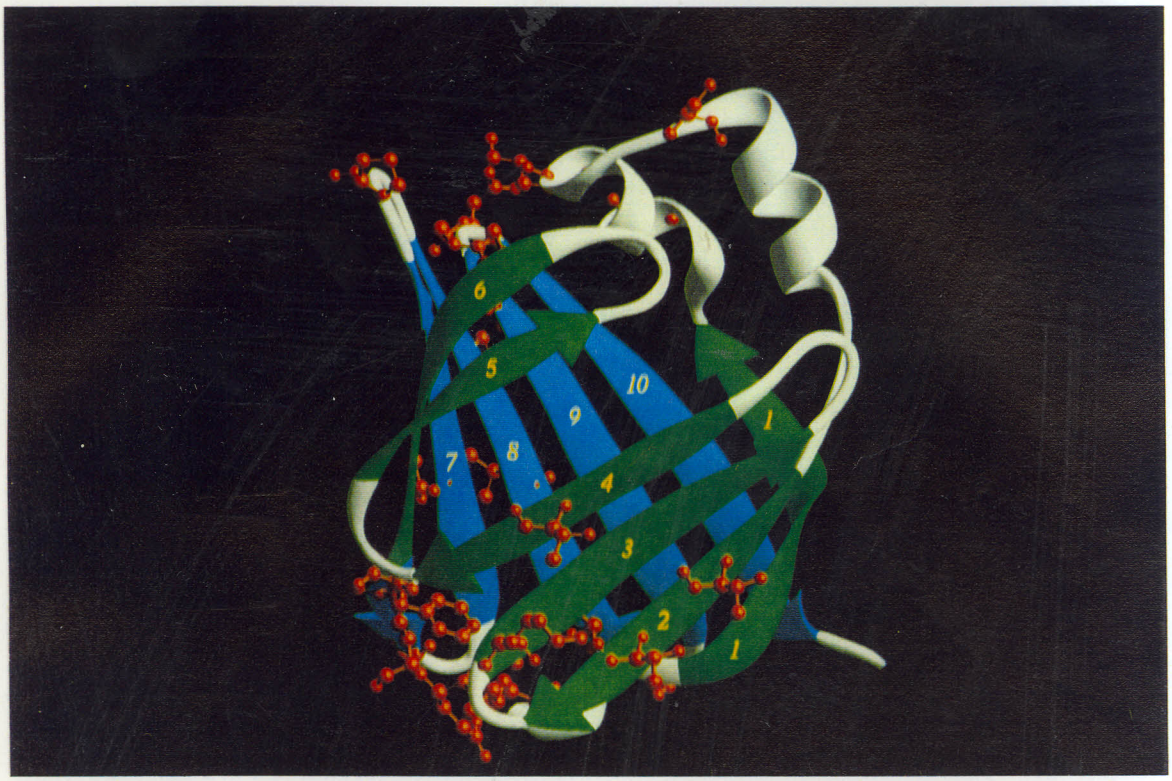
MODELAGEM MOLECULAR DE PROTEÍNAS POR HOMOLOGIA



Texto que sistematiza o trabalho científico do candidato para a obtenção do título de Livre Docência

USP / IFQSC / SBI
8-2-000679

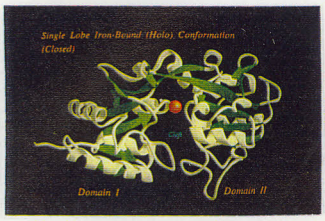
Prof. Dr. Richard Charles Garratt

Departamento de Física e Informática
Instituto de Física de São Carlos
Universidade de São Paulo

Março 1996

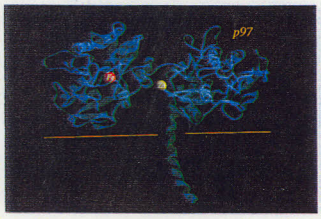
IFSC-USP SERVIÇO DE BIBLIOTECA E INFORMAÇÃO

ÍNDICE



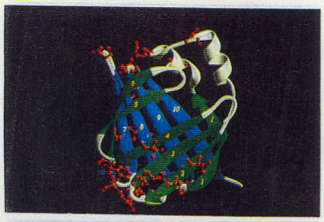
Sarra, R., Garratt, R.C., Gorinsky, B., Jhoti, H & Lindley, P.F. (1990) 'High Resolution X-Ray Studies on Rabbit Serum Transferrin at 2.3Å Resolution' Acta Cryst. **B46**, 763-771

Lindley, P.F., Bajaj, M., Evans, R.W., Garratt, R.C., Hasnain, S.S., Jhoti, H, Kuser, P., Neu, M., Patel, K., Sarra, R., Strange, R. & Walton, A. (1993) 'The mechanism of Iron-Uptake by Transferrins: The Structure of an 18kDa NII - Domain Fragment from Duck Ovotransferrin at 2.3Å resolution', Acta Cryst **D49**, 292-304

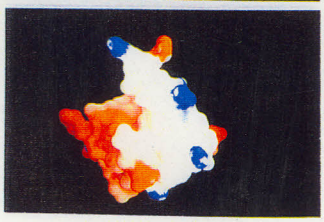


Evans, R.W., Crawley, J.B., Garratt, R.C., Grossmann, J.G., Neu, M., Aitken, A., Patel, K.J., Meilak, A., Wong, C., Singh, J., Bomford, A. & Hasnain, S.S. (1994) 'Characterization and Structural Analysis of a Functional Human Transferrin Variant and Implications for Receptor Recognition', Biochemistry **33**, 12512-12520

Garratt, R.C. & Jhoti, H. (1992) 'A model for the Three-Dimensional Structure of p97 Provides Evidence for a Postulated Zinc Binding Function', FEBS Letts **305**, 55-61

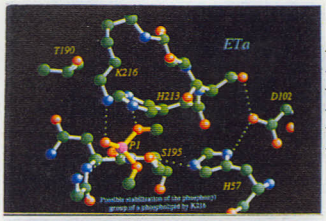


Tendler, M., Brito, C.A., Vilar, M.M., Serra-Freire, N., Diogo, C.M. Almeida, M.S., Delbem, A.C.B., da Silva, J.F., Savino, W. Garratt, R.C., Katz, N. & Simpson, A.J.G. (1996) Proc. Natl. Acad. Sci. **93**, 269-273



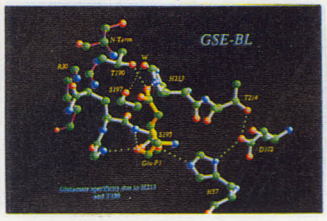
Franco, G.R., Garratt, R.C., Tanaka, M., Simpson, A.J.G. & Pena, S.D.J. (1996) 'Characterization of a *Schistosoma mansoni* Gene Encoding a Homologue of the Y-box binding protein', submitted to J. Biol. Chem.

IFSC-USP SERVIÇO DE BIBLIOTECA E INFORMAÇÃO



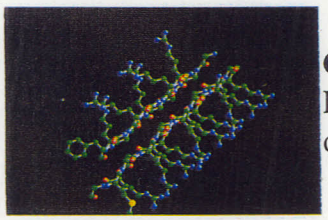
Dancer, S.J., Garratt, R.C., Saldanha, J.W. Jhoti, H. & Evans, R.W. (1990) 'The Epidermolytic Toxins are Serine Proteases', FEBS Letts **268**, 129-132

Barbosa, J.A.R.G., Garratt, R.C. & Saldanha, J.W. (1993) 'A structural model for the Glutamate-Specific Endopeptidase from *Streptomyces griseus* that Explains Substrate Specificity', FEBS Letts., **324**, 45-50

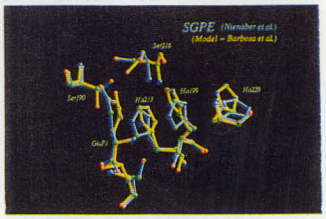


Barbosa, J.A.R.G., Saldanha, J.W. & Garratt, R.C. (1996) 'Novel Features of Serine Protease Active Sites ans Specificity Pockets: Sequence Analysis and Modelling Studies of Glutamate Specific Endopeptidases and Epidermolytic Toxins' Prot. Eng. in press

Garratt, R.C., Oliva, G., Caracelli, I. Leite, A. & Arruda, P (1993) 'Studies of the Zein-like α-Prolamins Based on an Analysis of Amino acid Sequences: Implications for their Evolution and Three-Dimensional Structure', PROTEINS: Structure, Function & Genetics, **15**, 88-99



Garratt, R.C., Thornton, J.M. & Taylor, W.R. (1991) 'An Extension of Secondary Structure Prediction Towards the Prediction of Tertiary Structure', FEBS Letts. **280**, 141-146



DEDICAÇÃO

To Dr. Robert Evans

For all that he has taught me and done for me.

AGRADECIMENTOS

Aproveito este espaço para reconhecer as pessoas que contribuiram de diversas maneiras para a realização deste trabalho. Meus agradecimentos:

- Ao Prof. Glaucius Oliva, pela amizade e total apoio em tudo que tenho feito no IFSC. Principalmente pela sua dedicação e competência nos trabalhos que realizamos juntos e pelo seu esforço em batalhar para conseguir os recursos para minha vinda original ao Brasil e minha contratação subsequente.
- Aos Profs. Yvonne P. Mascarenhas e Eduardo E. Castellano por terem me recebido no Grupo de Cristalografia do IFSC, pela amizade e pelo apoio no trabalho.
- Aos colegas professores pela coragem em terem aceito um bioquímico como professor no Instituto de Física.
- Aos colegas de trabalho: técnicos, pós-docs, professores visitantes e particularmente os alunos que me mantiveram atualizado em muitas áreas e que me ensinaram tanto.
- Aos colaboradores nos trabalhos científicos, sem suas idéias os trabalhos de modelagem teriam sido muito menos frutíferos.
- Aos meus pais e irmãos, apesar de estarem longe sempre me apoiaram.
- A Maria Helena pelo amor, companheirismo e apoio constante durante a redação desta tese e os trabalhos que a compõe.
- Aos meus grandes amigos Walcinyr e Wanessa.

INTRODUÇÃO

Esta tese trata de aspectos da modelagem teórica de estruturas proteicas apresentados através de uma coletânea de artigos publicados em revistas internacionais. Nestes artigos a técnica mais utilizada foi a chamada 'modelagem molecular por homologia' também conhecida como 'modelagem molecular comparativa'. Com a exceção do primeiro e do último, cada capítulo traz uma breve introdução ao assunto a ser abordado e os artigos a ele relacionados.

O primeiro capítulo traz uma introdução à assunto tratando dos tipos de abordagem existentes para a previsão de uma estrutura terciária a partir de uma sequência de aminoácidos. Os passos básicos da modelagem molecular por homologia são descritos e uma tentativa de situar a técnica no contexto atual da biologia molecular é feita.

Capítulo 2 aborda de uma superfamília de proteínas conhecida como transferrinas. Após uma breve introdução seguem-se quatro artigos que descrevem aspectos da estrutura e função das diversas transferrinas e os seus fragmentos proteolíticos. Diferente dos demais capítulos a do tema principal da tese, incluem-se dois artigos na área de cristalografia que ajudam no entendimento dos trabalhos de modelagem.

Capítulo 3 trata de dois estudos de modelagem molecular por homologia de proteínas do parasita *Schistosoma mansoni.* O primeiro é um estudo direcionado ao desenvolvimento de uma vacina contra esquistossomose e o segundo é resultado de uma colaboração com pesquisadores envolvidos no projeto genôma de *S. mansoni* que, entre outras coisas, visa a descoberta de genes importantes para o desenvolvimento, reprodução e sobrevivência do parasita.

Capitulo 4 descreve o meu trabalho na área de serino-proteases, particularmente as endopeptidases específicas para glutamato e as toxinas epidermolíticas, responsáveis pela síndrome da 'pele queimada'. O corpo do capítulo é constituído de três trabalhos de análise de sequências e modelagem comparativa.

Capitulo 5 descreve uma proposta para a estrutura básica das principais proteínas de reserva de milho, sorgo e *coix sp.*; as α-prolaminas. Neste caso a previsão é baseada em princípios fundamentais de estrutura tridimensional de proteínas e não utiliza a modelagem comparativa, pois até hoje não existe nenhuma estrutura homóloga desta família resolvida por técnicas experimentais.

Capítulo 6 apresenta um curto artigo que descreve um melhoramento num dos métodos mais utilizados para a previsão de estrutura secundária a partir de sequências. A modificação é simples mas traz o benefício de introduzir um componente de informação terciária à previsão.

O capítulo final traz conclusões tentando justificar a qualidade a utilidade dos resultados apresentados nos capítulos anteriores.

Capítulo 1

MODELAGEM MOLECULAR DE PROTEÍNAS HOMÓLOGAS

As motivações que levaram à escolha de modelagem molecular de proteínas por homologia como tema desta tese foram duas: em primeiro lugar representa um conjunto de trabalho inteiramente desenvolvido após o meu doutoramento e assim caracteriza uma nova linha de pesquisa tanto para mim quanto para o laboratório de cristalografia do Instituto de Física de São Carlos; em segundo lugar permite a formulação de uma tese sobre um tema único e coerente. Em contrapartida, os artigos aqui apresentados não representam a totalidade da minha produção científica ao longo destes anos.

1.1 Estrutura de proteínas

Sem qualquer dúvida o nosso conhecimento atual da bioquímica deve muito aos avanços na área de biologia molecular estrutural, em particular à determinação das estruturas tridimensionais das macromoléculas biológicas, principalmente as proteínas. O vínculo direto entre função e estrutura pode ser facilmente visto em áreas tais como imunologia, enzimologia, endocrinologia etc., bastando abrir qualquer livro texto moderno de bioquímica. As técnicas responsáveis por esta revolução são a difração de raios-X e a ressonância magnética nuclear.

Ambas as técnicas dependem fortemente de um conhecimento prévio da estrutura primária da proteína de interesse, ou seja a sua sequência de aminoácidos. Por isso e outros motivos os avanços nos métodos de sequênciamento de proteínas e cDNA também tiveram um impacto muito grande no desenvolvimento da área. O sequenciamento de uma proteína efetivamente se trata da elucidação da estrutura covalente da molécula enquanto a resolução da sua estrutura terciária envolve a determinação das interações não-covalentes (ligações de hidrogênio, pontes salinas, interações tipo Van der Waals etc.). Estas interações fracas, junto com contribuições entrópicas, são responsáveis pela manutenção da estrutura tridimensional da proteína e representam a base fundamental da bioquímica[1]. Além da sua importância no enovelamento de proteínas e ácidos nucléicos, são responsáveis por grande parte das interações intermoleculares importantes para reconhecimento molecular tipo enzima-substrato ou anticorpo-antígeno etc.

1.2 A necessidade da modelagem molecular

Com a chegada da revolução da biologia molecular e particularmente o advento de projetos genoma de diversas espécies[2], a diferença entre o número de sequências e o número de estruturas conhecidas continua a aumentar. A figura 1.1 demonstra a situação até o início desta década[3]. O uso da técnica de Etiquetas de Sequências Transcritas (Expressed Sequence Tags)[4] para a descoberta de genes humanos já levou ao acúmulo de mais de 300.000 ESTs humanos no banco dbEST ('release' 08/02/96). O sequenciamento dos genes de maior interesse e o acabamento do projeto genoma humano dentro do prazo estipulado de 15 anos[5] (um estimado 80.000 genes) levará a um excesso de estruturas primárias para as quais não existirá nenhuma informação estrutural. A incapacidade da cristalografia e ressonância magnética nuclear de acompanhar este ritmo, além das dificuldades em obter cristais (particularmente de proteínas de membrana) ou de lidar com proteínas de alto peso molecular (no caso de RMN), enfatizam a necessidade de técnicas teóricas capazes de extrair informações estruturais a partir de uma sequência de aminoácidos[6].

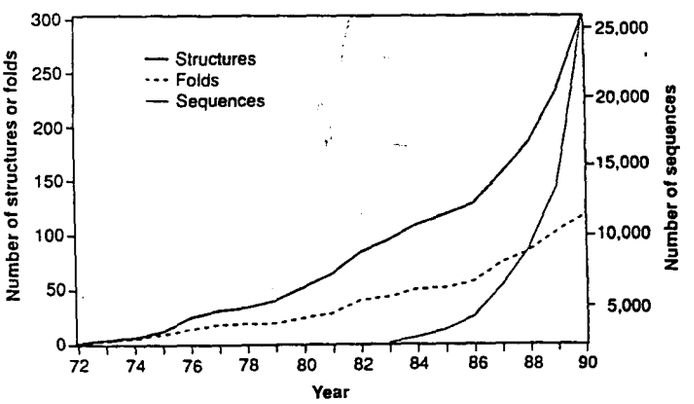


Figura 1.1 Crescimento no número de sequências, enovelamentos e estruturas conhecidos até 1990 [2].

Na ausência de dados experimentais, a modelagem molecular com base em uma ou mais estruturas de proteínas homólogas (modelagem molecular comparativa) é o único método confiável para obter informações estruturais[7]. O método está fundamentado na observação de que a estrutura terciária de proteínas se conserva mais ao longo do processo evolutivo do que a estrutura primária[8], e se reduz efetivamente à previsão das interações fracas acima mencionadas. A metodologia é 'baseada em conhecimento', sendo o modelo uma extrapolação do seu homólogo[9,10].

1.3 Tipos de abordagem

Uma reunião recente em Asilomar nos Estados Unidos tentou pela primeira vez avaliar os métodos de previsão de estrutura de proteínas numa larga escala[6]. Além de modelagem por homologia, mais dois tipos de abordagem foram avaliados; técnicas de 'threading' [11] e métodos *ab initio* incluindo a previsão de estrutura secundária[12].

A idéia de 'Threading' (também conhecido como alinhamento estrutura-sequência ou alinhamento 3D-1D) surgiu a partir da observação da figura 1.1 que mostra que o número de novos enovelamentos cresce muito mais lentamente que o número de novas estruturas e que isto não se deve apenas a membros homólogos da mesma família. Ou seja, proteínas com sequências aparentemente não-relacionadas adotam enovelamentos parecidos[3] e que o número destes enovelamentos deve ser limitado (estimativas variam em torno de uns milhares)[13,14]. Numa comparação tipo 'threading' uma sequência é comparada com um banco de estruturas ou enovelamentos representativos buscando o mais compatível. Com o crescimento do banco de enovelamentos conhecidos, espera-se um aumento no sucesso dos métodos atuais mas os resultados já obtidos, particularmente do programa THREADER[15], são impressionantes.

Métodos *ab initio* variam muito na sua metodologia, mas fundamentalmente todos tentam resolver o 'problema de enovelamento de proteínas', ou seja, dada um sequência de aminoácidos não-relacionada com outras de estrutura conhecida, qual é a estrutura tridimensional? Entre outros, métodos estatísticos, redes neurais e inteligência artificial para previsão de estrutura secundária usando alinhamentos múltiplos, similuções Monte-Carlo, comparações sequência-estrutura de fragmentos e simulações de dinâmica molecular tipo 'simulated annealing' têm sido empregados (para um resumo veja [12]).

Duas das conclusões principais da reunião em Asilomar foram que a modelagem molecular por homologia continua como a única metodologia teórica capaz de fornecer informações confiáveis a respeito da proteína de interesse e que a previsão acurada por métodos *ab initio* ainda não é possível. Os métodos de 'threading' parecem muito promissores e maiores avanços nesta área são esperados no futuro próximo.

1.4 Modelagem molecular por homologia

Diferenças entre estruturas homólogas aumentam com o decréscimo na identidade sequencial e consequentemente a precisão de modelos construídos por homologia também

cai[16,17]. No caso de um modelo construído à base de 90% de identidade, os erros podem chegar próximo aos das estruturas determinadas cristalograficamente. No caso de identidade sequencial em torno de apenas 20%, o gargalo principal está no alinhamento. Segundo Vriend, a tabela 1.1 representa as principais limitações atuais da modelagem por homologia[18].

O processo de modelagem envolve uma série de passos básicos, a ordem dos quais pode variar de caso em caso, mas inclui os seguintes componentes[19]:

- Detecção das sequências e estruturas homólogas através de buscas sistemáticas em bancos de dados adequados e disponíveis pela rede.
- Alinhamento das sequências
- Escolha da(s) estrutura(s) a ser(em) usada(s) como base do modelo
- Construção das regiões conservadas estruturalmente (cadeia principal)
- Construção das regiões variáveis (cadeia principal)
- Inclusão das cadeias laterais
- Remoção de impedimentos estéricos, minimização de energia (e dinâmica molécular)
- Avaliação do modelo através de indices de normalidade e energia potencial.

Identidade sequencial	Principal fator limitante para modelagem por homologia	Principais áreas de pesquisa atuais
100% - 75%	Tempo	Melhoramento na modelagem de cadeias laterais
75% - 50%	Qualidade	Melhoramento na modelagem de 'loops'e motodologias para avaliação dos resultados
50% - 25%	Alinhamento	Melhoramento nos algoritmos de alinhamento e a inclusão de informações estruturais. 'Threading'.
25% - 0%	Detecção	Desenvolvimento de métodos *ab initio*. 'Threading'.

Tabela 1.1 Estatus atual da modelagem molecular por homologia.

A detecção de uma similaridade sequencial que representa uma relação homologa de duas proteínas depende do grau de identidade e o comprimento do alinhamento[17]. Identidade de 25% pode ser suficiente caso o alinhamento contiver mais que 100 resíduos. As principais dificuldades no alinhamento em si resultam de baixa identidade e a incerteza na escolha da penalidade para as inserções. Frequentemente é possível superar tais problemas incluindo informações a respeito da estrutura secundária no alinhamento [20,21] e/ou 'threading'.

Caso tiver mais de uma estrutura homóloga conhecida pode-se usar a 'melhor' (definida usando critérios de similaridade sequencial, resolução e/ou qualidade estereoquímica etc.) como estrutura base do modelo. Alternativamente uma média ponderada das estruturas disponíveis[9,10] ou até mesmo fragmentos de diversas moléculas que são subsequentemente unidos. No caso mais simples, a cadeia principal das regiões conservadas estruturalmente (o 'framework') é copiada diretamente da estrutura base e os 'loops' incluídos depois. Para modelos de similaridade intermediária (25-75%) é a construção dos loops que constitui a maior dificuldade do procedimento como um todo. As metodologias mais utilizadas são:

- Aproveitamento da conformação do loop de uma estrutura homóloga sem ser a estrutura base
- O procedimento de Jones & Thirup que utiliza uma matriz de distâncias, calculada a partir do pontos fixos no 'framework' para buscar conformações em estruturas conhecidas[22]
- Técnicas *ab initio*
- Uso de preferências conformacionais para determinadas estruturas (por exemplo voltas-β em grampos-de-cabelo[23,24].
- Uso de estruturas canônicas (por exemplo em imunoglobulinas)[25].

As dificuldades persistentes desta etapa da modelagem são exemplificadas pelos resultados da reunião em Asilomar onde foi constatado que nenhum 'loop' grande foi corretamente modelado nas sete proteínas testadas[26].

As cadeias laterais são orientadas com referência às estruturas homólogas ou usando bibliotecas de rotâmeros que variam na sua sofisticação e implementação[27-29]. Vriend & Sander[30] descreve um método que permite o uso de rotâmeros sensíveis à conformação local da cadeia principal. Eventuais empedimentos estéricos gerados pela substituição das cadeias laterais podem ser removidos usando 'torsional drivers' e/ou técnicas padrões de minimização de energia[31].

Nos últimos anos muita enfâse tem sido colocado no desenvolvimento de métodos de verificação estrutural e avaliação da qualidade dos modelos. Cálculos energéticos têm sido utilizados por vários autores, mas a sua aplicação tem sido limitada[32,33]. Hoje em dia a maioria dos métodos empregados é baseada em índices de normalidade. O PROCHECK [34] é largamente utilizado tanto para estruturas experimentais quanto para modelos para avaliar estereoquímica. Regras de empacotamento[35], ambiente químico[36] e contatos atômicos[37,38] podem ser usados para detectar problemas locais ou até mesmo enovelamentos incorretos.

A precisão exigida do modelo depende criticamente da aplicação pretendida. A precisão necessária para identificar uma assimetria na distribuição de cargas na superfície molecular, por exemplo, será muito inferior àquela desejável para desvendar um mecanismo catalítico. As vezes um simples alinhamento de sequências contém uma boa parte da informação

requerida e portanto uma das decisões mais importantes do processo é saber em quais casos a construção do modelo vale a pena.

Para comparar o modelo e a estrutura base é conveniente dividir conceitualmente a estrutura em três regiões distintas: 1) regiões conservadas nas duas moléculas tanto em sequências quanto em estrutura, 2) regiões estruturalmente e sequencialmente variáveis e 3) regiões que apresentam variação em sequência, mas que são estruturalmente conservadas. Partes da molécula que pertencem à primeira categoria normalmente podem ser construídas com um elevado grau de confiança, mas trazem poucas informações novas pois o modelo acabará sendo uma réplica da estrutura base. Por outro lado, o segundo grupo geralmente conterá a maior parte daquilo que é diferente (e frequentemente mais interessante) sobre o modelo mas, devido às limitações das técnicas, será menos confiável. Resíduos que pertencem ao último grupo podem ser construídos com confiança, pois o problema se reduz ao modelamento de novas cadeias laterais num esqueleto fixo. Provavelmente por isto que a modelagem tem tido um certo sucesso na previsão de especificidade nas serino-proteases, por exemplo (para revisão veja [39]).

Como comentário final vale destacar uma das conclusões tiradas da reunião em Asilomar. Vários dos modelos que foram construídos automaticamente encontram-se entre os piores apresentados, enfatizando que a intervenção humana continua sendo um componente importante do processo de modelagem molecular por homologia[26].

Referências

[1] Schulz, G.E. & Schirmer, R.H. (1979) em 'Principles of Protein Structure', Springer-Verlag, New York

[2] Collins, F. (1995) Proc. Natl. Acad. Sci. **92**, 10821-10823

[3] Bowie, J.B., Lüthy, R. & Eisenberg, D. (1991) Science **253**, 164-170

[4] Adams, M.D. *et al.* Science **252**, 1651-1656

[5] Collins, F . & Galas, D. (1993) Science **262**, 43-46

[6] Moult, J., Pedersen, J.T., Judson, R. & Fidelis, K. (1995) PROTEINS, Structure, Function & Genetics **23(i),** ii-iv

[7] Swindells, M.B. & Thornton, J.M. (1991) Curr. Op. Struct. Biol. **1**, 219-223

[8] Bajaj, M. & Blundell, T.L. (1984) Ann. Rev. Biophys. Bioeng. **13**, 453-492

[9] Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E. & Thornton, J.M. (1987) Nature **326**, 347-352

[10] Blundell, T., Carney. D., Gardner, S., Hayes, F., Howlin, B., Hubbard, T., Overington, J., Singh, D.A., Sibanda, B.L. & Sutcliffe, M. (1988) Eur. J. Biochem. **172,** 511-520

[11] Lemer, C. M.-R., Rooman, M.J. & Wodak, S.J. (1995) PROTEINS, Structure, Function & Genetics **23**, 337-355

[12] Defay, T. & Cohen, F.E. (1995) PROTEINS, Structure, Function & Genetics **23**, 431-445

[13] Chothia, C. (1993) Nature **357,** 543-544

[14] Orengo, C.A., Jones, D.T. & Thornton, J.M. (1994) Nature **372,** 631-634

[15] Jones, D.T., Taylor, W.R. & Thornton, J.M. (1992) Nature **358**, 86-89

[16] Chothia, C. & Lesk, A. M. (1986) EMBO J. **5**, 823-836
[17] Sander, C. & Schneider, R. (1991) PROTEINS, Structure, Function & Genetics **9**, 56-68
[18] Holm, L., Rost, B., Sander, C., Schneider, R. & Vriend, G. (1994) em 'Molecular Modeling'apostila do curso 'Métodos Avançados para Análise Estrutural de Biomoléculas' São carlos, janeiro 1994
[19] Lee, R.H. (1992) Nature **356**, 534-544
[20] Barton, G.J. (1990) Meth. Enz. **183**, 403-428
[21] Gribskov, M. Lüthy, R. & Eisenberg, D. (1990) Meth. Enz. **183**, 146
[22] Jones, T.A. & Thirup, S. (1986) EMBO J. **5** 819-822
[23] Wilmot, C.M. & Thornton, J.M. (1988) J. Mol. Biol. **203**, 221-232
[24] Sibanda, B.L. & Thornton, J.M. (1985) Nature **316,** 170-174
[25] Chothia, C. *et al.* (1989) Nature 343, 877-883
[26] Mosimann, S. Meleshko, R. & James,M.N.G. (1995) PROTEINS, Structure, Function & Genetics **23,** 301-317
[27] Ponder, J.W. & Richards, F.M. (1987) J. Mol. Biol. **193,** 775-791
[28] McGregor, M.J., Islam, S.A. & Sternberg, M.J.E. (1987) J. Mol. Biol. **198,** 295-310
Dunbrack, R.L. & Karplus, M. (1993) J. Mol. Biol. **230**, 543-574
[29] Dunbrack, R.L. &Karplus, M. (1993) J. Mol. Biol. **230,** 543-574
[30] Vriend, G., Sander, C. & Stouten, P.F.W. (1994) Prot. Eng. **7**, 23-29
[31] McCammon, J.A. & Harvey, S.C. (1987) em 'Dynamics of Proteins and Nucleic Acids', Cambridge University Press, Cambridge
[32] Novotny, J., Bruccoleri, R. & Karplus, M. (1984) J. Mol. Biol. **177**, 787-818
[33] Sippl, M. (1993) PROTEINS, Structure, Function & Genetics **17,** 355-362
[34] Laskowski, R.A., MacArthur, M.W. & Thornton, J.M. (1993) J. Appl. Cryst. **26,** 283-291
[35] Gregoret, L.M. & Cohen, F.E. (1990) J. Mol. Biol. **211**, 959-974
[36] Lüthy, R., McLachlan, A.D. & Eisenberg, D. (1992) Nature **356**, 83-85
[37] Vriend, G. & Sander, C. (1993) J. Appl. Cryst. **26**, 47-60
[38] Singh, J. & Thornton, J.M. (1990) **211,** 595-615
[39] Stamato, F.M.L.G. Paulino, M., Garratt, R., Soares, C.M. & Tapia, O. (1995) Mol. Engin. **4,** 375-414

Capítulo 2

TRANSFERRINAS

Para combater os efeitos tóxicos do ferro no organismo, os vertebrados desenvolveram um sistema complexo de transporte e armazenamento, cujos componentes principais são as proteínas transferrina, o receptor de transferrina e ferritina. Este capítulo trata de um destes componentes, a molécula de transporte, transferrina.

A superfamília das transferrinas representa um grupo de proteínas derivado de um ancestral comum que inclui transferrina de soro, ovotransferrina (de clara de ovo de aves e répteis), lactoferrina de neutrófilos e secreções extracelulares e p97 (um antígeno de membrana associado ao melanoma)[1]. São gliçoproteínas de peso molecular de aproxamadamente 80,000 que (com a exceção de p97[2]) ligam-se reversivelmente a dois íons férricos. Um aspecto único da sua capacidade de ligar ferro é o requisito obrigatório da ligação concomitante de um ânion sinergístico [3].

Transferrina de soro age no intercambio de ferro entre sítios de absorção, armazenamento (principalmente no fígado) e utilização. Em pH neutro, a holo-transferrina possui uma alta afinidade pelo seu receptor específico nas membranas de células-alvo onde é internalizada. Em pH baixo dentro de um CURL (compartamento de desacoplamento entre receptor e ligante), o ferro é liberado mas nestas mesmas condições ácidas a apo-transferrina mantém-se ligada ao receptor até encontrar novamente o pH neutro do meio extracelular onde o

complexo transferrina-receptor se desfaz, liberando a transferrina para mais um ciclo de transporte (Figura 2.1). Para a célula-alvo poder distinguir entre holo- e apo-moléculas de transferrina, este ciclo necessita de um mecanismo molecular para comunicação entre os sítios de ligação de ferro e o receptor, assim garantindo a sua eficiência. Acredita-se que a linguagem desta comunicação é uma mudança conformacional em função da ligação do ferro (veja Lindley *et al*, (1993) e Evans *et al.* (1994), artigos em anexo).

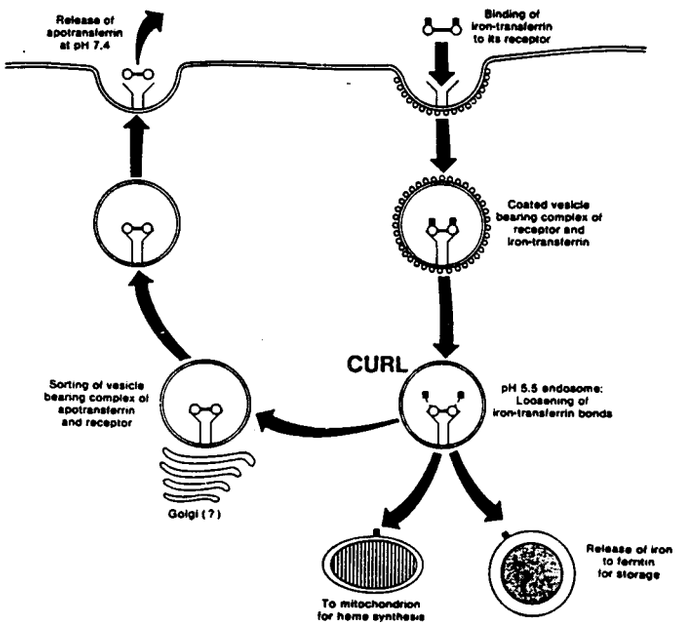


Figura 2.1. O ciclo transferrina

Acredita-se que a função principal da ovotransferrina e lactoferrina seja o sequestro de ferro no ovo e nas superfícies secretoras respectivamente. Desta maneira elas agem como agentes bacteriostáticos através da redução da biodisponibilidade do elemento para bactérias patogênicas. O p97 (também conhecido como melanotransferrina) é um antígeno de membrana expresso em baixos níveis em células humanas normais, mas sobre-expresso em melanomas[4]. A sua funcão permanece desconhecida mas a sua expressão específica chama atencão para a possibilidade de usá-lo como alvo para 'balas mágicas' no tratamento de câncer da pele.

Mais recentemente foram descritas duas proteínas que apresentam homologia sequencial com a superfamília das transferrinas, mas que interagem com outros ligantes: um inibidor de anidrase carbônica de plasma sanguíneo de porco[5] e saxifilina[6], uma proteína de ligação de saxitoxina de sapo. Estas últimas demostram que as transferrinas pertencem a uma família mais ampla de proteínas de ligação, suspeita que foi levantada pela primeira vez com a determinação da estrutura tridimensional da lactoferrina humana[7], que revelou uma topologia surpreendentemente similar à das proteínas de ligação periplasmáticas de bactéria Gram-negativas[8].

A determinação das estruturas cristalográficas de ovotransferrinas[9, Lindley *et al.* (1993) em anexo], transferrina de soro[10, Sarra *et al.* (1990) em anexo], lactoferrinas[7,11,12] e seus fragmentos de várias espécies, tanto na forma apo quanto na forma complexada com ferro, junto com outras técnicas bio- e físico-químicas, têm revelado muito a respeito do mecanismo de ligação e liberação de ferro.

2.1 Estrutura tridimensional

O gene de transferrina em vertebrados é o resultado de uma duplicação gênica num ancestral que codificava uma proteína da metade do comprimento[13]. Esta afirmação é evidente tanto na sequência quanto na estrutura tridimensional da molécula que apresenta dois lobos (N- e C-terminais), cada um com o mesmo enovelamento e composto de dois domínios (denominados I e II). Na fissura entre os domínios de cada lobo encontra-se um sítio de ligação de ferro. Os ligantes do íon férrico são as cadeias laterais de duas tirosinas, uma histidina e um ácido aspártico junto com o ánion carbonato que fornece dois oxigênios à esfera de ligação. Os resíduos que participam na formação do sítio de ligação do ferro provêm de ambos os domínios e da ponte entre eles, de tal maneira que é obrigatório o fechamento dos domínios para poder se ligar o metal.

Com a determinação da estrutura de lactoferrina humana na sua forma livre de ferro (apo)[12] ficou evidente que os dois domínios afastam-se na hora da liberação do metal, fato coerente com dados cinéticos anteriores que evidenciaram uma mudança conformacional durante a ligação reversível do ferro[14] (Figura 2.2). Medidas de transferrina em solução por espalhamento da raios-X a baixo ângulo corroboram os resultados cristalográficos demonstrando que a mudança conformacional observada não é apenas um artefato causado pelas forças de empacotamento dentro do cristal[15,16].

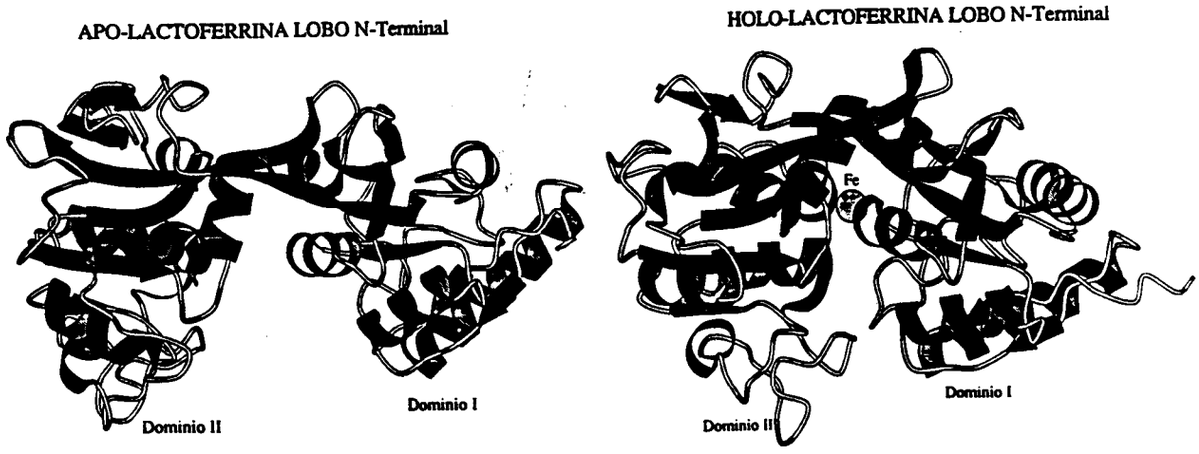


Figura 2.2. As conformações aberta e fechada do lobo N-terminal da lactoferrina

A divisão da molécula de transferrina em quatro domínios não é apenas uma conveniência de nomenclatura. Através de tratamento com enzimas proteolíticas é possível produzir fragmentos da molécula que correspondem aos lobos N- e C-terminais e em certos casos específicos o domínio N-II pode ser isolado do restante da molécula[17-19]. Este último continua com a capacidade de se ligar ao ferro com uma estequeometria de 1:1. Estruturas cristalográficas de diversos fragmentos de várias fontes já foram determinadas, os primeiros contando com a minha participação (Sarra *et al.* e Lindley *et al,* em anexo). Baseado em estudos do fragmento N-terminal de ovotransferrina de galinha, Dewan *et al.*[9] propõe um mecanismo para a abertura dos domínios que eles chamam de uma 'gatilho de lisinas'. A proposta se baseia na observação de duas lisinas inseridas na fissura entre os domínios da holo-proteína com uma separação de menos de 3Å entre os N_ζ. Em consequência da sua proximidade, os pK_a's das lisinas seriam radicalmente deslocados da norma (~10) e portanto não-carregadas a pH neutro. Em contato com o pH ácido do CURL durante endocitose na célula alvo, a protonação das lisinas levaria a uma repulsão eletrostática entre si e em consequência a abertura dos domínios facilitando a liberação do ferro.

2.2 Trabalhos apresentados em seguida

Apresento em seguida quatro artigos relacionados com a superfamília das transferrinas. Os dois primeiros descrevem as estruturas cristalográficas do fragmento N-terminal de transferrina de soro de coelho e o fragmento N-II de ovotransferrina de pato respectivamente. Apesar de não ser artigos de modelagem molecular (tema desta tese), eles foram incluídos pois contribuiram para um melhor entendimento do mecanismo de ligação e liberação de ferro. O terceiro artigo descreve a modelagem de uma variante natural de transferrina que apresenta baixa afinidade tanto pelo ferro quanto pelo seu receptor, em consequência de uma única mutação pontual. O artigo contribuiu para um melhor entendimento entre a ligação de ferro e o receptor à molécula de transferrina. O quarto artigo é um estudo de pura modelagem molecular por homologia, visando desvendar a função da p97.

Sarra, R., Garratt, R.C., Gorinsky, B., Jhoti, H. & Lindley, P.F. (1990) 'High Resolution X-ray Studies on Rabbit Serum Transferrin at 2.3Å Resolution', Acta Cryst. **B46**, 763-771

Lindley, P.F., Bajaj, M., Evans, R.W., Garratt, R.C., Hasnain, S.S., Jhoti, H., Kuser, P., Neu, M., Patel, K., Sarra, R., Strange, R. & Walton, A. (1993) 'The mechanism of Iron Uptake by Transferrins: The Structure of an 18kDa NII -Domain Fragment from Duck Ovotransferrin at 2.3Å Resolution', Acta Cryst. **D49,** 292-304

Evans, R.W., Crawley, J.B., Garratt, R.C., Grossmann, J.G., Neu, M., Aitken, A., Patel, K.J., Meilak, A., Wong, C., Singh, J., Bomford, A. & Hasnain, S.S. (1994) 'Characterization and Structural Analysis of a Functional Human Transferrin Variant and Implications for Receptor Recognition', Biochemistry **33**, 12512-12520

Garratt, R.C. & Jhoti, H. (1992) 'A Model for the Three-dimensional Structure of p97 Provides Evidence for a Potulated Zinc Binding Function', FEBS. Letts **305,** 55-61

Referências

[1] Baker, E.N. & Lindley, P.F. (1992) J. Inorg. Biochem. **47**, 147-160
[2] Baker E.N., Baker, H. M., Smith, C.A., Stebbins, M.R., Khan, M., Hellstrom, K.E & Hellstrom, I. (1992) FEBS Letts **289,** 215-218
[3] Schade, A.L., Reinhart, R.W., & Levy, H. (1949) Arch. Biochem. Biophys. **20,** 170-172
[4] Rose, T.M., Plowman, G.D., Teplow, D.B., Dreyer, W.J., Hellstrom, K.E., & Brown, J.P. (1986) Proc. Natl. Acad. Sci. **83**, 1261-1265
[5] Roush, E.D. & Fierke, C.A. (1992) Biochemistry **31**, 12536-12542
[6] Li, Y. & Moczydlowski, E. (1991) J. Biol. Chem. **266**, 15481-15487
[7] Anderson, B.F., Baker, H.M., Dodson, E.J., Norris, G.E., Rumball, S.V., Waters, J.M. & Baker, E.N. (1987) Proc. Natl. Acad. Sci. **84**, 1769-1773
[8] Quiocho, F.A., Vyas, N.K., Pflugrath, J.W., Saper, M.A., Vyas, M.N. & Sack, J.S. (1985) J. Biosci. **8**, 461-470
[9] Dewan, J.C., Mikami, B., Hirose, M. & Sacchettini, J.C. (1993), Biochemistry **32,** 11963-11968
[10] Bailey, S., Evans, R.W., Garratt, R.C., Gorinsky, B., Hasnain, S., Horsburgh, C., Jhoti, H., Lindley, P.F., Mydin, A., Sarra, R. & Watson, J.L. (1988) Biochemistry **27,** 5804-5812
[11] Anderson, B.F., Baker, H.M., Norris, G.E., Rice, D.W. & Baker, E.N. (1989) J. Mol. Biol. **209,** 711-787
[12] Anderson, B.F., Baker, H.M., Norris, G.E., Rumball, S.V., & Baker, E.N. (1990) Nature (London) **344,** 784-787
[13] Williams, J. (1982) Trends Biochem. Sci. **7**, 394-397
[14] Cowart, R., Kojima, N. & Bates, G. (1982) J. Biol. Chem. **257,** 7560-7565
[15] Kilar, F. & Simon, I. (1985) Biophys. J. **48,** 799-802
[16] Grossmann, j.G. Neu, M., Pantos, E., Schwab, F., Evans, R.W., Townes-Andrews, E., Lindley, P.F., Appel, H., Theis, W.-G., & Hasnain, S.S. (1992) J. Mol. Biol. **225,** 811-819
[17] Williams, J. (1974) Biochem. J. **141**, 745-752
[18] Williams, J. (1975) Biochem. J. **149**, 237-244
[19] Evans, R.W. & Madden, R.W. (1984) Biochem. Soc. Trans. **12**, 661-662

De acordo com as políticas editoriais, estes artigos não podem ser depositados em repositório de acesso aberto. Para acesso aos artigos completos entre em contato com o(a) autor(a) ou com o Serviço de Biblioteca e Informação IFSC - USP (bib@ifsc.usp.br).

SARRA, R.; GARRATT, R. C.; GORINSKY, B.; JHOTl, H.; LINDLEY, P. High-resolution x-ray studies on rabbit serum transferrin: preliminary structure analysis of the N-terminal half-molecule at 2·3 Â resolution. **Acta Crystallographica B**,Copenhagen, v.46, p.763-771, 1990.

LINDLEY, P.F.; BAJAJ, M.; EVANS, R.W.GARRATT, R.C.;HASNAIN, S.S.; JHOYI, H.; KUSER, P.; NEU, M.; PATEL, K.; SARRA, R.; STRANGE, R.; WALTON, A. The mechanism of iron uptake by transferrins: the Structure of an 18 kDa NII-Domain fragment from duck ovotransferrin at 2.3 Á resolution. **Acta Crystallographica D,** Copenhagen,v.49, p.292-304, 1993**.**

TAVARES, R. W.; CRAWLEY, J. B.; GARRATT, R.C.; GROSSMANN, J. G.; NEU, M.; AITKEN, A.; PATEL, K. J.; MEILAK, A.; WONG, C.; SINGH, J.; BOMFORD, A.; HASNAIN, S S. Characterization and structural analysis of a functional human serum transferrin variant and implications for receptor recognition. **Biochemistry,** New York, v.33, p.12512-20, 1994.

GARRATT, R. C.; JHOTI, H. Molecular model for the tumour-associated antigen, p97, suggests a Zn-binding function. **Febs Letters**, Amsterdam, v.305, n.1 , p.55-61, June 1992.

Capítulo 3

PROTEÍNAS DO PARASITA
Schistosoma mansoni

A esquistossomose é causada por platelmintos tremododos do gênero *Schistosoma* e afeta cerca de 200 milhões de pessoas mundialmente segundo estimativas da Organização Mundial de Saúde[1]. Países tropicais e subtropicais são os mais afetados onde as condições sócio-econômicas, principalmente a falta de saneamento básico e água tratada em regiões endêmicas, além da presença do hospedeiro intermediário (no Brasil, caramujo do gênero *Biomphalaria*) são responsáveis pela infecção e reinfecção da faixa mais pobre da população. A *Schistosoma mansoni* é a mais frequente mundialmente e a única espécie causadora da doença no Brasil[2].

Apesar da existência de drogas eficazes como oxaminiquini, metrifonato e praziquantel, a reinfecção de indivíduos após tratamento em áreas endêmicas é comum. Há portanto bastante interesse no desenvolvimento de vacinas eficazes contra o parasita[3,4]. Para tal, é necessário um melhor entendimento dos mecanismos protetores que levam à resistência e proteção contra reinfecção no hospedeiro e também os mecanismos usados pelos vermes adultos do parasita que vivem na corrente sanguínea, para escapar da resposta imune montada pelo hospedeiro. Várias evidências sugerem que uma vacina efetiva deveria ser direcionada contra a forma larval (esuistossômulo) do schistosoma. Não é uma tarefa fácil

lembrando que até hoje não existe nenhuma vacina de uso geral para doenças parasíticas humanas[5].

Nos últimos anos tem sido feito um eforço para encontrar antígenos candidatos à vacina e a sua clonagem e expressão em grandes quantidades por técnicas de biologia molecular. Atualmente um programa financiado pelo Programa Especial para Pesquisa e Treinamento em Doenças Tropicias (TDR) da Organização Mundial de Saúde está sendo realizado visando a identificação dos antígenos mais promissores para um subsequente desenvolvimento[3]. Foram selecionados seis antígenos, entre eles a proteína de ligação de ácidos graxos Sm14, cujo estudo de modelagem molecular encontra-se no artigo de Tendler *et al* (em anexo). A Tabela 3.1 apresenta alguns dados relevantes dos seis antígenos sendo testados.

		Proteção(%)		
Antígeno	Estágio	Camundongo	Rato	Outros
Sm28GST	Esquistossômulo e Verme Adulto	30-60	40-60	40 (Babuínos)
Paramiosina	Esquistossômulo e Verme Adulto	30	-	-
Miosina	Todos	50-70	95	25 (Babuínos)
TPI	Todos	30-60	-	-
Sm23	Todos	40-50	-	-
Sm14	Esuistossômulo	65 (outbred)	-	90-100 (coelhos)

Tabela 3.1 Candidatos à vacina anti-esquistossomose[3]

Em 1994 foi iniciado pelo TDR um programa de projetos genoma de parasitas, incluindo Filária, *Trypanosoma cruzi, Trypanosoma brucei, Leishmania* e *Schistosoma.* Os principais objetivos desta iniciativa eram o mapeamento dos genomas, a montagem de bancos de dados de sequências e o treinamento de cientistas de países em desenvolvimento através de colaborações internacionais[6]. O projeto de *S. mansoni* atualmente conta com a participação de laboratórios no Brasil, na Inglaterra, no Egito e no Japão e visa mapeamento físico, construção de bibliotecas de cDNA para as diversas fases do ciclo de vida e principalmente descoberta gênica. Um dos estímulos para a montagem do projeto foi a escassez de informação disponível ao nível molecular de um organismo de grande importância médica. Estima-se que o genoma de schistosoma contenha cerca de 20,000 genes expressos dos quais apenas 1% foi completamente sequenciado[7]. Além das informações valiosas a respeito da organização genomica do parasita, a expressão diferencial de genes durante as várias fases do ciclo de vida e a sua regulação é de se

esperar que o projeto genoma de *S. mansoni* contribua para a identificação de novos antígenos que possam ser aproveitados como componentes de vacinas potenciais.

Através de sequenciamento intensivo, Franco *et al.*[8] produziram mais de 600 etiquetas de sequências transcritas (Expressed Sequence Tags - ESTs) correspondentes a 169 genes do parasita. Destas, apenas 16% tinham sido descritas anteriormente, 22% eram homólogas a sequências de outros organismos e em 33% dos casos não foi detectada nenhuma homologia através de busca nos bancos de dados. Duas sequências completas foram determinadas para dois dos genes descobertos, sendo a modelagem da proteína correspondente (proteína de ligação ao elemento Y-box, SMYB1) descrito no artigo de Franco *et al.* (em anexo).

3.1 Trabalhos apresentados em seguida

Tendler, M., Brito, C.A., Vilar, M.M., Serra-Freire, N., Diogo, C.M., Almeida, M.S., Delbem, A.C.B., da Silva, J.F., Savino, W., Garratt, R.C., Katz, N. & Simpson, A.J.G. (1996) 'A *Schistosoma mansoni* fatty acid-binding protein, Sm14, is the potential basis of a dual-purpose anti-helminth vaccine', Proc. Natl. Acad. Sci. **93,** 269-273

Franco, G.R., Garratt, R.C., Tanaka, M., Simpson, A.J.G. & Pena, S.D.J. (1996) 'Characterization of a *Schistosoma mansoni* Gene Encoding a Homologue of the Y-box binding protein', submitted to J. Biol. Chem.

Referências

[1] World Health Organization (1991) - TDR 10th Programme Report, Geneva, pp41-47
[2] Johnson, D.A. *et al.* (1993) Parasitology Today **9,** 286-291
[3] Bergquist, N.R. (1995) Parasitology Today **11**, 191-194
[4] Dunne, D.W., Hagan, P. & Abath, F.G.C. (1995) Lancet **345**, 1488-1492
[5] Mackett, M & Williamson, J.D. (1995) in 'Human Vaccines and Vaccination', BIOS Scientific Publishers, Oxford, UK
[6] Pena, S.D.J. (1996) Trends Biotech. in press
[7] GenBank release 92, 18/12/95
[8] Franco, G.R., Adams, M.D., Soares, M.B., Simpson, A.J.G., Venter, C. & Pena, S.D.J. (1995) Gene **152**, 141-147

De acordo com as políticas editoriais, este artigo não pode ser depositado em repositório de acesso aberto. Para acesso ao artigo completo entre em contato com o(a) autor(a) ou com o Serviço de Biblioteca e Informação IFSC - USP (bib@ifsc.usp.br)

TENDLER, M.; BRITO, C. A.; VILAR, M. M.; SERRA-FREIRE, N.; DIOGO, C. M.; ALMEIDA, M. S.; DELBEM, A. C. B.; SILVA, J. F.; SAVINO, W.; GARRATT, R. C.; KATZ, N.; SIMPSON, A.J. G. Schistosoma mansoni fatty acid-binding protein, Sm14, is the potential basis of a dual-purpose anti-helminth vaccine. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, v. 93, p. 269-273, Jan. 1996.

Characterization of a *Schistosoma mansoni* Gene Encoding a Homologue of the Y-box Binding Protein*

Glória R. Franco§, Richard C. Garratt‡, Manami Tanaka¶, Andrew J. G. Simpson‖, and Sérgio D. J. Pena §**

From the §Departamento de Bioquímica e Imunologia, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil 31270-010, the ‡Departamento de Física e Informática, Universidade Federal de São Carlos, São Carlos, SP, Brazil 13560-970, the ¶ Institute of Basic Medical Sciences, University of Tsukuba, Tsukuba, Ibaraki 305, Japan and the ‖Instituto Ludwig de Pesquisas para o Câncer, São Paulo, SP, Brazil

** To whom correspondence should be addressed at: Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas - UFMG, Av. Antônio Carlos 6627, Belo Horizonte, MG, Brazil, 31270-010, Tel: (5531) 441-5611, FAX: (5531) 441-5963, email: spena@dcc.ufmg.br

RUNNING TITLE: Characterization of a *S. mansoni* YB protein encoding gene

SUMMARY

We have cloned and characterized a *Schistosoma mansoni* cDNA encoding a basic protein homologous to the human Y-box binding protein 1. The 1.3 Kb *S. mansoni YB-1* transcript, that was shown to be expressed in various stages of the parasite life cycle, codes for a 217 aa protein containing, towards its N-terminus, a nucleic acid binding motif, known as binding cold-shock domain. This domain is 64% identical to the cold-shock domains of other members of the Y-box binding protein family and 43% identical to the cold-shock protein CspA of *Escherichia coli*. In *S. mansoni* YB-1, the cold-shock domain possess some structural characteristics that permit dimer formation as occurs in the *Bacillus subtilis* cold-shock protein CspB. The C-terminal region of *S. mansoni* YB-1 differs from the other Y-box binding proteins due to the presence of tandem repeats of Arg and Gly, suggesting the formation of a fibroin-like β-sandwich structure. This novel folding pattern for the C-terminus of *S. mansoni* YB-1 might suggest a distinct specific function for this protein in the parasite.

INTRODUCTION

The Y-box element is a consensus sequence motif present in the major histocompatibility complex (MHC)[1] class II gene promoters. It contains a CCAAT inverted sequence, which has been shown to specifically interact with several families of transcription factors (1-3). One of these factors is the Y-box binding protein (YB-1), that was described in humans and was suggested to regulate negatively the expression of the MHC class II gene, *HLA-DRA* (4). However, YB-1 is not specific to lymphoid cells, being ubiquitously expressed in human tissues (5, 6). Proteins homologous to YB-1 have been cloned from rat, mouse, chicken, *Xenopus leavis* and from the marine invertebrate *Aplysia californica* by probing cDNA libraries with oligonucleotides containing the CCAAT inverted element derived from promoter sequences of different genes (7-12). Moreover, humans, mice and *Xenopus* all have more than one different Y-box binding protein, thus suggesting the existence of intraspecies families. Y-box binding proteins were shown to bind to both double stranded and single stranded DNA and also to mRNA (see refs. 13 and 14 for a review). They were demonstrated to act as transcription factors both activating or repressing genes in several organisms (7, 10, 15, 16) and also to regulate translation by binding and sequestering mRNA (13, 14, 17-20). Thus, diverse members of the Y-box binding protein family may have different biological roles.

All Y-box binding proteins have been shown to contain a conserved DNA-binding domain, known as the cold-shock domain (CSD), that shows substantial similarity to the *E. coli* (CspA) and *Bacillus subtilis* (CspB) cold-shock proteins which regulate transcriptional activation of cold-shock genes (21-23). The three-dimensional structures of CspA and CspB have been determined both for the crystallized proteins (24, 25) and in solution (26, 27), showing them to be folded into 5 anti-parallel β-strands with connecting turns and loops, creating a closed β-barrel (the so-called OB fold) (28). The crystal structure of CspB, but not that of CspA, appears to be that of a dimer stabilized by six hydrogen bonds linking two adjacent β4 stands (25). The surface of both cold shock proteins is characteristic for a protein interacting with single-strand nucleic

acid. The binding site is located on the side of the barrel where there is an arrangement of positively charged amino acids that can interact with nucleic acids together with the side chains of hydrophobic residues, all of which are conserved in the cold-shock domain of Y-box binding proteins (24-27, 29).

The trematode parasite *Schistosoma mansoni* is responsible for schistosomiasis, a chronic disease afflicting hundreds of millions of people in tropical and subtropical areas (30). As part of a gene discovery program in *S. mansoni* (31) we have identified a homologue of *YB-1* from an adult worm cDNA library. We have called this gene *SMYB1* (for *S. mansoni Y-box binding protein* gene). We here report the complete cDNA sequence, the chromosomal location, the partial genomic structure and the expression pattern of *SMYB1* in different developmental stages of *S. mansoni*. Based on the known crystal structure of CspA and CspB we have used computational modeling to propose a three dimensional structure for the cold-shock domain of the SMYB1 protein. We here present this model, together with evidence that SMYB1 may also form dimers. Moreover, we propose a fibroin-like structure for the C-terminus of SMYB1 in contrast with most other Y-box binding proteins that were suggested to fold as a charged zipper domain (8, 13-15). Our data support the notion that Y-box binding proteins form a functional and structural family having in common a similar cold-shock domain and different C-terminal domains that determine its specific functions.

MATERIALS AND METHODS

Random cDNA clone selection, production and identification of Expressed Sequence Tags (ESTs) - Clones from a plasmid adult worm cDNA library were randomly selected and used in single pass automated DNA sequencing to produce ESTs as described (31). Identification of ESTs obtained from one or both cDNA strands was performed by comparison with all DNA and protein sequences deposited in non-redundant databases, using the Basic Local Alignment Search Tool (BLAST) (32) email server at the National Center for Biotechnology Information site.

Full length cDNA sequencing - Five cDNAs randomly selected from the library matched the human *YB-1* gene. The largest cDNA insert (1.2 Kb) was obtained by double digestion of the recombinant clone SMPAC89 with the *Eco*RI and *Hin*dIII restriction enzymes. Fragments of this insert were produced after digestion at the specific restriction sites shown in Fig 1 and subcloned into the pUC18 vector (Pharmacia). DNA sequence analysis was carried out on both strands using either Auto Cycle or Auto Read Sequencing kits (Pharmacia) and the A.L.F. DNA sequencer (Pharmacia). Partial sequences were aligned to produce a consensus sequence using the DNAsis program. Additional 71 bp at the 5` end were obtained by polymerase chain reaction (PCR) amplification of the whole adult worm cDNA plasmid library using a vector-specific primer (M13R 5' CAGGAAACAGCTATGAC 3') and the primer Yboxext 5` AGCTGGTCTAGTGTCCGC 3', which was anchored at the beginning of the 5` end of the cDNA and directed to the vector region. The amplification reaction mixture contained 20 ng of the plasmid cDNA library, 0.2 μM of each primer, 2 mM MgCl2, 200 μM dNTPs and 1 U Taq DNA polymerase (Promega) in a 20 μl final volume of a specific reaction buffer (50 mM KCl, 10 mM Tris-HCl pH 9.0 and 0.1% Triton X-100). The thermal profile comprised 30 cycles of denaturation at 95 °C for 1 min, annealing at 50 °C for 1 min and extension at 72 °C for 1 min. Amplification products were analyzed on a silver-stained 6%

polyacrilamide gel. Fragments of different sizes were cloned into the pUC18 vetor (*Sma*I site), using the SureClone Ligation kit (Pharmacia) and sequenced as before.

Genomic sequencing - A fragment of genomic DNA, consisting of the complete coding region of the *SMYB1* gene, was obtained by PCR amplification using the upstream primer YboxF containing a custom *Eco*RI site (underlined) 5' CTGGAATTCACAATGGCGGACACTAGAC 3' and the downstream primer YboxR containing a custom *Pst*I site (underlined) 5' CTGGACGTCAAAATGCATATTTGATTACG 3'. The PCR reaction mixture was prepared using 200 ng of *S. mansoni* total DNA, 0.4 μM of each primer, 2 mM MgCl2, 200 μM dNTPs and 2.5 U Taq DNA polymerase (Promega) in a 50 μl final volume of a specific reaction buffer (50 mM KCl, 10 mM Tris-HCl pH 9.0 and 0.1% Triton X-100). The thermal profile was the same as that described above. The 2.0 Kb fragment, obtained after PCR amplification, was digested with *Eco*RI and *Pst*I restriction enzymes and cloned into the pUC18 vector. Segments of this insert were produced after digestion at specific restriction sites and subcloned into the same vector. DNA sequence analysis was carried out on both strands as described above.

Chromosomal mapping - A yeast artificial chromosome (YAC)-library was constructed with partially digested parasite genomic DNA. Two *SMYB1* cDNA clones identified by the EST strategy were used to screen gridded colony arrays on nylon filters of the YAC library (33). YAC clones hybridizing with cDNA inserts were localized into mitotic metaphasic chromosomes by chromosomal *in situ* suppression hybridization (CISS) as described (34).

Reverse transcribed PCR (RT-PCR) - To detect *SMYB1* transcripts in different stages of the parasite life cycle, the primers YboxM (5' GCATGTCTACGTTTCAACTCA 3') and YboxR were used to amplify a fragment of 458 bp from the 3' end of first-strand cDNAs. Oligo (dT)-primed first-strand cDNAs were produced by reverse transcription of mRNA obtained using the Micro-Fast Track mRNA Isolation kit (Invitrogen), as follows: 10 μl (< 1μg) of mRNA was mixed with 500 ng oligo (dT)12-18, 500 μM dNTPs, 10mM DTT and 200 U SuperScript II (Gibco,BRL) in a total reaction volume of 20 μl, containing 50 mM

Tris-HCl pH 8.3, 75 mM KCl and 3 mM MgCl2. The synthesis reaction was performed at 37 °C for 3 h. 1 µl of the first-strand cDNA was added to 30 µl of the PCR reaction mixture containing 0.4 µM of each primer, 2 mM MgCl2, 200 µM dNTPs and 1 U Taq DNA polymerase (Promega) in a specific reaction buffer (50 mM KCl, 10 mM Tris-HCl pH 9.0 and 0.1% Triton X-100). The thermal profile was the same as that described above. Amplification products were analyzed on silver-stained 6% polyacrilamide gels.

Modelling of the cold-shock domain - Residues 22 to 95 of SMYB1 present 40 and 43% sequence identity with the cold-shock proteins, CspB and CspA, from *B. subtilis* and *E.coli* respectively. This level of sequence identity is sufficient to infer a homologous relationship and therefore a similar three-dimensional structure (35) which permits the construction of a molecular model using homology modelling techniques. The model for the cold-shock domain of SMYB1 was based on the main-chain structure of CspA (25) (atomic coordinate set 1mjc of the Brookhaven databank (36)). The choice of CspA over CspB was based on an assessment of the crystallographic resolution and R-factor, the stereochemistry of the deposited coordinates, the quality of the structure as evaluated by atomic contact analyis (37) and the percentage sequence identity. In all cases CspA was superior to CspB, with the exception of the stereochemistry of the models which were comparable. For the construction of the molecular model initially the three sequences were aligned with the program MULTALIGN of the AMPS package (38), which resulted in no insertions or deletions within the elements of secondary structure. For the mostpart, the mainchain coordinates were taken directly from CspA with the exception of the fourth beta-strand and the large loop between strands 3 and 4. The C-terminal region of the fourth strand (residues 79-83) was based on the conformation of CspB and was modelled after least-squares superposition of the two structures. The loop was modelled by the procedure of Jones and Thirup (39), using the DGLP option of the molecular graphics program TOM (40) and subsequent manual/automated adjustment in order to regularize the geometry. Sidechains which were identical to those of CspA were left unaltered with the exception of Met22 which was omitted because its homologue in CspA is not observed in the crystal structure. Of the remaining 42 residues in the model, side chain rotamers were

copied from CspB where the residues were either identical or structurally analogous. The outstanding residues were orientated on the basis of the local backbone conformation (41) or on the known preferences for the given secondary structure (42). Residues presenting steric hindrance were identified by the program PROCHECK (43) and manually adjusted prior to energy minimization, which was performed using the GROMOS (44) option of the program WHATIF (45). A model for a potential dimer formed by two identical Ybox monomers in an analogous fashion to that observed in CspB (25), was also examined. For this purpose, the above model was inicially superimposed on the crystal structure for CspB (Brookhaven file 1csp) using only the C-alpha coordinates for the fourth beta-strand. The C-terminal half of this strand together with its symmetry-related partner form hydrogen bonds across the dimer interface. The second monomer was created by applying the rotation matrix corresponding to the crystallographic diad axis to the original model. The model was assessed using the programs PROCHECK (43), VERIFY_3D (46) and the quality control option of WHATIF (45).

The C-terminal (fibroin-like) domain - There is no known three-dimensional structure of a homologue to the C-terminal domain of SMYB1 protein. There is thus insufficient structural information for the construction of a full three-dimensional model. However, there are several regions within this domain in which glycine appears at every second position in the sequence, suggestive of the beta-sheet sandwich structure believed to occur in silk fibroin (47). Within these regions, arginine normally intercalates between the glycines. In order to visualize the distribution of the intercalating residues and in particular the charges on the arginines, the repetitive sequences were identified by manual inspection and one of many possible models for the core of this domain was built assuming the (RG)n repeats to form two stacked beta-sheets with the glycines at the interface. For this purpose, the antiparalel in-register silk II structure was used (coordinate set 2slk (48)), the appropriate residue substitutions being made where necessary.

RESULTS

Characterization of a S. mansoni adult worm cDNA encoding a Y-box binding protein - Using the EST strategy we have isolated and identified five cDNA clones from an adult worm library homologous to the human *Y-box binding protein 1* (31). In order to obtain the full-length sequence of the cDNA, the longest insert of 1.2 Kb (clone SMPAC89) was digested at specific enzyme restriction sites, shown in Fig. 1 and the fragments were subcloned into the pUC18 vector (Pharmacia) and entirely sequenced. Sequences from both DNA strands were assembled to obtain a consensus. The 5' end region of 71 bp flanking the original cDNA was obtained by PCR amplification, cloned into the pUC 18 vector and also sequenced. The complete cDNA insert is 1283 bp long and contains a poly (A) tail of 21 nt, but no characteristic polyadenilation signal (Figs. 1 and 2a). The putative ATG initiation codon is located at position 7 and is in agreement with the Kosak's consensus sequence that signals translation initiation (49). The cDNA was translated into the six possible frames and the length of the longest open reading frame (ORF) was 651 encoding a predicted 217 aa protein. There is also a long 3' untranslated region (Figs. 1 and 2a).

Two *SMYB1* cDNAs were also used to screen a *S. mansoni* YAC library in order to obtain clones for gene localization into metaphasic chromosomes. From the screening of gridded colony nylon filters, three YACs were selected and analyzed by pulsed-field gel electrophoresis (PFGE) in order to verify the size of the inserts and to confirm hybridization by Southern blotting (33). YAC clones identified by this way were localized into metaphasic chromosomes by CISS (34). The *SMYB1* gene was shown to be located in the middle region of the distal (q) arm of both chromosomes Z and W, immediately alongside the heterochromatic region (data not shown).

Analysis of the primary structure of the SMYB1 protein reveals six potential sites for phosphorylation by casein kinase II (CK-II), cAMP-dependent kinase and protein kinase C (PKC) (Fig. 2a). Several sites for modification by kinases have also been seen in other Y box binding proteins (7, 12, 13). No

N-linked glycosylation site was seen in the predicted protein. The sequence presents a cold-shock domain towards the N-terminus, aa 22 - 95 (Figs. 1 and 2a) which is homologous to the equivalent domain in other eukaryotic YB proteins (64% identity) and to the bacterial cold-shock protein CspA (43% similarity) (Fig. 3a and b). This extremely conserved domain folds as a β-barrel structure that contains a nucleic acid binding site, located on the surface of the N-terminal β-sheet. The binding site is formed by the side chains of basic and hydrophobic amino acids (24-27, 29). Some of these hydrophobic residues are present in the RNP1 (β-strand 2) and RNP2 (β-strand 3) consensus motifs characteristic of RNA binding proteins (50, 51) (Fig. 3a).

The carboxyl-terminal region of the protein is extremely basic as a result of a high arginine content (25/109 = 22.9%) and is rich in glycine (34/109 = 31.2%). Attempts to align this domain with that of the human protein YB1 for example leads to a sequence identity of only 21%, principally due to the high arginine content (Fig 3a) and this is the less similar region among all Y-box binding proteins. Although the similarity of the predicted protein with other Y-box binding proteins is very low in regions outside of the cold-shock domain, the characteristics of presenting a conserved cold-shock domain together with a tail composed mainly of basic residues lead us to designate this protein SMYB1 (*Schistosoma mansoni* Y-box binding protein 1). The C-terminal domain of the protein was searched for conserved motifs against the SBASE library of protein domains at Trieste, Italy, which resulted in significant similarities only with nuclear proteins enriched in arginine and glycine such as nucleolin (52). Stretches of basic amino acids, that are present in other Y-box binding proteins and are similar to previously described potential nuclear translocation signals (53), were not seen in the C-terminal domain of SMYB1 (Fig. 3a).

Isolation and sequencing of a genomic fragment of SMYB1 - A fragment of *S. mansoni* genomic DNA containing the *SMYB1* complete coding region was amplified using primers YboxF and YboxR. The 2.0 Kb amplified product was purified and digested at specific enzyme restriction sites. The fragments were subcloned into the pUC18 vector and totally sequenced. Fig. 2b shows the complete sequence of the *SMYB1* gene of 1980 bp containing three introns. The first intron is the smallest (32 bp long) and is located after the

cold-shock domain (nt 308-339) in the small linker region that connects the two protein domains. The second intron is 466 bp long and is located in the central part of the C-terminal domain (nt 505-970). The third intron, of 298 bp, is located towards the end of the coding region (nt 1131-1428). It is the only intron that interrupts the frame of the protein. The canonical donor/acceptor splice sites are conserved at all exon/intron junctions (Figs. 1 and 2b).

Developmental expression of SMYB1 mRNA - Fig. 4 shows the developmental expression of *SMYB1* assayed by RT-PCR. The 458 bp fragment was amplified from mRNA preparations of all stages of the parasite life cycle. The differences in the band intensities could be due to different amounts of mRNA present in each preparation, since they were not exactly quantified. The expression of *SMYB1* mRNA in male and female adult worms was also analyzed by Northern blot that demonstrated the presence of an abundant 1.3 Kb transcript in both sexes (data not shown).

The cold-shock domain - For the purposes of its description the sequence of SMYB1 protein was conveniently divided into four regions in common with other members of the Y-box binding protein family. We refer to these as the N-terminal region (residues 1-21), the cold-shock domain (22-95), the linker region (96-108) and the C-terminal or fibroin-like domain (109-217) (Fig. 2a).

The cold-shock domain is preceded by a 21 residue hydrophilic N-terminal tail which has no homologue in the bacterial cold-shock proteins and therefore could not be incorporated into the model. The model for the monomeric form of the SMYB1 cold-shock domain itself shows good stereochemistry as evaluated by the program PROCHECK (43). The atomic contact quality calculated by the method of Vriend and Sander (37) was -1.37, slightly lower than those of the starting structures CspA (-0.84) and CspB (-1.23) as expected for a model structure. The Eisenberg chemical environment index was measured as 29.8, close to the expected value of 32.9 for a polypeptide of 73 residues. All the above indices indicate both the stereochemistry and folding of the model to be reasonable. The atomic contact quality shows a marked improvement when the model is evaluated in its dimeric form, increasing from -1.37 to -1.21. All of the

residues involved in the antiparallel β-ladder formed by the fourth β-strand of each subunit at the dimer interface (residues 79 to 84) improve in quality on dimer formation. Of particular note are valines 80 and 81 which become buried on dimerization, the latter contacting its symmetry-related partner across the molecular two-fold axis.

The cold-shock domain possesses the OB fold, a barrel of five antiparalel strands which is conveniently divided into two sheets (28). The N-terminal sheet consists of the first three beta strands and the C-terminal sheet is composed of the fourth and fifth (Fig. 5). The cluster of basic and aromatic residues previously identified on the external surface of β-strands 1-3 as the nucleic acid binding site in the cold-shock domains (29) are largely conserved in SMYB1 and include the RNP1 and RNP2 RNA binding consensus sequences (50, 51) (Figs. 2a and 5). This suggests that the site of interaction with nucleic acid is probably conserved as has been suggested for other members of the Y-box family (14). Those residues expected to participate in nucleic acid recognition are Lys31, Lys36, His54, Arg68, Lys85, Trp32, Tyr39, Phe41 and Phe52 which correspond to Lys10, Lys16, His33, Lys43, Lys60, Trp10, Phe18, Phe20 and Phe31 in CspA.

The sequences of the SMYB1 β-strands are equally similar to CspA and CspB with the exception of the fourth strand which is more similar to CspB and is responsible for dimerization (Figs. 3b, 6a and 6b). In CspA, dimerization does not occur due to an N-terminal extension which blocks the dimerization site on the fourth strand. Several changes to the sequence of this strand are also observed which are consistent with the monomeric structure. In particular, the buried Val81 in CspB is replaced by glutamic acid in CspA. In SMYB1 the large N-terminal extension prior to the cold-shock domain might be expected to hinder dimerization. In fact, the N-terminal sequence of SMYB1 is rather different from that of CspA including two glutamic acids at positions 23 and 24 which could be successfully modeled leading into the solvent and avoiding the dimerization interface. The dimer structure naturally results in bringing Arg25 from one subunit and Asp84 of the other into close proximity favoring the formation of a salt bridge which would be expected to further stabilize the dimer (Fig. 6b). On dimer formation the aromatic clusters of the binding site point in

roughly opposite directions on the front surface of the molecule as seen in Fig. 6a. Curiously, the centers of the two clusters in the dimer are separated by ~35A, which is similar to the 34A separation observed between the recognition helices of dimeric DNA binding proteins containing the helix-turn-helix motif and which corresponds to the pitch of the B-form DNA double helix.

The C-terminal (fibroin-like) domain - The cold-shock domain of SMYB1 is connected to the C-terminal domain by a linker region (residues 96 to 108) which is conserved in all YB1 proteins, being nine of the 13 residues identical in both human YB-1 and SMYB1 (Fig. 3a). The size of this linker makes it difficult to predict the relative orientations of the cold-shock and C-terminal domains. However, the division of SMYB1 into two domains is consistent with the presence of the first intron, between residues K101 and G102, roughly at the center of the linker region.

In common with other members of the Y-box family, SMYB1 possesses a hydrophilic and highly charged C-terminal domain. The domains are also of very different length and show significant differences in the distribution of charges. The classical charge zipper domain as described for other Y-box binding proteins (7, 8, 13-15), is not as evident in SMYB1 where the acidic regions are less pronounced. Furthermore the positive charges in the *S. mansoni* protein (principally due to arginines) are almost exclusively located in stretches of (Arg-Gly)n repeats where n can be as large as 4 (Fig. 2a). This is in contrast to that of other members of the family where the arginines are typically located in continuous stretches of up to four residues and are not intercalated by glycines (13, 14), except for a single (Arg-Gly)n repeat in *A. californica* Y-box binding protein (12). This difference in the pattern of arginines may not influence greatly the overall charge of the domain but may have serious implications for its three-dimensional structure.

Sequences which repeat every other residue are suggestive of beta strands since the repetition corresponds to that expected for straight strands of planar β-sheets. Furthermore the presence of glycine at every second position raises the possibility that the (Arg-Gly)n repeats may form strands which participate in a β-sandwich similar to the classical structure proposed for silk II fibroin (47). Such a structure is planar, in-

register and antiparallel. The sequences 109-113, 121-128, 145-152, 167-173, 193-197, and 199-201 were identified from the sequence as being compatible with such a structure. Figs. 7a and 7b show one of the many imaginable topologies for the six sheet strands in which two beta-sheets pack against one another with the glycine side-chains at their interface. We imagine that the core of the C-terminal domain is based on this fibroin-like structure with the intervening regions forming the connections between the β-strands. However, Fig. 7 should not be interpreted as a definitive model for the core of the domain as we recognize that other topologies are equally valid. Indeed, due to the possibility of introducing β-bulges in the edge strands it is not even possible to unambiguously determine the extent of each strand. In particular, a two-fold symmetric arrangement for the β-strands is also attractive. Alternatively it should be remembered that an arrangement in which the strands form a single sheet (rather than two) seems equally attractive. In such a case the sandwich could only form on dimerization of two molecules, resulting in the burying of the glycines at the dimerization interface. Such an interaction may serve to stabilize further the dimer proposed above for the cold-shock domain. Whatever the details may be, it seems clear that the fibroin-like structure will lead to a sandwich in which the arginines decorate the two outer surfaces (Fig. 7b). Such a cluster of positive charges would be expected to be a strong nucleic acid attractor and may be functionally important for recognition.

DISCUSSION

Y-box binding proteins have been implicated in distinct regulatory functions and have been shown to be ubiquitous proteins containing the nucleic acid binding domain CSD, that is conserved from bacteria to man (13, 14). The CSD is sufficient to bind Y-box sequences *in vitro* and to induce the expression of bacterial cold-shock genes (54, 55), as well as to specifically recognize RNA (56) by interacting with the hexamer 5' AACAUC 3' (57). Furthermore, removal of this region from *Xenopus* YB-1 leads to a loss of DNA binding (15). The C-terminal region of the Y-box binding proteins is much less conserved among all members of this family and contains the so-called charge zipper domain that has been suggested to be responsible for protein-protein multimerization. It consist of a pattern of alternating positively and negatively charged amino acids that can form a charged zipper (7, 8, 13-15). The C-terminal region was also suggested to be important for RNA recognition and binding through interaction with islands of basic/aromatic amino acids (58), but was recently demonstrated to have only nonspecific interactions with these molecules (57). In *S. mansoni* we proposed the protein has also two distinct structural domains: a N-terminal region containing the CSD and a basic C-terminal region that may assume a completely distinct three dimensional conformation from the other Y-box binding proteins

The cold-shock domain of SMYB1 folds as a β-barrel in a very similar conformation as seen for CspA and CspB. This domain presents a slightly more pronounced segregation of charges than either CspA or CspB. The acidic side of the domain (beta4, beta5 and the N- and C- termini) may also serve to aid in directing DNA to the binding site by charge repulsion and/or to interact with the C-terminal domain of the protein. It seems probable that SMYB1 forms a dimer analogous to that observed in CspB. Noteworthy is the fact that all known Y-box binding proteins also possess sequences for the fourth β-strand which are consistent with dimerization. Thus the cold-shock domain may be at least partly responsible for the observed dimerization of such factors which may not be solely due to the C-terminal domain as proposed elsewhere (7,

8, 13-15). This might be expected to be of relevance to their function given that dimerization is a common feature of many families of nucleic acid binding proteins.

Other aspect that has to be pointed out is the probable regulation of the binding activity and dimer formation by phosphorylation. This seems to be very important for the function of the YB proteins. As an example, in the Y-box binding protein 2 of *Xenopus leavis* phosphorylation was reported to be required for the optmal RNA-binding activity (59). The cold-shock domain of SMYB1 possess three kinase recognition sites, two of them located at loop 2, that separates RNP1 and RNP2 motifs, responsible for nucleic acid binding and a third one located at the end of loop 3, immediately before the fourth β-strand, which is responsible for dimerization (Figs. 2a and 3a). Interestingly, all the three putative phosphorylation sites are conserved in the members of the Y-box binding protein family.

The C-terminal domain of the *S. mansoni* protein appears very different from its counterparts in other family members, the only real similarity being the high content of arginines. Thus, according to the presence of stretches of (Arg-Gly)n repeats, we proposed a model for the three dimensional structure of this domain based on the topology of the silk II fibroin molecule. This domain may be constituted of two β-sheets forming a β-sandwich with all glycine side-chains directed to the interface and arginine side-chains to the outer surface of the sandwich. Consistent with such a model is the fact that the second intron divides the C-terminal domain in two, suggestive of an internal gene duplication. The two halves of the domain present 38% sequence identity. Also interesting is the presence of three potential phosphorylation sites located at the proposed first β-sheet of this domain. The meaning of such arrangement is still unknown, but a probable important regulatory function can be suggested for this region.

The clustering of arginines residues suggests that the C-terminal domain may be responsible for non-specific interaction with DNA or RNA whilst the cold-shock domain mediates the specific interaction with the Y-box element itself. The high charge density of the C-terminal domain may therefore play a role in packaging or in bringing together two distant regions of nucleic acid or in helping to stabilize protein/nucleic

acid or protein/protein interactions. Alternatively, one side of the sandwich may be involved in interactions with the acidic face of the cold-shock domain as described above.

Examination of the sequence for the C-terminal domain shows a paucity of hydrophobic amino acids (Fig. 2a). It is thus extremely difficult to imagine an alternative fold for the domain as it would be almost impossible to form a hydrophobic core. Although the proposed model is only an outline for what we expect to be the three-dimensional structure of the C-terminal domain, it represents the most detailed concrete proposal thus far put forward for any member of the Y-box family. However, due to the marked differences in the pattern of arginines encountered in SMYB1, when compared to the remainder, the model should not be interpreted as being in any sense generic for the family as a whole. The high arginine content of the C-terminal domain would seem to be of general importance for function of these nucleic acid binding proteins but there may be more than one structural solution to the problem. That proposed here may well turn out to be valid only for the specific case of SMYB1. In addition, other RNA binding proteins such as the nucleolins also present C-terminal domains rich in arginines and glycines (52). However, once again the pattern does not correspond to the fibroin-like structure. The high positive charge density of such domains may be of importance for nucleic acid recognition but it would seem likely that they adopt different three-dimensional structures that may reflect different roles played by these factors.

ACKNOWLEDGMENTS

The authors thank Katia Barroso for automated DNA sequencing, Neuza A. Rodrigues for technical assistance and Emmanuel Dias Neto for supplying mRNA from egg, cercariae and schistosomula.

REFERENCES

1. Benoist, C. and Mathis, D. (1990) *Annu. Rev. Immunol.* **8,** 681-715

2. Dorn, A., Durand, B., Marfling, C., Le meur, M., Benoist, C., and Mathis, D. (1987) *Proc. Natl. Acad. Sci.* U. S. A. **84,** 6249-6253

3. Dorn, A., Bollekens, J., Staub, A., Benoist, C., and Mathis, D. (1987) *Cell* **50,** 863-872

4. Didier, D. K., Schiffenbauer, J., Woulfe, S. L., Zacheis, M., and Schwartz, B. D. (1988) *Proc. Natl. Acad. Sci.* U. S. A. **85,** 7322-7326

5. Spitkousky, D. D., Royer-Pokora, B., Delius, H., Kisseljov, F., Jenkins, N. A., Gilbert, D. J., Copeland, N. G., and Royer, H-D. (1992) *Nucleic Acids Res.* **20,** 797-803

6. Adams, M. D., Kerlavage, A. R., Fleischmann, R. D., Fuldner, R. A., Bult, C. J. et al. (1995) *Nature* **377** (suppl.), 3-174

7. Ito, K., Tsutsumi, K., Kuzumaki, T., Gomez, P. F., Otsu, K., and Ishikawa, K. (1994) *Nucleic Acids Res.* **22,** 2036-2041

8. Ozer, J., Faber, M., Chalkley, R., and Sealy, L. (1990) *J. Biol. Chem.* **265,** 22143-22152

9. Gai, X., Lipson, K. E., and Prystowsky, M. B. (1992) *Nucleic Acids Res.* **20,** 601-606

10. Grant, C. E., and Deeley, R. G. (1993) *Mol. Cell Biol.* **13,** 4186-4196

11. Tafuri, S. R., and Wolffe, A. P. (1990) *Proc. Natl. Acad. Sci.* U. S. A. **87,** 9028-9032

12. Skehel, P. A., and Bartsch, D. (1994) *Gene* **145,** 231-235

13. Wolffe, A. P. (1994) *BioEssays* **16,** 245-251

14. Wolffe, A. P., Tafuri, S., Ranjan, M., and Familari, M. (1992) *New Biologist* **4,** 290-298

15. Tafuri, S. R., and Wolffe, A. P. (1992) *New Biologist* **4,** 349-359

16. Kerr, D., Chang, C-F., Cheng, N., Gallia, G., Raj, G., Schwartz, B., and Khalili, K. (1994) *J. Virol.* **68,** 7637-7643

IFSC-USP SERVIÇO DE BIBLIOTECA E INFORMAÇÃO

17. Evdokimova, V. M., Wei, C. L., Sitikov, A. S., Simonenko, P. N., Lazarev, O. A., Vasilenko, K. S., Ustinov, V. A., Hershey, J. W., and Ovchinnikov, L. P. (1995) *J. Biol. Chem.* **270,** 3186-3192

18. Tafuri, S. R., and Wolffe, A. P. (1994) *J. Biol. Chem.* **268,** 24244-24261

19. Ranjan, M., Tafuri, S. R., and Wolffe, A. P. (1993) *Genes Dev.* **7,** 1725-1736

20. Bouvet, P., and Wolffe, A. P. (1994) *Cell* **77,** 931-941

21. Goldstein, J., Pollitt, S., and Inouye, M. (1990) *Proc. Natl. Acad. Sci.* U. S. A. **87,** 283-287

22. Willimsky, G., Bang, H., Fischer, G., and Marahiel, M. A. (1992) *J. Bacteriol.* **174,** 6326-6335

23. Wistow, G. (1990) *Nature* **344,** 823-824

24. Schindelin, H., Jiang, W., Inouye, M., and Heinemann, V. (1994) *Proc. Natl. Acad. Sci.* U. S. A. **91,** 5119-5123

25. Schindelin, H., Marahiel, M. A., and Heinemann, V. (1993) *Nature* **364,** 164-168

26. Newkirk, K., Feng, W., Jiang, W., Tejero, R., Emerson, S. D., Inouye, M., and Montelione, G. T. (1994) *Proc. Natl. Acad. Sci.* U. S. A. **91,** 5114-5118

27. Schnuchel, A., Wiltscheck, R., Czisch, M., Herrier, M., Willimsky, G., Graumann, P., Marahiel, M. A., and Holak, T. A. (1993) *Nature* **364,** 169-171

28- Murzin, A. G. (1993) *EMBO J.* **12,** 861-867

29. Schröder, K., Graumann, P., Schnuchel, A., Holak, T. A., and Marahiel, M. A. (1995) *Mol. Microbiol.* **16,** 699-708

30. World Health Organization: Epidemiology and Control of Schistosomiasis (1985). Technical Report Series 728, Geneva, Switzerland

31. Franco, G. R., Adams, M. D., Soares, M. B., Simpson, A. G. J., Venter, J. C., and Pena, S. D. J. (1995) *Gene* **152,** 141-147

32. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. (1990) *J. Mol. Biol.* **215,** 403-410

33. Tanaka, M., Hirai, H., LoVerde, P. T., Nagafuchi, S., Franco, G. R., Simpson, A. J. G., and Pena, S. D. J. (1995) *Mol. Biochem. Parasitol.* **69,** 41-51

34. Tanaka, M., Tanaka, T., Inazawa, J., Nagafuchi, S., Kitamura, L., Mitsui, Y., Dias Neto, E., Simpson, A. J. G., Kaukas, A., Johnston, D. A. and Rollinson, D. (1995) *Mol. Biochem. Parasitol.* **in press**

35. Blundell, T., Carney, D., Gardner, S., Hayes, F., Howlin, B., Hubbard, T., Overington, J., Singh, D. A., Sibanda, B. L., and Sutcliffe, M. (1988) *Eur. J. Biochem.* **172,** 513-520

36. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Myer, J. E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) *J. Mol. Biol.* **112,** 535-542

37. Vriend, G., and Sander, C. (1993) *J. Appl. Cryst.* **26,** 47-60

38. Barton, G., and Sternberg, M. J. E. (1987) *J. Mol. Biol.* **198,** 327-337

39. Jones, T. A., and Thirup, S. (1986) *EMBO J.* **5,** 819-822

40. Cambillau, C. C., and Horjales, E. (1987) *J. Mol. Graph.* **5,** 174-177

41. Vriend, G., Sander, C., and Stouten, P. F. W. (1994) *Prot. Eng.* **7,** 23-29

42. McGregor, M. J., Islam, S. A., and Sternberg, M. J. E. (1987) *J. Mol. Biol.* **198,** 295-310

43. Laskowski, R. A., MacArthur, M. W., and Thornton, J. M. (1993) *J. Appl. Cryst.* **26,** 283-291

44. Van Gunsteren, W. F., and Berendsen, H. J. C. (1987) *Groningen Molecular Simulation Library Manual*, BIOMOS B. V., Groningen, The Netherlands

45. Vriend, G. (1990) *J. Mol. Graph.* **8,** 52-56

46. Lüthy, R., Bowie, J. U., and Eisenberg, D. (1992) *Nature* **356,** 83-85

47. Marsh, R. E., Corey, R. B., and Pauling, L. (1955) *Biochem. Biophys. Acta.* **16,** 1-34

48. Fossey, S. A., Ne'methy, G., Gibson, K. D., and Scheraga, H. A. (1991) *Biopolymers* **31,** 1529-1541

49. Kosak, M. (1987) *J. Mol. Biol.* **196,** 947-950

50. Landsman, D. (1990) *Nucleic Acids Res.* **20,** 2861-2864

51. Burd, C. G., and Dreyfuss, G. (1994) *Science* **265,** 615-621

52. Heine, M. A., Rankin, M. L., and DiMario, P. J. (1993) *Mol. Biol. Cell* **4,** 1189-1204

53. Dingwall, C., and Laskey, R. A. (1986) *Annu. Rev. Cell Biol.* **2,** 367-390

54. La Teana, A., Brandi, A., Falconi, M., Spurio, R., Pon, C. L., and Gualerzi, C. O. (1991) *Proc. Natl. Acad. Sci.* U. S. A. **88,** 10907-10911

55. Graumann, P., and Marahiel, M. A. (1994) *FEBS Lett.* **338,** 157-160

56. Ladomery, M., and Sommerville, J. (1994) *Nucleic Acids Res.* **22,** 5582-5589

57. Bouvet, P., Matsumoto, K., and Wolffe, A. P. (1995) *J. Biol. Chem.* **270,** 28297-28303

58. Murray, M. T. (1994) *Biochemistry* **33,** 13910-13917

59. Marello, K., La Rovere, J., and Sommerville, J. (1992) *Nucleic Acids Res.* **20,** 5593-5600

FOOTNOTES

* This work was supported by FAPEMIG, PADCT and TDR-WHO. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The nucleotide sequences reported in this paper have been submitted to the GenBank™/EMBL Data Bank with the accession numbers U39883 and U42205

[1] The abbreviations used are: BLAST, Basic Local Alignment Search Tool, CISS, Chromosomal *in situ*-Suppression Hybridization; CK-II, casein kinase II; CSD, cold-shock domain; EST, Expressed Sequence Tag; MHC, major histocompatibility complex; ORF, open reading frame; PCR, polymerase chain reaction, PFGE, pulsed-field gel electrophoresis; PKC, protein kinase C; RT-PCR, reverse transcribed PCR, SMYB1, *Schistosoma mansoni* Y-box binding protein 1; YB-1, Y-box binding protein 1

FIGURE LEGENDS

FIG. 1. **Schematic representation of the *SMYB1* gene.** Diagram of the *SMYB1* gene. The longest ORF, the cold-shock domain, the polyA tail of 21 nucleotides and the enzyme restriction sites used in the subcloning are indicated, as well as the positions and sizes of the three introns. The enzyme restriction sites *Hin*dIII, *Not*I and *Eco*RI are from the vector cloning site.

FIG. 2. **Genomic and cDNA nucleotide sequence analysis of the *SMYB1* gene.** *Panel A*, nucleotide consensus sequence of the *SMYB1* cDNA and the deduced amino acid sequence of the longest ORF. The 5' sequence generated by PCR is in italics and the potential sites for phosphorylation by PKC [ST]-X-[RK], CK-II [ST]-X(2)-[DE] and cAMP-dependent kinase [RK]-X(2)-[ST] are in bold. The cold-shock domain is underlined. *Panel B*, nucleotide consensus sequence of the *SMYB1* gene fragment generated by PCR. Introns are boxed and the cold-shock domain is in italics.

FIG. 3. **Amino acid sequence comparisons of the predicted SMYB1 protein with Y-box binding proteins and cold-shock proteins.** *Panel A*, the deduced amino acid sequence of SMYB1 is compared to other members of the Y-box binding protein family using the program MULTALIGN of the AMPS package. Residues identical to SMYB1 are denoted by dots and gaps by hyphens. RNP1 and RNP2 RNA binding motifs are underlined. YB1_HUMAN (GenBank Accession No J03827), CBFX_HUMAN (Swiss Prot Accession No P16990), MOUSE_MSY1 (GenBank Accession No M62867), MOUSE_YB (GenBank Accession No M60419), CHKCSYB1 (GenBank Accession No L13032), YB1_XENLA (GenBank Accession No M38382), YB2_XENLA (GenBank Accession No M38383), and APLYSIA_YB (GenBank Accession No U02684). *Panel B*, the deduced amino acid sequence of the cold-shock domain of SMYB1 is compared with the cold-shock proteins CspA of *E. coli* (GenBank Accession No M30139) and CspB of *B. subtilis* (GenBank Accession No X59715) using the program MULTALIGN of the AMPS package. Identical residues are indicated by dots and gaps by hyphens.

FIG. 4. **Expression of *SMYB1* at different stages of the *S. mansoni* life cycle.** mRNA was isolated from the distinct *S. mansoni* stages, reverse transcribed using an Oligo (dT) primer and PCR amplified using YboxM and YboxR primers. The amplification products were submitted to electrophoresis on a 6% polyacrilamide gel and subsequently silver stained. The size in base pairs of the amplified fragment is indicated on the left. 1- egg; 2- miracidia, 3- cercariae, 4- 3-hour schistosomula, 5- 7-day schistosomula, 6- adult female, 7- adult male, 8- pool of adult male and female, 9- blank.

FIG. 5. **Schematic representation of the molecular structure of SMYB1 cold-shock domain.** Ribbon drawing showing the five antiparallel β-strands, forming a β-barrel and the location of hydrophobic and basic residues within the putative nucleic acid binding site on the surface of one β sheet (blue).

FIG. 6. **Structure of the SMYB1 cold-shock domain dimer.** *A*, Dimer formation occurs due to interactions between β-strand 4 of two molecules forming an antiparallel β-ladder. The structure is stabilized by hydrogen bonds and two possible salt bridges. *B*, Dimer interface showing the hydrogen bonds (dashed lines) and the two salt bridges.

FIG. 7. **Schematic representation of the SMYB1 fibroin-like domain.** *A*, proposed secondary structure for part of the SMYB1 C-terminal domain showing six antiparallel β-strands composed almost exclusively of Arg-Gly repeats. *B*, schematic drawing of the hypothetical β-sandwich formed by the interaction of two β-sheets, such as in the silk II fibroin structure. The glycine side-chains are directed to the interface of the sandwich and the arginines side-chains to the outside.

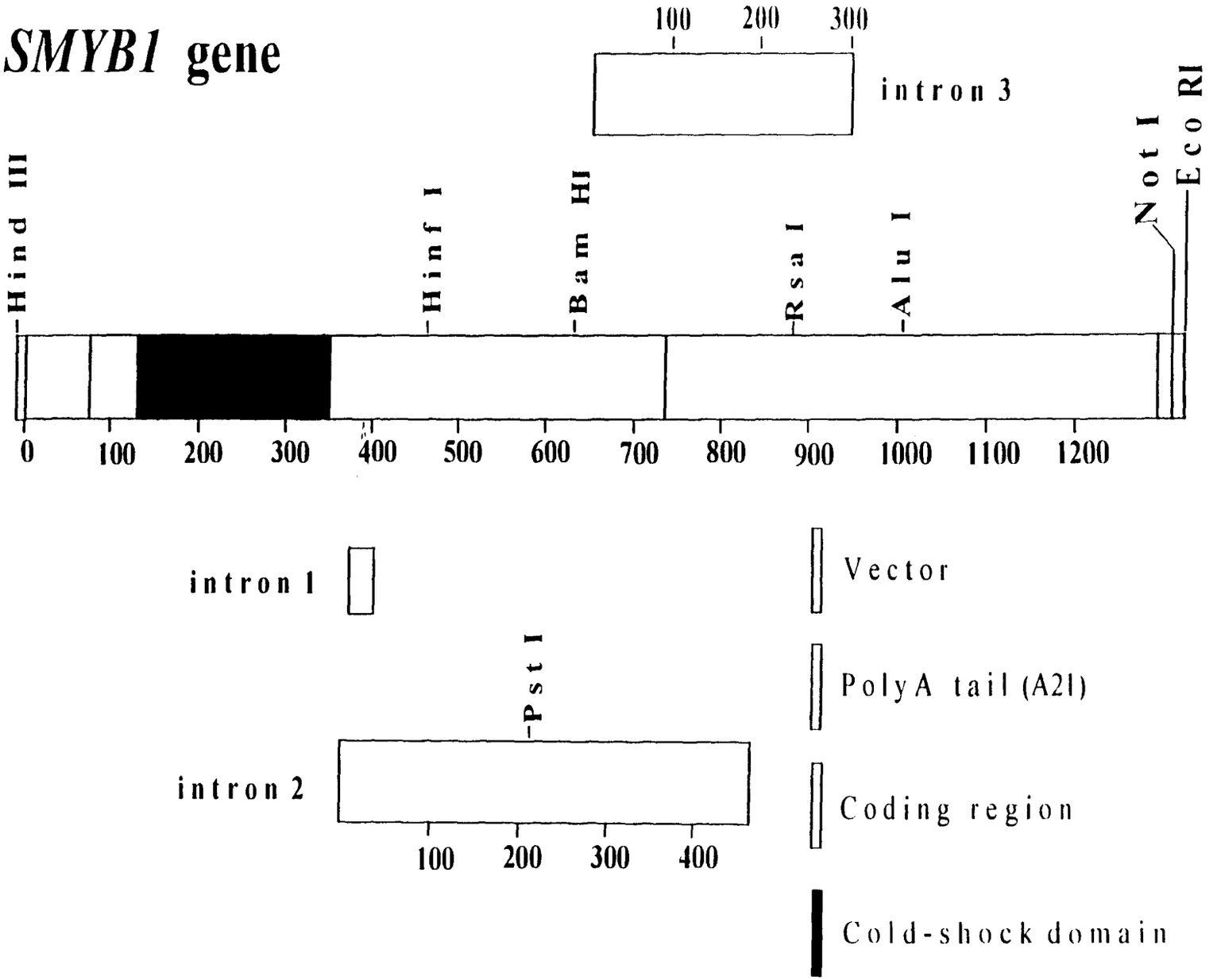

SMYB1 gene
100
200
300
intron 3
Hind III
Hinf I
Bam HI
Rsa I
Alu I
Not I
Eco RI
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
intron 1
Pst I
intron 2
100
200
300
400
Vector
PolyA tail (A21)
Coding region
Cold-shock domain

FIGURE 1

FIGURE 2

A.

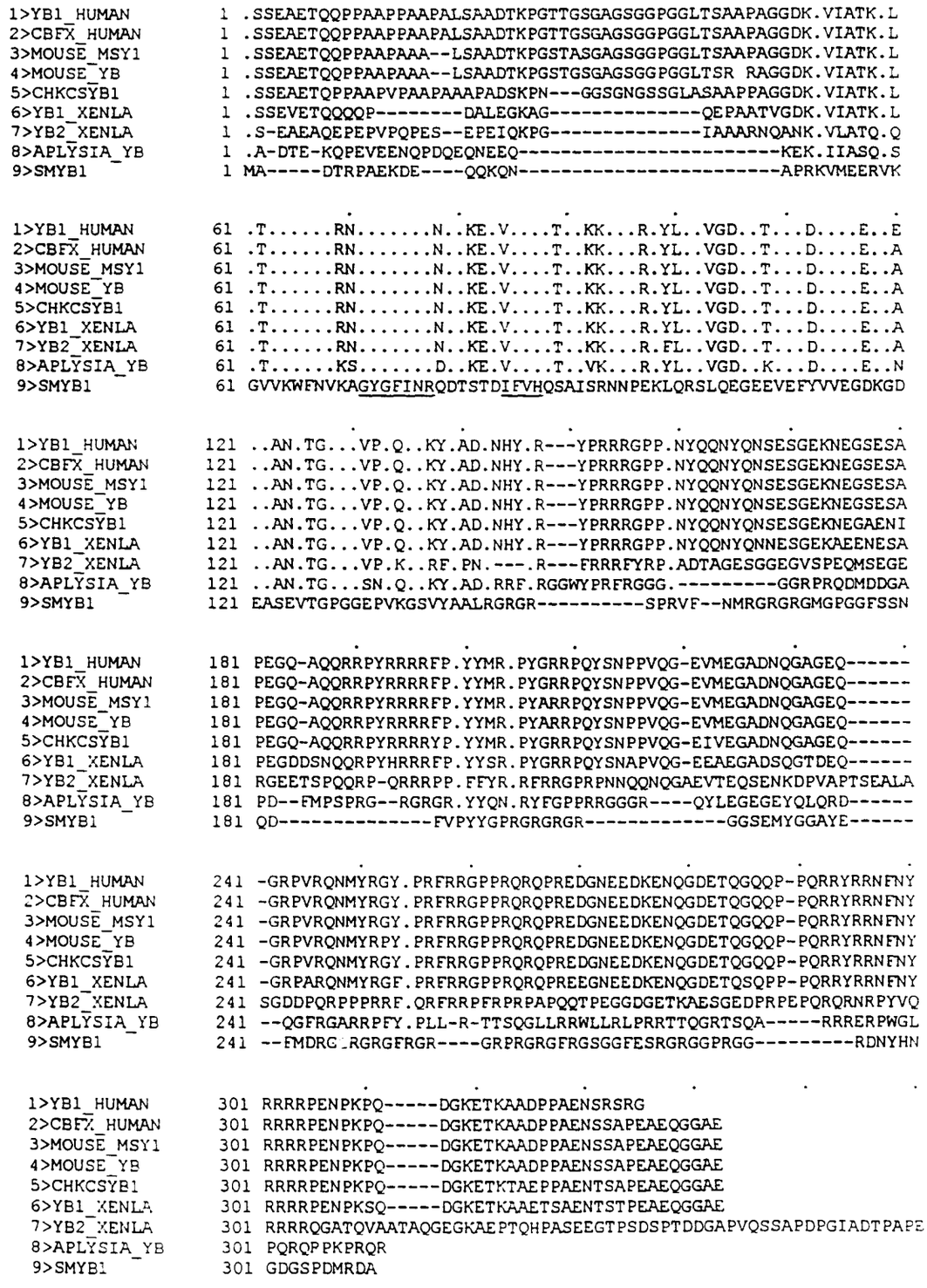
```
1    AT AGT TTT TGT CGT GGT CGT TTG TTC TAA AGG CGT GCC CAG TTG TTT GAT   50

51   TAC CAT CAG AGT GTG GGC TAA CAC ACA ATG GCG GAC ACT AGA CCA GCT GAA  101
1                                        M   A   D   T   R   P   A   E     8

102  AAA GAT GAA CAG CAG AAA CAA AAC GCG CCA CGC AAG GTG ATG GAA GAG CGT  152
9    K   D   E   Q   Q   K   Q   N   A   P   R   K   V   M   E   E   R    25

153  GTC AAA GGC GTG GTT AAG TGG TTC AAT GTG AAG GCT GGA TAT GGC TTC ATC  203
26   V   K   G   V   V   K   W   F   N   V   K   A   G   Y   G   F   I    42

204  AAC CGT CAG GAC ACA TCC ACG GAC ATA TTT GTA CAC CAG TCA GCA ATA TCA  254
43   N   R   Q   D   T   S   T   D   I   F   V   H   Q   S   A   I   S    59

255  CGT AAC AAT CCA GAG AAG CTC CAA CGC TCA CTT CAG GAG GGA GAA GAG GTG  305
60   R   N   N   P   E   K   L   Q   R   S   L   Q   E   G   E   E   V    76

306  GAG TTT TAT GTT GTT GAA GGA GAC AAA GGT GAC GAA GCA TCC GAA GTG ACA  356
77   E   F   Y   V   V   E   G   D   K   G   D   E   A   S   E   V   T    93

357  GGC CCT GGG GGT GAG CCT GTT AAG GGC AGT GTG TAC GCA GCT TTG CGT GGA  407
94   G   P   G   G   E   P   V   K   G   S   V   Y   A   A   L   R   G   110

408  CGT GGA CGC AGC CCG CGG GTA TTT AAC ATG CGT GGG AGA GGC CGC GGA ATG  458
111  R   G   R   S   P   R   V   F   N   M   R   G   R   G   R   G   M   127

459  GGA CCT GGC GGA TTC TCC AGC AAT CAA GAT TTC GTG CCT TAC TAC GGT CCT  509
128  G   P   G   G   F   S   S   N   Q   D   F   V   P   Y   Y   G   P   144

510  CGA GGT CGA GGC CGC GGC CGT GGG GGC TCA GAA ATG TAC GGA GGT GCT TAC  560
145  R   G   R   G   R   G   R   G   G   S   E   M   Y   G   G   A   Y   161

561  GAA TTT ATG GAT CGC GGT GGG AGA GGC CGT GGG TTC CGT GGA CGT GGA AGG  611
162  E   F   M   D   R   G   G   R   G   R   G   F   R   G   R   G   R   178

612  CCT CGT GGT CGT GGG TTT AGA GGA TCC GGT GGG TTC GAG TCT CGA GGT CGT  662
179  P   R   G   R   G   F   R   G   S   G   G   F   E   S   R   G   R   195

663  GGT GGC CCT CGT GGG GGA AGG GAT AAT TAT CAT AAC GGG GAT GGC TCA CCG  713
196  G   G   P   R   G   G   R   D   N   Y   H   N   G   D   G   S   P   212

714  GAT ATG CGA GAC GCT TAA AAT TCT CTG ATC AAA TAC TGA GAA ACG TCA TTT  764
213  D   M   R   D   A   *                                               217

765  TAG AAA TCT GTG TAT TCT GTC AGC ATC TAT ATA TGC ATG TCT ACG TTT CAA  815

816  CTC AAT TGT TTA ATC ATC CAA CTT GCA GCA CAC TGG ATT GTC ACT TGT CTT  866

867  TAT GGT ATT TCG GAT GTA CAC CAT CTT TTT CCC TTC CTC AGT TTC GAC GTC  917

918  TAA CTG GTC AAA CTT GTG TCT CAA ATT TTT GAA TCC AGA GTT CCT GTG GGA  968

969  ATG TTC AAA ATG TTT CAT TCG ATC TTA CAC CCT TAG CTT TCT GCG CTC TTC 1019

1020 GAT GCA TTT CTT GCC TAA ATG GGC TTT CCA TCT GAA CTT GAT TAC TTT CCT 1070

1071 ACC GTT TCA TAT CTA CTT ACT TCC GTC TGC CAA TCG CTT GAA GGT TCA GAC 1121

1122 CAT TTG AAA AAG TCA GAG ATC TGA CTC GTC AGC CAT TGG TGA AAC ACT TTC 1172

1173 CTG TAG AAT GTG CAT TCT TCT GTC ATC TTG TGT GGA TAT CTG GTG CTG TAT 1223

1224 GGT TTG ACT TTT TAC GTA ATC AAA TAT GCA TTT TGA TAC TTT ATA AAA TTT 1274

1275 CTT TAG AGT                                                        1283
```

B.

```
   1 CACAATGGCG GACACTAGAC CAGCTGAAAA AGATGAACAG CAGAAACAAA ACGCGCCACG
  61 CAAGGTGATG GAAGAGCGTG TCAAAGGCGT GGTTAAGTGG TTCAATGTGA AGGCTGGATA
 121 TGGCTTCATC AACCGTCAGG ACACATCCAC GGACATATTT GTACACCAGT CAGCAATATC
 181 ACGTAACAAT CCAGAGAAGC TCCAACGCTC ACTTCAGGAG GGAGAAGAGG TGGAGTTTTA
 241 TGTTGTTGAA GGAGACAAAG GTGACGAAGC ATCCGAAGTG ACAGGCCCTG GGGGTGAGCC
 301 TGTTAAGGTA AACTGAGTTG AGTGTTTATT CTCGTTTAGG GCAGTGTGTA CGCAGCTTTG
 361 CGTGGACGTG GACGCAGCCC GCGGGTATTT AACATGCGTG GGAGAGGCCG CGGAATGGGA
 421 CCTGGCGGAT TCTCCAGCAA TCAAGATTTC GTGCCTTACT ACGGTCCTCG AGGTCGAGGC
 481 CGCGGCCGTG GGGGCTCAGA AATGGTAAGT TGACGTAAAC TAGTATTCTG CTTATAGTGA
 541 GTCCAGAAGG ACACTGGTCG TCAATTTACA AACACAATAT ACGTAATCAC CGTTTTATTT
 601 TTGCCCATTG TGATGACGTT AAATCCAAAT CATCAGCCCA CATTTTCAAC ACAGTTTATC
 661 ATAGTCTTCG TCATATCGAT CACAACCATT GATGATAACC CATTGCAGGA CGCACCTGCA
 721 GTTCCGTTGA GTAAGTGTTG GGATTTCTTC CCCAGTGTAA ATCTTCAGTC CAAGTTTAAA
 781 GCTCAGTTAT GGTGCTTAAT TGCAAGTCCT GGCACCAGTC GTTCAAAGAC GCTCCTAATC
 841 CTACAGATGG ATTGATTAAA TATTGTTTAA AATCCATAGC CACGTTAAGA TTTTTATAGC
 901 TTCATAGGAA TATTGCATAA CATTTTGTTG GTTCTCGTTT CGATTAATTA TTACACTTTT
 961 TCGGCTTCAG TACGGAGGTG CTTACGAATT TATGGATCGC GGTGGGAGAG GCCGTGGGTT
1021 CCGTGGACGT GGAAGGCCTC GTGGTCGTGG GTTTAGAGGA TCCGGTGGGT TCGAGTCTCG
1081 AGGTCGTGGT GGCCCTCGTG GGGGGAGGGA TAATTATCAT AACGGGGATG GTAAGTTGCC
1141 AACCTGTTTC GATATATTTG CCTGTTATTA TCCTGAAGCG CATGCATTCC CTTTCTGAGC
1201 CGATTCGGAA CGATCATACG TTCCACAGCA TAATGAGCTA AAAACTTAAA GTCAAAAGTA
1261 TTGTGATTCC AAATACATTC TCTAGATGCC AAGACCAGCA CTAATCGGAT GTATGGCAAT
1321 GTGGTGGTTT TGTTTCTGAT ATTGCTAACG AGTATTTGCC AGGCAATGTT ACTTGTGGTT
1381 TTGCACTGAT GCTTCATTTT TGGGTATGCC TCTTAAATTT TTTTTTAGGC TCACCGGATA
1441 TGCGAGACGC TTAAAATTCT CTGATCAAAT ACTGAGAAAC GTCATTTTAG AAATCTGTGT
1501 ATTCTGTCAG CATCTATATA TGCATGTCTA CGTTTCAACT CAATTGTTTA ATCATCCAAC
1561 TTGCAGCACA CTGGATTGTC ACTTGTCTTT ATGGTATTTC GGATGTACAC CATCTTTTTC
1621 CCTTCCTCAG TTTCGACGTC TAACTGGTCA AACTTGTGTC TCAAATTTTT GAATCCAGAG
1681 TTCCTGTGGG AATGTTCAAA ATGTTTCATT CGATCTTACA GCCTTAGCTT TCTGCGCTCT
1741 TCGATGCATT TCTTGCCTAA ATGGGCTTTC CATCTGAACT TGATTACTTT CCTACCGTTT
1801 CATATCTACT TACTTCCGTC TGCCAATCGC TTGAAGGTTC AGACCATTTG AAAAAGTCAG
1861 AGATCTGACT CGTCAGCCAT TGGTGAAACA CTTTCCTGTA GAATGTGCAT TCTTCTGTCA
1921 TCTTGTGTGG ATATCTGGTG CTGTATGGTT TGACTTTTTA CGTAATCAAA TATGCATTTT
```

FIGURE 3

A.


1>YB1_HUMAN 1 .SSEAETQQPPAAPPAAPALSAADTKPGTTGSGAGSGGPGGLTSAAPAGGDK.VIATK.L
2>CBFX_HUMAN 1 .SSEAETQQPPAAPPAAPALSAADTKPGTTGSGAGSGGPGGLTSAAPAGGDK.VIATK.L
3>MOUSE_MSY1 1 .SSEAETQQPPAAPAAA--LSAADTKPGSTASGAGSGGPGGLTSAAPAGGDK.VIATK.L
4>MOUSE_YB 1 .SSEAETQQPPAAPAAA--LSAADTKPGSTGSGAGSGGPGGLTSR RAGGDK.VIATK.L
5>CHKCSYB1 1 .SSEAETQPPAAPVPAAPAAAPADSKPN---GGSGNGSSGLASAAPPAGGDK.VIATK.L
6>YB1_XENLA 1 .SSEVETQQQQP--------DALEGKAG----------QEPAATVGDK.VIATK.L
7>YB2_XENLA 1 .S-EAEAQEPEPVPQPES--EPEIQKPG----------IAAARNQANK.VLATQ.Q
8>APLYSIA_YB 1 .A-DTE-KQPEVEENQPDQEQNEEQ----------KEK.IIASQ.S
9>SMYB1 1 MA-----DTRPAEKDE----QQKQN----------APRKVMEERVK
1>YB1_HUMAN 61 .T......RN.......N..KE.V....T..KK...R.YL..VGD..T...D....E..E
2>CBFX_HUMAN 61 .T......RN.......N..KE.V....T..KK...R.YL..VGD..T...D....E..A
3>MOUSE_MSY1 61 .T......RN.......N..KE.V....T..KK...R.YL..VGD..T...D....E..A
4>MOUSE_YB 61 .T......RN.......N..KE.V....T..KK...R.YL..VGD..T...D....E..A
5>CHKCSYB1 61 .T......RN.......N..KE.V....T..KK...R.YL..VGD..T...D....E..A
6>YB1_XENLA 61 .T......RN.......N..KE.V....T..KK...R.YL..VGD..T...D....E..A
7>YB2_XENLA 61 .T......RN.......N..KE.V....T..KK...R.FL..VGD..T...D....E..A
8>APLYSIA_YB 61 .T......KS.......D..KE.V....T..VK...R.YL..VGD..K...D....E..N
9>SMYB1 61 GVVKWFNVKAGYGFINRQDTSTDIFVHQSAISRNNPEKLQRSLQEGEEVEFYVVEGDKGD
1>YB1_HUMAN 121 ..AN.TG...VP.Q..KY.AD.NHY.R---YPRRRGPP.NYQQNYQNSESGEKNEGSESA
2>CBFX_HUMAN 121 ..AN.TG...VP.Q..KY.AD.NHY.R---YPRRRGPP.NYQQNYQNSESGEKNEGSESA
3>MOUSE_MSY1 121 ..AN.TG...VP.Q..KY.AD.NHY.R---YPRRRGPP.NYQQNYQNSESGEKNEGSESA
4>MOUSE_YB 121 ..AN.TG...VP.Q..KY.AD.NHY.R---YPRRRGPP.NYQQNYQNSESGEKNEGSESA
5>CHKCSYB1 121 ..AN.TG...VP.Q..KY.AD.NHY.R---YPRRRGPP.NYQQNYQNSESGEKNEGAENI
6>YB1_XENLA 121 ..AN.TG...VP.Q..KY.AD.NHY.R---YPRRRGPP.NYQQNYQNNESGEKAEENESA
7>YB2_XENLA 121 ..AN.TG...VP.K..RF.PN.---.R---FRRRFYRP.ADTAGESGGEGVSPEQMSEGE
8>APLYSIA_YB 121 ..AN.TG...SN.Q..KY.AD.RRF.RGGWYPRFRGGG.---------GGRPRQDMDDGA
9>SMYB1 121 EASEVTGPGGEPVKGSVYAALRGRGR----------SPRVF--NMRGRGRGMGPGGFSSN
1>YB1_HUMAN 181 PEGQ-AQQRRPYRRRRFP.YYMR.PYGRRPQYSNPPVQG-EVMEGADNQGAGEQ------
2>CBFX_HUMAN 181 PEGQ-AQQRRPYRRRRFP.YYMR.PYGRRPQYSNPPVQG-EVMEGADNQGAGEQ------
3>MOUSE_MSY1 181 PEGQ-AQQRRPYRRRRFP.YYMR.PYARRPQYSNPPVQG-EVMEGADNQGAGEQ------
4>MOUSE_YB 181 PEGQ-AQQRRPYRRRRFP.YYMR.PYARRPQYSNPPVQG-EVMEGADNQGAGEQ------
5>CHKCSYB1 181 PEGQ-AQQRRPYRRRRYP.YYMR.PYGRRPQYSNPPVQG-EIVEGADNQGAGEQ------
6>YB1_XENLA 181 PEGDDSNQQRPYHRRRFP.YYSR.PYGRRPQYSNAPVQG-EEAEGADSQGTDEQ------
7>YB2_XENLA 181 RGEETSPQQRP-QRRRPP.FFYR.RFRRGPRPNNQQNQGAEVTEQSENKDPVAPTSEALA
8>APLYSIA_YB 181 PD--FMPSPRG--RGRGR.YYQN.RYFGPPRRGGGR----QYLEGEGEYQLQRD------
9>SMYB1 181 QD----------FVPYYGPRGRGRGR----------GGSEMYGGAYE------
1>YB1_HUMAN 241 -GRPVRQNMYRGY.PRFRRGPPRQRQPREDGNEEDKENQGDETQGQQP-PQRRYRRNFNY
2>CBFX_HUMAN 241 -GRPVRQNMYRGY.PRFRRGPPRQRQPREDGNEEDKENQGDETQGQQP-PQRRYRRNFNY
3>MOUSE_MSY1 241 -GRPVRQNMYRGY.PRFRRGPPRQRQPREDGNEEDKENQGDETQGQQP-PQRRYRRNFNY
4>MOUSE_YB 241 -GRPVRQNMYRPY.PRFRRGPPRQRQPREDGNEEDKENQGDETQGQQP-PQRRYRRNFNY
5>CHKCSYB1 241 -GRPVRQNMYRGY.PRFRRGPPRQRQPREDGNEEDKENQGDETQGQQP-PQRRYRRNFNY
6>YB1_XENLA 241 -GRPARQNMYRGF.PRFRRGPPRQRQPREEGNEEDKENQGDETQSQPP-PQRRYRRNFNY
7>YB2_XENLA 241 SGDDPQRPPPRRF.QRFRRPFRPRPAPQQTPEGGDGETKAESGEDPRPEPQRQRNRPYVQ
8>APLYSIA_YB 241 --QGFRGARRPFY.PLL-R-TTSQGLLRRWLLRLPRRTTQGRTSQA-----RRRERPWGL
9>SMYB1 241 --FMDRG_RGRGFRGR----GRPRGRGFRGSGGFESRGRGGPRGG---------RDNYHN
1>YB1_HUMAN 301 RRRRPENPKPQ-----DGKETKAADPPAENSRSRG
2>CBFX_HUMAN 301 RRRRPENPKPQ-----DGKETKAADPPAENSSAPEAEQGGAE
3>MOUSE_MSY1 301 RRRRPENPKPQ-----DGKETKAADPPAENSSAPEAEQGGAE
4>MOUSE_YB 301 RRRRPENPKPQ-----DGKETKAADPPAENSSAPEAEQGGAE
5>CHKCSYB1 301 RRRRPENPKPQ-----DGKETKTAEPPAENTSAPEAEQGGAE
6>YB1_XENLA 301 RRRRPENPKSQ-----DGKETKAAETSAENTSTPEAEQGGAE
7>YB2_XENLA 301 RRRRQGATQVAATAQGEGKAEPTQHPASEEGTPSDSPTDDGAPVQSSAPDPGIADTPAPE
8>APLYSIA_YB 301 PQRQPPKPRQR
9>SMYB1 301 GDGSPDMRDA

B.

```
1  >CSPA      1                          SGKMT.I.....ADK.....TPDDGSK.
2  >CSPB      1                            MLE.K.....SEK.....EV-EGQD.
3  >SMYB1     1 MADTRPAEKDEQQKQNAPRKVMEERVKGVVKWFNVKAGYGFINRQDTSTD

1  >CSPA     51 V...F...QNDG----YKS.D..QK.S.TIES.AK.PA.GN..SL
2  >CSPB     51 V...F...QGEG----FKT.E..QA.S.EIVE.NR.PQ.AN..KEA
3  >SMYB1    51 IFVHQSAISRNNPEKLQRSLQEGEEVEFYVVEGDKGDEASEVTGPG
```

FIGURE 4

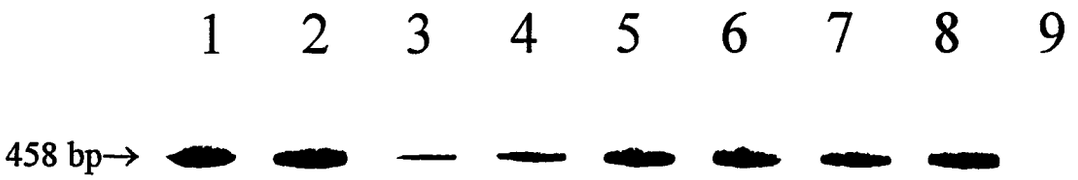


FIGURE 5

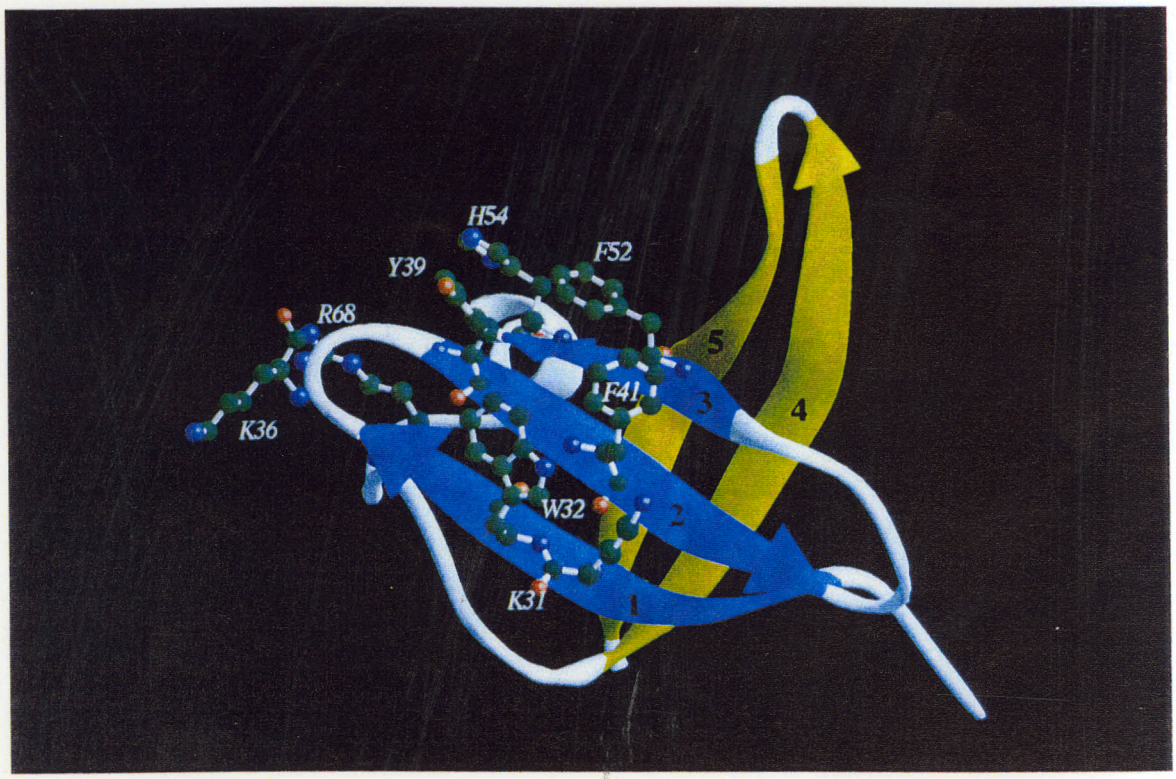
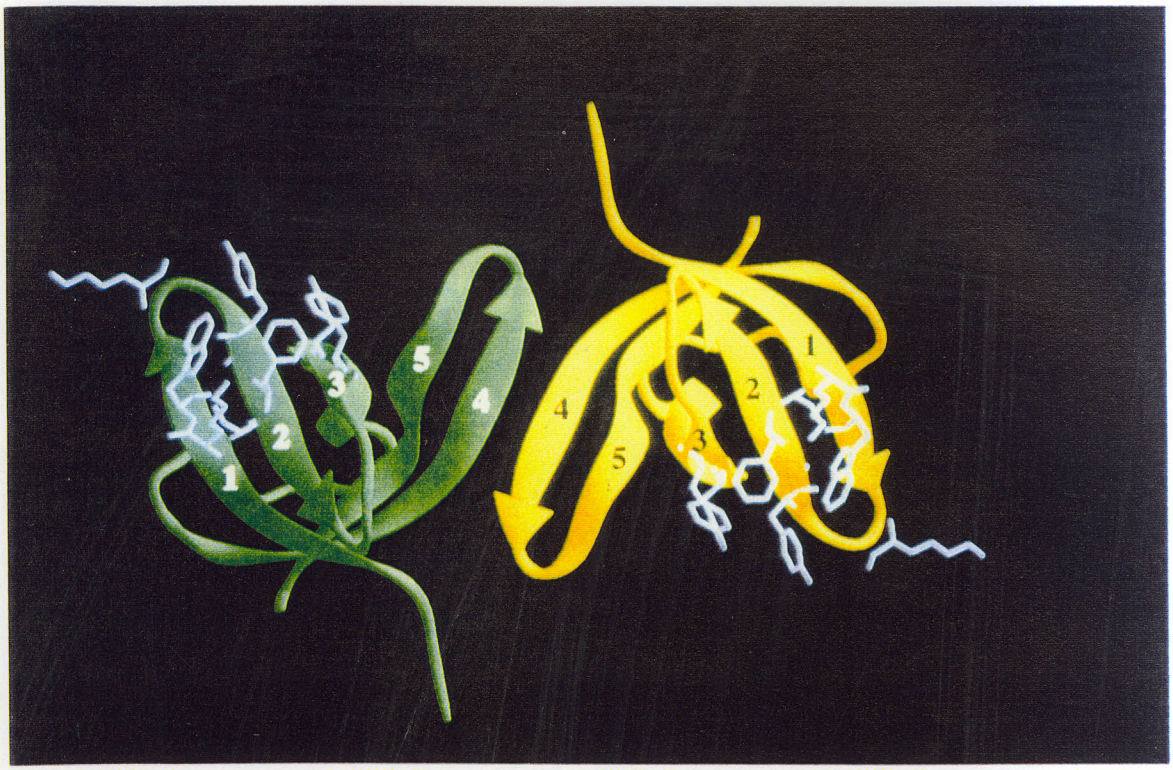


FIGURE 6

A



B

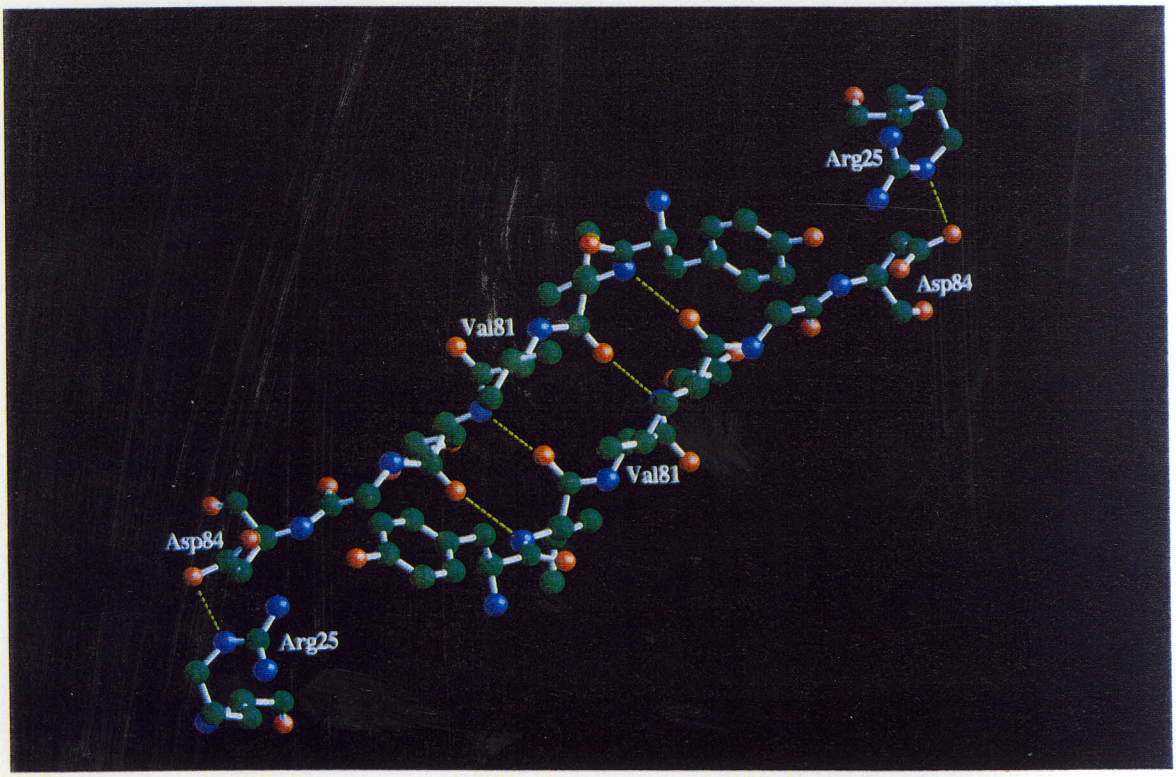
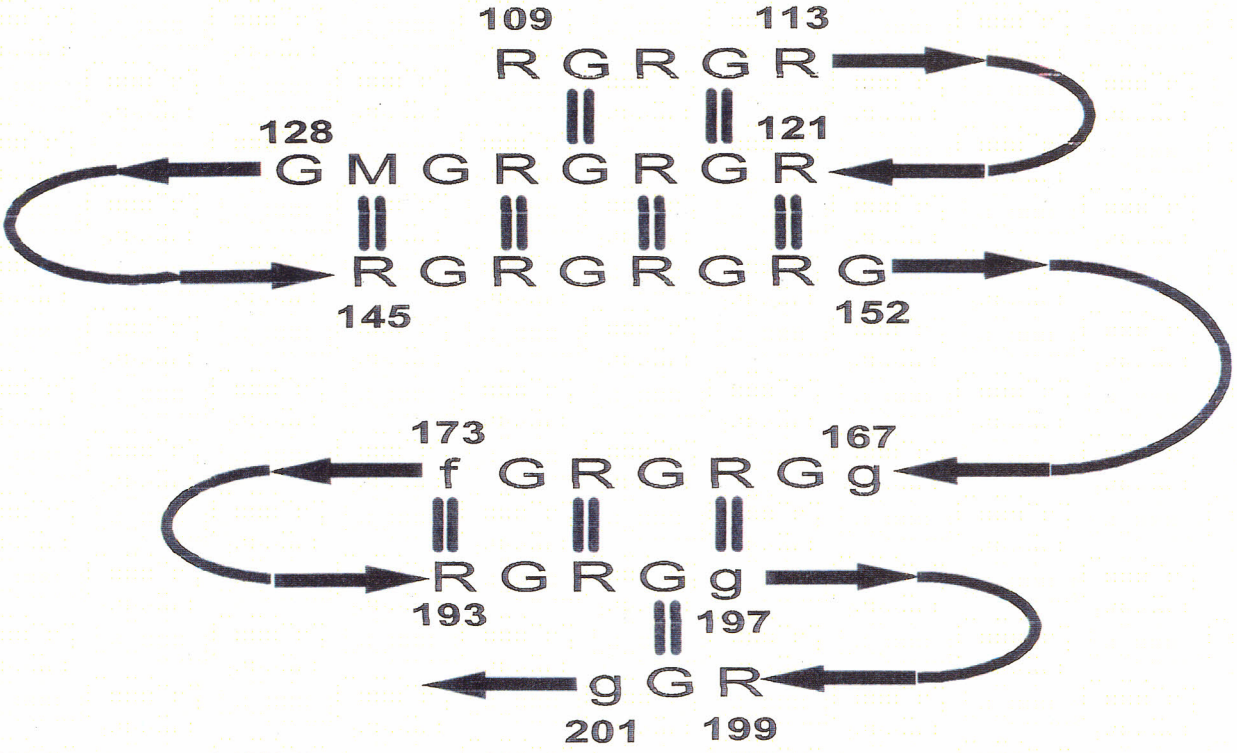
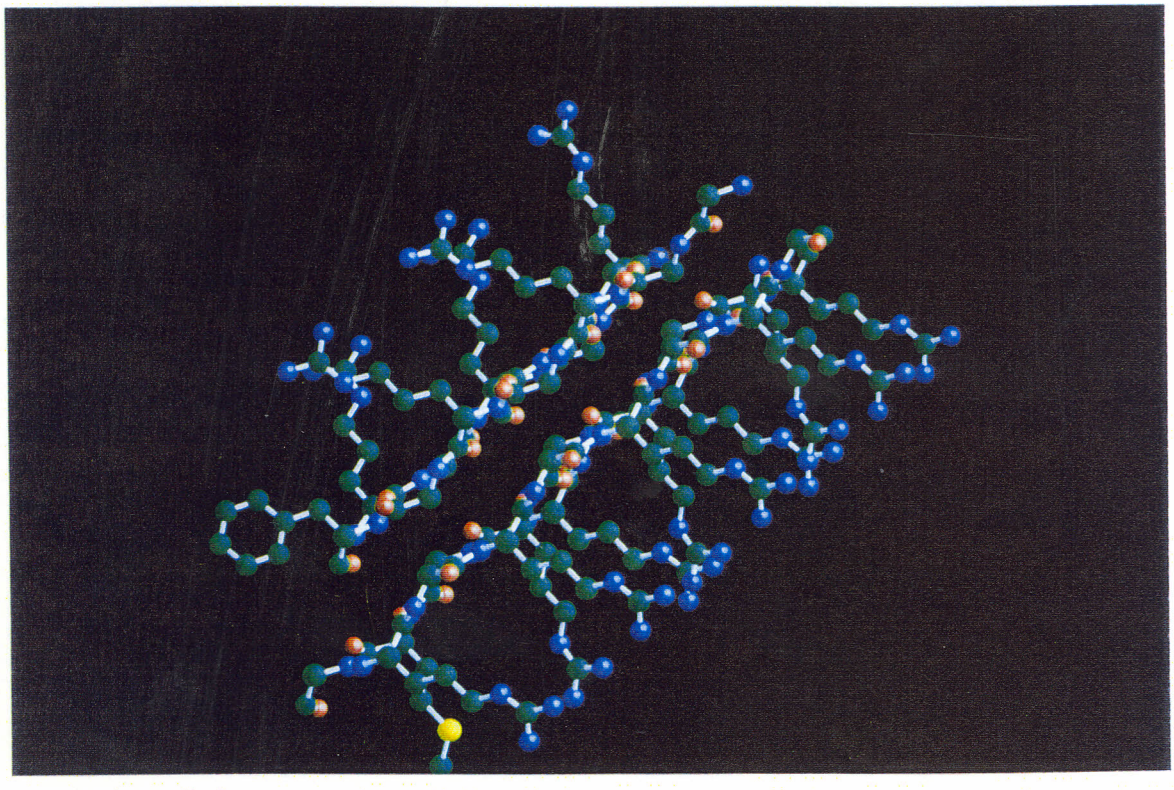


FIGURE 7

A


109
113
R G R G R
128
121
G M G R G R G R
R G R G R G R G
145
152
173
167
f G R G R G g
R G R G g
193
197
g G R
201
199

B



Capítulo 4

TOXINAS EPIDERMOLÍTICAS E ENDOPEPTIDASES ESPECÍFICAS PARA GLUTAMATO

Em 1990 (simultaneamente com Bailey & Smith [1]) publicamos um trabalho desecrevendo as toxinas epidermolíticas como membros da família de serino-proteases tipo tripsina (Dancer *et al.* em anexo). As toxinas epidermolíticas A e B (ETa e ETb respectivamente), que são secretadas por certas linhagens da bactéria *Staphylococcus aureus,* são responsáveis pela síndrome da pele `queimada'que tipicamente afeta recém-nascidos, frequentemente ainda na maternidade onde é transmitido pelas enfermeiras[2]. Histologicamente as toxinas proporcionam uma clivagem na epiderme ao nível do *stratum granulosum* gerando regiões avermelhadas que parecem uma queimadura de primeiro grau[3,4]. A ação das toxinas parece ser bastante específica pois passam pela corrente sanguínea sem manifestar efeito em outros tecidos. Nossa contribuição, pela primeira vez deu um indício do mecanísmo da síndrome da pele queimada ao nível molecular.

As serino-proteases mais parecidas com as toxinas em termos de sequência são as endopeptidases específicas para glutamato (GSEs) que, como o seu próprio nome diz, hidrolizam ligações peptídicas do lado C-terminal de resíduos de glutamato em cadeias polipeptídicas. A enzima mais conhecida deste grupo é a protease V8 de *Staphyloccus*

aureus (linhagem V8), também conhecida como Glu-C e que é largamente utilizada em sequenciamento de proteínas devido a sua alta especificidade[5]. A similaridade sequencial entre ETa, ETb e as GSEs (que é de 28% no máximo) despertou um interesse na base estrutural de especificidade destas enzimas.

As serino-proteases da família tripsina representam um dos grupos de enzimas mais estudadas do ponto de vista bioquímico e estrutural[6]. Estruturas cristalográficas de diversas enzimas de mamíferos, bactéria e vírus junto com os seus complexos com inibidores proteíco e não-proteíco e intermediários tipo acil-enzima revelam a existência de quatro sítios na enzima responsáveis pelo reconhecimento do substrato e a sua catálise[7] (Figura 4.1): 1) A **Triade Catalítica,** descrita pela primeria vez por Blow *et al.*[8], serve para aumentar a nucleofilicidade da serina catalítica (195 na sequência de quimotripsina) de tal maneira a facilitar o ataque no carbonila da ligação hidrolizável 2) O **Bolsão da Oxiánion** [9] que proporciona duas ligações de hidrogênio para estabilizar a carga negativa que se desenvolve no oxigênio do estado de transição. 3) O **Bolsão de Especificidade S1** (nomenclatura de [10]) responsável pelo reconhecimento da cadeia lateral na posição P1 do substrato. 4) O **Sítio Estendido de Ligação** (sítio não-específico de ligação) formado pela cadeia principal dos resíduos 215 e 216 e que interage através de ligações de hidrogênio com os resíduos P2 e P3 respectivamente do substrato, formando assim um pequeno trecho de folha-β antiparalela.

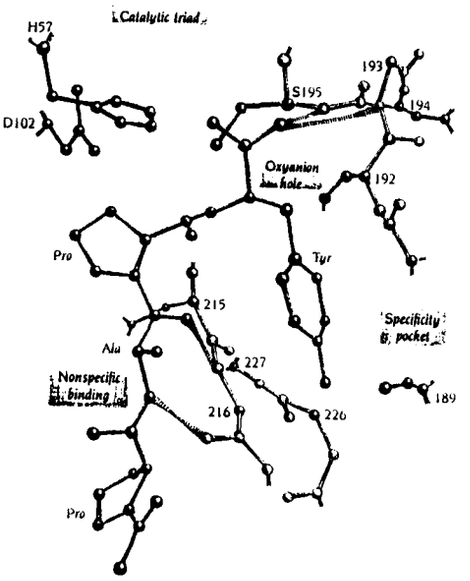


Figura 4.1 Os quatro sítios nas serino-proteases responsáveis pelo reconhecimento do substrato.

No artigo de Barbosa *et al.* (1996) em anexo encontra-se uma breve revisão da literatura das serino-proteases que não convém repetir aqui. Basta destacar a contribuição da cristalografia de proteínas e modelagem molecular em encontrar explicações para a especificidade de determinadas enzimas, baseada nos resíduos que decoram a cavidade

interna do bolsão S1. O exemplo clássico é a própria tripsina cuja especificidade pelos resíduos arginina e lisina na posição P1 se deve à presença do aspartarto 189 no fundo do bolsão[11]. Porém, está ficando cada vez mais evidente que a relação entre estrutura e especificidade é bastante complexa. Experimentos de mutagênese visando a conversão da enzima tripsina para uma especificidade tipo quimotripsina demonstram a importância de 'loops' fora do bolsão S1 e destacam a contribuição dos passos químicos da catálise ao invés de simples ligação na determinação de especificidade[12-14].

4.1 Trabalhos apresentados em seguida

Apresento em anexo três trabalhos (dois já publicados e um no prelo) descrevendo o uso de modelagem molecular no estudo das endopeptidases específicas para glutamato e as toxinas epidermolíticas visando entender a base estrutural da especificidade no primeiro caso e de desvendar algo do mecanismo molecular que leva à síndrome da pele queimada no segundo. O trabalho revelou aspectos inéditos dos sítios ativos e evolução das serino-proteases.

Na época em que iniciamos este trabalho não se conhecia nenhuma estrutura experimentalmente determinada para um membro deste grupo de enzimas. Subsequente à publicação do artigo de Barbosa *et al.* (1993, em anexo) que descreve um modelo para a endopeptidase específica para glutamato de *Streptomyces griseus,* a estrutura cristalográfica foi determinada por Nienaber *et al.* [15]. Uma breve discussão da comparação entre as estruturas encontra-se nas conclusões desta tese.

Dancer, S.J., Garratt, R., Saldanha, J., Jhoti, H. & Evans, R. (1990) 'The Epidermolytic Toxins are Serine Proteases' FEBS Letts **268,** 129-132

Barbosa, J.A.R.G., Garratt, R.C., & Saldanha, J.W. (1993) 'A Structural Model for the Glutamate-Specific Endopeptidase from *Streptomyces griseus* that Explains Substrate Specificity', FEBS Letts **324**, 45-50

Barbosa, J.A.R.G., Saldanha, J.W. & Garratt, R.C. (1996) 'Novel Features of Serine Protease Active Sites and Specificity Pockets: Sequence Analysis and Modelling Studies of Glutamate Specific Endopeptidases and Epidermolytic Toxins' Prot. Eng. in press

Referências

[1] Bailey, C.J. & Smith, T.P. (1990) Biochem. J. **269**, 535-537
[2] Dancer, S.J., Simmons, N.A., Poston, S.M. & Noble, W.C. (1988) J. Infect. **16**, 87-103
[3] Melish, M.E. & Glasgow, L.A. (1971) J. Paediatr. **78**, 958-967
[4] Elias, P.M. (1974) Arch. Dermatol. **110**, 295-296
[5] Drapeau, G.R. (1978) Can. J. Biochem. **56**, 534-544
[6] Warshel, A., Naray-Szabo, G., Sussman, F. & Hwang, J.-K. (1989) Biochemistry **28**, 3629-3637

[7] Branden, C. & Tooze, J. (1991) in 'Introduction to Protein Structure' Garland Publishing Inc. New York, pp 231-246
[8] Blow, D.M., Birktoft, J.J. & Hartley, B.S. (1969) Nature **221**, 337-340
[9]
[10] Schechter, I & Berger, A. (1967) Biochem. Biophys. Res. Commun. **27**, 157-162
[11] Bode, W & Schwager, P. (1975) J. Mol. Biol. **98,** 693-717
[12] Hedstrom, L., Szilagyi, L. & Rutter, W.J. (1992) Science **255,** 1249-1253
[13] Hedstrom, L. Perona, J.J. & Rutter, W.J. (1994) Biochemistry **33**, 8757-8763
[14] Hedstrom, L., Farr-Jones, S., Kettner, C.A. & Rutter, W.J. (1994) Biochemistry **33**, 8764-8769
[15] Nienaber, V.L., Breddam, K. & Birktoft, J.J. (1993) Biochemistry **32**, 11469-11475

De acordo com as políticas editoriais, estes artigos não podem ser depositados em repositório de acesso aberto. Para acesso aos artigos completos entre em contato com o(a) autor(a) ou com o Serviço de Biblioteca e Informação IFSC - USP (bib@ifsc.usp.br).

DANCER, S.J.; GARRATT, R.; SALDANHA, J.; JHOTI, H.; EVANS, R. The epidermolytic toxins are serine proteases. . **Febs Letters**, Amsterdam, v.268, n.1 , p.129-132, July 1990.

BARBOSA, J. A. R. G.; GARRATT, R. C.; SALDANHA, J. W. Structural model for the glutamate-specific endopeptidase from streptomyces griseus that explains substrate specificity. **Febs Letters**, Amsterdam, v.324, n.1, p.45-50, June 1993.

NOVEL FEATURES OF SERINE PROTEASE ACTIVE SITES AND SPECIFICITY POCKETS: Sequence Analysis and Modelling studies of Glutamate Specific Endopeptidases and Epidermolytic Toxins

Key words: active site/epidermolytic toxin/glutamate specificity/molecular modelling/serine protease

João Alexandre R.G. Barbosa[1,2], José W. Saldanha[3] and Richard C. Garratt[1]

[1] Instituto de Física e [2] Instituto de Química de São Carlos, Universidade de São Paulo, Caixa Postal 369, São Carlos - SP, CEP 13560-970, Brazil (Fax: 55 162 713616, RICHARD@IFQSC.SC.USP.BR)

[3] MRC Collaborative Centre, 1-3 Burtonhole Lane, Mill Hill, London NW7 1AD, UK

Running Title: Protease active sites and specificity pockets

Abstract

With a view to obtaining a better understanding of the structural determinants of P1 glutamate specificity in glutamate specific endopeptidases (GSE's), the active sites and specificity pockets of such enzymes from *Bacillus licheniformis* (gse-bl), *Bacillus subtilis* (mpr) and *Staphylococcus aureus* (v8 protease) have been modelled. This approach was extended to the epidermolytic toxins (ET's), responsible for the Staphylococcal Scalded Skin Syndrome. We identify a canonical structure for the S1 subsite, composed of H213 and T190 both of which we predict to interact directly with the P1 glutamate. The possible importance of R30 (for gse-bl and mpr) and of the N-terminus (for gse-bl, mpr and v8 protease) was also examined. In the case of mpr, a G193C substitution is predicted to participate in a novel disulphide bridge which stabilizes C193 in such a way as to maintain the oxyanion hole. In v8, the loss or substitution of several important structural components around D102 of the catalytic triad probably explain its reduced catalytic efficiency in comparison with other GSE's. In the case of the epidermolytic toxins K216 may be important for the previously reported phospholipase C-like activity, since the model predicts that it may stabilize the negative charge on the phosphonyl group.

Introduction

The trypsin-like serine proteases represent one of the most completely studied families of enzymes from both a biochemical and structural point of view. Three-dimensional structures at atomic resolution are available for enzymes of diverse function from mammalian, bacterial and viral sources and have also been determined for mutants, enzyme-inhibitor complexes, acyl-enzyme intermediates and for the inactive zymogens. Serine proteases also represent one of the few examples in the protein field of the application of kinetic crystallography by Laue diffraction (Singer *et al.*, 1993a, 1993b; Perona *et al.*, 1993), crystallographic cryoenzymology (Ding *et al.*, 1994) and neutron diffraction (Kossiakoff & Spencer, 1981).

This wealth of structural information together with that from spectroscopic, kinetic and other biochemical investigations has led to a general consensus concerning the overall reaction mechanism. Originally identified by Blow *et al.* (1969), the central catalytic machinery of the serine proteases consists of an Asp-His-Ser triad, where the Asp-His pair is believed to increase the nucleophilicity of the serine so as to facilitate its attack on the carbonyl carbon of the scissile bond. It now seems well established that the only charge transfer internal to the members of the catalytic triad is the abstraction of a proton from the serine by the histidine during the nucleophilic attack. This proposal has become known as the one-proton transfer mechanism in order to distinguish it from the two-proton tranfer or charge-relay system. The inumerable experimental and theoretical techniques which have been brought to bear on this question have been reviewed recently (Warshel *et al.*, 1989).

Recent discoveries continue to surprise and to provide further insight into the relationship between structure and function in the serine proteases. Experiments demonstrating that a mutant trypsin retains substantial activity even after all of the members of the catalytic triad have been eliminated by site-specific mutagenesis, is one such example (Corey & Craik, 1992). Such data have led to the suggestion that other structural determinants of the active site serve to mould the substrate into the required conformation for peptide bond hydrolysis, a process whose efficiency is subsequently dramatically improved by the presence of the members of the triad. A second example is the crystallographic demonstration that the Sindbis virus core protein (Choi *et al.*, 1991) and the picornaviral 3C **cysteine** proteinases (Allaire *et al.*, 1994) are structurally similiar to members of the trypsin family as predicted prior to their crystallographic determination (Bazan & Fletterick).

Crystallographic studies have also contributed considerably to our understanding of the structural basis of substrate specificity. Furthermore, this is also an area in which comparative molecular modelling techniques have also had considerable success. Perhaps the first example of such an application was the prediction of the importance of D189 in the determination of trypsin specificity based on the previously determined structure of chymotrypsin (Hartley, 1970). More recently, Murphy *et al.* (1988) were able to correctly predict the substrate specificity of the granzyme CCP1 prior to its experimental verification, using a molecular model based on the homologous Rat Mast Cell Protease II. In a further example Barbosa *et al.* (1993) proposed a structural model for the glutamate specific endopeptidase from *Streptomyces griseus* (sgpe) which was successful in predicting the most

IFSC-USP SERVIÇO DE BIBLIOTECA E INFORMAÇÃO

important interactions made between the glutamic acid of the substrate and the residues of the S1 pocket (nomenclature of Schechter & Berger, 1967) as subsequently observed in the crystal structure (Nienaber *et al.*, 1993).

In fairness however, modelling techniques have not always had such success. Several early examples of models for members of the trypsin family presented serious problems often associated with poor alignments due to low sequence identity (McLachlan & Shotton, 1971; Greer, 1981; Read *et al.*, 1984). Furthermore, it is becoming increasingly clear that the relationship between structure and specificity is more complex than originally imagined. Mutagenesis experiments aimed at converting trypsin into a chymotrypsin-like enzyme for example have shown the importance of loops outside the S1 pocket in determining primary specificity and emphasize the importance of the chemical steps of the reaction rather than simply binding in the determination of specificity (Hedstrom *et al.*, 1992, 1994a, 1994b). Studies of trypsin mutants including attempts to relocate the negative charge within the S1 pocket and the substitution of G226 by alanine have also shed light on the complex relationship between activity and specificity (Wilke *et al.*, 1991; Perona *et al.*, 1993).

Nevertheless, our recent success in the modelling of sgpe (Barbosa *et al.* 1993) stimulated the investigation of the active sites and specificity pockets of other glutamate specific endopeptidases (GSE's) none of whose crystal structures have as yet been determined. Probably the best known of the GSE's is the staphyloccocal serine protease v8 (Glu-C) which has been intensively employed in protein sequencing ever since its original description (Drapeau *et al.*, 1972). More recently, homologous proteins from *Bacillus*

licheniformis, gse-bl (Svendsen & Breddam, 1992; Kakudo *et al.*, 1992) and *Bacillus subtilis*, mpr (Sloma *et al.*, 1990) have been sequenced and partially characterized. There is also evidence that a glutamate specific enzyme may be responsible for activation of neutrophil serine protease zymogens (Salvesen & Enghild, 1990). Furthermore, two epidermolytic toxins (ET's) from *S. aureus* responsible for the Staphyloccocal Scalded Skin Syndrome have also been shown to present sequence homology with the v8 protease (Dancer *et al.*, 1990; Bailey & Smith, 1990). All the above molecules demonstrate hydrolytic activity either towards peptide or ester substrates containing glutamate residues in the P1 position.

The original reports of the sequences of the v8 protease and gse-bl demonstrated only local similarity to the members of the trypsin family in the vicinity of the members of the catalytic triad. Furthermore, mpr was classified as a metalloprotease on the basis of inhibition studies (Rufo *et al.*, 1990) and no comment was made concerning the epidermolytic toxins (O'Toole & Foster, 1987; Lee *et al.*, 1987; Sakurai *et al.*, 1988). We show that all are members of the trypsin family (although presenting at most 23% identity with trypsin itself) and, on the basis of sequence alignments and molecular models, attempt to define the structural basis for P1 glutamate specificity in terms of critical residues which decorate the S1 specifity pocket. This work therefore extends that of Bazan & Fletterick (1988, 1990) which was principally directed towards an understanding of the structural basis for glutamine P1 specificity in viral cysteine proteases but which also included a brief analysis of both v8 and the epidermolytic toxins. In so doing, we have uncovered some novel aspects of the structure of serine protease active sites.

Materials and Methods

Database searches

With a view to identifying as many recognised glutamate-specific endopeptidases of known sequence as possible the full sequence of v8 protease from *Staphyloccocus aureus* was used as probe in a search of the OWL database using the programs FASTA and SWEEP running on the Daresbury SEQNET facility. As well as identifying the epidermolytic toxins eta and etb, which are known homologues, the search detected the GSE from *B. licheniformis* (gse-bl). A subsequent search using gse-bl as the probe, identified the metalloprotease mpr from *B. subtilis*. Repeating the procedure with mpr found no futher sequences of interest.

Sequence alignments

Sequences of trypsin-like serine proteases for which three-dimensional structures were available were aligned by least-squares superposition of the C_α-coordinates. The structures included in this initial alignment, together with the respective PDB codes were as follows: chymotrypsin (4CHA), trypsin (2PTN), elastase (3EST), human neutrophil elastase (1HNE), kallikrein (2PKA), tonin (1TON), rat mast cell protease II (3RP2), trypsin (1SGT) and proteases A (2SGA) and B (3SGB) from *Streptomyces griseus* and α-lytic protease (2ALP). The glutamate specific endopeptidases and epidermolytic toxins typically present only approximately 20% sequence identity with the above mentioned enzymes. For their inclusion

in the alignment initially sequence profiles were derived from the structure-based alignment and also from that of Greer (1990) which includes a greater span of sequences not all of which are of known three-dimensional structure. All profiles were calculated according to the method of Gribskov *et al.*(1987), and in the case of the structure-based alignment, profiles were also calculated for the mammalian (including *S. griseus* trypsin) and bacterial enzymes independently. The amino acid sequences for the glutamate specific endopeptidases (v8, gse-bl and mpr) and two epidermolytic toxins (eta and etb) were aligned against these profiles using the program PROFILEGAP of the GCG package (Devereux *et al.*, 1984). At this stage several different combinations of gap and length penalties were tested with the various profiles in order to ascertain which parts of the alignment were stable. Ambiguities in the resulting alignments were resolved by manual inspection of the sequences; the C-terminal helix for example was aligned on the basis of the pattern of hydrophobic residues. Clearly such regions are, at best, dubious and where important for subsequent discussion are specifically identified.

The sequences for the Sindbis virus core protein (scp), the 3C cysteine-protease from hepatitis A (hav-3c) and the lysine specific protease I from *Achromobacter lyticus* (ach1) were subsequently included in the alignment as their crystal structures were reported.

Modelling

Given the low sequence identity between the five sequences of interest and those of known structure, the construction of complete molecular models would be a largely futile excercise. Instead it was decided to concentrate on the regions of the structure of principal interest, namely the active site and specificity pockets (in particular S1) which show the greatest level of structural conservation and are thus liable to produce partial models of greater reliability. Molecular models were constructed for gse-bl, mpr, v8 and eta using a silicon graphics IRIS Indigo and Control Data CYBER 910 workstations running the software O(Jones *et al.*, 1991), WHAT IF(Vriend, 1990) and TOM(Cambillau & Horjales, 1987). The starting structure used for all models was that of *Streptomyces griseus* trypsin at 1.7Å resolution (1SGT, Read & James, 1988), although this choice is somewhat arbitrary as no single structure is clearly more similar to the sequences to be modelled and all are available at better than 2.05Å resolution. Given the expected uncertainty in the atomic positions of the models the choice of starting structure is not expected to be critical for the conclusions drawn here.

From initial inspection of the sequences and from the results of the profile-based alignments it was apparent that strands 2, 3 and 6 of the six-stranded barrel in the N-terminal domain and strands 1, 4 and 5 from the second domain (here termed 1', 4' and 5') could be most easily identified. These strands correspond to the interface region between the two domains, in the vicinity of the active site. The construction of a complete model for gse-bl was attempted but its interpretation was restricted to the regions comprised of the above-

mentioned interface strands together with the N-terminus, strands 1 and 6' and the loop between strands 3' and 4'. Of particular relevance is the length of this latter loop, which was modelled on the conformation observed in the bacterial structures; α-lytic protease, and *S. griseus* proteases A and B. The rationale for this was a) the similarity in length of the bacterial sequences compared with the GSE's and b) the presence of D189 in the alignment of gse-bl, which would point into the S1 pocket if modelled on the basis of the mammalian structures. Such an orientation is unexpected for a glutamate specific enzyme due to the expected repulsion between the two acidic side-chains.

Within the S1 specificity pocket H213 and T190 were orientated on the basis of the conformations found for the equivalent histidine and serine in sgpe (Nienaber *et al.* 1993), as estimated from stereoscopic pictures. Remaining side-chains were either modelled on the basis of identical residues observed in other molecules or in order to make plausible interactions and to avoid uncommon rotamers. Docking of glutamic acid into S1 was achieved by superposition of the third domain of turkey ovomucoid inhibitor in complex with human neutrophil elastase (Bode *et al.*, 1986). The P1 sidechain was substituted and modelled based on sgpe as described above.

The model for mpr was based on that for gse-bl, with the modification that particular attention was paid to modelling the loop between strands 1' and 2'. The possibility of the formation of a disulphide bridge between the two unique cysteines at 143 and 193 was investigated. Mpr, along with the other sequences studied here, possesses a proline at 142 similar to that observed in protease I from *A. lyticus*. When built to adopt a similar

conformation, C143 of mpr is naturally brought close to C193.

In the modelling of the v8 protease particular attention was paid to the neighbourhood of the fourth member of the catalytic quartet (position 214) which is normally a serine but has been substituted by tryptophan in v8. Its environment includes the loops between strands 5 and 6 and between 2' and 3'. In this case the third domain of turkey ovomucoid was also used to model substrate binding to the S1' pocket whose specificity has been well characterized biochemically. For eta (which effectively represents both epidermolytic toxins) the process was restricted to the S1 pocket including K216. In this case the peptide substrate was replaced with a phosphoester by approximately superposing the phosphorus and three oxygen atoms of the phosphonyl onto the carbonyl and C_α of P1 and the nitrogen of P1'. Clearly such a superposition is only very approximate since the phosphorus is tetrahedral and the carbonyl carbon only planar. In this conformation the possibility of the formation of a salt bridge between the remaining phosphonyl oxygen and K216 was examined. For this purpose the relevant portion of the substrate-derived inhibitor (2-dodecanoyl-amino-1-hexanol-phosphoglycol) from its complex with porcine phospholipase A2 was used (Thunnissen *et al.*, 1990).

Dendrogram calculation.

A dendrogram describing the similarity relationships of the sequences included in the alignment was calculated using the Neighbour-Joining method of Saitou and Nei (1987) implemented in the program PHYLIP (Felsenstein, 1989). For such purposes the alignment was fixed to that previously described and the statistical confidence associated with the individual bifurcations determined using the bootstrap procedure (Felsenstein, 1985).

Results and Discussion

Sequence Alignment

FIG. 1 →

Figure 1 shows the sequence alignment derived from molecular superposition for the known crystal structures and from profile analyses in the case of the GSE's and epidermolytic toxins. Typical residue identities calculated between the sequences of interest and those of known 3D structure were of the order of 15-20%. In this context it is of interest to note that the Iterative Template Refinement method of Yi & Lander (1994) is able to directly identify the epidermolytic toxins and mpr as members of the trypsin family. Throughout we refer to residues according to the numbering scheme given in the alignment which is based on chymotrypsinogen (Hartley & Kauffmann, 1966). The most conserved regions lie close to the members of the catalytic triad (H57, D102 and S195). Six of the twelve strands comprising the two β-barrels of the trypsin fold were stable from one profile alignment to the next and

could be recognised from manual inspection of the sequences. These were strands 2, 3 and 6 from the N-terminal domain and 1', 4' and 5' from the C-terminal domain. Figure 2 shows them to form the core of the molecule at the domain interface in the vicinity of the active site. Besides the members of the catalytic triad (with the exception of C195 in hav-3c) only G196 is conserved across all sequences presumably due to the importance of maintaining a positive ϕ angle which is necessary for the correct orientation of the catalytic serine at 195. Several other residues are almost completely conserved, including the fourth member of the catalytic 'quartet' (S/T214), G193 which participates in the oxyanion hole, G140 which lies close to the N-terminal salt-bridge in mammalian enzymes and T/S54 whose O_γ forms an internal hydrogen bond stabilizing strands 2 and 3 close to the catalytic histidine. Hydrophobic residues are clearly conserved at several key structural positions.

FIG. 2 →

Residues of particular relevance to the modelling of the GSE's and epidermolytic toxins are indicated on Figure 1 and include a threonine and histidine at positions 190 and 213 respectively which are a hallmark of the GSE's. H213 is also conserved in sgpe where T190 has been substituted by serine. Other residues of particular note are the serine at 197; substitution of the aspartic acid at 194 in all GSE's and epidermolytic toxins; the N-terminal extension of one residue in gse-bl, mpr and v8; the arginine at position 30 in gse-bl and mpr; the substitution of residues 143 and 193 by cysteine in mpr; the replacement of S/T214 (the fourth member of the catalytic quartet) by tryptophan in v8 protease; and the presence of lysine at postion 216 in the epidermolytic toxins. The consequence of these substitutions for each of the individual structures is discussed below.

Also noteworthy is the unexpected alignment of D189 in both gse-bl and mpr since in trypsin-like enzymes the side chain of this acidic residue points into the S1 pocket and is responsible for the specificity for P1 basic residues and would therefore not be expected in glutamate specific enzymes. This apparent paradox is readily resolved if the short 3' to 4' loop adopts a conformation similar to that observed in the bacterial enzymes where the side chain of residue 189 points away from the S1 pocket. In all modelled structures it was treated as such, a supposition supported by its length.

The dendrogram shown in Figure 3 demonstrates that the majority of GSE's and ET's form a monophyletic group. The only glutamate specific enzyme to lie on an alternate branch of the dendrogram is sgpe and, as will be seen, there are some important differences in the specificity determining regions.

FIG 3

Modelling of gse-bl

The model for the S1 pocket of gse-bl clearly shows H213 in a position to form a hydrogen bond to the P1 glutamate via a hydrogen on $N_{\varepsilon 2}$ (Figure 4). A second hydrogen bond to the same carboxylate oxygen can be formed via the O_{γ} of T190 which is orientated in the normally unfavourable gauche$^-$ conformation. These interactions (including the side-chain rotamers) are analogous to those observed in sgpe by Nienaber *et al.*, (1993), although the alignment given by these authors is incorrect and assigns S190 to position 192.

FIG. 4

The importance of H213 and T/S190 is demonstrated by their conservation in all GSE's (Figure 1). In sgpe it has been proposed that glutamate specificity is conferred upon the enzyme by the presence of an unusual histidine triad which leads from the N-terminus of the C-terminal helix through the second β-barrel and results in a positive charge (or partial charge) on H213. No other basic residues are present within the S1 pocket. In GSE-BL the remaining histidines of the triad (199 and 228) are not conserved suggesting that an alternative explanation must be sought. Three possibilities are suggested. Firstly it is possible that polarization of H213 by the P1 glutamate itself is sufficient. This is consistent with the reduced pH optimum of gse-bl (and all other GSE's) compared with sgpe, 7.3-8.0 as opposed to 9.0 (Yoshida *et al.*, 1988; Svendsen & Breddam, 1992; Sorensen *et al.*, 1991). However, random mutagenesis studies of the S1 pocket of α-lytic protease have shown that the presence of H213 is not sufficient to produce a glutamate-specific enzyme (Graham *et al.*, 1993). Furthermore, the hav-3c cysteine proteinase also possesses H213 but is **glutamine** rather than glutamate specific (Jewell *et al.*, 1992). Transference of such results from different systems is a dangerous process but they suggest that alternative explanations for glutamate specificity should at least be investigated.

A second possibility is that the additional residue at the N-terminus of gse-bl in comparison with the mammalian enzymes might provide a positive charge to the S1 pocket. The N-terminal residue of mammalian enzymes forms an essential salt-bridge with D194 which is important for zymogen activation. Two striking features of the GSE sequences are the substitution of D194 by the neutral asparagine or glutamine and an N-terminal extension of one residue. This leaves an excess positive charge at the N-terminus which, due to its

extra length, can reach into the S1 pocket if one admits a small adjustment to the mainchain in the region of T190. However, due to the uncertainty in the alignment of the N-terminal region and due to steric clashes, principally with T190, produced on attempting such a model, this possibility is not favoured. Instead, it seems more likely that the N-terminus adopts a conformation similar to that recently observed in the lysine specific enzyme from *A. lyticus* in which P142 (conserved in the GSE's) causes a deviation of the mainchain in order to accommodate it.

The final possibility which the model raises is that R30, in the first β-strand of the N-terminal barrel, and which is buried in the interior of the structure, may be the ultimate source of the positive charge which stabilizes the glutamate at P1. This position, which is often occupied by a glutamine, is reasonably well determined by the alignment due to the presence of an aromatic cluster which precedes it (Blevins & Tulinsky, 1985). In this case the arginine would interact either directly or via water with S197 which in turn (either directly or indirectly) hydrogen bonds to the $N_{\delta 1}$ of H213 and thus to the P1 glutamate (Figure 4).

Either way, it seems likely that the $N_{\delta 1}$ of H213 interacts either directly or via a water molecule with S197 since the latter is also notable in that it is conserved in all GSE's with the exception of sgpe where $N_{\delta 1}$ participates in the histidine triad.

None of the above hypotheses support the suggestion of Kakudo *et al.* (1992) that R96 (R89 in the authors' numbering) is responsible for substrate specificity. This residue is neither conserved in GSE's nor is its assignment as the homologue of G216 of chymotrypsin

consistent with the residues identified by the authors and ourselves as the members of the catalytic triad. It seems likely that this possibility can be disgarded.

Peptide substrates bearing aspartic acid at P1 show values of k_{cat}/K_m approximately 1000-fold lower than the corresponding substrates with P1 glutamate (Breddam & Meldal, 1992). This reduction is principally due to an effect on k_{cat} rather than K_m, showing the enzyme to bind aspartate reasonably well in the S1 pocket. From the model it is clear that, due to the shorter side-chain, the interaction of the aspartate with H213/T190 would result in poor alignment of the sissile bond with respect to the catalytic serine. With respect to the S1' pocket the model suggests that the poor acceptance of aspartic acid at P1' (Breddam and Meldal, 1992) may be due to the close proximity of D61.

mpr from B. subtilis

Most of the above discussion also holds for mpr, including the possible explanations for substrate specificity, since all of the relevant residues are conserved in both enzymes (Figure 1). However one additional feature of the model is of particular note, namely the substitution of G193 by cysteine. The backbone amide of G193 participates in the formation of the oxyanion hole which is essential for stabilizing the charge which develops on the carbonyl oxygen of the tetrahedral intermediate during catalysis. As is therefore to be expected, it is one of the most conserved residues of the trypsin family.

Besides the cysteine at 193, mpr also contains a second unique cysteine which lies

spacially close to the first, in the loop between strands 1' and 2' (C143) and the model suggests that the formation of a disulphide bridge between these residues is feasible (Figure 5). This hypothesis helps to explain the apparent paradox between the clear identification made here that mpr is a member of the **serine** proteases and the reported biochemical evidence that led to its original classification as a metalloprotease.

FIG.5

Incubation of mpr with either DTT or EDTA/mercaptoethanol leads to the loss of catalytic activity (Rufo *et al.*, 1990) whereas PMSF (a serine protease inhibitor) is ineffective. Our model suggests that the disulphide bridge is responsible for stabilizing C193 in the disallowed conformation normally adopted by the glycine in this position (ϕ~100, ψ~-20), and as such the effect of the incubation with reducing agents is probably the loss of the disulphide bridge. The consequence is the removal of the compensatory force of the disulphide covalent bond which maintains the conformation of C193 and thus the integrity of the oxyanion hole. Upon reduction, C193 would be expected to reorientate with a subsequent loss of catalytic activity.

The possibility that the two cysteines participate in a metal-binding site can not be eliminated. This seems less likely given that the model did not suggest any obvious candidates for completion of the metal coordination sphere. However, either way, the mechanism of inhibition would be basically the same; a reorientation of C193 and disruption of the oxyanion hole, rather than chelation of a catalytic metal ion as implied by Rufo *et al.* (1990). The fact that PMSF is not an effective inhibitor of mpr does not eliminate the possibility that it is a serine protease as other enzymes, notable v8 protease, are similarly

resistent (Salvesen & Nagase, 1990).

v8 protease from S. aureus

The arguments used above for gse-bl and are also largely valid for the v8 protease, with the exception that R30 is not conserved in the latter, somewhat weakening the suggestion that it may be important in the determination of substrate specificity for GSE's as a whole. In the case of v8 protease, however, specific biochemical evidence can be invoked to support the importance of H213 in the S1 pocket. Two ionizable groups with pK_a's of 6.58 and 8.25 (Houmard, 1976) or 5.8 and 8.4 (Sϕrensen *et al.*, 1991) have been shown to be essential to the catalytic process. The former was attributed to the catalytic histidine but no assignment was made for the latter, which may correspond to H213. If this were so, a pK of above 8 for this histidine would be sufficient to account for glutamate specificity since it would be positively charged at the pH optimum of around 7.2 for v8 protease (Sϕrensen *et al.*, 1991). The model also explains the preference which v8 displays for large hydrophobic residues at P1' (Schuster *et al.*, 1990) since the S1' subsite as shown in Figure. 6 is both spacious and lined with hydrophobic residues, notably I41, A42 and V58, the latter two of which substitute a disulphide bridge (Figure 1).

FIG-6 →

Of greater interest in the case of the v8 model is the vicinity of the fourth member of the catalytic quartet, normally S/T214. This residue forms part of what has been described as the 'aspartate hole' (Warshel *et al.*, 1989) important for stabilizing D102 of the catalytic triad.

Besides S/T214, the catalytic aspartate receives hydrogen bonds from $N_{\delta 1}$ of H57 of the catalytic triad and the backbone amides of 56 and 57. Furthermore, in the mammalian enzymes the loop between strands 5 and 6 covers the D102-H57 couple so protecting it from solvent. In the bacterial enzymes α-LP, SGPA and SGPB, this loop is smaller and the equivalent role seems to be played by the enlarged loop between strands 2' and 3' as shown schematically in Figure 7. In both types of structure the aromatic 'lid' at position 94 appears to be important in contributing to this protection (Brayer *et al.*, 1978).

FIG. 7 →

In the v8 protease several alterations in this region may be pertinent to catalytic activity. The principal differences when compared with the majority of the family members are 1) the substitution of S/T214 by tryptophan; 2) the presence of small loops both between strands 5 and 6 (including the absence of the aromatic 'lid') and between 2' and 3'; and 3) the notable substitutions at positions 55 and 56 which are normally small (either glycine or alanine) for reasons of steric hindrance. In fact the alignment in the region from 1' to 4' is rather uncertain and thus it is difficult to assign exactly the size of the 2'-3' loop. However this whole region is shorter than in other sequences probably resulting in a smaller loop even if the exact alignment shown in Figure 1 is not perfect.

The loss of the large loops and the aromatic lid which serve to protect the D102-H57 couple from exposure to solvent and the substitution of the fourth member of the catalytic quartet (which normally helps to orientate D102) probably result in a reduction of the strength of the interaction of the D102-H57 couple. In turn this would be expected to lead to a reduction in the nucleophilicity of the catalytic S195, which may be the origin of the much

lower k_{cat}/K_m values for v8 protease on peptide substrates when compared with gse-bl and sgpe (Breddam & Meldal, 1992) and also its relatively poor reactivity with DFP.

Furthermore, although K56 can be easily accommodated due principally to the small loop between strands 5 and 6 and the loss of the aromatic 'lid', N55 presents greater problems. As can be seen from the alignment, the alanine at 55, together with its environment (composed of the residues of the catalytic quartet, I212 and the disulphide between residues 42 and 58) are all highly conserved. This makes accommodation of the asparagine in v8 impossible without some rearrangement of the mainchain. Although the nature of such changes are practically impossible to predict, N55 could potentially form several H-bonds, including with the catalytic aspartate if this were to adopt a rather different conformation. The A55N substitution, together with the appearance of the bulky W214 sidechain and the loss of the disulphide, argue strongly for a rearrangement in this region which may further weaken the interaction of D102 with the amides of residues 56 and 57 as seen in the Sindbis virus core protein where P55 and L56 also represent significant deviations from the norm. Furthermore it is conspicuous that the sequences that show large substitutions at these positions (v8, eta and etb) all lack the above mentioned disulphide and possess short loops between strands 5 and 6.

Site directed mutagenesis studies of subtilisin led Carter and Wells (1988) to conclude, on the assumption of a stepwise increase in turnover number, that the most likely sequence of evolutionary events leading to the present day catalytic triad would be the introduction of S195 followed by H57 and finally D102. A similar conclusion for trypsin is consistent with

the results of Corey & Craik (1992) who conclude that an oxyanion hole and a nucleophile are the essential components of a primitive protease. It seems reasonable to conclude that a further step in the optimization of catalytic efficiency would be the appearance of the aspartate hole and the loops which cover the D102-H57 couple in order to improve the strength of the ion-pair interaction which ultimately leads to the unusual nucleophilicity of the catalytic serine. The model for v8 protease suggests that we are seeing an example of a serine protease which lacks this 'final' stage of optimization. It is of interest to note that the recently determined structure of the 17E8 antibody which presents hydrolytic activity against norleucine and methionine phenyl esters possesses an active site which resembles that of the serine proteases including an equivalent of the oxyanion hole and a Ser-His couple (Zhou *et al.*, 1994). It may too be analogous to an intermediate in serine protease evolution in which the active site has not yet been optimized by the incorporation of an aspartic acid into the catalytic triad.

The model for v8 resembles in some respects the recently determined structures of ach1, scp and hav-3c which also show modifications to the environment of D102. In the case of hav-3c the aspartate is rotated out of the active site cleft such that it no longer interacts with H57, although this is not the case in the related protease from rhinovirus which has a glutamate as the third member of the catalytic triad (Matthews *et al.*, 1994). Besides the loss of the hydrogen bond between 214 and D102, which helps to orientate the latter, at least one of the hydrogen bonds to the backbone amides of residues 56 and 57 is also weakened in ach1 and scp. In both enzymes the aromatic 'lid' is absent and the D102-H57 hydrogen bond more exposed to solvent.

Epidermolytic Toxins

The epidermolytic toxins are responsible for a blistering disease in children known as the Staphylococcal Scalded Skin Syndrome in which confluent patches of reddened skin form as a result of breakdown of the intercellular cohesive forces at the level of the *stratum granulosum* (Lillibridge *et al.*, 1972). Principal among the effects of the toxins is the disruption of desmosomes, although the exact mechanism by which this occurs is unknown. In particular, the nature of the toxins' natural 'receptor' and/or substrate and the reason for their selectivity in site of action remain unanswered questions.

The model for eta shows many of the features already described above, although it possesses a larger N-terminal extension than the GSE's and, like v8, does not have R30. However, as shown in Figure 8, H213 and T190, the core elements of the S1 pocket, are present. No convincing evidence has so far been presented demonstrating the epidermolytic toxins to show proteolytic activity, however Bailey and Redpath (1992) have observed glutamate specific esterase activity which can presumably be attributed to the H213/T190 couple in the S1 pocket and the presence of a functional catalytic triad. Furthermore, the importance of S195 for epidermolysis is unquestionable since mutants in which S195 has been changed to glycine (Redpath *et al.*, 1991) or cysteine (Prevost *et al.* 1991) and chemical modification with DFP (Dancer *et al.*, 1990) show dramatic reductions in epidermolytic activity.

FIG. 8 →

One unique feature of the S1 pocket of eta (and by implication etb) is K216. It has

been suggested that the ET's also present phospholipase C-like activity (Wiley & Rogolsky, 1985) and the modelling of the phosphonyl group of a phosphodiester into the active site in place of the peptide, shows that K216 could stabilize the negative charge on the phosphonyl together with H213 (Figure 8). In the figure, H213 has been rotated further about χ_1 in comparison with the previous figures in order to interact directly with one of the oxygens of the phosphate moiety. Although highly speculative the model suggests that the mechanism of phosphoester bond cleavage by nucleophylic attack of the catalytic serine on the phosphorus may be similar to that believed to occur in alkaline phosphatase (Kim & Wyckoff, 1991).

It is not clear whether either the reported esterase or phospholipase activities of ET's are relevant to their physiological activity and in particular to the blistering disease which they cause, or if they are simply consequences of the serine protease-like active site geometry proposed here. After all, most serine proteases are also esterases, although to our knowledge do not present phospholipase activity, for which the presence of K216 may therefore be critical. Certainly K216 in addition to H213 does clearly suggest that the natural substrate, whatever it may be, should be expected to possess an anionic group which interacts with the S1 pocket. The model clearly suggests more interesting site-directed mutagenesis experiments than those which have been performed up until now. In particular the importance of H213, T190 and K216 for epidermolysis could be readily tested by such a methodology.

Conclusions

We have described a canonical feature of glutamate specific endopeptidases and epidemolytic toxins. Common to the S1 pocket of all such enzymes are H213 and T/S190, which we suggest are fundamental in determining the unique specificity of these molecules. What is less clear is whether they are in themselves the only determinants. The only crystal structure for a member of this group of enzymes is that from *S. griseus* (sgpe) which presents a histidine triad leading from H213 through to the C-terminal helix. Although this may be important for determining glutamate specificity in sgpe it is not conserved in any of the enzymes studied here.

It is clear that substrate specificity is a complex issue and that we are still some way from a full understanding of its structural basis. One thing that is clear is that different structural solutions can often be found which give rise to the same substrate specificity. The lysine specific protease I from *A. lyticus* clearly demonstrates this point. In this case the fundamental determinant of lysine specificity is D226 instead of D189 as found in trypsin and others. This is analogous to the different protease families which have arisen during the course of evolution as different solutions to the problem of peptide bond hydrolysis.

Although Figure 3 is not a phylogenetic tree, it probably implies that glutamate specificity has evolved at least twice independently in the trypsin family. H213 is a common feature to the solution found on both occasions but at position 190 threonine is found in the enzymes along the principal branch of the dendrogram whilst serine is found in sgpe. Only in

this latter enzyme is there a histidine triad which presumably helps in elevating the pK_a of H213 and perhaps therefore in raising the pH optimum in comparison with the remaining enzymes. Neither of the two solutions involves a simple substitution of D189 (as found in trypsin) by a basic residue, as might have been anticipated given the failure of attempts to do so by site directed mutagenesis (Graf *et al.*, 1987). In the case of the hav-3c from hepatitis A virus, the presence of a buried glutamic acid close to H213 is used as justification for the latter being neutrally charged and thus confering **glutamine** specificity on the enzyme.

Besides the question of substrate specificity, the modelling has uncovered a novel means for stabilizing the oxyanion hole via a disulphide bridge (or potentially a metal-binding site) and has emphasized the importance of the aspartate 'hole'. Recent results seem to suggest that we are seeing an extension to the parts of the molecule which are believed to be important for determining substrate specificity and/or catalytic activity. As a general rule the importance of such features attenuates with their distance from the active site. Nevertheless, more of the structure may have to be considered as being critical for one or other of these functions than was originally anticipated.

Acknowledgments

We thank the CNPq, FAPESP and FINEP (brazilian funding bodies) for financial support and gratefully acknowledge the Daresbury Laboratory for access to the SEQNET facility. J.A.R.G.B. is the recipient of a CNPq studentship.

References

Allaire, M., Chernaia, M.M., Malcolm, B.A. & James, M.N.G. (1994) *Nature* **369**, 72-76

Bailey, C.J. & Redpath, M.B. (1992) *Biochem. J.* **284,** 177-180

Bailey, C.J. & Smith, T.P. (1990) *Biochem. J.* **269**, 535-537

Barbosa, J.A.R.G., Garratt, R.C. & Saldanha, J.W. (1993) *FEBS Letts.* **324**, 45-50

Bazan, J.F. & Fletterick, R.J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 7872-7876

Bazan, J.F. & Fletterick, R.J. (1990) *Semin. Virol.* **1**, 311-322

Blevins, R.A. & Tulinsky, A. (1985), *J. Biol. Chem.* **260,** 4264-4275

Blow, D.M., Birktoft, J.J. & Hartley, B.S. (1969) *Nature* **221**, 337-340

Bode, W., Wei, A.Z., Huber, R., Meyer, E., Travis, J., Neumann, S. (1986) *EMBO J.* **5**, 2453-2462

Brayer, G.D., Delbaere, L.T.J. & James, M.N.G. (1978) *J. Mol. Biol.* **124**, 261-283

Breddam, K & Meldal, M. (1992) *Eur. J. Biochem.* **206,** 103-107

Cambillau, C.C. & Horjales, E. (1987) *J. Mol. Graph.* **5**, 174-177

Carter, P. & Wells, J.A. (1988) *Nature* **332**, 564-568

Choi, H-K., Tong, L., Minor, W., Dumas, P., Boege, U., Rossmann, M.G. & Wengler, G. (1991) *Nature* **354**, 37-43

Corey D.R. & Craik, C.S. (1992) *J. Am. Chem. Soc.* **114**, 1784-1790

Dancer, S.J., Garratt, R., Saldanha, J. Jhoti, H. & Evans, R. (1990) *FEBS Letts* **268**, 129-132

Devereux, J., Haeberli, P. & Smithers, O. (1984) *Nucl. Acid Res.* **12,** 387-395

Ding, X., Rasmussen, B.F., Petsko, G. & Ringe D. (1994) *Biochemistry* **33**, 9285-9293

Drapeau, G.R., Boily, Y. & Houmard, J. (1972) *J. Biol. Chem.* **247**, 6720-6726

Felsenstein, J. (1985) *Evolution* **39**, 783-791

Felsenstein, J. (1989) *Cladistics* **5**, 164-166

Graf, L., Craik, C.S., Patthy, A., Roczniak, S., Fletterick, R.J. & Rutter, W.J. (1987) *Biochemistry* **26**, 2616-2623

Graham, L.D., Haggett, K.D., Jennings, P.A., Le Brocque, D.S. & Whittaker, R.G. (1993) *Biochemistry* **32**, 6250-6258

Greer, J. (1981) *J. Mol. Biol.* **153**, 1027-1042

Greer, J. (1990) PROTEINS: *Structure, Function and Genetics* **7**, 317-334

Gribskov, M., McLachlan, A.D. & Eisenberg, D. (1987) *Proc. Natl. Acad. Sci.* **84**, 4355-4358

Hartley, B.S. (1970) *Phil. Trans. Roy. Soc. Lond. B* **257**, 77-87

Hartley, B.S. & Kauffman, D.L. (1966) *Biochem. J.* **101**, 229-231

Hedstrom, L., Farr-Jones, S., Kettner, C.A. & Rutter, W.J. (1994a) *Biochemistry* **33**, 8764-8769

Hedstrom, L., Perona, J.J. & Rutter, W.J. (1994b) *Biochemistry* **33**, 8757-8763

Hedstrom, L., Szilagyi, L. & Rutter, W.J. (1992) *Science* **255**, 1249-1253

Houmard, J. (1976) *J. Biol. Chem.* **68,** 621-627

Jewell, D.A., Swietnicki, W., Dunn, .M. & Malcolm, B.A. (1992) *Biochemistry* **31,** 7862-7869

Jones, T.A., Zou, J-Y., Cowan, S.W. & Kjeldgaard, M. (1991) *Acta Cryst.* **A47**, 110-119

Kakudo, S., Kikuchi, N., Kitadokoro, K. Fujiwara, T., Nakamura, E., Okamoto, H., Shin, M., Tamaki, M., Teraoka, H., Tsuzuki, H. & Yoshida, N. (1992) *J. Biol. Chem.* **267**, 23782-23788

Kim, E.E. & Wyckoff, H.W. (1991) *J. Mol. Biol.* **218,** 449-464

Kossiakoff, A.A. & Spencer, S.A. (1981) *Biochemistry* **20**, 6462-6473

Lee, C.Y., Schmidt, J.J., Johnson-Winegar, A.D., Spero, L. & Iandolo, J.J. (1987) *J. Bacteriol.* **169**, 3904-3909

Lillibridge, C.B., Melish, M.E. & Glasgow, L.A. (1972) *Paediatrics* **50,** 728-738

Matthews, D.A., Smith, W.W., Ferre, R.A., Condon, B., Budahazi, G., Sisson, W., Villafranca, J.E., Janson, C.A., McElroy, H.E., Gribskov, C.L., Worland, S. (1994) *Cell* 77, 761-771

McLachlan, A.D. & Shotton, D.M. (1971) *Nature* **229**, 202-205

Murphy, M.E.P., Moult, J., Bleackley, R.C., Gershenfeld, H., Weissman, I.L. & James, M.N.G. (1988) *Proteins: Structure, Function and Genetics* **4**, 190-204

Nienaber, V.L., Breddam, K. & Birktoft, J.J. (1993) *Biochemistry* **32**, 11469-11475

O'Toole, P.W. & Foster, T.J. (1987) *J. Bacteriol.* **169**, 3910-3915

Perona, J.J., Craik, C.S. & Fletterick, R.J. (1993) *Science* **261**, 620-621

Perona, J.J., Tsu, C.A., McGrath, M.E., Craik, C.S. & Fletterick, R.J. (1993) *J. Mol. Biol.* **230**, 934-949

Prevost, G., Rifal, S., Chaix, M.L. & Piemont, Y. (1991) *Infect. Immun.* **59**, 3337-3339

Read, R.J., Brayer, G.D., Jurasek, L. & James, M.N.G. (1984) *Biochemistry* **23**, 6570-6575

Read, R.J. & James, M.N.G. (1988), *J. Mol. Biol.* **200**, 523-551

Redpath, M.B., Foster, T.J. & Bailey, C.J. (1991) *FEMS Letts.* **81**, 151-156

Rufo, Jr., G.A., Sullivan, B.J., Sloma, A. & Pero, J. (1990) *J. Bacteriol.* **172**, 1019-1023

Sakurai, S., Suzuki, H. & Kondo, I (1988) *J. Gen. Microbiol.* **134**, 711-717

Saitou, N. & Nei, M. (1987) *Mol. Biol. Evol.* **4,** 406-425

Salvesen, G. & Enghild, J.J. (1990) *Biochemistry* **29,** 5304-5308

Salvesen, G. & Nagase, H. (1990) In Beynon, R.J & Bond, J.S. (eds) *Proteolytic Enzymes: a practical approach* IRL Press, Oxford, pp 83-104.

Schechter, I & Berger, A. (1967) *Biochem. Biophys. Res. Comm.* **27**, 157-162

Schuster, M., Aaviksaar, A., Schellenberger, V. & Jakubke, H-D. (1990) *Biochem. Biophys. Acta* **1036**, 245-247

Singer, P.T., Slamas, A., Carty, R.P., Mangel, W.F. & Sweet, R.M. (1993a) *Science* **259**, 669-673

Singer, P.T., Slamas, A., Carty, R.P., Mangel, W.F. & Sweet, R.M. (1993b) *Science* **261**, 621-622

Sloma, A., Rudolph, C.F., Rufo Jr., G.A., Sullivan, B.J., Theriault, K.A., Ally, D. & Pero, J. (1990) *J. Bacteriol.* **172**, 1024-1029

Sϕrensen, S.B., Sϕrensen, T.L & Breddam, K. (1991) *FEBS Letts.* **294**, 195-197

Svendsen, I. & Breddam, K. (1992) *Eur. J. Biochem.* **204**, 165-171

Thunnissen, M.M.G.M., AB, E., Kalk, K.H., Drenth, J., Dijkstra, B.W., Kuipers, O.P., Dijkman, R., de Haas, G.H. & Verheij, H.M. (1990) *Nature* **347**, 689-691

Vriend, G. (1990) *J. Mol. Graph.* **8**, 52-56

Warshel, A., Naray-Szabo, G., Sussman, F. & Hwang, J-K. (1989) *Biochemistry* **28**, 3629-3637

Wiley, B.B & Rogolsky, M.S. (1985) In Jeljaszewicz, J. (ed.) *The Staphylococci, Zbl. Bakt. Suppl. 14* Verlag, New York, pp 295-300

Wilke, M.E., Higaki, J.N., Craik, C.S. & Fletterick, R.J. (1991) *J. Mol. Biol.* **219**, 525-532

Yoshida, N., Tsuruyama, S., Nagata, K., Hirayama, K., Noda, K. & Makisumi, S. (1988) *J. Biochem (Tokyo)* **104**, 451-456

Yi, T-M. & Lander, E.S. (1994) *Protein Sci.* **3**, 1315-1328

Zhou, G.W. Guo, J., Huang, W., Fletterick, R.J. & Scanlan, T.S. (1994) *Science*, **265**, 1059-1064

Figure Legends

Fig. 1. Sequence alignment of the serine proteases considered in the present work. The enzymes are divided into three groups: a) the glutamate specific endopeptidases and epidermolytic toxins, subject of the present study; b) enzymes of known three-dimensional structure (the first five of which are of bacterial origin and the remainder of mammalian origin) and c) the sequences of two viral and one bacterial enzyme whose structures were determined more recently. The abreviations used correspond to the following proteins: mpr, metalloprotease from *B. subtilis*; gse-bl, GSE from *B. licheniformis*; v8, v8 protease from *S. aureus*; eta and etb, epidermolytic toxins; sgpe, sgpa, sgpb, proteases E, A and B from *S. griseus* respectively; alp, α-lytic protease; sgt, *S. griseus* trypsin; hne, human neutrophil elastase; chymo, chymotrypsin; tryp, trypsin; elast, porcine elastase; kall, kallikrein; rmcp2, rat mast cell protease II; scp, sindbis virus coat protein; hav-3c, 3C cystein protease from hepatitus A virus; achl, lysine specific protease I from *A. lyticus*. The regions of secondary structure are indicated by boxes for the sequences of known 3D structure and the strands which form the N- and C-terminal β-barrels are labelled 1 to 6 and 1' to 6' respectively. Two conserved helices are indicated H1 and H2. The members of the catalytic quartet are indicated by crosses and the principal residues discussed in the text by bullets. In the case of scp, hav-3c and achI, for simplicity the full sequences are not given but the length of the regions not shown explicitly are given in parentheses.

Fig. 2. Ribbon representation of the trypsin fold (from 2PTN) showing in bold the β-strands of the two barrels which could be readily identified from the sequences of GSE's and ET's.

These strands form the core of the molecule at the interface between the two domains where the catalytic triad (S195, H57, and D102) resides.

Fig. 3. Dendrogram of the sequences shown in Figure. 1 (with the exception of scp, hav-3c and ach1) calculated using the neighbour joining algorithm. The numbers represent the percentage of dendrograms in which the individual bifurcations were reproduced during bootstrap sampling.

Fig. 4. Stereo figure of the active site and S1 pocket of the model for gse-bl. Residues drawn with open bonds correspond to regions of the model to which reasonable confidence may be attached, including H213 and T190 which are predicted to form H-bonds directly with one of the carboxylate oxygens of the P1 glutamate. A water molecule (W) is shown bridging the $N_{\delta 1}$ of H213 to the O_{γ} of S197. For clarity only the sidechains are shown for residues T190 and S197. R30 and the N-terminus (whose possible significance is described in the text) are drawn in lighter print. R30 could ultimately be bridged to H213 via S197 and a water molecule but this would require that the latter side-chain be reorientated. If the N-terminus were to participate, the T190 sidechain for steric reasons would need to adopt the *trans* conformation rather than *gauche*$^-$ as shown.

Fig. 5. Stereo figure of mpr in the region of the S1 pocket and oxyanion hole. In this case the model suggests the formation of a disulphide bridge between C193 and C143 which stabilizes the unfavourable mainchain conformation of C193 such that its backbone amide can contribute to the formation of the oxyanion hole together with the amide of S195 as shown.

Fig. 6. Stereo figure of the S1 and S1' subsites of v8 protease with a modelled substrate of sequence Glu-Phe (in bold). In the S1' subsite I41, A42 and V58 are predicted to be responsible for the observed preference for large hydrophobics at P1'. Also shown is the tryptophan at 214 which replaces the fouth member of the catalytic quartet (normally serine or threonine). The side chain of W214 can be accomodated due to the reduction in the size of the loop between strands 5 and 6. However, there is insufficient space for the N55 side chain (which replaces a conserved alanine) due to the presence of the catalytic triad, V58, I212 and the mainchain of H213, suggestive of the need for an alteration in the mainchain of at least one of these regions, all of which lie close to the active site.

Fig. 7. (a) Schematic representation of the vicinity of D102 in the mammalian and bacterial structures of Figure. 1, showing S214, and the loops between strands 5 and 6 (or 2' and 3') which protect the aspartate from direct exposure to solvent. (b) The analogous region from v8 protease showing the substitution at positions 214, and the small size of the corresponding loops, leading to a reduced protection of D102.

Fig. 8. Model of the S1 subsite of eta (etb would be expected to be similar) with a phosphonyl group of a phospholipid modelled in place of the peptide substrate. As can be seen, K216 could aid in the stabilization of the negative charge on one of the phosphonyl oxygens (O2) whilst a slight alteration to the histidine side chain allows $N_{\varepsilon 2}$ to form a hydrogen bond with O4. O1 sits in the oxyanion hole.

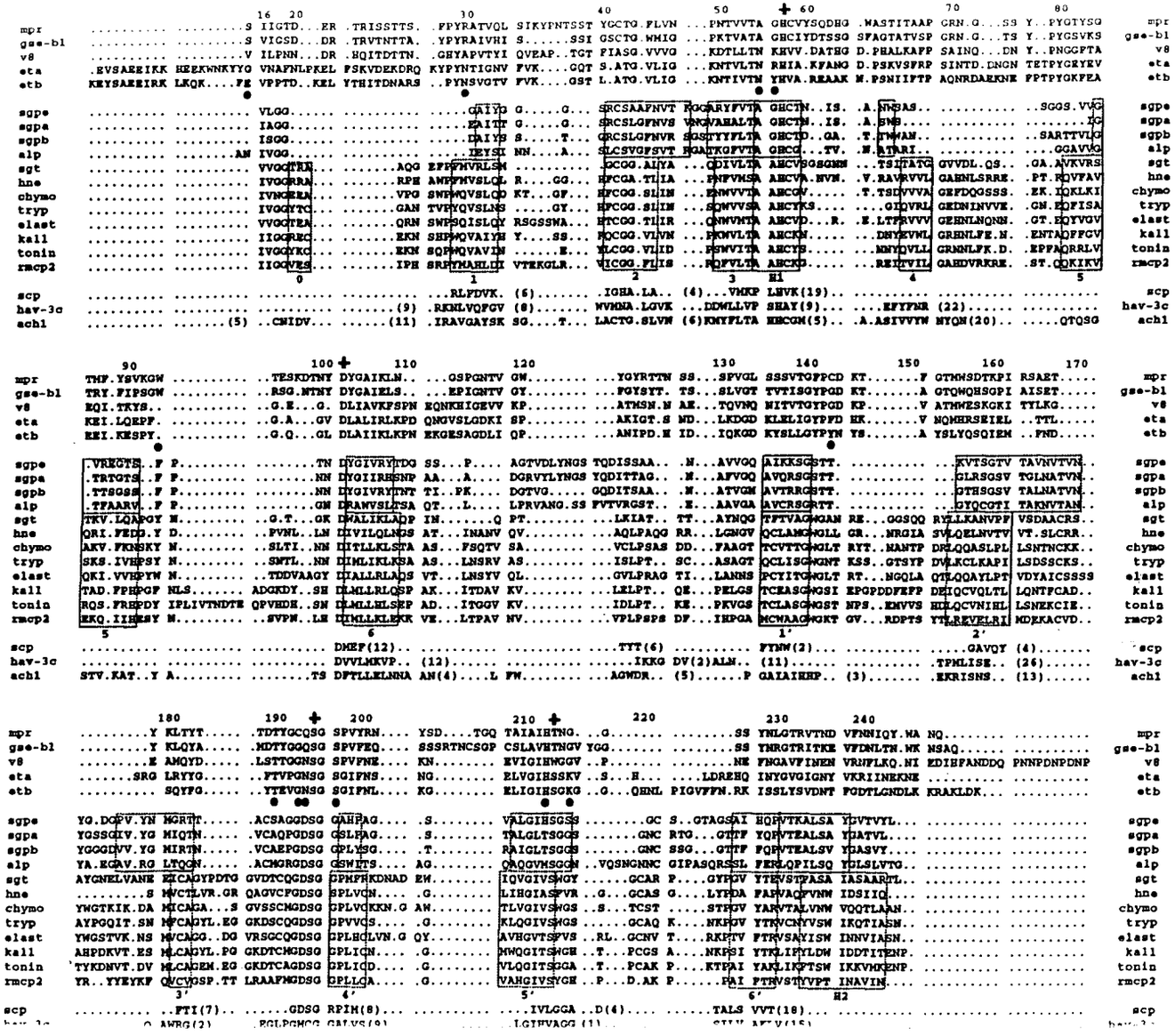


Fig. 1

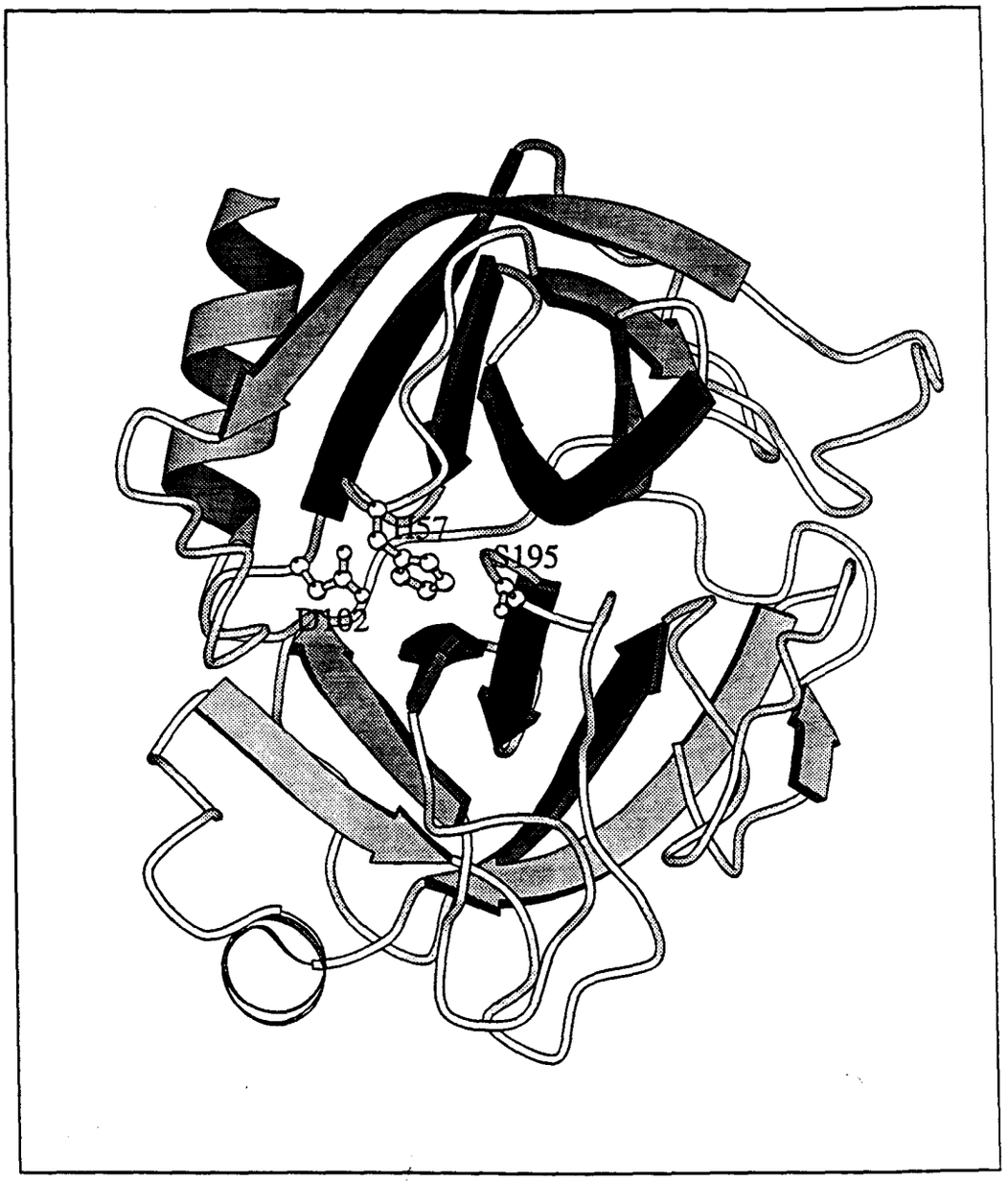


Fig. 2

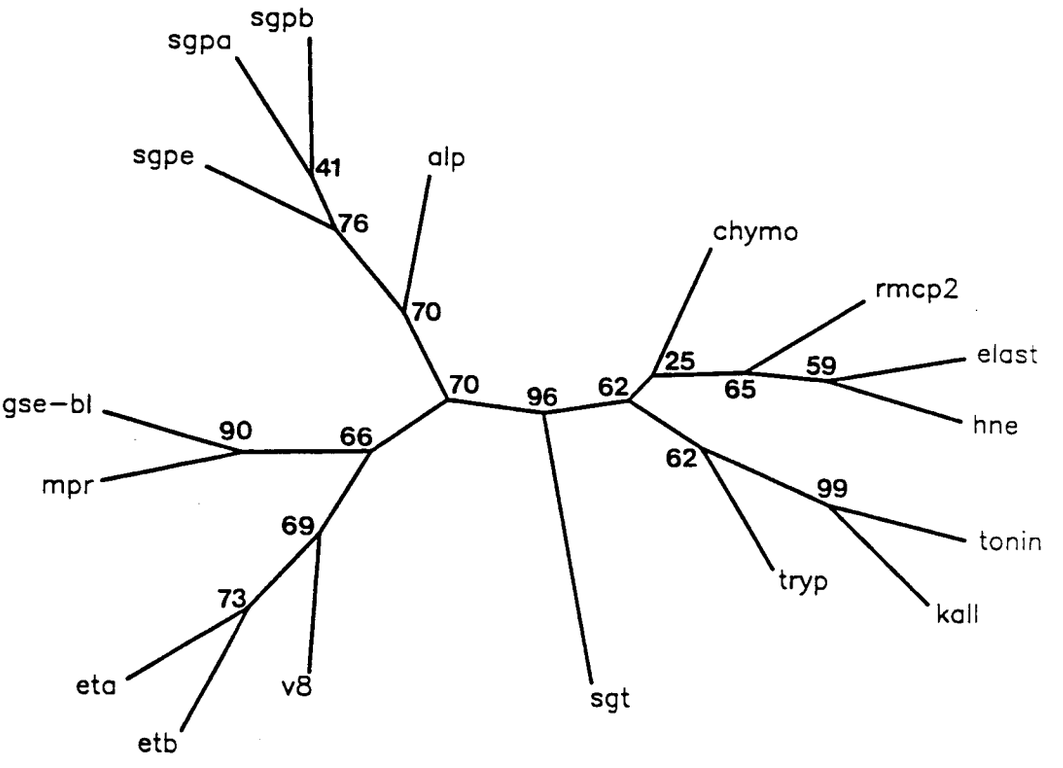

sgpb
sgpa
sgpe
41
alp
76
70
chymo
rmcp2
25
59
elast
65
70
96
62
gse-bl
90
66
hne
62
mpr
99
69
tonin
tryp
73
kall
eta
v8
sgt
etb

Fig. 3

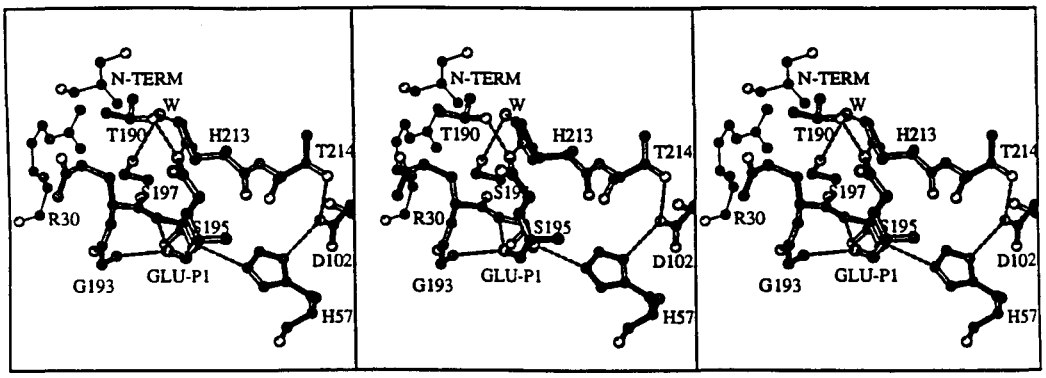

N-TERM
W
T190
H213
T214
S197
R30
S195
D102
G193
GLU-P1
H57
N-TERM
W
T190
H213
T214
S197
R30
S195
D102
G193
GLU-P1
H57
N-TERM
W
T190
H213
T214
S197
R30
S195
D102
G193
GLU-P1
H57

Fig. 4

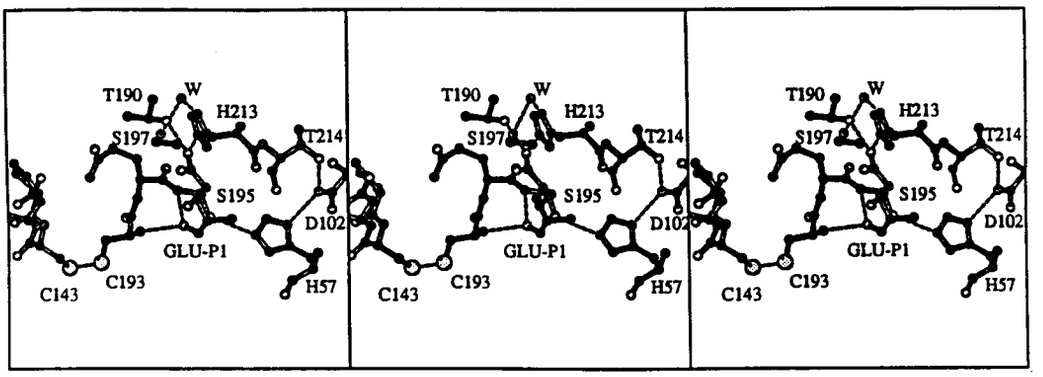

T190
W
H213
S197
T214
S195
D102
GLU-P1
C143
C193
H57
T190
W
H213
S197
T214
S195
D102
GLU-P1
C143
C193
H57
T190
W
H213
S197
T214
S195
D102
GLU-P1
C143
C193
H57

Fig. 5

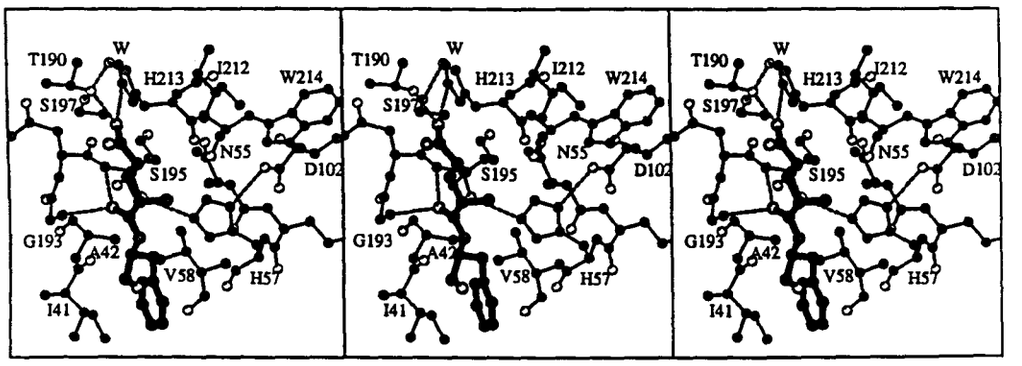


Fig. 6

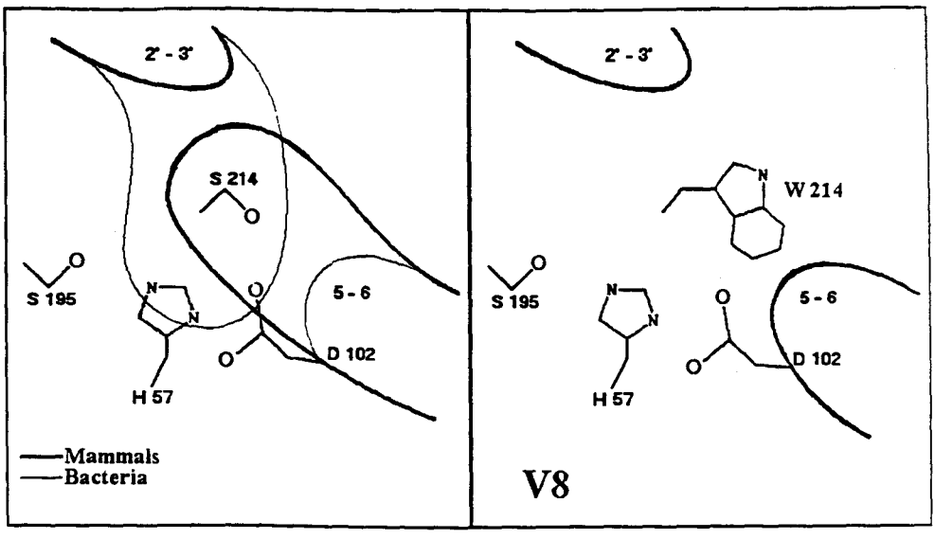

2' - 3'
S 214
S 195
5 - 6
D 102
H 57
Mammals
Bacteria
2' - 3'
W 214
S 195
5 - 6
D 102
H 57
V8

Fig. 7

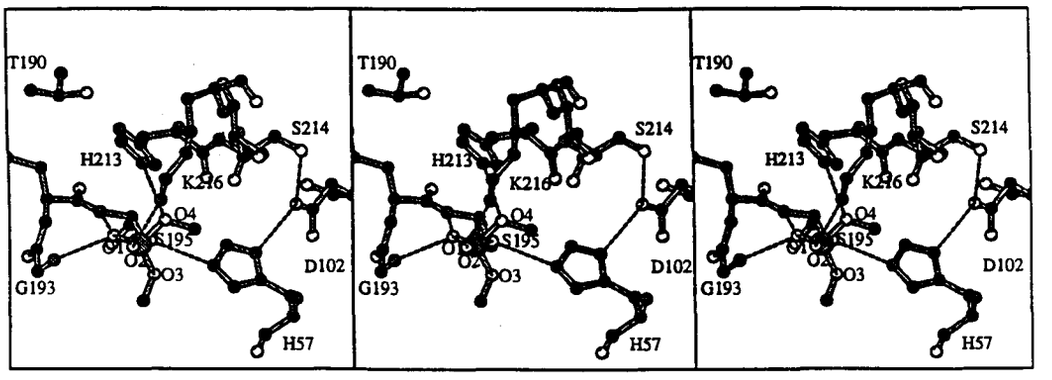


Fig. 8

Capítulo 5

α-PROLAMINAS

As prolaminas são proteínas de reserva responsáveis por aproximadamente a metade da proteína total de sementes maduras de todos os cereais importantes, com a exceção do arroz e de aveia[1]. Por isso elas determinam em grande parte o valor nutritivo e a qualidade física dos grãos e em consequência são de grande importância agronômica. Porém, sua composição incomum de aminoácidos é a fonte de deficiências nutritivas destes cereais como alimento. Em trigo, as prolaminas são os componentes principais de glúten e portanto afetam a qualidade tecnológica.

As prolaminas são definidas operacionalmente como sendo a fração da proteína total da semente solúvel em álcool[1]. Assim se distingue das albuminas (solúveis em água), glubulinas (solúveis em soluções salinas) e glutelinas (solúveis em extremos de pH). Hoje em dia se sabe que tais definições são de utilidade limitada, pois grande parte da proteína insolúvel em álcool ainda apresenta estruturas primárias parecidas (ou até mesmo idênticas) com as prolaminas. O fenômeno se deve à formação de pontes disulfeto inter-cadeia produzindo oligômeros insolúveis de alto peso molecular. Em termos evolucionários todas estas proteínas pertencem a uma única família, fato que só se tornou evidente com o sequenciamento das mesmas. Portanto, faz sentido classificar todas como prolaminas.

Dada a importância das prolaminas na determinação das propriedades nutricionais e físicas da semente, um conhecimento da sua estrutura tridimensional poderia ser de grande valor

no entendimento ou até mesmo melhoramento (através de técnicas da engenharia genética) das mesmas. Na obtenção de cristais para estudos cristalográficos existem algumas dificuldades básicas quando se trata das prolaminas, entre elas; 1) a heterogeneidade e microheterogeneidade das amostras devido à presença de muitos genes codificando produtos que diferem em poucas posições na estrutura primária; 2) a sua tendência de formar oligômeros heterogêneos e agregados e 3) a sua natureza hidrofóbica que dificulta a sua solubilizaçao adequada para cristalização.

Em milho, sorgo e *coix sp.* as prolaminas recebem os nomes triviais de zeínas, kafirinas e coixinas respectivamente e podem ser subdivididas em α,β,γ e δ baseado em solubilidade e peso molecular. As mais abundantes são as α-prolaminas, codificadas por uma família de genes[2] que em milho dão origem a dois grupos principais de moléculas chamados Z19 e Z22 conforme seu peso molecular aparente em géis de poliacrilamida[3]. A semente de coix (*Coix lacryma-jobi* var. Adlay) é mais rica em proteínas que o milho e as α-coixinas compreende quatro classes de polipeptídeo de peso molecular 27kDa, 25kDa, 17kDa e 15kDa[4] com sequências homólogas às α-zeínas, ricas em glutamina, leucina, prolina e alanina e pobres nos aminoácidos essenciais lisina e triptofano[4,5]. A maior parte das sequências consiste em repetições consecutivas de sequências semelhantes mas não-identicas[3]. O conhecimento da estrutura tridimensional desta família importante de proteínas poderia levar à possibilidade de melhorar o teor de aminoácidos essenciais da semente, através de modificações sítio dirigidas nas sequências visando a preservação da estrutura terciária e o empacotamento dentro do endosperma (Figura 5.1).

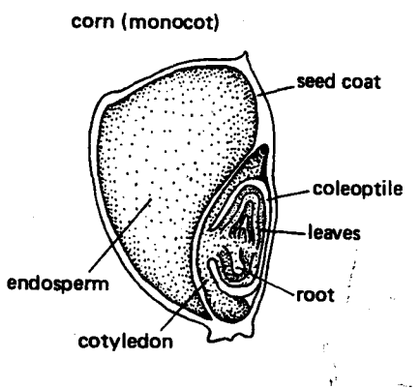


Figura 5.1 Esquema da semente de milho.

5.1 Trabalho apresentado em seguida

O único modelo estrutural proposto para as α-prolaminas apresenta certas dificuldades[3], entre elas; 1) a presença de um grande vão no meio de cada subunidade, que não é para se esperar numa molécula altamente hidrofóbica e que parece incompatível com a durabilidade da semente e 2) o modelo não se adapta bem a moléculas com menos repetições que devem ter existido em épocas passadas durante a sua evolução por

sucessivas duplicações génicas. Estas observações estimularam a investigação que levou a publicação:

Garratt, R.C., Oliva, 0., Caracelli, I. Leite, A & Arruda, P. (1993) 'Studies of the Zein-Like α-Prolamins Based on an Analysis of Amino Acid Sequences: Implications for Their Evolution and Three-Dimensional Structure', PROTEINS, Structure, Function and Genetics **15**, 88-99

Referências

[1] Tatham, A.S., Shewry, P.R. & Belton, P.S. (1990) Adv. Cer. Sci. Tech. **10**, 1-78

[2] Marks, M.D. & Larkins, B.A. (1982) J. Biol. Chem. **257**, 9976-9982

[3] Argos, P., Pederson, K., Marks, M.D., & Larkins, B.A. (1982) J. Biol. Chem. **257**, 9984-9990

[4] Ottoboni, L.M.M., Leite, A., Targon, M.L.N. Crozier, A. & Arruda, P. (1990) J. Agric. Food. Chem. **38**, 631-635

[5] Wall, J.S., (1964) Cereal Proteins in 'Proteins and Their Reactions: Symposium on Foods', Schutz, H.W., Anglemier, A. F. (eds) Westport: Avi Publ. Co., 315-341

De acordo com as políticas editoriais, estes artigos não podem ser depositados em repositório de acesso aberto. Para acesso aos artigos completos entre em contato com o(a) autor(a) ou com o Serviço de Biblioteca e Informação IFSC - USP (bib@ifsc.usp.br).

GARRATT, R. C.; OLIVA, G.; CARACELLI, I.; LEITE, A. ARRUDA, P. Studies of the zein -like 'ALFA'- prolamins based on an analysis of amino acid sequences: implications for their evolution and three- dimensional structure. **Proteins:** structure, function, and genetics, New York, v.15, p.88-99, 1993.

Capítulo 6

PREVISÃO DE ESTRUTURA SECUNDÁRIA

Em paralelo aos avanços nos métodos experimentais de sequenciamento, ressonância magnética nuclear e cristalografia, há o interesse continua em métodos teóricos para a previsão *ab initio* da estrutura terciária de proteínas a partir de suas sequências de aminoácidos. A solução do chamado 'problema de enovelamento de proteínas' representa um dos maiores desafios da biologia molecular moderna.

Ao longo dos anos mais de vinte métodos para a previsão da estrutura secundária de proteínas têm sido relatados na literatura. Os métodos originais se basearam em metodologias estatísticas empíricas[1,2] ou em critérios estereoquímicos[3]. Muitas modificações subsequentes foram feitas na tentativa de melhorar a qualidade das previsões, que ficou em torno de 60-65% quando avaliada pelo teste 'jacknife'[4]. A grande maioria dos métodos despreza interações de longo alcance, levando em conta apenas a influência da sequência local na determinação da estrutura secundária [5]. O artigo de Garratt *et al.* em anexo relata uma modificação ao método de Garnier *et al.*[2] que tenta levar em conta informação a respeito de estrutura terciária na forma de acessibilidades.

Mais recentemente, vários novos métodos baseados no uso de redes neurais de vários graus de complexidade têm sido utilizados com um certo sucesso[6-11]. O método PHD[10] por exemplo, disponível pela Internet atinge uma precisão de >70%. Provavelmente o fator principal que leva a este melhoramento é a inclusão de informação evolucionária a respeito de uma família de proteínas na forma de um alinhamento múltiplo. Tais alinhamentos carregam informação a respeito das regiões variáveis da estrutura o que ajuda melhorar a sua previsão. Vários outros métodos recentemente descritos alcançam graus de acerto similares[12-15]. Em alguns casos isto tem sido suficiente para a previsão do enovelamento tridimensional com algum sucesso[16].

6.1 O método de Garnier *et al.*

O método original de Garnier *et al.* [2,17-18] utiliza a teoria de informação para obter parâmetros de previsão de um banco de proteínas de estrutura conhecida para a sua aplicação subsequente a uma sequência alvo. A informação que um resíduo R na posição j (R_j) de uma sequência de aminoácidos carrega a respeito da estutura secundária S na posição $j+m$ (S_{j+m}) da mesma sequência é dada por

$$I(S_{j+m};R_j) = \ln [P(S_{j+m}|R_j)/ P(S_{j+m})]$$

onde $P(S_{j+m}|R_j)$ é a probabilidade do estado S na posição $j+m$, dado o resíduo R na posição j e $P(S_{j+m})$ é a probabilidade da conformação S na posição $j+m$ independente do resíduo na posição j.

Na previsão de estrutura secundária do resíduo j de uma sequência qualquer, pode-se calcular a informação carregada pela sequência local daquele resíduo para cada estado estrutural (tipicamente são definidos três estados; α-hélice, β-folha, e 'coil'). Na implementação original foram considerados como contribuintes para a conformação do resíduo j apenas os resíduos até $j\pm8$. Em cujo caso a informação a respeito do estado S é

$$I(S_j)=\sum_{m=-8}^{m=+8} I(S_j;R_{j+m})$$

onde os parâmetros $I(S_j;R_{j+m})$ são obtidos de tabelas derivadas de estruturas conhecidas. Uma vez calculada a informação carregada pela sequência de $j-8$ a $j+8$ sobre cada estado estrutural possível na posição j (α-hélice, β-folha e 'coil'), os valores são comparados e vale o estado com maior informação como a previsão de estrutura secundária. Para uma previsão completa da sequência inteira o processo é repetido para todos os resíduos.

6.2 Trabalho apresentado em seguida

Garratt, R.C., Thornton, J.M. & Taylor, W.R. (1991) 'An Extension of Secondary Structure Prediction Towards the Prediction of Tertiary Structure', FEBS Letts **280**, 141-146

Referências

[1] Chou, P.Y. & Fasman, G.D. (1974) Biochemistry **13**, 222-245
[2] Garnier, J., Osguthorpe, D.J. & Robson, B. (1978) J. Mol. Biol. **120,** 97-120
[3] Lim, V.I. (1974) J. Mol. Biol. **88**, 873-894
[4] Kabsch, W. & Sander, C. (1983) FEBS Letts **155**, 179-182
[5] Fasman, G.D. (1989) in 'Prediction of Protein Structure and the Principles of Protein Conformation' (Fasman, G.D. ed) Plenum Press, New York, pp193-316
[6] Qian, N. & Sejnowski, T.J. (1988) J. Mol. Biol. **202,** 865-884
[7] Holley, H.L. & Karplus, M. (1989) Proc. Natl. Acad. Sci. **86**, 152-156
[8] Zhang, X., Mesirov, J.P. & Waltz, D.L. (1992) J. Mol. Biol. **225**, 1049-1063
[9] Kneller, D.G., Cohen, F.E. & Lanfridge, R. (1990) J. Mol. Biol. **214**, 171-182
[10] Rost, B & Sander, C. (1993) J. Mol. Biol. **232**, 584-599
[11] Leng, B., Buchanan, B.G. & Nicholas, H.B. (1994) J. Comp. Biol. **1**, 25-38
[12] Zvelebil, M.J., Barton, G.J., Taylor, W.R. & Sternberg, M.J.E. (1987) J. Mol. Biol. **195**, 957-961
[13] Levin, J.M., Pascarella, S., Argos, P., & Garnier, J. (1993) Prot. Eng. **6**, 849-854
[14] Salamov, A.A. & Solovyev, V.V. (1995) J. Mol. Biol. **247,** 11-15
[15] Mehta, P.K., Heringa, J. & Argos, P. (1995) Prot. Sci. **4**, 2517-2525
[16] Defay, T. & Cohen, F.E. (1995) PROTEINS, Structure, Function and Genetics **23,** 431-445
[17] Robson, B. & Suzuki, E. (1976) J. Mol. Biol. **107,** 327-356
[18] Gibrat, J.- F., Garnier, J. & Robson, B. (1987) J. Mol. Biol. **198,** 425-443

De acordo com as políticas editoriais, estes artigos não podem ser depositados em repositório de acesso aberto. Para acesso aos artigos completos entre em contato com o(a) autor(a) ou com o Serviço de Biblioteca e Informação IFSC - USP (bib@ifsc.usp.br).

GARRATT, R. C.; THORNTON, J.M.; TAYLOR, W. R. An extension of secondary structure prediction towards the prediction of tertiary structure. . **Febs Letters**, Amsterdam, v.280, n.1 , p.141-146, Mar. 1991

Capítulo 7

CONCLUSÕES

A grande esperança da modelagem molecular é de que as técnicas melhorem até o ponto de poder prever estruturas tridimensionais a partir de sequências com o mesmo grau de confiança que atualmente atribuímos aos métodos experimentais de cristalografia e ressonância magnética nuclear. Ainda estamos longe deste objetivo e passamos por uma fase de aprendizagem que necessita de constante avaliação. Neste sentido, os resultados descritos nesta tese, na forma dos modelos apresentados, podem até parecer incompletos de um certo ponto de vista. Precisa-se de um teste rigoroso da qualidade dos modelos propostos. Só assim as contribuições de fatores como 1) os melhoramentos no conhecimento dos princípios básicos de estrutura protéica, 2) o crescimento no número de estruturas resolvidas, e 3) o aumento na sofisticação e automatização de programas de modelagem, podem ser avaliados com objetividade.

São muitas as maneiras de avaliar a precisão e utilidade de um modelo mas talvez a mais satisfatória seria a comparação com uma estrutura experimental determinada posteriormente. O congresso recentemente realizado em Asilomar, nos Estados Unidos, teve exatamente este propósito[1]. Um total de 43 modelos foram propostos por 13 grupos independentes de pesquisadores para 7 proteínas diferentes, cujas estruturas foram apresentadas pela primeira vez durante o congresso, após a submissão das coordenadas atômicas dos modelos[2]. O modelo proposto para a endopeptidase específica para glutamato de *Streptomyces griseus* (SGPE) no capítulo 4 desta tese pode ser sujeito a uma

avaliação objetiva similar, pois poucos meses após a publicação do nosso modelo, a estrutura cristalográfica foi resolvida[3].

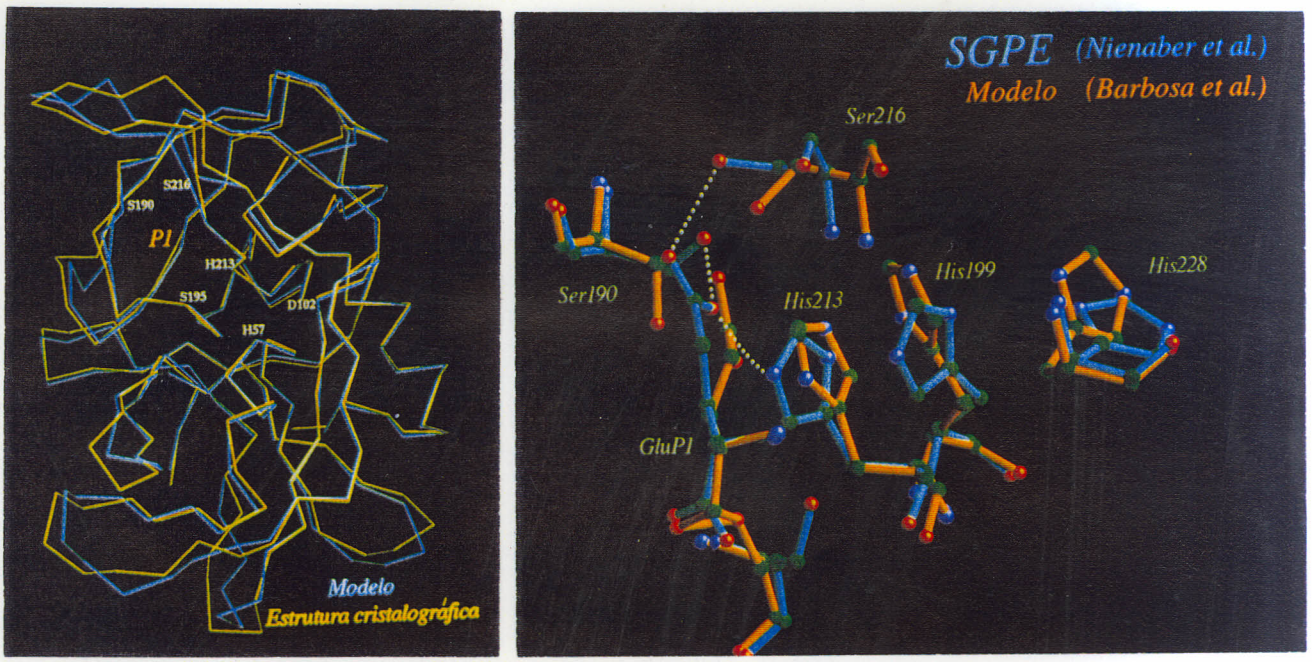


Figura 7.1 Comparação entre o modelo proposto pôr Barbosa *et al.* e a estrutura cristalográfica da endopeptidase específica para glutamato de *Streptomyces griseus*. **Esquerda**, traçado dos α-carbonos; **direita,** destaque do bolsão S1 junto com o resíduo P1 (glutamato) do substrato.

A figura 7.1 demonstra duas sobreposições entre o modelo de Barbosa *et al.* e a estrutura cristalográfica. O traçado dos C_α mostra uma excelente sobreposição dos elementos de estrutura secundária no interior da molécula mas discrepâncias maiores entre as duas estruturas nas regiões de 'loops'. A dificuldade na modelagem de tais regiões (que frequentemente correspondem às partes flexíveis da estrutura ou regiões envolvidas em contatos cristalinos) é reconhecida. Mais importante que a sobreposição global é a precisão do modelo nas regiões da estrutura mais relevantes para atividade biológica, ou seja, o sítio ativo e particularmente o bolsão de especificidade S1. A figura salienta que o rotâmero para o glutamato P1 e os três resíduos que interagem com ele (H213, S216 e S190) foram corretamente previstos pelo modelo. Entretanto, a conformação das cadeias laterais dos resíduos 190 e particularmente o rotâmero da H213 deixam algo a desejar. A modelagem de cadeias laterais continua como ponto fraco do processo, como julgado pelo levantamento dos modelos apresentados na reunião em Asilomar[2]. O erro na escolha da orientação da histidina 213 foi particularmente grave pois levou à formação da ligação de hidrogênio com o glutamato através do $N_{\delta 1}$ ao invés do $N_{\varepsilon 2}$. Em consequência, o modelo não foi bem sucedido em prever a existência de uma cadeia de três histidinas (213, 199 e 228) que levam do bolsão S1 até a hélice C-terminal. Segundo os autores, a cadeia seria responsável pela estabilização da carga negativa no glutamato P1 do substrato (Figura 7.2)[3,4].

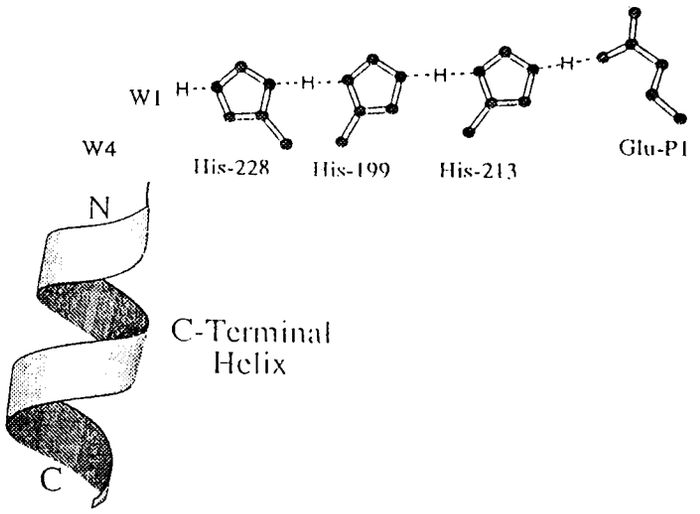


Figura 7.2 A cadeia de histidinas (triade de histidinas) descrita por Nienaber *et al.* [3] interage de um lado com o glutamato na posição P1 e, do outro lado, com o N-terminal de uma α-hélice através de duas moléculas de água.

A sobreposição do modelo com a estrutura cristalográfica leva a um desvio médio quadrático (dmq) de 1.12Å para os 187 C_α e 1.52Å para todos os átomos. Comparando com os resultados das avaliações dos modelos apresentados em Asilomar (Tabela 7.1), consta que o primeiro valor é comparável com a média de modelos baseados em estruturas de 76-77% de identidade sequencial apesar de que o grau de identidade no caso de SGPE seja de apenas 59% (com Protease A de *Streptomyces griseus*). O dmq em todos os átomos é superior à média respectiva (1.52Å contra 2.03Å) e comparável com os melhores dos modelos (1.39Å)[5], resultado notável lembrando o menor grau de identidade com a proteína base da modelagem.

Porcentagem entre modelo e estrutura base	Número de Modelos Avaliados	Número de Estruturas Incluídas na Análise	**DMQ**(Å) Valor **Médio**		**DMQ**(Å) Valor do **Melhor** Modelo	
			C_α	Todos os átomos	C_α	Todos os átomos
77%	7	1	1.00	2.03	0.53	1.39
41-42%	19	2	1.86	2.56	0.79	1.48
35%	11	1	4.28	5.38	2.85	4.00
22%	2	1	5.83	6.40	4.25	5.03

Tabela 7.1 Sumário dos resultados do teste objetivo de modelagem molecular em Asilomar[2].

Apesar da facilidade da comparação quantitativa apresentada acima, não se pode reduzir a avaliação da utilidade de um modelo à um simples valor de desvio médio quadrático. Muito mais importante é a capacidade do modelo de explicar e racionalizar resultados

anteriores e programar experimentos futuros, possibilidade que depende também do modelador e não apenas do modelo. A precisão necessária de um modelo (descrita por um simples dmq) depende da aplicação pretendida. No caso da SGPE, apesar dos erros na modelagem de algumas das cadeias laterais importantes para o reconhecimento do substrato, a *identificação* destes resíduos foi bem sucedida e a informação poderia ter sido utilizada no planejamento de experimentos de mutagênese sítio dirigida. No caso de Sm14, a identificação de epitopos provavelmente era menos sensível ainda aos eventuais erros nos detalhes da estrutura. Se a finalidade fosse, por exemplo, desvendar o mecanismo catalítico de uma enzima, maior precisão seria necessária.

Enquanto as técnicas de modelagem continuam 'sujas e rápidas', a validação dos resultados através da determinação das estruturas experimentais é imperativa. A maioria dos projetos aqui apresentados inclui também um componente experimental, sendo desenvolvido no próprio grupo de São Carlos ou através de colaborações com outras instituições. A endopeptidase específica para glutamato de *Bacillus licheniformis* já foi purificada e está em fase de ensaios de cristalização. Aguardamos amostras da toxina epidermolítica A (ETa) do Dr. Robert Evans (Guy's Hospital, Londres) com a mesma finalidade. Peptídeos de ETa escolhidos com base dos modelos das GSEs já foram sintetizados e utilizados na produção de anticorpos de alta atividade, que não apresentam reação cruzada com a toxina epidermolítica homóloga ETb. Os peptídeos têm aplicações no desenvolvimento de 'kits' para a identificação clínica das toxinas epidermolíticas.

A expressão da proteína SMYB1 de *S. mansoni* visando estudos físico-químicos e bio-químicos, inclusive cristalização, começará em breve assim como a produção do mutante G394R do lobo C-terminal de transferrina humana. Os trabalhos serão feitos através de colaborações com Drs. Sérgio Pena e Gloria Franco da UFMG e Dr. Robert Evans (Guy's Hospital) respectivamente.

O modelo de Sm14 já está sendo usado no planejamento de peptídeos para ensaios imunológicos visando aplicações como vacina e/ou em diagnóstico e os primeiros peptídeos foram produzidos pelo laboratório de Dr. Luis Juliano Neto (Universidade Federal de São Paulo). A molécula de fusão recombinante está entre os seis antígenos escolhidos pela Organização Mundial de Saúde como candidatos à vacina contra esquistossomose e está sendo testado em vários laboratórios em outros países. Simultaneamente ensaios em animais de grande porte devem começar ainda este ano na Austrália, visando a aplicação de Sm14 contra infecção de *Fasciola hepática* em gado. Portanto, existe ainda a esperança do uso de Sm14 como uma vacina bivalente contra dois parasitas de importância na saúde humana e na agronomia respectivamente.

Referências

[1] Moult, J., Pedersen, J.T., Judson, R. & Fidelis, K. (1995) PROTEINS: Structure, Function, Genetics **23(i)** , ii-iv

[2] Mosimann, S., Meleshko, R. & James, M.N.G. (1995) PROTEINS: Structure, Function, Genetics **23**, 301-317

[3] Nienaber, V.L., Breddam, K. & Birktoft, J.J. (1993) Biochemistry **32**, 11469-11475

[4] Stamato, F.M.L.G., Paulino, M., Garratt, R. Soares, C.M. & Tapia, O. (1995) Mol. Engin. **4**, 375-414

[5] Šali, A., Potterton, L., Yuan, F., van Vlijmen, H. & Karplus, M. (1995) PROTEINS: Structure, Function, Genetics **23**, 318-326