UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS

Lucas Bleicher

Implementação da análise de acoplamentos estatísticos e sua aplicação à família de proteínas tirosina
fosfatases

São Carlos – SP
2009

LUCAS BLEICHER

**Implementação da análise de acoplamentos estatísticos e sua aplicação à
família de proteínas tirosina fosfatases**

Tese apresentada ao Programa de Pós-Graduação em Física do Instituto de Física de São Carlos da Universidade de São Paulo para a obtenção do título de Doutor em Ciências.

Área de Concentração: Física Aplicada
Opção Física Biomolecular
Orientador: Prof. Dr. Igor Polikarpov

São Carlos

2009

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO, PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

Ficha catalográfica elaborada pelo Serviço de Biblioteca e Informação IFSC/USP

Bleicher, Lucas

Implementação da análise de acoplamentos estatísticos e sua aplicação à família de proteínas tirosina fosfatases./Lucas Bleicher; orientador Igor Polikarpov-- São Carlos, 2009.

167 p.

Dissertação (Mestrado – Programa de Pós-Graduação em Física - Área de concentração: Física Aplicada – opção Biomolecular) – Instituto de Física de São Carlos da Universidade de São Paulo.

1. Proteínas tirosina fosfatases. 2. Análise de acoplamentos estatísticos. 3. Arsenato redutases. 4. Cristalografia de proteínas. 5. Interleucina-22. 6. Laminarinase. I. Título.

**UNIVERSIDADE
DE SÃO PAULO**
Instituto de Física de São Carlos

Caixa:
13560-970 São
Av. Trabalhador São-car
13566-590 - São

Fone/Fax: (16) 3
www:
svposgrad@

FOLHA DE APROVAÇÃO

Lucas Bleicher

Tese apresentada ao Instituto de Física de São Carlos da Universidade de São Paulo para obtenção do título de Doutor em Ciências. Área de Concentração: Física Aplicada – Opção: Física Biomolecular.

Aprovado em: 09/03/2009

Comissão Julgadora

Prof. Dr. Igor Polikarpov

Instituição: IFSC/USP

Assinatura

Prof. Dr. Paulo Mascarello Bisch

Instituição: UFRJ

Assinatura

Prof. Dr. Munir Salomão Skaf

Instituição: UNICAMP

Assinatura

Prof. Dr. Antônio José da Costa Filho

Instituição: IFSC/USP

Assinatura

Prof. Dr. Shaker Chuck Farah

Instituição: IQ/USP

Assinatura

Agradecimentos

Aos meus pais Ervino Bleicher e Maria da Graça Bleicher, e à minha irmã Lana Bleicher por terem me estimulado desde cedo à busca pelo saber.

Ao prof. Igor Polikarpov, pela orientação.

À Patrícia de Moura e Leandra Watanabe, que ao longo de tanto trabalho envolvendo interleucinas e peroxidases, tornaram-se grandes amigas.

A Martyn Winn e Ron Keegan, pela calorosa ajuda quando da resolução da estrutura de IL-22 ligada a seu receptor IL-22R1, em que a utilização do software MrBump foi de importância crucial.

À Cecília Clementi, pela inspiração.

A José Ribamar dos Santos Ferreira Júnior, exemplo de ética e entusiasmo científico para qualquer jovem pesquisador.

À Camila Rodrigues, que mesmo conversando comigo praticamente todos os dias nos últimos cinco anos, não perde nem a paciência nem sua inesgotável fonte de novos assuntos.

A José Fernando Ruggiero Bachega, cujo estudo sobre superóxido dismutases teve importância crucial para o sucesso do presente trabalho. E que, mais que um colaborador, tornou-se um grande amigo.

A duas amigas que reencontrei em meu doutorado: Karen C. Weber, primeira amizade que fiz em São Carlos, e Cíntia Stecchini. E a uma que conheci já no finalzinho deste período, Maria Manuela Pereira Machado.

Aos velhos e bons amigos Manuel Mário Moura Nascimento, Vicente Câmara Rodrigues, Moésio Medeiros, Pedro Caram, Márcio Mesquita, Tuana Oliveira, Márcio Holanda, Juliana Campelo, Yuri Leite, Fanny Abitbol, Patrícia Landim, Luciana Braga.

Aos professores Eduardo Ernesto Castellano, Richard Charles Garratt, Luís Maurício Trambaioli da Rocha Lima, Javier Alcides Ellena, Glaucius Oliva, Munir Salomão Skaf, Antônio José da Costa Filho, Otávio Henrique Thiemann.

À Aline Faro, que tantas vezes veio às "mesas do fundo" da sala 15 para acompanhar um procedimento menos usual ou uma discussão mais acalorada, à Nathália Rodrigues, à Nayara Cavalcante e ao Alessandro Silva Nascimento. E também ao Milton Taidi Sonoda, à Victória Flório Pires de Andrade, ao Leonardo França, ao Alexandre Cassago, à Carolina de Guzzi, ao Mario de Oliveira Neto, ao Huita Couto Matozo, à Amanda Bernardes, ao Napoleão Valadares, à Ana Carolina Migliorini Figueira, à Daniela Barreto Barbosa Trivella, à Livia Sallum, ao Marcelo Liberato, ao Thiago Moda, ao Rafael Guido, à Maria Auxiliadora Morim e ao Huíta Matozo.

Aos amigos e “companheiros de república” João Renato Carvalho Muniz e Ricardo Nicolucci. E à Dona Zulmira e ao Seu Álvaro.

Aos sempre prestativos (mesmo com tanto trabalho) Luciana Lavezzo e Fernando Falvo.

À Larissa Grutes e à Izabel Cury. À Ana Luiza Valadão, companheira de alegrias e desassossegos, e à Ana Paula Montenegro.

Ao pessoal da música, em especial Flávia Bombonato, João Cordiano, Araceli Hackbarth e Ilza Zenker Leme Joly.

A alguns cientistas, de prestatividade à altura de sua competência, que muito contribuíram para minha formação em cursos e congressos: Beatriz Gomes Guimarães, Harry Powell, Jim Pflugrath, Ralf Grosse-Kunstleve, Zbigniew Dauter, Paul Adams, Tom Terwilliger, Robert Sweet, Min Yao, Andy Fitch, José Nelson Onuchic, Gábor Oszlányi.

Ao Andre Luis Bertelli Ambrosio, à Sandra Martha Gomes Dias, ao Rodrigo Villares Portugal, à Adriana Lucely Rojas, à Sandra Krauchenco, ao Ronaldo Alves Pinto Nagem e ao Mario Sanches.

Ao professor José Marcos Sasaki, maior responsável pela minha paixão pela ciência em geral e pela cristalografia em particular.

À Camila Freire Macedo, que me acompanha desde os preparativos para minha vinda à São Carlos, antes de longe, hoje sempre de muito perto, por tantas vezes ter tornado minha vida mais leve.

"There is no such thing on earth as an uninteresting subject; the only thing that can exist is an uninterested person."

(G. K. Chesterton)

RESUMO

BLEICHER, L. **Implementação da análise de acoplamentos estatísticos e sua aplicação à família de Proteínas Tirosina Fosfatases**. 2009. Tese (Doutorado) – Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2009.

A Análise de Acoplamentos Estatísticos é uma técnica computacional capaz de identificar resíduos importantes para a estrutura e função de proteínas em uma família por meio da quantificação de conservação posicional, correlação entre posições e identificação de grupos de resíduos correlacionados entre si. Neste trabalho, a análise de acoplamentos estatísticos foi implementada e aplicada ao estudo das proteínas tirosina fosfatases. Em conjunto com as proteínas tirosina quinases (PTKs), que adicionam um grupo fosforil a um resíduo de tirosina em uma proteína, as proteínas tirosina fosfatases (PTPs), que o removem, são responsáveis por diversos processos de sinalização celular. Elas são um caso de evolução convergente, onde um subgrupo (as proteínas tirosina fosfatases de baixo peso molecular) não apresenta homologia às chamadas PTPs "clássicas", capazes de defosforilar apenas resíduos de tirosina, e às fosfatases de especificidade dupla, capazes de defosforilar também resíduos de serina e treonina, além de substratos não-protéicos. Em comum, as três subfamílias apresentam apenas o motivo CX₅R, característico para todas as PTPs. Através do estudo das três subfamílias utilizando a análise de acoplamentos estatísticos, foi possível obter uma descrição detalhada de suas características conservadas e correlacionadas, relacionando-as ao conhecimento acumulado sobre proteínas tirosina fosfatases e a questões em aberto como a regulação por dimerização, a especificidade e mutações relacionadas a patologias. Foi possível também apresentar um método capaz de distinguir proteínas tirosina fosfatases de baixo peso molecular das arsenato redutases, derivadas das primeiras por evolução divergente. Adicionalmente, a técnica foi aplicada ao estudo das hexoquinases, às superóxido dismutases e às peroxidases. A tese descreve também estudos desenvolvidos pelo autor na área de cristalografia de proteínas – a determinação das estruturas da Transtirretina humana em complexo com genisteína, da holo-Hexoquinase PI de *S. cerevisiae*, do complexo IL-22/IL-22R1 e da Laminarinase de *R. marinus*.

Palavras-chave: Proteínas tirosina fosfatases. Análise de acoplamentos estatísticos. Arsenato redutases. Cristalografia de proteínas. Interleucina-22. Laminarinase.

ABSTRACT

BLEICHER, L. **Implementation of the statistical coupling analysis and its application to the Protein Tyrosine Phosphatases family**. 2009. Thesis (Doctoral) – Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2009.

The statistical coupling analysis is a computational technique which can identify important residues for the structure and function of proteins in a family by quantifying positional conservation, correlation between positions and identifying groups of self-correlating residues. Its implementation in this research group was applied to the study of the protein tyrosine phosphatases. Together with the protein tyrosine kinases (PTKs), which add a phosphoryl group to a tyrosine residue in proteins, the protein tyrosine phosphatases (PTPs), which remove it, are responsible for a variety of cell signaling processes. They are a case of convergent evolution, since one subgroup (the low molecular weight protein tyrosine phosphatases) are not homologous to the “classical” phosphatases, which can only dephosphorylate tyrosine residues, and the dual-specificity phosphatases, which can also dephosphorylate serine and threonine residues, and also non-proteinaceous substrates. All three sub-families have, in common, the CX₅R motif, a characteristic of all PTPs. By applying the statistical coupling analysis to the study of the three sub-families, it was possible to obtain a detailed depiction of their conserved and correlated characteristics, relating them to the accumulated knowledge on protein tyrosine phosphatases and open questions such as protein regulation by dimerization, specificity and disease-related mutations. It was also possible to present a method to distinguish between low molecular weight phosphatases and arsenate reductases, which are derived by the former by divergent evolution. In addition, the technique was applied to the study of hexokinases, superoxide dismutases and peroxidases. The thesis also describes studies developed by the author in the field of protein crystallography – the structure determination of human transthyretin in complex with genistein, holohexokinase PI from *S. cerevisiae*, the IL-22/IL-22R1 complex and the laminarinase from *R. marinus*.

Keywords: Protein tyrosine phosphatases. Statistical coupling analysis. Arsenate reductases. Protein crystallography. Interleukin-22. Laminarinase.

LISTA DE FIGURAS

Figura 1	Representação pictórica da relação entre distribuições de aminoácidos e valores obtidos da análise de acoplamentos estatísticos.	17
Figura 2	Obtenção de resíduos de importância funcional na família de receptores nucleares utilizando análise de acoplamentos estatísticos seguida de aglomeração hierárquica.	18
Figura 3	Organização dos elementos que participam da catálise em uma estrutura de proteína tirosina fosfatase modificada de forma a ter a capacidade de prender seu substrato.	24
Figura 4	Classificação das PTPs clássicas e sua organização estrutural.	25
Figura 5	Dois aglomerados de resíduos correlacionados em PTPs clássicas.	38
Figura 6	Aglomerado de posições correlacionadas em DSPs.	41
Figura 7	Superposição estrutural entre a arsenato redutase de <i>Bacillus subtilis</i> e a PPAC humana.	51
Figura 8	Padrões de correlação e conservação em PTPs clássicas, DSPs e LMW-PTPs.	53
Figura 9	Conservação posicional (ΔG^{stat}) em hexoquinase PI de <i>S. cerevisiae</i> .	56
Figura 10	Aglomerados de posições correlacionadas em hexoquinases.	57
Figura 11	Resultado da aglomeração hierárquica para o alinhamento de superóxido dismutases de Fe/Mn	58
Figura 12	Estrutura da peroxidase de <i>R. regia</i> .	60
Figura 13	Localização das 26 figuras mais conservadas na estrutura cristalográfica da peroxidase de <i>R. regia</i> .	62
Figura 14	Matriz de $\Delta\Delta G^{\text{stat}}$ calculada a partir do alinhamento de peroxidases, após aglomeração hierárquica e eliminação de linhas e colunas de baixo valor de $\Delta\Delta G^{\text{stat}}$.	63
Figura 15	Pseudo-alinhamento das classes obtidas a partir da análise de acoplamentos estatísticos para o alinhamento de peroxidases.	63
Figura 16	Estrutura da TTR selvagem obtida por difração de raios-X.	73
Figura 17	Conformações observadas para a genisteína nos dois sítios de ligação presentes na transtirretina.	74
Figura 18	Sítio ativo da Hexoquinase PI de <i>S. cerevisiae</i> .	77
Figura 19	Ajuste induzido na Hexoquinase PI de <i>S. cerevisiae</i> .	77
Figura 20	Complexo entre a IL-22 e a IL-22R1.	80

Figura 21	Um modelo para a unidade funcional heterotrimérica de IL-22.	81
Figura 22	Arranjo de uma célula P2 ₁ para permitir cristais geminados e emulação de uma célula C222 ₁ .	84
Figura 23	Estrutura da Laminarinase de <i>R. marinus</i> .	86
Figura 24	Detalhe dos resíduos Tyr91 e seus resíduos adjacentes.	86
Figura 25	O sítio ativo da Laminarinase de <i>R. marinus</i> .	87

LISTA DE TABELAS

Tabela 1	Motivos conservados em domínios de PTPs.	28
Tabela 2	Resíduos individualmente conservados em domínios de PTPs.	28
Tabela 3	Resíduos conservados nas duas classes obtidas do aglomerado de posições acopladas em LMW-PTPs.	43
Tabela 4	Um sumário das mutações em SHP2 que causam as síndromes de Noonan e LEOPARD.	49
Tabela 5	Posições mais conservadas em peroxidases de plantas.	61
Tabela 6	Distribuição de aminoácidos nas posições relativas às pontes dissulfeto cobertas pelo alinhamento da família de peroxidases no servidor PFAM.	64
Tabela 7	Dados cristalográficos da estrutura de transtirretina humana.	74
Tabela 8	Dados cristalográficos da estrutura de Hexoquinase PI de <i>S. cerevisiae</i> .	76
Tabela 9	Dados cristalográficos da estrutura do complexo IL-22/IL-10R2.	82
Tabela 10	Dados cristalográficos da estrutura da laminarinase de <i>R. marinus</i> .	85

LISTA DE ABREVIATURAS E SIGLAS

PDZ:	Domínio estrutural encontrado em proteínas de sinalização, cujo nome é um acrônimo para as proteínas onde foi primeiramente observado (PSD-95, DlgA e ZO-1)
GPCR:	<i>G-protein coupled receptor</i> (Receptor acoplado à proteína G)
PTP:	<i>Protein tyrosine phosphatase</i> (Proteína tirosina fosfatase)
PTK:	<i>Protein tyrosine kinase</i> (Proteína tirosina quinase)
DSP:	<i>Dual-specificity phosphatase</i> (fosfatase de especificidade dupla)
LMW-PTP:	<i>Low molecular weight protein tyrosine phosphatase</i> (Proteína tirosina fosfatase de baixo peso molecular)
RPTP:	<i>Receptor protein tyrosine phosphatases</i> (Proteína tirosina fosfatase com porção transmembrânica)
ArsC:	Arsenato redutase de classe C
SOD:	Superóxido dismutase
TTR:	Transtirretina
T4:	Tiroxina
IL-22:	Interleucina-22
IL-22R1:	Receptor 1 de IL-22
IL-22BP:	<i>IL-22 binding protein</i> (proteína de ligação à IL-22)
IL-10:	Interleucina-10
IL-10R1:	Receptor 1 de IL-10
IL-10R2:	Receptor 2 de IL-10
Rmsd:	<i>root mean square deviation</i> (desvio médio quadrático)

SUMÁRIO

1	Introdução e revisão bibliográfica	15
1.1	Análise de acoplamentos estatísticos	16
1.2	Proteínas Tirosina Fosfatases	21
2	Materiais e métodos	30
2.1	Análise de acoplamentos estatísticos	31
2.2	Ferramentas de análise estrutural	34
2.3	Protocolo de utilização da análise de acoplamentos estatísticos	35
3	Resultados	36
3.1	Proteínas Tirosina fosfatases clássicas	37
3.2	Proteínas Tirosina fosfatases de especificidade dupla	39
3.3	Proteínas Tirosina fosfatases de baixo peso molecular	41
4	Discussão	44
4.1	Interfaces de dimerização em proteínas tirosina fosfatases clássicas	45
4.2	Mutações em PTPs relacionadas a doenças	46
4.3	Especificidade em PTPs clássicas	50
4.4	Classes em LMW-PTPs e discriminação entre PTPs/ArsCs	50
4.5	Posições conservadas e correlacionadas em proteínas tirosina fosfatases	52
5	Colaborações no período em Análise de Acoplamentos Estatísticos	54
5.1	Hexoquinases	55
5.2	Superóxido dismutases	57
5.3	Peroxidases	59
6	Trabalhos desenvolvidos no período em Cristalografia de Proteínas	66
6.1	Preâmbulo: Cristalografia de proteínas – avanços e obstáculos recorrentes	67
6.2	Transtirretina humana	72
6.3	Hexoquinase PI de <i>Saccharomyces cerevisiae</i>	75
6.4	Complexo de Interleucina-22 humana e seu receptor IL-22R1	77
6.5	Laminarinase de <i>Rhodothermus marinus</i>	83
7	Conclusões	88
	Referências	93
	Anexos	103

1. Introdução e revisão bibliográfica

1.1 Análise de Acoplamentos Estatísticos

A Análise de Acoplamentos Estatísticos foi criada no grupo de pesquisa de Rama Ranganathan em 1999 (LOCKLESS; RANGANATHAN, 1999) como uma metodologia de estudo de famílias de proteínas tendo como base suas seqüências, usando descritores obtidos da Física Estatística. Ela se baseia na premissa de que a conservação em uma posição e o acoplamento de posições numa família de proteínas (representada por um alinhamento múltiplo de seqüências) são resultado de restrições evolutivas, enquanto posições que não têm papel importante na família estão livres dessas restrições e, portanto, são pouco conservadas e não apresentam correlações. Conservação e acoplamento são quantificados em parâmetros denominados ΔG^{stat} e $\Delta\Delta G^{\text{stat}}$, respectivamente. O primeiro parâmetro, calculado para uma dada posição no alinhamento, descreve o quanto a distribuição de aminoácidos naquela posição distancia-se daquela observada na natureza. Já o segundo é baseado no conceito de perturbação, em que se mede a variação de distribuição de frequências de aminoácidos numa posição quando um dado aminoácido é fixado numa outra (e.g.: $\Delta\Delta G^{\text{stat}}_{50|22=V}$ computa o quanto a distribuição de aminoácidos na posição 50 varia quando a posição 22 é ocupada por uma valina, sendo a perturbação, no caso, a imposição 22=V). Dessa forma, ele mede a correlação entre frequências de aminoácidos entre duas posições (o procedimento em detalhes do cálculo de ΔG^{stat} e $\Delta\Delta G^{\text{stat}}$ está descrito na seção Materiais e métodos). Uma representação pictórica da informação obtida pelos parâmetros é dada na Figura 1.

Neste primeiro artigo, os autores propuseram que o acoplamento estatístico entre duas posições em um alinhamento é capaz de prever o acoplamento termodinâmico entre resíduos, tal qual medido a partir de experimentos de ciclos de mutação em um domínio PDZ de PSD-95: ensaios de ligação ao peptídeo reconhecido por esta proteína utilizando espectroscopia de fluorescência foram feitos na presença de uma mutação A, na presença de uma mutação B, na presença das duas mutações e na proteína selvagem, sendo a diferença de energia $\Delta\Delta G^{\text{mut}(i,j)} = (\Delta G^i - \Delta G^{\text{wt}}) - (\Delta G^{i,j} - \Delta G^j)$ comparada ao valor de acoplamento estatístico entre as duas posições, mensurada pelo valor de $\Delta\Delta G^{\text{stat}}$. A correlação linear para estes dois parâmetros num conjunto de 14 mutações levou os autores a concluir que o acoplamento estatístico entre duas posições teria como aplicação imediata estimar o efeito de

mutações duplas em proteínas. Além disso, a observação de que resíduos estatisticamente acoplados formam “caminhos” conexos em proteínas sugeria que a técnica seria capaz de identificar as posições responsáveis pela transferência de energia ao longo de uma proteína.

Figura 1 - Representação pictórica da relação entre distribuições de aminoácidos e valores obtidos da análise de acoplamentos estatísticos. a) uma proteína arbitrária com posições *i*, *j*, *k* e *l*. b) a distribuição de aminoácidos em cada posição (V=Valina, H=Histidina, I=Isoleucina, L=Leucina, *mean* indica que a posição tem a distribuição média dos 20 aminoácidos encontrada na natureza). c) a distribuição de aminoácidos nas mesmas posições quando se impõe a perturbação *i*=H (histidina na posição *i*). Neste exemplo, as posições *i*, *j* e *l* teriam altos valores de ΔG^{stat} , pois apresentam preferência por dois ou três aminoácidos enquanto a posição *k*, com valor baixo para este parâmetro, não apresenta conservação. Aplicando-se a perturbação *i*=H, $\Delta \Delta G^{\text{stat}}_{j|i=H}$ terá um valor alto pelo fato da distribuição de aminoácidos na posição *j* variar quando se aplica tal perturbação, enquanto o valor de $\Delta \Delta G^{\text{stat}}_{k|i=H}$ e $\Delta \Delta G^{\text{stat}}_{l|i=H}$ serão baixos, pois a distribuição de aminoácidos em *k* e *l* não são alteradas quando a condição *i*=H é imposta. **Fonte:** Süel *et al.*, 2003.

A técnica foi estendida em 2003, sendo a nova metodologia descrita em (SUEL *et al.*, 2003) com a aplicação em três famílias distintas (GPCRs, serina proteases e globinas), e aplicada ao estudo da proteína G em (HATLEY *et al.*, 2003). Uma vez que é possível definir uma grande quantidade de perturbações, dado um alinhamento múltiplo de seqüências, pode-se construir uma matriz de acoplamentos estatísticos em que cada elemento *ij* contém o valor de $\Delta \Delta G^{\text{stat}}$ que quantifica a variação na distribuição de aminoácidos na posição *i* após a aplicação de uma perturbação *j*. Tais matrizes são dominadas por baixos valores – o que indica que, em geral, a maioria das posições numa família de proteínas tem distribuição de aminoácidos independente da ocorrência de determinados aminoácidos em outras posições – enquanto um subconjunto pequeno de posições apresenta correlação. Submetendo essa matriz a um procedimento de aglomeração hierárquica, verifica-se que os elementos da matriz aglomeram-se na forma de um pequeno grupo de posições auto-correlacionadas que, segundo a interpretação dos autores (SUEL *et al.*, 2003), representaria o núcleo de comunicação alostérica numa dada família de proteínas, uma vez que elas conectam as regiões importantes para sua função. A aplicação desta abordagem ao estudo da família de receptores nucleares (SHULMAN *et al.*, 2004) está mostrada na figura 2.

Figura 2 - Obtenção de resíduos de importância funcional na família de receptores nucleares utilizando análise de acoplamentos estatísticos seguida de aglomeração hierárquica. A) Matrizes de acoplamentos estatísticos ($\Delta\Delta G^{\text{stat}}$) antes (esquerda) e depois (direita) do procedimento de aglomeração hierárquica. As linhas das matrizes correspondem às posições na proteína, e as colunas às perturbações. Pontos com cores próximas ao vermelho correspondem a altos valores de acoplamento, i.e., a perturbação referente àquela coluna causa uma grande mudança em distribuição de aminoácidos na posição referente à linha. B) Sub-conjunto da matriz de acoplamentos estatísticos após a aglomeração hierárquica, contendo apenas as posições fortemente correlacionadas entre si. C) Representação das posições fortemente correlacionadas na estrutura do receptor nuclear RXR. Os resíduos identificados interligam regiões de grande importância nos receptores nucleares – a interface de dimerização, o sítio de ligação a hormônio, a interface de interação com co-ativadores e a hélice 12 (AF2). **Fonte:** Shulman *et al*, 2004.

A técnica foi aplicada também com sucesso no planejamento de proteínas artificiais. Em dois artigos numa mesma edição da revista *Nature* em 2005 (RUSS *et al.*, 2005; SOCOLICH *et al.*, 2005), o grupo de Rama Ranganathan efetuou uma análise de acoplamentos estatísticos para a família dos domínios WW e em seguida gerou alinhamentos de seqüências não existentes na natureza usando três métodos: o primeiro, usado como controle negativo, era composto por seqüências aleatórias de mesmo tamanho que um domínio WW. O segundo foi gerado de forma aleatória, mas respeitando a distribuição de aminoácidos observada em cada posição, de forma que o alinhamento final, embora composto apenas por seqüências aleatórias, continue apresentando as mesmas freqüências de aminoácidos individuais em cada posição no alinhamento contendo as seqüências observadas na natureza. Já o terceiro foi gerado aleatoriamente, mas de forma a reproduzir a matriz de $\Delta\Delta G$ s observada para a análise de acoplamentos estatísticos em domínios WW. No primeiro artigo (SOCOLICH *et al.*, 2005), mostrou-se que, enquanto as seqüências totalmente aleatórias e aquelas

cujo alinhamento conserva as frequências observadas para os domínios WW encontrados na natureza não resultam em seqüências viáveis, muitas das seqüências cujo alinhamento resultam em uma matriz de $\Delta\Delta G$ s similar àquela observada em domínios WW puderam de fato ser corretamente enoveladas, o que pode ser atestado por medição da estabilidade térmica e resolução estrutural por ressonância magnética nuclear. No segundo artigo (RUSS *et al.*, 2005), os autores mostram que além de se enovelarem corretamente, estas proteínas são capazes de exercer as mesmas funções observadas em proteínas da família encontradas na natureza – no caso, o reconhecimento específico de peptídeos ricos em prolina.

Em um trabalho mais recente, o grupo de Rama Ranganathan desenvolveu uma proteína quimérica composta de um domínio sensível à luz e um domínio catalítico, capaz de apresentar atividade catalítica dependente da luz (LEE *et al.*, 2008). Segundo os autores, o sucesso da obtenção de tal função para esta quimera se deve à justaposição de superfícies de forma a conectar seus “sítios alostéricos”, revelados pela análise de acoplamentos estatísticos.

Tal tipo de interpretação dos resultados da técnica sofreu pesadas críticas em um artigo de 2008 (CHI *et al.*, 2008), no qual a validação experimental dos dados obtidos para o domínio PDZ de PSD-95 foram feitos em maior detalhe de forma a serem confrontados com os dados publicados em 1999 (LOCKLESS; RANGANATHAN, 1999). A primeira crítica ao estudo, já reportada em 2004 por Fodor e Aldrich (FODOR; ALDRICH, 2004) e ressaltada por Chi e colaboradores (CHI *et al.*, 2008), diz respeito ao fato de que, uma vez que os pares de resíduos identificados por Lockless e Ranganathan estão próximos em estrutura, o fato de haver acoplamento energético entre eles é um resultado esperado, independente do cálculo de acoplamento estatístico. Além disso, o parâmetro físico medido (diferença entre o efeito de duas mutações) admite um único valor, enquanto o mesmo não acontece para o valor calculado – pela definição original de $\Delta\Delta G^{\text{stat}}_{ij}$, o acoplamento entre i e j não é necessariamente igual ao acoplamento entre j e i . O trabalho de Chi *et al* (CHI *et al.*, 2008) ressalta também o fato de que na interpretação de Lockless e Ranganathan, um parâmetro que diz respeito à uma família de proteínas ($\Delta\Delta G^{\text{stat}}$) é relacionado a um dado experimental referente à uma única proteína ($\Delta\Delta G^{\text{mut}}$), descartando a possibilidade de proteínas diferentes na mesma família terem

respostas diferentes às mutações efetuadas. Com efeito, os dados experimentais de Chi *et al* (CHI *et al.*, 2008) mostram que (1) o valor de acoplamento estatístico ($\Delta\Delta G^{\text{stat}}$) não é apropriado para estimar o acoplamento energético entre posições, sendo os resultados originais de Lockless e Ranganathan (LOCKLESS; RANGANATHAN, 1999) melhor explicados pela simples distância física entre as posições estudadas; (2) a análise de acoplamentos estatísticos não necessariamente está relacionada ao alosterismo em proteínas; (3) há diferenças entre $\Delta\Delta G^{\text{mut}}$ para resíduos equivalentes em proteínas diferentes (o que foi verificado pelos próprios autores ao efetuar mutações equivalentes no domínio PDZ-2 da proteína PTP-BL, o que resultou em um comportamento diferente do observado no domínio PDZ de PSD-95), invalidando a abordagem de correlacionar $\Delta\Delta G^{\text{stat}}$ (um parâmetro geral para uma família de proteínas) e $\Delta\Delta G^{\text{mut}}$ (uma medida específica para uma dada proteína da família).

Neste mesmo ano, o autor desta tese participou de um trabalho desenvolvido por Bachega e colaboradores (BACHEGA *et al.*, 2009) em que uma nova interpretação aos resultados da análise de acoplamentos estatísticos seguida de aglomeração hierárquica é apresentada. Ao aplicar a metodologia ao estudo de superóxido dismutases (SODs) de Fe/Mn, observou-se que os resíduos auto-correlacionados agruparam-se não em um, mas em três aglomerados independentes. Verificou-se que as perturbações que compunham tais aglomerados consistiam dos aminoácidos responsáveis por definir o estado oligomérico da proteína (dímero) e a preferência para o metal no sítio ativo (ferro ou manganês de acordo com o aglomerado). Tal separação entre os determinantes para estas características dentro da família de superóxido dismutases, decorrente da análise por aglomeração hierárquica a partir da matriz de $\Delta\Delta G^{\text{stat}}$ sem qualquer fornecimento prévio de tais características ao método, permitiu demonstrar que os resíduos identificados, além de, como demonstrado por Chi *et al.* (CHI *et al.*, 2008), não necessariamente corresponderem ao “núcleo de comunicação alostérica” da família de proteínas, podem ser resultado da existência de subconjuntos de proteínas, dentro da mesma família, que possuem características em comum e excludentes em relação às demais. Esta nova interpretação é perfeitamente compatível com a natureza dos dados obtidos por análise de acoplamentos estatísticos, que deveriam ser referentes não a uma proteína específica, mas refletir propriedades de toda uma família de proteínas.

Com efeito, se há numa mesma família de proteínas duas (ou mais) possibilidades para uma dada característica (como estado oligomérico, especificidade para ligantes, interação com outra macromolécula, etc.), espera-se que ela provavelmente não será determinada por um único aminoácido, mas por um conjunto deles. Se na mesma família há proteínas que tenham ou não tal característica, os aminoácidos responsáveis por ela irão necessariamente emergir da análise, visto que aparecerão em conjunto para as proteínas possuindo tal propriedade e não para as demais. Este fato força o aparecimento de um aglomerado de resíduos cuja distribuição de frequências está correlacionada entre si, com valores altos de $\Delta\Delta G^{\text{stat}}$. Estas observações eliminam os problemas apontados por Chi *et al.* (CHI *et al.*, 2008), ao mesmo tempo em que permitem novas aplicações da técnica.

No estudo sobre superóxido dismutases de ferro e manganês, esta nova interpretação para a análise de acoplamentos estatísticos permitiu a descoberta de novos resíduos, não descritos na literatura, responsáveis pela determinação de estado oligomérico e preferência por metal no sítio ativo em superóxido dismutases. A validação destes resíduos contra uma base de dados de superóxido dismutases com tais propriedades já determinadas por métodos experimentais mostrou que esta metodologia, que parte apenas de um alinhamento múltiplo de seqüências, tem poder preditivo comparável a métodos que necessitam de informações estruturais e de “treinamento” a partir de seqüências cujo estado oligomérico e especificidade por metal já são conhecidos, como o procedimento descrito recentemente pelo grupo de Marianne Rooman (WINTJENS *et al.*, 2008). Esta nova abordagem foi utilizada para o estudo das proteínas tirosina fosfatases de baixo peso molecular e das peroxidases, apresentadas no presente trabalho.

1.2 Proteínas tirosina fosfatases

As tirosinas fosfatases (PTPs, do inglês *protein tyrosine phosphatases*) são proteínas capazes de remover um grupo fosforil de resíduos de tirosina em proteínas. Juntamente com as proteínas tirosina quinases (PTKs, do inglês *protein tyrosine kinases*), de função inversa (fosforilam resíduos de tirosina), elas fazem parte de uma vasta gama de redes de sinalização celular. Enquanto as PTKs vêm

sido estudadas desde a década de 70 (uma revisão sobre PTKs pode ser encontrada em (HUBBARD; TILL, 2000)), o estudo das PTPs sofreu um atraso de aproximadamente dez anos – as primeiras purificações de PTPs foram realizadas em 1988 (PALLEN; PANAYOTOU *et al.*, 1988; PALLEN; SAHLIN *et al.*, 1988; TONKS *et al.*, 1988), e a primeira clonagem em 1990 (GUAN *et al.*, 1990). Talvez por este motivo, as PTKs estejam hoje bem mais caracterizadas e compreendidas que as PTPs.

Além disso, enquanto as PTKs são todas provenientes de um ancestral comum (GU; GU, 2003) e consideradas como uma única família de proteínas, as PTPs são geralmente divididas em três subfamílias: as PTPs “clássicas”, que defosforilam apenas resíduos de tirosina; as PTPs de especificidade dupla (também chamadas de DSPs, do inglês *dual specificity phosphatases*), capazes também de defosforilar resíduos de serina, de treonina, e substratos não-protéicos como lipídios e carboidratos; e as PTPs de baixo peso molecular (LMW-PTPs, do inglês *low molecular weight protein tyrosine phosphatases*), às vezes chamadas também de PTPs de classe II. Enquanto as PTPs clássicas e as DSPs têm mesmo enovelamento à exceção dos elementos de estrutura secundária não ordenados, as LMW-PTPs possuem um enovelamento completamente diferente das demais, apesar de apresentar a mesma função. É importante ressaltar que enquanto as PTPs clássicas e as DSPs evoluíram de um ancestral comum (ALONSO *et al.*, 2004), a comparação com as LMW-PTPs indicam um caso de evolução convergente, visto que estas têm em comum com as demais apenas o motivo característico de PTPs, um *loop* composto pela sequência CX₅R, comumente chamado de *P-loop*. Apesar de haver grande variação entre os resíduos no segmento X₅ entre as diferentes subfamílias de PTPs, a conformação deste *loop* é estritamente conservada, sendo facilmente superposto em estruturas diferentes de PTPs, resultando em desvios entre os carbonos- α menores que 1 Å (TABERNERO *et al.*, 2008). A catálise em PTPs necessita, além dos resíduos de cisteína e de arginina neste motivo, de um ácido que, em PTPs clássicas e DSPs está num motivo chamado *WPD-loop*, e em LMW-PTPs, num motivo denominado *DPYY-loop*. Nos dois casos, estes motivos estão distantes em sequência mas próximos em estrutura do *P-loop*, e curiosamente sua posição em sequência varia para os dois casos: em PTPs clássicas e DSPs, o *WPD-loop* encontra-se na direção N-terminal em relação ao *P-loop*, com o inverso ocorrendo em LMW-PTPs. Um terceiro loop de importância para a catálise ocorre em PTPs

e DSPs: o *Q-loop*, caracterizado por um resíduo de glutamina responsável por posicionar uma molécula de água catalítica no sítio ativo.

Entre as características comuns a todas as PTPs, está também a sensibilidade ao vanadato, a capacidade de hidrolisar p-nitrofenil fosfato, e um mecanismo de reação independente de metais no sítio ativo (ZHANG, 2002). No genoma humano há 107 PTPs identificadas, sendo 38 clássicas e 61 DSPs (ALONSO *et al.*, 2004).

Na reação catalisada por uma PTP, a cisteína do *P-loop* age como nucleófilo, atacando o grupo fosfato enquanto a arginina no mesmo *loop* tem papel na ligação ao substrato e estabilização do intermediário de reação. Enquanto isso, o *loop* WPD (DPYY em LMW-PTPs) se fecha sobre o substrato, com o grupo aspartato agindo como ácido geral no primeiro passo da catálise, quando o oxigênio fenólico da tirosina é protonado para clivar o fosfato da tirosina formando um intermediário cisteína-fosfato. No segundo passo, o aspartato age como base coordenando uma molécula de água para promover a hidrólise do intermediário. Um resíduo de glutamina altamente conservado também está envolvido na coordenação de uma molécula de água no sítio (TIGANIS; BENNETT, 2007).

A melhor compreensão estrutural da interação entre substrato e proteína foi obtida com mutantes capazes de “prender” o substrato, geralmente trocando a cisteína catalítica por uma serina ou alanina e/ou o aspartato do *loop* WPD por uma alanina. Tal abordagem tornou possível manter os valores de K_m enquanto se reduz consideravelmente os valores de k_{cat} , permitindo a cristalização de proteínas com substratos ligados (FLINT *et al.*, 1997). Um exemplo de estrutura obtida por essa abordagem, capaz de revelar a organização espacial dos elementos da catálise, está mostrada na figura 3.

Figura 3 - Organização dos elementos que participam da catálise em uma estrutura de proteína tirosina fosfatase modificada de forma a ter a capacidade de prender seu substrato (Código PDB: 1G1G). A figura mostra, em verde, a fosfatase PTP1B humana, ligada a um peptídeo contendo um resíduo fosfatado de tirosina (em azul claro). O *P-loop*, onde se encontram a cisteína e a arginina catalítica, está mostrado em amarelo (deve-se ressaltar que ao invés de uma cisteína, a figura mostra uma alanina iniciando a sequência CX₅R, modificação necessária para manter o substrato intacto no sítio ativo). À esquerda, vê-se um resíduo de ácido aspártico provindo do *loop* WPD e a glutamina do *Q-loop* que posiciona no sítio a molécula de água, mostrada em vermelho, que participa da reação.

As PTPs clássicas costumam ser sub-classificadas de acordo com critérios de identidade, organização de domínios e presença ou ausência de região transmembrânica. Uma revisão recente de Hendriks *et al.* (HENDRIKS *et al.*, 2008) apresenta um painel geral de proteínas tirosina fosfatases de acordo com estes aspectos, como mostrado na Figura 4.

Figura 4 - Classificação das PTPs clássicas e sua organização estrutural. Os representantes mostrados na primeira linha são transmembranares, porém mecanismos de processamento alternativo de RNA podem fazer com que um único gene codifique isoformas diferentes, em alguns casos com presença e ausência de região transmembranar para o mesmo gene (caso dos representantes da subfamília R7). Além dos domínios PTP, representados por D1, D2 e PTP (os domínios PTP reduzidos ou sem atividade são mostrados em azul), muitas proteínas apresentam também uma série de domínios diferentes: BRO1 (homólogo ao BRO de Baculovírus), CA (anidrase carbônica), Cad (caderina), CRB (homólogo à proteína celular de ligação à retinaldeído), FERM (homólogo a ezrina/radixina/moesina), FN (fibronectina tipo III), HD (domínio His), Ig (homólogo a imunoglobulina), KIM (motivo de interação a quinase), KIND (domínio semelhante ao lobo N de quinase), MAM (homólogo a meprina/A2/RPTP μ), PDZ (homólogo a proteína da densidade pós-sináptica 95/ZO1), Pro (sequência rica em prolina), SH2 (homólogo a src2). **Fonte: Hendriks et al, 2008.**

Uma vez que fosfatases possuem função oposta às quinases e estas possuem representantes oncogênicos, a descoberta das fosfatases levou à idéia de que elas poderiam funcionar como supressores de tumores (HENDRIKS *et al.*, 2008). Há pelo menos um caso bem descrito que confirma esta hipótese: a fosfatase de especificidade dupla PTEN apresenta tal característica, atribuída à sua capacidade de defosforilar lipídios (PULIDO; HOOFT VAN HUIJSDUIJNEN, 2008). Uma análise mutacional de larga escala em casos de câncer no cólon revelou mutações em seis proteínas tirosina fosfatases de classe I (clássicas): PTPRF, PTPRG, PTPRT, PTPN3, PTPN13 e PTPN14, sendo a PTPRT o caso mais comum de mutação. Além disso, a reintrodução desta fosfatase em células cancerosas inibiu o crescimento celular (WANG *et al.*, 2004). Por outro lado, outro estudo buscando mutações em PTPRT em diferentes tipos de câncer resultou em poucos casos encontrados (LEE *et al.*,

2007). Há também várias evidências indicando que a DEP-1, codificada pelo gene PTPRJ e também chamada de RPTP η , tenha atividade supressora de tumores (RUIVENKAMP *et al.*, 2002; RUIVENKAMP *et al.*, 2003; IULIANO *et al.*, 2004; LESUEUR *et al.*, 2005; VAN PUIJENBROEK *et al.*, 2005; LUO *et al.*, 2006). Esta fosfatase é objeto de estudo deste grupo de pesquisa, tendo sido caracterizada quanto ao seu estado oligomérico (MATOZO *et al.*, 2007) e com sua estrutura cristalográfica sendo atualmente determinada (MATOZO, NASCIMENTO *et al.*, 2009).

A regulação das PTPs se dá por diversos mecanismos, tendo sido descritos a expressão diferencial, o processamento alternativo de RNA, a clivagem proteolítica, a oxidação, modificações pós-traducionais como fosforilação e oxidação, a ligação a outras proteínas e a dimerização (DEN HERTOOG *et al.*, 2008). A regulação por dimerização recebeu considerável atenção na segunda metade dos anos 90 quando da resolução das estruturas tridimensionais obtidas por difração de raios-x da RPTP α (BILWES *et al.*, 1996) e da RPTP μ (HOFFMANN *et al.*, 1997). A primeira apresenta um dímero em que os sítios ativos estão ocluídos devido à própria formação da interface de dimerização, enquanto a segunda apresenta um arranjo dimérico diferente, em que os sítios ativos encontram-se expostos ao solvente. Desta forma, foram propostas duas possibilidades para a regulação por dimerização em fosfatases: as fosfatases com interface do “tipo α ” seriam capazes de se auto-inibir através da dimerização, enquanto este mecanismo não ocorreria em fosfatases com interfaces do “tipo μ ”. Já a compreensão do mecanismo de regulação por oxidação teve grandes avanços mais recentemente, quando se descobriu que, em PTP1B, a oxidação da cisteína catalítica não resulta num derivado de ácido sulfênico (estável), mas num anel sulfenil amida envolvendo um resíduo adjacente de serina (SALMEEN *et al.*, 2003; VAN MONTFORT *et al.*, 2003). Esta espécie não apenas é resistente à posterior oxidação (que seria provavelmente irreversível) como pode ser prontamente reduzida novamente para a forma livre da cisteína, cataliticamente ativa, podendo então consistir num mecanismo eficiente de ativação/inativação de acordo com a presença ou ausência de agentes oxidantes no meio.

A principal característica estrutural que diferencia as PTPs clássicas das PTPs de especificidade dupla é a profundidade da cavidade onde se encontra o sítio ativo. Enquanto em PTPs

clássicas ela mede aproximadamente 9Å, em DSPs a profundidade é de aproximadamente 6Å. A profundidade maior em PTPs clássicas leva à seletividade por substratos protéicos contendo fosfotirosina (TABERNERO *et al.*, 2008). Uma diferença estrutural no *P-loop* de DSPs foi observada para um representante desta subfamília: A estrutura cristalográfica da MKP3 apresenta um sítio ativo distorcido em que a cisteína (Cys293) e a arginina (Arg299) catalíticas estão desalinhadas, e o ácido do *loop* WPD (Asp262) encontra-se distante do sítio ativo (STUCKEY *et al.*, 1994). Enquanto entre PTPs é relativamente comum a ocorrência de um domínio inativo ou de baixa atividade em fosfatases contendo dois domínios PTP (D1 e D2), em DSPs foram descritas duas proteínas inativas: a STYX e a Sbf1, cuja função é ainda desconhecida (CAMPS *et al.*, 2000). DSPs são também caracterizadas por alta variação na região amino-terminal, cuja função pode estar relacionada à localização subcelular, uma vez que motivos relacionados a esse papel foram descritos para a região N-terminal da VH-3 humana (ISHIBASHI *et al.*, 1994), da VH-6 de rato (MOUREY *et al.*, 1996) e da DSP B59 (SHIN *et al.*, 1997), com a localização destas proteínas compatível com os motivos encontrados.

As proteínas tirosina fosfatases de baixo peso molecular (LMW-PTPs) caracterizam-se, além de seu diminuto tamanho (aproximadamente 18 kDa) e ausência de domínio transmembranar, por não possuírem especificidade por tecidos quanto à sua expressão (RAMPONI; STEFANI, 1997). Em humanos há um único gene para LMW-PTPs, para o qual foram encontrados quatro diferentes RNAs mensageiros obtidos por processamento alternativo de RNA, dos quais dois resultam em isoformas ativas, denominadas IF1 e IF2 (MODESTI *et al.*, 1998). Assim como as PTPs clássicas e DSPs, as LMW-PTPs podem ser reguladas por oxidação. Porém, o mecanismo proposto para esta classe (RAUGEI *et al.*, 2002) é diferente em relação às demais: ao invés da simples oxidação da cisteína catalítica com a formação de um sulfenil, a presença de uma cisteína adicional no loop CX₅R em boa parte das fosfatases de baixo peso molecular, incluindo a humana, permite a formação de uma ponte dissulfeto intramolecular, cuja existência foi confirmada pelos citados autores por espectroscopia de massa. Desta forma, a fosfatase estaria protegida do excesso de oxidação, ao mesmo tempo em que poderia voltar rapidamente à forma ativa através da redução da ponte formada. Embora estudos tenham sido descritos sugerindo que as LMW-PTPs são capazes de defosforilar fosfo-serina e fosfo-

treonina, na forma livre ou como resíduos de proteínas (OSTANIN *et al.*, 1995; ZHANG *et al.*, 1995), tais atividades catalíticas ocorrem em uma ou mais ordens de grandeza abaixo da atividade em fosfotirosina, o que torna questionável seu papel fisiológico.

Uma forma de se investigar as regiões importantes para a estrutura e a função de uma família de proteínas é a identificação de *motivos conservados* através de um alinhamento múltiplo de seqüências. Andersen e colaboradores (ANDERSEN *et al.*, 2001) utilizaram esta abordagem para PTPs clássicas, descrevendo dez motivos através da análise de um alinhamento de 113 seqüências, interpretando-os em termos de importância estrutural (como a formação do núcleo hidrofóbico da proteína), reconhecimento do substrato e catálise. Estes motivos estão mostrados na tabela 1.

Tabela 1 - Motivos conservados em domínios de PTPs (numeração utilizada: PTP1B). Resíduos sublinhados e em negrito ocorrem em todas as 151 fosfatases analisadas pelos autores. Resíduos apenas em negrito estão conservados em pelo menos 90% das seqüências, e resíduos sem formatação são conservados em pelo menos 80% das seqüências. **Fonte: Andersen, Mortensen *et al.*, 2001.**

Motivo e posição	Padrão conservado	Papel proposto
#1: 40-46	NXX(K/R)NRY	Reconhecimento de pTyr
#2: 53-59	DXXR(V/I)XL	Fita- β conservada
#3: 65-70	DYINA(N/S)	Núcleo estrutural
#4: 81-87	(F/Y)(I/V)AX QGP	Contorna o <i>P-loop</i>
#5: 91-101	TXXD FWX (M/L/V)X(W)(E/Q)	Hélice- α conservada
#6: 107-111	(I/L/V)(V/I) MXT	Núcleo hidrofóbico
#7: 120-126	KCXX YWP	Núcleo hidrofóbico
#8: 176-185	(Y/E)XX WPD XG XP	<i>WPD-loop</i>
#9: 210-223	PXX(V/D) HCSAGXGR (T/S) G	<i>P-loop</i>
#10: 261-269	(V/I/L)QTXXQYXF	<i>Q-loop</i>

Além dos motivos conservados, os autores supracitados identificaram também um conjunto de posições isoladamente conservadas, propondo o papel de cada um desses resíduos de forma a justificar sua alta conservação, como visto na Tabela 2.

Tabela 2 - Resíduos individualmente conservados em domínios de PTPs (numeração utilizada: PTP1B). A formatação utilizada para definir porcentagem de conservação é a mesma da tabela anterior. **Fonte: Andersen, Mortensen *et al.*, 2001.**

Posição	Aminoácido	Papel proposto
19	E	Define a estrutura da hélice $\alpha 2'$
115	E	Ligação de hidrogênio conservada com o resíduo Arg221
156	R	Define a estrutura da fita $\beta 10$
169	R	Define a estrutura da fita $\beta 11$
192	L	Define a estrutura da hélice $\alpha 3$
254	R	Ligação de hidrogênio com o <i>P-loop</i>
257	R	Ligação de hidrogênio com o <i>P-loop</i> , diminuindo o pK _a da cisteína catalítica

Informações adicionais podem ser obtidas através da identificação de pares acoplados de resíduos, i.e., duas posições em uma família de proteínas cuja distribuição de frequências de aminoácidos é altamente correlacionada (como poderia acontecer, por exemplo, em duas posições que possam formar pontes dissulfeto para parte dos membros da família) ou grupos de resíduos correlacionados, i.e., um conjunto de posições cuja distribuição de frequências de aminoácidos é dependente da ocorrência de certos aminoácidos em outros membros do grupo. A metodologia para tal tipo de investigação encontra-se na *análise de acoplamentos estatísticos*, descrita na seção 1.1. Uma vez que as três subfamílias de PTPs (clássicas, DSPs e LMW-PTPs) são muito bem representadas em bancos de dados de seqüências de proteínas, elas tornam-se apropriadas para serem estudadas à luz desta técnica, sendo este o objetivo principal deste trabalho.

2. Materiais e métodos

2.1 Análise de acoplamentos estatísticos

A análise de acoplamentos estatísticos tem como única entrada um *alinhamento múltiplo de seqüências*, grande o suficiente para permitir que os dados obtidos da análise tenham relevância estatística, e diversificado de forma a representar os diversos organismos em que a família estudada tenha membros presentes.

Dentre as formas de obter um alinhamento inicial, destacam-se duas: aquela que parte de uma única proteína a partir da qual são obtidas homólogas através de uma busca contra bancos de dados gerais de seqüências de proteínas, e aquela em que já se parte de um banco de dados específico para a família de proteínas estudada. A primeira abordagem é feita utilizando ferramentas como o PSI-Blast (ALTSCHUL *et al.*, 1997), que é capaz de identificar seqüências homólogas contra bancos de dados de seqüências de proteínas, submetidas em seguida a um procedimento de alinhamento múltiplo. A segunda abordagem é utilizada neste trabalho e utiliza o servidor PFAM (FINN *et al.*, 2008), que mantém um banco de dados atualizado regularmente de seqüências de proteínas já alinhadas e organizadas em famílias. A qualidade na construção de alinhamentos disponíveis no banco de dados PFAM é atestada, por exemplo, pelo sucesso em utilizá-los diretamente para obtenção de resíduos próximos em estrutura, como descrito por Fodor e Aldrich (FODOR; ALDRICH, 2004). É importante ressaltar, porém, que embora esta abordagem seja mais prática, ela necessita ainda de uma intervenção manual, uma vez que os alinhamentos incluem todos os produtos de seqüenciamento, incluindo redundância e fragmentos.

Para PTPs clássicas, DSPs e LMW-PTPs, foram utilizados os alinhamentos provenientes do servidor PFAM de códigos PF00102, PF00782 e PF01451, respectivamente. Após a retirada de seqüências redundantes e fragmentos, os alinhamentos finais continham respectivamente 479, 686 e 919 seqüências.

Uma vez obtido o alinhamento, quantifica-se *conservação posicional* e *acoplamento entre duas posições*, como mostrado a seguir:

A conservação (ΔG_i^{stat}) de um sítio i num alinhamento múltiplo de seqüências (MSA, do inglês *multiple sequence alignment*) de uma família é definida como:

$$\Delta G_i^{\text{stat}} = kT^* \sqrt{\sum_x \left(\ln \frac{P_i^x}{P_{\text{MSA}}^x} \right)^2} \quad (1)$$

A conservação, por essa definição, tem unidade de energia, sendo definida por analogia à formulação termodinâmica para diferença de energia entre dois estados numa distribuição de Boltzmann, onde o estado fundamental corresponderia a um sítio cuja distribuição de aminoácidos é proporcional àquela encontrada na natureza. A constante kT^* (onde k é a constante de Boltzmann) é arbitrária e não é utilizada nas análises, o somatório em x corresponde aos 20 possíveis aminoácidos, e P_i^x é a probabilidade binomial de se encontrar o número observado de aminoácidos no número de proteínas analisado:

$$P(x) = \frac{N!}{n_x!(N-n_x)!} p_x^{n_x} (1-p_x)^{N-n_x} \quad (2)$$

N é o número total de seqüências, n_x é o número de seqüências que apresentam o aminoácido x ($x=\text{ALA}, \text{CYS}$, etc.) na posição em questão, e p_x é a freqüência observada na natureza para o aminoácido x . P_{MSA}^x é a probabilidade de encontrar o número observado de aminoácidos em todo o alinhamento. Uma vez que a distribuição de aminoácidos numa dada família pode ser consideravelmente diferente daquela observada na natureza (e.g.: famílias de proteínas que comumente apresentem pontes dissulfeto provavelmente terão uma quantidade maior de cisteínas, um dos aminoácidos de mais baixa freqüência na natureza), utiliza-se como p_x a freqüência de aminoácidos observada no alinhamento. Desta forma, o valor de ΔG^{stat} será alto quando a distribuição de aminoácidos naquela posição for não usual, com freqüências distantes da esperada. É importante ressaltar que, apesar dos artigos originais de R. Ranganathan apresentarem os valores de ΔG^{stat} em unidades de energia, com a constante kT mostrada explicitamente, valores de ΔG^{stat} relativos a famílias de proteínas diferentes não devem ser comparados entre si, uma vez que a utilização da

probabilidade binomial resulta em valores de ΔG^{stat} mais altos quando alinhamentos maiores são utilizados. Uma alternativa ao cálculo de ΔG^{stat} foi proposta por Dima e Thirumalai (DIMA; THIRUMALAI, 2006), utilizando-se, ao invés da probabilidade binomial, apenas a frequência de aparecimento de um dado aminoácido na posição em questão sobre a sua frequência em todo o alinhamento. Neste trabalho, utilizou-se a definição original de R. Ranganathan (LOCKLESS; RANGANATHAN, 1999), baseada na probabilidade binomial.

O acoplamento estatístico entre duas posições é definido como:

$$\Delta\Delta G_{i,j}^{\text{stat}} = kT * \sqrt{\sum_x \left(\ln \frac{P_{i|\delta j}^x}{P_{MSA|\delta j}^x} - \ln \frac{P_i^x}{P_{MSA}^x} \right)^2} \quad (3)$$

Este parâmetro é calculado ao se fazer uma *perturbação* (i.e.: fixar um dado tipo de aminoácido numa posição) numa posição j , verificando-se o efeito desta na distribuição de aminoácidos na posição i . O símbolo $i|\delta j$ significa “sob o efeito da perturbação em j ” e refere-se ao subconjunto de seqüências que apresentam esta perturbação (assim, $P_{MSA|\delta j}^x$ é refere-se é definido como P_{MSA}^x , mas apenas com as frequências observadas nas seqüências em que a perturbação ocorre). Portanto, cálculos de $\Delta\Delta G^{\text{stat}}$ envolvem a construção de um sub-alinhamento, em que apenas tais seqüências são utilizadas. A ocorrência de altos valores de $\Delta\Delta G^{\text{stat}}$ em i após uma perturbação em j , portanto, significa que a distribuição de aminoácidos em i varia quando a restrição em questão é imposta em j , indicando um acoplamento estatístico entre as duas posições. Assim como descrito para a conservação posicional (ΔG^{stat}), o valor de acoplamento estatístico pode também ser calculado sem a utilização da probabilidade binomial (DIMA; THIRUMALAI, 2006).

Para validar a significância estatística de candidatos a perturbação adota-se o seguinte procedimento, descrito no material suplementar do artigo de Süel e colaboradores (SUEL *et al.*, 2003): tomam-se as cinco posições com mais baixo valor de ΔG^{stat} e monitora-se seu valor quando seqüências aleatórias são retiradas do alinhamento. Uma vez que posições sem conservação devem ter pressão evolutiva muito baixa, seu valor de ΔG^{stat} deve ser sempre próximo de zero. Ao retirar seqüências aleatórias do alinhamento, os valores deste parâmetro crescem após a retirada de quantidade suficiente

de seqüências, devido ao fato de o alinhamento não mais possuir variabilidade amostral para refletir a baixa conservação destas posições. Deve-se, portanto, escolher como tamanho mínimo para uma perturbação um valor mais baixo que aquele para o qual o valor de ΔG^{stat} começa a aumentar, i.e., uma perturbação válida é aquela para a qual o número de seqüências que obedece ao seu critério (e.g. “triptofano na posição 74”) tem no mínimo este valor. Os valores mínimos para perturbações válidas em PTPs clássicas, DSPs e LMW-PTPs, obtidos por este método, foram de 25%, 29% e 25% do alinhamento original, respectivamente.

Definido o número p de perturbações válidas, uma matriz $p \times n$ (onde n é o número de posições no alinhamento) é construída de forma que cada elemento ij da matriz contenha o valor de acoplamento ($\Delta\Delta G^{\text{stat}}$) da posição i devido à perturbação j . Esta matriz é então submetida a agrupamento hierárquico, organizando de forma conjunta posições que tenham perfis de acoplamento similares (i.e., que estão correlacionadas com as mesmas posições com amplitudes similares para cada par de posições). O agrupamento hierárquico é efetuado utilizando-se como distância entre linhas ou colunas a métrica de Manhattan (a distância entre dois vetores é o somatório dos módulos das diferenças entre cada par de coordenadas), com a ligação entre aglomerados feita pelo método de ligação completa (a distância entre dois aglomerados é definida como a maior distância existente entre um membro do primeiro aglomerado e um membro do segundo).

Eliminando as regiões em que não há acoplamento significativo entre posições e mantendo as posições fortemente correlacionadas, chega-se a um (em alguns casos mais de um) subconjunto de resíduos, que deve ser submetido à análise crítica sob a luz das características conhecidas a respeito da família de proteínas em questão.

2.2 Ferramentas de análise estrutural

Superposições espaciais entre estruturas de proteínas foram feitas utilizando o algoritmo SSM (do inglês “Secondary Structure Matching”), tal qual implantado no programa Superpose no pacote de programas de cristalografia CCP4 (KRISSINEL; HENRICK, 2004). Este método sobrepõe estruturas cristalográficas usando como critério a superposição de elementos de estrutura secundária, sendo

assim bastante apropriado para sobrepor estruturas cujo alinhamento de sequência é não-trivial, ou que contenham grandes diferenças quanto aos elementos não ordenados de estrutura secundária.

A obtenção de resíduos que participam de interfaces de dimerização foi feita utilizando o servidor PISA (KRISSINEL; HENRICK, 2007), que leva em conta a formação de ligações de hidrogênio, pontes salinas, pontes dissulfeto, área enterrada devido à dimerização e uma estimativa da diferença de energia livre devido à formação do dímero.

2.3 Protocolo de utilização da análise de acoplamentos estatísticos

Com base nos parâmetros descritos previamente, o protocolo padrão para o uso da análise de acoplamentos estatísticos segue os seguintes passos:

1. Obter um alinhamento de seqüências para a família, retirando fragmentos e duplicatas.
2. Calcular as freqüências de aminoácidos para cada posição e sua conservação posicional (ΔG^{stat}).
3. Determinar as perturbações possíveis, calcular as freqüências para cada posição nos sub-alinhamentos gerados, e em seguida uma matriz de acoplamentos estatísticos ($\Delta\Delta G^{\text{stat}}$) com base na diferença entre as freqüências.
4. Eliminar auto-correlações ($\Delta\Delta G^{\text{stat}}$ entre mesmas posições) na matriz de acoplamentos estatísticos, e submetê-la à aglomeração hierárquica para linhas e colunas utilizando como distância a métrica de Manhattan e o método de ligação completa.
5. Eliminar linhas e colunas de baixo sinal.

3.1 Proteínas tirosina fosfatases clássicas

A quantificação da conservação posicional (ΔG^{stat}) em PTPs clássicas teve como aplicação imediata a atualização da definição de motivos conservados. Uma vez que a definição de Andersen e colaboradores (ANDERSEN *et al.*, 2001) foi feita com base num conjunto de apenas 113 seqüências, nossa análise tornou possível uma descrição mais detalhada dos motivos conservados, devido à maior amostragem. Uma vez que nosso estudo parte de uma grande quantidade de dados genômicos, incluindo PTPs putativas, a maior parte das percentagens descritas por Andersen e colaboradores (ANDERSEN *et al.*, 2001) é reduzida, mesmo para resíduos previamente descritos como invariantes. Nosso alinhamento inclui, por exemplo, seqüências desprovidas da cisteína catalítica, um reflexo do fato de o alinhamento ser populado também por domínios de PTPs intrinsicamente inativos. O mesmo acontece para o *loop* WPD, que contém um aspartato necessário para a catálise por agir como ácido na reação. Há no alinhamento seqüências que apresentam nesse *loop* a alternativa WPE, que mantém a característica do resíduo catalítico, mas também resíduos não-ácidos nesta posição em algumas dezenas de seqüências. Desta forma, ao contrário de estudos anteriores descrevendo motivos e posições conservadas, não utilizaremos o termo *invariante* para descrever posições em PTPs, uma vez que em nenhum dos casos são encontradas posições com 100% de conservação para um dado aminoácido.

A análise de acoplamentos estatísticos a partir do alinhamento de PTPs clássicas levou à identificação de dois aglomerados de posições acopladas, como mostrado na figura 5. A razão do aparecimento do primeiro aglomerado, mostrado em vermelho, pode ser prontamente interpretada – ele contém resíduos estritamente necessários para um domínio ativo de fosfatase. O resíduo Y950 (a numeração utilizada para PTPs clássicas é a da rPTP η , exceto quando indicado) é parte do *loop* de reconhecimento de pTirosina, D1084 é o ácido necessário para a catálise no *loop* WPD e que se liga a K1027 quando da ligação ao substrato, A1120 e T1125 são resíduos do *loop* catalítico e Q1162 é a asparagina responsável por posicionar uma molécula de água que participa da reação.

Figura 5 - a) Os dois aglomerados de resíduos correlacionados em PTPs clássicas. A ocorrência de um resíduo em um dos aglomerados é acompanhada por um aumento de frequência para todos os outros resíduos no mesmo aglomerado. b) Mesma informação mostrada em a, após uma rotação de 180° em torno do eixo y. A numeração utilizada é a da fosfatase rPTP η .

O segundo aglomerado é mais disperso que o primeiro, e é causado por correlações entre resíduos que pertencem ou estão muito próximos aos motivos identificados por Andersen e colaboradores (ANDERSEN *et al.*, 2001). A glicina G993, por exemplo, que pertence ao motivo IAXQGP e está presente em 60% das seqüências, aumenta a frequência de treoninas na posição 1018, pertencente ao motivo IVMXT, de 53% para 75%. A posição 1129, que apresenta 43% de isoleucinas e 29% de cisteínas, está próxima em seqüência do *loop* catalítico. A presença de isoleucinas nesta posição aumenta a taxa de alaninas na posição 990 de 57% a 87%, e de metioninas na posição 1160 de 48% para 85%. Apesar de estas três posições estarem distantes em seqüência, elas estão próximas na estrutura. Além disso, a posição 990 pertence a um motivo conservado no núcleo da proteína, em volta do sítio catalítico, enquanto a M1160 precede imediatamente o motivo chamado de *Q-loop*, que contém a asparagina que posiciona uma molécula de água no sítio ativo para a catálise. Estes dados sugerem que este segundo aglomerado contenha resíduos cuja importância seja mais estrutural que funcional.

Também é possível encontrar pares de resíduos que mostram alta correlação entre si, mas que não são sensíveis a presença de outros aminoácidos, o que resulta em um alto valor de $\Delta\Delta G^{\text{stat}}$ que, na matriz submetida ao procedimento de aglomeração, mantém-se “isolado”. Como exemplo, as posições consecutivas 1007 e 1008 apresentam forte correlação, devido ao fato que a presença de triptofanos na posição 1007 aumenta de 57% para 85% devido à presença de um ácido glutâmico na posição 1008. Enquanto isso, a presença de uma glutamina ao invés do ácido glutâmico nesta posição é compatível

com a ocorrência de resíduos alternativos na posição 1007 como fenilalaninas (13%), tirosinas (14%) e leucinas (12%).

3.2 Proteínas tirosina fosfatases de especificidade dupla

Através do cálculo de ΔG^{stat} para proteínas tirosina fosfatases de especificidade dupla (DSPs), foi possível determinar resíduos e motivos conservados para esta subfamília, comparando-os aos motivos já descritos para as PTPs clássicas. O aspecto mais interessante desta análise foi o fato de, ao comparar os motivos encontrados por superposição estrutural, alguns motivos e resíduos conservados encontrados em DSPs equivalem estruturalmente àqueles observados em PTPs clássicas, mas com aminoácidos diferentes.

Para os resíduos 214-218 (a numeração utilizada para DSPs corresponde à proteína DUSP6) há um motivo L(Y/F)LG(N/S) que sobrepõe-se a cinco resíduos do motivo (F/Y)IATQGP (#4, pela definição descrita previamente). Na região estruturalmente equivalente ao longo motivo #5 TXXDFWX(M/L/V)X(W)(E/Q) das PTPs clássicas, as DSPs conservam apenas a posição 228, com 66% de leucinas. Há certa semelhança para o motivo #6, (I/L/V)(V/I)MXT em PTPs clássicas e equivalente a um motivo (I/V)(L/I)NT(T/A) em DSPs, que ainda é imediatamente precedido pela sequência conservada (I/V)X(H/A) – havendo fraca conservação para as posições equivalentes nas PTPs clássicas.

Diferenças consideráveis podem ser verificadas no *loop* WPD: enquanto em PTPs clássicas ele é descrito como (Y/F)XXWPDXGVP, em DSPs há apenas o padrão (I/V)P(V/I)XD. O único elemento em comum é o ácido aspártico, o que por sua vez é um resultado esperado por ele ser essencial para o mecanismo de reação.

Há também diferenças no *loop* catalítico: enquanto em PTPs clássicas ele é descrito como PXX(V/I)HCSAGXGR(T/S)G, em DSPs o motivo conservado é VLVHCXAG(V/I)SRS.

Finalmente, o *Q-loop* (V/I/L)QTXXQYXF em PTPs clássicas também possui um equivalente estrutural em DSPs, com padrão também diferente: PNXXFXXQL. Embora a região apresente tais

diferenças em resíduos conservados, as estruturas holo das DSPs CC14B (código no PDB: 1OHE) e VHR (1J4X) mostram que, assim como em PTPs clássicas, uma glutamina deste *loop* está orientada de forma a posicionar uma molécula de água no sítio ativo (capaz de ser vista no mapa de densidade eletrônica para a CC14B).

Pode-se descrever também um motivo YLM em DSPs para os resíduos 307-309 com ocorrência de 80%, 69% e 70% para tais resíduos, enquanto os equivalentes estruturais em PTPs clássicas (1132-1134) apresentam 51% de cisteínas, 21% isoleucinas e 22% de leucinas, respectivamente.

Quanto a resíduos isoladamente conservados, a posição 210 em DSPs, assim como na equivalente 979 em PTPs clássicas, é dominada por isoleucinas e valinas. Porém, a posição 255 apresenta 47% de tirosinas enquanto a equivalente em PTPs clássicas (1076) apresenta 48% de histidinas e 38% de glutaminas. Essa mudança em conservação é observada também para a posição 272, que apresenta 58% de fenilalaninas e 22% de leucinas enquanto a equivalente estrutural em PTPs clássicas (1095) é dominada por leucinas (58%), com baixa frequência de fenilalaninas (6%). A posição 319 é altamente conservada em DSPs com 78% de alaninas, enquanto nas clássicas há 24% de treonina e 23% de isoleucinas. A posição 324 é dominada pelos resíduos básicos lisina (41%) e arginina (28%) enquanto em PTPs clássicas (posição 1154) há quase exclusividade de argininas (90%). Já a posição 327 apresenta exatamente o mesmo comportamento de seu equivalente em PTPs clássicas (1157), sendo dominada por argininas (84% em DSPs, 95% em clássicas).

A análise de acoplamentos estatísticos seguida de aglomeração hierárquica usando dados do alinhamento de DSPs gerou um único aglomerado, contendo seis posições altamente acopladas a oito perturbações. Este aglomerado também conecta diferentes regiões com importância funcional (a serina 298 precede a arginina catalítica no motivo CX₅R, a prolina 332 e a glicina 335 estão no *Q-loop*) e outras posições vistas nos motivos conservados, como a isoleucina 279, a metionina 309 e a fenilalanina 272. Os resíduos neste aglomerado podem ser vistos na figura 6.

Figura 6 - Aglomerado de posições correlacionadas em DSPs. Assim como em PTPs clássicas, a ocorrência de um resíduo no aglomerado é acompanhada por um aumento de frequência para todos os outros resíduos mostrados.

3.3 Proteínas tirosina fosfatases de baixo peso molecular.

A principal característica verificada para a análise de proteínas tirosina fosfatases de baixo peso molecular (LMW-PTPs) é a ocorrência de longas seqüências de resíduos conservados. Como esperado, os motivos conservados em LMW-PTPs não têm qualquer correlação com aqueles observados em PTPs clássicas ou DSPs, visto que as LMW-PTPs possuem enovelamento completamente diferente. O motivo mais longo observado possui 19 posições consecutivas e se localiza no N-terminal da proteína. Ele inclui o *loop* catalítico contendo o padrão CX₅R, sendo X₅ também razoavelmente conservado e podendo ser descrito como X₂N(I/N)X considerando um corte de 80% de conservação. A asparagina presente na posição +3 em relação à cisteína catalítica tem 96% de conservação e tem importante função estrutural no sítio ativo (TABERNERO *et al.*, 2008), adotando uma conformação de hélice de mão esquerda estabilizada por ligações de hidrogênio com os resíduos Ser19, Ser43 e His72 (a numeração para LMW-PTPs corresponde à fosfatase PPAC), que apresentam respectivamente 96%, 97% e 36% de conservação, respectivamente. O resíduo que precede a arginina catalítica é uma cisteína para 67% das seqüências de LMW-PTPs. Há evidências em favor da formação de uma ponte dissulfeto entre esta cisteína e a cisteína catalítica (RAUGEI *et al.*, 2002), o que poderia estar relacionado à regulação por redox da atividade de proteínas tirosina fosfatases de baixo peso molecular. A presença desta cisteína adicional aumenta consideravelmente quando da presença de uma isoleucina na posição 16, uma prolina na posição 20 e uma tirosina na posição 132,

aumentando sua frequência para 99%, 91% e 95%, respectivamente. Foi mostrado que esta tirosina presente na posição 132 pode ser fosforilada na PPAC humana, mas sem diferença em atividade. Porém, a fosforilação da tirosina imediatamente anterior (Tyr131) na mesma proteína resulta em um aumento de 25% na atividade da enzima. A fosforilação da tirosina 132, apesar de não afetar a atividade, pode ter importância funcional devido à formação de um motivo (pYXNX) reconhecido pelo domínio SH2 de Grb2 (BUCCIANINI *et al.*, 1999).

Enquanto a análise de aglomerados para PTPs clássicas e DSPs mostrou que eles apresentam um pequeno conjunto de resíduos que tendem a estar presentes simultaneamente para todos os membros do aglomerado, as LMW-PTPs apresentaram um comportamento totalmente diverso: em primeiro lugar, o número de posições correlacionadas é muito grande (46 no total); em segundo, estas posições têm duas opções possíveis para conservação, particionando as LMW-PTPs em duas “classes”. Escolhendo as duas perturbações auto-exclusivas com os maiores valores de $\Delta\Delta G^{\text{stat}}$ em relação às outras posições (C92 e D92), é possível verificar as diferenças de conservação para cada uma das posições, como mostrado na tabela 3.

Tabela 3 - Resíduos conservados nas duas classes obtidas do aglomerado de posições acopladas em LMW-PTPs. Resíduos marcados com uma adaga estão presentes na PPAC humana, enquanto resíduos marcados com um asterisco estão presentes em ArsC de *B. subtilis*.

Posição	Classe I	Classe II
11	V [†]	L [*]
13	L [†]	T [*]
16	I [†]	S [*]
17	C [†]	A/C [*]
20	P [†]	Q [*] /I
26	F [†]	L
46	T [†]	<i>ausente</i>
52	G [†]	<i>não-conservada</i>
55	P [†]	V
56	D [†]	N [*] /H
57	<i>não-conservada</i>	P [*]
63	A/L	L/M [*]
70	L/I	I [*] /T
74	A [†]	S [*]
75	R [†]	K
78	<i>não-conservada</i>	D [*]
81	D [†]	<i>não-conservada</i>
85	F [†]	<i>não-conservada</i>
87	L/Y [†]	L [*] /F
88	I [†]	V [*]
89	L [†]	I/V [*]
90	A	T [*]
91	M [†]	V/L [*]
92	D [†]	C [*]
93	<i>não-conservada</i>	D
95	N [†]	A [*]
96	L [†]	A [*]
98	D [†]	E
99	L [†]	C
100	C/P	<i>não-conservada</i>
107	<i>não-conservada</i>	G
112	K [†]	<i>não-conservada</i>
113	I [†] /L/V	H
114	<i>não-conservada</i>	W
115	L [†]	G
119	F/Y [†]	P
129	D [†]	<i>não-conservada</i>
130	P [†]	<i>não-conservada</i>
131	Y [†]	<i>não-conservada</i>
132	Y [†]	<i>não-conservada</i>
139	E [†]	R
142	L/Y [†]	R [*]
144	L	E [*]
145	I/L/V	I [*] /L
148	A	R [*]
149	C [†]	I/V

4.1 Interfaces de dimerização em proteínas tirosina fosfatases clássicas

A estrutura cristalográfica de RPTP α , obtida em 1996 (BILWES *et al.*, 1996), mostrou um interessante arranjo no qual os sítios ativos apresentavam-se oclusos devido à dimerização. Desde então, a dimerização tem sido proposta como um mecanismo de regulação – PTPs com uma interface de dimerização do “tipo α ” apresentariam auto-inibição, enquanto um dímero ativo, com os sítios expostos, também era possível, como visto na interface totalmente diferente presente na estrutura de RPTP μ , determinada um ano depois (HOFFMANN *et al.*, 1997). Embora a conservação de resíduos nestas interfaces possa ser um indício para obter o provável estado de oligomerização de uma PTP, procedimento efetuado por Matozo e colaboradores (MATOZO *et al.*, 2007), existem exceções – a PTP κ , por exemplo, conserva resíduos na interface do tipo μ , mas não apresenta um homodímero com a mesma organização espacial. O comportamento particular com respeito ao aparecimento dessas duas possíveis interfaces deveria ser visto nos resultados da análise de acoplamentos estatísticos – se, por exemplo, a maior parte das PTPs apresentasse ou uma interface do tipo α ou do tipo μ , dois aglomerados deveriam emergir da análise, cada um deles contendo os resíduos responsáveis pelos contatos cruciais entre as unidades monoméricas em cada interface, como observado na análise de acoplamentos estatísticos de superóxido dismutases (BACHEGA *et al.*, 2009).

Um aglomerado contendo resíduos da interface de dimerização em PTP α é de fato um dos resultados de nossa análise para PTPs clássicas. Y950, D1084, A1120 e Q1162 são estruturalmente equivalentes a resíduos nesta interface de dimerização, e a presença de um deles é acompanhada de um aumento na frequência dos demais. Porém, como previamente descrito, eles também são elementos chave na função de PTPs mesmo em proteínas monoméricas, descartando, portanto, a interface de dimerização como razão do aparecimento do aglomerado.

Dezessete resíduos, dois deles formando pontes salinas inter-moleculares, estão presentes na interface de dimerização de RPTP μ . Uma baixa correlação entre eles já é esperada devido à ausência de um aglomerado de posições que os contenha. Com efeito, a análise de perturbações válidas nesta região mostrou não haver alta correlação, mesmo para os dois resíduos que formam a ponte salina

(R1054 e E1028, pela numeração de RPTP μ). A presença de uma arginina na posição 1054 não é seguida de um aumento relevante em resíduos ácidos na posição 1028, nem a presença de um ácido glutâmico nesta posição aumenta a ocorrência de resíduos básicos na posição 1054.

Além disso, a estrutura cristalográfica de RPTP κ apresenta um dímero cristalográfico diferente daquele observado em RPTP μ mesmo que os resíduos-chave na interface desta proteína sejam conservados e ambas pertençam à subclasse R2A de tirosinas fosfatases. Dímeros cristalográficos diferentes podem ser observados mesmo em PTPs praticamente idênticas, caso da PTPJ humana (códigos PDB 2NZ6 e 2CFV) e seu homólogo em rato rPTP μ (Matozo, Nascimento *et al.*, em preparação).

Tomados em conjunto, a atual informação estrutural sobre dímeros de PTPs e os dados de análise de acoplamentos estatísticos sugerem que a formação de dímeros não é estritamente uma opção entre “tipo α ” e “tipo μ ” – pelo contrário, diferentes possibilidades de homodímeros ativos e inativos podem existir.

4.2 Mutações em PTPs relacionadas a doenças

Algumas doenças estão bem documentadas como relacionadas a mutações em genes que codificam PTPs. O gene PTPN11, que codifica a fosfatase SHP2, tem descritas várias mutações relacionadas às síndromes de Noonan e LEOPARD, causadas pelo mal funcionamento da SHP2. A síndrome de Noonan se caracteriza pela má formação do coração, baixa estatura, problemas de aprendizagem, peito escavado e coagulação inapropriada (CURCIC-STOJKOVIC *et al.*, 1978), enquanto a síndrome de LEOPARD é caracterizada por manchas vermelhas na pele, anomalias cardíacas, alargamento ocular, estreitamento da artéria pulmonar, anomalias na genitália, crescimento retardado e surdez (GORLIN *et al.*, 1969). Sendo este gene aquele com mais mutações deletérias descritas na literatura, as quais se distribuem em várias regiões diferentes da proteína codificada, ele torna-se um caso-teste ideal para ser estudado à luz da análise de acoplamentos estatísticos, verificando como os dados de conservação e correlação provenientes desta técnica podem estar relacionados a proteínas com função prejudicada.

As posições para as quais havia registro de mutações deletérias foram avaliadas de acordo com sua conservação posicional (observando as percentagens observadas para os aminoácidos presentes na proteína selvagem e na mutada), efeitos estruturais/funcionais (localização em posições essenciais para catálise ou para formação de elementos estruturais) e correlação com outras posições do alinhamento. Como esperado, a maioria das mutações observadas estão em regiões bem conservadas, nos motivos descritos para PTPs clássicas, tornando óbvias as razões para o mal funcionamento das proteínas resultantes, como se pode observar na tabela 4. Apesar disso, alguns casos não são descritos como mera violação de uma restrição de conservação numa dada posição.

Um caso interessante são as mutações na posição N308 da SHP-2: este resíduo é parte de um motivo no núcleo estrutural da proteína e é altamente conservado, com asparaginas e serinas presentes cada uma em 46% das seqüências. A ocorrência do fenótipo da síndrome de Noonan, portanto, poderia ser explicado pelo rompimento deste núcleo estrutural, exceto pelo fato de que o fenótipo está presente não apenas nas mutações N308D e N308T, mas também em N308S, que está de acordo com o motivo conservado. Porém, a análise de acoplamentos estatísticos mostra que a escolha entre asparaginas e serinas não é totalmente independente de outras posições. Quando os efeitos das perturbações altamente correlacionadas a essa posição e presentes em SHP2 são avaliados, verifica-se que eles aumentam a preferência por asparaginas nesta posição. A posição 518, por exemplo, não é parte de nenhum motivo conservado, mas a presença de uma isoleucina nesta posição (o que ocorre em SHP2) aumenta a frequência de asparaginas na posição 308 de 46% a 65%.

Caso semelhante ocorre na posição M508V, também relacionada à síndrome de Noonan. Esta posição apresenta conservação moderada, com 48% de metioninas. Porém, a preferência por metioninas nesta posição aumenta quando uma alanina está presente na posição 283 (67%), uma tirosina na posição 327 (59%) e uma valina na posição 509 (65%), todas elas presentes em SHP2.

Na posição 282, há moderada preferência por isoleucinas (44%) e valinas (36%), o que sugeriria que a mutação I282V não seria incompatível com o padrão de conservação nesta posição. Porém, esta posição está altamente correlacionada com a presença de uma leucina na posição 283,

uma tirosina na posição 327 e uma valina na posição 509, cada uma elevando a preferência por isoleucinas na posição 282 para 56%, 53% e 53%, respectivamente, e todas elas presentes em SHP-2.

Apenas as posições 411 e 506 apresentam mutações relacionadas às síndromes de Noonan e Leopard que não podem ser associadas à modificação de um resíduo com alta conservação, uma posição correlacionada a resíduos específicos presentes em SHP2 ou uma mutação que teria efeito claro na estrutura ou função da proteína (caso da mutação G507R, em que uma glicina próxima do sítio ativo é trocada por um resíduo volumoso e carregado, ou mutações no resíduo P495, que inicia uma α -hélice).

Um sumário das mutações deletérias no domínio catalítico da SHP2 e sua relação com conservação posicional, correlação inter-posicional ou importância estrutural/funcional encontra-se na tabela 4.

Tabela 4 - Um sumário das mutações em SHP2 que causam as síndromes de Noonan e LEOPARD. A interpretação dos motivos da inviabilidade destas proteínas através dos dados de análise de acoplamentos estatísticos ou de informações estruturais/funcionais é apresentada, assim como as referências em que tais mutações foram primeiramente descritas.

Mutação	Síndrome	Detalhes	Referência
Y279S	LEOPARD	Loop de reconhecimento de tirosina fosfatada	(SARKOZY <i>et al.</i> , 2004)
Y279C	Noonan		(TARTAGLIA <i>et al.</i> , 2002)
I282V	Noonan	Dominada por I and V. Correlacionada a outros resíduos em SHP-2.	(TARTAGLIA <i>et al.</i> , 2001)
F285L	Noonan	Dominada por resíduos fenólicos	(TARTAGLIA <i>et al.</i> , 2002)
F285S	Noonan		(TARTAGLIA <i>et al.</i> , 2002)
F285C	Noonan		(TARTAGLIA <i>et al.</i> , 2006)
N308D	Noonan	Núcleo estrutural, estritamente N ou S. Correlacionada a outros resíduos em SHP-2.	(TARTAGLIA <i>et al.</i> , 2001)
N308T	Noonan		(ZENKER <i>et al.</i> , 2004)
N308S	Noonan		(TARTAGLIA <i>et al.</i> , 2002)
I309V	Noonan	Dominada por resíduos fenólicos.	(TARTAGLIA <i>et al.</i> , 2002)
T411M	Noonan	Não conservada, não correlacionada.	(BERTOLA <i>et al.</i> , 2004)
A465T	LEOPARD	Alanina conservada no <i>loop</i> catalítico.	(YOSHIDA <i>et al.</i> , 2004)
G468A	LEOPARD	Glicina conservada no <i>loop</i> catalítico.	(SARKOZY <i>et al.</i> , 2004)
T472M	LEOPARD	Subseqüente ao <i>loop</i> catalítico, dominada por T.	(DIGILIO <i>et al.</i> , 2002)
P495L	Noonan	Inicia uma α -hélice.	(BINDER <i>et al.</i> , 2005)
P495S	Noonan		(ZENKER <i>et al.</i> , 2004)
R502L	LEOPARD	90% de conservação, faz ligação de hidrogênio com o motivo CX ₅ R.	(SARKOZY <i>et al.</i> , 2004)
R502W	LEOPARD		(SARKOZY <i>et al.</i> , 2004)
R505K	Noonan	95% de conservação, responsável por diminuir o pKa da cisteína catalítica.	(TARTAGLIA <i>et al.</i> , 2002)
S506T	Noonan	Não conservada, não correlacionada.	(MAHESHWARI <i>et al.</i> , 2002)
S506A	Noonan		(KRATZ <i>et al.</i> , 2005)
S506L	Noonan		(BERTOLA <i>et al.</i> , 2006)
G507R	Noonan	Próxima à cisteína catalítica	(TARTAGLIA <i>et al.</i> , 2006)
M508V	Noonan	Moderadamente conservada, correlacionada a outros resíduos em SHP-2.	(TARTAGLIA <i>et al.</i> , 2001)
Q510P	LEOPARD	Glutamina que posiciona uma molécula de água no sítio ativo.	(CONTI <i>et al.</i> , 2003)
Q514E	Noonan	Parte do <i>Q-loop</i>	(TAKAHASHI <i>et al.</i> , 2005)
Q514P	LEOPARD		(KEREN <i>et al.</i> , 2004)
Q514R	Noonan		(BERTOLA <i>et al.</i> , 2005)

4.3 Especificidade em PTPs clássicas

O conhecimento acumulado até a atualidade sobre PTPs focando-se em seus substratos considera que, enquanto PTPs clássicas têm baixa especificidade, as DSPs podem ter ação extremamente específica (ZHANG, 2002). Há, porém, mecanismos de seletividade sugeridos para PTPs clássicas. Peters e colaboradores (PETERS *et al.*, 2004) propuseram que as diferenças em especificidade para a PTP1B e a PTP α estão relacionadas principalmente à uma região composta pelos resíduos 47, 48, 258 e 259 (numeração relativa à PTP1B). A posição 47 tem conservação muito baixa, sendo populada por todos os aminoácidos possíveis com exceção da cisteína, enquanto a posição 48 é dominada por ácidos aspárticos e asparaginas, sendo as duas possibilidades perturbações válidas para a análise de acoplamentos estatísticos. A posição 258 é populada por todos os aminoácidos possíveis, porém com uma pequena preferência por prolinas (24%), enquanto a posição 259 possui todos os possíveis aminoácidos exceto valinas, com uma preferência moderada por glicinas (29%), uma perturbação válida para a análise de acoplamentos estatísticos. De acordo com Peters e colaboradores (PETERS *et al.*, 2004), o resíduo 259 é o mais importante para a seletividade, visto que uma mutação Q259G em PTP α é capaz de fazer com que ela tenha especificidade semelhante à da PTP1B.

A perturbação D48 leva apenas a um aumento muito baixo de ocorrência de resíduos básicos na posição 47, e nenhuma correlação à posição 259. A perturbação N48 também tem correlação muito baixa aos resíduos 47 e 259, e o único efeito da perturbação G259 nestas posições é o aumento da ocorrência de lisinas na posição 47 de 16% para 29%. Estes resultados sugerem que os mecanismos de especificidade derivados de casos particulares (neste caso, PTP1B vs. PTP α) não podem ser facilmente generalizados para diferentes proteínas na família.

4.4 Classes em LMW-PTPs e discriminação entre PTPs/ArsCs

A análise de acoplamentos estatísticos mostrou que algumas posições, agrupadas devido à aglomeração hierárquica, têm dois padrões de conservação, definindo duas “classes” de LMW-PTPs, um resultado que levanta questões interessantes. Uma das classes contém o motivo DPYY (129-132), que age como o *loop* WPD das PTPs clássicas e de especificidade dupla, sendo um elemento essencial

para a função como fosfatase. Na segunda classe, os resíduos 129-132 são menos conservados – a posição 129 tem apenas 16% de ácidos aspárticos, sendo mais populada por glicinas (40%), a taxa de prolina na posição 130 cai para apenas 14%, enquanto as posições 131 e 132 são praticamente desprovidas de tirosinas. Dado que esta segunda classe possui conservação para uma grande quantidade de posições em outras regiões, é altamente provável que estes resíduos apresentando co-evolução sejam responsáveis por outra função que não a defosforilação de tirosinas em proteínas, pois um motivo essencial para esta função está muito melhor representado na primeira classe.

A estrutura cristalográfica da arsenato redutase de *Bacillus subtilis*, determinada a 1.6Å de resolução em 2001 (BENNETT *et al.*, 2001), é nitidamente similar à das LMW-PTPs, como visto na figura 7. Arsenato redutases da família ArsC também apresentam sítios catalíticos e mecanismo de reação similares aos das PTPs de baixo peso molecular, sugerindo que elas possam ter evoluído das LMW-PTPs (SHI *et al.*, 1998). Os mesmos autores levantaram a questão sobre a distinguibilidade de PTPs e ArsCs, uma vez que elas possuem em comum motivos de sequência essenciais. Uma solução para possíveis erros em anotação genômica foi sugerida também por Shi e colaboradores (SHI *et al.*, 1998): a análise de contexto genômico poderia ser usada para discriminar fosfatases de arsenato redutases, sendo reportado um caso de anotação errônea – a mt1355 de *M. thermoautotrophicum* foi anotada como “potencial arsenato redutase”, mas provavelmente é uma PTP devido à ausência de membros do operon *ars* perto de sua própria sequência.

Figura 7 - Superposição estrutural entre a arsenato redutase de *Bacillus subtilis* (código PDB: 1JL3), mostrada em amarelo, e a PPAC humana (código PDB: 5PNT), mostrada em verde. O valor de desvio médio quadrático para 115 carbonos alfa estruturalmente equivalentes é de 1.2Å.

Embora seja verdade que o padrão de sequência utilizado para buscas por PTPs em genomas, o motivo CX₅R, também esteja presente em arsenato redutases, Shi e colaboradores (SHI *et al.*, 1998) não consideraram a hipótese de que outros motivos, não necessariamente essenciais para a catálise, pudessem ser utilizados para distinguir PTPs de arsenato redutases. A análise de acoplamentos estatísticos de LMW-PTPs feita neste trabalho sugere fortemente que tal distinção possa efetivamente ser feita. Como descrito anteriormente, nossos resultados indicam que o alinhamento utilizado é claramente dividido em duas classes – há algumas posições que mostram alta conservação, mas com variação nos aminoácidos conservados em cada posição de acordo com a classe em questão. Ao verificar a presença dos resíduos de cada classe contra uma LMW-PTP e uma ArsC (a PPAC de *H. sapiens* e a ArsC de *B. subtilis*), torna-se claro que a maior parte dos resíduos da classe 1 está presente em PPAC, enquanto a maioria dos resíduos da classe 2 está presente em ArsC, como mostrado na tabela 3. Desta forma, acreditamos que os padrões de conservação mostrados nesta tabela podem ser usados como elementos de busca para anotação genômica, uma aplicação extremamente útil dos resultados da análise de acoplamentos estatísticos.

4.5 Posições conservadas e correlacionadas em proteínas tirosina fosfatases

Através da mensuração de conservação posicional e acoplamento estatístico entre posições, é possível prover uma descrição detalhada dos motivos conservados e suas inter-correlações em proteínas tirosina fosfatases. Uma lista de motivos conservados já havia sido descrita por Andersen e colaboradores (ANDERSEN *et al.*, 2001). Devido à maior amostragem utilizada neste trabalho e à análise de acoplamentos estatísticos, é possível apresentar diagramas para as três famílias de PTPs que não apenas indiquem os motivos conservados para cada grupo, mas também a correlação entre pares de posições (Figura 8). Juntamente com a informação obtida da análise por aglomeração hierárquica (figuras 5 e 6) e a definição das classes para LMW-PTPs (tabela 3), têm-se um retrato detalhado das particularidades em sequência das proteínas tirosina fosfatases.

Figura 8 - Padrões de correlação e conservação em (A) PTPs clássicas, (B) DSPS e (C) LMW-PTPs. Resíduos sublinhados e em negrito possuem conservação igual ou superior a 80%, apenas em negrito 70%, sem formatação 50%. Resíduos entre parênteses (geralmente posições onde mais de um resíduo é conservado) estão sublinhados e em negrito quando sua conservação é igual ou superior a 40%, apenas em negrito 30% e sem formatação 20%. Uma linha conecta resíduos diferentes quando a presença de um deles aumenta a ocorrência do outro em pelo menos 15%. Para facilitar a visualização, linhas conectando resíduos que já estejam nos aglomerados (para PTPs clássicas ou DSPS) ou na tabela 3 (para LMW-PTPs) não são mostradas.

**5. Colaborações no período em Análise de
Acoplamentos Estatísticos**

5.1 Hexoquinases

A hexoquinase (ATP: D-hexose 6-fosfotransferase, EC 2.7.1.1), primeira enzima da via glicolítica, é uma enzima que catalisa a formação de hexose-6-fosfato a partir da hexose, usando ATP como doador do grupo fosforil. Ela é encontrada em praticamente todos os seres vivos, muitas vezes em diferentes isoformas. Mamíferos, por exemplo, possuem quatro hexoquinases (KATZEN; SCHIMKE, 1965), uma delas composta de uma unidade de aproximadamente 50 kDa, e as demais possuindo duas unidades similares, num total de aproximadamente 100 kDa. Estas últimas foram provavelmente produto de duplicação e fusão gênica (WILSON, 1995).

Em leveduras, estão presentes duas isoformas com uma única unidade cada: a hexoquinase PI, com 485 resíduos, e a hexoquinase PII, com 486. Com a determinação da estrutura da hexoquinase PI ligada à glicose (STEITZ *et al.*, 1976; BENNETT; STEITZ, 1980) e da hexoquinase PII ligada ao inibidor competitivo orto-toluil-glucosamina (ANDERSON *et al.*, 1978; ANDERSON *et al.*, 1979), ao fim dos anos 70, elas tornaram-se um exemplo clássico de ajuste induzido (STRYER, 1995), devido às mudanças estruturais observadas quando em complexo com seus ligantes.

Após a implementação dos algoritmos de análise de acoplamentos estatísticos pelo autor e seus testes com os alinhamentos previamente utilizados nos trabalhos do grupo de Rama Ranganathan (LOCKLESS; RANGANATHAN, 1999; SUEL *et al.*, 2003), a família das hexoquinases foi a primeira aplicação inédita do método pelo autor. Neste trabalho, utilizou-se a metodologia original de Ranganathan sem quaisquer modificações, com o intuito de identificar regiões conservadas e correlacionadas nesta família de proteínas.

Um alinhamento de 317 seqüências foi obtido do banco de dados PFAM (FINN *et al.*, 2008). Valores de conservação (ΔG^{stat}) e acoplamento ($\Delta \Delta G^{\text{stat}}$) foram calculados como previamente definidos em 2.1, respectivamente para todas as posições do alinhamento e para 209 perturbações. O número de perturbações válidas foi obtido pelo método de remoção aleatória de seqüências, que resultou no limite de 35% do tamanho inicial do alinhamento. Após o processo de aglomeração hierárquica da matriz de $\Delta \Delta G^{\text{stat}}$, dois aglomerados foram obtidos.

A análise de conservação (ΔG^{stat}) mostrou que os resíduos mais conservados na proteína encontram-se em seu sítio ativo ou imediatamente próximos a ele, como visto na figura 9.

Figura 9 - Conservação posicional (ΔG^{stat}) em hexoquinase PI de *S. cerevisiae*. Resíduos com cores próximas ao azul são menos conservados, enquanto os próximos ao vermelho têm alta conservação. A figura foi gerada utilizando a opção de coloração por fator de temperatura, após modificação do PDB para substituir os fatores de temperatura de cada resíduo pelo seu valor de ΔG^{stat} normalizado. A normalização consistiu em atribuir o valor 100 para o resíduo mais conservado (posição 174, que apresenta 77.3% de triptofanos) e multiplicar os demais valores de ΔG^{stat} pelo fator $100/1510.28$, onde 1510.28 é o valor de ΔG^{stat} para a posição 174.

O primeiro dos dois aglomerados obtidos, contendo os maiores valores de $\Delta \Delta G^{\text{stat}}$, consistia de 30 posições, enquanto o segundo possuía 20 posições. A razão do aparecimento do primeiro deles é de fácil interpretação: ele contém resíduos que circundam a região do sítio ativo e também se estende à região de “dobradiça” no movimento de ajuste induzido. Um segundo aglomerado, ao invés de consistir numa rede conexa de aminoácidos, resultado esperado de acordo com a interpretação original de Lockless e Ranganathan (LOCKLESS; RANGANATHAN, 1999), é composto por posições esparsas na proteína. O significado funcional/estrutural deste conjunto de resíduos correlacionados é desconhecido. Os dois aglomerados, e também a região das “dobradiças” que levam ao movimento de ajuste induzido entre os dois domínios da proteína, podem ser vistos na figura 10.

Os resultados deste trabalho, juntamente com os dados estruturais obtidos por difração de raios-x (descritos na seção 6.3), foram publicados em artigo no periódico *Proteins: Structure, Function and Bioinformatics* (KUSER *et al.*, 2008), figurando na capa de sua edição 72, de agosto de 2008.

Superóxido dismutases (SODs) são enzimas capazes de converter radicais superóxido em peróxido de hidrogênio e oxigênio, tendo, portanto, grande importância no combate aos radicais livres. SODs de parasitos são objeto de pesquisa do CBME, devido ao seu potencial como alvos farmacológicos. A determinação das estruturas cristalográficas das SODs de *Plasmodium falciparum*, *Trypanosoma cruzi* e *Trypanosoma brucei*, parte dos estudos de pós-graduação de Marcos V. A. S. Navarro, Raquel K. Bortoleto-Bugs e José Fernando R. Bachega. Estas SODs pertencem à família de superóxido dismutases de Fe/Mn, que necessitam de um íon de ferro ou manganês em seu sítio ativo para seu correto funcionamento. Além disso, essa família inclui simultaneamente membros que funcionam na forma dimérica e outros que têm um tetrâmero como unidade biológica.

O então estudante de mestrado J. F. R. Bachega, interessado no estudo comparativo das proteínas da família, efetuou uma análise de acoplamentos estatísticos em colaboração com o autor desta tese. Os códigos para seqüências de superóxido dismutases de Fe/Mn foram obtidas do servidor PFAM (FINN *et al.*, 2008), e as seqüências obtidas a partir desses códigos do servidor ExpASY

(GASTEIGER *et al.*, 2003). Este procedimento foi necessário devido ao fato de os alinhamentos no PFAM corresponderem a apenas um dos domínios de superóxido dismutases. Uma vez obtidas as seqüências, elas foram alinhadas utilizando o programa MAFFT (KATOHI *et al.*, 2002). O alinhamento final, após a eliminação de fragmentos e redundância, continha 502 seqüências. Conservação e acoplamento estatístico foram calculados como previamente definido, assim como o tamanho mínimo de sub-alinhamento para definir uma perturbação, resultando em uma matriz de $\Delta\Delta G$ s de 208 linhas (posições no alinhamento) e 167 colunas (perturbações válidas). Essa matriz foi submetida a aglomeração hierárquica usando ligação completa com métrica de Manhattan.

O resultado da aglomeração hierárquica, inesperado até então, mostrou três aglomerados de posições correlacionadas, como mostrado na figura 11.

Figura 11 - Resultado da aglomeração hierárquica para o alinhamento de superóxido dismutases de Fe/Mn. As linhas correspondem às perturbações, sendo as colunas as posições no alinhamento, com valores de $\Delta\Delta G$ para cada par perturbação/posição determinado pela barra de cores à direita. O aglomerado 1 inclui resíduos específicos para proteínas diméricas, o aglomerado 2 inclui resíduos específicos de SODs de ferro, enquanto o aglomerado 3 inclui resíduos específicos de SODs de manganês. **Fonte:** (BACHEGA *et al.*, 2009)

Bachege verificou que algumas das posições observadas nos três aglomerados coincidiam com os resíduos descritos previamente por Wintjens e colaboradores (2008) como sendo determinantes de estado oligomérico (dímero ou tetrâmero) e metal ligado ao sítio ativo (ferro ou manganês). Tal observação levou a uma nova interpretação para os resultados de uma análise de acoplamentos

estatísticos – como superóxido dismutases de Fe/Mn podem ser ou dímeros ou tetrâmeros, é necessário que um conjunto de resíduos possibilite a formação ou não deste estado oligomérico; desta forma, a análise de acoplamento estatístico obrigatoriamente identificará tais resíduos, pois eles deverão estar correlacionados em frequências de aminoácidos. Em alguns casos, as características de dois subconjuntos na mesma família podem estar relacionadas à presença de diferentes aminoácidos na mesma posição, o que ocorre, por exemplo, nos aglomerados de resíduos relacionados à incorporação de Ferro e Manganês. A presença de uma alanina na posição 145 geralmente ocorre em SODs que incorporam ferro, enquanto uma glutamina na mesma posição em geral ocorre naquelas que incorporam manganês.

Uma vez que os aglomerados identificados neste estudo continham resíduos adicionais, além daqueles presentes no artigo de Wintjens e colaboradores (2008), sua capacidade de predição para estado oligomérico e metal associado foi testada. Verificou-se que o poder de predição dos resíduos identificados pela análise estatística de acoplamento era similar ao método de Wintjens, ainda que este utilize informações prévias sobre estrutura e treinamento a partir de seqüências com estado oligomérico e afinidade por metal já conhecidas, e aquele seja obtido apenas da análise estatística de um alinhamento múltiplo de seqüências.

Este estudo não apenas levou a uma nova interpretação dos resultados de uma análise de acoplamentos estatísticos (a formação de aglomerados em matrizes de $\Delta\Delta G$ s não necessariamente reflete um *núcleo alostérico* numa família de proteínas, mas pode ser fruto da existência de classes com características diferentes na mesma família), como também tornou possível novas e úteis aplicações para a técnica, como na anotação genômica.

5.3 Peroxidasas

As peroxidases são proteínas que catalisam reações oxidativas utilizando o peróxido de hidrogênio (H_2O_2) como substrato. Elas contêm um grupo heme, têm estrutura quaternária monomérica e estão presentes em animais, plantas, fungos e bactérias. Em colaboração com o grupo de Valery L. Shnyrov, da Universidade de Salamanca, a peroxidase da palmeira *Roystonea regia* foi

fornece a este grupo de pesquisa para estudos biofísicos e estruturais. A cristalização e posterior resolução estrutural desta proteína foram realizadas pelas pesquisadoras Leandra Watanabe e Patrícia Ribeiro de Moura, que demonstraram interesse em utilizar a técnica de análise de acoplamentos estatísticos para obter informações adicionais a respeito da família das peroxidases.

As peroxidases de plantas são divididas em três classes (WELINDER *et al.*, 1992), denominadas I, II e III. A Classe I compreende peroxidases intracelulares, enquanto a Classe II consiste em peroxidases secretórias em fungos, possuindo quatro pontes dissulfeto, dois sítios conservados para ligação ao cálcio e glicosilação do tipo N. As peroxidases da Classe III, além das características citadas para a Classe II, possuem elementos de estrutura secundária adicionais. Estruturas tridimensionais de peroxidases obtidas para as três classes mostram que todas elas possuem a mesma arquitetura, consistindo exclusivamente de hélices- α divididas em dois domínios entre os quais se encontra o grupo heme. A estrutura da peroxidase de *R. regia*, está mostrada na figura 12.

Figura 12 - Estrutura da peroxidase de *Roystonea regia*. O grupo heme está mostrado em verde, com o átomo de ferro em vermelho. Os íons cálcio estão mostrados em amarelo. As conexões mostradas em azul-claro e amarelo correspondem às pontes dissulfeto (carbonos em azul-claro e enxofres em amarelo).

A existência de classes com características próprias dentre as peroxidases tornou-se um atrativo adicional para a aplicação da análise de acoplamentos estatísticos, uma vez que esta técnica mostrou-se eficiente em obter descritores para subconjuntos de uma mesma família de proteínas (BACHEGA *et al.*, 2009).

Um alinhamento de peroxidases foi obtido através do servidor PFAM (FINN *et al.*, 2008). Após eliminação de seqüências redundantes e fragmentos, obteve-se um alinhamento final consistindo de 393 seqüências, truncado de forma a corresponder à numeração da peroxidase da palmeira *R. regia*. A conservação (ΔG^{stat}) e o acoplamentos estatísticos ($\Delta\Delta G^{\text{stat}}$) foram calculados como previamente descrito na seção Materiais e Métodos. A eliminação aleatória de seqüências do alinhamento mostrou que perturbações estatisticamente relevantes deveriam corresponder a posições ocorrendo em no mínimo 25% das seqüências do alinhamento original. A partir deste valor, um total de 263 perturbações válidas foi possível, sendo então construída uma matriz de $\Delta\Delta G^{\text{stat}}$ consistindo em 244 linhas (correspondendo ao número de resíduos na peroxidase de *R. regia*) e 263 colunas.

A análise de conservação posicional (ΔG^{stat}) indicou a presença de uma série de posições altamente conservadas. Escolhendo como critério de corte 50% do valor máximo de ΔG^{stat} , temos as 26 posições mostradas na tabela 4:

Tabela 5 - Posições mais conservadas em peroxidases de plantas.

Ordem	Posição	Resíduo(s)	ΔG^{stat}	Detalhes
1	169	His (100%)	1470.94	Sítio ativo, resíduo catalítico
2	42	His (98%)	1420.25	Sítio ativo, resíduo catalítico
3	152	Phe (99%)	1162.04	Localizado no plano do grupo heme
4	84	Lys (98%)	1155.45	Ponte salina com Glu88
5	123	Arg (100%)	1091.28	Ponte salina com D99
6	38	Arg (99%)	1079.96	Sítio ativo, resíduo catalítico
7	96	Ser (100%)	1050.74	Localizado em loop, vizinha à cisteína 97
8	175	His (81%)	1014.35	Localizado em fita- β , vizinha à cisteína 176
9	99	Asp (100%)	1010.43	Ponte salina com R123
10	125	Asp (100%)	1010.41	Localizado em loop
11	248	Asp (99%)	1002.66	Sítio ativo, resíduo catalítico
12	122	Gly (100%)	926.376	Localizado em loop
13	172	Gly (100%)	926.389	Localizado em loop, próximo ao canal de entrada que leva ao grupo heme
14	165	Leu (100%)	921.952	Localizado na hélice F, próximo ao grupo heme
15	170	Thr (89%)	845.978	Vizinha ao resíduo catalítico His169
16	167	Gly (96%)	819.907	Localizado na hélice F
17	234	Tyr/Phe (27%/70%)	793.212	Localizado em loop
18	163	Val (83%)	781.26	Localizado na hélice F
19	69	Pro (78%)	756.596	Localizado em loop, próximo ao resíduo catalítico His42
20	97	Trp/Cys (46%/53%)	754.873	Em classe II/III, é uma cisteína e forma ponte com Cys299
21	98	Ala (92%)	739.082	Vizinha à Cys97
22	227	Pro (77%)	731.079	Localizado em loop exposto ao solvente
23	91	Tyr/Cys (46%/53%)	720.178	Em classe II/III, é uma cisteína e forma ponte com Cys11
24	39	Met/Leu (34%/62%)	657.883	Localizado na hélice B
25	113	Gly (85%)	640.648	Localizado em loop exposto ao solvente
26	251	Leu (85%)	637.416	Localizado na hélice H

A localização espacial destes resíduos na estrutura da peroxidase da palmeira *R. regia* está mostrada na figura 13:

Figura 13 - Localização das 26 figuras mais conservadas na estrutura cristalográfica da peroxidase de *R. regia*. Os resíduos mais conservados estão mostrados em verde com legendas em vermelho, o grupo heme em branco e os açúcares ligados covalentemente à estrutura em amarelo.

Pela figura 13, observa-se que os resíduos mais conservados encontram-se em sua maioria próximos ao grupo heme, mas um pequeno conjunto de resíduos conservados é observado próximo à região N-terminal (porção superior esquerda da figura).

A matriz de $\Delta\Delta G^{\text{stat}}$ foi submetida à aglomeração hierárquica utilizando o método de ligação completa, com dissimilaridade medida pela métrica de Manhattan. Após a eliminação de linhas e colunas de baixo sinal, verificou-se que a matriz continuava consistindo de uma quantidade extremamente grande de posições com altos valores de acoplamento, ao contrário dos casos reportados por R. Ranganathan e similar aos resultados observados para as proteínas proteínas tirosina fosfatases de baixo peso molecular. A matriz final, após o procedimento de aglomeração e eliminação de linhas e colunas de baixo valor de $\Delta\Delta G^{\text{stat}}$, está mostrada na figura 14.

Figura 14 - Matriz de $\Delta\Delta G^{\text{stat}}$ calculada a partir do alinhamento de peroxidases, após aglomeração hierárquica e eliminação de linhas e colunas de baixo valor de $\Delta\Delta G^{\text{stat}}$.

A ocorrência de uma grande quantidade de resíduos correlacionados numa família de proteínas que, sabidamente, subdividem-se em três classes, sugere que o acoplamento se deva aos resíduos ocorrentes em classes específicas, como em LMW-PTPs. Ao se verificar o comportamento da distribuição de aminoácidos nas posições correlacionadas, observa-se que de fato trata-se do mesmo fenômeno, havendo uma clara separação entre duas “classes” dentro do alinhamento, cada uma conservando resíduos específicos. Este comportamento é mostrado na figura 15.

		10	20	30	40	
ClassI	LKXDLXQXXDXGXGCFIRMAWHSAGTYRXXXXDGR	40				
ClassII	VRXXVXXXXXXXXXXAAQLRLHFHDCFVRGCCASVLDLS	57				
		50	60	70	80	
ClassI	GEXRXAPLNSWFXDXLDKARLLWPIKQXXXXKYGKXIWA	80				
ClassII	TX.NXXEKDAPFNXXSLRGFJVIXAAALEACPEVVSC	96				
		90	100	110	120	
ClassI	DLMLTGNVALESMCFKTFAGGREDXWRVCGJRLENPL	120				
ClassII	ADILALARDSVVLXGGPXXVPLGRGRDGRVSAJSJXXN	136				
		130	140	150	160	
ClassI	AAVLAAAXDIREFARMAMNDEETVALAGGHIFGKHGA	160				
ClassII	LPXPPXXNXXLXXXFAKGLXXXDLVLLSGAHICXSHCX	176				
		170				
ClassI	GDAD.....LEQCLGWKNSECFD	178				
ClassII	SFSXRRLYNFXGXXDPTLDXYAAMLRXXCFXXGXXXX	216				
		180	190	200	210	
ClassI	ITSGLEGCSPTQWDNGFFENLFHXPXMLTDLSLRFDXY	218				
ClassII	XXVXLDPXSPXXFDNXXYYKNLXGKCLLXSQALXTDXT	256				
		220				
ClassI	EKIS	222				
ClassII	RXXV	260				

Figura 15 - Pseudo-alinhamento das classes obtidas a partir da análise de acoplamentos estatísticos para o alinhamento de peroxidases. Resíduos de mesmas propriedades físico-químicas conservados nas duas classes são sublinhados de acordo com o esquema de cores: vermelho: ácido; preto: alifáticos; cinza: alifáticos de cadeia lateral curta; verde: amida; marrom: aromático; azul: básico; magenta: hidroxila; laranja: imina; amarelo: cadeia lateral contendo enxofre. Posições marcadas com X indicam ausência de conservação naquela posição para a classe em questão, enquanto pontos (.) indicam ausência de resíduos nesta posição. As duas classes obtidas pela análise de acoplamentos estatísticos correspondem respectivamente às peroxidases de Classe I e às peroxidases das classes II e III.

A separação entre as duas classes no alinhamento pode ser feita ao utilizar como critério de separação a existência ou ausência de cisteínas nas posições onde há pontes dissulfeto, todas elas com conservação dominada por cisteína em aproximadamente 47% das seqüências e um ou mais resíduos nos demais casos, como sumarizado na tabela 6:

Tabela 6 - Distribuição de aminoácidos nas posições relativas às pontes dissulfeto cobertas pelo alinhamento da família de peroxidases no servidor PFAM. A presença de uma cisteína em qualquer uma dessas posições é seguida da presença de uma cisteína nas outras.

Posição	Cisteínas	Demais resíduos
27	47%	Alanina (53%)
32	47%	Isoleucina (22%), Treonina (14%), Valina (8%)
74	46%	Tirosina (53%)
80	46%	Triptofano (52%)
159	46%	Glicina (53%)
191	46%	Leucina (42%)

A ocorrência de uma grande quantidade de resíduos correlacionados que são conservados em apenas um dos casos, como verificado no pseudo-alinhamento mostrado na tabela 6, sugere que não apenas as cisteínas são conservadas nas classes II e III e ausentes na Classe I, mas uma série de outros resíduos apresenta conservação de acordo com a classe, como, por exemplo, Glu64, Phe41, Val95, Arg106 e Ser52 cuja frequência na presença das cisteínas (Classes II e III) é de 98.9%, 94.0%, 98.9%, 97.8% e 98.9%, respectivamente.

Além das pontes salinas cuja existência recorrente na família foi indicada pelos valores de conservação estrita (e.g. Asp99 e Arg123 apresentam, respectivamente, frequências de 100% e 97%), pontes salinas cuja ocorrência se dá apenas em parte das peroxidases pôde ser identificada por altos valores de correlação ($\Delta\Delta G^{\text{stat}}$). A presença de uma arginina na posição 182, por exemplo, elevou a frequência de ácidos aspárticos na posição 193 de 39.8% para 85.0%, indicando a importância desta ponte salina para manutenção da estrutura em peroxidases de classe II/III, em que a presença destes resíduos ocorre.

O fato da análise de acoplamentos estatísticos levar a uma bipartição do alinhamento entre classe I e classes II/III reforça a maior proximidade entre as classes II e III, como previamente descrito por Welinder (WELINDER *et al.*, 1992), de acordo com a presença de elementos estruturais

característicos. A presença de uma quantidade considerável de resíduos conservados em todas as classes, aliadas às observações mostradas acima sugere que a evolução se deu a partir de um ancestral comum, com uma divergência inicial separando classe I das demais classes, e posterior diferenciação entre classes II e III.

"Stick to crystallography, while you are young"

(Eric Westhof)

**6. Trabalhos desenvolvidos no período em
Cristalografia de Proteínas**

6.1 Preâmbulo: Cristalografia de Proteínas – avanços e obstáculos recorrentes

As particularidades referentes à natureza dos cristais de proteínas e à complexidade de suas estruturas resultaram em variados problemas enfrentados desde o início da aplicação da cristalografia ao estudo de macromoléculas biológicas – muitos deles resolvidos ou minimizados devido a avanços em técnicas experimentais e novos métodos nos procedimentos de resolução estrutural.

Embora os primeiros experimentos de difração de raios-x em cristais de proteínas tenham sido publicados por Bernal & Crowfoot em 1934 (BERNAL; CROWFOOT, 1934), as primeiras estruturas de proteínas, determinadas por cristalografia por Kendrew e Perutz, só foram publicadas em 1960 (KENDREW *et al.*, 1960; PERUTZ *et al.*, 1960). Naquela época, mesmo estruturas com menos de cem átomos eram de difícil resolução (DAUTER, 2006).

Entre os problemas minimizados por mudanças em metodologia está o dano da radiação em cristais. Já em seu artigo inaugural de difração de raios-x de 1934, Bernal e Crowfoot verificaram que cristais de proteínas tornavam-se irremediavelmente danificados e perdiam a capacidade de difratar raios-x quando secos, levando-se à necessidade de coletar dados de difração utilizando espécimes em capilares selados, em equilíbrio com a solução mãe (BERNAL; CROWFOOT, 1934). Uma vez que cristais coletados nestas condições são extremamente sensíveis aos danos causados por radiação, havia a recorrente necessidade de se utilizar vários cristais para obter conjuntos de dados completos, o que contribuía para o aumento dos erros obtidos para as intensidades medidas. Com a introdução do congelamento de cristal a baixíssimas temperaturas (comumente chamada de “criocristalografia”), em que a adição de agentes crioprotetores às soluções em que se encontram os cristais permite seu congelamento a temperaturas próximas de 100K, de forma que suas propriedades e poder de difração sejam preservados (HOPE, 1988). Tal procedimento reduziu drasticamente os danos por radiação, praticamente eliminando a necessidade de utilização de múltiplos cristais para a obtenção de conjuntos de dados de difração completos.

O fluxo de fótons necessário para experimentos de difração de raios-x também deixou de ser um problema com a evolução das fontes de raios-x. Enquanto no início da cristalografia de proteínas a

única fonte utilizada eram os tubos selados, hoje os ânodos rotatórios tornaram-se a fonte padrão para laboratórios e são capazes de produzir intensidade suficiente para a grande maioria dos casos. Além disso, as fontes de luz síncrotron evoluíram de tal forma que em muitos deles, experimentos de difração de raios-x em proteínas são feitos com atenuação do feixe. Em 2004, 70 linhas de luz estavam disponíveis no mundo para cientistas com interesse em efetuar experimentos em cristalografia de proteínas (JIANG; SWEET, 2004). Além do enorme ganho em fluxo de fótons quando comparada aos tubos selados utilizados nos anos 50, a luz síncrotron apresenta a vantagem de poder ser filtrada por comprimento de onda, permitindo a utilização de técnicas de faseamento como a MAD (do inglês “multi-wavelength anomalous dispersion”). Os avanços em biologia molecular também causaram grande impacto: enquanto no passado cristais de proteínas eram obtidos com proteína extraída a partir do próprio ser vivo que a produz, o que comumente tinha baixíssimo rendimento, hoje é possível produzir proteínas em grandes quantidades em organismos preparados para expressão heteróloga, e mesmo produzir proteínas que já contenham átomos com alto poder de espalhamento anômalo, caso das proteínas heterólogas contendo seleno-cisteínas.

Os avanços na computação também permitiram a utilização de programas com rotinas impensáveis no início da cristalografia de proteínas. Hoje é possível indexar, integrar e escalonar um grande conjunto de dados (que, com os padrões atuais de câmeras CCDs podem ocupar alguns gigabytes, unidades maiores em várias ordens de grandeza em relação à capacidade dos computadores utilizados no passado), efetuar seu faseamento e construir automaticamente um modelo cristalográfico utilizando um computador pessoal. Rotinas como a busca em seis dimensões de uma solução de substituição molecular ou a construção automática de um modelo utilizando o mapa de densidade eletrônica só foram possíveis graças aos avanços em poder computacional. Entre casos mais extremos inclui-se o algoritmo de *Charge Flipping* (OSZLANYI; SUTO, 2004), em que o faseamento de um conjunto de dados é efetuado pela simples repetição, por milhares de iterações, da transformação em espaço real e recíproco utilizando como modificação a simples troca de sinal em posições em que a densidade eletrônica está abaixo de um valor pré-definido, metodologia recentemente aplicada à cristalografia de proteínas (DUMAS; VAN DER LEE, 2008).

Casos de estruturas cujo modelo final era completamente equivocado (AMMON *et al.*, 1988; MCREE *et al.*, 1989; LUBKOWSKI *et al.*, 1994; BORGSTAHL *et al.*, 1995) levaram ao desenvolvimento de uma enorme quantidade de métodos de *validação*. Muitos deles verificam o modelo final e são capazes de detectar problemas comuns como estereoquímica incorreta (LASKOWSKI *et al.*, 1993) ou falhas de registro indicadas por ambiente químico improvável (BOWIE *et al.*, 1991; LUTHY *et al.*, 1992). Dentre os valores obtidos por análise geral dos dados obtidos, merece grande destaque o advento do R_{free} , proposto por Axel Brünger em 1992 (BRUNGER, 1992). Uma vez que o R_{factor} , que mede a discrepância entre fatores de estrutura experimentais e os calculados a partir do modelo, pode ser facilmente diminuído devido a ajustes no modelo que não correspondem à realidade, criou-se o R_{free} de forma a validar o ajuste calculado-experimental, a partir do cálculo idêntico ao do R_{factor} , mas utilizando um subconjunto das reflexões que não seja utilizada no refinamento da estrutura. A grande quantidade de métodos de validação de estrutura desenvolvidos na década de 90, revisados em 2000 por Gerard Kleywegt (KLEYWEGT, 2000), aliados aos crescentes avanços em automação em todas as fases de resolução estrutural leva a uma inevitável tendência de que, cada vez mais, proteínas sejam resolvidas por cristalografia com cada vez menos intervenção humana. Na opinião do cristalógrafo polonês Zbigniew Dauter, “embora a automação do processo de resolução de uma estrutura cristalina possa ser vista por cristalógrafos tradicionais como o fim de uma era heróica, ela certamente é um resultado bem-vindo para biólogos que querem utilizar o conhecimento de uma estrutura protéica para atacar diversos problemas biológicos. Porém, seria vantajoso que as ‘caixas pretas’ automatizadas tivessem embutidas algumas ‘luzes vermelhas’ com avisos como ‘a completeza dos dados está muito baixa’, ‘há muitas reflexões sobrecarregadas’, ‘o cristal apresenta alto grau de geminação’, ‘há muito dano por radiação para faseamento por MAD’, etc.” (DAUTER, 2006). No mesmo artigo, Dauter ressalta que “sempre haverá estruturas difíceis ou cristais apresentando vários tipos de comportamento patológicos, que darão oportunidades aos especialistas de ‘manterem seus cérebros ocupados’” (DAUTER, 2006).

Uma fonte comum de problemas é a determinação errônea de grupos espaciais. Este problema torna-se ainda mais frequentes em caso de cristais geminados. O tipo mais simples de geminação de

cristais é o epitaxial, em que dois (ou mais) domínios cristalinos estão presentes no cristal sem que haja operações de simetria relacionando-os. O padrão de difração de tais cristais, semelhantes ao de cristais partidos, é uma simples superposição desorientada dos padrões de difração de cada domínio, com intensidades proporcionais à quantidade em massa de cada um dos domínios. Conjuntos de dados obtidos de cristais apresentando geminação epitaxial podem ser processados pela simples escolha entre uma das redes (preferencialmente aquela que apresenta as maiores intensidades) e sua consequente integração e escalonamento. Opcionalmente, pode-se mesmo integrar separadamente as duas redes e em seguida escaloná-las em um único conjunto de dados.

Quando os domínios estão relacionados por uma operação de simetria, o tratamento dos conjuntos de dados de difração torna-se mais complicado. Nestes casos, há superposição de pontos de difração, que pode ser total ou parcial. Quando todos os pontos do padrão de difração correspondem à soma da contribuição dos domínios, a geminação é chamada de merohedral, sendo a geminação não-merohedral aquela em que há simultaneamente pontos de difração que correspondem à contribuição de cada domínio e também pontos relativos a um único fator de estrutura. Finalmente, a geminação pseudo-merohedral é aquela decorrente de combinações fortuitas de parâmetros de rede que levam a um padrão de difração idêntico ao daquele gerado pela geminação merohedral, mas que, pelo fato de muitas vezes a superposição entre as redes não ser perfeita, podem causar alargamento ou separação entre os pontos superpostos, especialmente na alta resolução (DAUTER, 2003). Um parâmetro importante na análise de cristais geminados é a fração de geminamento (α), que indica a proporção em massa dos dois domínios (e que também será o peso aplicado às intensidades resultantes da difração de cada domínio, resultando na intensidade total medida por cada ponto de difração em que ocorre superposição). Valores de α próximos de zero indicam que um dos domínios está presente em quantidade muito maior que o outro, o que resulta em efeito pouco pronunciado da geminação, e, por conseguinte, não levam a grandes dificuldades na resolução da estrutura (o que muitas vezes é feito ignorando a existência do fenômeno). O oposto acontece para valores de α próximos de 50%, o que no passado era considerado um problema intratável em cristalografia de proteínas. Hoje há uma grande quantidade de estruturas obtidas por cristais geminados. Este problema pode ser abordado de duas

formas: a primeira é a geração de um conjunto de dados “não-geminado” a partir dos dados originais de difração, o que é feito com base no fato de que, como as intensidades medidas são combinação linear das intensidades resultantes de cada domínio, com pesos α e $1 - \alpha$, é possível montar um sistema de equações lineares com as intensidades medidas, as intensidades de cada domínio e o valor de α de forma a obter todos estes valores. Tal sistema é sub-determinado para o caso de geminação perfeita, com $\alpha=0.5$, e mesmo nos demais casos deve-se levar em conta que a obtenção dos reais valores das intensidades independentes é afetada pelos erros intrínsecos das intensidades medidas. A segunda abordagem é o refinamento da estrutura incluindo o efeito do geminamento no cálculo dos fatores de estrutura, o que permite que o refinamento estrutural e o cálculo de α seja feito simultaneamente.

Uma vez que a qualidade de um modelo cristalográfico depende da qualidade dos cristais obtidos, o que leva a um limite em resolução efetiva independente da fonte de raios-x utilizada, a resolução de estruturas cristalinas a baixa resolução será sempre um problema recorrente em cristalografia. O primeiro grande problema é a interpretação dos mapas de densidade eletrônica – com resoluções mais baixas que 3Å, a identificação das cadeias laterais torna-se problemática, não podendo sequer ser construídas em muitos casos. Além disso, uma vez que a resoluções desta ordem há comumente menos parâmetros refinados que observações medidas, ocorre comumente o “sobre-refinamento”, i.e., o refinamento de mais parâmetros que o possível para uma dada resolução de forma a diminuir o valor de R_{factor} . Tal problema era ainda mais comum antes do advento do R_{free} – o que levou Gerard Kleywegt e Alwyn Jones, que ganharam notoriedade nos anos 90 devido a uma série de artigos sobre validação de estruturas, críticas a estruturas determinadas de forma inapropriada e defesa da deposição de dados experimentais e dos métodos de verificação como o próprio R_{free} , a afirmarem que a maior parte das estruturas determinadas a partir de dados de baixa resolução no passado foram de fato sobre-refinadas, o que era atestado pela verificação de que ao longo dos anos 90 os valores médios de R_{factor} aumentaram, ao invés de diminuir (KLEYWEGT; JONES, 2002). Ao mesmo tempo, a utilização do R_{free} ganhou ampla aceitação – em 1995, apenas 33% das estruturas depositadas reportavam o valor de R_{free} , contra 92% em 2000 (idem).

O autor deste trabalho, tendo se decidido pela carreira acadêmica por ocasião de seu projeto de iniciação científica em cristalografia (caracterização de materiais por difração de pó, difração de monocristal de alta resolução angular e computação cristalográfica), esteve sempre envolvido nesta área de pesquisa ao longo de sua formação acadêmica. Por esse motivo, que levou à participação de uma quantidade considerável de coletas de dados nas linhas de luz do Laboratório Nacional de Luz Síncrotron e nos difratômetros do Grupo de Cristalografia do Instituto de Física de São Carlos durante a vigência do doutorado, foi possível a participação na solução de algumas estruturas cristalográficas obtidas em seu grupo de pesquisa. Algumas delas apresentaram interessantes particularidades quando de sua resolução, o que justifica a existência desta seção.

6.2 Transtirretina humana

A transtirretina (TTR) é uma proteína transportadora capaz de se ligar ao hormônio tireoidiano (T_4) e ao retinol. Apesar da maior parte do T_4 ligado ser transportado pela globulina de ligação à tiroxina (TBG, do inglês *thyroxin binding globulin*), no cérebro a transtirretina é o principal transportador (BARTALENA; ROBBINS, 1993; ALDRED *et al.*, 1995). Ela é uma proteína homotetramérica rica em folhas β , que pode se decompor em monômeros capazes de formar fibras amilóides, relacionadas a doenças conhecidas como amiloidoses sistêmicas. Tais patologias ocorrem com mais frequência na presença de mutações pontuais, sendo mais de 80 delas documentadas como tendo efeito deletério até 2001 (SARAIVA *et al.*, 1984; SARAIVA, 2001). Uma das possíveis abordagens para tratamento é a tentativa de estabilizar o tetrâmero de TTR através do uso de pequenas moléculas capazes de ligar-se a seu sítio. A genisteína, um produto natural encontrado na soja, foi descrito como sendo capaz de promover tal efeito (GREEN *et al.*, 2005). A caracterização estrutural e biofísica da interação entre ligantes naturais e a TTR, tanto em sua forma selvagem como em mutantes pontuais, é parte do projeto da aluna de doutorado Daniela B. B. Trivella, com quem o autor teve participação direta na determinação de suas primeiras estruturas cristalográficas, em especial para a genisteína.

A TTR cristaliza-se no grupo espacial $P2_12_12$, com duas subunidades na unidade assimétrica. Devido a este fato, o tetrâmero fisiológico da proteína é obtida a partir da unidade assimétrica por uma operação de simetria – especificamente, a aplicação do eixo cristalográfico de ordem 2. Curiosamente, o sítio de ligação ao hormônio tireoidiano encontra-se exatamente sobre este eixo de simetria (a cavidade é formada por duas subunidades, havendo, portanto, dois sítios de ligação), sendo possíveis dois modos de ligação simétricos, uma vez que o próprio sítio de ligação apresenta uma simetria de ordem 2. A disposição das subunidades de TTR está mostrada na figura 16.

Figura 16 - Estrutura da TTR selvagem obtida por difração de raios-X. As subunidades mostradas à esquerda da figura estão relacionadas às da direita por um eixo cristalográfico de ordem 2. A unidade assimétrica contém duas subunidades.

Devido às duas possibilidades em modo de ligação da genisteína ao sítio em TTR e à presença do eixo cristalográfico nesta posição, o refinamento posicional do ligante foi possível selecionando-se uma conformação e impondo fator de ocupação 0.5 para seus átomos – desta forma, a operação cristalográfica gera automaticamente a segunda conformação, que pode então ser refinada normalmente. O efeito deste procedimento está mostrado na figura 17.

Figura 17 - Conformações observadas para a genisteína nos dois sítios de ligação presentes na transtirretina. O eixo cristalográfico de ordem dois atravessa os dois sítios, resultando em duas conformações diferentes para cada ligante. **Fonte: Trivella *et al.* (em preparação)**

A excelente resolução do conjunto de dados (1.75Å) permitiu a obtenção de um modelo cristalográfico de alta qualidade, com baixos valores de erro e interpretação inequívoca do mapa de densidade eletrônica para a grande maioria dos resíduos, inclusive em alguns casos de dupla conformação. O resumo das estatísticas de processamento e refinamento está mostrado na tabela 7. A estrutura cristalográfica da TTR humana ligada à genisteína (e também a do mutante pontual V30M ligado ao mesmo ligante) é parte dos resultados de um trabalho de autoria da doutoranda Daniela B. B. Trivella, em estágio final de preparação.

Tabela 7 - Dados cristalográficos da estrutura de transtirretina humana. Valores entre parênteses referem-se à última faixa de resolução (um total de 20 faixas foi utilizado). Os dados foram integrados e escalonados utilizando os softwares Denzo e Scalepack (OTWINOWSKI; MINOR, 1997), e a estrutura refinada utilizando o Refmac (MURSHUDOV *et al.*, 1997).

Resolução	29.66Å – 1.75Å (1.80Å – 1.75Å)
Grupo espacial	P2 ₁ 2 ₁ 2
Parâmetros de rede (Å)	a = 42.62 b = 82.86 c = 63.51
Completeza (%)	97.7 (87.4)
Reflexões utilizadas no refinamento	21460 (1424)
Total de reflexões	22830
Reflexões utilizadas para cálculo de R _{free}	1370 (82)
I/σ	18.5 (2.2)
Multiplicidade	3.3 (2.8)
R _{factor} (%)	17.84
R _{free} (%)	21.34
RMSD comprimentos de ligação (Å)	0.019
RMSD ângulos de ligação (°)	1.991
B-fator médio (modelo)	22.2
B-fator médio (gráfico de Wilson)	20.1

6.3 Hexoquinase PI de *Saccharomyces cerevisiae*

A hexoquinase, descrita na seção 5.1, tornou-se um conhecido exemplo do mecanismo de ajuste induzido com base em estruturas de proteínas que, à época de sua resolução, não tinham conhecidas suas seqüências de aminoácidos, sendo determinadas por inspeção da densidade eletrônica obtida por difração de raios-x. O estudo das hexoquinas por análise de acoplamentos estatísticos foi feito juntamente com um trabalho estrutural no qual a estrutura da hexoquinase PII de *S. cerevisiae* foi re-determinada, com a utilização da seqüência real de aminoácidos da proteína.

A cristalização e coleta de dados foram efetuadas no início dos anos 2000, no Laboratório Nacional de Luz Síncrotron. O modelo final gerado com os *softwares* de cristalografia disponíveis àquela época apresentava alguns problemas, que comumente ocorrem em estruturas de resolução de moderada a baixa (no caso em questão, de 2.95Å): uma diferença considerável entre os valores de R_{factor} e R_{free} (18.9% e 25.9%, respectivamente), altos valores de desvios quadráticos médios para comprimentos (0.026Å) e ângulos (2.313°) de ligação e um valor médio do fator de Debye-Waller (37.1Å²) relativamente distante daquele obtido pelo gráfico de Wilson (47.4Å²).

Uma conhecida fonte de problemas em resolução de estruturas de proteínas por difração de raios-x é a relação entre dados experimentais medidos e parâmetros refinados. Enquanto em estruturas de pequenas moléculas há comumente muito mais fatores de estrutura que parâmetros obtidos, em proteínas esta relação pode se tornar desfavorável, o que diminui a confiabilidade do modelo obtido (o que comumente pode ser detectado por altos valores de R_{free}). Na estrutura da hexoquinase PI de *S. cerevisiae*, foram medidas 14357 reflexões únicas, das quais 715 foram utilizadas para o cálculo de R_{free} , restando 13642 para serem utilizadas no refinamento. O número total de átomos refinados foi de 3661, o que leva a uma relação de observações/parâmetros de apenas 0.93 utilizando quatro parâmetros refinados por átomo (x, y, z e B_{iso}). Logo, uma abordagem mais realista para tal resolução é a utilização de fatores de temperatura únicos para cada resíduo. Desta forma, o refinamento foi refeito utilizando 477 grupos (471 resíduos de aminoácidos, uma molécula de glicose, um íon sulfato e

quatro moléculas de água), o que diminui o número de parâmetros refinados de 14644 para 11460, resultando numa relação observações/parâmetros de 1.19.

Além disso, em estruturas de baixa resolução é preferível utilizar um peso maior para as restrições geométricas utilizadas no refinamento, uma vez que a incerteza na posição dos átomos é maior. A aplicação de um protocolo de refinamento que incluiu o refinamento de fatores de temperatura por grupo e maior restrição geométrica foi capaz de gerar um modelo mais realista, o que se reflete na redução da diferença entre R_{factor} e R_{free} de 7% para 5%, desvios médios quadráticos de comprimentos e ângulos de ligação aceitáveis e fator de temperatura médio convergindo para um valor mais próximo àquele obtido pelo gráfico de Wilson, como mostrado na tabela 8.

Tabela 8 - Dados cristalográficos da estrutura de Hexoquinase PI de *S. cerevisiae*. Valores entre parênteses referem-se à última faixa de resolução (um total de 20 faixas foi utilizado). Os dados foram integrados e escalonados utilizando os softwares Denzo e Scalepack (OTWINOWSKI; MINOR, 1997), e a estrutura refinada utilizando o Phenix.refine (ADAMS *et al.*, 2002).

Resolução	12.92Å – 2.95Å (3.03Å – 2.95Å)
Grupo espacial	P2 ₁ 2 ₁ 2 ₁
Parâmetros de rede (Å)	a = 62.12 b = 78.87 c = 144.74
Completeza (%)	92.27 (94.44)
Reflexões utilizadas no refinamento	13642 (1007)
Total de reflexões	14357
Reflexões utilizadas para cálculo de R_{free}	715
I/σ	8.8 (2.49)
R_{merge} (%)	11.9 (47.77)
Multiplicidade	2.8 (2.7)
R_{factor} (%)	19.19
R_{free} (%)	24.45
RMSD comprimentos de ligação (Å)	0.002
RMSD ângulos de ligação (°)	0.470
B-fator médio (modelo)	44.8
B-fator médio (gráfico de Wilson)	47.4

A adoção do protocolo de refinamento utilizado neste trabalho foi capaz de gerar uma estrutura com qualidade compatível com a resolução obtida, capaz de identificar a ligação ao açúcar e a uma molécula de fosfato (figura 18), e de mostrar a distorção causada pela ligação ao substrato, no movimento de ajuste induzido (figura 19).

Figura 18 - Sítio ativo da Hexoquinase PI de *S. cerevisiae*. À esquerda, os contatos da molécula de glicose e do íon fosfato com resíduos da proteína são mostrados. À direita, os mesmos ligantes são mostrados juntamente com um mapa OMIT $2F_o - F_c$ contornado em 1σ a uma distância máxima de 2Å dos ligantes.

Figura 19 - Ajuste induzido na Hexoquinase PI de *S. cerevisiae*. Superpondo-se a estrutura da hexoquinase em sua forma apo com a forma holo, observa-se um alto valor de desvio em posição de carbonos α em certas posições da proteína (cada legenda indica a distância entre os C α dos resíduos entre parênteses). Estes desvios caem consideravelmente quando os dois domínios da proteína são superpostos separadamente, mostrando que o movimento de ajuste induzido pode ser descrito como a mudança de orientação dos dois domínios.

Como previamente descrito, este trabalho resultou na publicação de um artigo científico no periódico *Proteins: Structure, Function and Bioinformatics* (KUSER *et al.*, 2008).

6.4 Complexo de Interleucina-22 humana e seu receptor IL-22R1

A interleucina-22 (IL-22) é uma citocina produzida principalmente pelas células T, envolvida em processos inflamatórios e imunes, como a promoção de defesa antimicrobial, em diversos órgãos (DUMOUTIER *et al.*, 2000; KOTENKO, 2002). A estrutura cristalográfica da IL-22 recombinante foi determinada neste grupo de pesquisa por Nagem e colaboradores (NAGEM *et al.*, 2002), e consiste em um único domínio compacto formado por seis α -hélices anti-paralelas, com 179 resíduos, um peptídeo sinal, duas pontes dissulfeto e três sítios de glicosilação. A determinação da IL-22 expressa em células

de inseto (XU *et al.*, 2005) mostrou que a glicosilação destes sítios não leva a mudanças consideráveis em estrutura terciária.

A resposta celular à IL-22 requer sua ligação a um receptor membranar hetero-dimérico, formado por duas unidades da família de receptores de citocina de classe II. Há diversas combinações possíveis entre os membros desta família para a formação de receptores, e, no caso da IL-22, ele é composto de uma unidade do receptor IL-22R1 e outra do receptor IL-10R2. A primeira unidade é reconhecida também pelas citocinas interleucina-20 e interleucina-24, e a segunda está presente também no complexo reconhecido pela interleucina-10. Além disso, a IL-22 se liga a uma proteína solúvel denominada IL-22BP, que, pela ausência de região transmembrânica, possui papel ainda em aberto.

Há fortes evidências de que a formação do complexo IL-22/IL-22R1/IL-10R2 se dá através da ligação entre a IL-22 e a IL-22R1 com alta afinidade, e em seguida a associação deste complexo à IL-10R2, com afinidade mais baixa (LOGSDON *et al.*, 2002; LI *et al.*, 2004). Até o momento, não há estrutura resolvida para o complexo trimérico das três proteínas em questão.

Com o objetivo de compreender melhor a base estrutural da ação da IL-22, decidiu-se estudar sua interação com a IL-22R1 e com a IL-22BP através de cristalografia, num trabalho que contou com a colaboração do grupo de Didier Colau e Jean-Christophe Renauld, do Instituto Ludwig de Bruxelas, Bélgica, com as pós-doutorandas Leandra Watanabe e Patrícia Ribeiro de Moura, deste grupo de pesquisa.

Foram obtidos cristais dos complexos de interleucina-22 com IL-22R1 e com IL-22BP. Os dados de difração de raios-x foram coletados na linha de luz MX2 do Laboratório Nacional de Luz Síncrotron, em Campinas, à resolução de 1.9Å. Uma vez que a estrutura da própria IL-22 já era conhecida e havia um homólogo de estrutura disponível para a IL-22R1 (a IL-10R1, parte do receptor de IL-10), seria esperado tratar-se de um caso trivial de substituição molecular. Porém, enquanto a IL-22 foi facilmente encontrada na unidade assimétrica, não foi obtida solução válida para a IL-10R1. Além disso, o mapa inicial obtido pelo faseamento utilizando apenas a estrutura de IL-22

(aproximadamente uma vez e meia menor que a IL-22R1) não tinha qualidade suficiente fora da própria IL-22 para permitir que a cadeia da IL-22R1 fosse traçada manualmente.

A esta época, um *software* de auxílio ao procedimento de substituição molecular chamado MrBump (KEEGAN; WINN, 2008), de autoria de Ron Keegan e Martyn Winn, encontrava-se em fase de testes. O objetivo de tal programa é buscar melhores resultados em casos de substituição molecular através de uma busca extensa por homólogos, tratamento das estruturas molde (transformação em cadeias de poli-alaninas ou substituição de resíduos de acordo com alinhamento sequencial) e tentativas de substituição molecular por diferentes programas como Molrep (VAGIN; TEPLYAKOV, 1997) e Phaser (MCCOY *et al.*, 2007). A utilização do *software* MrBump permitiu a identificação de uma estrutura de busca diferente para a IL-22R1: o fator de tecido (código PDB: 1J7V), que apesar de ter identidade mais baixa em relação à IL-22R1 que a escolha mais óbvia do receptor de IL-10 (IL-10R1), acabou resultando numa substituição molecular bem sucedida.

Após a construção completa e refinamento do modelo de IL-22/IL-22R1, o motivo da falha na substituição molecular foi revelado. Uma vez que os receptores de citocina (e também o fator de tecido) apresentam-se na forma de dois domínios de fibronectina tipo III na forma de um V, pequenas variações no ângulo entre os dois domínios resultariam em grandes mudanças em distâncias interatômicas, o que dificultaria a identificação de uma solução de substituição molecular. E, de fato, a IL-10R1 apresenta um ângulo interdomínio de 115° (medido a partir dos resíduos Thr103, Gln30 e Gly178), enquanto a IL-22R1 tem para este mesmo ângulo o valor de 121° (medido pelos resíduos estruturalmente equivalentes Thr120, Thr48 e Glu196). A disposição destes domínios em IL-22R1 e sua localização em relação à IL-22, além dos contatos formados entre as duas proteínas estão mostrados na figura 20.

Figura 20 - A) Complexo entre a IL-22 e a IL-22R1. A IL-22R1, mostrada em vermelho, é composta de dois domínios fibronectina tipo III, D1 e D2, que formam um ângulo de 121°. Uma visão ampliada das interfaces de dimerização, contendo as cadeias laterais e suas interações, está mostrada em B e C. **Fonte: Bleicher, de Moura *et al.*, 2008.**

Experimentos de mutagênese sítio-dirigida na IL-22R1, em resíduos localizados nas interfaces mostradas na figura 20, efetuados no laboratório do Prof. Dr. Jean-Cristophe Renauld confirmaram que a unidade assimétrica mostrada no modelo cristalográfico é de fato a unidade biológica IL-22/IL-22R1, uma vez que tais mutações causam a perda da inibição por IL-22.

Com base nestes dados, foi possível especular a respeito do complexo hetero-trimérico de IL-22/IL-22R1/IL-10R2. Uma vez que algumas estruturas hetero-triméricas de interleucinas ligadas a seus receptores são conhecidas e uma estrutura de boa qualidade do complexo IL-22/IL-22R1 foi obtida, buscou-se identificar qual dos complexos ternários conhecidos seria mais compatível com a estrutura do complexo binário obtido por difração de raios-x. No início de 2008 (LAPORTE *et al.*, 2008), foram obtidos os complexos ternários de IL-4/IL-4R α / γ_c , IL-4/IL-13R α 1/IL-4R α e IL-13/IL-13R α 1/IL-4R α (códigos PDB: 3BPL, 3BPN e 3BPO, respectivamente). A superposição da estrutura de IL-22/IL-22R1 a esses complexos mostrou que a estrutura de IL-4/IL-4R α / γ_c é um modelo plausível para o complexo ternário IL-22/IL-22R1/IL-10R2, em que a IL-4R α ocuparia o lugar equivalente ao de IL-10R2 e γ_c teria a localização da IL-22R1. Esta hipótese está mostrada na figura 21. Para a

escolha desse complexo como modelo, foi utilizado como critério, além da superposição correta de uma citocina e um receptor, a localização do segundo receptor próximo aos resíduos Tyr51, Asn54, Tyr114 e Glu117 de IL-22, identificados por Logsdon e colaboradores (LOGSDON *et al.*, 2004) por ensaios de mutagênese sítio-dirigida como sendo importantes para a ligação à IL-10R2.

Figura 21 - Um modelo para a unidade funcional heterotrimérica de IL-22. A IL-22 está mostrada em amarelo, e a IL-22R1 em vermelho, superpostas ao complexo IL-4/IL-4R α / γ_c (verde, roxo e azul, respectivamente). Os resíduos identificados por Logsdon e colaboradores (LOGSDON *et al.*, 2004) como importantes para a ligação à IL-10R2 são mostrados em laranja.

A resolução estrutural do complexo IL-22/IL-22R1 resultou em publicação no periódico FEBS Letters (BLEICHER *et al.*, 2008). Durante o período entre a aceitação do artigo e sua publicação *online*, o grupo de Logsdon e Walter publicou a estrutura do mesmo complexo (JONES *et al.*, 2008), utilizando, assim como em seu trabalho estrutural sobre a IL-22, proteína expressa em células de inseto. Mais uma vez, a glicosilação não resultou em modificações estruturais em regiões de densidade eletrônica bem definida, sendo as principais diferenças entre elas a resolução máxima obtida (1.9Å neste trabalho contra 2.5Å do artigo de Jones e colaboradores) e a modelagem de loops flexíveis – neste trabalho os resíduos 131-139 em IL-22 e 149-152 e 172-175 em IL-22R1 não foram modelados devido à baixa definição do mapa de densidade eletrônica nessa região. Na estrutura de Jones e colaboradores (JONES *et al.*, 2008), todos os resíduos foram modelados, além de açúcares ligados covalentemente.

Poucas semanas após a publicação das duas estruturas, um artigo foi publicado no *Journal of Molecular Biology* (submetido antes da disponibilidade das estruturas dos complexos IL-22/IL-22R1)

com um longo estudo de mutagênese sítio-dirigida, em que todos os resíduos da IL-22 foram mutados para alanina (as posições em que originalmente já eram encontradas alaninas foram mutadas para glicinas) de forma a verificar o resultado da mutação na ligação à IL-22R1, à IL-10R2 e à IL-22BP (WU *et al.*, 2008). Os autores verificaram que cinco mutações em IL-22 (I161A, V72A, G165A, D71A e L169A) praticamente eliminam sua ligação à IL-22R1, e três delas (D67A, R73A e K172A) também diminuem a ligação à IL-22BP, concluindo que os sítios de ligação de IL-22R1 e IL-22BP em IL-22 são coincidentes ou superpostos. Quanto à IL-10R2, os autores identificaram seis resíduos em IL-22 (T56, Y51, R55, N54, F121 e E117) que interferem fortemente na ligação da citocina com a segunda cadeia de seu receptor. Tais resíduos, alguns deles já observados em trabalho previamente publicado por Logsdon e colaboradores (LOGSDON *et al.*, 2004), encontram-se na região da IL-22 que se ligaria à IL-10R2 de acordo com o modelo proposto quando de nossa publicação, com base no complexo IL-4/IL-4R α / γ_c .

As estatísticas de processamento e refinamento da estrutura do complexo IL-22/IL-22R1 encontra-se na tabela 9.

Tabela 9 - Dados cristalográficos da estrutura do complexo IL-22/IL-10R2. Valores entre parênteses referem-se à última faixa de resolução (um total de 10 faixas foi utilizado). Os dados foram integrados e escalonados utilizando o pacote HKL2000 (OTWINOWSKI; MINOR, 1997), e a estrutura refinada utilizando o Phenix.refine (ADAMS *et al.*, 2002).

Resolução	19.81Å – 1.90Å (1.97Å – 1.90Å)
Grupo espacial	C2
Parâmetros de rede (Å, °)	a = 49.2 b = 79.2 c = 92.0 β = 94.6
Completeza (%)	98.1 (87.0)
Reflexões utilizadas no refinamento	25984
Total de reflexões	27359 (2304)
Reflexões utilizadas para cálculo de R _{free}	1375
I/ σ	7.14 (2.82)
R _{merge} (%)	10.6 (30.5)
Multiplicidade	2.8 (2.4)
R _{factor} (%)	19.0
R _{free} (%)	23.4
RMSD comprimentos de ligação (Å)	0.012
RMSD ângulos de ligação (°)	1.56
B-fator médio (modelo)	30.24
B-fator médio (gráfico de Wilson)	26.61

6.5 Laminarinase de *Rhodothermus marinus*

Rhodothermus marinus é uma bactéria marinha termofílica, capaz de viver a temperaturas acima de 65°. Um gene de *R. marinus* que codifica para uma glicosil hidrolase da família 16 (*lamR*) foi clonado em 1998 por Krah e colaboradores (KRAH *et al.*, 1998), e sua expressão e caracterização, pelos mesmos autores, mostrou tratar-se de uma laminarinase, dominada por estruturas do tipo fita- β , e altamente termoestável.

A laminarinase de *R. marinus* foi cristalizada no início dos anos 2000 pela então doutoranda Adriana Lucely Rojas em colaboração com o pesquisador visitante Alexander Golubev. Embora tenha sido possível obter cristais de boa qualidade, o faseamento do conjunto de dados coletado não foi possível à época uma vez que não havia estruturas similares disponíveis e o conjunto coletado a partir de um cristal derivatizado com Césio não resultou em localização dos átomos pesados e conseqüente solução do problema das fases.

Motivado pelos recentes avanços nos algoritmos de substituição molecular – em especial o advento do programa *Phaser* (MCCOY *et al.*, 2007) – e o aumento em estruturas disponíveis no banco de dados PDB, o autor tentou fasear o conjunto de dados obtido previamente neste grupo de pesquisa. Primeiramente, observou-se que a indexação do padrão de difração, efetuada utilizando o programa *Mosflm* (Leslie, 1992), mostrou duas possibilidades para simetria: P2 ou C222 (ou os grupos obtidos a partir da troca de eixos de ordem 2 por eixos parafusos 2₁), sem aumento considerável de penalidade para a simetria mais alta C222. Uma busca no banco de dados PDB mostrou existir uma estrutura resolvida de laminarinase em 2006 (código PDB 2CL2). Tratava-se da laminarinase Lam16A de *Phaerochaete chrysosporium*, determinada pelo método de faseamento por dispersão anômala a um único comprimento de onda. A utilização do software *Phaser* (MCCOY *et al.*, 2007) permitiu uma substituição molecular bem-sucedida, porém sem convergência do modelo após a correção da seqüência, mesmo utilizando-se simetrias diferentes.

A análise do conjunto de dados pela ferramenta *phenix.xtriage*, parte do pacote de programas cristalográficos *Phenix* (ADAMS *et al.*, 2002), revelou tratar-se de um caso de cristal com geminação pseudo-merohedral, em que reflexões $h k l$ estavam superpostas a reflexões $-h, -k, h+l$.

Cristais geminados no sistema monoclinico são relativamente raros, e geralmente ocorrem quando se têm a próximo de c , ou β próximo de 90° . Porém, é possível que este efeito ocorra também no caso fortuito de uma célula unitária em que $c \cdot \cos(\beta) = -a/2$, em que as redes tornam-se geminadas pela superposição de a e b em direções opostas. Este fenômeno foi previamente observado por Declercq e Evrard (DECLERCQ; EVRARD, 2001) e Rudolph e colaboradores (RUDOLPH *et al.*, 2004) e seu efeito está mostrado na figura 22.

Figura 22 - Arranjo de uma célula $P2_1$ para permitir cristais geminados e emulação de uma célula $C222_1$. A célula unitária (em azul) está projetada ao longo de b . Uma rotação de 180° em torno de um eixo perpendicular a a e b (em vermelho) leva à superposição das duas redes primitivas quando $c \cdot \cos\beta = -a/2$. A célula do segundo domínio é mostrada em preto. O efeito é a emulação de centragem, com a célula resultante mostrada pela linha pontilhada. **Fonte:** Rudolph *et al.*, 2001.

Uma vez que cristais geminados de forma pseudo-merohedral produzem, para algumas reflexões, intensidades que são uma soma das intensidades produzidas pelas duas redes (pesadas pela fração de geminação α , que indica a proporção existente entre os dois domínios geminados), em muitos casos o refinamento estrutural é problemático. Frações de geminação próximas de 0 indicam que um dos domínios consiste na maior parte do cristal, e, nesses casos, em geral é possível resolver a estrutura sem a necessidade de tratamento adicional dos dados para correção de intensidades. O problema torna-se mais complicado para frações próximas de 50%, o que foi o caso da estrutura da laminarinase de *R. marinus*, em que as análises preliminares pelo programa *phenix.xtriage* (ADAMS

et al., 2002) indicaram valores próximos de 40%. Felizmente, existem atualmente diversas ferramentas para permitir a solução de estruturas de cristais geminados. Como previamente citado, há duas abordagens para este problema: gerar um conjunto de dados “corrigido” em que as intensidades corresponderiam à de um único domínio, ou utilizar funções de refinamento estrutural que já levem em conta a contribuição dos dois domínios.

O pacote de programas cristalográficos *Phenix* utiliza a segunda abordagem, que pode ser usada tanto quando do refinamento do modelo, como quando de sua completa reconstrução a partir do mapa de densidade eletrônica. A estrutura da laminarinase de *R. marinus* foi então reconstruída (utilizando sua sequência correta) utilizando esta abordagem.

A reconstrução do modelo com a sequência correta pôde ser feita facilmente utilizando funções que levam em conta a geminação, e o refinamento da estrutura resultou em um modelo de alta qualidade, com parâmetros estatísticos e geométricos comparáveis ao de estruturas de mesma resolução (1.95Å) obtidos através de conjuntos de dados de cristais não-geminados. As estatísticas de processamento de dados e refinamento estrutural da laminarinase de *R. marinus* estão mostrados na tabela 10.

Tabela 10 - Dados cristalográficos da estrutura da laminarinase de *R. marinus*. Valores entre parênteses referem-se à última faixa de resolução (um total de 10 faixas foi utilizado). Os dados foram integrados e escalonados utilizando o pacote HKL2000 (OTWINOWSKI; MINOR, 1997), e a estrutura refinada utilizando o Phenix.refine (ADAMS *et al.*, 2002).

Resolução	25.922Å – 1.95Å (2.06Å – 1.95Å)
Grupo espacial	P2 ₁
Parâmetros de rede (Å, °)	a = 52.218 b = 108.288 c = 64.588 β = 113.9
Completeza (%)	97.8 (94.9)
Reflexões utilizadas no refinamento	85360
Reflexões utilizadas para cálculo de R _{free}	4709
I/ σ	5.9 (1.9)
R _{merge} (%)	10.4 (37.2)
Multiplicidade	4.0 (4.1)
R _{factor} (%)	16.27
R _{free} (%)	19.15
Fração de geminação α (%)	40.6
RMSD comprimentos de ligação (Å)	0.005
RMSD ângulos de ligação (°)	0.947
B-fator médio (modelo)	18.9
B-fator médio (gráfico de Wilson)	16.5

Como previsto por Krah e colaboradores (KRAH *et al.*, 1998), a estrutura da laminarinase de *R. marinus* é dominada por estruturas β , apresentando-se como um sanduíche de folhas- β formadas cada uma por seis fitas- β antiparalelas, além de um pequeno segmento de hélice, como visto na figura 23. A estrutura é bastante similar à laminarinase de *P. chrysosporium*, sendo o rmsd observado após o alinhamento estrutural das duas estruturas de 1.83Å ao longo de 201 resíduos estruturalmente equivalentes.

Figura 23 - Estrutura da Laminarinase de *Rhodothermus marinus*.

A análise do gráfico de Ramachandran aponta apenas dois resíduos fora das regiões preferenciais. Tal fato não parece ser uma falha no modelo e sim uma característica da estrutura, uma vez que se trata do mesmo resíduo em cadeias diferentes (após um protocolo de refinamento em que não foram aplicadas restrições de simetria não-cristalográfica), e para os quais o mapa de densidade eletrônica apresenta-se bem definido, como visto na figura 24.

Figura 24 - Detalhe dos resíduos Tyr91 (na cadeia A à esquerda e na cadeia B à direita) e seus resíduos adjacentes. Apesar de estarem fora do gráfico de Ramachandran, os mapas de densidade eletrônica mostram que sua orientação está bem definida. O mapa 2Fo-Fc, com valor de $\sigma=1.0$, foi traçado até 2.0Å de distância aos átomos presentes nos resíduos 90-92 de cada cadeia.

As glicosil hidrolases da família 16, à qual pertence à laminarinase de *R. marinus*, são caracterizadas por um motivo WDEIDIE (em alguns casos o ácido aspártico após o triptofano é substituído por uma inserção de quatro resíduos variáveis), onde se encontra o sítio ativo da proteína. Os três últimos resíduos ácidos deste motivo são responsáveis pela catálise. A laminarinase de *Rhodothermus marinus* é um dos casos em que, na sequência consenso, o triptofano é separado do primeiro ácido glutâmico por uma inserção de quatro resíduos ao invés de um ácido aspártico, sendo a sequência na região do sítio ativo igual a WPDNGEIDIME. A análise desta região na estrutura cristalográfica mostra que, apesar da inserção, o triptofano mantém-se fisicamente próximo ao primeiro ácido glutâmico, como visto na figura 25.

Figura 25 - O sítio ativo da Laminarinase de *R. marinus*. Além dos resíduos ácidos catalíticos, o triptofano também é altamente conservado, e mantém-se próximo ao primeiro ácido glutâmico mesmo quando a distância em sequência entre os dois é maior, o que é o caso nesta laminarinase.

7. Conclusões

A implementação da análise de acoplamentos estatísticos e sua aplicação a diferentes famílias de proteínas permitiu, num primeiro momento, a melhor compreensão da gama de informações potencialmente adquiríveis com a técnica, além de suas limitações e possíveis artefatos. A utilização da técnica tal qual originalmente descrita por Rama Ranganathan e colaboradores, que a apresentaram pela primeira vez em 1999 (LOCKLESS; RANGANATHAN, 1999) e a desenvolveram ao longo de trabalhos subsequentes (SUEL *et al.*, 2003; SHULMAN *et al.*, 2004), permitiu observar, na família das hexoquinases, a correlação entre a distribuição de resíduos próximos ao sítio ativo destas proteínas e também àqueles nas regiões de dobradiça responsáveis pelo movimento de ajuste induzido (KUSER *et al.*, 2008).

O estudo de superóxido dismutases de Fe/Mn efetuado por Bachega e colaboradores (BACHEGA *et al.*, 2009), com participação do autor desta tese, revelou que uma quantidade muito maior de informações pode ser obtida pela técnica ao se observar não apenas o conjunto final de resíduos acoplados entre si, mas ao analisar o comportamento de frequências de tais resíduos de forma a compreender os motivos que levam ao acoplamento. Esta nova metodologia, que no estudo efetuado por Bachega e colaboradores permitiu identificar resíduos responsáveis pela discriminação entre superóxido dismutases específicas para ferro e as específicas para manganês, foi crucial para o sucesso obtido no estudo das Proteínas Tirosina Fosfatases, principal objeto de estudo desta tese de doutorado.

Uma vez que as proteínas tirosina fosfatases podem ser classificadas em três subfamílias diferentes (PTPs clássicas, DSPs e LMW-PTPs), a análise permitiu um estudo detalhado das particularidades de cada subfamília. Enquanto as PTPs clássicas já tinham sido previamente caracterizadas com respeito aos seus motivos conservados, as DSPs e LMW-PTPs não tinham sido estudadas sob este aspecto. Desta forma, a primeira contribuição deste trabalho para estudiosos de proteínas tirosina fosfatases é a apresentação de um quadro de motivos e resíduos conservados em todas as três subfamílias, atualizado no caso das PTPs clássicas e inédito para DSPs e LMW-PTPs, contendo não apenas a indicação explícita da conservação entre tais resíduos e motivos, mas também a interdependência entre posições distintas, informações que, como se confirmou a partir da discussão destes resultados, pode ser útil em algumas questões de interesse biológico.

A inspeção das posições correlacionadas em proteínas tirosina fosfatases clássicas mostrou que a dimerização nestas proteínas, embora tenha sido descrita como um mecanismo importante de regulação, não parece estar restrita à uma simples dicotomia entre “dimerização do tipo α ” e “dimerização do tipo μ ”, idéia prevalecente desde a descoberta, nos anos 90, de que a fosfatase RPTP α apresentava um dímero auto-inibitório pela oclusão de seus sítios ativos enquanto a RPTP μ apresentava um dímero com os sítios expostos ao solvente. A descoberta de novas estruturas de proteínas tirosina fosfatases tem mostrado que há outras possibilidades para dimerização, enquanto a análise de acoplamentos estatísticos também não dá suporte a esta idéia pela ausência de correlação entre os resíduos que formam a interface de dimerização em RPTP μ (embora haja correlação entre os resíduos relativos à interface de dimerização em RPTP α , ela é explicada pelo fato de que esses resíduos têm importância crucial para a função da proteína, e, portanto, obrigatoriamente apresentariam conservação e correlação).

Por outro lado, a inspeção de resíduos conservados e correlacionados tornou possível a interpretação de um conjunto de mutações na proteína tirosina fosfatase SHP-2 relacionadas a patologias. Embora boa parte dos casos seja resultado do já esperado fato de que mutações em posições altamente conservadas costumam resultar em proteínas com estrutura e/ou função prejudicada, alguns deles só puderam ser explicados ao observar casos de conservação não independente, em que a ocorrência de um dado aminoácido estava condicionada a existência de um ou mais aminoácidos em outras posições.

Uma vez que mecanismos de especificidade em PTPs clássicas são pouco conhecidos, tentou-se verificar a correlação entre posições previamente descritas como importantes para seletividade em casos específicos e outras posições do alinhamento, o que poderia sugerir novas posições a serem estudadas por mutagênese como candidatas a posições-chave para a especificidade. No entanto, observou-se pouca correlação entre tais posições, o que eliminou tal possibilidade.

Finalmente, a análise de acoplamentos estatísticos mostrou uma interessante aplicação quando do estudo das proteínas tirosina fosfatases de baixo peso molecular. Enquanto as PTPs clássicas e

DSPs apresentaram como posições altamente correlacionadas entre si um pequeno grupo de aminoácidos com tendência a estarem presentes de forma simultânea (dois aglomerados no caso de PTPs clássicas e um no caso de DSPs), no caso das LMW-PTPs uma quantidade considerável de posições apresentou correlação conjunta, e, ao contrário dos aglomerados previamente observados, cada posição possuía duas possibilidades para conservação, o que dividia o alinhamento em duas “classes”. Cada uma das classes apresentava um padrão determinado para a conservação destas posições, o que indicava que o alinhamento de LMW-PTPs deveria conter proteínas que deveriam possuir alguma característica particular presente apenas em uma das classes (ou duas características auto-excludentes). A revisão bibliográfica sobre proteínas tirosina fosfatases de baixo molecular revelou uma observação prévia de que arsenato redutases da família ArsC eram homólogas às LMW-PTPs, em um caso de evolução divergente. Por possuírem elementos de sequência similares aos observados em LMW-PTPs, elas deveriam ser diferenciadas destas quando de buscas em genomas através de seu contexto genômico, uma vez que as arsenato redutases geralmente são parte de um *operon* contendo outras proteínas relacionadas à degradação de xenobióticos. Com posse desta informação, compararam-se as particularidades de sequência observadas nas duas classes dos alinhamentos a uma LMW-PTP (PPAC humana) e a uma arsenato redutase (ArsC de *Bacillus subtilis*). Observou-se que uma das classes conserva resíduos presentes em PPAC e não na ArsC e vice-versa. Tal resultado sugere fortemente que LMW-PTPs e ArsCs possam, ao contrário do que se supunha previamente, ser diferenciadas apenas com base em sua sequência, hipótese que provavelmente não tenha sido levantada pelo fato de que os resíduos característicos das duas classes não estão necessariamente entre os resíduos essenciais para a catálise.

A análise de acoplamentos estatísticos foi aplicada também ao estudo das peroxidases, em um trabalho suplementar ao estudo estrutural da peroxidase da palmeira *R. Regia*, efetuado neste grupo de pesquisa. Neste trabalho, observou-se, assim como no alinhamento de LMW-PTPs, uma clara separação entre duas categorias no mesmo alinhamento, o que, neste caso, levava aos resíduos capazes de discriminar as peroxidases entre aquelas das Classes I e das Classes II e III, segundo a subclassificação usual desta família de proteínas.

O autor participou também de alguns trabalhos em cristalografia de proteínas, em graus variados de dificuldade metodológica. A transtirretina humana, embora consista em um caso trivial de substituição molecular em excelente resolução (1.75Å), apresenta a interessante propriedade de formação de um sítio de ligação na proteína a partir de duas unidades simetricamente relacionadas, o que levava a duas conformações possíveis (apresentando mesma simetria) para o ligante. A Hexoquinase PI de *Saccharomyces cerevisiae*, por outro lado, foi resolvida a partir de um conjunto de dados de resolução moderada (2.95Å), e, portanto, precisou ser refinada utilizando protocolos que fossem capazes de gerar um modelo cristalográfico realista em face de tal limitação. O complexo de Interleucina-22 e seu receptor IL-22R1, embora derivado de um cristal de boa resolução (1.9Å), apresentou empecilhos quando do seu faseamento, pela dificuldade em localizar na unidade assimétrica um modelo para o receptor IL-22R1, sem estrutura conhecida antes da publicação deste trabalho. A busca por moldes alternativos permitiu o correto faseamento da estrutura, cujo refinamento resultou em um modelo cristalográfico satisfatório (BLEICHER *et al.*, 2008). Finalmente, a resolução da estrutura da Laminarinase de *Rhodothermus marinus* levou à necessidade da utilização de protocolos para cristais apresentando geminação. Após a detecção e tratamento de tal problema, foi possível refinar sua estrutura até um modelo de qualidade comparável aos obtidos para cristais sem tal patologia em mesma faixa de resolução.

ADAMS, P. D. *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallographica Section D: Biological Crystallography*, v. 58, p. 1948-1954, Nov. 2002.

ALDRED, A. R. *et al.* The cerebral expression of plasma protein genes in different species. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, v. 111, n. 1, p. 1-15, May 1995.

ALONSO, A. *et al.* Protein tyrosine phosphatases in the human genome. *Cell*, v. 117, n. 6, p. 699-711, June 2004.

ALTSCHUL, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, v. 25, n. 17, p. 3389-402, Sept. 1997.

AMMON, H. L. *et al.* Preliminary crystal structure of *Acinetobacter glutaminasificans* glutaminase-asparaginase. *Journal of Biological Chemistry*, v. 263, n. 1, p. 150-6, Jan. 1988.

ANDERSEN, J. N. *et al.* Structural and evolutionary relationships among protein tyrosine phosphatase domains. *Molecular Cell Biology*, v. 21, n. 21, p. 7117-36, Nov. 2001.

ANDERSON, C. M. *et al.* Sequencing a protein by x-ray crystallography. II. Refinement of yeast hexokinase B co-ordinates and sequence at 2.1 Å resolution. *Journal of Molecular Biology*, v. 123, n. 1, p. 15-33, July 1978.

_____. Space-filling models of kinase clefts and conformation changes. *Science*, v. 204, n. 4391, p. 375-80, Apr. 1979.

BACHEGA, J. F. R. *et al.* Systematic structural studies of iron superoxide dismutases from human parasites and a statistical coupling analysis of metal binding specificity. *Proteins*, 2009. No prelo.

BARTALENA, L.; ROBBINS, J. Thyroid hormone transport proteins. *Clinical Laboratory Medical*, v. 13, n. 3, p. 583-98, Sept. 1993.

BENNETT, M. S. *et al.* *Bacillus subtilis* arsenate reductase is structurally and functionally similar to low molecular weight protein tyrosine phosphatases. *Proceedings of the National Academy of Science of the United States of America*, v. 98, n. 24, p. 13577-82, Nov. 2001.

BENNETT, W. S., JR.; STEITZ, T. A. Structure of a complex between yeast hexokinase A and glucose. I. Structure determination and refinement at 3.5 Å resolution. *Journal of Molecular Biology*, v. 140, n. 2, p. 183-209, June 1980.

BERNAL, J. D.; CROWFOOT, D. X-ray photographs of crystalline pepsin. *Nature*, v. 133, p. 794-795, Jan.-June 1934.

BERTOLA, D. R. *et al.* PTPN11 gene analysis in 74 Brazilian patients with Noonan syndrome or Noonan-like phenotype. *Genetic Testing*, v. 10, n. 3, p. 186-91, Fall 2006.

_____. Clinical variability in a Noonan syndrome family with a new PTPN11 gene mutation. *American Journal of Medical Genetics A*, v. 130A, n. 4, p. 378-83, Nov. 2004.

_____. Neurofibromatosis-Noonan syndrome: molecular evidence of the concurrence of both disorders in a patient. *American Journal of Medical Genetics A*, v. 136, n. 3, p. 242-5, July 2005.

BILWES, A. M. *et al.* Structural basis for inhibition of receptor protein-tyrosine phosphatase- α by dimerization. *Nature*, v. 382, n. 6591, p. 555-9, Aug. 1996.

BINDER, G. *et al.* PTPN11 mutations are associated with mild growth hormone resistance in individuals with Noonan syndrome. *Journal of Clinical Endocrinology Metabolism*, v. 90, n. 9, p. 5377-81, Sept. 2005.

BLEICHER, L. *et al.* Crystal structure of the IL-22/IL-22R1 complex and its implications for the IL-22 signaling mechanism. *FEBS Letters*, v. 582, n. 20, p. 2985-92, Sept. 2008.

BORGSTAHL, G. E. *et al.* 1.4 Å structure of photoactive yellow protein, a cytosolic photoreceptor: unusual fold, active site, and chromophore. *Biochemistry*, v. 34, n. 19, p. 6278-87, May 1995.

BOWIE, J. U. *et al.* A method to identify protein sequences that fold into a known 3-dimensional structure. *Science*, v. 253, n. 5016, p. 164-170, July 1991.

BRUNGER, A. T. Free R-value - a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, v. 355, n. 6359, p. 472-475, Jan. 1992.

BUCCIANINI, M. *et al.* The low Mr phosphotyrosine protein phosphatase behaves differently when phosphorylated at Tyr131 or Tyr132 by Src kinase. *FEBS Letters*, v. 456, n. 1, p. 73-8, July 1999.

CAMPS, M. *et al.* Dual specificity phosphatases: a gene family for control of MAP kinase function. *Faseb Journal*, v. 14, n. 1, p. 6-16, Jan. 2000.

CHI, C. N. *et al.* Reassessing a sparse energetic network within a single protein domain. *Proceedings of the National Academy of Sciences of the United States of America*, v. 105, n. 12, p. 4679-84, Mar. 2008.

CONTI, E. *et al.* A novel PTPN11 mutation in LEOPARD syndrome. *Human Mutation*, v. 21, n. 6, p. 654, June 2003.

CURCIC-STOJKOVIC, O. *et al.* Noonan's syndrome: male Turner's syndrome, Turner-like syndrome. *Medicinski Pregled*, v. 31, n. 7-8, p. 299-303, 1978.

DAUTER, Z. Twinned crystals and anomalous phasing. *Acta Crystallographica D: Biological Crystallography*, v. 59, n. Pt 11, p. 2004-16, Nov. 2003.

_____. Current state and prospects of macromolecular crystallography. *Acta Crystallographica D: Biological Crystallography*, v. 62, p. 1-11, Jan. 2006.

DECLERCQ, J. P.; EVRARD, C. A twinned monoclinic crystal form of human peroxiredoxin 5 with eight molecules in the asymmetric unit. *Acta Crystallographica D: Biological Crystallography*, v. 57, p. 1829-1835, Dec. 2001.

DEN HERTOOG, J. *et al.* Protein tyrosine phosphatases: regulatory mechanisms. *Febs Journal*, v. 275, n. 5, p. 831-47, Mar. 2008.

DIGILIO, M. C. *et al.* Grouping of multiple-lentigines/LEOPARD and Noonan syndromes on the PTPN11 gene. *American Journal of Human Genetics*, v. 71, n. 2, p. 389-94, Aug. 2002.

DIMA, R. I.; THIRUMALAI, D. Determination of network of residues that regulate allostery in protein families using sequence analysis. *Protein Science*, v. 15, n. 2, p. 258-68, Feb. 2006.

DUMAS, C.; VAN DER LEE, A. Macromolecular structure solution by charge flipping. *Acta Crystallographica D: Biological Crystallography*, v. D64, n. Pt 8, p. 864-73, Aug. 2008.

DUMOUTIER, L. *et al.* Human interleukin-10-related T cell-derived inducible factor: molecular cloning and functional characterization as an hepatocyte-stimulating factor. *Proceedings of the National Academy of Sciences of the United States of America*, v. 97, n. 18, p. 10144-9, Aug. 2000.

FINN, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Research*, v. 36, n. Database issue, p. D281-8, Jan. 2008.

FLINT, A. J. *et al.* Development of "substrate-trapping" mutants to identify physiological substrates of protein tyrosine phosphatases. *Proceedings of the National Academy of Sciences of the United States of America*, v. 94, n. 5, p. 1680-5, Mar. 1997.

FODOR, A. A.; ALDRICH, R. W. On evolutionary conservation of thermodynamic coupling in proteins. *Journal of Biological Chemistry*, v. 279, n. 18, p. 19046-50, Apr. 2004.

GASTEIGER, E. *et al.* ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research*, v. 31, n. 13, p. 3784-3788, July 2003.

GORLIN, R. J. *et al.* Multiple lentigenes syndrome. *American Journal of Diseases of Children*, v. 117, n. 6, p. 652-62, June 1969.

GREEN, N. S. *et al.* Genistein, a natural product from soy, is a potent inhibitor of transthyretin amyloidosis. *Proceedings of the National Academy of Sciences of the United States of America*, v. 102, n. 41, p. 14545-50, Oct. 2005.

GU, J.; GU, X. Natural history and functional divergence of protein tyrosine kinases. *Gene*, v. 317, n. 1-2, p. 49-57, Oct. 2003.

GUAN, K. L. *et al.* Cloning and expression of a protein-tyrosine-phosphatase. *Proceedings of the National Academy of Sciences of the United States of America*, v. 87, n. 4, p. 1501-5, Feb. 1990.

HATLEY, M. E. *et al.* Allosteric determinants in guanine nucleotide-binding proteins. *Proceedings of the National Academy of Sciences of the United States of America*, v. 100, n. 24, p. 14445-50, Nov. 2003.

HENDRIKS, W. J. *et al.* Protein tyrosine phosphatases: functional inferences from mouse models and human diseases. *Febs Journal*, v. 275, n. 5, p. 816-30, Mar. 2008.

HOFFMANN, K. M. *et al.* The crystal structure of domain 1 of receptor protein-tyrosine phosphatase mu. *Journal of Biological Chemistry*, v. 272, n. 44, p. 27505-8, Oct. 1997.

HOPE, H. Cryocrystallography of biological macromolecules: a generally applicable method. *Acta Crystallographica B*, v. 44 (Pt 1), p. 22-6, Feb. 1988.

- HUBBARD, S. R.; TILL, J. H. Protein tyrosine kinase structure and function. *Annual Review of Biochemistry*, v. 69, p. 373-98, 2000.
- ISHIBASHI, T. *et al.* A novel dual specificity phosphatase induced by serum stimulation and heat shock. *Journal of Biological Chemistry*, v. 269, n. 47, p. 29897-902, Nov. 1994.
- IULIANO, R. *et al.* The tyrosine phosphatase PTPRJ/DEP-1 genotype affects thyroid carcinogenesis. *Oncogene*, v. 23, n. 52, p. 8432-8, Nov. 2004.
- JIANG, J.; SWEET, R. M. Protein Data Bank depositions from synchrotron sources. *Journal of Synchrotron Radiation*, v. 11, n. Pt 4, p. 319-27, July 2004.
- JONES, B. C. *et al.* Structure of IL-22 bound to its high-affinity IL-22R1 chain. *Structure*, v. 16, n. 9, p. 1333-44, Sept. 2008.
- KATOH, K. *et al.* MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, v. 30, n. 14, p. 3059-3066, July 2002.
- KATZEN, H. M.; SCHIMKE, R. T. Multiple forms of hexokinase in the rat: tissue distribution, age dependency, and properties. *Proceedings of the National Academy of Sciences of the United States of America*, v. 54, n. 4, p. 1218-25, Oct. 1965.
- KEEGAN, R. M.; WINN, M. D. MrBUMP: an automated pipeline for molecular replacement. *Acta Crystallographica D: Biological Crystallography*, v. 64, n. Pt 1, p. 119-24, Jan. 2008.
- KENDREW, J. C. *et al.* Structure of myoglobin - 3-dimensional Fourier synthesis at 2 Å resolution. *Nature*, v. 185, n. 4711, p. 422-427, 1960.
- KEREN, B. *et al.* PTPN11 mutations in patients with LEOPARD syndrome: a french multicentric experience. *Journal of Medical Genetics*, v. 41, n. 11, p. e117, Nov. 2004.
- KLEYWEGT, G. J. Validation of protein crystal structures. *Acta Crystallographica D: Biological Crystallography*, v. 56, p. 249-265, Mar. 2000.
- KLEYWEGT, G. J.; JONES, T. A. Homo crystallographicus--quo vadis? *Structure*, v. 10, n. 4, p. 465-72, Apr. 2002.
- KOTENKO, S. V. The family of IL-10-related cytokines and their receptors: related, but to what extent? *Cytokine Growth Factor Review*, v. 13, n. 3, p. 223-40, June 2002.
- KRAH, M. *et al.* The laminarinase from thermophilic eubacterium *Rhodothermus marinus*--conformation, stability, and identification of active site carboxylic residues by site-directed mutagenesis. *European Journal Biochemistry*, v. 257, n. 1, p. 101-11, Oct. 1998.
- KRATZ, C. P. *et al.* The mutational spectrum of PTPN11 in juvenile myelomonocytic leukemia and Noonan syndrome/myeloproliferative disease. *Blood*, v. 106, n. 6, p. 2183-5, Sept. 2005.
- KRISSINEL, E.; HENRICK, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica D: Biological Crystallography*, v. 60, n. Pt 12 Pt 1, p. 2256-68, Dec. 2004.

_____. Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology*, v. 372, n. 3, p. 774-97, Sept. 2007.

KUSER, P. *et al.* Crystal structure of yeast hexokinase PI in complex with glucose: a classical "induced fit" example revised. *Proteins*, v. 72, n. 2, p. 731-40, Aug. 2008.

LAPORTE, S. L. *et al.* Molecular and structural basis of cytokine receptor pleiotropy in the interleukin-4/13 system. *Cell*, v. 132, n. 2, p. 259-72, Jan. 2008.

LASKOWSKI, R. A. *et al.* Procheck - a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, v. 26, p. 283-291, Apr. 1993.

LEE, J. *et al.* Surface sites for engineering allosteric control in proteins. *Science*, v. 322, n. 5900, p. 438-442, Oct. 2008.

LEE, J. W. *et al.* Mutational analysis of PTPRT phosphatase domains in common human cancers. *Apmis*, v. 115, n. 1, p. 47-51, Jan. 2007.

LESUEUR, F. *et al.* Allelic association of the human homologue of the mouse modifier Ptprij with breast cancer. *Human Molecular Genetics*, v. 14, n. 16, p. 2349-56, Aug. 2005.

LI, J. *et al.* Temporal associations between interleukin 22 and the extracellular domains of IL-22R and IL-10R2. *International Immunopharmacology*, v. 4, n. 5, p. 693-708, May 2004.

LOCKLESS, S. W.; RANGANATHAN, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, v. 286, n. 5438, p. 295-9, Oct. 1999.

LOGSDON, N. J. *et al.* The IL-10R2 binding hot spot on IL-22 is located on the N-terminal helix and is dependent on N-linked glycosylation. *Journal of Molecular Biology*, v. 342, n. 2, p. 503-14, Sept. 2004.

_____. Comparison of interleukin-22 and interleukin-10 soluble receptor complexes. *Journal of Interferon Cytokine Research*, v. 22, n. 11, p. 1099-112, Nov. 2002.

LUBKOWSKI, J. *et al.* Refined crystal structure of *Acinetobacter glutaminasificans* glutaminase-asparaginase. *Acta Crystallographica D: Biological Crystallography*, v. 50, n. Pt 6, p. 826-32, Nov. 1994.

LUO, L. *et al.* Loss of heterozygosity in human aberrant crypt foci (ACF), a putative precursor of colon cancer. *Carcinogenesis*, v. 27, n. 6, p. 1153-9, June 2006.

LUTHY, R. *et al.* Assessment of protein models with 3-dimensional profiles. *Nature*, v. 356, n. 6364, p. 83-85, Mar. 1992.

MAHESHWARI, M. *et al.* PTPN11 mutations in Noonan syndrome type I: detection of recurrent mutations in exons 3 and 13. *Human Mutation*, v. 20, n. 4, p. 298-304, Oct. 2002.

MATOZO, H. C. *et al.* Low-resolution structure and fluorescence anisotropy analysis of protein tyrosine phosphatase eta catalytic domain. *Biophysical Journal*, v. 92, n. 12, p. 4424-32, June 2007.

MCCOY, A. J. *et al.* Phaser crystallographic software. *Journal of Applied Crystallography*, v. 40, p. 658-674, Aug. 2007.

MCREE, D. E. *et al.* Crystallographic structure of a photoreceptor protein at 2.4 Å resolution. *Proceedings of the National Academy of Sciences of the United States of America*, v. 86, n. 17, p. 6533-7, Sept. 1989.

MODESTI, A. *et al.* Cloning, expression and characterisation of a new human low Mr phosphotyrosine protein phosphatase originating by alternative splicing. *FEBS Letters*, v. 431, n. 1, p. 111-5, July 1998.

MOUREY, R. J. *et al.* A novel cytoplasmic dual specificity protein tyrosine phosphatase implicated in muscle and neuronal differentiation. *Journal of Biological Chemistry*, v. 271, n. 7, p. 3795-802, Feb. 1996.

MURSHUDOV, G. N. *et al.* Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallographica D: Biological Crystallography*, v. 53, n. Pt 3, p. 240-55, May 1997.

NAGEM, R. A. *et al.* Crystal structure of recombinant human interleukin-22. *Structure*, v. 10, n. 8, p. 1051-62, Aug. 2002.

OSTANIN, K. *et al.* Cloning and characterization of a *Saccharomyces cerevisiae* gene encoding the low molecular weight protein-tyrosine phosphatase. *Journal of Biological Chemistry*, v. 270, n. 31, p. 18491-9, Aug. 1995.

OSZLANYI, G.; SUTO, A. Ab initio structure solution by charge flipping. *Acta Crystallographica A*, v. 60, n. Pt 2, p. 134-41, Mar. 2004.

OTWINOWSKI, Z.; MINOR, W. Processing of x-ray diffraction data collected in oscillation mode. In: CARTER JR, C. W.; SWEET, R. M. (Ed.). *Methods in Enzymology*. New York: Academic Press, 1997. p. 307-326.

PALLEN, C. J. *et al.* Purification of a phosphotyrosine phosphatase that dephosphorylates the epidermal growth factor receptor autophosphorylation sites. *Annals of the New York Academy of Science*, v. 551, p. 299-308, 1988.

_____. Purification and characterization of placental membrane phosphotyrosine phosphatases. *Cold Spring Harb Symposia Quantitative Biology*, v. 53 Pt 1, p. 447-54, 1988.

PERUTZ, M. F. *et al.* Structure of haemoglobin - 3-dimensional Fourier synthesis at 5.5-Å resolution, obtained by x-ray analysis. *Nature*, v. 185, n. 4711, p. 416-422, 1960.

PETERS, G. H. *et al.* Residue 259 in protein-tyrosine phosphatase PTP1B and PTPalpha determines the flexibility of glutamine 262. *Biochemistry*, v. 43, n. 26, p. 8418-28, July 2004.

PULIDO, R.; HOOFT VAN HUIJSDUIJNEN, R. Protein tyrosine phosphatases: dual-specificity phosphatases in health and disease. *Febs Journal*, v. 275, n. 5, p. 848-66, Mar. 2008.

RAMPONI, G.; STEFANI, M. Structural, catalytic, and functional properties of low M(r), phosphotyrosine protein phosphatases. Evidence of a long evolutionary history. *International Journal Biochemistry Cell Biology*, v. 29, n. 2, p. 279-92, Feb. 1997.

RAUGEI, G. *et al.* Low molecular weight protein tyrosine phosphatases: small, but smart. *Cell Molecular Life Science*, v. 59, n. 6, p. 941-9, June 2002.

RUDOLPH, M. G. *et al.* Combined pseudo-merohedral twinning, non-crystallographic symmetry and pseudo-translation in a monoclinic crystal form of the gamma delta T-cell ligand T10. *Acta Crystallographica D: Biological Crystallography*, v. 60, p. 656-664, Apr. 2004.

RUIVENKAMP, C. *et al.* LOH of PTPRJ occurs early in colorectal cancer and is associated with chromosomal loss of 18q12-21. *Oncogene*, v. 22, n. 22, p. 3472-4, May 2003.

RUIVENKAMP, C. A. *et al.* PtpRJ is a candidate for the mouse colon-cancer susceptibility locus Scc1 and is frequently deleted in human cancers. *Nature Genetics*, v. 31, n. 3, p. 295-300, July 2002.

RUSS, W. P. *et al.* Natural-like function in artificial WW domains. *Nature*, v. 437, n. 7058, p. 579-83, Sept. 2005.

SALMEEN, A. *et al.* Redox regulation of protein tyrosine phosphatase 1B involves a sulphenyl-amide intermediate. *Nature*, v. 423, n. 6941, p. 769-73, June 2003.

SARAIVA, M. J. Transthyretin mutations in hyperthyroxinemia and amyloid diseases. *Human Mutation*, v. 17, n. 6, p. 493-503, June 2001.

SARAIVA, M. J. *et al.* Amyloid fibril protein in familial amyloidotic polyneuropathy, Portuguese type. Definition of molecular abnormality in transthyretin (prealbumin). *Journal Clinical Investigation*, v. 74, n. 1, p. 104-19, July 1984.

SARKOZY, A. *et al.* Clinical and molecular analysis of 30 patients with multiple lentigines LEOPARD syndrome. *Journal Medical Genetics*, v. 41, n. 5, p. e68, May 2004.

SHI, L. *et al.* The serine, threonine, and/or tyrosine-specific protein kinases and protein phosphatases of prokaryotic organisms: a family portrait. *FEMS Microbiology Reviews*, v. 22, n. 4, p. 229-53, Oct. 1998.

SHIN, D. Y. *et al.* A novel human ERK phosphatase regulates H-ras and v-raf signal transduction. *Oncogene*, v. 14, n. 22, p. 2633-9, June 1997.

SHULMAN, A. I. *et al.* Structural determinants of allosteric ligand activation in RXR heterodimers. *Cell*, v. 116, n. 3, p. 417-29, Feb. 2004.

SOCOLICH, M. *et al.* Evolutionary information for specifying a protein fold. *Nature*, v. 437, n. 7058, p. 512-8, Sept. 2005.

STEITZ, T. A. *et al.* High resolution x-ray structure of yeast hexokinase, an allosteric protein exhibiting a non-symmetric arrangement of subunits. *Journal of Molecular Biology*, v. 104, n. 1, p. 197-22, June 1976.

STRYER, L. *Biochemistry*. New York: WH Freeman Company, 1995.

STUCKEY, J. A. *et al.* Crystal structure of Yersinia protein tyrosine phosphatase at 2.5 Å and the complex with tungstate. *Nature*, v. 370, n. 6490, p. 571-5, Aug. 1994.

SUEL, G. M. *et al.* Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural Biology*, v. 10, n. 1, p. 59-69, Jan. 2003.

TABERNERO, L. *et al.* Protein tyrosine phosphatases: structure-function relationships. *Febs Journal*, v. 275, n. 5, p. 867-82, Mar. 2008.

TAKAHASHI, K. *et al.* A novel mutation in the PTPN11 gene in a patient with Noonan syndrome and rapidly progressive hypertrophic cardiomyopathy. *European Journal Pediatrics*, v. 164, n. 8, p. 497-500, Aug. 2005.

TARTAGLIA, M. *et al.* PTPN11 mutations in Noonan syndrome: molecular spectrum, genotype-phenotype correlation, and phenotypic heterogeneity. *American Journal Human Genetics*, v. 70, n. 6, p. 1555-63, June 2002.

_____. Diversity and functional consequences of germline and somatic PTPN11 mutations in human disease. *American Journal Human Genetics*, v. 78, n. 2, p. 279-90, Feb. 2006.

_____. Mutations in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome. *Nature Genetics*, v. 29, n. 4, p. 465-8, Dec. 2001.

TIGANIS, T.; BENNETT, A. M. Protein tyrosine phosphatase function: the substrate perspective. *Biochemical Journal*, v. 402, n. 1, p. 1-15, Feb. 2007.

TONKS, N. K. *et al.* Purification of the major protein-tyrosine-phosphatases of human placenta. *Journal Biological Chemistry*, v. 263, n. 14, p. 6722-30, May 1988.

VAGIN, A.; TEPLYAKOV, A. MOLREP: an automated program for molecular replacement. *Journal of Applied Crystallography*, v. 30, p. 1022-1025, Dec. 1997.

VAN MONTFORT, R. L. *et al.* Oxidation state of the active-site cysteine in protein tyrosine phosphatase 1B. *Nature*, v. 423, n. 6941, p. 773-7, June 2003.

VAN PUIJENBROEK, M. *et al.* Mass spectrometry-based loss of heterozygosity analysis of single-nucleotide polymorphism loci in paraffin embedded tumors using the MassEXTEND assay: single-nucleotide polymorphism loss of heterozygosity analysis of the protein tyrosine phosphatase receptor type J in familial colorectal cancer. *Journal Molecular Diagnostic*, v. 7, n. 5, p. 623-30, Nov. 2005.

WANG, Z. *et al.* Mutational analysis of the tyrosine phosphatome in colorectal cancers. *Science*, v. 304, n. 5674, p. 1164-6, May 2004.

WELINDER, K. G. *et al.* Structure of plant and fungal peroxidases. *Biochemical Society Transactions*, v. 20, n. 2, p. 337-40, May 1992.

WILSON, J. E. Hexokinases. *Reviews Physiology Biochemistry Pharmacology*, v. 126, p. 65-198, 1995.

WINTJENS, R. *et al.* Mn/Fe superoxide dismutase interaction fingerprints and prediction of oligomerization and metal cofactor from sequence. *Proteins*, v. 70, n. 4, p. 1564-77, Mar. 2008.

WU, P. W. *et al.* IL-22R, IL-10R2, and IL-22BP binding sites are topologically juxtaposed on adjacent and overlapping surfaces of IL-22. *Journal Molecular Biology*, v. 382, n. 5, p. 1168-83, Oct. 2008.

XU, T. *et al.* Structure of insect-cell-derived IL-22. *Acta Crystallographica D: Biological Crystallography*, v. 61, n. Pt 7, p. 942-50, July 2005.

YOSHIDA, R. *et al.* Two novel and one recurrent PTPN11 mutations in LEOPARD syndrome. *American Journal Medical Genetics A*, v. 130A, n. 4, p. 432-4, Nov. 2004.

ZENKER, M. *et al.* Genotype-phenotype correlations in Noonan syndrome. *Journal Pediatrics*, v. 144, n. 3, p. 368-74, Mar. 2004.

ZHANG, Z. Y. Protein tyrosine phosphatases: structure and function, substrate specificity, and inhibitor development. *Annual Review Pharmacology Toxicology*, v. 42, p. 209-34, 2002.

ZHANG, Z. Y. *et al.* Purification and characterization of the low molecular weight protein tyrosine phosphatase, Stp1, from the fission yeast *Schizosaccharomyces pombe*. *Biochemistry*, v. 34, n. 33, p. 10560-8, Aug. 1995.

Crystal structure of yeast hexokinase PI in complex with glucose: A classical “induced fit” example revised

Paula Kuser,¹ Fabio Cupri,² Lucas Bleicher,³ and Igor Polikarpov^{3*}

¹EMBRAPA Informática Agropecuária, Av. Dr. André Tosello, s/n, Cidade Universitária “Zeferino Vaz,” C. P. 6041, CEP 13083-970, Campinas, São Paulo, Brazil

²Laboratório Nacional de Luz Síncrotron, C.P. 6192, CEP 13083-970, Campinas, São Paulo, Brazil

³Grupo de Cristalografia, Instituto de Física em São Carlos, USP, Av. Trabalhador Saocarlene, 400 CEP 13560-970, São Carlos, SP, Brazil

ABSTRACT

Hexokinase is the first enzyme in the glycolytic pathway that catalyzes the transfer of a phosphoryl group from ATP to glucose to form glucose-6-phosphate and ADP. Two yeast hexokinase isozymes are known, namely PI and PII. Here we redetermined the crystal structure of yeast hexokinase PI from *Saccharomyces cerevisiae* as a complex with its substrate, glucose, and refined it at 2.95 Å resolution. Comparison of the holo-PI yeast hexokinase and apo-hexokinase structures shows in detail the rigid body domain closure and specific loop movements as glucose binds and sheds more light on structural basis of the “induced fit” mechanism of reaction in the HK enzymatic action. We also performed statistical coupling analysis of the hexokinase family, which reveals two co-evolved continuous clusters of amino acid residues and shows that the evolutionary coupled amino acid residues are mostly confined to the active site and the hinge region, further supporting the importance of these parts of the protein for the enzymatic catalysis.

Proteins 2008; 72:731–740.
© 2008 Wiley-Liss, Inc.

Key words: hexokinase X-ray structure; statistical coupling analysis; *Saccharomyces cerevisiae*.

INTRODUCTION

Hexokinase (ATP: D-hexose 6-phosphotransferase, EC 2.7.1.1; HK) is a member of the kinase family of tissue-specific isoenzymes. *Saccharomyces cerevisiae* genome codify for two isoenzymes of hexokinase: PI and PII, which are about 76% similar in their amino-acid sequences.^{1,2} Isoenzyme PII has a total of 486 amino acid residues, whereas PI has 485 residues. Hexokinase PII is the predominant hexose kinase in *Saccharomyces cerevisiae* in growth on glucose³ and is required for catabolite repression by glucose of expression of other genes.^{4,5} The yeast HKs are known to exist as phosphoproteins *in vitro*⁶ and *in vivo*⁷; *in vivo* phosphorylation site was identified as Ser15.⁸ It was demonstrated that the transformed cells with the hexokinase PII S15A mutant gene could not provide glucose repression of invertase, suggesting that the phosphorylation of HK PII is essential for glucose signal transduction,⁹ however, it seems to be not the only factor involved.¹⁰ Nuclear localization of HK PII was shown to be mediated by an internal N-terminal decapeptide (amino acid residues 7–16) that is necessary to direct the protein to the nucleus, where it seems to be involved in the formation of specific DNA-protein complexes during glucose-induced repression of the SUC2 gene.¹¹ HK PI can only partially substitute HK PII in triggering glucose repression, but it is equally important in triggering fructose repression.^{5,12,13}

Hexokinase catalyses the formation of hexose-6-phosphate (mainly glucose-6-phosphate, Glc-6P) from hexose using ATP as a phosphoryl donor. The isoenzymes differ in their activity toward particular hexoses. The enzymatic activity of PI isoenzyme with fructose and mannose is 2.6 and 0.6 times than that with glucose, respectively, whereas with PII isoenzyme, the fructose:glucose ratio is only 1.3 and mannose:glucose ratio is 0.3.¹⁴ Yeast HKes exist in solution as dimers and monomers. The dissociation of dimers into monomers is promoted by glucose binding and an increase in pH and/or ionic strength.^{15–17}

HKs are found in practically all living beings. Mammals, for example, have four HK isoenzymes¹⁸ three of which have molecular weight of about 100 kDa and display a considerable sequence identity between their N and C-terminal

Grant sponsor: Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP); Grant number: 06/00182-8; Grant sponsor: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); Grant number: 473875/2003-9; Grant sponsor: Coordenação de Aperfeiçoamento de Pessoal do Nível Superior (CAPES).

*Correspondence to: Igor Polikarpov, Grupo de Cristalografia, Instituto de Física em São Carlos, USP, Av. Trabalhador Saocarlene, 400 CEP 13560-970, São Carlos, SP, Brazil. E-mail: ipolikarpov@ifsc.usp.br

Received 21 June 2007; Revised 8 November 2007; Accepted 29 November 2007

Published online 7 February 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21956

halves.¹⁶ The fourth mammalian HK isoenzyme has a molecular mass of about 50 kDa, similar to yeast and *Schistosoma mansoni* HKs. The sequence similarity between N- and C-terminal part of the mammalian isoenzymes I to II is believed to be a consequence of duplication and fusion of a primordial hexokinase gene.^{16,19,20} Yeast HKs, unlike human HKs and HK from *Schistosoma mansoni*, are insensitive to inhibition by the product of reaction Glc-6P.

Yeast HK isoenzyme is known to be a classical example of the theory of “induced fit” mechanism of enzymatic reaction originally proposed by Koshland.^{21,22} The crystallographic structures of yeast HK isoenzymes PI complexed with glucose^{23,24} and PII complexed with a competitive inhibitor, ortho-toluoylglucosamine^{25,26} have been determined at the end of the 70’s and used to illustrate the role of the “induced fit” in enzyme action and to portray structural changes of the enzyme associated to glucose binding.²² However, at the time of structure determination, the primary sequences of the yeast HK isoenzymes were not available, and the identity of the side-chains was guessed from inspection of crystallographically computed electron density. The models of PI and PII isoenzyme contain only 458 residues each and about 17% of these residues have uncertain identity. A comparison of the HK amino acid sequences found from X-ray studies with the actual ones^{1,2} revealed only 30% identity between them. This fact complicated detailed analysis of the conformational changes of the enzyme associated with the glucose binding and introduced uncertainty in quantitative parameters of the “induced fit” mechanism of reaction obtained from these structures.

One of the ways to assess and identify amino acid residues involved in the allosteric networks is offered by statistical coupling analysis (SCA). This technique, whose development started in 1999 by Rama Ranganathan’s group,²⁷ uses the amino acid distribution in positions of protein sequences alignments and the interdependence of these distributions between positions to identify amino acids with important roles in structure and function in a protein family. A later improvement of the technique consisted in iterative clusterization of $\Delta\Delta G^{\text{stat}}$ matrices.²⁸ The results obtained using clusters analysis show that the relatively restricted groups of residues found with the method connect the functionally important regions of protein families, suggesting that they could represent the core of protein allostery.²⁸ The technique has a great potential to identify “hot spots” in protein families, with the advantage of using no structural information – only a multiple sequence alignment of the protein/domain family is needed.^{27,28}

Here, we report the 2.95 Å resolution crystal structure of yeast holo-PI hexokinase (yhxPI) complexed with glucose in an attempt to present a better defined three-dimensional (3D) model of this yeast hexokinase isoen-

zyme, and to add more detailed structural information to the “induced fit” in yeast HK enzymatic mechanism. The model is based on the correct amino-acid sequence and the glucose binding site is described in detail. Comparing the present structure with the structure of apo-PII isoenzyme,²⁹ we were able to quantify conformational changes induced in HK by ligand binding more accurately. The mechanism of reaction, the lack of Glc-6P inhibition in yeast HKs, and the putative mode of ATP binding are also discussed. Furthermore, based on SCA we identified two clusters, which form a continuous 3D network of statistically coupled residues. The cluster with the strongest couplings is composed by amino acid residues of the active site and the hinge regions, further confirming cooperative function of these parts of the protein in the induced fit mechanism of the enzyme function.

MATERIALS AND METHODS

Crystallization

Hexokinase PI isoenzyme was purchased from Sigma-Aldrich and used for crystallization trials without any further purification. Crystals were grown in hanging drops at 277 K.³⁰ The protein solution was buffered with 100 mM potassium phosphate in the pH range 5.5–6.5 and 25–30% PEG as a precipitant. Hanging drops contained 3 µL of the protein at concentrations of 10 mg mL⁻¹ and 3 µL of precipitant. Needle shaped crystals appeared between 1 and 4 weeks and with typical crystal dimensions of 0.5 × 0.05 × 0.05 mm³. Data were collected from these crystals, which diffracted to 2.95 Å resolution.

X-ray diffraction data

Diffraction data were collected using a MAR image plate system at the dedicated protein crystallography beamline of the Brazilian synchrotron light source.^{31,32} The data were integrated, scaled, and merged using the programs DENZO and SCALEPACK.³³ A total of 14,357 reflections were collected, with a completeness of 92.27% for all data to 2.95 Å (94.44% for the highest resolution shell). The crystal belonged to the orthorhombic space group $P2_12_12_1$, with cell constants of $a = 62.12$ Å, $b = 78.87$ Å, and $c = 144.74$ Å. Estimation of the solvent content indicated that the crystal contained one molecule per asymmetric unit. The Matthews’ coefficient³⁴ assuming that the asymmetric unit contained a monomer was 3.1 Å³ Da⁻¹, which corresponds to a solvent content of 60%. The crystal structure of the yeast hexokinase PI was determined by molecular replacement using a P152K mutant of yeast hexokinase as the starting model. The program AMoRe³⁵ was used in calculation of rotation and translation functions based on data collected up to 3 Å. A unique solution was found after rigid-body refine-

ment, with a correlation coefficient of 47.9% and an R -factor of 0.53.

Refinement, model building, and overall quality of the model

The structure was then refined using the programs REFMAC³⁶ and Phenix³⁷ with working and free (5%) sets of data extended to include all data from 12.9 to 2.95 Å. Thirteen thousand six hundred forty-two reflections were used for the refinement, while 715 reflections, randomly selected and set aside during refinement process, were used for R -free calculation. Rounds of positional and group B -factors refinement were alternated with model verification using the O program.³⁸ Given the medium resolution of the X-ray data and the limited number of collected reflections, only one B -factor per residue has been refined to improve the observations-to-model parameters ratio. The refinement, which was forced to maintain tight geometry of the structure, was monitored by the decline in R -free. Refinement and rebuilding of the model lowered the R -factor to a final value of 19.2% and R -free to 24.5%. Details of the final stage of the refinement are given in Table I.

The final model contains all but the first 14 residues of N-terminal for which no discernible density could be observed. The average thermal parameter for the model is 44.8 Å² and the final ($2F_o - F_c$) electron density is well defined for most residues. The program LSQKAB of the CCP4 package was used for the analysis of root-mean-

squares deviations in the positions of C α atoms of different models. Ramachandran analysis³⁹ as implemented in PROCHECK⁴⁰ shows that 86.7% of the residues lie in the most favored regions, while 13.3% lie in the additionally and generously allowed regions. No residues are found in disallowed regions. The G -factor of the model is 0.2. The final model and structure factors were deposited in the Protein Data Bank under the PDB ID 3B8A.

Statistical coupling analysis

A multiple sequence alignment of hexokinases was obtained from the PFAM database,⁴¹ in a total of 317 sequences, including that of yeast hexokinase PI. From the alignment, it is possible to compute, in a parameter named ΔG^{stat} , the discrepancy between amino acid frequencies in a given position and the amino acid frequencies observed in nature, which shows how strongly this position is conserved in the protein family.²⁷ A second parameter, termed $\Delta\Delta G^{\text{stat}}$, quantifies how much the amino acid distribution of a certain position varies when a specific condition is imposed at another position (i.e., the later position should be occupied by a certain type of amino acid residue). If the amino acid distribution at the former position in the alignment varies (which is reflected in a high value for $\Delta\Delta G^{\text{stat}}$), this means that for some reason, normally related to the structure and/or function of the protein, these two positions coevolved together, as suggested by the observed statistical coupling.²⁷ Site conservation (ΔG^{stat}) and statistical coupling ($\Delta\Delta G^{\text{stat}}$) parameters between positions in the alignment were calculated as described in Lockless and Ranganathan, 1999.²⁷ ΔG^{stat} values for the 20 most conserved residues are shown in Table II. To determine the set of positions to be perturbed in order to calculate $\Delta\Delta G^{\text{stat}}$, we used the method used by Ranganathan's group, described in Süel *et al.*, 2003.²⁸ Monitoring the variance of mean $\Delta\Delta G^{\text{stat}}$ for the five least conserved positions of the alignment, which should, by definition, have no evolutionary restraint and therefore low coupling values, it is possible to define a limit for the size of the subalignment used to calculate $\Delta\Delta G^{\text{stat}}$. For our case, the subalignment should have at least 35% of the original sequences, resulting in 209 positions which could be perturbed. The initial matrix of $\Delta\Delta G^{\text{stat}}$ contained 209 columns (perturbed positions) and 485 lines (sequence size of HK PI). This matrix is sparse, suggesting that during evolution, few positions in a protein evolved together. The iterative analysis consists of clustering rows and columns with similar $\Delta\Delta G^{\text{stat}}$ distribution and eliminating low noise rows and columns. The matrix was subjected to successive rounds of clustering using the software package MATLAB and elimination of low-signal rows and columns. During the process, high signal positions grouped together in two clusters. The first cluster consists of a 2×29 matrix (30 amino acid residues positions), with mean $\Delta\Delta G^{\text{stat}} = 219$. The second

Table I
Crystallographic Information

Resolution range (Å)	12.92–2.95 (3.03–2.95)
Space group	$P2_12_12_1$
Unit cell parameters (Å)	$a = 62.12$, $b = 78.87$, $c = 144.74$
Completeness (%)	92.27 (94.44)
Reflections used in refinement	13,642 (1007)
Total reflections/Reflections for R_{free}	14,357/715
I/σ	8.8 (2.49)
R_{merge} (%)	11.9 (47.7)
Multiplicity	2.8 (2.7)
R_{factor} (%)	19.19
R_{free} (%)	24.45
Bond length RMS deviation (Å)	0.002
Bond angle RMS deviation (°)	0.470
Mean overall B -factor	44.8
Overall B -factor from Wilson plot	47.4
Mean B -factor for protein	44.9
Mean B -factor for glucose	29.1
Mean B -factor for SO ₄	48.71
Mean B -factor for water molecules	32.68

Values between parenthesis correspond to the value in the highest resolution bin

(a total of 20 bins was used). $R_{\text{factor}} = \frac{\sum_i |F_o - F_c|}{\sum_i |F_o|}$, R_{free} is defined the same as

R_{factor} except that the summation is over the 715 (about 5%) reflections not included in the refinement.

Table II
Statistical ΔG Parameter for 20 Most Conserved Residues

Ranking	Amino acid residue	ΔG^{stat}	Normalized ΔG^{stat}
1	Trp174	1510.28	100
2	Asn210	1141.41	75.58
3	Cys268	1128.30	74.71
4	Pro160	1114.98	73.83
5	Pro74	1091.75	72.29
6	Arg93	1043.57	69.10
7	Phe157	969.50	64.19
8	Ser158	965.34	63.92
9	Lys176	946.17	62.65
10	Phe240	932.62	61.75
11	Pro115	915.23	60.60
12	Asn237	894.36	59.22
13	Asp211	890.73	58.98
14	Gly307	888.12	58.80
15	Gly80	870.01	57.61
16	Gln163	865.02	57.28
17	Asn91	853.06	56.48
18	Gly88	848.74	56.20
19	Asp86	841.42	55.71
20	Trp128	841.12	55.69

$\Delta G_i^{\text{stat}} = \sqrt{\sum_x \left(\ln \frac{P_i^x}{P_{\text{MSA}}^x} \right)^2}$. P_i^x is the binomial probability of finding observed frequency of the amino acid x (x = alanine, cysteine, etc.) in the position i of the alignment, given the known frequency observed in nature, while P_{MSA}^x is the binomial probability of finding the frequency of the amino acid in all the multiple sequence alignment (MSA) in that position. As ΔG^{stat} is dependent of the number of sequences, its value has arbitrary units. The highest value for ΔG^{stat} , 1510.28, corresponds to a position which presents 77.3% of tryptophan residues. This is highly improbable to happen by chance, given that the expected frequency for tryptophan residue in proteins in nature is only 1%.

cluster consists in a 7×18 matrix (20 amino acid residues positions), with mean $\Delta \Delta G^{\text{stat}} = 182$. $\Delta \Delta G^{\text{stat}}$ values for all residues in the clusters are shown in Tables III and IV. Residue conservation is shown in Figure 1. This figure was rendered in Pymol⁴² after substituting the B -factor column of the PDB file with ΔG^{stat} values after normalization and setting the highest B -factor to one hundred. The same normalized values were used for all ΔG^{stat} reports throughout the text.

RESULTS AND DISCUSSION

The overall structure

Yeast hexokinase PI is a palm shape molecule and exhibits the same α/β fold observed in other hexokinase structures (Fig. 1).^{23–26,29,43–45} The polypeptide chain of 485 residues is distinctly folded into two domains of unequal size: the large domain (residues 17–76 and 212–458) and the small domain (residues 77–211 and 459–485). These are separated by a deep cleft containing the residues making the enzyme active site. The protein is found in closed conformation.²⁴ The most conserved residues, as revealed by SCA, are located in the protein core; near the glucose binding site (Fig. 1).

The crystallization of yeast HK isoenzyme PI was performed in the absence of a substrate. However, at the early stage of refinement a clear and continuous residual electron density appeared in both $(2F_{\text{obs}} - F_{\text{calc}})$ and $(F_{\text{obs}} - F_{\text{calc}})$ Fourier syntheses. This electron density could be readily interpreted in terms of a glucose molecule and a sulfate ion. Sulfate ions were present in 2M concentration in the crystallization conditions. The thermal parameters for ligands positioned in that density are comparable with the average thermal parameter for solvent molecules, inferring their substantial occupancy that was set to 1 during structure refinement.

Superposition of the previously determined yeast apo-HK PII model onto holo-HK PI+glucose complex gives an average rms deviation of 2.8 Å for C α atom pairs with maximum deviation of over 8 Å for several parts of the polypeptide backbone (Fig. 2). Separately comparing the large and the small domain of the structures give an average rms deviations of 1.17 and 1.12 Å, respectively. Glucose-

Table III
Statistically Coupled Clusters of Residues in HK: 2×29 Cluster
(Mean $\Delta \Delta G^{\text{stat}} = 219$)

Perturbation	210	174
G177	183.83	246
K176	208.42	258.82
D211	233.57	209.11
G185	173.03	286.07
D187	188.47	289.25
G169	173.56	242.36
L192	180.86	281.18
L171	184.38	308.2
F178	177.39	254.23
V189	151.94	285.2
V205	174.79	200.84
G214	196.79	320.99
W174	190.13	31.799
T175	184.08	385.97
V188	173.67	357.13
A207	208.62	233.29
T212	209.26	301.42
T215	212.64	254.05
I231	210.7	216.37
N237	222.91	201.56
T233	211.32	195.73
G235	216.96	188.49
N267	184.81	224.26
E269	183.42	202.55
S306	158.21	193.24
Y240	191.58	308.3
E242	186.93	295.07
N295	158.05	263.51
D353	144.84	216.15

Here $\Delta \Delta G_{ij}^{\text{stat}} = \sqrt{\sum_x \left(\ln \frac{P_{\text{MSA}/ij}^x}{P_{\text{MSA}}^x} - \ln \frac{P_i^x}{P_{\text{MSA}}^x} \right)^2}$, P_i^x and P_{MSA}^x as defined before. “ $i|j$ ”

stands for “upon perturbation of residue j ”, i.e., the probabilities are calculated for the whole alignment and for a subalignment where a condition is imposed (e.g., j is an alanine). $\Delta \Delta G^{\text{stat}}$ is also dependent of alignment size and has arbitrary units. The highest $\Delta \Delta G^{\text{stat}}$ found (385.97) was due to the perturbation T175: the presence of a threonine in this position led to the increase of the tryptophan frequency in position 174 from 77.3% to 99.2%.

Table IVStatistically Coupled Clusters of Residues in HK: 7×18 Cluster (Mean $\Delta\Delta G^{\text{stat}} = 182$)

Perturbation	219	226	239	424	421	385	210
L465	100.97	118.83	79.204	154.05	129.92	171.57	106.61
N295	124.1	134.44	212.13	124.03	167.66	217.75	158.05
D353	136.62	130.63	192.62	156.36	194.83	283.3	144.84
G274	238.74	248.74	223.4	278.99	220.37	249.28	90.786
G278	247.85	252.62	237.38	285.86	224.07	252.17	85.223
L368	198.93	252.52	201.67	241.47	191.14	208.59	85.223
V314	195.51	200.59	200.98	233.58	199.33	249.85	97.938
D377	175.37	183.82	226.66	230.85	196.6	260.85	114.86
F342	208.89	213.25	238.89	202.8	196.91	233.71	116.45
L327	146.75	154.7	194.07	173.36	171.5	232.54	96.96
V388	129.51	153.95	190.52	166.66	197.68	248.2	120.39
S455	193.34	206.49	192.6	212.87	216.34	213.56	97.51
C226	291.42	47.623	268.32	223.99	197.66	193.39	104.18
C385	159.71	167.62	200.91	200.7	232.22	46.027	120.08
Y421	180.68	189.4	215.52	218.43	45.66	261.25	109.43
C239	219.48	228.78	43.006	181.8	193.85	201.36	134.1
C219	46.932	302.82	265.92	220.21	193.91	189.78	113.82
T413	150.56	163.33	241.96	141.19	155.37	154.99	86.437

Here $\Delta\Delta G_{ij}^{\text{stat}} = \sqrt{\sum_x \left(\ln \frac{P_{ij}^x}{P_{MSA(ij)}^x} - \ln \frac{P_i^x}{P_{MSA}^x} \right)^2}$, P_i^x and P_{MSA}^x as defined before. “ $i|j$ ” stands for “upon perturbation of residue j ”, i.e., the probabilities are calculated for the whole alignment and for a subalignment where a condition is imposed (e.g., j is an alanine). $\Delta\Delta G^{\text{stat}}$ is also dependent of alignment size and has arbitrary units. The highest $\Delta\Delta G^{\text{stat}}$ found (385.97) was due to the perturbation T175: the presence of a threonine in this position led to the increase of the tryptophan frequency in position 174 from 77.3% to 99.2%.

induced conformational fit can be described as a rigid-body rotation and nonrigid body conformational changes.

Glucose-induced conformational changes

The main part of the HK fit induced by glucose is a rigid body rotation of one part of the molecule with respect to the other. Following the analysis by Hayward and Berendsen,⁴⁶ apo-HK PII and holo-HK PI+glucose complex

structures comparison permits identification of the dynamic domains of the protein. The whole structure can be divided into the fixed domain, comprising residues 20–59, 61–73, 104–106, 211, 214–259, 262–308, 310–458, 467–482 and a rotating domain made of residues 74–103, 107–210, 212–213, 459–466. The definition of the fixed and rotating domains approximately corresponds to the traditional definition of the large and small domain, respectively. Residues 73–74, 104–107, 211–212, 213–215, 458–

Figure 1

Overall fold of holo PI hexokinase. The protein is colored by residue conservation (ΔG): poorly conserved positions (low ΔG) are shown in “cold” colors, highly conserved positions (high ΔG) are shown in “hot” colors. The gray surface corresponds to the regions of the mechanical hinges, connecting the two domains.

Figure 2

Superposition of the previously determined yeast apo-HK PII model onto holo-HK PI+glucose complex. The superposition of the whole structure leads to high $C\alpha$ deviations in a considerable portion of the protein. These deviations fall dramatically when the two domains are superposed separately. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

459, and 466–467, noteworthy, highly conserved Asp211, Thr213, Gly215, and Gly458 (normalized ΔG^{stat} s are, respectively, 58.9, 30.5, 51.4, and 32.9), act as “mechanical hinges.” The moving domain rotates with respect to the axis that is approximately perpendicular to the straight line between the centers of masses of the two domains in the direction of the fixed domain by an angle of 17° . This is a surprisingly large rotation angle, taking into consideration that previous estimation, done on the basis of super-

position of earlier HK models, was 12° .^{22–26} Conformational changes result in movements of the polypeptide backbone of as much as 11.5 Å, which is also considerably larger than the maximum displacement of 8 Å reported before. To exclude possible computational differences related to the slightly different definition of the moving and fixed domains, we repeated the calculation of the rigid body rotation based on the earlier HK models.^{22–26} The result of such calculation, a rotation angle of 13.6° , is close to the value reported by Steitz et al. and Bennett and Steitz.^{23,24} The rest of the difference in the angle of HK domain closure, induced by binding of glucose, can be attributed to the fact that the early HK PI structure was determined as a complex with competitive inhibitor, ortho-toluoylglucosamine,^{25,26} which induces a slight closure of the domains,²⁹ leading to underestimation of the glucose-induced conformational changes of HK.

In addition to the rigid-body domain rotation, several regions of the polypeptide chain are involved in nonrigid conformational changes. These changes, however, are mostly confined to the surface loop regions, such as loops 142–149 and 470–478 of the small domain and loops 245–262, 338–342, and 445–448 of the large domain and might be induced by crystallographic contacts or are result of the single amino acid insertion (loop 445–448) (Fig. 2).

Surprisingly, virtually all the conformational changes in the active site region, including loop regions, could be described as a rigid-body rotation of the small domain of the protein with respect to its large domain.

The ligand binding site

The glucose snugly fits into the cleft between two HK domains (Fig. 3) and interacts with the HK binding site

Figure 3

A: Interactions between PI hexokinase, glucose and a sulfate anion in the binding site. B: The glucose molecule and the sulfate anion superimposed with the OMIT $2F_o - F_c$ maps, contoured at 1σ . [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

via well-defined interactions at the positions of 1-, 3-, 4-, 5-, and 6-OH groups. This is in agreement with substrate specificity studies of Sols and Crane⁴⁷, which implicate the participation of all but the 2-OH group of glucose in interactions with the mammalian HK. The β -enantiomer of glucose is more consistent with the electron density and ligand interactions. Although glucose α -enantiomer does not produce a steric hindrance, the distance between Glu 302 O ϵ 2 and the 1-OH group becomes too long for strong hydrogen bond interaction.

The highly conserved Asp211 (normalized $\Delta G = 58.9$, 13th highest for the protein), implicated as a catalytic base,^{24,25} makes a hydrogen bond with the hydroxyls at the positions 4 and 6 of glucose, whereas side chains of the amino acid residues Glu302, Glu269, and Lys176 are at the hydrogen bonding distances from the 1-OH, 3-OH, and 5-OH groups of glucose molecule, respectively. Lys176 makes an additional hydrogen bond with the 6-hydroxyl group of glucose as well. The absence of interactions with 2-OH group at the glucose-binding pocket is consistent with the fact that 2-substituted derivatives of glucose serve as HK substrate. Site-directed mutations of Asp211 cause loss of catalytical activity, but high sugar binding affinity is retained.⁴⁸ This indicates that sugar binding is mainly coordinated by Glu302 and Glu269, whereas Asp211 is indispensable for promoting a nucleophilic attack of the glucose 6-OH group on the γ -phosphate of ATP.²⁵ Residues Glu302 and Glu269 make part of the large (or fixed) domain, Asp211 belongs to a hinge region, whereas Lys176 comes from a small (rotating) domain. Obviously, the glucose binding pocket becomes complete only upon the closure of HK domains. Therefore, one can speculate that initial glucose binding to HK, mediated by interactions with the side chains of amino acid residues Asp211, Glu302, and Glu269, triggers large-scale conformational rotation of the HK domains with respect one another, bringing Lys176 side chain into hydrogen bonding distance from the glucose hydroxyl groups 4 and 6 and thus creating the competent glucose binding site.

The glucose molecule binds in a manner structurally analogous to that observed for other hexokinase complexes, maintaining the glucose binding site on our model in close agreement with the binding site as previously elucidated.^{16–19} However, several important differences exist. Strongly conserved Ser158 (normalized $\Delta G^{\text{stat}} = 63.9$, 8th mostly conserved residue in the protein), previously described as a conserved glucose-binding residue,²⁴ is at the distance of 3.86 Å from the 6-hydroxyl group of glucose, which is somewhat too long for a strong hydrogen binding. At the same time, directed mutation studies identify Ser158 as a critical determinant of the catalytical activity and conformation of HK.^{49,50} Furthermore, it was shown that binding of the certain nonphosphorylatable inhibitors of glucose such as D-xylose or D-lyxose promotes strikingly the short-lived hydrolytic activity of

the enzyme followed by inactivation of the enzyme via autophosphorylation of Ser158.^{17,51–53} All these evidences taken together support the idea that additional conformational changes might occur upon binding of ATP^{54–56} and that Ser158 might play an important role in the process of the phosphate transfer.

The sulfate group, observed in the active site of the holo-HK PI structure, interacts with conserved residues Ser419 and Thr234 (normalized ΔG^{stab} s are 37.6 and 47.8) of the fixed HK domain (Fig. 3). Its position is identical to the one of sulfate anion in the HK PII structure.³⁰ The presence of a sulfate ion in the structure of HK both in open and in close conformations proves that, first, glucose binding is not required for sulfate/phosphate binding and, second, sulfate/phosphate binding alone is not sufficient to provoke significant conformational changes of the enzyme. Furthermore, the fact that the same sulfate/phosphate anion binding site was repeatedly found in yeast^{23–26,29} and human^{43–45} hexokinase models shows that HK ATP binding site is able to bind monophosphates. This might have functional importance. It is known, for example, that glucose binding was strongly promoted in the presence of 0.05M phosphate.⁵² Phosphate binding might therefore somewhat restrict the conformations of the amino acid residues of the HK binding site and thus facilitating glucose binding.

Insights about the lack of HK PI inhibition by Glc-6P are obtained after comparison to the Glc-6P bound structure of human hexokinase I. For the human HK, Glc-6P contacts the loops 86–93 and also 454–460, which are on different conformations when compared with its structural analogs in yeast HK PI. The phosphate group of the Glc-6P interacts with the Thr536 OG1 and a main chain nitrogen atom of the same residue (loop 532–539) whereas in yeast HK interactions of the Glc-6P with the correspondent loop (residues 86–93). Correspondent Thr90 residue is more than 8 Å away. It has also been found that, for yeast HK PI, residue Asp211 is at a position which could form an hydrogen bond with the 4-OH group of Glc-6P, but the interaction between 2-OH of Glc-6P with Ser897 in human hexokinase I does not have an equivalent in yeast HK PI.

The putative ATP binding site

The ATP pattern recognition, identified on the basis of comparison of actin,⁵⁷ ATPase fragment of 70 K heat-shock cognate protein⁵⁸ and HK three-dimensional structures, was defined as consisting of five sequence motifs: Phosphate1 (residues 82:103), Connect1 (residues 203:223), Phosphate2 (residues 229:248), Adenosine (residues 411:439), and Connect2 (residues 452:472).⁵⁹ Residues Thr234 and Ser419 belong, respectively, to the Phosphate2 and Adenosine ATPase pattern recognition motifs. The α -phosphoryl group of ADP occupies the

Figure 4

Interactions between PI hexokinase ADP, as suggested by modeling an ADP molecule by analogy to the human hexokinase I-ADP/Glu complex. The putative interactions shown in the figure as dashed lines are between 2.5 and 3.5 Å distance limits. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

same site as the 6-phosphoryl group of glucose-6-phosphate in human HK PI structure.^{56,60}

An ADP molecule was modeled in yeast hexokinase PI by analogy to the position of an ADP molecule in the structure of the ADP/glucose complex of human hexokinase I (Fig. 4). The base moiety of ADP is sandwiched between helices $\alpha 9$ and $\alpha 11'$. The model suggests that the binding site of ADP is equivalent in both structures.

In yeast hexokinase PI both helices are suitably positioned to accommodate the ADP molecule, and there are no steric clashes between the protein and the ADP molecule. However, to maintain the same interactions with the side chain residues observed in the human hexokinase I, helix $\alpha 9$ would have to move closer to the ADP molecule. This movement could be triggered by the binding of ATP to the molecule.

The modeling also shows that in yeast hexokinase PI the β -phosphoryl group points out in the direction of the 6-hydroxyl group of glucose in a position and orientation adequate for productive binding without being blocked by loop L1 as it happens in the human hexokinase I. A γ -phosphoryl group could equally be modeled into a productive conformation.

Cluster analysis of SCA data

Our SCA data for position coupling in the hexokinase family were treated with elimination of low signal rows and columns alternated with clusterization in MATLAB. The positions with high $\Delta\Delta G^{\text{stat}}$ were grouped together in two different clusters. Highly coupled clusters in SCA studies comprise the positions that evolved together in a protein family, a fact that might have happened for structural and functional reasons. The highest $\Delta\Delta G^{\text{stat}}$ cluster we have found comprises the regions of the protein that we described as crucial to our induced fit model, that is, the binding site and the hinges which allow the rearrangement of the two domains (Fig. 1). The fact that the active site residues show strong statistical coupling with the hinge regions is very consistent with the “induced fit” mechanism of enzymatic catalysis of the HK which

Figure 5

A: A structural view of the clusters of highly evolutionarily coupled positions in hexokinases. The cluster of stronger coupled residues (mean $\Delta\Delta G^{\text{stat}} = 219$) is shown in blue, while the cluster with the weaker coupling values (mean $\Delta\Delta G^{\text{stat}} = 182$) is shown in red. The residues shown in mixed blue/red (R295, D353, N210) are present in both clusters; B: The same as (A) upon 180° rotation around Y axis.

implies a necessity of simultaneous conservation of these important parts of the protein. Another cluster, with lower coupling values between positions, was also found, and consists of disperse protein positions [Figs. 5(A,B)]. The structural and functional role of this second cluster remains elusive.

CONCLUSIONS

Yeast hexokinase is a classic example of enzyme with the “induced fit” mechanism of reaction. The present crystallographic structure of the yeast HK PI bound to glucose reveals the details of substrate recognition and shows that conformational changes of this enzyme upon ligand binding (“induced fit”) are somewhat larger than considered earlier (17° vs. 12°). The allosteric network involved in the induced fit mechanism, as suggested by our SCA analysis, is confined to the vicinity of the glucose binding site and the hinge region. A secondary network of residues which co-evolved with lower coupling values was also found, but its meaning remains to be determined. The fact that the residues in the binding site and in the mechanical hinges co-evolved together, as suggested by SCA, sheds light into the development of its mechanism of action – the ability to bind glucose and rotate the two domains were probably developed simultaneously during evolution of the protein family, suggesting that the induced fit mechanism was a necessity to protein function.

In addition, the X-ray structure of holo-PI HK allows for rationalization of the wealth of biochemical information accumulated to date. Moreover, the presence of a bound glucose molecule and, consequently, more compact conformation of the HK PI obtained from commercial preparations, requires a significant revision in the interpretation of past thermal denaturation experiments.

ACKNOWLEDGMENT

Authors acknowledge Rama Ranganathan's group for support with the statistical coupling analysis methods.

REFERENCES

- Kopetzki E, Entian K, Mecke D. Complete nucleotide sequence of the hexokinase PI gene (HXX1) of *Saccharomyces cerevisiae*. *Gene* 1985;39:95–101.
- Fröhlich K, Entian K, Mecke D. The primary structure of the yeast hexokinase PII gene (HXX2) which is responsible for glucose repression. *Gene* 1985;36:105–111.
- Gancedo JM, Clifton D, Fraenkel DG. Yeast hexokinase mutants. *J Biol Chem* 1977;252:4443–4444.
- Entian KD. Genetic biochemical evidence for hexokinase PII as a key enzyme involved in carbon catabolite repression in yeast. *Mol Gen Genet* 1980;178:633–637.
- De Winde JH, Crauwels M, Hohmann S, Thevelein JM, Winderickx J. Differential requirement of the yeast sugar kinases for sugar sensing in establishing the catabolite-repressed state. *Eur J Biochem* 1996;241:633–643.
- Fernandez R, Herrero P, Fernandez MT, Moreno F. Mechanism of inactivation of hexokinase PII of *Saccharomyces cerevisiae* by D-xylose. *J Gen Microbiol* 1986;132:3467–3472.
- Vojtek AB, Fraenkel DG. Phosphorylation of yeast hexokinases. *Eur J Biochem* 1990;190:371–375.
- Behlke J, Heidrich K, Naumann M, Muller EC, Otto A, Reuter R, Kriegl T. Hexokinase 2 from *Saccharomyces cerevisiae*: regulation of oligomeric structure by *in vivo* phosphorylation at serine-14. *Biochemistry* 1998;37:11989–11999.
- Randez-Gil F, Sanz P, Entian KD, Prieto JA. Carbon source-dependent phosphorylation of hexokinase pII its role in the glucose-signaling response in yeast. *Mol Cell Biol* 1998;18:2940–2948.
- Mayordomo I, Sanz P. Hexokinase PII: structural analysis glucose signaling in the yeast *Saccharomyces cerevisiae*. *Yeast* 2001;18:923–930.
- Herrero P, Martinez-Campa C, Moreno F. The hexokinase II protein participates in regulatory DNA-protein complexes necessary for glucose repression of the SUC2 gene in *Saccharomyces cerevisiae*. *FEBS Lett* 1998;434:71–76.
- Gancedo JM. Yeast carbon catabolite repression. *Microbiol Mol Biol Rev* 1998;62:334–361.
- Rose M, Albright W, Entian KD. Glucose repression in *Saccharomyces cerevisiae* is directly associated with hexose phosphorylation by hexokinases PI PII. *Eur J Biochem* 1991;199:511–518.
- Lazarus NR, Ramel AH, Rustum YM, Barnard EA. Yeast hexokinase. I. Preparation of the pure enzyme. *Biochemistry* 1966;5:4003.
- Feldman I, Kramp DC. Fluorescence-quenching study of glucose binding by yeast hexokinase isoenzymes. *Biochemistry* 1978;17:1541–1547.
- Wilson JE. Hexokinases. *Rev Physiol Biochem Pharmacol* 1995;126:65–198.
- Colowick SP. The hexokinases. In: Boyer PD, editor. *The Enzymes*, Vol. 9. New York: Academic; 1973. pp 1–48.
- Katzen HM, Schimke RT. Multiple forms of hexokinase in the rat: tissue distribution, age dependency, properties. *Proc Natl Acad Sci* 1965;54:1218–1225.
- Easterby JS, O'Brien MJ. Purification properties of pig-heart hexokinase. *Eur J Biochem* 1973;38:201–211.
- Holroyde MJ, Trayer IP. Purification properties of skeletal muscle hexokinase. *FEBS Lett* 1976;62:215–219.
- Koshland DE. Mechanisms of transfer enzymes. In: Boyer PD, Lardy H, Myrbäck K, editors. *The Enzymes*, 2nd ed., Vol. 1. New York: Academic Press; 1959. pp 305–346.
- Stryer L. *Biochemistry*. New York: WH Freeman Company; 1995. p 499.
- Steitz TA, Fletterick RJ, Anderson WF, Anderson CM. High resolution X-ray structure of yeast hexokinase, an allosteric protein exhibiting a non-symmetric arrangement of subunits. *J Mol Biol* 1976;104:197–222.
- Bennet WS, Jr, Steitz TA. Structure of a complex between yeast hexokinase A glucose. I. Structure determination refinement at 3.5 Å resolution. *J Mol Biol* 1980;140:183–209, 211–230.
- Anderson CM, Stenkamp RE, McDonald RC, Steitz TA. Sequencing a protein by X-ray crystallography. II. Refinement of yeast hexokinase B co-ordinates sequence at 2.1 Å resolution. *J Mol Biol* 1978;123:15–34.
- Anderson CM, Zucker FH, Steitz TA. Space-filling models of kinase clefts conformation changes. *Science* 1979;204:375–380.
- Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 1999;286:295–299.
- Süel GM, Lockless SW, Wall MA, Ranganathan R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol* 2003;10:59–69.
- Kuser PR, Krauchenco S, Antunes OAC, Polikarpov I. The high resolution crystal structure of yeast hexokinase PII with the correct primary sequence provides new insights into its mechanism of action. *J Biol Chem* 2000;275:20824–20831.

30. Kuser PR, Golubev AM, Polikarpov I. Crystallization preliminary crystal analysis of yeast hexokinase PI and PII. *Acta Crystallogr D Biol Crystallogr* 1999;55:2047–2048.
31. Polikarpov I, Perles LA, de Oliveira RT, Oliva G, Castellano EE, Garrat RC, Craievich A. Set-up experimental parameters of the protein crystallography beamline at the Brazilian National Synchrotron Laboratory. *J Synchrotron Rad* 1998;5:72–76.
32. Polikarpov I, Oliva G, Castellano EE, Garrat RC, Arruda P, Leite A, Craievich A. The protein crystallography beamline at LNLS, the Brazilian National Synchrotron Light Source. *Nucl. Instrum Methods A* 1998;405:159–164.
33. Otwinowski Z, Minor W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol* 1997;276:307–326.
34. Matthews BW. Solvent content of protein crystals. *J Mol Biol* 1968;33:491–497.
35. Navaza J. Implementation of molecular replacement in AMoRe. *Acta Crystallogr D Biol Crystallogr* 2001;57:1367–1372.
36. Murshudov GN, Vagin AA, Dodson EJ. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* 1997;53:240–255.
37. Adams PD, Grosse-Kunstleve RW, Hung L-W, Ioerger TR, McCoy AJ, Moriarty NW, Read RJ, Sacchettini JC, Sauter NK, Terwilliger TC. PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr D Biol Crystallogr* 2002;58:1948–1954.
38. Jones TA, Zou JY, Cowan SW, Kjeldgaard M. Improved methods for building protein models in electron density maps the location of errors in these models. *Acta Crystallogr A* 1991;47:110–119.
39. Ramachandran NG, Sasisekharan V. Conformation of polypeptides proteins. *Adv Protein Chem* 1968;23:283–437.
40. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr* 1993;26:283–291.
41. Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer ELL, Bateman A. Pfam: clans, web tools services. *Nucleic Acids Res* 2006;34:D247–D251.
42. DeLano WL. The PyMol Molecular Graphics System. San Carlos, CA, USA: DeLano Scientific; 2002. Also available at <http://www.pymol.org>.
43. Mulichack AM, Wilson JE, Padmanabhan K, Garavito RM. The structure of mammalian hexokinase-1. *Nat Struct Biol* 1998;5:555–560.
44. Aleshin AE, Zeng C, Bourenkov GP, Bartunik HD, Fromm HJ, Honzatko RB. The mechanism of regulation of hexokinase: new insights from the crystal structure of recombinant human brain hexokinase complexed with glucose glucose-6-phosphate. *Structure* 1998;6:39–50.
45. Rosano C, Sabini E, Rizzi M, Deriu D, Murshudov G, Biachi M, Serafini G, Magnani M, Bolonesi M. Binding of non-catalytic ATP to human hexokinase I highlights the structural components for enzyme-membrane control. *Structure* 1999;7:1427–1437.
46. Hayward S, Berendsen HJC. Systematic analysis of domain motions in proteins from conformational change: New results on citrate synthase T4 lysozyme. *Proteins* 1998;30:144–154.
47. Sols A, Crane RK. Substrate specificity of brain hexokinase. *J Biol Chem* 1954;210:581–595.
48. Kraakman LS, Winderickx J, Thevelein J, de Winde JH. Structure-function analysis of yeast hexokinase: structural requirements for triggering cAMP triggering catabolite repression. *J Biochem* 1999;343:159–168.
49. Arora KK, Filburn CR, Pedersen PL. Glucose phosphorylation. Site-directed mutations which impair the catalytic function of hexokinase. *J Biol Chem* 1991;266:5359–5362.
50. Xu LZ, Harrison RW, Weber IT, Pilakis SJ. Human β -cell glucokinase. Dual role of Ser-151 in catalysis hexose affinity. *J Biol Chem* 1995;270:9939–9946.
51. DelaFuente G, Sols A. The kinetics of yeast hexokinase in the light of the induced fit involved in the binding of its sugar substrate. *Eur J Biochem* 1970;16:234–239.
52. DelaFuente G. Specific inactivation of yeast hexokinase induced by xylose in the presence of a phosphoryl donor substrate. *Eur J Biochem* 1970;16:240–254.
53. Heidrich K, Otto A, Behlke J, Rush J, Wenzel KW, Kriegl T. Auto-phosphorylation-inactivation site of hexokinase 2 in *Saccharomyces cerevisiae*. *Biochemistry* 1997;36:1960–1964.
54. Shoham M, Steitz TA. Crystallographic studies model building of ATP at the active site of hexokinase. *J Mol Biol* 1980;140:1–14.
55. Shoham M, Steitz TA. The 6-hydroxymethyl group of a hexose is essential for the substrate-induced closure of the cleft in hexokinase. *Biochim Biophys Acta* 1980;705:380–384.
56. Aleshin AE, Kirby C, Liu X, Bourenkov GP, Bartunik HD, Fromm HJ, Honzatko RB. Crystal structures of mutant monomeric hexokinase I reveal multiple ADP binding sites conformational changes relevant to allosteric regulation. *J Biol Mol* 2000;296:1001–1015.
57. Kabsch W, Mannherz HG, Suck D, Pai EF, Holmes KC. Atomic structure of the actin:DNase I complex. *Nature* 1990;347:37–44.
58. Flaherty KM, DeLuca-Flaherty C, McKay DB. Three-dimensional structure of the ATPase fragment of a 70K heat-shock cognate protein. *Nature* 1990;346:623–628.
59. Bork P, Sander C, Valencia A. An ATPase domain common to prokaryotic cell cycle proteins. Sugar kinases, actin, hsp70 heat shock proteins. *Proc Natl Acad Sci* 1992;89:7290–7294.
60. Aleshin AE, Zeng C, Bartunik HD, Fromm HJ, Honzatko RB. Regulation of hexokinase I: crystal structure of recombinant human brain hexokinase complexed with glucose phosphate. *J Mol Biol* 1998;282:345–357.

PROTEINS:
Structure, Function, and Bioinformatics

Systematic structural studies of iron superoxide dismutases from human parasites and a statistical coupling analysis of metal binding specificity

Journal:	<i>PROTEINS: Structure, Function, and Bioinformatics</i>
Manuscript ID:	Prot-00651-2008
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	08-Oct-2008
Complete List of Authors:	Bachega, Jose Fernando; Universidade de São Paulo (USP), Instituto de Física de São Carlos (FFI), Física e Informática Navarro, Marcos Vicente; Universidade de São Paulo (USP), Instituto de Física de São Carlos (FFI), Física e Informática; Cornell University, Veterinary Medical Center Bleicher, Lucas; Universidade de São Paulo (USP), Instituto de Física de São Carlos (FFI), Física e Informática Bortoleto-Bugs, Raquel; Universidade de São Paulo (USP), Instituto de Física de São Carlos (FFI), Física e Informática; Centro de tecnologia da informação (CTI) Renato Archer, Centro de tecnologia da informação (CTI) Renato Archer Dive, Daniel; Institut Pasteur, INSERM U547 Hoffmann, Pascal; Université Paul Sabatier, UMR/CNRS 5068 Viscogliosi, Eric; Institut Pasteur,, INSERM U547 Garratt, Richard; Universidade de São Paulo (USP), Instituto de Física de São Carlos (FFI)
Key Words:	iron superoxide dismutase, metal specificity, Chagas' disease, statistical coupling analysis, Trypanosoma cruzi, Trypanosoma brucei, Plasmodium falciparum, crystal structure, malaria, sleeping sickness

**Systematic structural studies of iron superoxide dismutases from human parasites
and a statistical coupling analysis of metal binding specificity**

**Bachega, J.F.R. †*, Navarro, M. V. A. S. †‡*, Bleicher, L. †, Bortoleto-Bugs, R.
K. †♦, Dive, D. §, Hoffmann, P. ‡, Viscogliosi, E. §, and Garratt, R.C. †**

†Institute of Physics of São Carlos, University of São Paulo, Avenida Trabalhador
São-carlense 400, CEP 13560-970, São Carlos, SP, Brazil.

§ INSERM U547, Institut Pasteur, 1 rue du Professeur Calmette, BP 245, 59019 Lille
cedex, France.

‡ UMR/CNRS 5068, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse
cedex 4, France.

¶ Current address: C4-173 Veterinary Medical Center, Cornell University, Ithaca, NY
14853

♦Current address: Centro de tecnologia da informação (CTI) Renato Archer, Rodovia Dom
Pedro I, km 143, 6, Bairro Amaraís, CEP 13069-901 Campinas, SP, Brazil

Corresponding author: richard@if.sc.usp.br

* These authors contributed equally to this work.

KEYWORDS: *Plasmodium falciparum*, *Trypanosoma cruzi*, *Trypanosoma brucei*,
malaria, Chagas' diseases, sleeping sickness, iron superoxide dismutase, crystal
structure, statistical coupling analysis, metal specificity

ABSTRACT

Superoxide dismutases (SODs) are a crucial class of enzymes in the combat against intracellular free radical damage. They eliminate superoxide radicals by converting them into hydrogen peroxide and oxygen. In spite of their very different life cycles and infection strategies, the human parasites *Plasmodium falciparum*, *Trypanosoma cruzi* and *Trypanosoma brucei* are known to be sensitive to oxidative stress. Thus the parasite Fe-superoxide dismutases have become attractive targets for novel drug development. Here we report the crystal structures of FeSODs from the trypanosomes *T. brucei* at 2.0 Å and *T. cruzi* at 1.9 Å resolution, and that from *P. falciparum* at a higher resolution (2.0 Å) to that previously reported. The homodimeric enzymes are compared to the related human MnSOD with particular attention to structural aspects which are relevant for drug design. Although the structures possess a very similar overall fold, differences between the enzymes at the entrance to the channel which leads to the active site could be identified. These lead to a slightly broader and more positively charged cavity in the parasite enzymes. Furthermore, a statistical coupling analysis (SCA) for the whole Fe/MnSOD family reveals different patterns of residue coupling for Mn and Fe SODs, as well as for the dimeric and tetrameric states. In both cases, the statistically coupled residues lie adjacent to the conserved core surrounding the metal center and may be expected to be responsible for its fine tuning, leading to metal ion specificity.

INTRODUCTION

Tropical protozoal parasite infections such as malaria, African sleeping sickness and Chagas' disease represent a major global health problem, causing the death of approximately 1.3 million people per year¹ and having serious consequences on economic development in concerned countries. A variety of limiting factors associated with the current treatments, mainly drug resistance for malaria, low efficacy and toxicity² of drugs for sleeping sickness and Chagas' disease, and the limitation of future prospects in term of vaccination justifies the need for the continuing development of efficient drugs against these developing-world diseases. From among many plausible candidates, the enzymes involved in antioxidant defence have emerged as potential new drug targets³. Despite distinct life cycles, the causative agents of malaria, African sleeping sickness and Chagas disease in humans, *Plasmodium falciparum* (the most deadly *Plasmodium* species), *Trypanosoma brucei* and *Trypanosoma cruzi*, respectively, are all exposed to intense oxidative stress, during at least in one stage of their life cycles^{4,5}. However, their antioxidant defence systems can be overcome and an imbalance in the levels of reactive oxygen species (ROS), such as those caused by some antimalarial and antitrypanosomal drugs^{6,7,8}, could lead to parasite death. Additionally, the antioxidant defences found in mammalian hosts present major differences when compared to the parasitic organisms³. In particular, the enzyme responsible for the first line of defence against ROS, superoxide dismutase (SOD)⁹, is represented by different types (containing different metal co-factors) in hosts and parasites, leading to the belief that, the parasitic SODs may represent an alternative molecular drug target for development of specific chemotherapeutic agents. This proposition is further supported by recent antisense RNA experiments in trypanosomatids,

which demonstrate a significantly reduced survival of SOD-deficient parasites in macrophages¹⁰ as well as in culture¹¹.

SODs are metalloenzymes that catalyze the dismutation of superoxide radicals to produce molecular oxygen and hydrogen peroxide. In this reaction, the metal species plays a central role as the redox-active center^{12,13}:

Depending on metal selectivity and structural organization, the SODs can be divided into three different classes: manganese- or iron-containing SODs (MnSOD and FeSOD), which have a common ancestral origin and present a two-domain fold comprised mainly of α -helices and one three-stranded anti-parallel β -sheet^{14,15}; copper/zinc-containing SODs (Cu-ZnSOD), which contain both metals at the active site and adopt a flattened eight-stranded Greek-key β -barrel conformation^{16,17}; and Ni-containing SODs (NiSOD), which have been recently identified in *Streptomyces sp*¹⁸ and in several bacterial genomes^{19,20}, and present a distinct four-helix anti-parallel bundle topology^{21,22}. The different SOD types possess a widespread distribution in both eukaryotic and prokaryotic cells, appearing either in the cytosol, periplasm or compartmentalized in organelles such as mitochondria and chloroplasts^{15,23,24,25,26,27}. For instance, while the protozoan parasites *P. falciparum*, *T. cruzi* and *T. brucei* present only different isoforms of the iron-containing SOD, distributed in the cytosol, glycosomes, and mitochondria^{11,28,29,30,31,32}, their host human cells possess a mitochondrial MnSOD³³ and a cytosolic Cu-ZnSOD.

To date, a representative collection of about 20 unambiguous Mn/Fe-containing SOD structures are available in the protein data bank³⁴. Independent of the metal cofactor,

both forms can be present as either homodimers or homotetramers (a dimer of dimers). Their monomers display a similar overall fold and a virtually identical metal-binding site. However, functionality of the majority of Mn/FeSODs strictly depends on metal selectivity³⁵, with the exception of the “cambialistic” SODs, which are active when either of the two metals are present^{36,37}. Indeed, the interchange of metals (Mn or Fe) in native structures almost invariably results in nearly complete loss of activity^{38,39}. The most convincing attempts to explain these observations are based on redox tuning by the polypeptide which allows the metal center to undergo redox changes during catalysis^{41,42}. Furthermore, experiments have shown that the incorporation of iron into *E. coli* MnSOD leads to a significant change in the metal environment⁴⁰, providing a structural basis for the redox potential tuning hypothesis. This observation, together with the metal coordination invariance during the oxide-reduction cycle of native Mn/FeSODs (42), suggests that SOD sequences impose the correct geometry on specific metals, leading to catalytic active proteins. Although systematic and exhaustive analyses of Mn/FeSOD sequences and structures have identified specific fingerprints for the Mn- and Fe-containing proteins^{44,45}, the correlation between sequence and metal on the one hand and geometry and activity on the other, is still not well understood.

Here we present for the first time, the crystal structures of the *T. brucei* and *T. cruzi* FeSODs (TbSODB2 and TcSODB1, respectively) solved to 2.0 Å and 1.9 Å resolution, respectively, as well as a higher resolution structure of the *P. falciparum* enzyme (PfSOD1) (2.0 Å), when compared to that previously reported⁴⁶. A systematic comparison between the parasitic FeSODs and human MnSOD was performed to assess potential differences that could be exploited for the development of specific inhibitors. Finally, we applied statistical residue coupling analyses to the entire Mn/FeSOD family, based on a large multiple

sequence alignment in order to address the question of metal specificity. The identification of distinct Fe- and Mn- residue paths surrounding the SOD active site provides useful insights into the metal-tunability imposed by the polypeptide chain.

EXPERIMENTAL PROCEDURES

Protein Expression and Purification – Recombinant PfSOD1 was expressed in *E. coli* and purified as previously described⁴⁷. TcSODB1 was expressed in fusion with glutathione-S-transferase in *E. coli* strain ER2566, using the prokaryotic expression vector pGEX-2T (Amersham Biosciences) as previously described⁴⁸, and purified using a slightly modified protocol. Briefly, cells were grown in Luria-Bertani medium up to an optical density of $OD_{600\text{ nm}} = 0.6$, followed by induction with 0.4 mM IPTG for 20 h at 20°C. The culture was then cooled and bacterial cells were harvested by centrifugation at 4000g for 30 min. The cell pellet was resuspended in 20 mL of PBS buffer (140 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4 , 1.8 mM KH_2PO_4 , pH 7.4) containing 5 mM DTT. After sonication, the lysed cells were centrifuged at 17,000g for 20 min at 277 K. The supernatant containing the fusion protein was mixed with 4 mL of glutathione Sepharose 4 fast flow (Amersham Biosciences) pre-equilibrated with 0.15M PBS. The suspension was incubated with gentle agitation for 3 h at 298 K and centrifuged at 500g for 5 min. The supernatant was removed and the resin pellet washed with 10 mL of cleavage buffer (50 mM Tris, 150 mM NaCl, 2.5 mM CaCl_2 , pH 8.0). After three such washes, 4 mL of buffer and 80 units of bovine thrombin (Amersham Biosciences) were added. The mixture was incubated for 48 h at 293 K with two more additions of 80 units of thrombin being made after 16 and 32 h

respectively. Finally, the mixture was centrifuged under identical conditions and the supernatant containing the cleavage TcSODB1 was further purified by cation exchange chromatography using a monoQ HR 5/5 column (Amersham Biosciences) equilibrated in 50 mM Tris-HCl buffer. TbSODB2 was expressed in *E. coli* BL21(DE3) using the expression vector pET28a (Novagen). The protein expression and preparation of the cell lysate followed the same methods applied for the *T. cruzi* enzyme. The soluble protein was bound to nickel-agarose affinity resin (Qiagen) and washed with a buffer containing 20 mM Tris, pH 8.5, 250 mM NaCl, and 10 mM imidazole. His₆-tagged protein was then eluted with a buffer containing 20 mM Tris, pH 8.5, 250 mM NaCl, and 150 mM imidazole and subsequently treated with 5U/mg of thrombin for 16 hours to remove the 6xHis-tag. The purified proteins, which were >90% pure as judged by SDS-PAGE, were found homogenous according to dynamic light scattering (DLS) measurements and catalytically active, as determined by the pyragollol autooxidation method (48).

Crystallization and Data Collection – The proteins were concentrated to 5-8 mg/mL and the hanging drop method was used for crystallization screening. The best crystals were obtained by mixing the proteins with an equal volume of reservoir solution composed of 0.1 M Tris pH 8.5, 24-30% PEG4000 and 0.2 M MgCl₂ (TbSODB2 and TcSODB1) or 0.17 M Sodium Acetate (PfSOD1). Crystals grew to about 0.5 mm in the largest dimension and belonged to space groups P2₁, C2 and P2₁2₁2₁ for TbSODB2, TcSODB1 and PfSOD1, respectively. The unit-cell parameters are listed in Table 1, and the asymmetric units contained either one (TcSODB1 and PfSOD1) or two dimers (TbSODB2), with respective Matthews' coefficients of 2.3 Å³Da⁻¹, 2.2 Å³Da⁻¹ and 2.3 Å³Da⁻¹, which suggested an estimated solvent contents of 45%, 44% and 46%, for

TbSODB2, TcSODB1 and PfSOD1, respectively. Prior to data collection, the crystals were cryoprotected by rapid soaking in solution containing mother liquor with the addition of 20% (v/v) ethylene glycol and flash-frozen in liquid nitrogen. Diffraction data were collected at 100 K on a Rigaku rotating anode X-ray generator equipped with R-Axis IIC (TcSODB1 and PfSOD1) or MAR345dtb (TbSODB2) imaging plates. The data were processed and scaled using the programs MOSFLM⁵⁰ and SCALA⁵¹ from the CCP4 package (Collaborative Computational Project, Number 4, 1994). A summary of diffraction data statistics are shown in Table 1.

Structure Determination and Refinement – The structures were solved by molecular replacement using the program MOLREP⁵². The crystal structure of the *E. coli* FeSOD (Protein Data Bank code 1ISA) was used as search model in the cases of TcSODB1 and PfSOD1. The refined model of TcSODB1 was used to solve the structure of TbSODB2. For the refinement, cycles of interactive model building with the program COOT⁵³ was alternated with refinement carried out with the program REFMAC⁵⁴. Tight 2-fold noncrystallographic symmetry restraints (300 kcal/mol-Å²) were used in the early stages of refinement and gradually relaxed until significant differences were observed between the individual subunits in the asymmetric unit. The final values of *R*-work and *R*-free are 15.8% and 19.4%, respectively, for PfSOD1; 15.7% and 19.8% respectively for TcSODB1 and 18.3% and 21.9% respectively for TbSODB2. The refinement statistics are given in Table 1. The final models and structure factors for PfSOD1, TcSODB1 and TbSODB2 were deposited in the Protein Data Bank under the PDB IDs 2GOJ, 2GPC and 3ESF.

Electron paramagnetic resonance – PfSOD1 at a concentration of approximately 80μM in 50mM MES, 100 mM NaCl at pH 5.0 or alternatively in 50mM Tris-HCl, 100mM NaCl, pH 7.0 was used for CW-EPR measurements. These were undertaken at 4K using a Bruker-ELEXSYS E580 spectrometer equipped with a rectangular cavity in continuous wave mode at X-band frequency (9.5GHz).

Statistical coupling Analysis (SCA) –

In order to investigate the possible relationship between statistically coupled sequence positions and metal specificity/oligomeric state, we have applied Statistical Coupling Analysis (SCA -ref) to the family as a whole. This method relies on the availability of a large Multiple Sequence Alignment (MSA) allowing for the calculation of two statistical parameters: ΔG^{stat}_i , which reflects sequence conservation at each alignment position i , and $\Delta \Delta G^{\text{stat}}_{ij}$ which measures the dependence of the amino acid distribution at i , given a specific perturbation at j (ie, after fixing a given amino acid type at this position). When $\Delta \Delta G^{\text{stat}}_{ij}$ is large, this implies that there is a high correlation between the types of amino acid found at positions i and j , indicating that they are coupled. Groups of residues which are highly coupled on such a pairwise basis can subsequently be grouped into clusters, which often connect residues in the three-dimensional structure which are of functional importance^{55,56}.

Superoxide dismutase sequence codes were obtained from the Pfam server⁵⁷, and the sequences themselves were obtained from ExpASY⁵⁸. These were aligned in MAFFT⁵⁹, and then adjusted to the structured based format in Jalview⁶⁰. Reference sequences for which the oligomeric state and metal affinity are well known^{44,45}, including those solved in this paper, were added where necessary. The MSA was edited using the TbSODB2 as

reference sequence, excluding the gaps caused by non-conserved regions (leading to 208 positions in the final alignment), and eliminating sequences with more than 90 percent identity (resulting in 502 sequences). Residue numbers cited throughout this paper, therefore, refer to the TbSODB2 sequence. A sequence alignment is given in the supplementary material.

Positional conservation (ΔG^{stat}) and statistical coupling ($\Delta\Delta G^{\text{stat}}$) were calculated as defined by Lockless and Ranganathan⁵⁵, using our own programs, written in C++. To calculate the minimum sub-alignment size for each perturbation, we have used the procedure described in Süel et al⁵⁶ (supplementary material), which measures the increase of ΔG^{stat} for unconserved positions upon random elimination of sequences in the alignment. This demonstrated that a sub-alignment should have at least 30% of its original size to be statistically significant. Using this criterion, 167 perturbations were possible, some of which correspond to different amino acid perturbations at the same alignment position. In order to identify clusters of statistically coupled positions, the 208x167 $\Delta\Delta G^{\text{stat}}$ matrix (where each element $\Delta\Delta G^{\text{stat};ij}$ correspond to the response in position i upon the perturbation defined in j) was subjected to hierarchical clustering in Matlab Ver. 7.0 (Mathworks, Natick, MA), using the complete linkage method upon columns and rows.

RESULTS

Overall Description of the Structures –The refined models of TcSODB1 and PfSOD1 consist of two polypeptide chains, two iron ions and 458 and 305 water molecules, respectively. A magnesium ion (derived from the crystallization solution) can be observed in the structure of TcSODB1. The TbSODB2 model consists of four polypeptide chains arranged as two dimers. Each chain contains one iron ion and 708 solvent water molecules were identified in the structure. The electron density maps are contiguous for all modeled chains. However, some residues either at the N-terminus (first residue of PfSOD1 and TcSODB1) or C-terminus (final 3 residues of PfSOD1 and 11 residues of TbSODB2) could not be observed. Several residues were modeled in alternative conformations, including the PfSOD1 His143, which participates in an important salt bridge with the Asp144 at the dimeric interface. An unexpected oxidation was found at residue Cys83 from both chains of the *T. cruzi* structure. This artifact probably occurred during thrombin cleavage of the fusion protein which was performed in the absence of DTT, and has no physiological relevance.

The overall structures are similar to those of previously solved iron superoxide dismutases. The monomer consists of two structural domains: an N-terminal helical domain composed of two long antiparallel helices ($\alpha 1$ and $\alpha 3$), separated by a loop containing the small helix $\alpha 2$. This loop is replaced by extensions to the long helices $\alpha 1$ and $\alpha 3$ which project towards the solvent in the structures of *Sulfolobus solfataricus* and in the human manganese containing enzyme (HuMnSOD), and are important for tetramerization⁶⁰. This region, which is distant from both active site and dimer interface, presents the largest differences in the mainchain conformation for the three structures. The C-terminal domain presents an α/β -fold and consists of a three-stranded anti-parallel β -sheet with helices $\alpha 4$,

1
2
3 $\alpha 5$, $\alpha 6$, and $\alpha 7$ disposed on one side. Additionally two small single-turn helical structures
4
5 are observed immediately prior to helix $\alpha 6$ and strand $\beta 3$ (Fig. 1a). These are not
6
7 specifically labeled in order to retain consistency with the previously described structure of
8
9 PfSOD1⁴⁶.
10
11
12
13
14
15
16

17 ***The dimeric structure and the monomer-monomer interface*** – As expected, all three
18
19 structures present a similar dimeric arrangement to that observed previously for
20
21 homologous FeSODs (Fig 1b). The principal intersubunit contacts are made between
22
23 residues from $\alpha 1$ and the connections at one end of the β sheet. Overall 17 residues from
24
25 both chains are involved in the interface. The most notable features common to all three
26
27 structures include (1) the highly conserved Ser123, whose interaction with its symmetry
28
29 mate from the neighbouring subunit involves two water molecules locked into the middle
30
31 of the interface; (2) the salt bridge between Glu164 and His165, which is a ligand to the
32
33 iron and provides direct communication between the active site and the dimer interface and
34
35 (3) the salt bridge between His143 and Asp144 of the neighbouring subunit close to the
36
37 two-fold axis. This region has high solvent access and in PfSOD1 His143 presents two
38
39 conformations, in which the second orientation of the imidazole group interacts with
40
41 solvent, leading to breakage of the salt bridge. This disorder was not observed in the
42
43 previously reported structure of PfSOD1, possibly due to the lower resolution or
44
45 conceivably due to the lower pH of the crystals which would favour the maintenance of the
46
47 positive charge on the histidine sidechain. However, the latter seems unlikely given that
48
49 the FeSOD structure from *P.berghei*, which was crystallized at pH 4.0, presents similar
50
51
52
53
54
55
56
57
58
59
60

disorder. Furthermore, neither TbSODB2 nor TcSODB1 present evidence for a second conformation.

The two dimers in the asymmetric unit of TbSODB2 are linked by a water layer, and make only two direct hydrogen bonds. This minimal contact is presumably important for crystal packing but is unlikely to significantly affect the structure of the dimer, which is the biologically relevant unit.

The Active Site – The active site geometry is well defined by the electron density map (Fig. 2a) and is effectively identical to that reported for homologous SODs. Furthermore, there are no significant differences between the three structures reported here. All present the five ligands (His28, His76, His165, Asp161 and a water molecule) spatially arranged in a trigonal bipyramid around the iron atom²⁶. Beyond the ligands themselves, the active site neighbourhood also shows the well-conserved hydrogen bonding network involving Trp125, Gln72, Tyr36, His32 and Tyr168 from the vicinal subunit⁶². In the case of PfSOD1, this is consistent with the single EPR signal observed at pH 7.0 from a sample identical to that used for crystallization (Fig. 3). This signal is similar to those reported for many other FeSODs⁴⁷.

The active site lies at the bottom of a deep funnel formed mainly by $\alpha 1$ and $\alpha 3$ from one subunit and $\alpha 5$ from the other. Significant differences between the three parasite FeSODs reported here and the human MnSOD can only be observed towards the mouth of this funnel. Figure 4 shows the electrostatic potential surfaces in the vicinity of the active site. Significant differences can be seen in the charge distribution as a consequence of non-

conservative substitutions, the most significant being a Lys37 in the parasite enzymes which replaces asparagine in HuMnSOD, leading to a greater positive charge density at the channel entrance. The shape of the access channel is also altered, being slightly narrower in the human enzyme due to the presence of Ala in place of Gly at position 33 and Pro in place of Ala at 175 (although the *T. brucei* enzyme also have proline at this position).

Statistical coupling analysis

The SCA procedures yielded three clusters of coupled positions, as can be seen in Fig. 5. The appearance of more than one cluster after an SCA procedure has been previously reported⁵⁶, but no interpretation for this phenomenon was presented. For SODs, the most well-defined cluster, which lies at the left of the figure (cluster 1), has highly correlated positions which define the oligomeric state of the enzyme. Glu56, Asn68, Thr24, Pro148 and Phe121, as pointed out by Wintjens et al^{44,45}, are characteristic of SOD dimers and form a self-consistent cluster, as shown in Table 2.

The table shows that the presence of one of these residues in each position is generally accompanied by the presence of the others. This demonstrates that the appearance of a well-defined cluster derived from SCA can be caused by the existence of different characteristics (in this case, oligomeric states) in a single protein family, and this characteristic is largely determined by a small number of positions. On the top right portion of the figure, it is possible to identify a number of residues which, according to the analysis of Wintjens et al^{44,45}, are characteristic of either Fe dimers (Phe78, Ala71, Ala145, Gln72 and Phe67), or FeSOD in general (Asp170). We refer to this as cluster 2 and it includes two additional residues (Ala70 and Trp74), not identified by Wintjens *et al*, whose presence

correlates strongly with those previously described. The third and final cluster contains the positions which have been identified as Mn-specific: Gln145, Gly72, Asp146 and Met25^{44,45} as well as three additional residues: Ala43, Gly159 and Gly71. Figure 6 shows how the statistically coupled residues of each cluster map onto the three-dimensional structure (TbSODB2 taken as a reference). In the case of cluster 1, as expected, most of the statistically coupled positions lie towards the dimer interface. Positions 56 and 148 are the only exceptions and will be discussed below. In the case of clusters 2 and 3, these can be seen to map to regions which are distant from the active site but which make contact with the conserved core of the dimer. This core is defined as being composed of those residue positions whose ΔG^{stat} values are >80% of the maximum observed for the entire alignment. Clusters 2 and 3 (characteristic of Fe and Mn SODs respectively) are composed of different residue positions but converge on positions 72 and 145, where a glutamine residue performs the equivalent function in Fe and Mn enzymes respectively, stabilizing a water molecule bound to the metal ion.

Discussion

Based on the appearance of drug resistant strains and the general lack of investment in neglected diseases, it has been argued that there is a need to search for and validate new drug targets for most, if not all, tropical diseases. One foundation stone for such research is knowledge of the genome sequences of the etiological agents responsible for the diseases and this has been successfully accomplished in the case of *P. falciparum*, *T. cruzi*, and *T. brucei*^{63,64,65}. A second important contribution comes from the knowledge of the 3D

structures of potential targets. Within this context superoxide dismutase has been proposed to be of potential value and we have gone some way to contributing to this effort by solving two new FeSOD structures (from *T. cruzi* and *T. brucei*) and reporting a third structure, that from *P. falciparum*, to a higher resolution to that described previously.

As an alternative, it might be argued that automated comparative modeling provides a solution to the need for determining many new 3D structures. PlasmoDB⁶⁶ is an invaluable resource for research into *Plasmodium* biology which includes automatically generated homology models for many proteins using the MODELLER software⁶⁷. It is therefore of interest to evaluate the results of such efforts when crystal structures become available. In the case of PfSOD, the modeled structure deposited in PlasmoDB is of overall high quality, presenting an RMS deviation of 0.54Å (over 179 C α 's) from the crystal structure presented here. This is to be expected for a model based on a 54% identical homologue (*E. coli* FeSOD). Furthermore, only one alignment error, around position 150, leads to a badly modeled loop. However, some of the disadvantages of automated high-throughput modeling are also made apparent. Notably, several side chain conformations are incorrectly modeled at the dimer interface and at the active site. These are at least partly the result of not including the active site iron during model construction and of modeling a non-physiological oligomeric state, namely a monomer. We conclude that whilst undeniably useful in the absence of a crystal structure, such models should be treated with caution when generated by high-throughput automated methods alone.

Indeed, the active site geometry of all eight independent monomers solved in this study is very similar to those observed in other iron-containing SODs. This is of particular

interest in the case of PfSOD1 because an unusual EPR spectrum has been reported for this enzyme which would be expected to reflect local changes in the active site geometry which are not observed here. On the contrary, our results provide no structural basis which would justify the unusual signal at $g = 4.38$ reported previously⁴⁷. We therefore repeated the EPR measurements on the highly active sample used for crystallization. At pH 7.0, the spectrum showed effectively no evidence for the $g = 4.38$ signal (Fig. 3) and was very similar to that described for other FeSODs, consistent with the crystal structure. At lower pH however, the unusual signal appears, suggesting that the originally reported spectrum may arise from a mixture of intact and partially altered molecules.

The active site itself seems to provide little potential for selective inhibitor design, given its similarity with the Mn site in human MnSOD and its limited dimensions. However, leading towards the molecular surface from the active site itself lies the substrate channel. At the lip of this channel substitutions lead to an alteration in its charge distribution and shape (Fig. 4) and these alterations are largely common to all three structures described here. They offer the best prospects for exploitation within the region of the active site, with the advantage that a single compound may be effective against all three parasite enzymes. The mechanism of action of previously described FeSOD inhibitors is still unknown and further work will be necessary in order to determine of any such compounds interact with the entrance to the substrate channel⁷⁰⁷¹.

Other important differences when compared with the human MnSOD include alterations to the dimerization interface. As has been pointed out previously⁴⁶ this could represent an alternative strategy for the design of specific inhibitors. For example, the residues His143 and Asp144 form a salt bridge in all three parasite enzymes but are

1
2
3 replaced by Pro and Asn respectively in HuMnSOD. Although far from trivial, these
4 differences might be exploitable in selective design and it is worth pointing out that the
5 small molecule, 3-(2-benzothiazolylthio)-1-propanesulfonic acid, inhibits triosephosphate
6 isomerase from *T.cruzi*, by acting in such a manner⁷².
7
8
9
10
11

12
13
14 One issue when considering the usefulness of a potential target is the existence of
15 natural genetic variability and the potential risk of resistance. In the case of PfSOD1 a
16 systematic survey of 26 isolates revealed 11 positions subject to polymorphic variation⁷³.
17 Fig. 7 shows how these map to the crystal structure. In all cases the affected positions lie
18 distant to both the dimerization interface and the active site and its associated substrate
19 channel, suggesting reasons for optimism in the use of SOD as a target.
20
21
22
23
24
25
26
27
28

29 A further aspect pertinent to the structures of proteins from *P. falciparum* relates to
30 its well-known genome-wide nucleotide bias in which A:T represents approximately 80%
31 of the entire genome. In organisms which show such a bias, this is known to result in an
32 altered amino acid composition of the encoded proteins⁶⁸. For example, A:T rich genomes,
33 such as *P. falciparum*⁶⁵, present an increased content of the amino acids FYMINK.
34 Somehow, such proteins must be capable of dealing with a distorted amino acid distribution
35 whilst retaining both three-dimensional structure and function. On comparison of the
36 amino acid content of the three superoxide dismutase structures described here, with
37 respect to the six amino acids cited above, we note that only asparagine and isoleucine are
38 significantly over-represented in the *P. falciparum* enzyme. There are 23 asparagines and
39 15 isoleucines in PfFeSOD compared with 12 Asn and 7 Ile in TcSODB1 and 10 Asn and 9
40 Ile in TbFeSOD. It is not surprising that the excess asparagine residues are distributed
41 widely over the protein surface where substitutions amongst polar amino acids are more
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

readily accommodated. What is more intriguing is the situation with isoleucine. The average solvent accessible area of the isoleucines in PffFeSOD which are conserved in at least one of the remaining two enzymes is only $7.7 \text{ \AA}^2$ whilst that of the PffFeSOD-unique isoleucines is $29 \text{ \AA}^2$. Although not completely exposed at the protein surface, the average solvent exposure is significantly greater for the latter group. This seems to suggest that the use of partially solvent exposed positions is an important factor aiding in accommodating bias at the amino acid level, even in the case of hydrophobic amino acids.

Interest in the development of inhibitors capable of distinguishing between the parasite FeSODs on the one hand and the human manganese enzyme on the other, inspired interest in better understanding the long-standing conundrum of the structural basis of metal specificity within the Fe/MnSOD family as a whole. Statistical coupling analysis provides a means to examine coevolving and structurally coupled residues which might be characteristic for a given metal, and necessary for the fine-tuning of the active site. As expected, the residues comprising the conserved core surrounding the metal ligands do not show elevated values of $\Delta\Delta G^{\text{stat}}$ as they are independent of other sequence positions. However, clusters 2 and 3, which specifically characterize Fe and Mn enzymes respectively, make direct contact with this core, and might therefore be expected to influence the metal site indirectly. Such fine tuning may require only relatively subtle changes to the active site, as shown by the study of Edward *et al.*⁴⁰ in which changes to N-Fe-N/O bond angles of the order of 20° in Fe-substituted MnSOD were sufficient to allow the coordination of a hydroxide ligand to a sixth coordination position and consequently reduced activity at neutral pH.

Further evidence that long-range fine-tuning of the active site takes place comes from site-directed mutagenesis studies of the cambialistic SOD from *Porphyromonas gingivalis* which was converted to an Fe-specific enzyme by substitution of Gly155 (Gly159 in the numbering scheme adopted here) by Thr. This residue, which is located approximately 10Å from the metal, is part of cluster 3 suggesting that other members of the cluster may also play a role in determining metal ion specificity. This hypothesis could be readily tested by future mutagenesis studies.

The results of the SCA go some way towards addressing what is sometimes seen to be a conundrum in protein structure, namely the disparity between the size of a full protein structure compared with that of its active site, which typically only involves a handful of residues. Often this is understood in terms of the provision of an appropriately stable three-dimensional framework which is necessary for placing the catalytic residues in the correct orientation. In the case of Fe/Mn SODs, those residues identified by statistical coupling form a further shell of residues extending well beyond the metal site itself. These residues which are probably implicated in the fine tuning of metal specificity, help justify the need for such a large polypeptide chain even for the binding of a biological minute substrate. Similar observations have been made with respect to the specificity-determining regions of serine proteases⁶⁹.

Our SCA results follow on from comprehensive studies by Wintjens *et al*^{44,45} in which a different approach was used to identify residues which characterize both metal specificity and/or oligomerization state and then apply the findings in an algorithm to predict these characteristics from amino acid sequence alone. There is overall good agreement between the results of both methods. In the case of the Fe-specific cluster two

additional residues not specifically described by Wintjens *et al*^{44,45} (A70 and A74) were identified by SCA whereas for the Mn cluster, three extra residues, including Gly159 described above, were statistically coupled to the main group.

The appearance of these additional positions suggests that they may also be involved in Mn/Fe discrimination and may present predictive value. We analyzed the presence of each of these residues in the SODs whose binding site metal has already been experimentally determined (Wintjens *et al*, supplementary material) and evaluated their capacity to discriminate between Fe and MnSOD. For FeSODs, residue A70 leads to only 12% false positives (i.e. only 8 out of the 69 sequences which have an alanine at position 70 are not Fe-SODs) and 29% false negatives (ie. 24 out of 85 FeSODs do not present an alanine at this position). This rate of false negatives is very similar to that observed for the Fe-specific residues identified previously (27% on average). However, the level of false positives is slightly higher for A70, being virtually zero for the previously described residues. Residue W74 on the other hand, is a particular case: although it presents no false positives, it has a very high degree of false negatives (65%). Its presence is therefore a strong indicator of iron-containing SODs. Indeed Meier *et al*⁷⁵, have observed that W74 seems to be related to inactivation by H₂O₂, a feature that is absent in Mn-SODs.

In the case of the Mn cluster, the three previously unidentified residues A43, G159 and G71, have an average of 14% false positives and 20% false negatives, which is very similar to the predictive power of the residues described previously (D146, M25, G72 and Q145) which on average present 10% false positives and 23% false negatives (Full results for all residues described here are included as supplementary materials). Overall these data suggest that the novel residues identified here, which are statistically coupled to those

1
2
3 already described as characteristic for a given metal, are approximately equally powerful
4
5 discriminators of metal specificity and may be useful in giving additional predictive power
6
7 to the novel method described by Wintjens et al.
8
9

10
11 The third cluster we identify (cluster 1) shows very strong statistical coupling and
12
13 includes those positions which are characteristic of dimeric structures as opposed to
14
15 tetramers. One additional residue to those described previously (E56) is distant from the
16
17 dimer interface and is found within a region which forms part of a second inter-subunit
18
19 interface in tetramers. E56 is can therefore probably best be described as anti-tetrameric
20
21 and therefore statistically (but not structurally) coupled to the remained of the dimeric
22
23 cluster.
24
25
26
27
28

29 Previous works using the statistical coupling analysis have usually generated a
30
31 single cluster of coupled residues, which has been interpreted as the “allosteric core” of the
32
33 protein family. Our analysis shows that clusters can emerge in a SCA procedure because of
34
35 the existence of partitioned characteristics in a single protein family (in our case, the choice
36
37 for Fe or Mn as the binding site metal, and the choice between dimer or tetramer as the
38
39 oligomeric state). Since these clusters are obtained using no previous knowledge of the
40
41 function or structure of the proteins in question, groups of the residues with predictive
42
43 power to distinguish between these characteristics are obtained without the notion of a
44
45 “training set”. This unexpected new interpretation for the Statistical Coupling Analysis,
46
47 therefore, can have useful applications in the classification of novel sequences during gene
48
49 annotation.
50
51
52
53
54
55
56
57
58
59
60

Figure captions

Figure1: Overall structure of parasite FeSODs. (a) a cartoon representation of a TbSODB2 monomer in ‘chainbows’ format. The nomenclature of the secondary structure elements as used in the test is shown as are the ligands to the metal ion. (b) Global superposition of the three dimers; TbSODB2 (blue), TcSODB1 (red) and PfSOD1 (green).

Figure 2: $2F_{\text{obs}} - F_{\text{calc}}$ electron density map around the iron binding site. The direct ligands to the metal, including one solvent-derived ligand, are indicated.

Figure 3: EPR spectrum of the PfSOD1 sample used for the crystallographic analysis. The spectrum at pH 7.0 is similar to that observed for other FeSODs and shows little sign of the unusual signal at $g=4.38$ reported previously. On the other hand, at pH 5, a strong signal at $g=4.37$ is observed suggestive of pH-induced alterations to the local iron geometry.

Figure4: Electrostatic potential around the access channel leading to the active site: (a) HuMnSOD, (b) TbSODB2, (c) TcSODB1 and (d) PfSOD1. Features common to the

1
2
3 parasite enzymes, such as the positive potential due to Lys37/38/39 can be seen to differ
4
5 significantly from the human Mn enzyme. The substitution of a glycine in the parasite
6
7 enzymes by an alanine leads to minor changes to the shape of the cavity.
8
9

10
11 Figure5: Final cluster of statistically coupled residues. On the vertical axis the
12 perturbations are given and on the horizontal axis, the residue positions to which they
13 couple. Three self-consistent and largely independent clusters can be seen. Cluster 1
14 corresponds to residues associated with the dimeric state, cluster 2 with FeSODs and cluster
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Figure 6: SCA clusters mapped onto the 3D structure of TbSODB2. The highly conserved core ($\Delta G \geq \Delta G_{\max}$) is shown as a mesh. (a) Cluster 1 shows that most of the residues are located close to the dimer interface. Others, such as E56 which are statistically coupled to dimers, probably disfavour the formation of tetramers. (b) cluster 2 shows that the majority of the FeSOD associated residues lie adjacent to the conserved core where they are likely to play a role in fine-tuning the active site redox potential. (c) cluster 3 shows a similar phenomenon for the MnSOD associated residues. In (b) and (c) the active site metal and ligands are explicitly shown.

Figure 7: Positions affected by polymorphism mapped onto the structure of PfSOD1, in all cases the affected positions lie distant to both the dimerization interface and the active site and its associated substrate channel.

ACKNOWLEDGEMENTS

We would like to thank Andre L.B. Ambrosio, Otavio Thiemann, Ana Paula Ulian de Araújo and Antônio José da Costa Filho for help with X-ray data collection, protein expression, molecular biology and EPR spectroscopy respectively. We gratefully acknowledge the Brazilian funding agencies FAPESP (via the CEPID project, 98/14138-2), CAPES and CNPq, for financial support.

REFERENCES

1. World Health Organization – World Health Report 2004: Changing history. Geneva: World Health Organization. Available: <http://www.who.int/whr/2004/en>. Accessed 18 February 2008.
2. Nwaka S, Ridley RG. Virtual drug discovery and development for neglected diseases through public-private partnerships. *Nat Rev Drug Discov.* 2003;2:919-928.
3. Turrens JF. Oxidative stress and antioxidant defenses: a target for the treatment of diseases caused by parasitic protozoa. *Mol Aspects Med.* 2004;25:211-220.
4. Becker, K., L. Tilley, J. L. Vennerstrom, D. Roberts, S. Rogerson, and H. Ginsburg. Oxidative stress in malaria parasite-infected erythrocytes: host-parasite interactions. *Int. J. Parasitol* 2004;34:163–189.
5. Beaman L, Beaman BL. The role of oxygen and its derivatives in microbial pathogenesis and host defense. *Annu Rev Microbiol.* 1984;38:27-48.
6. Vennerstrom JL, Eaton JW. Oxidants, oxidant drugs, and malaria. *J Med Chem.* 1988;31:1269-77.
7. Boveris A, Docampo R, Turrens JF, Stoppani AO. Effect of beta-lapachone on superoxide anion and hydrogen peroxide production in *Trypanosoma cruzi*. *Biochem J.* 1978;175:431-439.

8. Docampo R, Moreno SN. Free radical metabolites in the mode of action of chemotherapeutic agents and phagocytic cells on *Trypanosoma cruzi*. *Rev Infect Dis*. 1984 Mar-Apr;6(2):223-238.
9. McCord JM, Fridovich I. Superoxide dismutase. An enzymatic function for erythrocuprein (hemocuprein). *J. Biol. Chem*. 1969; 244:6049-6055.
10. Ghosh S, Goswami S, Adhya S. Role of superoxide dismutase in survival of *Leishmania* within the macrophage. *Biochem J*. 2003;369:447-452.
11. Wilkinson SR, Prathalingam SR, Taylor MC, Ahmed A, Horn D, Kelly JM. Functional characterization of the iron superoxide dismutase gene repertoire in *Trypanosoma brucei*. *Free Radic. Biol. Med*. 2006; 40:193-195.
12. Bull C, Fee JA. Steady-state kinetic studies of superoxide dismutases: properties of the iron containing protein from *Escherichia coli*. *J Am Chem Soc*. 1985;107:3295-3304.
13. Lah MS, Dixon MM, Patridge KA, Stallings WC, Fee JA, Ludwig ML. Structure-function in *Escherichia coli* iron superoxide dismutase: comparisons with the manganese enzyme from *Thermus thermophilus*. *Biochemistry* 1995;34:1646-1660.
14. Stoddard BL, Howell PL, Ringe D, Petsko GA. The 2.1-Å resolution structure of iron superoxide dismutase from *Pseudomonas ovalis*. *Biochemistry*. 1990;29:8885-8893.
15. Smith MW, Doolittle RF. A comparison of evolutionary rates of the two major kinds of superoxide dismutases. *J. Mol. Evol*. 1992; 34:175-184.
16. Tainer JA, Getzoff ED, Beem KM, Richardson JS, Richardson DC. Determination and analysis of the 2 Å-structure of copper, zinc superoxide dismutase. *J Mol Biol*. 1982;160:181-217.
17. Cardoso RM, Silva CH, Ulian de Araújo AP, Tanaka T, Tanaka M, Garratt RC. Structure of the cytosolic Cu,Zn superoxide dismutase from *Schistosoma mansoni*. *Acta Crystallogr D Biol Crystallogr*. 2004;60:1569-1578.
18. Youn HD, Kim EJ, Roe JH, Hah YC, Kang SO. A novel nickel-containing superoxide dismutase from *Streptomyces* spp. *Biochem J* 1996;318:889-896.
19. Palenik B, Brahamsha B, Larimer FW, Land M, Hauser L, Chain P, Lamerdin J, Regala W, Allen EE, McCarren J, Paulsen I, Dufresne A, Pastensky F, Webb EA, Waterbury J. The genome of a motile marine *Synechococcus*. *Nature* 2003;424:1037-1042.
20. Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR, Johnson ZI, Land M, Lindell D, Post AF, Regala W, Shah M, Shaw SL, Steglich C, Sullivan MB, Tolonen A, Webb EA, Zinser ER, Chisholm SW. Genome divergence in two *Prochlorococcus* ecotype reflect oceanic niche differentiation. *Nature* 2003;424:1042-1047.

21. Barondeau DP, Kassmann CJ, Bruns CK, Tainer JA, Getzoff ED. Nickel superoxide dismutase structure and mechanism. *Biochemistry* 2004;43:8038–8047.
22. Wuerges J, Lee JW, Yim YI, Yim HS, Kang SO, Carugo KD. Crystal structure of nickel-containing superoxide dismutase reveals another type of active site. *Proc Natl Acad Sci USA* 2004;101:8569–8574.
23. James ER. Superoxide dismutase. *Parasitol Today*. 1994;10:481-484.
24. Bordo D, Pesce A, Bolognesi M, Stroppolo ME, Falconi M, Desideri A. Cu, Zn superoxide dismutase in prokaryotes and eukaryotes. In: Messerschmidt A, Huber R, Poulos T, Wieghardt K, editors. *Handbook of Metalloproteins*. Chichester: Wiley; 2001. pp. 1284–1300.
25. Stroupe ME, DiDonato M, Tainer JA. Manganese superoxide dismutase. In: Messerschmidt A, Huber R, Poulos T, Wieghardt K, editors. *Handbook of Metalloproteins*. Chichester: Wiley; 2001. pp. 941–951.
26. Miller, A-F. Fe-superoxide dismutase. In: Messerschmidt A, Huber R, Poulos T, Wieghardt K, editors. *Handbook of Metalloproteins*. Chichester: Wiley; 2001. pp. 668–682.
27. Lynch MC, Kuramitsu HK. Role of superoxide dismutase activity in the physiology of *Porphyromonas gingivalis*. *Infect Immun*. 1999;67:3367-3375.
28. Sienkiewicz N, Daher W, Dive D, Wrenger C, Viscogliosi E, Wintjens R, Jouin H, Capron M, Müller S, Khalife J. Identification of a mitochondrial superoxide dismutase with an unusual targeting sequence in *Plasmodium falciparum*. *Mol Biochem Parasitol*. 2004;137:121-132.
29. Bécuwe P, Gratepanche S, Fourmaux MN, Van Beeumen J, Samyn B, Mercereau-Puijalon O, Touzel JP, Slomianny C, Camus D, Dive D. Characterization of iron-dependent endogenous superoxide dismutase of *Plasmodium falciparum*. *Mol Biochem Parasitol*. 1996;76:125-134.
30. Ismail SO, Paramchuk W, Skeiky YA, Reed SG, Bhatia A, Gedamu L.; Molecular cloning and characterization of two iron superoxide dismutase cDNAs from *Trypanosoma cruzi*. *Mol Biochem Parasitol*. 1997;86:187-197.
31. Dufernez F, Yernaux C, Gerbod D, Noël C, Chauvenet M, Wintjens R, Edgcomb VP, Capron M, Opperdoes FR, Viscogliosi E. The presence of four iron-containing superoxide dismutase isozymes in trypanosomatidae: characterization, subcellular localization, and phylogenetic origin in *Trypanosoma brucei*. *Free Radic Biol Med*. 2005;40:210-225.
32. Kabiri M, Steverding D. Identification of a developmentally regulated iron superoxide dismutase of *Trypanosoma brucei*. *Biochem J*. 2001;360:173-177.

33. Borgstahl GE, Parge HE, Hickey MJ, Beyer WF, Hallewell RA, Tainer JA. The structure of human mitochondrial manganese superoxide dismutase reveals a novel tetrameric interface of two 4-helix bundles. *Cell*. 1992;71:107-118.
34. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol*. 1977; 112:535-542.
35. Whittaker JW. The irony of manganese superoxide dismutase. *Biochem Soc Trans* 2003;31:1318-1321.
36. Guan Y, Hickey MJ, Borgstahl GE, Hallewell RA, Lepock JR, O'Connor D, Hsieh Y, Nick HS, Silverman DN, Tainer JA. Crystal structure of Y34F mutant human mitochondrial manganese superoxide dismutase and the functional role of tyrosine 34. *Biochemistry*. 1998;37:4722-4730.
37. Auerbach G, Ostendorp R, Prade L, Korndörfer I, Dams T, Huber R, Jaenicke R. Lactate dehydrogenase from the hyperthermophilic bacterium *thermotoga maritima*: the crystal structure at 2.1 Å resolution reveals strategies for intrinsic protein stabilization. *Structure*. 1998;6:769-781.
38. Yamakura F, Suzuki K. Cadmium, chromium, and manganese replacement for iron in iron-superoxide dismutase from *pseudomonas ovalis*. *J Biochem* 1980;88:191-196.
39. Ose DE, Fridovich I. Superoxide dismutase. Reversible removal of manganese and its substitution by cobalt or zinc. *J Biol Chem* 1976;251:1217-1218.
40. Edward RA, Whittaker MM, Whittaker JW, Jameson GB, Baker EN. Distinct Metal Environment in Fe-Substituted Manganese Superoxide Dismutase Provides a Structural Basis of Metal Specificity. *J. Am. Chem. Soc.* 1998;120:9684-9685.
41. Vance CK, Miller AF. A simple proposal that can explain the inactivity of metal-substituted superoxide dismutase. *J Am Chem Soc* 1998;120:461-467.
42. Vance CK, Miller AF. Novel insights into the basis for *Escherichia coli* superoxide dismutase's metal ion specificity from Mn-substituted FeSOD and its very high E(m). *Biochemistry* 2001;40:13079-13087.
43. Ludwig ML, Metzger AL, Patridge KA, Stallings WC. Manganese superoxide dismutase from *Thermus thermophilus*. A structural model refined at 1.8 Å resolution. *J Mol Biol*. 1991;219:335-358.
44. Wintjens R, Noël C, May AC, Gerbod D, Dufernez F, Capron M, Viscogliosi E, Rooman M. Specificity and phenetic relationships of iron- and manganese-containing superoxide dismutases on the basis of structure and sequence comparisons. *J Biol Chem* 2004;279:9248-9254.

45. Wintjens R, Gilis D, Rooman M. Mn/Fe superoxide dismutase interaction fingerprints and prediction of oligomerization and metal cofactor from sequence. *Proteins*. 2008;70:1564-1577.
46. Boucher IW, Brzozowski AM, Brannigan JA, Schnick C, Smith DJ, Kyes SA, Wilkinson AJ. The crystal structure of superoxide dismutase from *Plasmodium falciparum*. *BMC Struct Biol*. 200;6:20.
47. Gratepanche S, Ménage S, Touati D, Wintjens R, Delplace P, Fontecave M, Masset A, Camus D, Dive D. Biochemical and electron paramagnetic resonance study of the iron superoxide dismutase from *Plasmodium falciparum*. *Mol Biochem Parasitol*. 2002;120:237-46.
48. Ismail SO, Paramchuk W, Skeiky YA, Reed SG, Bhatia A, Gedamu L. Molecular cloning and characterization of two iron superoxide dismutase cDNAs from *Trypanosoma cruzi*. *Mol Biochem Parasitol*. 1997;86:187-197.
49. Marklund S, Marklund G. Involvement of the superoxide anion radical in the autoxidation of pyrogallol and a convenient assay for superoxide dismutase. *Eur J Biochem*. 1974;47:469-474.
50. Leslie AG. The integration of macromolecular diffraction data. *Acta Crystallogr D Biol Crystallogr*. 2006;62:48-57.
51. Evans P. Scaling and assessment of data quality. *Acta Crystallogr D Biol Crystallogr*. 2006;62:72-82.
52. Vagin A, Teplyakov A. An approach to multi-copy search in molecular replacement. *Acta Crystallogr D Biol Crystallogr*. 2000;56:1622-1624.
53. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr*. 2004;60:2126-2132.
54. Murshudov GN, Vagin AA, Dodson EJ. Refinement of Macromolecular Structures by the Maximum-Likelihood Method. *Acta Crystallogr D Biol Crystallogr*. 1997;53:240-255.
55. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 1999; 286:295-299.
56. Süel GM, Lockless SW, Wall MA, Ranganathan R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol* 2003;10:59-69.
57. Finn RD, Tate J, Mistry J, Coghill PC, Sammut JS, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A. The Pfam protein families database. *Nucleic Acids Res*. 2008; 36D:281-288.

58. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: The proteomics Server for in-depth protein knowledge and analysis. *Nucleic Acid Res.* 2003;31:3784-3788.
59. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 2002;30:3059-3066.
60. Clamp M, Cuff J, Searle SM, Barton GJ. The Jalview Java Alignment Editor. *Bioinformatics.* 2004;20:426-427.
61. Flückiger S, Mittl PR, Scapozza L, Fijten H, Folkers G, Grütter MG, Blaser K, Cramer R. Comparison of the crystal structures of the human manganese superoxide dismutase and the homologous *Aspergillus fumigatus* allergen at 2-Å resolution. *J Immunol.* 2002;168:1267-1272.
62. Muñoz IG, Moran JF, Becana M, Montoya G. The crystal structure of an eukaryotic iron superoxide dismutase suggests intersubunit cooperation during catalysis. *Protein Sci.* 2005;14:387-394.
63. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B, Böhme U, Hannick L, Aslett MA, Shallom J, Marcello L, Hou L, Wickstead B, Alsmark UC, Arrowsmith C, Atkin RJ, Barron AJ, Bringaud F, Brooks K, Carrington M, Cherevach I, Chillingworth TJ, Churcher C, Clark LN, Corton CH, Cronin A, Davies RM, Doggett J, Djikeng A, Feldblyum T, Field MC, Fraser A, Goodhead I, Hance Z, Harper D, Harris BR, Hauser H, Hostetler J, Ivens A, Jagels K, Johnson D, Johnson J, Jones K, Kerhornou AX, Koo H, Larke N, Landfear S, Larkin C, Leech V, Line A, Lord A, Macleod A, Mooney PJ, Moule S, Martin DM, Morgan GW, Mungall K, Norbertczak H, Ormond D, Pai G, Peacock CS, Peterson J, Quail MA, Rabinowitsch E, Rajandream MA, Reitter C, Salzberg SL, Sanders M, Schobel S, Sharp S, Simmonds M, Simpson AJ, Tallon L, Turner CM, Tait A, Tivey AR, Van Aken S, Walker D, Wanless D, Wang S, White B, White O, Whitehead S, Woodward J, Wortman J, Adams MD, Embley TM, Gull K, Ullu E, Barry JD, Fairlamb AH, Opperdoes F, Barrell BG, Donelson JE, Hall N, Fraser CM, Melville SE, El-Sayed NM. The genome of the African trypanosome *Trypanosoma brucei*. *Science.* 2005;309:416-422.
64. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, Ghedin E, Worthey EA, Delcher AL, Blandin G, Westenberger SJ, Caler E, Cerqueira GC, Branche C, Haas B, Anupama A, Arner E, Aslund L, Attipoe P, Bontempi E, Bringaud F, Burton P, Cadag E, Campbell DA, Carrington M, Crabtree J, Darban H, da Silveira JF, de Jong P, Edwards K, Englund PT, Fazelina G, Feldblyum T, Ferella M, Frasch AC, Gull K, Horn D, Hou L, Huang Y, Kindlund E, Klingbeil M, Kluge S, Koo H, Lacerda D, Levin MJ, Lorenzi H, Louie T, Machado CR, McCulloch R, McKenna A, Mizuno Y, Mottram JC, Nelson S, Ochaya S, Osoegawa K, Pai G, Parsons M, Pentony M, Pettersson U, Pop M, Ramirez JL, Rinta J, Robertson L, Salzberg SL, Sanchez DO, Seyler A, Sharma R, Shetty J, Simpson AJ, Sisk E, Tammi MT, Tarleton R, Teixeira S, Van Aken S, Vogt C, Ward PN, Wickstead B, Wortman J, White O, Fraser CM, Stuart KD, Andersson B. The genome

sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science*. 2005;309:409-415.

65. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan MS, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DM, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*. 2002;419:498-511.

66. Bahl A, Brunk B, Crabtree J, Fraunholz MJ, Gajria B, Grant GR, Ginsburg H, Gupta D, Kissinger JC, Labo P, Li L, Mailman MD, Milgram AJ, Pearson DS, Roos DS, Schug J, Stoeckert CJ Jr, Whetzel P. PlasmoDB: the *Plasmodium* genome resource. A database integrating experimental and computational data. *Nucleic Acids Res*. 2003;31:212-215.

67. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*. 1993;234:779-815.

68. Singer GA, Hickey DA. Nucleotide Bias Causes a Genomewide Bias in the Amino acid Composition of Proteins, *Mol. Biol. Evol*. 2000;17:1581-1588.

69. Hedstrom L, Szilagyi L, Rutter WJ. Converting trypsin to chymotrypsin: the role of surface loops. *Science*. 1992;255:1249-1253.

70. Soulère L, Viodé C, Périé J, Hoffmann P. Selective inhibition of Fe- versus Cu/Zn-superoxide dismutases by 2,3-dihydroxybenzoic acid derivatives. *Chem Pharm Bull*. 2002;50:578-582.

71. Soulère L, Delplace P, Davioud-Charvet E, Py S, Sergheraert C, Périé J, Ricard I, Hoffmann P, Dive D. Screening of *Plasmodium falciparum* iron superoxide dismutase inhibitors and accuracy of the SOD-assays. *Bioorg Med Chem*. 2003;11:4941-4944.

72. Téllez-Valencia A, Olivares-Illana V, Hernández-Santoyo A, Pérez-Montfort R, Costas M, Rodríguez-Romero A, López-Calahorra F, de Gómez-Puyou MT, Gómez-Puyou A. Inactivation of triosephosphate isomerase from *Trypanosoma cruzi* by an agent that perturbs its dimer interface. *J Mol Biol*. 2004;341: 1355–1365.

73. Baert CB, Deloron P, Viscogliosi E, Dauchez M, Camus D, Dive D. Analysis of genetic diversity at the iron-containing superoxide dismutase locus in *Plasmodium falciparum* wild isolates. *FEMS Microbiol Lett*. 1999;181:237-243.

74. Vance CK, Miller AF. Novel insights into the basis for *Escherichia coli* superoxide dismutase's metal ion specificity from Mn-substituted FeSOD and its very high E(m). *Biochemistry*. 2001 Oct 30;40(43):13079-87.

1
2
3 75. Meier B, Sehn AP, Michel C, Saran M. Reaction of hydrogen peroxide with
4 superoxide dismutase from *Propionibacterium shermanii* – an enzyme which is equally
5 active with iron or manganese – are independent of the prosthetic metal. Arch Biochem
6 Biophys 1994;313:96-303
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For Peer Review

Table1 : Data collection and refinement statistics

	PfSOD	TcSOD	TbFeSOD
Data collection statistics			
Space group	P2 ₁ 2 ₁ 2 ₁	C2	P2 ₁
Unit cell dimensions (Å, °)			
a	55.40	109.63	68.3
b	78.41	38.98	76.5
c	88.37	91.07	77.0
β		105.85	92.70
Resolution range (Å)	31.6 - 2.0	29.5 - 1.9	30.0 - 2.0
No. of observed reflections	161346	122072	224290
No. of unique reflections	25131	27127	52771
Completeness (%) [*]	95.3 (73.3)	94.6 (69.9)	98.4 (89.7)
Redundancy	6.4 (2.8)	4.4 (2.9)	5.7(2.9)
$R_{sym}(\%)^{\dagger}$	9.8 (37.0)	6.3 (27.6)	8.0 (29.8)
$I/\sigma(I)$	7.5 (2.0)	10.6 (2.7)	17.4 (3.8)
Refinement statistics			
No. of protein atoms	3215	3109	7034
No. of solvent molecules	303	453	722
No. of Fe ³⁺ ions	2	2	4
No. of Mg ²⁺ ions		1	-
$R_{factor}(\%)^{\ddagger}$	15.6 (20.1)	15.4 (18.7)	17.3
$R_{free}(\%)^{\parallel}$	19.4 (24.2)	19.8 (24.3)	22,53
Avarege B factors (Å ²)	21.77	22.56	22.26
R.m.s deviations			
Bonds lengths (Å)	0.009	0.008	0.014
Angles (°)	1.107	1.118	1.421
Ramachandran plot [¥]			
Most favored	93.6	91.7	92.2
Additionally allowed	5.2	7.1	6.7
Generally allowed	0.6	0.6	0.6
Disallowed	0.6	0.6	0.5

^{*} Values in parentheses are for the data in last resolution shell, which were 2.11-2.00 Å for the PfSOD crystal and 1.99-1.9 Å for the TcSOD crystal.

[†] $R_{sym} = \sum |I_i - \langle I \rangle| / \sum I_i$, where I_i is the intensity of a reflection and $\langle I \rangle$ is the average intensity of that reflection

[‡] $R_{factor} = \sum ||F_{obs}| - |F_{calc}|| / \sum |F_{obs}|$

^{||} R_{free} is the R_{factor} calculated by using 5% of reflection data chosen randomly and omitted from the start of refinement

[¥]Calculated with the program PROCHECK

Table 2: Amino acid frequency changes after dimer-related perturbations. Frequencies for the full alignment at the given position are shown in parentheses in the first column.

	E56	P148	F121	N68	T24
24 (T=60.8%)	94.3%	90.0%	91.2%	92.6%	X
68 (N=61.8%)	96.7%	95.3%	96.9%	X	94.1%
148 (P=63.6%)	98.0%	X	97.5%	98.1%	95.1%
121 (F=63.4)	98.8%	97.2%	X	99.4%	94.1%

Figure 1
270x103mm (600 x 600 DPI)

Figure 2
135x101mm (600 x 600 DPI)

Figure 3
2383x1666mm (600 x 600 DPI)

Figure 4
287x211mm (600 x 600 DPI)

Figure 5
43x57mm (600 x 600 DPI)

Figure 6
135x304mm (600 x 600 DPI)

Figure 7
135x101mm (600 x 600 DPI)

Crystal structure of the IL-22/IL-22R1 complex and its implications for the IL-22 signaling mechanism

Lucas Bleicher^a, Patricia Ribeiro de Moura^a, Leandra Watanabe^a, Didier Colau^b,
Laure Dumoutier^{b,c}, Jean-Christophe Renauld^{b,c}, Igor Polikarpov^{a,*}

^a Instituto de Física de São Carlos, Universidade de São Paulo, Av. Trabalhador São-Carlense, 400, CEP 13560-970 São Carlos, SP, Brazil

^b Ludwig Institute for Cancer Research, Brussels Branch, Belgium

^c Experimental Medicine Unit, Christian de Duve Institute, Université Catholique de Louvain, Brussels, Belgium

Received 13 June 2008; revised 2 July 2008; accepted 6 July 2008

Available online 7 August 2008

Edited by Hans Eklund

Abstract Interleukin-22 (IL-22) is a member of the interleukin-10 cytokine family, which is involved in anti-microbial defenses, tissue damage protection and repair, and acute phase responses. Its signaling mechanism involves the sequential binding of IL-22 to interleukin-22 receptor 1 (IL-22R1), and of this dimer to interleukin-10 receptor 2 (IL-10R2) extracellular domain. We report a 1.9 Å crystal structure of the IL-22/IL-22R1 complex, revealing crucial interacting residues at the IL-22/IL-22R1 interface. Functional importance of key residues was confirmed by site-directed mutagenesis and functional studies. Based on the X-ray structure of the binary complex, we discuss a molecular basis of the IL-22/IL-22R1 recognition by IL-10R2.

Structured summary:

MINT-6693956:

IL22 (uniprotkb:Q9GZX6) and *IL22R1* (uniprotkb:Q8N6P7) bind (MI:0407) by X-ray crystallography (MI:0114)

© 2008 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.

Keywords: Cytokines; IL-22R1; IL-22; Interleukins; Immunology; X-ray crystallography

1. Introduction

Interleukin-22 (IL-22) is an inflammatory cytokine that is mainly produced by activated T-cells [1,2]. It was first described as an IL-9-inducible gene and called IL-TIF (interleukin-10, IL-10-related T-cell derived inducible factor) [1]. IL-22 stimulates the production of acute phase reactants and promotes the antimicrobial defense, what indicates IL-22 involvement in the inflammatory and immune responses in several organs [3,4]. It has been reported that IL-22 can suppress inter-

leukin-4 (IL-4) production, suggesting that IL-22 may have therapeutic utility in the treatment of asthma [2].

The crystal structure of recombinant human IL-22 was determined in 2002 [5], and it consists of a compact bundle of a single domain composed of six anti-parallel α -helices. Human IL-22 shares a limited amino acid sequence identity (22%) with human IL-10, and possesses 179 residues, an N-terminal hydrophobic signal peptide, two disulfide bridges, ~20 kDa in size, and three potential N-linked glycosylation sites [1]. In 2005, the IL-22 crystal structure was determined from insect-cell expressed protein [6], and no relevant change of the overall tertiary structure is detected whether glycosylated and non-glycosylated structures are compared.

IL-22 cellular responses require its binding to specific cell surface receptors belonging to the class II cytokine receptor family, as observed in all IL-10 family members [7]. These two distinct receptor chains are interleukin-22 receptor 1 (IL-22R1) (cytokine receptor family class 2 member 9, CRF2-9) and interleukin-10 receptor 2 (IL-10R2) (the second chain of the IL-10 receptor complex, CRF2-4) [2,4,8], and they are structurally related to the interleukin-10 receptor 1 (IL-10R1) receptor chain [3]. IL-22R1 is also able to bind interleukin-20 (IL-20) and interleukin-24 (IL-24), forming a heterodimer with IL-20R2 [9], while IL-10R2 is also the second chain of the IL-10 heterodimeric receptor. There is strong experimental evidence that IL-22 initially binds IL-22R1 with high affinity, and then the ternary complex is formed by sequential association of the IL-22/IL-22R1 with IL-10R2 [10,11]. The IL-22/IL-22R1/IL-10R2 ternary complex enables JAK/STAT activation, which induces the expression of IL-22 specific genes [1].

So far, no IL-22/IL-22R1/IL-10R2 crystal structure has been reported and the structural basis of this ternary complex formation remains elusive. Nevertheless, low-resolution structures of the IL-22/IL-22R1 complex were assessed by small angle X-ray scattering [12], and the IL-10/IL-10R1/IL-10R2 ternary complex has been generated by homology modeling [13]. Although experimental data about the interactions between the IL-22 binary complex and IL-10R2 are scarce, it has been suggested that IL-22/IL-22R1 complex formation would lead to conformational changes in IL-22 that would allow IL-10R2 binding, and the sequential binding of IL-10R1 would lead to effective signal transduction [10].

Here we report the crystal structure of the IL-22/IL-22R1 binary complex refined to 1.9 Å resolution. The three-dimensional structure reveals key residues that are crucial for IL-22 binding to IL-22R1. We also conducted mutagenesis and

*Corresponding author. Fax: +55 16 33739881.

E-mail address: ipolikarpov@ifsc.usp.br (I. Polikarpov).

Abbreviations: IL-22R1, interleukin-22 receptor 1; IL-22, interleukin-22; IL-10, interleukin-10; IL-10R2, interleukin-10 receptor 2; CRF2-9, cytokine receptor family class 2 member 9; CRF2-4, the second chain of the IL-10 receptor complex; IL-20, interleukin-20; IL-24, interleukin-24; IL-10R1, interleukin-10 receptor 1; FBN-III, fibronectin-III domain; IL-4, interleukin-4; IL-4R α / γ_c , IL-4 receptor α common γ chain (γ_c); IL-13R α_1 , interleukin-13 receptor α ; STAT3, signal transducer and activator of transcription 3; HEK-293, human embryonic kidney 293 cells

functional studies to confirm the importance of these residues for IL-22 binding. Furthermore, based on the X-ray structure of the binary complex, we propose a plausible model of IL-22/IL-22R1/IL-10R2 ternary assembly.

2. Materials and methods

2.1. Protein expression and purification

The recombinant human IL-22 (residues 29–179) and the extracellular domain of human IL-22R1 (residues 18–228) were expressed in *Escherichia coli* and purified from inclusion bodies, as described previously [5,12]. The IL-22 receptor complex was formed by overnight incubation of IL-22 with IL-22R1 in a 1.2:1 molar ratio at 277 K. The complex was purified by gel filtration chromatography on a Superdex 200 column, and the purified complex was concentrated to 5 mg/ml for crystallization trials.

2.2. Crystallization and X-ray data collection

Crystallization experiments were performed at 291 K and employed the hanging drop vapour diffusion method. Briefly, equal amounts of purified protein and reservoir solution were mixed and 2 µl-hanging drops were allowed to equilibrate against the reservoir solution. Suitable crystals were obtained after 5 days from solution consisting of 0.9 M sodium acetate, 1 mM Triton X-100 and 0.1 M HEPES at pH 7.5. Prior to data collection, crystals were soaked into a cryo-cooling solution composed of mother liquor containing 20% (v/v) glycerol, and flash frozen in a 100 K nitrogen stream in the W01B-MX2 beam line at Laboratório Nacional de Luz Síncrotron. A total of 139 images were collected using 1° of oscillation step, to a maximum resolution beyond 1.9 Å. The dataset was indexed, reduced, merged and scaled using HKL2000 [14].

2.3. Structure determination and refinement

Successful molecular replacement solutions for both chains (IL-22 and IL-22R1) were possible using the MrBump pipeline [15], Phaser for molecular replacement [16] and Chainsaw [17] for PDB modification. A molecule of IL-22 was easily found in the asymmetrical unit using the coordinates of hIL-22 (PDB accession code 1M4R, [5]) as a search model. In the case of IL-22R1, which has no solved structure, the coordinates of the IL-10R1 receptor (PDB accession code 1J7V, [18]) were initially used as a search model, but this approach was not successful. By allowing MrBump to seek for all possible search models in the PDB, the extracellular domain of human tissue factor (PDB accession code 1TFH, [19]) was identified as a reasonable starting model, and the molecular replacement was accomplished. The reason for successful molecular replacement using human tissue factor (which has lower sequence identity with IL-22R1) instead of IL-10R1 is related to the fact that the angle between the two fibronectin-III domain (FBN-III) in IL-22R1 is slightly larger than in IL-10R1, leading to a better Cα–Cα superpositioning, and the poor IL-10R1/IL-22R1 main-chain alignment would compromise the molecular replacement. The model was refined alternating cycles of model manipulation in Coot [20] and refinement in Phenix [21]. Stereochemical analysis was performed with PROCHECK [22]. General information and statistics for data processing and model refinement are given in Table 1.

2.4. IL-22R1 mutagenesis

The IL-22R1 cDNA was amplified by RT-PCR from the HepG2 hepatoma cell line before cloning into the pEF-BOSpuro expression vector. Mutagenesis on IL-22R1 was performed by using *Quick Change II XL Site-Directed Mutagenesis* kit (Stratagene) with specific primers: **K58A** sense [5' GTC TAC AGC ATC GAG TAT GCC ACG TAC GGA GAG AGG G 3'] and anti-sense [5' CCC TCT CTC CGT ACG TGG CAT ACT CGA TGC TGT AGA C 3']; **K58E** sense [5' ACA CGG TCT ACA GCA TCG AGT ATG AGA CGT ACG GAG 3'] and anti-sense [5' CTC CGT ACG TCT CAT ACT CGA TGC TGT AGA CCG TGT 3']; **Y60A** sense [5' GCA TCG AGT ATA AGA CGG CCG GAG AGA GGG ACT GGG 3'] and anti-sense [5' CCC AGC CCC TCT CTC CGG CCG TCT TAT ACT CGA TGC 3']; **Y60R** sense [5' GCA TCG AGT ATA AGA CGG CCG GAG AGA GGG ACT GGG 3'] and anti-sense

Table 1

Crystallographic and structural information^a

Data collection parameters	
Space group	C2
Cell parameters (Å; °)	49.2, 79.2, 92.0; 90.0, 94.6, 90.0
Wavelength (Å)	1.5418
Resolution limits (Å)	50.0–1.90 (1.97–1.90)
Completeness (%)	0.98 (0.95)
Redundancy	2.8 (2.4)
R_{merge} (%) ^b	10.6 (30.5)
I/σ	7.14 (2.82)
Total number of reflections	27 359
Reflections used for refinement/ R_{free}	25 984/1375
Refinement parameters	
Bond length mean deviation (Å)	0.012
Bond angles mean deviation (°)	1.56
R_{factor} (%)	19.0
R_{free} (%)	23.4

^aValues between parentheses refer to the last resolution shell.

^b $R_{\text{merge}} = \sum |I_i - \langle I \rangle| / \sum I_i$, where I_i is the intensity of the i th observation and $\langle I \rangle$ is the mean intensity of the reflections.

[5' CCC AGT CCC TCT CTC CGC GCG TCT TAT ACT CGA TGC 3']. Clones obtained were sequenced by using *BigDye Terminator v.3.1 Ready Reaction* (Applied Biosystem).

2.5. Transient cell transfection and luciferase assay

Human embryonic kidney (HEK) 293-EBNA cells were grown in DMEM medium supplemented with 10% FCS. The pGL3-Pap1 luciferase reporter plasmid was used as a specific promoter for signal transducer and activator of transcription 3 (STAT3) activation, whereas the pRL-TK plasmid (Promega), encoding *Renilla* luciferase, was co-transfected as an internal control of the transfection process. Transient human embryonic kidney 293 cells (HEK-293) cell transfections were carried out as previously described [23]. Briefly, HEK-293-EBNA cells were seeded in 48-well plates (1×10^5 cells/well) 1 day before transfection, and all transfections assays were carried out using the Lipofectamine 2000 method (Invitrogen), according to the manufacturer's recommendations, and using 250 ng of receptor constructs cDNA, 250 ng of pGL3-Pap1 and 100 ng of pRL-TK vector. After 5 h, transfected cells were stimulated with control medium or with human IL-22 protein for 16 h. Recombinant human IL-22 was produced in *E. coli* as previously described [9], and luciferase assays were performed using the *Dual Luciferase Reporter Assay* kit (Promega).

3. Results and discussion

3.1. Model quality

The IL-22/IL-22R1 complex structure was solved by molecular replacement and refined at 1.9 Å resolution to R_{factor} of 19.0% and R_{free} of 23.4% (Table 1). There is a single IL-22/IL-22R1 complex in the asymmetric unit cell, resulting in a solvent content of 41.9%. The final model of the complex includes IL-22R1 (residues Asp20 to Thr228) and residues Ala40 to Ile179 of IL-22, and also 309 water molecules. Polypeptide fragments ¹⁴⁹GDHG and ¹⁷²NRTY of IL-22R1, and ¹³¹TCHI-EGDDL of IL-22 were not included in the final model because of the poor or disordered electron density. The Ramachandran plot indicates that a total of 94.5% and 5.5% of the amino acid residues are in the most favoured and additionally allowed regions, respectively. No residues have been found in the disallowed regions.

3.2. Overall IL-22/IL-22R1 complex structure

Human IL-22 and IL-10 have 22% amino acid sequence identity (Fig. 1A), bind to different high affinity receptors

Fig. 1. Amino acid sequence alignment of human IL-22 and IL-10 (A), and the extracellular domains of human IL-22R1 and IL-10R1 receptor chains (B). Secondary structure elements are shown in blue. Residues are colored according to their side chain chemical properties: acidic in red, aliphatic in black and grey (small), amide in green, aromatic in brown, basic in navy, hydroxyl in pink, imino in orange and sulfur in yellow. Sequence alignments were performed with TEXshade software.

chains (IL-22R1 and IL-10R1, respectively) but share a common signaling receptor (IL-10R2) for the assembly of their active ternary complexes. The IL-22R1, IL-10R1 and IL-10R2 receptor chains all belong to the cytokine receptor family class 2 (CRF-2). The amino acid sequence of extracellular domain of IL-22R1 has 24% identity with IL-10R1 receptor (Fig. 1B).

The three-dimensional structure of the IL-22/IL-22R1 binary complex is shown in Fig. 2. IL-22 consists of a compact bundle of six anti-parallel α -helices (termed A–F) with two disulfide bridges (Cys40/Cys132 and Cys89/Cys178) [5,6]. Comparison of the receptor-bound and free IL-22 structures reveals that the largest main chain deviations are observed for residues located in the pre-A and A helices region (Asp43–Phe57), and in the AB loop (Asn68–Arg73) (Fig. 2A). The superposition of monomers A and B observed in the asymmetric unit of free hIL-22 X-ray structure (PDB code 1M4R, [5]) onto receptor-bound IL-22 yielded r.m.s. deviations of 0.41 and 0.72 Å for 118 and 120 pairs of C α atoms, respectively.

The extracellular IL-22R1 chain presents two FBN-III type domains in tandem. Each FBN-III domain contains a sandwich of two anti-parallel β -sheets comprising seven β -strands (A, B, E and C, C', F, and G), and these two FBN-III domains are connected by a small helix extending from Ser114 to His118. The N-terminal domain (D1) and the C-terminal do-

main (D2) display a slightly larger inter-domain angle than that observed in the IL-10R1 receptor chain. In IL-22R1, the angle between residues Thr120, Thr48, and Glu196 is 121°, while the structural equivalent residues Thr103, Gln30, and Gly178 in IL-10R1 form an angle of ~115°. Despite the difference observed in inter-domain angles between IL-10R1 and IL-22R1, the latter contacts IL-22 using five receptor binding loops (L2–L6), as observed for the IL-10/IL-10R1 complex [18]. The AB loop and the helix F interact with the receptor loops L2–L4, and the N-terminus of helix A makes contacts with the receptor loops L5–L6 of D2 domain (Fig. 2A).

The topology and the secondary structure organization of IL-22R1 and IL-10R1 are very similar, but in the IL-22R1 N-terminal there is a small helix (Pro21–Leu24) which is absent in IL-10R1 structure. Moreover, two disulfide bonds present in IL-10R1 (Cys35/Cys54 and Cys181/Cys202) are not structurally equivalent to those found in the IL-22R1 (Cys71/Cys79 and Cys128/Cys217) (Fig. 1B). In the IL-22R1, the first disulfide bond connects Cys71 (on β -strand C') with Cys79 (on β -strand E) in the D1 domain, and the second links β -strand A (Cys128) to β -strand G (Cys217) in the D2 domain. On the other hand, there is a counterpart of IL-22R1 Cys71/Cys79 connectivity to the Cys49/Cys57 disulfide bridge present in the extracellular domain of human tissue factor (PDB accession code 1TFH).

Fig. 2. Three-dimensional structure of IL-22/IL-22R1 complex determined by X-ray crystallography. (A) Ribbon representation of the IL-22/IL-22R1 complex, showing the IL-22 molecule (in yellow) and the IL-22R1 extracellular domain (in red). Stereo view of IL-22/IL-22R1 binding interface, showing contacts between IL-22 and IL-22R1 fibronectin-III D1 (B) and D2 (C) domains. Disulfide bridges are shown in cyan, water molecules are green, and intermolecular hydrogen bonds are shown as dotted lines.

3.3. The IL-22/IL-22R1 binding interface

Comparative analysis of the IL-22/IL-22R1 and IL-10/IL-10R1 interfaces indicates that the binding regions of both complexes are grossly similar, but differ in details. Most of the residues present in the IL-22/IL-22R1 binding interface are polar residues that establish an extensive network of hydrogen bonds between IL-22 and the receptor loops. Moreover, the presence of water molecules in the binding interface might be involved in the IL-22/IL-22R1 complex stabilization.

The formation of the IL-22/IL-22R1 complex buries 828 Å² of IL-22 and 886 Å² of IL-22R1 solvent-accessible areas, and a total of 25 amino acid residues are involved in the IL-22/IL-22R1 binding interface (Fig. 2B and C). IL-22R1 contacts IL-22 residues situated on pre-helix A (Phe47, Gln49), helix

A (Thr53, Ser64), the AB loop (Asp67, Thr70, Asp71, Arg73), and helix F (Lys162, Gly165, Glu166, Asp168, Arg175). The most important residues of IL-22R1 involved in the binding interface are located on the receptor loops L2–L6: Lys58, Tyr60, Gly61, Glu62, Leu88, Thr89, Glu90, Tyr93, Arg112, Ser114, Pro206, and Thr207.

Each FBN-III domain of IL-22R1 shows direct and water-mediated interactions with IL-22 (Table 2). At the N-terminal FBN-III domain (D1 domain), the hydroxyl group of Tyr60 makes direct contacts with IL-22 Lys162 and Glu166 residues, which are part of the IL-10 signature (¹⁶²KAIGELDLL) situated at the beginning of helix F. Besides, the backbone oxygen of Tyr60 interacts with the side chain of IL-22 Arg73 through a water molecule (w4), and the Arg73 side chain is also hydrogen

Fig. 2 (continued)

bonded to both O ϵ 1 side chain of Glu90 and the backbone oxygen of Thr89 from IL-22R1. The IL-22 Arg73 residue is lo-

cated on the AB loop, together with Asp67 and Asp71 residues that interact with Arg112 and Gly61 residues of IL-22R1,

Table 2
Water mediated interactions in the IL-22/IL-22R1 interface

IL-22			Water mediated interactions			IL-22R1		
Region	Residue	Atom	Distance (Å)	Water	Distance (Å)	Atom	Residue	Region
Pre-A	Phe47	O	3.1	w9	2.9	O	Pro206	L6
Pre-A	Phe47	O	2.6	w10	2.7	O	Thr207	L6
Pre-A	Gln49	O	3.2	w9	2.8	O	Pro206	L6
Helix A	Thr53	OG	2.9	w9	2.8	O	Pro206	L6
Helix A	Thr53	OG	2.7	w10	2.7	O	Thr207	L6
Helix A	Ser64	O	2.8	w3	2.5	OH	Tyr93	L3
AB loop	Asp67	O	2.9	w3	2.5	OH	Tyr93	L3
AB loop	Arg73	NH1	2.9	w4	2.7	O	Tyr60	L2
AB loop	Arg73	N	2.9	w5	2.7	O	Gly61	L2
AB loop	Arg73	NH2	2.9	w6–w7	2.8	O	Leu88	L3
AB loop	Arg73	NH2	2.9	w6–w7	2.7	O	Glu90	L3
AB loop	Arg73	NH1	2.9	w4	2.6	OE1	Glu90	L3
Helix F	Lys162	O	2.8	w8	2.9	NH1	Arg112	L4
Helix F	Lys162	O	2.8	w8	2.9	NH2	Arg112	L4
Helix F	Gly165	O	3.2	w1	2.8	NH1	Arg112	L4
Helix F	Asp168	OD1	2.9	w1	2.8	NH1	Arg112	L4
Helix F	Asp168	OD2	2.9	w2	2.8	OG	Ser114	L4

respectively. The backbone nitrogen of Gly61 interacts with IL-22 Thr70 residue, which is also hydrogen bonded to IL-22R1 Lys58, and the Arg112 side chain interacts with Gly165 and contacts Asp168 through a water molecule (w1) (Fig. 2B).

Among these interactions found in the IL-22/IL-22R1 binding interface, there are three salt bridges (Asp71/Lys58, Asp67/Arg112, and Arg73/Glu90), and also four additional contacts, which are also conserved, in the IL-10/IL-10R1 complex (Glu166/Tyr60, Asp71/Gly61, Gly165/Arg112, and Asp168/Arg112 – IL-22/IL-22R1 numbering) (Fig. 2B). However, there are two salt bridges in the IL-10/IL-10R1 complex that are not conserved in the IL-22/IL-22R1 binding interface. The first one involves IL-10 Asp162 and IL-10R1 Arg117 residues, which are equivalent to IL-22 Asp168 and IL-22R1 Arg112, respectively. Our X-ray structure of IL22/IL-22R1 reveals that Asp168 and Arg112 are not oriented to each other and they present a water mediated interaction (w1). Besides, in the IL-10/IL-10R1 heterodimer, Asp62 and Arg97 residues form a salt bridge, but their corresponding residues Thr70 and Tyr93 do not interact with each other in the IL-22/IL-22R1 complex.

At the C-terminal FBN-III domain of IL-22R1 (D2 domain), the backbone oxygen of IL-22R1 Thr207 interacts with both Phe47 and Thr53 residues from IL-22 through a water molecule (w10), and the O γ 1 side chain of Thr207 residue makes a direct contact with N ϵ atom of Arg175 located on the IL-22 helix F. Moreover, the backbone oxygen of Pro206 from IL-22R1 forms interactions with IL-22 Phe47, Gln49, and Thr53 residues mediated by a water molecule (w9) (Fig. 2C). The positions of Phe47 and Arg175 residues in the IL-22 structure alone are different from their positions in the IL-22/IL-22R1 complex, indicating that there is a conformational change in the N-terminal region of IL-22 helix A upon IL-22R1 binding.

3.4. Structure-based site-directed mutagenesis studies reveal functionally important residues at the IL-22R1 interface

To pinpoint the hot spots of interaction between IL-22 and IL-22R1, we mutated key amino acid residues at the interface of complex. The IL-22R1 Tyr60 side chain is hydrogen bonded to both Lys162 and Glu166 residues of IL-22 at the interface (Fig. 3A). These interactions have their counterparts in IL-10/IL-10R1 complex, where IL-10R1 Tyr64 form hydrogen bonds with IL-10 Lys156 and Glu160 side chains. In order to check the importance of Tyr60 residue for IL-22/IL-22R1 binding, IL-22R1 Tyr60 residue was mutated to either an Arg or Ala residue. Both Y60R and Y60A mutants did not affect IL-22R1 cell surface expression, assessed by fluorescent automatic cell sorter analysis (data not shown). In order to assess the functional role of IL-22R1 Tyr60 point mutants, STAT3 dependent luciferase reporter gene assays were performed in HEK-293 cells transiently transfected with wild type or mutated IL-22R1 cDNAs. Wild type IL-22R1 transfections showed clear functional response. By contrast, HEK-293 cells transfected with IL-22R1 Y60R or Y60A mutants completely failed to respond to IL-22, demonstrating that Tyr60 is a critical IL-22R1 residue for IL-22 binding (Fig. 3B).

A similar mutagenesis strategy was used to assess the importance of the interaction between Lys58 of IL-22R1 with IL-22 (Fig. 3A). When Lys58 was mutated into a negatively charged

Fig. 3. The influence of IL-22R1 mutants in the IL-22 binding. (A) Ribbon representation of the IL-22/IL-22R1 interface, showing IL-22R1 Lys58 and Tyr60 residues and their IL-22 contact residues (residues are colored as described in previous figure). (B) STAT activation by IL-22R1 mutants: 10^5 HEK-293 cells were seeded in 48-well plates, and cells were transfected with the indicated receptor mutants, together with the STAT-dependent pGL3-Pap1-Luc and the constitutive pRL-TK reporter plasmids. Cells were stimulated with IL-22 or with control medium for 16 h, before a luciferase assay was performed. The experiments were repeated at least five times in duplicate with similar results.

residue such as Glu, IL-22R1 could not be detected at the cell surface (data not shown), suggesting that K58E point mutation does not allow a normal protein folding, and cell surface expression. However, the mutation of Lys58 to an Ala residue did not affect IL-22R1 cell surface expression, which allowed us to assess its functional role in IL-22 activity. As shown in Fig. 3B, the IL-22R1 K58A mutant was still capable to mediate an IL-22 response, but with a 100-fold drop in specific activity, reflecting its importance in the IL-22/IL-22R1 binding.

3.5. The IL-22/IL-22R1/IL-10R2 ternary complex model

Very recently, three ternary complexes have been published: IL-4/IL-4 receptor α common γ chain (γ_c) (IL-4R α / γ_c), IL-4/interleukin-13 receptor α (IL-13R α_1)/IL-4R α and IL-13/IL-

Fig. 4. The IL-22/IL-22R1/IL-10R2 ternary complex model. Stereo view of the IL-22/IL-22R1 complex (IL-22 is shown in yellow, IL-22R1 is marked in red) as superposed to the IL-4/IL-4R α / γ_c (green, purple, and blue, respectively). The residues identified by Logsdon and co-workers as a “hot spot” for IL-10R2 binding are shown as an orange mesh.

13R α 1/IL-4R α (PDB ids: 3BPL, 3BPN and 3BPO, respectively) [24]. The superposition of IL-22 from our IL-22/IL-22R1 crystal structure onto IL-4 and IL-13-receptor complexes reveals that IL-4/IL-4R α / γ_c is a plausible model for IL-22/IL-22R1/IL-10R2 ternary complex, where γ_c occupies a similar position to IL-22R1 and IL-4R α occupies a similar position to IL-10R2 (Fig. 4).

We speculate that in the putative IL-22/IL-22R1/IL-10R2 complex model, the IL-22R1 and IL-10R2 interact with each other via their D2 domains, and the IL-22 molecule interacts with both receptor chains through distinct binding sites. The IL-22R1 binding site includes IL-22 residues located on helix A, the AB loop and helix F, while the IL-10R2 binding site is formed by helices A and D that contains five hot spot residues for IL-10R2 binding (Tyr51, Asn54, Arg55, Tyr114, and Glu117), which were identified by alanine scanning mutagenesis assays [25].

It has been shown that IL-22 first binds to the IL-22R1 extracellular domain with high affinity, and the IL-10R2 sequentially recognizes and binds to the IL-22/IL-22R1 binary complex [10]. The IL-22/IL-22R1 interactions induce conformational changes in IL-22R1 and/or IL-22 and create a surface binding site for IL-10R2, which is composed by IL-22 hot spot residues and IL-22R1 FBN-III D2 domain (residues 174–190). Functionally, after IL-10R2 binding to the IL-22/IL-22R1 complex, the ternary complex activates the JAK/STAT signaling pathway what leads to diverse biological effects related to IL-22 [7,8].

Although proposed IL-22 ternary complex architecture seems to be consistent with the available structural and functional data, this putative model should be challenged by further structural and functional studies.

3.6. Protein data bank accession code

Atomic coordinates and structure factors of IL22/IL-22R1 complex have been deposited in the Protein Data Bank under accession code 3DLQ.

Acknowledgements: L.B. thanks Martyn Winn and Ronan M. Keegan at the Daresbury Laboratory for the outstanding support for MrBump. This work was supported by FAPESP (Grants #06/00182-

8, #06/01534-5 and #06/60860-0), CNPq (Grants #154559/2006-7 and #473875/2003-9), CAPES, Belgian Federal Service for Scientific, Technical, and Cultural Affairs, and by the *Actions de Recherche Concertées* of the *Communauté Française de Belgique*.

References

- [1] Dumoutier, L., Louahed, J. and Renauld, J.C. (2000) Cloning and characterization of IL-10-related T cell-derived inducible factor (IL-TIF), a novel cytokine structurally related to IL-10 and inducible by IL-9. *J. Immunol.* 164, 1814–1819.
- [2] Xie, M.H., Aggarwal, S., Ho, W.H., Foster, J., Zhang, Z., Stinson, J., Wood, W.I., Goddard, A.D. and Gurney, A.L. (2000) Interleukin (IL)-22, a novel human cytokine that signals through the interferon receptor-related proteins CRF2-4 and IL-22R. *J. Biol. Chem.* 275, 31335–31339.
- [3] Kotenko, S.V. (2002) The family of IL-10-related cytokines and their receptors: related, but to what extent? *Cytokine Growth Factor Rev.* 13, 223–240.
- [4] Dumoutier, L., Van Roost, E., Colau, D. and Renauld, J.C. (2000) Human interleukin-10-related T cell-derived inducible factor: molecular cloning and functional characterization as a hepatocyte-stimulating factor. *Proc. Natl. Acad. Sci. USA* 97, 10144–10149.
- [5] Nagem, R.A.P., Colau, D., Dumoutier, Renauld, J.C., Ogata, C. and Polikarpov, I. (2002) Crystal structure of recombinant human interleukin-22. *Structure* 10, 1051–1062.
- [6] Xu, T., Logsdon, N.J. and Walter, M.R. (2005) Structure of insect-cell-derived IL-22. *Acta Crystallogr. D: Biol. Crystallogr.* 61, 942–950.
- [7] Fickenscher, H., Hör, S., Küpers, H., Knappe, A., Wittman, S. and Sticht, H. (2002) The interleukin-10 family of cytokines. *Trend Immunol.* 23, 89–96.
- [8] Kotenko, S.V., Izotova, L.S., Mirochnitchenko, O.V., Esterova, E., Dickensheets, H., Donnelly, R.P. and Pestka, S. (2001) Identification of the functional interleukin-22 (IL-22) receptor complex: the IL-10R2 chain (IL-10Rbeta) is a common chain of both the IL-10 and IL-22 (IL-10-related T cell-derived inducible factor, IL-TIF) receptor complexes. *J. Biol. Chem.* 276, 2725–2732.
- [9] Dumoutier, L., Leemans, C., Lejeune, D., Kotenko, S.V. and Renauld, J.C. (2001) Cutting edge: STAT activation by IL-19, IL-20 and mda-7 through IL-20 receptor complexes of two types. *J. Immunol.* 167, 3545–3549.
- [10] Li, J., Tomkinson, K.N., Tan, X.-Y., Wu, P., Yan, G., Spaulding, V., Deng, B., Annis-Freeman, B., Heveron, K., Zollner, R., de Zutter, G., Wright, J.F., Crawford, T.K., Liu, W., Jacobs, K.A., Wolfman, N.M., Ling, V., Pittman, D.D., Veldman, G.M. and

- Fouser, S. (2004) Temporal associations between interleukin 22 and the extracellular domains of IL-22R and IL-10R2. *Int. Immunopharmacol.* 4, 693–708.
- [11] Logsdon, N.J., Jones, B.C., Josephson, K., Cook, J. and Walter, M.R. (2002) Comparison of interleukin-22 and interleukin-10 soluble receptor complexes. *J. Interferon Cytokine Res.* 22, 1099–1112.
- [12] Oliveira-Neto, M., Ferreira-Júnior, J.R.S., Colau, D., Fischer, H., Nascimento, A.S., Craievich, A.F., Dumoutier, L., Renauld, J.C. and Polikarpov, I. (2007) Interleukin-22 forms dimers that are recognized by two interleukin-22R1 receptor chains. *Biophys. J.* 94 (5), 1754–1765.
- [13] Pletnev, S., Magracheva, E., Wlodawer, A. and Zdanov, A. (2005) A model of the ternary complex of interleukin-10 with its soluble receptors. *BMC Struct. Biol.* 5, 10.
- [14] Otwinowski, Z. and Minor, W. (1997) Processing of X-ray data collected in oscillation mode. *Method Enzymol.* 276, 307–326.
- [15] Keegan, R.M. and Winn, M.D. (2007) Automated search-model discovery and preparation for structure solution by molecular replacement. *Acta Crystallogr. D* 63, 447–457.
- [16] McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C. and Read, R.J. (2007) Phaser crystallographic software. *J. Appl. Crystallogr.* 40, 658–674.
- [17] Schwarzenbacher, R., Godzik, A., Grzechnik, S.K. and Jaroszewski, L. (2004) The importance of alignment accuracy for molecular replacement. *Acta Crystallogr. D* 60, 1229–1236.
- [18] Josephson, K., Logsdon, N.J. and Walter, M.R. (2001) Crystal structure of the IL-10/IL-10R1 complex reveals a shared receptor binding site. *Immunity* 14, 35–46.
- [19] Huang, M., Syed, R., Stura, E.A., Stone, M.J., Stefanko, R.S., Ruf, W., Edgington, T.S. and Wilson, I.A. (1998) The mechanism of an inhibitory antibody on TF-initiated blood coagulation revealed by the crystal structures of human tissue factor, Fab 5G9 and TF.G9 complex. *J. Mol. Biol.* 275, 873–894.
- [20] Emsley, P. and Cowtan, K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* 60, 2126–2132.
- [21] Adams, P.D., Gopal, K., Grosse-Kunstleve, R.W., Hung, L.-W., Ioerger, T.R., McCoy, A.J., Moriarty, N.W., Pai, R.K., Read, R.J., Romo, T.D., Sacchettini, J.C., Sauter, N.K., Storoni, L.C. and Terwilliger, T.C. (2004) Recent developments in the PHENIX software for automated crystallographic structure determination. *J. Synchrotron Rad.* 11, 53–55.
- [22] Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* 26, 283.
- [23] Dumoutier, L., Lejeune, D., Colau, D. and Renauld, J.C. (2001) Cloning and characterization of IL-22 binding protein, a natural antagonist of IL-10-related T-cell inducible factor/IL-22. *J. Immunol.* 166, 7090–7095.
- [24] LaPorte, S.L., Juo, Z.S., Vaclavikova, J., Colf, L.A., Qi, X., Heller, N.M., Keegan, A.D. and Garcia, K.D. (2008) Molecular and structural basis of cytokine receptor pleiotropy in the interleukin-4/13 system. *Cell* 132, 259–272.
- [25] Logsdon, N.J., Jones, B.C., Allman, J.C., Izotova, L., Schwartz, B., Pestka, S. and Walter, M.R. (2004) The IL-10R2 binding hot spot on IL-22 is located on the N-terminal helix and is dependent on N-linked glycosylation. *J. Mol. Biol.* 342, 503–514.

Crystallization and Preliminary Crystallographic Analysis of Laminarinase from *Rhodothermus marinus*: A Case of Pseudomerohedral Twinning

Alexander M. Golubev^{1,2}, Adriana L. Rojas^{2,†}, Alessandro S. Nascimento², Lucas Bleicher², Anna A. Kulminskaya¹, Elena V. Eneyskaya¹ and Igor Polikarpov^{2,*}

¹Petersburg Nuclear Physics Institute, Gatchina, St. Petersburg, 188300, Russia; ²Instituto de Física de São Carlos, Universidade de São Paulo, Av Trabalhador Saocarlense 400, CEP 13560-970, São Carlos, SP, Brazil

Abstract: Thermophilic endo-1,3(4)- β -glucanase (laminarinase) from *Rhodothermus marinus* was crystallized by the hanging-drop vapor diffusion method. The needle-like crystals belong to space group $P2_1$ and contain two protein molecules in the asymmetric unit with a solvent content of 51.75 %. Diffraction data were collected to a resolution of 1.95 Å and resulted in a dataset with an overall R_{merge} of 10.4% and a completeness of 97.8%. Analysis of the structure factors revealed pseudomerohedral twinning of the crystals with a twin fraction of approximately 42%.

Keywords: Laminarinase, *Rhodothermus marinus*, crystallization, pseudomerohedral twinning.

INTRODUCTION

Rhodothermus marinus is a thermophilic halophilic bacterium found in Icelandic marine hot springs which presents a surprisingly reduced glycoside hydrolase activity [1,2]. However, *R. marinus* produces a number of thermophilic hydrolases, including cellulase [3], xylanase [4] and also two endoglucanases, which were isolated from the bacterium and characterized. The structure of the highly thermostable cellulase Cel12A from *R. marinus* belonging to glycoside hydrolase family (GHF) 12 has already been determined [5]. One of the two endoglucanases, laminarinase 1,3(4)- β -glucanase (LamR) is an enzyme that hydrolyzes β -glycosidic bonds in soluble and insoluble glucans (laminarin, lichenin and curdlan) producing DP3 and DP4 oligosaccharides as main products of the reaction. The laminarinase displays an exceptionally high thermostability, reaching its highest enzymatic activity at 88 °C and pH 5.5 [2]. On the basis of amino acid sequence comparison, the enzyme was classified as a member of GHF 16.

NMR analysis of the enzymatic reaction revealed that the hydrolysis occurs with retention of the anomeric configuration and that LamR catalyzes transglycosylation producing both 1,3- β -glycosidic and 1,4- β glycosidic linkages [6]. Studies of enzymatic activity of mutants confirmed that the acidic residues E129, D131 and E134 play a role in the catalysis [2]. Recently, it was shown that the *R. marinus* laminarinase might be a suitable vehicle for exploring the balance between hydrolysis and transglycosylation. Based on pri-

mary structure analysis, the GHF 16 can be divided in two subgroups. The first group, which includes laminarinases, was shown to possess an active site sequence tag Glu- ϕ -Asp- ϕ -Met-Glu, where the first glutamate acts as the nucleophile, the second acts as the acid/base catalyst and ϕ are hydrophobic residues like isoleucine [7]. The second group does not contain Met residues in the active site [7]. Evolutionary analysis, also based on sequence alignments, showed that *R. marinus* laminarinase is closer to bacterial glucanases of the second group, which prompted us to study the role played by the Met residue in the active site. By mutational analysis, it was found out that, among M133A, M133W and M133C mutants, the latter demonstrated near 20% greater rate of transglycosylation activity in comparison with M133A and M133W [7]. However, for a better comprehension of the enzymatic activities detected for the mutants, the 3-dimensional structure of the protein is needed. Deeper understanding of the structural features which influence the transglycosylation activity of the enzyme may be useful for optimization of hemo-enzymatic applications of glycoside hydrolases.

Here we report the crystallization and preliminary diffraction analysis of the *R. marinus* laminarinase as a first step toward structural characterization of the native protein and its mutants with increased transglycosylation activity.

PROTEIN PREPARATION AND PURIFICATION

Recombinant plasmids pMKM1 [2] derived from the wild type strain of *R. marinus* were used for transformation of *Escherichia coli*. The culture of *E. coli* DH5 α used as a host strain for the plasmids was grown in Luria/Bertani (LB) medium with addition of 150 mg/mL of ampicillin at 37 °C for 7 h with continuous shaking. Recombinant laminarinase was purified by heat treatment, ion-exchange chromatography on a Toyopearl SP-650M column and gel-permeation

*Address correspondence to this author at the Instituto de Física de São Carlos, Universidade de São Paulo, Av Trabalhador Saocarlense 400, CEP 13560-970, São Carlos, SP, Brazil; Tel: +55-16-3373-8088; Fax: +55-16-3373-9881; E-mail: ipolikarpov@ifsc.usp.br

[†]Current Address: Laboratory of Molecular Biology, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, US Department of Health and Human Services, Bethesda, MD 20892, USA

chromatography on a SephadexG-50F column as described [8] and lyophilized.

The crystals were obtained at 293 K using the hanging-drop vapor diffusion method. In the first crystallization trials, the lyophilized protein was dissolved in a 5 mM potassium phosphate buffer, pH 6.0 at a concentration of 8 mg/ml. A crystallization solution contained PEG3350 in a 10 mM phosphate buffer solution, pH 5.0. Different concentrations of PEG (8-32%) and protein (2.0-6.4 mg/ml) were screened by varying the volumes mixed in the crystallization drop. A second crystallization screening assay, based on the results of this first screen, was performed by varying the potassium phosphate buffer concentration and pH and fixing the PEG concentration in 22%. The best crystals grown in the drops containing 6% PEG 3350 and 120 mM potassium buffer, pH 6.0, using protein at 8 mg/ml, which were equilibrated against 25% PEG 3350, pH 6.0, 50 mM potassium phosphate buffer as a well solution (Fig. (1A)). Additives were added to this condition in an attempt to increase the crystal size and better crystals appeared using 2% ethanol supplemented to the well solution. The crystals appeared in one week and completed their growth in about 2 weeks (Fig. (1B)). The crystal shown in Fig. (1B) was isolated, transferred to a solution containing 20%(v/v) glycerol as a cryoprotectant, picked up in a nylon loop after 20s and then flash-cooled in a nitrogen-gas stream at 100 K for X-ray data collection.

X-RAY CRYSTALLOGRAPHIC STUDIES

X-ray diffraction data were collected from a single crystal at the protein crystallography beamline (MX-1) of the Brazilian National Synchrotron Light Laboratory (LNLS), Campinas, Brazil [9]. The wavelength of the synchrotron radiation was set 1.42 Å to optimize the flux and the diffraction efficiency, given the synchrotron ring characteristics [10, 11]. The distance between the crystal and the detector was 80 mm. A total of 225 frames were recorded with an oscillation of 1 degree per image and 20 min exposure time (Fig. (2)). The recorded dataset was indexed and integrated using the program MOSFLM [12] and then scaled with SCALA [13]. The crystals belong to monoclinic $P2_1$ space group with the unit cell parameters $a = 52.22$ Å, $b = 108.29$ Å, $c = 64.59$ Å, $\beta = 113.90^\circ$. The Matthews's coefficient [14] of 2.55 Å³/Da⁻¹ indicated a dimer in the asymmetric unit with a solvent content of 51.75 %.

Analysis of the diffraction data quality was carried out using the program XTRIAGE available in the PHENIX suite [15]. The twin analysis indicated the presence of pseudomerohedral twinning, with a twin law $-h, -k, h+l$. The H test [16] computed using the same program indicated a twin fraction close to 42% for both mean $|H|$ (41.6%) and for cumulative distribution of H (41.8%). The Britton test [17] also estimated the twin fraction to be 41.9% with a correlation of 0.9963. Finally, the third test, based on $|L|$ statistics ($|L| = |I_1 - I_2| / |I_1 + I_2|$ [18]), using a maximum likelihood approach estimated the twin fraction to be 37.0% [Zwart PH, unpublished results, 19].

The observed phenomenon is a rare case of twinning in the monoclinic system, in which twinning usually occurs when a is close to c , or when β is close to 90° . However, another fortuitous cell metric which can lead to twinning is

Figure 1. A. LamR crystals grown in the drop containing 6% PEG 3350 and 90 mM potassium phosphate buffer at pH 6.5. Well Solution contained 25% PEG 3350 and 50 mM sodium phosphate, pH 6.0; B. LamR crystals used for data collection. In the drop: 6% PEG 3350 and 120 mM potassium phosphate buffer, pH 6.0. Well solution: 25% PEG 3350, 50 mM potassium phosphate, pH 6.0 and 2% ethanol.

Figure 2. Diffraction image recorded from a lamR crystal on the MX-1 beamline at LNLS. The maximum resolution edge of the image is 1.87 Å.

Table 1. X-Ray Data Collection Statistics

Wavelength	1.42 Å
Resolution range	30.20-1.95 (2.06-1.95) Å
Space Group	P2 ₁
Unit cell parameters	a=52.22Å b=108.29Å c=64.59Å β=113.90°
Completeness	97.8 % (94.9 %)
Multiplicity	4.0 (4.1)
R _{merge} [*]	10.4 % (37.2 %)
I/σI	5.9 (1.9)
Total reflections	188058
Unique reflections	46598
R _{pim} [†]	5.8 % (20.3 %)

^{*} $R_{\text{merge}} = \frac{\sum_{\text{hkl}} \sum_i |I_i(\text{hkl}) - \langle I(\text{hkl}) \rangle|}{\sum_{\text{hkl}} \sum_i I_i(\text{hkl})}$.

[†] $R_{\text{pim}} = \frac{\sum_{\text{hkl}} [1/(N-1)]^{1/2} \sum_i |I_i(\text{hkl}) - \langle I(\text{hkl}) \rangle|}{\sum_{\text{hkl}} \sum_i I_i(\text{hkl})}$.

when $c \cos \beta = -a/2$. This particular case of twinning has previously been observed by Declercq & Evrard [20] and Rudolph *et al.* [21], who also performed a search in the PDB for cases where this phenomenon could have been present and was overlooked, and found several other cases with strong indications of twinning.

This specific case of monoclinic cell parameters can lead to mimicking of orthorhombic metrics and confuse the data processing software (in our case, it was also possible to index data in C222). The twin law, $-h, -k, h+l$, corresponds to a two-fold rotation perpendicular to axes a and b (i.e., along c^*). Rudolph *et al.* [21] noted that, in their case, the NCS axis was parallel to the twin operation, an indication that the NCS rotation would promote twinning. In our case, however, the two molecules in the asymmetric unit are related to a 2-fold NCS axis parallel to a .

Despite the clear twinning, the crystal structure was solved by molecular replacement using the program MOLREP [22] with the structure of endo-β-1,3-glucanase from alkaliphilic *Nocardiopsis* sp. (PDB entry 2HYK) as the search model, which shares 44% identity and 57% similarity, according to BLAST parameters. Two clear peaks were found in the rotation function with 8.26 σ and 8.22 σ, in contrast to 4.30 σ for the third peak. After translation of the first monomer a score of 0.368 with an R_{fact} of 54.7% was achieved. Fixing this monomer, a second solution was found resulting in a final score of 0.418 and an R_{fact} of 53.1%. The model containing two molecules in the asymmetric unit was used in a simulated annealing protocol with a start temperature of 5,000K followed by a slow cooling, as implemented in PHENIX [15]. The model had an R_{fact} of 36.5% and an

R_{free} of 39.3%. Further structural refinement cycles lead to a model with an R_{fact} of 23.4% and an R_{free} of 28.1%. Final structural refinement steps are currently in progress.

ACKNOWLEDGEMENTS

This work was supported by grants from the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) via projects #04/08070-9 and #06/00182-8 and the Program for Basic Research in Molecular and Cell Biology from the Presidium of Russian Academy of Sciences.

ABBREVIATIONS

LamR = *Rhodothermus marinus* laminarinase (1,3(4)-β-glucanase)

GHF = Glycoside hydrolase family

REFERENCES

- [1] Alfredsson, G.A., Kristjansson, J.K., Hjorleifsdottir, S. and Stetter, K.O. (1988) *J. Gen. Microbiol.*, 134, 299.
- [2] Krah, M., Misselwitz, R., Politz, O., Thomsen, K.K., Welfle, H. and Borriß, R. (1998) *Eur. J. Biochem.*, 257, 101.
- [3] Halldorsdottir, S., Thorolfsson, E.T., Spilliaert, R., Johansson, M., Thorbjarnardottir, S.H., Palsdottir, A., Hreggvidsson, G.O., Kristjansson, J.K., Holst, O. and Eggertsson, G. (1998) *Appl. Microbiol. Biotechnol.*, 49, 277.
- [4] Karlsson, E.N., Holst, O. and Tocaj, A. (1999) *J. Biosci. Bioeng.*, 87, 598.
- [5] Crennell, S.J., Hreggvidsson, G.O. and Nordberg, K.E. (2002) *J. Mol. Biol.*, 19, 883.
- [6] Petersen, B.O., Krah, M., Duus, J. and Thomsen, K.K. (2000) *Eur. J. Biochem.*, 267, 361.
- [7] Neustroev, K.N., Golubev, A.M., Sinnott, M.L., Borriß, R., Krah, M., Brumer, H. 3rd., Eneyskaya, E.V., Shishlyannikov, S.M., Shabalin, K.A., Pesheonov, V.T., Korolev, V.G. and Kulminskaya, A.A. (2006) *Glycoconj. J.*, 23, 501.
- [8] Borriß, R., Krah, M., Brumer, H.3rd., Kerzhner, M.A., Ivanen, D.R., Eneyskaya, E.V., Elyakova, L.A., Shishlyannikov, S.M., Shabalin, K.A. and Neustroev, K.N. (2003) *Carbohydr. Res.*, 338, 1455.
- [9] Polikarpov, I., Perles, L.A., Oliveira, R.T., Oliva, G., Castellano, E.E., Garratt, R.C. and Craievich, A. (1998) *J. Synchrotron Rad.*, 5, 72.
- [10] Polikarpov, I., Teplyakov, A. and Oliva, G. (1997) *Acta Cryst.*, D53, 734.
- [11] Teplyakov, A., Oliva, G. and Polikarpov, I. (1998) *Acta Cryst.*, D54, 610.
- [12] Leslie, A.G.W. (1999) *Acta Cryst.*, D55, 1696.
- [13] Collaborative Computational Project, Number 4 (1994) *Acta Cryst.*, D50, 760.
- [14] Matthews, B.W. (1968) *J. Mol. Biol.*, 33, 491.
- [15] Adams, P.D., Grosse-Kunstleve, R.W., Hung, L.-W., Ioerger, T.R., McCoy, A.J., Moriarty, N.W., Read, R.J., Sacchettini, J.C., Sauter, N.K. and Terwilliger, T.C. (2002) *Acta Cryst.*, D58, 1948.
- [16] Yeates, T.O. (1988) *Acta Cryst.*, A44, 142.
- [17] Fisher, R.G. and Sweet, R.M. (1980) *Acta Cryst.*, A36, 755.
- [18] Padilla, J.E. and Yeates, T.O. (2003) *Acta Cryst.*, D59, 1124.
- [19] Zwart, P.H., Grosse-Kunstleve, R.W. and Adams, P.D. (2005) *CCP4 newsletter*, 42, 10.
- [20] Declercq, J. P. and Evrard, C. (2001) *Acta Cryst.*, D57, 1829.
- [21] Rudolph, M. G., Wingren, C., Crowley, M. P., Chien, Y.-H. and Wilson, I.A. (2004) *Acta Cryst.*, D60, 656.
- [22] Vagin, A. and Teplyakov, A. (1997) *J. Appl. Cryst.*, 30, 1022.