Universidade Federal do Rio Grande do Sul
Centro de Biotecnologia
Programa de Pós-Graduação em Biologia Celular e Molecular

Análise computacional dos determinantes de atividades mio- e neurotóxicas de fosfolipases A_2 do veneno de serpentes

Fabiano Pasin

Porto Alegre
Maio de 2006

Universidade Federal do Rio Grande do Sul
Centro de Biotecnologia
Programa de Pós-Graduação em Biologia Celular e Molecular

Análise computacional dos determinantes de atividades mio- e neurotóxicas de fosfolipases A_2 do veneno de serpentes

Fabiano Pasin

Orientador: Prof. Dr. Jorge Almeida Guimarães

Dissertação submetida ao Programa de Pós-Graduação em Biologia Celular e Molecular (PPGBCM) do CBiot/UFRGS como parte dos requisitos necessários à obtenção do grau de Mestre

Porto Alegre, maio de 2006

Banca Examinadora

Dra. Célia Regina Ribeiro da Silva Carlini

Departamento de Biofísica do Instituto de Biociências e Centro de Biotecnologia,

Universidade Federal do Rio Grande do Sul

Dra. Leila Ribeiro

Departamento de Informática Teórica do Instituto de Informática,

Universidade Federal do Rio Grande do Sul

Dr. João Alexandre Ribeiro Gonçalves Barbosa

Centro de Biologia Molecular Estrutural, Laboratório Nacional de Luz Síncrotron

Maio 2006

Dedico este trabalho a todos aqueles que participaram comigo deste empreendimento científico-pessoal, cada um da sua maneira, mas cuja contribuição permitiu-me concluí-lo com sucesso.

Índice

Universidade Federal do Rio Grande do Sul 1
Apresentação 11
Abreviaturas e Convenções 14
Introdução 15
1. Revisão Bibliográfica 17
1.1 As fosfolipases A_2 17
1.1.1 FLA_2 intracelulares 18
1.1.2 FLA_2 extracelulares 19
1.1.2.1 Atividade catalítica das FLA_2 extracelulares 21
1.1.2.2 Diversidade funcional das FLA_2 22
1.1.2.3 "Sítios farmacológicos" 25
1.1.2.4 FLA_2 neurotóxicas e miotóxicas 27
1.2 A Bioinformática na busca por padrões em seqüências biológicas 28
1.2.1 Diferentes pontos de partida, mesma meta: interpretar dados brutos e convertê-los em conhecimento aplicável 30
1.3 Modelos Ocultos de Markov (MOMs) 35
1.3.1 Fundamentos dos MOMs aplicados à bioinformática 37
1.3.2 MOMs aplicados ao reconhecimento de atividade biológica de FLA_2 39
2. Objetivos 40
3. Resultados 41
3.1 Aplicação de metodologia para deteção de mio- e neurotoxicidade em FLA_2 42
3.2 Predição de neuro- e miotoxicidade em FLA_2 enfocando a busca por seus "sítios farmacológicos" 47
3.3 Ferramenta de classificação automática online 74
4. Conclusão 76
5. Referências Bibliográficas 77
6. Referências de Endereços da Internet 87
7. Perspectivas 90
APÊNDICE 1: Conceitos básicos para a compreensão da bioinformática deste trabalho 91
APÊNDICE 2: Mecanismo catalítico de FLA_2 93
CURRICULUM VITÆ 95

Agradecimentos

Agradeço particularmente aos meus pais e à minha irmã, pelo exemplo e pelo amor que me fornecem.

Também aos meus orientadores, Prof. Dr. Jorge Almeida Guimarães e Prof. Dr. Hermes Luís Neubauer Amorim, pela paciência, dedicação e todo o conhecimento que comigo compartilharam.

Não haveria como listar todas as pessoas que merecem meus agradecimentos por estarem presentes e atuantes neste período em que desenvolvi minhas atividades de Mestrado, a estes, que não cito a seguir, meu mais sincero abraço e agradecimento.

A seguir, uma pequena lista das pessoas que, de maneira mais significativa, contribuíram para que este trabalho fosse concluído e que eu mantivesse o foco e a motivação necessária para a execução dos estudos, trabalhos e atividades:

Jorge Almeida Guimarães
Hermes L. N. Amorim
Giancarlo Pasquali
Carlos Termignoni
Sílvia Centeno
Luciano Saucedo
Fernanda S. Oliveira
Rafael Cáceres
Renata Terra
Rosemari Dias
Guilherme A. Roesler
Tiago Charão
Sérgio Carlos Pazzini
Inelve Pavan Pazzini
Jerry M. Dressler
Cristina Russo
Sandro Silva de Souza
Sandro Pavan
Cristiano Masutti

A todos os amigos e colaboradores do Centro de Biotecnologia

Ao programa de Pós-Graduação em Biologia Celular e Molecular (PPGBCM)

Resumo

A superfamília das fosfolipases A_2 (FLA_2) é composta por proteínas dotadas de propriedades diferenciadas que as tornam capazes de apresentar distintas atividades biológicas, além de sua atividade catalítica. Esta diversidade funcional é intrigante devido à alta similaridade de seqüência primária e de estrutura entre essas proteínas. O principal objetivo deste trabalho é o desenvolvimento de uma metodologia para a predição de atividades biológicas específicas em FLA_2 de peçonha de serpentes a partir da análise de seqüências primárias. A metodologia desenvolvida compreende: a) seleção de seqüências anotadas quanto à função ou atividade biológica que desempenham; b) detecção e validação estatística de motivos de seqüência relacionados à atividade estudada; e c) construção de Modelos Ocultos de Markov (MOMs) representando cada motivo. MOM consiste em uma modelagem estatística que tem sido aplicada com sucesso em diversas áreas onde se faz necessária a detecção e representação de padrões de informação; por sua base matemática robusta e formal, pode ser utilizada na automação deste tipo de processo. A metodologia foi testada para duas atividades de FLA_2 de peçonha de serpente: neurotoxicidade e miotoxicidade. Para as FLA_2 neurotóxicas, foram detectados seis motivos conservados, dos quais três foram validados estatisticamente como sendo adequados na discriminação de seqüências neurotóxicas de não neurotóxicas. Para as FLA_2 miotóxicas, foram detectados seis motivos conservados, dos quais quatro foram validados. Os MOMs dos motivos validados podem ser usados na predição de atividade neurotóxica e miotóxica. As relações entre seqüência, estrutura, função e evolução das FLA_2s são discutidas. Os dados obtidos foram coerentes com a hipótese apresentada por Kini (2003), da existência de diferentes sítios farmacológicos na superfície das FLA_2,

interagindo independente ou cooperativamente com diferentes receptores, para gerar as diversas funções biológicas observadas. Por não haver, até o momento, qualquer ferramenta automatizada para a predição de função biológica de FLA_2, os resultados deste trabalho foram a base para a construção de uma ferramenta (disponível em www.cbiot.ufrgs.br/bioinfo/phospholipase) para a identificação de miotoxicidade e neurotoxicidade em FLA_2.

Palavras-chave: fosfolipase A_2 – Modelos Ocultos de Markov – peçonha de serpente – motivos conservados

Abstract

The phospholipase A_2 (PLA_2) superfamily is composed by proteins that show distinct properties that make them able to present distinct biological functions, besides its catalytic activity. This functional diversity is intriguing since these proteins present high similarity of primary sequence and structure. The main purpose of this work is the development of a methodology that could be used to predict specific biological activity of snake venom's PLA_2 through primary sequence analysis. The methodology developed includes: a) the selection of sequences with biological function annotation; b) detection and statistic validation of amino acid sequence motifs related to the biological function to be modeled; and c) construction of Hidden Markov Models (HMMs) representing the validated motifs. HMM is a statistical model that has been successfully applied to several areas where there is a need to detect and represent information patterns. Thanks to its robust and formal mathematical basis, it could be used for automating this kind of process. The methodology was tested for two biological activities of snake venom PLA_2: neurotoxicity and myotoxicity. For the neurotoxic PLA_2, six conserved motifs have been detected, and three have been validated. For the myotoxic PLA_2, six conserved motifs have been detected, and four have been validated. The HMMs constructed from these validated motifs showed to be useful for the prediction of neurotoxic and myotoxic activities. The relationships among sequence, structure, function and evolution of PLA_2 are discussed. The findings were consistent with the hypothesis presented by Kini (2003), according to the idea of the existence of many pharmacological sites in PLA_2 surface, interacting with different receptors, independently or cooperatively, to generate the diversity of activities observed. Since there is no automatic tool for PLA_2 functional

prediction, the results of this study were the basis for the construction of a tool (available at www.cbiot.ufrgs.br/bioinfo/phospholipase) for PLA_2 myotoxicity and neurotoxicity identification.

Keywords: phospholipase A_2 – Hidden Markov Models – snake venom – conserved motifs

Apresentação

A Dissertação de Mestrado aqui apresentada é fruto de um trabalho desenvolvido junto ao Grupo de Bioinformática Estrutural, vinculado ao Laboratório de Bioquímica Farmacológica do Centro de Biotecnologia (CBiot) da Universidade Federal do Rio Grande do Sul (UFRGS), sob orientação do Professor Doutor Jorge Almeida Guimarães.

Já sendo bacharel em Ciências da Computação pela UFRGS, cursei os dois primeiros anos de graduação em Ciências Biológicas, também na UFRGS, onde obtive os conhecimentos fundamentais que me deixaram mais confiante para iniciar a busca por um orientador e pleitear uma vaga no Mestrado do Programa de Pós-Graduação em Biologia Celular e Molecular (PPGBCM) do CBiot.

Por indicação do Professor Tarso B.L. Kirst (à época um dos nomes publicamente ligados à nascente área da bioinformática na UFRGS), cheguei até o Grupo de Bioinformática Estrutural onde, aceito como orientado pelo Professor Jorge A. Guimarães e como mestrando do PPGBCM, iniciei meu trabalho de pesquisas sobre a família de proteínas das fosfolipases A_2 (FLA_2), a partir da proposta dos Professores Jorge A. Guimarães e Hermes L.N. Amorim. À época, a então estudante de Iniciação Científica Cristina Russo (também oriunda do Curso de Graduação em Ciências da Computação, UFRGS), estudava a família de proteínas das serpinas, aplicando em seu trabalho uma abordagem estatística bem sucedida no reconhecimento de padrões (neste caso, aplicado a padrões de distribuição de aminoácidos): Modelos Ocultos de Markov (MOMs).

O objetivo inicial para meu projeto de Mestrado foi o de melhorar a resolução dos métodos existentes à época para a identificação de seqüências primárias de FLA_2, separando-as automaticamente de acordo com a principal classificação utilizada, formulada por Edward A. Dennis (1994), até o nível de subgrupo. Até aquele momento, os métodos existentes permitiam detectar, no máximo, se a seqüência de aminoácidos era ou não FLA_2 e, eventualmente, permitiam fornecer alguma informação referente à classificação de Dennis.

Para essa classificação, quatro critérios são utilizados para enquadrar cada FLA_2 em um dos 14 grupos (I-XIV) e 20 subgrupos (representados por uma letra após o algarismo romano do grupo – ex. IIB): (i) a enzima deve hidrolisar um substrato fosfolipídico natural na posição sn-2; (ii) a seqüência completa de aminoácidos deve ser conhecida, (iii) os membros de cada grupo devem ter homologia de seqüência facilmente identificável; e (iv) isoformas que retêm atividade catalítica são classificadas no mesmo grupo e subgrupo, mas são distinguidas por números arábicos.

A metodologia inicialmente desenvolvida por nós estava baseada no conceito de "janelas deslizantes", descrita em Truong e Ikura (2002). A partir de um Alinhamento Múltiplo de Seqüência (AMS) – no caso deste trabalho, um AMS por subgrupo de FLA_2 – são gerados vários MOMs, cada um construído a partir de um trecho específico do AMS, permitindo a identificação das regiões de seqüência primária mais discriminativas de cada subgrupo e utilização dos respectivos MOMs para a classificação automática.

Os primeiros resultados do trabalho, ainda concentrado na classificação das FLA_2 conforme Dennis, foram apresentado na VI Reunião Anual do PPGBCM, em 2004. Logo a seguir, o foco do trabalho tornou-se mais ambicioso. Constatamos que esta classificação

carecia de significância biológica, posto que agrupava, sob uma mesma classificação, FLA_2 com funções biológicas ou atividades bem distintas. Assim, iniciou-se um estudo para verificar se a mesma abordagem poderia ser utilizada para capturar padrões de mais alta ordem, que permitissem a identificação de FLA_2 conforme grupos biologicamente significativos. isto é, em grupos onde todas as FLA_2 integrantes apresentassem a mesma função ou atividade biológica.

Como conseqüência, o trabalho foi reformatado, concentrando-se exclusivamente nas FLA_2 de peçonha de serpentes e buscando o desenvolvimento de um protocolo de bioinformática para a análise e classificação destas, capaz de identificar automaticamente as atividades biológicas apresentadas por uma FLA_2 qualquer de peçonha, partindo apenas da informação de seqüência primária da mesma. A aplicação de MOMs continuou sendo o núcleo da metodologia, de maneira a estabelecer relações entre distribuições específicas de aminoácidos nas FLA_2 e atividades biológicas destas proteínas.

Os resultados deste trabalho estão apresentados sob a forma de artigos e resumos em congressos. Os artigos estão em anexo, ao final desta Dissertação, na seção de *Resultados*. No capítulo *Perspectivas*, estão comentadas as linhas possíveis na continuidade deste estudo e, no *Curriculum Vitæ*, estão listados os trabalhos dos quais participei como co-autor durante o período do Mestrado.

Abreviaturas e Convenções

A seguir, a lista das abreviaturas e acrônimos utilizados ao longo deste documento:

AMS: Alinhamento Múltiplo de Seqüências

FLA_2: Fosfolipase A_2

MEME: Multiple Em (Expectation of Maximization) for Motif Elicitation

HMMER: Hidden Markov Model Builder

MOM / MOMs: Modelo Oculto de Markov / Modelos Ocultos de Markov

PLA_2: Phospholipase A_2

RNM: Ressonância Nuclear Magnética

Nota: *Ao longo do texto, sempre que uma ferramenta ou base de dados possuir página na Internet, a mesma terá ao lado de sua referência o símbolo "⏭", significando que esta consta na lista "Referências com endereço na Internet", com maiores informações e a página correspondente*.

Introdução

Por meio do presente trabalho, visei a criação de uma metodologia para estudo e classificação automática das FLA_2 de peçonha de serpente quanto às atividades neuro- e miotóxicas, utilizando uma abordagem bioinformática.

A bioinformática é uma área relativamente nova, seguindo uma forte tendência da ciência atual: a interdisciplinaridade. Fundindo conhecimentos de áreas tão diversas quanto biologia, computação, farmácia química, entre outras, seu desafio é integrar ferramentas, abordagens, informações, transformando a imensa massa de dados gerada pelos atuais métodos de biologia molecular em *conhecimento*.

Ao longo deste trabalho, na forma de orientadores, orientados, colaboradores, participaram pesquisadores cujas áreas de origem são tão diversas quanto Veterinária/Bioquímica/Farmacologia (Prof. Dr. Jorge A. Guimarães), Química (Prof. Dr. Hermes L.N. Amorim), Biologia (estudante de Iniciação Científica, Fernanda Oliveira), além de Ciências da Computação, área primária de formação do autor deste trabalho.

No intuito de facilitar a interlocução entre os domínios destas distintas áreas e o entendimento deste trabalho, o *Apêndice I* explica alguns dos principais conceitos básicos relacionados à bioinformática.

Ainda sobre o trabalho que será aqui exposto, uma breve nota quanto às referências às FLA_2s: podendo as mesmas serem encontradas na natureza em dois "ambientes orgânicos" principais - o meio intracelular ou extracelular (as FLA_2 destinadas à secreção) - e sendo o foco deste trabalho as FLA_2 de peçonha de serpente

que pertencem a esta última categoria, salvo quando explicitado em contrário, estarei me referindo às FLA_2 extracelulares.

1. Revisão Bibliográfica

1.1 As fosfolipases A_2

As fosfolipases foram inicialmente caracterizadas como enzimas que hidrolisam diversos substratos lipídicos fosfatados. As primeiras publicações que se referem a enzimas com esta atividade datam da década de 1960 (Hanahan *et al*., 1960; Doery e Pearson, 1961; Saito e Hanahan, 1962; Moore e Williams, 1964). Posteriormente, mais proteínas com esta atividade catalítica foram descritas e as atividades de hidrólise catalisadas por elas foram melhor caracterizadas.

Verificou-se que as diversas fosfolipases apresentavam distinta seletividade com relação à posição na cadeia da ligação-alvo da hidrólise no fosfolipídio-substrato. Este foi o principal critério que deu origem à atual divisão das famílias de fosfolipases (**Figura 1**).

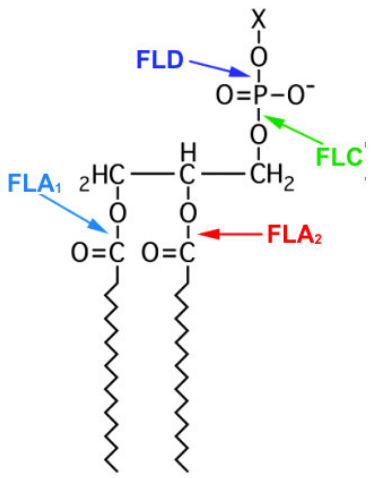


Figura 1: Equema representando substrato fosfolipídico típico e as ligações sobre as quais atuam diferentes fosfolipases, representantes de suas distintas famílias.
FLD: Fosfolipase D
FLC: Fosfolipase C
FLA_1: Fosfolipase A_1
FLA_2: Fosfolipase A_2

As fosfolipases A_2 (FLA_2) foram, então, classificadas como pertencentes à família de enzimas que têm por especificidade catalítica a hidrólise de fosfolipídeos na posição sn-2 (Valentin e Lambeau, 2000). Desta reação, são liberados um ácido graxo e um

lisofosfolipídeo (**Figura 2**). O lisofosfolipídeo liberado na reação tem ação na desestabilização de membranas celulares, e o ácido graxo participa de diversas funções celulares na forma de precursores de moléculas sinalizadoras e de outros compostos bioativos.

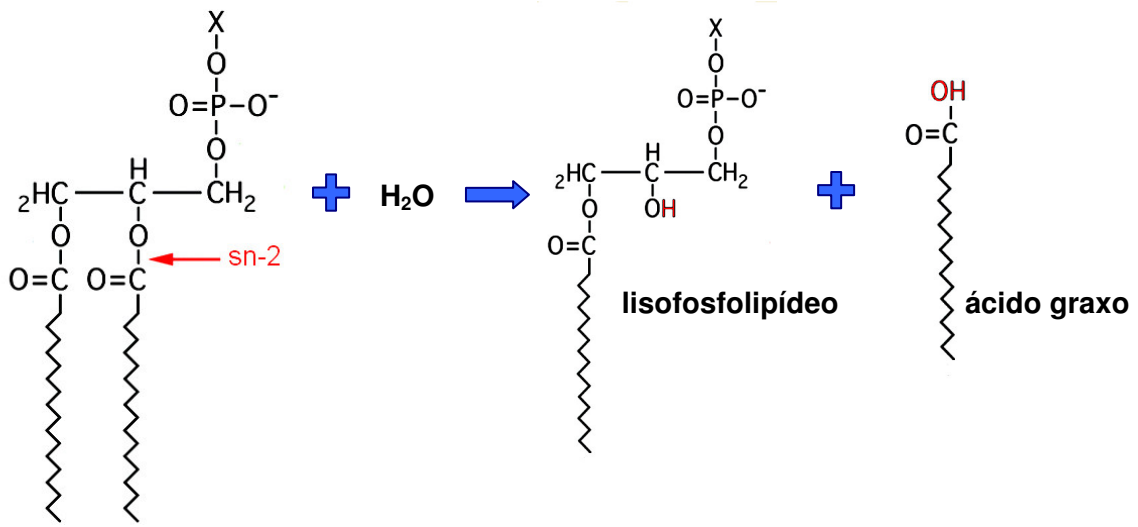


Figura 2: Esquema ilustrativo da catálise promovida pelas FLA2 em um substrato típico, com a posição sn-2, na qual ocorre a catálise, indicada por uma seta e os produtos derivados da reação. A letra "X" na figura representa grupos laterais diversos.

As FLA_2 são classificadas como hidrolases que atuam em ligações éster carboxílicas. Seu código na classificação de enzimas da *Joint Comission on Biochemical Nomenclature* (JCBN, IUPAC-IUBMB) é E.C. 3.1.1.4. (fosfatidilcolina 2-acil hidrolases).

1.1.1 FLA_2 intracelulares

A respeito das FLA_2, pode-se fazer uma clara distinção entre dois tipos de enzimas: as que atuam no meio intracelular e as que atuam no meio extracelular. Elas possuem marcantes diferenças, principalmente no que concerne à estrutura: diferentemente das extracelulares, as FLA_2 intracelulares são proteínas maiores, com massa molecular variando entre 26 kDa (~231 aminoácidos) e 114 kDa (~1012

aminoácidos) (Six e Dennis, 2000; Hirabayashi e Shimizu, 2000) . A **Figura 3** apresenta um modelo da estrutura tridimensional de uma representante dessa classe.

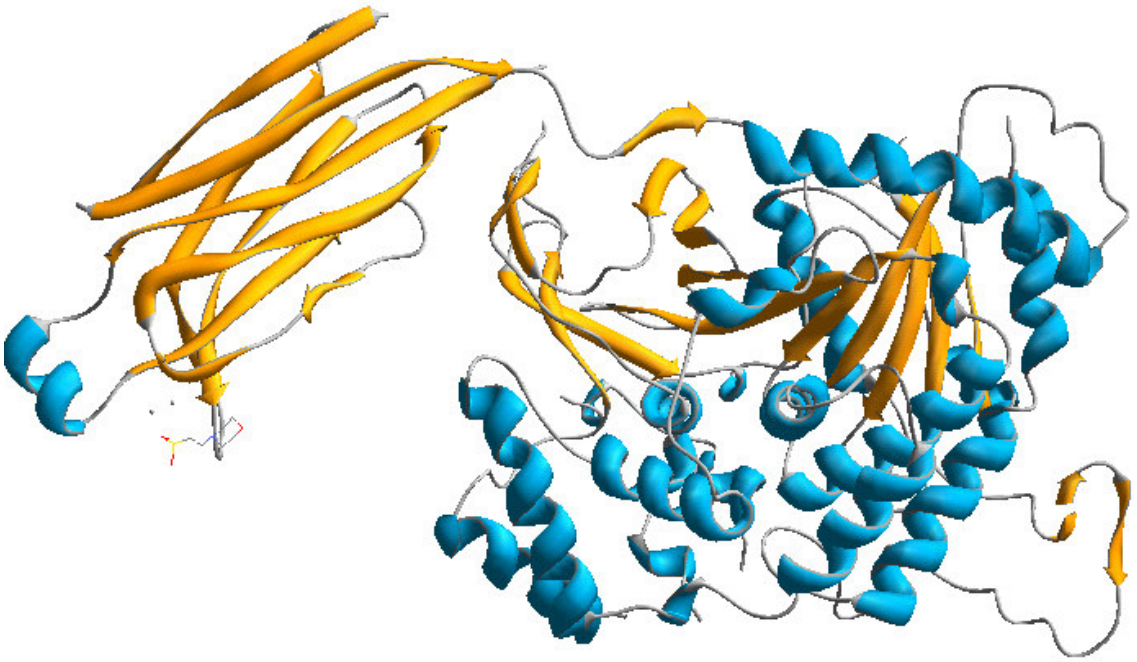


Figura 3: FLA2 citosólica humana. (Modelo de estrutura protéica obtido na base de dados *RCSB Protein Data Bank*⏩, indexada pelo código 1CJY. Figura gerada pelo autor, utilizando o software *SPDBViewer*⏩).

Essas FLA_2 estão geralmente envolvidas nos processos de sinalização celular e remodelagem de membrana (Brown *et al*., 2003). Um dos produtos de sua atividade catalítica, o ácido araquidônico, participa de rotas de síntese de compostos como os tromboxanos, leucotrienos e prostanóides (prostaglandinas e prostaciclinas), importantes intermediários em processos celulares diversos como resposta inflamatória, proteção de mucosas e do endotélio, neurotransmissão, controle de temperatura (Murakami e Kudo, 2004) e mediação da dor (Booting, 2004).

1.1.2 FLA_2 extracelulares

As FLA_2 extracelulares são proteínas menores com massa molecular variando entre 13 kDa (~110 aminoácidos) e 18 kDa (~170 aminoácidos) (Six e Dennis, 2000),

possuindo, portanto, massa molecular mais baixa em relação às extracelulares e alto conteúdo de pontes dissulfeto. Estas proteínas só adquirem sua conformação ativa quando no meio extracelular. No meio intracelular as FLA2 dos grupos I, II, V e X apresentam, na extremidade N-terminal, um peptídeo sinal para sua exportação ao meio extracelular (Murakami e Kudo, 2004; Bingham e Austen, 1999). As FLA_2 estão amplamente distribuídas em todos os principais ramos filogenéticos existentes conforme podemos ver na **Figura 5**.

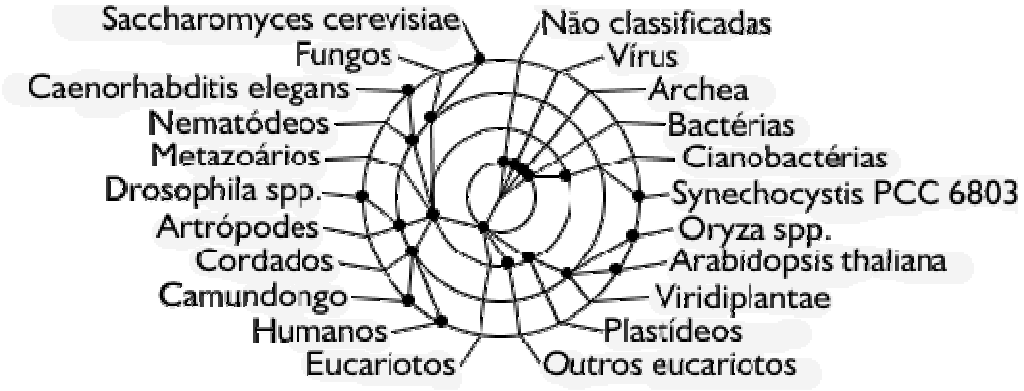


Figura 5: Diagrama de filogenia radial, evidenciando a distribuição das FLA_2 nos principais ramos evolutivos. Imagem modificada do site Interpro, acessada em 16 de janeiro de 2006. (http://www.ebi.ac.uk/interpro/DisplayIproEntry?ac=IPR001211)

Apesar da sua presença distribuída ao longo da evolução, as FLA_2 extracelulares são proteínas com alta similaridade estrutural entre si. As principais regiões estruturais compartilhadas entre elas são três grandes alfa-hélices centrais; duas fitas-beta formando uma região conhecida como asa-beta; uma alça de ligação com cálcio e uma alça C-terminal (**Figura 6**).

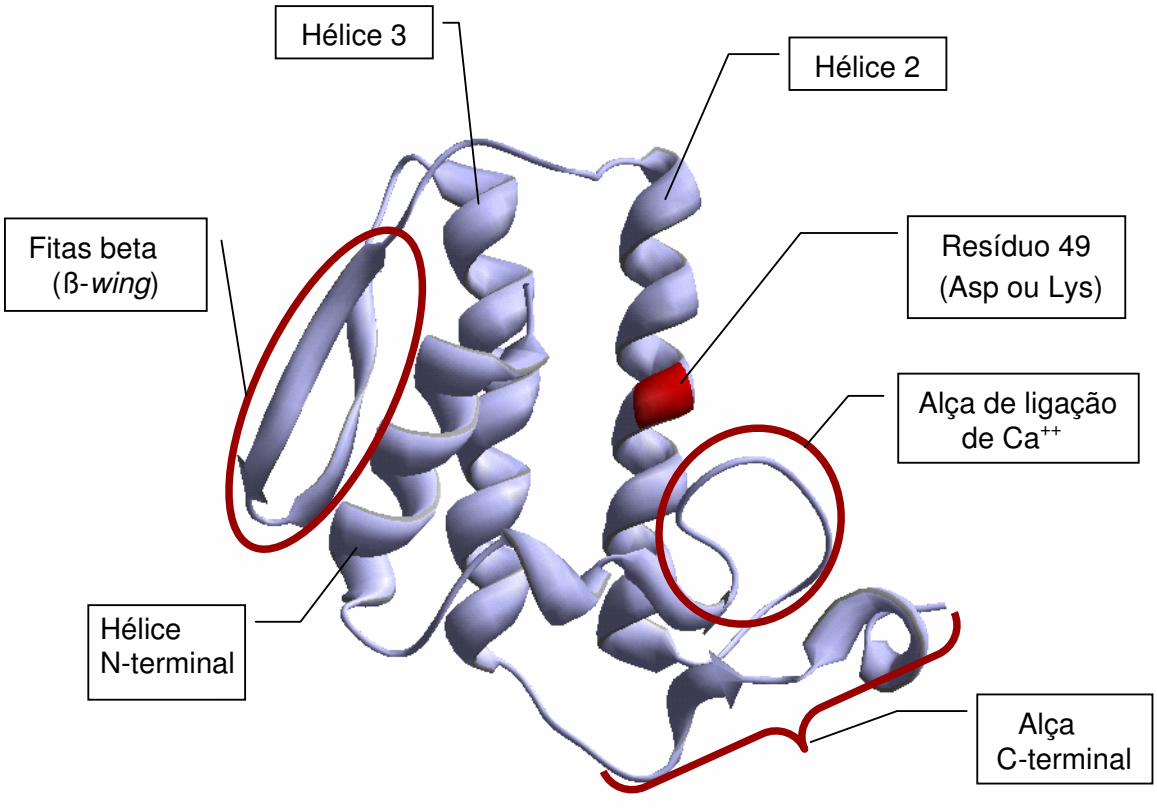


Figura 6: Esquema das regiões características de uma FLA_2 extracelular. Estrutura de FLA_2 de peçonha de *Notechis scutatus*, (Modelo de estrutura protéica obtido na base de dados *RCSB Protein Data Bank*⏭, indexada pelo código 1AE7. Figura gerada pelo autor, utilizando o software *SPDBViewer*⏭).

1.1.2.1 Atividade catalítica das FLA_2 extracelulares

Os determinantes da atividade catalítica das proteínas FLA_2 já são bem conhecidos. A capacidade de hidrólise é determinada pelo resíduo presente na posição 49 da proteína madura. As FLA_2 com alta atividade catalítica apresentam um ácido aspártico nesta posição. Este resíduo possibilita a ligação do íon cálcio à enzima, o que é essencial para a liberação dos produtos do sítio catalítico após a reação (Lee *et al.*, 2001).

A substituição mais freqüente do Asp49 entre as FLA_2 conhecidas é a lisina (Lys49). Esta mutação impede a ligação do íon cálcio à proteína e, desta forma, a hidrólise ocorre, porém a liberação dos produtos fica comprometida. Sem o íon cálcio, o

ácido graxo permanece no sítio catalítico e, desta forma, após a primeira catálise, o *turnover* da enzima não ocorre e ela torna-se inativa (Lee *et al.*, 2001). No **Apêndice 2** desta Dissertação, há um esquema detalhado do mecanismo catalítico das FLA_2 e das conseqüências funcionais da presença de um ácido aspártico ou de uma lisina na posição 49.

As FLA_2-Lys49 são classificadas como FLA_2, apesar de não terem atividade catalítica detectável, devido à alta similaridade de seqüência de aminoácidos e estrutura tridimensional entre elas e as FLA_2 cataliticamente ativas (Lomonte *et al.*, 2003). Ainda existem outras substituições do resíduo 49 menos freqüentes, como serina. As FLA_2 Ser49 também possuem baixa atividade catalítica (Polgar *et al*., 1996).

As FLA_2 extracelulares apresentam grande diversidade de substrato preferencial, determinado principalmente pela topologia de superfície da região de interação da proteína com o substrato. O mecanismo típico de catálise é conhecido como catálise interfaceal, na qual o substrato é preferencialmente uma superfície lipídica, como micelas ou bicamadas (Scott *et al*., 1990).

1.1.2.2 Diversidade funcional das FLA_2

As FLA_2 extracelulares estão envolvidas em processos fisiológicos como digestão, mediação inflamatória e sinalização celular; também atuam no sistema imunológico e participam de secreções tóxicas de diversos animais (Dennis, 1994). Além da atividade catalítica (que pode ou não estar presente, como foi visto acima), estas proteínas podem apresentar diversas ações biológicas tais como ação pró- e antiinflamatória, pró- e anticoagulante, indutora de edema, indutora de dano tissular,

inibidora de agregação plaquetária, cardiotóxica, convulsivante, hemolítica, mio-, neuro- e citotóxica (Kini, 2005).

Algumas das atividades apresentadas acima são verificadas na homeostase de sistemas e processos fisiológicos regulares, e são características das FLA_2 de diversos tecidos (**Figura 7**). Por outro lado, as FLA_2 que apresentam efeitos tóxicos são as que fazem parte de sistemas de defesa de alguns animais, principalmente serpentes e artrópodes, na forma de peçonhas. Estas FLA_2 são, freqüentemente, os principais componentes ativos dos venenos destes animais e a causa da alta toxicidade provocada nos outros animais e no homem quando expostos ao contato com tais secreções (Wickramaratna *et al.*, 2003; Chwetzoff *et al.*, 1989; Lambeau e Lazdunski, 1999). Em alguns casos, a toxicidade se restringe a apenas um ramo evolutivo, como no caso da toxina Phaiodactylipina, do escorpião *Anuroctonus phaiodactylus*, letal apenas para insetos (Valdez-Cruz *et al.*, 2004) . Estes fatos tornam as FLA_2 de peçonhas um objeto de estudos de interesse em muitos ramos de pesquisa farmacêutica, biomédica e biotecnológica.

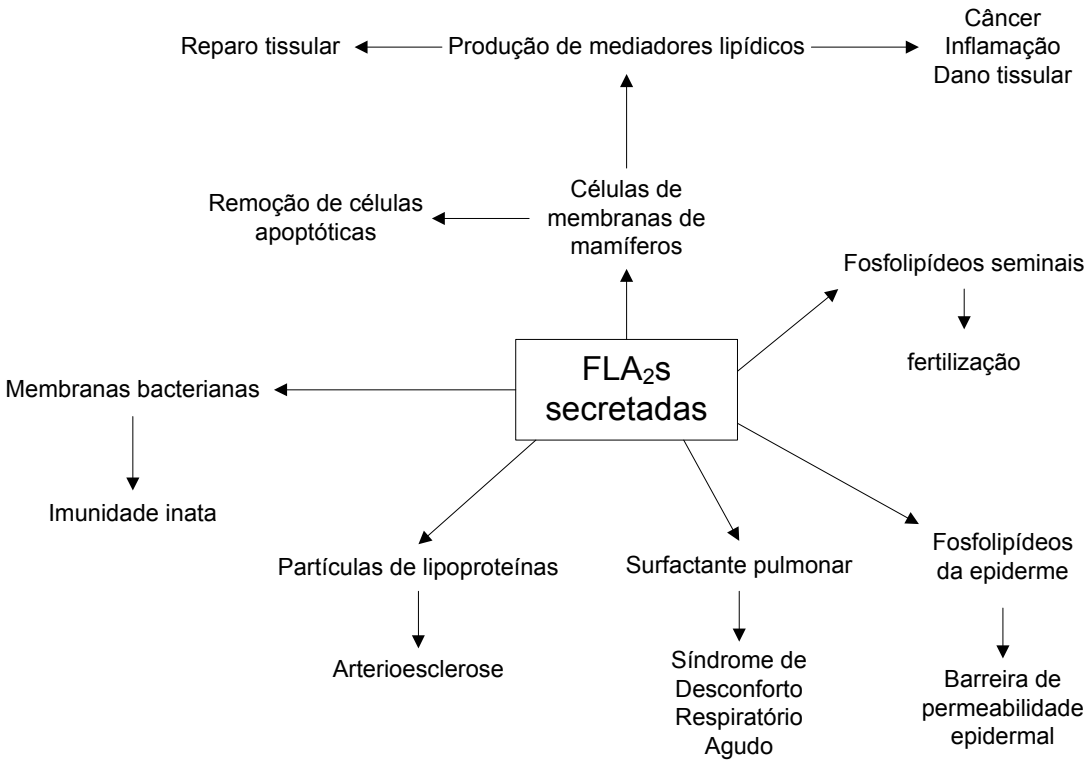


Figura 7: Processos fisiológicos dos quais as FLA_2 secretadas de mamíferos participam (adaptado de Murakami e Kudo (2004))

Acredita-se que as FLA_2 de peçonha de serpente tenham evoluído de certas enzimas pancreáticas que apareceram cedo na evolução dos vertebrados, por duplicação gênica e evolução acelerada dos éxons nas regiões codificantes da proteína madura (Ohno *et al*., 2003), adquirindo a capacidade de produzir efeitos biológicos distintos, como neurotoxicidade, cardiotoxicidade, miotoxicidade, efeitos pró- e anticoagulante e anti-agregação plaquetária. A especificidade das enzimas a vários tecidos e células foi alterada por substituições de aminoácidos, principalmente na superfície molecular (Kini e Chan, 1999). O alto conteúdo de cisteínas formadoras de pontes dissulfeto torna o núcleo destas proteínas muito mais compacto e estável que o das proteínas globulares, nas quais a estrutura é estabilizada principalmente por forças não covalentes e torna possível a

ocorrência de um número muito maior de mutações sem comprometer a viabilidade da proteína (Fry, 2005).

Por fim, cabe uma observação acerca da relevância das FLA_2 no estudo evolutivo das serpentes. Sendo um dos principais componentes tóxicos do mecanismo de defesa das serpentes peçonhentas, as FLA_2 fazem parte do complexo arsenal bioquímico que viabilizou a sobrevivência destes animais tão atípicos em termos de biomecânica, habilidades de deslocamento, defesa e predação (Ogawa *et al.*, 1996; Creer *et al.*, 2003, Li *et al.*, 2005).

Uma distinção importante para a compreensão deste trabalho e da literatura relativa aos diferentes papéis desempenhados pelas FLA_2 em sistemas vivos é a diferença entre função biológica e atividade biológica. Considerando as FLA_2 atuando no próprio organismo onde foram produzidas, considera-se que estas estão desempenhando uma *função* biológica, e quando atuando em organismos que não o de origem apresentam uma determinada *atividade* biológica – este é o caso dos efeitos tóxicos de peçonhas.

1.1.2.3 "Sítios farmacológicos"

Um fato intrigante em relação às diversas funções e atividades apresentadas pelas FLA_2 é, como já foi citado, a similaridade de seqüência e de estrutura existente entre elas. Considerando-se as FLA_2 secretadas de serpentes, observa-se de 40 a 99% de identidade na seqüência de aminoácidos entre distintas FLA_2. Quando se incluem as FLA_2 de mamíferos nessa análise, verificam-se também índices de identidade elevados entre várias das FLA_2 de mamíferos e serpentes. Além disto percebe-se considerável conservação da estrutura tridimensional (Valentin e Lambeau, 2000b) (**Figura 8**). Algumas regiões-chave, como a alça de ligação com cálcio, são conservadas mesmo

quando se consideram grupos evolutivamente mais distantes, como as FLA_2 de insetos (Nicolas *et al*., 1997).

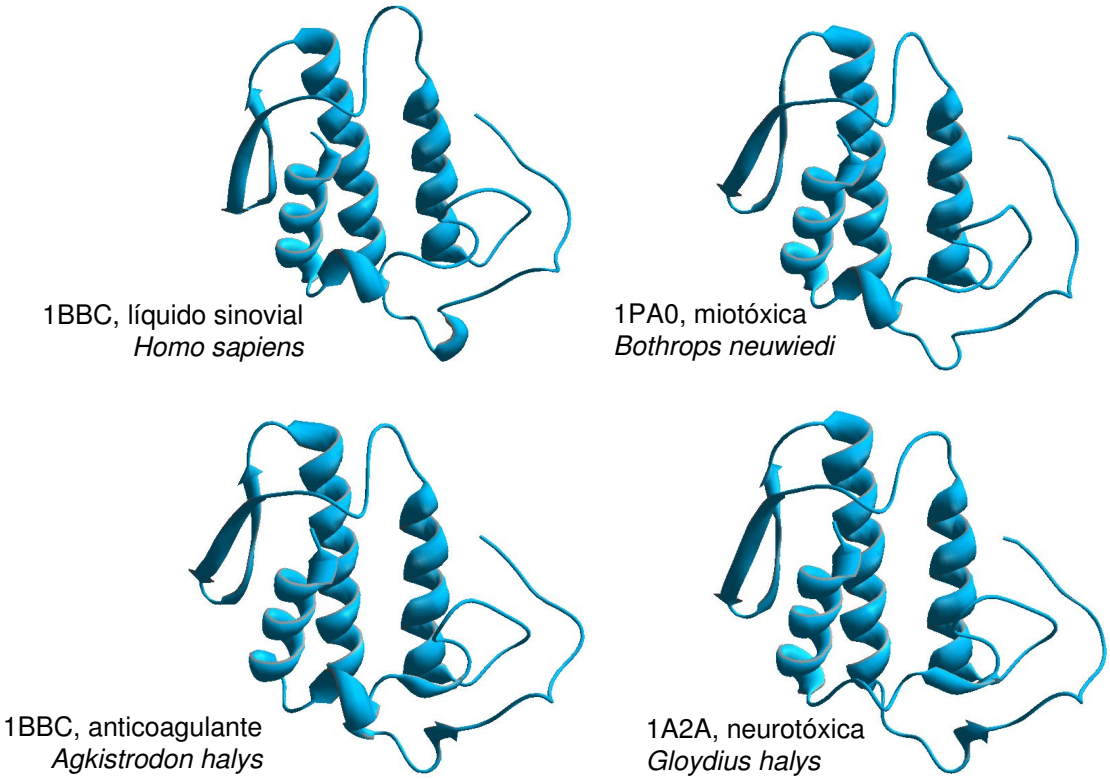


Figura 8: FLA_2 secretadas de diferentes organismos, destacando que a diversidade funcional não é evidente a partir da observação da estrutura, que se mostra bastante similar. O número de registro de cada estrutura no *RCSB Protein Data Bank*⏩, a principal atividade biológica e o organismo de origem estão indicados ao lado de cada modelo. (Figura gerada pelo autor, utilizando o software *SPDBViewer*⏩).

Uma explicação possível para esse fenômeno de alta diversidade funcional, contrastando com a alta similaridade das FLA_2 foi proposta por Kini (2003). O autor sugere a ocorrência de diversos sítios farmacológicos na superfície das FLA_2 que interagem com receptores-alvo, gerando respostas biológicas distintas. O receptor-alvo e o sítio farmacológico devem ser complementares um ao outro em termos de forma, potencial eletrostático, hidrofobicidade e forças de Van der Waals. No que concerne o

mecanismo de ação, é importante observar que estes sítios farmacológicos podem atuar independente ou cooperativamente. Uma FLA_2 pode exibir o mesmo efeito tóxico ou farmacológico por mais de um mecanismo, ligando-se a diferentes proteínas-alvo receptoras. Além disto, o efeito biológico das FLA_2 pode ser modulado em função da sua concentração em determinado tecido (Mora *et al*., 2005). Uma outra evidência que reforça esta hipótese é o fato de não haver necessidade de um aparato catalítico ativo para que estas atividades se manifestem (Lomonte *et al.*, 2003).

Já foram identificadas várias proteínas solúveis e de membrana que se ligam a diferentes FLA_2, dando força à hipótese de que estas proteínas poderiam atuar como ligantes de alta afinidade (Valentin e Lambeau, 2002).

1.1.2.4 FLA_2 neurotóxicas e miotóxicas

O presente trabalho foi conduzido tendo-se como foco o estudo de duas atividades biológicas apresentadas por FLA_2 de peçonha de serpente: neurotoxicidade e miotoxicidade.

As FLA_2 neurotóxicas têm atuação tanto pré-sináptica como pós-sináptica. As pré-sinápticas ligam-se às terminações nervosas axonais e podem causar um aumento ou bloqueio da liberação de neurotransmissores na fenda sináptica (Chang, 1985). As FLA_2 neurotóxicas pós-sinápticas ligam-se aos receptores acetilcolinérgicos nas extremidades dendríticas, causando um impedimento da despolarização e da conseqüente transmissão de estímulo nervoso (Kini, 1997).

As FLA_2 miotóxicas são caracterizadas por causar dano mionecrótico em tecido muscular esquelético de animais superiores por injeção intramuscular. Estas proteínas estão presentes em alta concentração nas peçonhas de serpentes viperídeas e crotalídeas

(Lomonte *et al.*, 2003). Além disto, algumas miotoxinas apresentam atividade neurotóxica. Estas são mais potentes que as FLA_2 que apresentam apenas atividade miotóxica e agem pré-sinapticamente na junção neuromuscular, além de gerar mionecrose expressiva em musculatura esquelética (Lomonte *et al*., 2003). Outros sintomas que acompanham a miotoxicidade local, *in vivo*, são um edema moderado (Angulo *et al*., 2000), hiperalgesia (Chacur *et al*., 2003) e liberação de citocinas pró-inflamatórias, como a interleucina-6 (Lomonte *et al*., 1993). A ocorrência de muitas FLA_2 miotóxicas sem atividade catalítica aponta para a existência de um mecanismo de ação miotóxica independente de catálise (Lomonte *et al*., 1999).

O estudo concomitante de neuro- e miotoxicidade é especialmente interessante pois existem evidências que sugerem alguma relação molecular entre estas atividades. Como foi citado acima, ambas podem estar presentes na mesma proteína, e há o aumento de potência da miotoxicidade nas FLA_2 miotóxicas que também apresentam neurotoxicidade pré-sináptica.

1.2 A Bioinformática na busca por padrões em seqüências biológicas

A aplicação de abordagens computacionais para auxiliar na anotação funcional de proteínas é relativamente recente. As tentativas iniciais mais conhecidas situam-se da segunda metade da década de 1960 a de 1970. Em 1965, Margareth Dayhoff compilou as primeiras matrizes de substituição, indicando as freqüências típicas de ocorrência de aminoácidos, nas proteínas já seqüenciadas até então. Estes dados serviram de base (e ainda hoje são utilizados, bem como variantes da mesma idéia) para o algoritmo de Needleman-Wunsch, para realização de Alinhamentos Múltiplos de Seqüências (AMS), cujos detalhes foram publicados em 1970.

Desde 1990, quando o denominado “Projeto Genoma Humano” foi lançado, até hoje, 16 anos depois, a evolução dos métodos de biologia molecular causou um descompasso entre a capacidade de obter dados brutos, a partir de fontes biológicas, e a capacidade de interpretá-los biologicamente, em um contexto sistêmico que permita a expansão das fronteiras do conhecimento e a geração de novas aplicações de interesse para o ser humano. Uma idéia aproximada da velocidade na qual os dados acerca de genes e proteínas tem sido produzidas pode ser visualizada na **Figura 9**, que exibe o crescimento dos dois principais bancos de dados utilizados em bioinformática, *GenBank* (seqüências primárias) e *RCSB Protein Data Bank* (estruturas tridimensionais).

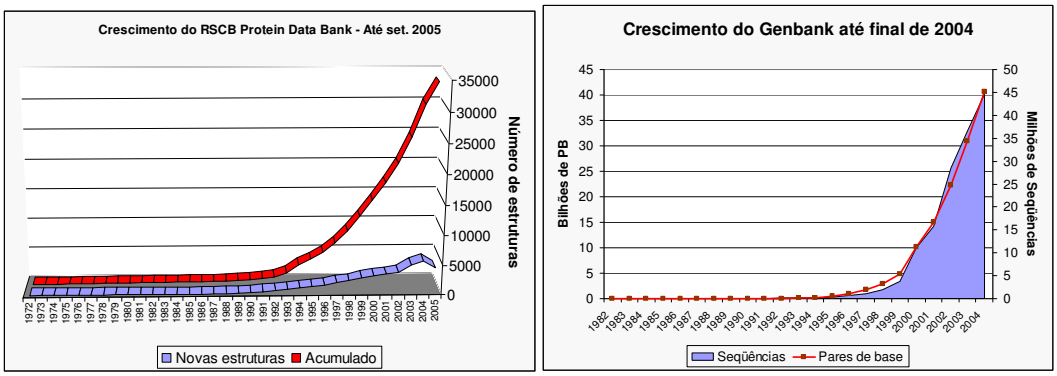


Figura 9 - Crescimento das principais bases de dados de bioinformática: *RSCB Protein Data Bank* e *GenBank*.

Até então, a caracterização funcional de proteínas dependia essencialmente de métodos de “bancada úmida” (ensaios *in vivo*, caracterizações físico-químicas utilizando reagentes) e equipamentos/métodos com baixa velocidade de geração de dados. Com a introdução de tecnologias e métodos tais como a reação em cadeia da DNA-polimerase (PCR), seqüenciadores automatizados e linhas de luz síncrotron para a análise de cristais de proteínas, tornou-se imperativa a necessidade de armazenar e distribuir de forma eficiente estes dados para a comunidade científica, bem como produzir ferramentas

computacionais que automatizassem parcial ou totalmente tanto o processo de correlação destes dados entre si, quanto com conhecimentos biológicos já existentes.

1.2.1 Diferentes pontos de partida, mesma meta: interpretar dados brutos e convertê-los em conhecimento aplicável

Essencialmente, as abordagens e ferramentas que usualmente recebem o título de "bioinformáticas" operam sobre algum dos seguintes tipos de dados (ou uma combinação dos mesmos):

Dados primários, obtidos diretamente a partir de material biológico, após tratamento adequado para transformá-los em uma abstração computável:

- Dados de seqüência: onde aminoácidos ou bases estão representados como uma seqüência de letras, na mesma ordem em que aparecem na seqüência primária;
- Dados de estrutura tridimensional: nesta classe situam-se tanto as informações de estrutura secundária ou seja, que regiões de uma proteína correspondem a que motivo secundário, quanto de estrutura terciária, na forma de coordenadas atômicas, representando a posição de cada átomo da molécula no espaço tridimensional;
- Dados físico-químicos: parâmetros físico-químicos tanto relacionados a cada aminoácido quanto a proteínas inteiras tais como ponto isoelétrico (pI), massa molecular, entre outros.

Dados secundários, derivados da análise dos dados primários, e normalmente disponibilizados conjuntamente com ferramentas que permitem sua correlação com dados primários:

- Domínios (ex. lectina tipo C) - Prodom⏭;
- Assinaturas de famílias de proteínas (ex. globinas) - PFam⏭, Prosite⏭;

- Modelos estatísticos tais como Matrizes de substituição de aminoácidos (ex. PAM, Gonnet), densidades de distribuição (ex. *Dirichlet Mixtures*) ou Modelos Ocultos de Markov;
- Motivos estruturais (ex. ß-α-ß) - CATH⏩;

Utilizando-se esses dados como base, diferentes abordagens podem ser realizadas e combinadas entre si com o objetivo de enfrentar o desafio de inferência da função de proteínas recém identificadas.

A maioria dos métodos utiliza uma abordagem comparativa: partindo de uma proteína de função desconhecida, procura-se relacioná-la a proteínas cujas funções já estejam caracterizadas ou, então, enquadrá-la em uma "família" que pode ser definida por diferentes parâmetros tais como homologia, estrutura tridimensional, perfis de hidrofobicidade, entre outros, e cujas características gerais sejam compartilhadas entre seus membros.

A partir da obtenção de uma nova seqüência de proteína, caso a mesma ainda não tenha sido estrutural e funcionalmente caracterizada, o ponto de partida usual para sua análise é a sua comparação com outras seqüências já estudadas e anotadas em busca de evidências de homologia, uma vez que seqüências com alta similaridade têm, usualmente, função similar. Além disto, se a seqüência de busca apresentar um nível de identidade superior a 40% com outra cuja estrutura seja conhecida, pode-se também supor que suas estruturas sejam similares e então construir um modelo tridimensional por homologia. (Martí-Renom *et al.*, 2000)

O método mais simples de comparação entre dados de seqüência é o alinhamento de seqüências, que pode ser de dois tipos: par-a-par ou alinhamento múltiplo de seqüências (AMS).

No primeiro, apenas duas seqüências são comparadas entre si, e o resultado é o alinhamento que melhor alinha regiões semelhantes ou comuns entre ambas as seqüências, atribuindo um valor indicativo de similaridade. A ferramenta mais utilizada atualmente para esse tipo de pesquisas é o *Basic Local Alignment Search Tool* ou, simplesmente, BLAST ⏩ (Altschul, 1990; McGinnis & Madden, 2004) que, embora não tenha a mesma precisão de outros métodos, apresenta desempenho adequado para a busca de similaridades em bases de seqüências com as dimensões das atuais.

No caso do AMS, um grupo de seqüências é fornecido como entrada para o algoritmo de alinhamento, e as seqüências são alinhadas umas em relação às outras, em busca de um alinhamento que represente eventos mutacionais de inserção, deleção, substituição, etc., ou seja, estabeleça uma hipótese de evolução utilizando as seqüências de entrada. Nesse método, buscam-se conjuntos de aminoácidos que aparecem na mesma ordem ou em padrões similares de distribuição entre as seqüências que estão sendo comparadas e, conforme os algoritmos próprios de cada método, são gerados valores que expressam a medida da similaridade entre as seqüências comparadas. Um dos mais conhecidos programas para esta tarefa é o ClustalW (Thompson *et al.*, 1994; Chenna *et al.*, 2003) ⏩.

Uma vez que é comum na natureza a estrutura tridimensional de uma proteína estar estreitamente relacionada com a sua função[1], uma abordagem análoga à comparação de seqüências pode ser utilizada, mas utilizando-se dados da estrutura tridimensional como entrada. Diferentes abordagens computacionais são utilizados com o objetivo de se encontrar proteínas com maior similaridade a partir da estrutura de interesse como, por exemplo, os baseados no alinhamento de carbonos-alfa (Ex. MAMMOTH⏭), baseados em elementos da estrutura secundária (Ex. SSAP⏭), ou utilizando matrizes de distância entre os carbonos-alfa de uma proteína (Ex. DaliLite⏭).

Na análise comparativa também enquadram-se as ferramentas operadas sobre dados secundários ou padrões. Por "padrão" entende-se qualquer característica observada em um grupo específico de proteínas e que pode ser utilizada para discriminá-lo de outros grupos. Como exemplos de padrões podem-se citar motivos de seqüência ou estruturais, domínios, perfis estatísticos de composição de aminoácidos.

Computacionalmente, esses padrões podem ser representados de diversas formas, tais como: expressões regulares (ex. Prosite⏭, representando domínios e famílias de proteínas); modelos matemáticos (ex. PFam⏭, que utiliza MOMs para representar famílias de proteínas); bibliotecas de motivos estruturais (ex. SCOP, CATH⏭, que agrupa proteínas em famílias, de acordo com ocorrência de padrões estruturais); entre outros. Diversas ferramentas permitem utilizar os próprios padrões como ponto de partida na busca por seqüências ou estruturas similares como, por exemplo, a ferramenta HMMSEARCH, que faz parte do pacote HMMER⏭ e que recebe, como entrada, um MOM representando um grupo ou família de proteínas, ou a ferramenta

[1] Um contra-exemplo deste padrão são as próprias fosfolipases A_2 sem atividade catalítica, mas que apresentam atividade tóxica. Isto lhes confere uma função que não está diretamente relacionada com a estrutura, enquanto esta última mantém-se ainda como reminescência do ancestral com atividade catalítica.

SCANPROSITE⏩, que permite que seja informada uma expressão regular como entrada para a busca.

Outra abordagem possível busca enriquecer o conjunto de dados associados à seqüência de interesse, por meio do acréscimo de informações físico-químicas e da aplicação de métodos de predição de estrutura secundária ou terciária, o que pode fornecer pistas importantes ao pesquisador quanto à função desempenhada pela proteína (ex. STING Protein Dossier⏩).

Existem também abordagens não-triviais, combinando uma ou ambas as abordagens já expostas com técnicas de inteligência artificial, tais como redes neurais – RONN⏩ (Yang *et al.* 2005); YASPIN ⏩ (Lin *et al.*, 2005) - , aprendizado de máquina (Bazzan *et al.*, 2002; Larrañaga *et al.*, 2006) e algoritmos genéticos (Contreras-Moreira *et al.* 2003; Zviling *et al.*, 2005).

1.3 Modelos Ocultos de Markov (MOMs)

Os MOMs são uma metodologia estatística cuja teoria básica foi publicada por Baum *et al*. entre o fim da década de 1960 e o início da década de 1970 (Baum e Petrie, 1966; Baum e Egon, 1967; Baum e Sell, 1968; Baum *et al*., 1970; Baum, 1972). Inicialmente, foi um modelo que recebeu atenção apenas dos pesquisadores da área da matemática. Rabiner (1989) relata a utilização dos MOMs em programas de reconhecimento de voz, ainda na década de 70. A partir da década de 80, cresceu a aplicação dos MOMs em diversas áreas nas quais se desejava automatizar o processo de detecção e representação de padrões de informação. Alguns exemplos ilustram a grande faixa de aplicações possíveis deste modelo teórico:

- classificação automática de programas de televisão, a partir da comparação entre imagens e MOMs que representam diferentes perfis (de imagens) típicos de diferentes tipos de programa (Lu *et al*., 2001);
- reconhecimento de rostos (Nefian e Hayes, 1998);
- detecção e reconhecimento de gestos (Liu e Lovell, 2003);
- previsão de conflitos sociais (Schrodt, 2000);

O MOM é considerado "oculto" pois representa um padrão de informação que possui determinantes desconhecidas. Não se sabendo quais são e como se comportam as variáveis que regem o padrão observado, o modelo vai basear-se na informação fornecida ou observada para gerar uma representação estatística do comportamento deste padrão. Para facilitar o entendimento desta explicação, pode-se exemplificar com uma situação em que *não* é necessária a utilização de um modelo "oculto", ou seja, o arremessar de dados não viciados. Sabe-se que cada face do dado tem 1/6 de probabilidade de cair voltado para cima; neste caso, não é necessária a observação de uma seqüência de

arremessos para conhecermos estas probabilidades, pois as variáveis que as regem são conhecidas (Rabiner, 1989).

Os modelos utilizados nas Ciências Biológicas são, em geral, estatísticos e baseados em informações observadas devido à complexidade das variáveis que compõem os sistemas biológicos. Por isto, das diversas áreas de pesquisa nas quais os MOMs e suas variantes podem ser utilizados com eficácia, as Ciências Biológicas são uma das mais expressivas. Processos ecológicos (Li *et al.*, 2001), epidemiologia (Cooper e Lipsitch, 2004), genética de populações (Lazzeroni e Lange, 1997), análises de dados bioinformáticos (Viklund e Elofsson, 2004; Truong e Ikura, 2002; Qian e Goldstein, 2003; Bateman *et al.*, 2004), por exemplo, são fortemente ligadas à aplicação de métodos teóricos, tanto para maximizar a capacidade de extrair conhecimento a partir dos dados observados, quanto para gerar modelos característicos e preditivos destes fenômenos.

Como foi revisado na seção anterior, a bioinformática tem um amplo espectro de objetos de estudo. Dados de seqüência de nucleotídeos e aminoácidos são muito adequados à representação como informação e, a partir de seu tratamento estatístico e informático, inferências em relação à funcionalidade e à evolução são possíveis.

Hoje existem diversas ferramentas de bioinformática que permitem automatizar a aplicação dos MOMs ao tratamento das informações de seqüência e estrutura de proteínas. A sua utilização em estudos funcionais e evolutivos por análise de seqüências biológicas tem crescido e produzido resultados confiáveis (Krogh *et al*., 1994; Baldi *et al*., 1994; Karplus *et al*., 1998; Troung e Ikura, 2002; Qian e Goldstein, 2003). Por exemplo, os MOMs têm sido empregados na detecção de regiões com especial significância em genomas, tais como genes (Stanke e Waak, 2003; Krogh 1997), sítios de

splicing (Pertea *et al.* 2001) e motivos de particular significância em proteínas e genes (Bateman *et al*., 2004).

1.3.1 Fundamentos dos MOMs aplicados à bioinformática

Os MOMs, quando aplicados a um AMS, permitem a construção de perfis estatísticos das probabilidades de ocorrência de cada aminoácido e de eventos de deleção ou inserção em cada posição do alinhamento (Durbin *et al*., 1998). As probabilidades finais geradas pelo modelo são o resultado (i) da aplicação de uma teoria estatística com forte base matemática às sequências geradoras do modelo (Eddy, 1998, 2001 e 2004); e (ii) da introdução de conhecimento *prévio* sobre as substituições usuais entre aminoácidos, por características de tamanho e comportamento físico-químico (Sjölander *et al*., 1996) .

No caso dos MOMs, a introdução de conhecimento prévio ou, *a priori*, acerca de distribuições de aminoácidos, é feita por uma abordagem conhecida como Distribuições de Dirichlet. Esta técnica é utilizada para correções nas probabilidades de ocorrência de aminoácidos, de acordo com o padrão observado em cada posição e correlações com os padrões de distribuição de seqüências de aminoácidos que ocorrem universalmente nas proteínas (Sjölander *et al*., 1996).

As Distribuições de Dirichlet estão baseadas na informação de bancos de dados de AMS. A partir desta informação, são identificados os padrões universais ou prototípicos de distribuição de aminoácidos *por posição* de alinhamento. Além disto, são identificados os diferentes *contextos* moleculares possíveis e isto é levado em conta na correção das probabilidades calculadas a partir dos dados fornecidos. Por exemplo, o aminoácido isoleucina é muito comum em fitas-beta em núcleos de proteínas e, neste contexto, valina

e leucina a substituem com freqüência – se identificado este contexto em uma posição, as probabilidades de valina e leucina serão corrigidas de forma a aumentar sua probabilidade de ocorrência. Assim, as correções efetuadas pelas Distribuições de Dirichlet são posição- e contexto-dependentes (Sjölander *et al*., 1996).

Além disso, a influência das Distribuições de Dirichlet acontecem em função da quantidade de seqüências utilizada no conjunto de entrada. Quanto maior a amostra, menor a influência das Distribuições de Dirichlet. Porém, se a amostra for pequena, a informação *a priori* terá mais peso nas probabilidades finais fornecidas pelo MOM; ou seja, o modelo pode ser bem representativo ainda que poucas sequências tenham sido utilizadas em sua construção (Sjölander *et al*., 1996).

Dentre as principais vantagens dos MOMs em aplicações de bioinformática, pode-se citar o embasamento estatístico: os MOMs estão apoiados em teorias estatísticas amplamente utilizadas e cujas propriedades são bem descritas e conhecidas, fornecendo uma base confiável para ferramentas que os utilizem. (Eddy, 1998).

Uma importante limitação da aplicação de MOMs é que este tipo de modelo desconsidera relações de mais alta ordem que a posição da seqüência de aminoácidos. Para o cálculo das probabilidades de uma posição, o modelo leva em conta as probabilidades da posição anterior apenas e, por esta razão, não considera a possível relevância de pequenas regiões de aminoácidos que apresentem uma distribuição específica, ou aminoácidos que formem, eventualmente, uma região importante na estrutura tridimensional da proteína, mas que não sejam contíguas na seqüência primária (Eddy, 2001).

1.3.2 MOMs aplicados ao reconhecimento de atividade biológica de FLA_2

Neste trabalho, a escolha dos MOMs é justificada pela importância de se utilizar de um modelo estatístico sofisticado o suficiente para captar diferenças sutis entre proteínas com elevada similaridade de seqüência, pertencentes à mesma família. Além disto, a alta capacidade de generalização, decorrente da aplicação de informação *a priori*, permite que os modelos gerados tenham boa eficácia na predição automática de função de qualquer proteína submetida à pesquisa, ainda que não seja possível a construção dos modelos a partir de um grande número de seqüências (Sjölander *et al*., 1996).

2. Objetivos

Pelo presente trabalho, tive como objetivo a elaboração de uma metodologia capaz de predizer determinadas propriedades funcionais de uma família de proteínas derivadas de peçonha de serpentes e reportadas como pertencentes à família de fosfolipases A_2. Como objetivos específicos, selecionei, como atividades a serem testadas, a neurotoxicidade e a miotoxicidade destas proteínas, aplicando a elas o método desenvolvido. Os principais passos do método são descritos a seguir:

1) detecção de padrões de conservação de aminoácidos em seqüências primárias de FLA_2 de uma mesma atividade biológica, permitindo relacionar tais padrões conservados com a atividade;

2) conversão destes padrões em Modelos Ocultos de Markov - uma representação ao mesmo tempo rica em conteúdo informacional e facilmente utilizável em ferramentas de bioinformática como dado de entrada;

3) elaboração de uma validação estatística que permitisse selecionar os motivos conservados e, ao mesmo tempo, *discriminativos* de cada atividade.

Com os motivos selecionados na validação estatística, ainda visei a elaboração de uma ferramenta de identificação automática das atividades estudadas. Além disto, busquei colaborar para um melhor entendimento dos mecanismos moleculares e evolutivos envolvidos nas diversidades funcionais apresentadas pelas FLA_2 de peçonha de serpente, proteínas estruturalmente tão similares e, por isto, tão intrigantes.

3. Resultados

Neste capítulo estão apresentados dois manuscritos de artigos científicos gerados a partir dos resultados do desenvolvimento da metodologia e sua aplicação no reconhecimento das duas atividades biológicas das FLA_2 de serpente. O primeiro manuscrito foi publicado no periódico *Advances in Bioinformatics and Computational Biology - Lecture Notes in computer Science* em março de 2005, enquanto o segundo está na forma de manuscrito revisado, a ser submetido ao periódico Toxicon.

Para todo o tratamento envolvendo os MOMs deste trabalho, foi utilizada a ferramenta HMMER⏩ (um acrônimo para ***H****idden* ***M****arkov* ***M****odel Build****er***). Disponibilizada gratuitamente pelo pesquisador e desenvolvedor Sean R. Eddy, trata-se de um pacote de programas para construção e tratamento de MOMs em um contexto de bioinformática. Utilizando-se os programas deste pacote, é possível construir MOMs a partir de AMS, testar seqüências contra MOMs já existentes, além de várias outras funcionalidades.

Nos respectivos manuscritos estão pormenorizadamente descritas a metodologia aplicada, os resultados obtidos e sua discussão, de forma que estas seções não serão repetidas na Dissertação.

3.1 Aplicação de metodologia para deteção de mio- e neurotoxicidade em FLA_2

Artigo 1 – Pazzini, F. ; Oliveira, F. S. ; Guimaraes, J. A. ; Amorim, H.L.N. (2005)

Prediction of myotoxic and neurotoxic activities in Phospholipases A_2 from primary sequence. *Lecture Notes in Computer Science*, Berlin 3594, 194-197.

Este artigo, descreve a metodologia criada e a capacidade preditiva alcançada na identificação das atividades neuro- e miotóxica. Por ser uma publicação voltada ao público das Ciências da Computação, o enfoque foi o de estabelecer um protocolo eficiente de reconhecimento de função biológica a partir de informação de seqüência de aminoácidos.

Prediction of Myotoxic and Neurotoxic Activities in Phospholipases A2 from Primary Sequence Analysis

Fabiano Pazzini[1], Fernanda Oliveira[1], Jorge A. Guimarães[1] and Hermes Luís Neubauer de Amorim[1,2]

[1] Biotechnology Center, Federal University of Rio Grande do Sul – UFRGS, 91501-970, RS, Brazil
fabianopasin@cbiot.ufrgs.br
[2] Department of Chemistry, Lutheran University of Brazil - ULBRA, 92420-280, RS, Brazil

Abstract. We developed a methodology to predict myotoxicity and neurotoxicity of proteins of the family of Phospholipases A2 (PLA2) from sequence data. Combining two bioinformatics tools, MEME and HMMER, it was possible to detect conserved motifs and represent them as Hidden Markov Models (HMMs). In ten-fold cross validation testing we have determined the efficacy of each motif on prediction of PLA2 function. We selected motifs whose efficacy in predict function were above 60 % at the Minimum Error Point (MEP), the score in which there are fewest both false positives and false negatives. Combining HMMs of the best motifs for each function, we have achieved a mean efficacy of 98 ± 4 % on prediction of myotoxic function and 77.4 ± 4.8% on prediction of neurotoxicity. We have used the results of this work to build a web tool (available at www.cbiot.ufrgs.br/bioinfo/phospholipase) to classify PLA2s of unknown function regarding myotoxic or neurotoxic activity.

1 Introduction

One of the most important tasks of the bioinformatics is to give meaning to the large amount of data from genomic and proteomic projects. Part of this task comprises automatic prediction of the function of proteins. However, the most currently used algorithms and databases (such as BLAST [1], PFAM [2] and PROSITE [3]) strive to classify protein sequences into broad families, which not necessarily share the same biological function. The Phospholipase A2 (PLA2) family (E.C. 3.1.1.4), initially classified according to its ability to catalyze the cleavage of membrane phospholipids, represents an interesting challenge. Despite the high level of sequence similarity and structure conservation of the family [4], the proteins of this group are involved in distinct biological functions such as digestion, cell signaling and inflammation. They also present myotoxic, neurotoxic and cytotoxic activities.

Based on the analysis of conserved amino acids and protein motifs and using Hidden Markov Models (HMMs) to capture the particular characteristics of the Multiple Sequence Alignments (MSAs), we developed a methodology to discriminate

J.C. Setubal and S. Verjovski-Almeida (Eds.): BSB 2005, LNBI 3594, pp. 194-197, 2005.

between neurotoxic PLA2 (nPLA2) and myotoxic PLA2 (mPLA2). Based in the results of this work we provided a tool, now available at www.cbiot.ufrgs.br/bioinfo/phospholipase, which allows the identification of PLA2s displaying myotoxicity and neurotoxicity. To our knowledge no other method is available allowing classification of the biological function of these PLA2s.

2 Methodology

We collected sequences which were used to build and test the models representing the biological function of interest. For each biological function there are two main sets of sequences: one represents sequences with biological function and other with sequences without the function (negative control).

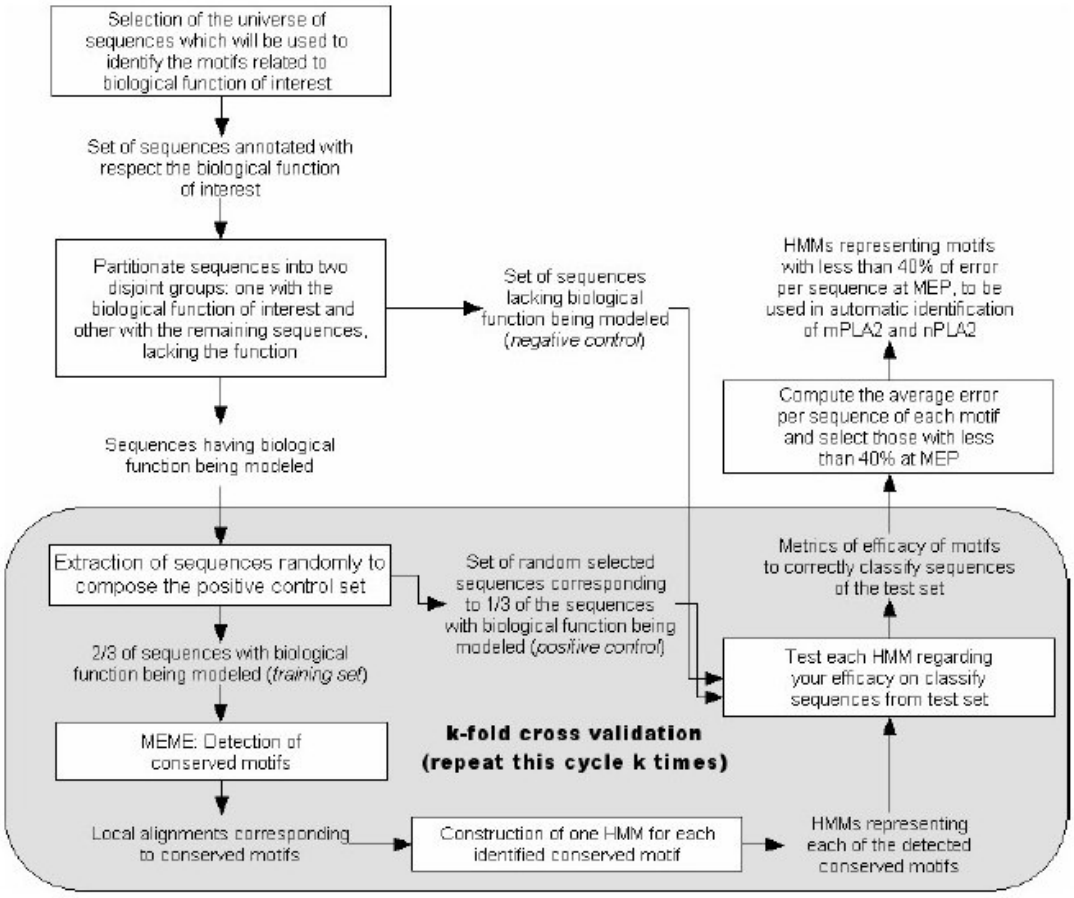


Fig. 1. Flow diagram of the methodology to detect sequence motifs specific to some biological function.

In our approach we used MEME [5] to detect conserved motifs and HMMER [6] to construct HMMs of each motif found. The complete process is depicted in **Fig. 1**.

On each iteration, the sequences that have the biological function are split in training set (used to construct HMMs) and positive control (which together with negative control forms the *test set*).

The first efficacy metric that is calculated is the *Error Rate* (ER) at each score. It shows how well some HMM classify the sequences of the test set:

$$ER = (FP + FN) / \text{size of test set} \quad (1)$$

FP represents false positives, and FN, false negatives.

Based on the ER at each score, it is possible to determine the *Minimum Error Point* (MEP), the score which ER value is minimum. The efficacy value at the MEP is the best possible for the motif. Note that we define Prediction Accuracy (PA) as the complement of the error rate, *i.e.*,

$$PA + ER = 1 \quad (2)$$

The *coverage* measures how much of the true positives (TP) were correctly classified above some score. It is calculated as the ratio between TP above some score and the total number of sequences of the positive control.

As the biological function can be associated to more than one motif, all motifs with PA greater than 60% were selected to be used in function prediction.

In the case of occurrence of multiple motifs to detect the same biological function, it is possible to combine them, improving the PA of the respective function. If each of these motifs recognizes different subsets of true positives, combining their results will increase the coverage, but the impact on PA must be calculated considering both TP and FP of the maximal set composed by all sequences with score above MEP of the respective motif.

3 Results

Table 1 shows the motifs detected and the corresponding average accuracy values, computed after 10-fold cross validation process.

Table 1. Motifs with mean predictive accuracy (PA) greater than 60 % at MEP

Group	Motif	MEP score	PA at MEP (%)	Coverage at MEP (%)
mPLA2	N-terminal	28.03	86	72
mPLA2	C-terminal region	18.17	80	60
nPLA2	N-terminal	82.67	63.5	39.8
nPLA2	near catalytic site	47.67	67.3	75

In order to improve the PA of the final model, all detected motifs related to the same biological function were combined, maximizing their capability to correctly recognize their target biological function. The parameters of the 10-fold cross validation and the mean efficacy for the best motifs are in **Table 2**.

Table 2. Parameters and results of k-fold cross validation

Group	k	Size of functional set	Size of negative control	Number of motifs with ER < 40%	Coverage for combined motifs (%)	Best mean PA for combined motifs
mPLA2	10	20	8	2	96,0	98,0±4,0%
nPLA2	10	16	13	2	78,6	69.5±7.6%

4 Concluding Remarks

The use of conserved motifs, instead the entire sequences, to construct each HMM helps to minimize the bias induced by the small training sets [7]. Additionally the utilization of Dirichlet Mixtures by HMMER also increases the generalization power of the resulting HMM [8].

Considering the biochemical and pharmacological importance of the PLA2s, especially those exhibiting toxicological effects, we expect that the methodology described here can contribute for the advance of the knowledge in this area of research.

References

1. Scott McGinnis, Thomas L.Madden: BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucl. Acids. Res.,Vol. 32. (2004) W20-W25
2. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L., Studholme, D.J., Yeats, C., and Eddy, S.R.: The Pfam protein families database. Nucl. Acids. Res.,Vol. 32. (2004) D138-D141
3. Sigrist C.J.A., Cerutti L., Hulo N., Gattiker A., Falquet L., Pagni M., Bairoch A., and Bucher P.: PROSITE: A documented database using patterns and profiles as motif descriptors. Briefings in Bioinformatics,Vol. 3. (2002) 265-274
4. Manjunatha Kini, R.: Excitement ahead: structure, function and mechanism of snake venom phospholipase A2 enzymes. Toxicon,Vol. 42. (2003) 827-840
5. Timothy L. Bailey and Charles Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California, 1994.
6. Eddy, S.R.: Profile hidden Markov models. Bioinformatics, Vol. 14. (1998) 755-763
7. Grundy, W.N. *et al*: Meta-MEME : motif-based hidden Markov models of protein families. CABIOS, Vol. 13. (1997) 397-406
8. Haussler, D. *et al*: Dirichlet Mixtures: A Method for Improving Detection of Weak but Significant Protein Sequence Homology. CABIOS, Vol. 12. (1996) 327-345

3.2 Predição de neuro- e miotoxicidade em FLA_2 enfocando a busca por seus "sítios farmacológicos"

Artigo 2 – Pasin, F., Oliveira, F. S., Guimarães, J.A., Amorim, H.L.N., 2006. **Computational analysis of the determinants of the myotoxic and neurotoxic activities of snake venom phospholipases A_2: is it possible to predict these specific functions from sequence data?**

Este artigo explicita a metodologia desenvolvida em sua forma final, incorporando vários aprimoramentos que permitiram alcançar melhores indicadores de eficácia metodológica e uma melhor padronização nos procedimentos realizados, além de um aprofundamento na compreensão dos dados obtidos, determinantes para a caracterização das atividades biológicas das FLA_2.

Como o manuscrito ainda não foi submetido, sua formatação segue o mesmo padrão do corpo da dissertação. A única exceção são as referências bibliográficas que estão delimitadas por colchetes, para evitar confusão com as referências incluídas no corpo da Dissertação, que se encontram delimitadas por parênteses.

Computational analysis of the determinants of the myotoxic and neurotoxic activities of snake venom phospholipases A_2: is it possible to predict these specific functions from sequence data?

Fabiano Pasin[1], Fernanda Oliveira[1], Jorge Almeida Guimarães[1], Hermes Luís Neubauer de Amorim[1, 2]

[1]Centro de Biotecnologia - Universidade Federal do Rio Grande do Sul (UFRGS), Av. Bento Gonçalves, 9500. Bloco IV, Prédio 43421, Lab 202. Porto Alegre, RS, Brazil P.O.Box 15.005.
[2]Departamento de Química - Universidade Luterana do Brasil (ULBRA), Av. Farroupilha 1001, Prédio 14, Sala 420C - Canoas, RS – Brazil. 92450-900.

Running Title: Toxicity Prediction of Snake Phospholipases A_2

Address correspondence to: Hermes Luís Neubauer de Amorim, Centro de Biotecnologia - Universidade Federal do Rio Grande do Sul (UFRGS), Av. Bento Gonçalves, 9500. Bloco IV, Prédio 43421, Lab 202. Porto Alegre, RS, Brazil. P.O.Box 15.005, Tel: 0055-51-33167770; Fax: 0055-51-33167309; E-mail: hamorim@cbiot.ufrgs.br

Keywords: Hidden Markov Model; phospholipase A_2; snake venom; protein motif; biological activity prediction

Abbreviations: PLA_2, Phospholipase A_2; $mPLA_2$, myotoxic phospholipase A_2; $nPLA_2$, neurotoxic phospholipase A_2; *sv*PLA_2, snake venom Phospholipase A_2; HMM, Hidden Markov Model; MEP, Minimum Error Point; MSA, Multiple Sequence Alignment; PA, Predictive Accuracy; TP, True Positive, FP, False Positive, TN, True Negative.

Summary

Despite the high sequence and structural similarities, Phospholipase A_2 (PLA_2) protein superfamily exhibits a large spectrum of biological functionalities. Using sequence data and bioinformatics tools, we have developed a method to predict myotoxicity and neurotoxicity of the proteins of the snake venom PLA_2 family (*sv*PLA_2). Combining the bioinformatics tools MEME and HMMER, it was possible to identify conserved protein motifs and represent them as Hidden Markov Models (HMMs). After a ten-fold cross validation test, we were able to determine the average efficacy of each motif, allowing the prediction of *sv*PLA_2 function, and select those able to correctly predict the function of at least 60% of the sequences of the test

set. Using the HMMs built from the selected motifs for each function, we have achieved a mean efficacy of 99.34 ± 0.66 % on the prediction of myotoxic activity and 66.72 ± 3.00 % for neurotoxicity. Two distinct motif distributions, named motif pattern I and motif pattern II, could be observed in the group of neurotoxic and myotoxic *sv*PLA_2. Differences between motif patterns for each *sv*PLA_2 group were detected in N- and C-terminal regions. Additionally, other sequence motifs were recognized in regions known to be important for the expression of *sv*PLA_2 function. More specifically, the catalytic region, the β-wing and the α-helix 3 of such protein class. The results detailed in this work were used to build up a web tool (available at www.cbiot.ufrgs.br/bioinfo/phospholipase) to indicate, for any PLA_2 query sequence, the occurrence of motifs related to either neurotoxic or myotoxic activities.

Introduction

Phospholipases A_2 (PLA_2, EC 3.1.1.4) are enzymes that catalyze the hydrolysis of the sn-2 fatty acyl bond of phospholipids to release free fatty acids and lysophospholipids. The reaction products generated from PLA_2 enzymatic activity play a key role in various biological processes, from homeostasis of cellular membranes to lipid digestion and production of lipid mediators [Murakami & Kudo, 2004]. Secretory PLA_2 (*s*PLA_2) can be divided into several groups based on their primary sequences and disulfide bond arrangement, most of them classified into groups I and II [Valentin & Lambeau, 2000]. For example, members of group I PLA_2 are found in the venoms of *Elapidae* (*Elapinae* and *Hydrophiinae*) snakes, whereas members of group II PLA_2 include those from viperid (*Viperinae* and *Crotalinae*) snake venoms and the group II of mammalian *s*PLA_2 which is found in inflammatory exudates. Members of group II PLA_2 are divided into two subgroups: (a) the D49-PLA_2 which has an aspartic acid residue at position 49 and high catalytic activity on synthetic phospholipid substrates and (b) the K49-PLA_2, which has a lysine residue at position 49 and very low or no hydrolytic activity on synthetic substrates. In the D49-PLA_2, the critical

step in the mechanism of phospholipid hydrolysis is the nucleophilic attack of a water molecule over the sn-2 ester bond of the phospholipid. The tetrahedral intermediary formed in this step is stabilized by a calcium ion, which is coordinated by an aspartic amino acid residue (Asp49), one water molecule and the backbone oxygens of Gly30, Trp31 and Gly32 [Scott, D. L., 1990]. The Asp49 residue has the key role to assist the stabilization of the tetrahedral intermediate during the catalysis and to contribute for the increase of nucleophilicity of His 48. In the K49-PLA_2, the substitution of the aspartate residue by lysine causes steric hindrance of Ca^{2+} binding, because of the ε–amino group of the Lys49. The consequence of loss of Ca^{2+} binding in the K49-PLA_2 is the observed lack of the hydrolytic activity against both synthetic and natural phospholipids. However, as described bellow, the range of known biological effects of K49-PLA_2 is broader than those of D49-PLA_2 [Ownby *et al.*, 1999; Lomonte *et al.*, 2003].

Among the proteins of the PLA_2 superfamily, the group of snake venom PLA_2 (*sv*PLA_2) represents an interesting puzzle to scientists: despite the high structural and amino acid sequence similarity among its representatives, there is a great variability regarding the biological effects they are able to produce. Moreover, while mammalian *s*PLA_2 are generally non-toxic and do not induce potent pharmacological effects, *sv*PLA_2 are toxic. The toxicity of *sv*PLA_2 is resultant from the broad spectrum of biological effects induced by these proteins, including several activities such as pro- and anticoagulant, convulsivant, antiplatelet, edema-inducing, tissue-damaging, as well as myotoxic and neurotoxic effects. This variability in function was rationalized in a model [Kini, 2003] that points to the occurrence of multiple "pharmacological sites" located at *sv*PLA_2 surfaces that, operating independently or cooperatively, interact with target receptors, generating distinct biological responses. The receptor target sites and the "pharmacological sites" should be complementary to each other in terms of shape, electrostatics, hydrophobicity and van der Waals contact surfaces. It is also

important to consider that different *sv*PLA_2 can exhibit the same biological effects through different mechanisms, and hence binding to different target receptors. Yet, the biological effects of *sv*PLA_2 can change as a function of protein concentration [Mora *et al.*, 2005]. Considering the biological and the toxicological relevance of the *sv*PLA_2 family, the facts above mentioned address the question of how to identify *sv*PLA_2 functionalities from primary sequence of these proteins.

The currently used algorithms and databases, such as BLAST [McGinnis & Madden, 2004], PFAM [Eddy, 1998] and PROSITE [Sigrist *et al.*, 2002] were developed to classify protein sequences into broad families, however these proteins do not necessarily share the same biological function. Despite the efforts to automatically classify proteins with respect to their biological activities, there are no tools available to allow classification of PLA_2 with reasonable accuracy. So far in this work, we present a method capable to predict the activity of a *sv*PLA_2 based on its sequence analysis. Assuming that the subtle determinants of interaction between *sv*PLA_2 and their receptors are, at least in part, related to specific amino acid motifs, we carried out an analysis to detect conserved motifs on two groups of *sv*PLA_2 presenting either neuro- or myotoxic activities. Our work differs from previous studies aiming to correlate sequence and function [de Araujo *et al.*, 1996; Ward *et al.*, 1998; Chioato & Ward, 2003] in two fundamental aspects: (a) is the first time that this particular method is applied for analysis of *sv*PLA_2 sequences and (b) by the fact that the sequences of the HMM training set were selected from literature thus guarantee that the sequence grouping was made according to biological function. The latter concerns with the fact that many proteins classified as myotoxic exhibit neurotoxic activity (and *vice-versa*), being biochemical assays the only way to differentiate if a given *sv*PLA_2 presents only one or both activities. In resume, the method used combines (a) a careful selection of sequences of proteins in order to establish if a given sequence it is related exclusively with only one of the functions or both, (b) multiple

sequence alignments of the selected sequences for the detection of highly conserved motifs, (c) building of Hidden Markov Models (HMM) profiles based in the motif sequences and (d) selection of representative motifs after a ten-fold cross validation procedure. Based in the results presented here, we designed a tool for the identification of myotoxicity and neurotoxicity which is available at www.cbiot.ufrgs.br/bioinfo/phospholipase.

Results

The motifs generated as results of the method are designated as N and M, being related, respectively, with the sequences of neurotoxic and myotoxic *sv*PLA_2.

Detected motifs on neurotoxic *sv*PLA_2 (nPLA_2)

From motifs detected for neurotoxic *sv*PLA_2 (see Table 1), two distinct sequence motif distributions, named motif pattern *N*-I and motif pattern *N*-II, could be noted. Fig. 1 presents consensus sequences of PLA_2 in which the differences are situated at both N-terminal (motifs N1 and N6) and C-terminal (motifs N5 and N7) regions. Common to both motif patterns *N*-I and *N*-II, three motifs were found in the following regions of PLA_2 general fold (Fig. 1): Motif N2, named catalytic motif, constituted by the segment that extends from the amino acid residue located at position 9 to the amino acid located at position 49 of the mature protein. This region displays two important elements of the *sv*PLA_2 fold: (i) the α-helix 2 in which are located His48 and Asp49, both residues displaying critical function for the enzymatic activity in catalytic PLA_2; (ii) the calcium binding loop, responsible for an essential coordination of Ca^{2+} which enables the substrate hydrolysis. Motif N3: named β-wing motif, formed by the segment that comprises the amino-acids residues 52 to 83. This region presents a short antiparallel β-sheet, generally referred as the β-wing, and the 69-loop (also referred as "elapid" or "pancreatic" loop). The first structural element forms a part of the homodimer interface [Arni *et al.*, 1995] whereas the 69-loop is probably involved in catalysis, since deletions in this region can change PLA_2 enzymatic activity [Kuipers *et al.*, 1989]. Studies

had pointed that in the 69-loop the hydroxyl of the conserved residue Tyr69 interacts with the *sn*-2 or *sn*-3 phosphate groups of synthetic substrates [Bahnson, 2005]. Motif N4: named H3 motif, is formed by a segment stretching from residues 85 to 107. In this region is located the α-helix 3, a structural element conserved in all secreted PLA_2. Since N1 motif of the motif pattern I (Fig. 1, Panel A) extends over the six first amino acid residues of the N-terminal region, which corresponds to the region where the N6 motif (the N-terminal motif of the motif pattern *N*-II (Fig. 1, Panel B) is located, we generated a single HMM combining the MSA of both N1 and N6 motifs.

Detected motifs on myotoxic *sv*PLA_2 (mPLA_2)

For detected motifs in myotoxic PLA_2 (Table 2), two distinct distributions could be noted: motif pattern *M*-I and motif pattern *M*-II. Fig. 2 presents the consensus sequences of *sv*PLA_2 in which the differences between motif patterns are situated at the C-terminal region (motifs M5 and M6). Common to both motif patterns *M*-I and *M*-II, four motifs were found in the following regions of PLA_2 general fold (Fig. 2): Motif M1, named propeptidic motif, formed by a segment of fifteen amino-acids located in the propeptidic region; Motif M2: named catalytic motif, constituted by the segment that extends from the amino acid residue located at position 5 to the amino acid located at position 54 of the mature protein; Motif M3: named β-wing motif, formed by amino-acids residues located at the pleated β-sheet (β-wing), stretching from residues 58 to 78; Motif M4: defined as H3 motif, is formed by a segment structurally located in the α-helix 3 of the phospholipases A_2 and corresponds to the amino acid residues situated between positions 81 and 109.

The structural localization of motifs 2, 3 and 4 for neurotoxic and myotoxic *sv*PLA_2 is depicted in Fig. 3. The models are orientated in a way that each *sv*PLA_2 approximates 'the lipids eye view' [Arni & Ward, 1996]. In the example, the neurotoxic protein shown is a K49 PLA_2 from *Bothrops neuwiedi pauloensis* which does not have enzymatic activity, therefore

its neurotoxic effect probably results from its interaction with cellular receptors through the represented face.

Phylogram of the C-terminal region for both functional sets

It has been extensively reported that the C-terminal region of *sv*PLA_2 is involved in both neurotoxic and myotoxic activities. From this notion we could expect that the C-terminal sequence of these proteins would be highly conserved. However, in this work we observed that the C-terminal region of *sv*PLA_2 presents a more entropic structure and thus it appears in the form of two distinct motifs for each analyzed function (N5 and N7 for neurotoxic, and M5 and M6 for myotoxic *sv*PLA_2). To further visualize how the C-terminal region of *sv*PLA_2 has evolved, we built a phylogram of both mPLA_2 and nPLA_2 C-terminal region (Fig. 4) using MEGA package [Kumar *et al.*, 1994]. The sequences containing M5 and N7 show a clear separation on the tree, clustering accordingly to their respective activity, however the same did not occur with sequences containing N5 and M6 motifs.

Discussion

The most striking characteristic of *sv*PLA_2 proteins is the fact that they exhibit diverse physiological activities, despite sharing the same structural fold,. The same fold is also highly conserved among secreted PLA_2 of all vertebrates, and structural conservation on key regions (like the Ca^{2+} binding loop) could also be noted in insect PLA_2 [Nicolas, 1997]. Until now, the molecular mechanisms by which the proteins of PLA_2 superfamily have evolved and acquired their diverse functions remain unclear. It is known that *sv*PLA_2 isozymes have evolved in an accelerated manner to acquire capability to generate different actions upon contact with a cellular medium, thus causing several biological effects. It is also possible that the amino acid substitutions have occurred more frequently in particular sites of the *sv*PLA_2 structure, thus originating the so called "pharmacological sites" [Kini, 2003]. These sites would be located at the protein surface and would be apt to operate independently or cooperatively, through the

interaction with target receptors, generating distinct biological responses. One of the objectives of our current research is to identify putative "pharmacological sites" on different PLA_2, in the form of protein sequence motifs, and to correlate them with the biological functions of these proteins. Besides helping the understanding of the structure-function relationships of *sv*PLA_2 proteins, the identification of such "pharmacological sites" could be useful for the development of new pharmaceutical drugs and therapeutic products.

The strategy of motif identification utilized in this work has a proper advantage. As suggested by Troung & Ikura (2002), unlike single-sequence similarity search, it permits exploiting additional information, such as the position and identity of residues that are conserved throughout the family, as well as variable insertion and deletion probabilities resulting from evolution. The diagnostic success of these specified signatures over the possible wide range signatures lies in the number of true positives recognized over the minimal or nil false positives picked from the non redundant databases [Giri *et al.*, 2004].

In myotoxic *sv*PLA_2, the intronic regions are unusually conserved when compared to the protein-coding regions [Ohno *et al.*, 2003]. As well as the introns, the propeptidic (M1, N1) motifs of myotoxic and neurotoxic *sv*PLA_2 analyzed are found highly conserved. However, these motifs cannot be, apparently, directly related with specific activities in *sv*PLA_2, because they are present in all PLA_2, including those non-toxic ones. In these cases, it is possible that the conservation of propeptidic sequences in both myotoxic and neurotoxic *sv*PLA_2 represents the more efficient way to the activation of the toxic form of these proteins by common peptidases. On the other hand, certain regions of the *sv*PLA_2 are subject to a high rate of variability in the amino acid sequence. Probably, these variable regions have some association with the diversity of functions characteristic of *sv*PLA_2.

The first step for the effective establishment of the constraints that govern the *sv*PLA_2 activities is the detection and validation of the particular amino acid sites that, located at

protein surface, are responsible for some specific functions. From our method we could distinguish three putative sites that fit in prerequisites for each function analyzed, as follows: N2, N3 and N4 for neurotoxic (Fig. 1) and M2, M3 and M4 for myotoxic (Fig. 2) *sv*PLA_2. In all *sv*PLA_2 these regions are located at synonymous sites, but present different lengths and (consensus) amino acid sequences: thus, N2 and M2 are located at the catalytic site region, N3 and M3 are located at β-wing, N4 and M4 are located at α-helix 3. The analysis of each motif can be made taking into account the amino acid residues that are common or exclusive to the two synonymous sites. For example, common to both neurotoxic N2 and myotoxic M2 motifs, the amino acid residues Tyr22, Gly26, Cys27, Cys29, Gly30, Asp42, Cys44 and Cys45 were identified. Exclusive for N2 motif, the more conserved residues are Tyr25, Tyr28, Gly32, Gly33, Pro37, Asp39, Arg43 and Asp49, whereas for M2 motif, Lys16, His48, Cys50, Cys51, Tyr52 and Lys54 appear as exclusive residues. It is however still difficult to conclude on the higher importance of a single amino acid, that exclusively appears in each one of the motifs, as involved in the expression of either neurotoxic or myotoxic function. Nevertheless, this information will be useful for the design of experimental protocols in order to test biochemical activity of each consensus sequence and also the influence of the mutation of key residues.

It is well known that *sv*PLA_2 have evolved in an accelerated fashion mainly due to substitutions in fully exposed residues rather than in the buried residues [Kini, 2003]. Thus, it is reasonable to consider that *sv*PLA_2 have acquired capability to interact with different substrates, ligands or receptors toward the modification of the molecular surface rather than the modification of elements of the secondary or tertiary structure.

The occurrence of beta-wing neurotoxic motif in all myotoxic sequences shows that the beta-wing region (mapped onto homonymous motifs) is highly conserved amongst neurotoxic and myotoxic sequences and could not be used alone to discriminate between the functional

classes of proteins analyzed in this work. Additional activities predicted by the theoretical model would not be surprising if we consider that the biological effects were experimentally determined from studies in isolated tissues or *in vitro* methods rather than in animals. The *in vitro* studies sometimes show non specific effects because of inherent phospholipids activity. Besides that, the proper enzymatic activity of *sv*PLA_2 is critically dependent on the quality of phospholipid surface [Kini, 2005]. Yet, the identification of a variety of membrane and soluble proteins that bind to different PLA_2 suggests that the *sv*PLA_2 enzymes could also function as high affinity ligands [Valentin & Lambeau, 2000]. The discriminative motifs found can represent sites of interaction with receptors and thus explain the distinct physiological activities that *sv*PLA_2 proteins have acquired during evolution. Also, it is important to note that some specific biological effect cannot be necessarily consequence of the presence of one single motif at protein surface, but in certain cases dependent on a combination of overlapping motifs. Therefore, the presence of neurotoxic motifs in *sv*PLA_2 that were experimentally characterized as myotoxic can be consequence of structural constraints. This assumption is in accordance with the hypothesis that nature has used few molecular templates to generate proteins that exhibit diverse functions.

Finally, despite the remarkable differences in *sv*PLA_2 physiologic activities, they share high identity (40-99 %) in their amino acid sequences and hence significant similarity in their three dimensional fold. Thus, the functional differences among PLA_2 enzymes cannot be easily correlated to their structural differences, making the structure-function relationships complicated and challenging [Kini, 2003]. From the evolutionary point of view, PLA_2 acquired the capability to operate as interfacial proteins that make contact with the organized interfaces along a well defined protein surface or a more complex receptor, thus this interaction is probably accomplished by a combination of protein-interface complex structural changes. Therefore to draw additional conclusions on the importance of each PLA_2 motif, it

would be necessary to generate more experimental data concerning the neurotoxic and/or myotoxic activities of each consensus sequence and the influence of the mutation of key residues. Additionally, the knowledge of molecular mechanisms related with *sv*PLA_2 activities can be paradigmatic to other protein families such as the serpins (serine proteinase inhibitors) superfamily, whose proteins present several distinct, and sometimes antagonistic, biochemical functions [van Gent *et al.*, 2003].

Experimental Procedures

Data set collection

In order to use the method, a rigorously annotated *sv*PLA_2 database was created for those proteins whose myotoxic and/or neurotoxic function was experimentally characterized according to literature reports. A fully annotated database was needed in order to validate the results. For each *sv*PLA_2 biological function modelled was necessary to built two datasets of sequences: one composed by sequences presenting the biological function and the other composed by those sequences that did not show such specific biological function (negative control). The sequences with the biological function modelled (functional set) were divided into two subsets when constructing the models: positive control and training set. The positive control was used to test the generated models, built using the training set, regarding their ability to recognize true positives (see section *Statistical validation and motifs selection*). The size of the dataset was limited to sequences available at the time of collection. For myotoxic *sv*PLA_2 ($mPLA_2$), the functional set and the negative control were composed by 20 and 8 sequences, respectively, and for neurotoxic *sv*PLA_2 ($nPLA_2$) the respective sets contain 16 and 13 sequences.

Detection of conserved motifs and conversion into Hidden Markov Models

The searches for conserved motifs were carried out by submitting the functional set of sequences to MEME tool[2] [Bailey & Elkan, 1994], using default values for all parameters (zero or one occurrences of a single motif, per sequence; minimum width = 6; maximum width = 50) except for *number of different motifs*, which was set to 7. Higher values for this parameter would generate non-significant motifs for proteins with the size of the *sv*PLA_2 (ranging from 110 to 140 amino-acids) used on our datasets. Smaller values could implicate low resolution due to the loss of important motifs, since these could be diluted if sequence segments of major length were chosen. Motifs with low statistical significance (e-value < 0.001) were discarded.

To increase the generalization capability and to facilitate the use of motif information by automated procedures we generated one Hidden Markov Model for each motif, using HMMER package[3] [Eddy, 1998]. Prior to the HMM generation, the sequences were aligned using ClustalW [Chenna *et al.*, 2003]. The resulting multiple sequence alignments (MSAs) were manually corrected when needed. This step aims to improve the resulting HMM since one of the advantages of HMMER is modelling gaps adequately. The motifs found were submitted to k-fold cross validation, utilizing the generated HMMs, in order to identify the significant motifs (i.e., those that can discriminate *sv*PLA_2 with the biological function being modelled from the *sv*PLA_2 lacking it). This procedure will be further explained in the next topic. The regions corresponding to each significant motif were extracted from the MSA and the respective HMM was generated and calibrated to a 5,000 sample universe. In another step, these HMMs have been used in the automatic identification of function.

[2] MEME version 3.0 (Release date: 2004/08/18 09:07:01 - http://meme.sdsc.edu/meme/website/meme.html)
[3] HMMER 2.3.2 (Oct 2003 - http://hmmer.wustl.edu/)

Statistical validation and motif selection

Each HMM was associated with only one type of biological function, neurotoxic or myotoxic, depending on the respective functional control from where it was built. To assess the capability of each motif to capture *sv*PLA_2 sequences with the specific biological function, a 10-fold cross validation was carried out as depicted in Fig. 5. This validation procedure is suitable for that purpose, because it increases the accuracy of the results despite the small number of sequences utilized [Kahsay *et al.*, 2005], and it is quite similar to *jacknife*, another procedure that could also be applied to survey the efficacy of this kind of method [Bagos *et al.*, 2004]. At each iteration, the sequences from the functional set are randomly divided into two groups: 2/3rd of them constituted the training set, used in motif detection and HMM training steps. The remaining 1/3rd of the sequences became the positive control. An e-value cut-off of 0.001 was used to eliminate sequences whose alignment with selected motifs was statistically non-significant. The main parameter used to select the motifs, preferably or exclusively found in some functional group, was the Predictive Accuracy (PA) defined as the ratio between the number of correctly predicted sequences and the total number of sequences in the test set, and is calculated as

$$PA = (\%TP + \%TN) / 2$$

where %TP denotes the fraction of true positives (sequences of the positive control correctly classified, that were not used as training set) and %TN the fraction of true negatives (sequences of the negative control correctly classified). The Error Rate (ER) is complementary to PA as

$$PA + ER = 1$$

and is used to determine Minimum Error Point (MEP). The MEP is the score at which PA value reaches its best value (lesser number of classification errors). Another parameter used to assess the quality of each HMM with respect to biological function prediction is the Precision (PR)

$$PR = TP / TP + FP$$

that gives a measure of how the True Positives (TP) concentrates at the higher scores, where FP denotes de absolute number of False Positives. From all detected motifs, only those reaching an average PA value equal or above 60% were considered as discriminative concerning *sv*PLA_2 biological functions. An overview of the complete cycle going from the generation to selection of discriminative motifs is depicted in Fig. 5.

This work was supported by grants from CAPES-MEC (Brazil).

References

1. Arni R.K., Ward, R.J., Gutierrez J.M. & Tulinsky A. (1995) **Structure of a calcium-dependent phospholipase-like myotoxic protein from *Bothrops asper* venom**. *Acta Crystallographica D* 51, 311-317.

2. Arni R.K. & Ward R.J. (1996) **Phospholipase A_2 – A structural review**. *Toxicon* 34(8), 827-841.

3. Bagos P.G., Liakopoulos T.D., Spyropoulos I.C., Hamodrakas S.J. (2004) **A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins** *BMC Bioinformatics* 15, 5-29

4. Bahnson B.J. (2005) **Structure function and interfacial allosterism in phospholipase A_2: insight from the anion-assisted dimmer**. *Archives of Biochemistry and Biophysics* 433, 96-106.

5. Bailey T.L. & Elkan C. (1994) **Fitting a mixture model by expectation maximization to discover motifs in biopolymers**. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 28-36, AAAI Press.

6. Chenna R., Sugawara H., Koike T., Lopez R., Gibson T.J., Higgins D.G. & Thompson J.D. (2003) **Multiple sequence alignment with the Clustal series of programs**. *Nucleic Acids Research* 31(13), 3497-500.

7. Chioato L., Ward R.J. (2003) **Mapping structural determinants of biological activities in snake venom phospholipases A_2 by sequence analysis and site directed mutagenesis** *Toxicon* 42(8), 869-883.

8. de Araujo H.S., White S.P., Ownby C.L. (1996) **Sequence analysis of Lys49 phospholipase A_2 myotoxins: a highly conserved class of proteins** *Toxicon* 34(11-12), 1237-1242.

9. Eddy S.R. (1998) **Profile Hidden Markov Models**. *Bioinformatics* 14(9), 755-763.

10. Giri A.V., Anishetty S. & Gautam P. (2004) **Functionally specified protein signatures distinctive for each of the different blue copper proteins**. *BMC Bioinformatics* 5(1), 127-135.

11. Kini, M.R. (2003) **Excitement Ahead: Structure, Function and Mechanism of Snake Venom Phospholipase A_2 Enzymes**. *Toxicon* 42(8), 827-840.

12. Kini M.R. (2005) **Structure-function relationships and mechanism of anticoagulant phospholipase A_2 enzimes from snake venoms**. *Toxicon* 45, 1147-1161.

13. Kuipers O.P., Dijkman R., Pals V.G.E.M., Verheij H.M., Haas G. (1989) **Evidence for the involvement of tyrosine-69 in the control of stereospecificity of porcine pancreatic phospholipase A_2**. *Protein Engineering* 2, 467-471.

14. Kumar S,, Tamura K., Nei M. (1994) **MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers**. *Computer Application in the Biosciences* 10(2), 189-191.

15. Lomonte B., Angulo Y., Calderon L. (2003) **An overview of lysine-49 phospholipase A_2 myotoxins from crotalid snake venoms and their structural determinants of myotoxic action** *Toxicon* 42(8), 885-901.

16. McGinnis S., Madden T.L. (2004) **BLAST: at the Core of a Powerful and Diverse Set of Sequence Analysis Tools**. *Nucleic Acids Research* 32 (Web server issue), W20-W25.

17. Mora R., Valverde B., Diaz C., Lomonte B. & Gutierrez J.M. (2005) **A Lys49 Phospholipase A_2 Homologue From *Bothrops Asper* Snake Venom Induces Proliferation, Apoptosis and Necrosis in a Lymphoblastoid Cell Line**. *Toxicon* 45(5), 651-660.

18. Murakami, M., Kudo, I. (2004) **Secretory Phospholipase A_2**. *Biological and Pharmaceutical Bulletin* 27(8), 1158-1164.

19. Nicolas J.P., Lin Y., Lambeau G., Ghomashchi F., Lazdunski M., Gelb M.H. (1997) **Localization of Structural Elements of Bee Venom Phospholipase A_2 Involved in N-Type Receptor Binding and Neurotoxicity**. *Journal of Biological Chemistry* 272(11), 7173-7181.

20. Ohno M., Chijiwaa T., Oda-Uedaa N., Ogawab T., Hattori S. (2003) **Molecular evolution of myotoxic phospholipases A_2 from snake venom**. *Toxicon* 42, 841-854.

21. Ownby C.L., Selistre de Araujo H.S., White S.P., Fletcher J.E. (1999) **Lysine 49 phospholipase A_2 proteins**. *Toxicon* 37(3), 411-45.

22. Kahsay, R.Y., Gao G., Liao L. (2005) **An improved hidden Markov model for transmembrane protein detection and topology prediction and its applications to complete genomes**. *Bioinformatics* 21(9), 1853–1858.

23. Scott, D.L., White, S.P., Otwinowski, Z., Yuan, W., Gelb, M.H., Sigler, P.B. (1990). **Interfacial Catalysis: The Mechanism of Phospholipase A_2**. *Science* 250, 1541-1546.

24. Sigrist C.J.A., Cerutti L., Hulo N., Gattiker A., Falquet L., Pagni M., Bairoch A. & Bucher P. (2002) **PROSITE: A Documented Database Using Patterns and Profiles As Motif Descriptors**. *Briefings in Bioinformatics* 3(3), 265-274.

25. Truong, K., Ikura, M. (2002) **Identification and characterisation of subfamily specific signatures in a large protein super family by a hidden Markov model approach**. *BMC Bioinformatics* 3(1),1-15.

26. Valentin, E., Lambeau G. (2000) **Increasing Molecular Diversity of Secreted Phospholipases A_2 and their Receptors and Binding Proteins**. *Biochimica et Biophysica Acta* 1488, 59-70.

27. van Gent D., Sharp P., Morgan K., Kalsheker N. (2003) **Serpins: structure, function and molecular evolution**. *The International Journal of Biochemistry & Cell Biology* 35(11), 1536-1547.

28. Ward R.J., Alves A.R., Ruggiero Neto J., Arni R.K., Casari G. (1998) **SequenceSpace analysis of Lys49 phospholipases A_2: clues towards identification of residues involved in a novel mechanism of membrane damage and in myotoxicity** *Protein Engineering,* 11(4), 285-94.

Table 1. Conserved motifs detected on neurotoxic *sv*PLA$_2$ of the functional control. The score, PA and coverage values are averages over the results of the 10-fold cross validation. PA and coverage were computed at each iteration after determination of respective MEP.

[a]no alignment with e-value < 0.001
[b]n.d.: not determined

Motif	Designation	MEP score	Predictive accuracy at MEP (%)	Coverage at MEP (%)	e-value
N1	N-terminal 1	no alignment[a]	n.d.[b]	n.d.[b]	6.1 e-93
N2	catalytic region	95.76 ± 5.80	69.77 ± 0.05	58.0 ± 0.20	3.6 e-345
N3	H3	45.51 ± 2.61	66.62 ± 0.07	54.0 ± 0.21	1.7 e-142
N4	beta-wing	35.55 ± 7.56	63.77 ± 0.09	96.0 ± 0.08	6.3 e-118
N5	C-terminal-1	20.0 ± 2.13	51.28 ± 0.00	33.33 ± 0.00	5.3 e-33
N6	N-terminal 2	no alignment[a]	n.d.[b]	n.d.[b]	7.4 e-14
N7	C-terminal-2	20.0 ± 1.80	57.37 ± 0.02	20.0 ± 0.00	1.9 e-20

Table 2. Conserved motifs detected on myotoxic *sv*PLA$_2$ of the functional control. The score, PA and coverage values are averages over the results of the 10-fold cross validation. PA and coverage were computed at each iteration after determination of respective MEP.

Motif	Designation	MEP score	Predictive accuracy at MEP (%)	Coverage at MEP (%)	e-value
M1	Propeptidic	28.84 ± 1.38	78.57 ± 0.0	57.14 ± 0	1.4 e-129
M2	catalytic region	109.70 ± 11.46	98.66 ± 0.03	98.57 ± 0.04	1.0 e-670
M3	H3	57.13 ± 2.15	100.00 ± 0.0	100.0 ± 0.0	3.2 e-349
M4	beta-wing	34.44 ± 4.63	99.37 ± 0.02	100.0 ± 0.0	5.2 e-251
M5	C-terminal-1	15.89 ± 2.31	69.58 ± 0.03	39.17 ± 0.06	1.8 e-055
M6	C-terminal-2	19.94 ± 3.19	55.94 ± 0.01	11.88 ± 0.01	1.9 e-20

FIGURE LEGENDS

Fig. 1. Conserved motifs detected on neurotoxic *sv*PLA$_2$ of the positive control. Motifs N2, N3 and N4 occur in all sequences of functional set. The remaining motifs (highlighted) occur in pairs and are distributed in distinct patterns for neurotoxic activity: in the motif pattern I appears both motifs N1 and N6 whereas in the motif pattern II appears both motifs N5 and N7. The amino acids positions of the motif 1 that are inside propeptidic region are numbered as negative. Each detected motif is inserted in a box. Numbering sequence represents the first and the last amino acids. The symbols of amino acids in capital letter represent those that are highly conserved (i.e. appearing ≥ 50 % in a given position of the motifs) whereas the symbols in small letter represent more entropic positions (lesser conserved). The location corresponding to catalytic aspartic acid (here defined as in position 49) is indicated by the arrow in both sequences. **(A)** Sequence of the motif pattern I (for example, the sequence of ammodytoxin C precursor (acc. P11407) is recognized by the application of this pattern for the search of neurotoxic proteins). **(B)** Sequence of the motif pattern II (for example, the sequence of Notechis II-5 (acc. P00609) is recognized by the application of this pattern for the search of neurotoxic proteins).

Fig. 2. Conserved motifs detected in myotoxic *sv*PLA$_2$. Motifs 2, 3 and 4 occur in all sequences of functional set whereas motif 1 occurs only in the sequences containing propeptidic region. Motifs 5 and 6 (highlighted) appear as mutually exclusive at the C-terminal region. Each detected motif is inserted in a box. Numbering sequence indicates the first and the last amino-acids. The symbols of amino acids in capital letter represent those that are highly conserved (i.e. appearing ≥ 50 % in a given position of the motifs) whereas the symbols in small letter represent positions with more entropic organization (lesser conserved). The location corresponding to catalytic aspartic acid (here defined as in position 49) is indicated by the arrow in both sequences. **(A)** Sequence of the motif pattern I (for example, myotoxin III precursor (acc. Q9PVE3) from *Bothrops asper* is a *sv*PLA$_2$ recognized by the application of this pattern for the search of myotoxic proteins). **(B)** Sequence of the motif pattern II (for example, trimucrotoxin (acc. Q90W39) from *Protobothrops mucrosquamatus* is a *sv*PLA$_2$ recognized by the application of this pattern for the search of myotoxic proteins).

Fig. 3. "Pharmacological" motifs in a neurotoxic (left) and myotoxic (right) *sv*PLA$_2$. Neurotoxic: D49 PLA$_2$ from *Agkistrodon halys pallas* (entry 1A2A of PDB); Myotoxic: K49 PLA$_2$ from *Bothrops asper pallas* (entry 1CLP of PDB). The structural localization of related motifs 2, 3 and 4 (numbered as indicated in Fig. 1) are depicted by different colours: green for motif 3, yellow for motif 2 and blue for motif 4. Also, the C- and N-terminal regions are labelled. Panel A: ribbon representation. Panel B: schematic representation of the solvent accessible surface area – in this case, structures in upper and bottom differ from a rotation of 180 degrees along the vertical axis.

Fig. 4. Linearized tree for the C-terminal regions of functional sets from both *sv*PLA$_2$ groups (mPLA$_2$ and nPLA$_2$). Each sequence has the name of the source organism, the accession number and is tagged according to motif detected by MEME:

N5 and N7 are, respectively, the C-terminal motifs N5 and N7 of neurotoxic *sv*PLA_2; M5 and M6 represent the C-terminal motifs M5 and M6 of myotoxic *sv*PLA_2. The bootstrap values are shown next to branches and mutation rate are indicated by the horizontal ruler. The phylogram was built using MEGA package [Kumar *et al.*, 1994].

Fig. 5. Flowchart of the methodology to model *sv*PLA_2 biological activities into HMMs.

Fig. 1.

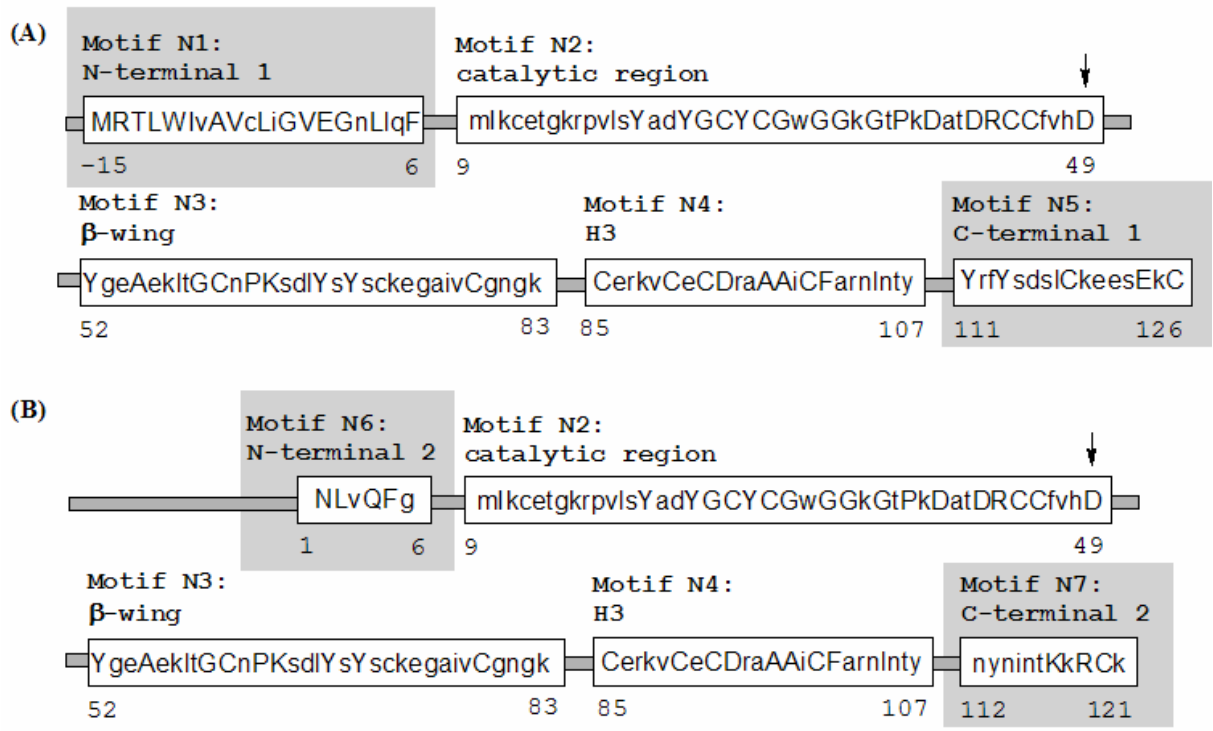

(A)
Motif N1:
N-terminal 1
Motif N2:
catalytic region
MRTLWIvAVcLiGVEGnLIqF
mIkcetgkrpvIsYadYGCYCGwGGkGtPkDatDRCCfvhD
-15
6
9
49
Motif N3:
β-wing
Motif N4:
H3
Motif N5:
C-terminal 1
YgeAekItGCnPKsdIYsYsckegaivCgngk
CerkvCeCDraAAiCFarnInty
YrfYsdsICkeesEkC
52
83
85
107
111
126
(B)
Motif N6:
N-terminal 2
Motif N2:
catalytic region
NLvQFg
mIkcetgkrpvIsYadYGCYCGwGGkGtPkDatDRCCfvhD
1
6
9
49
Motif N3:
β-wing
Motif N4:
H3
Motif N7:
C-terminal 2
YgeAekItGCnPKsdIYsYsckegaivCgngk
CerkvCeCDraAAiCFarnInty
nynintKkRCk
52
83
85
107
112
121

Fig. 2.

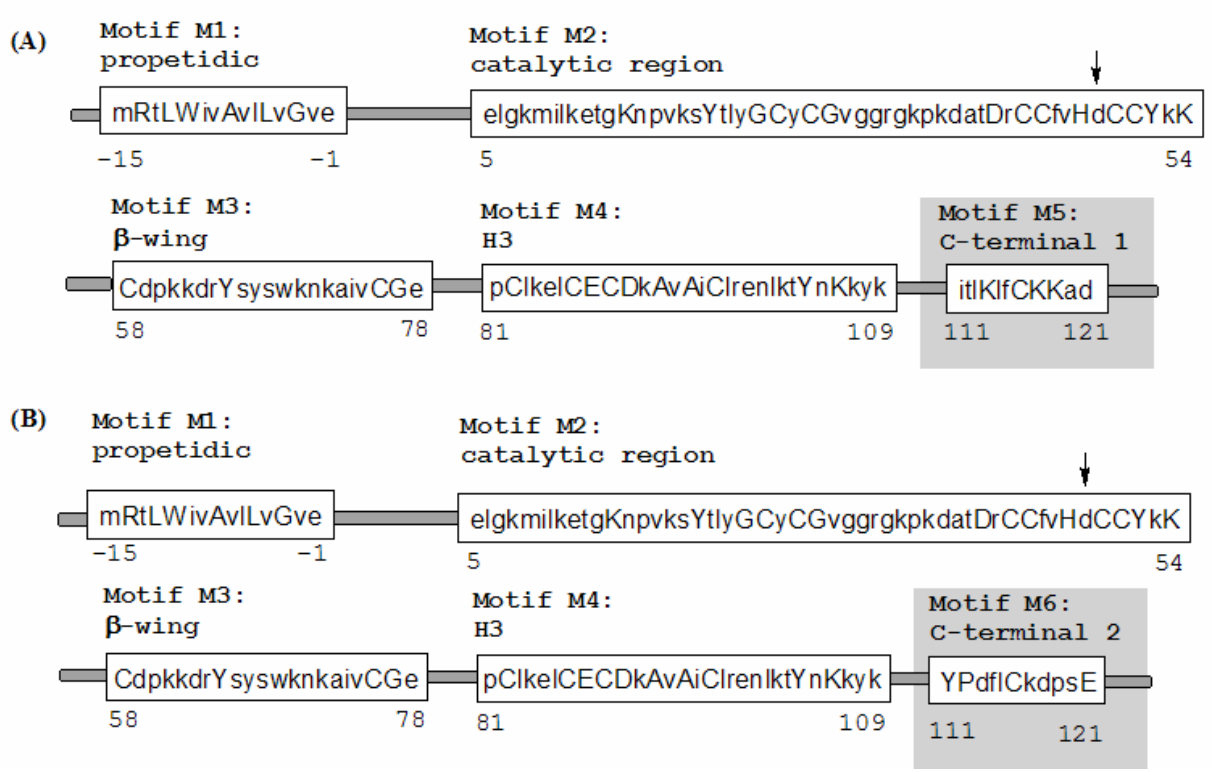

(A)
Motif M1:
propetidic
Motif M2:
catalytic region
mRtLWivAvlLvGve
elgkmilketgKnpvksYtlyGCyCGvggrgkpkdatDrCCfvHdCCYkK
-15
-1
5
54
Motif M3:
β-wing
Motif M4:
H3
Motif M5:
C-terminal 1
CdpkkdrYsyswknkaivCGe
pClkelCECDkAvAiClrenlktYnKkyk
itlKlfCKKad
58
78
81
109
111
121
(B)
Motif M1:
propetidic
Motif M2:
catalytic region
mRtLWivAvlLvGve
elgkmilketgKnpvksYtlyGCyCGvggrgkpkdatDrCCfvHdCCYkK
-15
-1
5
54
Motif M3:
β-wing
Motif M4:
H3
Motif M6:
C-terminal 2
CdpkkdrYsyswknkaivCGe
pClkelCECDkAvAiClrenlktYnKkyk
YPdflCkdpsE
58
78
81
109
111
121

Fig. 3.

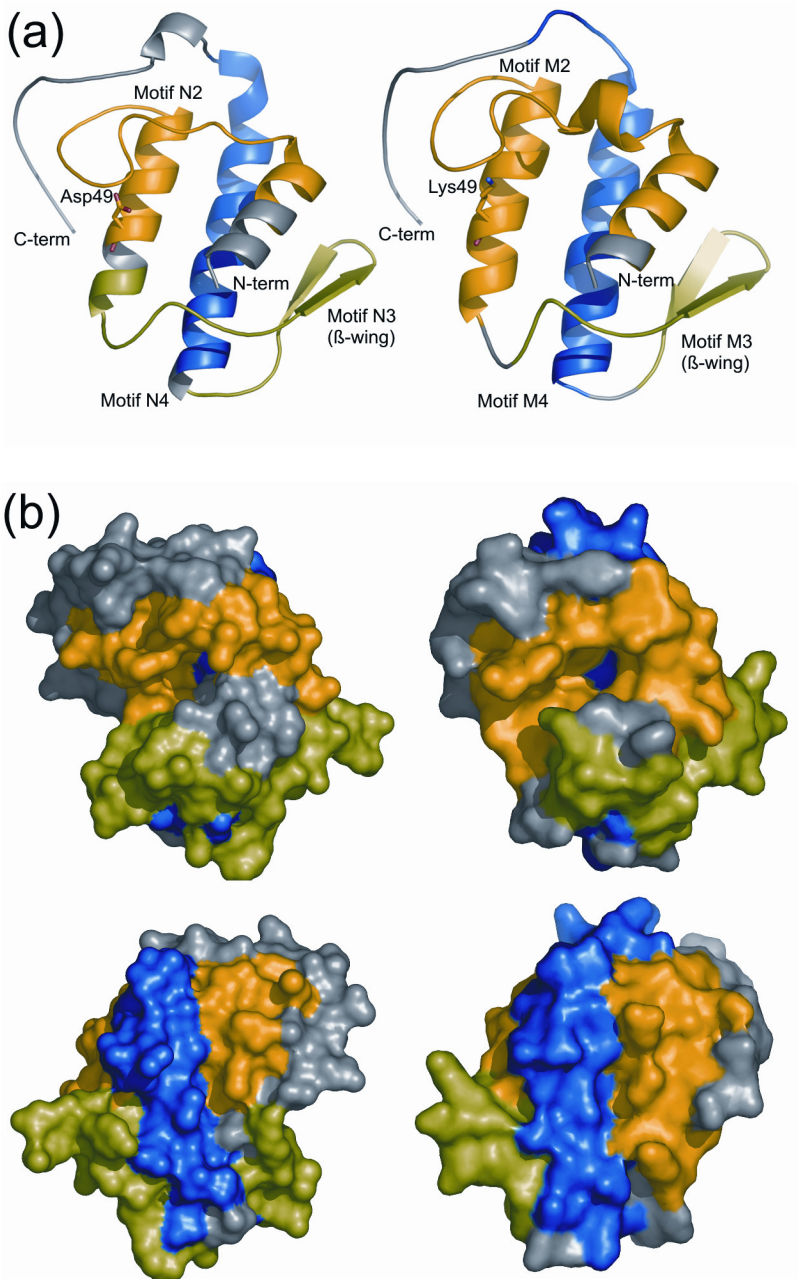

(a)
Motif N2
Asp49
C-term
N-term
Motif N3
(ß-wing)
Motif N4
Motif M2
Lys49
C-term
N-term
Motif M3
(ß-wing)
Motif M4
(b)

Fig. 4.

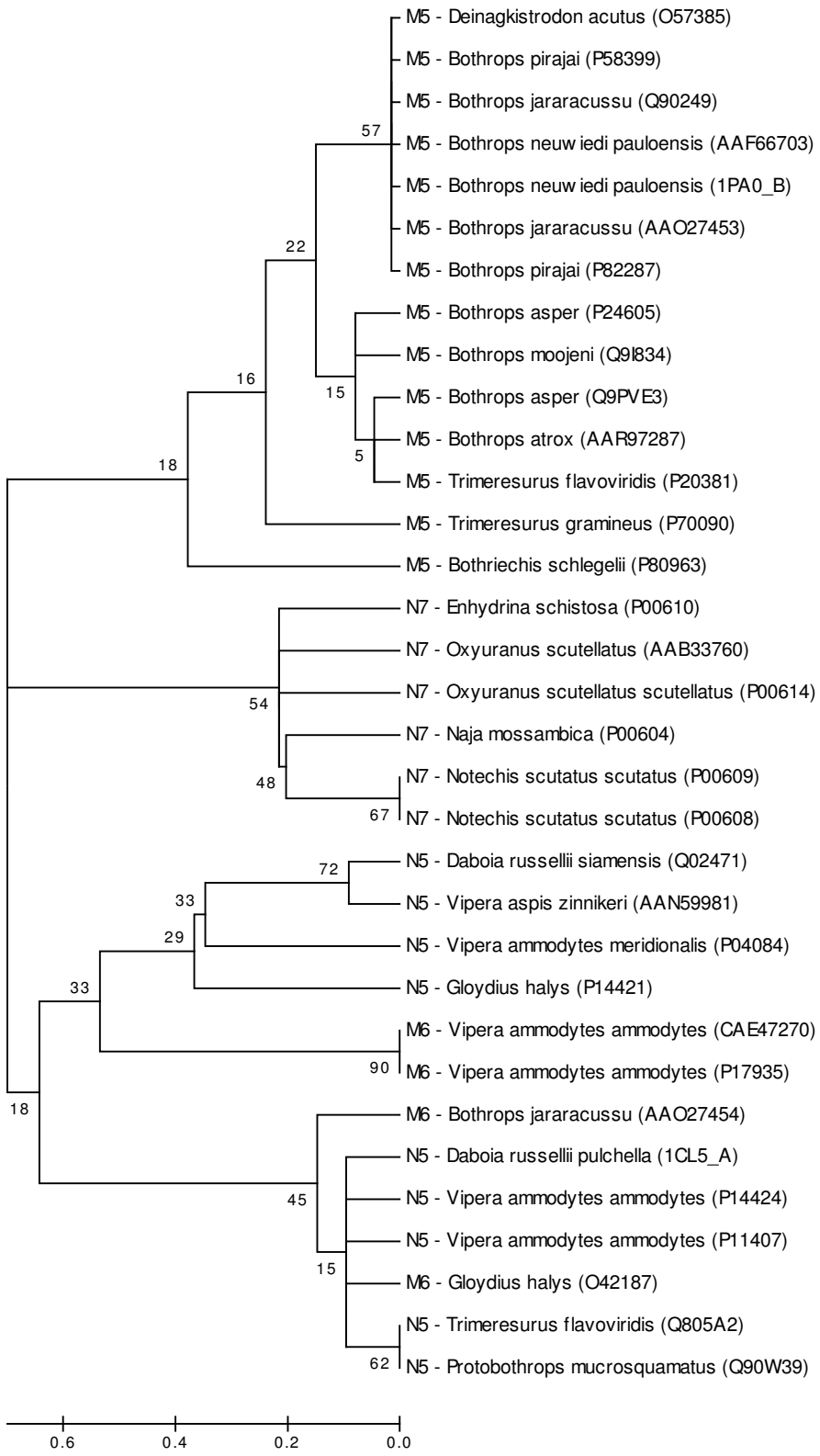
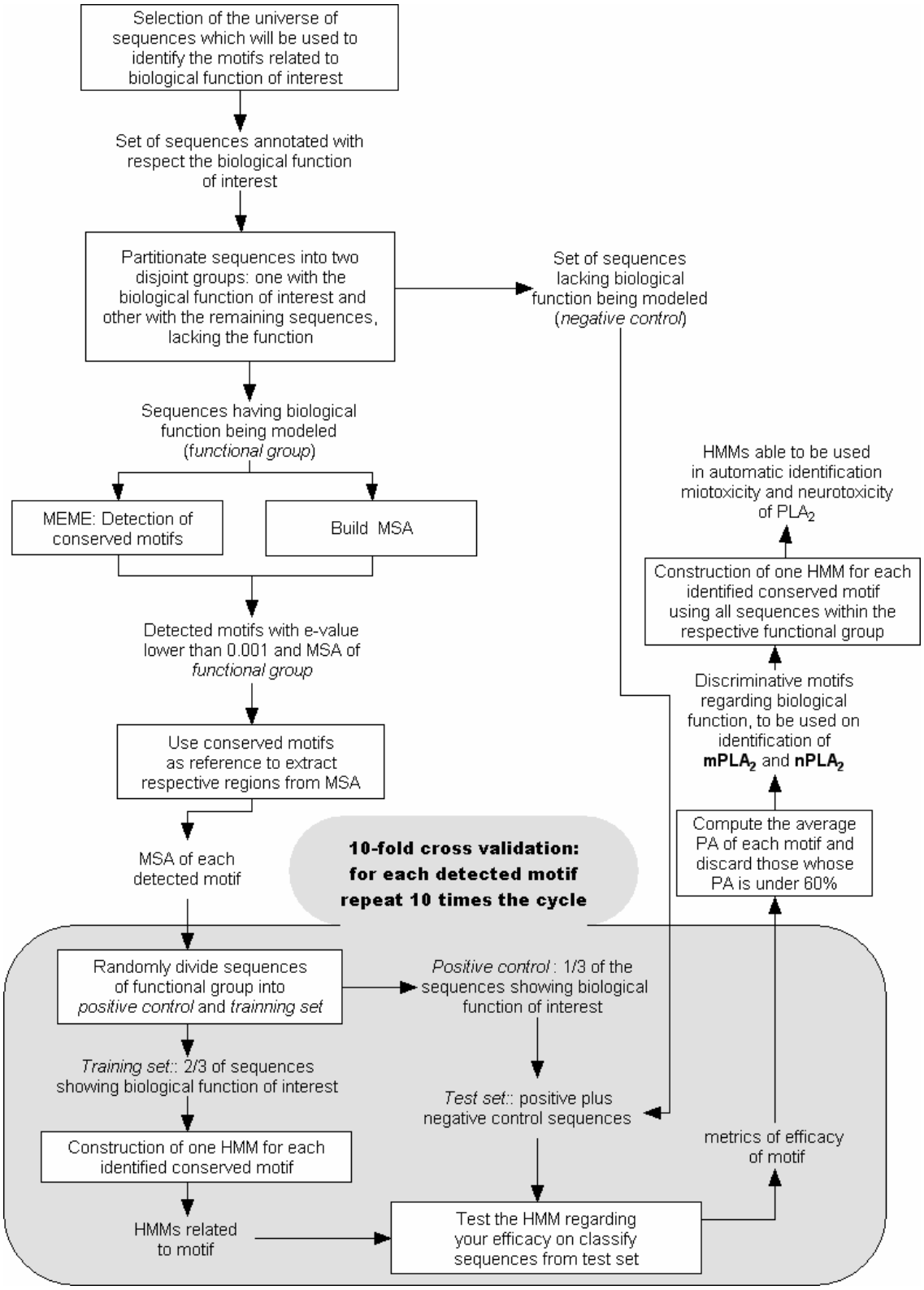

M5 - Deinagkistrodon acutus (O57385)
M5 - Bothrops pirajai (P58399)
M5 - Bothrops jararacussu (Q90249)
M5 - Bothrops neuwiedi pauloensis (AAF66703)
M5 - Bothrops neuwiedi pauloensis (1PA0_B)
M5 - Bothrops jararacussu (AAO27453)
M5 - Bothrops pirajai (P82287)
M5 - Bothrops asper (P24605)
M5 - Bothrops moojeni (Q9I834)
M5 - Bothrops asper (Q9PVE3)
M5 - Bothrops atrox (AAR97287)
M5 - Trimeresurus flavoviridis (P20381)
M5 - Trimeresurus gramineus (P70090)
M5 - Bothriechis schlegelii (P80963)
N7 - Enhydrina schistosa (P00610)
N7 - Oxyuranus scutellatus (AAB33760)
N7 - Oxyuranus scutellatus scutellatus (P00614)
N7 - Naja mossambica (P00604)
N7 - Notechis scutatus scutatus (P00609)
N7 - Notechis scutatus scutatus (P00608)
N5 - Daboia russellii siamensis (Q02471)
N5 - Vipera aspis zinnikeri (AAN59981)
N5 - Vipera ammodytes meridionalis (P04084)
N5 - Gloydius halys (P14421)
M6 - Vipera ammodytes ammodytes (CAE47270)
M6 - Vipera ammodytes ammodytes (P17935)
M6 - Bothrops jararacussu (AAO27454)
N5 - Daboia russellii pulchella (1CL5_A)
N5 - Vipera ammodytes ammodytes (P14424)
N5 - Vipera ammodytes ammodytes (P11407)
M6 - Gloydius halys (O42187)
N5 - Trimeresurus flavoviridis (Q805A2)
N5 - Protobothrops mucrosquamatus (Q90W39)
57
22
16
15
5
18
54
48
67
72
33
29
33
90
18
45
15
62
0.6
0.4
0.2
0.0

Fig. 5.


Selection of the universe of sequences which will be used to identify the motifs related to biological function of interest
Set of sequences annotated with respect the biological function of interest
Partitionate sequences into two disjoint groups: one with the biological function of interest and other with the remaining sequences, lacking the function
Set of sequences lacking biological function being modeled (*negative control*)
Sequences having biological function being modeled (*functional group*)
MEME: Detection of conserved motifs
Build MSA
Detected motifs with e-value lower than 0.001 and MSA of *functional group*
Use conserved motifs as reference to extract respective regions from MSA
MSA of each detected motif
10-fold cross validation: for each detected motif repeat 10 times the cycle
Randomly divide sequences of functional group into *positive control* and *trainning set*
Positive control : 1/3 of the sequences showing biological function of interest
Training set:: 2/3 of sequences showing biological function of interest
Test set:: positive plus negative control sequences
Construction of one HMM for each identified conserved motif
HMMs related to motif
Test the HMM regarding your efficacy on classify sequences from test set
metrics of efficacy of motif
Compute the average PA of each motif and discard those whose PA is under 60%
Discriminative motifs regarding biological function, to be used on identification of **mPLA$_2$** and **nPLA$_2$**
Construction of one HMM for each identified conserved motif using all sequences within the respective functional group
HMMs able to be used in automatic identification miotoxicity and neurotoxicity of PLA$_2$

3.3 Ferramenta de classificação automática online

Uma ferramenta para classificação automática de funções miotóxica e neurotóxica, a partir dos resultados deste trabalho está disponível no endereço http://www.cbiot.ufrgs.br/bioinfo/phospholipase. Os MOMs dos motivos validados para FLA_2 miotóxicas e neurotóxicas também estão disponibilizados na página, de acordo com os resultados do Artigo 2.

O resultado da pesquisa de uma seqüência fornece a atividade detectada e a taxa de erro estimado para esta predição, o(s) motivo(s) detectado(s) na seqüência com o *e-value* associado.

A ferramenta automática processa o resultado da seguinte forma: o cálculo da probabilidade de erro para cada MOM considera a relação entre os falsos positivos e o escore. Quando acima de um escore não ocorriam falsos positivos, a este escore se atribuía taxa de erro = 0%. Quando, para um mesmo escore, haviam diferentes taxas de erro associadas (em função da validação cruzada), o valor final corresponde a uma média dos valores obtidos para o mesmo escore.

Uma vez obtida a lista escore *versus* taxa de erro, foi executada uma análise de regressão linear, para detectar a reta que melhor representava os valores de erro em função do escore, tornando-os diretamente relacionados. Como o resultado da regressão resultou em valores de erro negativos nos escores mais altos, todos os valores negativos de taxa de erro foram agrupados como 0.

Da mesma maneira, escores com valores de erro acima de 100% ficaram representados pelo primeiro escore onde este valor de erro era atingido.

4. Conclusão

Neste trabalho, busquei detectar motivos de seqüência primária em FLA_2 neurotóxicas e miotóxicas que fossem discriminantes destas atividades, com vistas à sua utilização em uma ferramenta de classificação automática. A forma de representação destes motivos escolhida foram os MOMs.

Para tal, desenvolvi uma metodologia baseada nas ferramentas MEME⏩ e HMMER⏩. Através dela, foram encontrados três motivos discriminativos para as FLA_2 neurotóxicas e cinco motivos discriminativos para as FLA_2 miotóxicas. A representação dos motivos como MOMs mostrou-se apropriada para a distinção entre proteínas com alta similaridade. Além disso, a utilização de um pequeno número de seqüências, devido à pouca disponibilidade de dados de seqüência e atividade para as mesmas proteínas, não inviabilizou a eficiência dos MOMs gerados, pela boa capacidade de generalização destes modelos.

Os resultados obtidos estão de acordo com a hipótese de existência de sítios farmacológicos como causa da multiplicidade de atividades das FLA_2 de peçonha de serpente (Kini, 2003). Espera-se que os motivos discriminativos detectados para FLA_2 neuro- e miotóxicas possam colaborar na busca pelos "sítios farmacológicos" envolvidos nestas atividades.

5. Referências Bibliográficas

1. Altschul, S.F., Gish W., Miller W., Myers, E.W. e Lipman D.J. (1990) **Basic local alignment search tool.** *Journal of Molecular. Biology.* 215(3), 403–410.

2. Angulo, Y., Olamendi-Portugal, T., Possani, L.D. e Lomonte, B. (2000) **Isolation and characterization of myotoxin II from *Atropoides* (*Bothrops*) *nummifer* snake venom, a new Lys49 phospholipase A_2 homologue.** *The International Journal of Biochemistry and Cell Biology* 32, 63-71.

3. Baldi, P., Chauvin, Y., Hunkapiller, T. e McClure, M. (1994) **Hidden Markov Models of biological primary sequence information.** *Proceedings of the National Academy of Science of the USA*, 91, 1059-1063.

4. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L., Studholme, D.J., Yeats, C., and Eddy, S.R. (2004) **The Pfam protein families database**. *Nucl. Acids. Res.*, 32, D138-D141.

5. Baum, L.E. e Petrie, T. (1966) **Statistical inference for probabilistic functions of finite state Markov chains.** *The Annals of Mathematical Statistics* 37, 1554-1563.

6. Baum, L.E. e Egon, J.A. (1967) **An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology.** *Bulletin of the American Meteorological Society* 73, 360-363.

7. Baum, L.E. e Sell, G.R. (1968) **Growth functions for transformations on manifolds.** *Pacific Journal of Mathematics* 27(2), 211-227.

8. Baum, L.E., Petrie, T., Soules, G. e Weiss, N. (1970) **A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains.** *Annals of Mathematical Statistics* 41(1), 164-171.

9. Baum, L.E. (1972) **An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes.** *Inequalities* 3, 1-8.

10. Bazzan, A.L.C., Engel, P.M., Schroeder L.F., da Silva, S.C. (2002) **Automated annotation of keywords for proteins related to mycoplasmataceae using machine learning techniques**. *Bioinformatics* 18(suppl_2), S35-S43.

11. Bingham, C.O., Austen, K.F. (1999) **Phospholipase A2 Enzymes in Eicosanoid Generation**. *Proceedings of the Association of American Physicians* 111(6), 516-524.

12. Booting, R. (2004) **Antipyretic therapy.** *Frontiers In Bioscience* 9, 956-66.

13. Brown, William J., Chambers, Kimberly e Doody, Anne (2003) **Phospholipase A_2 (PLA_2) Enzymes in Membrane Trafficking: Mediators of Membrane Shape and Function.** *Traffic* 4 (4), 214-221.

14. Chacur, M., Longo, I., Picolo, G., Gutiérrez, J.M., Lomonte, B., Guerra, J.L., Teixeira, C.F.P.e Cury, Y. (2003) **Hyperalgesia induced by Asp49 and Lys49 phospholipases A_2 from *Bothrops asper* snake venom: pharmacological mediation and molecular determinants.** *Toxicon* 42, 667-678.

15. Chang, C.C. (1985). **Neurotoxins with phospholipase A_2 activity in snake venoms.** *Proc. Natl. Sci. Counc.* 9, 126-142.

16. Chenna R., Sugawara H., Koike T., Lopez R., Gibson T.J., Higgins D.G. & Thompson J.D. (2003) **Multiple sequence alignment with the Clustal series of programs**. *Nucleic Acids Research* 31(13), 3497-500.

17. Chwetzoff, S., Tsunasawa, S., Sakiyama, F., Menez, A. (1989) **Nigexine, a phospholipase A2 from cobra venom with cytotoxic properties not related to esterase activity. Purification, amino acid sequence, and biological properties.** *The Journal of Biological Chemistry* 264(22), 13289-13297.

18. Contreras-Moreira, B., Fitzjohn, P.W., Offman M., Smith, G.R., Bates, P.A. (2003) **Novel use of a genetic algorithm for protein structure prediction: Searching template and sequence alignment space**. Proteins 53(S6), 424-429.

19. Cooper, B., Lipsitch, M. (2004) **The analysis of hospital infection data using hidden Markov models**. Biostatistics 5(2), 223-237.

20. Creer, S., Malhotra, A., Thorpe, R.S., Stöcklin, R.S., Favreau, P.S., Chou, W.S.H. (2003) **Genetic and Ecological Correlates of Intraspecific Variation in Pitviper Venom Composition Detected Using Matrix-Assisted Laser Desorption Time-of-Flight Mass Spectrometry (MALDI-TOF-MS) and Isoelectric Focusing**. *Journal of Molecular Evolution* 56(3), 317-329.

21. Dennis, E.A. (1994) **Diversity of Group Types, Regulation, and Function of Phospholipase A_2.** *Journal of Biological Chemistry* 269(18), 13057-13060.

22. Doery, H.M. e Pearson, J.E. (1961) **Haemolysins in venoms of Australian snakes. Observations on the haemolysins of the venoms of some Australian snakes and the separation of phospholipase A from the venom of *Pseudechis porphyriacus*.** *Biochem. J.* 78, 820-827.

23. Durbin, R., Eddy, S.R., Krogh, A. e Mitchison, G.J., 1998. **Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.** Cambridge University Press, Cambrigde UK.

24. Eddy, S. R. (1998) **Profile Hidden Markov Models.** *Bioinformatics* 14(9), 755-763.

25. Eddy, S.R. (2001) **HMMER: Profile hidden Markov models for biological . sequence analysis.** (Manual e tutorial) Não publicado. Disponível *online* no endereço: http://hmmer.wustl.edu.

26. Eddy, S.R. (2004) **What is a Hidden Markov Model?** *Nature Biotechnology* 22 (10), 1315-1316.

27. Fry, B.G. (2005) **From Genome to "Venome": Molecular Origin and Evolution of the Snake Venom Proteome Inferred From Phylogenetic Analysis of Toxin Sequences and Related Body Proteins**. *Genome Research* 15(3), 403-420.

28. Hanahan, D.J., Brockerhoff, H. e Barron, E.J. (1960) **The site of attack of phospholipase (lecithinase) A on lecithin: a re-evaluation. Position of fatty acids on lecithins and triglycerides.** *J. Biol. Chem.* 235, 1917-1923.

29. Hirabayashi, T. e Shimizu, T. (2000) **Localization and regulation of cytosolic phospholipase A_2**. *Biochim. Biophys. Acta*. 1488(1-2), 124-138.

30. Karplus, K., Barrett, C. e Hughey, R. (1998) **Hidden Markov Models for Detecting Remote Protein Homologies**. *Bioinformatics* 14, 846-856.

31. Kini, M.R. (1997) **Phospholipase A_2: a complex multi-functional protein puzzle.** In: *Venom phospholipase A_2 Enzymes: Structure, Function and Mechanism* Wiley, Manchester, 1-28.

32. Kini, M.R. e Chan. Y.M. (1999) **Accelerated evolution and molecular surface of venom phospholipase A_2 enzymes.** *J. Mol. Evol.* 48, 125-132.

33. Kini, M.R., (2003) **Excitement Ahead: Structure, Function and Mechanism of Snake Venom Phospholipase A_2 Enzymes.** *Toxicon* 42 (8), 827-840.

34. Kini, M.R. (2005) **Structure-function relationships and mechanism of anticoagulant phospholipase A_2 enzimes from snake venoms.** *Toxicon* 45, 1147-1161.

35. Krogh, A., Brown, M., Mian, I.S., Sjölander, K. e Haussler, D. (1994) **Hidden Markov Models in computational biology.** *Journal of Molecular Biology* 235, 1501-1531.

36. Krogh, A. (1997) **Two methods for improving performance of an HMM and their application for gene finding**. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 5, 179–186.

37. Lambeau, G., Lazdunski, M. (1999) **Receptors for a growing family of secreted phospholipases A2**. *Trends in Pharmacological Sciences* 20(4), 162-170.

38. Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J.A., Armananzas, R., Santafe, G., Perez, A. e Robles, V. (2006) **Machine learning in bioinformatics**. *Briefings in Bioinformatics* 7(1), 86-112.

39. Lazzeroni,, L.C., Lange K. (1997) **Markov chains for Monte Carlo tests of genetic equilibrium in multidimensional contingency tables**. *Annals of Statistics* 25(1), 138–168

40. Lee, W.H., da Silva Giotto, M.T., Marangoni, S., Toyama, M.H., Polikarpov, I., Garrat, R.C. (2001). **Structural basis for low catalytic activity in Lys49 phospholipases A_2—a hypothesis: the crystal structure of piratoxin II complexed to fatty acid**. *Biochemistry* 40, 28–36.

41. Li, C., Biswas, G., Dale, M. B., Dale, P. (2001) **Building Models of Ecological Dynamics Using HMM Based Temporal Data Clustering - A Preliminary Study**. *Lecture Notes in Computer Science* 2189, 53-62.

42. Li, M., Fry, B.G., Kini, R.M. (2005) **Putting the Brakes on Snake Venom Evolution: The Unique Molecular Evolutionary Patterns of *Aipysurus eydouxii* (Marbled Sea Snake) Phospholipase A_2 Toxins**. Mol Biol Evol 22 (4):934-941, 2005.

43. Liu, N., Lovell, B.C. (2003) **Gesture Classification Using Hidden Markov Models and Viterbi Path Counting**. *Proceedings of the Seventh International Conference on Digital Image Computing: Techniques and Applications* 273-282.

44. Lin, K., Simossis, V.A., Taylor, W.R., Heringa, J. (2005) **A Simple and Fast Secondary Structure Prediction Algorithm using Hidden Neural Networks**. *Bioinformatics* 21(2), 152-159.

45. Lomonte, B., Tarkowski, A. e Hanson, L.A. (1993) **Host response to *Bothrops asper* snake venom: analysis of edema formation, inflammatory cells, and cytikine release ina mouse model.** *Inflammation* 17, 93-105.

46. Lomonte, B., Angulo, Y, Rufini, S., Cho, W., Giglio, J.R., Ohno, M., Daniele, J.J., Geoghegan P. e Gutiérrez, J.M. (1999) **Comparative study of cytolytic activity of myotoxic phospholipases A_2 on mouse endothelial (tEnd) and skeletal muscle (C2C12) cells *in vitro*.** *Toxicon* 37, 145-158.

47. Lomonte, B., Angulo, Y. e Calderon, L. (2003) **An overview of lysine-49 phospholipase A_2 myotoxins from crotalid snake venoms and their structural determinants of myotoxic action.** *Toxicon*, 42(8), 885-901.

48. Lu, C., Drew, M.S., Au, J. (2001) **Classification of summarized videos using hidden markov models on compressed chromaticity signatures**. In *Proceedings of the Ninth ACM international Conference on Multimedia* ACM Press, New York, NY, vol. 9, 479-482

49. Martí-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., Sali, A. (2000) **Comparative protein structure modeling of genes and genomes**. *Annual Review of Biophysics and Biomolecular Structure* 29(1), 291-325.

50. Moore, J.H. e Williams, D.L. (1964) **Some observations on the specificity of phospholipase A.** *Biochim. Biophys. Acta* 84, 41-54.

51. Mora, R., Valverde, B., Diaz C., Lomonte B. e Gutierrez J. M. (2005) **A Lys49 phospholipase A_2 homologue from *Bothrops asper* snake venom induces proliferation, apoptosis and necrosis in a lymphoblastoid cell line.** *Toxicon* 45(5), 651-660.

52. Murakami, M., Masuda, S., Shimbara, S., Bezzine, S., Lazdunski,M., Lambeau, G., Gelb, M. H., Matsukura, S., Kokubu, F., Adachi,M., Kudo,I. (2003) **Cellular Arachidonate-releasing Function of Novel Classes of Secretory Phospholipase A_2s (Groups III and XII)**. *Journal of Biological Chemistry* 278(12), 10657-10667.

53. Murakami, M. e Kudo, I. (2004) **Secretory Phospholipase A_2**. *Biological and Pharmaceutical Bulletin* 27(8), 1158-1164.

54. Nefian, A.V., Hayes, M. H. (1998) **Hidden markov models for face recognition**. *in ICASSP98,* 2721-2724.

55. Nicolas, J.P., Lin, Y., Lambeau, G., Ghomashchi, F., Lazdunski M. e Gelb, M.H. (1997) **Localization of Structural Elements of Bee Venom Phospholipase A_2 Involved in N-Type Receptor Binding and Neurotoxicity.** *Journal of Biological Chemistry* 272(11), 7173-7181.

56. Ogawa, T., Nakashima, K.I., Nobuhisa, I., Deshimaru, M., Shimohigashi, Y., Fukumaki, Y., Sakaki, Y., Hattori, S., Ohno, M. (1996) **Accelerated evolution of snake venom phospholipase A2 isozymes for acquisition of diverse physiological functions**. *Toxicon* 34(11-12), 1229-1236.

57. Ohno, M., Chijiwaa, T., Oda-Uedaa, N., Ogawab, T. e Hattori, S. (2003) **Molecular evolution of myotoxic phospholipases A_2 from snake venom.** *Toxicon* 42, 841-854.

58. Qian, B. e Goldstein, R.A. (2003) **Detecting distant homologs using phylogenetic tree-based HMMs.** *Proteins: Structure, Function and Genetics* 52, 446-453.

59. Pertea M., Lin X., Salzberg S.L. (2001) **GeneSplicer: a new computational method for splice site prediction**. *Nucleic Acids Res.* 29(5)-1185-90.

60. Polgar, J., Magnenat, E.M., Peitsch, M.C., Wells, T.N.C e Clementson, K.J. (1996) **Asp-49 Is Not an Absolute Prerequisite for the Enzymic Activity of**

Low-M(r) Phoshpolipases A_2: Purification, Characterization and Computer Modelling of an Enzymically Active Ser-49 Phospholipase A_2, Ecarpholin S, From the Venom of *Echis Carinatus Sochureki* (Saw-Scaled Viper). *Biochemical Journal* 319(3), 961-968.

61. Rabiner, L.R. (1989) **A tutorial on Hidden Markov Models and selected applications in Speech Recognition.** *Procedings of the IEEE* 77 (2), 257-285.

62. Saito, K. e Hanahan, D.J. (1962) **A study of the purification and properties of the phospholipase A of *Crotalus adamanteus* venom.** *Biochemistry* 1, 521-532.

63. Scott, D.L, White, S., Otwinowski, Z., Yuan W., Gelb, M.H. e Sigler, P.B. (1990) **Interfacial Catalysis: The Mechanism of Phospholipase A_2.** *Science* 250(4987), 1541-1546.

64. Schrodt, P.A. (2000) **Forecasting Conflict in the Balkans using Hidden Markov Models** in American Political Science Association meeting 2000. Disponível *online* no endereço: http://www.ku.edu/~keds/papers.dir/forecasting.html

65. Six, D.A. e Dennis, E.A. (2000) **The expanding superfamily of phospholipase A_2 enzymes: classification and characterization**. *Biochimica et Biophysica Acta* 1488, 1-19.

66. Sjölander, K., Kevin, K., Brown, M., Hughey, R., Krogh, A., Mian, L.S. e Haussler. D. (1996) **Dirichlet Mixtures: A Method for improved Detection of Weak but Significant Protein Sequence Homology.** *Computer Applications in the Biosciences* Aug, 12(4), 327-345.

67. Stanke, M., Waak, S. (2003) **Gene prediction with a hidden Markov model and a new intron submodel**. *Bioinformatics* 19(suppl_2), ii215-ii225.

68. Thompson, J.D., Higgins, D.G., Gibson, T.J.(1994) **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence**

weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22(22), 4673-4680.

69. Truong, K. e Ikura, M. (2002) **Identification and Characterization of Subfamily-Specific Signatures in a Large Protein Superfamily by a Hidden Markov Model Approach**. *BMC Bioinformatics 3,* 1.

70. Valdez-Cruz, N.A., Batista, C.V.F., Possani, L.D. (2004). **Phaiodactylipin, a glycosylated heterodimeric phospholipase A_2 from the venom of the scorpion *Anuroctonus phaiodactylus***. *European Journal of Biochemistry* 271(8), 1453-1464.

71. Valentin, E. e Lambeau, G. (2000a) **What can venom phospholipases A_2 tell us about the functional diversity of mammalian secreted phospholipases A_2?** *Biochimie* Sep-Oct, 82(9-10), 815-831.

72. Valentin, E. e Lambeau, G. (2000b) **Increasing Molecular Diversity of Secreted Phospholipases A_2 and their Receptors and Binding Proteins.** *Biochimica et Biophysica Acta* 1488, 59-70.

73. Viklund, H. e Elofsson, A. (2004) **Best α-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information**. *Protein Science* 13(7), 1908-1917.

74. Wickramaratna, J.C., Fry, B.G., Aguilar, M.I., Kini, R.M., Hodgson, W.C. (2003) **Isolation and pharmacological characterization of a phospholipase A2 myotoxin from the venom of the Irian Jayan death adder (*Acanthophis rugosus*)**. *British Journal of Pharmacology* 138(2), 333-342.

75. Yang, Z.R., Thomson, R., McMeil, P., Esnouf, R.M. (2005) **RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins**. *Bioinformatics* 21 (16), 3369-3376.

76. Zviling, M., Leonov, H., Arkin, I.T. (2005) **Genetic algorithm-based optimization of hydrophobicity tables**. *Bioinformatics* 21(11), 2651-2656.

6. Referências de Endereços da Internet

Ao longo do texto deste trabalho, foram referenciadas ferramentas, bases de dados e outras fontes de recursos de bioinformática, indicadas pelo símbolo . A seguir é apresentada a lista com os endereços das páginas na Internet destas referências, bem como uma breve explicação sobre cada referência[4].

Título	Descrição resumida	Endereço na Internet
Blast	**B**asic **L**ocal **A**lignement **S**earch **T**ool. Ferramenta para identificação de similaridade entre duas seqüências	http://0-www.ncbi.nlm.nih.gov.csulib.ctstateu.edu/BLAST/
CATH	Classificação de proteínas, conforme parâmetros estruturais (topologia, arquitetura, etc.)	http://cathwww.biochem.ucl.ac.uk/latest/
ClustalW	Ferramenta para Alinhamento Múltiplo de Seqüências	http://www.ebi.ac.uk/clustalw/
DaliLite	**D**istance Matrix **Ali**gnment. Ferramenta para busca de similaridade de estruturas tridimensionais utilizando matrizes de distância entre carbonos-alfa.	http://ekhidna.biocenter.helsinki.fi:9801/dali_lite/start
Genbank	Maior banco de dados de seqüencias primárias (proteínas, DNA, RNA, etc.), administrado pelo NCBI.	http://www.ncbi.nlm.nih.gov/Genbank/index.html
HMMER	Ferramenta utilizada nestre trabalho, para a geração dos MOMs a partir de seqüências biológicas.	http://hmmer.wustl.edu/
MAMMOTH	**MA**tching **M**olecular **M**odels **O**btained from **T**heory. Ferramenta para busca de similaridade de estruturas tridimensionais, através do alinhamento do esqueleto carbônico.	http://fulcrum.physbio.mssm.edu:8083/mammoth
MEME	**M**ultiple **E**m (Expectation of Maximization) for **M**otif **E**licitation. Ferramenta utilizada neste trabalho para a identificação dos	http://meme.sdsc.edu

[4] Como a Internet é um ambiente dinâmico e "volátil", alguns destes endereços podem não mais estar disponíveis após certo tempo ou terem sido alterados. Nestes casos sugere-se a utilização de ferramentas de busca, tais como o Google (www.google.com), utilizando-se o termo da coluna *título* além dos termos mais significativos da coluna *descrição* como palavras chave da busca, para tentar localizar o novo endereço.

	motivos de interesse.	
PFam	Protein Families Database: Banco de domínios e famílias de proteínas, representados na forma de MOMs.	http://www.sanger.ac.uk/Software/Pfam/
Prodom	Banco de dados e ferramenta de busca de domínios conservados em famílias de proteínas	http://protein.toulouse.inra.fr/prodom/current/html/home.php
Prosite	Representa domínios conservados e famílias de proteínas, definidas através de assinaturas específicas de aminoácidos.	http://www.expasy.org/prosite/
RCSB Protein Data Bank	Maior banco de dados de estruturas tridimensionais de proteínas e outras macromoléculas de origem biológica	http://www.rcsb.org
RONN	**R**egional **O**rder **N**eural **N**etwork. Predição de regiões pouco estruturadas de uma proteína a partir da seqüência primária.	http://www.strubi.ox.ac.uk/RONN
ScanProsite	Ferramenta de busca sobre base Prosite, que permite expressões regulares como parâmetro de busca.	http://ca.expasy.org/tools/scanprosite/
Sting Protein Dossier	Ferramenta de análise de proteínas, construída pelo Centro de Bioinformática da Embrapa	http://sms.cbi.cnptia.embrapa.br/SMS/STINGm/proteindossier2/
SSAP	**S**equential **S**tructure **A**lignment **P**rogram. Ferramenta para busca de similaridade de estruturas tridimensionais, através do alinhamento de elementos de estrutura secundária.	http://www.biochem.ucl.ac.uk/cgi-bin/cath/GetSsapRasmol.pl
SPDBv	**Swiss-PDBViewer.** Ferramenta para visualização e manipulação de modelos de estruturas tridimensionais de moléculas.	http://www.expasy.org/spdbv/
YASPIN	Realiza predição de estrutura secundária de proteínas utilizando uma técnica chamada Rede Neural Oculta.	http://ibivu.cs.vu.nl/programs/yaspinwww/index.php

7. Perspectivas

Pode-se traçar duas linhas de interesse principais a partir da pesquisa realizada por este trabalho: o aprimoramento da ferramenta de classificação automática de função; e o aprofundamento na compreensão dos mecanismos moleculares das atividades apresentadas pelas FLA_2 de peçonha de serpente.

Como perspectiva dentro da primeira linha, espero conseguir melhorar os fundamentos da ferramenta já disponível, de forma a maximizar sua eficácia e as informações fornecidas no resultado de busca, assim como descrever e disponibilizar outras atividades de FLA_2 para pesquisa.

Uma ressalva acerca da metodologia descrita neste trabalho é o fato de o trabalho necessário, até chegar-se aos MOMs dos motivos validados, ser bastante manual e repetitivo. Isto é devido ao fato de a metodologia ter sido inteiramente desenvolvida ao longo deste trabalho, durante o que aconteceram muitas modificações no intuito de aumentar sua eficácia. O foco deste trabalho esteve em aumentar a eficácia do protocolo criado. Atestada a sua eficácia, a automatização dos seus passos torna-se uma perspectiva natural.

Na segunda linha de interesse, tenho como perspectiva a utilização dos motivos discriminativos estudados neste trabalho em uma busca mais detalhada pelos "sítios farmacológicos" envolvidos nas atividades mio- e neurotóxica. Para isto, diversos métodos poderiam ser empregados, como estudos da influência da mutação de resíduos chave, e verificação da atividade de peptídeos derivados das regiões dos motivos.

APÊNDICE 1: Conceitos básicos para a compreensão da bioinformática deste trabalho

Seqüências biológicas
Seqüência de letras representando aminoácidos ou bases, conforme a origem seja, respectivamente, proteína ou DNA/RNA, obtida por métodos de seqüenciamento.

Alinhamento Múltiplo de Seqüências (AMS)
Um artifício importante no estudo de relações entre sequências, o Alinhamento Múltiplo de Seqüências busca evidenciar as regiões conservadas em um conjunto de proteínas. É muito útil para a predição de função, na identificação de novos membros de uma família de proteínas e no desenvolvimento de estudos para testar e modificar a função de uma proteína específica. Além disso, é a imformação de entrada para diversas ferramentas de bioinformática, como o HMMER.

E-Value
Valor calculado a partir do escore em bits obtido quando se compara uma seqüência contra um modelo de referência tais como AMS ou MOMs. Indica o número esperado de Falsos Positivos com um escore igual ou superior ao que é obtido pela seqüência que está sendo avaliada. Este valor é sensível à quantidade de seqüências utilizadas, posto que quanto maior for o número de seqüências utilizadas, maior a probabilidade de ocorrência de falsos positivos.

Bits
Bits (***Bi****nary digi****ts***) é uma medida comumente utilizada para indicar quantidade de informação. No contexto da bioinformática, refere-se ao log base 2 (ou "0"s e "1"s) mínimo necessário para representar uma dada informação. Por exemplo, para representar os números de 0 a 7, são necessários 3 dígitos binários, conforme demonstrado a seguir:

```
representação em bits: 000, 001, 010, 011, 100, 101, 110, e 111
representação decimal:  0 ,  1 ,  2 ,  3 ,  4 ,  5 ,  6 , e  7
```

Escore (*score* - em *bits*)
Quando se analisa uma determinada seqüência em relação a outras ou a algum modelo estatístico um dos resultados mais comuns da classificação é o chamado escore, dado em bits. De maneira geral, este valor indica o quão bem uma determinada seqüência alinhou contra outras seqüências ou modelo utilizado. No caso da seqüência ter alinhado melhor com o modelo nulo ou com não homólogos, a classificação é negativa.

Modelo nulo
A calibração é outro procedimento necessário para a utilização de um HMM como modelo de um MSA. Para a calibração se utiliza o modelo nulo, que consiste em

seqüências aleatórias de mesmo número de posições do MSA de entrada, utilizadas como grupo controle, em contraponto ao grupo de seqüências que originaram o HMM.
No momento de avaliar a similaridade entre uma seqüência de consulta e o grupo representado pelo HMM (por exemplo, o HMM com das PLA_2 neurotóxicas), os parâmetros do HMM representam a situação ideal de similaridade, enquanto que o HMM do modelo nulo representa o outro extremo, desta forma servindo de referência no cálculo de conteúdo de informação (escore) e também no cálculo do e-value.

Seqüência consenso
Representação simplificada de um AMS, na forma de uma única seqüência (de bases ou aminoácidos), onde o caractere existente em cada posição corresponde ao de maior freqüência no AMS, para a respectiva posição. No caso de não haver prevalência de um caractere específico em alguma posição do AMS, nesta posição é utilizado o caractere "X", para AMS de proteínas, ou "N", para AMS de nucleotídeos.

Ex:

Dado o seguinte conjunto de seqüências de bases:

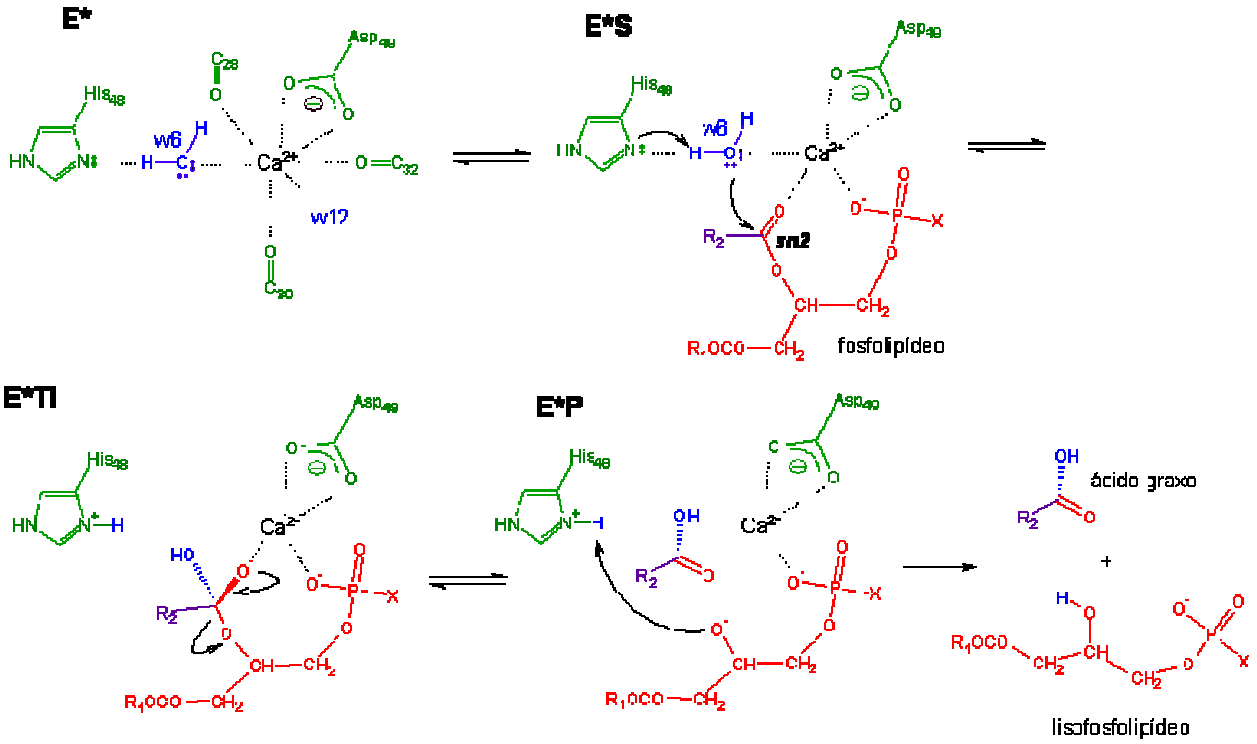
```
A C A - - - A T G
T C A A C T A T C
A C A C - - A G C
A G A C - - A T C
A C C G - - A T C
```

A seqüência consenso derivada seria:

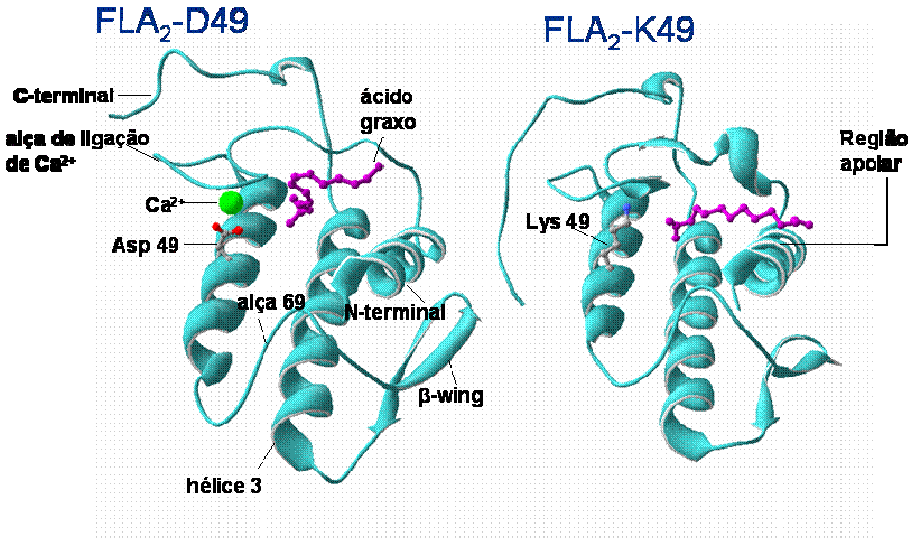
```
A C A C - - A T C
```

APÊNDICE 2: Mecanismo catalítico de FLA_2

Apêndice 2.1: Esquema do mecanismo de reação na hidrólise catalítica de fosfolipídeos pela FLA_2. Ilustradas em verde a díade catalítica de FLA_2, Asp49 e His48. E*: FLA_2 com o íon cálcio e uma molécula de água ligados; E*S: fosfolipase com o substrato, um fosfolipídeo, no sítio catalítico. Indicada a ligação sn2, onde ocorre a hidrólise; E*TI: ataque nucleofílico da água sobre o substrato; E*P: enzima e produtos da clivagem liberados após a catálise.



Apêndice 2.2: Modelos da estrutura de uma FLA_2 D49 e uma K49. Ambas são compostas estruturalmente pelas mesmas regiões, sendo que a diferença principal ocorre no resíduo 49. Nas D49, o resíduo da posição 49 é um aspartato, e estas FLA_2 são capazes de ligar o íon cálcio. Nas K49, o resíduo na posição 49 é uma lisina, e esta mutação impede a ligação do íon cálcio. Sem o íon cálcio, a região apolar na face interna da hélice N-terminal é suficiente para manter o ácido graxo no sítio catalítico, e desta forma, após a primeira catálise o *turnover* da proteína não ocorre e ela fica inativa.



CURRICULUM VITÆ

(resumido)

Fabiano Pasin

CURRICULUM VITÆ

Fabiano Pasin

DADOS PESSOAIS

Nascimento: 12 de março de 1974 — Porto Alegre, Rio Grande do Sul, Brasil

Endereço residencial: R. Marquês do Maricá, 300 – Porto Alegre/RS CEP 91750-460

Telefone residencial: (0xx51) 3263.1938 **E-mail:** fabianopasin@hotmail.com

Endereço profissional: Rua São Luís, 172 - 301 – Porto Alegre/RS CEP 90620-170

Telefone profissional: (0xx51) 3061.4450 **E-mail:** pazzini@dbki.com.br

Fax: (0xx51) 3061.4450

FORMAÇÃO

Pós-graduação
2004/1-2006/1 – Mestrado no Programa de Pós-graduação em Biologia Celular e Molecular (PPGBCM), no Centro de Biotecnologia da Universidade Federal do Rio Grande do Sul (UFRGS).

Graduação
1993-2000 – Bacharelado em Ciências da Computação, ênfase em software de aplicação, no Instituto de Informática, da Universidade Federal do Rio Grande do Sul (UFRGS).

Domínio de línguas

Inglês: compreende (bem), fala (razoavelmente), lê (bem), escreve (razoavelmente)
Espanhol: compreende (bem), fala (razoavelmente), lê (bem)
Italiano: compreende (bem), fala (bem), lê (bem), escreve (bem),
Esperanto: compreende (razoavelmente), fala (razoavelmente), lê (bem)

ESTÁGIOS E BOLSAS

2005 (abril)-2005 (dezembro) **-** Bolsa de Mestrado da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) no Laboratório de Bioquímica Farmacológica do Centro de Biotecnologia do Estado do Rio Grande do Sul, Universidade Federal do Rio Grande do Sul (UFRGS). Orientação de Jorge Almeida Guimarães, linha de pesquisa "Estudo das relações estrutura-função das proteínas pertencentes às subfamílias funcionais da família das fosfolipases A_2 (FLA_2) e suas implicações no desenvolvimento de novos fármacos".

2003 (outubro)-2004 (fevereiro) – Bolsa DTI (Desenvolvimento Tecnológico Industrial), do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), no Instituto de Informática (Universidade Federal do Rio Grande do Sul e Governo do Estado do Rio Grande do Sul). Sob

orientação de Ana Cetertich Bazan, no planejamento e construção de ferramenta de anotação automatizada de genomas (ATUCG), utilizando técnicas de inteligência artificial.

1996 (julho)-1997(junho) – Integrante/empreendedor do projeto SmartBeans, de inovação tecnológica, incubado no Centro de Empreendimentos em Informática do Instituto de Informática, UFRGS, cujo foco era construção de componentes em linguagem Java (JavaBeans) para implementar transações em bases de dados distribuídas e comunicação segura sobre a internet.

1995 (julho)-1996 (julho) – Bolsa de Iniciação Científica do CNPq, no grupo de inteligência artificial distribuída do Instituto de Informática, UFRGS. Orientação de Antônio Carlos Rocha Costa, projeto "Definição de linguagem para programação de multi-agentes distribuídos e estudos para a implementação de plataforma para migração de agentes"

1995 (janeiro)-1995 (julho) – Bolsa de auxílio à pesquisa, da Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS), no Instituto de Filosofia e Ciências Humanas (IFCH), da UFRGS. Orientador: Ivaldo Gehlen, no projeto: "Digitalização e indexação do acervo do Centro de Documentação Social"

ATIVIDADES DIDÁTICAS

2005 (abr-mai): 16 horas/aula na disciplina de Bioinformática, para alunos do 1º semestre do Curso de Graduação em Biomedicina da UFRGS.

2004 (nov): 16 horas/aula na disciplina de Microbiologia Molecular, para alunos do Curso de Graduação em Ciências Biológicas da UFRGS.

2004 (ago): 12 horas/aula, no curso "Bioinformática e Modelagem Molecular Aplicadas ao Desenvolvimento de Novos Agentes Terapêuticos", módulos de filogenia e de manipulação de modelos tridimensionais de estruturas protéicas.

ARTIGOS PUBLICADOS

Pazzini, F., Oliveira, F.S., Guimarães, J.A., Amorim, H.L.N. (2005) Prediction of Myotoxic and Neurotoxic Activities in Phospholipases A_2 from Primary Sequence Analysis In: **Advances in Bioinformatics and Computational Biology - Lecture Notes in Computer Science**, v. 3594, p.194. Berlin, Germany.

Pasin, F., Oliveira, F.S., Guimarães, J.A., Amorim, H.L.N. (2006) A bioinformatics approach to predict myotoxic and neurotoxic activities of Snake Venom Phospholipases A_2. In: **Toxicon**, *in submission*.

Fuentefria, A. M., **Pazzini, F.**, Faganello, J., Valente, P., Valente, P., Vainstein, M. H. (2006) Phenotypic typing of *Cryptococcus neoformans* and *Cryptococcus gattii* by killer sensitive patterns as a

complementary tool to PCR fingerprinting standard method. **Journal of Applied Microbiology,** *in submission*.

Fuentefria, A.M., Perez, L.R.R., d'Azevedo, P., **Pazzini, F.**, Schrank, A., Vainstein, M., Valente, P. (2006) Typing of multi-resistant *Staphylococcus epidermidis* clinical strains by selective panel of Brazilian killer yeasts. **Folia microbiologica**, *in submission*.

RESUMOS E TRABALHOS APRESENTADOS EM CONGRESSOS

Pazzini, F., Oliveira, F.S., Guimarães, J.A., Amorim, H.L.N. (2005) Prediction of Myotoxic and Neurotoxic Activities in Phospholipases A_2 from Primary Sequence Analysis. In: **Proceedings of IV Brazilian Symposium on Bioinformatics.** Springer-Verlag, v. 3594, p. 194-197. São Leopoldo, RS, Brasil. Apresentação oral.

Pazzini, F., Guimarães, J.A., Amorim, H.L.N. (2005) Function prediction of Phospholipases A_2 through a bioinformatics approach. In: **Livro de Resumos da XXXIV Reunião Anual da Sociedade Brasileira de Bioquímica e Biologia Molecular – SBBq**. Águas de Lindóia, SP, Brasil. Apresentação oral e cartaz.

Bastolla, F.M., **Pazzini, F.**, Kirch, R.P., Carazzole, M.F., Brondani, R.V., Coelho, A.S.G., Brommonschenkel, S.H., Schwambach, J., Fett-Neto, A.G., Pappas Jr, G.J., Pereira, G.A.G., Cascardo, J.C.M., Pasquali, G. (2005) GENOLYPTUS Project: sequencing and functional analysis of the Eucalyptus transcriptome. In: **Anais da XXXIV Reunião Anual da Sociedade Brasileira de Bioquímica e Biologia Molecular - SBBq**, v. 1. p. E103-E103, Águas de Lindóia, SP, Brasil.

Fuentefria, A.M., Perez, L.R.R., **Pazzini, F.**, Azevedo, P., Schrank, A., Vainstein, M., Valente, P. (2005) Poder discriminatório de um método fenotípico de tipagem de isolados clínicos de Staphylococcus coagulase-negativo multi-resistentes utilizando um painel de leveduras killer. In: **XXIII Congresso Brasileiro de Microbiologia**, Santos, SP, Brasil. Cartaz.

Pasquali, G., Bastolla, F.M., **Pazzini, F.**, Kirch, R.P., Pizzoli, G., Carazzole, M.F., Brondani, R.V., Coelho, A.S.G., Grattapaglia, D., Brommonschenkel, S.H., Pappas Jr, G.J., Pereira, G.A.G., Cascardo, J.C.M. (2005) Sequencing and differential expression of xylem specific genes from two Eucalyptus species with highly contrasting wood properties. In: **Book of Abstracts and Programme of the IUFRO Tree Biotechnology 2005 Meeting**, p. S1-S1, University of Pretoria, Pretoria, South Africa.

Oliveira, F.S., **Pazzini, F.**, Guimarães, J.A., Amorim, H.L.N. (2005) Methodology Development For The Identification Of Myotoxic Phospholipases A_2 (PLA_2) Utilizing Primary Sequences. In: **Livro de**

Resumos da XXXIV Reunião Anual da Sociedade Brasileira de Bioquímica e Biologia Molecular – SBBq, v. 1. p. V7-V7. Águas de Lindóia, SP, Brasil.

Bastolla, F.M., **Pazzini, F.**, Carazzole, M.F., Brondani, R.V., Coelho, A.S.G., Grattapaglia, D., Brommonschenkel, S.H., Pappas Jr, G.J., Pereira, G.A.G., Pasquali, G., Cascardo, J.C.M. (2005) Differential expression of xylem specific genes involved in cellulose quality from two contrasting Eucalyptus species. In: **51º Congresso Brasileiro de Genética**, CD - Zeppelkini Editorial & Comunicação p. GP058-GP058. Águas de Lindóia, SP, Brasil

Pizzoli, G., Bastolla, F.M., **Pazzini, F.**, Kirch, R.P., Roesler, G.A., Miranda, R.P., Carazzole, M.F., Pappas Jr, G.J., Pereira, G.A.G., Cascardo, J.C.M., Pasquali, G. (2005) Projeto GENOLYPTUS: seqüenciamento e análise de transcritos de tecidos vasculares de Eucalyptus grandis. In: **51º Congresso Brasileiro de Genética**, CD - Zeppelini Editorial & Comunicação, p. GP057-GP057. Águas de Lindóia, SP, Brasil.

Kirch, R.P., Schwambach, J., Bastolla, F.M, **Pazzini, F.,** Pizzoli, G., Roesler, G.A., Miranda, R.P., Carazzole, M.F., Brondani, R.V., Coelho, A.S.G., Grattapaglia, D., Brommonschenkel, S.H., Pappas Jr, G.J., Pereira, G.A.G., Fett-Neto, A.G., Cascardo, J.C.M., Pasquali, G. (2005) Projeto GENOLYPTUS: seqüenciamento e análise de transcritos de plântulas de Eucalyptus grandis submetidos a diferentes estímulos. In: **51º Congresso Brasileiro de Genética**, CD - Zeppelini Editorial & Comunicação, p. GP055-GP055. Águas de Lindóia, SP, Brasil.

Pazzini, F., Schnack, W.R., Carneiro, M.L., Comparsi, F., Hagen, E. (2000) Videoconferência e Vídeo Sob Demanda em Redes de Alta Velocidade: Suporte Tecnológico à Educação à Distância. In: **Anais do VII Congreso Internacional de Informática en la Educación**. C. de La Habana, Cuba.

Pazzini, F., Comparsi, F., Torres, G., Schnack, W. R., Angonese, P. R., Carneiro, M. L., Hagen, E. (1999) REDES ATM - suporte à comunicação multimídia. In: **Anais do Telemática 99**. SUCESU, Porto Alegre, RS, Brasil.

Pazzini, F., Silva, A.R. (1995) Criação de uma Linguagem para Definição de Sistemas Multiagentes Distribuídos. In: **Livro de resumos do VII Salão de Iniciação Científica**. Porto Alegre, RS, Brasil.

Pazzini, F., Guimarães, M. (1994) Informatização do Acervo do CDS (Centro de Documentação Social). In: **Livro de resumos do VI Salão de Iniciação Científica**. Porto Alegre, RS, Brasil.

ORIENTAÇÕES

Orientação da aluna de Iniciação Científica **Fernanda da Silva Oliveira,** em seu trabalho de conclusão do curso de Ciências Biológicas (2005)**: Predição de Atividade Miotóxica em Fosfolipases A_2 de Peçonha de Serpente Através de Análise de Seqüência Primária**.

Orientação de estagiário **Guilherme Arsego Roesler**, do curso de Ciências Biológicas (UFRGS), em suas atividades de suporte à bioinformática do projeto **Genolyptus** de análise funcional do transcriptoma do Eucalipto, junto ao Laboratório de Biologia Molecular Vegetal, prof. Giancarlo Pasquali, Centro de Biotecnologia, UFRGS.

DEMAIS CURSOS E PARTICIPACOES EM ATIVIDADES CIENTIFICAS

Curso de Cristalografia de Proteínas, LNLS, Campinas, SP, 21-23 de junho de 2004.

Palestrante sobre o tema "Bioinformática", da **Jornada Acadêmica de Biologia da UFRGS**, em 09 de novembro de 2005, Porto Alegre, RS.

Transformando Biotecnologia em Bionegócios, 30 de junho de 2005, Porto Alegre, RS.

Participação em congressos, simpósios e outros eventos científicos.