UNIVERSIDADE ESTADUAL DE CAMPINAS – UNICAMP
Instituto de Física Gleb Wataghin – IFGW

TESE DE DOUTORADO

Otimização das condições para a aquisição de dados de derivados e para a determinação das fases dos fatores de estrutura de cristais de proteínas por meio da difração da luz síncrotron

Doutorando: Ronaldo Alves Pinto Nagem
Orientador: Prof. Dr. Igor Polikarpov
Co-orientador: Prof. Dr. José Antônio Brum

Setembro 2003
Campinas – SP

Ronaldo Alves Pinto Nagem

Otimização das condições para a aquisição de dados de derivados e para a determinação das fases dos fatores de estrutura de cristais de proteínas por meio da difração da luz síncrotron

Tese apresentada ao Curso de Doutorado do Instituto de Física Gleb Wataghin da Universidade Estadual de Campinas como requisito parcial à obtenção do título de Doutor em Física.

Área de concentração: Cristalografia de Proteínas

Orientador: Prof. Dr. Igor Polikarpov
Instituto de Física de São Carlos – USP

Co-orientador: Prof. Dr. José Antônio Brum
Instituto de Física Gleb Wataghin – UNICAMP

Campinas – SP
Instituto de Física Gleb Wataghin da UNICAMP
2003

Ficha Catalográfica elaborada pela Biblioteca do IFGW
Universidade Estadual de Campinas - UNICAMP

N133o

Nagem, Ronaldo Alves Pinto
Otimização das condições para a aquisição de dados de derivados e para a determinação das fases dos fatores de estrutura de cristais de proteínas por meio da difração da luz síncrotron / Ronaldo Alves Pinto Nagem. -- Campinas, SP : [s.n.], 2003.

Orientadores: Igor Polikarpov, José Antônio Brum.
Tese (doutorado) - Universidade Estadual de Campinas, Instituto de Física "Gleb Wataghin".

1. Cristalografia de raio-x. 2. Proteínas. 3. Radiação sincrotrônica. 4. Espalhamento. I. Polikarpov, Igor. II. Brum, José Antônio. III. Universidade Estadual de Campinas. Instituto de Física "Gleb Wataghin". IV. Título.

UNICAMP PÓS GRADUAÇÃO IFGW Instituto de Física Gleb Wataghin
C.P. 6165
CEP: 13083-970
Tel. (19) 3788-5305
e-mail: secpos@ifi.unicamp.br

MEMBROS DA COMISSÃO JULGADORA DA TESE DE DOUTORADO DE **RONALDO ALVES PINTO NAGEM – RA 991278** APRESENTADA E APROVADA AO INSTITUTO DE FÍSICA "GLEB WATAGHIN", DA UNIVERSIDADE ESTADUAL DE CAMPINAS, EM 09 / 09 / 2003.

COMISSÃO JULGADORA:

Prof. Dr. Igor Polikarpov (Orientador do Candidato) – GC/IFSC/USP

Prof. Dr. Raghuvir Krishnaswamy Arni – DF/IBILCE/UNESP

Prof. Dr. Glaucius Oliva – CEPID/IFSC/USP

Prof. Dr. Antonio Rubens Britto de Castro - DFMC/IFGW/UNICAMP

Profa. Dra. Iris Concepción Linares de Torriani – DFMC/IFGW/UNICAMP

Dedicatória

Cinqüenta anos se passaram desde a descoberta da estrutura em fita dupla do DNA em 1953, por James Dewey Watson e Francis Harry Compton Crick, até o anúncio da elucidação completa do genoma humano em 14 de abril de 2003. Esta conquista, realizada por vários grupos de pesquisa de todo o mundo, é prova inequívoca de que o esforço individual de muitos seres humanos, engajados em um objetivo comum, pode produzir resultados maravilhosos. Este trabalho é dedicado a essas pessoas que estão permitindo um porvir extraordinário para a Biologia Molecular e Estrutural.

Agradecimentos

Devido à impossível tarefa de mencionar, em tão pouco espaço, todas as pessoas que contribuíram, desde muito cedo, direta ou indiretamente, para a minha formação e para a realização deste trabalho, agradeço a Míriam Pompeu Nagem, Maria Helena Alves Pinto Nagem e Ronaldo Luiz Nagem, em nome de minha família e amigos. Agradeço também a Neide Abreu Barbosa (i*n memoriam*), Igor Polikarpov e Zbigniew Dauter, em nome de meus professores e supervisores, e aos integrantes do grupo de Cristalografia de Proteínas do Laboratório Nacional de Luz Síncrotron, do qual fiz parte de 1999 a 2003. Expresso ainda meu agradecimento especial a Ruth Léa Nagem pela revisão ortográfica e gramatical desta tese.

Por fim, deixo registrado meu enorme apreço e minha consideração às seguintes entidades e instituições de ensino e pesquisa:

Colégio Logosófico González Pecotche,
Universidade Federal de Minas Gerais (UFMG),
Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq),
Instituto Tecnológico de Aeronáutica (ITA),
Laboratório Nacional de Luz Síncrotron (LNLS),
Universidade Estadual de Campinas (UNICAMP),
Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP),
Brookhaven National Laboratory (BNL),
European Molecular Biology Laboratory (EMBL) e
Instituto de Tecnologia Química e Biológica (ITQB).

Sumário

Lista de figuras

Lista de tabelas

Resumo

Atualmente, os modelos tridimensionais de macromoléculas biológicas obtidos por meio das técnicas de difração de raios-X (cristalografia) ou de ressonância magnética nuclear constituem uma das bases principais para o desenho racional de fármacos. Nesse contexto, a área de cristalografia de proteínas é responsável por cerca de 85% de todos os modelos estruturais de macromoléculas biológicas encontrados nos bancos de dados. Este fato, por si só, tem permitido um grande avanço da área tanto em instituições de pesquisa científica quanto em empresas farmacêuticas. Desde que a linha de luz CPr do Laboratório Nacional de Luz Síncrotron (LNLS, Campinas, SP, Brasil) foi construída e aberta à comunidade científica em 1997, centenas de conjuntos de dados de difração de raios-X por cristais de macromoléculas biológicas têm sido coletados. Grande parte desses conjuntos foi usada para resolver a estrutura de proteínas pela técnica de substituição molecular; técnica esta que utiliza um modelo tridimensional já conhecido semelhante àquele que se pretende determinar. Contudo, até 1999, quando o projeto desta tese foi escrito, não mais do que duas estruturas cristalográficas inéditas haviam sido resolvidas por grupos de pesquisa no Brasil. Em vista disso, o projeto teve como objetivo principal otimizar as condições necessárias para a resolução de estruturas inéditas de proteínas usando as facilidades da linha CPr do LNLS. O processo envolveu não só a preparação das amostras para a coleta de dados, mas também o processamento e a análise dos resultados. Durante os quatro anos desta tese, inúmeros conjuntos de dados foram coletados na linha CPr. Os primeiros experimentos utilizando uma proteína teste (lisozima) foram essenciais para avaliar tanto o método de crio derivatização rápida para a preparação das amostras quanto a qualidade dos conjuntos de dados coletados. Os experimentos iniciais possibilitaram aumentar ainda mais as potencialidades do método de crio derivatização rápida e ajustá-lo às condições da linha de luz CPr. A experiência adquirida com este trabalho inicial foi aplicada, em seguida, na resolução de algumas estruturas inéditas de proteínas de organismos variados. Os resultados permitiram concluir que a aplicação do método de crio derivatização rápida para a preparação dos cristais e o uso de radiação monocromática (1,54 Å), proveniente do síncrotron brasileiro, podem aumentar, significativamente, as chances de sucesso para a determinação de estruturas inéditas de macromoléculas biológicas no país. Apesar de todo o trabalho ter sido realizado com uma fonte de radiação síncrotron (LNLS), os resultados obtidos no final desta pesquisa permitem adiantar e prever o uso de geradores convencionais de raios-X junto com o método de crio derivatização rápida para a resolução de estruturas inéditas em laboratórios de pequeno porte. Isso poderá proporcionar, sem sombra de dúvida, um grande avanço nos projetos pós-genoma em desenvolvimento no país.

Palavras-chave: Cristalografia de Proteínas, Radiação Síncrotron, Método de Crio Derivatização Rápida e Estruturas Macromoleculares Inéditas.

Abstract

Nowadays, three-dimensional models of macromolecules obtained by X-ray diffraction (crystallography) and nuclear magnetic resonance techniques are one of the main bases for rational drug design. The protein crystallography field is responsible for almost 85% of all macromolecule models found in databases. This fact, by itself, is responsible for great progress of the field in research institutions and pharmaceutical industries. From the time when the CPr beamline of the Brazilian National Synchrotron Light Source (LNLS, Campinas, SP, Brazil) was built and its use by the scientific community was authorized in 1997, hundreds of X-ray diffraction datasets have been collected. Most of the datasets collected at the CPr beamline were used to solve the structure of proteins by the molecular replacement method. This method involves the use of a known search model, closely similar to the investigated macromolecule. However, until 1999, when the project for this thesis was written, no more than two novel crystallographic structures had been solved in Brazil. As a result of this, the project had the main goal to optimize the conditions required to the solution of novel macromolecule structures using the CPr beamline at the LNLS. This project involved not only preparation of samples for data acquisition, but also processing and analysis of results. During four years, several X-ray diffraction datasets were collected at the CPr beamline. First experiments using a test protein (lysozyme) were essential to check the quick cryo soaking approach for sample preparation as well as the quality of datasets. These initial experiments allowed to increase the applicability of the quick cryo soaking approach and to adjust it to the CPr beamline. The experience acquired as a result of this initial work was used to solve the structure of a few novel proteins from different organisms. The main conclusion achieved after this work is that the use of the quick cryo soaking approach to prepare crystals for data collection and the use of monochromatic radiation (1,54 Å) from the Brazilian synchrotron can noticeably increase the chances of success in the determination of novel protein structures in our country. Even thought only synchrotron radiation was used in this work, results obtained at the end of this research indicate and forecast the use of conventional X-ray sources with the quick cryo soaking approach to solve the structure of novel proteins in small laboratories. This may promote, with no doubt, great advances in pos-genomic projects under development in our country.

Key words: Protein Crystallography, Synchrotron Radiation, Quick Cryo Soaking Approach and Novel Macromolecule Structures.

1 Apresentação

Esta tese foi desenvolvida ao longo dos últimos quatro anos (1999-2003), durante o curso de Doutorado do Instituto de Física Gleb Wataghin (IFGW) da Universidade Estadual de Campinas (UNICAMP). O trabalho experimental foi realizado, em sua maior parte, no grupo de Cristalografia de Proteínas (CPr) do Laboratório Nacional de Luz Síncrotron (LNLS). Porém, alguns experimentos foram realizados nos Estados Unidos, no *National Synchrotron Light Source* (NSLS). A formação acadêmica, científica e intelectual do aluno foi complementada com a participação em 10 congressos nacionais e internacionais realizados no país e 2 cursos internacionais realizados pelo *European Molecular Biology Laboratory* (EMBL) no *European Synchrotron Radiation Facility* (ESRF) em Grenoble (França) e pelo Instituto de Tecnologia Química e Biológica (ITQB) em Oeiras (Portugal). Além disso, o aluno realizou um doutorado sanduíche, com duração de seis meses, no grupo do pesquisador Dr. Zbigniew Dauter no *Brookhaven National Laboratory* em Upton (Estados Unidos).

O conteúdo desta tese abrange tanto aspectos relacionados à Física quanto à Biologia e, por isso, pode-se classificá-la como pertencente a área de Biofísica. Uma classificação mais específica permite, ainda, colocá-la dentro da subárea conhecida por Cristalografia de Proteínas, já que seu conteúdo aborda a aplicação de uma técnica física de medida (difração de raios-X por cristais) para estudar a estrutura tridimensional de macromoléculas biológicas como as proteínas.

Este projeto teve o apoio financeiro da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) e está cadastrado sob o número 98/06218-6.

O autor deste trabalho, Ronaldo Alves Pinto Nagem (rapnagem@hotmail.com), nasceu em Belo Horizonte (MG), no ano de 1975. Em 1993, ainda em sua cidade natal, completou o segundo grau no Colégio Logosófico González Pecotche. No ano seguinte, iniciou um curso universitário no Instituto Tecnológico de Aeronáutica (ITA) em São José dos Campos (SP), mas, por motivos pessoais, retornou a Belo Horizonte, no mesmo ano, para fazer o curso de Física pela Universidade Federal de Minas Gerais (UFMG). Seu primeiro contato com a área de Cristalografia de Proteínas foi em 1997, quando foi selecionado para uma bolsa de verão no Laboratório Nacional de Luz Síncrotron. Desde novembro de 1998, Ronaldo Alves Pinto Nagem é bacharel em Física pela Universidade Federal de Minas Gerais.

2 Introdução

Em 1912, Max von Laue interpretou, pela primeira vez, os resultados de um experimento de difração de raios-X por um cristal (FRIEDRICH, KNIPPING & LAUE, 1912; GIACOVAZZO, 1992). O estudo marcou o início da cristalografia de raios-X e permitiu com que os princípios físicos que governam esses experimentos pudessem ser determinados. Normalmente, em um experimento típico de difração, um feixe contínuo de raios-X é incidido sobre um cristal e, devido à periodicidade do cristal e à interação dos raios-X com a matéria, principalmente com os elétrons, os raios incidentes são difratados em várias direções. A análise das intensidades relativas dos raios difratados permite obter informações estruturais sobre as moléculas constituintes do cristal. Em 1913, a técnica foi utilizada por William Lawrence Bragg e von Laue para resolver a estrutura cristalográfica de pequenas moléculas inorgânicas como o cloreto de sódio, iodeto de potássio, entre outras (BRAGG, 1913).

As proteínas, ao contrário das moléculas inorgânicas, são extremamente maiores podendo ter milhares de átomos covalentemente ligados. Essas macromoléculas biológicas, presentes aos milhares em todas as células, possuem também as mais diversas formas e funções. Essa diversidade de funções só é possível porque cada macromolécula apresenta uma forma tridimensional específica que lhe confere uma atividade biológica também específica. Apesar de os cristais de proteína serem conhecidos desde o início do século XX, foi apenas em meados de 1960 que Max Ferdinand Perutz,

John Cowdery Kendrew e outros pesquisadores determinaram as primeiras estruturas cristalográficas de macromoléculas biológicas (PERUTZ *et al*., 1960; KENDREW *et al*., 1960; BLAKE *et al*., 2001). Desde então, grandes avanços metodológicos e tecnológicos foram e têm sido observados nesta área.

As informações estruturais obtidas a partir do modelo tridimensional de uma macromolécula biológica constituem uma das bases para o desenho racional de fármacos, pois permitem, juntamente com os estudos bioquímicos, determinar com precisão os mecanismos de reação das moléculas bem como localizar suas regiões funcionais. Esses fatos têm permitido um grande avanço da área tanto em instituições de pesquisa científica como em empresas farmacêuticas.

Atualmente, a área de Cristalografia de Proteínas é responsável por cerca de 85% das estruturas macromoleculares encontradas no *Protein Data Bank* (PDB; BERMAN *et al*., 2000). O banco de dados (http://www.rcsb.org/pdb/) contém as coordenadas espaciais de diversas estruturas macromoleculares resolvidas até o momento, tanto por difração de raios-X quanto por ressonância magnética nuclear. Como se pode observar pela figura 2.1, o número de estruturas resolvidas anualmente vem aumentando rapidamente. No início de 2003, o banco de dados armazenava cerca de 20.000 modelos estruturais de macromoléculas biológicas de vários organismos. Apesar da grande quantidade de estruturas resolvidas até essa data, sabe-se que o valor corresponde a uma fração ainda pequena do total de proteínas encontradas nos seres vivos. Assim, grandes investimentos financeiros e esforços humanos têm sido feitos para permitir um rápido crescimento do número de estruturas macromoleculares resolvidas anualmente em todo o mundo. Exemplos desses esforços são vistos nos vários projetos "Genoma Estrutural" desenvolvidos em alguns países (STEVENS, YOKOYAMA & WILSON, 2001).

Até onde se sabe, a contribuição brasileira para o PDB é ainda modesta, uma vez que o número de grupos de pesquisa em cristalografia ou ressonância magnética nuclear dedicados ao estudo de macromoléculas biológicas é pequeno, quando comparado com outras áreas da Ciência. Entretanto, alguns investimentos, em âmbito nacional, têm permitido o crescimento dessas áreas de pesquisa. Um dos exemplos mais marcantes e que foi vital para o desenvolvimento deste trabalho é a construção de uma estação experimental inteiramente dedicada à Cristalografia de Proteínas no Laboratório Nacional de Luz Síncrotron. A estação, mais conhecida como linha CPr (POLIKARPOV *et al*., 1998a, 1998b), foi liberada em 1997 para a comunidade científica que passou a dispor de uma fonte intensa de raios-X para experimentos de difração por cristais de proteína.

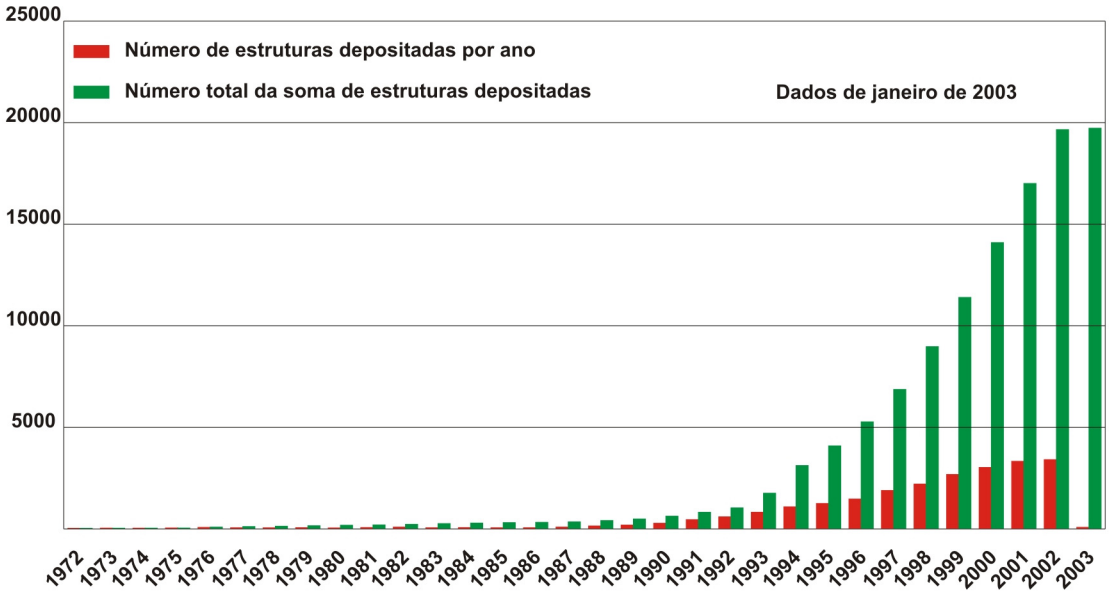


Figura 2-1. Evolução do número de estruturas macromoleculares resolvidas ao longo dos anos utilizando as técnicas de cristalografia de proteínas e ressonância magnética nuclear. [Fonte: *Protein Data Bank*, http://www.rcsb.org/pdb/].

Quando o projeto desta tese foi escrito, no início de 1999, vários conjuntos de dados já haviam sido coletados na linha CPr do LNLS. Entretanto, quase todos foram utilizados para resolver estruturas de proteínas que já apresentavam alguma semelhança com proteínas previamente depositadas no PDB. Essa situação não contribuía, portanto, para a determinação de estruturas ainda "desconhecidas" pela comunidade científica. Era necessário, então, verificar a viabilidade da resolução de estruturas inéditas utilizando as facilidades da linha de luz CPr do LNLS.

Foi estabelecido, portanto, como objetivo principal para este projeto, adquirir os conhecimentos técnicos e científicos necessários para a resolução de estruturas de proteínas inéditas, além de otimizar as condições necessárias para a execução da tarefa na linha de luz CPr.

Além do que já foi exposto, uma explicação a respeito do título desta tese (*Otimização das condições para a aquisição de dados de derivados e para a determinação das fases dos fatores de estrutura de cristais de proteínas por meio da difração da luz síncrotron*) merece ainda ser dada. Inicialmente, deve-se ter em conta que o termo "luz síncrotron" é usado para designar toda radiação eletromagnética proveniente de um anel de armazenamento de partículas carregadas, tradicionalmente, elétrons relativísticos. Neste caso específico, a radiação de interesse se concentra na faixa dos raios-X e, em vista disso, a substituição do termo "luz síncrotron" por "raios-X" poderia ser feita.

A obtenção da estrutura tridimensional de um cristal de proteína por meio da técnica de difração de raios-X requer, inevitavelmente, a determinação das fases dos fatores de estrutura do

cristal. No caso de cristais de proteínas inéditas, sem homologia com nenhuma outra proteína de estrutura conhecida, as fases são obtidas, normalmente, com a coleta dos dados de difração de um cristal devidamente preparado chamado “derivado”. Assim, pode-se dizer que este trabalho foi desenvolvido sob a perspectiva da otimização das condições necessárias para a resolução da estrutura de cristais de proteínas inéditas por meio da técnica de difração de raios-X utilizando radiação proveniente de uma fonte síncrotron.

É importante ressaltar que o termo “estruturas de proteínas inéditas”, utilizado ao longo desta tese, refere-se àquelas estruturas cristalográficas de proteínas que não podem ser resolvidas pelo método de substituição molecular a partir de um modelo tridimensional já conhecido.

Para uma melhor compreensão da pesquisa desenvolvida, esta tese foi dividida em duas partes principais. Na primeira, capítulos 3 a 5, são abordados alguns pontos teóricos. Na segunda, principalmente o sexto capítulo, são apresentados os resultados alcançados neste projeto. Alguns dos trabalhos já publicados ou mesmo aqueles já concluídos e em processo de submissão para revistas internacionais, relacionados diretamente com a pesquisa sugerida no projeto de tese, serão incorporados ao texto permitindo uma visão mais completa do assunto.

3 Peptídios e proteínas

Neste capítulo será apresentado um resumo sobre o tema peptídios e proteínas salientando os aspectos essenciais para a compreensão do trabalho experimental desenvolvido nesta tese. Não é objetivo, portanto, mostrar todos os detalhes sobre o tema uma vez que o assunto é amplamente descrito em inúmeros livros de bioquímica. A introdução começará com a descrição das unidades básicas dos peptídios e das proteínas: os aminoácidos. Será apresentado, também, como os peptídios e as proteínas são formados pelos aminoácidos e como estes são responsáveis pela formação das estruturas primária, secundária, terciária e quaternária. Para aqueles que possuem os conhecimentos básicos sobre o tema, a leitura deste capítulo pode ser omitida. Já para aqueles que desejam saber mais sobre o assunto, os livros *Lehninger Principles of Biochemistry* (NELSON & COX, 2000) e *Introduction to Protein Structure* (BRANDEN & TOOZE, 1991), utilizados para a elaboração deste capítulo, abordam outros aspectos importantes que não são tratados aqui.

3.1 Aminoácidos

As proteínas e os peptídios são macromoléculas biológicas compostas por unidades fundamentais chamadas aminoácidos. O segredo da estrutura e da função das macromoléculas está na maneira como os aminoácidos interagem e se dispõem no espaço.

De uma forma geral, as proteínas são constituídas a partir de um mesmo conjunto de 20 aminoácidos, unidos covalentemente em seqüências características. O notável é que as células podem organizar os 20 aminoácidos em variadas combinações e seqüências para obter peptídios e proteínas com propriedades e atividades diferentes. Entretanto, todos os 20 aminoácidos possuem uma estrutura geral conforme mostra a figura 3.1.

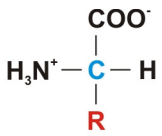


Figura 3-1. Estrutura geral de um aminoácido. Os grupos amina e carboxila aparecem ionizados, como ocorre em pH 7.

Pode-se ver que existem quatro grupos ligados covalentemente ao carbono central (Cα em azul): um grupo carboxila, um grupo amina, um hidrogênio e um grupo R. O grupo R ou cadeia lateral varia em estrutura, tamanho, carga elétrica e solubilidade em água e é, portanto, a parte responsável pela diferenciação entre os 20 aminoácidos.

Uma característica interessante de um aminoácido se refere à sua quiralidade[1]. Sabe-se que de todos os 20 aminoácidos apenas um não possui um centro quiral (Cα). Um centro quiral pode existir tanto na forma L quanto na forma D, conforme mostrado na figura 3.2. Entretanto, os aminoácidos presentes em proteínas apresentam apenas a forma L.

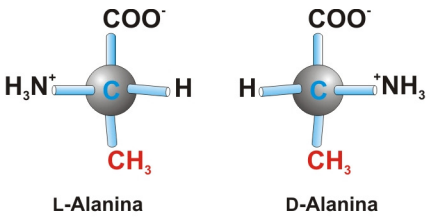


Figura 3-2. Exemplos da forma L e D em aminoácidos. No caso da alanina, a cadeia lateral é formada pelo grupo CH_3. A forma L apresenta, no sentido horário a partir da ligação Cα-H, os grupos $^{+}NH_3$, COO^- e R enquanto a forma D apresenta, no mesmo sentido, os grupos R, COO^- e $^{+}NH_3$.

A figura 3.3 mostra as estruturas dos 20 aminoácidos encontrados normalmente em proteínas e peptídios, seus nomes e abreviações e a classificação deles quanto à polaridade do grupo R.

[1] Uma molécula é dita quiral quando sua imagem espelhada não pode ser superposta à sua imagem real. Para maiores detalhes aconselha-se a leitura do capítulo 7 do livro "*Fundamentals of Crystallography*" (GIACOVAZZO *et al*., 1992).

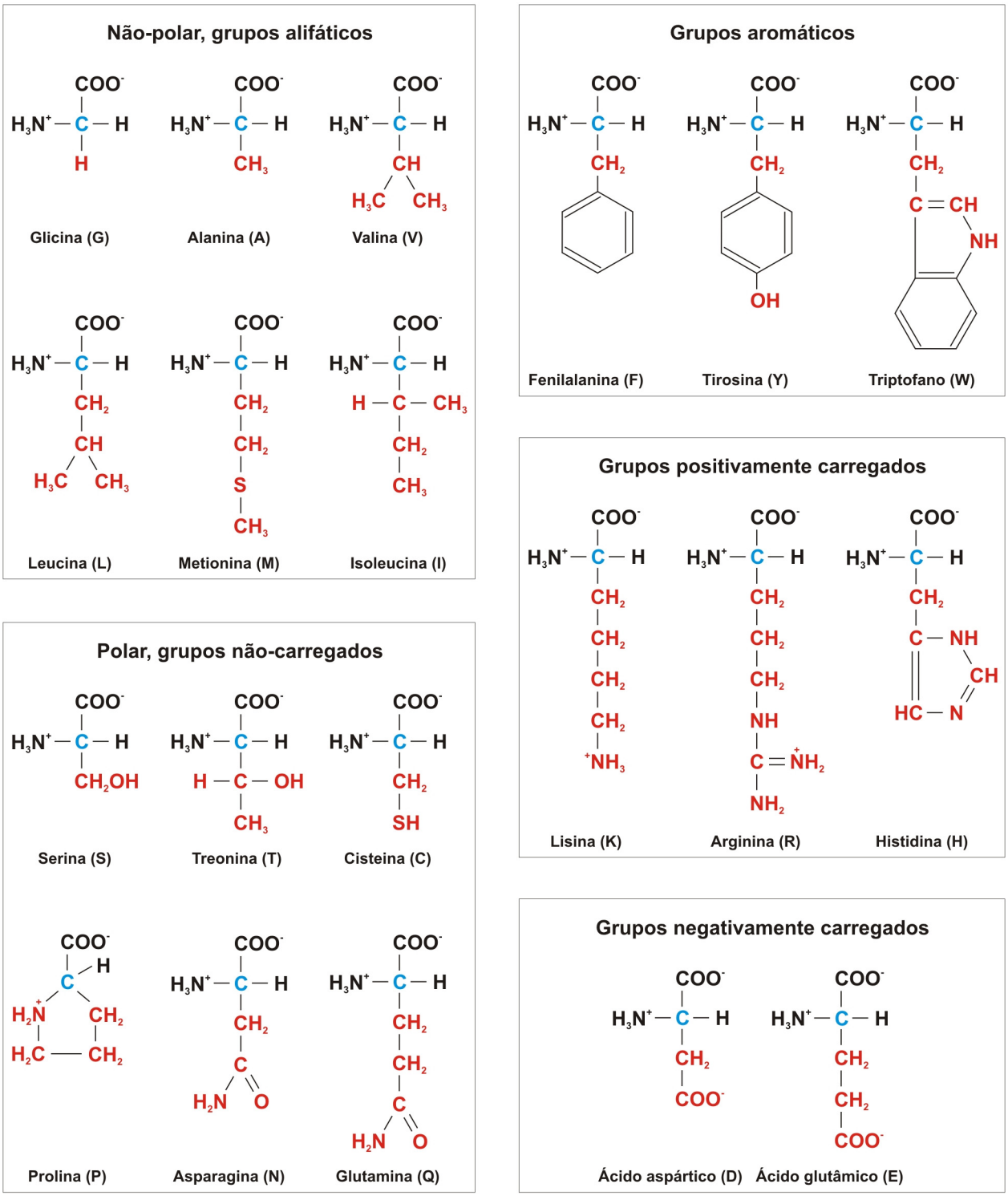


Figura 3-3. Estrutura, nomenclatura e classificação quanto à polaridade dos grupos R dos 20 aminoácidos comumente encontrados em proteínas e peptídios. Eles são mostrados com os grupos amina e carboxila ionizados, como acontece em pH 7. A prolina (P) é o único aminoácido desta lista que não apresenta a estrutura geral mostrada anteriormente. A glicina (G), por sua vez, é a única que não apresenta um centro quiral. [Adaptado de NELSON & COX, 2000].

3.2 *Ligações peptídicas*

A ligação covalente entre dois aminoácidos é chamada de ligação peptídica. A reação química ocorre pela retirada do ânion OH^- do grupo carboxila de um aminoácido e pela retirada do cátion H^+ do grupo amina do outro aminoácido resultando na formação de um dipeptídio (polipeptídio) e de uma molécula de água, como mostra a figura 3.4. De maneira similar, três ou mais aminoácidos podem se unir para formar polipeptídios maiores e, dependendo do tamanho das cadeias polipeptídicas, estas macromoléculas recebem o nome de proteínas.

Os aminoácidos que formam o polipeptídio ou a proteína passam a ser chamados de "resíduos" já que perderam alguns átomos durante a reação química. O resíduo de aminoácido localizado no extremo do polipeptídio com um grupo amina livre é chamado de resíduo N-terminal, enquanto o resíduo localizado na extremidade oposta (grupo carboxila livre) é chamado de resíduo C-terminal. Assim, os polipeptídios são reconhecidos pela seqüência de seus resíduos, começando pelo resíduo N-terminal. A formação de sucessivas ligações peptídicas gera uma cadeia de resíduos que adota uma conformação espacial característica para aquela seqüência. A cadeia formada por todos os átomos com exceção daqueles pertencentes às cadeias laterais (grupos R) constitui a chamada cadeia principal da proteína.

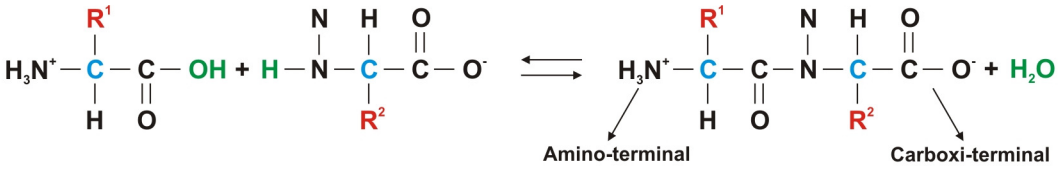


Figura 3-4. Esquema de uma ligação peptídica para formação de um dipeptídio.

Dentre todos os 20 resíduos de aminoácidos, há um que requer uma atenção especial: o resíduo cisteína. Ele pode estar presente nas proteínas em duas formas: como cisteína propriamente dita ou como cistina. A segunda forma ocorre quando dois resíduos de cisteína estão unidos covalentemente por uma ponte dissulfeto (S-S). A ponte dissulfeto estabiliza a estrutura tridimensional do polipeptídio, mas só pode ocorrer se os resíduos de cisteína estiverem espacialmente próximos um do outro.

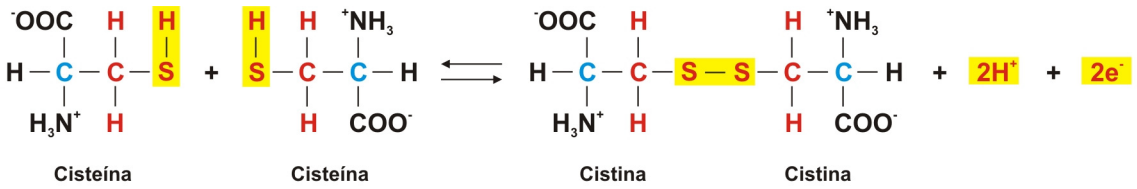


Figura 3-5. Ponte dissulfeto entre os átomos de enxofre das cadeias laterais de dois resíduos de cisteína. Cada um dos resíduos é, então, chamado de cistina.

Além da divisão por resíduos, existe uma outra forma de analisar um polipeptídio em termos de unidades repetitivas. A cadeia peptídica pode ser definida por unidades que vão de um átomo Cα até o próximo átomo Cα. A razão de se dividir a cadeia peptídica dessa maneira é que os átomos estão fixos em um plano com as distâncias e os ângulos de ligação muito similares, já que as unidades peptídicas não envolvem as cadeias laterais. Sendo essas unidades grupos rígidos unidos por ligações covalentes com os átomos Cα, podem girar apenas em torno das ligações N-Cα e Cα-C, como é mostrado na figura 3.6. Foi adotado, por convenção, que o ângulo de giro em torno da primeira ligação (N-Cα) se chamaria ϕ (phi) e o ângulo em torno da segunda ligação (Cα-C), para o mesmo átomo Cα, seria chamado de ψ (psi).

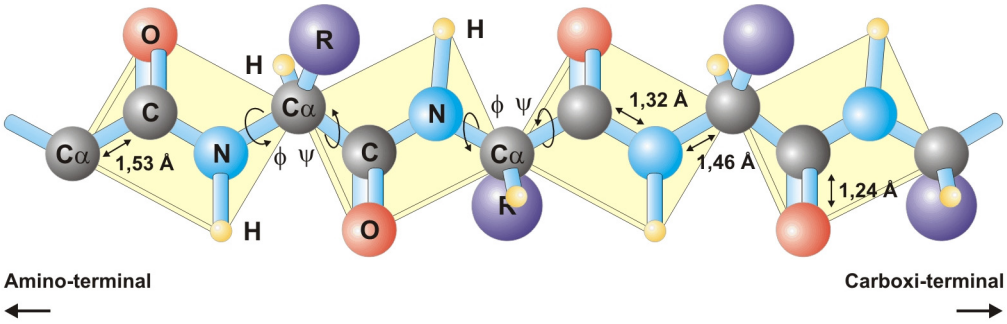


Figura 3-6. Cadeia polipeptídica onde os átomos da cadeia principal estão representados por unidades peptídicas rígidas ligadas pelos átomos Cα. As distâncias atômicas entre as principais ligações químicas também são mostradas. [Adaptado de NELSON & COX, 2000].

Assim, a conformação da cadeia principal do polipeptídio estará completamente determinada quando os ângulos ϕ e ψ estiverem definidos para cada um dos resíduos. Contudo, muitas combinações dos ângulos ϕ e ψ não são permitidas porque poderia haver colisões entre os átomos pertencentes às cadeias lateral e principal.

3.3 Proteínas

As proteínas são as macromoléculas mais abundantes nas células constituindo até 50% do peso destas quando secas. Não se conhecem, com certeza, os tamanhos mínimo e máximo da cadeia polipeptídica das proteínas, entretanto não é difícil encontrar algumas com 60 ou até 1000 resíduos de aminoácidos. Certas proteínas possuem apenas uma cadeia polipeptídica, mas outras, chamadas de oligoméricas, apresentam duas ou mais cadeias. As proteínas apresentam diversas funções biológicas e, por isso, milhares de tipos diferentes podem ser encontrados em uma única célula. Apenas como ilustração, o ser humano é dotado de mais de 30.000 genes capazes de codificar proteínas diferentes que exercem desde o transporte do oxigênio dos pulmões para as células até a regulação da atividade celular ou fisiológica (NCBI, 2003).

Além dos resíduos de aminoácidos, as proteínas podem conter, em sua estrutura tridimensional, grupos prostéticos como lipídios, açúcares, metais específicos e outros. Estas macromoléculas são denominadas proteínas conjugadas.

3.3.1 Estrutura tridimensional

A estrutura tridimensional de uma proteína é normalmente dividida em quatro partes, como mostrado na figura 3.7. A primeira parte, chamada de estrutura primária, é a seqüência de resíduos de aminoácidos da cadeia polipeptídica. Com o avanço das técnicas de biologia molecular, a obtenção da seqüência de aminoácidos se tornou uma rotina nos laboratórios de pesquisa. Contudo, somente com os estudos de difração de raios-X por cristais de proteína ou com experimentos de ressonância magnética nuclear é que informações tridimensionais precisas sobre a estrutura das proteínas puderam ser identificadas. Estes estudos permitiram verificar que diferentes regiões da seqüência de resíduos de aminoácidos de várias proteínas adotam estruturas locais regulares conhecidas como hélices-α, folhas-β, entre outras. Estes elementos estruturais locais foram chamados de estrutura secundária de uma proteína e, devido à sua grande importância, serão tratados em separado na próxima seção. A terceira parte, chamada de estrutura terciária, é formada pelo "empacotamento" dos elementos de estrutura secundária em unidades chamadas domínios. Quando uma proteína é formada por várias cadeias polipeptídicas, os domínios se agrupam em uma estrutura quaternária. Pela formação de todas essas estruturas, a seqüência de resíduos de aminoácidos é colocada em um arranjo tridimensional capaz de formar regiões funcionais conhecidas como sítios

ativos, onde estão estabelecidas as condições físico-químicas necessárias para o exercício de uma função específica.

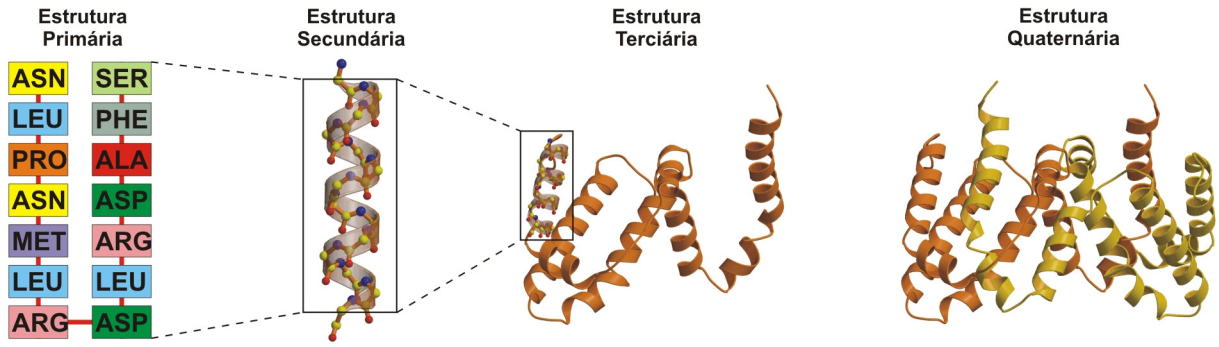


Figura 3-7. Esquema das estruturas primária, secundária, terciária e quaternária de uma proteína. [Adaptado de NELSON & COX, 2000].

3.3.1.1 Elementos de estrutura secundária

Com a determinação, em alta resolução, das estruturas moleculares de algumas proteínas, foi possível comprovar que regiões distintas da cadeia principal adotavam conformações locais similares. Essas conformações foram chamadas de estruturas secundárias e são caracterizadas pelo grande número de pontes de hidrogênio entre os grupos N-H e C=O dos resíduos de aminoácidos que as formam. Além disso, elas são obtidas quando vários números consecutivos de resíduos possuem os mesmos ângulos ϕ e ψ. As estruturas secundárias mais conhecidas são as chamadas hélices-α e as folhas-β.

A hélice-α recebe esse nome porque o trajeto espacial adotado pela cadeia principal que a compõe segue a forma helicoidal. A cadeia principal se apresenta torcida ao longo de um eixo imaginário que atravessa longitudinalmente o centro da hélice, enquanto as cadeias laterais dos resíduos ficam dispostas para fora do eixo central. Essa estrutura é obtida quando um conjunto consecutivo de resíduos apresentam ϕ e ψ em torno de –60 e –50 graus, respectivamente. A figura 3.8 mostra alguns aspectos da estrutura secundária.

A hélice-α possui 3,6 resíduos por volta e pontes de hidrogênio entre o grupo C=O do resíduo n e o grupo N-H do resíduo n+4. Assim, todos os grupos N-H e C=O estão ligados por pontes de hidrogênio exceto o primeiro grupo N-H e o último grupo C=O da hélice. Em média, uma hélice-α contém cerca de 10 resíduos, podendo ser encontradas hélices com mais de 40. Um fato

interessante a respeito das hélices-α está relacionado com o sentido do giro adotado pela cadeia principal. Apesar de L-aminoácidos serem capazes de formar hélices-α que giram tanto para a esquerda quanto para a direita, as proteínas apresentam apenas hélices-α com giro para direita.

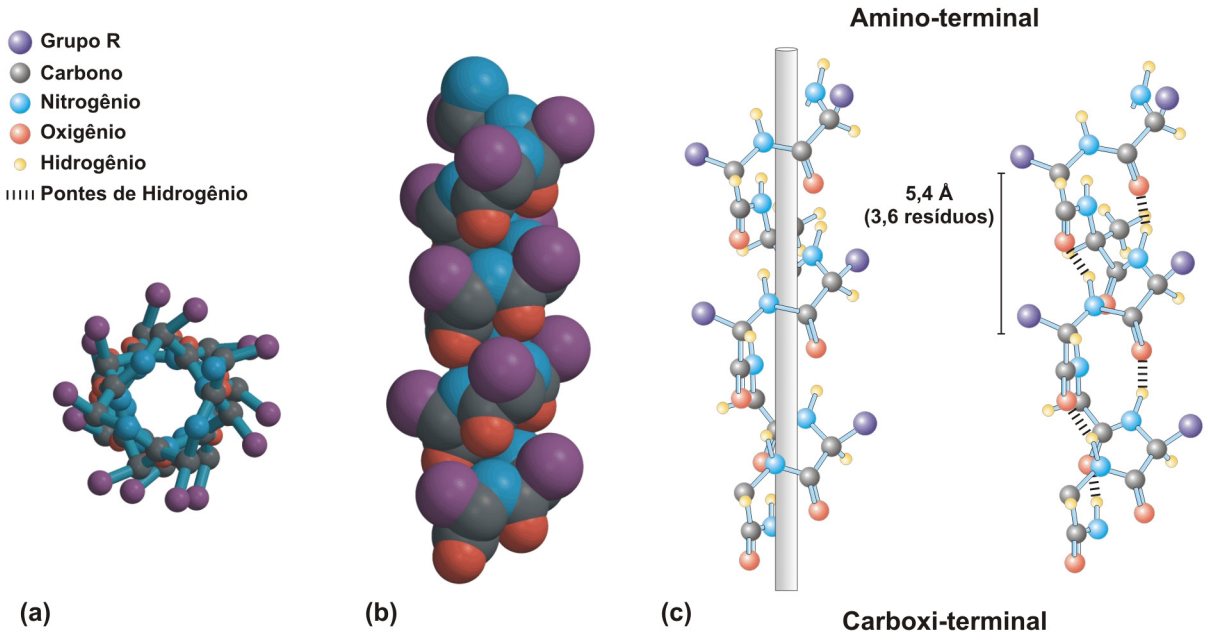


Figura 3-8. Diversas representações de uma hélice-α mostrando algumas de suas características. (a) Esta visão superior da hélice-α pode dar a falsa impressão de que o centro da hélice é vazia, contudo os raios de van der Waals dos átomos não estão corretamente representados. (b) Representação lateral utilizando átomos com raios de van der Waals corretos. Os átomos de hidrogênio foram omitidos nas representações (a) e (b). (c) Esquematização do traçado adotado pela cadeia principal em uma hélice-α. [Adaptado de NELSON & COX, 2000].

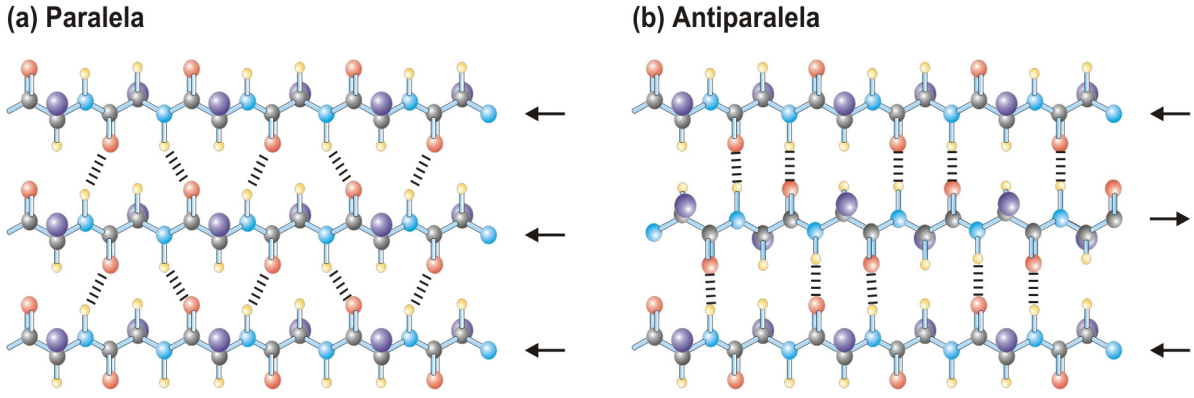


Figura 3-9. Esquema das pontes de hidrogênio entre os grupos N-H e C=O nas estruturas secundárias tipo folhas-β paralela e antiparalela. [Adaptado de NELSON & COX, 2000].

As folhas-β constituem outro grupo importante de estruturas secundárias normalmente observadas em proteínas. Ao contrário das hélices-α que são estruturas consecutivas, as folhas-β são

formadas pela combinação de diversas regiões da cadeia polipeptídica que se alinham lado a lado de forma paralela (com a mesma orientação amina-carboxila) ou antiparalela (com orientação amina-carboxila oposta). Esse tipo de estrutura secundária é formado basicamente por segmentos de 5 a 10 resíduos (fitas β) com ângulos ϕ e ψ em torno –30 a –160 e 90 a 180 graus, respectivamente.

Em cada um dos dois tipos de folhas-β (paralela e antiparalela), ocorre um padrão diferente de pontes de hidrogênio entre os grupos N-H e C=O, como pode ser visto pela figura 3.9.

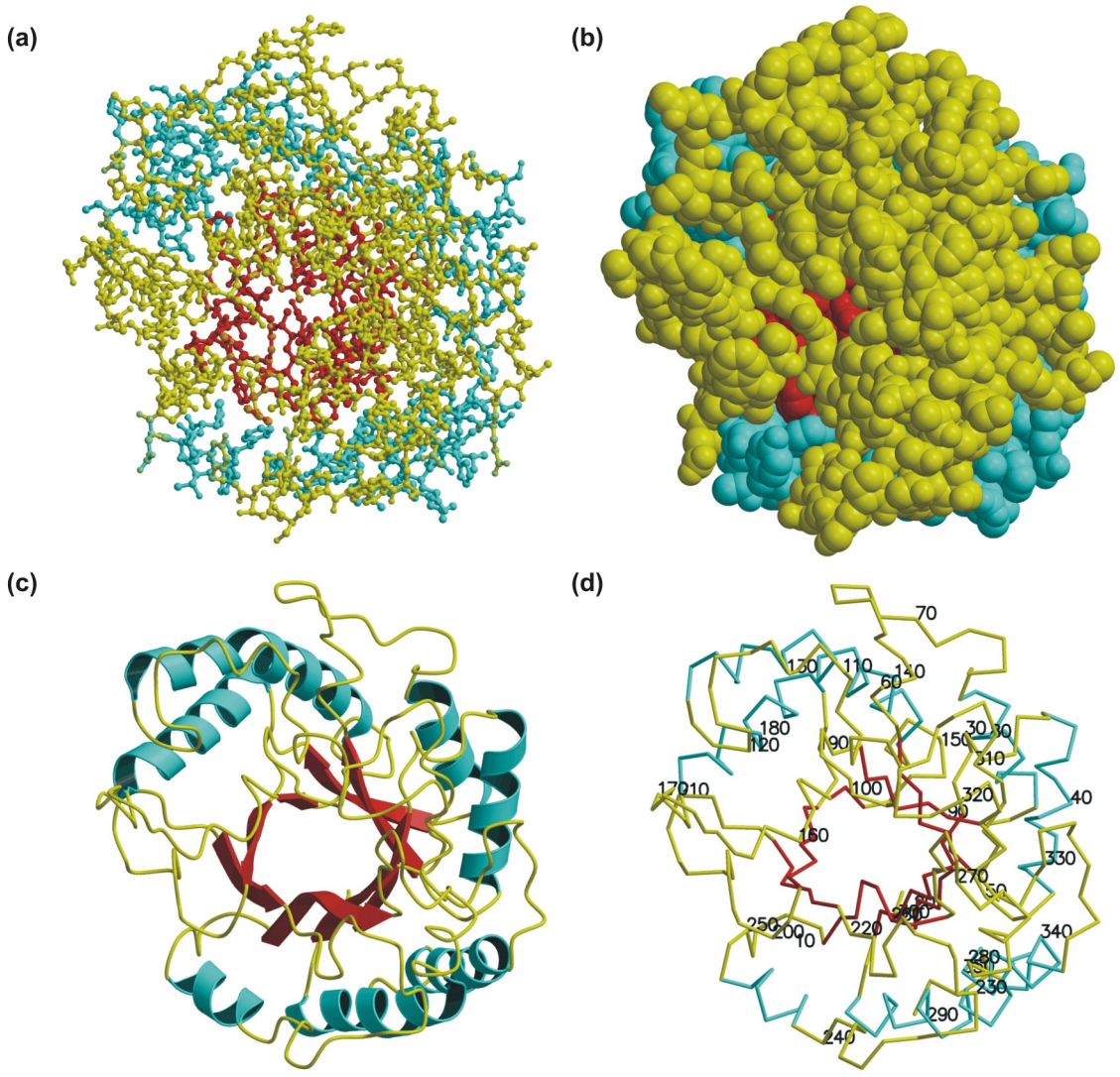


Figura 3-10. Exemplos de representações para a estrutura tridimensional de uma proteína. Representação gráfica (a) para alguns átomos e ligações químicas; (b) para todos os átomos; (c) para as estruturas secundárias (hélices-α em ciano; folhas-β em vermelho; "*loops*" em amarelo); (d) para apenas os Cα (cadeia principal). Representação gráfica para o domínio "TIM barrel" da β-galactosidase de *Penicillium sp.*.

A maior parte das proteínas é formada pela combinação dos dois tipos de estruturas secundárias (hélices-α e folhas-β). Entretanto, outros tipos podem ser encontrados, mas uma descrição de todos eles estaria bem além dos objetivos deste trabalho. Vale ressaltar, porém, a existência das estruturas secundárias tipo "*loop*" ou "β *turns*" que são elementos de conexão entre as estruturas secundárias descritas e todas elas são usadas para análise e descrição dos modelos tridimensionais das proteínas estudadas.

Para facilitar a visualização do modelo tridimensional de uma proteína e permitir uma melhor compreensão da interação entre os elementos de estrutura secundária, foram criadas as diferentes representações gráficas mostradas na figura 3.10.

4 Cristais de proteínas

Este capítulo tem o objetivo principal de permitir àqueles que não tiveram uma formação específica na área de Cristalografia de Proteínas conhecer algumas definições básicas que serão usadas, com freqüência, no decorrer deste trabalho. O capítulo começará com uma rápida introdução ao processo de crescimento dos cristais de proteínas e a preparação deles para a coleta dos dados. Em seguida, serão introduzidos alguns pontos fundamentais sobre os cristais como rede cristalina, célula unitária, índices de Miller entre outros. Uma abordagem mais detalhada sobre o tema pode ser encontrada nos seguintes livros: *Fundamentals of Crystallography* (GIACOVAZZO *et al*., 1992) e *Principles of Protein X-ray Crystallography* (DRENTH, 1999).

4.1 Princípios de cristalização de uma proteína

Um dos passos essenciais para determinar a estrutura tridimensional de uma proteína por difração de raios-X consiste no crescimento de cristais dessas macromoléculas. A cristalização de proteínas é considerada por muitos como um processo de tentativa e erro no qual as proteínas são precipitadas, lentamente, a partir de uma solução que as contém. O processo envolve algumas etapas importantes:

i. A pureza da proteína é determinada e, se necessário, alguns processos de purificação são usados. De uma forma geral, quanto mais pura estiver a proteína, maiores as chances de crescimento de cristais.

ii. A proteína é dissolvida em uma solução tampão contendo aditivos e agentes precipitantes como sulfato de amônio, cloreto de sódio e outros.

iii. A solução deve atingir um estado de supersaturação para que pequenos agregados de proteínas possam ser formados. Uma vez que os núcleos são formados, o crescimento dos cristais pode começar.

Os cristais de proteína devem ser crescidos bem lentamente para alcançarem um maior grau de ordem em sua estrutura. Entretanto, para uma precipitação mais rápida e eficiente, pode-se alterar o grau de supersaturação da solução variando a temperatura ou aumentando a concentração da proteína na solução, bem como alterar as forças de repulsão e atração entre as moléculas de proteína com o uso de um solvente orgânico ou com a mudança do pH da solução.

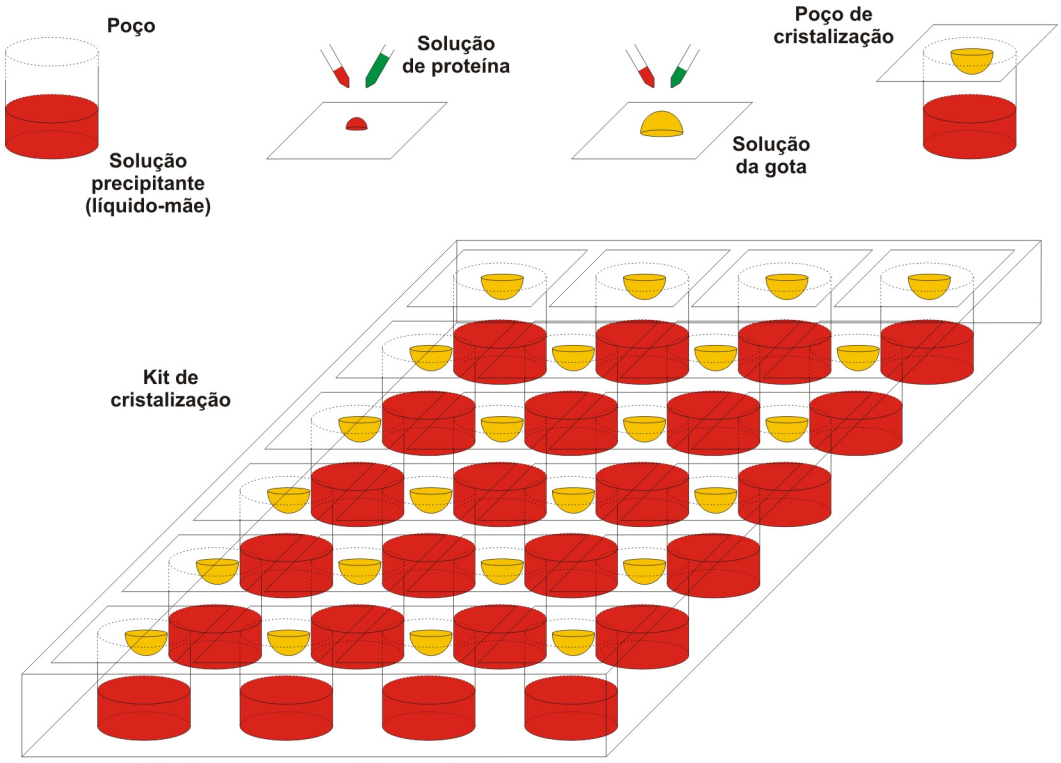


Figura 4-1. Método da gota suspensa para a cristalização de proteínas. O equilíbrio é alcançado pela difusão de vapor entre a gota e a solução precipitante.

O método de cristalização de proteínas mais conhecido atualmente e que foi utilizado em todos os experimentos descritos nesta tese é conhecido como o método da gota suspensa.

Normalmente neste método, uma única gota, formada pela mistura de 1 a 5 μl de uma solução de proteína com a mesma quantidade de uma solução precipitante, deve ser colocada sobre uma lamínula de vidro devidamente preparada. A lamínula deve ser invertida e colocada em cima de um pequeno recipiente, vedando-o. O recipiente, conhecido como poço, deve estar preenchido com a mesma solução precipitante utilizada anteriormente, como mostra a figura 4.1. A difusão do vapor do solvente entre a solução precipitante contida no poço e a gota promoverá o estado de supersaturação na gota, permitindo a formação dos primeiros núcleos de cristalização.

Dentre os parâmetros que mais afetam o processo de cristalização, pode-se citar a natureza e a concentração da proteína e do precipitante, o tipo da solução tampão e o seu pH, além da temperatura. Dessa forma, para se otimizar o processo de cristalização, muitos desses parâmetros precisam ser testados e analisados em todas as suas extensões, o que torna a cristalização de proteínas um processo que envolve inúmeras experiências sob diversas circunstâncias.

4.2 Preparação do cristal para a coleta de dados

Durante a coleta dos dados de difração de raios-X, os cristais de proteína sofrem um desgaste devido à formação de radicais livres. Sabe-se, entretanto, que o rápido resfriamento dos cristais a uma temperatura próxima à do nitrogênio líquido pode ser usado para atenuar o processo (GARMAN & SCHNEIDER, 1997). O procedimento usual de congelamento consiste, inicialmente, em retirar o cristal da gota de cristalização e transferi-lo para uma solução crio protetora (por exemplo, uma solução à base de glicerol ou etilenoglicol). Em seguida, o cristal deve ser levado rapidamente para o equipamento de raios-X, onde um fluxo contínuo de vapor de nitrogênio (100 K) incidirá sobre o cristal durante toda a coleta dos dados. Para detalhes, veja a figura 4.2.

4.2.1 Obtenção de cristais derivados

Dependendo de como a estrutura do cristal deverá ser resolvida, é necessária a preparação de cristais derivados, ou seja, cristais que apresentam, além das proteínas, algum tipo de átomo pesado ou espalhador anômalo em sua estrutura cristalina (mais explicações serão dadas no próximo capítulo). A maneira clássica para a obtenção de derivados sugere que, antes do processo de congelamento, os cristais sejam banhados por um tempo prolongado (horas ou mesmo dias) em soluções diluídas (0,01 – 0,001 M) de vários sais de átomos pesados, como K_2PtCl_4, $KAu(CN)_2$,

$Hg(CH_3COO)_2$, dentre outros. Esses procedimentos consomem muito tempo, mas como os cristais de proteína apresentam uma grande quantidade de solvente e de canais, espera-se que alguns sítios de ligação específicos sejam preenchidos pelos átomos pesados.

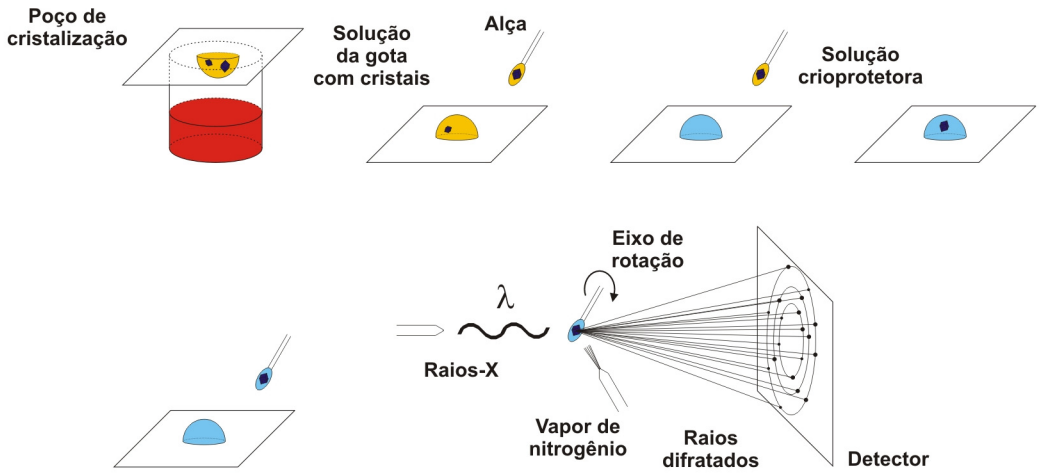


Figura 4-2. Procedimento usual de congelamento de um cristal para coleta dos dados de difração.

Recentemente, uma nova técnica de derivatização de cristais de macromoléculas biológicas, chamada "*quick cryo soaking with halides*", foi sugerida por Dauter e colaboradores (DAUTER, DAUTER & RAJASHANKAR, 2000). A técnica procura acelerar a preparação de cristais derivados já que combina, em um único passo, o processo de derivatização e a proteção criogênica. Imediatamente antes de congelar o cristal, ele é imerso, durante um curto intervalo de tempo (15 – 300 segundos), em uma gota de solução crioprotetora com alta concentração (0,25 – 1,0 M) de sais à base de iodo ou bromo (NaI, KI e NaBr principalmente). Comparado com o procedimento clássico de derivatização, que combina longos tempos de imersão e baixas concentrações de átomos pesados, este procedimento é capaz de gerar bons derivados significativamente mais rápido. Uma ilustração comparativa entre a nova técnica e o procedimento clássico de derivatização pode ser vista na figura 4.3.

Dentre os fatores principais que tornam a nova técnica extremamente eficiente, destacam-se os seguintes:

i. Permite a rápida incorporação dos átomos pesados à superfície da molécula, ocupando posições anteriormente preenchidas por moléculas ordenadas de água.
ii. Permite visualização imediata e fácil controle da deterioração do cristal.

iii. Permite que o derivado seja congelado imediatamente e usado na experiência, uma vez que a solução derivatizante já possui em sua composição agentes crio protetores.

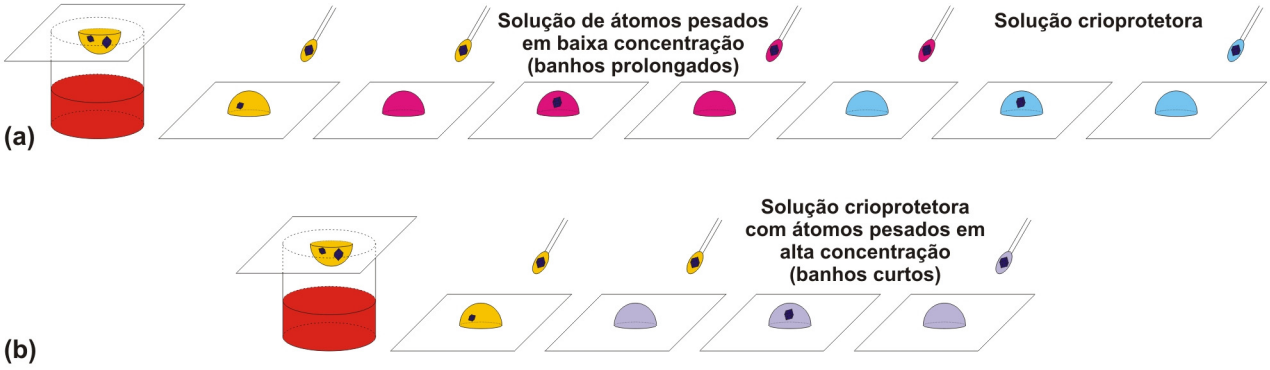


Figura 4-3. Preparação de derivados para a coleta de dados. (a) Processo tradicional para a preparação de um derivado. (b) Uso da técnica *quick cryo soaking* para obtenção de derivados.

4.3 A respeito dos cristais

Os cristais existem em uma enorme variedade de formas, cores e tamanhos e, por isso, ainda hoje, impressionam muito as pessoas. A presença de faces naturalmente planas – reflexo do ordenamento no arranjo espacial de suas moléculas, átomos ou íons – permite sua distinção das substâncias amorfas. Entretanto, o arranjo espacial não pode ser observado sem o auxílio de equipamentos e técnicas especiais já que as entidades constituintes dos cristais são muito pequenas.

Segundo DRENTH (1999), o estudo dos cristais começou em 1669 quando Nicolaus Steno propôs que, durante o crescimento dos cristais, os ângulos entre as faces permaneciam constantes. A regularidade interna dos cristais foi sugerida, mas, somente em 1912, ela foi provada por Max von Laue, Walter Friedrich e Paul Knipping na primeira experiência de difração de raios-X.

Uma das características fundamentais do estado cristalino é a repetição regular, em três dimensões, de um motivo (objeto) composto por moléculas, íons ou átomos, estendendo por distâncias correspondentes a milhares de dimensões atômicas. As moléculas (íons ou átomos), ao se precipitarem quando em solução, procuram alcançar um estado de mais baixa energia livre. Este processo vem acompanhado por uma disposição regular delas, o que torna possível a formação do cristal. Porém, de uma forma geral, os cristais possuem inúmeros defeitos ou contêm impurezas, sem entretanto perderem sua ordem. Para o estudo teórico desenvolvido aqui e no próximo capítulo, supõe-se a existência de um cristal ideal, sem defeitos nem impurezas.

4.3.1 Célula unitária

A periodicidade translacional encontrada nos cristais pode ser convenientemente estudada considerando a geometria das repetições ao invés das propriedades do motivo que se repete. Suponha-se, portanto, que uma determinada substância, representada pela figura 4.4 a, foi cristalizada nas formas ilustradas pelas figuras 4.4 b e 4.4 c (a analogia em duas dimensões foi utilizada para facilitar a compreensão). Pode-se notar que o motivo que se repete em cada uma das duas formas cristalinas é diferente apesar de ambos serem constituídos pela mesma substância. Se o motivo é repetido periodicamente em intervalos a, b e c ao longo de 3 direções não-coplanares, a geometria da repetição pode ser completamente descrita por uma seqüência periódica de pontos, separados por intervalos a, b e c ao longo das mesmas 3 direções. Ao conjunto desses pontos dá-se o nome de rede cristalina. É importante salientar que o ambiente atômico ao redor de cada ponto da rede é o mesmo, o que torna cada ponto indistinguível de qualquer outro. As redes cristalinas para os dois cristais bidimensionais das figuras 4.4 b e 4.4 c podem ser vistas nas figuras 4.4 d e 4.4 e, respectivamente.

Se um ponto da rede for escolhido como a origem da rede cristalina, a posição de qualquer outro ponto será definida pelo vetor $\mathbf{T}_{n_1,n_2,n_3} = n_1\mathbf{a} + n_2\mathbf{b} + n_3\mathbf{c}$, onde $\mathbf{a}$, $\mathbf{b}$ e $\mathbf{c}$ são os vetores de módulo a, b e c ao longo das 3 direções não-coplanares e n_1, n_2 e n_3 são números inteiros[2]. Os vetores $\mathbf{a}$, $\mathbf{b}$ e $\mathbf{c}$, bem como os ângulos α, β e γ entre eles, definem um paralelepípedo chamado de célula unitária, como é mostrado na figura 4.5. As linhas nas direções de $\mathbf{a}$, $\mathbf{b}$, e $\mathbf{c}$ são chamadas de eixos x, y e z, respectivamente.

A escolha da célula unitária pode ser feita de várias maneiras, como mostram as figuras 4.4 d e 4.4 e. Células unitárias contendo apenas um ponto da rede são chamadas primitivas, e aquelas contendo mais de um ponto são chamadas múltiplas ou centradas. Dentre as regras básicas adotadas, por convenção, para a escolha correta da célula unitária, destacam-se as seguintes:

i. O sistema de eixos deve ser *right-handed*, ou seja, os vetores $\mathbf{a}$, $\mathbf{b}$ e $\mathbf{c}$ devem adotar direções tais que um observador localizado ao longo da direção positiva de z veja o eixo x se movendo em direção ao eixo y por uma rotação no sentido anti-horário.

ii. Os vetores $\mathbf{a}$, $\mathbf{b}$ e $\mathbf{c}$ devem coincidir, o máximo possível, com as direções de mais alta simetria no cristal.

[2] Em alguns casos, dependendo da escolha dos vetores **a**, **b** e **c**, os índices n_1, n_2 e n_3 podem assumir valores racionais. Um exemplo disso pode ser visto em cristais que apresentam células unitárias múltiplas ou centradas.

iii. A célula unitária escolhida deve ser a menor possível, desde que seja capaz de satisfazer a condição ii.

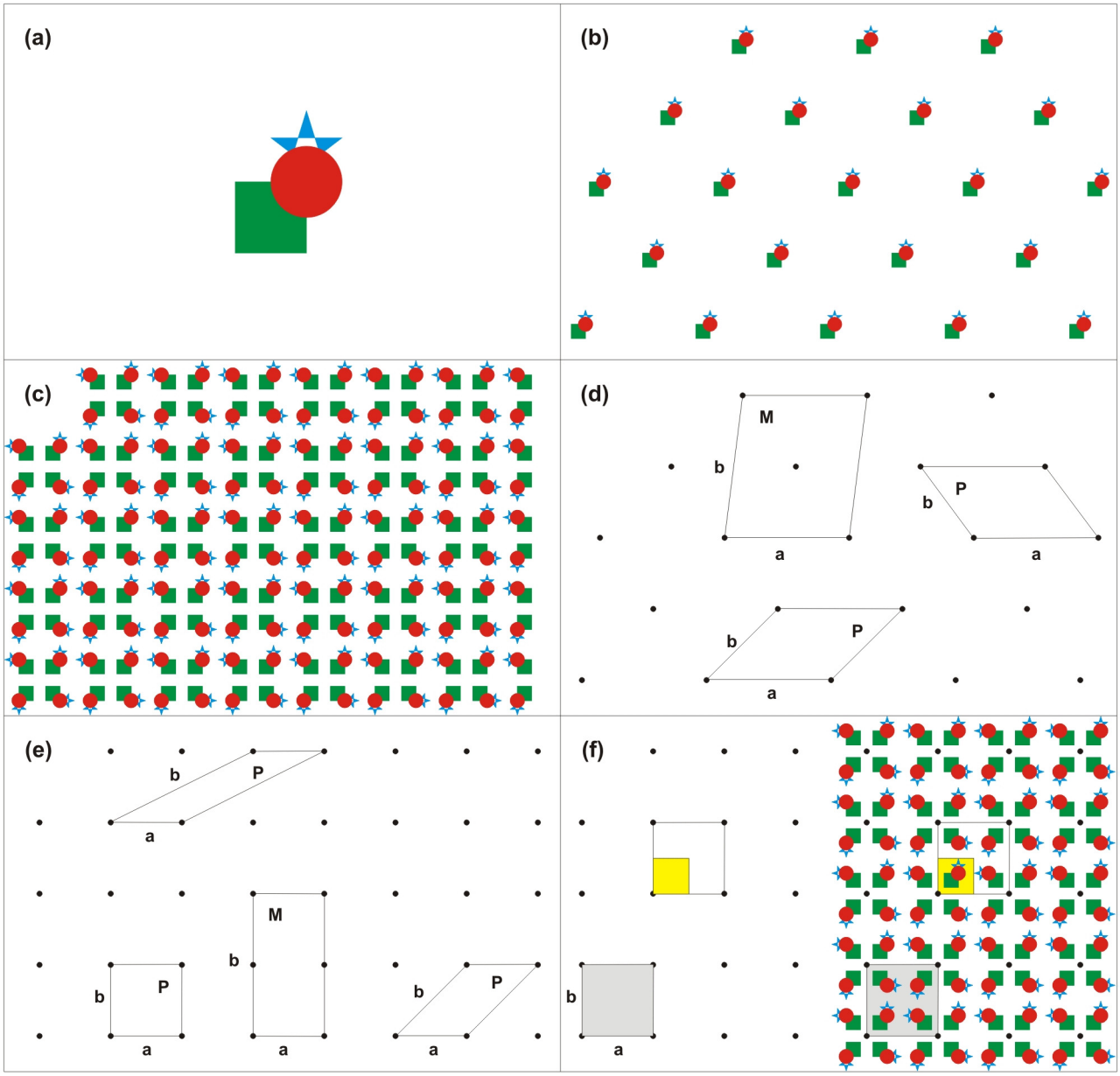


Figura 4-4. Ilustração, em duas dimensões, simulando a formação de um cristal. (a) Substância utilizada para a cristalização. (b) Ilustração para a primeira forma cristalina. O motivo que se repete contém apenas uma unidade da substância usada para a cristalização. (c) Ilustração para a segunda forma cristalina. O motivo que se repete contém quatro unidades da substância usada para a cristalização, relacionadas por um eixo de ordem 4. (d) Rede cristalina para o cristal representado em (b). (e) Rede cristalina para o cristal representado em (c). Algumas possíveis células unitárias são desenhadas em (d) e (e). As células marcadas com a letra “P” são primitivas, e as com letra “M” são múltiplas ou centradas. (f) Definições corretas para a célula unitária (em cinza) e para a unidade assimétrica (em amarelo) do cristal representado em (c).

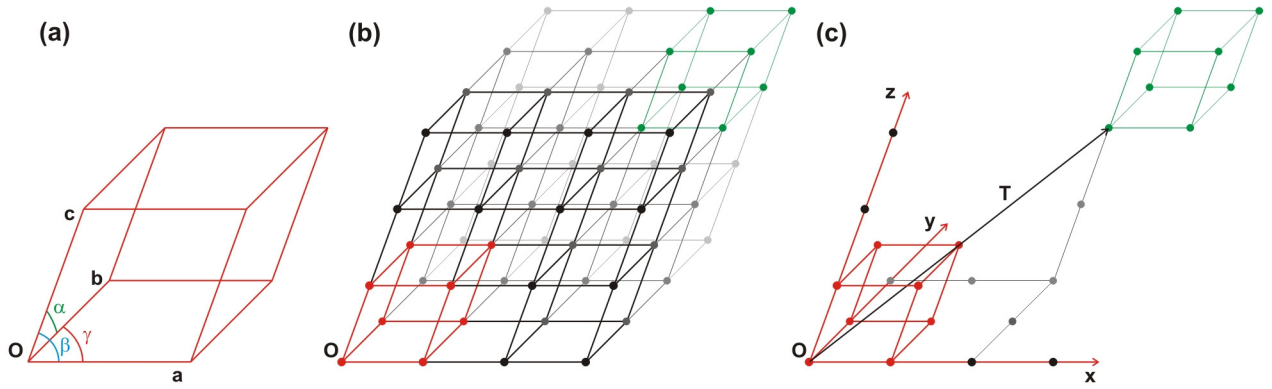


Figura 4-5. Representação de uma célula unitária geral. (a) Célula unitária com eixos **a**, **b** e **c** e ângulos α, β e γ situada na origem **O**. (b) A rede cristalina é obtida com o arranjo tridimensional das células unitárias. (c) Qualquer célula unitária do cristal está relacionada com a célula da origem por uma operação de translação **T**.

Para finalizar esta seção, convém mencionar que o conteúdo de uma célula unitária varia de cristal para cristal e pode conter desde poucos íons até algumas dezenas de moléculas com milhares de átomos. Além disso, pode-se ver que a estrutura de um cristal é obtida pela repetição, em cada ponto da rede, da célula unitária, ou seja, pela aplicação da operação de translação $\mathbf{T} = n_1\mathbf{a} + n_2\mathbf{b} + n_3\mathbf{c}$ a uma célula unitária (figura 4.5 c).

4.3.2 Índices de Miller

Imagine-se uma célula unitária de vetores **a**, **b** e **c** e ângulos α, β e γ localizada em uma origem **O** da rede cristalina. Suponha-se, então, que um único plano, cortando todo o cristal, intercepte os vetores **a**, **b** e **c** nos pontos a/h, b/k e c/l, respectivamente, sendo h, k e l inteiros. Os números h, k e l são chamados de índices de Miller e podem ser usados para dar nome ao plano. Pode-se mostrar que a menor distância d_{hkl} entre este plano e a origem **O** da rede cristalina é dada pela equação 4.1.

$$d_{hkl} = \left(1-\cos^2(\alpha)-\cos^2(\beta)-\cos^2(\gamma)+2\cos(\alpha)\cos(\beta)\cos(\gamma)\right)^{1/2}\left(\frac{h^2}{a^2}sen^2(\alpha)+\frac{k^2}{b^2}sen^2(\beta)+\frac{l^2}{c^2}sen^2(\gamma)+\cdots\right.$$
$$\left.\cdots\frac{2kl}{bc}\left(\cos(\beta)\cos(\gamma)-\cos(\alpha)\right)+\frac{2lh}{ca}\left(\cos(\gamma)\cos(\alpha)-\cos(\beta)\right)+\frac{2hk}{ab}\left(\cos(\alpha)\cos(\beta)-\cos(\gamma)\right)\right)^{-1/2}$$

Eq. 4-1

Se uma seqüência de planos paralelos ao plano inicial é criada, mantendo a mesma distância d_{hkl} entre eles, uma família de planos (hkl) atravessará todos os pontos da rede cristalina como pode ser facilmente observado na figura 4.6. Os planos cristalinos que não cortam um dos vetores **a** , **b** ou **c** possuem um índice de Miller, para aquela direção, igual a zero.

Um dos motivos principais para a definição dos planos imaginários é o fato de que o processo de difração dos raios-X por um cristal pode ser visto como uma "reflexão" da radiação nos planos. Esta visão apresenta algumas limitações, mas simplifica enormemente a teoria cinemática da difração, que calcula basicamente os efeitos de interferência entre as ondas espalhadas dentro do volume do cristal. Ela desconsidera, por exemplo, os efeitos de atenuação do feixe ao penetrar no cristal bem como os efeitos de interferência entre as ondas incidentes e espalhadas. A teoria que leva em conta todos esses fenômenos é chamada de teoria dinâmica da difração.

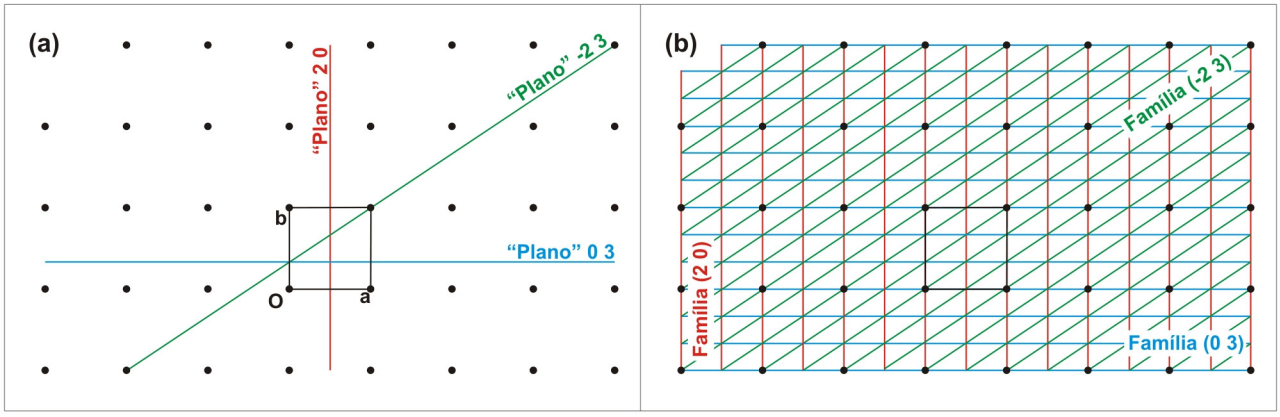


Figura 4-6. Uma rede cristalina bidimensional ilustrando os índices de Miller para um certo conjunto de "planos". (a) Rede cristalina com uma célula unitária situada na origem O e alguns "planos" cruzando os eixos **a** e **b**. (b) As três famílias de "planos" são obtidas com a repetição dos "planos" iniciais.

4.3.3 Simetrias adicionais

Como foi visto nas seções anteriores, todo cristal possui uma simetria tridimensional de translação, correspondente à repetição das células unitárias. Além disso, a disposição regular das moléculas durante o crescimento do cristal permite o surgimento de outras simetrias como eixos de rotação e roto-translação, planos de espelho, centros de inversão, entre outras. Um exemplo pode ser visto na figura 4.4 c, onde, além da simetria translacional, eixos de rotação de ordem 2 (180^o) e 4 (90^o) são encontrados na estrutura cristalina. A combinação de todas as operações de simetria

admitidas em um cristal permite classificá-lo entre 230 grupos espaciais diferentes. Para cristais de proteína, entretanto, o número de grupos espaciais encontrados é de apenas 65. Isso se dá porque a aplicação de certas operações de simetria pode mudar a assimetria de um aminoácido. Um exemplo disso é que um L-aminoácido seria transformado pelas simetrias em um D-aminoácido, embora isso nunca tenha sido observado em cristais de proteínas. Os detalhes de todos os 230 grupos espaciais podem ser encontrados no livro *International Tables for Crystallography volume A* (HAHN, 2002). Ao conjunto das simetrias presentes em um determinado grupo espacial dá-se o nome de simetrias cristalográficas.

Devido à presença das simetrias cristalográficas, diversos objetos simetricamente relacionados irão coexistir dentro de uma única célula unitária. Portanto pode-se definir a unidade assimétrica como o menor volume fechado de uma célula unitária que, com a aplicação das operações de simetria, é capaz de gerá-la novamente. Essa definição pode ser mais bem compreendida com a análise da figura 4.4 f, que mostra uma das unidades assimétricas do cristal representado pela figura 4.4 c. Assim, para que a estrutura de um cristal possa ser determinada, deve-se conhecer seu grupo espacial e os parâmetros de sua célula unitária, bem como o conteúdo da unidade assimétrica e a disposição espacial do conteúdo.

Com a análise dos primeiros dados de um experimento de difração de raios-X por um cristal, é possível determinar os parâmetros de sua célula unitária e, em alguns casos, seu grupo espacial. Entretanto, a determinação do conteúdo da unidade assimétrica e da disposição dos átomos dentro dela, requer, em muitos casos, a coleta de mais de um conjunto de dados além de alguns meses de análise. Detalhes a respeito desse assunto serão abordados no próximo capítulo.

5 Determinação de estruturas cristalográficas

Neste capítulo serão introduzidos os princípios básicos de um experimento de difração de raios-X por cristais, bem como alguns dos métodos utilizados, atualmente, para a determinação das estruturas cristalográficas. Uma seção deste capítulo será utilizada para a demonstração da resolução do problema das fases, ponto-chave na área de Cristalografia de Proteínas e no contexto desta tese. Esta seção estará focada principalmente nos métodos utilizados ao longo desta tese para a resolução de diversas estruturas de macromoléculas biológicas. Uma abordagem mais detalhada sobre os temas acima e a dedução de algumas equações apresentadas ao longo do texto podem ser encontradas nos seguintes livros: *Fundamentals of Crystallography* (GIACOVAZZO *et al*., 1992) e *Principles of Protein X-ray Crystallography* (DRENTH, 1999).

5.1 Difração de raios-X pelo método de rotação

Diversos métodos experimentais, utilizando raios-X, são empregados atualmente para a análise de amostras biológicas. Em Cristalografia de Proteínas, o método de rotação é o mais utilizado. Neste método, um determinado cristal é submetido a uma fonte monocromática de raios-X

com comprimento de onda entre 0,7 e 1,8 Å por um certo intervalo de tempo. Durante esse tempo, o cristal é girado ao redor de um eixo de rotação por alguns graus (0,5^{o} – 2,0^{o}) e a radiação espalhada pelo cristal é medida por um detector bidimensional de raios-X. O resultado da medida é chamado de um padrão de difração e pode ser visto na figura 5.1.

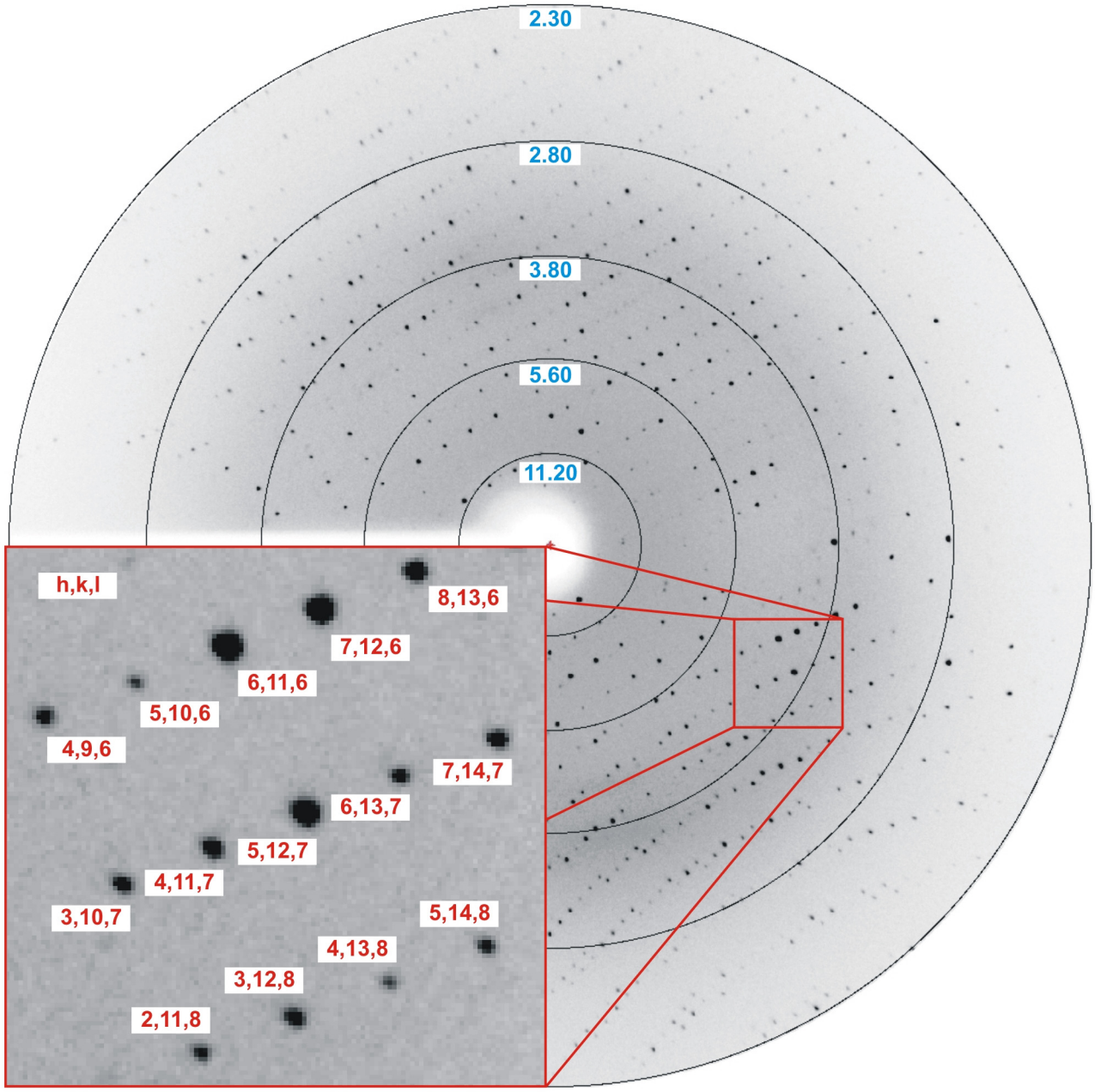


Figura 5-1. Imagem de um padrão de difração de um cristal de lisozima. O cristal pertence ao grupo espacial $P4_32_12$ com parâmetros de célula a = 78,62, b = 78,62, c = 37,05 Å e $\alpha = \beta = \gamma = 90.00^{o}$. Algumas reflexões são mostradas com seus respectivos índices de Miller. Os círculos mostrados na figura indicam as faixas de resolução do conjunto de dados. Os dados foram coletados no Laboratório Nacional de Luz Síncrotron usando radiação monocromática de 1,54 Å.

O procedimento é repetido várias vezes, girando o cristal, no mesmo intervalo em graus a partir do ponto de parada da medida anterior. Essa maneira de medir permite acompanhar, em imagens sucessivas, o desenvolvimento do padrão de difração à medida que o cristal é girado.

Na figura 5.2 são mostradas algumas das instalações do Laboratório Nacional de Luz Síncrotron, único laboratório da América do Sul capaz de produzir radiação síncrotron utilizada em diversos experimentos científicos. Na mesma figura, são mostrados alguns detalhes a respeito da linha de luz CPr, permitindo a identificação de vários equipamentos usados durante uma coleta de dados de difração de raios-X por cristais de proteínas.

Figura 5-2. Instalações do Laboratório Nacional de Luz Síncrotron (LNLS) localizado na cidade de Campinas (SP). (a) Vista aérea do campus do LNLS. No maior prédio do campus fica o anel de armazenamento de elétrons. (b) Vista superior do anel de armazenamento de elétrons no LNLS. (c) Linha de luz CPr no anel de armazenamento. Montagem experimental para a coleta de dados de difração de raios-X pelo método de rotação. (d) Equipamentos para automatização da aquisição e processamento dos dados de difração na linha de luz CPr.

5.2 Lei de Bragg

Freqüentemente, uma das perguntas que surgem após a análise de um padrão de difração é a seguinte: Por que se observa uma grande área do detector com muito pouca intensidade (áreas em claro) e alguns pontos com elevada intensidade (pontos mais escuros) ?

Respostas com maior ou menor grau de informação podem ser encontradas em inúmeros livros de cristalografia. Contudo, todas estão baseadas nos efeitos de espalhamento de raios-X pela matéria (principalmente elétrons) e na interferência das ondas espalhadas para a formação do padrão de difração.

Uma das maneiras mais simples de se compreender o resultado do experimento de difração, ainda que não inteiramente completa sob o ponto de vista quantitativo, é pela formulação de William Lawrence Bragg (BRAGG & BRAGG, 1913). Bragg considerou a difração como uma conseqüência da "reflexão" de vários feixes de raios-X por inúmeros planos de Miller pertencentes a uma mesma família (hkl) (seção 4.3.2). Suponha-se, portanto, que um feixe de raios-X com direção $\mathbf{s}_o$ incida sobre um cristal, fazendo um ângulo de incidência θ (ângulo de Bragg) com uma família de planos (hkl)[3], conforme mostrado na figura 5.3.

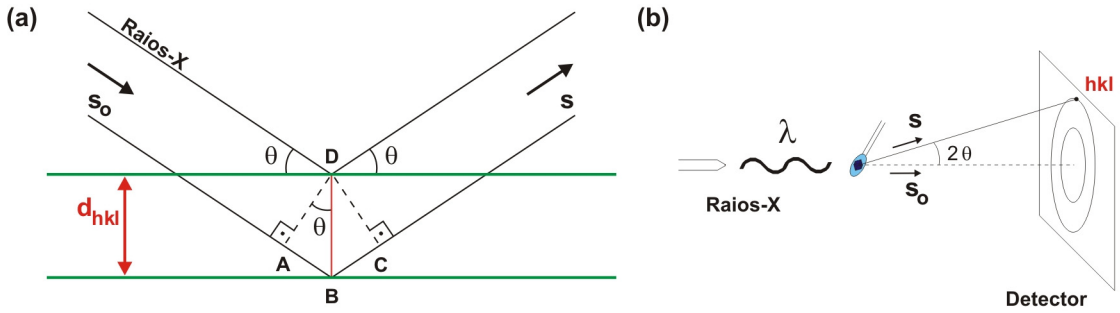


Figura 5-3. Reflexão dos raios-X por dois planos pertencentes a uma mesma família (*hkl*). A distância entre os planos é dada por d_{hkl}. (a) Visão de "dentro" de um cristal. (b) Visão do experimentador.

A diferença de caminho entre os feixes "refletidos" em D e em B é igual a $AB + BC = 2d_{hkl} sen(\theta)$, já que o ângulo de reflexão também vale θ (a direção do feixe refletido é dada por $\mathbf{s}$). Se a diferença de caminho entre os feixes for um múltiplo inteiro do comprimento de onda da radiação usada no experimento, então as ondas que representam cada um dos feixes

[3] Por hora vamos considerar famílias de planos com índices *h*, *k* e *l* sem nenhum fator inteiro maior que 1 em comum.

"refletidos" irão se combinar com um máximo de interferência construtiva. Essa afirmação é conhecida como a Lei de Bragg e a equação que a quantifica é a seguinte:

$$2\mathrm{d}_{hkl}sen(\theta) = n\lambda \ . \qquad \text{Eq. 5-1}$$

Como os raios-X penetram profundamente nos cristais, um grande número de planos refletirá os feixes incidentes. Porém, somente aquelas ondas refletidas por planos que obedecem à equação 5.1 irão interferir construtivamente enquanto as demais interferirão destrutivamente. À medida que o cristal é girado no método de rotação, o ângulo θ mudará para cada família de planos (hkl) de maneira que certos planos que estavam refletindo construtivamente os feixes incidentes passarão a refleti-los destrutivamente, enquanto outros planos farão justamente o contrário. Assim, o que se observa no padrão de difração à medida que o cristal é girado é uma mudança dos planos que estão contribuindo para a difração naquela orientação do cristal.

A equação 5.1 pode ser reescrita para qualquer família de planos $(h'k'l')$ [4] como

$$2\mathrm{d}_{h'k'l'}sen(\theta) = \lambda \qquad \text{Eq. 5-2}$$

uma vez que famílias de planos com índices $h' = nh$, $k' = nk$ e $l' = nl$ apresentam uma distância entre seus planos $\mathrm{d}_{h'k'l'} = \mathrm{d}_{hkl}/n$ (equação 4.1).

Em vista dessa abordagem simplificada para o processo de difração, convencionou-se chamar de "reflexão" a cada um dos pontos máximos de interferência das ondas refletidas. Além disso, cada reflexão pode ser identificada pelos índices de Miller da família de planos (hkl) que a gerou. Na figura 5.1 são mostradas algumas reflexões com seus respectivos índices de Miller. A distância $\mathrm{d}_{h'k'l'}$ entre os planos $(h'k'l')$ é utilizada, com freqüência, para designar a resolução de uma reflexão. Reflexões provenientes de planos afastados, ou seja, com grande $\mathrm{d}_{h'k'l'}$ são consideradas de baixa resolução, enquanto aquelas provenientes de planos próximos, com pequeno $\mathrm{d}_{h'k'l'}$, são de alta resolução. Na figura 5.1 o conjunto de dados é separado por faixas de resolução, o que permite identificar as resoluções máxima e mínima de um conjunto de dados.

[4] Podemos considerar agora família de planos com índices *h'*, *k'* e *l'* com fatores inteiros maior ou igual a 1 em comum.

Ao final de uma coleta de dados, inúmeras imagens de difração são obtidas; cada uma referente a uma determinada orientação do cristal. Cada imagem deve ser analisada individualmente de modo que cada reflexão possa ser identificada e o valor de sua intensidade possa ser medido. Os processos pelos quais essas tarefas são executadas são conhecidos como "indexação" e "integração" das imagens de difração. Por fim, todas as reflexões de todas as imagens são comparadas em uma etapa conhecida como "escalonamento", em que as intensidades das reflexões são ajustadas a uma escala comum, e os erros associados a cada uma das intensidades podem ser estimados com maior precisão. Assim, de uma forma bem simplificada, o resultado de um experimento de difração de raios-X por um cristal é um conjunto de reflexões (hkl) com suas intensidades e respectivos erros.

5.3 Métodos para obtenção da estrutura cristalográfica

Na seção anterior, a teoria adotada para explicar o experimento de difração permite prever as condições necessárias para que uma determinada reflexão apareça em um padrão de difração, porém não permite saber qual a relação entre o conteúdo de um cristal e a intensidade de cada uma das reflexões medidas. Além disso, não proporciona informação que permita a resolução da estrutura cristalina.

Para que a estrutura cristalográfica de um determinado composto (biológico ou inorgânico) possa ser determinada pela técnica de difração de raios-X, é necessário que se estabeleçam relações matemáticas entre o conteúdo do cristal e os dados medidos. É possível mostrar que as relações de fato existem. As grandezas físicas que as expressam recebem o nome de "fatores de estrutura" ($\mathbf{F}_{hkl}$). A definição geral para essas grandezas é mostrada na equação 5.3.

$$\mathbf{F}_{hkl} = \sum_{j=1}^{N} \mathrm{f}_j(hkl) \exp[2\pi i(hx_j + ky_j + lz_j)] = \mathrm{F}_{hkl} \exp[i\alpha_{hkl}], \qquad \text{Eq. 5-3}$$

onde F_{hkl} e α_{hkl} são chamados, respectivamente, de amplitude e fase do fator de estrutura $\mathbf{F}_{hkl}$. Os índices h, k e l que aparecem aqui são os índices de Miller, vistos anteriormente. Os termos x_j, y_j, z_j e $\mathrm{f}_j(hkl)$ são as coordenadas atômicas e o fator de espalhamento atômico do j-ésimo átomo (de um total de N átomos) dentro da célula unitária do cristal. O fator de espalhamento atômico $\mathrm{f}_j(hkl)$ é uma grandeza que pode ser calculada para cada tipo de átomo, pois é definida por

$$\mathrm{f}_j(hkl) = \int_V \rho_j(x'y'z')\exp[2\pi i(hx' + ky' + lz')]dv' = \mathrm{f}_j^{\circ}, \qquad \text{Eq. 5-4}$$

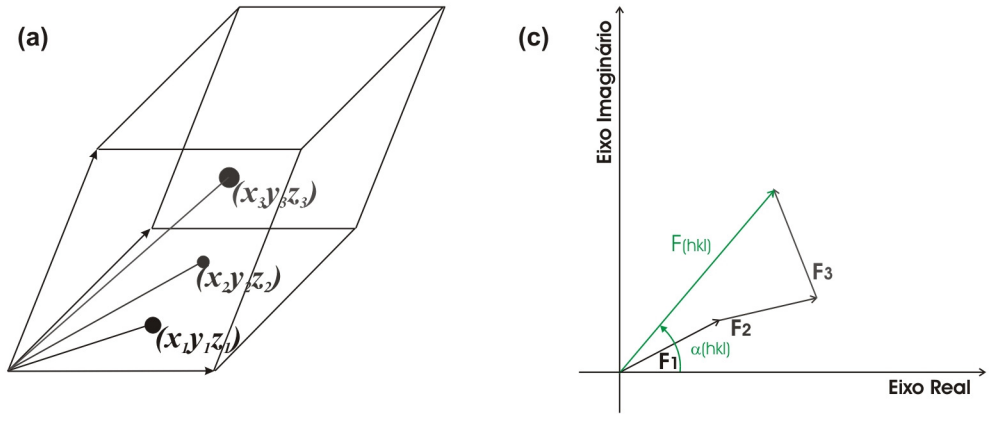
onde $\rho_j(x'y'z')$ é a densidade eletrônica do átomo j e a integral é feita sobre o volume V' do átomo. Assim, se a estrutura cristalográfica de um determinado composto é conhecida *a priori*, o termo central da equação 5.3 permite o cálculo exato do valor de $\mathbf{F}_{hkl}$ (ou F_{hkl} e α_{hkl}) para cada valor de h, k e l. Note-se que $\mathbf{F}_{hkl}$ (equação 5.3) é um número complexo que depende da posição atômica e do fator de espalhamento atômico de cada átomo dentro do cristal, bem como dos índices h, k e l. Caso a ocupância e a agitação térmica de cada átomo sejam levadas em conta para o cálculo do fator de estrutura $\mathbf{F}_{hkl}$, a equação 5.3 terá um termo adicional Occ_j que indica a freqüência da presença do átomo j nas células unitárias do cristal e um termo $\exp(-B_j sen^2\theta/\lambda^2)$, onde B_j é conhecido como o fator de temperatura do átomo j, θ é o ângulo de Bragg para a família de planos (hkl) e λ é o comprimento de onda da radiação incidente. Assim, a equação completa para $\mathbf{F}_{hkl}$ seria dada por

$$\sum_{j=1}^{N} \mathrm{Occ}_j \mathrm{f}_j(hkl)\exp[-B_j sen^2\theta/\lambda^2]\exp[2\pi i(hx_j + ky_j + lz_j)]. \qquad \text{Eq. 5-5}$$

Entretanto, esses dois novos termos podem ser incorporados no cálculo de $\mathrm{f}_j(hkl)$ e a omissão deles na equação 5.5 pode ser feita sem perda de generalidade. Na figura 5.4, são mostrados o cálculo de um fator de estrutura qualquer $\mathbf{F}_{hkl}$ para um cristal contendo três átomos distintos na célula unitária e a representação de $\mathbf{F}_{hkl}$ em um plano de Argand.

Acontece entretanto que, em um experimento de difração de raios-X, o resultado medido, ou seja, a intensidade de cada máximo de difração é um número real conhecido como I_{hkl}. Então, como relacionar $\mathbf{F}_{hkl}$ com I_{hkl}? Pode-se mostrar que o quadrado do módulo do fator de estrutura é diretamente proporcional à intensidade da reflexão medida, de forma que

$$\mathrm{I}_{hkl} \propto \mathbf{F}_{hkl} \times \mathbf{F}_{hkl}^{*} = \left|\mathbf{F}_{hkl}\right|^2 = \mathrm{F}_{hkl}{}^2. \qquad \text{Eq. 5-6}$$



(b) **Fator de Estrutura**

$$\mathbf{F(hkl)} = F_{(hkl)} \exp[i\alpha_{(hkl)}] = \Sigma\, \mathbf{f}_j \exp[2\pi i(hx_j + ky_j + lz_j)]$$
$$= \mathbf{f}_1 \exp[2\pi i(hx_1 + ky_1 + lz_1)] + \mathbf{f_2} \exp[2\pi i(hx_2 + ky_2 + lz_2)]$$
$$+ \mathbf{f}_3 \exp[2\pi i(hx_3 + ky_3 + lz_3)]$$

Figura 5-4. Cálculo do fator de estrutura. (a) Representação de um cristal com três átomos na célula unitária. (b) Cálculo de um fator de estrutura $\mathbf{F}_{hkl}$ para o mesmo cristal. (c) Representação no plano de Argand de $\mathbf{F}_{hkl}$.

Outra relação importante a ser considerada é a que estabelece a conexão entre o fator de estrutura $\mathbf{F}_{hkl}$ e a densidade eletrônica $\rho(xyz)$ dentro da célula unitária de um cristal. A partir das equações 5.3 e 5.4, pode-se demonstrar que

$$\rho(xyz) = \frac{1}{V} \sum_{h=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \mathbf{F}_{hkl} \exp[-2\pi i(hx + ky + lz)]\,; \qquad \text{Eq. 5-7}$$

ou ainda que

$$\rho(xyz) = \frac{1}{V} \sum_{h=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \mathrm{F}_{hkl} \exp[i\alpha_{hkl}] \exp[-2\pi i(hx + ky + lz)] \qquad \text{Eq. 5-8}$$

onde V é o volume da célula unitária e x, y e z são as coordenadas atômicas de um ponto qualquer dentro da célula unitária.

A determinação da estrutura tridimensional de uma determinada substância, seja ela uma molécula inorgânica ou mesmo uma proteína, é, em última análise, um dos objetivos finais do trabalho de um cristalógrafo. Se os valores de F_{hkl} e α_{hkl} para um conjunto amplo de reflexões (hkl)

pudessem ser medidos por meio dos experimentos de difração ou de qualquer outra técnica, a densidade eletrônica dentro da célula unitária do cristal seria facilmente calculada pela equação 5.8. A partir da densidade eletrônica, seria possível construir um modelo tridimensional do cristal estudado e, por fim, determinar a estrutura molecular da substância pesquisada. Acontece, entretanto, que, no processo de difração, apenas F_{hkl} (amplitude ou módulo do fator de estrutura) pode ser obtido (equação 5.6). Os valores das fases α_{hkl} foram perdidos no processo. De fato, cabe ao cristalógrafo determinar a fase de cada um dos $\mathbf{F}_{hkl}$. Este problema ficou conhecido como o problema das fases.

Felizmente, existem três métodos básicos para a solução do problema das fases:

i. Métodos diretos: são usados normalmente para a solução de estruturas cristalográficas pequenas (aproximadamente 200 átomos) quando os dados de difração estendem até resolução atômica (superiores ou próxima a 1 Å).
ii. Método de substituição molecular: utiliza um modelo tridimensional já conhecido para resolver a estrutura de interesse, supondo, obviamente, que exista uma grande semelhança entre as estruturas.
iii. Método dos átomos pesados: as fases α_{hkl} são obtidas devido ao espalhamento diferenciado dos raios-X por um pequeno número de átomos de elevado peso atômico ou espalhadores anômalos presentes nos cristais.

Dos três métodos, o terceiro é o que permite, geralmente, a resolução de estruturas cristalográficas inéditas de macromoléculas biológicas. Tendo em vista que esse método, em suas diversas configurações, foi usado para a solução das estruturas de proteínas descritas nesta tese, uma abordagem mais detalhada sobre ele será dada mais à frente.

Antes, porém, é importante salientar alguns aspectos físicos relacionados com as equações introduzidas até o momento. Sabe-se que, quando a freqüência da radiação incidente usada para a coleta dos dados de difração se aproxima de uma das freqüências naturais de oscilação de um sistema elétron-núcleo presente no cristal, um efeito de ressonância passará a ocorrer. Sob essa condição de ressonância, o espalhamento da radiação pela matéria é chamado de espalhamento anômalo. Analiticamente, uma conseqüência imediata desse fenômeno é que o fator de espalhamento atômico, anteriormente descrito pela equação 5.4, passa a contar com mais dois termos não-nulos, como mostra a equação 5.9.

$$\mathbf{f}_j(hkl) = \int_V \rho_j(xyz)\exp[2\pi i(hx+ky+lz)]dv = \mathrm{f}_j^{\mathrm{o}} + \Delta\mathrm{f}_j' + i\mathrm{f}_j'' = \mathrm{f}_j' + \mathrm{if}_j''. \qquad \text{Eq. 5-9}$$

O termo $\mathrm{f}_j^{\mathrm{o}}$, também presente na equação 5.4, é o fator de espalhamento atômico real enquanto $\Delta\mathrm{f}_j'$ e f_j'' são as correções reais e imaginárias para o fator de espalhamento atômico devido ao efeito da dispersão anômala da radiação. De maneira simplificada, pode-se dizer que o valor de $\mathrm{f}_j(hkl)$, de agora em diante $\mathbf{f}_j(hkl)$, poderá assumir valores reais ou imaginários dependendo do tipo do átomo j bem como do comprimento de onda da radiação incidente sobre o átomo.

Se o valor de $\mathbf{f}_j(hkl)$ para todos os átomos em um cristal for um número real, então fica estabelecido o que se convencionou chamar lei de Friedel:

$$\begin{aligned} \mathrm{F}_{hkl} &= \mathrm{F}_{-h-k-l} = \mathrm{F}_{\bar{h}\bar{k}\bar{l}} \\ \alpha_{hkl} &= -\alpha_{-h-k-l} = -\alpha_{\bar{h}\bar{k}\bar{l}} \end{aligned} \qquad \text{Eq. 5-10}$$

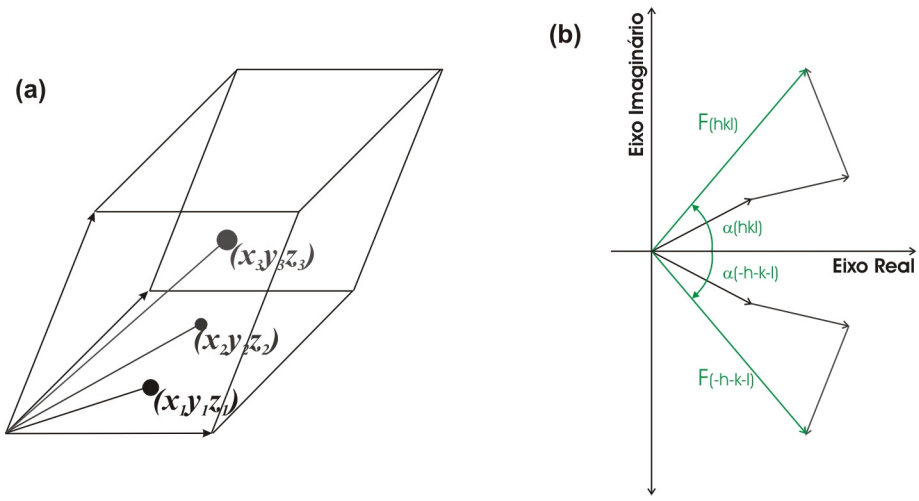


Figura 5-5. Representação para a lei de Friedel no plano de Argand. (a) Representação de um cristal com três átomos com fatores de espalhamento atômico reais na célula unitária. (b) Representação da lei de Friedel para as reflexões (*hkl*) e (*-h-k-l*). Reflexões deste tipo são conhecidas como pares de Friedel.

Isso significa que, na ausência de espalhadores anômalos em um cristal, as amplitudes dos fatores de estrutura ou as intensidades de duas reflexões (hkl) e $(\bar{h}\bar{k}\bar{l})$ [5] serão iguais e que as fases dos fatores

[5] Reflexões com índices negativos podem ser representadas com um traço sobre os índices.

de estrutura para as reflexões serão opostas. Na figura 5.5, a lei de Friedel é representada no plano de Argand.

Durante um experimento típico de difração de raios-X com cristais de proteína, a contribuição anômala para o espalhamento total do cristal é praticamente nula e a lei de Friedel é conservada. Como esses cristais são formados basicamente por átomos de hidrogênio, carbono, nitrogênio, oxigênio e enxofre, pode-se assumir *a priori* que $\mathbf{f}_{hid}(hkl)$, $\mathbf{f}_{car}(hkl)$, $\mathbf{f}_{nit}(hkl)$, $\mathbf{f}_{oxi}(hkl)$ e $\mathbf{f}_{enx}(hkl)$ assumem valores reais e nenhuma contribuição anômala deve ser esperada desses átomos.

5.3.1 Método dos átomos pesados

O método dos átomos pesados requer, inevitavelmente, a presença de átomos de elevado peso atômico ou espalhadores anômalos na estrutura cristalina estudada, uma vez que utiliza o fato de que eles espalham a radiação incidente de forma diferenciada, quando comparados com os elementos tradicionais encontrados em cristais de proteína (hidrogênio, carbono, nitrogênio, oxigênio e enxofre).

Os elementos químicos de elevado peso atômico (zinco, mercúrio, urânio, iodo, césio, ferro, etc) apresentam grande número de elétrons e, por isso, contribuem mais para o espalhamento total da radiação. Contudo, não é muito freqüente encontrar átomos de elevado peso atômico em proteínas, daí a necessidade de se preparar cristais derivados (seção 4.2.1). A preparação dos cristais derivados constitui uma das etapas essenciais sendo, em muitos casos, tão importante quanto a própria cristalização da proteína.

A incorporação de espalhadores anômalos na estrutura cristalina é feita da mesma forma que a dos átomos pesados, e o termo derivado também é usado. Como já mencionado, o sinal anômalo em cristais de proteínas é praticamente nulo. Contudo, como o efeito da dispersão anômala da radiação depende do comprimento de onda dos raios-X usados no experimento, o correto ajuste da radiação incidente permite que determinados elementos químicos, naturalmente presentes na estrutura cristalina ou incorporados via preparação de derivados, tenham seu poder de dispersão anômala aumentado, contribuindo ainda mais para o sinal anômalo total.

Como será visto a seguir, as diferenças observadas em um padrão de difração devido à presença dos átomos pesados ou dos espalhadores anômalos são, em princípio, suficientes para a resolução da estrutura tridimensional de macromoléculas biológicas.

5.3.1.1 Técnicas de substituição isomorfa

As técnicas de substituição isomorfa foram introduzidas na comunidade científica no início da década de 60 por David W. Green, Vernon M. Ingram e Max F. Perutz (GREEN, INGRAM & PERUTZ, 1954) quando, pela primeira vez, a estrutura de uma macromolécula biológica foi resolvida (ROSSMANN, 2001). A técnica de substituição isomorfa simples (*Single Isomorphous Replacement* – SIR) consiste em tentar resolver o problema das fases usando os dados de um cristal nativo e de um cristal derivado. Já a técnica de substituição isomorfa múltipla (*Multiple Isomorphous Replacement* – MIR) utiliza o mesmo cristal nativo, porém mais de um tipo de derivado.

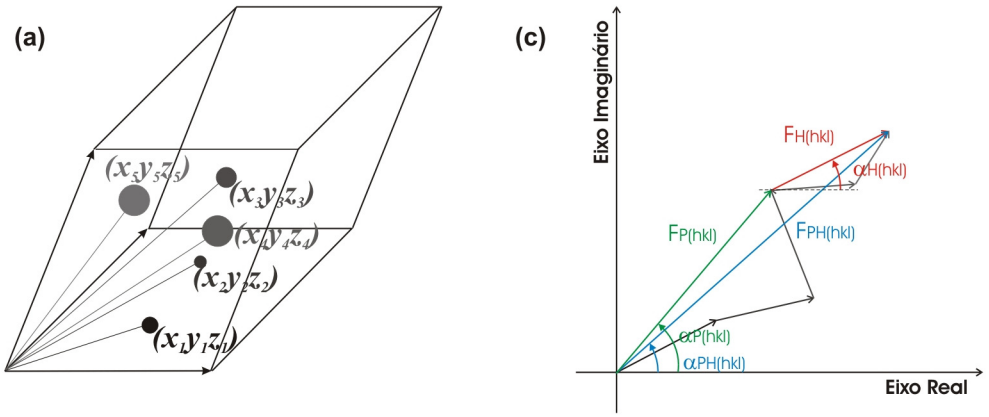
Para que se possa entender a solução para o problema das fases pelas técnicas de substituição isomorfa, suponha-se que a introdução dos átomos pesados em um cristal derivado não tenha causado mudanças bruscas na orientação das moléculas do cristal nem deformação da estrutura cristalina. Assim, pode-se assumir que a diferença entre os fatores de estrutura de um cristal nativo e de um derivado será devida à presença dos átomos pesados. Nesse caso, os dois cristais são ditos isomorfos e a equação 5.3 nos permite escrever que

$$\mathbf{F}_{PH} = \mathbf{F}_P + \mathbf{F}_H \qquad \text{Eq. 5-11}$$

para cada um dos índices h, k e l. $\mathbf{F}_{PH}$, $\mathbf{F}_P$ e $\mathbf{F}_H$ são os fatores de estrutura para um cristal derivado, um cristal nativo e um cristal formado apenas pela subestrutura dos átomos pesados, respectivamente. Os índices h, k e l foram omitidos apenas para facilitar a escrita. A representação da equação 5.11 em um diagrama de Argand pode ser vista na figura 5.6.

Suponha-se que dois conjuntos de dados, um nativo e outro derivado, tenham sido coletados e os valores de F_{PH} e F_P são conhecidos para cada um dos índices h, k e l (equação 5.6). Com a análise da figura 5.6, pode-se observar que a aplicação da regra dos cossenos à F_{PH} permite escrever a seguinte relação:

$$\alpha_P = \alpha_H + \cos^{-1}\left(\frac{(\mathrm{F}_{PH})^2 - (\mathrm{F}_P)^2 - (\mathrm{F}_H)^2}{2(\mathrm{F}_P)(\mathrm{F}_H)}\right) = \alpha_H \pm \alpha'. \qquad \text{Eq. 5-12}$$



(b) Fator de Estrutura

$$\mathbf{F_{PH(hkl)}} = F_{PH(hkl)} \exp[i\alpha_{PH(hkl)}] = \Sigma\, \mathbf{f}_j \exp[2\pi i(hx_j + ky_j + lz_j)]$$

$$= f_1 \exp[2\pi i(hx_1 + ky_1 + lz_1)] + f_2 \exp[2\pi i(hx_2 + ky_2 + lz_2)]$$

$$+ f_3 \exp[2\pi i(hx_3 + ky_3 + lz_3)] + f_4 \exp[2\pi i(hx_4 + ky_4 + lz_4)]$$

$$+ f_5 \exp[2\pi i(hx_5 + ky_5 + lz_5)]$$

$$= F_{P(hkl)} \exp[i\alpha_{P(hkl)}] + F_{H(hkl)} \exp[i\alpha_{H(hkl)}] = \mathbf{F_{P(hkl)}} + \mathbf{F_{H(hkl)}}$$

Figura 5-6. Representação dos fatores de estrutura, mostrando as contribuições de $\mathbf{F}_P$ e $\mathbf{F}_H$ para o valor de $\mathbf{F}_{PH}$. Nesta figura, os átomos 1, 2 e 3 representam a estrutura de uma proteína enquanto os átomos 4 e 5 representam a subestrutura dos átomos pesados. Os átomos 4 e 5 podem ser de elementos químicos diferentes, porém normalmente apenas um tipo de átomo é usado no processo de derivatização.

De acordo com a equação 5.3, observa-se que, se as poucas posições dos átomos pesados pudessem ser determinadas de algum modo, e de fato podem[6], então α_H e F_H poderiam ser calculados e conseqüentemente, via equação 5.12, as fases α_P do cristal nativo poderiam ser obtidas pelo método SIR. Entretanto, o valor de α_P obtido por este método apresenta uma ambigüidade, pois depende de um termo trigonométrico que pode assumir 2 valores distintos. Convém mencionar que, se efeitos de dispersão anômala pudessem ser medidos no cristal derivado, poder-se-ia usar a técnica de substituição isomorfa simples com espalhamento anômalo (*Single Isomorphous Replacement with Anomalous Scattering* – SIRAS) para resolver, sem ambigüidade alguma, o problema das fases (ver mais adiante).

Caso dois ou mais derivados sejam preparados (MIR), o valor de α_P poderia ser calculado sem ambigüidade, pois seriam obtidas duas ou mais relações do tipo da equação 5.12. Caso efeitos anômalos fossem considerados nesta abordagem, a técnica empregada receberia o nome de

[6] Detalhes a este respeito serão dados no final deste capítulo.

substituição isomorfa múltipla com espalhamento anômalo (*Multiple Isomorphous Replacement with Anomalous Scattering* – MIRAS).

Outra maneira de verificar a ambigüidade das fases no procedimento SIR e a não-ambigüidade das fases obtidas por MIR consiste na análise dos círculos de fase de Harker (HARKER, 1956), mostrados na figura 5.7. A partir da origem de um diagrama de Argand, desenha-se $-\mathbf{F}_{H1}$, ou seja, o negativo do fator de estrutura para índices h, k e l quaisquer, obtido apenas com a subestrutura de átomos pesados do cristal derivado 1. Ainda no mesmo diagrama, um círculo de raio F_P centrado na origem e outro círculo de raio F_{PH1} centrado na ponta de $-\mathbf{F}_{H1}$ são desenhados. F_P e F_{PH1} são as amplitudes dos fatores de estrutura com mesmos índices h, k e l do cristal nativo e derivado 1, respectivamente. Os locais onde as duas circunferências se cruzam definem os prováveis valores de fase para α_P. Pode-se ainda reproduzir, no mesmo diagrama, o efeito de um segundo derivado (derivado 2) de maneira análoga ao primeiro. Desta vez, o local onde as três circunferências se cruzarem irá definir o valor mais provável para a fase α_P.

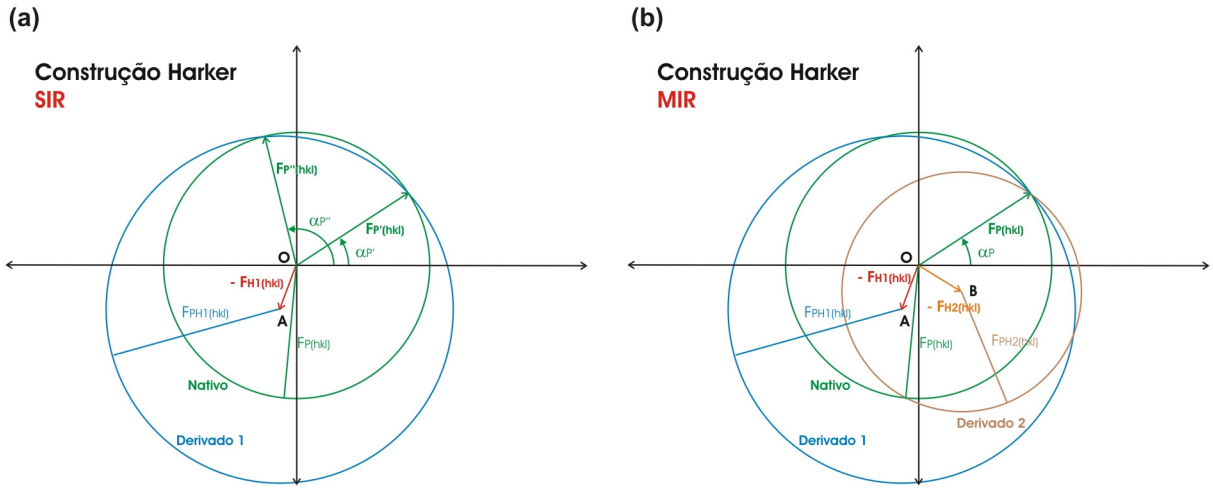


Figura 5-7. Construção dos círculos de Harker para os procedimentos (a) SIR e (b) MIR. As fases $\alpha_{P'}$ e $\alpha_{P''}$ estão localizadas simetricamente em relação a $\mathbf{F}_H$.

5.3.1.2 Técnicas que utilizam o sinal anômalo

Para atingir os objetivos propostos com esta seção, é conveniente lembrar que o espalhamento anômalo da radiação pode ser caracterizado pelo surgimento de um termo complexo não-nulo na expressão do fator de espalhamento atômico de um átomo, como mostra a equação 5.9.

Suponha-se, portanto, que um conjunto de dados de um cristal derivado, preparado com um único tipo de átomo pesado, tenha sido coletado. Além disso, o comprimento de onda da radiação incidente foi ajustado para aumentar o sinal anômalo apenas desses átomos pesados. Mesmo não tendo coletado um conjunto de dados de um cristal nativo, a equação 5.11 ainda é válida. Como $\mathbf{F}_P$ só levaria em conta os P átomos leves da estrutura nativa, tem-se que

$$\mathbf{F}_P(hkl) = \mathbf{F}_P(+) = \sum_{j=1}^{P} \mathrm{f}_j^{\mathrm{o}} \exp[2\pi i(hx_j + ky_j + lz_j)] \text{ para os índices } h\text{, } k \text{ e } l \qquad \text{Eq. 5-13}$$

e

$$\mathbf{F}_P(\bar{h}\bar{k}\bar{l}) = \mathbf{F}_P(-) = \sum_{j=1}^{P} \mathrm{f}_j^{\mathrm{o}} \exp[-2\pi i(hx_j + ky_j + lz_j)] \text{ para os índices } \bar{h}\text{, } \bar{k} \text{ e } \bar{l}\text{,} \qquad \text{Eq. 5-14}$$

de acordo com a lei de Friedel. Já a contribuição dos H átomos pesados deve ser escrita de forma similar, porém, levando-se em conta as novas contribuições $\Delta\mathrm{f}'_j$ e if''_j para o fator de espalhamento atômico dos átomos. Assim, tem-se que

$$\mathbf{F}'_H(+) = \sum_{j=1}^{H} \mathrm{f}'_j \exp[2\pi i(hx_j + ky_j + lz_j)],\ \mathbf{F}''_H(+) = \sum_{j=1}^{H} \mathrm{f}''_j \exp[2\pi i(hx_j + ky_j + lz_j)] \qquad \text{Eq. 5-15}$$

e

$$\mathbf{F}'_H(-) = \sum_{j=1}^{H} \mathrm{f}'_j \exp[-2\pi i(hx_j + ky_j + lz_j)],\ \mathbf{F}''_H(-) = \sum_{j=1}^{H} \mathrm{f}''_j \exp[-2\pi i(hx_j + ky_j + lz_j)]. \qquad \text{Eq. 5-16}$$

Combinando as equações 5.11 com as equações 5.13 até 5.16 tem-se que

$$\mathbf{F}_{PH}(+) = \mathbf{F}_P(+) + \mathbf{F}'_H(+) + i\mathbf{F}''_H(+) \qquad \text{Eq. 5-17}$$

e

$$\mathbf{F}_{PH}(-) = \mathbf{F}_P(-) + \mathbf{F}'_H(-) + i\mathbf{F}''_H(-). \qquad \text{Eq. 5-18}$$

A representação dessas equações em um diagrama de Argand é mostrada na figura 5.8. Como os átomos pesados são de um mesmo tipo, $\mathbf{F}'_H$ e $\mathbf{F}''_H$ apresentam uma mesma fase; logo $i\mathbf{F}''_H$ será ortogonal a $\mathbf{F}'_H$.

Uma das primeiras conseqüências da existência do efeito anômalo é a violação da lei de Friedel, ou seja, as reflexões com índices (hkl) e $(\bar{h}\bar{k}\bar{l})$ já não apresentam, obrigatoriamente, as mesmas intensidades, pois

$$\mathrm{F}_{PH}(+) = \mathrm{F}_{PH}(hkl) \neq \mathrm{F}_{PH}(\bar{h}\bar{k}\bar{l}) = \mathrm{F}_{PH}(-)\,. \qquad \text{Eq. 5-19}$$

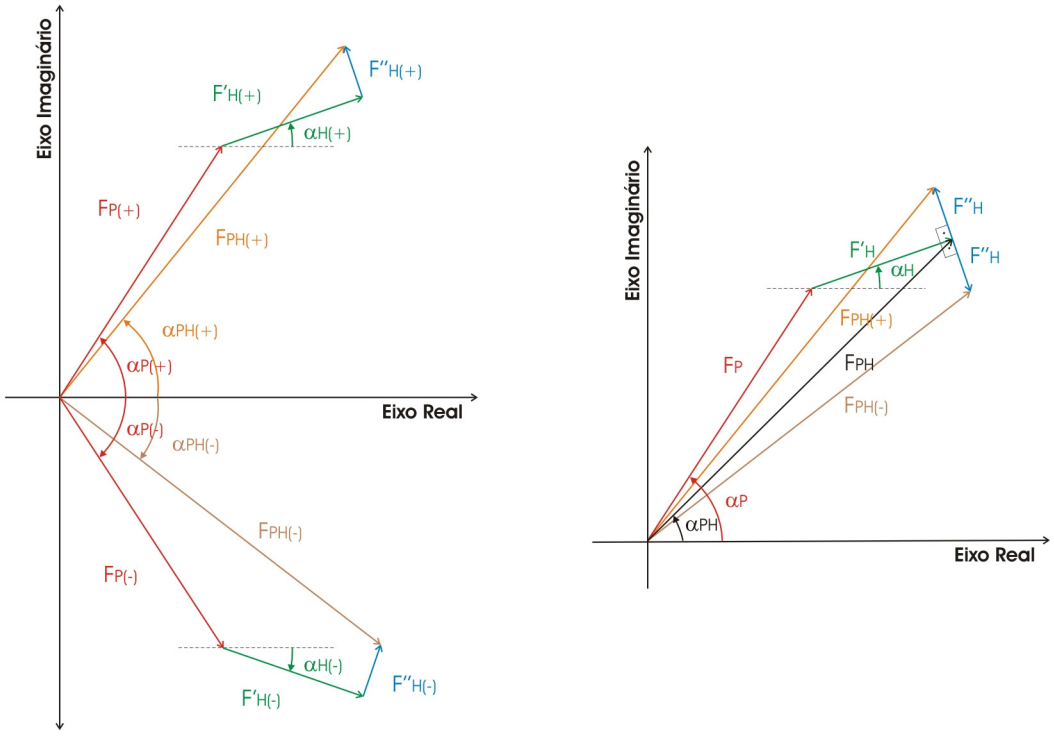


Figura 5-8. Representação dos fatores de estrutura, mostrando as contribuições de $\mathbf{F}_P(+)$, $\mathbf{F'}_H(+)$ e $\mathbf{F''}_H(+)$ para o valor de $\mathbf{F}_{PH}(+)$. Idem para $\mathbf{F}_{PH}(-)$. Na figura da direita $\mathbf{F}_{PH}(-)$ foi espelhado em relação ao eixo real.

É possível ver pela figura 5.8 que a aplicação da lei dos cossenos à $\mathrm{F}_{PH}(+)$ e à $\mathrm{F}_{PH}(-)$ resulta nas equações

$$\left(\mathrm{F}_{PH}(+)\right)^2 = \left(\mathrm{F}_{PH}\right)^2 + \left(\mathrm{F}''_H\right)^2 - 2\left(\mathrm{F}_{PH}\right)\left(\mathrm{F}''_H\right)\cos\left(\alpha_{PH} - \alpha_H + \frac{\pi}{2}\right) \qquad \text{Eq. 5-20}$$

e

$$\left(F_{PH}(-)\right)^2 = \left(F_{PH}\right)^2 + \left(F''_H\right)^2 - 2\left(F_{PH}\right)\left(F''_H\right)\cos\left(\alpha_{PH} - \alpha_H - \frac{\pi}{2}\right), \qquad \text{Eq. 5-21}$$

que, combinadas, determinam a seguinte expressão:

$$\left(F_{PH}(+)\right)^2 - \left(F_{PH}(-)\right)^2 = 4\left(F_{PH}\right)\left(F''_H\right)\cos\left(\alpha_{PH} - \alpha_H - \frac{\pi}{2}\right). \qquad \text{Eq. 5-22}$$

Após o rearranjo dos termos da equação 5.22, é possível verificar que os valores das fases α_{PH} para o cristal derivado serão dados, conforme a equação 5.23.

$$\alpha_{PH} = \alpha_H + \frac{\pi}{2} + \cos^{-1}\left(\frac{\left[\left(F_{PH}(+)\right) - \left(F_{PH}(-)\right)\right]}{2\left(F''_H\right)}\right). \qquad \text{Eq. 5-23}$$

Note-se que, na equação 5.23, não há nenhum termo relacionado aos fatores de estrutura de um cristal nativo, mas apenas do cristal derivado. Mais uma vez, se as posições atômicas dos espalhadores anômalos pudessem ser determinadas, os termos α_H e F''_H poderiam ser calculados e os valores das fases α_{PH} para os fatores de estrutura do cristal derivado seriam obtidos.

A técnica que utiliza apenas um cristal com sinal anômalo não-desprezível para a obtenção das fases dos fatores de estrutura recebe o nome de difração anômala simples (*Single Anomalous Diffraction* – SAD). De maneira similar à técnica SIR, as fases obtidas por SAD apresentam uma ambigüidade devido a um termo trigonométrico que pode assumir dois valores opostos. No entanto, a ambigüidade pode ser solucionada com a coleta de dados de um cristal nativo, sem espalhadores anômalos. A técnica recebe, então, o nome de substituição isomorfa simples com espalhamento anômalo (*Single Isomorphous Replacement with Anomalous Scattering* – SIRAS). Analogamente, a utilização de um cristal nativo e mais de um cristal derivado com sinal anômalo recebe o nome de substituição isomorfa múltipla com espalhamento anômalo (*Multiple Isomorphous Replacement with Anomalous Scattering* – MIRAS).

A representação dessas técnicas nos círculos de fase de Harker é obtida de maneira similar àquela adotada para as técnicas SIR e MIR. Na figura 5.9, as técnicas SAD e SIRAS estão representadas.

Uma outra técnica conhecida como difração anômala com múltiplos comprimentos de onda (*Multiwavelength Anomalous Diffraction* – MAD) também utiliza o efeito anômalo para obter as fases necessárias para a construção de um mapa de densidade eletrônica. A técnica, introduzida há quase 20 anos por Wayne A. Hendrickson (HENDRICKSON, 1985, 1991), é muito utilizada em todo o mundo e consiste basicamente na coleta de três ou quatro conjuntos de dados de difração de um único cristal utilizando raios-X com diferentes comprimentos de onda. Devido à dependência do espalhamento anômalo da radiação com o comprimento de onda dos raios-X incidentes, é possível obter informações diferentes de cada um dos conjuntos. A combinação das informações permite, ao final da coleta dos dados, obter as fases dos fatores de estrutura para o cristal estudado.

A técnica MAD requer a utilização de fontes de radiação com alta resolução energética ($\Delta\lambda/\lambda < 10^{-3}$) e todo um aparato experimental que permita a mudança do comprimento de onda da radiação ao longo do experimento. Como a linha de luz CPr do LNLS não foi desenvolvida especialmente para esta técnica, nenhum experimento desse tipo foi ainda realizado no país. Em vista disso, uma abordagem mais detalhada estaria além dos objetivos deste trabalho e, por isso, não será apresentada aqui.

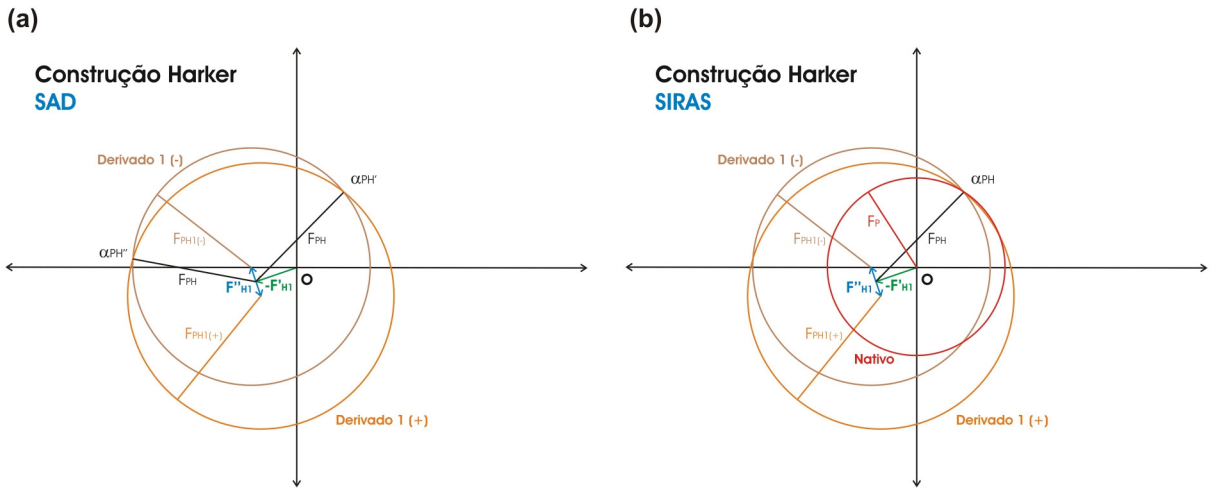


Figura 5-9. Construção dos círculos de Harker para os procedimentos (a) SAD e (b) SIRAS. As fases $\alpha_{PH'}$ e $\alpha_{PH''}$ estão localizadas simetricamente em relação à **F"**$_H$.

Antes de terminar esta seção, é conveniente esclarecer um ponto que envolve todas as técnicas de átomos pesados apresentadas neste capítulo: SIR, MIR, SAD, SIRAS, MIRAS e MAD. Em todas elas, a determinação das fases dos fatores de estrutura passa, obrigatoriamente, pela determinação das coordenadas atômicas dos átomos pesados ou dos espalhadores anômalos. Esta

etapa mostra que o problema de resolver a estrutura total de um cristal foi, inicialmente, reduzido a resolver a subestrutura dos átomos pesados ou espalhadores anômalos do cristal. Como o número desses átomos é muito menor que o número total de átomos de um cristal, o problema pode ser resolvido mais facilmente. Atualmente, diversos algoritmos e formulações são utilizados em muitos programas para resolver esse problema (BLESSING & SMITH, 1999; WEEKS & MILLER, 1999; SHELDRICK, 1998). A teoria envolvida em cada uma dessas formulações pode ser encontrada na literatura e, por isso, não será abordada aqui. Deve-se, entretanto, ter em mente que essas formulações utilizam apenas as diferenças observadas nos padrões de difração devido à introdução dos átomos pesados para a correta identificação dos sítios atômicos destes átomos.

5.4 Considerações finais

A obtenção das fases dos fatores de estrutura é, sem sombra de dúvidas, uma das etapas fundamentais de todo o processo de determinação da estrutura cristalográfica de uma macromolécula. Entretanto, outras etapas precisam ser vencidas para que se possa chegar de fato ao modelo tridimensional de uma macromolécula.

Inicialmente, o mapa de densidade eletrônica obtido experimentalmente deve ser interpretado e, na medida do possível, um modelo tridimensional deve ser construído de forma a explicar a densidade eletrônica observada dentro do cristal. Entretanto, é muito comum que os mapas de densidade eletrônica experimentais apresentem falhas e não possam ser interpretados visualmente nem por algoritmos computacionais. Muitas falhas estão associadas diretamente aos erros experimentais, fato desconsiderado até o momento em todos os cálculos. Sabe-se que toda medida experimental apresenta um erro e quando ela é usada para o cálculo de outras grandezas, o erro se propaga no cálculo influenciando o resultado final. Assim, a qualidade do conjunto de dados de difração e de todas as grandezas obtidas a partir desse conjunto deve ser avaliada ao longo das etapas que envolvem a determinação estrutural de uma macromolécula biológica para que o objetivo possa ser cumprido com êxito e o resultado seja confiável.

Em relação ao conjunto de dados de difração, além dos erros experimentais associados à cada intensidade (hkl), outros parâmetros são usados para quantificar a qualidade das medições. Dentre eles, destacam-se:

i. $\langle I_{hkl}/\sigma(I_{hkl})\rangle$. O valor médio, por faixa de resolução, da razão entre as intensidades das reflexões (hkl) e seus respectivos erros. É usado normalmente para definir a resolução máxima de um conjunto de dados.

ii. As resoluções mínima e máxima de um conjunto de dados. São definidas como os maiores e menores valores de d_{hkl}, respectivamente, associados com as reflexões (hkl) encontradas em todo o conjunto.

iii. Número de reflexões totais. Número de todas as reflexões medidas e aceitas na faixa de resolução estabelecida para aquele conjunto de dados.

iv. Completeza. Porcentagem das reflexões medidas num total de reflexões teoricamente esperadas para a faixa de resolução estabelecida.

v. Número de reflexões únicas. Número de reflexões distintas em um conjunto de dados. As reflexões idênticas ou relacionadas pela simetria do cristal são consideradas reflexões equivalentes e teoricamente devem possuir a mesma intensidade.

vi. Redundância. Razão entre o número de reflexões totais e o número de reflexões únicas, por faixa de resolução.

vii. $$R_{merge}(I) = \frac{\sum_{hkl}\sum_{i}\left|I_i(hkl) - \langle I(hkl)\rangle\right|}{\sum_{hkl}\sum_{i} I_i(hkl)}.$$

Mede a consistência dos dados em relação às intensidades das reflexões idênticas ou simetricamente equivalentes. Juntamente com o primeiro, esse parâmetro é usado para definir a máxima resolução dos dados coletados.

Uma vez que os conjuntos de dados são coletados, a determinação das fases pode ser efetuada. Com a análise das equações 5.12 e 5.23, observa-se que as fases α_P ou α_{PH} só poderão ser determinadas com acerto se as coordenadas dos átomos pesados ou espalhadores anômalos e os valores de F_{PH} e F_P ou $F_{PH}(+)$ e $F_{PH}(-)$ forem determinados com precisão. Na verdade, tanto a obtenção correta das fases quanto das coordenadas atômicas dos átomos pesados ou espalhadores anômalos depende da precisão e do erro associado às quantidades

$$\Delta F^{ISO} = F_{PH} - F_P \qquad \text{Eq. 5-24}$$

e

$$\Delta F^{ANO} = F_{PH}(+) - F_{PH}(-) \qquad \text{Eq. 5-25}$$

que podem ser obtidas dos experimentos de difração (os índices h, k e l foram omitidos). ΔF^{ISO}, conhecido como diferença isomorfa, é obtido a partir de dois conjuntos de dados (um nativo e um

derivado). Esta quantidade reflete as diferenças dos fatores de estrutura mediante a incorporação dos átomos pesados. Já ΔF^{ANO}, conhecido como diferença anômala ou diferença Bijvoet, é calculado com um conjunto de dados apenas, pois reflete a quebra da lei de Friedel devido ao espalhamento anômalo da radiação por átomos específicos.

Com as equações 5.24 e 5.25, pode-se definir o que se convencionou chamar de sinais isomorfo e anômalo.

Sinal isomorfo: $$\frac{\left\langle\left|\Delta F^{ISO}\right|\right\rangle}{\left\langle|F|\right\rangle} \approx \frac{f'_A}{Z_{eff}}\sqrt{\frac{N_A}{2N}}\,. \qquad \text{Eq. 5-26}$$

Sinal anômalo: $$\frac{\left\langle\left|\Delta F^{ANO}\right|\right\rangle}{\left\langle|F|\right\rangle} \approx \frac{2f''_A}{Z_{eff}}\sqrt{\frac{N_A}{2N}}\,. \qquad \text{Eq. 5-27}$$

Os termos à direita nas equações 5.26 e 5.27 (HENDRICKSON, SMITH & SHERIFF, 1985) permitem estimar, teoricamente, qual o sinal isomorfo e anômalo, respectivamente de uma determinada estrutura cristalina com N átomos totais (exceto hidrogênio) e N_A átomos pesados ou espalhadores anômalos. Z_{eff} é o número atômico médio efetivo dos átomos da estrutura e, no caso de proteínas, assume o valor aproximado de 6,7. Os termos f'_A e f''_A representam as contribuições real e imaginária para o fator de espalhamento atômico dos átomos pesados ou espalhadores anômalos.

A análise dessas quantidades a partir de um conjunto de dados permite tanto prever a presença ou não de átomos pesados ou espalhadores anômalos quanto estimar o sucesso na obtenção das fases. Naturalmente, os valores dos sinais isomorfo e anômalo dependem da substância cristalizada, do tipo e quantidade de átomos pesados ou espalhadores anômalos e do comprimento de onda da radiação incidente. Normalmente, esses valores ficam em torno de 10 – 25% para o sinal isomorfo e 1 – 8% para o sinal anômalo.

Retornando à questão dos erros experimentais, é possível ver, na figura 5.10 a, as principais fontes de erros associadas ao cálculo das fases dos fatores de estrutura pelos métodos de substituição isomorfa. Os termos σ representam os erros das intensidades atribuídos à imprecisão das medidas; os termos η indicam os erros associados às posições atômicas, ocupâncias e fatores de temperatura

dos átomos pesados, e os termos μ indicam a falta de isomorfismo entre os conjuntos de dados devido à incorporação dos átomos pesados.

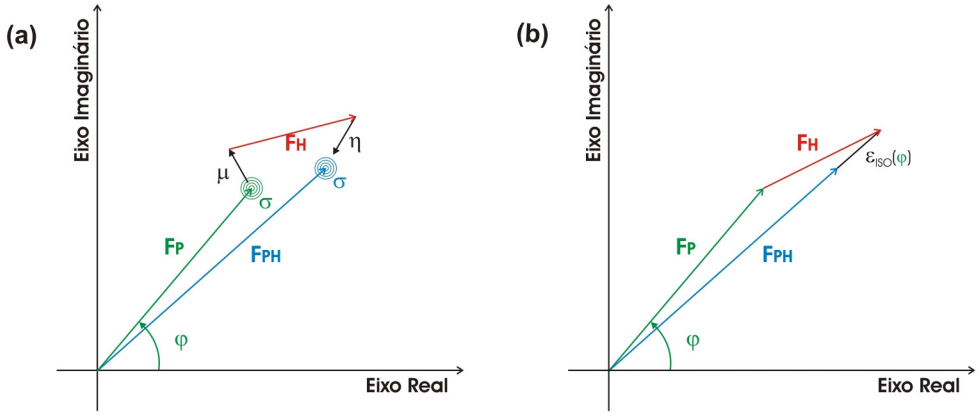


Figura 5-10. Fontes de erros no método de substituição isomorfa. (a) Representação das principais fontes de erros σ, η e μ. (b) Simplificação para as principais fontes de erros.

O tratamento dos erros nessa situação pode ser simplificado enormemente se for assumido que F_P e $\mathbf{F}_H$ são conhecidos com precisão e que todo o erro ε_{ISO} reside no valor medido de F_{PH} como mostra a figura 5.10 b (BLOW & CRICK, 1959). Assim, para uma fase arbitrária φ, o triângulo definido por $\mathbf{F}_{PH}$, $\mathbf{F}_P$ e $\mathbf{F}_H$ não fechará completamente. Se $\mathbf{F}_C$ for definido como a soma de $\mathbf{F}_H$ e $F_P \exp(i\varphi)$, então o erro ε_{ISO} do triângulo de fase (*lack of closure*) será dado pela equação 5.28.

$$\varepsilon_{ISO}(\varphi) = \left|\mathbf{F}_H + F_P \exp(i\varphi)\right| - F_{PH} = F_C - F_{PH} \,. \qquad \text{Eq. 5-28}$$

Se E_{ISO} for definido como o erro associado com essa medida (feita em intervalos de alguns graus), e a distribuição dos erros for assumida gaussiana, então, é possível encontrar, para cada fator de estrutura (hkl), uma função $P_{ISO}(\varphi)$ que descreve a probabilidade de a fase φ ser a fase correta. Esta função $P_{ISO}(\varphi)$ é dada pela equação 5.29.

$$P_{ISO}(\varphi) = N \exp\left(-\varepsilon_{ISO}^2(\varphi)/2E_{ISO}^2\right) \qquad \text{Eq. 5-29}$$

onde N é um fator de normalização. Um exemplo da equação 5.29 pode ser visto na figura 5.11.

Caso o sinal anômalo medido tenha sido usado para o cálculo das fases, uma expressão similar à equação 5.29 pode ser usada para definir a função de probabilidade $P_{ANO}(\varphi)$. Esta função apresenta, contudo, termos $\varepsilon_{ANO}(\varphi)$ e E_{ANO} calculados com base nas medidas de $F_{PH}(+)$ e $F_{PH}(-)$ (MATTHEWS, 1966; HENDRICKSON, 1979; TERWILLIGER & EISENBERG, 1987).

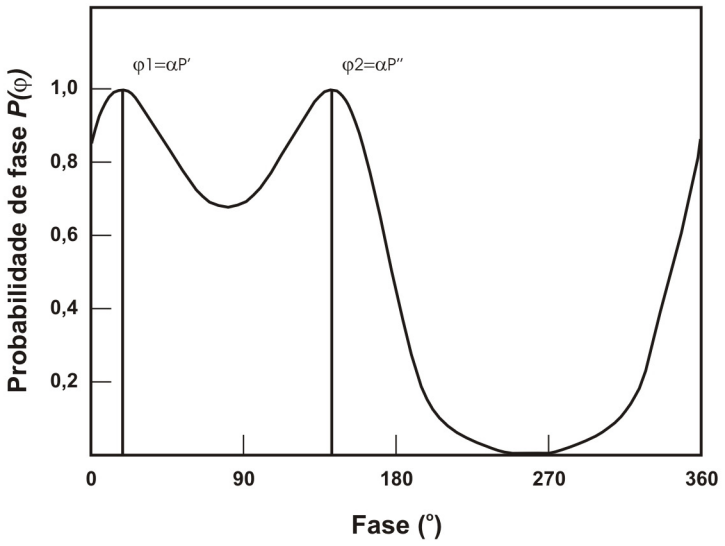


Figura 5-11. Função de probabilidade de fase não-normalizada para um fator de estrutura (*hkl*) qualquer pelo método de substituição isomorfa simples. Neste método duas fases, igualmente prováveis, podem ser obtidas.

Quando as fases dos fatores de estrutura puderem ser obtidas tanto a partir do sinal isomorfo quanto do sinal anômalo, ou ainda com o uso dos dados de difração de outros cristais, a função total de probabilidade de fase $P_{TOTAL}(\varphi)$ pode ser obtida pela multiplicação de cada uma das funções, como mostrado na equação 5.30.

$$P_{TOTAL}(\varphi) = N\left(\prod P_{ISO}(\varphi)\right)\left(\prod P_{ANO}(\varphi)\right). \qquad \text{Eq. 5-30}$$

A abordagem adotada para o tratamento do erros durante o cálculo das fases permite definir um parâmetro conhecido como figura de mérito m que mede a confiança com que uma fase foi determinada. Esse parâmetro, definido pela equação 5.31, assume valores entre 0 e 1 e é utilizado com freqüência para caracterizar o processo de obtenção das fases. Teoricamente, figuras de mérito próximas de 1 indicam que as fases φ apresentam erro próximo de 0^o.

$$m = \frac{\int_0^{2\pi} P(\varphi)\exp(i\varphi)d\varphi}{\int_0^{2\pi} P(\varphi)d\varphi} \qquad \text{Eq. 5-31}$$

Uma vez que um mapa de densidade eletrônica é calculado com as fases experimentais, a interpretação dele pode começar a ser feita. Contudo, esse passo pode ser afetado por alguns fatores dos quais destacam-se:

i. Confiabilidade dos termos F_{hkl} e α_{hkl}. A correta determinação de F_{hkl} e de α_{hkl} é fundamental para o cálculo da densidade eletrônica. Erros, principalmente em α_{hkl}, resultam em mapas de densidade eletrônica completamente não-interpretáveis.
ii. Completeza do conjunto de dados. Conjuntos de alta completeza resultarão, em princípio, em mapas de densidade eletrônica completos e sem falhas.
iii. Resolução do conjunto de dados. Conjuntos de dados provenientes de cristais que difratam até alta resolução apresentam mais reflexões (*hkl*) que um conjunto, do mesmo cristal, com menor resolução. Assim, segundo a equação 5.8, a presença de mais termos (F_{hkl} e α_{hkl}) contribuirá para um detalhamento maior à função de densidade eletrônica.

Em muitos casos, alguns procedimentos matemáticos conhecidos como "modificação de densidade eletrônica" são utilizados para melhorar a qualidade dos mapas obtidos experimentalmente. Esses procedimentos modificam os valores das fases em torno de seu valor experimental α_{hkl}, permitindo a identificação da estrutura atômica no mapa de densidade eletrônica. Nas figuras 5.12 a e 5.12 b, dois mapas de densidade eletrônica são mostrados: um calculado com as fases experimentais e o outro obtido após o procedimento de modificação de densidade eletrônica. Os algoritmos que permitem a modificação da densidade eletrônica estão implementados em vários programas utilizados em cristalografia de proteínas e, basicamente, o funcionamento deles consiste em supor que existe uma concentração de densidade eletrônica na região ocupada pelas macromoléculas, enquanto na região ocupada pelo solvente a densidade eletrônica é praticamente nula (LESLIE, 1987; ABRAHAMS, 1997). Essa suposição encontra suporte no fato de que, na grande maioria dos casos, algo em torno de 50% do volume de um cristal de proteína é composto

por moléculas de água desordenadas que não contribuem para o espalhamento coerente da radiação e, portanto, não são vistas nos mapas de densidade eletrônica.

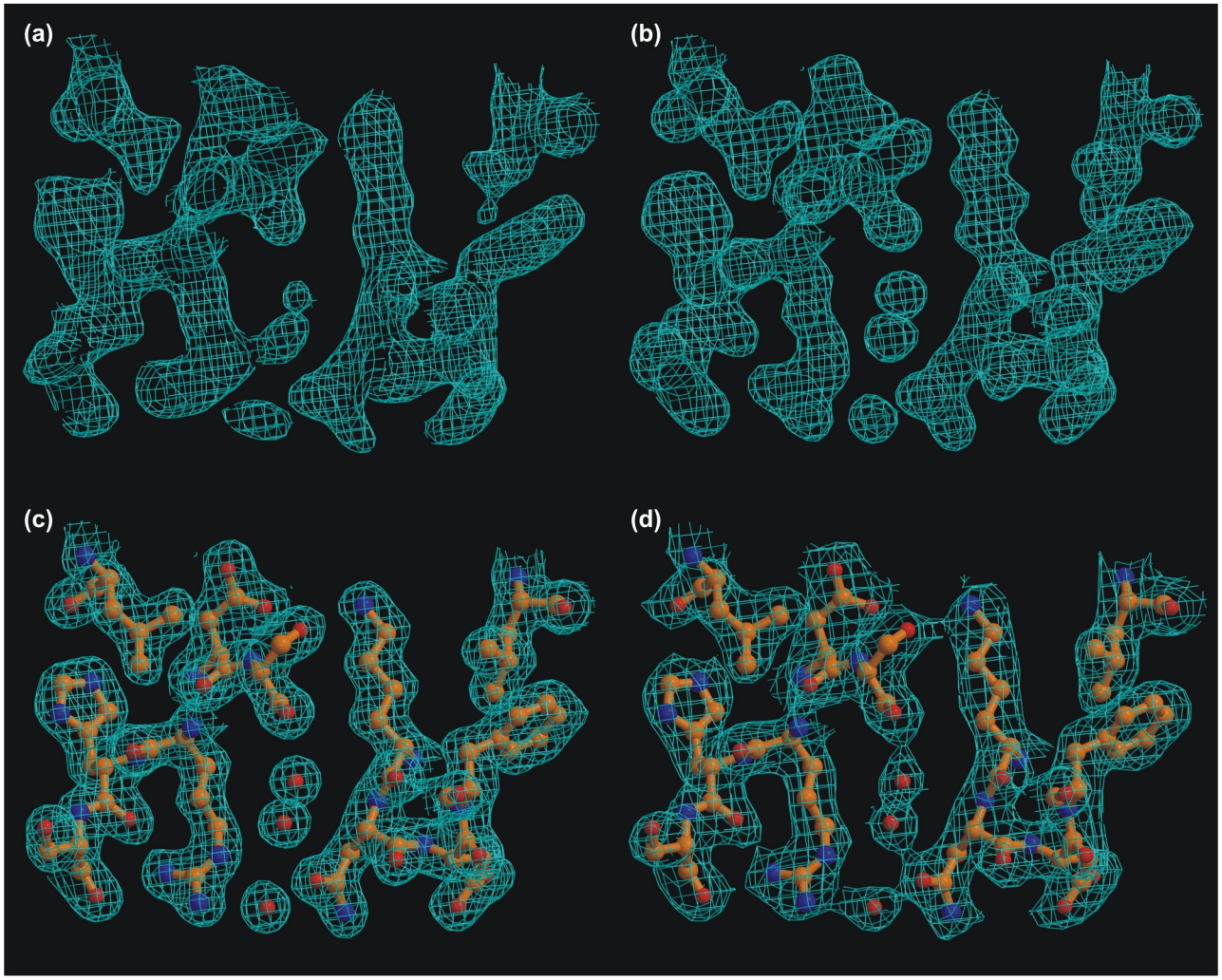


Figura 5-12. Interpretação dos mapas de densidade eletrônica. (a) Mapa de densidade eletrônica calculado com fases MIRAS experimentais. (b) O mesmo mapa anterior após os procedimentos de modificação de densidade eletrônica. (c) Construção de um modelo tridimensional em um mapa de densidade eletrônica com resolução de 1.9 Å. (d) O equivalente do anterior em um mapa de 2.5 Å de resolução. Átomos de carbono, oxigênio e nitrogênio estão representados por esferas laranjas, vermelhas e azuis, respectivamente. Os átomos de hidrogênio não estão representados nesta figura. Normalmente, em cristalografia de proteínas, estes átomos não podem ser identificados.

Dependendo da qualidade dos mapas de densidade eletrônica, pode ser feita interpretação automática ou manual. Um modelo tridimensional pode ser então construído, preenchendo as regiões de mais alta densidade eletrônica com os átomos do modelo. Entretanto, a escolha dos átomos deve obedecer à estrutura primária previamente determinada por estudos bioquímicos.

Nas figuras 5.12 c e 5.12 d são mostrados dois mapas de densidade eletrônica nos quais um modelo tridimensional pôde ser construído. A ausência de informação sobre a estrutura primária de

uma proteína ou sobre a composição química de uma molécula poderia levar ao erro de assinalar um determinado átomo na posição ocupada por outro tipo de átomo. Em conjuntos de dados de alta resolução, esta identificação pode ser feita de forma correta já que, além de se poder identificar com precisão às posições atômicas dos átomos de hidrogênio, pode-se ter uma estimativa do valor real da densidade eletrônica em um determinado local da célula unitária.

O modelo tridimensional inicialmente construído, normalmente apresenta erros. Em uma etapa conhecida como "refinamento", o modelo é ajustado, ou seja, tanto as coordenadas atômicas quanto os fatores de temperatura e a ocupância de cada um dos átomos são refinados em torno das posições iniciais procurando a obtenção de um modelo 3D ideal. Este modelo deve ser entendido como aquele que melhor representa os dados experimentais, ou seja, um modelo que permita com que os fatores de estrutura calculados a partir da equação 5.5 sejam os mais próximos dos fatores de estrutura medidos com o experimento. Esta tarefa pode ser atingida com a minimização da função

$$R_{factor} = \frac{\sum_{hkl} \left| |\mathrm{F}_{obs}| - k|\mathrm{F}_{calc}| \right|}{\sum_{hkl} |\mathrm{F}_{obs}|} \times 100\%, \qquad \text{Eq. 5-32}$$

onde k é um termo de escala entre as amplitudes dos fatores de estrutura medidos no experimento de difração (F_{obs}) e os mesmos fatores de estrutura calculados a partir do modelo (F_{calc}). Normalmente, estruturas macromoleculares refinadas apresentam R_{factor} inferior a 20%.

Entretanto, certos parâmetros estereoquímicos como as distâncias e os ângulos entre as ligações químicas, a planeza de certos grupos de átomos e os contatos de van der Waals para átomos não-ligados também devem ser levados em conta durante o processo de refinamento para que os modelos tridimensionais possam fazer algum sentido físico e biológico! Assim, a função que deve ser minimizada (função-alvo) durante o refinamento deve incluir não só o termo cristalográfico (equação 5.32), mas termos que expressem as restrições estereoquímicas apresentadas acima. Diversas variações para a formulação da função-alvo são encontradas na literatura (JACK & LEVITT, 1978; SUSSMAN, 1985; BRÜNGER, KURIYAN & KARPLUS, 1987; TRONRUD, EYCK & MATTHEWS, 1987; TRONRUD, 1992); contudo, não serão apresentadas nesta tese.

6 Resultados

Ao longo dos últimos quatro anos, todo o trabalho desenvolvido nesta tese teve como objetivo principal a obtenção dos conhecimentos necessários para a resolução de estruturas cristalográficas inéditas de macromoléculas biológicas utilizando as facilidades da linha de luz CPr do Laboratório Nacional de Luz Síncrotron. Neste período, diversos experimentos foram realizados, o que permitiu encontrar excelentes condições para a coleta dos dados de difração e conseqüente determinação estrutural. Além disso, o projeto contribuiu para que as estruturas de algumas macromoléculas biológicas inéditas fossem resolvidas em menos de três anos de pesquisa. As estruturas resolvidas neste projeto estão dentre as primeiras em todo o Brasil. Estima-se que não mais do que duas estruturas cristalográficas inéditas de macromoléculas biológicas tenham sido resolvidas por grupos de pesquisa brasileiros, anteriormente. Nesta seção, serão mostrados em uma ordem cronológica, sempre que possível, cada uma das etapas alcançadas no projeto, ressaltando as mais significativas e de importância fundamental para o contexto da tese. Não serão descritos, portanto, todos os passos adotados para a resolução de cada uma das estruturas cristalinas nem os detalhes biológicos específicos sobre cada uma dessas macromoléculas. Em alguns casos, apenas uma sucinta descrição dos fatos será apresentada já que alguns resultados ainda estão sendo analisados. Alguns dos trabalhos já publicados ou mesmo os já concluídos e em processo de

submissão para revistas internacionais, relacionados diretamente com a pesquisa sugerida no projeto de tese, serão incorporados ao texto permitindo uma visão mais completa do assunto.

6.1 Experimentos iniciais com uma proteína teste

Com o objetivo de testar as facilidades da linha de luz CPr do Laboratório Nacional de Luz Síncrotron em relação à sua capacidade de gerar dados de difração de raios-X confiáveis, capazes de permitir a solução de uma estrutura inédita de uma macromolécula, foram coletados, inicialmente, os dados de difração de cristais de lisozima, uma proteína com uma única cadeia polipeptídica com 129 resíduos de aminoácidos.

A lisozima (*hen egg-white lysozyme,* HEWL), uma das proteínas mais conhecidas em todo o mundo, é usada normalmente em estudo-teste, pois é facilmente adquirida e sua cristalização é rápida e eficiente. Além disso, seus cristais apresentam alto poder de difração, atingindo facilmente 2,0 Å de resolução. A estrutura tridimensional da HEWL já é conhecida desde 1962 (BLAKE *et al*., 2001), contudo, este experimento permitiu verificar a possibilidade de resolver esta estrutura pelo método dos átomos pesados, simulando assim a determinação de uma estrutura inédita.

Iniciou-se o trabalho com a cristalização desta macromolécula seguindo protocolos conhecidos. Todos os cristais de lisozima foram preparados pelo método da gota suspensa (seção 4.1) utilizando proteína purificada comprada da empresa SIGMA. A gota de cristalização (2 – 4 μl) consistia de 20 – 30 mg/ml de proteína, 0,5 M de NaCl e 0,025 M de tampão de acetato de sódio em pH 4,5. Já a solução do poço era composta por 1,0 M de NaCl e 0,050 M de tampão de acetato de sódio. As placas de cristalização foram mantidas a uma temperatura de 18 ^{o}C por alguns dias até que os cristais fossem formados.

Conforme mencionado na seção 5.3.1, o método dos átomos pesados requer, inevitavelmente, a presença de átomos de elevado peso atômico ou espalhadores anômalos na estrutura cristalográfica que se pretende resolver. Já é do conhecimento dos cristalógrafos que, infelizmente, a HEWL apresenta em sua estrutura apenas átomos de hidrogênio, carbono, oxigênio, nitrogênio e enxofre. Assim, para se utilizar o método dos átomos pesados era necessário preparar pelo menos um cristal derivado (seção 4.2.1) e coletar os dados de difração. Entretanto, algumas dúvidas precisavam ser respondidas antes do início dos experimentos. Qual método de derivatização seria adotado? Qual composto químico deveria ser utilizado para a derivatização? Qual o melhor comprimento de onda da radiação incidente?

Ficou então decidido que os cristais derivados seriam preparados pelo método de crio derivatização rápida com halogênios (*quick cryo soaking with halides*), em especial compostos a base de iodo. A facilidade e a rapidez com que esse novo método pode ser aplicado foram essenciais para sua escolha.

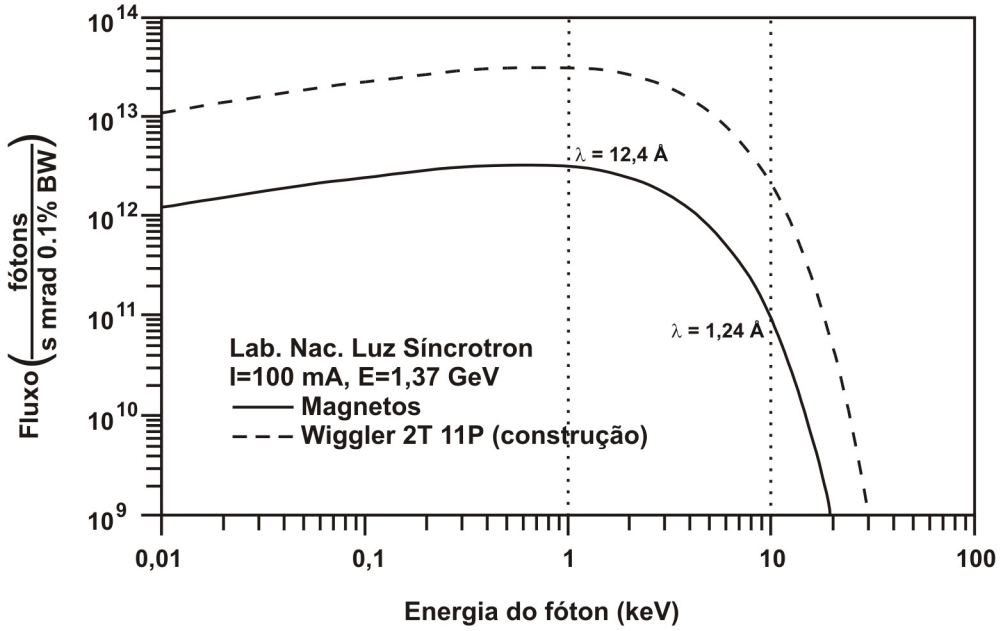


Figura 6-1. Espectro da radiação produzida no anel de armazenamento de elétrons do Laboratório Nacional de Luz Síncrotron (LNLS). A construção de um *Wiggler*, projeto em andamento no LNLS, permitirá um aumento do fluxo e da energia dos fótons produzidos como mostrado pela curva tracejada.

A decisão de usar compostos à base de iodo ao invés de bromo está intimamente relacionada com as propriedades da radiação incidente e os efeitos de dispersão anômala destes átomos. A análise do espectro da radiação obtido na linha de luz CPr do LNLS (figura 6.1) indica que o maior fluxo de fótons, normalmente utilizados em cristalografia de proteínas, está concentrado na faixa de comprimento de onda que vai de 1,3 até 1,7 Å. Por meio de um cálculo teórico para o fator de espalhamento atômico dos átomos de iodo e bromo (CROMER, 1983) utilizando radiação com comprimento de onda dentro dessa faixa, observa-se que os átomos de iodo apresentam sinal anômalo cinco vezes maior que os átomos de bromo (figura 6.2). Já que o sinal anômalo pode ser usado para a determinação das fases dos fatores de estrutura, a otimização deste sinal foi considerada essencial para atingir os objetivos propostos.

Ainda na figura 6.2 é possível fazer uma comparação entre o sinal anômalo de vários átomos pesados que foram e podem ser usados para o processo de derivatização.

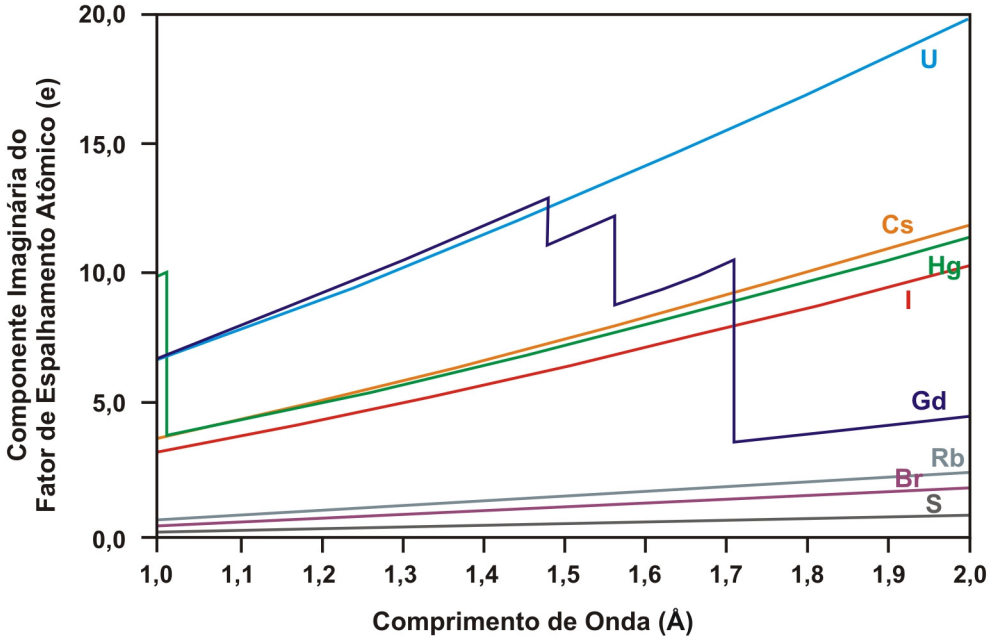


Figura 6-2. Sinal anômalo teórico para os átomos de enxofre, bromo, rubídio, gadolínio, iodo, mercúrio, césio e urânio dentro da faixa de comprimento de onda da radiação produzida na fonte de luz CPr.

Optou-se por utilizar radiação incidente com comprimento de onda igual a 1,54 Å, valor similar ao obtido em fontes convencionais de raios-X encontrados em laboratórios de todo o mundo. A escolha, além de permitir o alto fluxo de fótons na linha de luz CPr, possibilitaria que os resultados obtidos pudessem ser repetidos, teoricamente, com fontes convencionais de raios-X.

Nesta primeira etapa, os dados de difração de um cristal nativo de HEWL e um derivado de iodo foram coletados. Os dados estatísticos de cada um desses conjuntos bem como o procedimento adotado na preparação dos cristais para a coleta dos dados podem ser vistos na tabela 6.1. Na tentativa de simular uma coleta de dados rotineira, a resolução do conjunto de dados do cristal nativo foi limitada até 2,0 Å e a do cristal derivado até 2,3 Å. Na segunda etapa, a ser descrita na próxima seção, outros dois conjuntos de dados foram coletados com derivados de césio e gadolínio preparados pelo método de crio derivatização rápida.

Após a coleta e o processamento dos dados de difração dos cristais nativo e derivado de iodo, foi possível dar início ao processo de obtenção das coordenadas atômicas dos átomos de iodo (espalhadores anômalos) e, posteriormente, à determinação das fases dos fatores de estrutura do cristal nativo através da técnica SIRAS. As posições atômicas dos átomos foram obtidas pelo programa *SnB*[7] – *Shake-and-Bake* (WEEKS & MILLER, 1999) que utilizou apenas as diferenças

[7] Programa para localização de átomos pesados e espalhadores anômalos que utiliza uma poderosa formulação de métodos diretos que alterna entre o refinamento das fases no espaço recíproco e uma seleção de soluções no espaço real.

anômalas do conjunto de dados derivado. Essas coordenadas atômicas iniciais foram então usadas pelo programa *SHARP*[8] *– Statistical Heavy Atom Refinement and Phasing* (DE LA FORTELLE & BRICOGNE, 1997) junto com as amplitudes dos fatores de estrutura dos cristais nativo e derivado (iodo) para o cálculo das fases SIRAS. Após o cálculo, o primeiro mapa de densidade eletrônica com as fases experimentais foi calculado, conforme mostra a figura 6.3 a. O procedimento de modificação de densidade eletrônica foi aplicado, em seguida, com o uso do programa *SOLOMON*[9] (ABRAHAMS & LESLIE, 1996). O resultado pode ser observado na figura 6.3 b.

Tabela 6-1. Detalhes sobre a preparação e a estatística dos dados de difração de raios-X de quatro cristais de lisozima (um nativo e três derivados) usados neste projeto.

Cristalização				
Líquido-mãe	1,0 M de NaCl, 50 mM de acetato de sódio (tampão NaOAc), pH 4,5			
Temperatura	18 ^{o}C			
Solução de proteína	20-30 mg/ml			
Preparação do cristal	**Nat-HEWL**	**I-HEWL**	**Cs-HEWL**	**Gd-HEWL**
Solução crioprotetora	1,0 M NaCl tampão NaOAc 15% etilenoglicol	0,75 M NaI tampão NaOAc 15% etilenoglicol	1,0 M CsCl tampão NaOAc 15% etilenoglicol	0,625 M NaCl 0,250 M $GdCl_3$ tampão NaOAc 15% etilenoglicol
Tempo de banho* (s)	60	60	300	300
Coleta de dados	**Nat-HEWL**	**I-HEWL**	**Cs-HEWL**	**Gd-HEWL**
Comprimento de onda (Å)	1,54	1,54	1,54	1,54
Grupo espacial	$P4_32_12$	$P4_32_12$	$P4_32_12$	$P4_32_12$
Parâmetros de célula (Å)	a = 78,58 c = 36,89	a = 78,62 c = 37,05	a = 78,64 c = 36,79	a = 78,59 c = 36,82
Resolução (Å)	21,8-2,00 (2,06-2,00)	21,8-2,30 (2,33-2,30)	18,52-2,30 (2,37-2,30)	18,0-2,30 (2,37-2,30)
N. de reflexões	105547	147472	150896	152044
N. de reflexões únicas**	7449	9828	9895	9909
$\langle I/\sigma(I)\rangle$	27,3 (19,6)	26,3 (17,6)	51,1 (36,6)	35,0 (26,7)
Multiplicidade	14,2 (4,5)	15,0 (13,5)	15,2 (14,7)	15,3 (14,3)
Completeza	99,9 (99,8)	98,7 (95,0)	99,8 (100,0)	99,9 (100,0)
R_{merge}	6,6 (7,6)	13,0 (16,5)	6,7 (9,1)	8,6 (11,8)
Oscilação total (graus)	220	360	360	360
Fonte de raios-X	Linha CPr, LNLS	Linha CPr, LNLS	Linha CPr, LNLS	Linha CPr, LNLS
Detector de raios-X	mar345 (ip)#	mar345 (ip)	mar345 (ip)	mar345 (ip)

* *Tempo de imersão do cristal na solução crioprotetora. Os derivados foram preparados pelo método de crio derivatização rápida.*

** *A multiplicidade dos conjuntos derivados (nativo) foi calculada tratando os pares de Friedel como reflexões distintas (equivalentes).*

\# *Image plate (ip).*

Os valores estatísticos para as camadas com mais alta resolução são mostrados entre parênteses.

[8] Utiliza o método da máxima verossimilhança para o refinamento dos parâmetros dos átomos pesados (coordenadas, ocupâncias e fatores de temperatura) e para o cálculo das fases dos fatores de estrutura.

[9] Utiliza o método conhecido como "*solvent flipping*" para os procedimentos de modificação de densidade eletrônica. Ao contrário de outros métodos, a densidade eletrônica na região do solvente é invertida e não ajustada a um valor constante.

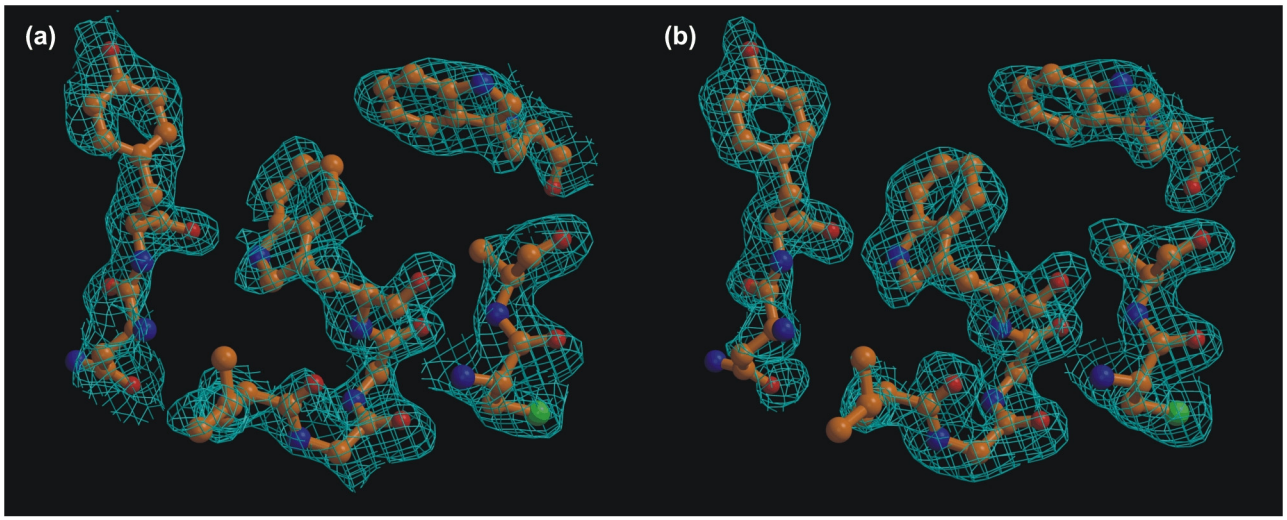


Figura 6-3. Mapas de densidade eletrônica para uma região arbitrária do cristal nativo de HEWL. (a) Mapa experimental SIRAS obtido com os dados de difração do cristal nativo e do derivado de iodo. (b) Resultado da utilização dos procedimentos de modificação de densidade eletrônica no mapa representado em (a). O modelo tridimensional da HEWL também está representado. Os átomos de carbono, nitrogênio, oxigênio e enxofre são representados por esferas alaranjadas, azuis, vermelhas e verdes, respectivamente. Os átomos de hidrogênio não são representados.

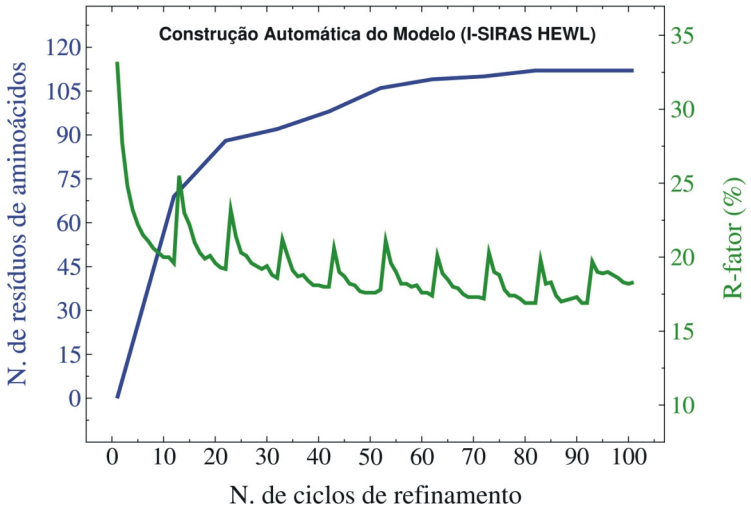


Figura 6-4. Interpretação e construção automática do modelo tridimensional da HEWL pelo programa ARP/wARP utilizando um mapa de densidade eletrônica com 2,0 Å de resolução.

Uma vez que a estrutura tridimensional da HEWL já é conhecida, a construção de um modelo tridimensional para o cristal nativo teve o único objetivo de verificar a possibilidade da interpretação automática do mapa de densidade eletrônica. Esta tarefa, executada pelo programa *ARP/wARP* (PERRAKIS, MORRIS & LAMZIN, 1999), e seu desempenho ao longo dos ciclos de interpretação e construção automática podem ser vistos na figura 6.4.

A análise do modelo inicial obtido por *ARP/wARP*[10], através do programa gráfico *O* (JONES *et al*., 1991), permitiu corrigir alguns erros da construção automática. Por fim, o modelo foi então refinado contra o conjunto de dados nativo até valores aceitáveis de R_{factor}.

O modelo nativo, já refinado, foi utilizado para resolução da estrutura do derivado de iodo. Uma vez que ambos os conjuntos são praticamente idênticos (nativo e derivado), exceto pelos espalhadores anômalos, o processo de refinamento foi extremamente rápido.

A conclusão do processo de refinamento permitiu analisar, em detalhes, o ambiente químico onde se encontravam os íons de iodo. De maneira similar ao que já havia sido reportado na literatura (DAUTER, DAUTER & RAJASHANKAR, 2000), os íons de iodo (halogênios) adotam, mediante o processo de crio derivatização rápida, inúmeras posições atômicas em torno das macromoléculas. Estes íons, com as mais variadas ocupâncias, passam a ocupar as regiões anteriormente preenchidas por moléculas ordenadas de água, interagindo com resíduos de aminoácidos carregados e com o nitrogênio da cadeia principal.

Na figura 6.5 um mapa de Fourier de diferença anômala é mostrado sobre a estrutura da HEWL. O mapa tem a propriedade de ressaltar os espalhadores anômalos da estrutura cristalográfica.

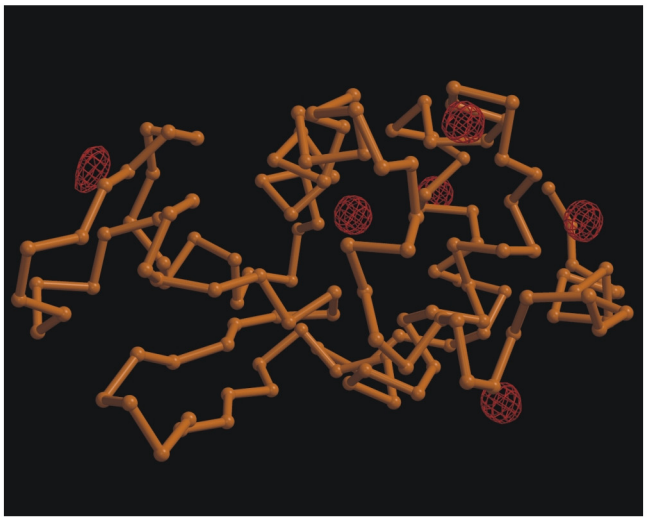


Figura 6-5. Sobreposição do modelo refinado para o derivado I-HEWL (cadeia Cα em laranja) e do mapa de Fourier de diferença anômala (em vermelho) contornado a cinco sigmas. Neste mapa os íons de iodo podem ser vistos ao redor da estrutura cristalográfica da proteína, ocupando as primeiras camadas de solvente ao redor da macromolécula.

[10] Este programa tem como objetivo principal a construção e o refinamento de uma estrutura tridimensional combinando a interpretação do mapa de densidade eletrônica por meio do conceito de modelo híbrido, o reconhecimento de padrões de ligações químicas e o refinamento da estrutura pelo método de máxima verossimilhança.

6.2 Complementação do método de crio derivatização rápida

A utilização de iodeto de sódio para o processo de crio derivatização rápida provou ser extremamente eficiente para a obtenção das fases dos fatores de estrutura pela técnica SIRAS. De fato, a tentativa de obtenção das fases por SAD, utilizando apenas o conjunto de dados do cristal derivado, também resultou em um mapa de densidade eletrônica interpretável.

Contudo, na tentativa de ampliar as possibilidades da técnica de crio derivatização rápida, foram realizados outros experimentos testes com HEWL. Os sais de halogênios (NaI, KI e NaBr) usados no processo de derivatização rápida se dissociam em solução aquosa permitindo com que seus íons interajam com os demais compostos em solução. Neste processo, os ânions de iodo ou bromo passam a interagir com as moléculas ordenadas de água e com os resíduos de aminoácidos carregados expostos ao solvente. De maneira análoga, espera-se que os cátions presentes na solução também possam interagir com regiões específicas das macromoléculas. Deve-se, contudo, ter em mente que, devido à diferença de carga entre estes íons, a interação deles com as macromoléculas deve ocorrer em nichos atômicos diferentes.

Sob o ponto de vista da obtenção da estrutura cristalográfica, essa diferença se torna uma complementaridade. Se dois ou mais derivados diferentes pudessem ser preparados, então as informações de fase provenientes de cada um deles poderiam ser combinadas pelas técnicas MIR e MIRAS (seções 5.3.1.1 e 5.3.1.2). Para testar esta possibilidade foram preparados, pelo método de crio derivatização rápida, dois derivados de HEWL com compostos à base de metais alcalinos e terras raras. A escolha dos sais de cloreto de césio (CsCl) e cloreto de gadolínio III ($GdCl_3$) foi baseada, principalmente, em suas propriedades fisico-químicas. O primeiro composto (CsCl), altamente solúvel em água, se dissocia em Cs^+ e Cl^- permitindo que estes íons interajam com outros compostos da solução. O cátion de césio apresenta praticamente o mesmo sinal anômalo do ânion de iodo dentro da mesma faixa de comprimento de onda disponível na linha de luz CPr do LNLS (figura 6.2). Já o íon Gd^{3+} foi escolhido porque, além de proporcionar um elevado sinal anômalo, ele apresenta, em solução, uma coordenação específica formada por outros íons ou moléculas. Deve-se, portanto, esperar sítios de ligação distintos tanto para os derivados preparados com terras raras quanto para os preparados com metais alcalinos e com halogênios.

Após algumas tentativas, os dados de difração desses dois derivados foram coletados na linha de luz CPr. A estatística dos conjuntos de dados e os detalhes da preparação dos cristais pelo processo de crio derivatização rápida são mostrados na tabela 6.1. Mais uma vez, a resolução dos conjuntos de dados foi limitada a 2,3 Å.

Apesar de os procedimentos adotados ao longo dos três experimentos de crio derivatização rápida serem praticamente os mesmos, foi possível verificar, logo no início da preparação dos derivados, que o uso dos compostos à base de terras raras causava uma rápida degradação do cristal. Ao contrário, os compostos à base de halogênios e metais alcalinos praticamente não afetavam o cristal, mesmo em altas concentrações. Assim, inúmeros testes foram realizados em várias concentrações de terras raras até que se obtivesse um derivado que suportasse o tempo de imersão na solução crio derivatizante.

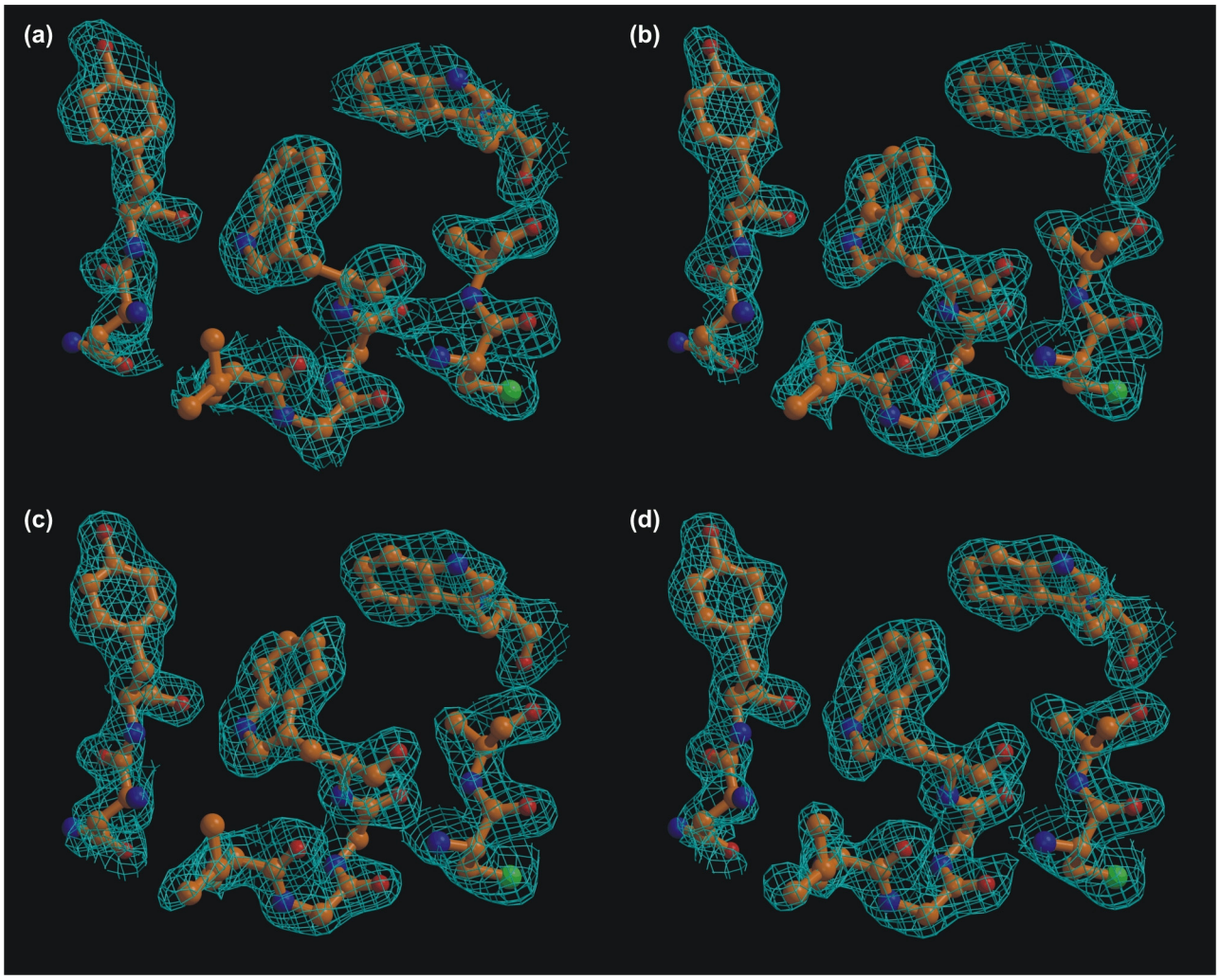


Figura 6-6. Mapas de densidade eletrônica para o cristal nativo de lisozima. Os mapas foram obtidos por (a) SIRAS com o derivado de césio, (b) SIRAS com o derivado de gadolínio, (c) MIRAS com os derivados de césio e de gadolínio e (d) MIRAS com os derivados de césio, gadolínio e iodo após os procedimentos de modificação de densidade eletrônica.

As posições atômicas dos átomos de césio e gadolínio foram identificadas, em cada um dos derivados, utilizando os mesmos procedimentos adotados para o caso do derivado de iodo. Após

esta etapa, as fases SIRAS para o cristal nativo (coletado anteriormente) foram obtidas ora com os dados de difração do derivado de césio, ora com as do derivado de gadolínio. Além disso, mais dois conjuntos de fases MIRAS foram calculados com o uso combinado dos quatro conjuntos de dados. Os mapas de densidade eletrônica obtidos após os protocolos de modificação de densidade eletrônica são mostrados nas figuras 6.6 a até 6.6 d. Nesses quatro mapas, é mostrada a mesma região da figura 6.3.

A interpretação dos mapas de densidade eletrônica pôde ser feita de forma automática em quase todos os casos testados e, em certas regiões da unidade assimétrica, a interpretação visual foi tão precisa quanto a interpretação automática.

Neste momento, foram feitas várias análises dos conjuntos de dados coletados e dos processos utilizados para a obtenção das fases dos fatores de estrutura e conseqüente determinação do modelo tridimensional da HEWL. As análises permitiram encontrar alguns parâmetros importantes que ajudariam a prever o sucesso ou o fracasso na resolução da estrutura cristalográfica.

Um dos primeiros parâmetros analisados em todos os quatro conjuntos de dados foi a razão $I_{hkl}/\sigma(I_{hkl})$ por faixa de resolução. É fácil compreender que essa razão esteja intimamente relacionada com a redundância dos dados coletados (tabela 6.1). Nos quatro conjuntos, a coleta de 360^o fez com que várias reflexões simetricamente equivalentes fossem medidas. Sob o ponto de vista estatístico, isso implicou uma determinação mais precisa do valor das intensidades das reflexões e de seus respectivos erros já que o processo foi baseado em uma amostragem maior de reflexões equivalentes. Se, entretanto, 90^o de dados fossem coletados e analisados, seria atingida também uma completeza próxima a 100%, porém a redundância dos dados seria significativamente menor e conseqüentemente a estimativa do erro de cada intensidade de cada reflexão seria obtida com uma amostragem menor de reflexões equivalentes. Tradicionalmente, as coletas de dados de difração eram feitas utilizando uma abordagem minimalista, ou seja, coletava-se a menor quantidade de dados (menor rotação total) até que o conjunto atingisse uma completeza bem próxima a 100%. Contudo, os valores de pequenos sinais, como o das diferenças anômalas, só poderão ser interpretados com segurança se os erros forem suficientemente inferiores. Dessa forma, sob o ponto de vista da obtenção das fases dos fatores de estrutura com o uso do sinal anômalo, é importante que a coleta seja, além de completa, redundante.

O efeito benéfico da redundância dos dados pode ser facilmente observado com a análise da figura 6.7. Nessa figura, as razões $\Delta F^{ANO}/F$ e $\Delta F^{ANO}/\sigma\left(\Delta F^{ANO}\right)$ por faixa de resolução são analisadas à medida que a redundância dos dados é aumentada para os três conjuntos derivados de

HEWL. É claramente visível que o valor do sinal anômalo ($\Delta F^{ANO}/F$) é superestimado com uma baixa redundância dos dados. À medida que a amostragem de reflexões equivalentes aumenta, uma estimativa correta do sinal real pode ser feita. Assim, é possível verificar se de fato houve a incorporação dos espalhadores anômalos na estrutura cristalográfica. A análise dessa quantidade é, portanto, um potencial indicador para a presença dos espalhadores anômalos. Já a razão $\Delta F^{ANO}/\sigma\left(\Delta F^{ANO}\right)$ vai aumentando de valor à medida que a redundância dos dados aumenta. Esse efeito é reflexo de uma melhor estimativa dos erros associados às diferenças anômalas ($\sigma\left(\Delta F^{ANO}\right)$) e o seu comportamento permite avaliar a qualidade do sinal anômalo medido.

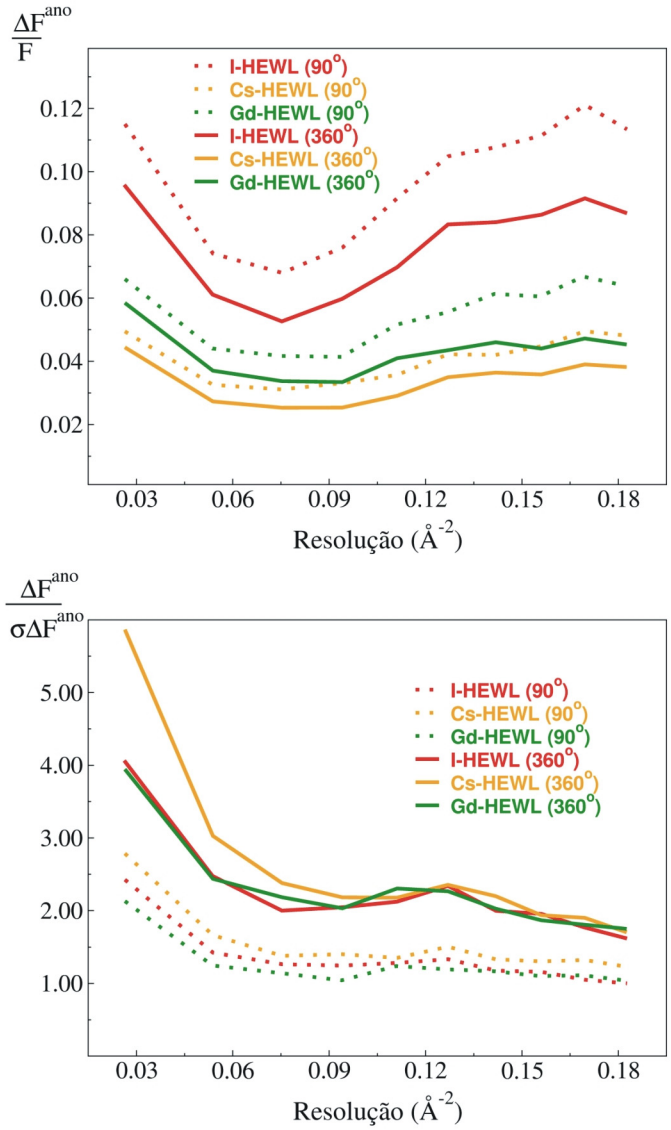


Figura 6-7. Efeito da redundância do conjunto de dados sobre as razões $\Delta F^{ano}/F$ e $\Delta F^{ano}/\sigma(\Delta F^{ano})$ nos três conjuntos de dados derivados de HEWL (derivados de iodo, césio e gadolínio).

De maneira similar, o sinal isomorfo (seção 5.4) também foi analisado. Esse sinal, normalmente bem maior que o sinal anômalo, é muito usado para a obtenção das fases. Contudo, deve-se ter em mente que a presença dele não é necessariamente uma confirmação inequívoca da presença dos átomos pesados ou espalhadores anômalos na estrutura do derivado. Isso se dá porque o processo de derivatização com átomos pesados ou espalhadores anômalos pode promover um rearranjo da estrutura cristalina sem, necessariamente, permitir a incorporação desses átomos. Dessa forma, os cristais nativo e pseudo-derivado (cristais não-isomorfos) apresentarão grandes diferenças isomorfas que não poderão ser usadas para obtenção das fases. O comportamento do sinal isomorfo em cada um dos três pares de dados é apresentado na figura 6.8.

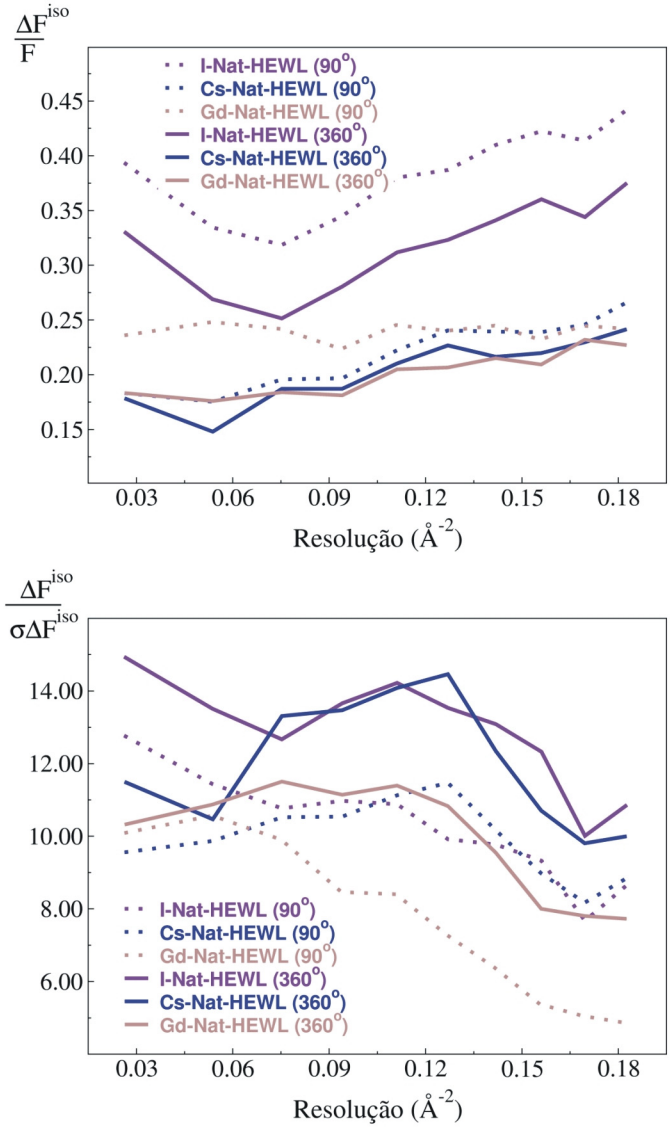


Figura 6-8. Efeito da redundância do conjunto de dados sobre as razões $\Delta F^{iso}/F$ e $\Delta F^{iso}/\sigma(\Delta F^{iso})$ nos três pares de conjuntos de dados de HEWL (nativo com os derivados de iodo, césio e gadolínio).

Os parâmetros analisados até o momento permitem quantificar a qualidade dos dados de difração obtidos durante o experimento, principalmente sob o ponto de vista da obtenção das fases dos fatores de estrutura. Contudo, é comum dizer que a estrutura cristalográfica só poderá ser determinada se os mapas de densidade eletrônica forem suficientemente interpretáveis. Esse julgamento subjetivo pode ser feito de forma quantitativa com a análise da grandeza conhecida como figura de mérito (seção 5.4). A figura de mérito m, para uma reflexão (hkl) qualquer, mede a probabilidade de acerto no cálculo da fase para esta reflexão. Assim, uma maneira expressiva de julgar a qualidade das fases obtidas é por meio do perfil médio da figura de mérito por faixa de resolução.

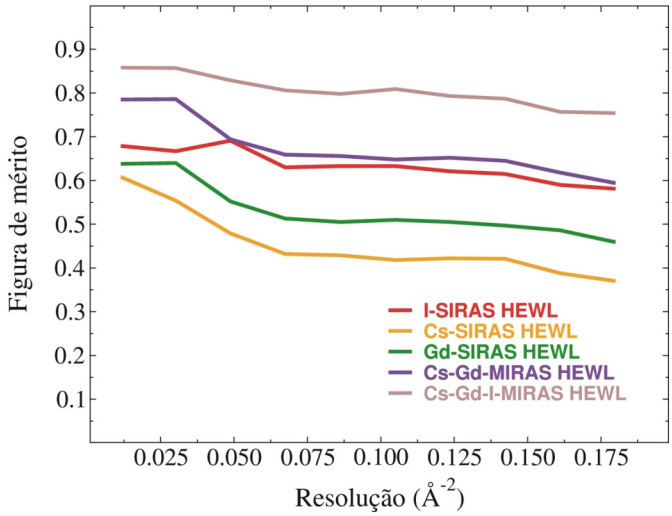


Figura 6-9. Valores médios das figuras de mérito por faixa de resolução para as fases obtidas por SIRAS e MIRAS com os dados de difração dos cristais de HEWL.

Na figura 6.9, os valores médios das figuras de mérito por faixa de resolução são mostrados para cada uma das técnicas de obtenção de fases utilizadas com os cristais de HEWL. Com a análise da figura, é possível verificar o efeito complementar dos diversos derivados ao permitir com que as fases dos fatores de estrutura fossem determinadas com maior precisão e confiabilidade.

O efeito complementar dos derivados é resultado do fato de que a incorporação dos espalhadores anômalos ocorre em regiões diferentes da superfície das macromoléculas (figura 6.10). Os cinco sítios principais dos íons de césio foram encontrados a uma distância de 2,7 a 3,7 Å dos grupos carbonis da cadeia principal, das cadeias laterais negativamente carregadas dos resíduos de ácido aspártico e ácido glutâmico e das cadeias laterais polares dos resíduos asparagina e serina. Já os átomos de gadolínio foram encontrados próximos aos átomos de oxigênio das cadeias laterais de

asparagina e ácido aspártico. O sítio principal deste íon estava coordenado pelo grupo carboxil de um ácido aspártico (a uma distância de 1,8 Å) e cinco moléculas de água (2,4 – 2,8 Å). Este último resultado pode explicar o motivo da rápida degradação dos cristais quando imersos nas soluções concentradas de $GdCl_3$. A difusão desses íons pelos canais de solvente do cristal poderia estar promovendo um grande deslocamento de moléculas de água ao longo dos canais, perturbando a ordem cristalina.

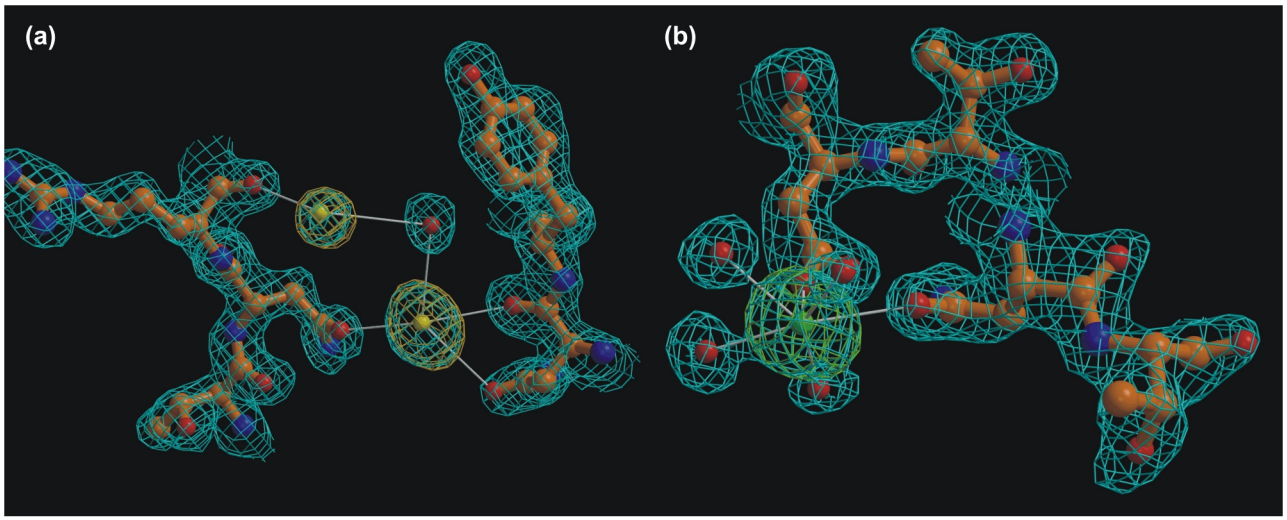


Figura 6-10. Sítios atômicos para os átomos de (a) césio e de (b) gadolínio nos derivados de lisozima. Estes sítios representam, de forma genérica, o ambiente químico no qual os espalhadores anômalos foram encontrados. Os derivados foram preparados pelo método de crio derivatização rápida.

Para finalizar esta parte, é importante ressaltar que, apesar de não haver garantia de que o bom comportamento dos parâmetros analisados até o momento (redundância, $\Delta F^{ANO}/F$, $\Delta F^{ANO}/\sigma\left(\Delta F^{ANO}\right)$, $\Delta F^{ISO}/F$, $\Delta F^{ISO}/\sigma\left(\Delta F^{ISO}\right)$, figuras de mérito) seja suficiente para o sucesso na determinação da estrutura cristalográfica de uma macromolécula biológica, um desempenho similar ao apresentado nos testes da HEWL é, pelo menos, um bom indicativo de que o experimento está sendo conduzido com acerto. Os valores de 4 a 8 % para o sinal anômalo, 15 a 25 % para o sinal isomorfo e figuras de mérito acima de 0,65 na baixa resolução são bons indícios de sucesso na determinação da estrutura cristalográfica.

Todas as proposições referentes à possibilidade de resolver uma estrutura inédita utilizando o processo de crio derivatização rápida e, além disso, utilizar alguns parâmetros experimentais para julgar o sucesso de cada uma das etapas desse procedimento precisavam ser testadas em casos verdadeiramente inéditos. Esta tarefa foi realizada logo em seguida e será apresentada nas próximas

seções. Apesar de a utilização de compostos à base de terras raras ter dado resultados satisfatórios para o teste com HEWL, a preparação dos derivados foi mais laboriosa e, por isso, optou-se por utilizar apenas os compostos à base de halogênios e metais alcalinos nos casos inéditos.

6.3 Primeiros experimentos com uma proteína "inédita"

A primeira proteína "inédita" utilizada neste trabalho foi um inibidor de proteinase; mais especificamente um inibidor de tripsina extraído das sementes de uma planta conhecida como *Copaifera langsdorffii*. Apesar de já haver um grande número de estruturas cristalográficas para inibidores de proteinases, esta proteína foi resolvida como uma proteína inédita uma vez que o método de substituição molecular (seção 5.3) não pôde ser aplicado. Esta proteína é o objeto de estudo de um projeto de doutorado, ainda em andamento, no grupo de pesquisa em que esta tese foi desenvolvida. Em vista disso, os resultados de caráter biológico e bioquímico obtidos com a resolução desta estrutura não serão apresentados aqui.

Os inibidores de proteinases podem ser definidos como proteínas capazes de inibir a ação de enzimas hidrolíticas tanto *in vitro* quanto *in vivo* pela formação de complexos macromoleculares estáveis. Os inibidores são encontrados em várias formas em inúmeros tecidos e fluídos de plantas, animais e microorganismos. As macromoléculas são classificadas em famílias de acordo com a classe das enzimas que inibem. Das famílias de inibidores de tripsina, as mais importantes são as do tipo Kunitz e as do tipo Bowman-Birk. As primeiras apresentam peso molecular em torno de 20 kDa e duas pontes dissulfeto. As últimas, por sua vez, são menores (8-10 kDa) e apresentam sete ligações S-S (KRAUCHENCO *et al*., 2001).

Apesar de as estruturas tridimensionais de alguns inibidores de tripsina já serem conhecidas, a proteína, de agora em diante denominada TRIN, não apresentava, até este estudo, um modelo tridimensional para sua estrutura. Pelo fato de que sua estrutura primária não era conhecida, a procura por um modelo tridimensional homólogo já existente não poderia ser feita de forma consistente. Além disso, os primeiros estudos de cristalização dessa proteína indicaram que se tratava, na verdade, de um heterodímero composto por duas cadeias polipeptídicas não-ligadas de 9 e 11 kDa. Em vista dessas singularidades, optou-se por utilizar o método de átomos pesados.

De maneira similar aos procedimentos adotados para a HEWL, os primeiros cristais da TRIN foram obtidos por meio do método da gota suspensa com condições de cristalização estabelecidas no grupo de Cristalografia de Proteínas do Laboratório Nacional de Luz Síncrotron (KRAUCHENCO

et al., 2001). Na tentativa de resolver a estrutura cristalina dessa macromolécula, dois cristais derivados, um de iodo e outro de césio, foram preparados pelo método de crio derivatização rápida. Na tabela 6.2, são apresentados os dados estatísticos do cristal nativo de TRIN e dos derivados de iodo e césio, bem como a forma como eles foram preparados para a coleta de dados.

Tabela 6-2. Detalhes sobre a preparação e a estatística dos dados de difração de raios-X dos cristais de TRIN coletados na linha de luz CPr do Laboratório Nacional de Luz Síncrotron.

Cristalização			
Líquido-mãe	20-25% PEG 8000, 100 mM de acetato de sódio (tampão NaOAc), pH 4,8		
Temperatura	18 ^{o}C		
Solução de proteína	10 mg/ml		
Preparação do cristal	**Nat-TRIN**	**I-TRIN**	**Cs-TRIN**
Solução crioprotetora	líquido-mãe 20% etilenoglicol	0,50 M NaI líquido-mãe 20% etilenoglicol	1,0 M CsCl líquido-mãe 20% etilenoglicol
Tempo de banho* (s)	60	180	300
Coleta de dados	**Nat-TRIN**	**I-TRIN**	**Cs-TRIN**
Comprimento de onda (Å)	1,54	1,54	1,54
Grupo espacial	$P4_32_12$	$P4_32_12$	$P4_32_12$
Parâmetros de célula (Å)	a = 58,71 c = 93,75	a = 58,42 c = 93,91	a = 58,33 c = 93,80
Resolução (Å)	25,3-1,83 (1,87-1,83)	25,3-1,92 (1,96-1,92)	22,9-2,00 (2,05-2,00)
N. de reflexões	123604	285176	312518
N. de reflexões únicas**	15024	23500	20853
$\langle I/\sigma(I)\rangle$	23,6 (3,3)	33,7 (13,6)	33,7 (13,1)
Multiplicidade	8,2 (5,4)	12,1 (12,0)	15,0 (14,7)
Completeza	99,7 (97,6)	99,5 (93,9)	99,8 (99,5)
R_{merge}	8,9 (50,8)	7,3 (17,0)	9,3 (23,8)
Oscilação total (graus)	130	292	360
Fonte de raios-X	Linha CPr, LNLS	Linha CPr, LNLS	Linha CPr, LNLS
Detector de raios-X	mar345 (ip)#	mar345 (ip)	mar345 (ip)

* *Tempo de imersão do cristal na solução crioprotetora.*
** *A multiplicidade dos conjuntos derivados (nativo) foi calculada tratando os pares de Friedel como reflexões distintas (equivalentes).*
\# *Image plate (ip).*
Os valores estatísticos para as camadas com mais alta resolução são mostrados entre parênteses.

Pode-se observar que os conjuntos de dados foram coletados com alta redundância, uma vez que a abordagem mostrou ser fundamental para todo o processo de obtenção das fases. Inicialmente, todas as imagens de difração dos três conjuntos foram processadas e escalonadas com cuidado, procurando, principalmente, eliminar as reflexões com elevada probabilidade de apresentarem erros de medida ou escalonamento, sobreposição com outras reflexões ou mesmo influência de reflexões provenientes de cristais de gelo.

A presença dos espalhadores anômalos nos derivados de iodo e césio foi verificada inicialmente com a análise dos pares de Friedel. Observou-se claramente, nos dois conjuntos de dados, a quebra da lei de Friedel (seção 5.3.1.2) como pode ser visto na figura 6.11. A confirmação da presença dos espalhadores anômalos foi obtida com uma análise similar para as diferenças isomorfas (figura 6.11).

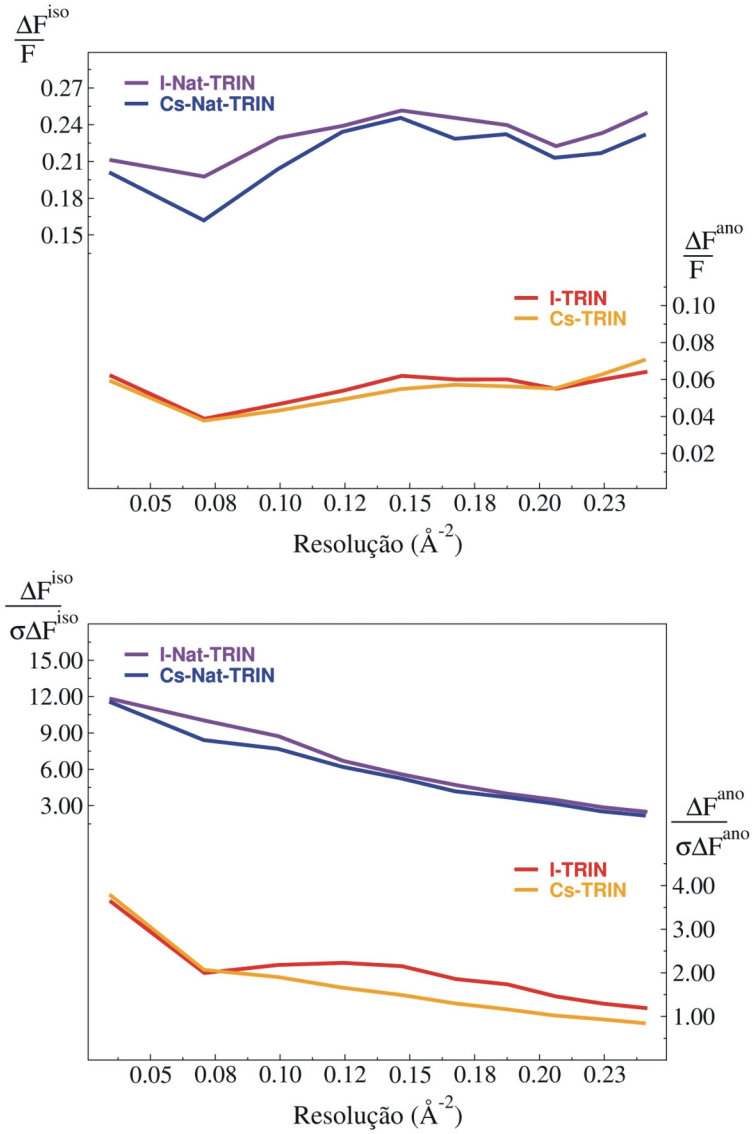


Figura 6-11. Comportamento das razões $\Delta F^{iso}/F$, $\Delta F^{ano}/F$, $\Delta F^{iso}/\sigma(\Delta F^{iso})$ e $\Delta F^{ano}/\sigma(\Delta F^{ano})$ para os conjuntos de dados dos cristais de TRIN.

A obtenção das coordenadas dos átomos de iodo e césio foi feita pelo programa *SnB* que, de maneira similar ao caso da HEWL, utilizou apenas as diferenças anômalas dos conjuntos de dados derivados para esta tarefa. O cálculo das fases dos fatores de estrutura foi feito com o programa

SHARP utilizando as técnicas I-SIRAS, Cs-SIRAS e MIRAS. A análise quantitativa da qualidade das fases obtidas por cada um dos métodos indica que, em princípio, todos os três cálculos resultaram em fases capazes de gerar mapas de densidade eletrônica interpretáveis, contudo, as fases obtidas por MIRAS foram as que tiveram as melhores figuras de mérito, como pode ser visto na figura 6.12.

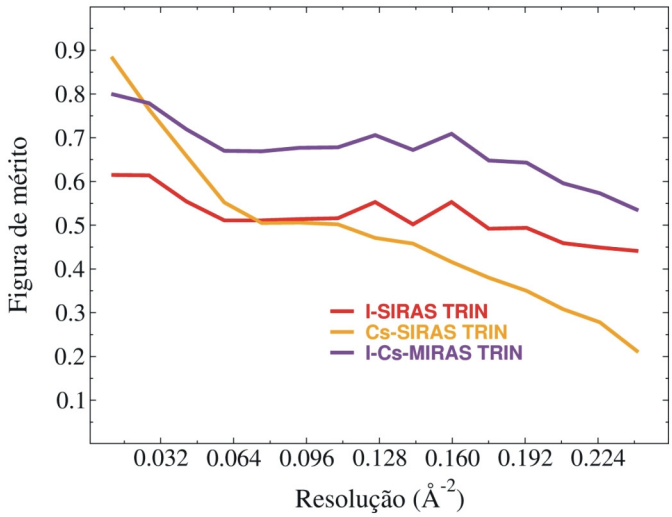


Figura 6-12. Valores médios das figuras de mérito por faixa de resolução para as fases obtidas por SIRAS e MIRAS com os dados de difração dos cristais de TRIN.

Os mapas de densidade eletrônica calculados de forma independente com as fases obtidas por SIRAS e MIRAS foram submetidos aos procedimentos de modificação de densidade eletrônica e os resultados podem ser vistos nas figuras 6.13 a, 6.13 b e 6.13 c.

Devido à excelente qualidade de todos os três mapas de densidade eletrônica, uma interpretação quase inteiramente automática pôde ser feita pelo programa *ARP/wARP*. O modelo inicial foi analisado no programa de visualização gráfica *O*, o que permitiu corrigir alguns erros da construção automática e assinalar a seqüência mais provável de aminoácidos das duas cadeias polipeptídicas da macromolécula. Ao final dessa etapa, o modelo foi atualizado inúmeras vezes durante vários ciclos de refinamento e visualização gráfica. O modelo tridimensional final pode ser visto, em parte, na figura 6.13 d superposto ao mapa de densidade eletrônica refinado e, por completo, na figura 6.14 b que representa as estruturas secundárias observadas nas duas cadeias polipeptídicas. Os detalhes biológicos e bioquímicos resultantes deste trabalho cristalográfico estão sendo utilizados para a elaboração de um trabalho científico que deve ser publicado em breve.

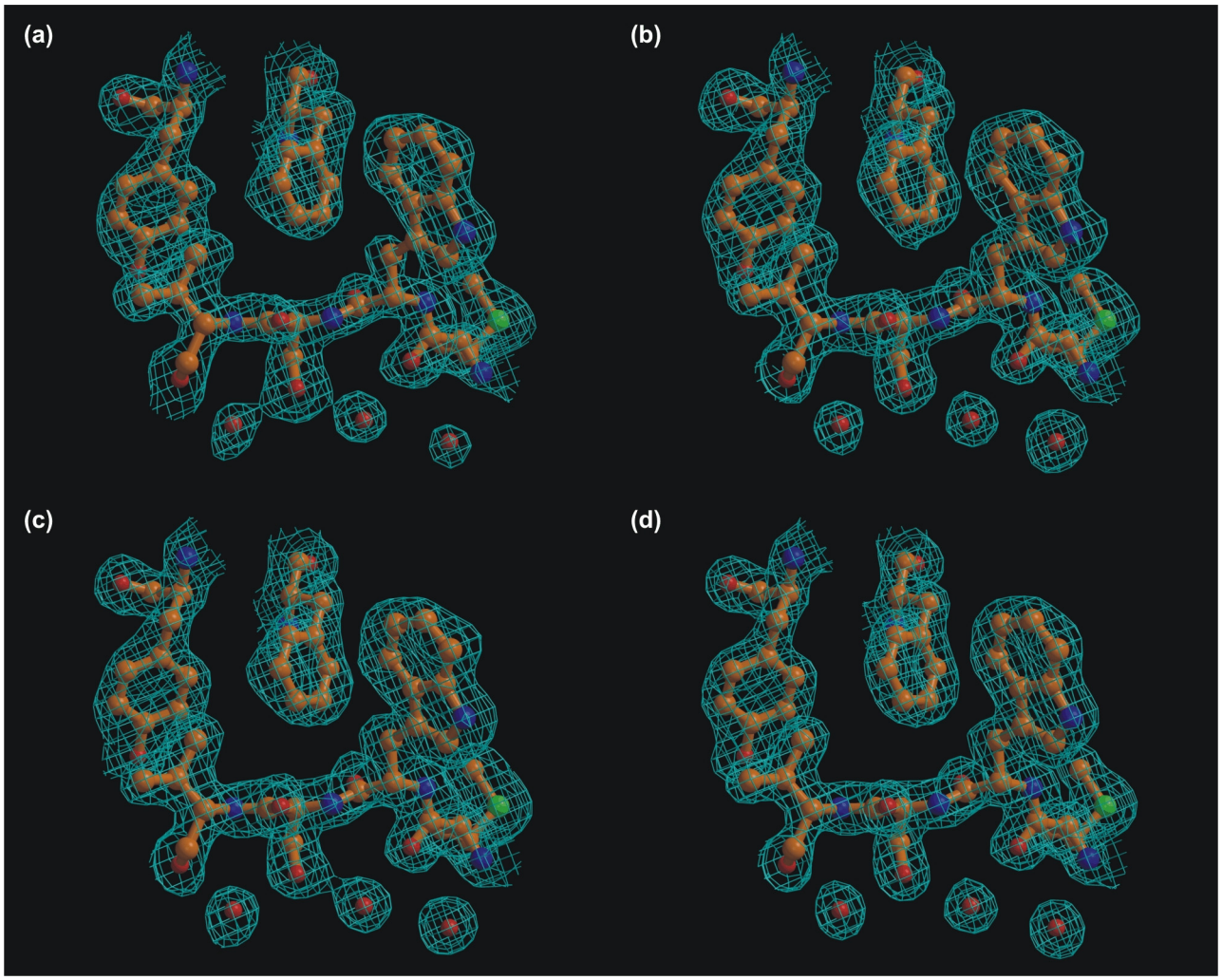


Figura 6-13. Mapas de densidade eletrônica para o cristal de TRIN. Mapas obtidos pelas técnicas (a) SIRAS com césio, (b) SIRAS com iodo e (c) MIRAS com iodo e césio após os procedimentos de modificação de densidade eletrônica. O mapa representado em (d) foi obtido após o refinamento da estrutura. O modelo tridimensional do inibidor de tripsina também está representado em todas as figuras.

Apenas a título de curiosidade, dois mapas de Fourier de diferença anômala são mostrados sobre a estrutura cristalográfica da TRIN na figura 6.15. É possível visualizar nos mapas a densidade eletrônica de todos os espalhadores anômalos presentes nas estruturas cristalográficas dos derivados. Pode-se observar que os elementos iodo (vermelho) e césio (amarelo) adotam, como esperado, posições atômicas completamente diferentes em torno da superfície das macromoléculas biológicas, interagindo com os resíduos de aminoácidos expostos ao solvente. Mais detalhes do ambiente químico em torno dos íons usados para o processo de derivatização estão nas próximas seções.

O método de crio derivatização rápida com halogênios e metais alcalinos, este último desenvolvido a partir dos trabalhos iniciais desta tese, foi e tem sido aplicado para a determinação de

outras estruturas cristalográficas. Outros exemplos, obtidos ao longo deste trabalho de tese, serão mostrados na próxima seção. Antes, porém, será apresentado um trabalho científico que permitirá visualizar o contexto geral do método de crio derivatização rápida no que diz respeito à resolução de estruturas cristalográficas inéditas.

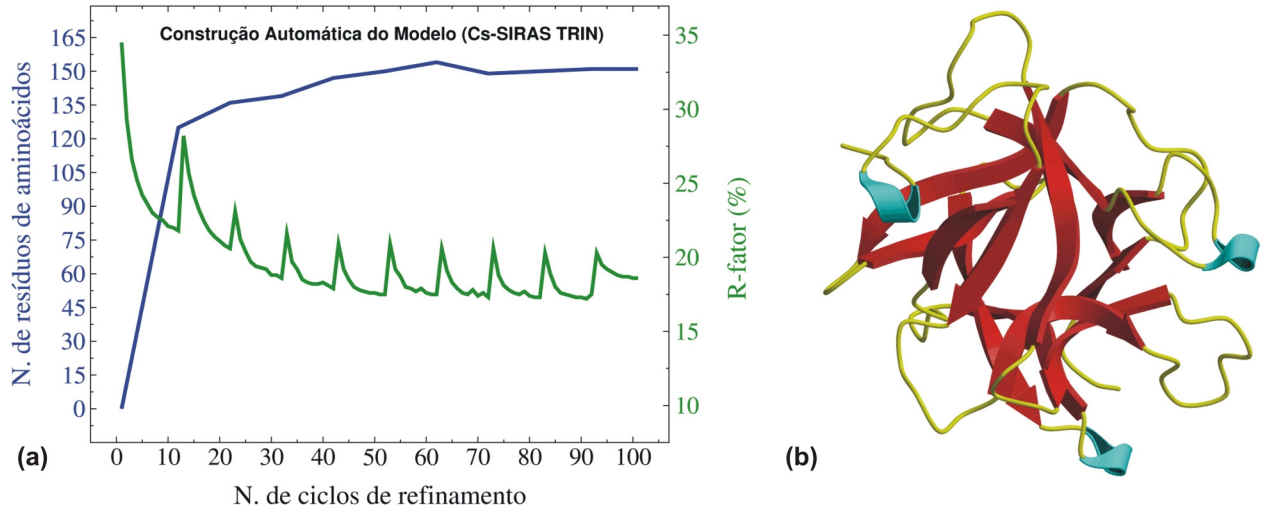


Figura 6-14. Modelo tridimensional refinado da estrutura cristalográfica da TRIN. (a) Construção automática do modelo tridimensional da TRIN usando o método SIRAS com o derivado de césio. (b) Representação das estruturas secundárias observadas na estrutura cristalina.

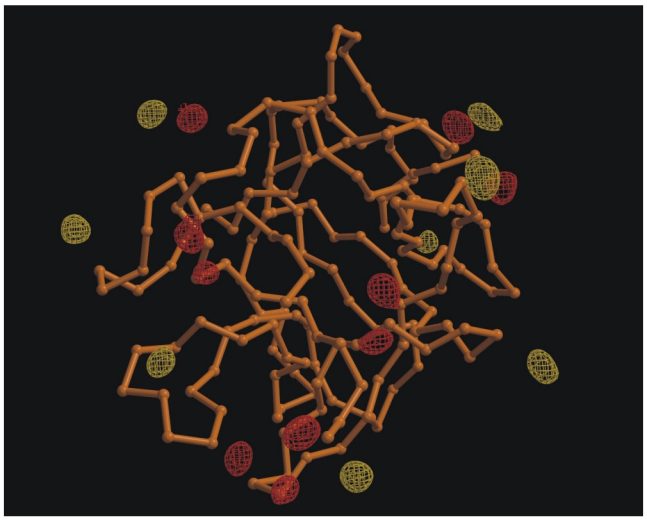


Figura 6-15. Sobreposição do modelo nativo refinado para a TRIN e de dois mapas de densidade eletrônica de diferença anômala contornados a cinco sigmas (iodo em vermelho e césio em amarelo).

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Protein crystal structure solution by fast incorporation of negatively and positively charged anomalous scatterers

Ronaldo A. P. Nagem,[a,b] Zbigniew Dauter[c] and Igor Polikarpov[a]*

[a]Laboratório Nacional de Luz Síncrotron, Caixa Postal 6192, CEP 13084-971, Campinas SP, Brazil, [b]Departamento Física, UNICAMP, Caixa Postal 6165, CEP 13084-971, Campinas SP, Brazil, and [c]Synchrotron Radiation Research Section, National Cancer Institute and Brookhaven National Laboratory, Building 725A-X9, Upton, NY 11973, USA

Correspondence e-mail: igor@lnls.br

Received 12 February 2001
Accepted 1 May 2001

The preparation of derivatives by the traditional methods of soaking is one of the most time-consuming steps in protein crystal structure solution by X-ray diffraction techniques. The 'quick cryosoaking' procedure for derivatization with halides (monovalent anions) offers the possibility of significantly speeding up this process [Dauter *et al.* (2000), *Acta Cryst.* D**56**, 232–237]. In the present work, an extension of this technique is proposed and the use of two different classes of compounds (monovalent and polyvalent cations) that can be successfully utilized in the quick cryosoaking procedure for the derivatization and phasing of protein crystals is described. This approach has been tested on hen egg-white lysozyme and has been successfully used to solve the structure of a novel trypsin inhibitor. The possibility of using cations in the fast cryosoaking procedure gives additional flexibility in the process of derivatization and increases the chances of success in phase determination. This method can be applied to high-throughput crystallographic projects.

1. Introduction

The speed of determination of novel macromolecular structures at both in-house and synchrotron X-ray sources is limited to a large extent by the process of preparation of heavy-metal derivatives. The traditional multiple isomorphous replacement (MIR) method (Green *et al.*, 1954; Bragg & Perutz, 1954) for solving protein crystal structures by X-ray diffraction techniques requires several (at least two) isomorphous derivatives. These normally are prepared by introducing different heavy-metal ions into the native crystals. Sometimes, after months of work preparing derivatives, only a few or none at all are useful for determining the crystallographic protein structure. The pace of structure solution could be improved by making use of the anomalous signal in single/multiple isomorphous replacement with anomalous scattering (SIRAS/MIRAS) methods (North, 1965; Matthews, 1966). Single anomalous dispersion (SAD) and multiwavelength anomalous dispersion (MAD; Hendrickson, 1991; Smith, 1991) require only one derivative crystal for crystallographic structure solution, which results in a great reduction in time and effort. In this situation, one can collect only one data set with an optimized anomalous signal or collect only a few data sets using X-rays of different wavelengths. In addition, the most common derivatization procedure applied to MAD includes the use of genetic engineering to replace methionines by selenomethionines in proteins prior to crystallization.

Recently, a new procedure for obtaining derivatives, named 'quick cryosoaking', has been proposed (Dauter *et al.*, 2000). According to this procedure, a quick soak of a protein crystal

Table 1
Details of the preparation and data-collection statistics of HEWL and TRIN crystals.

Statistical values for the highest resolution shells are shown in parentheses. All five derivatives were prepared following the quick cryosoaking procedure for derivatization.

	Nat-HEWL	Cs-HEWL	Gd-HEWL	I-HEWL	Nat-TRIN	Cs-TRIN	I-TRIN
Space group	$P4_32_12$	$P4_32_12$	$P4_32_12$	$P4_32_12$	$P4_32_12$	$P4_32_12$	$P4_32_12$
Unit-cell parameters (Å)	a = 78.58, c = 36.89	a = 78.64, c = 36.79	a = 78.59, c = 36.82	a = 78.62, c = 37.05	a = 58.71, c = 93.75	a = 58.33, c = 93.80	a = 58.42, c = 93.91
Resolution (Å)	21.8–2.00 (2.06–2.00)	18.52–2.30 (2.37–2.30)	18.0–2.30 (2.37–2.30)	21.8–2.30 (2.33–2.30)	25.3–1.83 (1.87–1.83)	22.9–2.00 (2.05–2.00)	25.3–1.92 (1.96–1.92)
No. of observations	105547	150896	152044	147472	123604	312518	285176
No. of unique observations†	7449	9895	9909	9828	15024	20853	23500
$\langle I/\sigma(I)\rangle$	27.3 (19.6)	51.1 (36.6)	35.0 (26.7)	26.3 (17.6)	23.6 (3.3)	33.7 (13.1)	33.7 (13.6)
Multiplicity	14.2 (4.5)	15.2 (14.7)	15.3 (14.3)	15.0 (13.5)	8.2 (5.4)	15.0 (14.7)	12.1 (12.0)
Completeness (%)	99.9 (99.8)	99.8 (100.0)	99.9 (100.0)	98.7 (95.0)	99.7 (97.6)	99.8 (99.5)	99.5 (93.9)
$R_{\rm merge}$‡ (%)	6.6 (7.6)	6.7 (9.1)	8.6 (11.8)	13.0 (16.5)	8.9 (50.8)	9.3 (23.8)	7.3 (17.0)
Data collected (°)	220	360	360	360	130	360	292
Cryoprotectant solution	1.0 *M* NaCl, NaOAc buffer, 15% ethylene glycol	1.0 *M* CsCl, NaOAc buffer, 15% ethylene glycol	0.625 *M* NaCl, 0.250 *M* $GdCl_3$, NaOAc buffer, 15% ethylene glycol	0.75 *M* NaI, NaOAc buffer, 15% ethylene glycol	Mother liquor, 20% ethylene glycol	Mother liquor, 1.0 *M* CsCl, 20% ethylene glycol	Mother liquor, 0.5 *M* NaI, 20% ethylene glycol
Soaking time (s)	60	300	300	60	60	300	180

† Multiplicity of derivative (native) data sets calculated with Friedel-related reflections treated separately (as equivalent). ‡ $R_{\rm merge} = \sum_{hkl} |I - \langle I\rangle| / \sum_{hkl} I$.

in a cryoprotectant solution containing bromide or iodide anions just before freezing the crystal in a cryogenic nitrogen stream leads to incorporation of these anomalous scatterers into the ordered solvent regions around the protein molecules. Halides provide a rapid and convenient way of protein crystal derivatization. Halides are negatively charged ions that bind preferably to positively charged surface areas of the proteins, competing in binding with waters. It was shown that halide derivatization can be successfully used in rapid solution of macromolecular structures (Dauter *et al.*, 2000).

In the current paper, we suggest the extension of the method of rapid cryoderivatization by incorporating short cryosoaks with monovalent cations (such as Cs^+ or Rb^+) and polyvalent cations (such as lanthanides). One might expect that if monovalent cations (Cs^+ or Rb^+) were used instead of anions in the cryoprotectant solution they would preferentially bind to the negatively charged niches on the surface of the macromolecules, which will be distinct from the halide sites. Additional sites, different from those of both monovalent cations and halides, can also be expected if lanthanide ions (Gd^{3+}, Eu^{3+}, Sm^{3+} or Ho^{3+}) are used in the cryoderivatization procedure. Lanthanide ions require a specific coordination sphere for binding and therefore will not compete with monovalent cations and anions in binding to a protein molecule. Since the overall charge and surface-charge distribution of a given protein molecule depends on the pH of the cryoderivatization solution, positively or negatively charged anomalous scatterers might better suit a particular situation.

Different heavy-ion sites permit combined use of these derivatives in MIR(AS) phasing, even when none of them are strong enough to produce interpretable electron-density maps *via* SAD or SIR(AS) methods.

Just for the sake of comparison, an Se atom, one of the most frequently used heavy atoms for SAD/MAD synchrotron data collection, has an anomalous scattering-factor component $\delta f''$ of up to 10 electrons at the white line of the K absorption edge (0.98 Å). A Gd atom away from the edge at 1.54 Å has an $\delta f''$ of 12.3 electrons. At the same copper-anode characteristic wavelength (1.54 Å), $\delta f''$ of Cs and I atoms are equal to 7.9 and 6.8 electrons, respectively, as estimated by the program *CROSSEC* (Cromer, 1983). This opens the way to using the fast cryoderivatization procedure combined with relatively long wavelength data collection (1.5–1.6 Å), which can be performed far from the absorption edges of the anomalous scatterers at the synchrotron (including medium-energy machines) as well as rotating-anode X-ray sources.

In the present work, we describe the use of the quick cryosoaking procedure with positively and negatively charged anomalous scatterers to phase hen egg-white lysozyme (HEWL) and a novel trypsin inhibitor (TRIN). Caesium chloride and gadolinium (III) chloride were used in the cryoprotectant solution as a source of positively charged anomalous scatterers and sodium iodide as a source of negatively charged anomalous scatterers. Diffraction data were collected at the Brazilian National Synchrotron Light Laboratory, a medium-energy synchrotron ring, using X-rays with the wavelength 1.54 Å.

2. Materials and data acquisition

HEWL (Sigma) was crystallized following standard protocols (hanging drops) at 291 K. The crystallization solution consisted of 20–30 mg ml^{-1} protein, 0.5 *M* NaCl in 0.025 *M* sodium acetate buffer pH 4.5 and the well solution was 1.0 *M* NaCl in 0.050 *M* sodium acetate buffer.

A novel trypsin inhibitor has been extracted from the seeds of *Copaifera langsdorffi* (also known as Copaíba), a tree which is widespread in the central region of Brazil. The inhibitor has a molecular weight of 18 kDa. It was purified to homogeneity using ammonium sulfate precipitation followed by anion-exchange and affinity chromatography. The hanging-drop method performed at pH 4.8 with PEG 8000 as a precipitant was used to obtain well diffracting crystals of TRIN (work to be published).

Sodium iodide, caesium chloride and gadolinium (III) chloride used in the quick cryosoaking derivatization procedure were of analytical grade and were purchased from Sigma and Hampton Research Corp.

Native and derivative data sets for both proteins were collected at the Protein Crystallography beamline (Polikarpov, Oliva *et al.*, 1997; Polikarpov, Perles *et al.*, 1997) at the Brazilian National Synchrotron Light Laboratory (Campinas, SP, Brazil) using a MAR345 image plate.

X-ray diffraction images were integrated with *DENZO* (Otwinowski & Minor, 1997) in *P*1 to allow a better fit of predicted spots and measured reflections. *SCALEPACK* (Otwinowski & Minor, 1997) was used to scale images in the correct space group.

The diffraction data were collected with a high multiplicity resulting from the wide total angular range of rotation (130–360°) and the high symmetry of the crystals. No attempts were made to measure Bijvoet-related reflections close in time or on the same rotation image.

For all derivatives, a single crystal was transferred for a short period of time (60–300 s) to a cryoprotectant solution containing heavy-atom compounds (NaI, CsCl or $GdCl_3$) at different concentrations (0.25–1.0 *M*) and then quickly frozen in a fibre loop in a nitrogen-gas stream at 100 K. The synchrotron-radiation wavelength was set to 1.54 Å. Details of the preparation of the native crystals and derivatives for data acquisition as well as data-collection statistics are given in Table 1. In order to simulate a typical X-ray diffraction experiment, native and derivative HEWL data sets were only collected to 2.0 and 2.3 Å resolution, respectively.

The quality of each data set and the presence of the anomalous signals of Cs, I and Gd atoms were controlled following the $|\Delta F|/F$ and $|\Delta F|/\sigma(\Delta F)$ ratios as a function of resolution for each derivative. The results are summarized in Fig. 1.

3. Results and discussion

3.1. The heavy-atom substructures

The use of anomalous difference Patterson maps to locate heavy-atom sites was successfully employed for the Gd-HEWL derivative. In all other cases direct methods were used.

Scaled intensities for each derivative were submitted to *SnB* 2.1 (Weeks & Miller, 1999), where normalized anomalous differences (diff*E*) were calculated with the program *DREAR* (Blessing & Smith, 1999). Normalized anomalous differences were then used by *SnB* to find the relative positions of the caesium, iodide and gadolinium ions for HEWL derivatives and the caesium and iodide ions for TRIN derivatives.

The location of major heavy-atom sites by direct methods was a straightforward task when all data collected were included in a search. Default parameters suggested by *SnB* were used in most of the cases. In some cases increases in the number of reflections, triple invariants and *SnB* cycles were necessary to successfully find heavy-atom substructure within the equivalent sets of possible origin shifts or enantiomers.

In the case of TRIN, the caesium-derivative data set was used in the heavy-atom search. A total of 500 diff*E* values (Blessing & Smith, 1999) were used to generate 5000 triple invariants. Four random atom positions were generated for each of 1000 trial structures and each trial structure was subjected to 30 *SnB* cycles, each cycle consisting of phase refinement and Fourier filtering. The bimodal distribution of the $R_{\rm min}$ histogram was used to identify the correct solution (Debaerdemaeker & Woolfson, 1983; Hauptman, 1991; De

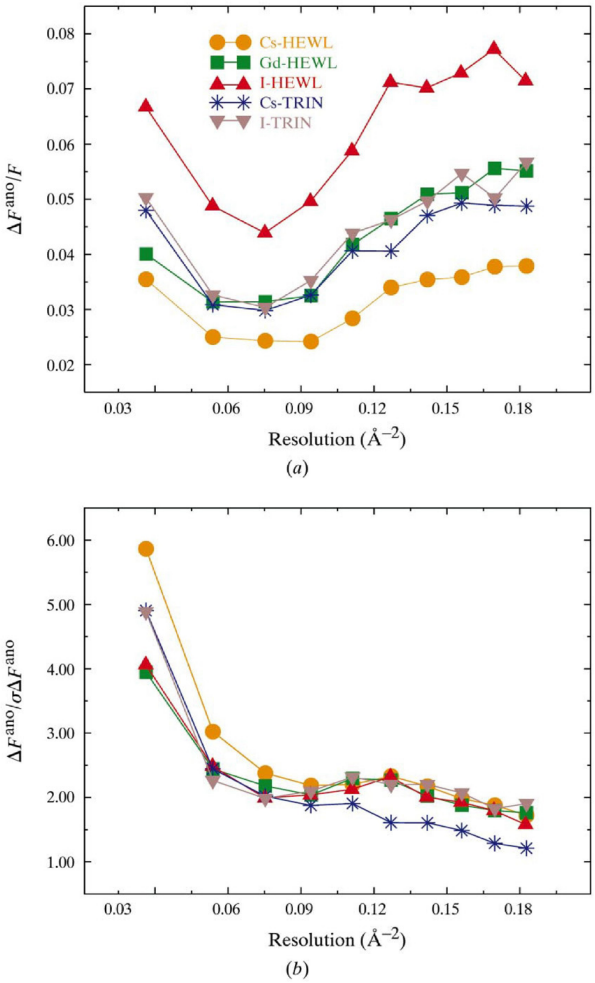


Figure 1
Statistics of the anomalous signal in derivative data sets. (*a*) $|\Delta F|/F$; (*b*) $|\Delta F|/\sigma(\Delta F)$ as a function of resolution. All derivatives were prepared according to the quick cryosoaking procedure for derivatization.

Titta *et al.*, 1994). The relative positions of iodide anions were then located in cross-phased Fourier maps in respect to the origin of caesium sites using the *CNS* package (Brunger *et al.*, 1998).

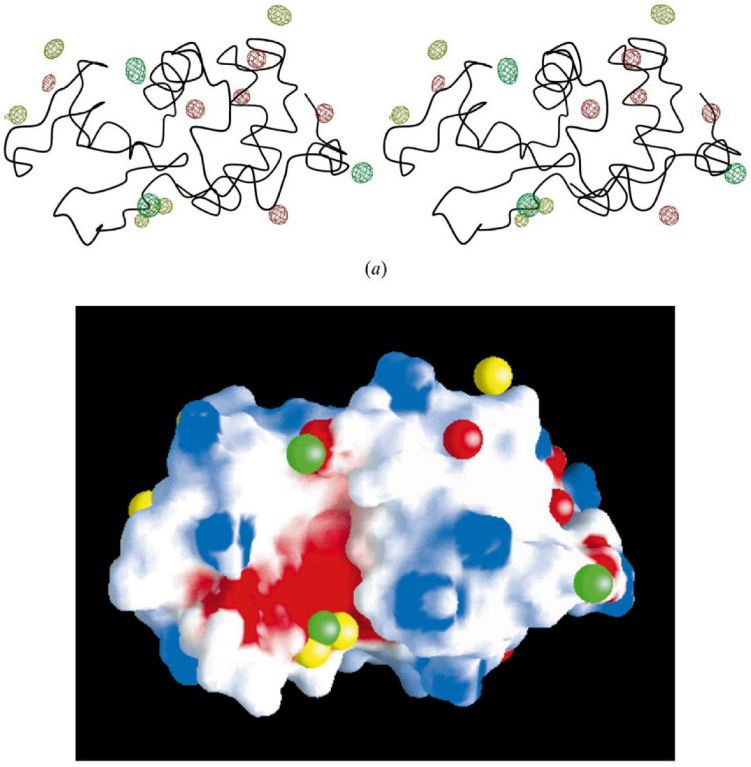
3.2. Phasing

For all derivatives, the heavy-atom substructure obtained directly from *SnB* was initially refined with the *CNS* package using anomalous and isomorphous difference Fourier maps. Refined coordinates were then input into *SHARP* (de La Fortelle & Bricogne, 1997) and different methods (SAD, SIRAS and MIRAS) were used for phase calculation. Density modification with solvent flattening was then performed using the program *SOLOMON* (Abrahams & Leslie, 1996).

Independently, to check the correctness of heavy-ion positions, a lysozyme model (PDB code 1lz8) was refined in *CNS* and *SHELXL* (Sheldrick & Schneider, 1997) against each HEWL data set. The resulting *R*-factor values were in the range 18–19% and $R_{\rm free}$ values were in the range 19–22%. The heavy-atom positions released by *SnB* and refined in *CNS* could then be checked using anomalous difference Fourier maps ($\Delta F^{\rm ano}$, $\varphi_{\rm calc} - 90^\circ$). Cs-HEWL and I-HEWL anomalous difference Fourier maps showed a number of anomalous peaks surrounding the protein surface. On the other hand, only three major sites were confirmed by the Gd-HEWL anomalous map. The superposition of all three anomalous difference Fourier maps and the ion distribution around the HEWL electrostatic surface are shown in Figs. 2(*a*) and 2(*b*), respectively. The figures of merit and a correlation of the electron-density maps obtained by SAD, SIRAS and MIRAS methods to the refined maps ($2mF_{\rm obs} - DF_{\rm calc}$,$\varphi_{\rm calc}$) for HEWL and TRIN are given in Table 2.

The SIRAS phases obtained with the positions of five caesium cations in the Cs-TRIN derivative were used to find six iodide anions in the second derivative. A MIRAS electron-density map was calculated with *SHARP*/*SOLOMON*. A representative part of the map is shown in Fig. 3(*a*). All 11 heavy-atom sites were included in the phasing procedure. Based on this map, we used the program *ARP*/*wARP* (Perrakis *et al.*, 1999) in mode '*warpNtrace*' to build a hybrid model of TRIN. A specific region of the output *wARP* electron-density map and with the built hybrid model is shown in Fig. 3(*b*).

A total of 20 automatic building cycles were performed in *wARP*. In the last cycle, up to 92% of the structure was traced in six chains. Evolution of model building and quality of the model can be followed in Fig. 4. The number of traced residues and the *R* factor for the hybrid model as a function of automatic building cycles is shown.



(*a*)

(*b*)

Figure 2
(*a*) Stereoview of a coil representation of HEWL and anomalous difference Fourier maps of Cs, Gd and I derivatives prepared by the quick cryosoaking procedure, shown in yellow, green and red, respectively. The three maps are contoured at the 10σ level. The figures were prepared using *MOLSCRIPT* (Kraulis, 1991), *Bobscript* (Esnouf, 1997) and *Raster3D* (Merritt & Bacon, 1997). (*b*) Distribution of ions around HEWL electrostatic surface as calculated by *GRASP* (Nicholls *et al.*, 1991). Diffusion of ions into protein crystal channels is relatively fast. Even within a short period of time (a few seconds), anions (or cations) preferentially bind positively (or negatively) charged patches of the macromolecule surface. Caesium, gadolinium and iodide ions are represented as yellow, green and red spheres, respectively. Both figures are in the same orientation.

3.3. Environment of anomalous sites

As can be seen from Fig. 2, gadolinium and caesium cations bind in negatively charged niches at the protein surface, interacting mostly with glutamic and aspartic acid side chains, waters and main-chain carbonyl groups. On the other hand, iodide anions are normally located in other surface areas, preferentially bound to lysine and arginine. For a detailed description of the environment of iodide (halide) anions see Dauter *et al.* (2000). The binding sites are distinct and particular for all three HEWL derivatives.

Caesium cations, similar to iodide anions, are distributed around the protein surface of HEWL. A comparison of the native and caesium-derivative electron-density maps show that caesium cations indeed replace coordinated water molecules, leading to elliptical electron-density patterns where caesium and waters share sites that are close to each other. The five

Table 2
Relevant parameters of electron-density maps obtained by SAD, SIRAS and MIRAS methods using HEWL and TRIN derivatives prepared by the quick cryosoaking procedure.

Phasing method	Data sets used in phasing	Resolution range	FOM before *SOLOMON*	FOM after *SOLOMON*	Correlation†
SAD	I-HEWL	18.00–2.30	0.48	0.91	0.62
SIRAS	Nat/Gd-HEWL	18.00–2.30	0.31	0.96	0.71
SIRAS	Nat/I-HEWL	18.00–2.30	0.62	0.95	0.81
MIRAS	Nat/Cs/Gd-HEWL	18.00–2.30	0.45	0.95	0.64
MIRAS	Nat/Cs/I-HEWL	18.00–2.30	0.63	0.95	0.79
MIRAS	Nat/Cs/Gd/I-HEWL	18.00–2.30	0.75	0.95	0.85
SAD	I-TRIN	22.90–2.00	0.39	0.93	0.54
SIRAS	Nat/Cs-TRIN	22.90–2.00	0.48	0.95	0.51
SIRAS	Nat/I-TRIN	22.90–2.00	0.53	0.96	0.74
MIRAS	Nat/Cs/I-TRIN	22.90–2.00	0.63	0.97	0.84

† Correlation coefficients of TRIN maps were calculated using a reference electron-density map ($2mF_{obs} - DF_{calc}, \varphi_{calc}$) obtained on the basis of current structure built and refined by *wARP* (Perrakis *et al.*, 1999). The structure is more than 92% complete.

prominent caesium sites were found between 2.7 and 3.7 Å from main-chain carbonyl groups, from the negatively charged side chains of aspartate and glutamate residues and from the polar side chains of asparagine and serine residues. A typical example of a caesium site is illustrated in Fig. 5(*a*).

Gadolinium cations require a coordination sphere and are therefore more specific in binding compared with caesium cations and iodide anions. Nevertheless, gadolinium cations are able to bind to negatively charged and polar residues on the surface of HEWL (and other proteins) even in quick cryosoaking experiments. Fig. 5(*b*) shows the major gadolinium site. It is coordinated by six nearest neighbours: the carboxyl group of Asp52 (at a distance of 1.8 Å) and five waters (2.4, 2.6, 2.7, 2.7 and 2.8 Å from the Gd atom).

All other gadolinium cations are found inside channels between symmetry-related proteins or close to O atoms in asparagine and aspartate side chains. Several water molecules take part in the coordination sphere.

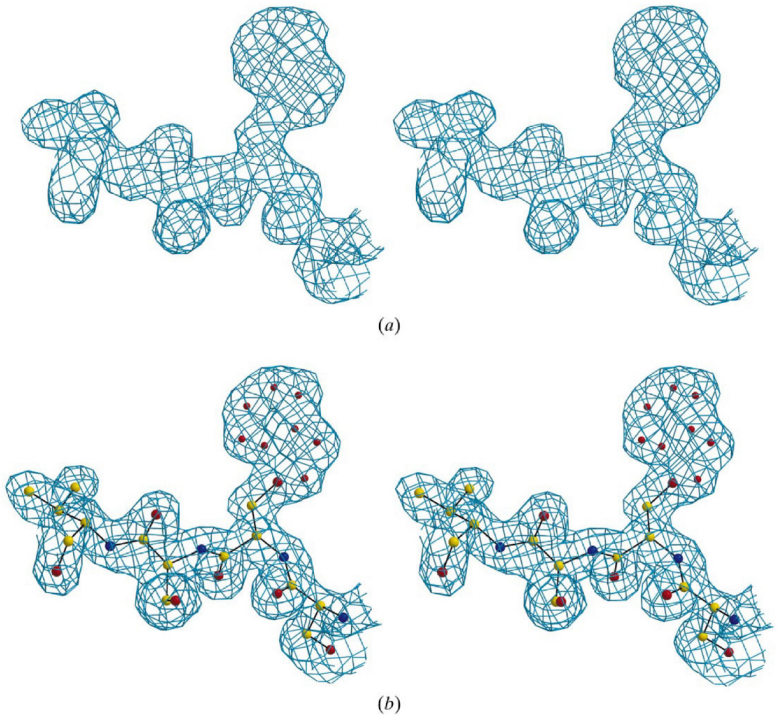


Figure 3
(*a*) Stereoview of a representative part of the TRIN electron-density map after density modification performed by *SOLOMON*. The solvent-flattened electron-density map was used as an input to *wARP* for automatic model building. (*b*) A part of a hybrid model of TRIN built by *wARP* in a ($2mF_{obs} - DF_{calc}, \varphi_{calc}$) electron-density map. Inserted waters (small red spheres) clearly define a tryptophan side chain.

4. Conclusions

In the present work, we show that the monovalent and polyvalent cations Cs^{+} and Gd^{3+} can be successfully used in the quick cryosoaking procedure for derivatization and phasing of protein crystals. Caesium cations as well as monovalent anions seem to be less destructive to protein crystals. Their concentration in the cryoprotectant solution can be increased to 1.0 *M* or even higher without significant deterioration of the protein crystal diffraction quality. On the other hand, the use of lanthanides requires more care. Their concentration in the cryoprotectant solution must be correctly chosen to avoid crystal degradation. Nevertheless, lower concentrations of lanthanides (0.25 *M*; about 100 times higher than the concentration of the heavy metal in traditional soaks) associated with a slightly longer time of soaking (up to 300 s) seem to be effective.

The addition of cations to the quick cryosoaking derivatization procedure opens up the possibility of fast and rational searching for derivatives. For a particular protein, its pI and the pH range of the crystallization solution are normally fixed and therefore cations and/or anions may be chosen to best suit the derivatization experiment. The residues most frequently involved in binding of the anomalous scatterers in quick cryosoaking experiments with monovalent anions are lysine, arginine and histidine (positively charged side chains) and with monovalent cations are aspartate and glutamate (negatively charged side chains). It should be

emphasized that at lower pHs and for basic proteins, halides may be particularly suitable for quick cryosoaking procedures. On the other hand, at alkaline pHs and for proteins with high contents of glutamic and aspartic acid residues, positively charged anomalous scatterers may be more suitable for quick cryoderivatization.

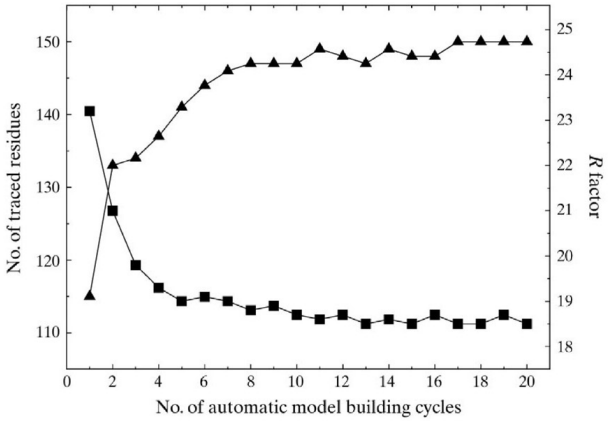


Figure 4
Number of traced residues and the crystallographic *R* factor as a function of *warpNtrace* cycles on the example of TRIN. In the last cycle 150 residues were traced in six chains. More than 92% of the structure were correctly built.

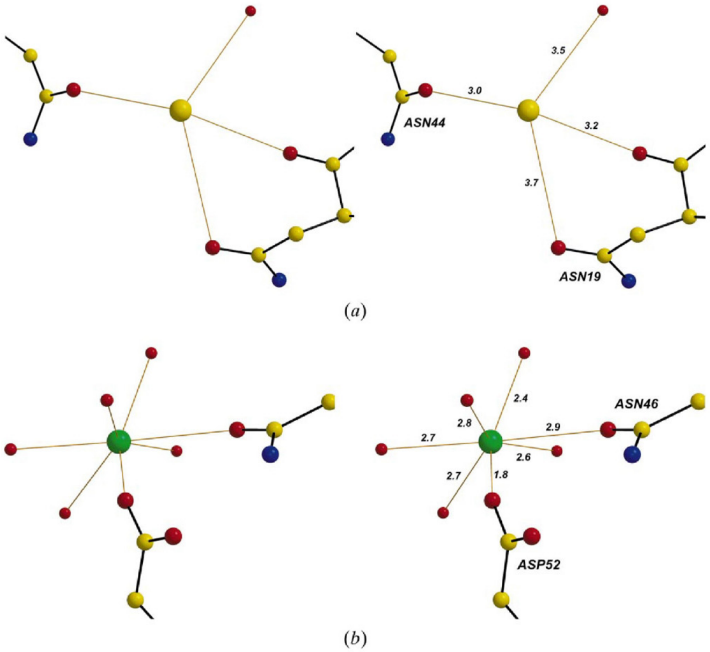


Figure 5
Stereoview of the caesium- and gadolinium-binding sites in HEWL derivatives prepared by quick cryosoaking. (*a*) A caesium cation (yellow sphere) relatively close to the O atoms (red spheres) of the Asn44 and Asn19 side chains, a carbonyl group and a water molecule. (*b*) The major gadolinium site (green sphere) is coordinated by six neighbouring atoms: Asp52 carboxyl group (at a distance of 1.8 Å) and five water molecules (at a distances of 2.4, 2.6, 2.7, 2.7 and 2.8 Å).

The absorption edges of caesium and iodide ions lie beyond the easily accessible wavelength range for typical diffraction experiments. However, even at the wavelength of 1.54 Å, caesium, iodine and gadolinium have a significant anomalous signal which can be utilized in the SAD or SIRAS/MIRAS methods. Since copper-anode X-ray sources can be used in such experiments, it may very well be possible for experimenters to validate their quick cryosoaked derivatives prior to synchrotron visits.

Synchrotron radiation can be used to increase the anomalous effect in gadolinium derivatives since its *L* absorption edge can be easily reached at standard beamlines. Besides, Rb (rubidium) can be used in place of caesium cations in the same way as bromine was used instead of iodine in MAD experiments (Dauter *et al.*, 2000). At the dedicated MAD beamlines of high-energy synchrotron X-ray sources, the *K* absorption edge of rubidium (0.82 Å) can easily be reached (A. Joachimiak, personal communication). Although the anomalous scattering factors of atoms which can potentially be used for quick cryosoaking derivatization procedure are comparable or even higher than those of common Se-Met derivatives, the occupancies of quick cryosoaked ions are generally less than unity (normally close to 0.7–0.5). Therefore, the ultimate molecular weight of the protein that can be solved using quick cryosoaking procedure needs further experimental studies.

The quick cryosoaking approach with halides and cations requires little preparative effort and may be particularly applicable for high-throughput crystallographic projects. Novel proteins that require rapid three-dimensional structure elucidation and diffract beyond 2.0 Å resolution are potential targets for these approach. The high-resolution data increase the chances of an automatic building of the three-dimensional model resulting in a structure solution within the shortest period of time.

The authors are grateful to J. R. Brandão Neto for help with data collection and Valeria Forrer and Renata Carmona e Ferreira for help with crystallization of the proteins. Financial help from CNPq and FAPESP *via* grants 99/03387-4 and 98/06218-6 is acknowledged.

References

Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30–42.
Blessing, R. H. & Smith, G. D. (1999). *J. Appl. Cryst.* **32**, 664–670.
Bragg, W. L. & Perutz, M. F. (1954). *Proc. R. Soc. London Ser. A*, **225**, 315.
Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.
Cromer, D. T. (1983). *J. Appl. Cryst.* **16**, 437–438.
Dauter, Z., Dauter, M. & Rajashankar, K. R. (2000). *Acta Cryst.* D**56**, 232–237.

Debaerdemaeker, T. & Woolfson, M. M. (1983). *Acta Cryst.* A**39**, 193–196.
De Titta, G. T., Weeks, C. M., Thuman, P., Miller, R. & Hauptman, H. A. (1994). *Acta Cryst.* A**50**, 203–210.
Esnouf, R. M. (1997). *J. Mol. Graph.* **15**, 133–138.
Green, D. W., Ingram, V. M. & Perutz, M. F. (1954). *Proc. R. Soc. London Ser. A*, **225**, 287.
Hauptman, H. A. (1991). *Crystallographic Computing 5: from Chemistry to Biology*, edited by D. Moras, A. D. Podjarny & J. C. Thierry, pp. 324–332. Oxford: IUCr/Oxford University Press.
Hendrickson, W. A. (1991). *Science*, **254**, 51–58.
Kraulis, P. (1991). *J. Appl. Cryst.* **24**, 946–950.
La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.
Matthews, B. W. (1966). *Acta Cryst.* **20**, 82–86.
Merritt, E. A. & Bacon, D. J. (1997). *Methods Enzymol.* **277**, 505–524.
Nicholls, A., Sharp, K. A. & Honig, B. (1991). *Proteins: Struct. Funct. Genet.* **11**, 281–296.
North, A. C. T. (1965). *Acta Cryst.* **18**, 212–216.
Otwinowski, Z. & Minor, M. (1997). *Methods Enzymol.* **276**, 307–326.
Perrakis, A., Morris, R. J. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
Polikarpov, I., Oliva, G., Castellano, E. E., Garratt, R., Arruda, P., Leite, A. & Craievich, A. (1997). *Nucl. Instrum. Methods A*, **405**, 159–164.
Polikarpov, I., Perles, L. A., de Oliveira, R. T., Oliva, G., Castellano, E. E., Garratt, R. & Craievich, A. (1997). *J. Synchrotron Rad.* **5**, 72–76.
Sheldrick, G. M. & Schneider, T. R. (1997). *Methods Enzymol.* **277**, 319–343.
Smith, J. L. (1991). *Curr. Opin. Struct. Biol.* **1**, 1002–1011.
Weeks, C. M. & Miller, R. (1999). *J. Appl. Cryst.* **32**, 120–124.

6.4 Aumento da capacidade de resolução

Apesar de a estrutura cristalográfica de uma proteína inédita ter sido resolvida por meio da técnica de crio derivatização rápida com halogênios e/ou metais alcalinos com o uso da radiação síncrotron proveniente da linha de luz CPr do LNLS, é extremamente importante que o mesmo procedimento seja aplicado para a resolução da estrutura tridimensional de outros cristais obtidos das mais diversas condições de cristalização e formados por macromoléculas de diferentes organismos com os mais variados tamanhos e funções. Esse fato permitiria não só adquirir a experiência e a segurança necessárias para a resolução de estruturas inéditas como também verificar a aplicabilidade do método de crio derivatização rápida em outros casos.

Para testar essas idéias, procurou-se aplicar o método em outros casos estudados no grupo de Cristalografia de Proteínas do LNLS. Duas proteínas inéditas foram então selecionadas: interleucina-22 humana e β-galactosidase de *Penicillium sp*. Todas as etapas de processamento e análise dos dados de difração, de obtenção das coordenadas dos espalhadores anômalos, do cálculo e da análise das fases dos fatores de estrutura e, finalmente, da interpretação dos mapas de densidade eletrônica foram feitas com extremo cuidado nos dois casos. Contudo, a maneira como essas etapas foram vencidas não foi, necessariamente, idêntica. Em alguns momentos, vários programas foram usados para uma mesma tarefa e até mesmo uma combinação de procedimentos mostrou ser eficiente para a resolução das estruturas. O detalhamento das etapas para cada um dos casos seria uma tarefa extremamente laboriosa e documentativa, porém, não estaria diretamente focalizada no ponto principal da seção: mostrar a aplicabilidade do método de crio derivatização rápida para a resolução de estruturas inéditas dos mais variados tipos. Em vista disso, optou-se por apresentar apenas alguns detalhes fundamentais e característicos de cada um dos conjuntos.

A primeira macromolécula, interleucina-22 (IL-22), é uma proteína nova identificada primeiramente em células de ratos (mIL-22) (DUMOUTIER, LOUAHED & RENAULD, 2000) e, logo em seguida, em células humanas (hIL-22) (DUMOUTIER *et al*., 2000). Ela foi inicialmente classificada como uma citocina devido a várias características: a presença de um peptídeo sinal hidrofóbico em seu N-terminal, um tamanho aproximado de 20 kDa e uma identidade de aminoácidos, ainda que baixa, com interleucina-10 (IL-10); uma citocina bem conhecida. De fato, a proteína faz parte da família de citocinas IL-10, que também inclui IL-19, IL-20, IL-24 além da própria IL-10 e vários de seus homólogos virais. As únicas estruturas tridimensionais conhecidas de membros da família IL-10, até o início deste estudo, eram as de IL-10 humana e do vírus Epstein-Barr (EBV). Sob o ponto de vista biológico, existem alguns indícios de que essa nova citocina esteja

relacionada com alguns processos de resposta inflamatória (DUMOUTIER *et al*., 2000). Além disso, outros estudos apontam provável ligação dessa proteína com alergia e asma (TEMANN *et al*., 1998; MCLANE *et al*., 1998).

Tabela 6-3. Detalhes sobre a preparação e a estatística dos dados de difração de raios-X dos cristais nativo e derivados de interleucina-22 humana.

Cristalização			
Líquido-mãe	0,9 M sódio *tartrate*, detergente Triton X-100,0,1 M HEPES (tampão HEPES), pH 7,5		
Temperatura	18 ^{o}C		
Solução de proteína	10 mg/ml, 20 mM MES (tampão MES), pH 5,4		
Preparação do cristal	**Nat-hIL-22**	**I-hIL-22**	**Hg-hIL-22**[*]
Solução crioprotetora	líquido-mãe 15% etilenoglicol	0,125 M NaI líquido-mãe 15% etilenoglicol	5 mM $HgCl_2$ líquido-mãe 15% etilenoglicol
Tempo de banho[**] (s)	30	180	10 horas
Coleta de dados	**Nat-hIL-22**	**I-hIL-22**	**Hg-hIL-22**
Comprimento de onda (Å)	1,54	1,54	1,54
Grupo espacial	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
Parâmetros de célula (Å)	a = 55,43 b = 61,61 c = 73,47	a = 56,05 b = 61,78 c = 73,63	a = 56,04 b = 61,71 c = 74,61
Resolução (Å)	21,7-2,00 (2,05-2,00)	21,8-1,92 (1,96-1,92)	22,4-1,90 (1,97-1,90)
N. de reflexões	61846	182876	55855
N. de reflexões únicas[#]	16382	37777	29854
$\langle I/\sigma(I)\rangle$	14,5 (3,8)	13,4 (3,1)	8,2 (2,1)
Multiplicidade	3,8 (3,4)	4,8 (4,3)	1,9 (1,7)
Completeza	92,7 (82,3)	99,9 (99,7)	75,9 (77,9)
R_{merge}	8,2 (35,0)	11,7 (43,9)	10,0 (49,9)
Oscilação total (graus)	103	248	70
Fonte de raios-X	Linha CPr, LNLS	Linha CPr, LNLS	Linha X4A, NSLS[##]
Detector de raios-X	mar345 (ip)[+]	mar345 (ip)	quantum4 (ccd)[++]

[*] *Durante o processo de obtenção das fases, este derivado foi identificado como um quasi-nativo já que não apresentou nenhuma incorporação dos átomos de mercúrio na estrutura cristalográfica.*

[**] *Tempo de imersão do cristal na solução crioprotetora.*

[#] *A multiplicidade dos conjuntos derivados (nativo) foi calculada tratando os pares de Friedel como reflexões distintas (equivalentes).*

[##] *National Synchrotron Light Source, Upton – N.Y., Estados Unidos.*

[+] *Image plate (ip).*

[++] *Charge coupled device (ccd).*

Os valores estatísticos para as camadas com mais alta resolução são mostrados entre parênteses.

Os primeiros cristais de interleucina-22 humana foram obtidos pela técnica da gota suspensa (NAGEM *et al*., 2002). Os primeiros conjuntos de dados coletados na linha de luz CPr do LNLS permitiram identificar que esses cristais pertenciam ao grupo espacial $P2_12_12_1$ e difratavam além de 2,0 Å de resolução. A estrutura tridimensional desta macromolécula foi resolvida pela técnica SIRAS, com a coleta dos dados de difração de um cristal nativo e um derivado de iodo preparado

pelo método de crio derivatização rápida (tabela 6.3). Um segundo derivado de mercúrio, obtido pelo método de derivatização tradicional, foi coletado na linha de luz X4A no *National Synchrotron Light Sourc*e. Contudo, este último derivado não apresentou uma incorporação efetiva dos átomos de mercúrio e, por isso, não foi utilizado para a obtenção de fases.

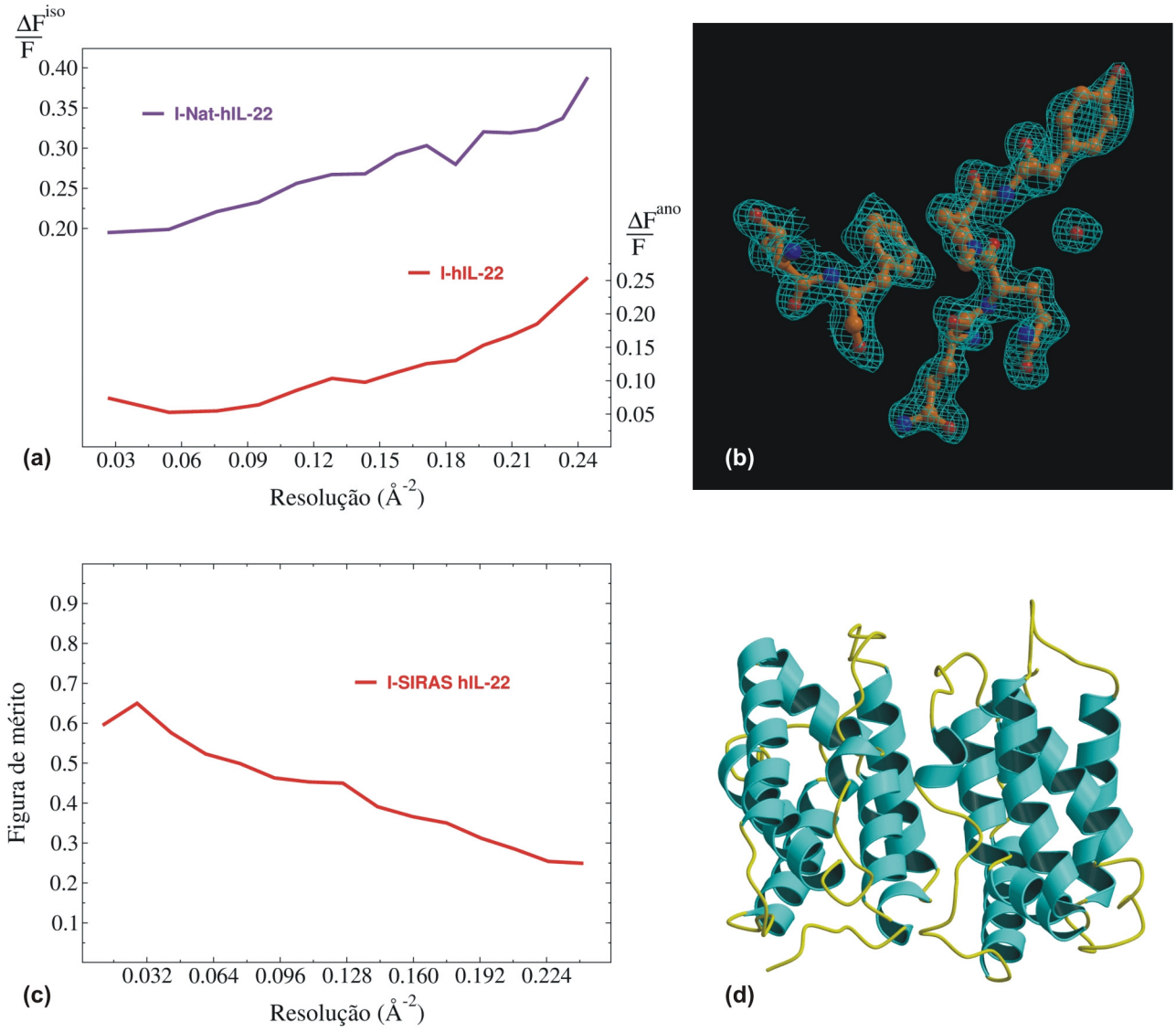


Figura 6-16. Obtenção da estrutura cristalográfica da interleucina-22 humana. (a) Comportamento das quantidades $\Delta F^{iso}/F$ e $\Delta F^{ano}/F$ por faixa de resolução para os cristais coletados de interleucina-22. (b) Mapa de densidade eletrônica em uma região arbitrária da molécula. Este mapa foi obtido com as fases experimentais após os procedimentos de modificação de densidade eletrônica. (c) Valor médio da figura de mérito por faixa de resolução para as fases I-SIRAS. (d) Modelo tridimensional do dímero cristalográfico da hIL-22. Ao contrário do dímero da IL-10, esta proteína não apresenta um entrelaçamento de domínios. Para mais detalhes, veja o artigo anexo, no final desta seção.

Os processos para obtenção das coordenadas atômicas dos espalhadores anômalos, determinação das fases dos fatores de estrutura, cálculo dos mapas de densidade eletrônica e

posterior construção e refinamento do modelo tridimensional da proteína foram executados de maneira similar aos que já haviam sido feitos para os demais cristais. Contudo, um detalhe curioso a respeito da obtenção dessa estrutura merece ser relatado.

É possível verificar, pela tabela 6.3, que o conjunto derivado I-hIL-22 é, além de redundante, completo. O conjunto nativo (Nat-hIL-22), ao contrário, não foi coletado com máxima completeza devido a um erro durante a coleta dos dados. Acredita-se que a falta de completeza tenha sido o principal fator que impossibilitou a construção automática do modelo nativo. Contudo, a utilização do conjunto de dados do derivado de iodo permitiu que o modelo pudesse ser construído e, logo em seguida, o mesmo modelo serviu como base para a estrutura nativa. Além da redundância, a completeza exerceu um papel fundamental na interpretação dos mapas de densidade eletrônica. A ausência de determinadas reflexões, principalmente as mais fortes, podem causar falhas de conectividade nos mapas de densidade eletrônica.

Na figura 6.16 são mostrados, além dos parâmetros utilizados para analisar o conjunto de dados e as fases dos fatores de estrutura, o mapa de densidade eletrônica e o modelo já refinados da estrutura do cristal nativo da interleucina-22 humana em uma região arbitrária da célula unitária. Além disso, uma representação tridimensional da estrutura terciária do dímero de hIL-22 encontrado na unidade assimétrica é mostrada. Outros detalhes sobre a cristalização, a estrutura tridimensional e as implicações biológicas dessa nova proteína podem ser encontrados ao final desta seção, em dois artigos científicos já publicados com os resultados deste trabalho.

A outra proteína, β-galactosidase de *Penicillium sp.* (Psp-β-gal), é classificada como uma carboidrase. As β-galactosidases (E.C 3.2.1.23) são enzimas que hidrolisam as ligações β-1,3 e β-1,4 da galactose em poli e oligossacarídeos. É sabido que muitas carboidrases catalisam não só a reação de hidrólise, mas também a reação reversa de condensação ou transglicosilação. Esta propriedade, observada em β-galactosidases de vários organismos, tem sido usada para a síntese enzimática de oligossacarídeos que contêm galactose. A β-galactosidase extracelular de *Penicillium sp*. possui uma alta atividade de transglicosilação para p-nitrofenil, β-D-galactopiranose, lactose e metil-β-D-galactopiranose, aparentando ser uma ferramenta poderosa para diversas reações de síntese enzimática (NEUSTROEV *et al*., 2000).

Apesar do conhecimento de algumas funções biológicas e propriedades físico-químicas da Psp-β-gal, sua seqüência de aminoácidos ainda é desconhecida. Mesmo sendo uma proteína relativamente grande (aproximadamente 110 kDa), ela foi cristalizada com sucesso antes mesmo do início deste estudo (NEUSTROEV *et al*., 2000). Contudo, foi necessária a preparação de novos

cristais da proteína com o uso do método da gota suspensa e utilizando PEG 8000 como precipitante. Após inúmeras tentativas de cristalização e coleta de dados, quatro conjuntos foram usados ao longo da resolução da estrutura tridimensional dessa macromolécula: um conjunto de dados de um cristal nativo e três de derivados. Dois cristais derivados foram preparados pelo método de crio derivatização rápida com iodeto de sódio e cloreto de césio e o outro derivado foi preparado pelo método tradicional de derivatização com UO_2Ac_2. Os detalhes dessas coletas de dados podem ser vistos na tabela 6.4.

Tabela 6-4. Detalhes sobre a preparação e a estatística dos dados de difração de raios-X dos cristais nativo e derivados da β-galactosidase de *Penicillium sp*..

Cristalização				
Líquido-mãe	15% PEG 8000, 50 mM fosfato de sódio (tampão fosfato), pH 4,0			
Temperatura	18 °C			
Solução de proteína	5-10 mg/ml			
Preparação do cristal	**Nat-β-GAL**	**I-β-GAL**	**Cs-β-GAL**	**U-β-GAL**
Solução crioprotetora	líquido-mãe 15% etilenoglicol	0,5 M NaI líquido-mãe 15% etilenoglicol	0,33 M CsCl líquido-mãe 15% etilenoglicol	50 mM UO_2Ac_2 líquido-mãe 15% etilenoglicol
Tempo de banho* (s)	30	330	60	12 horas
Coleta de dados	**Nat-β-GAL**	**I-β-GAL**	**Cs-β-GAL**	**U-β-GAL**
Comprimento de onda (Å)	1,54	1,54	1,54	1,54
Grupo espacial	$P4_3$	$P4_3$	$P4_3$	$P4_3$
Parâmetros de célula (Å)	a = 110,96 c = 161,05	a = 110,62 c = 159,89	a = 110,83 c = 160,92	a = 111,01 c = 161,33
Resolução (Å)	22,3-1,90 (1,93-1,90)	23,4-2,10 (2,15-2,10)	27,0-2,10 (2,15-2,10)	22,2-2,40 (2,46-2,40)
N. de reflexões	639065	944162	614561	608673
N. de reflexões únicas**	155025	224036	223388	148232
$\langle I/\sigma(I)\rangle$	13,9 (3,2)	12,5 (3,1)	8,4 (2,4)	10,0 (2,8)
Multiplicidade	4,1 (3,9)	4,2 (4,2)	2,8 (2,7)	4,1 (4,0)
Completeza	99,9 (100,0)	99,9 (100,0)	99,6 (99,6)	98,6 (95,7)
R_{merge}	10,5 (40,0)	10,7 (48,1)	12,3 (47,0)	13,0 (53,1)
Oscilação total (graus)	105	240	134	215
Fonte de raios-X	Linha CPr, LNLS	Linha CPr, LNLS	Linha CPr, LNLS	Linha CPr, LNLS
Detector de raios-X	mar345 (ip)#	mar345 (ip)	mar345 (ip)	mar345 (ip)

* Tempo de imersão do cristal na solução crioprotetora.
** A multiplicidade dos conjuntos derivados (nativo) foi calculada tratando os pares de Friedel como reflexões distintas (equivalentes).
\# Image plate (ip).
Os valores estatísticos para as camadas com mais alta resolução são mostrados entre parênteses.

Apesar de os quatro conjuntos de dados terem sido coletados, a estrutura da Psp-β-gal foi inicialmente obtida com apenas dois conjuntos pela técnica I-SIRAS. Entretanto, resultados similares foram obtidos com os demais conjuntos de dados pelas técnicas Cs-SIRAS e U-SIRAS.

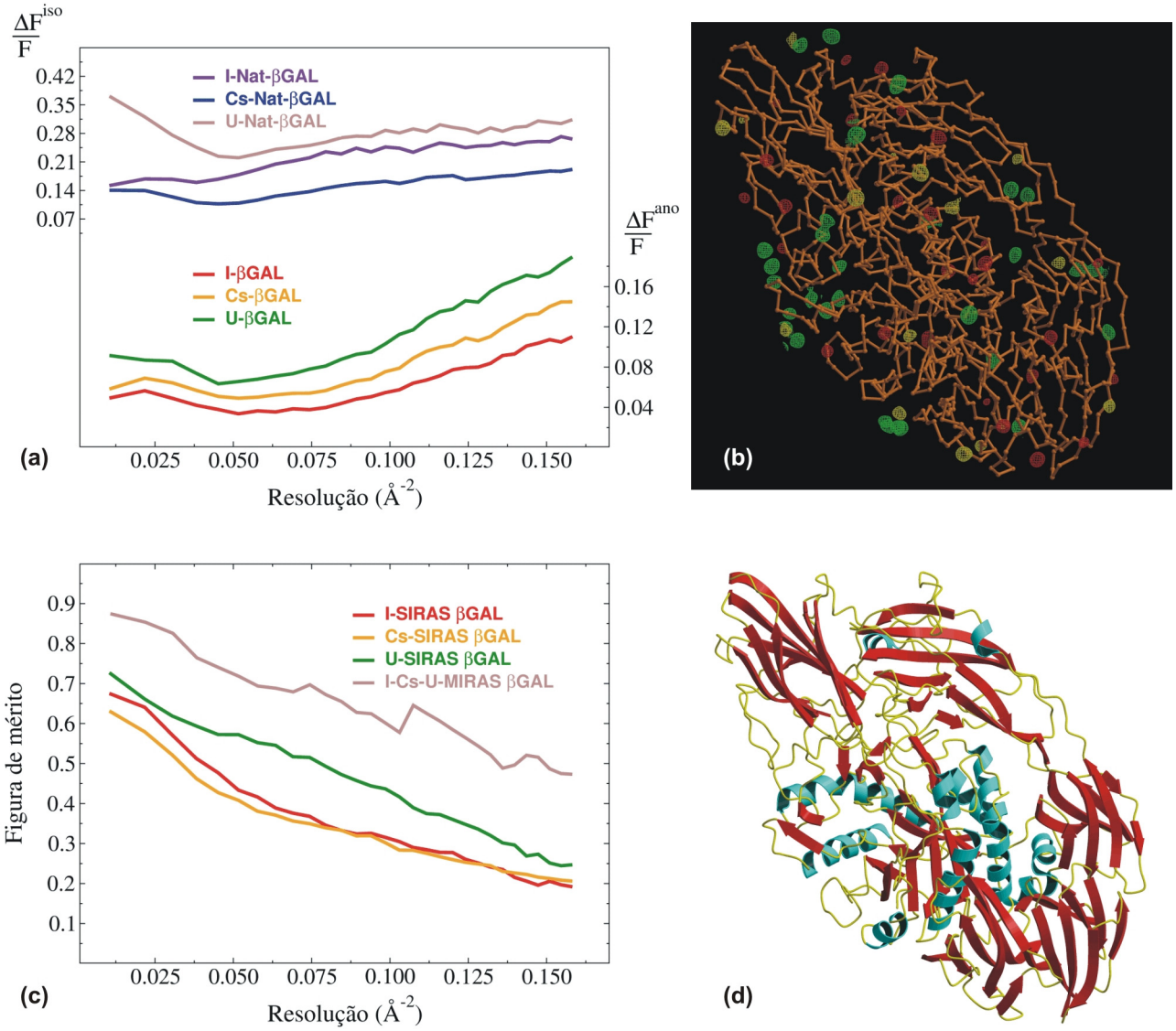


Figura 6-17. Detalhes sobre a obtenção do modelo tridimensional da β-galactosidase de *Penicillium sp*. por difração de raios-X. (a) Comportamento das quantidades $\Delta F^{iso}/F$ e $\Delta F^{ano}/F$ por faixa de resolução para os cristais coletados de β-galactosidase. (b) Superposição de três mapas de Fourier de diferença anômala e o modelo tridimensional da β-galactosidase. Os átomos de iodo, césio e urânio podem ser vistos em vermelho, amarelo e verde, respectivamente. (c) Valor médio da figura de mérito por faixa de resolução para as fases I-SIRAS, Cs-SIRAS, U-SIRAS e MIRAS. (d) Modelo tridimensional do monômero da β-galactosidase. Esta é a maior estrutura resolvida, até o momento, com o uso da técnica de crio derivatização rápida.

A incorporação dos átomos pesados na estrutura dos cristais nativo foi altamente eficaz (figura 6.17). A presença de um conteúdo de solvente de aproximadamente 70% fez com que a difusão dos íons fosse extremamente rápida, o que possibilitou com que inúmeros sítios de ligação fossem observados. O fato permitiu que a estrutura cristalográfica da proteína fosse resolvida também pela técnica SAD com a coleta de um único conjunto de dados de um derivado. Apesar de a estrutura não ter sido obtida inicialmente dessa forma, a possibilidade de resolver uma estrutura

deste tamanho (quase 1000 resíduos de aminoácidos) por SAD com a técnica de crio derivatização rápida amplia a faixa de aplicação da nova técnica. De fato, a estrutura tridimensional da Psp-β-gal é a maior resolvida em todo o mundo até o momento com o uso da técnica de crio derivatização rápida.

Um detalhe curioso a respeito desta proteína é que, de maneira similar ao inibidor de tripsina de *Copaifera langsdorffii*, sua seqüência de aminoácidos ainda é desconhecida. Contudo, com o uso de um mapa de densidade eletrônica obtido por MIRAS com os quatro conjuntos de dados, 971 aminoácidos da cadeia polipeptídica foram inicialmente identificados e usados para uma busca por proteínas homólogas nos bancos de dados. O resultado permitiu, além de identificar com maiores chances de acerto a seqüência de aminoácidos da β-galactosidase de *Penicillium sp.*, classificá-la como uma glicosil hidrolase da família 35. De fato, essa é a primeira estrutura de uma carboidrase dessa família. Devido à presença de uma galactose no sítio ativo da β-galactosidase em dois dos cristais, foi possível identificar os prováveis aminoácidos responsáveis pela reação enzimática potencializada pela enzima. Mais detalhes podem ser encontrados em um artigo, já submetido para publicação, que se encontra no final desta seção.

Além dos trabalhos descritos até o momento, o autor desta tese participou, direta e indiretamente, de alguns projetos junto a outros pesquisadores. Apesar de os resultados alcançados em muitas dessas colaborações não constituírem o objetivo principal deste projeto, eles foram essenciais para a formação científica do doutorando. Em vista disso, as publicações originadas dessas parcerias serão apresentadas no final desta tese.

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Crystallization and synchrotron X-ray diffraction studies of human interleukin-22

R. A. P. Nagem,[a,b] K. W. Lucchesi,[a] D. Colau,[c] L. Dumoutier,[c] J.-C. Renauld[c] and I. Polikarpov[a,d]*

[a]Laboratório Nacional de Luz Síncrotron, Caixa Postal 6192, CEP 13083-970, Campinas, SP, Brazil, [b]Instituto de Física Gleb Wataghin, UNICAMP, Caixa Postal 6165, CEP 13083-970, Campinas, SP, Brazil, [c]Ludwig Institute for Cancer Research, Brussels Branch and the Experimental Medicine Unit, Christian de Duve Institute of Cellular Pathology, Université de Louvain, Brussels, Belgium, and [d]Instituto de Física de São Carlos, Universidade de São Paulo, Av. Trabalhador Sãocarlense 400, CEP 13560-970, São Carlos, SP, Brazil

Correspondence e-mail: ipolikarpov@if.sc.usp.br

Received 11 December 2001
Accepted 19 January 2002

Human interleukin-22, a novel member of the cytokine family, has been crystallized in hanging drops using the vapour-diffusion technique. Preliminary X-ray diffraction experiments using synchrotron radiation reveal that the protein crystallizes in space group $P2_12_12_1$, with unit-cell parameters a = 55.44, b = 61.62, c = 73.43 Å, and diffracts beyond 2.00 Å resolution.

1. Introduction

Interleukin-22 (IL-22), also known as interleukin-10-related T-cell-derived inducible factor (IL-TIF), a protein which shares 22% identity with interleukin-10 (IL-10), has been recently identified in interleukin-9 (IL-9) induced murine T cells (Dumoutier, Louahed *et al.*, 2000) and subsequently in human cells (Dumoutier, Van Roost, Colau *et al.*, 2000) as a novel member of the cytokine family. *In vitro* experiments showed that expression of IL-22 is rapidly induced by IL-9 in T cells, with maximal levels reached in 1 h, and that this induction does not require protein synthesis as the upregulation of the IL-22 mRNA is not blocked by cycloheximide (Dumoutier, Louahed *et al.*, 2000). This new protein, like other cytokines, has an approximate size of 20 kDa and an N-terminal hydrophobic signal peptide. Human and mouse IL-22 (hIL-22 and mIL-22) share 79% amino-acid sequence identity and both have 179 amino-acid residues.

Cytokines exert their actions by binding to specific cell-surface receptors, which leads to the activation of cytokine-specific signal transduction pathways. Two distinct receptor chains from the class II cytokine receptor family have been identified as taking part in the IL-22–receptor complex (Xie *et al.*, 2000; Kotenko *et al.*, 2001). These receptors are the CRF2-9 (IL22-R) and the CRF2-4 (IL-10R2 or IL-10Rβ), the latter also being a functional component of the IL-10 signalling complex (Kotenko *et al.*, 1997). This is the first scientific observation within the class II cytokine receptor family of a receptor being utilized as a component of multiple distinct cytokine signalling complexes.

It has been demonstrated *in vivo* that mIL-22 expression is rapidly increased in several mice organs after lipopolysaccharide (LPS) injection, indicating that the role of IL-22 may not be restricted to the immune system and that it is also involved in inflammatory responses (Dumoutier, Van Roost, Colau *et al.*, 2000). The latter is corroborated by the fact that IL-22 stimulation of HepG2 human hepatoma cells up-regulated the production of acute phase reactants such as serum amyloid A, α1-antichymotrypsin and haptoglobin (Dumoutier, Van Roost, Colau *et al.*, 2000). Moreover, there are data linking IL-22 to asthma and allergy owing to the probable involvement of IL-9 in these two pathologies (Temann *et al.*, 1998; McLane *et al.*, 1998; Levitt *et al.*, 1999). Other evidence for this linkage could be obtained at the DNA level. The hIL-22 gene is located on chromosome 12q (Dumoutier, Van Roost, Ameye *et al.*, 2000), where several loci potentially linked to asthma have been identified (Cookson, 2000).

In this paper, we describe the crystallization and results of preliminary X-ray diffraction studies of recombinant hIL-22 (rhIL-22) at 2.00 Å resolution.

2. Protein purification

Human IL-22 (corresponding to amino acids Gln29–Ile179; without the signal peptide) was cloned and expressed in *Escherichia coli* strain BL21 codon plus-(DE3)-RIL (Stratagene) as described in Dumoutier, Van Roost, Colau *et al.* (2000). Cells containing rhIL-22 were disrupted and inclusion bodies were collected by centrifugation. They were washed extensively first with 50 m*M* Tris–HCl, 100 m*M* NaCl, 1 m*M* EDTA, 1 m*M* DTT, 0.5%(*w*/*v*) deoxycholate (DOC) pH 8 and finally with the same buffer without detergent. Inclusion bodies were solubilized overnight in 8 *M* urea, 50 m*M* MES, 10 m*M* EDTA, 0.1 m*M* DTT pH 5.5. The rhIL-22 protein was refolded by direct dilution of the solubilized inclusion bodies in the following folding mixture: 100 µg ml^{-1} rhIL-22, 100 m*M* Tris–HCl, 2 m*M* EDTA, 0.5 *M* L-arginine, 1 m*M* reduced glutathione, 0.1 m*M* oxidized glutathione pH 8. The solution was incubated for 72 h. The folding

mixture was then concentrated by ultrafiltration in an Amicon chamber with a YM3 membrane before purification on a Superdex75 (Amersham Pharmacia Biotech) gel-filtration column. The protein was eluted with 25 m*M* MES, 150 m*M* NaCl pH 5.4. Recombinant human IL-22 peak fractions were concentrated to 5 mg ml^{-1} with a YM3 Amicon membrane and desalted using a Hi-Prep 26/10 column (Amersham Pharmacia) with elution buffer containing 10 m*M* MES pH 5.4. The protein was concentrated again to 5 mg ml^{-1} and lyophilized in 1 mg fractions.

3. Crystallization and data collection

Initial crystallization conditions were screened by the sparse-matrix method using the macromolecular crystallization reagent kits Crystal Screen I and II (Hampton Research) at 291 K. Small crystals were found in conditions number 18, 26 and 29 of the Crystal Screen I kit. In each trial, a hanging drop of 1 µl protein solution (10 mg ml^{-1} in 20 m*M* MES buffer pH 5.4) was mixed with 1 µl precipitant solution and equilibrated against a reservoir containing 500 µl precipitant solution. Several attempts to improve crystal quality were performed, including pH and precipitant concentration refinement, detergent addition and macroseeding. Well diffracting crystals were finally obtained using a precipitant solution consisting of 0.9 *M* sodium tartrate, Triton X-100 detergent and 0.1 *M* HEPES pH 7.5. Two synchrotron X-ray diffraction data sets, one native and one derivative, have been collected for this study. Both data sets were collected using a 345 mm MAR Research image-plate detector at the Brazilian National Synchrotron Light Laboratory protein crystallography beamline (Polikarpov, Perles *et al.*, 1997; Polikarpov, Oliva *et al.*, 1997) by the oscillation method at 100 K. The native crystal was immersed for 30 s in a cryocooling solution (precipitant solution with 15% ethylene glycol), mounted in a rayon loop and flash-cooled to 100 K in a cold nitrogen stream. An iodine derivative was prepared following the quick cryosoaking derivatization procedure (Dauter *et al.*, 2000; Nagem *et al.*, 2001). A single crystal was immersed for 3 min in a cryocooling solution (the same as used for the native crystal) additionally containing 0.125 *M* sodium iodide and then frozen in a rayon loop in a cold nitrogen-gas stream (100 K). The synchrotron radiation wavelength was set to 1.54 Å to optimize the X-ray flux and to increase the anomalous signal for the iodine derivative. The initial images of each data set were subjected to the autoindexing routine of *DENZO* (Otwinowski & Minor, 1997), which suggested a primitive orthorhombic cell to be the best solution. Following an optimum strategy of data collection suggested by the program *marHKL*, totals of 103 and 248° of native and iodine-derivative data were collected, respectively. The images were processed and scaled with the programs *DENZO* and *SCALEPACK* (Otwinowski & Minor, 1997).

The iodine derivative, contrary to most derivatives prepared by traditional soaks, did not suffer a considerable loss of diffraction power, modification of unit-cell parameters or degradation during its preparation and has a great chance of being isomorphous (Table 1).

A search for additional heavy-atom derivatives and phase calculations using the SIRAS method are currently under way.

Table 1
Native and iodine-derivative crystal data and data-collection statistics.

Values for the highest resolution shell are shown in parentheses.

	Native	Iodine derivative
Space group	$P2_12_12_1$	$P2_12_12_1$
Unit-cell parameters (Å)	*a* = 55.44, *b* = 61.62, *c* = 73.47	*a* = 56.05, *b* = 61.78, *c* = 73.63
Resolution (Å)	21.7–2.00 (2.05–2.00)	21.8–1.92 (1.96–1.92)
No. of reflections	61846	182876
No. of unique reflections†	16382	37777
$\langle I/\sigma(I)\rangle$	14.5 (3.8)	13.4 (3.1)
Multiplicity	3.8 (3.4)	4.8 (4.3)
Completeness (%)	92.7 (82.3)	99.9 (99.7)
$R_{\rm merge}$‡ (%)	8.2 (35.0)	11.7 (43.9)
$R_{\rm fac}$§ (%)		21.7
Data collected (°)	103	248
Cryoprotectant solution	Mother liquor, 15% ethylene glycol	Mother liquor, 15% ethylene glycol, 0.125 *M* NaI
Soaking time (s)	30	180

† Multiplicity of derivative and native data sets calculated with Friedel-related reflections treated separately and as equivalent, respectively. ‡ $R_{\rm merge} = \sum_{hkl}|I_{hkl} - \langle I_{hkl}\rangle|/\sum_{hkl} I_{hkl}$. § $R_{\rm fac} = \sum_{hkl}|F_{PH} - F_P|/\sum_{hkl} F_P$.

This work was supported by grants 99/03387-4 and 98/06218-6 from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brazil to IP and RAPN, respectively, and by CNPq. We are grateful to J. Brandão for his help with data collection.

References

Cookson, W. (2000). *Nature (London)*, **402**, B5–B11.
Dauter, Z., Dauter, M. & Rajashankar, K. R. (2000). *Acta Cryst.* D**56**, 232–237.
Dumoutier, L., Louahed, J. & Renauld, J. C. (2000). *J. Immunol.* **164**, 1814–1819.
Dumoutier, L., Van Roost, E., Ameye, G., Michaux, L. & Renauld, J. C. (2000). *Genes Immun.* **1**, 488–494.
Dumoutier, L., Van Roost, E., Colau, D. & Renauld, J. C. (2000). *Proc. Natl Acad. Sci. USA*, **97**, 10144–10149.
Kotenko, S. V., Izotova, L. S., Mirochnitchenko, O. V., Esterova, E., Dickensheets, H., Donnelly, R. P. & Pestka, S. (2001). *J. Biol. Chem.* **276**, 2725–2732.
Kotenko, S. V., Krause, C. D., Izotova, L. S., Pollack, B. P., Wu, W. & Pestka, S. (1997). *EMBO J.* **16**, 5894–5903.
Levitt, R. C., McLane, M. P., MacDonald, D., Ferrante, V., Weiss, C., Zhou, T. Y., Holroyd, K. J. & Nicolaides, N. C. (1999). *J. Allergy Clin. Immunol.* **103**, S485–S491.
McLane, M. P., Haczku, A., van de Rijn, M., Weiss, C., Ferrante, V., MacDonald, D., Renauld, J. C., Nicolaides, N. C., Holroyd, K. J. & Levitt, R. C. (1998). *Am. J. Respir. Cell Mol. Biol.* **19**, 713–720.
Nagem, R. A. P., Dauter, Z. & Polikarpov, I. (2001). *Acta Cryst.* D**57**, 996–1002.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Polikarpov, I., Oliva, G., Castellano, E. E., Garratt, R. C., Arruda, P., Leite, A. & Craievich, A. (1997). *Nucl. Instrum. Methods A*, **405**, 159–164.
Polikarpov, I., Perles, L. A., de Oliveira, R. T., Oliva, G., Castellano, E. E., Garratt, R. C. & Craievich, A. (1997). *J. Synchrotron Rad.* **5**, 72–76.
Temann, U. A., Geba, G. P., Rankin, J. A. & Flavell, R. A. (1998). *J. Exp. Med.* **188**, 1307–1320.
Xie, M. H., Aggarwal, S., Ho, W. H., Foster, J., Zhang, Z., Stinson, J., Wood, W. I., Goddard, A. D. & Gurney, A. L. (2000). *J. Biol. Chem.* **275**, 31335–31339.

Structure, Vol. 10, 1051–1062, August, 2002, ©2002 Elsevier Science Ltd. All rights reserved. PII S0969-2126(02)00797-9

Crystal Structure of Recombinant Human Interleukin-22

Ronaldo Alves Pinto Nagem,[1,2] Didier Colau,[3] Laure Dumoutier,[3] Jean-Christophe Renauld,[3] Craig Ogata,[4] and Igor Polikarpov[1,5,6]
[1]Laboratório Nacional de Luz Síncrotron
Caixa Postal 6192
CEP 13084-971
Campinas, São Paulo
Brazil
[2]Universidade Estadual de Campinas
Deptartamento de Física
Caixa Postal 6165
CEP 13084-971
Campinas, São Paulo
Brazil
[3]Ludwig Institute for Cancer Research
Brussels Branch and
The Experimental Medicine Unit
Christian de Duve Institute of Cellular Pathology
Université de Louvain
Brussels
Belgium
[4]Brookhaven National Laboratory
Howard Hughes Medical Institute
NSLS
Upton, New York 11973
[5]Instituto de Física de São Carlos
Universidade de São Paulo
Avenida Trabalhador Sãocarlense 400
CEP 13560-970
São Carlos, São Paulo
Brazil

Summary

Interleukin-22 (IL-10-related T cell-derived inducible factor/IL-TIF/IL-22) is a novel cytokine belonging to the IL-10 family. Recombinant human IL-22 (hIL-22) was found to activate the signal transducers and activators of transcription factors 1 and 3 as well as acute phase reactants in several hepatoma cell lines, suggesting its involvement in the inflammatory response. The crystallographic structure of recombinant hIL-22 has been solved at 2.0 Å resolution using the SIRAS method. Contrary to IL-10, the hIL-22 dimer does not present an interpenetration of the secondary-structure elements belonging to the two distinct polypeptide chains but results from interface interactions between monomers. Structural differences between these two cytokines, revealed by the crystallographic studies, clearly indicate that, while a homodimer of IL-10 is required for signaling, hIL-22 most probably interacts with its receptor as a monomer.

Introduction

Interleukin-22 (IL-22) is a novel protein that was recently identified in murine cells [1] and was subsequently found in human cells [2]. In vitro experiments showed that IL-22 expression is induced by interleukin-9 (IL-9) in T cells and mast cells [1]. IL-9 induction is rapid (within 1 hr) and does not require protein synthesis. IL-22 was initially classified as a cytokine due to several characteristic features: the presence of an N-terminal hydrophobic signal peptide, ~20 kDa in size, and a low, but detectable, amino acid identity with interleukin-10 (IL-10). This fact was later confirmed by the finding that the IL-22 receptor complex consists of two members of the class II cytokine receptor family, namely, CRF2-9 and CRF2-4 or IL-10Rβ [3, 4]. Other members of the class II family are the two interferon-γ (IFN-γ) receptor chains (Rα and Rβ), the two chains of the IL-20 receptor (Rα and Rβ), the two chains of the IFN-α/β receptor, the interleukin-10 receptor (IL-10R1), and the tissue factor [5]. On the other hand, the growth hormone (GH) receptor and the prolactin receptor, among others, are members of the class I cytokine receptor family.

IL-22 is a member of interleukin-10 family of cytokines, which also includes IL-19 [6], IL-20 [7], and IL-24 [8] in addition to IL-10 [9] and a number of its viral homologs [10]. The members of the IL-10 family have sequence identities of 20%–27% with human cellular paralogs, whereas amino acid sequence identity between homologs from different organisms can be as high as 70%–80% [10].

Human and mouse IL-22 (hIL-22 and mIL-22) have 179 residues, including four cysteines, and share 79% amino acid sequence identity. On the other hand, the hIL-22 (mIL-22) sequence has only 25% (22%) identity with human IL-10 (hIL-10). Most of the conserved residues between IL-22 and IL-10 are located in the C-terminal half of the protein, which has been found to be critical for IL-10 activity, leading to the hypothesis that IL-22 and IL-10 may share common or related biological activities. The only three-dimensional structures of members of the IL-10 family of cytokines known at present are those of IL-10 [11, 12, 13] and its viral homolog from Epstein-Barr virus (EBV) [14].

The fact that mIL-22 is induced in various organs by lipopolysaccharide (LPS) suggests that the role of this new cytokine is not restricted to the immune system [2]. It was also found that hIL-22 activates the signal transducers and activators of transcription factors 1 and 3 in several hepatoma cell lines. IL-22 stimulation of HepG2 human hepatoma cells upregulated the production of acute phase reactants, such as serum amyloid A, α1-antichymotrypsin, and haptoglobin [2]. A similar acute phase reactant induction was also observed in mouse liver upon IL-22 injection, suggesting involvement of IL-22 in the inflammatory response. In addition, IL-22 might play a role in allergy and asthma due to the involvement of IL-9 in these two pathologies [15, 16]. Furthermore, the IL-22 gene is located on human chromosome 12q [17], where several loci potentially linked to asthma have been identified by genetic studies [18].

[6]Correspondence: ipolikarpov@if.sc.usp.br

Key words: crystal structure; IL-22; IL-TIF; IL-10; IFN-γ; interleukin

Cytokines exert their actions by binding to specific cell surface receptors, leading to the activation of cytokine-specific signal transduction pathways. Recent results [3, 4] show that the functional IL-22 receptor complex consists of two receptor chains, the CRF2-9 (IL-22R) chain and the CRF2-4 (IL-10R2 or IL-10Rβ) chain. The latter has been demonstrated to be a functional component of the IL-10 signaling complex [19]. This is the first example within the class II cytokine receptor family of a receptor being utilized as a component of multiple distinct cytokine signaling complexes. A similar sharing is also observed in IL-2, IL-4, IL-7, IL-9, and IL-15 receptor complexes (γ common chain, γ_c).

Here we describe the structure of recombinant hIL-22 refined to 2.0 Å resolution and its comparison with the crystallographic models of hIL-10 [11, 12, 13] and hIFN-γ [20]. The possible receptor binding sites were inferred on the basis of structural comparison with the hIFN-γ/hIFN-γRα complex [20], the recently determined complex of IL-10 with its receptor IL-10R1 [21], and amino acid sequence alignments. Structural comparisons with IL-10 and IFN-γ receptor complexes clearly indicate that, while a homodimer of IL-10 and/or IFN-γ is required for signaling, hIL-22 most probably interacts with its receptor as a monomer.

Results and Discussion

Structure Determination

Purified recombinant hIL-22 was crystallized at the Protein Crystallography Laboratory of the Brazilian National Synchrotron Light Laboratory (LNLS) using the hanging drop method. Several attempts to improve crystal quality were performed, including pH and precipitant concentration refinement, detergent addition, and macroseeding. Well-diffracting crystals were obtained using sodium tartrate and TRITON X-100 detergent in HEPES buffer at pH 7.5. The crystal space group was determined to be $P2_12_12_1$ (see Experimental Procedures for details).

The structure of hIL-22 was solved by X-ray diffraction using the SIRAS method. The data sets of an iodine derivative (I-hIL-22) and a native crystal (Nat-hIL-22) were collected at the Protein Crystallography beamline [22, 23] at the LNLS (Campinas, São Paulo, Brazil). The I-hIL-22 derivative was prepared according to the quick-cryosoaking procedure [24, 25] for fast derivatization of protein crystals. One additional Hg derivative data set was collected at the X4A beamline at NSLS (Upton, New York). These data did not provide a strong isomorphous or anomalous signal and were therefore used only at the later stages of refinement and during the construction of disordered loops, as a quasi-native data set. Details of the preparation of native and derivative crystals for data collection as well as data statistics are given in Table 1.

SIRAS-derived phases using native and iodine derivative data have a mean figure of merit of 0.45 in the resolution range from 21.7 to 2.4 Å. Due to the high resolution and completeness of the I-hIL-22 data set and the quality of the solvent-flattened electron density map, automatic construction of an hIL-22 hybrid model could be performed by the ARP/wARP program [26]. The inferred amino acid sequence derived from the cDNA [2] was used in the final model side chain assignment.

Quality of the Model

The initial structure of hIL-22 was improved by a number of cycles of refinement and rebuilding using the CNS package [27]. The final model is characterized by an R factor of 0.188 and an R_{free} of 0.22 for the Nat-hIL-22 data in the resolution range from 21.7 to 2.0 Å.

The isolated cDNA of hIL-22 encodes a protein of 179 amino acids, the first 33 of which are predicted to function as a signal sequence [3]. The N-terminal amino acid analysis of hIL-22 confirms that the mature protein begins at amino acid residue 34. The refined model of hIL-22, a dimer in the asymmetric unit (Figure 1), includes monomer A, with 142 amino acid residues (Ser38–Ile179), monomer B, with 141 amino acid residues (His39–Ile179), and 189 water molecules. A total of 93.8% and 6.2% of the amino acid residues adopt a conformation corresponding to the most favored and additionally allowed regions of the Ramachandran plot, respectively (see Table 2 for further information about refinement and geometry statistics). No residues have been encountered in the disallowed regions of the Ramachandran plot. Pro113 is in the *cis* conformation. Alternative conformations of the side chains were found for residues Met172 (monomer A) and Asp43, Ser45, Arg55, Ile75, His81, Arg124, Ile161, and Leu174 (monomer B). The electron density for part of loop DE (residues 127–132) is weak, resulting in high temperature factors for this part of the protein (Figure 2).

Each monomer of the hIL-22 model, as shown in Figure 1B, is characterized by six α helices (A–F) that fold in a compact bundle. Helix A (amino acid residues Lys44–Ser64) is linked to a short helix, B (Glu77–Phe80), by a large loop, AB (Leu65–Gly76). Helix A has a kink at Gln48–Gln49, presumably due to a hydrogen bond between Nε-Gln49 and O-Ser45 (2.79 Å and 2.55 Å in monomers A and B, respectively). This divides helix A into two unequal parts: A_1 and A_2. Loop BC (His81–Glu87) connects helix B to helix C (Arg88–Glu102). Helix C is joined to helix F by a disulfide bond between Cys89 and Cys178. Another loop (CD; Val103–Tyr114) links helix C to helix D (Met115–Leu129). According to PROCHECK [28], a small difference in secondary structure between monomers is observed at the loop CD region. A small α helix is observed between amino acid residues Phe105 and Gln107 of monomer B. Helix D is connected to helix E by a disordered loop (DE; Ser130–Asp138). This loop is stabilized, at least in the vicinity of Cys132, by another disulfide bond between Cys132 and Cys40, the latter in the N-terminal coil. Finally, a simple junction EF (Gly156) joins the last two helices, E (Leu139–Leu155) and F (Glu157–Cys178). Probably, as a consequence of a disulfide bond between Cys89 and Cys178, the latter belonging to the C-terminal of helix F, a kink at Glu166 divides helix F into two parts: F_1 and F_2.

Dimer Formation

A significant part (61%) of the volume of the asymmetric unit (6.27×10^4 Å^3) is occupied by a dimer of hIL-22. A small fraction of this volume (8%) is filled with ordered

Table 1. Details of the Preparations and Data Collection Statistics of hIL-22 Crystals

	Nat-hIL-22	I-hIL-22	Hg-hIL-22 (Quasi-Native)
Wavelength (Å)	1.54	1.54	1.54
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
Unit cell parameters (Å)	a = 55.43, b = 61.61, and c = 73.47	a = 56.05, b = 61.78, and c = 73.63	a = 56.04, b = 61.71, and c = 74.61
Resolution (Å)	21.7–2.00 (2.05–2.00)	21.8–1.92 (1.96–1.92)	22.4–1.90 (1.97–1.90)
Number of reflections	61,846	182,876	55,855
Number of unique reflections[a]	16,382	37,777	29,854
$<I/\sigma(I)>$	14.5 (3.8)	13.4 (3.1)	8.2 (2.1)
Multiplicity	3.8 (3.4)	4.8 (4.3)	1.9 (1.7)
Completeness	92.7 (82.3)	99.9 (99.7)	75.9 (77.9)
R_{merge}[b]	8.2 (35.0)	11.7 (43.9)	10.0 (49.9)
Data collected (degrees)	103	248	70
Cryoprotectant solution	mother liquor 15% ethylene glycol	mother liquor 15% ethylene glycol 0.125 M NaI	mother liquor 15% ethylene glycol 5 mM $HgCl_2$
Soaking time	30 s	180 s	10 hr

Statistical values for the highest resolution shells are shown in parentheses.
[a] Multiplicity of derivative (native) data sets calculated with Friedel-related reflections treated separately (as equivalent).
[b] $R_{merge} = \Sigma_{hkl}|I_{hkl} - <I_{hkl}>|/\Sigma_{hkl}\ I_{hkl}$.

water molecules. The monomers are essentially equal; however, a number of significant differences in the main chain conformation in the vicinity of amino acid residues Gln48, Asn69, Gly136, and Lys154 are observed (Figure 2). These differences could mostly be explained by crystallographic and noncrystallographic contacts. The reason for a significant positional difference between monomers around Gln48 is the fact that this region in monomer A is involved in interface interactions, while the same region in monomer B is exposed to the solvent. Besides this, the presence of two intramolecular interactions (Oδ1-Asp43/Oγ-Ser45 at a distance of 2.64 Å in monomer A and O-Asn46/Nε2-Gln49 at 2.55 Å in monomer B) contributes to a relative change in main chain atomic positions between residues Leu42 and Pro50. The second conformational difference, around residue Asn69, is a consequence of a crystallographic contact between side chain atoms of Asn69 and Thr70 of monomers A and B, respectively. Gly136 is localized in disordered loop DE. This fact explains the rmsd (root mean square deviation) of around 2.0 Å in the vicinity of this residue. Finally, the last major difference between monomers is found close to Lys154. In this region three distinct interactions of Lys153 and Lys154 from monomer B (Oε1-Glu102/Nζ-Lys153 at a distance of 2.68 Å, Oδ1-Asn46/Nζ-Lys153 at 2.78 Å, and Oε1-Glu160/Nζ-Lys154 at 2.80 Å), which are absent in monomer A, are responsible for a high rmsd of main chain atoms.

Unlike hIL-10, the hIL-22 dimer does not result from the intertwining of the main chain of each monomer

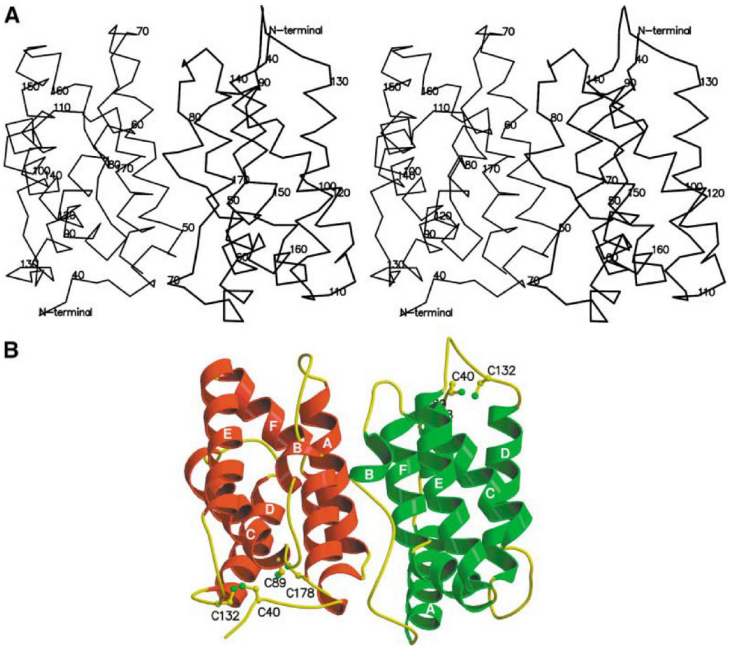


Figure 1. Structure of the IL-22 Dimer

(A) Stereo view of the C_α trace of the dimeric structure of hIL-22.

(B) Schematic representation of the secondary structure of the hIL-22 dimer, according to PROCHECK [28], showing the location of the two disulfide bonds (Cys40-Cys132 and Cys89-Cys178) represented by ball and stick diagrams. The figures were prepared using Molscript [29], Bobscript [30], and Raster3D [31].

Table 2. Refinement Statistics and Quality of the hIL-22 Model

Refinement Statistics	
Resolution range (Å)	21.7–2.0
Total number of reflections	15,684
Working set number of reflections	14,892
R factor (%)	18.8
Test set number of reflections	792
R_{free} (%)	22.0
Total number of protein atoms	2330
Total number of water molecules	189
Stereochemical Parameters	
Rmsd bond distances (Å)	0.006
Rmsd bond angles (°)	1.1
Average B factors	
Residue atoms (Å²) (A, B)	24.3 (22.3, 26.2)
Main chain atoms (Å²) (A, B)	22.1 (20.2, 24.1)
Side chain atoms (Å²) (A, B)	26.3 (24.4, 28.1)
Water molecules (Å²)	37.3
Average B factor rmsd	
Residue atoms (Å²) (A, B)	2.5 (2.6, 2.5)
Main chain atoms (Å²) (A, B)	1.0 (1.0, 1.0)
Side chain atoms (Å²) (A, B)	1.9 (2.0, 1.8)
Water molecules (Å²)	11.4
Noncrystallographic Symmetry[a]	
Rmsd coordinates	
Cα atoms (Å)	0.911
Main chain atoms (Å)	0.884
All bonded atoms (Å)	1.670
Rmsd B factors	
Cα atoms (Å²)	10.04
Main chain atoms (Å²)	10.05
All bonded atoms (Å)	10.77

[a] Between subunits A and B.

(Figure 1). An interface area of approximately 2250 Å², which corresponds to 30% of the total surface area of a monomer, is involved in the dimer formation. The buried surface for the chosen dimer conformation is at least twice that of any other possible dimer generated as a result of crystal packing (~960 Å² or less). Besides this, the dimer interface, which is formed mostly by residues Arg41–Phe80 and Asp168–Ile179 in monomer A and

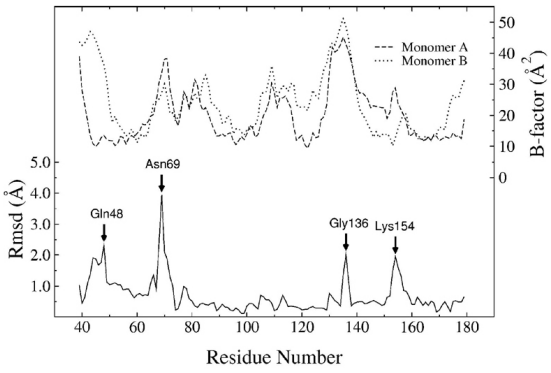


Figure 2. Least-Square Fit of Monomer A to Monomer B and Temperature Factors Plots

The root mean square deviation (rmsd) and B factors are shown as a function of residue number. Only main chain atoms were used in the calculation.

Table 3. Intermolecular Contacts

Monomer A		Monomer B		
Residue	Atom	Residue	Atom	Distance (Å)
Arg175	Nη2	Glu166	Oε1	2.57
Phe57	O	Asn176	Nδ2	2.64
Arg73	Nη2	Val83	O	2.71
Lys44	Nζ	Ser64	Oγ	2.85
Arg175	Nη1	Asp168	Oδ2	2.86
Asn176	Nδ2	Ile75	O	2.91
Gln48	O	Lys61	Cε	2.96
Lys44	Nζ	Glu166	Oε1	2.98
Lys61	Nζ	Ile179	OT1	3.12
Gln49	Oε1	Lys61	Nζ	3.15

The distance cutoff of 3.2 Å was used.

Thr53–Arg88 and Glu166–Ile179 in monomer B, has a significant number of hydrophobic residues. Intermolecular interface contacts closer than 3.2 Å are listed in Table 3. The electrostatic and hydrophobic distribution of the hIL-22 surface, together with the position of the principal amino acid residues involved in the formation of the dimer, are given in Figure 3.

Potential Glycosylation Sites

According to the predicted primary structure, human IL-22 has three potential glycosylation sites (Asn-Xaa-Thr/Ser) localized in helix A (Asn54-Arg55-Thr56), loop AB (Asn68-Asn69-Thr70), and helix C (Asn97-Phe98-Thr99). Since the recombinant hIL-22 used in crystallization is not glycosylated, we attempted an analysis of the possible interactions between oligosaccharides and hIL-22 by calculating the accessible area of each residue in all three putative glycosylation sites. The results demonstrate that site 2, localized in loop AB, has a larger accessible area in both the IL-22 dimer and monomer. A solvent-accessible area of approximately 37 Å² was found for the Nδ2 atom of Asn68 and for the Oγ1 atom of Thr70, indicating that there is no steric hindrance to their participation in N-glycosyl and O-glycosyl links, respectively. On the other hand, sites 1 and 3 seem to be able to participate only in N-glycosyl links. The accessible areas of Oγ1-Thr56 and Oγ1-Thr99 are 0 and 6 Å², respectively, both in the monomer and dimer of IL-22, whereas the Nδ2-Asn54 and Nδ2-Asn97 atoms possess, respectively, surface-accessible areas of 24 and 18 Å². This structural analysis is in agreement with biochemical studies suggesting that these three sites are of the N-glycosyl type [4]. The present structure shows that putative glycosylation sites 1 and 2 reside close to the dimer interface, and glycosyl linkages at these positions might hamper dimer formation. It is difficult to conclude whether glycosylation of the protein will interfere with its interactions with receptor chains (see Potential Receptor Binding Sites below) on the basis of the current analysis alone. Further biochemical studies are required to address this question.

Comparison with IL-10 and IFN-γ

The crystallographic structure of hIL-22, as seen in Figure 1B, is a compact dimer with a buried surface area of approximately 2250 Å². Several intermolecular inter-

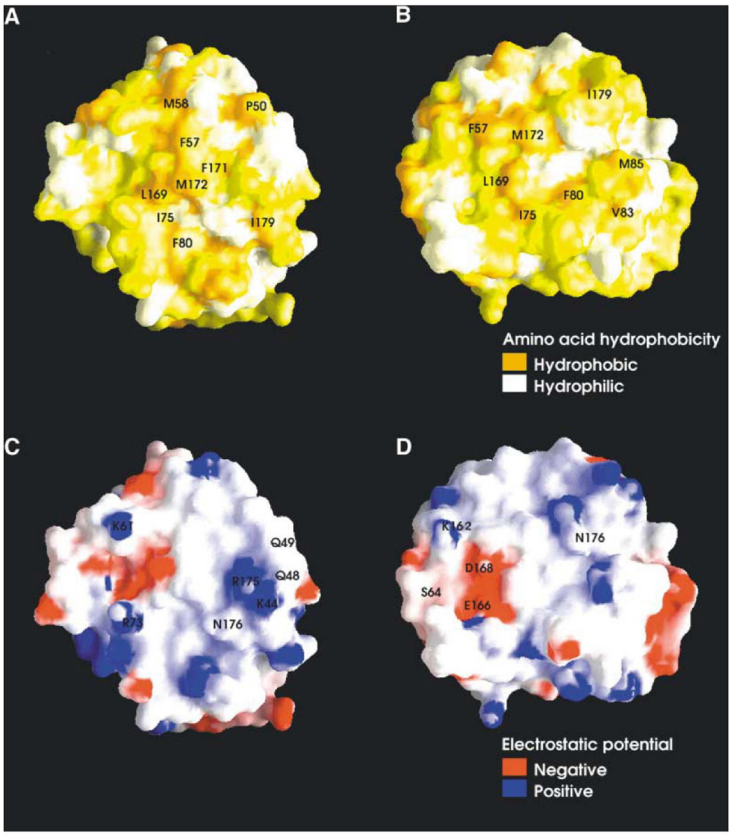


Figure 3. The Contact Surface of the hIL-22 Dimer

The figure is colored according to residue hydrophobicity (A and B) and electrostatic potential (C and D). Interface views of monomer A (A and C) and monomer B (B and D) are shown. In parts (A) and (B), the darker the yellow, the greater the hydrophobicity. In parts (C) and (D), areas of negative, positive, and neutral electrostatic potential are depicted in red, blue, and white, respectively. The figures were prepared with GRASP [32].

actions along the interface surface keep the monomers together. Each monomer, composed of a single domain, is formed by six α helices (A–F) from the same polypeptide chain. In contrast, the crystallographic structures of hIL-10 [11, 12, 13] and hIFN-γ [20, 33] revealed the presence of a homodimer composed of two α-helical domains formed by the intertwining of α helices donated by the first and the second monomer composing a dimer. The first four helices of one chain (A–D) together with helices E′ and F′ from the second chain form the first domain, while helices A′–D′, E, and F form the second domain.

There are, however, significant structural similarities between IL-22, IL-10, and INF-γ (Figure 4). In all these proteins, helices A–D of each monomer form a rigid framework with a highly hydrophobic depression in its middle. This depression is covered in hIL-22 by helices E and F from the same monomer, while, in hIL-10 and hIFN-γ, this is accomplished by helices E′ and F′ (from the second monomer). The basic reason for these differences is in loop DE. There are two cysteine residues, Cys126 and Cys132, in the hIL-10 DE loop that make two distinct disulfide bonds with residues Cys30 and Cys80, respectively. (Here we adopt the residue numbering according to the hIL-10 cDNA sequence). These two disulphide bridges restrict the flexibility of the polipeptide chain and the length of loop DE in such a manner that helices E and F cannot fold onto their respective monomer to occupy position of their counterparts, E′ and F′. This leads to the intertwined dimer formation [11, 12, 13]. A monomeric form of hIL-10 could only be formed if the Cys80-Cys132 disulfide bond were to be reduced or if a small insertion were made after Cys132 [11]. The latter approach has been applied with success to hIL-10, where the insertion of a small polypeptide linker in the loop connecting the swapped secondary structure elements led to the formation of a monomeric protein [34]. Similarly to IL-10, the hIFN-γ intertwined dimer is formed because loop DE is not long enough to allow the folding of helices E and F into the same domain.

In hIL-22 just one disulfide bond (Cys40-Cys132) exists within loop DE. This renders sufficient flexibility and extension of the loop to bring helices E and F into a close interaction with helices A–D and to complete the folding of the monomer. A second disulfide bond (Cys89-Cys178), in the C-terminal of helix F, adds to the rigidity of the final hIL-22 structure.

The best superposition of hIL-22 onto hIL-10 and hIFN-γ, respectively, was obtained using a single domain of the hIL-10 and hIFN-γ dimers and a monomer of hIL-22 (Figures 4A and 4C). The superposition of hIL-10 and hIFN-γ onto hIL-22 yielded an rmsd of 1.9 Å and 2.3 Å for 432 and 300 pairs of main chain atoms, respectively. Helices A–D of the hIL-22 monomer superimpose with helices A–D of one of the monomers of hIL-10 and hIFN-γ. Helices E and F fit nicely into the spatial position occupied by helices E′ and F′ of the second monomer. The 3D superposition of the structures al-

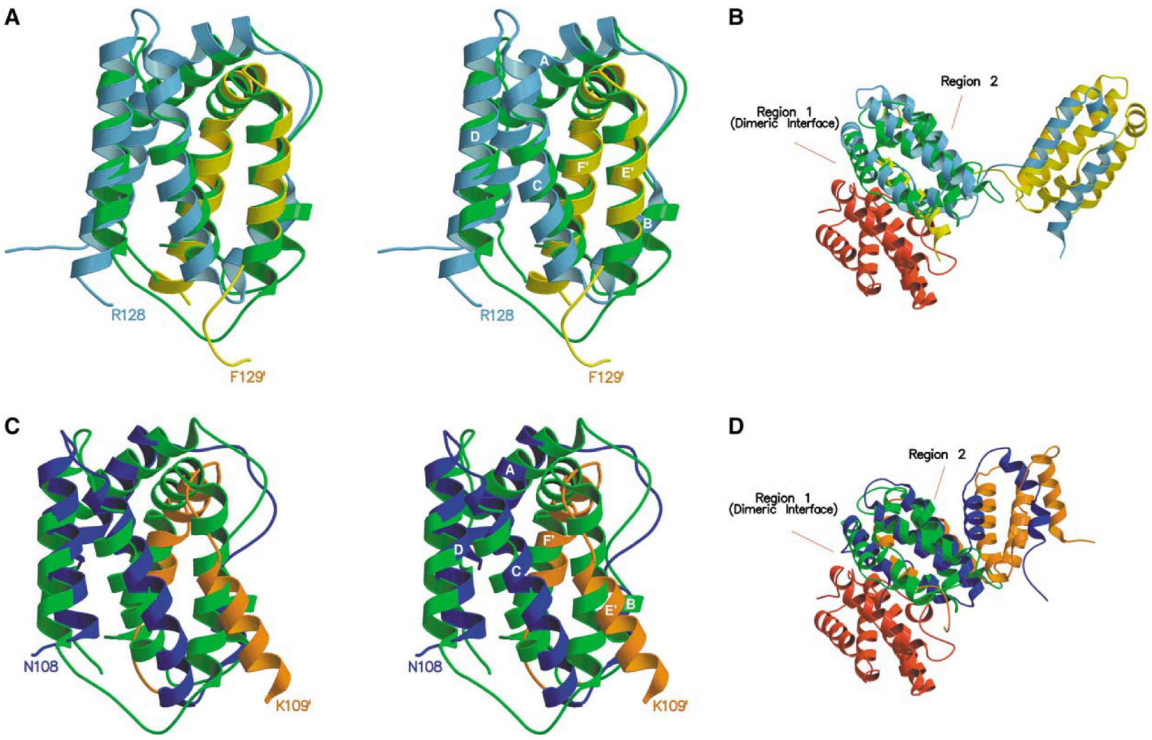


Figure 4. Structural Comparison of hIL-22, hIL-10, and hINF-γ

Ribbon diagram showing the superposition of the hIL-22 monomer (green) onto (A) a single hIL-10 domain (helices A–D, light blue; helices E′–F′, yellow; helices A′–D′, E, and F were omitted) and (C) a single hIFN-γ domain (helices A–D, dark blue; helices E′–F′, orange; helices A′–D′, E, and F were omitted). Superposition of the hIL-22 dimer (individual monomers colored in red and green) onto (B) the hIL-10 dimer (light blue and yellow) and (D) the hIFN-γ dimer (dark blue and orange).

lowed us to perform the structure-based sequence alignment for hIL-22, hIL-10, and hIFN-γ shown in Figure 5. Inspection of the hIL-22 and hIL-10 structure superposition revealed strong similarities in the conformation of the main chain trace of helices E (E′) and F (F′) and, to a lesser extent, several parts of loop AB, helix C, and helix D. As can be seen in the sequence alignment, all these regions have high sequence similarity. Some significant differences in the regions of the N-terminal coil, helix A, helix B, loop BC, loop CD, and, clearly, loop DE were also observed.

Reasonable superposition of hIL-10 or hIFN-γ dimers onto the hIL-22 dimer has proven to be impossible. The dimer formation in each case is so different that only one monomer of hIL-22 could be superimposed with one domain from hIL-10 or hIFN-γ, while the second domain of these structures occupies a completely different spatial position (Figures 4B and 4D). In hIL-10 and hIFN-γ the intertwining of α helices is essential for the formation and integrity of the molecules, which assume the form of V-shaped dimers, while dimer formation is not required for folding in hIL-22. It must be stressed that the buried surface on the hIL-22 interface coincides with the part of the external surface of the hIL-10 and hIFN-γ V-shaped dimer surfaces (Figures 4B and 4D).

Potential Receptor Binding Sites

Two receptor chains have been identified for IL-22, namely, CRF2-9 and CRF2-4. The second receptor chain is common to IL-22 and IL-10 and is necessary for signaling, whereas the first one is specific for IL-22 [3, 4] and shows some primary sequence homology with another receptor chain of IL-10 (the IL-10R1). The binding affinity of IL-22 and IL-10 to CRF2-4 seems different. CRF2-4 alone is sufficient to bind IL-22, while the presence of a second receptor chain is required for efficient IL-10 binding. Moreover, CRF2-9 and CRF2-4 present significant sequence homology to the INF-γ receptor, INF-γRα. The three-dimensional structure of hINF-γRα has been solved as a complex with its ligand [20]. The structure of hIL-22 was superimposed onto the structure of hINF-γ/hINF-γRα complex in order to identify the residues putatively involved in the hIL-22/receptor interactions. A similar structural comparison using the hGH/hGHBP complex had been used previously in the putative receptor binding sites analysis of IL-10 [11].

The superposition of hIL-22 onto the hINF-γ/hINF-γRα complex indicates that one possible receptor binding site is localized in the region formed by helix A, loop AB, and helix F of hIL-22 (region 1, R1; see Figures 4D, 6A, and 6B). Among the 17 residues involved in hINF-γ/hINF-γRα interactions (closer than 3.4 Å), only 2 residues do not have their hIL-22 structural counterparts localized in R1. Nine of the 17 residues localized in R1 are not sufficiently close to their hINF-γ counterparts, which may explain the inability of hIL-22 to bind to hINF-γRα. The major differences between hINF-γ and hIL-22 within R1 are observed in loop AB. Distances of more than

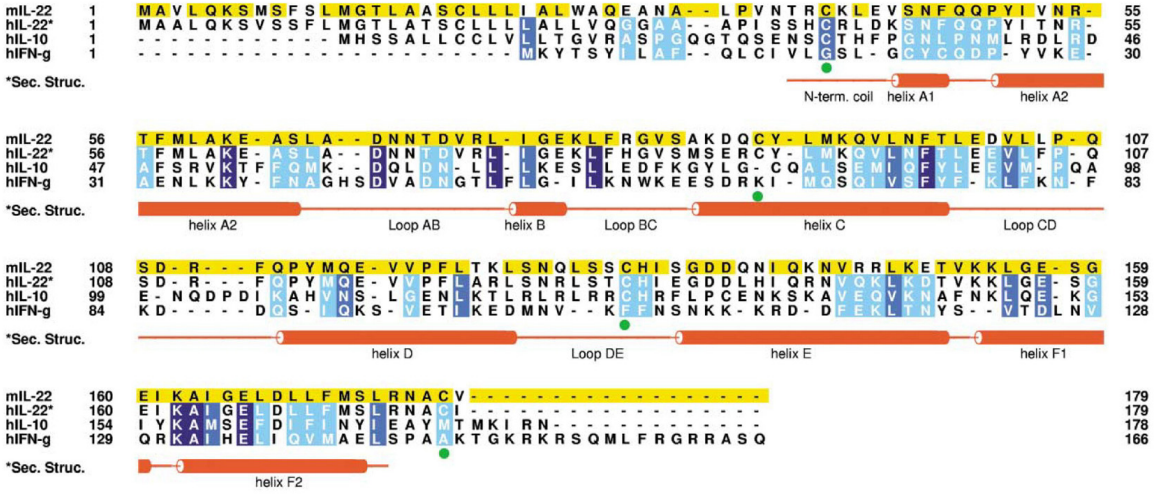


Figure 5. Primary Structure Alignment of Murine IL-22, Human IL-22, Human IL-10, and Human IFN-γ

Whenever possible the three-dimensional information was used to improve the alignment. Disulfide bonds in hIL-22 are marked with green circles. The amino acid similarity between hIL-22, hIL-10, and hIFN-γ, as calculated by the program ALSCRIPT [35], is shown in three different shades of blue. The darker color corresponds to higher sequence similarity. Residues conserved in mIL-22 and hIL-22, yellow; human IL-22 secondary structure elements, red. The figure was drawn using the program ALSCRIPT.

7 Å are found between their main chains. However, six relatively conserved residues (Lys61, Thr70, Asp71, Lys162, Glu166, and Leu169) occupy almost the same spatial position as six hINF-γ residues (Lys35, Asp47, Asn48, Lys131, Glu135, and Gln138; cDNA numbering), as can be seen in Figure 6A.

A comparison with the hIL-10 putative receptor binding site [11] shows that the same region 1 (helix A, loop AB, and helix F′ in the case of hIL-10) should be involved in receptor interactions. Amino acid residues Gln60, Asp62, Asn63, Lys156, Glu160, Asp162, Asn166, and Glu169 of the hIL-10 binding site have their hIL-22 counterparts in residues Asn68, Thr70, Asp71, Lys162, Glu166, Asp168, Met172, and Arg175. Among these eight residues, four of them (Thr70, Asp71, Lys162, and Glu166) were also found in the hIL-22:INF-γ/INF-γRα structure comparison.

After the current structure was determined and its structural comparison with the hINF-γ/hINF-γRα and hGH/hGHBP complexes had been performed, the crystallographic structure of the IL-10/IL-10R1 became available [21]. The structure of this complex confirms that the prime binding site of the IL-10 with its high-affinity receptor (site 1a) includes residues Gln56, Gln60, Asp62, Asn63, Lys156, Ser159, Glu160, and Asp162, corresponding to amino acid residues Ser64, Asn68, Thr70, Asp71, Lys162, Gly165, Glu166, and Asp168 in hIL-22. Of these, three amino acid residues, Ser64, Glu166, and Asp168, are involved in intermolecular contacts at the homodimer interface of the current IL-22 structure (Table 3). Besides this, Asp168 and Glu166 of monomer B form salt bridges with Arg175 of the neighboring monomer, A. Arg175, therefore, occupies a position equivalent to that of Arg96 of the IL-10R1 receptor in the IL-10/IL-10R1 complex. Arg96 is one of the most important residues involved in the IL-10/Il-10R1 complex formation, as it forms an extensive hydrogen bond network with Asp162, Gln56, and the main chain carbonyl oxygen of Ser159 from IL-10 [21]. Remarkably, sequence alignment shows that the CRF2-9 receptor chain contains an arginine residue homologous to Arg96 of IL-10R1. One may infer, therefore, that Arg 175 of IL-22 monomer A binds to residues Glu166 and Asp168 of IL-22 monomer B in a way that resembles CRF2-9 binding to this cytokine (Figure 6C).

Amino acid residues Pro38, Arg42, Arg45, and Glu169 form the 1b subsite in the IL-10/IL-10R1 structure. These residues are equivalent to Gln48, Ile52, Arg55, and Arg175 of the IL-22 molecule. Gln48 and Arg175 of IL-22 monomer A again form part of the dimer interface interacting with the residues Lys61, Glu166, and Glu168 of monomer B (Table 3). This means that a large part of the IL-22 homodimer interface is formed by interactions of the amino acid residues potentially involved in receptor binding.

The structural comparison of hIL-22 with the hINF-γ/hINF-γRα and IL-10/IL-10R1 complexes leaves little doubt that region 1 is a putative receptor binding site. The three-dimensional and amino acid sequence similarities observed between hIL-22, hIL-10, and INF-γ, respectively, in region 1, especially between helices F and F′, indicate that this region might serve as the high-affinity receptor binding site. Further support for this hypothesis comes from the similarities in amino acid sequence of IL-10R1, CRF2-9, and INF-γRα, particularly in the loop regions involved in the interactions with the respective cytokines [21]. However, the differences observed in loop AB, including the presence of a potential glycosylation site in hIL-22 and, also, the lack of sufficient biochemical data, do not allow us to definitively infer which receptor chain binds region 1 of hIL-22. In fact, considering the hypothesis that the glycosylation site in the hIL-22 AB loop may affect receptor recognition and also that IL-10R1 and INF-γRα have low, but

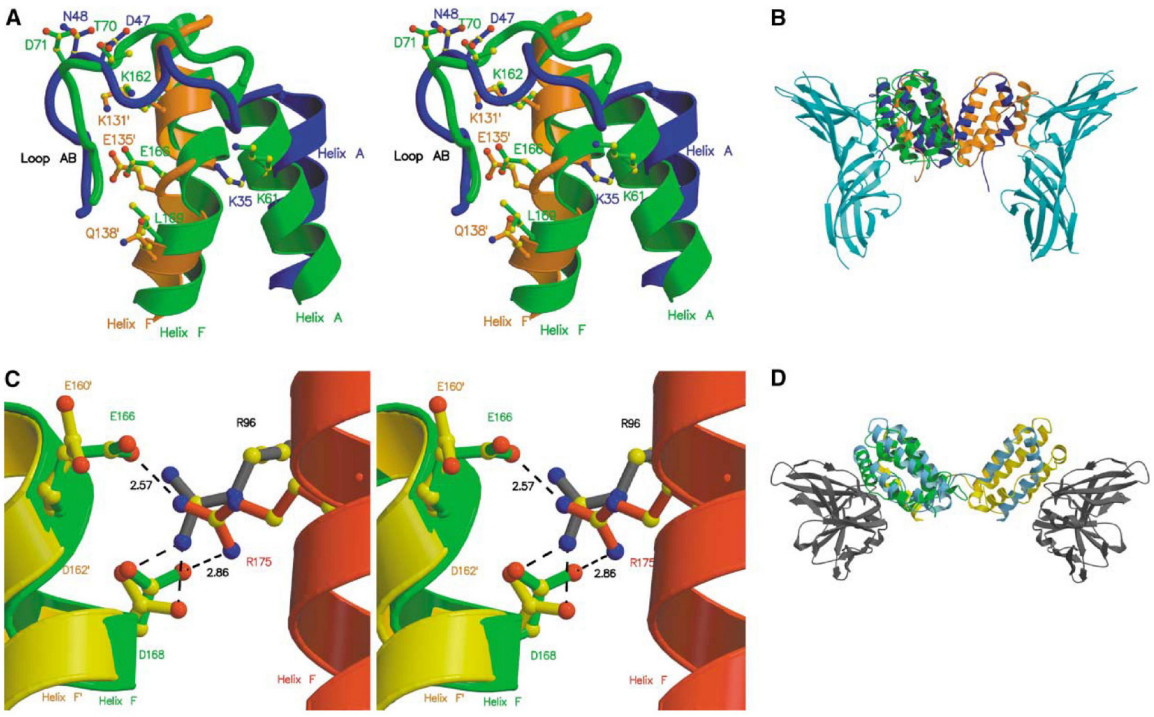


Figure 6. Molecular Details of the Putative Receptor Binding Site (R1) of IL-22

(A) Superposition of the hIFN-γ/hIFN-γRα complex. hIFN-γ, dark blue and orange; hIL-22 monomer, green. Amino acid residues forming the putative receptor binding site of IL-22 and the binding site of hIFN-γ to the hIFN-γRα receptor are represented by ball and stick diagrams, colored according to the molecule coloring scheme and labeled.

(B) Ribbon diagram of hIFN-γ bound to the hIFN-γRα receptor. hIFN-γ, dark blue and orange; hIFN-γRα, cyan. The hIL-22 monomer (green) is superimposed with one of the hIFN-γ domains.

(C) Superposition of hIL-22 onto hIL-10. Only residues Glu160 and Asp162 of the hIL-10 receptor binding site 1a and residue Arg96 of hIL-10R1 are shown superimposed onto the hIL-22 putative binding site residues Glu166 and Asp168 of monomer B and amino acid residue Arg175 of monomer A. Note that the amino acid residue Arg175 at the hIL-22 homodimer interface in the hIL-22 dimer structure mimics hIL-10R1 Arg96 interactions with the correspondent residues of the hIL-10 1a binding site in the hIL-10/Il-10R1 complex [21]. hIL-10, gold; hIL-22 monomer A, red; hIL-22 monomer B, green; IL-10R1 residue Arg96, gray. Numbers in black show distances in Å.

(D) A ribbon diagram of hIL-10 bound to the hIL-10R1 receptor. hIL-10, light blue and gold; hIL-10R1, gray; hIL-22 monomer, superimposed with one of the hIL-10 domains, green.

detectable, similarity to CRF2-4, one may suggest that R1 could also be the recognition/binding site for CRF2-4. This hypothesis also finds support in the model for the IL-10/IL-10R1/CRF2-4 complex, where both receptors share the same binding site [21].

In the present crystallographic model, region 1 of each monomer is hidden at the dimer interface. Moreover, a number of potential receptor binding residues are directly involved in the dimer formation (see Table 3). Therefore, an hIL-22 receptor chain could only bind a monomer of hIL-22, which would require dissociation of the dimer observed in the present crystallographic structure. In contrast, the hIL-10 and hIFN-γ dimers do not require dissociation in order to interact with their receptor, since their high affinity receptor binding sites are located on the external part of the V-shaped dimer surface.

To further assess the oligomerization state of hIL-22 in solution, gel filtration and dynamic light-scattering (DLS) studies were undertaken. The state of oligomerization of the protein was found to be concentration dependent. The initial protein concentration applied to the gel filtration column was 1 mg/ml, and the eluted recombinant IL-22 sample had an apparent molecular weight of 15 kDa and was fully active.

DLS measurements performed at a protein concentration of 5 mg/ml resulted in an apparent Stokes radius of 2.08 ± 0.06 nm, which corresponds to a standard molecular weight of approximately 19 kDa. This demonstrates that hIL-22 was present in solution predominantly in a monomeric form, in line with the results of the gel filtration studies. An excellent correlation with the Stokes radius of 2.11 nm calculated using the HYDROPRO program [36] on the basis of the crystal structure of the monomer determined in the present work further confirms this conclusion. Therefore, both gel filtration and DLS studies demonstrate that hIL-22 is a monomer at physiologically relevant concentrations (≤5 mg/ml).

The DLS measurements performed on the protein sample concentrated to 10 mg/ml resulted in a Stokes radius of 2.74 ± 0.06 nm and an apparent molecular weight of 35.5 kDa, suggesting that a dimer is the predominant oligomeric form under these conditions. The

apparent Stokes radius of 2.58 nm computed by HYDROPRO from the crystallographic model of the dimer occupying the asymmetric unit is in agreement with the experimental results. This indicates not only that, under conditions of crystallization, hIL-22 predominantly forms dimers in solution, but also that the calculated hydrodynamic parameters of the dimers observed in the crystal are close to the experimentally determined hydrodynamic parameters of the dimers observed in solution.

Assuming that the CRF2-4 and CRF2-9 receptor chains have separate binding sites, there should be at least two distinct sites for receptor binding to the IL-22 molecule. A second possible binding site in hIL-22 could not be easily identified from inspection of the interactions between hINF-γ and hINF-γRα or hIL-10 and hIL-10R1. Nevertheless, the C-terminal parts of helices C and E of each hIL-22 monomer could be another potential binding site (region 2, R2) for the common receptor chain (CRF2-4). A sequence comparison between hIL-22 and several IL-10s shows that several amino acids are conserved within R2 (sequences FTLEEVL and KLGE in hIL-22 helices C and E, respectively). These regions (R2) are localized at the surface of the IL-22 opposite from R1. Localization of each putative binding region (R1 and R2) on the opposite sides of the hIL-22 molecule would allow for interaction with two receptor chains simultaneously. The hIL-10 R2 counterparts are localized at the inner part of the V-shaped dimer surface. The angle between each hIL-10 domain in the V-shaped dimer could be large enough to allow for the interaction of two CRF2-4 receptor chains with the two putative binding sites at R2 (Figure 4B).

Biological Implications

Cytokines are proteins that exert their action by binding to specific cell surface receptors, leading to the activation of cytokine-specific signal transduction pathways. These proteins play several important biological roles, including multiple and generally immunosuppressive activities, immunomodulatory and antiviral effects, and stimulation of T cell growth, among others [37]. IL-22 is a new cytokine expressed in T cells upon activation by IL-9 and in mast cell, thymus, and brain upon activation by concanavalin A (ConA), indicating that this factor might exhibit pleiotropic activities, within and outside of the immune system [1]. Two different receptor chains, CRF2-9 and CRF2-4, have been identified as components of the IL-22 signaling complex [3, 4]. A role for this protein in inflammatory processes has been suggested after observing that IL-22 induces STAT activation and upregulates acute phase reactant production by liver cells [2].

Here we report the first crystallographic structure of this new cytokine, human IL-22. Possible receptor binding sites were identified via comparison with the related complex structures of hIFN-γ/hIFN-γRα and hIL-10/hIL-10R1.

hIL-22 is a compact molecule of a single domain composed of six α helices. A dimer of human IL-22 was found in the asymmetric unit of the crystal. Unlike IL-10 and IFN-γ, which are V-shaped homodimers of two interpenetrating polypeptide chains, the hIL-22 dimer is comprised of two independent monomers held together by intermolecular interactions. No intertwining of α helices was observed in the crystallographic structure of hIL-22.

The superposition of hIL-22 onto the hIFN-γ/hIFN-γRα and hIL-10/hIL-10R1 complexes allowed us to identify a possible hIL-22 receptor binding site (region 1, R1). This site is comprised of helix A, loop AB, and helix F, similarly to the first binding site reported for hIL-10 [11, 21]. The overall similarity of hIL-22 to hIL-10, especially in the region of this putative receptor binding site and, also, in their respective high-affinity receptors (CRF2-9 and IL-10R1), indicates that this site could represent the binding site for CRF2-9. However, the lack of biological studies does not allow us to unambiguously conclude whether this is a binding site for the CRF2-9 or CRF2-4 receptor chains or the common binding site for both of these receptors [21]. Remarkably, the receptor binding site in R1 is occluded in the hIL-22 structure as a consequence of dimer formation. This leads us to propose that the active species of hIL-22 is a monomer. This is in a perfect agreement with the results of gel filtration and DLS studies demonstrating that hIL-22 is a monomer at physiologically relevant concentrations. Several other cytokines have already been shown to dimerize at high concentrations while being biologically active as monomers (see, for example, [38] and [39] and the references therein). An observation that an engineered human IL-10 monomer is biologically active in cellular proliferation assays further corroborates this hypothesis [34].

Another putative binding region (region 2, R2) for CRF2-4 (or IL-10R2) was identified on hIL-22 by surface-mapped sequence comparison with hIL-10. R2 is localized on the opposite side from the first binding site (region 1, R1), leaving enough room for binding for a second receptor. Primary structure comparison of several IL-10s and IL-22s from different organisms indicates that the residues forming the second putative binding site are highly conserved.

The structure of hIL-22 presented here brings us closer to an understanding of its interactions with its receptor chains and the mechanism of signal transduction exerted by this molecule. Moreover, this structural information may help in the solution of the structure of the IL-22:receptor complex and also offers the possibility of structure-based site-directed-mutagenesis mapping of the receptor binding sites.

Experimental Procedures

Protein Expression and Purification

Recombinant human IL-22 was produced in *E. coli* as follows. The IL-22 sequence (corresponding to amino acids Q29–I179) was amplified by PCR from a cDNA clone using primers hTIFa (5′-GGCCCTC TTGGTACATATGCAGGGAGGAGCAGCTGCG-3′) and hTIFb (5′-CAG CTTTGCTCTGGGGATCCTTATCAAATGCAGGCATTTCTCAG-3′). The PCR product was digested with NdeI and BamHI and cloned into the pET3A plasmid (Stratagene, La Jolla, CA). *E. coli* strain BL21-codon plus-(DE3)-RIL (Stratagene) was used as the expression host. The cells were grown in LB medium supplemented with 100 μg/ml ampicillin and 34 μg/ml chloramphenicol. Expression of IL-22 was induced with 1 mM IPTG at a cell density (600 nm) of $\sim$1. Cells were collected by centrifugation 4 hr after induction. The cell pellet was

disrupted with a high-pressure cell homogenizer, and the IL-22 inclusion bodies were collected by centrifugation. Inclusion bodies were washed extensively, first with 50 mM Tris-HCl, 100 mM NaCl, 1 mM EDTA, 1 mM DTT, and 0.5% (w/v) DOC (pH 8) and finally with the same buffer without detergent. Inclusion bodies were solubilized overnight at 4°C in 8 M urea, 50 mM MES, 10 mM EDTA, and 0.1 mM DTT (pH 5.5). The solution was centrifuged for 1 hr at 100,000 × g, and the supernatant was stored at −80°C until use. The purity of the IL-22 was estimated at ~80% based on SDS-PAGE and Coomassie blue staining analysis. The concentration of protein was estimated by UV absorbance in urea solution using a calculated $\epsilon_{280} = 3840$. The IL-22 protein was refolded by direct dilution of the solubilized inclusion bodies in the following folding mixture: 100 μg/ml IL-22, 100 mM Tris-HCl, 2 mM EDTA, 0.5 M L-Arginine, 1 mM reduced glutathion, and 0.1 mM oxidized glutathion (pH 8). The solution was incubated for 72 hr at 4°C. The folding mixture was then concentrated by ultrafiltration in an AMICON chamber with a YM3 membrane before purification on a Superdex75 (Amersham Pharmacia Biotech) gel filtration column. The protein was eluted with 25 mM MES and 150 mM NaCl (pH 5.4). The protein bioactivity was assessed following procedures previously described [40]. The recombinant protein was found to be fully active. Human IL-22 peak fractions were concentrated to 5 mg/ml with a YM3 AMICON membrane and desalted using a Hiprep 26/10 column (Amersham-Pharmacia) with elution buffer containing 10 mM MES (pH 5.4). Human IL-22 was concentrated again to 5 mg/ml and lyophilized in 1 mg fractions.

Protein Crystallization

Preliminary screening of the crystallization conditions was performed using a sparse-matrix screen at 291 K (Crystal Screen I and II; Hampton Research). Small crystals were found in the conditions 18, 26, and 29 of the Crystal Screen I kit. Several attempts to enhance crystal quality were performed, including pH and precipitant concentration refinement, detergent addition, and macroseeding. Well-diffracting crystals were obtained in hanging drops equilibrated against a reservoir solution consisting of 0.9 M sodium tartrate, TRITON X-100 detergent, and 0.1 M HEPES (pH 7.5). The crystallization drops contained equal volumes (1 μl) of reservoir and purified hIL-22 (10 mg/ml in 20 mM MES buffer at pH 5.4) solutions. The protein crystallized in the space group $P2_12_12_1$, with unit cell dimensions a = 55.43, b = 61.61, and c = 73.43 Å.

Data Collection

Crystals were soaked in different cryosoaking solutions, mounted in rayon loops, and, finally, flash-cooled to 80 K in a cold nitrogen stream. Data collection was performed at the Protein Crystallography beamline (LNLS, Campinas, Brazil [22, 23]) and at the X4A beamline (NSLS, Upton, New York) using a MAR345 image plate and a Quantum-4 CCD detector, respectively. Three diffraction datasets were collected to a maximum resolution beyond 1.95 Å. Diffraction images were processed and scaled with the programs DENZO and SCALEPACK [41].

Heavy-Atom Derivatives and Phasing

The structure was solved by SIRAS. An iodine derivative was obtained by soaking the crystal for 180 s in 2 μl of cryoprotectant solution containing 0.125 M sodium iodide following a novel derivatization procedure named "quick cryosoaking" [24, 25]. One weak mercury derivative was also obtained using traditional methods of derivatization. However, this derivative did not provide independent phase information and was used as a quasi-native dataset. The heavy-atom positions of the iodine derivative were determined by direct methods with the programs DREAR [42] and SnB 2.1 [43]. The bimodal distribution of the R_{min} histogram was used to identify the correct solution [44, 45]. The heavy-atom substructure obtained directly from SnB was initially refined with the CNS package using anomalous and isomorphous difference Fourier maps. Refined coordinates were then input into SHARP [46] for phase calculation, resulting in an overall figure of merit of 0.45 for all reflections in the range of 21.7–2.40 Å. Density modification with solvent flattening was performed using the program SOLOMON [47].

Model Building and Refinement

A solvent-flattened electron density map and structure factor amplitudes from the iodine derivative were used by the ARP/wARP program [26] for an automatic build of a hybrid model of hIL-22. Due to the high resolution and completeness of the I-hIL-22 data set, an initial model was obtained without manual intervention after six ARP/wARP jobs and more than 4000 REFMAC [48] cycles. In the last cycle, after almost 72 hr of uninterrupted CPU time on a Pentium III 500 MHz, 81.6% of the total amino acid residues were correctly traced.

The initial model contained 231 amino acid residues (in nine distinct chains) and 809 water molecules. To avoid modeling of a large number of partially occupied iodine ions that were present at the concentration of 125 mM in the cryosolution of this derivative, subsequent refinement was performed against the Nat-hIL-22 data set. Construction of disordered loops and filling of main chain gaps were performed manually using the program O [49], and the CNS refinement against the Nat-hIL-22 and Hg-hIL-22 (quasi-native) data sets was performed when judged necessary. This allowed for a complete trace of the main chain atoms through disordered regions. Independent refinement of the model against I-hIL-22 data resulted in higher R factor and R_{free} values, due to a large number of halides with partial occupancies bound to the protein found in the electron density maps. After several iterations of energy minimization, B factor refinement, and bulk-solvent and anisotropic corrections, the final R factor and R_{free} against the Nat-hIL-22 data set were 0.188 and 0.220, respectively. The final model includes 283 residues divided into two chains and 189 water molecules.

Protein Oligomerization State

The protein oligomerization state has been assessed by gel filtration and dynamic light-scattering (DLS) studies. Gel filtration experiments were conducted in the conditions described in Protein Expression and Purification. DLS measurements have been performed with a DynaPro MS200 instrument (Protein Solutions) at 20°C using a 12 μl cuvette. The protein samples were concentrated to 5 mg/ml and 10 mg/ml in 0.1 M HEPES buffer at pH 7.5 prior to measurements.

Acknowledgments

The authors are grateful to J.R. Brandão Neto and Zbigniew Dauter for help with data collection, Katucha W. Lucchesi and Miroslawa Dauter for help with crystallization of the proteins, Dr. Nilson I.T. Zanchin and Profs. Munro Neville, Ricardo R. Brentani, and Rogério Meneghini for comments and suggestions, and Prof. Richard C. Garratt for style and grammar corrections. Financial help from CNPq and FAPESP (via grants 99/03387-4 and 98/06218-6) is acknowledged.

Received: November 16, 2001
Revised: April 30, 2002
Accepted: May 14, 2002

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. 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 an hepatocyte-stimulating factor. Proc. Natl. Acad. Sci. USA *97*, 10144–10149.
3. 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.
4. Kotenko, S.V., Izotova, L.S., Mirochnitchenko, O.V., Esterova, E., Dickensheets, H., Donnelly, R.P., and Pestka, S. (2001). Iden-

tification of the functional interleukin-22 (IL-22) receptor complex. J. Biol. Chem. *276*, 2725–2732.
5. Kotenko, S.V., and Pestka, S. (2000). Jak-stat signal transduction pathway through the eyes of cytokine class II receptor complexes. Oncogene *19*, 2557–2565.
6. Gallagher, G., Dickensheets, H., Eskdale, J., Izotova, L.S., Mirochnitchenko, O.V., Peat, J.D., Vazquez, N., Pestka, S., Donnelly, R.P., and Kotenko, S.V. (2000). Cloning, expression and initial characterization of interleukin-19 (IL-19), a novel homologue of human interleukin-10 (IL-10). Genes Immun. *1*, 442–450.
7. Blumberg, H., Conklin, D., Xu, W.F., Grossmann, A., Brender, T., Carollo, S., Eagan, M., Foster, D., Haldeman, B.A., Hammond, A., et al. (2001). Interleukin-20: discovery, receptor identification and role in epidermal function. Cell *104*, 9–19.
8. Jiang, H., Lin, J.J., Su, Z.Z., Goldstein, N.I., and Fisher, P.B. (1995). Subtraction hybridization identifies a novel melanoma differentiation associated gene, mda-7, modulated during human progression. Oncogene *11*, 2477–2486.
9. Moore, K.W., de Waal Malefyt, R., Coffman, R.L., and O'Garra, A. (2001). Interleukin-10 and the interleukin-10 receptor. Annu. Rev. Immun. *19*, 683–765.
10. Fickenscher, H., Hor, S., Kupers, H., Knappe, A., Wittmann, S., and Sticht, H. (2002). The interleukin-10 family of cytokines. Trends Immunol. *23*, 89–96.
11. Zdanov, A., Schalk-Hihi, C., Gustchina, A., Tsang, M., Weatherbee, J., and Wlodawer, A. (1995). Interleukin-10: crystal structure reveals the functional dimer with an unexpected topological similarity to interferon gamma. Structure *3*, 591–601.
12. Zdanov, A., Schalk-Hihi, C., and Wlodawer, A. (1996). Crystal structure of human interleukin-10 at 1.6 Å resolution and a model of a complex with its soluble receptor. Protein Sci. *5*, 1955–1962.
13. Walter, M.R., and Nagabhushan, T.L. (1995). Crystal-structure of interleukin-10 reveals an interferon gamma-like fold. Biochemistry *34*, 12118–12125.
14. Zdanov, A., Schalk-Hihi, C., Menon, S., Moore, K.W., and Wlodawer, A. (1996). Crystal structure of Epstein-Barr virus protein BCRF1, a homolog of cellular interleukin-10. J. Mol. Biol. *268*, 460–467.
15. Temann, U.A., Geba, G.P., Rankin, J.A., and Flavell, R.A. (1998). Expression of interleukin-9 in the lungs of transgenic mice causes airway inflammation, mast cell hyperplasia, and bronchial hyperresponsiveness. J. Exp. Med. *188*, 1307–1320.
16. McLane, M.P., Haczku, A., van de Rijn, M., Weiss, C., Ferrante, V., MacDonald, D., Renauld, J.C., Nicolaides, N.C., Holroyd, K.J., and Levitt, R.C. (1998). Interleukin-9 promotes allergen-induced eosinophilic inflammation and airway hyperresponsiveness in transgenic mice. Am. J. Respir. Cell Mol. Biol. *19*, 713–720.
17. Dumoutier, L., Van Roost, E., Ameye, G., Michaux, L., and Renauld, J.C. (2000). IL-TIF/IL-22: genomic organization and mapping of the human and mouse genes. Genes Immun. *1*, 488–494.
18. Renauld, J.C. (2001). New insights into the role of cytokines in asthma. J. Clin. Pathol. *54*, 577–589.
19. Kotenko, S.V., Krause, C.D., Izotova, L.S., Pollack, B.P., Wu, W., and Pestka, S. (1997). Identification and functional characterization of a second chain of the interleukin-10 receptor complex. EMBO J. *16*, 5894–5903.
20. Thiel, D.J., le Du, M.H., Walter, R.L., D'Arcy, A., Chene, C., Fountoulakis, M., Garotta, G., Winkler, F.K., and Ealick, S.E. (2000). Observation of an unexpected third receptor molecule in the crystal structure of human interferon-gamma receptor complex. Structure *8*, 927–936.
21. Josephson, K., Logsdon, N.J., and Walter, M.R. (2001). Crystal structure of the IL10-IL-10R1 complex reveals a shared receptor binding site. Immunity *14*, 35–46.
22. Polikarpov, I., Perles, L.A., de Oliveira, R.T., Oliva, G., Castellano, E.E., Garratt, R.C., and Craievich, A. (1997). Set-up and experimental parameters of the protein crystallography beamline at the Brazilian National Synchrotron Laboratory. J. Synchrotron Radiat. *5*, 72–76.
23. Polikarpov, I., Oliva, G., Castellano, E.E., Garratt, R.C., Arruda, P., Leite, A., and Craievich, A. (1997). The protein crystallography beamline at LNLS, the Brazilian National Synchrotron Light Source. Nucl. Instrum. Methods Phys. Res. A *405*, 159–164.
24. Dauter, Z., Dauter, M., and Rajashankar, K.R. (2000). Novel approach to phasing proteins: derivatization by short cryo-soaking with halides. Acta Crystallogr. D Biol. Crystallogr. *56*, 232–237.
25. Nagem, R.A.P., Dauter, Z., and Polikarpov, I. (2001). Protein crystal structure solution by fast incorporation of negatively and positively charged anomalous scatterers. Acta Crystallogr. D Biol. Crystallogr. *57*, 996–1002.
26. Perrakis, A., Morris, R., and Lamzin, V.S. (1999). Automated protein model building combined with iterative structure refinement. Nat. Struct. Biol. *6*, 458–463.
27. Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., et al. (1998). Crystallography and NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr. D Biol. Crystallogr. *54*, 905–921.
28. 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*, 283–291.
29. Kraulis, P.J. (1991). Molscript—a program to produce both detailed and schematic plots of protein structures. J. Appl. Crystallogr. *24*, 946–950.
30. Esnouf, R.M. (1997). An extensively modified version of Molscript that includes greatly enhanced coloring capabilities. J. Mol. Graph. *15*, 133–138.
31. Merritt, E.A., and Bacon, D.J. (1997). Raster3D: photorealistic molecular graphics. Methods Enzymol. *277*, 505–524.
32. Nicholls, A., Sharp, K.A., and Honing, B. (1991). Protein folding and association—insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins: Struct. Funct. Genet. *11*, 281–296.
33. Ealick, S.E., Cook, W.J., Vijay-Kumar, S., Carson, M., Nagabhushan, T.L., Trotta, P.P., and Bugg, C.E. (1991). Three-dimensional structure of recombinant human interferon-gamma. Science *252*, 698–702.
34. Josephson, K., DiGiacomo, R., Indelicato, S.R., Iyo, A.H., Nagabhushan, T.L., Parker, M.H., Walter, M.R., and Ayo, A.H. (2000). Design and analysis of an engineered human interleukin-10 monomer. J. Biol. Chem. *275*, 13552–13557.
35. Barton, G. (1993). ALSCRIPT: a tool to format multiple sequence alignments. Protein Eng. *6*, 37–40.
36. Garcia de La Torre, J., Huertas, M.L., and Carrasco, B. (2000). Calculation of hydrodynamic properties of proteins from their atomic level structures. Biophys. J. *78*, 719–730.
37. Trèze, J. (1999). The Cytokine Network and Immune Functions (Oxford: Oxford University Press).
38. Goger, B., Halden, Y., Rek, A., Mosl, R., Pye, D., Gallagher, J., and Kungl, A.J. (2002). Different activities of glycosaminoglycan oligosaccharides for monomeric and dimeric interleukin-8: a model for chemokine regulation at inflammatory sites. Biochemistry *41*, 1640–1646.
39. Laurence, J.S., Blanpain, C., Burgner, J.W., Parmentier, M. and LiWang, P.J. (2000). CC chemokine MIP-1β can function as a monomer and depends on Phe13 for receptor binding. Biochemistry *39*, 3401–3409.
40. Dumoutier, L., Lejeune, D., Colau, D., and Renauld, J.-C. (2001). Cloning and characterization of Interleukin-22 binding protein (IL-22BP), a natural antagonist of IL-TIF/IL-22. J. Immunol. *166*, 7090–7095.
41. Otwinowski, Z., and Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol. *276*, 307–326.
42. Blessing, R.H., and Smith, G.D. (1999). Difference structure-factor normalization for heavy-atom or anomalous-scattering substructure determinations. J. Appl. Crystallogr. *32*, 664–670.
43. Weeks, C.M., and Miller, R. (1999). The design and implementation of SnB version 2.0. J. Appl. Crystallogr. *32*, 120–124.
44. Debaerdemaeker, T., and Woolfson, M.M. (1983). On the application of phase-relationships to complex structures. Techniques for random phase refinement. Acta Crystallogr. A *39*, 193–196.
45. De Titta, G.T., Weeks, C.M., Thuman, P., Miller, R., and Hauptman, H.A. (1994). Structure solution by minimal-function phase

refinement and Fourier filtering. Theoretical basis. Acta Crystallogr. A *50*, 203–210.
46. de La Fortelle, E., and Bricogne, G. (1997). Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. Methods Enzymol. *276*, 472–494.
47. Abrahams, J.P., and Leslie, A.G.W. (1996). Methods used in the structure determination of bovine mitochondrial F-1 ATPase. Acta Crystallogr. D Biol. Crystallogr. *52*, 30–42.
48. Murshudov, G.N., Vagin, A.A., and Dodson, E.J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr. D Biol. Crystallogr. *53*, 240–255.
49. Jones, T.A., Zou, J.Y., Cowan, S.W., and Kjeldgaard, M. (1991). Improved methods for building protein models in electron-density maps and the location of errors in these models. Acta Crystallogr. A *47*, 110–119.

Accession numbers

The atomic coordinates and the structure factors are deposited with the Protein Data Bank under accession code 1M4R (RCSB ID code RCSB016596).

Crystal structures of β-galactosidase from *Penicillium sp.* and its complex with galactose

Rojas, A. L.[*§], Nagem, R. A. P.[†#§], Neustroev, K. N.[‡], Golubev, A. M.[‡], Eneyskaya, E. V.[‡], Kulminskaya, A. A.[‡] and Polikarpov, I. [*†]

[*] *Instituto de Física de São Carlos, Universidade de São Paulo, Av. Trabalhador Sãocarlense 400, CEP 13560-970, São Carlos, SP, Brazil*
[†] *Laboratório Nacional de Luz Síncrotron, Caixa Postal 6192, CEP 13084-971, Campinas, SP, Brazil*
[#] *Instituto de Física Gleb Wataghin, UNICAMP, CEP 13083-970, Campinas, SP, Brazil*
[‡] *Petersburg Nuclear Physics Institute, Gatchina, St. Petersburg, 188300, Russia*

[§] *A.L.R. and R.A.P.N. contributed equally to this work.*

[¶] *Corresponding author.*
Email: ipolikarpov@if.sc.usp.br

The crystallographic structures of *Penicillium sp.* β-galactosidase and its complex with galactose were solved by SIRAS quick cryo-soaking technique at 1.90 Å and 2.10 Å resolution, respectively. A native structure, containing a single molecule of 971 amino acid residues in the asymmetric unit cell, was refined to a final R_{free} = 18.2% (R_{factor} = 16.5%). Even though the amino acid sequence of this enzyme is unknown from molecular biology studies, the excellent quality of X-ray data enabled us to derive the protein primary structure directly from the electron density. On the base of the primary structure alignments with close homologues, it was shown that the β-galactosidase belongs to the family 35 of glycosyl hydrolases. This is the first 3D model of a member of this family. The X-ray structure of enzyme-galactose complex was used to identify two putative glutamic acid residues involved in the reaction mechanism with retention of anomeric configuration. Superposition of the enzyme-galactose complex with other β-galactosidase complexes from different hydrolase families allowed identification of residues Glu160 as the proton donor and Glu259 as the nucleophile of the catalysis. The *Penicillium sp.* β-galactosidase is a glycoprotein containing seven N-linked oligosaccharide chains. Among the β-galactosidase structures deposited in the Protein Data Bank, this is the only model of a glycosilated enzyme.

Keywords: glycosyl hydrolases, β-galactosidases, crystal structure, *Penicillium sp.*, quick cryo soaking

Introduction

β-Galactosidase (E.C. 3.2.1.23) is an enzyme that hydrolyses β(1-3) and β(1-4) galactosyl bonds in poly- and oligosaccharides. Many carbohydrases have been found to catalyze not only the hydrolysis reaction but also the reaction of condensation or transglycosylation. This property has been observed in β-galactosidases from several sources and is being used for the enzymatic synthesis of galactose containing oligosaccharides. The high degree of stereospecificity enables these enzymes to be applied for chemo-enzymatic synthesis of galactooligosaccharides with β(1-3) and β(1-4) bond configurations.[1,2] The enzymes from *Bacillus circulans* and from bovine testes were successfully utilized to the synthesis of Galβ1-3GlcNAc, allowing the study of sialyl-Lewis antigen key oligosaccharide.[3,4] The same transglycosylation activity was used for the synthesis of Galβ1-3GalNAc, an important constituent of the mucin type glycoproteins, and for the synthesis of N-acetyllactosamine.[5] Sialyl N-acetyllactosamine was obtained with β-galactosidase from *Diplococcus pneumoniae* and sialidases from various origins.[6] The transglycosylation ability of β-galactosidase was also used to obtain β-galactosyl-serine derivatives.[7]

β-Galactosidases are used for the hydrolysis of lactose in dairy products, such as milk and cheese whey, and thermostable β-galactosidases recently attracted special interest due to their potential usefulness in the industrial processing of lactose-containing products.[8] β-Galactosidases have been used as a tool in molecular biology due to the fairly strict specificity toward the galactosyl bonds but not in respect to wide variety of aglycons. This feature allows the use of simple, colorimetric assays based on substrates such as para-nitrophenyl-β-D-galactopyranoside (PNPG) and β-D-galactopyranoside with chromogenic aglycons (XGal). The extracellular β-galactosidase from *Penicillium sp.* has a high transglycosylation activity toward p-nitrophenyl β-D-galactopyranoside (PNPG), lactose and methyl β-D-galactopyranoside, and is a promising tool for a number of enzymatic syntheses.

Traditionally, glycosidases were grouped together based on the ability to hydrolyze similar substrates; for example, enzymes that cleave off galactose from

lactose, PNPG or XGal were classified as β-galactosidases. Based on amino acid sequence similarities, the glycosyl hydrolases (GHs) have been classified into 90 families.[9-13] Enzymes exhibiting β-galactosidase activity were divided into four families termed GH-1, GH-2, GH-35, and GH-42. Based on hydrophobic cluster analysis (HCA)[11] GHs have also been classified into several superfamilies called "clans". A clan is a group of families that are thought to have a common ancestry and are recognized by significant similarities in tertiary structure together with conservation of the catalytic residues and catalytic mechanism. The largest member of these superfamilies is the glycoside hydrolase clan GH-A, which comprises families 1, 2, 5, 10, 17, 26, 30, 35, 39, 42, 51 and 53. GH-A consists of enzymes that have a TIM barrel fold in the catalytic domain and cleave the glycosidic bonds through the retaining mechanism. As a rule, this mechanism involves two glutamic acid residues that are the proton donor and the nucleophile, with an asparagine always preceding the proton donor.[14] This clan is sometimes described as the 4/7 superfamily because the proton donor and nucleophile are found on strands 4 and 7 of the $(\beta/\alpha)_8$ barrel, respectively.[15] An updated list of enzymes within clan GH-A with known three-dimensional structure is available through the ExPASy WWW server[16] at the URL: 'http://bo.expasy.org/cgi-bin/lists?glycosid.txt'.

Nowadays, only 3D models of enzymes from families 1, 2, 5, 10, 17 and 42 are available. On the other hand, enzymes from families 26, 30, 35, 39, 51, 53 have not been structurally characterized yet.

The β-galactosidase from *Escherichia coli* (Ec-β-gal), which belongs to GH-2, is one of the most studied β-galactosidases. However, many other important β-galactosidases are now being discovered in phylogenetically diverged organisms. The three-dimensional structure of Ec-β-Gal is known and the structural basis for its reaction mechanism was reported.[17-19] Apart from that, *Thermus thermophilus* A4 β-galactosidase (A4-β-gal) is the unique GH-42 enzyme whose crystallographic structure was very recently determined.[8]

In this work we describe the first crystal structure of β-galactosidase from *Penicillium sp.* (Psp-β-gal) and its complex with galactose determined by quick cryo-soaking approach. Amino acid sequence of the protein is unknown but the excellent quality of X-ray data enabled us to derive primary structure from the experimentally computed electron density maps. Amino acid sequence comparison[20] shows that *Penicillium sp.* β-galactosidase model is the first three-dimensional structure of a member of GH family 35. The Psp-β-gal catalytic residues Glu160 and Glu259 were identified as a proton donor and as a nucleophile of reaction. In spite of low similarity in primary structure with the other members of the superfamily 4/7, Psp-β-gal has the distinct structural features of the superfamily including the spatial location of two catalytic residues and the characteristic distance between their carboxylic groups.

Results and Discussion

Description of the 3D model

The native crystallographic structure of Psp-β-gal was solved at 1.90 Å resolution with final R_{factor} and R_{free} of 16.5 and 18.2 %, respectively. The model has good stereochemical parameters, with a root mean square deviation (rmsd) of 0.006 Å for bond distances, a rmsd of 1.4° for bond angles and an average b-factor of 22.2 $Å^2$ for all atoms. Only 2 amino acid residues (Asn100 and Asp461) are found in the disallowed regions of the Ramachandran plot. A sequence of hydrogen bond interactions between Asn100 Oδ1 and Tyr98 OH atoms (2.61 Å) and between Tyr98 OH and Asn768 N atoms (2.69 Å) may explain the otherwise energetically unfavorable conformation residue Asn100. Asp461 is located on the protein surface and an electron density observed for this amino acid residue is poor. Four cysteine residues form two disulfide bridges (Cys165-Cys166 and Cys227-Cys276). Two alternative conformations were found for the side chains of the residues Ile22, Val613 and Tyr890.

Obtained Psp-β-gal model comprises 971 amino acid residues and its tertiary structure can be divided into 5 domains (Figure 1). Primary structure comparison with β-galactosidases from *Aspergillus candidus* and *Aspergillus niger* and two β-galactosidase's fragments from *Penicillium canescens* and *Talaromyces emersonii* indicates that approximately thirty N-terminal amino acid residues might be missing in the current model. However, this region, if present, is disordered and is not observed in the electron density maps. In addition to 971 amino acid residues, the crystallographic model contains 16 mannoses (MAN) and 12 N-acetylglucosamines (NAG) molecules in 7 oligosaccharide chains, 1256 water molecules, 3 sodium ions, 4 phosphate ions and 9 ethylene glycol molecules.

The first domain of Psp-β-gal, containing the catalytic site, is a distorted TIM barrel. It comprises 355 amino acid residues from Ala1 to Gly355. Unlike normal TIM barrels that consist of eight β/α repeats, the β/α barrel in Psp-β-gal lacks the fifth helix, and the seventh parallel strand of the barrel is distorted. This feature was also observed in the crystal structure of Ec-β-gal.[17] In Ec-β-gal structure, however, the TIM barrel is the third domain (see section 2.3 for details). An anti-parallel β-sheet, formed by N-terminal amino acid residues (Ala1 to Leu20), seals the base of the TIM barrel. The following two domains are also located at the bottom of the TIM barrel. The second domain, comprising amino acid residues Tyr356 to Tyr536, consists of 16 consecutive antiparallel β-strands and an

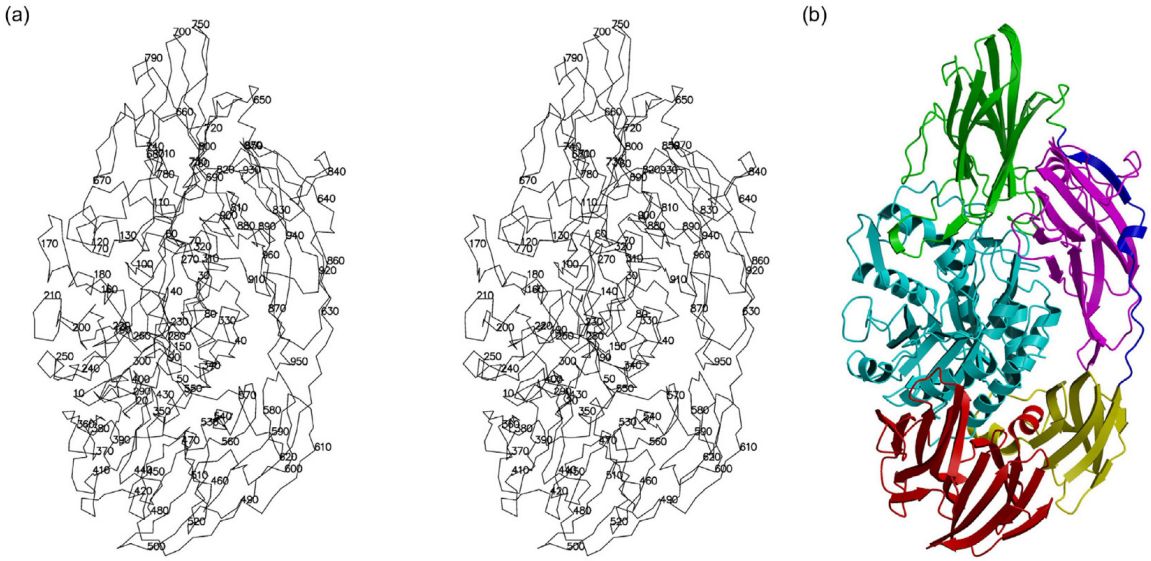


Figure 1. Psp-β-gal is a 120kDa monomer composed of five distinct structural domains. The overall structure is built around the first, TIM barrel, domain. Domain 3 has a jelly-roll barrel fold and domains 4 and 5 display immunoglobulin folds. (a) Stereo view of the C_α trace of Psp-β-gal. (b) Ribbon representation of the secondary structure elements of Psp-β-gal, according to PROCHECK.[43] Domains 1, 2, 3, 4 and 5 are colored in cyan, red, yellow, green and magenta respectively. The long polypeptide chain connecting domains 3 and 4 is depicted in blue. The figures were drawn using the programs Molscript,[44] Bobscript [45] and Raster3D. [46]

α-helix at the C-terminus. On the other hand, the third domain (Trp537 to Tyr625) is two times smaller than the second domain and consists of an α-helix at the N-terminus followed by 8 consecutive antiparallel β-strands arranged in a jelly-roll fold. The third and fourth domains are joined together by a long polypeptide chain formed by 23 amino acid residues (Thr626 to Pro647). This chain starts at the third domain at the bottom of the Psp-β-gal and goes along the molecule surface to the top of the protein where domains 4 and 5 are located. However, this long polypeptide chain does not go straight to the fourth domain, instead, it passes inside the fifth domain (Tyr822 to Tyr971) and forms a β-strand that interacts with the others β-strands of this domain. The fourth domain that comprises amino acid residues Glu648 to Leu821 has a total of 12 β-strands. Ten of these build up the main core of the domain, while two of them participate in interactions with the first domain. Both fourth and fifth domain display immunoglobulin fold.

Primary structure identification

All attempts to solve the crystal structure of Psp-β-gal using the Molecular Replacement method (MR) with *E. coli* and *T. thermophilus* β-galactosidases as search models were unsuccessful. The phase problem was solved by single isomorphous replacement with anomalous scattering (SIRAS) method. An initial hybrid model consisted of a poly-glycine peptide chain and water molecules was built automatically in the SIRAS-derived electron density map. Since the amino acid sequence of this enzyme was unknown, the identity of amino acid residues was derived from the experimental electron density map. To improve the phase estimates and a quality of the electron density map, a complete multiple isomorphous replacement with anomalous scattering (MIRAS) phase calculation using four data sets up to 1.9 Å of resolution (one native and three derivatives) was performed with SHARP.[21]

The resulting phases had an overall figure of merit of 0.61 in the range of 27.0 – 2.40 Å. The MIRAS derived phases together with the molecular envelop derived from the poly-glycine model were used in a density modification procedure with SOLOMON.[22] The final phases were of excellent quality and allowed for electron density map calculations that were used for primary structure identification. Moreover during side chain assignment process, amino acid sequence alignments were performed between β-galactosidases from different organisms and the protein under study. This information was used in conjunction with anomalous difference Fourier maps, potential hydrogen bonds and protein-oligosaccharide bonds to identify most of the residues in the model. In Figure 2 the X-ray derived primary structure of Psp-β-gal is aligned with a number of homologues.

Several relevant crystallographic parameters for each amino acid residue are also displayed. This comparison allowed us to undoubtedly assign Psp-β-gal as a member of family 35.

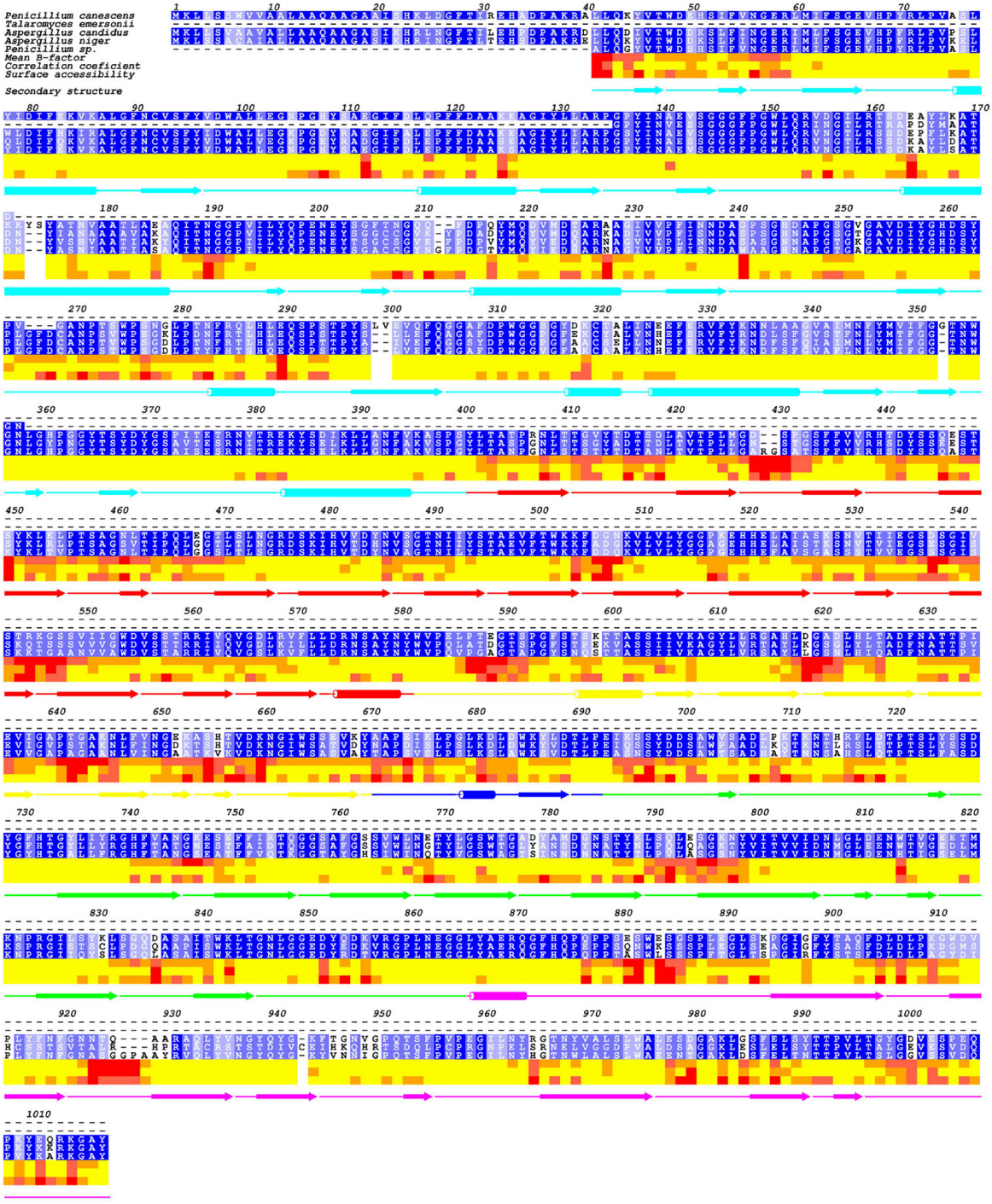


Figure 2. Primary structure alignment of β-galactosidase from *Penicillium canescens* (171 aa), *Talaromyces emersonii* (218 aa), *Aspergillus candidus* (1005 aa), *Aspergillus niger* (1006 aa) and a putative amino acid sequence of *Penicillium sp.* β-galactosidase derived from electron density. The amino acid similarity between primary structures as calculated by the program ALSCRIPT [47] is shown in three different shades of blue. The darker color corresponds to higher sequence similarity. Mean B_{factor}, correlation coefficient (CC) between predicted and observed electron density and total exposed surface area per residue are shown in a color scale from yellow to red. Four different colors were used for different ranges of B_{factor}, CC and surface area. Red (Mean B_{factor} higher than 30 Å^2; CC lower than 95%; Total exposed surface area higher than 90 Å^2). Light red (Mean B_{factor} higher than 25 and lower than 30 Å^2; CC lower than 96 and higher than 95%; Total exposed surface area higher than 60 and lower than 90 Å^2). Orange (Mean B_{factor} higher than 20 and lower than 25 Å^2; CC lower than 97 and higher than 96%; Total exposed surface area higher than 30 and lower than 60 Å^2). Yellow (Mean B_{factor} lower than 20 Å^2; CC higher than 97%; Total exposed surface area lower than 30 Å^2). Secondary structure elements for Psp-β-gal are shown in cyan, red, yellow, green and magenta for domains 1, 2, 3, 4 and 5, respectively. The long polypeptide chain connecting domains 3 and 4 is depicted in blue.

Comparison of the Psp-β-gal and two known β-galactosidase structures

The crystallographic structure of Psp-β-gal is the first 3D model of a β-galactosidase from fungi. Psp-β-gal is a five-domain monomer with molecular mass estimated by SDS-PAGE to be approximately 120 kDa. The first domain exhibits a TIM barrel fold whilst domains 4 and 5 have immunoglobulin-like structure and domain 3 has a jelly-roll fold. All β domain 2, to our knowledge, has no clear similarity to any known fold.

To date, only two other β-galactosidases, both from bacteria, were described: *E. coli* β-galactosidase (Ec-β-gal) and *T. termophilus* β-galactosidase (A4-β-gal). Ec-β-gal is a 464 kDa homotetramer and each subunit is also composed of five domains. However, the main difference between Ec-β-gal and Psp-β-gal monomers is the position of each domain. The overall structure of Ec-β-gal is built around the third domain (a TIM barrel). The other domains consist mainly of β-sheets with domain 1 having a jelly-roll fold and domains 2 and 4 having immunoglobulin folds. Domain 5 is a β-sandwich. According to Jacobson and co-workers,[17] domains 1, 2, 4 and 5 of Ec-β-gal are essentially independent folding modules that serve to supplement or modify the main role that is played by domain 3. In Psp-β-gal, apart from domain 5, that is formed by two non-consecutive parts of the main chain (Tyr625-Pro647 and Leu821-Tyr971), all other domains can be also considered as independent folding modules.

Contrary to both Psp-β-gal and Ec-β-gal, the recombinant A4-β-gal structure is trimeric. Each A4-β-gal monomer is composed of only three domains: a TIM barrel fold domain, an α/β fold domain and a β fold domain.[8]

As a result of the differences observed between the crystallographic structures of *Penicillium sp.*, *E. coli* and *T. thermophilus* β-galactosidases, a reasonable superposition of the monomers was proved to be impossible. Nevertheless, a superposition of TIM barrel domains of all three β-galactosidases could be performed (Figure 3). A root mean square deviation of 2.8 and 3.3 Å was obtained for 236 and 291 pairs of Cα atoms from Psp-β-gal/Ec-β-gal and Psp-β-gal/A4-β-gal TIM barrel fold domains respectively.

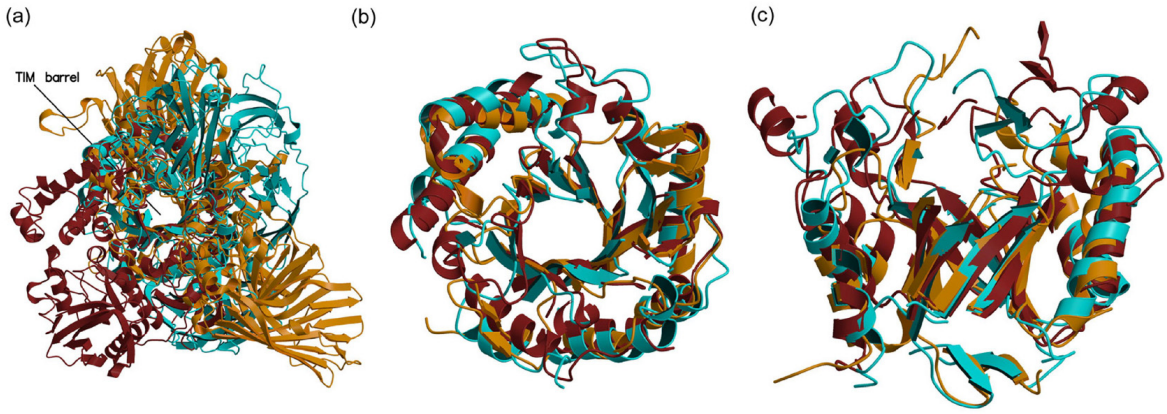


Figure 3. (a) Overall view of the superposition of Psp-β-gal (in cyan), Ec-β-gal (in orange) and A4-β-gal (in brown) monomers. (b) Top and (c) side view of the superposition of their respective TIM barrel domains.

Galactose-binding site

Analysis of the galactose-binding site was performed based on comparison of the native structure and the structure of the galactose-containing complex of enzyme. A single galactose molecule is bound to the TIM barrel domain of Psp-β-gal in the chair conformation with its O1 atom in the β-anomer configuration. Figure 4 shows the bound galactose molecule and the neighboring amino acid residues. Close contacts between protein atoms, water molecules and OH groups of galactose are given in Table 1. The electron density around the sugar molecule, depicted in Figure 4, undoubtedly indicates its presence in the catalytic site.

Primary structure comparison between β-galactosidases from different organisms (Figure 2) indicates that of the 11 residues involved in galactose binding, 9 are well conserved among species. The only difference is observed in *T. emersonii* β-galactosidase where residues Ala101 and Glu259 are substituted to valine and glutamine, respectively.

The enzymatic hydrolysis of the bond between sugar molecules catalyzed by glycoside hydrolases takes place via a general acid catalysis that requires two critical residues: a proton donor and a nucleophile/base.[23]

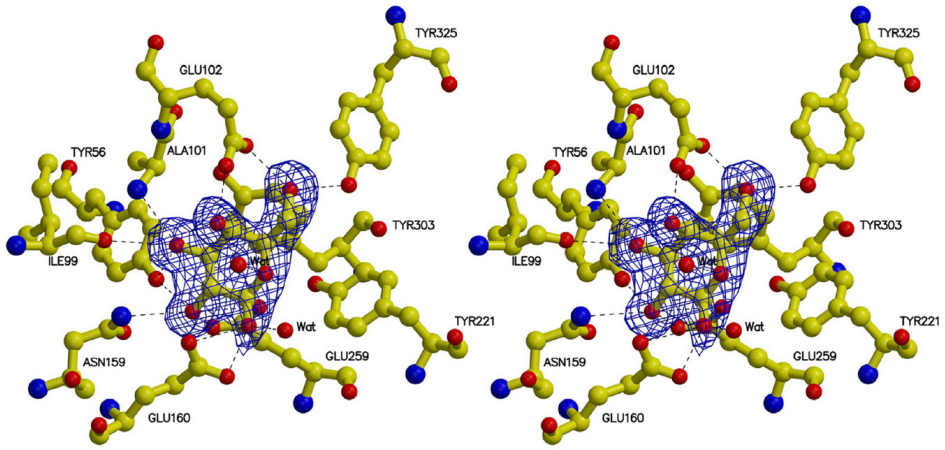


Figure 4. Stereo view representation of galactose-protein interactions in the catalytic site of Psp-β-gal. Galactose and the neighboring residues are shown with an omit electron density map ($mF_{obs} - DF_{calc}$, φ_{calc}) around galactose countered at 3 sigma.

Table 1. Close contacts between protein and galactose non-carbon atoms. A distance cut-off of 3.3 Å was used.

	Close contacts	
Sugar atom	**Protein atom**	**Distance (Å)**
O1	Glu259 OE1	3.3
	Glu160 OE1	3.1
	Glu160 OE2	2.5
	Water (612)	3.0
O2	Glu259 OE1	3.2
	Tyr56 OH	3.2
	Asn159 ND2	3.0
	Glu259 OE2	2.7
O3	Ala101 N	2.9
	Tyr56 OH	2.8
	Ile99 O	3.1
O4	Glu102 OE2	2.7
	Asn100 ND2	2.9
	Water (586)	3.3
O5	Water (586)	3.3
O6	Tyr325 OH	2.7
	Glu102 OE1	2.7

This hydrolysis occurs via two major mechanisms giving rise to either an overall retention or an inversion of the anomeric configuration of substrate.[24] In both the retaining and the inverting mechanisms, the positions of the proton donor are identical, which is found within hydrogen bond distance of the glycosidic oxygen. In retaining enzymes, the nucleophilic catalytic base is in close vicinity of the sugar anomeric carbon. This base, however, is more distant in inverting enzymes that must accommodate a water molecule between the base and the sugar. Consequently, the average distance between the two catalytic residues is around 5.5 and 10 Å in the retaining and inverting enzymes respectively.[25]

Based on a 3D superposition of Psp-β-gal and Ec-β-gal (Figure 5), the two putative Psp-β-gal catalytic residues were identified as Glu160 and Glu259. The first was also inferred to be the proton donor while the last proposed as the nucleophile of the catalysis. Glu160 and Glu259 are located, respectively, at the C-terminal of β-4 and β-7 strands of the TIM barrel and are close to C1 position of the galactose molecule. The presence of catalytic residues at the beta strands 4 and 7 is one of the most important characteristics of the 4/7 superfamily, which classifies Psp-β-gal as its member. The average distance between all four pairs of side-chain oxygen atoms of Glu160 and Glu259 is 4.8 Å that is in an appropriate range for retaining enzymes. This strongly suggests that Psp-β-gal belongs to the group of enzymes that retain the overall anomeric configuration of the substrate. The presence of an asparagine residue preceding the proton donor Glu160 reinforces the classification of Psp-β-gal as a glycoside hydrolase with a retaining mechanism of catalysis.[14]

It has been reported that several members of the 4/7 superfamily including β-galactosidases from *E. coli* and *T. termophilus* have Trp/Phe at the end of β-8.[19] The Trp/Phe residue constitutes subsite-1 with the aromatic side-chain and forms a cis-peptide bond with the next residue. Psp-β-gal contains a tyrosine residue (Tyr303) at this position. It forms the only non-proline cis-peptide bond in the entire molecule. This residue as well as Met304 is well conserved among β-galactosidases (Figure 2) and the electron density map around these residues leaves no doubt about their identity and configuration. Inspection of the superimposed structures of all three available β-galactosidases (Psp-β-gal, Ec-β-gal, A4-β-gal) shows the aromatic rings of Tyr303 (Psp-β-gal), Phe350 (A4-β-gal) and Trp568 (Ec-β-gal) in the same orientation with respect to the bound galactose. We cannot predict the effect of this substitution on the enzymatic activity of Psp-β-gal solely on the base of the current structure. The directed mutagenesis experiments in association with kinetic studies may help to answer this question.

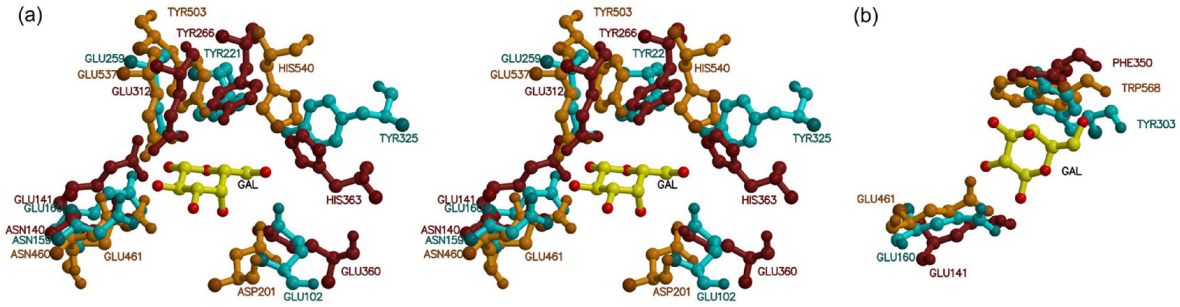


Figure 5. (a) Stereo view of the superposition of the catalytic sites in Psp-β-gal (in cyan), Ec-β-gal (in orange) and A4-β-gal (in brown). (b) Detailed view of the catalytic subsite-1 showing the orientation of the aromatic rings from Tyr303 (Psp-β-gal), Phe350 (A4-β-gal) and Trp568 (Ec-β-gal) in respect to the bound galactose. In both pictures only the galactose from Psp-β-gal complex is shown.

Carbohydrate components

Even thought an inspection of the Protein Data Bank indicates that as many as 70% of the deposited proteins have potential N-glycosylation sites (Asn-X-Ser/Thr, where X is not proline), crystallographic structures of the glycosilated proteins are not so common. The crystallographic studies of oligosaccharides and the glycan components of glycoproteins are known to be notoriously difficult.[26] Crystals of oligosaccharides are difficult to obtain whereas crystallographic disorder frequently leads to a lack of identifiable electron density for the glycan components of glycoproteins hampering their structural characterization. As a result, relatively few oligosaccharide structures have been determined, hindering the process of structural identification of general conformational rules of glycosidic linkages. About 50% of glycoprotein structures contain between one and three resolved sugar residues, as a result of a high degree of glycan mobility or disorder. There are very few glycoproteins in which seven or more residues of a single glycan were resolved.[27]

Seven N-glycosylation sites have been localized in the electron density map of Psp-β-gal, and three of them have 5, 7 and 9 monosaccharide residues, respectively. Several oligosaccharides are wrapped around domains 3, 4 and 5 of β-galactosidase. A number of hydrogen bonds between amino acid residues at the protein surface and the atoms of carbohydrate moieties are observed. The electron density for a representative N-liked carbohydrate chain is shown in Figure 6.

Search of potential N-glycosylation sites in β-galactosidases used for primary structure comparison (Figure 2), indicates that all of them posses more than 7 NXS/T repeats (β-galactosidase fragments were not considered). However, only two bacterial non-glycosilated β-galactosidase structures have been solved and deposited in the Protein Data Bank (PDB) so far. Therefore, this is the first crystallographic report of the native fungal glycosilated β-galactosidase structure with several clearly defined oligosaccharides attached to it.

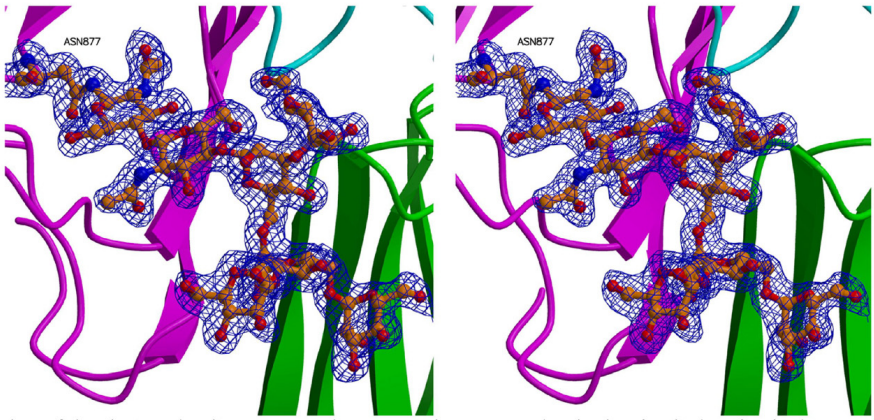


Figure 6. Stereo view of the electron density map around representative sugar molecule showing its location in the crystallographic structure of native Psp-β-gal.

The conformational behavior of the linkage between an N-glycan and corresponding asparagine side chain has been well characterized by NMR studies of glycopeptides in solution. These studies demonstrate that the asparagines linkage is relatively rigid and planar, with a tendency to extend the first glycan residue away from the peptide backbone and into the solvent.[28,29] In Psp-β-gal five different glycosidic linkages (Manβ1-4GlcNAc, GlcNAcβ1-4GlcNAc, Manα1-3Man, Manα1-2Man, Manα1-6Man) were identified. Their topology is shown in Figure 7.

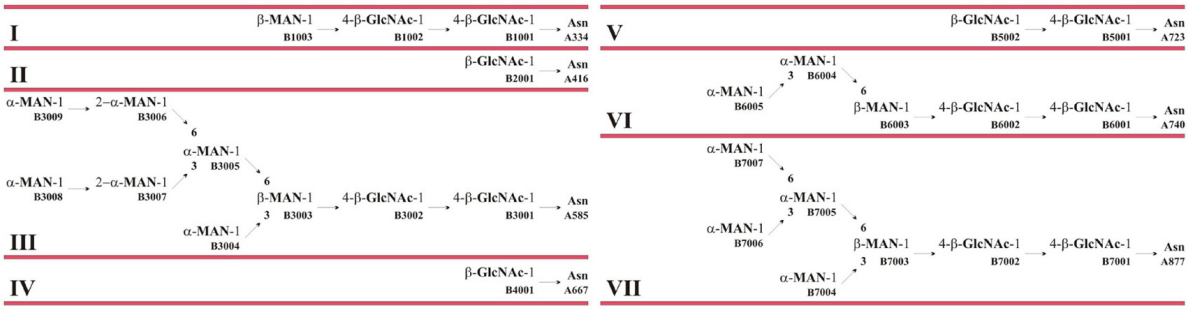


Figure 7. Structures of all seven N-linked oligosaccharides found in Psp-β-gal crystallographic structure. Five different glycosidic linkages (Manβ1-4GlcNAc, GlcNAcβ1-4GlcNAc, Manα1-3Man, Manα1-2Man and Manα1-6Man) were identified on the basis of electron density maps and previous biochemical studies.

In the last decade, Petrescu and co-workers[27] composed a database with all crystallographic information on glycans using several structures of oligosaccharides covalently attached to proteins deposited in the PDB. Using simple statistical analysis they were able to identify distinct conformers for each linkage type, providing average structures for glycosidic linkages.

A comparison between Psp-β-gal oligosaccharides and Petrescu's database indicates that most of the glycosidic linkages found in the crystallographic structure of β-galactosidase are commonly found in other proteins. The glycosidic linkages Manβ1-4GlcNAc, GlcNAcβ1-4GlcNAc and Manα1-3Man are identified as having a single conformer in Petrescu's classification, while Manα1-2Man and Manα1-6Man are found to have two and three distinct conformers, respectively (see Table 2 for details).

Table 2. Comparison between glycoside linkages of Psp-β-gal and Petrescu's conformers classification.

	Avg. linkage torsion angles for distinct conformers Petrescu's classification‡			**Linkage torsion angles for distinct conformers Psp-β-gal**			
Glycosidic linkage	**ϕ**	**ψ**	**ω**	**Residues**	**ϕ**	**ψ**	**ω**
GlcNAcβ1-4GlcNAc	-73.7±8.4	116.8±15.6	-	1001/1002	-83.0*	94.8*	-
				3001/3002	-76.2	117.3	-
				5001/5002	-63.2*	135.8*	-
				6001/6002	-69.6	134.4*	-
				7001/7002	-80.1	110.0	-
Manβ1-4GlcNAc	-88.0±10.8	107.9±20.3	-	1002/1003	-84.5	136.4*	-
				3002/3003	-68.7*	134.3*	-
				6002/6003	-96.9	97.9	-
				7002/7003	-62.0**	132.2*	-
Manα1-2Man	62.2±8.3	-175.0±10.3	-	3006/3009	82.2	-100.1	-
	71.9±13.1	-104.4±15.4	-	3007/3008	65.8	-125.5*	-
Manα1-3Man	72.5±11.0	-112.3±22.5	-	3003/3004	-108.6**	-131.1*	-
				3005/3007	66.9	-131.1	
				6004/6005	94.9**	-78.6*	-
				7003/7004	72.8	-113.8	-
				7005/7006	79.8	-94.1	-
Manα1-6Man	65.4±9.0	182.6±5.1	66.4±10.2	3003/3005	70.0	99.8	39.5**
				7005/7007	53.1**	170.6**	52.3**
	66.5±10.8	180.7±15.1	185.0±11.2	3005/3006	77.7*	224.3**	181.4
	67.4±14.4	109.1±13.7	203.0±22.7	6003/6004	72.4	190.1	185.7
				7003/7005	52.2*	187.8*	48.0*

‡ϕ = O5-C1-O-C(x)', ψ = C1-O-C(x)'-C(x-1) for 1-2, 1-3 and 1-4 linkages x = 2, 3 or 4 and ϕ = O5-C1-O-C(6)', ψ = C1-O-C(6)'-C(5), ω = O-C6'-C5'-C4' for 1-6 linkages.

* The values slightly deviating from average values.

** The values deviating more than 2σ from average values.

Conclusions

The extracellular β-galactosidase from *Penicillium sp.* (Psp-β-gal) is an enzyme, belonging to the GH group that hydrolyses β(1-3) and β(1-4) galactosyl bonds in poly- and oligosaccharides. The enzyme shows high transglycosylation activity toward PNPG, lactose and methyl β-D-galactopyranoside turning itself into a promising tool for several enzymatic syntheses. In this work we report the crystal structures of glycosilated Psp-β-gal and its complex with galactose at 1.90 Å and 2.10 Å resolution, respectively. The crystallographic structure comprises a single Psp-β-gal monomer with several covalently N-linked oligosaccharides in the asymmetric unit.

The primary structure of this enzyme is not known and it was inferred on the base of an experimentally determined electron density map. A total of 971 amino acid residues of a single polypeptide chain of this 120 kDa enzyme could be identified. As this is not a common procedure for primary structure identification and is subjected to errors, we used several crystallographic parameters such mean B_{factor}, correlation coefficient between predicted and observed electron density and total exposed surface area per residue, as well as primary structure alignments with known amino acid sequences to improve the quality of assignment and the level of confidence in the identity of each amino acid residue. It was found, after cautiously conducted amino acid sequence assignment, that Psp-β-gal belongs to a family 35 of GHs. This is the first 3D model of a member of this family. Furthermore the sequence similarity with β-galactosidases from *Aspergillus candidus*, *Aspergillus niger*, *Penicillium canescens* and *Talaromyces emersonii* suggest that these proteins, members of the family 35 of GHs, might share significant structural homology with Psp-β-gal.

Also a number of structural features of Psp-β-gal model, including the existence of the TIM barrel domain are similar to both crystallographically studied bacterial β-galactosidases. However, there are several important differences. β-galactosidases from *Penicillium sp.* and *E. coli* have five structural domains each, but their topological arrangement is completely different. In contrast, *T. termophilus* β-galactosidase contains only three distinct domains. In addition, the enzymes differ in their multimeric arrangement. *E. coli* β-galactosidases is tetramer under physiological conditions, recombinant β-galactosidases from *T. termophilus* forms trimers, while Psp-β-gal is a monomer.

Structural comparison of native enzyme with enzyme-galactose complex allowed the identification of Glu160 and Glu259 as the two catalytically important residues. The former was identified as the proton donor and the latter as the nucleophile in the reaction mechanism. The location of Glu160 and Glu259 at the C-terminal of β-4 and β-7 of the TIM barrel allow the classification of Psp-β-gal as a glycoside hydrolase belonging to 4/7 superfamily. The average distance of 4.8 Å between all four pairs of side-chain oxygen atoms of these two catalytic residues implies that hydrolysis occurs via an overall retention of the anomeric configuration of substrate. Besides, the presence of a tyrosine, a phenylalanine and a tryptophan residue forming subsite-1 in Psp-β-gal, A4-β-gal and Ec-β-gal, respectively indicates that an aromatic group, forming a stacking base with the galactosyl moiety, may be important to β-galactosidase activity. These structural observations are experimentally testable by side directed mutagenesis experiments and enzymatic kinetic studies.

One of the factors that may have contributed to the large size of β-galactosidases is that the addition of domains may be associated with the development of a second enzymatic activity.[19] This suggestion cannot be fully confirmed by this study, nevertheless the fact that Psp-β-gal and Ec-β-gal, with five domains each, show glycosylation and transglycosylation activities provides support to this hypothesis. On the other hand, A4-β-gal does not show transglycosylation activity, which might indicate that its three domains are sufficient for substrate cleavage but not for oligosaccharide synthesis and that this activity might be mapped to the other two domains.

Materials and Methods

Protein purification, crystallization and data collection.

Extracellular β-galactosidase from *Penicillium sp.* was purified to homogeneity and crystallized as described.[30] The estimation of molecular mass by SDS-PAGE yielded a value of 120 ± 5 kDa. This value coincides with the mass calculated from the elution profile on a Sephacryl S-300 column assuming that the enzyme exists as a monomer in solution.

Preliminary X-ray diffraction studies revealed that β-galactosidase from *Penicillium sp.* crystallized in the tetragonal space group $P4_1$ or $P4_3$, with unit cell parameters a = b = 110.96, c = 161.05 Å, and the crystals diffracted beyond 2.0 Å of resolution. Calculation of the Matthews coefficient[31] initially suggested the presence of a dimer (V_M = 2.3 $Å^3Da^{-1}$) in the asymmetric unit with a solvent content of approximately 45%. However, a search for non-crystallographic symmetry did not reveal the dimer in the asymmetric unit. Recalculated Matthews coefficient suggested a solvent content of 70% with one monomer in asymmetric unit.

In total, four complete data sets (one native and three derivatives) were collected at the Protein Crystallography beamline[32,33] of the Brazilian National Synchrotron Light Laboratory using an MAR345 image

plate detector. Crystals were immersed in different soaking solutions for different periods of time (Table 3), and mounted in rayon loops. Finally, samples were flash-cooled to 100 K in a cold nitrogen stream. The iodine and the cesium derivatives (Table 3) were prepared according to the quick cryo-soaking approach for derivatization[34,35] while the uranium derivative was obtained by soaking the crystal in UO_2Ac_2 solution overnight. Diffraction images were processed and scaled with the programs DENZO and SCALEPACK.[36] Data collection statistics of each data set are also shown in Table 3.

Table 3. Details of the preparation, data-collection and refinement statistics of β-galactosidase crystals. Statistical values for the highest resolution shells are shown in parentheses.

	Nat-β-gal	**I-β-gal**[3]	**Cs-β-gal**[3]	**U-β-gal**
Crystal preparation				
Cryoprotectant solution	mother liquor 15% ethylene glycol	mother liquor 0.5 M NaI 15% ethylene glycol	mother liquor 0.33 M CsCl 15% ethylene glycol	mother liquor 50 mM UO_2Ac_2 15% ethylene glycol
Soaking time	30 seconds	330 seconds	60 seconds	12 hours
Data collection				
Wavelength (Å)	1.54	1.54	1.54	1.54
Space group	$P4_3$	$P4_3$	$P4_3$	$P4_3$
Unit cell parameters (Å)	a=b=110.96 c=161.05	a=b=110.62 c=159.89	a=b=110.83 c=160.92	a=b=111.01 c=161.33
Resolution (Å)	22.3 - 1.90 (1.93 - 1.90)	23.4 - 2.10 (2.15-2.10)	27.0 - 2.10 (2.15 - 2.10)	22.2 - 2.40 (2.46 - 2.40)
No. of reflections	639065	944162	614561	608673
No. of unique ref.[1]	155025	224036	223388	148232
<I/(I)>	13.9 (3.2)	12.5 (3.1)	8.4 (2.4)	10.0 (2.8)
Multiplicity	4.1 (3.9)	4.2 (4.2)	2.8 (2.7)	4.1 (4.0)
Completeness	99.9 (100.0)	99.9 (100.0)	99.6 (99.6)	98.6 (95.7)
R_{merge}[2]	10.5 (40.0)	10.7 (48.1)	12.3 (47.0)	13.0 (53.1)
Data collected (degrees)	105	240	134	215
Structure refinement				
R-factor (%)	16.5	16.7		
R-free (%)	18.2	18.3		
# protein atoms	7374	7374		
# water molecules	1256	1070		
# sugar atoms	344	356		
# phosphate ions	4	1		
# sodium ions	3	1		
# iodine ions [4]	0	29		
# ethyl. glycol molecules	9	11		
Rmsd bond distances (Å)	0.006	0.006		
Rmsd bond angles (°)	1.4	1.3		
Average B factors (Å^2)				
all atoms	22.23	22.02		
residue atoms	19.85	19.4		
main chain atoms	19.61	19.09		
side chain atoms	20.12	19.78		
water molecules	33.40	35.04		
sugar atoms	31.17	33.18		

[1]Multiplicity of derivative (native) data sets calculated with Friedel-related reflections treated separately (as equivalent).
[2]$R_{merge} = \Sigma_{hkl} |I_{hkl} - \langle I_{hkl} \rangle| / \Sigma_{hkl} I_{hkl}$
[3]Data set collected from a crystal of the protein complexed with galactose.
[4]Iodine sites were not fully occupied.

Phasing procedure

The attempt to determine the phases of β-galactosidase crystals by using the Molecular Replacement (MR) method using the only two known crystal structures of bacterial β-galactosidase did not provide reasonable solutions. Therefore, the phase problem was solved by the SIRAS method, using the native and the first derivative (iodine) data sets. The major heavy-atom positions of the iodine derivative were determined by direct methods with the programs DREAR and SnB.[37,38] This heavy-atom substructure was then set as input into SHARP for phase calculation. Density modification with solvent flattening was performed with the program SOLOMON. At this point, the correct space group was found to be $P4_3$. Heavy atom sites in two new derivatives with cesium and uranium were obtained by difference Fourier maps analyses using the density modified SIRAS derived phases.

The last two derivatives were not used at all for initial model building but they were used with the first

two data sets (native and iodine) for a complete MIRAS phase calculation. Using a total of 74 heavy atoms with varying occupancies (0.8 – 0.1), a complete MIRAS phase calculation has been done using SHARP. This process took approximately seven days in a Pentium IV 1.9 GHz. More than 155.000 reflections were phased with an overall figure of merit of 0.61 in the range of 27.0 – 2.40 Å. Henceforward we used the MIRAS derived phases with a poly-glycine model (obtained after exclusion of the side chains in our initial model) to perform a density modification with SOLOMON.

Initial model building and structure refinement

Solvent flattened I-SIRAS derived electron density map and structure factor amplitudes from the native data set were used by the ARP/wARP program[39] for an automatic build of the β-galactosidase crystal structure. A hybrid model of β-galactosidase was finally obtained after a few thousand cycles of iterative refinement with REFMAC[40] and hundreds of automatic main chain trace. Preliminary inspection of this model revealed a few main chain gaps where cis-Pro residues could be placed. A Fourier map with coefficients $2mF_{obs}$-DF_{calc} and model derived phases enabled the identification of side chains in the Psp-β-gal model. Main chain gaps construction and initial side chain assignment were performed using the program O.[41] Comparison of such obtained primary structure (971 residues) with sequences from databases using the BLAST program revealed no identity. Nevertheless, the results showed that sequence of β-galactosidase from *Penicillium sp*. is similar to those of β-galactosidases from *Aspergillus candidus* (1005aa) and *Aspergillus niger* (1006 aa) and two β-galactosidases fragments from *Penicillium canescens* (171 aa) and *Talaromyces emersonii* (218 aa). None of these four β-galactosidases have their 3-dimensional structure elucidated so far.

After careful identification of the amino acid sequence based on a non-side chain biased electron density map using MIRAS derived phases, the three-dimensional model went through many cycles of refinement, simulated annealing protocols, iterations of energy minimization, water molecules insertion, oligosaccharides addition using the CNS package.[42] The final native model includes one single β-galactosidase molecule with 971 amino acid residues, 7 oligosaccharides chains comprising 16 MAN and 12 NAG, three sodium, four phosphate ions, 9 ethylene glycol molecules and 1256 water molecules in the asymmetric unit.

Analysis of the electron density maps of the iodine and cesium derivatives revealed clear electron density for the galactose bound to the active site of the enzyme, whereas native crystal and uranium derivative did not. The native β-galactosidase in crystals used in iodine and cesium derivatizations was obtained from separate purification procedure, leading to a conclusion that in this case galactose was fortuitously bound to the enzyme in fungal extract and was not washed out in the process of purification.

The structure of the galactose-containing complex was refined against iodine derivative data set using the same protocols as a native structure. The final refinement statistics of the native structure and the structure of the enzyme complexed with galactose is given in Table 3.

Acknowledgments

This work was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) grants #99/03387-4, 98/06218-6 and 01/07014-0, the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) to IP (300220/96-0); and Grant of Presidium of the Russian Academy of Sciences on Basic Research program "Physical-chemistry and biology" and Grant of Russian Foundation Basic Research #03-04-48756 to KNN and AMG.

References

1. Usui, T., Kuboto, S. & Ohi, H. (1983). A convenient synthesis of β-D-galactosyl disaccharide derivatives using the β-D-galactosidase from *Bacillus circulans*. *Carbohydr.Res.* **244**, 315-323.
2. Mori, T., Fujita, S. & Okahata, Y. (1997). Transglycosylation in a two-phase aqueous-organic system with catalysis by a lipid-coated β-D-galactosidase. *Carbohydr. Res.* **298**, 65-73.
3. Fujimoto, H., Miyasato, M., Ito, Y., Sasaki, T. & Ajisaka, K. (1998). Purification and properties of recombinant β-galactosidase from *Bacillus circulans*. *Glycoconj. J.* **15**, 155-160.
4. Hedbys, L., Johansson, E., Mosbach, K., Larsson, P. O., Gunnarsson, A., Svensson, S. & Lonn, H. (1989). Synthesis of Gal beta 1-3GlcNAc and Gal beta 1-3GlcNAc beta-SEt by an enzymatic method comprising the sequential use of beta-galactosidases from bovine testes and Escherichia coli. *Glycoconj. J.* **6**, 161-168.
5. Vetere, A. & Paoletti, S. (1996). High-Yield Synthesis of *N*-Acetyllactosamine by Regioselective Transglycosylation. *Biochem. Biophys. Res. Commun.* **219**, 6-13.
6. Ajisaka, K., Fujimoto, H. & Isomura, M. (1994). Regioselective transglycosylation in the synthesis of oligosaccharides: comparison of β-galactosidases and sialidases of various origins. *Carbohydr. Res.* **259**, 103-115.
7. Cantacuzene, D. & Attal, S. (1991). Enzymic synthesis of galactopyranosyl-L-serine derivatives using galactosidases. *Carbohydr. Res.* **211**, 327-331.
8. Hidaka, M., Fushinobu, S., Ohtsu, N., Motoshima, H., Matsuzawa., Shoun, H. & Wakagi T. (2002). Trimeric crystal structure of the glycoside hydrolase family 42 beta-galactosidase from Thermus thermophilus A4 and the structure of its complex with galactose. *J. Mol. Biol.* ***322***, 79-91.
9. Henrissat, B. (1991). A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem. J.* **280**, 309–316.
10. Henrissat, B. & Bairoch, A. (1993). New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem. J.* **293**, 781–788.
11. Henrissat, B. & Bairoch, A. (1996). Updating the sequence based classification of glycosyl hydrolases. *Biochem. J.* **316**, 695–696.

12. Henrissat, B. & Davies, G (1997). Structural and sequence-based classification of glycoside hydrolases. *Curr. Opin. Struct. Biol.* 7,637-644.
13. Davies, G. & Henrissat, B. (1995). Structures and mechanisms of glycosyl hydrolases. *Struct.* **3**, 853–859.
14. Durand, P., Lehn, P., Callebaut I, Fabrega, S., Henrissat, B. & Mornon, J. P., (1997). Active-site motifs of lysosomal acid hydrolases: invariant features of clan GH-A glycosyl hydrolases deduced from hydrophobic cluster analysis. *Glycobiol.* **7**, 277-284.
15. Jenkins, J., Leggio, L. L., Harris, G. & Pickersgill, R. (1995). β-glucosidase, β-galactosidase, family A cellulases, family F xylanases and two barley glucanases form a superfamily of enzymes with 8-fold β/α architecture and with two conserved glutamates near the carboxy-terminal ends of β-strands four and seven. *FEBS Lett.* **362**, 281-285.
16. Appel, R. D., Bairoch, A. & Hochstrasser, D. F. (1994). A new generation of information retrieval tools for biologists: the example of the ExPASy WWW server. *Trends Biochem. Sci.* **19**, 258-60.
17. Jacobson, R. H., Zhang, X-J., DuBose, R. F. & Matthews, B. W. (1994). Three-dimensional structure of β-galactosidase from *E. coli. Nature* **369**, 761-766.
18. Juers, D. H., Huber, R. E. & Matthews, B. W. (1999). Structural comparisons of TIM barrel proteins suggest functional and evolutionary relationships between β-galactosidase and other glycohydrolases. *Protein Sci.* **8**, 122–136.
19. Juers, D. H., Jacobson, R. H., Wigley, D., Zhang, X. J., Huber, R. E., Tronrud, D. E. & Matthews, B. W. (2000). High resolution refinement of β-galactosidase in a new crystal form reveals multiple metal-binding sites and provides a structural basis for α-complementation. *Protein Sci.* **9**, 1685–1699.
20. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
21. La Fortelle, E. & de Bricogne, G. (1997). Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. *Methods Enzymol.* **276**, 472-494.
22. Abrahams, J. P. & Leslie, A. G. W. (1996). Methods used in the structure determination of bovine mitochondrial F-1 ATPase. *Acta Cryst.* D**52**, 30-42.
23. Sinnott M. L., (1990). Catalytic mechanism of enzymic glycosyl transfer. *Chem. Rev.* **90**, 1171-1202.
24. Koshland, D.E. (1953). Stereochemistry and the mechanism of enzymatic reactions. *Biol. Rev. Camb. Philos. Soc.* **28**, 416-436.
25. McCarter, J. D. & Withers, S.G. (1994). Mechanisms of enzymatic glycosidase hydrolysis. *Curr. Opin. Struct. Biol.* **4**, 885-892.
26. Wormald, M. R. & Raymod, A. D. (1999). Glycoproteins: glycan presentation and protein-fold stability. *Struct.* **7**,155-160.
27. Petrescu, A. J., Petrescu, S. M., Dwek, R.A. & Wormald, M. R. (1999). A statistical analysis of N-and O-glycan linkage conformations from crystallographic data. *Glycobiol.* **9**, 343-352.
28. Davis, J. T., Hirani, S., Barlett, C. & Reid, B. R. (1994). ^{1}H NMR studies on an Ans-linked glycopeptide. GlcNAc-1 C2-N2 bond is rigid in H_2O. *J. Biol. Chem.* **269,** 3331-3338.
29. Imberty, A. & Perez, S., (1995). Stereochemistry of the N-glycosylation sites in glycoproteins. *Prot. Eng.* **8**, 699-709.
30. Neustroev, K. N., de Sousa, E. A., Golubev, A. M., Brandão Neto, J. R., Eneyskaya, E. V., Kulminskaya, A. A. & Polikarpov, I. (2000). Purification, crystallization and preliminary diffraction study of beta-galactosidase from *Penicillium sp. Acta Cryst.* D**56**, 1508-1509.
31. Matthews, B. W. (1968). Solvent content of protein crystals. *J. Mol. Biol.* **33**, 491-497.
32. Polikarpov, I., Oliva, G., Castellano, E. E., Garratt, R., Arruda, P., Leite, A. & Craievich, A. (1997). The protein crystallography beamline at LNLS, the Brazilian National Synchrotron Light Source. *Nucl. Instrum. Methods A* **405**, 159-164.
33. Polikarpov, I., Perles, L. A., de Oliveira, R. T., Oliva, G., Castellano, E. E., Garratt, R. & Craievich, A. (1997). Set-up and experimental parameters of the protein crystallography beamline at the Brazilian National Synchrotron Laboratory. *J. Synchrotron Rad.* **5**, 72-76.
34. Dauter, Z., Dauter, M. & Rajashankar, K. R. (2000). Novel approach to phasing proteins: derivatization by short cryo-soaking with halides. *Acta Cryst.* D**56**, 232-237.
35. Nagem, R. A. P., Dauter, Z. & Polikarpov, I. (2001). Protein crystal structure solution by fast incorporation of negatively and positively charged anomalous scatterers. *Acta Cryst.* D**57**, 996-1002.
36. Otwinowski, Z. & Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307-326.
37. Blessing, R. H. & Smith, G. D. (1999). Difference structure-factor normalization for heavy-atom or anomalous-scattering substructure determinations. *J. Appl. Cryst.* **32**, 664-670.
38. Weeks, C. M. & Miller, R. (1999). The design and implementation of SnB version 2.0. *J. Appl. Cryst.* **32**, 120-124.
39. Perrakis, A., Morris, R. & Lamzin, V. S. (1999). Automated protein model building combined with iterative structure refinement. *Nature Struct. Biol.* **6**, 458-463.
40. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. *Acta Cryst.* D**53**, 240-255.
41. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). Improved methods for building protein models in electron-density maps and the location of errors in these models. *Acta Cryst.* A**47**, 110-119.
42. Brünger A.T, Adams P. D, G., Clore M., DeLano W. L., Gros P., Grosse-Kunstleve R. W., Jiang J.S., Kuszewskic J., Nilges M., Pannu N. S.,Readi R. J., Rice L. C, Simonson T., & Warren G. L., (1998). Crystallography & NMR System: A New Software Suite for Macromolecular Structure Determination. *Acta Cryst.* D**54**, 905 - 921
43. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283-291.
44. Kraulis, P. J. (1991). Molscript – A program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* **24**, 946-950.
45. Esnouf, R. M. (1997). An extensively modified version of Molscript that includes greatly enhanced coloring capabilities. *J. Mol. Graph.* **15**, 133-138.
46. Merritt, E. A. & Bacon, D. J. (1997). Raster3D: Photorealistic molecular graphics. *Methods Enzymol.* **277**, 505-524.
47. Barton G. (1993). ALSCRIPT: a tool to format multiple sequence alignments. *Protein Eng.* **6**, 37-40.

6.5 Subestrutura dos átomos pesados e qualidade dos dados

A determinação da subestrutura dos átomos pesados ou dos espalhadores anômalos em um cristal de proteínas é tão importante para a obtenção da estrutura cristalográfica pelas técnicas SAD, MAD, SIR(AS) e MIR(SA) quanto a própria cristalização da macromolécula biológica. Somente com a localização desses átomos dentro da célula unitária é que se pode obter as fases dos fatores de estrutura necessários para o cálculo dos mapas de densidade eletrônica. Ao longo dos anos, esta tarefa vem sendo aperfeiçoada por diversos pesquisadores teóricos e experimentais e, atualmente, inúmeros softwares de domínio público estão disponíveis.

Apesar de ser uma atividade fundamental à área de cristalografia de proteínas, optou-se por não descrever as idéias, as teorias e os princípios físicos que dão solidez aos métodos atuais de determinação da subestrutura dos átomos pesados, pois a área vem sendo constantemente aperfeiçoada e mais informações podem ser encontradas na literatura. Dois métodos são usados normalmente para a determinação da subestrutura dos átomos pesados: os que utilizam mapas de Patterson (PATTERSON, 1934; STUBBS & HUBER, 2001) e os métodos diretos (SHELDRICK *et al*., 2001). Os mapas de Patterson, principalmente os de diferença isomorfa ou de diferença anômala, foram muito utilizados no início do estabelecimento da área de cristalografia de proteínas em todo o mundo. Apesar de serem, ainda hoje, bastante utilizados em algumas ocasiões, eles perdem espaço para os métodos diretos, porque, ao contrário dos primeiros, os métodos diretos funcionam muito bem para casos em que o sinal isomorfo e/ou anômalo não são tão elevados e porque são muito eficientes quando a subestrutura procurada contém diversos átomos.

Apesar dos grandes avanços já alcançados na área, um fato é certo: a taxa de sucesso na determinação da subestrutura dos átomos pesados está diretamente relacionada com a qualidade dos dados de difração medidos. Essa afirmação é fundamentada não apenas nos resultados dos diversos experimentos mostrados até o momento, mas em uma variedade de outros casos, observados em todo o mundo. Durante o período em que o autor desta tese realizou seu doutorado sanduíche nos Estados Unidos, no grupo do Dr. Zbigniew Dauter, foi realizada uma avaliação da qualidade dos dados de difração e sua importância para a localização da subestrutura dos átomos pesados ou dos espalhadores anômalos, por meio da aplicação dos métodos diretos. Os principais resultados dessa avaliação são apresentados no final desta seção, sob a forma de um artigo publicado recentemente pela revista *Zeitschrift für Kristallographie*.

Z. Kristallogr. **217** (2002) 694-702

Direct way to anomalous scatterers

Z. Dauter*,[I] and R. A. P. Nagem[I,II]

[I] Synchrotron Radiation Research Section, National Cancer Institute, Brookhaven National Laboratory, Bldg. 725A-X9, Upton, NY 11973, USA
[II] CBME, Laboratório Nacional de Luz Síncrotron and Instituto de Física Gleb Wataghin, UNICAMP, Caixa Postal 6192, CEP 13084-971, Campinas, SP, Brazil

Received July 5, 2002; accepted August 29, 2002

Abstract. The first step in solving macromolecular crystal structures by multi- or single-wavelength anomalous diffraction methods is the location of the anomalous scatterers. This can be done by direct methods, using either Bijvoet differences within the single data set, or anomalous scattering amplitudes estimated from measurements at several wavelengths. The calculations suggest that Bijvoet differences are equally successful for this purpose as anomalous amplitudes, F_A, which theoretically should be more suitable. The calculation of F_A values is susceptible to the accumulation of errors contained in the individual intensity measurements at several wavelengths and in the inaccurate estimation of the anomalous atomic scattering corrections, f' and f''. Direct methods often give better results at resolution lower than the full extent of the diffraction data limit. This may be attributed to the enhanced accuracy of measurements of the strong, low resolution reflections and to more effective phase refinement and propagation through P_2 relations.

Introduction

The classic way of solving macromolecular crystal structures, the Multiple Isomorphous Replacement (MIR) method, relies on differences between structure amplitudes measured from the native protein crystal and from derivative crystals, containing a number of heavy atoms, usually incorporated by soaking native crystals in solutions of appropriate reagents (Blundell and Johnson, 1976; Drenth, 1997). The classic MIR procedure assumes that derivatives are isomorphous with the native crystals, which means that the lattice and the scattering contribution of the protein molecules in all crystals are unchanged, and the observed differences result only from the additional scattering of the heavy atoms. In practice, derivative atoms always perturb to some extent the protein structure, and non-isomorphism remains a problem in MIR phasing. As a consequence the lattice and the scattering contribution of the protein atoms in the native and derivative crystals are not unchanged, and the measured intensity differences do not result only from the additional scattering of the heavy atoms.

* Correspondence author (e-mail: dauter@bnl.gov)

The potential of anomalous scattering for phasing protein crystals was realized early (Rossmann, 1961; North, 1965; Matthews, 1966a, 1966b). However, at first the anomalous scattering effect of heavy atoms was used only as an auxiliary source of phase information, since the Bijvoet differences, usually smaller than isomorphous differences, were difficult to estimate accurately. The introduction of crystal freezing, availability of tunable synchrotron beam lines and fast and accurate X-ray detectors enhanced the use of anomalous scattering for phasing macromolecular diffraction data. The current most widely used method of phasing novel structures, the multi-wavelength anomalous diffraction (MAD) approach (Hendrickson, 1991, 1999), relies entirely on the anomalous scattering effect. Since the data are usually collected from the same specimen at various wavelengths, the non-isomorphism is almost totally alleviated in MAD, and may result only from the crystal radiation damage. If diffraction data contain accurately measured anomalous scattering signal, it is possible to obtain a structure solution by single-wavelength anomalous diffraction (SAD) approach (Hendrickson and Teeter, 1981; Wang, 1985; Dauter, Dauter and Dodson, 2002).

The first step in MAD and SAD phasing involves the solution of the partial structure of anomalous scatterers, since knowledge of the calculated anomalous atoms diffraction contribution, F_A, is necessary for estimation of the protein phases. The number of anomalously scattering atoms in the macromolecule is usually small and they are located at mutual distances longer than chemical bonds in a macromolecule. The partial structure of the anomalous scatterers consists therefore of non-overlapping atoms even at resolution of the diffraction data much lower than 1 Å. The condition of "atomicity", required by direct methods, is fulfilled at resolutions practically achieved in macromolecular crystallography. Since first proposed for this purpose (Mukherjee, Helliwell and Main, 1989), direct methods proved to be extremely useful for locating anomalous scatterers in macromolecular crystals.

Anomalous scattering contributions and Bijvoet differences

Figure 1 illustrates the vector diagram for a pair of Friedel-related structure factors, F^+_T and F^-_T, containing a

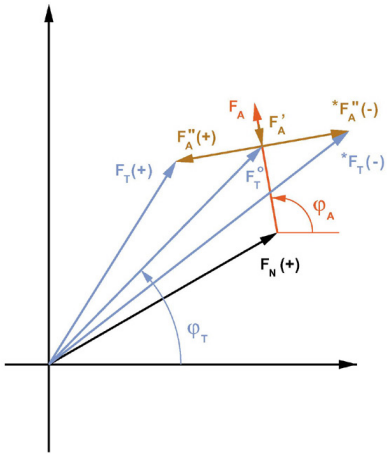


Fig. 1. Vector diagram showing the relations between the normal and anomalous scattering contributions to the entire scattering factor, $F_T = F_N + F_A + F_A' + iF_A''$.

contribution of normally scattering atoms, F_N, and of anomalous scatterers, F_A. It is assumed that all anomalous scatterers are of the same kind; hence, the imaginary contribution of the anomalous scatterers, F''_A, is perpendicular to their normal scattering vector, F_A. Since the normal scattering component of anomalous scatterers is represented by F_A, these values should be used for partial structure solution. However, the amplitudes $|F_A|$ cannot be estimated from the single-wavelength data, because the relation between the experimentally measured Bijvoet difference, $\Delta F^{\pm} = |F^+{}_T| - |F^-{}_T|$, and $|F_A|$, depends on the unknown difference of phases, φ_T and φ_A, according to the relation:

$$\begin{aligned}\Delta F^{\pm} &= 2F''_A \sin(\varphi_T - \varphi_A) \\ &= 2(f''/f^0)\, F_A \sin(\varphi_T - \varphi_A)\,.\end{aligned}$$

Only if Bijvoet differences measured at more than one wavelength are available, as in the MAD experiment, it is possible to estimate F_A, assuming that the atomic scattering factors f^0, f' and f'' are known. Several methods have been proposed to estimate values of F_A from multi-wavelength data.

The classic MAD approach (Karle, 1980; Hendrickson, 1991) provides the value of F_A (as well as F_T and $\varphi_T - \varphi_A$) from the solution of the system of algebraic equations. In the REVISE procedure (Fan, Woolfson and Yao, 1993) the Bijvoet differences at different wavelengths are checked for consistency and corrected, to make the ratio $(|F^+|^2 - |F^-|^2)/f''$ wavelength-independent, and simultaneously the F_A values are optimally evaluated. The recently introduced (Burla, Carrozzini, Cascarano, Giacovazzo, Polidori & Siliqi, 2002) rigorous method of evaluating F_A, based on the joint probability distribution functions, may potentially provide better estimates than the traditional techniques.

In the SAD case, the only available quantities are Bijvoet differences. From the sine relation between $\Delta F^{\pm}$ and F''_A it can be inferred that for reflections with the largest anomalous differences, $(\varphi_T - \varphi_A) = \pm\pi/2$, and the magnitude of Bijvoet difference is proportional to F_A. Those reflections are quite appropriate for use for a direct methods solution of the anomalous substructure, since these methods use only a subset of the largest normalized structure amplitudes.

Diffraction data

Almost all diffraction data presented in Table 1 were collected at the beam line X9B of NSLS (Brookhaven National Laboratory) using the Quantum-4 ADSC CCD detector and were processed with HKL2000 (Otwinowski and Minor, 1997). The only exception is ferredoxin, with data collected at EMBL Hamburg. The following data sets were used previously for structure solution and the results were published: ferredoxin (Dauter et al., 1997), native lysozyme (Dauter et al., 1999), *E. coli* thioesterase (Li et al., 2000), human thioesterase + Br (Devedjiev et al., 2001), and *Pseudomonas* serine-carboxyl proteinase + Br (PSCP, Dauter et al., 2001).

Data included in Table 1 contain various amounts of an anomalous signal from different anomalous scatterers. In some cases, the X-ray wavelength was optimized to maximize the anomalous diffraction signal of elements such as Br, Se, Ta and Lu, which are often used for MAD phasing; in other data sets, the anomalous signal originates from S, Cl, Fe and Mn, at wavelengths far remote from the absorption edges of these elements. In two MAD cases (subtilisin + Lu and *E. coli* thioesterase), the typical four-wavelengths MAD data were collected, whereas for lysozyme soaked in 1M solution of NaBr and glucose isomerase soaked in 1mM solution of $Ta_6Br_{12}{}^{-2}$ cluster data were collected at several wavelengths through the absorption edge of bromine or tantalum. The latter multi-MAD data were acquired especially to estimate the F_A values as accurately as possible and to investigate the variation of the anomalous scattering contributions in the vicinity of the absorption edge. The values of f' and f'' were estimated with XPREP (Bruker Analytical X-ray Systems) for MAD data and are plotted in Fig. 2. In all MAD cases, estimation of the anomalous scattering factors f' and f'' by XPREP gave quite reasonable results. Whereas this procedure gives the absolute values of f'', the f' values cannot be estimated absolutely and only the relative values at different wavelengths are significant. For lysozyme crystal soaked in NaBr, the synchrotron ring injection occurred between collection of the fifth and sixth data sets, and the X-ray wavelength drifted by a small amount, which is evidenced by the perturbation in the curve of the estimated f'' values.

Comparison of the use of $\Delta F^{\pm}$ and F_A

The F_A values can only be estimated approximately, and their accuracy depends on the precision of intensity measurements at various wavelengths, as well as on the stability of wavelength during the whole data collection session. The effect of crystal deterioration may also play an important role. The large value of F_A may result from a very small Bijvoet difference if the F_T and F_A vectors are nearly parallel, and in such cases the estimation may be

Table 1. Diffraction data with anomalous scattering contribution. Values of f' and f'' for MAD data were estimated with XPREP (Bruker Analytical X-ray Systems).

Protein	Size (amino acids)	Resolution (Å)	Anomalous atoms	Wavelength (Å)	f'	f''
MAD data						
Lysozyme + Br	127	1.9	5Br*	0.92357	−0.68	0.58
				0.92042	−2.92	1.23
				0.92028	−3.25	1.89
				0.92021	−3.56	1.99
				0.92014	−3.51	3.18
				0.92007	−3.86	3.15
				0.92001	−2.99	4.80
				0.91987	−2.09	5.20
				0.91674	−0.24	3.98
Gluc. Isom. + Ta_6Br_{12}	388	1.7	30Ta*	1.28065	−1.34	1.05
				1.25549	−6.40	1.75
				1.25511	−8.22	2.44
				1.25486	−9.00	3.82
				1.25460	−6.79	5.77
				1.25435	−3.32	6.05
				1.25410	−0.55	4.92
				1.25384	−0.31	3.95
				1.25360	−0.56	3.48
				1.25336	−0.32	3.24
E. coli Thioesterase	590	2.5	8Se	0.98008	−1.82	1.14
				0.97931	−5.99	3.98
				0.97870	−3.32	6.40
				0.97469	−1.17	5.12
Subtilisin + Lu	275	1.75	4Lu*	1.37758	−1.45	1.32
				1.34120	−4.29	4.87
				1.34080	−4.39	7.54
				1.31280	−1.80	2.99
SAD data						
Human Thioesterase	418	1.8	22Br [a]	0.92	∼ −3.0	∼5.0
PSCP	375	1.8	20Br [a]	0.92	∼ −3.0	∼5.0
Ferredoxin	55	0.94	8Fe	0.88	0.28	1.25
Thermolysin	316	1.45	1Zn + 4Ca	1.28	∼ −6.0	∼4.0
Gluc. Isom. II form	776	1.50	2Mn + 18S	1.54	−0.57 + 0.32	2.80 + 0.56
Gluc. Isom.	388	1.50	1Mn + 9S	1.34	−0.06 + 0.28	2.23 + 0.43
Gluc. Isom.	388	1.60	1Mn + 9S	1.08	0.24 + 0.21	1.54 + 0.28
Gluc. Isom.	388	1.45	1Mn + 9S	0.98	0.29 + 0.18	1.30 + 0.23
Lysozyme	127	1.55	10S + 8Cl[a]	1.54	0.32 + 0.35	0.56 + 0.70
Thaumatin	207	1.7	17S	1.54	0.32	0.56

a: These atoms are partially occupied in the crystal structure.

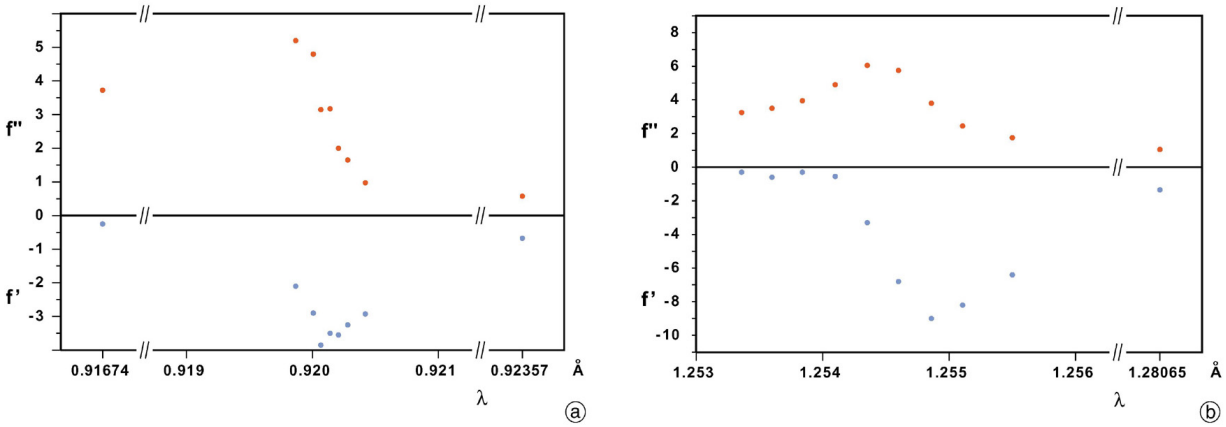


Fig. 2. Values of f' and f'' estimated by XPREP (Bruker Analytical X-ray Systems) for MAD data sets: **a**) lysozyme + Br, **b**) glucose isomerase + Ta_6Br_{12}.

very inaccurate. It may be safer to use Bijvoet differences within a single wavelength data set, rather than F_A set erroneously estimated from multi-wavelength data. The F_A data set may contain centrosymmetric reflections, but those reflections do not occur in the $\Delta F^{\pm}$ set, since Friedel-related centrosymmetric reflections are equivalent.

The MAD data sets, shown in Table 1, collected from crystals containing various anomalous scatterers, were used for solution of anomalous scatterers' positions with program SHELXD (Sheldrick, 1998), comparing two approaches, using either Bijvoet differences, $\Delta F^{\pm}$, within the data set containing the highest f'' contribution (SAD data at peak wavelength) or amplitudes F_A estimated by XPREP from all available data sets collected with various wavelengths (MAD data). One hundred multisolution trials were run, with the same parameters used in both approaches, including the resolution and number of highest normalized amplitudes (about 1500), although the reflections selected automatically were not the same. The SAD and MAD data sets are compared in Figure 3, which shows the fraction of reflections common for both sets in batches (multiples of 50) of the strongest normalized amplitudes. The number of centrosymmetric reflections in the F_A data is also shown.

The number of centrosymmetric reflections in the F_A data obviously depends on the crystal symmetry and it may be expected that in higher symmetry space groups their percentage is larger than in lower symmetry space groups. In the examples shown in Fig. 3, the centrosymmetric reflections amount to 10–20%, up to about 30% among the largest group of reflections for lysozyme in $P4_32_12$ symmetry.

In general, the number of common reflections in the SAD and MAD data is rather low, especially among the largest amplitudes, less than 20%, and not exceeding 50% overall. Nevertheless, a comparison of the SHELXD results, Figure 4 (a–h), shows that the use of SAD and MAD data leads to similar results. The success rate of solutions with a high E_o/E_c correlation coefficient (CC) in both cases is comparably high in various resolution ranges.

Effect of resolution

In contrast to the normal atomic scattering factors, f^0, which diminish with increasing diffraction angle but do not depend on wavelength, the anomalous corrections, f' and f'', do not depend on the diffraction angle but vary with wavelength of X-rays: $f(\theta,\lambda) = f^0(\theta) + f'(\lambda) + if''(\lambda)$. It may therefore be expected that the relative amount of anomalous signal should be larger at a higher resolution. Unfortunately, in macromolecular data this tendency is frustrated by the fact that the high-resolution re-

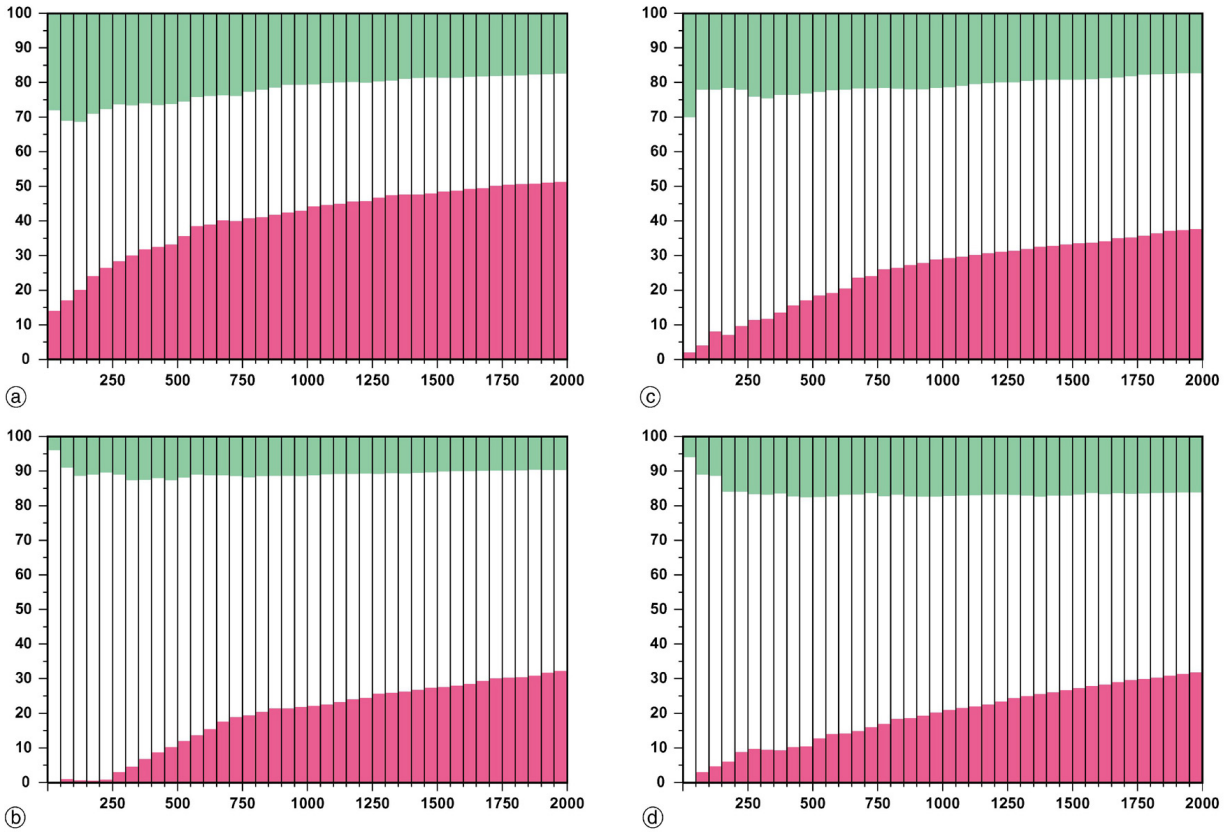


Fig. 3. The percentage of common reflections in the F_A and $\Delta F^{\pm}$ data sets among a number (in batches of multiples of 50) of the largest normalized amplitudes in these sets, in: (**a**) lysozyme + Br, (**b**) glucose isomerase + Ta_6Br_{12}, (**c**) subtilisin + Lu, (**d**) *E. coli* thioesterase. The fraction of centrosymmetric reflections in the F_A data is shown in green at the top of the histogram.

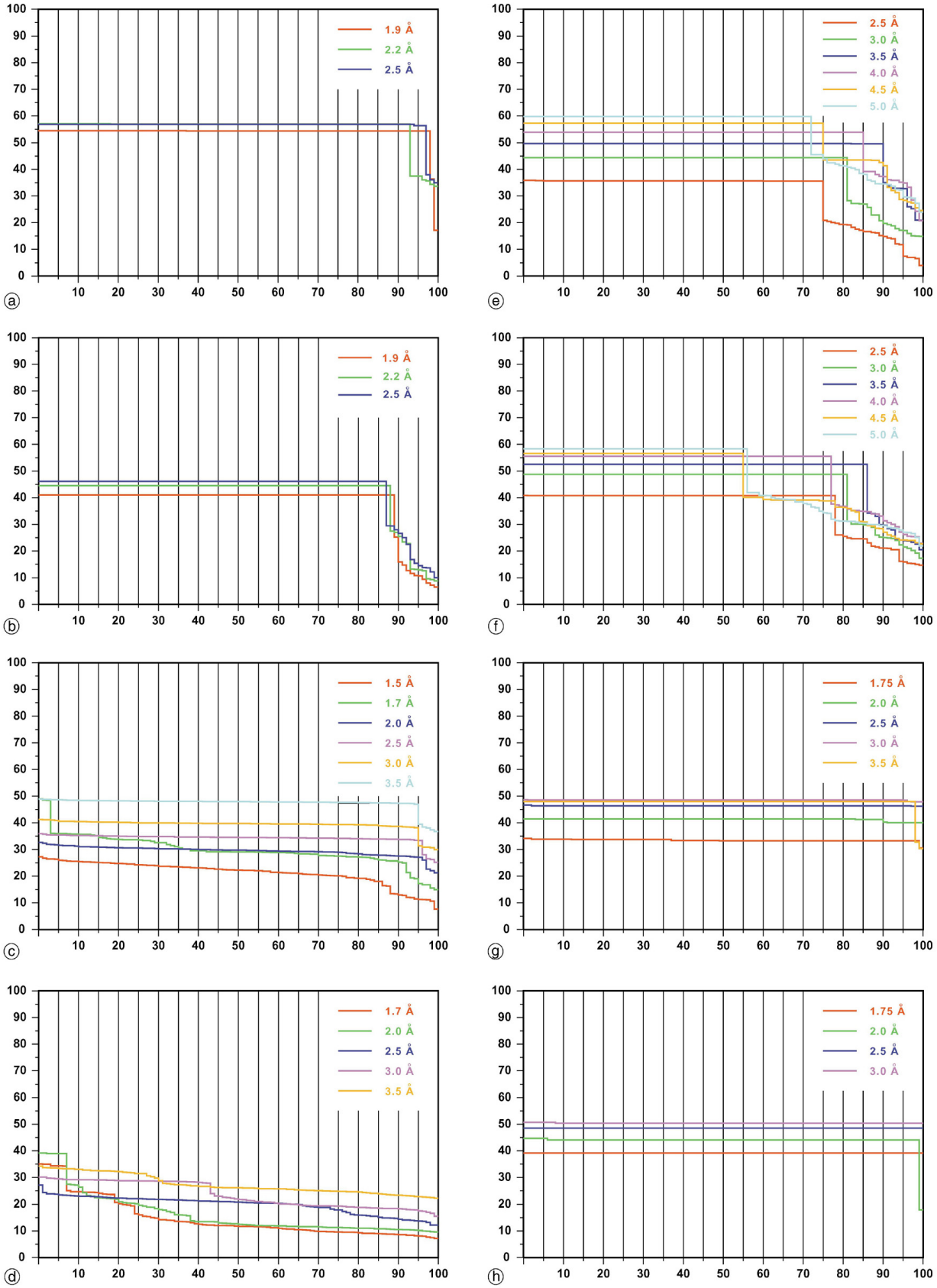

1.9 Å
2.2 Å
2.5 Å
a
2.5 Å
3.0 Å
3.5 Å
4.0 Å
4.5 Å
5.0 Å
e
1.9 Å
2.2 Å
2.5 Å
b
2.5 Å
3.0 Å
3.5 Å
4.0 Å
4.5 Å
5.0 Å
f
1.5 Å
1.7 Å
2.0 Å
2.5 Å
3.0 Å
3.5 Å
c
1.75 Å
2.0 Å
2.5 Å
3.0 Å
3.5 Å
g
1.7 Å
2.0 Å
2.5 Å
3.0 Å
3.5 Å
d
1.75 Å
2.0 Å
2.5 Å
3.0 Å
h

Fig. 4.

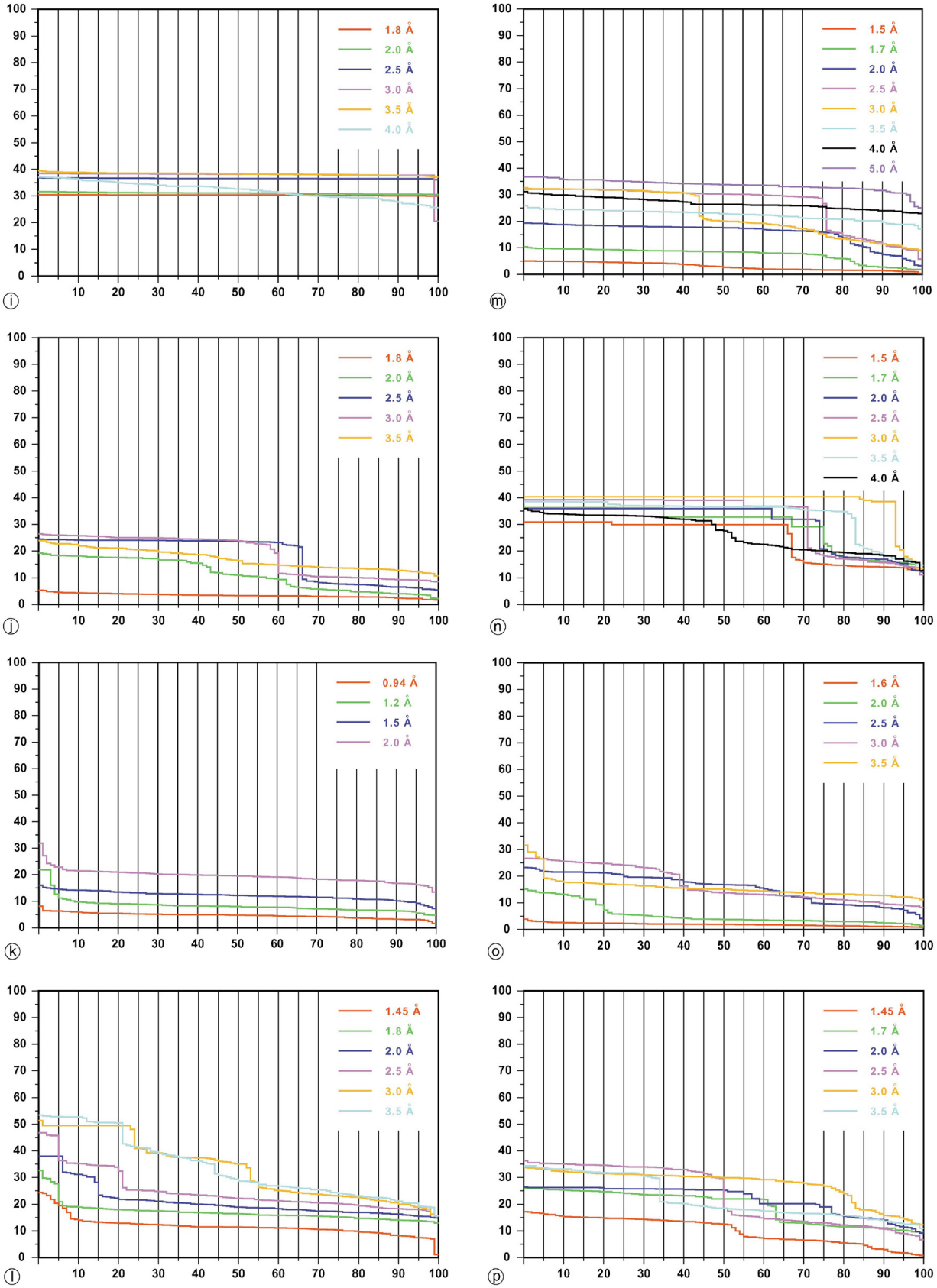
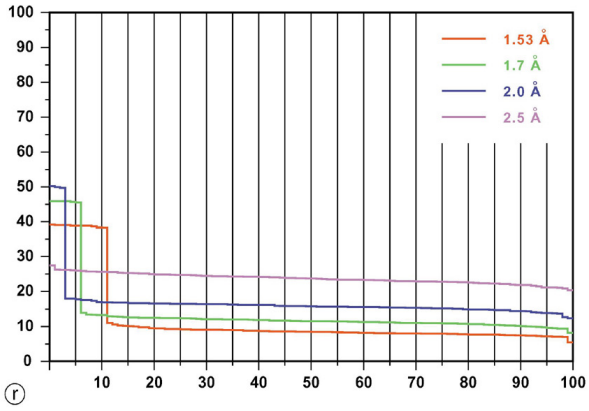

1.8 Å
2.0 Å
2.5 Å
3.0 Å
3.5 Å
4.0 Å
ⓘ
1.5 Å
1.7 Å
2.0 Å
2.5 Å
3.0 Å
3.5 Å
4.0 Å
5.0 Å
ⓜ
1.8 Å
2.0 Å
2.5 Å
3.0 Å
3.5 Å
ⓙ
1.5 Å
1.7 Å
2.0 Å
2.5 Å
3.0 Å
3.5 Å
4.0 Å
ⓝ
0.94 Å
1.2 Å
1.5 Å
2.0 Å
ⓚ
1.6 Å
2.0 Å
2.5 Å
3.0 Å
3.5 Å
ⓞ
1.45 Å
1.8 Å
2.0 Å
2.5 Å
3.0 Å
3.5 Å
ⓛ
1.45 Å
1.7 Å
2.0 Å
2.5 Å
3.0 Å
3.5 Å
ⓟ

Fig. 4.



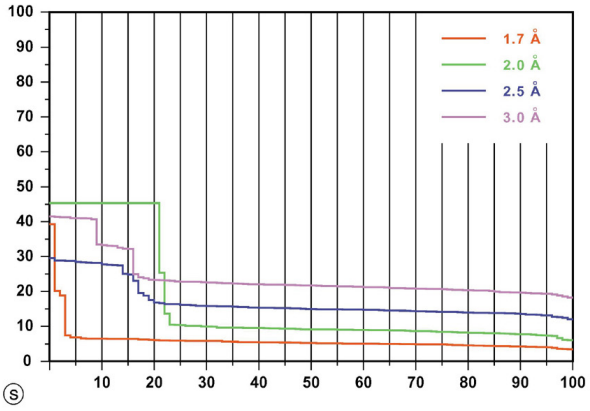


Fig. 4. Results of SHELXD for various data sets, showing the E_o/E_c correlation coefficient for 100 multisolution trials run for each data. Various colours correspond to different data resolution cut-offs. (**a**) lysozyme + Br, F_A data; (**b**) lysozyme + Br, $\Delta F^{\pm}$ data; (**c**) glucose isomerase + Ta_6Br_{12}, F_A data; (**d**) glucose isomerase + Ta_6Br_{12}, $\Delta F^{\pm}$ data; (**e**) *E. coli* thioesterase, F_A data; (**f**) *E. coli* thioesterase, $\Delta F^{\pm}$ data; (**g**) subtilisin + Lu, F_A data; (**h**) subtilisin + Lu, $\Delta F^{\pm}$ data; (**i**) human thioesterase; (**j**) PSCP; (**k**) ferredoxin; (**l**) thermolysin; (**m**) glucose isomerase, II form, $\lambda = 1.54$ Å; (**n**) glucose isomerase, $\lambda = 1.34$ Å; (**o**) glucose isomerase, $\lambda = 1.08$ Å; (**p**) glucose isomerase, $\lambda = 0.98$ Å; (**r**) lysozyme; (**s**) thaumatin.

flections are much weaker and generally measured with less accuracy than the strong, low-resolution data. The anomalous differences usually amount to 1–5% of the total structure amplitude, and may be easily lost in the noise contained in the high resolution, weak intensities.

Another effect, characteristic for direct methods applied to macromolecular data at high resolution, is the sparse distribution of a relatively small subset of strong reflections (*i.e.*, with large F_A or $\Delta F^{\pm}$) in the large reciprocal diffraction sphere, densely populated by weak reflections. As a consequence, the number of Σ_2 interactions between reflection triplets is small at high resolution, particularly in lower symmetry space groups. The only remedy is to increase the number of reflections at the cost of computing time. At low resolution the diffraction sphere is smaller, and the same number of strong reflections interrelate through more Σ_2 triplets. However, at very low resolution there may not be enough strong reflections above the minimal $E = 1.0$ limit.

Inspection of Fig. 4 shows that for most data sets the solution of the partial structure of anomalous scatterers can be obtained at high as well as low resolution, although at low resolution the chance of success is higher. For PSCP, ferredoxin, glucose isomerase at $\lambda = 1.54$ Å and 1.08 Å, attempts at full diffraction data resolution were unsuccessful with CC below 10%, but at resolution lower than 2.0 Å, in all these cases the correct solution was obtained. The higher rate of success at low resolution is particularly visible for data with a small amount of the anomalous signal resulting from weak anomalous scatterers, such as iron in ferredoxin at $\lambda = 0.88$ Å, manganese in glucose isomerase, zinc in thermolysin or sulfur in thaumatin or lysozyme. The apparent lower success for native lysozyme at 2.5 Å than at a higher resolution results from the fact that eight out of ten sulfurs in lysozyme form disulfide bridges with S–S distance of about 2.1 Å; and at a resolution lower than that value, those sulfurs cannot be successfully resolved.

Data redundancy

Practically all diffraction data from macromolecular crystals are collected with the rotation method and two-dimensional detectors. Usually the minimalist approach is applied, with the rotation range selected to give a complete data set, with (almost) each reflection measured at least once. However, the two-dimensionality of the detector and the crystal symmetry lead to most reflections being measured several times. This makes it possible to merge and scale the individual intensities and to estimate their uncertainties. If the redundancy of measurements is low, the accuracy of estimated intensities is poor. The accuracy of estimated intensities can be enhanced by increasing the multiplicity of measurements of individual reflections, including their symmetry equivalents. Since the anomalous scattering signal is usually estimated as a small difference between large amplitudes, the data accuracy is especially important in MAD and SAD.

The effect of data redundancy on the success rate of solving the anomalous substructure was investigated for the native lysozyme. The original data were collected in four passes with various exposure times, each through 180° of total rotation (Dauter et al., 1999). These data were reprocessed in narrower batches of total rotation and the statistics is given in Table 2.

The increased redundancy has a clear effect of the data quality, as evidenced by the increasing $I/\sigma(I)$ ratio. The standard R_{merge} increases with redundancy, which only underlines the known fact that it is a poor data quality criterion (Weiss and Hilgenfeld, 1997; Diederich and Karplus, 1997). Fig. 5 shows the Bijvoet ratio as a function of resolution for these data sets. For the most accurate data from 180° rotation the Bijvoet ratio is actually smallest, but closest to 1.4%, the value theoretically expected from the presence of ten sulfurs and eight chlorides among 1001 atoms of lysozyme.

As evidenced in Table 2, the success rate of SHELXD solutions critically depends on data quality and parallels

Total rotation (°)	45	60	90	135	180
Redundancy[a]	3.4	4.1	6.1	8.3	10.4
Completeness (%)[b]	86.1 (64.5)	94.5 (68.5)	96.6 (69.3)	98.1 (83.5)	100.0 (99.8)
R_{merge} (%)[b]	3.1 (11.4)	3.2 (10.8)	3.4 (12.9)	3.8 (14.3)	4.0 (15.7)
$I/\varrho(I)$	34.8 (5.8)	39.1 (6.5)	47.7 (7.8)	50.2 (7.4)	53.5 (6.8)
Solutions[c]				45	144

a: Redundancy and completeness refer to individual Friedel mates
b: In parentheses are values for the highest resolution shell, 1.56–1.53 Å
c: The number of correct solutions in 1000 SHELXD phasing trials

Table 2. Batches of data for native lysozyme processed with various redundancy.

the redundancy of measurements. When the diffraction data contains only a small amount of the anomalous signal, as in case of the native lysozyme, it is important to enhance the accuracy of intensity estimation by recording data with high redundancy of measurements.

Conclusions

As evidenced in Fig. 4, it was possible to obtain by the direct methods program SHELXD the positions of anomalous scatterers from the anomalous signal contained in all the diffraction data quoted in Table 1, collected from various crystals at different resolutions. The amount of anomalous signal in these data varies from a few percent to about 0.5% for glucose isomerase at $\lambda = 0.98$ Å. For the successful use of very weak anomalous signal it is important to collect highly redundant diffraction data. In all of the cases discussed above, it was also possible not only to locate the anomalous scatterers, but also subsequently to solve the protein model by SAD phasing.

From the comparison of the performance of phasing based on $\Delta F^{\pm}$ and F_A, it seems that both types of data can lead to a partial structure solution with a similar chance of success. From the theoretical point of view, the phasing process based on Bijvoet differences may be expected to be inferior, since the $\Delta F^{\pm}$ data lack all centrosymmetric reflections and their values only approximate the scattering contribution of the anomalous atoms. However, the F_A values, which theoretically are more appropriate for this purpose, may contain additional errors resulting from difficulties in their accurate estimation from the multi-wavelength data with uncertain values of f' and f''. As a result, the large part of the theoretical advantage of using F_A may be lost in practical applications. It may be expected that the full advantage of using F_A values will result from the improved methods of their estimation, such as based on the appropriate probability distributions (Burla et al., 2002).

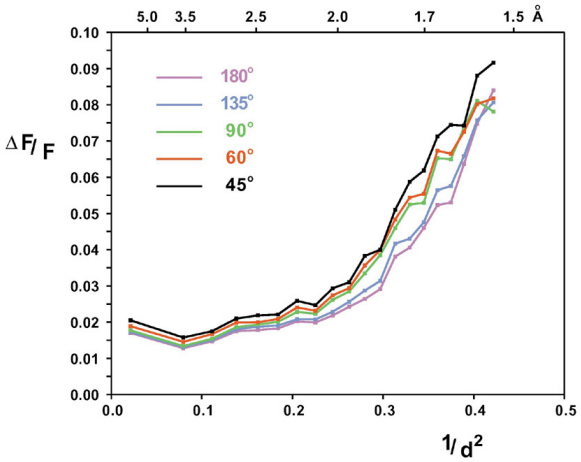


Fig. 5. The Bijvoet ratio $\langle \Delta F^{\pm} \rangle / \langle F \rangle$ as a function of resolution for native lysozyme data collected with various total rotation ranges. The theoretically expected value is 1.4% for ten sulfur and eight chlorine atoms among 1001 atoms of the lysozyme molecule.

The results of the direct methods trials using the same data at various resolutions suggest that it may be more productive to search for anomalous scatterers at low resolution, which is faster and seems to be more successful. This effect may be attributed to the fact that low-resolution intensities are stronger and can be measured with higher accuracy. Moreover, at low resolution the reflections with large normalized amplitudes are interconnected through more Σ_2 relationships, which may lead to more effective phase estimation and refinement.

All diffraction data are available upon request from ZD.

References

Blundell, T. L.; Johnson, L. N.: Protein Crystallography. New York, Academic Press, (1976).

Burla, M. C.; Carrozzini, B.; Cascarano, G. L.; Giacovazzo, C.; Polidori, G.; Siliqi, D.: MAD phasing: probabilistic estimate of $|F_{oa}|$. Acta Crystallogr. **D58** (2002) 928–935.

Dauter, Z.; Dauter, M.; Dodson, E. J.: Jolly SAD. Acta Crystallogr. **D58** (2002) 494–506.

Dauter, Z.; Dauter, M.; de La Fortelle, E.; Bricogne, G.; Sheldrick, G. M.: Can anomalous signal of sulfur become a tool for solving protein crystal structures? J. Mol. Biol. **289** (1999) 83–92.

Dauter, Z.; Wilson, K. S.; Sieker, L. C.; Meyer, J.; Moulis, J. M.: Atomic resolution (0.94 Å) structure of *Clostridium acidurici* ferredoxin. Detailed geometry of [4Fe–4S] clusters in a protein. Biochemistry **36** (1997) 16065–16073.

Dauter, Z.; Li, M.; Wlodawer, A.: Practical experience with the use of halides for phasing macromolecular structures: a powerful tool for structural genomics. Acta Crystallogr. **D57** (2001) 239–249.

Devedjiev, Y.; Dauter, Z.; Kuznetsov, S. R.; Jones, T. L. Z.; Derewenda, Z. S.: Crystal structure of the human acyl protein thioesterase I from a single X-ray data set to 1.5 Å. Structure **8** (2000) 1137–1146.

Diederichs, K.; Karplus, P. A.: Improved R-factor for diffraction data analysis in macromolecular crystallography. Nature Struct. Biol. **4** (1997) 269–275.

Drenth, J.: Principles of protein X-ray crystallography. New York, Springer, (1994).

Fan, H. F.; Woolfson, M. M.; Yao, J. X.: New techniques of applying multi-wavelength anomalous scattering data. Proc. R. Soc. Lond. **A442** (1993) 13–32.

Hendrickson, W. A.: Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. Science **254** (1991) 51–58.

Hendrickson, W. A.: Maturation of MAD phasing for the determination of macromolecular structures. J. Synchrotron Rad. **6** (1999) 845–851.

Hendrickson, W. A.; Teeter, M. M.: Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulfur. Nature **290** (1981) 107–113.

Karle, J.: Some developments in anomalous dispersion for the structural investigation of macromolecular systems in biology. Int. J. Quant. Chem. **7** (1980) 357–367.

Li, J.; Derewenda, U.; Dauter, Z.; Smith, S.; Derewenda, Z. S.: Crystal structure of the *Eschericha coli* thioesterase II, a homolog of the human Nef binding enzyme. Nature Struct. Biol. **7** (2000) 555–559.

Matthews, B. W.: The extension of the isomorphous replacement method to include anomalous scattering measurements. Acta Crystallogr. **20** (1966a) 82–86.

Matthews, B. W.: The determination of the position of anomalously scattering heavy atom groups in protein crystals. Acta Crystallogr. **20** (1966b) 230–239.

Mukherjee, A. K.; Helliwell, J. R.; Main, P.: The use of *MULTAN* to locate the positions of anomalous scatterers. Acta Crystallogr. **A45** (1989) 715–718.

North, A. C. T.: The combination of isomorphous replacement and anomalous scattering data in phase determination of non-centrosymmetric crystals. Acta Crystallogr. **18** (1965) 212–216.

Otwinowski, Z.; Minor, W.: Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol. **276** (1997) 307–326.

Rossmann, M. G.: The position of anomalous scatterers in protein crystals. Acta Crystallogr. **14** (1961) 383–388.

Sheldrick, G. M.: In: *Direct Methods for Solving Macromolecular Structures.* (Fortier S., ed.) Dordrecht, The Netherlands: Kluwer Academic Publishers, pp. 401–411 (1998).

Wang, B. C.: Resolution of phase ambiguity in macromolecular crystallography. Methods Enzymol. **115** (1985) 90–112.

Weiss, M. S.; Hilgenfeld, R.: On the use of merging R factor as a quality indicator for X-ray data. J. Appl. Cryst. **30** (1997) 203–205.

7 Conclusões e perspectivas futuras

A construção de uma estação experimental inteiramente dedicada à Cristalografia de Proteínas no Laboratório Nacional de Luz Síncrotron, bem como a introdução e o estabelecimento da técnica de crio derivatização rápida com halogênios e metais alcalinos fizeram com que o número de estruturas cristalográficas inéditas resolvidas em todo o Brasil aumentasse significativamente em pouco tempo. Até o início deste trabalho, acredita-se que não mais do que duas estruturas cristalográficas de proteínas inéditas haviam sido resolvidas por grupos de pesquisa no Brasil. Em 1995, Oliva e colaboradores publicaram, até onde se sabe, a primeira estrutura de uma proteína inédita no país (OLIVA *et al*., 1995).

Com o uso do método de crio derivatização rápida, a obtenção do modelo tridimensional pode ser alcançada em poucos meses ou mesmo semanas. Portanto, este método tem-se mostrado de crucial importância para o avanço de projetos pós-genoma como Genoma Estrutural e Proteoma. O método de crio derivatização rápida aumentou consideravelmente a capacidade brasileira para a determinação de estruturas de macromoléculas biológicas inéditas uma vez que os experimentos, até então, só podiam ser realizados nos Estados Unidos, na Europa ou no Japão, ou exigiam muito tempo e grandes esforços com o uso de técnicas tradicionais.

Cabe ressaltar que esta técnica está tendo excelente aceitação nos síncrotrons de todo o mundo e dezenas de estruturas já foram resolvidas com o seu uso. Alguns exemplos sobre a aplicação da técnica poderão ser encontrados na próxima edição do livro "*Methods in Enzymology - Volume C: Macromolecular Crystallography*" em um artigo intitulado "*Phasing on rapidly soaked ions*" com a participação do autor desta tese (ver artigo no final desta seção).

Os resultados obtidos ao longo dos quatro anos de pesquisa permitem concluir resumidamente que:

i. O método de crio derivatização rápida pode ser feito tanto com halogênios (anions monovalentes) quanto com metais alcalinos (cátions monovalentes) permitindo a rápida resolução de estruturas cristalográficas inéditas.
ii. Os derivados à base de halogênios (iodo e bromo) têm sítios de ligação complementares aos dos derivados à base de metais alcalinos (césio e rubídio) e, portanto, o uso combinado desses derivados pode ser usado para a obtenção das fases dos fatores de estrutura dos cristais de proteína.
iii. A linha de Cristalografia de Proteínas existente no LNLS deve ser utilizada para a coleta de dados de difração e as técnicas MIRAS/SIRAS são adequadas para a solução das estruturas cristalográficas inéditas.
iv. A preparação de derivados pela técnica de crio derivatização rápida é extremamente barata e não necessita de nenhuma técnica de biologia molecular ou engenharia genética como acontece em derivados de Se-metionina.
v. A técnica de crio derivatização rápida abre novas perspectivas para o uso da nova linha MAD que está sendo construída no LNLS, já que os átomos de bromo e rubídio apresentam bordas de absorção próximas a 1,0 Å; região do espectro que passará a ter um elevado fluxo de fótons.

Além das três estruturas cristalográficas resolvidas ao longo do desenvolvimento desta tese, um grande número de estruturas macromoleculares de uma variedade de organismos com funções diferentes já foi resolvido utilizando o método de crio derivatização rápida. Apesar de não haver uma receita geral para todos os tipos de proteínas, algumas instruções, adquiridas com a prática, podem ser seguidas para se conseguir derivados úteis para a obtenção das fases dos fatores de estrutura.

Normalmente, a solução crio derivatizante é preparada a partir do líquido-mãe. A simples adição de um crioprotetor e um sal apropriado em concentrações elevadas é freqüentemente suficiente para produzir uma boa solução derivatizante. Dependendo da condição de cristalização, uma substituição completa ou parcial dos sais que contêm ânions por sais de bromo ou iodo pode ser feita. De maneira análoga, lítio, sódio e potássio podem ser substituídos por césio e/ou rubídio.

Os resultados obtidos até o momento indicam que o método crio derivatização rápida pode ser usado com um variado número de condições de cristalização. Diferentes tipos de precipitantes já foram usados incluindo PEG de vários tamanhos, sulfato de amônio em várias concentrações e outros. O uso de outros aditivos tais como detergente e azida parecem não afetar de forma drástica o processo de derivatização. Além disso, é apropriado mencionar que, mesmo que o conteúdo da unidade assimétrica só possa ser estimado durante a coleta de dados e tal informação não possa ser usada para verificar a aplicabilidade do método crio derivatização rápida, estruturas tridimensionais de cristais de macromoléculas com um conteúdo de solvente tão baixo quanto 35% já foram resolvidas. Para mais detalhes recomenda-se a leitura do artigo que se encontra no final desta seção.

Como foi amplamente mostrado nesta tese, a preparação dos derivados constitui uma das etapas do processo de obtenção do modelo tridimensional da macromolécula pesquisada. Para que essa tarefa possa ser realizada com êxito, a análise de algumas grandezas como a redundância dos dados, as quantidades $\Delta F^{ANO}/F$, $\Delta F^{ANO}/\sigma\left(\Delta F^{ANO}\right)$, $\Delta F^{ISO}/F$ e $\Delta F^{ISO}/\sigma\left(\Delta F^{ISO}\right)$ e as figuras de mérito, ao longo dessas etapas, permite julgar a qualidade dos resultados obtidos e prever o sucesso ou o insucesso do trabalho.

O método de crio derivatização rápida é, até o momento, a única técnica auxiliar para a resolução de estruturas cristalográficas inéditas de macromoléculas biológicas desenvolvida com a participação direta de pesquisadores brasileiros. Apesar disso, sua aplicação em território nacional ainda é pouco explorada, devido, principalmente, ao limitado número de grupos de cristalografia de proteínas em todo o Brasil. Uma das perspectivas futuras em relação a este trabalho é que a técnica seja difundida e que, por conseqüência, outros grupos de pesquisa no Brasil possam se beneficiar dos resultados obtidos nesta tese. A construção da linha MAD, prevista no LNLS para os próximos anos, irá ampliar ainda mais o potencial da técnica, pois permitirá um aumento significativo do fluxo de raios-X disponível e a possibilidade de regular o sinal anômalo através do ajuste do comprimento de onda da radiação síncrotron para as bordas de absorção dos átomos de bromo e rubídio (aproximadamente 1,0 Å).

Embora esta técnica pareça perfeitamente adequada para a solução de estruturas de proteínas com o uso de um anodo rotatório (gerador convencional de raios-X), nenhum artigo científico foi publicado até o momento relatando esta aplicação. O uso combinado de geradores convencionais de raios-X com a técnica de crio derivatização rápida poderia aumentar enormemente a eficiência dos projetos pós-genoma, incluindo aqueles em desenvolvimento no país, pois permitiria a resolução de estruturas cristalográficas inéditas, em laboratórios de pequeno porte, ou mesmo validar derivados

potenciais antes de uma coleta de dados definitiva em um síncrotron. Esta abordagem seria ainda mais eficiente caso a estrutura cristalográfica das macromoléculas pudesse ser obtida com a coleta de um único conjunto de dados pela técnica SAD.

Esta última perspectiva não deve ser considerada apenas como uma possibilidade remota e difícil de ser alcançada. Experimentos recentes realizados pelo autor desta tese no final de seu projeto permitiram com que a estrutura de uma proteína de aproximadamente 55 kDa fosse resolvida por SAD com os dados de difração coletados em um anodo rotatório pertencente ao Grupo de Cristalografia de Proteínas do Instituto de Física de São Carlos da Universidade de São Paulo. Os resultados desse trabalho não foram apresentados nesta tese, mas serão usados, no momento oportuno, para elaboração de um artigo científico.

Phasing on Rapidly Soaked Ions

By RONALDO A. P. NAGEM, IGOR POLIKARPOV AND ZBIGNIEW DAUTER

Introduction

There are three basic ways of solving macromolecular crystal structures: the molecular replacement method, the direct methods and the heavy-atom method. Molecular replacement involves the use of a known search model, closely similar to the investigated macromolecule. The direct methods are routinely used to solve the structures of small molecules, where the diffraction data extend to atomic resolution. As a consequence, these two methods cannot be used to solve the novel structures with crystals diffracting to a resolution lower than atomic. The most general method of solving novel crystals structures is therefore a heavy atom approach in its various modifications. In general, in this approach the initial phases are derived from the differences in crystal scattering caused by the presence of a small number of heavy and/or anomalously diffracting atoms, which can be inherently present in the native macromolecule or introduced by chemical derivatization. Various types of the diffraction differences can be utilized. In the Single or Multiple Isomorphous Replacement (SIR or MIR) methods only the differences between the intensities of the native and derivative crystals are used. If the anomalous differences are used in addition, the method becomes the Isomorphous Replacement with Anomalous Scattering (SIRAS or MIRAS). If only the differences between Friedel-related intensities at one or more X-ray wavelengths are used, the technique is termed Single- or Multiple-wavelength Anomalous Dispersion (SAD or MAD) respectively. The details of each of these approaches are comprehensively discussed in many classic textbooks[1,2].

The heavy atoms, providing the initial phase information, can be present in the original macromolecule, such as certain transition metals in metalloproteins or, more recently proposed as general tools, sulfur in proteins[3,4] and phosphorus in nucleic acids[5]. However, most general is the derivatization before or after the crystallization. The example of the former is the production of the selenomethionine protein variants[6] by genetic engineering for MAD technique. The classic derivatization approach involves prolonged soaking of the native crystals in the diluted solutions of various heavy metal salts and coordination compounds. Such soaking procedures are time consuming and often unsuccessful due to the lack of heavy atom binding or the deterioration of the crystal quality.

It has been recently proposed that certain simple anions or cations suitable for phasing, such as halides or alkali metals, can be introduced into protein crystals by rapid soaks in the appropriate cryo derivatization solutions[7,8] ("quick cryo soaking approach"). This procedure combines in one rapid single step the derivatization and the cryogenic protection. Immediately before freezing the crystal for data collection, a native crystal is immersed for a short period of time in a cryo-protectant solution drop containing in addition a high concentration of the appropriate salt. Comparing to classic soaks that combine low heavy metal concentration and long immersion times, this procedure is able to generate very good isomorphous derivatives significantly faster. This approach is based on somewhat different chemical behavior of halides and alkaline ions in comparison with the classic heavy atom compounds.

Protein crystals contain a significant proportion of the liquid solvent phase, filling the voids between the more or less globular protein molecules. Various small chemical compounds can diffuse through the solvent channels within protein crystals, and this has often been used to obtain *e.g.* enzyme complexes with inhibitors, cofactors etc. This diffusion is quick, which is evidenced by the frequent presence of the glycerol molecules bound to the protein surface after a very short, two to five second immersion of the crystal in the cryoprotecting solution containing glycerol[9].

The rapid soak approach uses this property of protein crystals, which allows small ions to diffuse within a short time to the solvent regions surrounding the protein molecules and adopt ordered sites at their surface.

Ions used for rapid soaks

Both negatively charged heavy halides and positively charged heavy alkali ions have been proposed for this fast derivatization approach.

The presence of chloride anions has been observed in the structures of proteins crystallized from solutions containing a significant concentration of sodium chloride, *e.g.* in tetragonal lysozyme[3,10]. When lysozyme was

[1] Blundell, T.L. & Johnson, L.N. (1976). *Protein Crystallography.* Academic Press, New York.

[2] Drenth, J. (1999). *Principles of protein X-ray crystallography*, 2nd ed. Springer, Heidelberg.

[3] Dauter, Z, Dauter, M., de La Fortelle, E., Bricogne, G. & Sheldrick, G.M. (1999). *J. Mol. Biol.*, **289**, 83-92.

[4] Micossi, E., Hunter, W.N. & Leonard, G.A. (2002). *Acta Crystallogr.*, **D58**, 21-28.

[5] Dauter, Z. & Adamiak, D.A. (2001). *Acta Crystallogr.*, **D57**, 990-995.

[6] Hendrickson, W.A. & Ogata, C.M. (1997). *Meth. Enzym.*, **276**, 494-523.

[7] Dauter, Z., Dauter, M. & Rajashankar, K.R. (2000). *Acta Crystallogr.*, **D56**, 232-237.

[8] Nagem, R.A.P., Dauter, Z. and Polikarpov, I. (2001). *Acta Crystallogr.*, **D57**, 996-1002

[9] Lubkowski, J., Dauter, Z., Yang, F., Alexandratos, J., et al., & Wlodawer, A. (1999). *Biochemistry*, **38**, 13512-13522.

[10] Blake, C.C.F., Mair, G.A., North, A.C.T., Phillips, D.C. & Sarma, V.R. (1967). *Proc. Roy. Soc.*, **B167**, 365-377.

crystallized from the solution containing NaBr[11,12] or NaI[13], a number of sites occupied by these halides appeared at the protein surface. This observation and the analysis of data collected on a few test crystals led to the proposal of using soaked bromides and iodides for phasing[7].

The two heavier halides, bromine and iodine, display a significant anomalous signal in the range of wavelength easily accessible at most of the synchrotron beam lines. Bromine has the K absorption edge at 0.92 Å (13474 eV) and is appropriate for phasing through MAD method. Bromine has one more electron than selenium and it has been used for MAD solution of the oligonucleotide structures after substituting thymine by the almost isostructural bromouracil[14]. It can be considered as the nucleic acid equivalent of the SeMet substitution in protein crystallography[6].

The iodine absorption edges (K at 0.37 Å and L_I at 2.39 Å) are not easily accessible and iodine is therefore not suitable for the MAD work. However, it retains a significant anomalous signal (f' = 6.8 electron units) at the copper characteristic wavelength of 1.54 Å. Iodine has been used as a heavy atom in protein crystallography after chemical modification of the tyrosine aromatic rings[15,16].

Chlorine, the halide lighter than bromine or iodine, has its K edge at a very long wavelength (4.39 Å), and displays only a small anomalous effect at more accessible wavelength (f" = 0.70 at 1.54 Å and f" = 0.88 at 1.74 Å). Nevertheless, with the accurately measured data it is possible to use its anomalous signal for phasing[3,17,18].

The use of heavy alkali metals for rapid cryosoaks was proposed as an extension to the quick cryo soaking approach with halides[8]. The pair of heavier alkali metals, rubidium and caesium, have two electrons more than the pair of halides, bromine and iodine, respectively. This difference, extremely important from the chemical point of view, allows Cs or Rb cations to occupy different positions in the crystal structure when compared to the positions occupied by I or Br anions. This fact adds an additional flexibility to the procedure of quick cryo soaking and permits combined use of such derivatives in MIR(AS) phasing, even when none of them are strong enough to produce interpretable electron-density maps individually.

The K absorption edge of rubidium (0.82 Å) is in the same range as Br and Se K edges, which makes it a suitable atom for MAD phasing. Indeed, it was recently tested as a useful MAD phasing source[19].

On the other hand caesium atoms, even though possessing a strong anomalous signal with f" = 7.90 electrons at 1.54 Å wavelength, are not suitable for MAD experiments. The Cs absorption edges (K at 0.34 Å and L_I at 2.17 Å) are far away from the wavelength range accessible at most synchrotron beam lines. The anomalous signal of caesium has been used for phasing in the past, *e.g.* for gramicidin[20].

Differences between classic reagents and ions used for quick soaks

The ions used for quick soak approach differ in their chemical properties from the classic heavy atom reagents. Halides in water solution occur as not coordinated, monoatomic anions, although they interact with water through hydrogen bonds. Alkali cations are coordinated by water molecules, but the coordination is not very strong, and the metal aquo ligands can be easily exchanged e.g. by the carboxyl or carbonyl oxygen atoms. In contrast, in many standard heavy atom reagents the ligands are strongly coordinated or covalently bound to the metal[21]. To bind to the appropriate protein sites such reagents have to undergo a partial hydrolysis or a similar chemical substitution. They usually form strong complexes or bind covalently to certain chemical functions of the protein, such as *e.g.* mercury reagents with the cysteine sulfhydryl groups. If present in higher concentration, these reagents often disrupt the protein intermolecular interactions, damaging the crystalline order and adversely influencing the crystal diffraction. The usual procedures involve long (several hours or days) soaks at low, milimolar concentration of the appropriate reagent, allowing the chemical reactions to proceed slowly. Bromide and iodide anions are soft, monoatomic and polarizable. They are attracted to the protein surface through various types of relatively weak, noncovalent interactions. Due to their negative charge they can form ion pairs with the positively charged arginine and lysine side chain functions. Secondly, they can accept hydrogen bonds from various proton donors, such as protein amides (in the main or side chains) and hydroxyls, as well as solvent water molecules.

11 Lim, K., Nadarajah, A., Forsythe, E.L. & Pusey, M.L. (1998). *Acta Crystallogr.*, **D53**, 240-255.

12 Dauter, Z. & Dauter, M. (1999). *J. Mol. Biol.*, **289**, 93-101.

13 Steinrauf, L.K. (1998). *Acta Crystallogr.*, **D54**, 767-779.

14 Smith, J.L. & Hendrickson, W.A. (2001). In *International Tables for Crystallography* (M.G. Rossmann & E. Arnold, eds.), Vol. F, Chapter 14.2.1., pp. 299-303.

15 Chen, L.Q., Rose, J.P., Breslow, E., Yang, D., et al., & Wang, B.C. (1991). *Proc. Natl. Acad. Sci. USA*, **88**, 4240-4244.

16 Brady, L., Brzozowski, A.M., Derewenda, Z.S., Dodson, E., et al., & Menge, U. (1990). *Nature*, **343**, 767-770.

17 Lehmann, C. (2000). Ph.D. Thesis, Göttingen, Germany.

18 Loll, P.J. (2001). *Acta Crystallogr.*, **D57**, 977-980.

19 Korolev, S., Dementieva, I., Sanishvili, R., et al., & Joachimiak, A. (2001). *Acta Crystallogr.*, **D57**, 1008-1012.

20 Wallace, B.A., Hendrickson, W.A. & Ravikumar, K. (1990). *Acta Crystallogr.*, **B46**, 440-446.

21 Carvin, D., Islam, S.A., Sternberg, M.J.E. & Blundell, T.E. (2001). In *International Tables for Crystallography* (M.G. Rossmann & E. Arnold, eds.), Vol. F, Chapter 12.1, pp. 247-255.

TABLE I

ANOMALOUS SCATTERERS SITES IN SOME CRYSTAL STRUCTURES DESCRIBED IN TABLE 2

TEN STRONGEST SITES ARE LISTED WITH THEIR CORRESPONDING PEAK HEIGHTS GIVEN IN σ AND NORMALIZED TO THE HIGHEST PEAK IN THE ANOMALOUS DIFFERENCE FOURIER MAP

β-galactosidase (iodine)							**β-galactosidase (caesium)**						
					Peak height							*Peak height*	
Site		*Fractional coordinates*			*Sigma*	*Norm.*	*Site*		*Fractional coordinates*			*Sigma*	*Norm.*
I	1	0.0405	0.3618	0.7993	31.31	1.00	Cs	1	0.7358	0.5808	0.4773	31.06	1.00
I	2	0.1530	0.1618	0.8228	30.81	0.98	Cs	2	0.8417	0.5499	0.4459	25.31	0.81
I	3	0.4482	0.1546	0.7447	19.90	0.64	Cs	3	0.9259	0.5141	0.7811	23.20	0.75
I	4	0.4558	0.0530	0.8035	19.21	0.61	Cs	4	0.4873	0.5272	0.6456	20.40	0.66
I	5	0.1338	0.4844	0.9018	14.92	0.48	Cs	5	0.7780	0.3222	0.7513	9.67	0.31
I	6	0.0706	0.4947	0.6388	14.73	0.47	Cs	6	0.9676	0.3061	0.7619	7.95	0.26
I	7	0.3282	0.3609	0.9262	14.24	0.45	Cs	7	0.5367	0.1733	0.4704	7.88	0.25
I	8	0.4323	0.4568	0.9342	13.94	0.45	Cs	8	0.8344	0.5700	0.6374	6.53	0.21
I	9	0.4217	0.5679	0.9902	13.57	0.43	Cs	9	0.5718	0.3767	0.6471	5.81	0.19
I	10	0.5587	0.0003	0.9921	12.77	0.41	Cs	10	0.0482	0.2172	0.6737	5.39	0.17

Acyl protein thioesterase I (bromine)							**Trypsin inhibitor (caesium)**						
					Peak height							*Peak height*	
Site		*Fractional coordinates*			*Sigma*	*Norm.*	*Site*		*Fractional coordinates*			*Sigma*	*Norm.*
Br	1	0.0724	0.2361	0.6181	41.92	1.00	Cs	1	0.4544	0.1594	0.3748	25.22	1.00
Br	2	0.3298	0.1346	0.8744	41.50	0.99	Cs	2	0.6969	0.1170	0.2326	21.74	0.86
Br	3	0.9559	0.9969	0.1666	30.58	0.73	Cs	3	0.7012	0.2988	0.2500	20.00	0.79
Br	4	0.7732	0.3760	0.0022	24.19	0.58	Cs	4	0.5693	0.4503	0.3392	18.00	0.71
Br	5	0.1386	0.1800	0.6227	19.59	0.47	Cs	5	0.1248	0.5987	0.3385	17.46	0.69
Br	6	0.8183	0.0550	0.3299	19.17	0.46	Cs	6	0.4357	0.5200	0.3344	15.53	0.62
Br	7	0.6100	0.3168	0.1365	18.81	0.45	Cs	7	0.5996	0.3860	0.1891	12.87	0.51
Br	8	0.3252	0.1924	0.8098	18.35	0.44	Cs	8	0.0984	0.4384	0.2413	10.35	0.41
Br	9	0.8060	0.3409	0.4412	17.33	0.41	Cs	9	0.6120	0.1409	0.1027	10.15	0.40
Br	10	0.4539	0.1646	0.3777	16.59	0.40	Cs	10	0.3589	0.5542	0.2645	7.56	0.30

TABLE II
EXAMPLES OF CRYSTAL STRUCTURES SOLVED BY CRYO SOAKING

Protein	Size (a.u) kDa	Solvent cont. (%)	Cryoderivatization conditions	Soak time (s)	Phasing Method	# of sites	Resol. (Å)	Ref.
Insulin-like growth factor-1 from *Homo sapiens*	1 x 6	55	25 % (w/v) PEG 3350, 30 % MPD, 0.2 M sodium cacodylate pH 6.5 2.8 mM deoxy big CHAPS, 10 M NaBr	30	MAD	1 Br + 6 Cys Sγ	2.00 (1.80)	22
β-defensin-2 from *Homo sapiens*	4 x 4	40	36 % PEG 4000, 0.32 M lithium sulfate, 0.16 M MOPS pH 7.1 10 % glycerol, 0.25 M KBr (0.25 M KI)	60	MIRAS	9 Br (9 I)	2.00 (1.40)	23
C-terminal domain of TonB from *Escherichia coli*	2 x 8	35	28-30 % PEG 3350, 0.1 M Tris pH 7.5 50-100 mM calcium chloride, 1.0 M KBr	50	MAD	4 Br	2.50 (1.55)	24
Peroxiredoxin 5 from *Homo sapiens*	1 x 17	65	1.6 M ammonium sulfate, 0.1 M sodium citrate pH 5.3, 0.2 M potassium sodium tartrate, 1 mM 1,4-dithio-dl-threitol, 0.02 % (w/v) azide, 20 % (v/v) glycerol, 1.0 M NaBr	30	MAD	5 Br	1.90 (1.50)	25
Trypsin inhibitor from *Copaifera langsdorffi*	1 x 18	45	20-25 % PEG 8000, 0.1 M sodium acetate pH 4.5, 20 % ethylene glycol, 1.0 M CsCl	300	SIRAS	5 Cs	2.00 (2.00)	8
Interleukin-22 from *Homo sapiens*	2 x 17	33	0.9 M sodium tartrate, 0.1 M Hepes pH 7.5, TRITON X-100 detergent, 15 % ethylene glycol, 0.125 M NaI	180	SIRAS	10 I	1.92 (1.92)	26
GAF domain YKG9 from *Saccharomyces cerevisiae*	2 x 18	65	2.5 M ammonium sulfate, 0.05 M lithium sulfate, 30% sucrose, 0.1 M Tris-HCl pH 8.0, 10 % (v/v) glycerol, 0.5 M NaBr	45	MAD	7 Br	2.80 (1.90)	27
Carboxyl proteinase from *Pseudomonas sp.* 101	1 x 41	56	1.0 M ammonium sulfate, 0.005 M guanidine, 0.1 sodium citrate pH 3.3, 18 % glycerol, 1.0 M NaBr	30	SAD	9 Br	1.80 (1.40)	28
α-galactosidase from *Trichoderma reesei*	1 x 47	47	15 % PEG 3350, 100 mM potassium phosphate 10 % glycerol, 0.26 M CsCl	480	SIRAS	10 Cs	1.60 (1.60)	29
Acyl protein thioesterase I from *Homo sapiens*	2 x 25	38	42 % saturation ammonium sulfate, 0.1 M sodium acetate pH 5.0 20 % (v/v) glycerol, 1.0 M NaBr	20	SAD	22 Br	1.80 (1.50)	30
Thiamin pyrophosphokinase from *Saccharomyces cerevisiae*	2 x 35	49	25 % PEG MME 2000, 0.1 M ammonium sulfate, 0.1 M sodium acetate pH 5.1, 50 mM sodium chloride, 1.0 M NaBr	45	MAD	12 Br	2.00 (1.80)	31
Cysteine-rich domain of Sfrp-3 from *Mus musculus*	6 x 14	45	0.1 M HEPES pH 6.6, 33 % PEG 3350 0.5 M NaBr	40	MAD	9 Br	1.90 (1.90)	32
SptP:SicP complex (2:4) from *Salmonella Typhimurium*	2 x 18 4 x 12	58	5-10 % PEG 6000, 15 % glycerol 2.0 M NaBr	30	MAD	31 Br	2.50 (1.90)	33
β-galactosidase from *Penicillium sp.*	1 x 110	72	15 % PEG 8000, 50 mM sodium phosphate pH 4.0 30 % ethylene glycol, 0.25 M NaI (CsCl)	180 (300)	SIRAS	13 I (12 Cs)	1.95 (1.85) 2.04 (1.85)	34

Thirdly, they can interact with the protein hydrophobic surfaces. All these interactions are observed in the protein crystals containing halide sites. In respect of chemical interactions with proteins, halides are not highly specific. The halide bonding interactions do not require any slow chemical reaction to take place and can be formed very quickly.

The cations are more specific in the character of their binding. Alkali ions have a preference for oxygen functions such as carboxyl (negatively charged), carbonyls or water. Their sites are in the vicinity of a few side chain carboxyls or main and side chain carbonyls available at the protein surface and usually have a number of water ligands. Rubidium and caesium ions are not strongly demanding toward the coordination geometry and can have five to eight ligands[19]. Because they do not coordinate water molecules very strongly, the substitution of water ligands by the protein oxygens takes place rapidly.

As stated above, binding of halides and alkali ions is not very strong and their sites around the protein surface are partially occupied, even if their concentration in the mother liquor is higher than 1 M. All these ions can share the sites with water molecules, and their occupancy results from the competition between them and water in binding to various protein functions with variable strength in the state of equilibrium. It is difficult to estimate accurately the absolute occupancy factors of these sites. The relative occupancies of the strongest anomalous sites in a few example structures are given in Table 1. Figure 1 illustrates various, most typical sites and coordination of soaked ions from the crystal structures solved by their use.

Procedure

A number of macromolecular structures from a variety of organisms with different functions have been solved using the quick cryo soaking approach with halides and alkaline metals. In Table 2 we show several derivatization aspects of some of these structures, including size of protein, soaking time, cryo derivatization conditions etc. It is easy to see from Table 2 that there is not a general recipe for all types of proteins. However, a few instructive steps, acquired with practice, could be followed to enhance the chance of obtaining suitable derivatives for phasing.

Normally, the cryo derivatization solution is prepared from the original mother liquor. The simple addition of a cryoprotectant and an appropriate salt in high concentration is often enough to produce a good cryoderivatization solution. Depending on the crystallization condition, a complete or partial substitution of salts containing various anions by bromides or iodides can be also performed. As in the SptP:SicP complex structure determination[33], the sodium chloride used during crystallization was completely replaced by sodium bromide during derivatization. Analogously, lithium, sodium and potassium can be replaced by caesium or rubidium. In cases with a saturated crystallization solution a partial substitution of reagents is likely to work.

The results indicate that the quick cryo soaking approach can be used with success in a number of adverse crystallization conditions. Different types of precipitants have been used so far, including several PEG sizes, ammonium sulfate in various concentration and others. The use of other additives as detergent[26], azide[25], sucrose[27] and even ammonium sulfate in high concentration[30] seems to not drastically affect the derivatization.

The choice between ethylene glycol and glycerol for cryogenic protection is in principle not too much relevant for the derivatization process itself. It should rather be chosen for each solution in order to provide complete cryogenic protection.

Another decision that must be taken before preparation of the derivative is the choice of the correct salt. If a MAD data collection can be performed the use of bromide and rubidium salts like NaBr, KBr or RbCl are recommended. The pH of the mother liquor may suggest the most appropriate salt. In principle, in low pH the protein molecules are positively charged which therefore makes halides a better option. At high pH the alkaline metals may be recommended. These recommendations are more based on the chemical intuition and more extensive studies with different pH values are required for a final conclusion. On the other hand, if MAD approach is not applicable, the use of iodide (LiI, NaI or KI) or caesium (CsCl) salts combined with a longer X-ray wavelength is advised.

It seems appropriate to mention that even though the content of the asymmetric unit can only be estimated during data collection, and such information cannot be used to verify the applicability of the quick cryo

[22] Vajdos, F.F., Ultsch, M., Schaffer, M.L., Deshayes, et al., & de Vos, A.M. (2001). *Biochemistry* **40**, 11022-11029.

[23] Hoover, D.M., Rajashankar, K.R., Blumenthal, R., et al., & Lubkowski, J. (2000). *J. Biol. Chem.*, **275**, 32911-32918.

[24] Chang, C., Mooser, A., Plückthun, A. & Wlodawer, A. (2001). *J. Biol. Chem.*, **276**, 27535-27540.

[25] Declercq, J.-P., Evrard, C., Clippe, A., Stricht, D.V., Bernard, A. & Knoops, B. (2001). *J. Mol. Biol.*, **311**, 751-759.

[26] Nagem, R.A.P., Colau, D., Dumoutier, L., Renauld, J.-C., et al., & Polikarpov, I. (2002). *Structure*, **10**, 1051-1062.

[27] Ho, Y.-S.J., Burden, L.M. & Hurley, J.H. (2000). *EMBO J.*, **19**, 5288-5299.

[28] Wlodawer, A., Li, M., Dauter, Z., Gustchina, A., et al., & Oda, K. (2001). *Nature Struct. Biol.*, **8**, 442-446.

[29] Golubev, A.M. & Polikarpov, I. *In preparation.*

[30] Devedjiev, Y., Dauter, Z., Kuznetsov, S.R., Jones, T.L.Z. & Derewenda, Z.S. (2000). *Structure*, **8**, 1137-1146.

[31] Baker, L.-J., Dorocke, J.A., Harris, R.A. & Timm, D.E. (2001). *Structure*, **9**, 539-546.

[32] Dann, C.E., Hsieh, J.-C., Rattner, A., Sharma, D., Nathans, J. & Leahy, D.J. (2001). *Nature* **412**, 86-89.

[33] Stebbins, C.E. & Galán, J.E. (2001). *Nature*, **414**, 77-81.

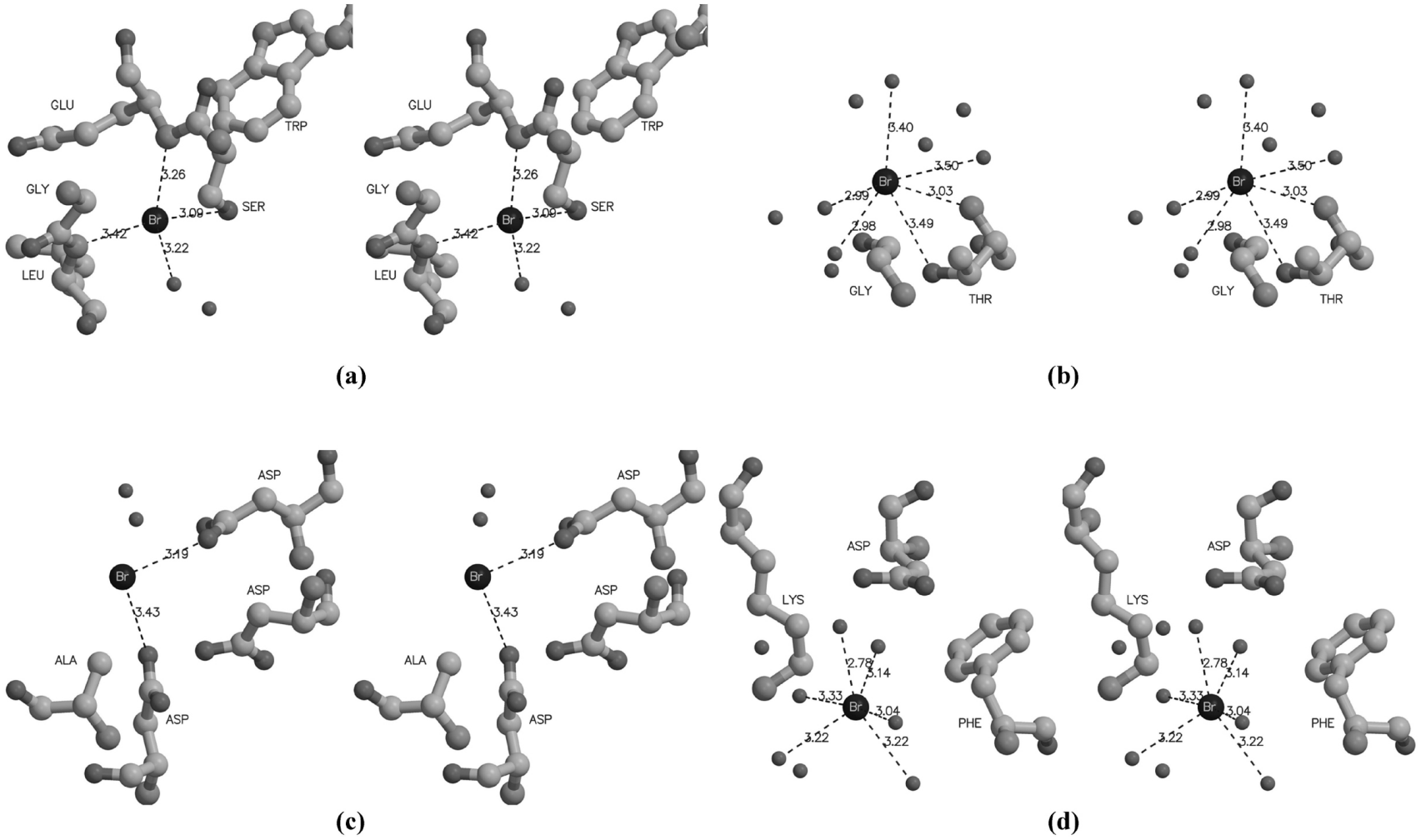


FIG. 1. Stereo plot of the representative sites of cryo soaked ions. Surrounding residues are shown if they have at least one atom closer than 4.5 Å from the ion. Coordination distances to polar atoms closer than 3.5 Å are marked. (a) to (d) bromide sites in acyl protein thioesterase I, (e) to (h) caesium sites in trypsin inhibitor.

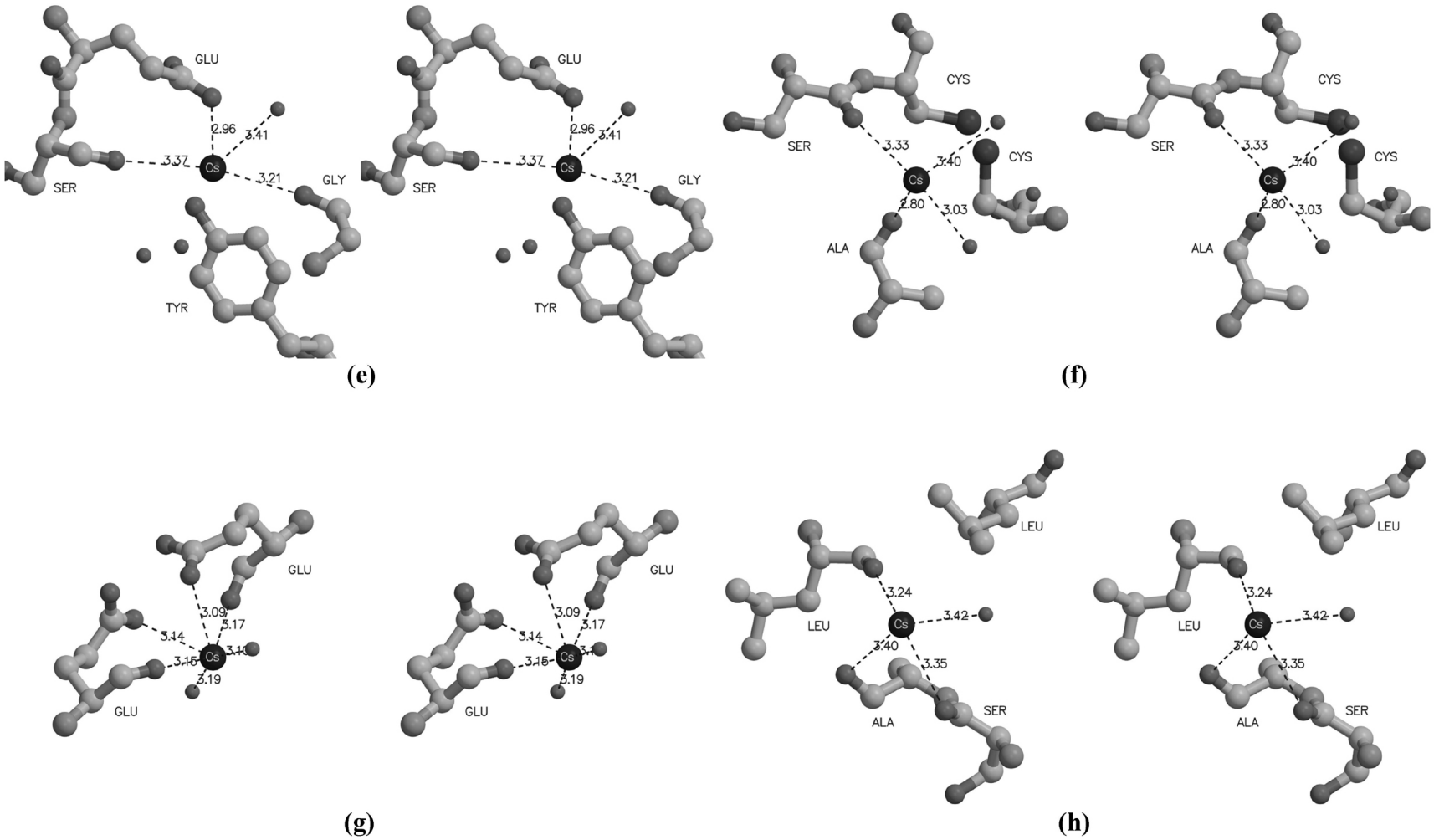


FIG. 1. (Continued)

soaking approach, the results indicate that even larger macromolecules like β-galactosidase[34] or SptP:SicP complex[33] can be solved by this technique. Similarly, macromolecular crystals with a solvent content as low as 35 % have already been solved[26,24].

Bromides and iodides do not require long soaking time. The experience shows that soaking for longer than 30 seconds is not necessary and may in fact lead to the deterioration of the diffraction power. Surprisingly, it has been observed that short halide soaks can improve the crystal diffraction[35] and sometimes even cause the crystal phase transition to a different symmetry[36]. Metal ions seem to require somewhat longer soaking for successful derivatization.

Examples

α-galactosidase from *Trichoderma reesei*

The firsts crystals of α-galactosidase from *Trichoderma reesei* were obtained a few years ago[37]. Since then, many efforts were spent on solving the phase problem for this protein. In spite of using a number of heavy metals for derivatization (nearly 20 chemicals of Pt, Ag, Au, U, W, rare earth elements), all "derivatives" suffered the absence of heavy atom binding.

The method of quick cryo soaking was then used to overcome this difficulty. Native crystals of α-galactosidase ($P2_12_12_1$, a=46.5 b=79.1 c=119.4 Å) were firstly soaked in a cryoprotectant solution containing in addition 0.1 - 0.5 M KI and then used for X-ray diffraction data collection. Initially, these diffraction data, collected at the PCr beamline[38] at the Brazilian National Synchrotron Light Source (LNLS), could not be used for phasing because the search for iodine binding sites failed (non-isomorphism was also observed; a=41.9 b=79.9 c=120.0 Å).

A second quick derivative, prepared using 0.2 M CsCl in the cryoprotectant solution was then used for 1.6 Å resolution data collection at the same beamline. The incorporation of Cs atoms was successful and RSPS[39] and SnB[40] programs found a few equivalent Cs sites independently using solely the anomalous signal of Cs atoms. However, similarly to the I-soaked crystal, the Cs-soaked crystal (a=42.1 b=80.4 c=120.0 Å) was non-isomorphic to the native. Both soaked crystals showed a difference of almost 10% in the **a** cell parameter compared to the native crystals.

Data for I-pseudo-derivative and Cs-derivative of α-galactosidase were used as native and derivative, respectively, for initial SIRAS phasing. Ten Cs sites were used by SHARP[41] in SIRAS approach, followed by DM[42], and the resulting phases were good enough for an automatic model building by wARP[43].

β-galactosidase from *Penicillium sp.*

One of the highest molecular weight protein structures solved so far using a derivative prepared according to the quick cryo soaking procedure was the β-galactosidase from *Penicillium sp.* with 110 kDa (one molecule) in the asymmetric unit. Initial X-ray diffraction studies revealed that β-galactosidase crystallized in space group $P4_3$ with unit cell parameters a=b=110.9 Å, c=161.0 Å and diffracted to 1.85 Å resolution.

An iodine derivative was prepared immersing a native crystal in mother liquor solution containing in addition 0.25 M NaI and 30 % ethylene glycol. Crystals were visually stable in the derivatization solution and did not suffer large changes in cell parameters or loss of diffraction power compared to native crystals. The SnB[40] program used the normalized anomalous differences of this derivative to locate the halide substructure. Phase calculation performed by SHARP[41] in SIRAS approach with 13 iodine sites gave a mean figure of merit of 0.37 in the 27.0 - 2.60 Å resolution range. The final electron density map obtained after density modification with SOLOMON[44] was used by wARP[43] for an automatic model building. The number of built residues was increasing in each cycle, however, the convergence was very slow and three days were required to obtain the final model (95% complete).

Even though just one halide derivative was used for phasing and solving the crystal structure it can be mentioned that a second quick cryo soaked derivative was obtained during model building. At this time, CsCl was used instead of NaI in the derivatization solution. Twelve caesium sites were used for SIRAS phasing. Similar to the first phase calculation this one gave a mean figure of merit of 0.37 in the same resolution range. When native and both derivative data sets were combined in MIRAS phasing, the resulting

[34] Rojas, A.L., Nagem, R.A.P., Neustroev, K.N., Golubev, A.M., Eneyskaya, E.V., et al., & Polikarpov, I. *Submitted.*

[35] Harel, M., Kasher, R., Nicolas, A., Guss, J.M., Balass, M., Fridkin, et al., & Fuchs, S. (2001). *Neuron*, **32**, 265-75.

[36] Dauter, Z., Li, M. & Wlodawer, A. (2001). *Acta Crystallogr.*, **D57**, 239-49.

[37] Golubev, A.M. & Neustroev, K.N. (1993). *J. Mol. Biol.*, **231**, 933-934.

[38] Polikarpov, I., Perles, L.A., de Oliveira, R.T., Oliva, G., et al., & Craievich, A. (1997). *J. Synchrotron Rad.* **5**, 72-76.

[39] CCP4, (1994). *Acta Crystallogr.*, **D50**, 760-763.

[40] Weeks, C.M. & Miller, R. (1999). *J. Appl. Cryst.* **32**, 120-124.

[41] de La Fortelle, E. & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472-494.

[42] Cowtan, K. (1999). *Acta Crystallogr.*, **D55**, 1555-1567.

[43] Perrakis, A., Morris, R. & Lamzin, V.S. (1999). *Nature Struct. Biol.* **6**, 458-463.

[44] Abrahams, J.P. & Leslie, A.G.W. (1996). *Acta Crystallogr.* **D52**, 30-42.

figure of merit was 0.52 and the electron density map showed significant improvement compared to either of the SIRAS maps.

Acyl protein thioesterase

Crystals of human acyl protein thioesterase I were grown from the solution containing a high concentration of ammonium sulfate in the monoclinic cell with two molecules of 228 amino acid residues each in the asymmetric unit. They were soaked for 20 s in the mother liquor with added 1M NaBr. The diffraction data were collected to 1.8 Å resolution at the near-remote wavelength, at the energy 50 eV higher than the Br absorption edge. The data displayed a clear anomalous signal and the structure was solved by SAD, using only this data set.

The initial seven Br sites were located by SnB[40]. They were input to SHARP[41], which after two iterations identified 22 Br sites in the residual maps and produced a phase set with an overall figure of merit 0.40. The strongest 18 Br sites formed two groups of almost identical constellations, clearly identifying the presence of non-crystallographic two-fold axis. The application of the density modification with DM[42] increased the figure of merit to 0.85 and produced an easily interpretable map. The majority of the residues were built automatically using wARP[43]. At the end of the refinement the anomalous difference Fourier map identified a total of 40 bromide sites located at the surface of the two independent protein molecules. They were included in the refinement of the final model with either full or half occupancies, depending on the appearance of the corresponding peaks in the anomalous difference map.

Conclusions

The quick cryo soaking approach was established a few years ago (2000) and since then a number of macromolecular crystalline structures have been solved with this method. The results obtained so far indicate that due to its several intrinsic aspects it can be particularly applicable for high-throughput crystallographic projects.

This approach can be used with a great number of different crystallization solutions and the presence of different compounds, like sugars, additives or even precipitants in high concentration, do not impede the fast incorporation of halide or alkaline metal ions to the solvent regions surrounding the protein molecules. Moreover, halides or alkaline metals are less likely to react with certain compounds that are used during crystallization than some heavy metal salts (*e.g.* they do not precipitate with phosphate anions).

Another interesting point refers to the fact that very little preparative effort and a short time is required to produce a potential derivative. All the equipment and chemical compounds used in this approach can be found in a simple protein crystallization laboratory. A single 2-5 μl derivatization solution drop is used in each trial and the crystal soak time is usually less than a minute. In addition one can immediately see if the crystal is stable or not in solution and modify the salt concentration and/or soaking time to obtain a better derivative.

The choice between a halide and an alkaline metal salt has to be made before the derivatization procedure; however, this selection does not mean that a second salt cannot be also used for another derivative. Sometimes this decision can be easily taken depending on the types of compounds used in the crystallization solution. The iodides or bromides can replace chlorides. Similarly, lithium, sodium and potassium can be substituted by caesium or rubidium. Since preparation of derivatives is very fast and data collection is normally performed with frozen crystals, it can be suggested that both types of derivatives can be prepared and immediately frozen for data collection. Moreover, the use of halide and alkaline metal salts during derivatization opens up the possibility of using two essentially different derivatives to solve a protein structure through MIR(AS) when none of them alone is able to do so.

Bromide and rubidium salts have an advantage over iodide and caesium salts, that the former can be easily used for MAD because their K absorption edges are in the similar energy range as the Se edge, the most often used scatterer for MAD phasing. The latter derivatives, on the other hand, with K absorption edges in the vicinity of 0.35 Å are not suitable for MAD phasing, even though some X-ray diffraction experiments at this energy have been done[45]. Nevertheless iodine and caesium atoms possess significant anomalous signal at longer wavelengths that makes them appropriate for SIRAS, MIRAS and SAD phasing. Specifically at the copper-anode characteristic wavelength (1.54 Å), f'' of I and Cs atoms are 6.8 and 7.9 electrons, respectively. In general this new approach was proposed as an alternative way of preparing derivatives when a protein does not bind heavy-metal atoms or is not amenable to the preparation of a SeMet variant.

Acknowledgments

The authors would like to thank A. Golubev for providing substantial information about α-galactosidase phasing. This work was supported in part by grants 98/06218-6 from FAPESP (Brazil). The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organization imply endorsement by the U.S. Government.

[45] Takeda, K., et al., & Kamiya, N. (2001). Poster sect. in 7th Int. Conf. on Biol. Synchr. Rad., São Pedro, SP, Brazil.

8 Referências bibliográficas

ABRAHAMS, J. P.. Bias reduction in phase refinement by modified interference functions: introducing the gamma correction. **Acta Crystallographica D**, n. 53, p. 371-376, 1997.

ABRAHAMS, J. P. & LESLIE, A. G. W.. Methods used in the structure determination of bovine mitochondrial F-1 ATPase. **Acta Crystallographica D**, n. 52, p. 30-42, 1996.

BERMAN, H. M. *et al*.. The protein data bank. **Nucleic Acids Research**, v. 28, n. 1, p. 235-242, 2000.

BLAKE, C. C. F. *et al*.. How the structure of lysozyme was actually determined. In: ROSSMANN, M. G. & ARNOLD, E. (Ed.). **International Tables for Crystallography**. Dordrecht: Kluwer Academic Publishers, 2001. v. F, p. 745-772.

BLESSING, R. H. & SMITH, G. D.. Difference structure-factor normalization for heavy-atom or anomalous-scattering substructure determinations. **Journal of Applied Crystallography**, n. 32, p. 664-670, 1999.

BLOW, D. M. & CRICK, F. H. C.. The treatment of errors in the isomorphous replacement method. **Acta Crystallographica**, n. 12, p. 794-802, 1959.

BRAGG, W. L.. The structure of some crystals as indicated by their diffraction of X-rays. **Proceedings of the Royal Society of London A**, n. 89, p. 248-277, 1913.

BRAGG, W. H. & BRAGG, W. L.. The reflection of X-rays in crystals. **Proceedings of the Royal Society of London A**, n. 88, p. 428-438, 1913.

BRANDEN, C. & TOOZE, J.. **Introduction to Protein Structure**. New York: Garland Publishing, 1991.

BRÜNGER, A. T.; KURIYAN, J. & KARPLUS, M.. Crystallographic R factor refinement by molecular dynamics. **Science**, v. 235, p. 458-460, jan. 1987.

CROMER, D. T.. Calculation of anomalous scattering factors at arbitrary wavelengths. **Journal of Applied Crystallography**, n. 16, p. 437-438, 1983.

DAUTER, Z.; DAUTER, M. & RAJASHANKAR, K. R.. Novel approach to phasing proteins: derivatization by short cryo-soaking with halides. **Acta Crystallographica D**, n. 56, p. 232-237, 2000.

DE LA FORTELLE, E. & BRICOGNE, G.. Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. **Methods in Enzymology**. Nova York: Academic Press, 1997. v. 276, p. 472-494.

DRENTH, J.. **Principles of Protein X-ray Crystallography**. 2. ed. New York: Springer, 1999.

DUMOUTIER, L.; LOUAHED, J. & RENAULD, J. C.. 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. **The Journal of Immunology**, n. 164, p. 1814-1819, 2000.

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 Science USA**, n. 97, p. 10144-10149, 2000.

FRIEDRICH, W.; KNIPPING, P. & LAUE, M. V.. Interferenz-Erscheinungen bei Röntgenstrahlen. **Proceedings of the Bavarian Academy of Sciences**, p. 303-322, 1912.

GARMAN, E. F. & SCHNEIDER, T. R.. Macromolecular cryocrystallography. **Journal of Applied Crystallography**, n. 30, p. 211-237, 1997.

GIACOVAZZO, C. *et al*.. **Fundamentals of Crystallography**. Oxford: Oxford University Press, 1992.

GIACOVAZZO, C.. The diffraction of X-rays by crystals. In: GIACOVAZZO, C. *et al*.. **Fundamentals of Crystallography**. Oxford: Oxford University Press, 1992. p. 141-228.

GOLUBEV, A. M. & NEUSTROEV, K. N.. Crystallization of α-galactosidase from *Trichoderma reesei*. **Journal of Molecular Biology**, v. 231, n. 3, p. 933-934, 1993.

GREEN, D. W.; INGRAM, V. M. & PERUTZ, M. F.. The structure of haemoglobin. IV. Sign determination by the isomorphous replacement method. **Proceedings of the Royal Society of London A. Mathematical and Physical Sciences**, London, v. 225, p. 287-307, set. 1954.

HARKER, D.. The determination of the phases of the structure factors of non-centrosymmetric crystals by the method of double isomorphous replacement. **Acta Crystallographica**, n. 9, p. 1-9, 1956.

HAHN, T. (Ed.). **International Tables for Crystallography**. 5. ed. Dordrecht: Kluwer Academic Publishers, 2002. v. A.

HENDRICKSON, W. A.. Analysis of protein structure from diffraction measurements at multiple wavelengths. **Transactions of the American Crystallographic Association**, n. 21, 11-21, 1985.

HENDRICKSON, W. A.; SMITH, J. L. & SHERIFF, S.. Direct phase determination based on anomalous scattering. In: WYCKOFF, H. W.; HIRS, C. H. W. & TIMASHEFF, S. N. (Ed.). **Methods in Enzymology**. Nova York: Academic Press, 1985. v. 115 B, p. 41-55.

HENDRICKSON, W. A.. Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. **Science**, v. 254, n. 5028, p. 51-58, out. 1991.

HENDRICKSON, W. A.. Phase information from anomalous-scattering measurements. **Acta Crystallographica A**, n. 35, p. 245-247, 1979.

JONES, T. A. *et al*.. Improved methods for building protein models in electron-density maps and the location of errors in these models. **Acta Crystallographica A**, n. 47, p. 110-119, 1991.

JACK, A. & LEVITT, M.. Refinement of large structures by simultaneous minimization of energy and R factor. **Acta Crystallographica A**, n. 34, p. 931-935, 1978.

KENDREW, J. C. *et al*.. Structure of myoglobin. A three-dimensional Fourier synthesis at 2 Å resolution. **Nature**, London, n. 185, p. 422-427, 1960.

KRAUCHENCO, S. *et al*.. Crystallization and preliminary X-ray diffraction analysis of a novel trypsin inhibitor from seeds of *Copaifera langsdorffii*. **Acta Crystallographica D**, n. 57, p. 1316-1318, 2001.

LESLIE, A. G. W.. A reciprocal-space method for calculating a molecular envelope using the algorithm of B. C. Wang. **Acta Crystallographica A**, n. 43, p. 134-136, 1987.

MATTHEWS, B. W.. The extension of the isomorphous replacement method to include anomalous scattering measurements. **Acta Crystallographica**, n. 20, p. 82-86, 1966.

MCLANE, M. P. *et al*.. Interleukin-9 promotes allergen-induced eosinophilic inflammation and airway hyperresponsiveness in transgenic mice. **American Journal of Respiratory Cell and Molecular Biology**. n. 19, p. 713-720, 1998.

NAGEM, R. A. P. *et al*.. Crystallization and synchrotron X-ray diffraction studies of human interleukin-22. **Acta Crystallographica D**, n. 58, p. 529-530, 2002.

NCBI - National Center for Biotechnology Information. Human genome resources. Disponível em: <http://www.ncbi.nlm.nih.gov/genome/guide/human/>. Acesso em: 26 abril 2003.

NELSON, D. L. & COX, M. M.. **Lehninger Principles of Biochemistry**. 3. ed. New York: Worth Publishers, 2000.

NEUSTROEV, K. N. *et al*.. Purification, crystallization and preliminary diffraction study of β-galactosidase from *Penicillium sp*.. **Acta Crystallographica D**, n. 56, p. 1508-1509, 2000.

OLIVA, G. *et al*.. Structure and catalytic mechanism of glucosamine 6-phosphate deaminase from *Escherichia coli* at 2.1 Å resolution. **Structure**, n. 3, p. 1323-1332, 1995.

PATTERSON, A. L.. A Fourier series method for the determination of the components of interatomic distances in crystals. **Physical Review**, n. 46, p. 372-376, 1934.

PERRAKIS, A.; MORRIS, R. & LAMZIN, V. S.. Automated protein model building combined with iterative structure refinement. **Nature Structural Biology**, n. 6, p. 458-463, 1999.

PERUTZ, M. F. *et al*.. Structure of haemoglobin. A three-dimensional Fourier synthesis at 5.5 Å resolution, obtained by X-ray analysis. **Nature**, London, n. 185, p. 416-422, 1960.

POLIKARPOV, I. *et al*.. Set-up and experimental parameters of the protein crystallography beamline at the Brazilian National Synchrotron Laboratory. **Journal of Synchrotron Radiation**, v. 5, p. 72-76, mar. 1998.

POLIKARPOV, I. *et al*.. The protein crystallography beamline at LNLS, the Brazilian National Synchrotron Light Source. **Nuclear Instruments and Methods in Physics Research A**, n. 405, p. 159-164, 1998.

ROSSMANN, M. G.. Historical background. In: ROSSMANN, M. G. & ARNOLD, E. (Ed.). **International Tables for Crystallography**. Dordrecht: Kluwer Academic Publishers, 2001. v. F, p. 4-9.

SHELDRICK, G. M. *et al*.. Ab initio phasing. In: ROSSMANN, M. G. & ARNOLD, E. (Ed.). **International Tables for Crystallography**. Dordrecht: Kluwer Academic Publishers, 2001. v. F, p. 333-345.

SHELDRICK, G. M.. SHELX: applications to macromolecules. In: Fortier, S.. **Direct methods for solving macromolecular structures.** Dordrecht: Kluwer Academic Publishers, 1998. p. 401-411.

STEVENS, R. C.; YOKOYAMA, S. & WILSON, I. A.. Global efforts in structural genomics. **Science**, v. 294, n. 5540, p. 89-92, out. 2001.

STUBBS, M. T. & HUBER, R.. Locating heavy-atom sites. In: ROSSMANN, M. G. & ARNOLD, E. (Ed.). **International Tables for Crystallography**. Dordrecht: Kluwer Academic Publishers, 2001. v. F, p. 256-260.

SUSSMAN, J. L.. Constrained-restrained least-squares (CORELS) refinement of proteins and nucleic acids. In: WYCKOFF, H. W.; HIRS, C. H. W. & TIMASHEFF, S. N. (Ed.). **Methods in Enzymology**. Nova York: Academic Press, 1985. v. 115 B, p. 271-303.

TEMANN, U. A. *et al*.. Expression of interleukin-9 in the lungs of transgenic mice causes airway inflammation, mast cell hyperplasia, and bronchial hyperresponsiveness. **The Journal of Experimental Medicine**. n. 188, p. 1307-1320, 1998.

TERWILLIGER, T. C. & EISENBERG, D.. Isomorphous replacement: effects of errors on the phase probability distribution. **Acta Crystallographica A**, n. 43, p. 6-13, 1987.

TRONRUD, D. E.. Conjugate-direction minimization: an improved method for the refinement of macromolecules. **Acta Crystallographica A**, n. 48, p. 912-916, 1992.

TRONRUD, D. E.; EYCK, L. F. T. & MATTHEWS, B. W.. An efficient general-purpose least-squares refinement program for macromolecular structures. **Acta Crystallographica A**, n. 43, p. 489-501, 1987.

WEEKS, C. M. & MILLER, R.. The design and implementation of SnB version 2.0. **Journal of Applied Crystallography**, n. 32, p. 120-124, 1999.

9 Anexo 1 – Outras publicações

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Crystallization and preliminary X-ray diffraction studies of human catalase

R. A. P. Nagem,[a,b] E. A. L. Martins,[c] V. M. Gonçalves,[c] R. Aparício[a,b] and I. Polikarpov[a]*

[a]Laboratório Nacional de Luz Síncrotron, Caixa Postal 6192, CEP 13083-970, Campinas SP, Brazil, [b]Depto. Física, UNICAMP, Caixa Postal 6165, CEP, 13083-970, Campinas SP, Brazil, and [c]Centro de Biotecnologia, Instituto Butantan, Av. Vital Brasil 1500, CEP 05503 900, São Paulo, Brazil

Correspondence e-mail: igor@lnls.br

Received 1 May 1999
Accepted 14 July 1999

The enzyme catalase (H_2O_2–H_2O_2 oxidoreductase; E.C. 11.1.6) was purified from haemolysate of human placenta and crystallized using the vapour-diffusion technique. Synchrotron-radiation diffraction data have been collected to 1.76 Å resolution. The enzyme crystallized in the space group $P2_12_12_1$, with unit-cell dimensions a = 83.6, b = 139.4, c = 227.5 Å. A molecular-replacement solution of the structure has been obtained using beef liver catalase (PDB code 4blc) as a search model.

1. Introduction

The enzyme catalase (H_2O_2–H_2O_2 oxidoreductase; E.C. 11.1.6) plays an important role in cellular defence against active oxygen species (Aebi, 1984; Halliwell & Gutteridge, 1986; Michiels *et al.*, 1994). Its mechanism of decreasing the hydrogen peroxide concentration has been well described (Jones, 1982; Feinsteins *et al.*, 1971; Almarsson, 1993).

Catalase is found in almost all aerobic organisms (Murthy *et al.*, 1982), and some microbial catalases are used in various industrial processes in which H_2O_2 is utilized for bleaching or disinfecting (Godfrey & West, 1996). Studies in mammalian cells have shown that a lack of or decreased amount of catalase activity is related to many diseases: respiratory distress syndrome (Metnitz *et al.*, 1999), peptic ulcer (Majani & Das, 1998), DNA damage and carcinogenesis (Ohkuma & Kawanishi, 1999), as well as the ageing process (Shah *et al.*, 1999; Casado *et al.*, 1998). Cases of acatalasaemia and hypocatalasaemia have been described (Kishimoto *et al.*, 1992). These studies have shown that gene mutations inhibiting catalase expression or affecting molecules in such a way as to cause instability of tetramer formation decrease the enzyme activity.

Mammalian catalases have been proposed for clinical application in the treatment of many diseases in which oxidative injury has some importance (Greenwald, 1990), such as myocardial ischaemia reperfusion oxidative injury (Simpson *et al.*, 1987; Zughaib *et al.*, 1994), arthritis and inflammatory diseases (Greenwald, 1990), and ageing (Halliwell & Gutteridge, 1986). The human catalase cDNA from kidney has been cloned and its nucleotide sequence has been determined (Quan *et al.*, 1986).

The active mammalian catalase is a homotetramer of 4 × 60 kDa, with one site for haem (Schonbaun & Chance, 1976) and one site for NADP (Kirkiman & Gaetani, 1984; Gouet *et al.*, 1995) per monomer. Crystal structures of several catalases have been described: beef liver (Reid *et al.*, 1981; Fita *et al.*, 1986), *Penicillium vitale* (Melik-Adamyan *et al.*, 1986), *Proteus mirablis* (Gouet *et al.*, 1995) and *Escherichia coli* (Bravo *et al.*, 1995).

In the present work, we describe crystallization and preliminary X-ray diffraction studies of human catalase at 1.76 Å resolution.

2. Protein purification

Various procedures for catalase purification from different organisms, different tissues and blood erythrocytes have been described. For this work, catalase was purified from haemolysate of human placenta by combination of the Cohn precipitation method and chromatography (Gonçalves *et al.*, 1999). Essentially, human placentas frozen immediately after childbirth were defrosted, ground and haemolysate was extracted with saline and ethanol. The cellular mass was separated by centrifugation. Haemoglobin was precipitated by ethanol/chloroform and separated by filtration. The clarified solution was submitted to two steps of chromatography: anion-exchange chromatography on Q-Sepharose and affinity chromatography on blue-Sepharose. The purified catalase was desalted, concentrated to 30 mg ml^{-1} in an Amicon system and used for crystallization trials.

3. Crystallization and data collection

Preliminary screening of the crystallization conditions was performed using a sparse-matrix screen at 291 K (Crystal Screen I and II, Hampton Research Corp.). Small crystals were found in the condition number 37 of the Crystal Screen II kit (10% PEG 8000, 8% ethylene glycol, 0.1 *M* HEPES pH 7.5). A search for refined crystallization conditions has been carried out. New crystals were grown at

Table 1
Crystal data and data-collection statistics.

Statistical values for highest resolution shell (1.79–1.76 Å) are shown in parentheses.

Space group	$P2_12_12_1$
Unit-cell dimensions (Å)	a = 83.6, b = 139.4, c = 227.5
Resolution (Å)	15.0–1.76
No. of observations	604975
No. of unique reflections	232698
$\langle I/\sigma(I)\rangle$	9.9 (1.4)
Multiplicity	2.6 (1.5)
Completeness (%)	88.7 (59.1)
$R_{\rm merge}$† (%)	10.2 (33.4)

† $R_{\rm merge} = \sum_{hkl} |I - \langle I\rangle| / \sum_{hkl} I$.

room temperature using the hanging-drop vapour-diffusion technique, by mixing equal volumes (2 µl + 2 µl) of a protein solution concentrated to 30 mg ml^{-1} and a reservoir solution containing 6–9% PEG 8000, 0.1 *M* MES in the pH range 5.5–8.0. Crystals of two different morphologies appeared after 2 d under similar crystallization conditions and frequently in the same drop: well shaped bipyramidal crystals of dimensions 0.5 × 0.4 × 0.4 mm and rectangular-shaped crystals measuring 0.4 × 0.2 × 0.2 mm. It was subsequently found that the bipyramidal crystals did not diffract X-rays; all the data were therefore collected from the rectangular-shaped crystals.

X-ray diffraction data were collected from crystals immersed for 1 min in a cryo-cooling solution (22% ethylene glycol, 10% PEG 8000, 0.1 *M* HEPES pH 7.5), mounted in a rayon loop and flash-cooled to 80 K in a cold nitrogen stream. Data collection was performed at the Protein Crystallography beamline (Polikarpov, Oliva *et al.*, 1997; Polikarpov, Perles *et al.*, 1997) at the Laboratório Nacional de Luz Síncrotron (Campinas, SP, Brazil), using a MAR345 image plate. The synchrotron-radiation wavelength was set to 1.38 Å to optimize X-ray flux and minimize absorption errors. Several crystals were tested until a good set of diffraction data was obtained. The first image was subjected to the autoindexing routine of *DENZO* (Otwinowski, 1993), from which the best refined solution was a primitive orthorhombic cell. Following an optimum strategy of data collection suggested by the program *marHKL*, a total of 81° of data was collected in steps of 0.5°. The collected images were processed and scaled with the programs *DENZO* and *SCALEPACK* (Otwinowski, 1993). Data sets were collected and processed for two crystals. Details of the significantly better one, which had a mosaic spread of 0.5°, are given in Table 1; the other had a high mosaicity (1.5°) and gave data to a resolution of 2.05 Å. Both showed considerable radiation decay.

Calculations using the Matthews coefficient (Matthews, 1968) suggested the presence of one tetramer per asymmetric unit (V_m = 2.76 Å^3 Da^{-1}), which was subsequently verified by molecular replacement. A *BLAST* search with the human catalase primary sequence (total length of 527 amino-acid residues) against the PDB database showed that beef liver catalase (506 amino-acid residues) represents 91% identity and 95% similarity in primary sequence, followed by *Proteus mirablis* catalase (484 amino-acid residues), which displays 52% primary sequence identity and 65% similarity. The crystal structure of human catalase was solved by molecular-replacement method with the program *AMoRe* (Navaza, 1994), using a tetramer of beef liver catalase (PDB code 4blc) as a search model.

The rotation function was calculated using diffraction data in the resolution range 8.0–3.5 Å using a Patterson radius of 40 Å. The rotation function gave a clear solution [correlation coefficient (CC) of 0.357], where the next highest peak had CC = 0.158. The translation search was performed in the same resolution range using the Crowther & Blow (1967) translation function. The translation search gave a strong peak with CC = 0.614 and an *R* factor of 36.7%, and this solution was then subjected to ten cycles of rigid-body refinement against all data between 8.0 and 3.5 Å resolution (fitting function of *AMoRe*). The resulting *R* factor was 33.4%, with a correlation coefficient of 0.689. Refinement was undertaken using the program *REFMAC* (Murshudov *et al.*, 1997). Three cycles of rigid-body refinement against data in the resolution range 15–1.76 Å were initially performed, treating each of the monomers with its corresponding haem group as a separate rigid body, followed by 40 steps of positional and *B*-factor refinement. Non-crystallographic symmetry restraints have been applied to the monomers, requiring them to be related by local 222 symmetry throughout the refinement. At present, $R_{\rm free}$ and the *R* factor of the model are 29.4 and 25.8%, respectively. No water molecules have so far been introduced. The current model has essentially the same fold as beef liver catalase, but represents significant conformational differences in the C-terminal and N-terminal regions. Further steps of model building and refinement are under way.

We are grateful to J. Brandão for his help with data collection and P. R. de Moura for help with crystallization. Financial support from CNPq and FAPESP is acknowledged.

References

Aebi, H. (1984). *Methods Enzymol.* **105**, 121–127.

Almarsson, O. (1993). *J. Am. Chem. Soc.* **115**, 7093–7102.

Bravo, J., Verdaguer, N., Tormo, J., Betzel, C., Switala, J., Loewen, R. C. & Fita, I. (1995). *Structure*, **3**, 491–502.

Casado, A., de la Torre, R., Lopez-Fernandez, E., Carrascosa, D. & Venarucci, D. (1998). *Gac. Med. Mex.* **134**, 539–544.

Crowther, R. A. & Blow, D. M. (1967). *Acta Cryst.* **23**, 544–549.

Feinsteins, R. N., Savol, R. & Howard, J. B. (1971). *Enzymologia*, **41**, 345–352.

Fita, I., Silva, A. M., Murthy, M. R. N. & Rossmann, M. G. (1986). *Acta Cryst.* B**42**, 497–515.

Godfrey, T. & West, S. (1996). Editors. *Industrial Enzymology 2.* London: MacMillan.

Gonçalves, V. M., Leite, L. C. C., Raw, I. & Cabrera-Crespo, J. (1999). *Biotechnol. Appl. Biochem.* **29**, 73–77.

Gouet, P., Jouvre, H. M. & Dideberg, O. (1995). *J. Mol. Biol.* **249**, 933–954.

Greenwald, R. A. (1990). *Free Radic. Biol. Med.* **8**, 201–209.

Halliwell, B. & Gutteridge, J. M. (1986). *Free Radicals in Biology and Medicine.* Oxford: Clarendon.

Jones, P. (1982). *The Biological Chemistry of Iron*, edited by H. B. Dunford, pp. 427–438. Dordrecht: Reidel..

Kirkiman, H. N. & Gaetani, G. F. (1984). *Proc. Natl Acad. Sci. USA*, **81**, 4343–4347.

Kishimoto, Y., Murakami, Y., Hayashi, K., Takahara, S., Sugimura, T. & Sekiya, T. (1992). *Hum. Genet.* **88**, 487–490.

Majani, V. & Das, U. N. (1998). *Prostag. Leuk. ESS*, **59**, 401–406.

Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.

Melik-Adamyan, W. R., Barynin, V. V., Vagin, A. A., Borisov, V. V., Vainshtein, B. K., Fita, I., Murthy, M. R. N. & Rossmann, M. G. (1986). *J. Mol. Biol.* **88**, 63–72.

Metnitz, P. G., Bartens, C., Fischer, M., Fridrich, P., Steltzer, H. & Druml, W. (1999). *Intensive Care Med.* **25**, 180–185.

Michiels, C., Raes, M., Toussaint, O. & Remacle, J. (1994). *Free Radic. Biol. Med.* **17**, 235–248.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

Murthy, M. R. N., Reid, T. J. III, Sicignano, A., Tanaka, N. & Rossmann, M. G. (1982). *The Biological Chemistry of Iron*, edited by H. B. Dunford, pp. 439–458. Dordrecht: Reidel.

Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.

Ohkuma, Y. & Kawanishi, S. (1999). *Biochem. Biophys. Res. Commun.* **257**, 555–560.

Otwinowski, Z. (1993). *Proceedings of the CCP4 Study Weekend. Data Collection and Processing*, edited by L. Sawyer, N. Isaacs & S. Bailey, pp. 56–62. Warrington: Daresbury Laboratory.

Polikarpov, I., Oliva, G., Castellano, E. E., Garratt, R., Arruda, P., Leite, A. & Craievich, A. (1997). *Nucl. Instrum. Methods A*, **405**, 159–164.

Polikarpov, I., Perles, L. A., de Oliveira, R. T., Oliva, G., Castellano, E. E., Garratt, R. & Craievich, A. (1997). *J. Synchrotron Rad.* **5**, 72–76.

Quan, F., Korneluk, R. G., Tropak, M. B. & Gravel, R. A. (1986). *Nucleic Acids Res.* **14**(13), 5321–5335.

Reid, T. J. III, Murthy, M. R. N., Sicignano, A., Tanaka, N., Musick, W. D. L. & Rossmann, M. G. (1981). *Proc. Natl Acad. Sci. USA*, **78**, 4767–4771.

Shah, P. C., Brolin, R. E., Amenta P. S. & Deshmukh, D. R. (1999). *Mech. Ageing Dev.* **107**, 37–50.

Schonbaun, G. R. & Chance, B. (1976). *The Enzymes*, Vol. 13, 3rd ed., edited by P. D. Boyer, pp. 383–408. New York: Academic Press.

Simpson, P. J., Mickelson, J. K. & Lucchesi, B. R. (1987). *Fed. Proc.* **46**, 2413–2421.

Zughaib, M. E., Tang, X. L. & Bolli, R. (1994). *Ann. NY Acad. Sci.* **723**, 218–228.

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Crystallization and preliminary X-ray study of haem-binding protein from the bloodsucking insect *Rhodnius prolixus*

R. A. P. Nagem,[a,b] J. R. Brandão Neto,[a] V. P. Forrer,[a] M. H. Sorgine,[c] G. O. Paiva-Silva,[c] H. Masuda,[c] R. Meneghini,[a] P. L. Oliveira[c] and I. Polikarpov[a]*

[a]National Synchrotron Light Laboratory, LNLS, Caixa Postal 6192, CEP 13083-970, Campinas, SP, Brazil, [b]Gleb Wataghin Physics Institute, State University of Campinas, UNICAMP, Caixa Postal 6165, CEP 13083-970, Campinas, SP, Brazil, and [c]UFRJ, ICB, Departamento Bioquímica Médica, Rio de Janeiro, Brazil

Correspondence e-mail: igor@lnls.br

Received 11 December 2000
Accepted 12 March 2001

Rhodnius haem-binding protein (RHBP) from the bloodsucking insect *Rhodnius prolixus*, a 15 kDa protein, has been crystallized using polyethylene glycol as a precipitant. X-ray diffraction data have been collected at a synchrotron source. The crystals belong to the space group $P4_{1(3)}2_12$, with unit-cell parameters $a = b = 64.98$, $c = 210.68$ Å, and diffract beyond 2.6 Å resolution.

1. Introduction

Haem, iron protoporphyrin-IX, participates in several fundamental biochemical reactions such as respiration, oxygen transport in extracellular fluids and photosynthesis. On the other hand, haem may be a catalyst of the formation of reactive oxygen species and thus lead to oxidation of lipids, proteins and DNA (Aft & Mueller, 1983; Tappel, 1955; Vincent, 1989). Owing to its potential toxicity, haem is normally found associated with proteins such as haemopexin and albumin, free haem frequently being related to pathological conditions (Rytter & Tyrrel, 2000).

Haematophagous arthropods ingest enormous amounts of blood. The bloodsucking insect *R. prolixus* takes between five to ten times its body weight each single meal (Lehane, 1991). As vertebrate blood has about 10 m*M* of haem bound to haemoglobin, the digestion of blood in the midgut of this insect generates potentially toxic amounts of haem. Furthermore, after a blood meal water is massively excreted, leading to even higher haem concentrations in the gut lumen. Therefore, in order to use blood as the sole food source, haematophagous arthropods must develop efficient ways to counteract haem toxicity and a whole array of antioxidant defences designed to prevent radical formation or to scavenge oxygen reactive species. Among these defences, a haem-binding protein (RHBP; *Rhodnius* haem-binding protein) has been described in the haemolymph and oocytes of *R. prolixus*.

RHBP has a single 15 kDa polypeptide chain and in the haemolymph can be found free (apoRHBP) or associated with one haem molecule (holoRHBP). The binding of haem to circulating apoRHBP protects this insect from haem-induced lipid peroxidation (Dansa-Petretski *et al.*, 1995).

In this paper, we describe the purification, crystallization and results of preliminary X-ray diffraction study of holoRHBP as a first step toward its crystal structure solution. The knowledge of the three-dimensional structure of the protein will help in the understanding of the structural basis of its function.

2. Materials and methods

HoloRHBP was purified to homogeneity from *R. prolixus* oocytes as described (Oliveira *et al.*, 1995). Initial crystallization conditions were screened by the sparse-matrix method (Jancarik & Kim, 1991) using the macromolecular crystallization reagent kits I and II (Hampton Research). In each trial, a hanging drop of 1 µl of protein solution (10 mg ml^{-1} in water) was mixed with 1 µl of precipitant solution and then equilibrated against a reservoir containing 500 µl of precipitant solution. Small plane-like dark red crystals grew at 291 K in precipitant solution No. 9 [30%(*w*/*v*) polyethylene glycol 4000, 0.2 *M* ammonium acetate, 0.1 *M* trisodium citrate dihydrate pH 5.6] from Hampton reagent kit number I in three weeks. Further optimization at 291 K, including pH refinement (McPherson, 1995), leads to new values of precipitant concentration (25% PEG 4000) and pH (6.0). Well formed bipyramidal crystals of dimensions $0.05 \times 0.04 \times 0.03$ mm grew in three to six weeks.

All data sets were collected using a 345 mm MAR Research image-plate detector at the LNLS Protein Crystallography beamline (Polikarpov, Oliva *et al.*, 1998; Polikarpov, Perles *et al.*, 1998) by the oscillation method at 100 ± 1 K. The exposure time was approximately 5 min per frame. For the native data set the crystal-to-detector distance was set to 190 mm and two oscillation ranges were used: 0.8 and 0.6°. The crystal-to-detector distance for the derivative data set was 155 mm and an oscillation range of 0.7° was used. The incident radiation wavelength for native and derivative data sets was 1.55 Å. The wavelength was chosen in order to maximize the flux in the spectrum available at the beamline (Polikarpov *et al.*, 1997; Rossmann & Blow, 1962).

The data were processed using *DENZO* and *SCALEPACK* packages (Otwinowski & Minor, 1997). Results of data processing for the native and iodine derivative are summarized in Table 1.

3. Results and discussion

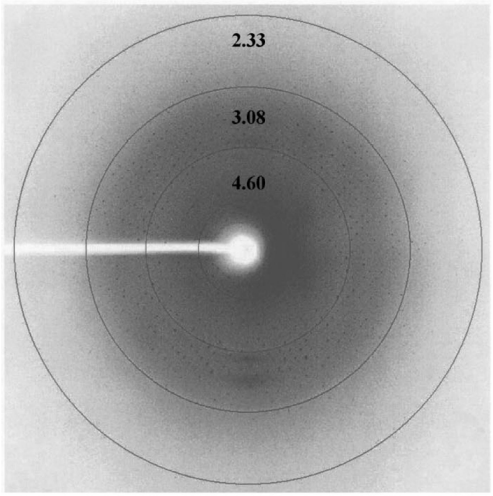
The first crystals obtained diffracted to 3.2 Å resolution and presented considerable radiation decay when exposed to the X-ray beam at room temperature, resisting for only a few images. Cryocrystallographic techniques (Garman & Schneider, 1997) were employed to overcome radiation damage. Flash-freezing of the crystal in a gaseous nitrogen stream (Oxford Cryosystems) was performed after dipping the crystal in a mixture of the mother liquor with 20% ethylene glycol.

Three native data sets were collected with crystals obtained in the first refined crystallization round. The best native diffraction data set, extending to 3.2 Å resolution, is presented in Table 1. Slightly larger crystals were grown in subsequent crystallization trials, as evidenced by the first iodine-derivative data set recently collected to 2.6 Å resolution (Fig. 1). The iodine derivative was prepared by a rapid cryosoaking procedure (Dauter *et al.*, 2000) using a cryoprotectant solution containing in addition 1.0 *M* sodium iodide. Flash-freezing significantly improved the crystal resistance to X-rays, allowing a higher X-ray dose per image. An improvement in the $I/\sigma(I)$ ratio along with a high-redundancy data set led to better statistics and to a higher resolution cutoff.

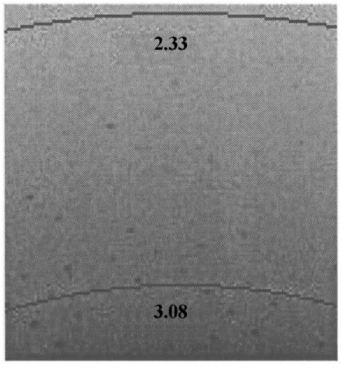
The calculated unit-cell volume is 8.82×105 Å^3. Assuming three monomers per asymmetric unit, the calculated Matthews coefficient (V_M; Matthews, 1968) is 2.45 Å^3 Da^{-1}, which corresponds to 47.8% solvent content. On the other hand, if we presume a tetramer in the asymmetric unit, V_M is 1.84 Å^3 Da^{-1} and the solvent content is 30.5%. A self-rotation function calculation using the *POLARRFN* program from the *CCP*4 suite (Collaborative Computational Project, Number 4, 1994) using different integration radii and resolution ranges did not give conclusive information about the asymmetric unit content. No consistent evidence for a threefold or fourfold non-crystallographic axis was observed, which might indicate that the non-crystallographic axis is approximately parallel to one of the cell axes. To elucidate this question, a complete structure determination will be required.

Sequence alignment has shown no significant homology with other proteins of known three-dimensional structure; therefore, a heavy-atom multiple isomorphous replacement method for the structure solution will have to be used. An extensive search for additional heavy-atom derivatives is currently under way.

This work was supported by grants 99/03387-4 from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brazil to IP. RAPN, JRBN, VPF, RM and IP gratefully acknowledge financial support from FAPESP and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil.



(*a*)



(*b*)

Figure 1
(*a*) A typical diffraction pattern from the iodine-derivative data set. The crystal-to-detector distance was set to give a resolution limit of 2.33 Å at the outer edge of the image. (*b*) Enhanced contrast detail image from the highest resolution shell.

Table 1
Native and iodine-derivative crystal data and data-collection statistics.

Values for the highest resolution shell are shown in parentheses.

	Native	Iodine derivative
Space group	$P4_{1(3)}2_12$	$P4_{1(3)}2_12$
Unit-cell parameters (Å)	$a = b = 65.24$, $c = 207.14$	$a = b = 64.98$, $c = 210.68$
Resolution (Å)	35.0–3.20 (3.35–3.20)	22.0–2.60 (2.72–2.60)
No. of observations	24684 (3062)	48064 (5386)
No. of unique reflections	7642 (940)	19721 (2400)
$\langle I/\sigma(I)\rangle$	8.3 (3.9)	10.4 (2.8)
Multiplicity	3.2 (3.3)	2.4 (2.2)
Completeness (%)	96.0 (99.1)	74.7 (73.1)
R_{merge}† (%)	12.9 (31.3)	8.7 (40.8)
R_{fac}‡ (%)		36.9

† $R_{merge} = \sum(I - \langle I\rangle)/\sum I$. ‡ $R_{fac} = \sum |F_{PH} - F_P|/\sum F_P$.

References

Aft, R. L. & Mueller, G. C. (1983). *J. Biol. Chem.* **258**, 12069–12072.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Dansa-Petretski, M., Ribeiro, J. M., Atella, G. C., Masuda, H. & Oliveira, P. L. (1995). *J. Biol. Chem.* **270**, 10893–10896.
Dauter, Z., Dauter, M. & Rajashankar, K. R. (2000). *Acta Cryst.* D**56**, 232–237.
Garman, E. F. & Schneider, T. R. (1997). *J. Appl. Cryst.* **30**, 211–237.
Jancarik, J. & Kim, S.-H. (1991). *J. Appl. Cryst.* **24**, 409–411.
Lehane, M. J. (1991). *Biology of Bloodsucking Insects*, pp. 79–110. London: Harper Collins Academic.
McPherson, A. (1995). *J. Appl. Cryst.* **28**, 362–365.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
Oliveira, P. L., Kawooya, J. K., Ribeiro, J. M. C., Meyer, T., Poorman, R., Alves, E. W., Walker, F. A., Machado, E. A., Nussenzveig, R. H., Padovan, G. J. & Masuda, H. (1995). *J. Biol. Chem.* **270**, 10897–10901.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Polikarpov, I., Oliva, G., Castellano, E. E., Garratt, R. C., Arruda, P., Leite, A. & Craievich, A. (1998). *Nucl. Instrum. Methods A*, **405**, 159–164.
Polikarpov, I., Perles, L. A., de Oliveira, R. T., Oliva, G., Castellano, E. E., Garratt, R. C. & Craievich, A. (1998). *J. Synchrotron Rad.* **5**(2), 72–76.
Polikarpov, I., Teplyakov, A. & Oliva, G. (1997). *Acta Cryst.* D**53**, 734–737.
Rossmann, M. G. & Blow, D. M. (1962). *Acta Cryst.* **15**, 24–31.
Rytter, S. W. & Tyrrel, R. M. (2000). *Free Radic. Biol. Med.* **28**, 289–309.
Tappel, A. L. (1955). *J. Biol. Chem.* **217**, 721–733.
Vincent, S. H. (1989). *Semin. Hematol.* **26**, 105–113.

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Comparison of different crystal forms of 3-dehydroquinase from *Salmonella typhi* and its implication for the enzyme activity

Wen-Hwa Lee,[a]† Luis Augusto Perles,[b] Ronaldo A. P. Nagem,[a,b] Annette K. Shrive,[c]‡ Alastair Hawkins,[d] Lindsay Sawyer[c] and Igor Polikarpov[a,e]*

[a]Laboratório Nacional de Luz Síncrotron/LNLS, Campinas, SP, Brazil, [b]Departamento de Física, Universidade Estadual de Campinas (UNICAMP), Campinas, SP, Brazil, [c]Structural Biochemistry Group, ICMB, University of Edinburgh, Michael Swann Building, King's Buildings, Mayfield Road, Edinburgh EH9 3JR, England, [d]Department of Biochemistry and Genetics, New Medical School, Framlington Place, University of Newcastle upon Tyne, Newcastle-upon-Tyne NE2 4HH, England, and [e]Grupo de Cristalografia, Instituto de Física em São Carlos, Universidade de São Paulo, Av. Trabalhador Sãocarlense, 400, CEP 13560-970, São Carlos, SP, Brazil

† Present address: Computational Biology and Bioinformatics, Department of Molecular Biology TPC-28, The Scripps Research Institute, La Jolla, CA, USA.
‡ Current address: School of Life Sciences, Keele University, Keele, Staffordshire ST5 5BG, England.

Correspondence e-mail: ipolikarpov@if.sc.usp.br

Received 27 November 2001
Accepted 28 February 2002

PDB Reference: type I DHQase, 1l9w, r1l9wsf.

The type I 3-dehydroquinate dehydratase (DHQase) which catalyses the reversible dehydration of 3-dehydroquinic acid to 3-dehydroshikimic acid is involved in the shikimate pathway for the biosynthesis of aromatic compounds. The shikimate pathway is absent in mammals, which makes structural information about DHQase vital for the rational design of antimicrobial drugs and herbicides. The crystallographic structure of the type I DHQase from *Salmonella typhi* has now been determined for the native form at 1.78 Å resolution (R = 19.9%; R_{free} = 24.7%). The structure of the modified enzyme to which the product has been covalently bound has also been determined but in a different crystal form (2.1 Å resolution; R = 17.7%; R_{free} = 24.5%). An analysis of the three available crystal forms has provided information about the physiological dimer interface. The enzyme relies upon the closure of a lid-like loop to complete its active site. As the lid-loop tends to stay in the closed position, dimerization appears to play a role in biasing the arrangement of the loop towards its open position, thus facilitating substrate access.

1. Introduction

The enzyme 3-dehydroquinase (3-dehydroquinate dehydratase; DHQase; EC 4.2.1.10) catalyses the reversible dehydration of 3-dehydroquinic acid to 3-dehydroshikimic acid. This reaction occurs in two distinct metabolic pathways: (i) the biosynthetic shikimate pathway of aromatic compounds in microorganisms and plants (Bentley, 1990; Haslam, 1993) and (ii) the catabolic quinate pathway, a carbon-scavenging pathway common to many microbial species (Giles *et al.*, 1985).

The DHQases can be divided into two different classes according to the mechanism of action, stereochemistry, overall structure and sequence homology (White *et al.*, 1990; Servos *et al.*, 1991; Kleanthous *et al.*, 1992; Harris *et al.*, 1996). The type I enzymes, which are involved only in biosynthesis, occur either as homodimers of subunit M_r = 27 000 or as components of multifunctional enzymes with other shikimate-pathway enzymes (Lumsden & Coggins, 1977; Charles *et al.*, 1986; Bentley, 1990; Deka *et al.*, 1994). They use a covalent imine intermediate to catalyze a *syn* elimination (Butler *et al.*, 1974; Chaudhuri *et al.*, 1991). The type II enzymes are dodecamers consisting of identical subunits of M_r 16 000. They work by an entirely different mechanism and catalyze a *trans* elimination that almost certainly involves an enolate-type transition-state mechanism (Kleanthous *et al.*, 1992; Gourley *et al.*, 1994; Bottomley *et al.*, 1996; Harris *et al.*, 1996).

Recently, the structures of both a type I (from *Salmonella typhi*) and a type II (from *Mycobacterium tuberculosis*) DHQase were solved (Gourley *et al.*, 1999). The overall structure of type I DHQase from *S. typhi* consists of a single domain that folds into the commonly observed eight-stranded α/β (or TIM) barrel.

A comparative analysis of the native type I enzyme with that of the borohydride-reduced product–enzyme complex has suggested that there are critical structural changes that affect the activity of the enzyme. One of the loops of the enzyme, that between strand *h* and helix *H* (see below), acts as a lid for the active-site cleft, shielding the active site from the solvent environment.

Small differences in the crystallization conditions lead to significantly different crystal forms. Here, we present the structure solution of two new crystal forms of the type I DHQase enzyme from *S. typhi* and compare them with the original crystal form of the same enzyme (Gourley *et al.*, 1999). We have used the following arbitrary nomenclature: crystal form I (native enzyme; new; $P2_12_12$, one chain per asymmetric unit), crystal form II ($P2_1$, two chains per symmetric unit; PDB code 1qfe) and crystal form III (borohydride-reduced enzyme–product complex; Chaudhuri *et al.*, 1991; new; $P2_1$, four chains per asymmetric unit).

2. Materials and methods

Samples of DHQase from *S. typhi* were purified to homogeneity from an overproducing strain of *Escherichia coli* and subjected to a sparse-matrix crystallization screen. Several crystallization trials were performed and the best conditions were found using PEG 4000 as precipitant and 100 m*M* citrate–phosphate as buffer in the pH range 5.0–6.5 (Boys *et al.*, 1992). Soaking with different heavy atoms yielded derivatives for MIR phase determination (Gourley *et al.*, 1999).

X-ray diffraction data sets were collected at room temperature using synchrotron radiation at stations X31 and X11 at the EMBL Outstation, Hamburg and an Enraf–Nonius FR571 X-ray generator in-house. The images were collected using a MAR image-plate detector. The data were processed, indexed and merged using the programs *DENZO* and *SCALEPACK* (Otwinowski & Minor, 1997). Data-collection statistics are summarized in Table 1. The structure of crystal form II was determined by multiple isomorphous replacement as described in Gourley *et al.* (1999). The structures of crystal forms I and III were solved by molecular replacement (*AMoRe*; Navaza & Saludjian, 1997) using the structure of crystal form II as search model. The single main peaks in the rotational function had correlation coefficients of 0.31 and 0.19 for crystal forms I and III, respectively, with the next highest peaks in both forms having values of less than 0.1. These rotation solutions were used in the translation-function search and the best solutions were subjected to ten cycles of rigid-body refinement, yielding an *R* factor of 21.6% for crystal form I (88.2% correlation factor) and 27.1% for crystal form III (81.3% correlation factor). Crystallographic refinement was carried out using the maximum-likelihood method implemented in the program *REFMAC* (Murshudov *et al.*, 1997; Collaborative Computational Project, Number 4, 1994). Successive rounds of cycles of refinement interspersed with manual adjustment using the program *O* (Jones & Kjeldgaard, 1993) were employed to improve the quality of the original molecular-replacement solution. Non-crystallographic symmetry restraints were in place at this stage. The model was inspected using both $F_o - F_c$ and $2F_o - F_c$ electron-density maps. The progress of the refinement was monitored using both the conventional and free *R* factors (Brünger, 1992). The refinement rounds continued until both *R* factors reached a minimum value, with no further improvements as new rounds were requested. Water molecules were then added with the program *ARP* (Lamzin & Wilson, 1993) before continuing refinement with the program *REFMAC*.

The *R* factor of the final model of crystal form I was 17.6%, with an $R_{\rm free}$ of 22.6% for data in the resolution range 10–1.78 Å. Removal of the NCS restraints at the end of the

Table 1
Crystallographic data and statistics for DHQase.

	Crystal form I	Crystal form II	Crystal form III
Data and parameters			
Space group	$P2_12_12$	$P2_1$	$P2_1$
Unit-cell parameters			
a (Å)	48.78	60.49	42.61
b (Å)	112.33	45.39	158.56
c (Å)	42.94	85.47	85.89
β (°)	90.00	95.48	93.61
Z (No. of chains in a.u.)	1	2	4
$V_{\rm M}$ (Å^3 Da^{-1})	2.12	2.12	2.61
Solvent content† (%)	41.62	41.50	52.57
Completeness (%)	97.2	86.1	76.45
$R_{\rm merge}$ (%)	5.1	10.4	6.9
Maximum resolution (Å)	1.78	2.1	2.1
R factor (%)	19.9	17.6	17.7
$R_{\rm free}$ (%)	24.7	22.6	24.5
Redundancy	4.12	2.84	3.42
No. unique reflections	22395	23517	49700
Ramachandran plot			
Most favoured (%)	93.9	94.1	93.9
Allowed (%)	4.4	5.0	5.2
Generously allowed (%)	1.7	0.9	0.8
Disallowed (%)	0.0	0.0	0.1
RMS deviation			
Bond length (Å)	0.013	0.010	0.018
Bond angle (°)	1.7	1.5	2.4

† Assuming a protein density of 1.34 g cm^{-3}.

Table 2
Interactions occurring in the dimer interface.

		Distance (Å)			
				Crystal form III	
Residue atom	Residue atom	Crystal form I	Crystal form II	Subunits 1 and 2	Subunits 3 and 4
Lys178 NZ	Val218 O	2.89	2.85	3.03	2.93
Lys207 NZ	Ala252 O	2.32	2.54	2.99	2.61
Val218 O	Lys178 NZ	2.89	3.06	2.97	3.00
Ala252 O	Lys207 NZ	2.32	2.67	3.27	3.02
Gly235 N	Ala252 O	2.64	n.a.	n.a.	n.a.

refinement followed by a few further cycles led to the R_{free} rising slightly, thereby indicating that their removal was unjustified (Kleywegt & Jones, 1997). The model for crystal form III had an R factor of 17.7% in the resolution range 10–2.1 Å (R_{free} = 24.5%). Both new structures have acceptable quality statistics as reported by *PROCHECK* (Laskowski *et al.*, 1993) and Ramachandran plots; statistics for the final crystallographic models are shown in Table 1 together with those from form II for comparison.

3. Results and discussion

3.1. General description of the *S. typhi* type I DHQase structure

The geometry of the molecules in each of the crystal forms is good, with all but one of the residues falling at least within the generously allowed regions of the Ramachandran plots. The residue lying in a disallowed region is Lys7 from subunit 4 of crystal form III, the distortion of which is caused by the interaction of its carbonyl group with the charged side chain of Lys160 of subunit 1.

The DHQase molecule has a typical $(\alpha/\beta)_8$ (TIM barrel) structure (Fig. 1), with two short antiparallel β-strands located at the N-terminal end of the barrel that block it off. The opposite end of the barrel provides the means by which substrate can reach the active site of the enzyme. While the short loops connecting the β-strands and α-helices consist of six residues on average, the *hH* loop (residues 227–239) contains 13 residues, four of which, including Gln236 which makes direct contact with the substrate, are strictly conserved. In addition, this loop is located at the C-terminal end of the barrel, adjacent to the entrance of the active site. When substrate is in the active site, the loop is closed; in the absence of substrate, the loop swings open and appears to adopt several conformations, as suggested by the poorly defined electron density and larger average *B* factors.

A particular feature of the family of TIM-barrel enzymes is the packing of side chains in the core. Normally, the core of a TIM barrel is arranged in three layers, where each layer contains four side chains from alternate β-strands (Branden & Tooze, 1999). In the structure of DHQase, Ile44, Met112, Ile168 and Ala223 of strands *b*, *d*, *f* and *h*, respectively, form the first layer, closest to the N-terminal end of the barrel. The second, or middle, layer comprises Ile19, Leu78, Val139 and Ile201 of strands *a*, *c*, *e* and *g*, respectively. Finally, the third layer, nearest the C-terminal end of the barrel, contains the active site but is made up of the polar residues Ser21, Glu46, Arg48, Thr80, Arg82, Asp114 and Lys170 on strands *a*, *b*, *b*, *c*, *c*, *d* and *f*, respectively.

In DHQase, closure of the *hH* loop forms an additional hydrophobic layer on top of the third hydrophilic one, composed of residues Phe145, Ala172, Met205 on strands *e*, *f* and *g*, respectively, and residues Ala233 and Pro234 of the *hH* loop.

3.2. The physiological dimer and crystal form I

The main interface of the molecule responsible for specific dimer formation is present in solution as the physiological unit (Reilly *et al.*, 1994) and is also found in all three lattices. It is formed by the side chains of residues located on helices *F*, *G* and *H* of one monomer that pack against the same helices of the other monomer. The barrels are arranged side-by-side,

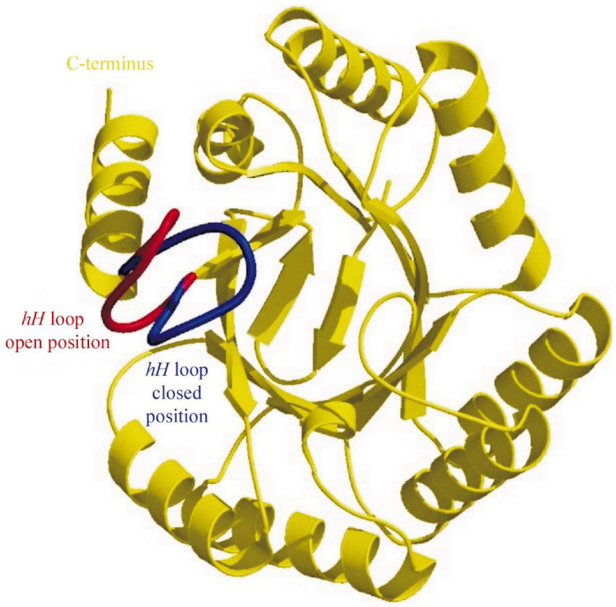


Figure 1
The overall fold of the DHQase polypeptide chain (yellow) with the lid loop depicted in the open position (red, crystal form I) and closed position (blue, crystal form II). The diagram was produced using *MOLSCRIPT* (Kraulis, 1991) and *Raster3D* (Merritt & Bacon, 1997).

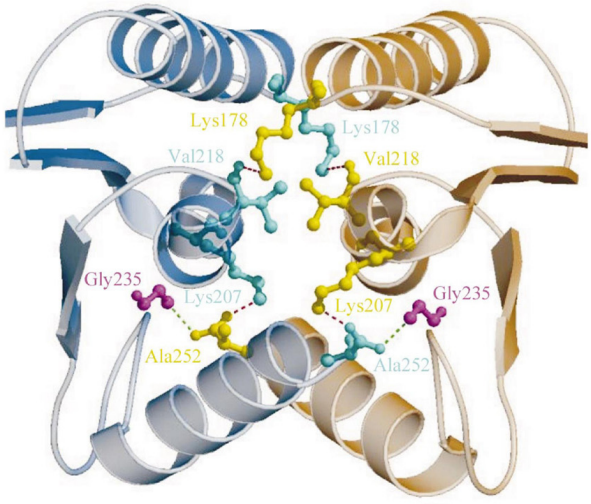


Figure 2
The physiological dimer interface. Blue and yellow shades represent elements belonging to different subunits. Magenta depicts the loop residue Gly235 interacting with Ala252.

with the active sites facing in opposite directions (Fig. 2). There are four residues involved in this packing, making only four interactions (Table 2). It is interesting to note that although the enzyme is a dimer, there are relatively few hydrogen bonds in the dimer interface. While generic dimeric interfaces have on average 0.88 ± 0.40 hydrogen bonds per 100 Å^2 buried surface area per subunit (Jones & Thornton, 1995), the dimer interface of the DHQase has only 0.34 hydrogen bonds per 100 Å^2 buried surface area per subunit. However, this is the highest value found of all the interfaces present in the three crystal forms. Moreover, the dimer-dissociation constant can be estimated from the data in Kleanthous *et al.* (1992) to be about 18 μM (or $\Delta G^o \simeq$ 25 kJ mol^{-1}), which is not an atypical value for a dimer-dissociation constant. The ΔG^o value is also consistent with the number of hydrogen bonds observed.

The native crystal, form I, has the subunit as the asymmetric unit, indicating a strict molecular as well as crystallographic twofold symmetry. However, in crystal forms II and III there is no such restriction. Superimposing the dimers from the different lattices showed no significant RMS deviation in the C$^\alpha$-atom positions. This suggests that the dimer interface is relatively stable, without significant movements between the monomers (such as a hinge movement), despite the few formal interactions that hold the two chains together. In addition, the RMS deviation of the side-chain atoms involved in the dimer interface shows there is no significant difference in their positions. Further supporting evidence of the high stability of this interface can be inferred from the solvent-accessible area buried on dimerization. The dimer interface buries ~1150 Å^2 of the accessible surface area per dimer and similar values are found in all three crystal forms. The percentage of buried surface at this interface is thus constant at ~10.6% for each lattice and lies well within the range typically found for strongly associated dimers (Jones & Thornton, 1995; Tables 2 and 3).

Table 2 shows that in crystal form I, the native enzyme, the C-terminal carboxyl group makes an additional hydrogen bond to the NH group of Gly235 in the open *hH* loop (Fig. 2). In the structures of the enzyme with labelled Lys170 in the active site (crystal forms II and III, loop closed), this Gly235–Ala252* interaction is absent. The presence of the *hH* loop in the interface in crystal form I increases the buried surface area by a small amount, reflecting the presence of the single hydrogen bond from Gly235.

3.3. The active site

Binding of the substrate to the native enzyme induces closure of the *hH* loop, breaking the Gly235–Ala252* hydrogen bond. The energy required is offset to some extent by the interactions between Ser234, Gln236 and the substrate. Several water molecules are also displaced during binding, from both the cavity and from the loop. For example, there are waters equivalent to the substrate 4- and 5-hydroxyl groups and a line of waters between Ser21 and Lys170, all of which occupy the space required by the C$_3$–C$_5$ part of the substrate

Table 3
Accessible areas of the functional dimer in three crystal forms, arranged in pairs.

The last pair (crystal form III, subunits 2 and 3) illustrates an ordinary crystal-packing interaction. Accessible areas were calculated using the program *AREAIMOL* (Collaborative Computational Project, Number 4). Structural waters were ignored in the calculations.

Crystal form	Area per subunit† (Å^2)	Area per subunit (after dimerization)‡ (Å^2)	Average buried area§ (Å^2)	Average % buried area
I (subunit 1)	11274	10094	1181.5	10.5
I (subunit 2)	11302	10119		
II (subunit 1)	10881	9698	1161	10.8
II (subunit 2)	10666	9527		
III (subunit 1)	10634	9537	1484.5	13.6
III (subunit 2)	10452	9357		
III (subunit 3)	10500	9389	1108	10.6
III (subunit 4)	10463	9358		
III (subunit 2)	10452	10115	329	3.1
III (subunit 3)	10500	10183		

† Subunit accessible areas were calculated separately from their respective pairs in the crystal forms. ‡ Accessible area calculated with the assembled dimer. § Calculated from the difference between 'area per subunit' and 'area per subunit after dimerization', hence giving the hidden area for every subunit arising from dimerization.

ring. However, there is no clearly defined solvent observed in the volume occupied by the C$_6$, C$_1$ and C$_2$ C atoms or the carboxylate of the substrate. This may be because of the largely hydrophobic nature of the side chains of Phe225 and Met203 which dominate this part of the active site. Further, the side chain of Arg213, which forms a hydrogen bond with the main-chain carbonyl group of Phe225 in the free enzyme structure, moves to form an ion pair with the carboxyl group when the substrate binds. As observed elsewhere in crystal form II (Gourley *et al.*, 1999), the substrate is exquisitely located in the active site and the final hydrogen bonds with Gln236 and Ser234 are presumably formed as the lid loop *hH* swings over to close off the active site from the solvent.

In the absence of substrate, this loop is mobile and the hydrophobic nature of the fourth layer of residues should favour the closed conformation, but access to the active site requires it to be open. Thus, it seems possible that the Gly235–Ala252* hydrogen bond in crystal form I will enhance the activity of the enzyme by counteracting the slight bias towards the closed position of the loop. This explains why in the absence of substrate we see the loop, albeit sketchily, in the open substrate-receptive position. In fact, the average *B* factors for the *hH* loop in crystal form I are significantly higher than those of crystal forms II and III (Table 4).

In the active site of crystal form III, the dehydroshikimate (product) molecule has the same environment in each of the four independent chains of the asymmetric unit. Each subunit contains one borohydride-reduced dehydroshikimate, covalently linked to residue Lys170. As in crystal form II, the substrate is coordinated by Ser21 OG, Glu46 OE2, Arg48 NH1, Arg213 NH1 and NH2, Ser232 OG and Gln236 NE2. The active sites in all crystal forms are alike, with the exception of the Met203 and Arg213 side chains in the native crystal form I. In the native enzyme, both of these side

Table 4
B factors (Å^2) of the lid-loop residues.

Standard deviations are given in parentheses after the average values.

		Crystal form II (Å^2)		Crystal form III (Å^2)			
	Crystal form I (Å^2)	Subunit 1	Subunit 2	Subunit 1	Subunit 2	Subunit 3	Subunit 4
Loop (residues 227–239)	46.9 (23.2)	15.9 (5.2)	16.2 (5.1)	38.4 (8.4)	38.5 (8.7)	39.9 (8.3)	38.2 (8.3)
Subunit overall	26.7	15.3	15.9	29.2	28.7	31.3	30.8

chains are displaced toward the exit of the active site. When the substrate is present (crystal forms II and III) both Met203 and Arg213 adopt a more 'introspective' position. This is related to the closure of the loop, which breaks the interactions of Arg213 in order to form new ones with the substrate carboxyl group. Met203 must then move back in order to accommodate the substrate carboxylate in the active site. No other differences were noticed in the positions of the main and side chains.

The placement of the substrate in the active site of cystal form III was guided by the electron density. However, the electron density for the same regions was poorer than in crystal form II. Although the electron density was continuous, its volume was larger than in form II, suggesting less precise location. In fact, the electron density was slightly stretched midway between atom C3 of the substrate and the ring of Phe145. The program *ARP* (Collaborative Computational Project, Number 4, 1994) assigned two water molecules to this position. Interestingly, the same water molecules are present in the native form, which suggests that the active site may not have been fully occupied by covalently bound substrate in this crystal form. However, in subunits 1 and 3 of form III even the *B* factors are very similar to those found in the native form (subunit 1, form III, Wat169, 43.5 Å^2, Wat208, 32.5 Å^2; subunit 3, Wat204, 40.5 Å^2, Wat211, 33.5 Å^2; form I, Wat103, 51.6 Å^2, Wat32, 32.4 Å^2), yet the product 3-dehydroshikimate is clearly visible.

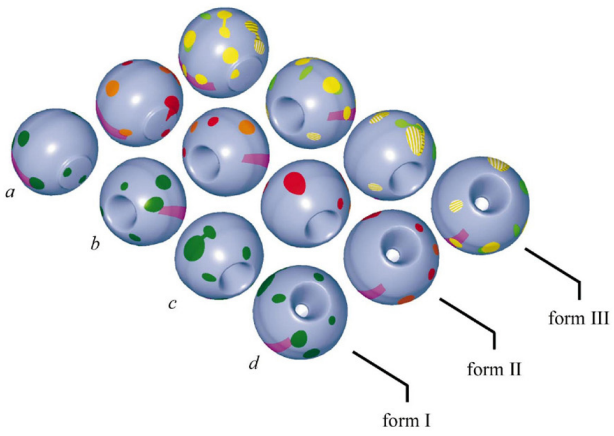


Figure 3
Schematic drawings of the different packing patterns of each crystal form. Contact areas are depicted as coloured patches colour coded for each crystal form (green for I, red for II and yellow for III). The active site is represented as a cavity and the bottom of the barrel as a flat raised disc on the sphere, clearly seen in row *a*. The magenta stripe represents the position of helix *H*. Comparison of the contact areas utilized by the three crystal forms is shown in four different orientations, *a*, *b*, *c* and *d*, but observed from the same viewpoint in each crystal form. In crystal forms II and III, where there is more than one subunit per asymmetric unit, the patches are colour coded according to the subunit to which the patch belongs. In crystal form II, subunit 1 has red contact patches, while those of subunit 2 are orange. In crystal form III the colour scheme is the following: subunit 1 (yellow with black stripes), 2 (green), 3 (yellow with white stripes) and 4 (plain yellow).

In addition to the several interactions with structural water molecules, a network of intramolecular hydrogen bonds also stabilizes the residues involved in binding the substrate. For example, Gln236 interacts with Ser232 in order to orient it towards the DHQ molecule. There are also hydrophobic interactions among the C atoms of the DHQ ring and several residues, two of which are implicated in the fourth hydrophobic layer (Met203 and Ala233, see above).

Finally, there is a significant increase in the stability of the enzyme when the substrate is bound (Kleanthous *et al.*, 1991): the concentration of GuHCl required to unfold the native enzyme is about 1.5 *M* and this increases to 4 *M* when the substrate or product is bound. Exclusion of water from the active site together with formation of the fourth hydrophobic layer locks the core of the structure, making it resistant both to chaotropic agents and to thermal denaturation.

3.4. Crystal form III

Two physiological dimers constitute the asymmetric unit of the crystal form III. These two dimers interact solely through subunits 2 and 3, making a structure similar to the Greek capital letter lambda (Λ). There are two hydrogen bonds between these two subunits, namely 2-Arg38 NH1 to 3-Asn135 OD1 (2.34 Å) and 2-Ala71 N to 3-Ala133 O (3.19 Å). In addition, there is an interaction mediated by a water molecule, bridging residues 2-Glu39 OE1–Wat88–3-Asp167 OD1 (2.78 and 3.28 Å). One face of subunit 2 of the 1–2 dimer interacts with the edge of a face of subunit 3 of the 3–4 dimer. In fact, the C-terminal ends of helices *A* (Arg38) and *B* (Ala71) in subunit 2 interact with the end of helix *D* (Ala133 and Asn135) in subunit 3. This interface buries an area of some 327 Å^2 which is well below the value expected for a physiologically relevant interface (Jones & Thornton, 1995).

Inspection of the 'lambda' interface and superposition of several chains from different crystal forms reveal that the side chains do not differ significantly in conformation, except for the side chain of Arg38 of subunit 2. In all other subunits, this side chain points straight out into the solvent. In crystal form III, residue 2-Arg38 assumes a relatively well defined conformation (*B* factor = 37.5 Å^2, compared with higher values for the residues *n* either side) that allows formation of the two hydrogen bonds that bind the subunits 2 and 3 together. None of the residues involved in the lambda interface is conserved except Asp167, suggesting that these interactions occur merely as a result of the crystallization conditions.

Table 5
Residues involved in crystallographic hydrogen-bonding contacts.

Residues making contact in the all three crystal forms are listed in the first column. The following columns specify the subunit and the residue(s) that are involved in the crystallographic contact(s) in each crystal form. The numbers in parentheses after residues identify the subunit to which they belong.

Residue	Form I (17)†		Form II (16)†		Form III (18)†	
	Subunit	Contacting	Subunit	Contacting	Subunit	Contacting
Arg38	1	Asp121	2	Ala252	2, 4	Asn135 (2), Asp167 (4)
Gln59	1	Asn8	1	Arg66	2	Ala133
Asp123	1	Lys2	1	Lys2	2	Thr93

† The values in parentheses are the total number of interacting residues in each crystal form.

3.5. Crystallographic contacts

Each crystal form presents several distinct crystallographic contacts, few of which are common to the three crystal forms. However, in each case different residues combine to form different patches that make the crystallographic contacts. There are few of these patches and in general they comprise at most three residues. The main patch is that forming the physiological dimer and is only present as a crystallographic contact in crystal form I. A schematic drawing of the areas of contact for each of the crystal forms is presented in Fig. 3, while Table 5 summarizes the crystallographic contacts made by residues Arg38, Gln59 and Asp123 in all three crystal forms. It is clear that these three residues do not form a patch common to all forms. The number of residues involved in crystallographic interactions of each of the three forms is similar (17 in form I, 16 in form II and 18 in form III) and often the contact is made through a single residue. Further, despite the similarity of the unit-cell lengths ($a^{\rm I} \simeq c^{\rm I} \simeq b^{\rm II} \simeq a^{\rm III}$, $2b^{\rm II} \simeq c^{\rm II}$; $2a^{\rm III} \simeq {\rm c}^{\rm III}$, $c^{\rm II} \simeq c^{\rm III}$), which is presumably a reflection of the approximately cylindrical nature of the dimers, the arrangement of molecules is not related. Thus, for example, the *bc* face of form II bears little resemblance to the *ac* face of form III, a consequence of the widespread distribution of potential crystal contacts all over the molecular surface.

The lack of a consistent pattern of interacting patches or residues could explain the significant polymorphism that has hindered crystallographic studies. The borohydride-reduced Schiff-base intermediate of the enzyme that was eventually solved was expected to lock the structure and produce a unique crystal form, but it is in fact only the native form that makes some use of an active site-related crystal contact and involves the base of the lid loop *hH*: the side chains of residues at the base of the loop (Lys229 and Asn240) interact with a crystallographic neighbour.

4. Conclusions

Careful analysis of the different crystal forms shows that there are no very specific patches on the molecule surface that lead to a unique crystal packing. The effect of this is that there are several possible packing arrangements which can form from essentially the same conditions.

The environment of the substrate–product molecule is one that provides high specificity and is unaffected by the different crystal forms. Despite the large number of specific interactions between substrate and enzyme, slight variations in position do appear to be possible. However, it is not clear from this study that these minor differences have any functional significance.

Careful comparison of the three different crystal forms suggests that the functional dimer interface is implicated in the enzyme mechanism. It appears that the open position of the *hH* loop is stabilized by the dimer interface, thus facilitating access to the active site by the substrate. The stabilization, in the manner of a weak catch, is achieved through a hydrogen bond formed between the C-terminal residue from one subunit, Ala252, and the conserved Gly235 residue in the loop. When substrate binds, new hydrogen bonds as well as several hydrophobic interactions (the fourth hydrophobic layer) are formed, completing a solvent-shielding layer that allows the reaction to proceed efficiently.

We are grateful to Professor John Coggins for helpful discussions. The EMBL Outstation at Hamburg is thanked for providing data-collection facilities on stations X11 and X31. The financial assistance of the Brazilian funding bodies CNPq, FAPESP (*via* project 99/03387-4) and CAPES and of the Biotechnology and Biological Science Research Council is gratefully acknowledged.

References

Bentley, R. (1990). *Crit. Rev. Biochem. Mol. Biol.* **25**, 307–384.
Bottomley, J. R., Hawkins, A. R. & Kleanthous, C. (1996). *Biochem. J.* **319**, 269–278.
Boys, C. W. G., Bury, W. M., Sawyer, L., Moore, J. D., Charles, I. G., Hawkins, A. R., Deka, R., Kleanthous, C. & Coggins, J. R. (1992). *J. Mol. Biol.* **227**, 352–355.
Branden, C. & Tooze, J. (1999). *Introduction to Protein Structure*, 2nd ed. New York/London: Garland Publishing Inc.
Brünger, A. T. (1992). *Nature (London)*, **355**, 472–474.
Butler, J. R., Alworth, W. L. & Nugent, M. J. (1974). *J. Am. Chem. Soc.* **96**, 1617–1618.
Charles, I. J., Keyte, J. W., Brammar, W. J., Smith, M. & Hawkins, A. R. (1986). *Nucleic Acids Res.* **14**, 2201–2213.
Chaudhuri, S., Duncan, K., Graham, L. D. & Coggins, J. R. (1991). *Biochem. J.* **275**, 1–6.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Deka, R. K., Anton, I. A., Dunbar, B. & Coggins, J. R. (1994). *FEBS Lett.* **349**, 397–402.
Giles, N. H., Case, M. E., Baum, J. A., Geever, R. F., Huiet, L., Patel, V. B. & Tyler, B. M. (1985). *Microbiol. Rev.* **49**, 338–358.
Gourley, D. G., Coggins, J. R., Isaacs, N. W., Moore, J. D., Charles, I. G. & Hawkins, A. R. (1994). *J. Mol. Biol.* **241**, 488–491.
Gourley, D. G., Shrive, A. K., Polikarpov, I., Krell, T., Coggins, J. R., Hawkins, A. R., Isaacs, N. W. & Sawyer, L. (1999). *Nature Struct. Biol.* **6**, 521–525.

Harris, J. M., Gonzalez-Bello, C., Kleanthous, C., Hawkins, A. R., Coggins, J. R. & Abell, C. (1996). *Biochem. J.* **319**, 333–336.
Haslam, E. (1993). *Shikimic Acid: Metabolism and Metabolites.* Chichester: J. Wiley & Sons.
Jones, S. & Thornton, J. M. (1995). *Prog. Biophys. Mol. Biol.* **63**, 31–65.
Jones, T. A. & Kjeldgaard, M. (1993). *O Version* 5.9, *The Manual.* Uppsala University, Sweden.
Kleanthous, C., Deka, R., Davis, K., Kelly, S. M., Cooper, A., Harding, S. E., Price, N. C., Hawkins, A. R. & Coggins, J. R. (1992). *Biochem. J.* **282**, 687–695.
Kleanthous, C., Reilly, M., Cooper, A., Kelly, S., Price, N. C. & Coggins, J. R. (1991). *J. Biol. Chem.* **266**, 10893–10898.
Kleywegt, G. J. & Jones, T. A. (1997). *Methods Enzymol.* **277**, 208–230.
Kraulis, J. (1991). *J. Appl. Cryst.* **24**, 946–950.
Lamzin, V. S. & Wilson, K. S. (1993). *Acta Cryst.* D**49**, 129–147.
Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.
Lumsden, J. & Coggins, J. R. (1977). *Biochem. J.* **161**, 599–607.
Merritt, E. A. & Bacon, D. J. (1997). *Methods Enzymol.* **277**, 505–524.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Navaza, J. & Saludjian, P. (1997). *Methods Enzymol.* **276**, 581–594.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Reilly, A., Morgan, P., Davis, K., Kelly, S. M., Greene, J., Rowe, A. J., Harding, S. E., Price, N. C., Coggins, J. R. & Kleanthous, C. (1994). *J. Biol. Chem.* **269**, 5523–5526.
Servos, S., Chatfield, S., Hone, D., Levine, M., Dimitriadis, G., Pickard, D., Dougan, G., Fairweather, N. & Charles, I. (1991). *J. Gen. Microbiol.* **137**, 147–152.
White, P. J., Young, J., Hunter, I. S., Nimmo, H. G. & Coggins, J. R. (1990). *Biochem. J.* **265**, 735–738.