – “Single Point Charge”(pg. 39)

SPC/E – “Single Point Charge Extended” (pg. 39)

1. INTRODUÇÃO

Proteínas são polímeros estruturais e componentes essenciais da célula viva onde desempenham múltiplas funções, entre elas a regulação de processos fisiológicos dos organismos vivos. A família das serino-proteases é constituída por inúmeras proteínas dotadas de atividade enzimática (proteínas hidrolíticas que apresentam um resíduo nucleofílico de serina na tríade catalítica), estando envolvidas em uma ampla variedade de processos fisiológicos, incluindo aí a digestão protéica, a regulação da pressão arterial, a coagulação sangüínea, e muitos outros.

Por suas propriedades funcionais, que necessitam refinado mecanismo fisiológico para sua regulação, as serino-proteases estão freqüentemente associadas à gênese de condições patológicas graves como a trombose vascular, o embolismo pulmonar, infarto agudo do miocárdio, isquemia cerebral, bem como o quadro da hemofilia B, ocasionada pela deficiência de fator IX, um importante componente da cascata da coagulação sangüínea. Assim, a melhor compreensão da interação das serino-proteases com seus inibidores é essencial para desvendar mecanismos de certas doenças e para o planejamento de fármacos mais específicos e eficientes.

Os inibidores das serino-proteases constituem também uma família de proteínas e peptídeos de grande interesse, em virtude de seu reconhecido potencial terapêutico. Tais proteínas, conhecidas como serpinas, tem sua denominação cunhada no idioma inglês como serpin, para "SERine Protease INhibitor". As serpinas produzem inibição quase sempre específica de certas serino-proteases, promovendo desta forma sua regulação seletiva. Assim, a α-1-antitripsina, uma serpina típica, exerce no alvéolo pulmonar importante papel regulador do eventual efeito nocivo resultante do excesso da elastase neutrofílica localmente liberada. No exemplo, essa serpina impede a formação de enfisema pulmonar que ocorreria devido à destruição dos alvéolos pulmonares provocada pela excessiva e prolongada ação da elastase. Por outro lado, alterações pontuais na estrutura de uma serpina, sejam de origem mutagênica, conformacional ou por reatividade com grupos funcionais de sua seqüência primária, podem resultar na perda da sua função

inibitória. Uma disfunção comum na α-1-antitripsina é a formação de metionina-sulfona por agentes oxidantes, uma condição freqüente entre fumantes crônicos. Disto resulta a inativação da serpina como reguladora da elastase e o conseqüente agravamento do quadro de enfisema pulmonar. Igualmente em outros processos fisiológicos como a hemostasia, a disfunção ou a deficiência da serpina regulatória como, por exemplo, na deficiência de antitrombina, resulta em um quadro de hipercoagulabilidade, levando à formação de coágulos espontâneos, o que propicia o risco da ocorrência do grave processo trombo-embólico.

O sistema hemostático oferece diversos alvos para a atuação de possíveis agentes anti-hemostáticos, tanto naturais como sintéticos. Portanto, o descobrimento de novos princípios ativos e o desenvolvimento de drogas e produtos capazes de propiciar a adoção de procedimentos clínicos e terapêuticos como instrumentos de intervenção anti-trombótica constitui hoje a estratégia de eleição para a prevenção e o tratamento do grave quadro trombo-embólico. Não é sem razão que a busca de novos princípios anti-hemostáticos de origem natural seja objeto de acirrada competição na pesquisa acadêmica e na indústria farmacêutica.

As estratégias de pesquisa que buscam identificar substâncias capazes de inibir a formação do trombo incluem o estudo e a caracterização de inibidores específicos para proteínas envolvidas na cascata de coagulação. A heparina, o agente natural mais comumente usado na terapia anticoagulante, inibe a trombina (uma serino-protease) via ativação da antitrombina (uma serpina). Porém, a heparina é inespecífica, ligando-se a várias outras proteínas. Uma pista para a busca de inibidores específicos é a procura na saliva de animais hematófagos: estes precisam de seus anticoagulantes naturais para bloquear as defesas hemostáticas do hospedeiro.

O melhor conhecimento da estrutura e da função das serpinas é, assim, de grande importância para estudos de modelagem de novas drogas terapêuticas destinadas não apenas ao tratamento da trombose, como também de outros problemas de saúde humana, incluindo desde quadros inflamatórios até o câncer. O alto grau de especificidade na atividade desses inibidores sobre as serino-proteases faz com que as serpinas sejam

ferramentas muito úteis para o estudo dos mecanismos de ação, da regulação e das relações entre estrutura e função de tais enzimas.

1.1. Emprego de Métodos de Bioinformática

Uma proteína é um polímero formado de aminoácidos. Processos de seqüenciamento têm sido empregados na identificação das seqüências de proteínas desde a explosão de dados genômicos. Estruturas de proteínas e de outros biopolímeros estão sendo determinadas pela genômica estrutural e por outros esforços. Os projetos genoma (e.g. o Projeto Genoma Humano, assim como o seqüenciamento de vários outros organismos incluindo bactérias, plantas e outros eucariotos) disponibilizaram uma grande quantidade de informações, as quais encontram-se distribuídas em bancos de dados na internet. Porém, grande parte do processo de anotação funcional está incompleto ou errado. Ainda são desconhecidos dados específicos de regulação dos genes, mapas de vias metabólicas ou sinalização celular. Seqüências sem similaridade não podem ser anotadas. Soma-se a isso a dificuldade em encontrar e extrair informações dos bancos de dados públicos.

A associação dos dados de seqüência disponíveis à modelos físicos pode esclarecer muito sobre a função e o comportamento de várias proteínas. Partindo daí, pode-se incorporar estudos de descoberta de fármacos e de outras aplicações para previsão do comportamento de sistemas biológicos. Para tanto, aproveita-se do crescente poder computacional para o tratamento, classificação, modelagem e simulação de sistemas.

Dados de função biológica estão altamente conectados à estrutura tridimensional que a seqüência de aminoácidos assume quando exposta às condições fisiológicas. Técnicas experimentais, como cristalografia de raios-X e ressonância magnética nuclear (RMN), vem sendo empregadas na determinação de um grande número de estruturas. Estas estruturas estão depositadas em bancos de dados como o Protein Data Bank, PDB

(BERMAN *et al.*, 2000; BERMAN *et al.*, 2003). De acordo com estatísticas do PDB, obtidas em janeiro de 2006, o número de estruturas depositadas é de 34777.

Deve-se considerar que muitas das estruturas de proteínas depositadas no PDB correspondem a uma mesma proteína, cuja diferença pode ser a complexação com ligantes diferentes ou a presença de mutações sítio dirigidas. Em vista disto, estima-se que o número total de diferentes moléculas representadas no PDB seja em torno de 4.000. Assim, há ainda um grande número de estruturas a ser descoberto. Outras questões podem ainda ser consideradas. Por exemplo, se a estrutura cristalográfica representa a estrutura da proteína no seu estado fisiológico e funcional, se a resolução obtida é adequada para análise detalhada e precisa da estrutura, sendo que muitas proteínas não cristalizam nas condições conhecidas (GRANT & RICHARDS, 1995). Além disso, a cristalografia é uma técnica cuja obtenção de resultados algumas vezes pode ser medida em anos.

Estas condições ensejaram o desenvolvimento de uma nova área da Bioinformática, conhecida como Bioinformática Estrutural (BIE). A BIE pode ser definida como uma área do conhecimento que envolve a intersecção de outras três grandes áreas, a Biologia Molecular Estrutural, a Modelagem Molecular e a Bioinformática (DURBIN *et al.*, 1998). A evolução do poder computacional tem contribuído cada vez mais para o crescimento da Bioinformática Estrutural, pois computadores pessoais tornam-se mais rápidos e acessíveis, a uma velocidade que dobra a cada dois anos (BRONSON & BRAND, 2004).

Conforme resultados experimentais substituem cada vez mais a noção de que proteínas são estruturas rígidas, a dinâmica molecular (DM) passou a ser usada para investigar flutuações e tendências dos movimentos globais e locais de macromoléculas biológicas. Comparações experimentais indicam que simulações de dinâmica molecular são capazes de fornecer dados detalhados da evolução de um sistema biomolecular e de sua trajetória (KARPLUS, 2003). Assim, a DM permite o estudo dos fenômenos biológicos em tempo real. A previsão do comportamento é foco em problemas como especificidade de ligantes, mecanismos de ativação, regulação de moléculas, entre outros.

1.2. Serino-proteases

A família das serino-proteases envolve um conjunto de proteínas com atividade proteolítica e alto grau de homologia de seqüência e de similaridade estrutural. Mais de 500 serino-proteases já foram seqüenciadas, e mais de 20 estruturas determinadas (FERSHT, 2002). Estas proteínas exercem muitas atividades fisiológicas, como digestão (cujos representantes mais comuns são tripsina, quimotripsina e elastase), ativação do sistema imunológico e inflamação (por exemplo, kalikreínas e bradicinina), coagulação sangüínea (trombina, fVIIa, fIXa, fXa, fVIIIa, fVa, proteína C).

As serino-proteases são altamente conservadas com relação à estrutura nos módulos serino-proteinase (domínio catalítico). A característica mais comum destas enzimas é a tríade catalítica constituída pelos resíduos His57, Asp102, e Ser195 (conforme sistema de numeração baseado no quimotripsinogênio (BODE *et al.*, 1989)). Várias serino-proteases existem na forma de multi-domínios. Por exemplo, os fatores VIIa, IXa e Xa da cascata da coagulação sangüínea apresentam, além do domínio serino-protease, os domínios EGF-like (EGF: fator de crescimento epidérmico “Epidermal Growth Factor”) e Gla (domínio rico em resíduos de ácido γ-carboxiglutâmico), os quais podem apresentar mais de 60% de homologia (NORLEDGE, 2003).

O padrão conservado de enovelamento destas proteínas é ilustrado na figura 1, onde é mostrada a representação esquemática em fitas de fIXa. Apesar de muito semelhantes quanto à estrutura, as serino-proteases apresentam diferentes especificidades em relação ao substrato, decorrentes de diferenças topológicas e eletrostáticas na região do sítio ativo.

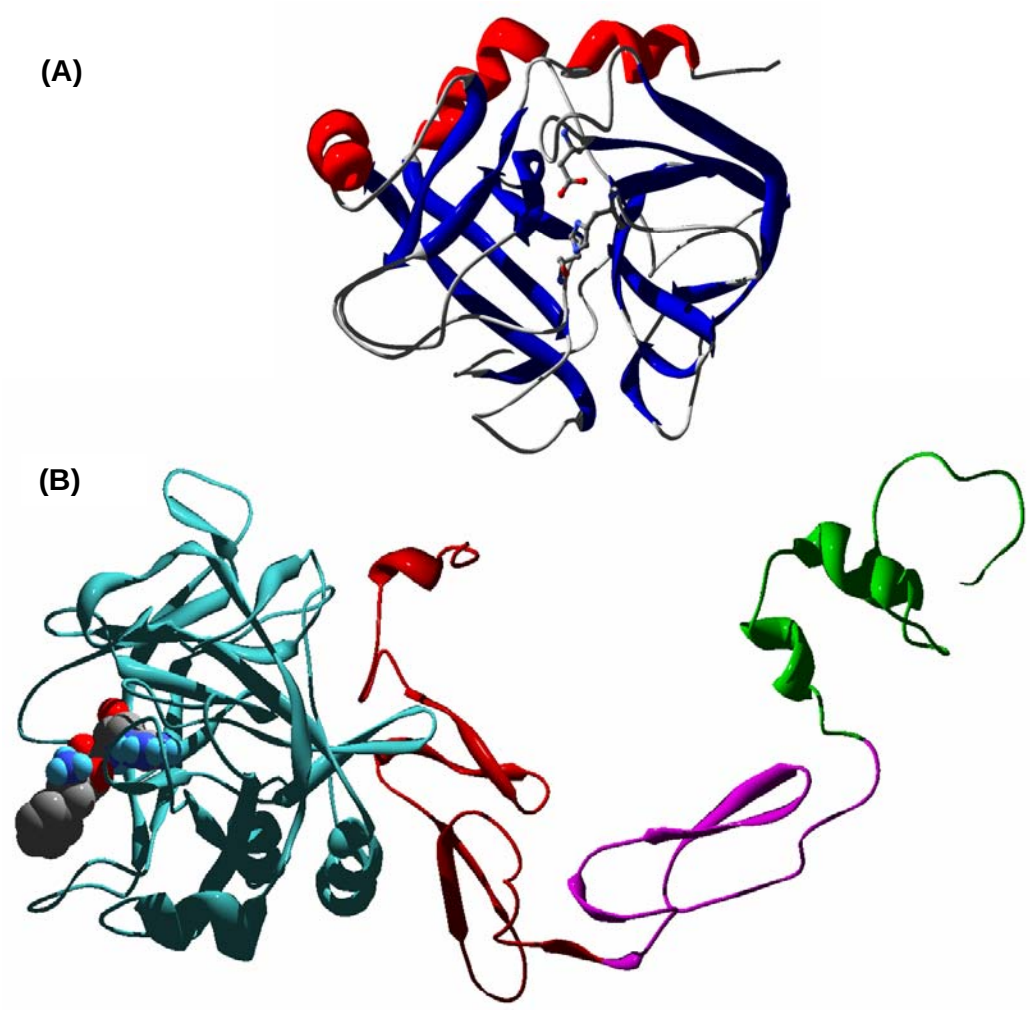

Figura 1. Representação esquemática em fitas do fator IXa. Painel A: domínio catalítico (serino-protease) do fator IXa humano, colorido em função da estrutura secundária (código 1RFN do PDB). Painel B: fator IXa suíno (código 1PFX do PDB). Verde: módulo Gla; lilás: módulo EGF-2; vemelho: mólulo EGF-1; azul claro: módulo catalítico. A estrutura do inibidor D-Phe-Pro-Arg é representada pelo modelo de preenchimento espacial (*space filling*).

1.2.1. Serino-proteases da coagulação sangüínea

A coagulação sangüínea é o resultado final de uma complexa série de reações seqüenciais, em cascata, envolvendo várias glicoproteínas (e seus co-fatores) presentes no plasma. O resultado final da ativação da cascata da coagulação é a formação da rede de fibrina polimerizada. A figura 2 apresenta um esquema da cascata da coagulação, a qual pode ser dividida em duas vias, a intrínseca (fisiológica) e a extrínseca. A ativação de cada uma das vias, que pode ocorrer simultaneamente ou separadamente, dependerá da natureza do estímulo desencadeador. A via mais comum e mais rápida é via extrínseca, a qual é ativada pela lesão de vasos sangüíneos ou a partir de estímulos tissulares. Na via extrínseca, o fator tissular (TF, fator III ou tromboplastina tecidual) é liberado no local da lesão do endotélio vascular e desencadeia as reações de coagulação. O fator tissular encontra-se expresso em células epiteliais, macrófagos e outros tipos de células que estão normalmente separadas do plasma sangüíneo e de seus fatores de coagulação circulantes. Em baixas concentrações, este pode ainda estar expresso na superfície de monócitos e em micropartículas derivadas de leucócitos. Estas fontes intravasculares de fator tissular podem servir para “amarrar”, e concentrar, plaquetas (ativadas) e células endoteliais no local da lesão e inflamação. Como efeito desencadeador da coagulação, a liberação de uma maior concentração de fator tissular no plasma pode ocorrer em conseqüência de eventos como a perfuração de uma parede vascular ou, ainda, a partir da ativação do endotélio por processos inflamatórios, substâncias químicas ou citocinas.

No local da lesão, o fator tissular se combina com o fator VII (acelerador da conversão da protrombina no plasma) e, na presença do fator IV (cálcio), promove a rápida ativação deste zimogênio em fator VII ativo (FVIIa). Por sua vez, o complexo TF-FVIIa pode atuar tanto na ativação direta do fator X quanto na ativação indireta deste através da ativação do fator IX, reforçando a cascata da via intrínseca.

A via intrínseca pode ser iniciada *in vitro* pelo contato de quatro proteínas da coagulação (fator XII, XI, pré-calicreína e cininogênio de alto peso molecular) com uma

superfície negativa (silicatos, pó de vidro, celite) que simule a superfície das estruturas presentes no revestimento das paredes vasculares ou nas células do sangue. A interação complexa das quatro proteínas com a superfície negativa, denominada reação de ativação pelo contato, resulta na conversão do fator XI a fator XIa que, por sua vez promove também a ativação do fator IX. *In vivo*, o fator XI é ativado pela trombina.

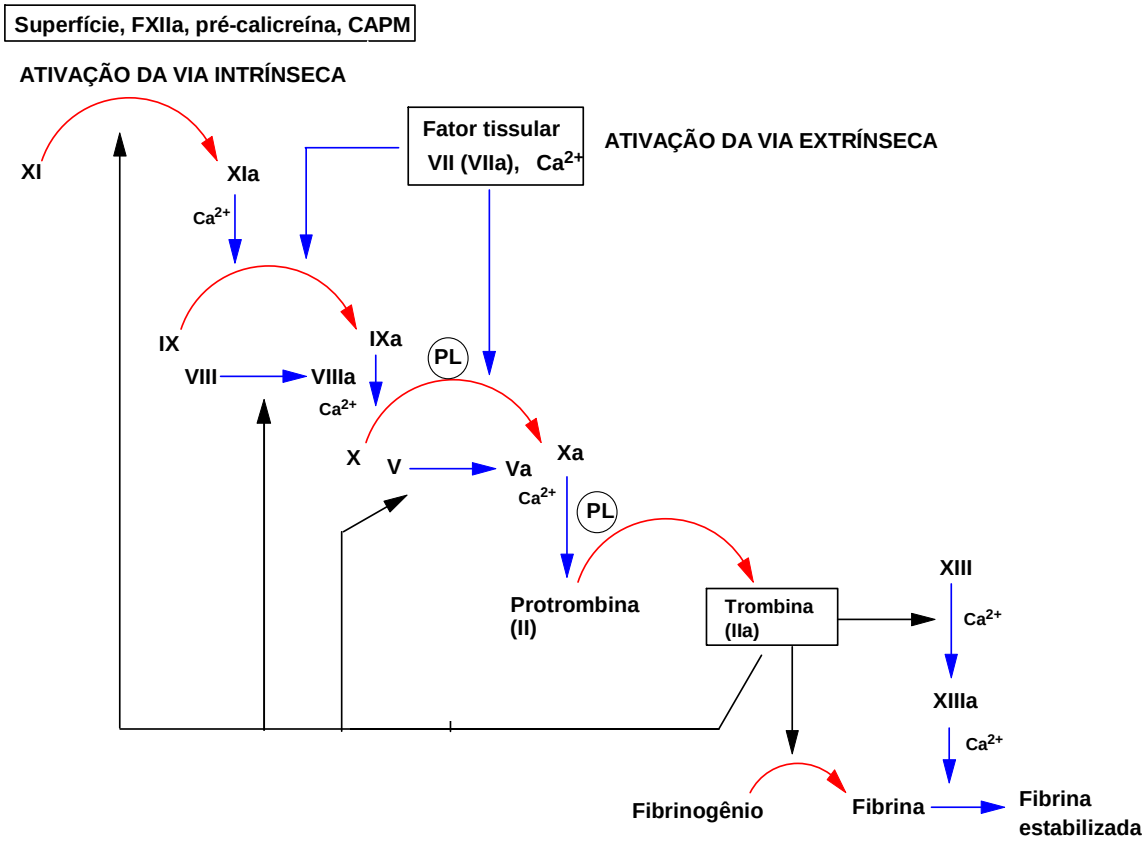

Figura 2. Esquema geral da cascata da coagulação sangüínea.

Independente da via de ativação, o fator IXa forma então um complexo (denominado complexo tenase ou *Xase*) com fator VIII ativo (FVIIIa), cálcio e um fosfolipídeo (ou fator plaquetário III). Neste complexo multimolecular, FIXa catalisa a

ativação do fator X. Portanto, as vias extrínseca e intrínseca confluem na etapa de ativação do fator X. Finalmente, o complexo protrombinase, formado pelo fator Xa, juntamente com fator V ativo (pró-acelerina), cálcio e uma superfície fosfolipídica, converte rapidamente o fator II (protrombina) em trombina (fator IIa), a última enzima da cascata da coagulação, responsável pela transformação do fibrinogênio em monômero de fibrina. Na presença de Ca^{++}, a α-trombina é também envolvida na ativação do fator XIII, gerando o fator XIIIa . O fator XIIIa estabiliza a coagulação do sangue como conseqüência da ligação cruzada realizada entre este fator e a fibrina através da formação de pontes N^{ε}–(γ-glutamil)lisina, resultando na fibrina polimerizada.

1.2.2. Nitroforina-2

Animais hematófagos valem-se de mecanismos anticoagulantes para se alimentar do sangue do hospedeiro. Estes animais – em sua maioria artrópodos, como moscas, mosquitos e barbeiros, mas que incluem sanguessugas e morcegos – contém em sua saliva inibidores de proteínas coagulantes. As abordagens para impedir a coagulação são as mais diversas, tornando os hematófagos uma boa fonte para pesquisa de drogas anticoagulantes.

A proteína nitroforina-2 (NP-2), também denominada prolixina-S, presente nas glândulas salivares do inseto barbeiro *Rhodnius prolixus*, é um inibidor específico do complexo tenase intrínseco (ZHANG *et al.*, 1998). Com 179 aminoácidos e 20 kDa, inibe fIXa ligado à superfície fosfolipídica ou fIXa ligado à fVIIIa. A partir de estudos experimentais realizados em nosso laboratório, onde foram sintetizados diversos peptídeos da seqüência de NP-2, foi detectado o fragmento com ação anticoagulante, o qual corresponde à seqüência KKAVLKEADEKNSYTLTVL, localizada entre os resíduos 90 a 110. Como pode ser observado na figura 3, esta seqüência corresponde à uma região que compreende uma alça e uma folha-β de NP-2 e foi o ponto inicial no estudo da interação entre NP-2 e fIXa.

(A) **(B)**

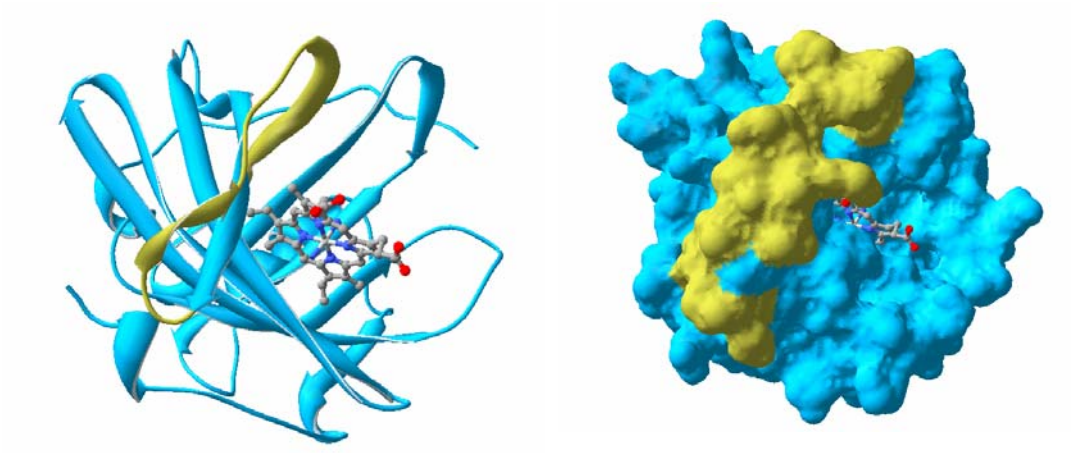

Figura 3. Representação esquemática em fitas (painel A) e da superfície acessível ao solvente (painel B) da estrutura da nitroforina-2. A região com atividade anticoagulante (resíduos 90-110) está destacada.

1.3. Serpinas

A superfamília de proteínas denominada serpinas engloba um conjunto de moléculas que apresentam seqüência e estruturas similares, encontradas em vários organismos (nematódios, vírus, insetos, plantas e vertebrados superiores, incluindo o *Homo sapiens*). (SILVERMAN *et al.*, 2001; YE &GOLDSMITH, 2001.)

Inicialmente, o acrônimo para o nome serpina ("serpin": "SERine Protease INhibitor") foi assim determinado a partir da ação funcional inibitória das proteínas pertencentes a esta família sobre serino-proteases. As proteínas da superfamília das serpinas estão ligadas à várias funções fisiológicas. Disfunções como enfisema pulmonar, cirrose, certos tipos de demência e patologias associadas à coagulação sangüínea podem decorrer de mutações na cadeia polipeptídica de serpinas.

As serpinas contém de 370 a 500 aminoácidos por seqüência, e estrutura terciária conservada, contendo 7-9 α-hélices (nomeadas de A a I) e 3 feixes de folhas-β (A, B e C).

A alça do centro reativo (RCL, *reactive center loop*), constituída de aproximadamente 17 resíduos localizados entre os feixes de folhas-β A e C (figura 4), está diretamente envolvida nos rearranjos estruturais que ocorrem nas serpinas durante a atividade inibitória. As serpinas podem ser convertidas em uma forma mais estável que a nativa, denominada latente, a partir da inserção da alça do centro reativo no meio do feixe de folhas-β A. Neste processo, ocorre a formação de uma nova folha-β antiparalela entre a fita 1 do feixe A e o RCL. Concomitantemente, a fita 1 do feixe de folhas-β C (f1C) separa-se do feixe C para formar uma alça à partir da extremidade inferior da serpina. Serpinas na forma latente não são inibitórias. O estado mais estável de uma serpina inibitória corresponde ao estado RCL-clivado. Nesta conformação, a alça central reativa encontra-se completamente inserida no feixe de folhas-β A, mas, ao contrário do estado latente, não ocorre a separação de f1C do feixe de folhas-β C.

Além do estado nativo (ativo), latente e clivado, as serpinas podem ainda ser encontradas em δ e polimérico. O estado δ é um estado intermediário entre nativo e latente, enquanto que polímeros de α-1-antitripsina indicam a polimerização da alça central reativa.

As serpinas inibem as serino-proteases por um mecanismo irreversível de inibição suicida. Inicialmente, forma-se um complexo não-covalente com a protease a partir da interação com resíduos adjacentes (P1-P1') à ligação cindível do RCL. O ataque da serina nucleofílica (Ser195) da protease sobre a ligação cindível da serpina conduz à formação de uma ligação éster entre Ser195 e a carbonila da ligação peptídica do resíduo em P1. Como resultado, a ligação peptídica é clivada. Eliminadas as restrições espaciais, o RCL inicia o processo de inserção no feixe A. Depois de uma completa inserção da alça do centro reativo, a protease é translocada por cerca de 70 Å, sendo seu sitio ativo distorcido.

Os arranjos estruturais aos quais as serpinas estão sujeitas envolvem algumas regiões importantes: a) “hinge” (dobradiça), o qual fornece mobilidade à alça central reativa; b) “breach” (lacuna), localizado no topo da folha-β A, é o ponto de inserção da alça central; c) “shutter” (cortina), situado perto do centro da folha-β A, o qual, em

conjunto com o “shutter”, facilita a abertura da folha-β A para a inserção da alça do centro reativo; d) “gate” (portão), envolve as fitas 3C e 4C (fitas 3 e 4 do feixe de folhas-b C), este arranjo estrutural está presente no estado nativo, servindo para inserir a alça reativa no feixe A sem clivagem.

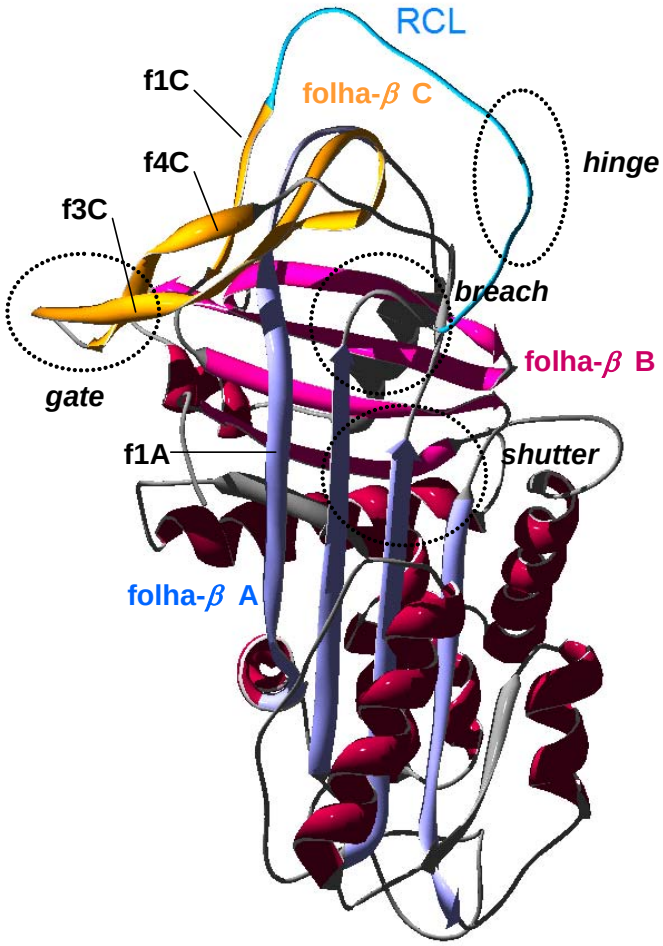

Figura 4. Representação esquemática em fitas de uma serpina com destaque para as principais regiões.

Entretanto, nem todas as serpinas apresentam atividade inibitória de serino-proteases. Evolutivamente, muitas perderam esta função, passando a ter as mais diversas atividades: são encontradas serpinas transportadoras de hormônio (como a globulina

ligadora de cortisol, ou CBG), chaperonas (proteína de choque térmico de 47-kD, ou HSP47), e de estoque (ovalbumina). Proteínas como a ovalbumina apresentam extremidade C e N mais curtas do que α-1-antitripsina, e são deficientes no peptídeo sinal. Além disso, a alça do centro reativo é substituída por uma α-hélice.

Atualmente, os repositórios de seqüências e de famílias de proteínas como Pfam (SONNHAMMER *et al.*, 1997) e PROSITE (Bucher &Bairoch, 1994), não classificam as serpinas por sua função. Estas continuam agrupadas no formato de superfamília, mesmo quando suas funções são tão diversificadas, e muitas proteínas já perderam a função-título de inibição.

1.4. Bioinformática Estrutural

A explosão de dados genômicos (BENSON *et al.*, 2005) e a resolução crescente de estruturas tridimensionais impulsionaram a área da Bioinformática, e de sua extensão, a Bioinformática Estrutural. A Bioinformática surgiu como uma junção da Biologia Molecular e da Ciência da Computação, para tratar a grande quantidade de dados seqüenciados. Sua tarefa inicial foi de classificar, automatizar, buscar informação dentro do um vasto conjunto de bancos de dados públicos (ALTSCHUL, *et al.*, 1997; PAGNI *et al.*, 2001). Ultimamente tem sido usada para inferir função, criar modelos estatísticos e achar padrões; tal como alinhamento múltiplos (PEARSON &LIPMAN, 1988; THOMPSON *et al.*, 1994), modelos por homologia, simulação de sistemas e previsão de seu comportamento.

A seqüência de aminoácidos que forma uma proteína é denominada de estrutura primária. No espaço tridimensional, tais aminoácidos se enovelam formando uma estrutura tridimensional definida e estável no ambiente fisiológico (processo também conhecido como "folding"). A conformação local do esqueleto ("backbone") protéico é chamada de estrutura secundária, sendo as α-hélices e as folhas-β as formas mais conhecidas. O empacotamento do conjunto das estruturas secundárias e segmentos de conformação randômica forma domínios modulares, definindo a chamada estrutura

terciária. A associação de subunidades terciárias em uma proteína forma sua estrutura quaternária. A importância do estudo destes dois últimos níveis envolve a descoberta de interações entre aminoácidos que se encontravam separados na estrutura primária.

A descoberta das estruturas tridimensionais requer métodos de espectroscopia vibracional, como cristalografia de raios-X e RMN. Tais técnicas apresentam limitações. Na cristalografia, por exemplo, é preciso primeiramente obter um cristal da proteína, para que depois a localização de seus átomos seja determinada pela difração de raios-X. Neste processo, a determinação de condições para cristalização pode ser muito custosa e demorada. Algumas proteínas, como as de membrana, dificilmente cristalizam na forma de cristais ordenados.

Métodos de Bioinformática podem superar algumas destas limitações, sendo que um de seus principais objetivos é prever o comportamento de sistemas quanto à estrutura/função. Também podem ser aplicados na predição da estrutura terciária e no processo de enovelamento de proteínas. Pode-se ainda empregar esta abordagem no planejamento e modelagem de fármacos seletivos a partir da identificação e análise de sítios-ativos e sítios de ligação (DAURA *et al.*, 2000), "docking" e desenho de ligantes (RUSSO *et al.*, 2002; RUSSO *et al.*, 2004), análise de interações proteína-proteína (SOUZA &ORNSTEIN, 1999) e simulação de sistemas biológicos. Estes métodos são um auxílio à forma tradicional de descoberta de fármacos, onde vários compostos aleatórios são purificados e testados em laboratório em um processo de tentativa e erro muito demorada. Por este motivo, a Bioinformática Estrutural está sendo impulsionada por investimentos provenientes de companhias farmacêuticas.

A quantidade de estruturas determinadas em relação à quantidade de proteínas existentes é muito pequena. Entretanto, o número de padrões de enovelamento exclusivos ("templates", ou diferentes tipos de "fold") encontrados é pequeno (estimado em 1000-2000 em GRANT & RICHARDS, 1995). Esta convergência em número e formato das estruturas é uma pista para aquelas de estrutura desconhecida. A Bioinformática traz métodos, como a modelagem estatística envolvendo estruturas conhecidas e buscando padrões comuns entre elas (modelos ocultos de Markov ou homologia); ou modelagem de

estruturas terciárias e aplicações de forças físicas influindo no resultado (modelagem e dinâmica molecular). Além disso, possibilita a análise de imagens biológicas, criação de mapas de vias metabólicas, mapeamento de funções filogenéticas. Estes são exemplos de como recursos computacionais podem ser aplicados no para expandir limitações de técnicas de espectroscopia.

2. OBJETIVOS

2.1. Objetivos Gerais

Análise de potenciais alvos farmacológicos e de seus inibidores visando o planejamento racional de novos fármacos anticoagulantes.

Utilização de técnicas de bioinformática e de recursos computacionais para identificação de relações seqüência-estrutura-função na família de serpinas.

2.2. Objetivos Específicos

Esta dissertação é composta de dois estudos, especificados a seguir.

A) Análise computacional da interação entre a nitrofronina-2 (NP-2) e o complexo tenase intrínseco, visando o desenvolvimento de fármacos anticoagulantes;

B) Modelagem da superfamília de proteínas serpinas visando a classificação destas em 4 subfamílias:
    a. serpinas com atividade inespecífica;
    b. serpinas com atividade inibitória inespecífica;
    c. serpinas com atividade inibitória específica contra enzimas envolvidas na cascata da coagulação sangüínea;
    d. proteínas classificadas como serpinas mas que não inibem serino-proteases;

3. MÉTODOS

3.1. Mecânica Molecular

Descrições detalhadas de átomos podem ser obtidas a partir da mecânica quântica. Esta faz uso de descrições rigorosas de moléculas do ponto de vista estrutural e eletrônico, cujo tempo computacional para análise pode ser proibitivo (principalmente no caso de macromoléculas). Por outro lado, o campo da espectroscopia vibracional possibilita o emprego de funções de energia potencial para descrever o comportamento geral da molécula. Assim, a representação do movimento das moléculas é dada em termos de campos de forças clássicos relacionados com métodos de mecânica molecular.

Na área da mecânica molecular, o átomo é tratado como uma esfera. Considera-se uma distribuição eletrônica fixa associada a cada átomo. Assim, núcleo e elétrons são condensados numa partícula única. Uma molécula é representada como uma coleção de esferas, ligadas por molas. A vibração da mola pode ser descrita pela constante de Hook ($H = kx^2/2$, sendo k a constante de Hook, e x a distância ou comprimento da ligação), a qual está relacionada com a energia de ligação.

O movimento destes átomos pode ser descrito por leis da física clássica e por simples funções de energia potencial. Assim, a complexidade dos cálculos de energia é reduzida, sendo possível trabalhar com um sistema de até 10^5 átomos.

3.1.1. Função de Energia

A função de energia é a base do campo de forças, usada para descrever o conjunto de esferas em vibração. A energia de uma molécula é calculada pela soma de interações ligadas e não ligadas:

$$E_{tot} = E_l + E_\theta + E_\omega + E_{nc}$$

onde E_l , E_θ, E_ω e E_{nc} são, respectivamente, comprimento da ligação, ângulo, diedro e forças não-covalentes, conforme descritos a seguir.

- Termos para átomos ligados. Incluem o comprimento da ligação, descrito pela função de Morse:

$$E_l = \Sigma \ k_l (l - l_0)^2$$

onde k_l é a constante de elasticidade definindo a vibração da ligação e l_0 o valor de equilíbrio para a ligação. O ângulo entre as ligações pode ser descrito da mesma forma:

$$E_\theta = \Sigma \ k_\theta (\theta - \theta_0)^2$$

sendo k_θ a constante de força, e θ_0 o valor de equilíbrio para o ângulo. Diedrais são ângulos que devem permitir uma rotação ao longo do eixo da molécula, e descritos como:

$$E_\omega = \Sigma \ V_m (1 + s \cos n\omega)$$

onde é a barreira rotacional, m a periodicidade das rotações, e s =1 ou -1.

- Termos para interação entre átomos não ligados: representam todos os pares de átomos não ligados diretamente, para os termos atrativo e repulsivo de van der Waals, e para a interação coulômbica, onde

$$E_{nc} = E_{vdw} + E_c$$

A interação de van der Waals (energia de dispersão eletrônica) pode ser descrita por um termo atrativo-repulsivo, como o potencial de Lennard-Jones:

$$E_{vdw} = \Sigma \ \varepsilon [(A/r^{12}) - (B/r^6)]$$

onde A e B são constantes, ε a constante dielétrica do meio, r^{-12} representa o termo de repulsão e r^{-6} as forças de London de dispersão-atração. A energia coulômbica é geralmente modelada com cargas pontuais, o que resulta em um potencial eletrostático, localizado a um ponto relativo ao átomo. Pode ser descrita pela lei de Coulomb,

$$E_c = \Sigma \ q_i q_j / Dr_{ij}$$

onde D é a constante dielétrica do meio; q_i , q_j , as cargas dos átomos e r_{ij} , a distância entre elas.

3.1.2. Cálculo de Mecânica Molecular

Para o cálculo de mecânica molecular são necessárias as coordenadas e a topologia do sistema. Coordenadas fornecem a descrição de como os átomos são organizados espacialmente. As coordenadas são fornecidas por métodos experimentais, como cristalografia de raios-X ou ressonância magnética nuclear. Um exemplo de formato de coordenadas são os arquivos PDB. Após uma série de informações sobre sistema, autores e métodos da cristalografia, o PDB apresenta os dados da geometria dos átomos. A figura 5 mostra um exemplo deste formato.

```
ATOM      1  N   VAL A  16      18.541  22.060  59.057  1.00 15.10      N
ATOM      2  CA  VAL A  16      20.010  22.030  59.296  1.00 16.28      C
ATOM      3  C   VAL A  16      20.329  21.945  60.787  1.00 17.16      C
ATOM      4  O   VAL A  16      20.024  22.853  61.570  1.00 15.87      O
ATOM      5  CB  VAL A  16      20.706  23.285  58.712  1.00 14.93      C
ATOM      6  CG1 VAL A  16      22.202  23.181  58.897  1.00 12.95      C
ATOM      7  CG2 VAL A  16      20.362  23.436  57.246  1.00 15.28      C
ATOM      8  N   VAL A  17      20.940  20.832  61.170  1.00 17.75      N
ATOM      9  CA  VAL A  17      21.332  20.600  62.544  1.00 17.18      C
ATOM     10  C   VAL A  17      22.753  21.114  62.648  1.00 18.17      C
ATOM     11  O   VAL A  17      23.531  20.954  61.716  1.00 19.54      O
```

Figura 5. Segmento de um arquivo de coordenadas do fator IXa (entrada PDB 1RFN).

Após a palavra-chave “ATOM”, os campos indicam: número do átomo; tipo do átomo (seguindo formato específico do PDB); nome do aminoácido; indicador da cadeia; posição do aminoácido na seqüência; coordenada do átomo nos eixos x, y e z do plano cartesiano (em Å), ocupação, fator de temperatura e símbolo do elemento. Caso as coordenadas não estejam disponíveis, pode-se aplicar métodos de modelagem *ab initio* ou homologia.

A topologia compreende um conjunto de informações de conectividade entre os átomos. Normalmente, é gerada a partir das coordenadas, e varia dependendo do programa de simulação.

3.1.3. Campo de Forças

Assim que as coordenadas e a função de energia (descrita acima) forem definidas, a primeira e segunda derivadas parciais da função de energia (com respeito às coordenadas atômicas) vão originar as forças que agem em cada átomo, ou seja, o campo de forças. Algoritmos baseados em tal tipo de tratamento podem ser empregados para localizar mínimos energeticamente favoráveis (mais estáveis) para os átomos.

A parametrização dos termos da função de energia é, em geral, baseada em dados experimentais (podendo também ser estimada *ab initio*). As constantes de força são provenientes de dados de espectroscopia, geometrias obtidas de cristais com alta resolução e ressonância nuclear.

Nas simulações realizadas neste trabalho foi empregado o campo de forças GROMOS 96 43A1 (VAN GUNSTEREN *et al.*,1996), incluído no pacote de programas GROMACS.

3.1.4. Minimização de Energia

A energia do sistema é medida em função dos graus de liberdade entre ligações covalentes, ângulos e ângulos diedro (os termos energéticos levados em conta são coulômbico, van der Waals, ligações covalentes, ângulos diedro próprios e impróprios). Uma busca por conformações é feita para encontrar todas as conformações energeticamente favoráveis de uma molécula (GRANT & RICHARDS, 1995).

Para obter geometrias mais confiáveis, deve-se minimizar a energia do sistema. Para isto, otimizam-se as posições atômicas sujeitas à forças do campo de forças usado. Este passo envolve o cálculo da derivada da energia potencial. A condição para o mínimo de uma curva (ponto x_{min}) é que a primeira derivada seja zero:

$f'(x_{min}) = 0$

O mínimo energético é estabelecido mediante a aplicação de um gradiente até que estados conformacionais apresentem uma mudança pequena ou nula de energia. O gradiente fornece a melhor indicação de convergência. Um dos métodos de minimização de energia utilizado é o “Steepest Descent” (SD). Este método é originado apenas por constantes de força ao longo da superfície de potencial. Se imaginada a superfície de energia como uma paisagem em relevo (figura 6), a forma mais eficaz de encontrar um “vale” (mínimo de energia), é seguir o gradiente descendente. Por exemplo, quando aplicado à uma função unidimensional, o método fica na forma iterativa:

$$x_i = x_{i-1} - \in f'(x_{i-1})$$

para um ponto de partida x_0 e para algum $\in > 0$, até que um ponto fixo seja atingido. As coordenadas são ajustadas a cada passo, na direção negativa do gradiente. Se a energia diminui, o tamanho do passo é aumentado em 20% para acelerar a convergência.

Deve-se observar a diferença entre mínimo local e mínimo global. Dado que existem três graus de liberdade por átomo, uma molécula de N átomos terá quase 3N variáveis a serem minimizadas. O cálculo do mínimo global é um problema matemático e que não pode ser solucionado para o mínimo de energia.

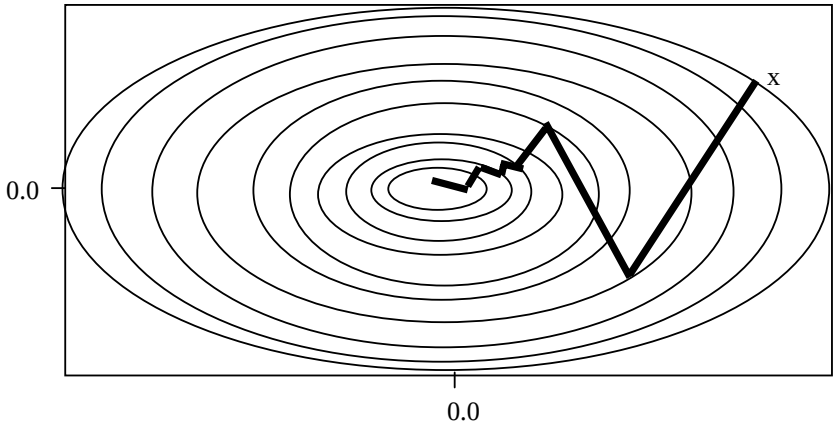

Figura 6. Esquema representando a trajetória de minimização para uma superfície de energia usando o método SD.

3.2. Dinâmica Molecular

Conforme mencionado anteriormente, pelo tratamento da mecânica clássica os sistemas moleculares são descritos como esferas rígidas com ligações rotacionáveis. Nos sistemas reais, as moléculas não são completamente rígidas, e nem se encontram no mínimo energético. Ligações (e ângulos) encontram-se em constante vibração. Por este motivo, resultados das técnicas de espectroscopia fornecem apenas uma média das posições atômicas. O estudo do processo dinâmico de sistemas moleculares é o caminho para a compreensão de processos complexos (KARPLUS, 2003). Este método teórico pode ser usado para avaliar a importância dos movimentos moleculares sobre a função biológica. Pode ser aplicado no estudo da flutuação da acessibilidade de sítio ativos (abertura e fechamento), flexibilidade, e mudança de conformação induzida na ativação de proteínas como as serpinas.

A partir da dinâmica molecular, pode-se fazer a análise do comportamento dinâmico de um sistema. Em simulações de DM, o cálculo do movimento de moléculas individuais é feito a partir da observação da variação temporal de posições, velocidades e

orientações dos componentes do sistema. Tais dados podem esclarecer as propriedades específicas de um sistema mais facilmente que experimentos. Assim, é possível conectar informações termodinâmicas com as interações atômicas envolvidas.

3.2.1. As equações de movimento

Dado o cálculo de forças nos átomos de uma molécula, o próximo passo usa tais forças na equação de Newton para calcular a dinâmica.

Seja um sistema de coordenadas com N átomos, o objetivo da DM é assinalar velocidades iniciais e/ou temperatura uniforme conhecida, em um número finito de passos para cálculo da força resultante em cada átomo em tempo $t+\Delta t$. Conforme a segunda lei de Newton,

$$F = m\,a$$

onde F é a força agindo sobre uma partícula, m é sua massa e a a aceleração. Esta equação pode ser resolvida para cada átomo, obtendo-se suas novas velocidades e posições. Se for possível calcular a próxima configuração do conjunto de partículas, então se tem um método que traça a evolução do sistema ao longo do tempo.

Reorganizando a equação acima, colocando a aceleração como segunda derivada da posição (d) em função do tempo, $\delta^2 d/\delta t^2$, tem-se:

$$\delta^2 d/\delta t^2 = F/m$$

Para simulação do comportamento do sistema, deve-se resolver esta equação para cada partícula. Integrando em relação ao tempo, tem-se:

$$\delta d/\delta t = (F/m)t + C_1$$

Para tempo zero ($t = 0$), o termo F/m desaparece ($F/m = 0$). Então,

$$\delta d/\delta t = C_1$$

E pela variação da distância sobre variação de tempo tem-se a velocidade. Logo, $C_1 = v$. A velocidade inicial (v_0) pode ser dada então por

$$\delta d/\delta t = a\ t + v_0$$

para cada tempo t.

Integrando novamente em relação ao tempo t resulta em

$$d = v_0\ t\ +\ ½\ a\ t^2 +\ C_2$$

onde a constante é a nova posição. Então, pode-se calcular a distância a partir da velocidade inicial v_0 e a aceleração ($a = F/m$).

Assim, variações sucessivas de conformação de uma molécula são geradas no formato de uma trajetória. O número de passos da DM representa o número de avaliações de energia. Sendo um método estatístico, a trajetória gerada deve convergir para uma propriedade específica (mínimo de energia ou determinada conformação).

A dinâmica molecular pode ser gerada por pacotes específicos. Alguns exemplos mais usados atualmente são CHARMM (Chemistry at Harvard Macromolecular Mechanics, disponível em http://www.charmm.org/), (BROOKS *et al.*,1983), Amber (Assisted Model Building with Energy Refinement, http://amber.scripps.edu/) (PEARLMAN *et al.*,1995) e Gromacs (http://www.gromacs.org/). Em geral, estes pacotes realizam a DM segundo o processo:

1) Inicialização: fornece posição inicial e velocidades iniciais para os átomos;
2) Termalização: aumento da energia cinética do sistema;

3) Ambientação: a energia cinética e a potencial são uniformemente distribuídas ao longo do sistema, quando a temperatura média se estabiliza;
4) Produção: nesta etapa são coletados e analisados os dados das trajetórias.

3.2.2. Tratamento do solvente

A maioria dos sistemas biológicos encontra-se envolto em solvente. Por isto, é necessária uma representação do ambiente em solução aquosa. Não é possível determinar experimentalmente se flutuações de solvente influenciam o movimento interno da proteína. Porém, a inclusão do solvente nos cálculos teóricos é uma forma apropriada para descrever efeitos como pontes de hidrogênio envolvendo átomos do soluto e do solvente; efeitos de cargas; energia de solvatação e custo entrópico.

Uma das formas de representação do solvente é chamada de “explícita”, onde cada molécula de água (ou de solvente orgânico) é definida individualmente, e suas interações incluídas nos parâmetros da mecânica molecular. (O método implícito inclui uma constante dielétrica e cálculos de eletrostática ao longo do sistema.)

A modelagem do solvente requer uso de parâmetros disponíveis em modelos como TIP3 (“Transferable Interaction Potential Three Point”), SPC (“Single Point Charge”) e SPC/E (“Single Point Charge Extended”, usado neste trabalho). As diferenças dos modelos estão nos parâmetros Lennard-Jones, cargas e geometrias internas. Em geral, fornecem uma boa descrição para os sistemas, dadas as limitações computacionais.

Os limites são em relação à quantidade de moléculas de solvente que se pode simular, e isto é resolvido com o artifício das condições periódicas de contorno. (Para isso, usa-se um sistema retangular, ou de outra geometria escolhida, envolto por cópias de si mesmo. Assim, as moléculas da borda do retângulo interagem com as moléculas de sua imagem na cópia vizinha, ao invés de interagirem com vácuo.)

3.2.3. Análise da trajetória

Após o cálculo da DM e geração da trajetória, obtém-se uma coleção de estruturas ao longo do tempo. Informações podem ser extraídas destes dados, como a diferença entre as estruturas iniciais e finais, ou uma estrutura média sobre toda a trajetória. A superfície acessível ao solvente pode ser obtida com a varredura da superfície com uma sonda de uma molécula hipotética, como água, de raio 1.4 Å. Pode-se também calcular as pontes de hidrogênio, para átomos que estiverem a uma distância < 2.4 Å e ângulo de ~130°.

Além disso, pode-se testar o quão estável (logo, próxima do real) a simulação de DM é. Para isso, deve-se observar alguns fatores, como:

- Estabilidade da temperatura. Fatores de temperatura ("B-factors") indicam a vibração térmica de um átomo na estrutura cristalográfica. A diferença entre os fatores de temperatura experimentais dos da simulação podem resultar em uma trajetória não-equilibrada. Concordância entre os fatores-B indica uma simulação estável;
- Inspeção do gráfico de Ramachandran para observar propriedades estruturais. A análise deve retornar ângulos ϕ e ψ que adotem valores restritos.

3.3. Modelos Ocultos de Markov

Conhecido como HMM (do nome original "Hidden Markov Models"), este método estatístico é usado para classificação de dados e descoberta de padrões. O estudo de HMM iniciou nos anos 70, na área da ciência da computação, sendo aplicado, por exemplo, em reconhecimento de fala. Com a explosão de dados genômicos, a biologia se utilizou deste método para modelar seqüências de proteínas e de DNA. (DURBIN *et al.*, 1998, KROGH *et al.*, 1994). Desta forma, passou a ser usado para pesquisa de seqüências homólogas, ou modelagem de famílias de seqüências (TRUONG &IKURA, 2002; KARPLUS *et al.*, 1998, KARCHIN *et al.*, 2002).

Para sua implementação, usa-se um conjunto de estados e de transições entre eles, como num grafo (estrutura de representação comum na matemática e ciência da computação). Para um conjunto de seqüências (por exemplo, uma família de proteínas) que se deseja modelar, passa-se pelo HMM cada uma delas.

Inicialmente, cada posição da seqüência de aminoácidos irá gerar um estado no HMM. Cada estado contém uma tabela com probabilidades para cada aminoácido na seqüência aparecer naquela posição (no caso do DNA, para cada base nucleotídica). Esta tabela é desconhecida do usuário, o que dá ao método seu nome de oculto. Novas seqüências passadas ao HMM irão influenciar as tabelas em cada estado, e diferenças entre elas irão criar estados adicionais para acomodá-las.

Os estados são definidos como “match”, inserção e deleção. O estado “match” contém aminoácidos consenso, ou seja, que aparecem em todas ou em grande maioria das proteínas daquele conjunto alinhado. Cada um destes estados vai conter a tabela com a distribuição de probabilidade para os aminoácidos que ali aparecem. Esta tabela é recalculada, tendo suas probabilidades modificadas, a cada nova seqüência que atravessa o HMM.

Os estados de inserção incluem aminoácidos que podem aparecer entre colunas do consenso. Biologicamente, isto representa a variação natural que existe entre as seqüências. Analogamente, os estados de deleção representam “gaps” que apareceram evolutivamente entre nas seqüências, onde aminoácidos entre colunas foram deletados.

Para computar a probabilidade de uma seqüência inteira, deve-se considerar um caminho pelos estados do HMM. Cada vez que um caminho passa por um estado, um aminoácido da seqüência é usado, e sua probabilidade é recalculada pela tabela do estado.

A forma do grafo HMM, com seus estados representando um consenso para um modelo de proteínas é determinado por sua arquitetura. (Figura 7). As arquiteturas mais comuns são a linear e a circular (esta última conhecida como “wheel”). Outras arquiteturas podem ser obtidas a partir da combinação das primeiras (como por exemplo, a paralela),

mas não são muito utilizadas. A arquitetura Linear estabelece um grafo cujo caminhamento é sempre para adiante, de um estado "match" para o próximo, e assim por diante. É recomendada para estudos com proteínas, já que suas estruturas primárias seguem um sentido linear (enquanto que, para alguns ácidos nucléicos, como RNA, a necessidade de repetições, "gaps" ou "hairpins" demanda o uso de uma arquitetura circular).

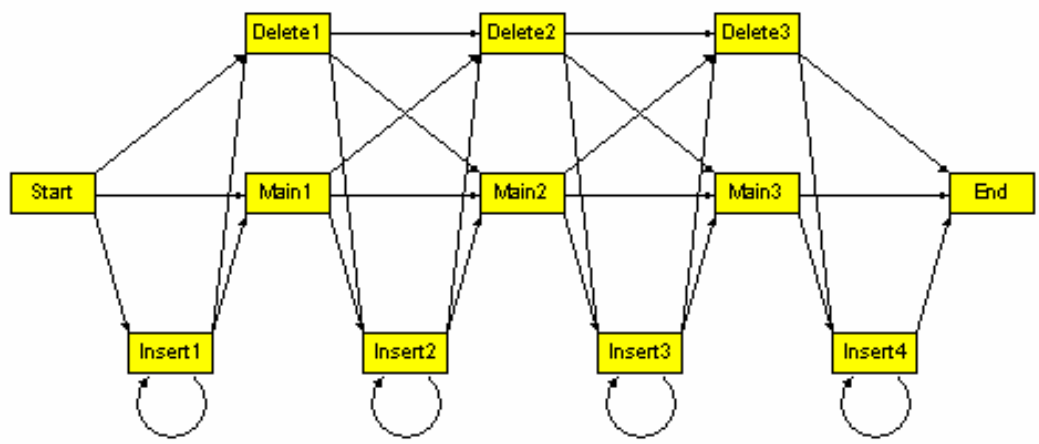

Figura 7. Esquema apresentando a arquitetura linear.

3.3.1. Algoritmos

Existem alguns algoritmos para traçar caminhos pelo grafo. O algoritmo de Viterbi gera alinhamentos, percorrendo o grafo e obtendo o caminho cujos estados apresentavam maior probabilidade. Outros algoritmos, como o que usa o método de Baum-Welch (ou Enhancement Maximization), adicionam probabilidades através de todos os caminhos.

Exemplos de pacotes para geração de HMMs incluem o HMMpro (desenvolvido pelo Net-ID e disponível em http://www.netid.com/html/hmmpro.html), HMMER (DURBIN *et al.*, 1998) disponível em http://hmmer.wustl.edu/) e SAM (Sequence Alignment and Modeling Sysem, http://www.cse.ucsc.edu/compbio/sam.html), sendo que os dois primeiros foram utilizados neste trabalho (KARPLUS *et al.*,1997, McCLURE *et al.*, 1996).

4. RESULTADOS E DISCUSSÃO

4.1 Interação fator IXa – nitroforina-2

A modelagem molecular foi empregada no estudo da interação entre a serino-protease fIXa com um polipeptídeo ativo derivado do inibidor (nitroforina-2). Foi possível identificar candidatos a sítios de interação localizados no módulo catalítico de fIXa humano pelo reconhecimento e caracterização de cavidades. A análise da interface entre proteína e peptídeo revelou a existência de complementaridade eletrostática e geométrica. Os aspectos de complementaridade serviram de base para que o fragmento ativo de NP-2 fosse manualmente orientado no sitio de ligação com fIXa.

As simulações de dinâmica molecular foram estáveis, conforme ficou evidenciado após a análise das trajetórias de energia e dos parâmetros estruturais. Os dados da simulação do polipeptídeo nativo orientaram a escolha de substituições de resíduos a fim de obter-se interações mais efetivas com fXIa.

O sistema cujos dados apontaram maior afinidade é caracterizado pela menor distância entre os centros de massas e pela maior quantidade de ligações de hidrogênio entre o polipeptídeo e a proteína. Assim, foi obtida uma nova seqüência, teoricamente com maior potencial anticoagulante: LKEADE. Este trabalho necessita de comprovação de bancada experimental, onde as seqüências simuladas deverão ser sintetizadas e testadas quanto à atividade inibitória. Detalhes de resultados e discussão deste trabalho estão apresentados na forma do artigo a seguir: "MOLECULAR MODELING STUDIES OF AN ANTICOAGULANT PEPTIDE DERIVED FROM NITROPHORIN-2 IN THE ACTIVE SITE OF FACTOR IXa: PERSPECTIVES TO DRUG DESIGN" C. Russo, A. F. M. Pinto, M. A. Juliano, H. L. N. de Amorim, J. A. Guimarães. January 2004. *Proceedings of the winter international synposium on Information and communication technologies WISICT '04*. Publisher: Trinity College Dublin.

O trabalho cuja citação encontra-se acima, foi apresentado no WISICT (*Winter International Seminar on Information Technology*), na sessão Systeomatics (*Dynamic Biological Systems Informatics*) e está publicado no ACM (Association of Computer Machinery) digital library: http://portal.acm.org/, disponível em "ACM International Conference Proceeding Series; Vol. 58. Proceedings of the winter international synposium on Information and communication technologies"

MOLECULAR MODELING STUDIES OF AN ANTICOAGULANT PEPTIDE DERIVED FROM NITROPHORIN-2 IN THE ACTIVE SITE OF FACTOR IXa: PERSPECTIVES TO DRUG DESIGN

Russo, C.[1]; Pinto, A.F.M.[1]; Juliano, M.A.[2].; de Amorim, H.L.N.[1,3]; Guimarães; J.A.[1]
[1]Centro de Biotecnologia - UFRGS, Porto Alegre/RS; [2]Depto. de Biofísica - UNIFESP/SP;
[3]Depto. de Química - ULBRA ,Canoas/RS

ABSTRACT

The extrinsic tenase complex is insufficient to sustain hemostasis because tissue factor pathway inhibitor rapidly inactivates TF-bound FVIIa. To overcome this limitation, the intrinsic tenase, a phospholipid membrane bound complex of factor VIIIa (FVIIIa) and factor IXa (FIXa), must be activated. The critical role of intrinsic tenase makes this enzyme complex an attractive target for inhibition. Nitrophorin-2 (NP-2), present in the salivary glands of the blood sucking bug *Rhodnius prolixus*, is a specific inhibitor of intrinsic tenase complex. In previous study, peptides from nitrophorin-2 sequence were synthesized and the anticoagulant activity determined. One of these peptides was found exhibiting anticoagulant activity against the intrinsic tenase complex. As a first approach, the location of FIXa binding site was determined by describing its pockets. The orientation and conformation of the active NP-2 peptide was adjusted using a combination of rigid docking and flexible geometry optimization with GROMOS96 force field. The docking was performed by three molecular dynamics (MD) simulations with different NP-2 peptides. One MD simulation was carried out with free FIXa, for reasons of comparison. The complex stability was evaluated based on analysis of MD trajectories. The results showed that the sequence LKEADE, which is common to the three peptides in each simulation, seems to be the basis of interaction on the FIXa active site. New studies will be carried out using the structure LKEADE as a starting point for drug design.

INTRODUCTION

The blood coagulation cascade acts in order to stop the loss of blood that follows vascular injury. It starts with the exposition of blood to the tissue factor (TF) on the wound. Then, extrinsic and/or intrinsic pathways are activated. Both pathways involve a series of coagulation factors circulating in the plasma, which function as a cascade.

The extrinsic tenase complex, a phospholipid membrane-bound complex of tissue factor and factor VIIa (FVIIa), is insufficient to sustain hemostasis because tissue factor pathway inhibitor rapidly inactivates TF-bound FVIIa. To overcome this limitation, the intrinsic tenase, a phospholipid membrane-bound complex of factor VIIIa (FVIIIa) and factor IXa (FIXa), must be activated. Coagulation factor VIIIa (FVIIIa) binds to the serine proteinase FIXa, to ion Ca^{+2} and to a phospholipid layer (PL), in order to assemble the FX activation (FXase) complex [1]. It releases FXa, which associates to FVa in the presence of PL, forming the prothrombinase complex. Then, the prothrombinase complex activates thrombin, the last coagulation cascade enzyme. Factor X activation is crucial to hemostasis due to its intersection between intrinsic and extrinsic pathways. The critical role of intrinsic tenase makes this enzyme complex an attractive target for inhibition.

Factor IX is a single chain of 415 amino acids. It is activated by FXIa or by TF and FVIIa, in the presence of Ca^{+2}. By the time it is activated, two peptide bonds are cleaved, forming the serine protease FIXa and releasing an activation peptide [2]. The serino proteinase domain of FIXa contains the catalytic triad His221, Asp269, Ser195.

Blood-sucking insets have developed mechanisms to interfere in the host's blood coagulation. From the salivary glands of the blood sucking bug *Rhodnius prolixus* was isolated the protein nitrophorin-2 (NP-2), which shows anticoagulant properties. Zhang and coworkers showed that NP-2 acts on the three major components of the intrinsic tenase complex. It inhibits FIXa bound to PL, to activated platelet surface or to FVIIIa [3]. Thus, it interferes on the assembly of intrinsic tenase complex. The complete inhibition mechanisms have not been yet determined.

In previous study, twenty peptides from nitrophorin-2 sequence were synthesized and the anticoagulant activity determined. One of these peptides was found exhibiting anticoagulant activity against the intrinsic tenase complex. This work presents an analysis of FIXa binding pockets along with its interaction to nitrophorin-2 in a dynamical way. The FIXa active binding pockets were calculated. Also, NP-2 inhibitory peptide was

mutated: it had its orientation and conformation adjusted. So, we present the results of four molecular dynamics (MD) simulations. Three MD simulations present a binding between FIXa and NP-2 peptide, and one, free FIXa. From the NP-2 peptide simulations, one was carried out using the native peptide, while the other two, mutated peptides. We describe the effects of the peptide models. The simulation data is expected to facilitate the design of high affinity selective inhibitors of FIXa.

METHODS

As a first approach, the location of FIXa binding site was determined by searching for the protein pockets. The pockets and cavities were calculated by computational geometry methods based on alpha shape and discrete flow theory (implemented in CASTp, [4]). Besides, the pockets were mapped according to electrostatic potential. Three pockets (named here as Pocket 1, Pocket 2 and Pocket 3) of FIXa were then identified as components of its active site.

We have modeled two other peptides based on NP-2 peptide, including mutations on both. The original peptide, called FIXpA, has the sequence EAVLKEADEK. Two mutated peptides, FIXpB and FIXpC, with the sequences GAVLKEADE and KAVLKEADEK, respectively, were designed. The orientation and conformation of the three active peptides were adjusted using a combination of rigid docking and flexible geometry optimization. The manual examinations of the molecular structures were done using the molecular modeling software SPDBV [5] (available at http://www.expasy.org/spdbv/). Then, the peptide models were submitted to 200 steps of steepest descent energy minimization (EM) in vacuum. After, a visual check to detect eventual steric clashes was made.

The areas and volumes of the peptides were measured. Also, the area and volumes of the three protein pockets combined were measured, as seen in Table 1. This was done in order to obtain a possible fit between the peptide and the binding pockets.

	Area (Å^2)	Volume (Å^3)
LKEADE **FIXpA**	507	478
LKEADE **FIXpB**	583	608
LKEADE **FIXpC**	594	624
Pocket 1	324	362
Pocket 2	56	41
Pocket 3	84	121
∑ Pockets (1, 2, 3)	464	524

Table 1. Calculations of area and volume of the pockets and cavities from the protein active site compared to the area and volume of the peptides interacting with them.

The NP-2 peptide presenting anticoagulant activity showed to be steric and electrostatic complementary to the protein pockets. Also, volume measurements of the peptide are compatible to those of the combination of Pocket 1, Pocket 2 and Pocket 3. The docking between FIXa and the peptide was performed by different molecular dynamic simulations. To establish comparisons between the nitrophorin-2 peptides, we have carried four MDs of: (i) free factor IX a (FIXa); (ii) FIXa complexed with peptide A (FIXpA), (iii) FIXa complexed with peptide B (FIXpB) and (iv) FIXa complexed with peptide C (FIXpC). All four 1.2 ns simulations where performed using the Gromacs package (available at http://www.gromacs.org).

The starting coordinates of FIXa were obtained from the entry 1RFN of Protein Data Bank (PDB). The starting conformation used to model the peptide was taken from entry 1EUO of PDB. The simulations used GROMOS96 43A1 force field [6], calculated in the NPT ensemble at 310 K with Berendsen temperature coupling and constant pressure (1 atm). Bond lengths were constrained using the LINCS algorithm.

The system of free FIXa contains 29197 atoms. The systems FIXpA, FIXpB and FIXpC contain 29166, 29055 and 30324 atoms respectively. Ions were added to neutralize the charge density of the complex. The solvent was introduced into the system by adding

8941 SPC water molecules for FIXa and 8896, 8865 and 9281 for FIXpA, FIXpB and FIXpC respectively.

RESULTS AND DISCUSSION

First, the simulations were analyzed in order to evaluate its stability by means of measuring the MD's convergence. The root mean square deviation (RMSD) is a suitable way to evaluate the stability of the simulation. We have measured the RMSD for all atoms and for the backbones in all simulations, for 1.2 ns of the simulations, where the initial time on the plots (Figure 1) corresponds to the beginning of the simulations. The all-atoms RMSD is 0.25nm ± 0.1nm and the backbone-atoms RMSD is 0.2nm ± 0.125nm. The data shows stability in the trajectories, where the FIXpA complex is perhaps the most stable one in the period measured. The system FIXpC shows a peak at around 800 ps to 1100 ps.

We calculated the distance between the centers of mass of FIXa and the corresponding peptide in the systems FIXpA, FIXpB and FIXpC (Figure 2). FIXpA seems to have the most stable measure between the three systems, of 1.6nm ± 0.1nm, while FIXpC shows a major peak at 1000 ps. In all the systems, FIXa and peptide are moving apart untill 500 ps. Then, they move closer in FIXpB, indicating a better affinity between these two elements. This movement can be also seen in FIXpC, after 1000 ps. In order to analyze the relative size of the molecules, we measured the radius of gyration (mass-weighted root mean-square distance from a set of atoms to its center of gravity). The radius of gyration for free FIXa is 1.66nm ± 0.01nm (Figure 2). FIXpB shows a slight increase during all the simulation. FIXpA and FIXpC are moving closer, reaching 1.66nm and 1.64nm, respectively.

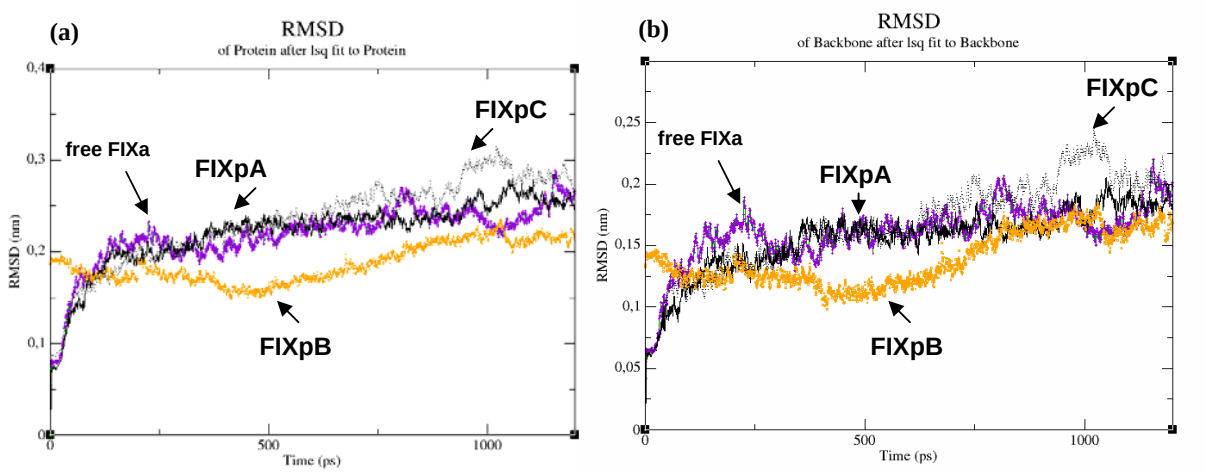

Figure 1. Time dependence of the root mean square deviation (RMSD) of structures from the four simulations, for (a) all the atoms and (b) only the backbone atoms. The continuous line represents FIXpA, squares represent FIXpB; the dashed line, FIXpC, the "+" sign, free FIXa.

The average number of hydrogen bonds is 7 (Figure 2). FIXpC shows an increased number and two major peaks (at around 400 ps and around 850 ps).

The MD simulations showed that the three peptides interacting at the active site of the FIXa bind to the pockets calculated. The system FIXpA seems to be the most stable one, since its RMSD, distance and radius of gyration measures does not show much variation. From the three peptide-FIXa systems, the peptide in FIXpC showed the best affinity. Peptide C seems to be moving closer to FIXa, as its distance and radius of gyration measures decreases and its hydrogen bond number increases.

The sequence LKEADE, which is common to the three peptides, seems to be the basis of interaction on the FIXa active site. As show in Table 2, the segment LKE interacts with Pocket 1, while AD interacts to Pocket 2 and E to Pocket 3. Figure 3 shows the protein surface and the peptides conformation.

CONCLUSIONS

We presented an analysis of FIXa binding pockets along with its interaction to nitrophorin-2 in a dynamical way. The FIXa active binding pockets were estimated. The NP-2 peptide presenting anticoagulant activity showed to be steric and electrostatic

complementary to the protein pockets. Also, volume measurements of the peptide are compatible to those of the combination of Pocket 1, Pocket 2 and Pocket 3.

The three MD simulations presented a binding between FIXa and NP-2 peptide (or a mutated NP-2 peptide). From the peptide models analysis, it was shown that the system FIXpA seems to be the most stable one, since its RMSD, distance and radius of gyration measures does not show much variation. Besides, the peptide in FIXpC showed the best affinity. A next step would include a modelling of FVIIIa, and its MD simulation along with FIXa and NP-2, to check the increased affinity.

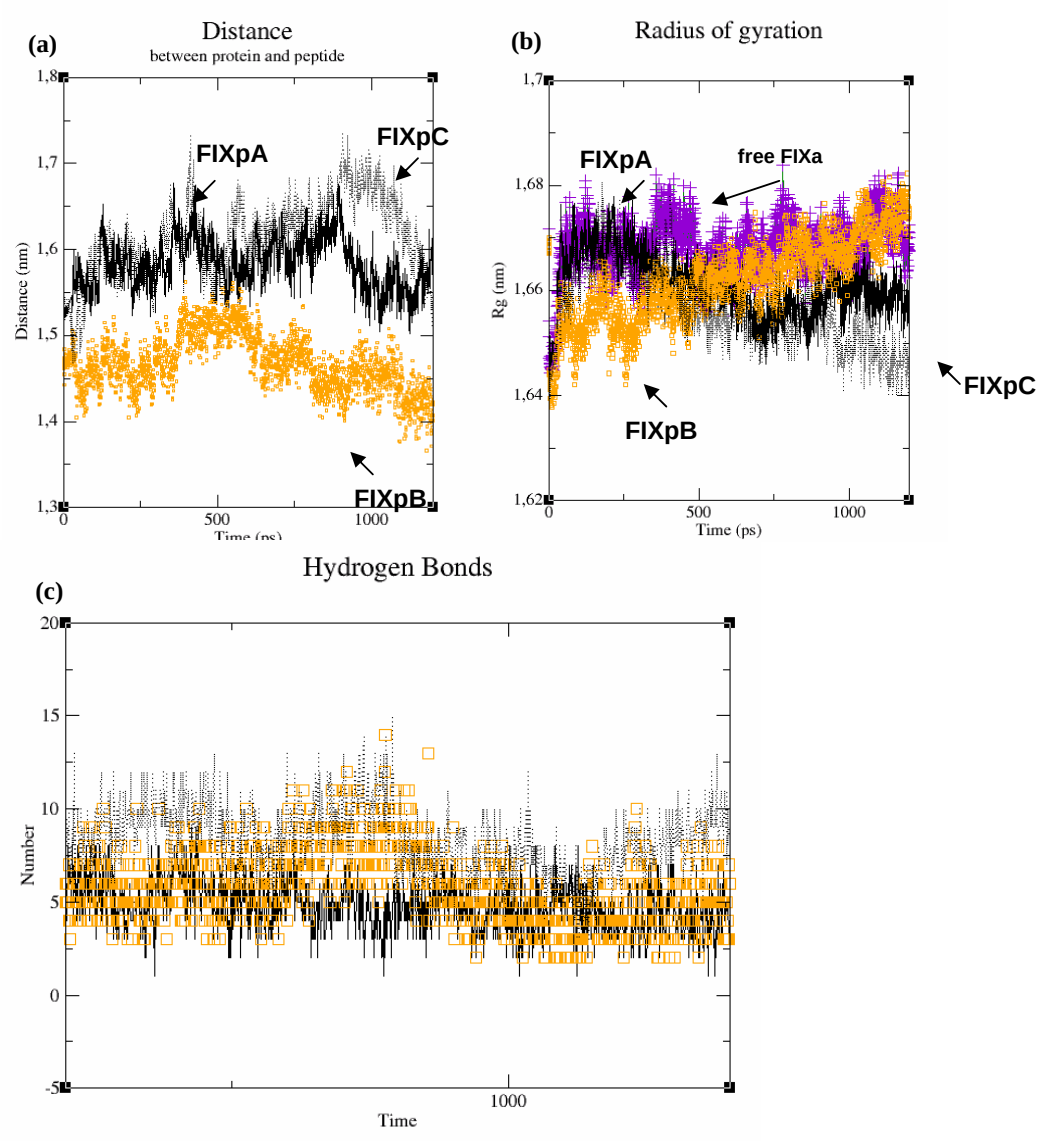

Figure 2. (a) Comparison between the distance of the center of mass of FIXa and peptide as a function of time. (b) Analysis of the radius of gyration of structures from the simulations. (c) Time dependence plot of the number of hydrogen bonds in the peptide-FIXa simulations. For all plots, the continuous line represents FIXpA; squares represent FIXpB and the dashed line, FIXpC. The "+" sign in the (c) plot shows the radius of gyration for free FIXa.

System	Peptides used on system
FIXpA	L K E A D
FIXpB	V L K E A D E
FIXpC	V L K E A D E
Related Pockets	1 2 3

Table 2. Interaction between residues of peptide fragments and protein pockets

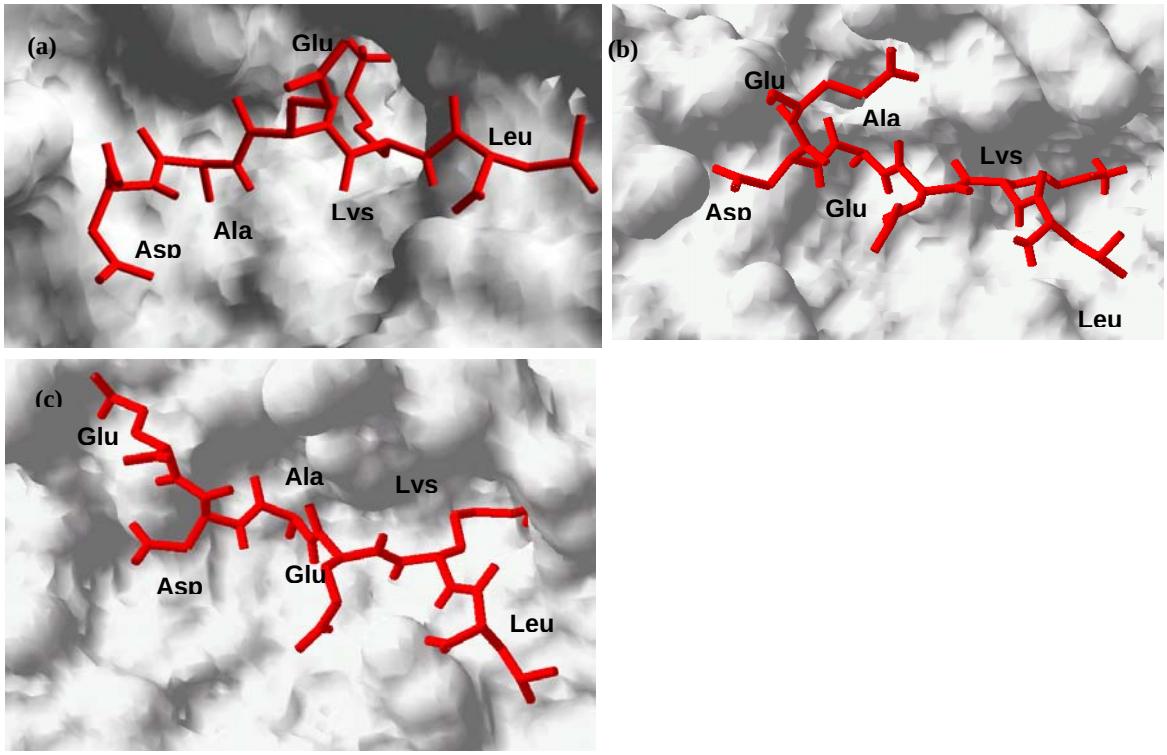

Figure 3. Images showing the protein surface and the peptides conformation in the active binding site of FIXa. The images represent, by clockwise, FIXpA, FIXpB and FIXpC.

New studies will be carried out using the structure LKEADE as a starting point for drug design (in the search for peptide-like ligands orally active).

Also, previous studies [3] indicate that Nitrophorin-2 inhibitory effect increases in the presence of FVIIIa and of PL. NP-2 shows more affinity by the enzyme-substrate complex (FIXa/FVIIIa/Ca^{+2}/PL/FX complex) than by the enzimatic complex (FIXa/FVIIIa/Ca^{+2}/PL). The next step would include a modeling of FVIIIa, and its MD simulation along with FIXa and NP-2, to check the increased affinity.

REFERENCES

1) Stoylova, et al. *The Journal of Biological Chemistry*, **274**: 36573 – 36578, 1999.
2) Bajaj, et al. *The Journal of Biological Chemistry*, **276**: 16302 – 16309, 2001.
3) Zhang, et al. *Biochemistry*, **37**: 10681 – 10690, 1998.
4) Liang, et al. *Protein Science*, **7**:1884 – 188, 1998.
5) Guex, et al. *Electrophoresis*, **18**: 2714-2723, 1997;
6) van Gunsteren, et al. *Vdf Hochschulverlag AG an der ETH Zurich*, 1996.

4.2. Subfamílias de Serpinas

O segundo artigo, que fez uso de modelos ocultos de Markov, subdividiu as serpinas em relação à sua função biológica, criando assinaturas para cada grupo: seqüências consenso e padrões de aminoácidos que são distintos para cada modelo/função correspondente.

Para a superfamília de serpinas (representando qualquer serpina, independente de atividade inibitória), foi identificada o padrão AMLS (localizado entre os resíduos 87 e 90 da seqüência da α1-antitripsina). Para a subfamília de serpinas que apresentam atividade inibitória, foi identificado o padrão HKAVL (resíduos 358-362). Já o padrão proveniente do modelo de serpinas que inibem serino-proteases envolvidas com a coagulação sangüínea é descrito pela assinatura IFFSPVSI (posições 74-81).

Em busca no BLASTp usando a seqüência IFFSPVSI, 70% das proteínas resultantes são serpinas da coagulação sangüínea. Este dado foi empregado na criação da expressão regular [IVTLM]-[FLVA]-F-S-P-[VLWYF]-[SG]-[IV]. Esta expressão pode indicar a função de inibição da coagulação, quando encontrada em uma proteína. Em pesquisa em banco de dados, aproximadamente 70% das proteínas resultantes são serpinas da coagulação. Este padrão codifica a seqüência de uma região específica ("shutter") envolvida em importantes transições conformacionais.

Resultados e discussão mais detalhada é apresentada na forma do manuscrito a seguir: "IDENTIFICATION AND CHARACTERIZATION OF SUBFAMILY SIGNATURES OF SERPINS IN THEIR PROTEIN SUPERFAMILY" Cristina Russso; Ana Bazzan, Hermes Luís Neubauer de Amorim, Jorge Almeida Guimarães. Trabalho recentemente submetido para publicação na revista BMC Bioinformatics.

Identification and Characterization of Subfamily Signatures of Serpins in their Protein Superfamily

Cristina Russso[1]; Ana Bazzan [2], Hermes Luís Neubauer de Amorim[1,3] and Jorge Almeida Guimarães[1],

[1]Centro de Biotecnologia - Universidade Federal do Rio Grande do Sul (UFRGS), Av. Bento Gonçalves, 9500, P.O.Box 15.005, 91500-970 Porto Alegre, RS, Brazil
[2]Instituto de Informática, Universidade Federal do Rio Grande do Sul (UFRGS), Av. Bento Gonçalves, 9500, P.O.Box 15064, Porto Alegre, RS, Brazil
[3]Departamento de Química - Universidade Luterana do Brasil (ULBRA), Av. Farroupilha 1001, Prédio 14, Sala 420C - 92450-900 Canoas, RS – Brazil.

Address correspondence to: Hermes Luís Neubauer de Amorim, Centro de Biotecnologia - Universidade Federal do Rio Grande do Sul (UFRGS), Av. Bento Gonçalves, 9500, P.O. Box 15005, Bloco IV, Prédio 43421, 91500-970 Porto Alegre, RS, Brazil.
Tel: 0055-51-33167770; Fax: 0055-51-33167309; E-mail: hamorim@cbiot.ufrgs.br

ABSTRACT

Background. The superfamily of serpins (Serine Protease Inhibitors) is a group of regulatory proteins that share the same biochemical mechanism and structural fold, although presenting different biological and biochemical functions. There is a significant amount of information on the mechanism of action of serpins; however, automatic classification of serpin subfamilies is still not available.

Results. To investigate the presence of sequence signatures that can be related with four serpins subfamilies, we analyzed sequences of 340 members of this protein superfamily and classified them according to specific biological functions. We built Hidden Markov Models (HMMs) and used Multiple Sequence Alignments to establish appropriate comparisons. The four models distinguish: (a) serpins; (b) general inhibitory serpins, (c) serpins that inhibit serine proteases of blood coagulation cascade, and (d) serpins with unrelated and/or unknown inhibitory function. It was observed that the HMMs generated

reasonable information, matching the four subfamilies of serpins. HMMs for each serpin subfamily showed different sites displaying high sequence conservation. A conserved site found at positions 74 to 81 (human α1-antitrypsin numbering) for serpins related to the clotting cascade could be mapped into the shutter region, which connects to the RCL (reactive center loop) during the inhibition mechanism.

Conclusions. The emergence of patterns could present biological relevance. Singnatures can be used for discriminating between subfamilies. Also, the sites found could help identifying structure-sequence-function relationships, since some of them are related to controlling and modulating serpin conformational changes in inhibitory serpins.

BACKGROUND

Serine Protease Inhibitors, the serpins, correspond to a superfamily of regulatory proteins (350-500 amino acids in length). These proteins perform a variety of biological roles, including, but not limited to, inhibition of chymotrypsin- and trypsin-like serine proteases. However, not all proteins classified as serpins present inhibitory activity. The serpins present increasing interest because of their medical importance (point mutations can cause a number of disease states, including blood clotting disorders, pulmonary emphysema, cirrhosis, and mental diseases) and also for their unusual mechanism of behavioral function. Such mechanism involves a conformational change, known as the stressed→relaxed (S→R) transition, between conformational states of different folding topologies [1].

The conserved serpin fold includes three β-sheets and at least seven (most typically nine)

α-helices. The reactive center loop (RCL, or reactive site loop, RSL) is crucial for the inhibitory function. This loop contains a stretch of ~17 residues tethered between A and C β-sheets. Additional requirements for the serpins to express effective inhibitory action upon proteases include a critical RCL length. Also, it requires the presence of appropriate residues within the loop that are compatible with rapid and favorable burial into the β-sheet A. Several regions are important in controlling and modulating serpin conformational changes, such as the hinge, breach, shutter and gate, illustrated in Figure 1A [2]. Five conformational states, namely: native, cleaved, latent, δ, and polymeric states appear in serpin crystal structures. These states differ primarily in the structure of RCL.

Several hundred serpins have been identified in higher eukaryotes and viruses [3]. They represent an expanding superfamily of structurally similar but functionally diverse proteins. Most serpin inhibit serine proteases of the chymotrypsin/trypsin family. However, cross-class inhibitors and several members that no longer function as protease inhibitors have been also identified. They perform other roles in different physiological systems, such as hormone transport, corticosteroid binding globulin and blood pressure regulation. Thus, understanding the biological function of most serpins remains an ongoing challenge for the biomedical researchers.

Since the beginning of the post-genomic era, functional and structural characterization of new proteins represents a critical task for computational biology. Most profile and motif databases tend to classify protein sequences into a broad spectrum of protein families. However, protein classification through systems capable of distinguishing subfamilies structures as well as the identification of functionally diverse members of the superfamily has not yet been implemented. The most popular methods of sequence analysis, like single-sequence similarity algorithms such as BLAST [4] or FASTA [5], are not sensitive enough

to distinguish and capture small differences resulting from protein sequences [6-8]. Such traditional pairwise alignment uses position-independent scoring parameters, missing important information about the degree of conservation at various positions in the multiple alignment. The goal of this work is to contribute with the elucidating efforts of structure-function relationships in proteins. Specifically, we explore a Hidden Markov Model methodology for preliminary identification and characterization of subfamily signatures in serpin superfamily. We created four models in order to distinguish: (a) proteins classified as being part of the serpin superfamily; (b) proteins classified as being part of a serpin subfamily exhibiting general inhibitory activity; (c) proteins classified as being part of a serpin superfamily which inhibit specifically the serine-proteases of blood coagulation cascade, and (d) proteins classified as being part of the serpin superfamily but with unrelated and/or unknown inhibitory function. A consensus model for each set of serpins was created; well representing each subfamily. Cutoff values were assigned to distinguish each subfamily, allowing automatic classification, database search, and unidentified protein sequence classification. The emergence of patterns that could present biological relevance were observed: the conserved sites on the models can lead to a signature for the serpin superfamily, thus indicating a pattern of inhibition. A regular expression was created to represent the serpins related to the coagulation cascade.

RESULTS

Subfamily Models

In order to characterize protein members of subfamilies of serpins, models representing each subfamily were created. Table 1 presents the list of examples of serpins collected

Table 1 - Serpins classification on the input sets.

Serpin	MT*	MI**	MC***	MN****
α1-Antichymotrypsin	X	X	X	-
α1-Antichymotrypsin II	X	X	X	-
α1-Antitrypsin	X	X	X	-
α2-Antiplasmin	X	X	X	-
Angiotensinogen	X	-	-	X
Antithrombin-III	X	X	X	-
Bomapin	X	X	-	-
Plasma protease C1 inhibitor	X	X	X	-
Cytoplasmic antiproteinase 2	X	X	X	-
Collagen-binding protein 2	X	-	-	X
Contrapsin-like protease inhibitor	X	X	X	-
Corticosteroid-binding globulin	X	-	-	X
Glia-derived nexin	X	-	-	X
47 kDa Heat-shock protein	X	-	-	X
Heparin cofactor II	X	X	X	-
Hurpin	X	-	-	X
Kalistatin	X	X	X	-
Kallikrein-binding protein	X	X	X	-
Leukocyte elastase inhibitor	X	X	-	-
Manduca sexta alaserpin	X	X	-	-
Maspin precursor	X	-	-	X
Megsin	X	-	-	X
Neuroserpin precursor	X	X	-	-
Ovalbumin	X	-	-	X
Pigment epithelium-derived factor	X	-	-	X
Placental thrombin inhibitor	X	X	X	-
Plasma serine protease inhibitor	X	X	X	-
Plasminogen Activator Inhibitor-1	X	X	X	-
Plasminogen Activator Inhibitor-2	X	X	X	-
Protein C Inhibitor	X	X	X	-
Protein Z-dependent protease inhibitor	X	X	-	-
Squamous Cell Carcinoma Antigen-1	X	X	-	-
Squamous Cell Carcinoma Antigen-2	X	X	-	-
Thyroxine-binding globulin	X	-	-	X
Uteroferrin-associated protein	X	-	-	X
Uterine milk protein	X	-	-	X
Viral serpin CrmA	X	X	-	-

* Model with general serpins, **inhibitory serpins, *** with serpins related to the clotting cascade and **** serpins with unknown inhibitory activity.

from the Swiss-Prot Database and used in the input sets. From this representative list of serpins, the four groups MT, MI, MC and MN could be organized. Assigning cutoff values

to such models allowed automatic classification, database search, and unidentified protein sequence classification. In some cases, a particular pattern could be assembled from the emergence of sequence patterns presenting biological relevance. Finally, we designed a regular expression potentially describing serpins acting on enzymes of the blood coagulation cascade. So far, all consensus models, cutoff values, patterns and expressions have been showing consistent results.

Database search and cutoff values

A database search using the entire models over the SWISSPROT database was performed (with the program hmmsearch). The search sweeps the database to find sequences matching the models. If the models are accurate, only the serpins presenting that specific function (for example, general inhibitory activity) would be expected in the results. Indeed, most models displayed highly accurate results. At the time of this search (October, 2003), most proteins from the database fitting the model MT were serpins (Table 2). The model presented very accurate results, since it gave only four false positives. For MI, all top proteins (158 proteins) captured by the model are serpins presenting inhibitory activity, up to 511.6 bits of raw score (or E-value 1.9e-149). Again, the model presented consistent results for MI serpins, with only seven false positives and 95.56% accuracy. Similar results were observed with MC proteins, where an even more accurate cuttof of 554 bits or 3.4e-162 of E-value. Up to this cutoff, it displayed only serpins (119 proteins) all of them able to inhibit enzymes of the coagulation cascade, two false positives and 98.31% accuracy. When applied to MN (non-inhibitory serpins), the model displayed a cutoff of 544.5 bits or E-value 2.5e-159. This is the less accurate between the serpin models, displaying 65.62% accuracy. Among the top 68 serpins with non-inhibitory role, the model has captured 22

Table 2 - hmmsearch over SWISSPROT with the serpin models.

Model	Cutoff score (bits)	Cutoff e-value	Number of false positives	Number of sequences found	Accuracy
MT	-	-	4	288	98.6%
MI	511.6	1.9e-149	7	158	95.56%
MC	554.0	3.4e-162	2	119	98.31%
MN	544.5	2.5e-159	22	64	65.62%

Scores and e-values for models MT, MI, MC and MN found with the hmmsearch feature of HMMER. All the proteins returned after a search with the entire MT over SWISSPROT were serpins. The number decreases a little bit for the following models, with the specific cutoff indicated (number of proteins found inside the range of the cutoff). The accuracy here was calculated based on the number of true positives found.

false positives, i.e. proteins not related to serpins. Since all these non-related serpins from MN seem to be somewhat divergent, it could explain the decreased success in the search for similar proteins. The results obtained for inhibitory and for coagulation serpins confirm that the related models are accurate, since a search in large databases match them only with specific serpins (all results displayed in Table 2). It is then concluded that the models can be used to distinguish with high accuracy proteins with non-identified serpin function.

Patterns of Sequence Conservation

Using human α1-antitrypsin as a model, information concerning the sequence conservation was obtained from the MSAs (mapped to the amino acid numbering of the sequence human α1-antitrypsin). The model MT presents two match-sites of residue conservation higher than 75%. The first one, located on the amino acid positions 87 to 90 of the human α1-antitrypsin sequence (Swiss-Prot entry A1AT_HUMAN), is AMLS (with slight variations). It is found in 45,7% of MT sequences; in 43,9% of MI sequences; and in 64,9% of MC. The highly conserved site HKAVL present by the MI model is shared by MT and MC models and found the positions 358-362 of the human α1-antitrypsin amino acid chain (this region can be seen conserved in a Multiple Sequence Alignment portion,

on Table 3, and mapped to the tertiary structure, on Figure 1B). These conserved regions were mapped to variations of the pattern in the corresponding tertiary structure of α1-antichimotrypsin and antithrombin, as shown in Figure 1. It appears in 47,18% of MT sequences; in 58% of MI and in 71% of MC sequences. Since this site appears only on the models containing serpins with inhibitory activity, it could indicate a signature for inhibitory serpins. Due to its divergence on time, the model MN, for serpins with unrelated and/or unknown inhibitory function, did not show these conserved regions.

Subfamily Signatures

The differences found between the models MT, MI, MC and MN indicate signatures of the subfamilies. As Table 3 depicts the sequence alignment of serpins classified as MI, Table 4 shows the MC alignment for serpins able to act upon the coagulation cascade. The MC group, shows a conserved site located at amino acid positions 74 to 81 of α1-antitrypsin. Such conservation suggests its biological relevance for the proteins activity in controlling and modulating serpin conformational changes [1,2]. The site, IFFSPVSI, appears in more than 82% of the MC sequences (Table 4 shows a portion of a Multiple Sequence Alignment, displaying this region; and Figure 1C shows a mappingo to the corresponding tertiary structure on antithrombin III). A BLAST search for short nearly exact matches using "IFFSPVSI" as a query took place. From the top hits returned, with score from 29 to 23 bits, more than 70% were serpins related to the coagulation cascade. Confirming such results, the search on the HITS database resulted in 75% of serpins related to the coagulation cascade. The search returned six false positives; despite not being related to the coagulation cascade, these proteins were all serpins. From the human α1-antitrypsin amino acid pattern IFFSPVSI, a regular expression was created, following

the convention used on PROSITE [9], in order to represent serpins related to the coagulation cascade signature: [IVTLM]-[FLVA]-F-S-P-[VLWYF]-[SG]-[IV]. Despite the first residue on the pattern being isoleucine (I) for more than 82% of the sequences; valine (V), threonine (T), leucine (L) and methione (M) also appear in much smaller probabilities. Tested against HITS protein domain database with Pattern Search, the regular expression returned a group of proteins, from which almost 70% were serpins related to the coagulation cascade (68,75%). The results of the tests using the "raw" pattern and the regular expression on BLAST and Pattern Search are compared in the Table 5.

Consensus Model

In order to validate the models, the best of each consensus sequence created was tested using BLASTp program [4,8]. The results are shown in Table 6. As expected, MT consensus sequence gave BLASTp results allowing its comparison with a large group of serpins with score ranging from 383 to 36.2 bits (e-value of 5e-106 to 0.16). For the second group, the MI model, BLASTp provided again another group of serpins where only the serpins with inhibitory role had the higher scores: from 301 to 183 bits, or e-value from 3e-81 to 6e-46. Of course, in this case proteins with scores lower than 183 bits were also serpins, but they did not present inhibitory activity. For the third set, the MC model, the BLASTp results with higher scores (499 bits to 183 bits, and e-values of 4e-141 to 9e-46), were serpins able to inhibit the coagulation cascade. The last model was the most difficult to classify, since its training set contained only 34 sequences. The BLASTp results for the MN model indicated non inhibitory serpins, having a range between 366 and 327 bits (e-values of 7e-101 to 5e-89).

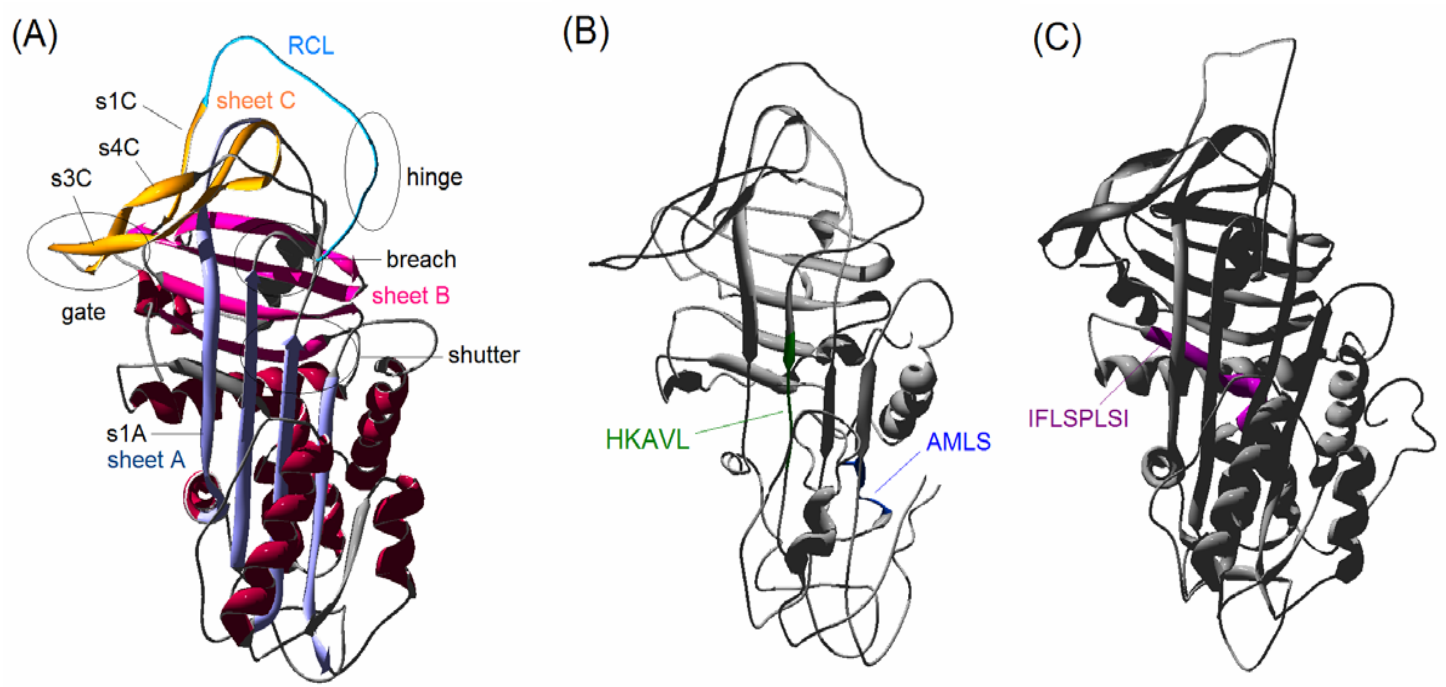

***Figure 1*.Schematic representation of serpins**. A: ribbon structure of native α_1-antitrypsin (entry 1ATU of PDB). B: ribbon structure of uncleaved α_1-antitrypsin (entry 1ATU of PDB). C: antithrombin III (entry 2ANT of PDB). In panel A relevant structural features of inhibitory serpins are shown. RCL: reaction center loop; s1C, s3C and s4C are, respectively, sheets 1, 3 and 4 of β-sheet C; s1A corresponds to sheet 1 of β-sheet C. Structural localization of MT and MC signatures are depicted in Panel B and C.

Table 3 - Signature pattern for inhibitory serpins.

Serpin (originated from)	**Alignment portion**
α1-Antitrypsin (*Homo sapiens*)	APLKLSKAV**HKAVL**TIDEKGTEAAGAMFLE
α1-Antitrypsin (*Tamias sibiricus*)	APLTVSKAL**HKAVL**DIDEEGTEAAGGTVLG
α1-Antitrypsin 1 (*Mus musculus*)	APLKLSQAV**HKAVL**TIDETGTEAAAVTVLQ
α1-Antitrypsin 3 (*Mus musculus*)	APLKLSQAV**HKAVL**TMDETGTEAAAATVLL
α1-Antiproteinase F (*Oryctolagus cuniculus*)	EPLKASQAL**HKAVL**TIDERGTEAAGATYME
α1-Antiproteinase F (*Cavia porcellus*)	MPLKISKGL**HKA**L**L**TIDEKGTEAAGATELE
Contrapsin (*Cavia porcellus*)	MPLKISKGL**HKA**L**L**TIDEEGTEAAAATVLE
α1-Antitrypsin-related (*Homo sapiens*)	APLKLSKAV**H**V**AVL**TIDEKGTEATGAPHLE
α1-Antiproteinase (*Didelphis marsupialis virginiana*)	TNLKLSQAV**HKAV**VNIDEKGTEASGATFAE
Contrapsin-like protease inhibitor 1 (*Mus musculus*)	KKLSVSQVV**HKAVL**DVAETGTEAAAATGVI
Kallikrein-binding protein (*Mus musculus*)	KDLIVSQMV**HKAVL**DVAETGTEGVAATGVN
Contrapsin-like protease inhibitor 1 (*Ratus norvegicus*)	KNLHVSQVV**HKAVL**DVDETGTEGAAATAVT
α1-Antichymotrypsin (*Homo sapiens*)	RNLAVSQVV**HKAVL**DVFEEGTEASAATAVK
Plasma serine protease inhibitor (*Homo sapiens*)	SNIQVSEMV**HKAV**VEVDESGTRAAAATGTI

Small portion of MI's MSA, where MI represents the model for serpins showing inhibitory activity. The pattern HKAVL, in bold, appears highly conservative, indicating a signature for inhibitory serpins.

Table 4 - Signature for serpins related to the clotting cascade.

Serpin	Alignment portion
α1-Antitrypsin (*Homo sapiens*)	TN**IFFSPVSI**ATAFAMLSLGTKADTHDEIL
α-1-Antitrypsin (*Tamias sibiricus*)	TN**IFFSPVSI**ATALAMLSLGTKGDTHTQIL
α-1-Antitrypsin 1 (*Mus musculus*)	SN**IFFSPVSI**ATAFAMLSLGSKGDTHTQIL
α-1-Antitrypsin 3 (*Mus musculus*)	SN**IFFSPVSI**ATAFAMLSLGSKGDTHTQIL
α-1-Antiproteinase F (*Oryctolagus cuniculus*)	TN**IFFSPVSI**ALAFAMLSLGAKGDTHTQVL
α-1-Antiproteinase F (*Cavia porcellus*)	SN**IFFSPVSI**ATALAMVSLGAKGDTHTQIL
Plasma serine protease inhibitor (*Homo sapiens*)	QN**IFFSPVSI**SMSLAMLSLGAGSSTKMQIL
Kallistatin (*Homo sapiens*)	KN**IFFSP**L**SI**SAAYAMLSLGACSHSRSQIL
Contrapsin-like protease inhibitor 6 (*Ratus norvegicus*)	KNVV**FSP**L**SI**SAALAVVSLGAKGNSMEEIL
Kallikrein-binding protein (*Mus musculus*)	TN**IV FSP**L**SI**SAALAIVSLGAKGNTLEEIL
Contrapsin-like protease inhibitor 1 (*Ratus norvegicus*)	KNVV**FSP**L**SI**SAALAILSLGAKDSTMEEIL

Small portion of MC's MSA, where MC represents the model for serpins related to the clotting cascade. The pattern IFFSPVSI, in bold, appears highly conservative, indicating a signature for those serpins.

Table 5 - Comparison between pattern testing result (subset of output proteins).

Protein shown in the results (originated from)	**Serpin related to clotting cascade**	**Found by BLAST (raw pattern: IFFSPVSI)**	**Found by Pattern Search (raw pattern IFFSPVSI)**	**Found by Pattern Search (regular expression: [IVTLM]-[FLVA]-F-S-P-[VLWYF]-[SG]-[IV])**
Contrapsin (*Cavia porcellus*)	X	X	X	X
α1-Antitrypsin-like protein (*Tamias sibiricus*)	X	X	X	X
α1-Antiproteinase F (*Cavia porcellus*)	X	X	X	X
Hibernation specific plasma protein (*Tamias sibiricus*)	-	X	X	X
α1-Antitrypsin-like protein (*Tamias sibiricus*)	X	X	X	X
Thyroxine-binding globulin (*Homo sapiens*)	-	X	X	X
α1-Antiproteinase (*Bos taurus*)	X	X	X	X
Thyroxine-binding globulin (*Bos taurus*)	-	X	X	X
Thyroxine-binding globulin (*Sus scrofa*)	-	X	X	X
α1-Antitrypsin (*Papio anubis*)	X	X	X	X
α1-Antitrypsin 1-3 (*Mus musculus*)	X	X	X	X
α1-Antitrypsin-like protein CM55-MS (*Tamias sibiricus*)	X	X	X	X
Plasma serine protease inhibitor (*Homo sapiens*)	X	X	X	X
Thyroxine-binding globulin (*Ovis aries*)	-	X	X	X
α1-Antiproteinase (*Ovies aries*)	X	X	X	X

α1-Antitrypsin-like protein (*Tamias sibiricus*)	X	X	X	X
α1-Antitrypsin 1-4 (*Mus musculus*)	X	X	X	X
Thyroxine-binding globulin (*Ratus norvegicus*)	X	X	X	X
α1-Antiproteinase F (*Oryctolagus cuniculus*)	X	X	X	X
α1-Antitrypsin 1-1 (*Mus musculus*)	X	X	-	X
α1-Antitrypsin (*Homo sapiens*)	X	X	X	X
Estrogen-regulated protein EP45 (*Xenopus laevis*)	-	X	-	-
α1-Antiproteinase (*Didelphis marsupialis virginiana*)	X	X	-	-
α1-Antiproteinase (*Callosciurus caniceps*)	X	X	-	X
Hurpin (*Homo sapiens*)	-	X	-	X
α1-Antiproteinase (*Ratus norvegicus*)	X	X	-	-
Kallistatin precursor (*Homo sapiens*)	X	X	-	X
Corticosteroid-binding globulin (*Homo sapiens*)	-	X	-	-
α1-Antiproteinase (*Mus caroli*)	X	X	-	-
Squamous cell carcinoma antigen 1 (*Homo sapiens*)	-	X	-	-
Squamous cell carcinoma antigen 2 (*Homo sapien*s)	-	X	-	-
Chalcone synthase (*Gerbera hybrida*)	-	-	-	X
Plasminogen activator inhibitor (*Bos taurus*)	X	-	-	X
Neuroserpin Precursor (*Homo sapiens*)	-	-	-	X

Table 6 - Consensus sequence analysis and validation.

Model	Highest BLASTp Score (bits)	Top BLASTp e-value	Cutoff score (bits)	Cutoff e-value
MT*	383	5e-106	36.2	0.16
MI**	301	3e-81	183	6e-46
MC***	499	4e-141	183	9e-46
MN****	366	7e-101	327	5e-89

Scores and e-values, provided by BLASTp, for the consensus sequence of each model. Above the cutoff, the proteins provided by BLASTp belong to the subfamily analyzed by the model.

* Model with general serpins, **inhibitory serpins, *** with serpins related to the clotting cascade and **** serpins with unknown inhibitory activity.

DISCUSSION

While superfamily classification is best done by methods that can generalize the features shared by a diverse group of examples, subfamily discrimination requires separating examples that may differ only slightly. In the case of serpins and other proteins that are classified as part of superfamilies, the diagnosis and characterization of subfamilies deserves even greater importance. The reason being related to the fact that such proteins classified within the same superfamily can often present several different functional roles. For example, despite the classification as serine-protease inhibitors, there are several members of serpins with no longer function as proteinase inhibitors. Moreover, the classification at subfamily level is relevant also because of the enormously increasing amount of data generated by several genome projects in development. In order to achieve the subfamily classification for proteins of serpin superfamily, signatures (windows of residues that are distinct among subfamilies) were created, representing each subfamily. A signature that appears conserved in all the subfamilies suggests a signature for the entire superfamily. The amino acids covered in the signature may be of biological relevance. The information about conservation or variation of the proteins could be obtained from the models MSA. All models containing serpins with inhibitory activity showed a conserved

site located at positions 358-362 of the human α1-antitrypsin amino acid chain. This site, HKAVL, appears only in the models representing serpins with inhibitory activity upon different proteases, indicating a signature for inhibitory serpins. The serpins displaying inhibitory activity to enzymes of the coagulation cascade showed a conserved site located at positions 86-89, namely the sequence IFFSPVSI. In this work we suggest that it constitutes a signature that characterizes the serpin function as inhibitors of components of the blood clotting cascade. The pattern could not be found on the other models. This site could be mapped into shutter region of inhibitory serpins, suggesting its biological relevance for the protein activity in controlling and modulation serpin conformational changes The residues IFF are located in a β-strand and seem to be related to the conformational change presented by those serpins. It is thus important to allow the S→R transition to take place. The two serines and the proline in the pattern are part of the shutter, which connects to the RCL during inhibition. Furthermore the model indicates that the Ser80 (α1-antitrypsin numbering) appears highly conserved only on the MC sequences, suggesting it has some important structural role in this site. The conserved residues in MT, MI and MC models seem to represent those mobile parts of the protein related to the conformational change S→R, since they are mapped to the shutter and to other important regions of the protein. Actually, Irving and co-workers [2] showed that site-directed mutation designed to switch a single amino acid residue in this region produced profound differences on the S→R conformational change. Thus, the observation that the MN serpin sequences do not show the same conserved sites in HMM is according with experimental results. It was observed that the HMMs generated reasonable information concerning the match of the subfamilies of serpins.

CONCLUSIONS

The method could be used for building models and profiles for a superfamily of proteins with diverse structural information. Besides information about the superfamily, the methodology brings inferences about the protein activity. It uses position-specific information, thus bringing data concerning conservation. Such a method seems to be more sensitive than the PROSITE. Each of the models generated a consensus sequence that would represent the most probable sequence for all the subfamily. Then, a serpin could be classified to a subfamily by testing its similarity to the subfamily consensus sequence. The testing of the sequences with BLASTp as a validation tool showed that they represent correctly the members of the corresponding model. Accordingly, cutoff E-values for each subfamily were assigned. HMMs limitations are due to the higher-order correlations, since the identification of a position is treated independently from the others, making it hard to model RNA (as base pair interactions will be missing). Although HMMs have higher processing time and memory costs, they provide a solid foundation for MSA information modeling. For further work, we plan to create an automated and faster clustering technique in order to classify proteins into distinct functional groups.

METHODS

Hidden Markov Models (HMMs) were used in order to create consensus models that represent each serpin subfamily and to observe the emergence of sequence patterns presenting biological relevance. HMM methods capture position-specific information able to indicate how conserved each column of the alignment is, and which residues are likely

to appear in a given position of the protein sequence. Being a profile method, HMMs can be used for searching databases using multiple sequence alignments instead of single query sequences. In this work we built four HMMs, named MT, MI, MC and MN. The first model, MT, contains the primary sequences of unspecific serpins; MI contains sequences of serpins known to function as general protease inhibitors; MC contains sequences of enzyme inhibitors of the blood clotting cascade system; and MN uses sequences of serpins with unrelated and/or unknown inhibitory function. The models were generated on the packages HMMpro and HMMER [10, 11] with a set of parameters and a set of input data. The primary sequences were obtained from Swiss-Prot Database [12]. The models were trained and calibrated and the best ones among those created by HMMpro and HMMER were chosen. Multiple Sequence Alignments (MSAs) using HMMpro and CLUSTALw [13] were also created and introduced into the analytical systems. The results of the four HMMs were tested using BLASTp, against Swiss-Prot database, with BLOSUM62 matrix, gap extension 1, and expect value for inclusion in subsequent rounds of 10. The input set comprises primary sequences of serpins collected from the Swiss-Prot Database for each model, as follow: MT, 142; MI, 107; MC, 57 and MN, 34 amino acid sequences. Table 1 lists a number of serpins used in these models. Although different in size, the number of sequences on the training set has little correlation to the quality of the HMMER model [14]. The smallest sequence is 305 amino acids long, and the longest one, 504. The four models created with HMMpro used the same parameters: protein alphabet, linear architecture, and size 400 (which is an average value between the sequences size). As the sequences length varied from 305 to 504, the residue positions were mapped to the 400 positions models via Delete or Insert states. The algorithms used for generating probabilities and training set for the models were Viterbi, Gradient and Online because of

their specificity to proteins. The models created with HMMER used the Plan7 Architecture, mixture Dirichlet priors and lengths of 407, 404, 424 and 417 residues for MT, MI, MC and MN, respectively. The program hmmcalibrate was used in order to estimate E-value (Expectation values) scores. The E-value describes the amount of false positives expected to be above the bit score. We used hmmsearch from HMMER package to search the database for testing the models accuracy and to identify possible new homologs. For validation, we used BLAST [15] searches over Swiss-Prot and Pattern Search queries over the HITS database of protein domain [16] (http://hits.isb-sib.ch/). We calculated accuracy as (true hits)/(true hits + false positives) over Swiss-Prot.

AUTHORS' CONTRIBUTION

CR and HLNDA designed the study and JAG supervised it. CR worked on the Hidden Markov Models with the supervision of AB. CR designed the protein signatures and drafted the manuscript, which was revised by HLNDA and JAG.

ACKNOWLEDGMENTS

The financial support for this project was provided by the Brazilian agencies CAPES (Foundation for Improving Higher Education), CNPq (National Counsel of Technological and Scientific Development) and FAPERGS (Foundation for the Support of Research in Rio Grande do Sul).

REFERENCES

1. Ye S, Goldsmith EJ: **Serpins and other covalent protease inhibitors**. *Curr. Opin. Struc. Biol.* 2001, **11**:740-745.

2. Irving JA, Pike RN, Lesk AM, Whisstock JC: **Phylogeny of the serpin superfamily: Implications of patterns of amino acid conservation for structure and function.** *Genome Res*. 2000, **10**(12):1845-1864.
3. Silverman GA, Bird PI, Carrell RW, Church FC, Coughlin PB, Gettins PG, Irving JA, Lomas DA, Luke CJ, Moyer RW, Pemberton PA, Remold-O'Donnell E, Salvesen GS, Travis J, Whisstock JC: **The serpins are an expanding superfamily of structurally similar but functionally diverse proteins. Evolution, mechanism of inhibition, novel functions, and a revised nomenclature.** *J Biol Chem* 2001, **276**:33293-33296.
4. Altschul S, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucl. Acid Res*. 1997, **25**:3389-3402.
5. Pearson W, Lipman D: **Improved tools for biological sequence comparison.** *Proc Natl Acad Sci U S A* 1988, **85**:2444-2448.
6. Truong K, Ikura M: **Identification and characterization of subfamily-specific signatures in a large protein superfamily by a hidden Markov model approach.** *BMC Bioinformatics* 2002, 3:1.
7. Karchin R, Karplus K, Haussler D: **Classifying G-protein coupled receptors with support vector machines.** *Bioinformatic*s 2002, **18**(1):147-159.
8. Karplus K, Sjolander K, Barret C, Cline M, Haussler D, Hughey D, Holm L, Sander C: **Predictin**g **protein structure using hidden markov models.** *Proteins: Structure, Function and Genetics* 1997, **1**:134-139.
9. Bucher P, Bairoch A: **A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation.** *Proc Int Conf Intell Syst Mol Biol*. 1994, **2**:53-61.
10. Durbin R, Eddy S, Krogh A, Mitchison G: **Biological sequence analysis: probabilistic models of proteins and nucleic acids.** Edited by Cambridge University Press; 1998.
11. **HMMER: profile HMMs for protein sequence analysis** [http://hmmer.wustl.edu]
12. **Swiss-Prot Protein knowledgebase** [http://www.expasy.org/sprot/]
13. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: **improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice.** *Nucleic Acids Res.* 1994, **22**:4673-4680.
14. McClure M, Smith C, Elton P: **Parameterization studies for the SAM and HMMER methods of hidden Markov model generation**. In: *ISMB-9 1996, Menlo Park, CA.* Edited by D.J. States, P. Agarwal, T. Gaasterland, L. Hunter & R. F. Smith (AAAI Press); 1996:155-164.
15. **SIB BLAST Network Service** [http://www.expasy.org/tools/blast]
16. Pagni M, Iseli C, Junier T, Falquet L, Jongeneel V, Bucher P: **trEST, trGEN and Hits: access to databases of predicted protein sequences.** *Nucleic Acids Res*. 2001, **29**:148-151.

5. CONCLUSÕES

5.1 Estudo preliminar da interação fIXa – NP-2

- Foi possível identificar os possíveis sítios de interação localizados no módulo catalítico de fIXa humano pelo reconhecimento e caracterização de cavidades.
- A análise da interface entre fIXa e o peptídeo derivado de NP-2 revelou a existência de complementaridade eletrostática e geométrica.
- Os aspectos de complementaridade serviram de base para que o fragmento ativo de NP-2 fosse manualmente orientado no sitio de ligação com fIXa.
- A partir da análise da interação entre o fragmento anticoagulante de NP-2 e fIXa, foram desenhadas novas seqüências potencialmente ativas.
- O fragmento LKEADE foi identificado nas simulações como aquele que apresenta maior afinidade por fIXa.

5.2 Assinaturas de seqüência encontradas usando métodos ocultos de Markov na análise de serpinas

- Foram criadas expressões que permitem o reconhecimento de serpinas com precisão nível de resolução e precisão maior que aquela obtida pelos métodos de classificação disponíveis.
- A assinatura de seqüência AMLS está presente em todas as proteínas atualmente caracterizadas como serpinas.
- A subfamília de serpinas que possuem atividade inibitória apresenta a seqüência exclusiva HKAVL.
- Serpinas que inibem serino-proteases envolvidas com a coagulação sangüínea são descritas pela assinatura IFFSPVSI.
- Os dados obtidos foram usados na criação da expressão regular [IVTLM]-[FLVA]-F-S-P-[VLWYF]-[SG]-[IV] a qual pode ser empregada na identificação de serpinas associadas à coagulação sangüínea.

6. PERSPECTIVAS

O modelo do complexo fIXa-nitroforina-2 será expandido com a inclusão de fator VIIIa para estudo de demais interações. O peptídeo LKEADE será sintetizado e testado em relação à sua característica anticoagulante.

Quanto às assinaturas de serpinas, serão feitas análise mais detalhadas nos modelos gerados, para que mais características estruturais possam ser previstas. Além disso, o conjunto de proteínas será aumentado, para obtenção de modelos estatísticos mais refinados. Sugere-se que a seqüência do modelo de coagulação IFFSPVSI, assim como a estrutura secundária correspondente, sejam melhor investigadas, por sugerirem um alvo para inibição ou ativação. Os modelos para subfamília de serpinas serão implementados e disponibilizados na internet, como forma de mecanismo automático de busca e pesquisa.

7. REFERÊNCIAS

1. ALTSCHUL, S., MADDEN, T. L., SCHAFFER, A. A., ZHANG, J., ZHANG, Z., MILLER, W., LIPMAN, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acid Res*. 25:3389-3402, 1997.
2. BENSON, D. A., KARSCH-MIZRACHI, I., LIPMAN, D. J., OSTELL, J., WHEELER, D. L. GenBank. *Nucleic Acids Res*. (Database issue):D34-8, 2005.
3. BERMAN, H. M., WESTBROOK. J., FENG, Z., GILLILAND, G., BHAT, T. N., WEISSIG, H., SHINDYALOV, I. N., BOURNE, P. E. The Protein Data Bank. *Nucleic Acids Res.* 28:235-242, 2000.
4. BERMAN, H. M., HENRICK, K., NAKAMURA, H.: Announcing the worldwide Protein Data Bank. *Nature Structural Biology* 10(12):980, 2003.
5. BODE, W., BAUMAN, M. I., HUBER, R., STONE, S. R., HOFSTEENGE, J. The refined 1.9 Å crystal structure of human alpha-thrombin: interaction with D-Phe-Pro-Arg chloromethylketone and significance of the Tyr-Pro-Pro-Trp insertion segment. *EMBO J*. **8**(11):3467-3475, 1989.
6. BROOKS, B. R., BRUCCOLERI, R. E., OLAFSON, B. D., STATES, D. J., SWAMINATHAN, S., KARPLUS, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculation. *Journal of Computational Chemistry*, 4(2):187-217, 1983.
7. BUCHER, P. and BAIROCH A. A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation.In: *Proceedings of the 2nd ISMB Conference*, 53-61, 1994.
8. BRONSON, M. L., and BRAND. L.. Numerical computer methods. *Elsevier Academic Press*, 2004.
9. DAURA, X., HAAKSMA, E., VAN GUNSTEREN, W. F. Factor Xa: Simulation studies with an eye to inhibitor design. *Journal of Computer-Aided Molecular Design.* 14:507-529, 2000.
10. DURBIN, R., EDDY, S., KROGH, A., MITCHISON, G. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press, 1998.
11. FERSHT, A. Structure and mechanism in protein science. 4a ed. W. H. Freeman and Company, 2002.
12. GRANT, G. H. and RICHARDS, W. G. Computational chemistry. Oxford University Press, 1995.
13. HIGGINS, D. G., THOMPSON, J. D., GIBSON, T. J. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol*. 266: 383-402, 1996.
14. IRVING, J. A., PIKE, R .N., LESK, A. M., WHISSTOCK, J. C. Phylogeny of the serpin superfamily: Implications of patterns of amino acid conservation for structure and function. *Genome Res*. 10(12):1845, 2000.
15. KARCHIN, R., KARPLUS, K., AND HAUSSLER, D. Classifying G-protein coupled receptors with support vector machines. *Bioinformatic*s 18(1):147-159, 2002.
16. KARPLUS, K., SJOLANDER, K., BARRET, C., CLINE, M., HAUSSLER, D., HUGHEY, D., HOLM, L., SANDER, C. Predicting protein structure using hidden markov models. *Proteins: Structure, Function and Genetics* 1:134-139, 1997.

17. KARPLUS, K.; BARRET, C.; HUGHEY, R. Hidden Markov models for detecting.remote protein homologies. *Bioinformatics* 14:846-856, 1998.
18. KROGH, A., BROWN, M., MIAN, I. S., SJOLANDER, K., HAUSSLER, D. Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology* , 235:1501-1531, 1994.
19. KARPLUS, M. Molecular dynamics of biological macromolecules: a brief history and perspective. *Biopolymers*, 68: 350 – 358, 2003;
20. MCCLURE, M., SMITH, C., ELTON, P. Parameterization studies for the SAM and HMMER methods of hidden Markov model generation. In: *ISMB-96* 155-164, 1996.
21. NORLEDGE, B. V., PETROVAN, R. J., RUF, W. & OLSON, A. J. The tissue factor/factor VIIa/factor Xa complex: a model built by docking and site-directed mutagenesi. *Proteins: Structure, Function and Genetics*, 53: 640 – 648, 2003.
22. PAGNI, M., ISELI, C., JUNIER, T., FALQUET, L., JONGENEEL, V., BUCHER, P. trEST, trGEN and Hits: access to databases of predicted protein sequences. *Nucleic Acids Res*. 29:148-151, 2001.
23. PARK, J., KARPLUS, K., BARRET, C., HUGHEY, R., HAUSSLER, D., HUBBARD, T., CHOTHIA, C. Sequence comparisons using multiple sequences detect three times as many remote homologs as pairwise methods. *J. Mol. Biol.* 284: 1201-1210, 1998.
24. PEARLMAN, D. A., CASE, D. A., CALDWELL, J. W., ROSS, W. R., CHEATHAM, T. E., DEBOLT, S. III, FERGUSON, D., SEIBEL, G., KOLLMAN, P. AMBER, a computer program for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to elucidate the structures and energies of molecules. *Comp. Phys. Commun.* 91:1-41, 1995.
25. PEARSON, W. AND LIPMAN, D. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* 85:2444-2448, 1988.
26. RUSSO, C., PINTO, F. M., JULIANO, M. A., de AMORIM, H. L. N., GUIMARÃES, J. A. Docking Studies of an Aticoagulant Peptide Fragment Derived from Nitrophorin-2 in the Active Site of Factor IXa: Perspectives to Drug Design. *SBBq/Programme and Abstracts*, p. 170, 2002;
27. RUSSO, C., PINTO, F. M., JULIANO, M. A., AMORIM, H. L. N., GUIMARÃES, J. A. Molecular modeling studies of an anticoagulant peptide derived from nitrophorin-2 in the active site of factor IXa: perspectives to drug design. *Proceedings of the winter international synposium on Information and communication technologies WISICT '04*, 249 – 254, 2004;
28. SILVERMAN, G. A., BIRD, P. I., CARRELL, R. W., CHURCH, F. C., COUGHLIN, P. B., GETTINS, P. G., IRVING, J. A., LOMAS, D. A., LUKE, C. J., MOYER, R. W., PEMBERTON, P. A., REMOLD-O'DONNELL, E., SALVESEN, G. S., TRAVIS, J., WHISSTOCK, J. C. The serpins are an expanding superfamily of structurally similar but functionally diverse proteins. Evolution, mechanism of inhibition, novel functions, and a revised nomenclature. *J Biol Chem* 276:33293-6, 2001.
29. SONNHAMMER, E. L. L., EDDY, S. R., DURBIN, R. Pfam: A Comprehensive Database of Protein Families Based on Seed Alignments. *Proteins* **28**:405-420, 1997.

30. SOUZA, O. N., ORNSTEIN, R. L. Molecular dynamics simulations of a protein-protein dimer: Particle-Mesh Ewald electrostatic model yields far superior results to standard cutoff model.. *Journal of Biomolecular Structure & Dynamics* 16: 1205-1218, 1999.

31. THOMPSON, J.D., HIGGINS, D.G., GIBSON, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673-80, 1994.

32. TRUONG, K. AND IKURA, M. Identification and characterization of subfamily-specific signatures in a large protein superfamily by a hidden Markov model approach. *BMC Bioinformatics* 3:1, 2002.

33. VAN GUNSTEREN, W. F., BILLETER, S. R., EISING, A. A., HÜNENBERGER, P. H., KRUGER, P., MARK, A. E., SCOTT, W. R. P., TIRONI, I. G. Biomolecular simulation: The GROMOS96 manual and user guide. Hochschulverlag AG an der ETH Zurich and BIOMOS b.v., Zurich, Groningen.1996

34. VAKSER, I. A. Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex. *Proteins* Suppl 1:226-30, 1997.

35. YE, S., AND GOLDSMITH, E. J. Serpins and other covalent protease inhibitors. *Curr. Opin. Struc. Biol.*11:740-745, 2001.

36. ZHANG, Y., RIBEIRO, J. M., GUIMARÃES, J. A. & WALSH, P. N. Nitrophorin-2: A novel mixed-type reversible specific inhibitor of the intrinsic factor-X activating complex. *Biochemistry,* 28;37(10): 10681 – 90, 1998.