

# Development of X-Ray Powder Diffraction Methods for Biomolecules

THÈSE N° 5138 (2011)

PRÉSENTÉE LE 9 SEPTEMBRE 2011  
À LA FACULTÉ SCIENCES DE BASE  
LABORATOIRE DE CRISTALLOGRAPHIE  
PROGRAMME DOCTORAL EN CHIMIE ET GÉNIE CHIMIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

**Sebastian BASSO**

acceptée sur proposition du jury:

Dr S. Gerber, présidente du jury  
Prof. M. Schiltz, directeur de thèse  
Dr R. Cerny, rapporteur  
Prof. A. Fitch, rapporteur  
Prof. P. Leiman, rapporteur



ÉCOLE POLYTECHNIQUE  
FÉDÉRALE DE LAUSANNE

Suisse  
2011



## Abstract

The comparable order of magnitude between interatomic distances in a crystal and the wavelength of X-rays make X-ray crystallography the ideal analytical tool to gain insight into the structure of crystalline material, including biomolecules. Nevertheless, biomolecular crystallography has until now relied on the successful growth of single crystals of suitable size and quality. These remain the exception rather than the rule, since biomolecules often produce polycrystalline precipitate instead. Yet, an interest in making use of the once-discarded polycrystalline material, through the technique of powder diffraction, has only recently emerged. This can be accounted for by the information deficit which powder diffraction data suffers from in comparison with that of single crystal. The paucity of information in powder data stems from the compression of the three-dimensional reciprocal space onto the one-dimension of a powder pattern. In spite of this, powder diffraction holds the potential for application in biomolecular crystallography as is shown in the two different studies presented herein. Both studies were carried out with methods which do not rely on employing previously determined crystal-structures as molecular models. This therefore allowed the objective assessment of the quality of information that powder diffraction data can contribute to the structural investigation of biomolecules.

In the first project, the traditional single-crystal structure-solution process is applied to data extracted from protein powder diffraction patterns measured on a synchrotron source. The use of models is avoided by employing the *de novo* phasing method of isomorphous replacement. With two protein test-cases, namely hen egg white lysozyme and porcine pancreatic elastase, it is demonstrated that protein powder diffraction data can afford structural information up to medium resolution. Indeed, a single isomorphous replacement analysis generated molecular envelopes accurately describing the crystal packing of both protein systems, while a multiple isomorphous replacement experiment, carried out only on lysozyme, revealed an electron density map in which elements of the secondary structure could be located. In fact, the resolution of the latter was discovered to be sufficient to determine the chirality of the protein molecule it represented. In addition to being encouraging, these results do not reflect the full potential of biomolecular powder diffraction, due, in large part, to the ultimately unsuitable nature of one type of phasing method, single crystal, being applied to another type of data, powder.

An alternative approach to extract information from protein powder diffraction data is to employ powder-specific structure-solution techniques, such as global optimization methods. Although these methods make use of a starting ‘model’, it is a molecular description of the system under study based on known chemical quantities rather than a related molecular configuration based on a previously determined crystal structure. Since their conception, global optimization methods have continuously been developed to enable the tackling of increasingly complex crystal structures. However, the immense complexity of biomacromolecules has kept proteins well out of reach of such methods. In an attempt to further reduce the gap separating the two levels of complexity, the second study reported herein puts forth the implementation of Ramachandran plot restraints into the algorithm of a global optimization method, *i.e.* that of simulated annealing. More specifically, the Ramachandran plot was approximated using a two-dimensional Fourier series which was subsequently expressed as a penalty function and incorporated into the search algorithm.

**Keywords:** powder diffraction; protein; secondary structure; isomorphous replacement; phasing; peptide; global optimization methods; Ramachandran plot; rigid body.

## Resumé

L'ordre de grandeur des distances interatomiques dans un cristal étant comparable à celui des longueurs d'onde des rayons X, la diffraction de rayons X est devenu l'outil d'analyse idéal pour obtenir une connaissance de la structure de la matière cristalline, biomolécules incluses. Toutefois, la cristallographie biomoléculaire s'est jusqu'ici appuyée sur la réussite de la préparation de monocristaux de taille et de qualité suffisantes. Ces derniers sont encore l'exception plutôt que la norme, car les biomolécules produisent souvent un précipité polycristallin. Pourtant, ce n'est que récemment qu'un intérêt a été porté à l'utilisation de ce matériel polycristallin, autrefois écarté, au travers de la technique de diffraction par les poudres. Ceci peut s'expliquer par la perte d'information dont souffrent les données de diffraction par les poudres au regard de celles obtenues d'un monocristal. La carence d'information dans des données de diffraction par les poudres provient de la compression de l'espace réciproque tridimensionnel sur un diagramme de diffraction de poudres à une dimension. Malgré cela, la diffraction par les poudres trouve des applications potentielles en cristallographie biomoléculaire, illustrées ici par la réalisation de deux études distinctes. Toutes deux ont été réalisées en utilisant des méthodes ne s'appuyant pas sur l'emploi de structures cristallines en tant que modèles moléculaires prédéterminés. Ceci a donc permis d'évaluer d'une façon objective la qualité d'information que des données de poudres seraient en mesure de fournir à la recherche sur la structure des biomolécules.

Dans un premier temps, le procédé classique de résolution de structures à partir d'un monocristal est appliqué aux données extraites de diagrammes de diffraction par les poudres mesurés à l'aide d'une source de rayonnement synchrotron. L'utilisation de modèles est évitée en employant la méthode *de novo* de phasage qu'est le remplacement isomorphe. Démonstration est ici faite, à partir de deux protéines de référence que sont le lysozyme de blanc d'œuf de poule et l'élastase pancréatique porcine, que les données de diffraction par les poudres peuvent offrir une information structurelle jusqu'à moyenne résolution. En effet, une analyse par remplacement isomorphe simple a permis de générer des enveloppes moléculaires décrivant convenablement l'empilement moléculaire des deux protéines, alors qu'une expérience de remplacement isomorphe multiple, pratiquée sur le lysozyme uniquement, a révélé une carte de densité électronique dans laquelle des éléments de la structure secondaire ont pu être mis en évidence. Par ailleurs, il a été montré que la

résolution de cette carte est suffisante pour déterminer la chiralité de la protéine qu'elle représente. En plus d'être encourageants, ces résultats ne reflètent pas pleinement le potentiel de la diffraction par les poudres de biomolécules, à cause, en grande partie, de la nature au final inadaptée de l'application d'un type de méthode de phasage, celle pour les monocristaux, à un type de données différent, celles provenant des poudres.

Une approche alternative pour extraire de l'information des données de poudres de protéine consiste à utiliser des techniques de résolution de structures spécifiques aux poudres, telles que des méthodes d'optimisation globale. Bien que ces méthodes utilisent un modèle de départ, il s'agit d'une description moléculaire basée sur des quantités chimiques connues plutôt que d'une configuration moléculaire basée sur une structure cristalline déterminée antérieurement. Depuis leur conception, les méthodes d'optimisation globale se sont largement développées pour permettre d'aborder des structures cristallines de plus en plus complexes. Cependant, l'immense complexité des biomacromolécules a tenu les protéines hors de portée de ce type de méthodes. Essayant de réduire d'avantage l'écart séparant ces deux niveaux de complexité, la seconde étude reportée ici est basée sur l'implémentation de contraintes sur le diagramme de Ramachandran à l'intérieur de l'algorithme d'une méthode d'optimisation globale, *i.e.* celle du recuit simulé. Plus spécifiquement, le diagramme de Ramachandran a été approximé en utilisant une série de Fourier à deux dimensions qui a ensuite été exprimé en tant que fonction de pénalité et incorporé dans l'algorithme de recherche.

**Mots-clés** : diffraction par les poudres, protéine(s), structure secondaire, remplacement moléculaire, phasage, peptide(s), méthodes d'optimisation globale, diagramme de Ramachandran, corps rigide.

# Contents

<b>Acronyms</b>	<b>vii</b>
<b>1 Introduction - Biomacromolecular crystallography</b>	<b>1</b>
<b>2 De novo phasing of macromolecules with X-ray powder diffraction data</b>	<b>9</b>
2.1 Introduction . . . . .	9
2.1.1 Phasing methods . . . . .	12
2.1.1.1 Anomalous scattering . . . . .	13
2.1.1.2 Isomorphous replacement . . . . .	14
2.2 Experimental . . . . .	18
2.2.1 Protein crystallization . . . . .	18
2.2.1.1 Heavy-atom derivatization . . . . .	21
2.2.2 Protein systems . . . . .	22
2.2.2.1 Hen egg white lysozyme . . . . .	22
2.2.2.2 Porcine pancreatic elastase . . . . .	24
2.2.3 X-ray data collection . . . . .	25
2.3 Results . . . . .	28
2.3.1 Anisotropic lattice changes . . . . .	28
2.3.2 Intensity extraction . . . . .	33
2.3.3 Heavy-atom phasing . . . . .	37
2.3.4 Density modification . . . . .	43
2.3.5 Detection of secondary-structure elements . . . . .	47
2.3.6 Handedness of protein molecules . . . . .	47
2.4 Discussion . . . . .	49
2.5 Outlook . . . . .	51

---

<b>3 Implementation of Ramachandran plot restraints in a global optimization algorithm</b>	<b>53</b>
3.1 Introduction . . . . .	53
3.2 Global optimization methods . . . . .	55
3.2.1 Simulated annealing . . . . .	57
3.3 Rigid bodies . . . . .	58
3.3.1 Amino acids . . . . .	61
3.4 The Ramachandran plot . . . . .	65
3.4.1 2D Fourier series approximation . . . . .	68
3.5 Peptide systems . . . . .	76
3.5.1 Fmoc-Ala-Ala-Ala . . . . .	76
3.5.2 Phe-Gly-Gly-Phe . . . . .	78
3.5.3 Tyr-Ala-Gly-Phe-Leu . . . . .	79
3.5.4 ACTH-(4-10) . . . . .	80
3.5.4.1 Crystallization . . . . .	81
3.6 Preliminary results . . . . .	83
3.7 Outlook . . . . .	85
<b>4 Conclusion</b>	<b>89</b>
<b>A X-ray diffraction</b>	<b>91</b>
A.1 Fundamental principles . . . . .	91
A.2 Powder diffraction . . . . .	92
A.3 Synchrotron radiation . . . . .	93
<b>B Rigid body - definition of <math>\varphi, \psi</math> &amp; <math>\omega</math></b>	<b>95</b>
<b>C Least squares approximation</b>	<b>97</b>
<b>Bibliography</b>	<b>101</b>
<b>Acknowledgements</b>	<b>119</b>
<b>CURRICULUM VITÆ</b>	<b>121</b>

# Acronyms

**DNA** deoxyribonucleic acid. 2

**ESRF** European Synchrotron Radiation Facility. 10

**FWHM** full width at half maximum. 23

**HEWL** hen egg white lysozyme. 20

**MAD** multiple-wavelength anomalous diffraction. 12

**MC** Monte Carlo. 46

**MIPPIE** multi-pattern Pawley intensity extraction. 28

**MIR** multiple isomorphous replacement. 13

**MR** molecular replacement. 48

**PPE** porcine pancreatic elastase. 21

**R-PLOT** Ramachandran plot. 52

**SA** simulated annealing. 46

**SDPD** structure determination from powder diffraction data. 44

**SIR** single isomorphous replacement. 13

**S.U.** standard uncertainties. 29



# Chapter 1

## Introduction - Biomacromolecular crystallography

Although the study of crystals as symmetrical objects dates back to as early as the 17<sup>th</sup> century (Kepler 1611), the investigation of the three-dimensional atomic architecture of crystalline materials did not emerge prior to the discovery of X-rays. This short wavelength electromagnetic radiation was discovered by Röntgen (1895) and, seventeen years later, Max von Laue suggested and showed that it could be exploited in diffraction studies of crystals (Friedrich et al. 1912). Soon thereafter, the crystal structures of a collection of simple inorganic salts, including NaCl, were solved using X-ray radiation by W. Lawrence Bragg (Bragg 1913). The following decade saw the structure determination of small organic compounds, such as hexamethylbenzene (Lonsdale 1928), as well as the start of macromolecular crystallography. Indeed, in 1926, the crystallization of the first enzyme, namely urease, was reported by Sumner (1926). Half a decade later, the diffraction patterns of a series of natural fibrous proteins, including silk and keratin, were shown to be roughly classifiable into two categories referred to as  $\alpha$  and  $\beta$  (Astbury 1933). In the succeeding year, Bernal & Crowfoot (1934) reported the first single-crystal X-ray photographs of a globular protein, namely pepsin.

However, the first structural results of an X-ray experiment on a protein molecule did not come until after the end of World War II. In the late 1940s, the ambitious work of Max Perutz on haemoglobin gave rise to the suggestion that the globular protein was composed of bundles of  $\alpha$ -fibrous-like rods (Boyes-Watson et al. 1947, Perutz 1949). Shortly after (Pauling & Corey 1951, Pauling et al. 1951), Linus Pauling showed, with the use of helical polypeptide models, that a helix composed of 3.6 residues per turn, the  $\alpha$ -helix, roughly explained the fibrous  $\alpha$ -diffraction patterns obtained by Astbury and perhaps

also the rods observed in haemoglobin. The  $\alpha$ -helix was contemporaneously confirmed to be present in haemoglobin by Perutz (Perutz 1951*a*, Perutz 1951*b*). A subsequent study on helical structures was reported in the following year (Cochran et al. 1952) and paved the way to the famous discovery of the double-helical structure of deoxyribonucleic acid (DNA) by Watson & Crick (1953), supported by diffraction photographs recorded by Franklin & Gosling (1953).

As part of a perpetual effort to obtain structural information on protein crystal structures at increasingly higher resolutions, the 1950s and '60s produced a series of structural reports on two globular proteins, *i.e.* haemoglobin and myoglobin. An initial breakthrough came with the first application of the isomorphous replacement method on a protein molecule, namely haemoglobin (Green et al. 1954). With the subsequent development of this method, especially the multiple isomorphous replacement variant, the structures of both globular macromolecules were determined at medium resolution: 6 Å for myoglobin (Kendrew et al. 1958) and 5.5 Å for haemoglobin (Perutz et al. 1960). However, due to the smaller molecular size of the former compared to the latter, myoglobin was the first protein of which the structure was solved at near-atomic resolution (Kendrew et al. 1960). Indeed, the complexity of the structure-solution problem for haemoglobin was such that an atomic model was only determined eight years later (Perutz et al. 1968). In the meantime, the structure of only one additional macromolecule, namely lysozyme, was solved (Blake et al. 1965). In the years following this pioneering work, the rate at which protein structures were being determined increased astonishingly rapidly. This can be seen by looking at the cumulative number of structures deposited in the Protein Data Bank during the first thirty years of its existence: 13 in 1976, 213 in 1986 and 4992 in 1996<sup>1</sup>.

The accelerated progression in the number of structurally solved macromolecules was in large part due to the technical advances made in the second half of the 20<sup>th</sup> century. Among such advances, is the development of computers. Although Kendrew et al. (1960) determined the structure of myoglobin using a computer built onsite, namely at the University of Cambridge, commercially available mainframe computers started to emerge around the same time. Over the years, the technology of computers brought with it a continuous increase in calculating power, to tackle gradually more complex problems, in quality of graphics, for model building purposes, and in programmability, to develop the

---

<sup>1</sup><http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total&seqid=100> (valid in May 2011)

---

necessary tools for determining a given crystal structure.

A second, and perhaps the most important, technical contributor to the field of protein crystallography was synchrotron radiation. Synchrotrons are circular accelerators of elementary particles which came into existence in the 1970s. They were originally built for high-energy particle physics experiments designed to discover new subatomic particles formed during collisions between the accelerated particles. However, when the full potential associated with the electromagnetic radiation generated as a byproduct of the acceleration of charged particles was realized, a wave of second-generation synchrotrons were constructed for the sole purpose of generating and exploiting X-ray radiation. At the time when synchrotrons arrived to the scene, a measurement of a protein crystal could take up to weeks on a standard laboratory X-ray diffractometer. The comparatively high intensity of synchrotron radiation reduced that time down to hours. A further advantage of a synchrotron source is the tunability of its wavelength. The availability of this feature probably represents the most important turning point in macromolecular crystallography as it provided a practical means to carry out multiwavelength anomalous diffraction experiments. Although the possibility of exploiting measurements at multiple wavelengths for phasing purposes was recognized as early as the 1950s (Okaya & Pepinsky 1956, Mitchell 1957), a laboratory diffractometer only offered a very limited set of wavelengths, dictated by the fluorescence emission lines of the metal anode in the X-ray tube. The importance of this phasing method came with the advantages it holds with respect to its traditional alternative, *i.e.* isomorphous replacement. These mainly stem from the fact that all necessary data are measured on only one single crystal, thereby substantially simplifying the sample preparation procedures as well as enhancing the accuracy of the recovered phase angles as a consequence of perfect isomorphism between the data sets. Nevertheless, the theoretical foundations (Karle 1980) and the practical algorithm (Hendrickson 1985) of the multiwavelength anomalous diffraction method were not fully developed until the 1980s. Subsequently, the initial reluctance of macromolecular crystallographers to move away from the established method of isomorphous replacement, partially contributed to the multiwavelength anomalous diffraction method only being employed routinely, and as the method of choice, in the second half of the 1990s (Hendrickson 1999).

The other cause responsible for the scarce use of the method in its infancy, was the lack of means to properly apply it. Yet, the advent of synchrotron facilities brought about a series of ‘forced’ technological advances, in response to the strength of their X-ray beams, from which the method of multiwavelength anomalous diffraction greatly

benefited. Among these is the development of detectors. Up until the end of the 1980s, all X-ray diffraction data were still recorded on photographic film. However, with the reduction in required exposure time to obtain a data set at a synchrotron source, detectors needed to perform at greater speeds. This need was, over the years, satisfied by the development of image plate detectors, charge-coupled devices and, recently, pixel array detectors (Gruner 2010). This evolution in detector technology was necessarily accompanied by an improved signal-to-noise ratio. The improvement in quality of the measured intensity was instrumental in the use of phasing methods to accurately exploit the intensity differences between data sets. This is especially true for the method of multiwavelength anomalous diffraction given that the intensity differences generated as a result of probing the anomalous scattering phenomenon are relatively weak.

An additional technological advance, part of the aforementioned series, is the development of cryopreservation techniques. The main reason to collect data at cryogenic temperatures ( $\sim 100$  K) is to reduce the rate of radiation damage associated with the high flux of synchrotron sources (Garman 2010). Since protein crystals are partially composed of water-based solvent which is susceptible to form ice upon cooling, various strategies have been devised to achieve such low temperatures while avoiding ice formation (Rodgers 1994). Of these, the method of flash-cooling has proven to be most effective; it consists in rapidly cooling the crystal down to reach a level of viscosity of the solvent high enough for it to turn into a glass state rather than into ice (Pflugrath 2004). However, ice formation can persist regardless of the cooling speed and it has thus become common practice to use cryoprotectants, such as ethylene glycol and glycerol, during the crystallization procedure to synthetically increase the viscosity of the solvent within the crystal lattice. Advantages to cryocooling crystals aside from reducing radiation damage, which in turn allows for longer exposure times, include limiting the thermal disorder, all of which can lead to the extension of the attainable angular resolution. It is thus apparent to see how cryogenic cooling contributed to the progress of macromolecular crystallography in general and in particular to the multiwavelength anomalous diffraction experiments since it requires a single crystal to survive long enough in the beam for multiple data sets to be measured on that one crystal. Nevertheless, the process of cooling can in some instances damage the crystal and, as a result, the development of cryocooling techniques remains an active field of research (Alcorn & Juers 2010).

Up until the 1970s, only natural macromolecules available in appropriate abundance were accessible to be studied crystallographically. However, since then, advances made in the technology of recombinant DNA allowed many proteins, especially prokaryotic ones,

---

to be artificially produced in high yields. Once again, this technical development contributed significantly to the field as a whole through the dramatic effect it had on the success of the multiwavelength anomalous diffraction method. The effect on the phasing method stems from the fact that the recombinant technology can be employed to synthetically incorporate into proteins relatively strong anomalously scattering atoms through the use of methionine analogs, *i.e.* selenomethionine (Hendrickson et al. 1990) and telluromethionine (Budisa et al. 1997). More specifically, the sulfur atom in natural methionine, which is associated with a relatively weak anomalous scattering signal and has only been exploited in special cases (Hendrickson & Teeter 1981), is replaced by a stronger anomalously scattering atom. This procedure thus allowed the many proteins that did not naturally contain a strong anomalously scattering atom to be structurally investigated using the method of multiwavelength anomalous diffraction.

Yet another area of protein crystallography in which considerable progress was made is that of crystallization itself. In single-crystal macromolecular diffraction studies, the most crucial step is the production of single crystals of sufficient size and quality. As a result, numerous crystallization methods have been developed over the years (§2.2.1). In spite of these, the crystallization step remains one of the major bottlenecks as it often proves to be a time-consuming trial-and-error process. However, the effort required to optimize single-crystal growth conditions has, to some extent, been reduced with the advent of dedicated microfocus beamlines at 3<sup>rd</sup> generation synchrotrons, on which crystals of sizes of the order of 10  $\mu\text{m}$  can be measured successfully (Evans et al. 2011). Nevertheless, these types of measurements represent a significant technical challenge, in particular with respect to radiation damage. Indeed, the corresponding intensity of these highly focused X-ray beams, inflicted on such small crystals, causes substantial radiation damage even if cryogenic cooling is employed. Consequently, data from multiple microcrystals is often necessary to collect a complete data set.

On the other hand, the preparation of polycrystalline samples is much less demanding. In fact, many failed attempts to crystallize suitable single crystals produce powder samples which are often simply discarded. Furthermore, such samples can be obtained for several proteins by batch precipitation (Von Dreele 2003). The size of the individual crystallites constituting such precipitates is generally too small for single-crystal diffraction experiments while optimal for powder-diffraction experiments. Indeed, in the protein powder diffraction studies reported so far (see §2.1), the estimated crystallite size was of the order of 0.1–1  $\mu\text{m}$ . In a powder diffraction measurement, the recorded signal is generated by many thousands of simultaneously irradiated crystallites and hence the

effects of radiation damage are distributed over a much larger sample volume than would be the case for a single microcrystal.

In view of the protein molecules which resist forming suitable single crystals, the exploitation of the X-ray powder diffraction technique widens the spectrum of protein molecules which could be crystallographically investigated. However, powder diffraction studies suffer from a loss of information due to the compression of the three dimensions of the reciprocal space onto the one dimension of a powder diffraction pattern. The collapse of dimensions leads to significant peak overlap which accounts for the information loss since the recorded quantity, *i.e.* the intensity of individual Bragg reflections, can no longer be accurately measured. This is an especially serious problem for protein powder samples, since the large unit cells in which protein molecules crystallize lead to a high density of peaks in the resulting powder pattern and, in turn, to an enhanced degree of peak overlap. The size of their unit cells, as well as their high solvent content, is also responsible for the often low intensity of diffraction associated with protein crystals. As a result of these intrinsic issues, the use of synchrotron radiation is almost a necessity. Indeed, the intense brilliance of such radiation ensures a measurable diffraction signal while its highly collimated characteristic allows for the measurement of narrow diffraction peaks which are less likely to overlap.

Moreover, powder diffraction data afford a variety of complimentary information with respect to single-crystal measurements, for instance, the shape of the diffraction peaks depends on the microstructure of the material under study and accurate lattice parameters can be determined without difficulty. A further advantage of the powder diffraction is the persistence of the samples to form and survive under more varied conditions than a corresponding single crystal. This property of polycrystalline powders has enabled the use of the powder technique to investigate the phase diagrams of various protein molecules (Basso et al. 2005, Norrman et al. 2006, Collings et al. 2010). Such studies are particularly important for polycrystalline biopharmaceuticals (Basu et al. 2004) since polymorphism and crystal packing govern crystal solubility and hence the drug-release rate into the human body. Even though insulin remains to date the only protein drug available in a crystalline form on the market, the cited examples do exhibit the potential of protein powders as materials in their own right. In addition, their application is not restricted to medicine, as their chemical and physical properties have been shown to be useful across a broad range of disciplines (Margolin & Navia 2001).

In light of the above, it is apparent that the technique of powder diffraction will have a part to play in the future of macromolecular crystallography. However, in order

to ascertain the importance of its future role, it is necessary to objectively qualify the information that powder data can contribute to the structural investigation of protein molecules. It is with regards to this aim that the two projects constituting the work presented herein have been carried out. In order to achieve this aim, both projects involve employing methods which do not rely on the use of previously determined crystal-structures as molecular models. More specifically, in the first project the traditional single-crystal structure-solution process is applied to protein powder diffraction data in an isomorphous replacement experiment. This study thus quantifiably offers a sense of how much information is in fact lost in the aforementioned collapse in dimensions. In the second project, on the other hand, Ramachandran plot restraints are incorporated into a simulated annealing protocol to improve the success rate of the structure-solution process of oligopeptides using powder-diffraction data. Since structural investigation of short polypeptide chains can prove vital in gaining insight into the structural properties of those chains within protein molecules, the development of methods to study such peptides will be instrumental in the determination of the relationship between sequence and structure of a protein.



## Chapter 2

# De novo phasing of macromolecules with X-ray powder diffraction data

### 2.1 Introduction

As an independent field of study, protein powder diffraction is considerably young. Yet, the first reported protein powder diffraction patterns date back to the early parts of the twentieth century. Indeed, such patterns were measured on the macromolecules chymotrypsinogen (Corey & Wyckoff 1936) and tobacco mosaic virus (Wyckoff & Corey 1936). Subsequent studies went a step further and indexed the powder patterns to afford unit-cell parameters of various plant viruses (Bernal & Fankuchen 1941) and two forms of the protein tubulin (Amos et al. 1984).

Despite this early activity in the field, protein powder diffraction did not appear as a viable technique prior to the turn of the century with the pioneering work of Robert B. Von Dreele, which relied heavily on the progress made in the discipline of powder diffraction with the advent of the Rietveld method (Rietveld 1969) and of high-angular resolution powder-dedicated instruments on synchrotrons. The work of Von Dreele initiated with the first structure refinement of a protein molecule, namely metmyoglobin, using high-resolution synchrotron powder-diffraction data (Von Dreele 1999). To perform such a refinement, the software *GSAS* (Larson & Von Dreele 2004) was modified so as to be able to carry out stereochemically restrained Rietveld refinements. A year later, Von Dreele (2000) put forth the structure of the first protein solved using the method of molecular replacement in conjunction with powder-diffraction data. More specifically, the structure of a new variant of the T<sub>3</sub>R<sub>3</sub> human insulin-Zn complex was determined

using, as a model, the single-crystal structure of the  $T_3R_3$  complex. The final study in the pioneering series demonstrated the possibility of using powder-diffraction data to detect ligands in protein-ligand systems (Von Dreele 2001). Indeed, the location of various N-acetylglucosamine oligosaccharides bound to hen egg white lysozyme was detected using difference Fourier maps generated with structure factors extracted from the powder data.

Although two ulterior ligand detection studies have been reported in the literature (Von Dreele 2005, Allaire et al. 2009), the technique of protein powder diffraction has especially been used to solve structures with the method of molecular replacement. Following the structure solution of the  $T_3R_3$  variant, molecular replacement was successfully applied to powder data using starting models of decreasing similarity with the sought-after crystal structure. The first protein to have its structure determined *via* molecular replacement using powder data and a model other than that derived from its single-crystal structure is turkey egg white lysozyme (Margiolaki et al. 2005). In this experiment, the model was derived from the structure of a closely related protein molecule with 95% sequence homology, namely hen egg white lysozyme. Within the following four years, the structure of the second SH3 domain of the ponsin protein was determined with a model of 38% sequence homology (Margiolaki, Wright, Wilmanns et al. 2007) and it was shown to be possible to solve the structure of lysozyme with one of 60% sequence homology (Doebbler & Von Dreele 2009). It is important to note that hen egg white lysozyme comprises twice as many amino-acid residues as does the second SH3 domain of ponsin.

The advances made in structure solution from molecular replacement applied to powder-diffraction data are in part due to the improved quality of the extracted data generated from the multi data set approach. This approach consists in introducing small anisotropic changes in the lattice constants, by varying one or more experimental parameters, in order to induce peak shifts and hence improve the peak separation of accidentally overlapping peaks (§2.3.1). Although this method was originally employed to generate more accurate partitioning of overlapped peaks in multi-pattern Rietveld refinements (Basso et al. 2005, Von Dreele 2007), it has increasingly been used to extract better quality single-crystal-like intensities in multi-pattern Pawley intensity extractions (§2.3.2) for a more efficient use of traditional single-crystal software. For example, the aforementioned molecular replacement studies of the ponsin domain and of hen egg white lysozyme both exploited this approach to good effect. In the former, the anisotropic variation of the unit-cell parameters was prompted by measuring powder patterns at varying degrees of X-ray radiation doses, while, in the latter, the individual powder patterns were obtained from different samples which had been crystallized using various concentrations of the

crystallizing agent.

Taking the above into consideration, it is apparent that the structure solution *via* molecular replacement and the refinement of relatively small protein molecules is well within the reach of the powder-diffraction technique. Nevertheless, both processes require the availability of a good approximate molecular model of the protein structure of interest. Alternatively, *de novo* methods provide a means of obtaining structural information about molecular systems for which no model exists. In such cases, the standard *de novo* phasing methods used in single-crystal protein crystallography are that of isomorphous replacement and anomalous scattering (§ 2.1.1). These exploit intensity differences produced by heavy atoms synthetically embedded in protein molecules (§ 2.2.1.1) and/or by anomalously scattering atoms, respectively. The first report of such protein derivatives in the context of powder diffraction is that published by (Von Dreele 2006). In this study, atoms of Xe were bound to hen egg white lysozyme and their positions within the protein molecule were detected from a difference Fourier map calculated using structure factors from the powder data and phases from the single-crystal protein structure. A second such study investigated the effect of employing the multi-pattern Pawley intensity extraction method on the ease with which U-atoms could be located in porcine pancreatic elastase using the Patterson method (Besnard et al. 2007).

With the use of this multi-pattern method and that of isomorphous replacement, the work presented herein addresses the need to qualitatively assess the information that can be extracted from protein powder diffraction data without the use of a molecular model. The *de novo* method of isomorphous replacement is appropriate to perform this assessment given that it does not require prior knowledge of the protein under study. Single and multiple isomorphous replacement analyses were performed using two protein test cases, namely hen egg white lysozyme and porcine pancreatic elastase. The results of the single isomorphous replacement analysis, carried out on both protein systems, were published in Wright et al. (2008) as part of a collaboration between our laboratory and the group of the ID31 beamline at the European Synchrotron Radiation Facility (ESRF). On the other hand, the multiple isomorphous replacement analysis, only performed on hen egg white lysozyme, was reported in Basso et al. (2010). Although the work published in Wright et al. (2008) was completed prior to the start of this thesis, it is an integral part of the project and hence is also described herein. Furthermore, the methodology employed in both studies is the same with the exception of the heavy-atom detection.

### 2.1.1 Phasing methods

As their name suggests, phasing methods are employed to recover the phase information ‘lost’ in a diffraction experiment. The main *de novo* phasing methods include that of isomorphous replacement and anomalous scattering. These rely on measurable differences in the intensity of diffraction between various data sets. Such differences are in both cases produced by a subset of atoms incorporated within the protein structure. This subset of atoms is referred to as the substructure and is composed of relatively heavy atoms for isomorphous replacement experiments and of anomalously scattering atoms for anomalous scattering experiments. In this way, intensity differences are measured between data sets generated from multiple samples (one without and one or more with embedded heavy atoms) in an isomorphous replacement experiment and from one sample at various wavelengths in an anomalous scattering experiments. Using either the Patterson method or direct methods, these intensity differences can be exploited to determine the position of the atoms constituting the substructure. With these, the structure factor of the substructure can be calculated for each reflection and hence be used as a reference wave to retrieve phase information for that of the protein structure.

While the method of isomorphous replacement has historically been the workhorse for phase determination in single-crystal protein crystallography, that of anomalous scattering has recently become the most popular of the two. One of the main reasons for this is the lack of non-isomorphism between the various data sets in a multi-wavelength anomalous scattering experiment. The issue of non-isomorphism arises from the structural changes a protein molecule can potentially undergo in the process of heavy-atom derivatization during the sample preparation for an isomorphous replacement experiment. If these changes are relatively significant, the protein structures in the different samples on which the two sets of data are collected are no longer identical, *i.e.* not isomorphous. Therefore, it is apparent to see that the phasing method of anomalous scattering is inherently immune to non-isomorphism effects given that all data sets are collected on a single sample. Furthermore, these samples are easily prepared since anomalous scatterers can either naturally be present within a protein molecule or synthetically embedded into it by introducing amino acid analogs, such as selenomethionine, into the growth medium of bacteria.

For reasons discussed in the following section, the phasing method of choice for the work reported herein was that of isomorphous replacement. In consequence, the technique of anomalous scattering is only briefly described. A more in-depth account of its phasing

capabilities in the context of macromolecular crystallography can be found in various published work, such as the recent book by Rupp (2009).

### 2.1.1.1 Anomalous scattering

The phenomenon of anomalous scattering is associated with the scattering properties of the crystal structure under study as a function of wavelength. Indeed, when the wavelength approaches an absorption edge of an atom of the crystal structure, the scattering properties of that atom change. An absorption edge represents the wavelength at which an electronic transition occurs within the anomalous scatterer. In the vicinity of this particular wavelength, the observed scattering varies in both strength and phase. This fluctuation arises from the fact that the anomalous-scattering correction term of the scattering factor varies as a function of wavelength in proximity of the absorption edge. As a result, the scattering factor,  $f$ , will vary according to

$$f = f_0 + f' + if'', \quad (2.1)$$

where  $f_0$  is the regular scattering-factor term which is independent of the wavelength, and  $f'$  and  $if''$  are the real and imaginary part, respectively, of the anomalous-scattering correction term. It is important to note that the scattering contribution of  $if''$  is  $90^\circ$  out of phase with respect to the other two terms in the sum.

The method of multiple-wavelength anomalous diffraction (MAD) to phase protein structures was pioneered by Wayne Hendrickson in the 1980s (Hendrickson 1991). This work was made possible by the advent of synchrotron facilities which offered a tunable wavelength of the X-ray beam and thus provided a means to efficiently collect various data sets on a single sample. The different wavelengths at which data are measured are typically selected so as to minimize  $f'$  and subsequently maximize  $f''$  in order to generate the maximum intensity differences between the datasets (known as dispersive differences). Although such intensity differences are smaller than those associated with a heavy-atom derivative, a pair of MAD data sets obtained from a single crystal is in principle enough to successfully phase a protein structure. This stems from the exploitability of the breakdown of the Friedel law associated with the phenomenon of anomalous scattering. As a result of the out-of-phase contribution of the last term in equation 2.1, the Friedel law no longer holds and hence  $|F(hkl)|^2 \neq |F(\bar{h}\bar{k}\bar{l})|^2$  (known as anomalous differences). The intensity differences between the Friedel pairs can thus be exploited in the same way as

those between different data sets measured at different wavelengths.

Although there are no conceptual reasons why the method of MAD could not be applied to powder diffraction data, there is one major drawback: in a powder pattern the Friedel pairs overlap exactly. Consequently, the breakdown of the Friedel law is unexploitable, since only the average intensity of the pairs can be measured, thus making a pair of MAD data sets insufficient for phasing. Although the dispersive differences could still potentially be exploited in the phasing process (Helliwell et al. 2005), the utility of those provided by data sets measured at additional wavelengths would be fairly limited since they would inevitably be inferior to that obtained with the initial two. This is due to the fact that  $f'$  and  $if''$  would have been chosen to already maximize the intensity differences, as mentioned above. In light of this, the method of isomorphous replacement was selected in the present study for the phasing of protein structures *via* the technique of powder diffraction.

### 2.1.1.2 Isomorphous replacement

The method of isomorphous replacement exploits intensity differences generated from the scattering power of heavy atoms. Such atoms, commonly defined as those with more electrons than iodine, can be introduced into protein crystals through various methods described in §2.2.1.1. All samples containing heavy atoms are referred to as derivatives while the one devoid of such atoms is called the native sample. A distinction is made in the name of the method depending on whether one or several derivatives are used. In the former case the method is referred to as single isomorphous replacement (SIR) and, in the latter, multiple isomorphous replacement (MIR).

A few years after the first successful applications of the isomorphous method to small molecular systems (Lipson & Beevers 1935, Robertson 1936), the idea that heavy atoms embedded into a protein crystal could be used to phase the X-ray reflections of that protein was suggested by Bernal (1939). However, it was not until fifteen years later that the first practical application of the isomorphous replacement method to a macromolecule was reported (Green et al. 1954). In this study, a SIR experiment was conducted on horse haemoglobin. Although the method of MIR was briefly described as part of a study of the organic molecule of strychnine sulfate a few years earlier (Bokhoven et al. 1951), the details of the phasing procedure using two derivatives were only fully understood with the publication of the article by Harker (1956). Within the same year, two additional noteworthy contributions were made to the development of the MIR method in Perutz

(1956) and Crick & Magdoff (1956). Furthermore, the detailed account of the principals of the method provided in Harker (1956) was introduced through what is now known as the Harker construction or phase circle construction. This type of construction is very useful as it visually illustrates how the various experimentally obtained data can be exploited to recover the phase angles of the protein structure factors for each measured reflection.

In an isomorphous replacement experiment, the experimental data consists of the intensity of the native protein,  $|\mathbf{F}_P|^2$ , and that of the derivative(s),  $|\mathbf{F}_{PH_x}|^2$  (where  $x \in \mathbb{N}$ , to denote the various derivatives). Given the vectorial relationship

$$\mathbf{F}_{PH_x} = \mathbf{F}_P + \mathbf{F}_{H_x}, \quad (2.2)$$

where  $\mathbf{F}_{H_x}$  is the structure factor of the heavy-atom substructure,  $|\mathbf{F}_{PH_x}|^2$  can be decomposed according to

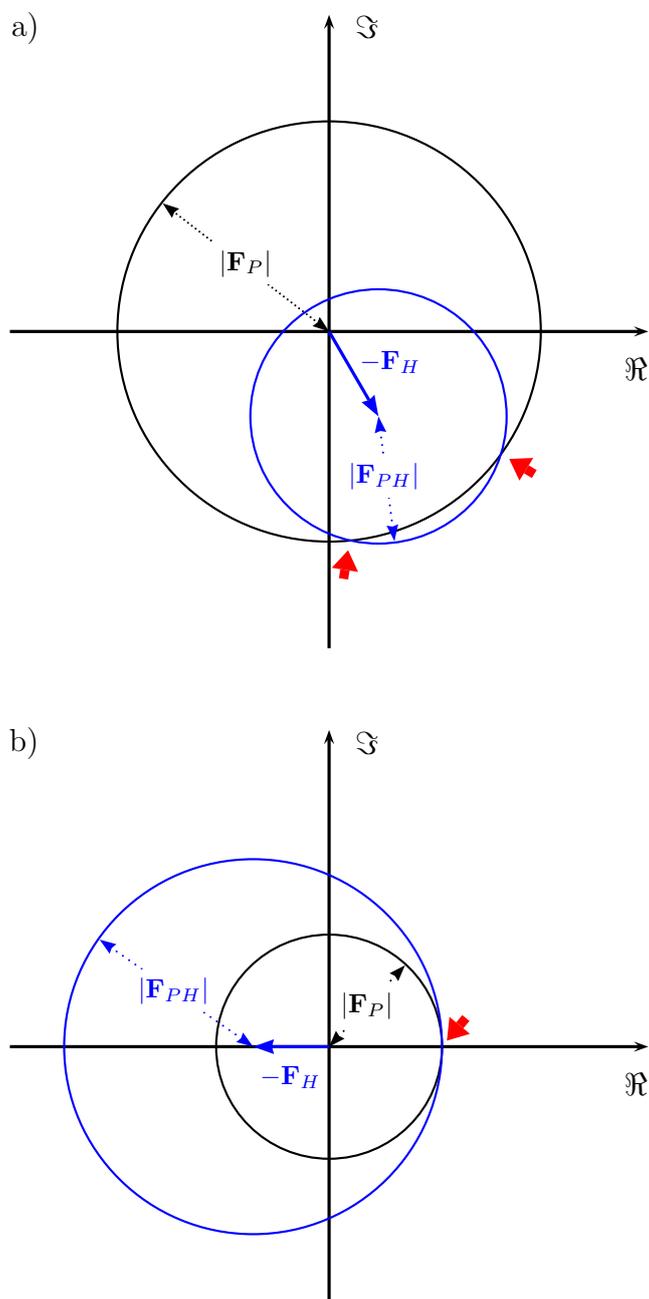
$$|\mathbf{F}_{PH_x}|^2 = |\mathbf{F}_P|^2 + |\mathbf{F}_{H_x}|^2 + |\mathbf{F}_P| |\mathbf{F}_{H_x}| \cos(\varphi_P - \varphi_{H_x}). \quad (2.3)$$

In order to use  $\mathbf{F}_{H_x}$  as a reference wave to determine the phase of  $\mathbf{F}_P$  for each reflection,  $\mathbf{F}_{H_x}$  has to be computed from the atomic positions of the heavy atoms within the crystal. As a result, these have to be determined. This can be performed by using the isomorphous differences,  $|\mathbf{F}_{PH_x}|^2 - |\mathbf{F}_P|^2$ , as an approximation of  $|\mathbf{F}_{H_x}|^2$  in either the Patterson method or direct methods.

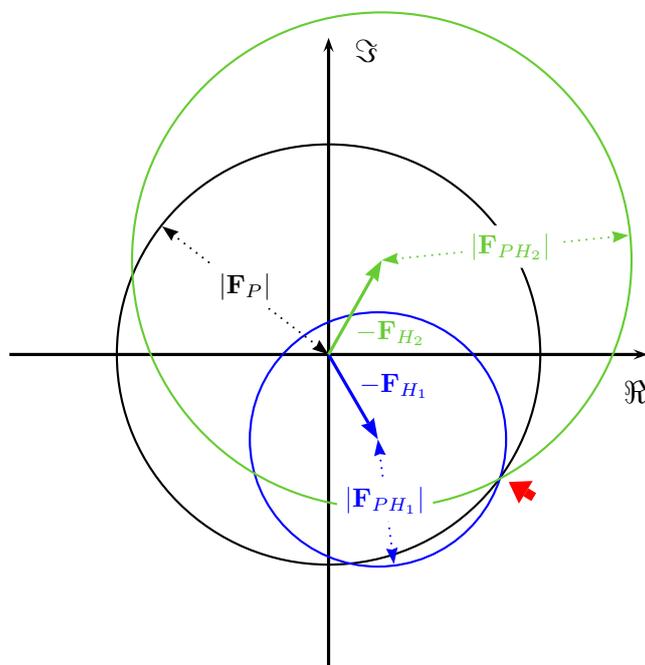
Once the heavy-atom substructure has been determined,  $\mathbf{F}_P$  can be phased for each reflection with the use of  $|\mathbf{F}_P|$ ,  $|\mathbf{F}_{PH}|$  and  $\mathbf{F}_{H_x}$ . In a Harker construction of a given reflection, these are assembled together to generate a set of circles, each representing a measured reflection intensity. The radius of a circle is defined by the amplitude of the corresponding structure factor. Furthermore, equation 2.2 can only be satisfied when the various circles intersect. In light of this, it can be seen from Figure 2.1a that the phasing of an acentric reflection with the SIR method will yield two possible solutions. For centric reflections<sup>1</sup>, on the other hand, the SIR method identifies a unique solution since the phases of such reflections are intrinsically separated  $0^\circ$  or  $180^\circ$  (Figure 2.1b).

---

<sup>1</sup>Centric reflections are reflections of which the structure factors are restricted to lie on a straight line in the complex plane, *i.e.* their phase angles are limited to two possible values separated by  $180^\circ$



**Figure 2.1** Harker constructions for a) an acentric reflection and b) a centric reflection using a SIR data set. The double-headed dotted arrows represent the vector norms derived from experimental data:  $|\mathbf{F}_P|$  (black) and  $|\mathbf{F}_{PH}|$  (blue) correspond to that of the native and derivative, respectively. As the phases are unknown for both vectors, a circle is drawn to represent all phase possibilities. The solid blue arrows represent the vectors  $\mathbf{F}_H$  calculated from the experimentally determined substructure. The thick red arrows point to phase solutions for the vector  $\mathbf{F}_P$ . While the SIR method finds a unique solution for centric reflections, for acentric reflections a twofold ambiguity remains.



**Figure 2.2** Harker constructions for an acentric reflection using a MIR data set composed of two heavy-atom derivatives. The double-headed dotted arrows represent the vector norms derived from experimental data:  $|\mathbf{F}_P|$ ,  $|\mathbf{F}_{PH_1}|$  and  $|\mathbf{F}_{PH_2}|$  correspond to that of the native, and of the first and second derivative, respectively. As the phases are unknown for these vectors, a circle is drawn to represent all phase possibilities. The solid blue and green arrows represent the vectors  $\mathbf{F}_{H_1}$  and  $\mathbf{F}_{H_2}$ , respectively, which are calculated from the corresponding experimentally determined substructure. The thick red arrow points to the unique phase solution for the vector  $\mathbf{F}_P$ .

However, given that all protein structures crystallize in noncentrosymmetric space groups, the majority of the resulting reflections are acentric. Two options exist to eliminate the bimodal ambiguity inherent to the SIR method for acentric reflections. The first consists in exploiting the breakdown of the Friedel law induced by anomalous scattering of the heavy atoms to perform a SIRAS analysis (single isomorphous replacement with anomalous scattering). On the other hand, the second simply makes use of an additional heavy-atom derivative. Although the two options are analogous, since they both result in the addition of an extra circle in the Harker construction, the first is not applicable to powder diffraction data for reasons mentioned above. Figure 2.2 depicts how a unique solution is found using the MIR method.

In reality, however, the solution termed ‘unique’ is in fact more accurately represented by a relatively broad phase probability distribution owing to the fact that the circles do not exactly coincide in one point. This is due to experimental errors, such as the

uncertainties on the measured intensities and on the heavy-atom substructure, and to the degree of non-isomorphism between the native and derivative data sets. A further complication occurs in the case of reflections for which  $\mathbf{F}_{H_x}$  is relatively small. In this instance, the resulting derivative circle will virtually be concentric with that of the native and hence there will not be well defined points of intersection. Nevertheless, probabilistic approaches, such as the one developed by Blow & Crick (1959), can be used to ensure that the ‘best’ phases are selected for  $\mathbf{F}_P$ .

## 2.2 Experimental

### 2.2.1 Protein crystallization

As a whole, the process of crystallization is a multiparametric one. In this respect, the crystallization of a protein molecule is no exception. However, due to its biochemical properties and the unique nature of the resultant biomacromolecular crystals, methods employed to produce them have developed away from small-molecule crystal growth techniques. Indeed, given the labile nature of protein molecules and their susceptibility to changes in their natural environment, the conditions in which crystals grow lie within a narrow range of pH, temperature and ionic strength. Furthermore, macromolecular crystals are composed of anywhere between 25 to 90% of disordered liquid (*i.e.* solvent), with an average value of 50% (Matthews 1985). As such, the solvent occupying the channels and spaces around the protein molecules is as much part of the crystal as the macromolecules themselves. Consequently, crystals need to be kept in contact with the mother liquor throughout the crystallization process as well as during data collection so as to prevent dehydration which would in turn lead to the collapse of the crystal structure.

In general, and contrary to most small-molecule crystals, macromolecular crystals are not grown from seeds but are nucleated *ab initio* at elevated levels of supersaturation (McPherson 1982). Once a nucleus has formed, often triggered from impurities in the solvent or defects in solid surfaces in contact with the solvated protein molecules, its growth subsequently propagates under highly unfavourable conditions of intense supersaturation. However, in most crystallization techniques, the state of supersaturation is reached gradually.

After many decades of fine-tuning and spinoff development by macromolecular crystallographers, the four types of methods most commonly used today include the following: batch, vapor diffusion, dialysis and interface diffusion methods. All share the same basic

principle of dissolving the protein molecule of interest in a water-based buffer solution (or organic solvent) which is mixed with an aqueous precipitant (often small salts or PEG polymers).

(a) *Batch methods*

Batch methods are certainly the simplest, essentially consisting in crudely mixing the precipitant solution with the protein one. In order to avoid dehydration and guarantee reproducibility, more elaborate variations of this method include submerging the mixture under a layer of oil or even suspending it between two layers of different oils (Chayen 1996, Lorber & Giegé 1996). The latter is particularly suitable for obtaining relatively larger crystals as it reduces the number of nucleation sites by removing any contact between the mixture and the walls of its supporting vessel.

(b) *Vapor diffusion methods*

The first recorded use of crystallization by vapor diffusion was reported by Hampel et al. (1968) in the structural study of tRNA. The basic principle is to equilibrate a drop, from 2 to 25  $\mu\text{L}$ , composed of a mixture of the buffered protein-solution and precipitant, against a reservoir containing the aqueous precipitant but at a greater concentration (usually 1:2) than that in the drop. Given the non-equilibrium state of the starting point, the volatile species (e.g. water or organic solvent) will diffuse from the drop to the reservoir until the vapor pressure of both is identical. This will lead to a decrease in volume of the drop and, in turn, an increase in concentration of all its constituents, to eventually reach supersaturated conditions. Vapor diffusion methods have become the methods of predilection among single-crystal crystallographers especially in the form of the ‘hanging drop’ and ‘sitting drop’. In the former, the drop is suspended from a siliconized glass slide covering a reservoir, whereas the latter involves the drop resting on an elevated support above the precipitant solution which sits at the bottom of a sealed reservoir.

(c) *Dialysis methods*

Similarly to the first two techniques described above, dialysis methods can be used in a variety of different setups all sharing the same underlying process. Much like in vapor diffusion methods, a protein-precipitant solution is separated from a large

volume of solvated precipitant, only in this case, a semipermeable membrane divides them. The role of this membrane is to prevent the passage of the macromolecules while allowing that of the smaller precipitant ones. As a result, the equilibration progress is dependent not on the vapor pressure of the solvent but rather on the molecular-weight exclusion size of the membrane.

(d) *Interface diffusion methods*

As their name suggests, these methods restrict the mixing of the protein solution with the precipitant to an interface, and hence equilibration is achieved through simple diffusion of the molecules (Salemme 1972). To minimize the phenomenon of advection, as well as to reduce the interface surface, crystallization experiments are often carried out in capillaries. Additionally, the less dense solution is, in general, carefully laid over the denser one in order to obtain a well-defined interface. In cases in which too many small crystals are obtained, the protein can be dissolved in a buffer/gel media (Robert & Lefauchaux 1988), using, for example, silica gels (Cudney et al. 1994), essentially decreasing the diffusion rate of the precipitant molecules in the macromolecular environment as well as reducing the number of nucleation sites.

Of the above four, batch crystallization techniques are the only set of methods which immediately start off from supersaturated conditions. Often, the main disadvantage this entails is the high number of nucleation sites which rapidly emerge and subsequently develop into relatively small crystals. Although the use of oils does to some extent alleviate the problem, batch crystallization remains a set of methods not the most adequate to obtain a large crystal with which to measure good quality single-crystal data. Alternatively however, they could make suitable methods to straightforwardly produce polycrystalline samples. This is especially true since additional experimental parameters, such as temperature and pH of the crystallization solution, can be varied to further influence both the nucleation rates (McPherson 1982) and the crystal size (Judge et al. 1999).

Nevertheless, powder samples of the two test cases employed in this study were not prepared according to conditions established *via* an investigation in which these experimental parameters were systematically varied to produce an ideal powder sample. Instead, microcrystalline slurries were obtained following a procedure previously published by Von Dreele (2003). This method consists in initially producing relatively large crystals (several hundred  $\mu\text{m}$  in size) from standard recipes and means, followed by grinding

them using an agate pestle and mortar. The crushing of the crystals was performed in sufficient quantities of mother liquor to prevent dehydration of the sample. The resultant microcrystalline slurries were subsequently vortexed prior to the transfer into capillaries so as to avoid preferred orientation. However barbaric an approach it may seem, the crystal-grinding process was shown not to induce significant strain in the crystals. This was done by carefully inspecting the peak width observed in the powder diffraction patterns measured at the ESRF. Indeed, the quality of the powder diffraction instruments at synchrotron sources is such that the broadness of peaks is virtually only due to the crystal size and micro-strain effects (§ 2.2.3). Consequently, the narrow diffraction line width of the observed peaks testifies to the diffraction quality of the sample as a powder.

### 2.2.1.1 Heavy-atom derivatization

The preparation of a derivative requires the binding of a heavy atom to a specific position within the protein molecule, usually on its surface and often by displacement of a solvent molecule or an ion, without distorting the protein or crystal lattice. Despite the wealth of knowledge now available on the subject<sup>2</sup> (Blundell & Johnson 1976, Blundell & Jenkins 1977), synthesizing a successful heavy-atom derivative remains a trial and error process. Yet, a technique which has had a significant impact on macromolecular crystallography is the one in which methionine residues are replaced by amino acids in which the sulphur atoms have been substituted with selenium atoms (Hendrickson et al. 1990) or tellurium atoms (Budisa et al. 1997). Although this technique has been primarily used to prepare samples for MAD experiments (§ 2.1.1.1), it has occasionally been employed in experiments which combine the use of the isomorphous replacement method with the anomalous scattering signal of the selenium and tellurium atoms.

The majority of derivatives employed in purely isomorphous replacement experiments are produced using the method of soaking. As its name suggests, the technique simply consists in bathing native protein crystals in a solution of variable concentration, typically between 0.01mM and 1M, of a heavy-atom-containing complex. Taking advantage of the particularly high solvent content in macromolecular crystals, metal compounds can diffuse along the solvent channels and into the crystal to reach the protein molecules. Although the insertion of these ions within the crystal lattice can occur in as little as a few minutes, the binding process with the macromolecules can take up to days or even

---

<sup>2</sup>See also the Heavy Atom Databank (Islam et al. 1998) at <http://www.sbg.bio.ic.ac.uk/had/>

weeks. In this approach the substitution mechanism is complex and largely entropically driven, and hence often leads to partially occupied sites. Nevertheless, many studies, including the one reported herein, have successfully exploited the phasing power of these partially occupied derivatives.

In some cases, soaking crystals leads to a degradation in crystal quality, especially in highly concentrated solutions of heavy-atom complex. Therefore, there is a tradeoff between high and low concentrations to obtain good levels of site occupancy while not damaging the crystal lattice. Alternatively, the method of co-crystallization can be employed. It involves using any of the techniques described in the previous section but with the addition of a heavy-atom complex to the precipitant solution. In this way, crystallization occurs once the heavy-atom compound is already bound to the protein molecule. The only major drawback of this procedure is that the addition of a heavy-atom complex, can, in some instances, disrupt the process of crystallization to the point where no crystals are obtained. As a result, the initial conditions determined for the growth of native protein-crystals may have to be modified in order to obtain a corresponding derivative.

## 2.2.2 Protein systems

The protein systems selected for this project were hen egg white lysozyme and porcine pancreatic elastase. These were chosen for their well-characterized nature, as a result of having been extensively studied, given that our interest lies in the methodology and not in the systems themselves. They make ideal protein test-cases for applying the method of isomorphous replacement to powder diffraction data as they are relatively small biomacromolecules and have known heavy-atom derivatives. In addition, due to the numerous structural studies carried out on both selected systems, the crystallization conditions have, over time, been perfected and hence give rise to rapid crystallization (a few days). The preparation of crystalline powders has been described in §2.2.1 and the crystallization protocols for the samples employed in this study is reported in the following sections.

### 2.2.2.1 Hen egg white lysozyme

The protein hen egg white lysozyme (HEWL) was first discovered, as well as described, by Fleming (1922). It consists of one polypeptide chain with a molecular weight of 14.6 kg/mol and is composed of 129 amino acids cross-linked by four intramolecular disulphide bonds. HEWL is an enzyme which catalyzes the hydrolysis of specific kinds of

polysaccharides comprising the cell walls of bacteria. The first structural investigation on this enzyme put forth a molecular model of the tetragonal form at a resolution of 6 Å (Blake et al. 1962). Since then, a vast number of crystallographic studies have been reported at an increasingly better resolution on this polymorph. Indeed, the structure of tetragonal HEWL was solved to a resolution of 0.94 Å as part of a crystal growth experiment under microgravity (Sauter et al. 2001). Despite not having been worked on as exhaustively, other polymorphs of the enzyme were also reported: hexagonal (Brinkmann et al. 2006), orthorhombic (Rypniewski et al. 1993), monoclinic (Harata & Akiba 2006) and triclinic (Hodsdon et al. 1990).

For this study, the protein was purchased from Sigma-Aldrich in lyophilized form and subsequently used without further purification. All polycrystalline samples of HEWL were synthesized in batch mode (§ 2.2.1) and at two different pH values - for reasons discussed in § 2.3.1 - within the pH range at which HEWL crystallizes in the tetragonal space group  $P4_32_12$  (Basso et al. 2005). Two corresponding heavy-atom derivatives were obtained by co-crystallization (§ 2.2.1.1). The specific crystallization conditions for each sample type is given below:

- NATIVE - HEWL

Samples were prepared by mixing a solution of dissolved lyophilized HEWL in a 0.1M ammonium acetate buffer with a 2M NaCl solution, to reach a final protein concentration of 50 mg/ml. These steps were carried out for two pH values: 3.5 and 4.5.

- GADOLINIUM - HEWL

Gadolinium derivatives were obtained using the same procedure described for the native samples with the addition of the lanthanide complex Gd-HPDO3A (Girard et al. 2002) at a concentration of 50 mM. A total of two polycrystalline samples were produced: one at each pH value.

- HOLMIUM - HEWL

For the holmium derivatives, the heavy-atom containing compound used was  $\text{HoCl}_3$  (Jakoncic et al. 2006). It was diluted at a concentration of 400mM in a 0.05M NaAc buffer already containing 1.2M NaCl and subsequently added to the protein solution. A total of two samples were crystallized at pH values of 4.5 and 5.5 and a

final protein concentration of 37.5 mg/ml.

A second gadolinium derivative was produced using a concentration of 10 mM for the lanthanide complex, but given its low site occupancy, and hence low phasing power, it was not used in either the SIR or MIR analysis. Additional heavy atoms were attempted to be bound to HEWL without success. These include iridium, mercury and uranium using both the soaking and co-crystallization methods. Further synthesis of heavy-atom derivatives of HEWL were discontinued once the holmium derivative was obtained.

### 2.2.2.2 Porcine pancreatic elastase

The protein porcine pancreatic elastase (PPE) belongs to a family of enzymes known as the serine protease family, which is one of the most substantially studied assembly of enzymes. It is comprised of one polypeptide chain and with almost twice the number of amino acids with respect to its test-case counterpart, namely 240, it has a molecular weight of 25.9 kg/mol cross-linked by four disulphide bridges. PPE possesses the ability to hydrolyze a wide variety of protein substrates, including native elastin, a substrate not attacked by all other proteases. Its tertiary structure was first elucidated by Watson et al. (1970) to a resolution of 3.5 Å. It was the second serine protease to have its structure solved, which, since then, has been determined to atomic resolution, *i.e.* 1.1 Å (Würtele et al. 2000).

As part of this study, lyophilized PPE was purchased from SERVA Chemicals and used without further purification. All elastase crystals were prepared using the hanging drop method, as described in § 2.2.1.1, at pH 5.5 at which PPE crystallizes in the orthorhombic space group  $P2_12_12_1$ . The crystals contained in the various drops were amassed and crushed to a microcrystalline slurry. A heavy-atom derivative was subsequently synthesized through soaking (§ 2.2.1.1) of the native slurry. The specific crystallization conditions for each sample type is given below:

- NATIVE - PPE

A native sample was obtained by dissolving the protein in a 0.05M NaAc buffer at pH 5.5, already containing 0.2M Na<sub>2</sub>SO<sub>4</sub>, to form drops with an initial protein concentration of 50 mg/ml equilibrated over a reservoir solution of 0.4M Na<sub>2</sub>SO<sub>4</sub>.

- URANIUM - PPE

A uranium-containing derivative was produced by soaking a native microcrystalline slurry in a 5mM  $\text{UO}_2(\text{NO}_3)_2$  solution for a few days.

In order to conduct a MIR analysis, a second heavy-atom derivative was sought. Substantial work was carried out in an effort to bind mercury atoms to elastase through the use of the heavy-atom containing complex: p-chloromercuribenzenesulfonic acid (Shotton & Watson 1970). Crystallization was attempted by soaking as well as by co-crystallization and, unfortunately, Hg atoms were never successfully bound. As part of a collaboration with Richard Kahn's group at the IBS<sup>3</sup>, in Grenoble, we were provided with lanthanide-containing complexes: Gd-DO3A, Gd-HPDO3A, Gd-DTPA and  $\text{Na}_3[\text{Eu}(\text{DPA})_3]$ . With these, soaking experiments were performed using microcrystalline slurries of PPE. All four heavy-atom complexes have, by now, been successfully bound to single crystals of HEWL (Girard et al. 2003) and/or PPE (Pompidor et al. 2010). However, the analysis of the powder diffraction data collected on the soaked microcrystalline-slurries revealed that none of the four complexes had successfully bound lanthanide atoms to the PPE molecule.

### 2.2.3 X-ray data collection

Given the size of protein molecules and the high solvent content within protein crystals, the corresponding unit cells are relatively large. Hence, for a given sample volume, a protein crystal will contain fewer unit cells than a crystal of a small molecule and, in turn, generate a relatively weak diffraction signal since the integrated reflection intensities increase with the number of unit cells exposed to the X-ray beam. While this issue affects any type of diffraction experiment, large unit cells also amplify an issue specific to the powder-diffraction technique, namely peak overlap. Indeed, large unit cells entail large  $d$ -spacings which give rise to a high density of peaks and thus to a high degree of peak overlap. This is especially problematic at high Bragg angles as the density of peaks increases with the cube of the distance from the reciprocal space origin. As a result of these intrinsic problems, the use of synchrotron radiation is essentially inevitable for all but the most basic structural investigations (see Appendix A.3). Indeed, the extremely high intensity of such radiation ensures a measurable diffraction signal while

---

<sup>3</sup>Institut de Biologie Structurale - J.P. Ebel, 41, rue Jules Horowitz, F-38027 Grenoble

its exceptional collimation allows for the recording of narrow diffraction peaks which are less likely to overlap.

Moreover, dedicated high-resolution powder-diffraction beamlines are equipped to produce even higher angular resolution to further reduce the likelihood of peak overlap (Margarolaki, Wright, Fitch et al. 2007). Two such beamlines exist at the ESRF, namely the Swiss-Norwegian beamline BM01B and the ID31 beamline (Fitch 2004). Although these are very similar beamlines, the main difference between the two is that the former is situated on a bending magnet (bent section of the storage ring) and the latter on an insertion device (straight section of the storage ring). There are two types of insertion devices: wigglers and undulators. While wigglers produce a radiation within a wider spectral range than undulators, they both produce a significantly higher flux as compared to a bending magnet. The X-ray radiation available at ID31 is produced from three undulators and hence is considerably more intense than that generated at BM01B. Nevertheless, both beamlines provide diffraction data of very high angular resolution since they are equipped with Si(111) analyzer crystals. Each analyzer crystal precedes a detector, of which there are six on BM01B and nine on ID31. The set of detectors are separated by approximately  $1.1^\circ$  and  $2.0^\circ$  for the former and latter beamlines, respectively. With such instrumentation, an angular resolution as low as  $0.01^\circ$  full width at half maximum (FWHM) has been obtained for protein powder samples. The instrumental contribution to the FWHM is minimal, roughly  $0.003^\circ$ , while the rest is essentially due to crystallite size and microstrain effects. The achieved resolution is due to the ability of analyzer crystals to discard the contribution due to the sample size and, consequently, larger sample volumes can be employed (of the order of  $3 \text{ mm}^3$ ). This is a crucial aspect for protein diffraction experiments carried out on beamlines such as ID31. Indeed, given their high-photon flux, coupled with the intrinsic weak scattering power and sensitivity to radiation damage of protein crystals, protein powder samples need to be large enough to allow data to be recorded at multiple points unaffected by radiation damage. These multiple data sets can then be added to one another to generate a single powder diffraction pattern in which the counting statistics have been maximized.

In this study, high-resolution powder diffraction data were collected at room temperature (291K) on both ID31 and BM01B beamlines. However, prior to the measurements the slurries of crushed crystals were transferred into 1.5 mm diameter glass capillaries. The capillaries were subsequently centrifuged so as to improve the crystallite packing density (Von Dreele 2003). Most of the excess mother liquor was then removed to be able to cut the capillary for a precise alignment of it. Once cut, the capillaries were sealed

**Table 2.1** X-ray powder-data acquisition parameters for the different sample types of HEWL and PPE.

Samples	HEWL						PPE	
	Nat		Gd		Ho		Nat	U
	pH 3.5	pH 4.5	pH 3.5	pH 4.5	pH 4.5	pH 5.5	pH 5.5	
Beamline	ID31		ID31		BM01B		ID31	
Wavelength (Å)	1.54999(3)		1.54999(3)		0.8000(1)		1.25085(3)	
Scan range $2\theta$ (°)	0 - 30		0 - 30		0 - 12		0 - 30	
Scan speed (°/min)	10		10		0.3		10	
Binning step (°)	0.004		0.004		0.002		0.002	
Nbr. of merged scans	8		8		8	7	6	

with wax in order to prevent dehydration of the microcrystals. For samples recorded on ID31, measurements were performed at multiple points on the sample by translating the capillaries and thereby exposing a ‘fresh’ part of the slurry. As a result, for those samples, the steps prior to the cutting of the capillary were repeated until the sample length would allow multiple translations to be made. While the process of translation is usually used to enable the addition of a number of scans unaffected by radiation damage, it can also be employed to add scans which have been exposed to an equivalent dose of irradiation (see § 2.3.1).

In total, powder diffraction data were collected for six polycrystalline slurries of HEWL and for two of PPE. Table 2.1 contains a summary of the experimental parameters used for the data collection of the various sample types: native samples (Nat), and gadolinium (Gd), holmium (Ho) and uranium (U) derivatives. During each measurement, capillaries were spun at high speed to ensure sufficient averaging of crystallite orientations. For HEWL samples measured on ID31, *i.e.* Nat and Gd, the capillaries were translated while recording several scans at each ‘fresh’ position. Of those scans, the eight unaffected by radiation damage were added to one another resulting in increased counting statistics. The holmium-HEWL data were measured on BM01B. Given the relatively low photon-flux of this beamline, radiation damage was not significant enough to justify translating the capillaries. In consequence, the data were obtained from a single position on each capillary, merging only the scans unaltered by radiation damage which ultimately emerges. PPE capillaries, measured on ID31, were translated to a total of six positions while collecting four and five scans per position for the native and uranium-derivative samples, respectively. The scans subjected to the same degree of irradiation, at each capillary position,

were added to one another leading, once again, to increased counting statistics. Finally, the powder diffraction data used in the isomorphous replacement analyses included six patterns for HEWL (two at different pH values for each sample type: Nat, Gd and Ho) and nine for PPE (four and five for Nat and U, respectively).

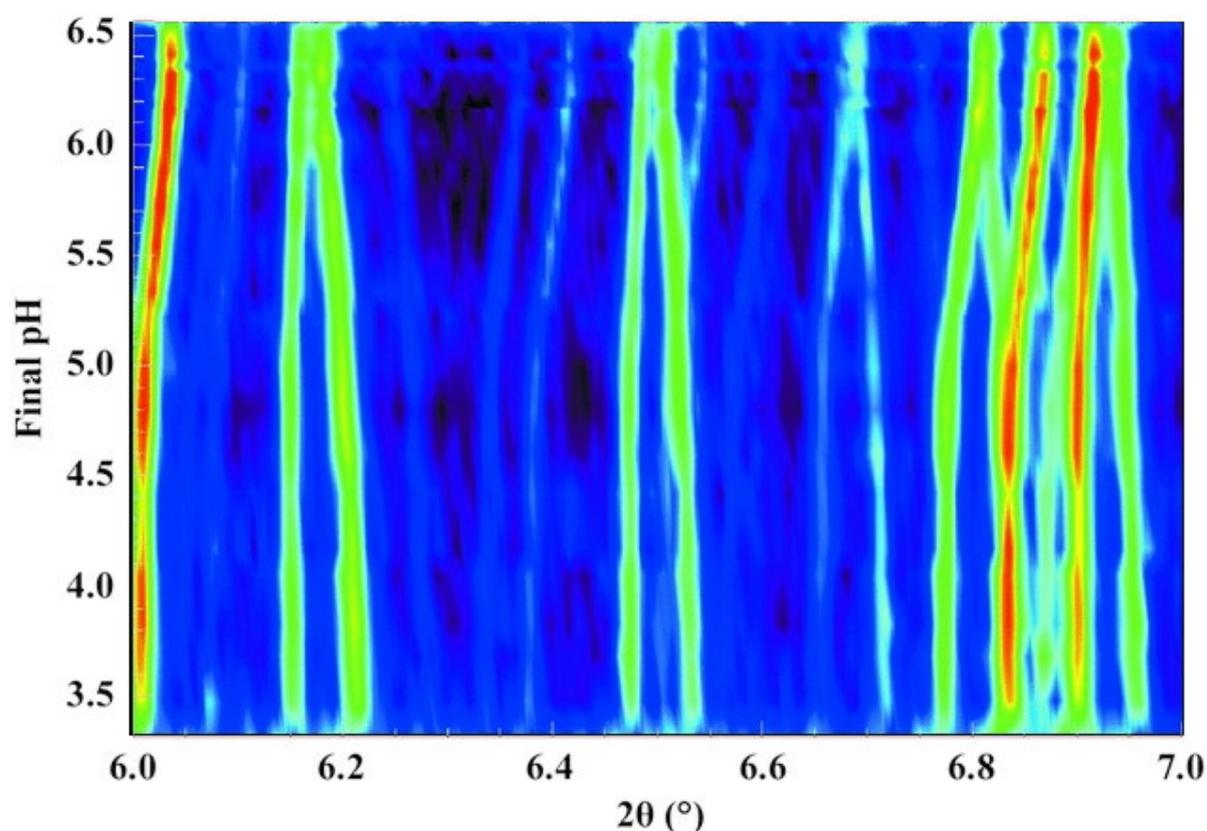
### Single-crystal reference data

Single-crystal data were collected on single crystals of both PPE sample-types for comparison purposes. The data were measured at the BM01A beamline using an image-plate area detector. They were subsequently processed with *MOSFLM* (Leslie 1993) and reduced with programs from the *CCP4* software suite (Collaborative Computational Project, Number 4 1994).

## 2.3 Results

### 2.3.1 Anisotropic lattice changes

The method of isomorphous replacement relies on generating phase information from measured intensity differences for each reflection in a set of data (§ 2.1.1.2). Therefore, the powder-specific problem of overlapping peaks (see Appendix A.2) impedes its success when applied to powder data, even if synchrotron radiation is employed. The situation is further aggravated by the fact that there are often small differences in the lattice parameters between the native sample and the heavy-atom derivatives (non-isomorphism). As a result, a reflection that is well resolved in the native pattern may be overlapping in the derivative data or vice versa. In turn, the measured intensity differences between the two peaks can no longer simply be attributed to the addition of heavy atoms within the protein molecules. One way to partially circumvent this issue is to intentionally induce further anisotropic changes in the unit-cell parameters (without affecting the crystal and lattice symmetries) of each sample type. In doing so, the separation of accidentally overlapping reflections can be improved and thus the intensity difference between peaks of different sample types can be determined from more accurately extracted intensities (Shankland, David & Sivia 1997). These lattice variations can be prompted through various experimental parameters such as temperature, radiation damage and pH of the crystallization solution. In small-molecule crystallography, anisotropic thermal expansion has been exploited to evaluate the individual contributions of reflections that are overlapping at one temperature but are resolved at another (Shankland, David &



**Figure 2.3** Colour representation of powder diffraction data from the pH-variation experiment (Basso et al. 2005). HEWL samples were crystallized at 277 K from pH 6.56 to 3.33. At this temperature and pH range the anisotropic shift in the peak position is apparent.

Sivia 1997, Brunelli et al. 2003). However, performing measurements while varying the temperature is not suitable for protein powders for two reasons. The first concerns the temperature range in which protein molecules are stable, essentially excluding the use of relatively high temperatures (above  $\sim 60^\circ$ ). The second concerns the fact that protein crystals contain water-based solvent which is susceptible to turn into ice at temperatures below  $0^\circ\text{C}$ . In single-crystal experiments, this issue has been resolved with the advent of flash cooling techniques. Nevertheless, this method is inadequate to cool powder samples since the thickness of a powder sample will prevent the cooling to be fast enough for crystallites buried in the middle of the capillary (Jenner et al. 2007).

For HEWL, anisotropic changes in the lattice parameters were induced by varying the pH of the crystallization buffer. Indeed, for the tetragonal polymorph of HEWL there is a systematic dependence of the lattice parameters on the pH of the crystallization solution (Basso et al. 2005). This is clearly visible in Figure 2.3, taken from the aforementioned article, which shows the pronounced anisotropic variation in peak position as a function

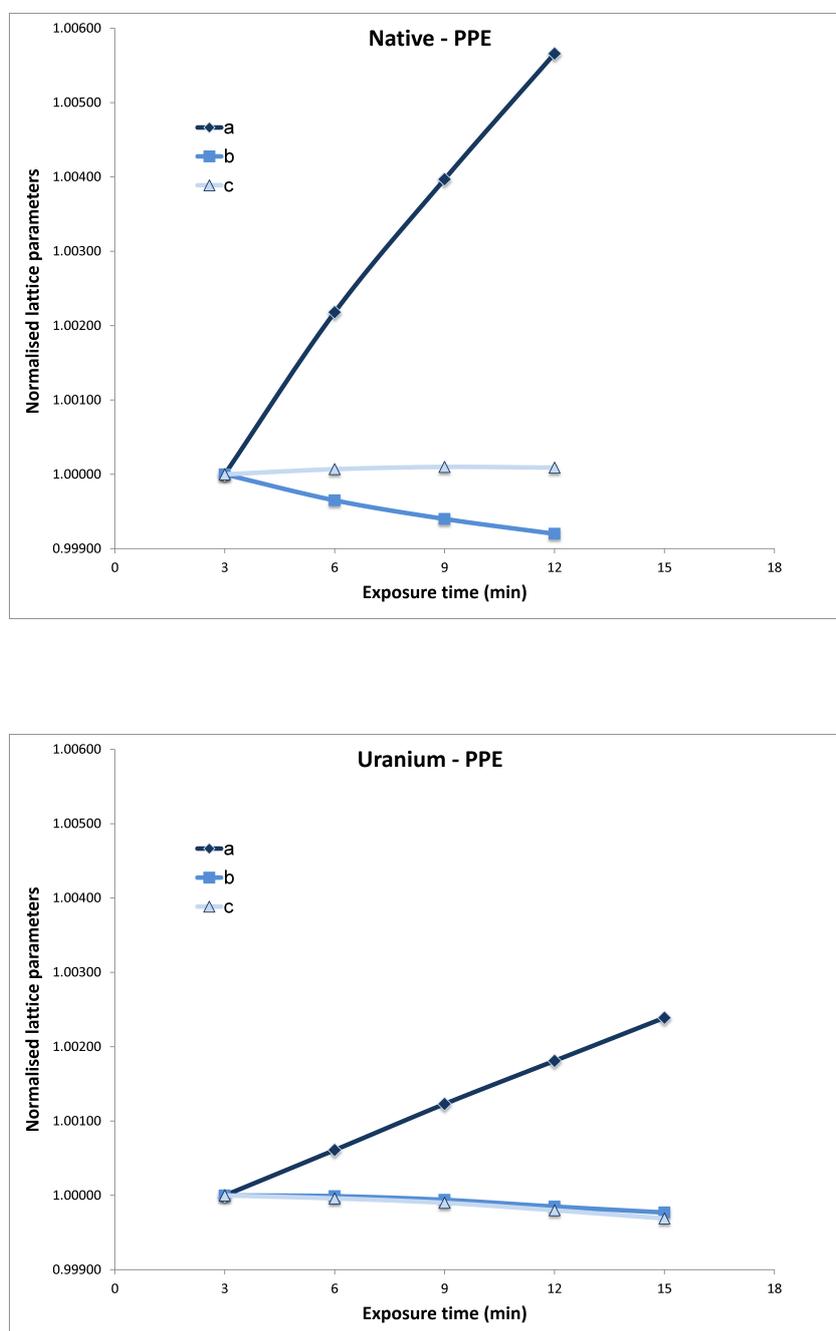
**Table 2.2** Lattice parameters for the different sets of HEWL and PPE polycrystalline slurries: native samples (Nat), and gadolinium (Gd) and holmium (Ho) derivatives. The relative change in lattice parameters as a function of pH variation is given by  $\Delta a/a$  and  $\Delta c/c$ .

Samples	HEWL					
	Nat		Gd		Ho	
	pH 3.5	pH 4.5	pH 3.5	pH 4.5	pH 4.5	pH 5.5
$a$ (Å)	79.226(8)	79.244(8)	78.990(8)	78.988(8)	79.228(8)	79.235(8)
$c$ (Å)	37.956(4)	37.931(4)	38.448(4)	38.446(4)	37.857(4)	37.885(4)
$\Delta a/a$ (%)	+0.023		−0.003		+0.010	
$\Delta c/c$ (%)	−0.066		−0.005		+0.076	

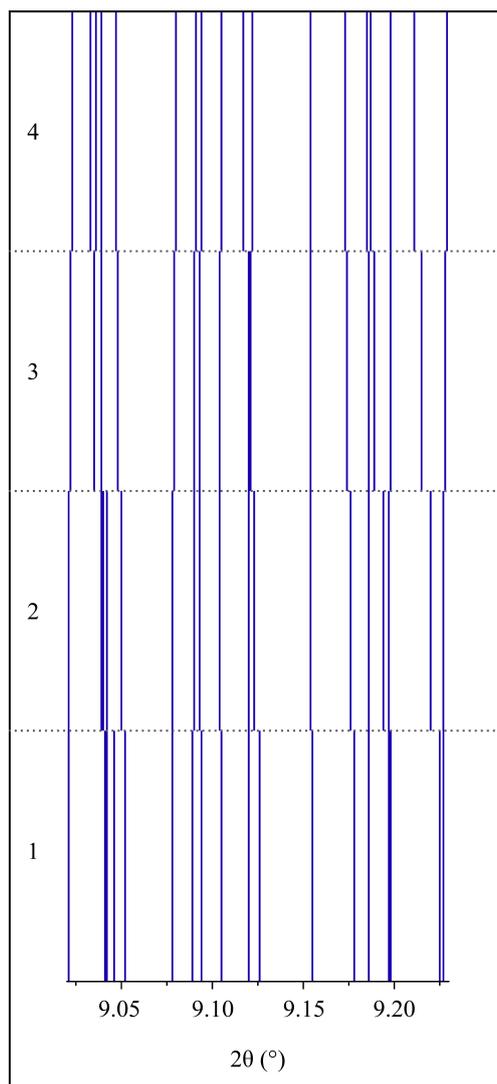
of pH using 24 HEWL powder-samples crystallized at 277K. The correlation between pH and unit-cell dimensions observed for the various sample types of this study is reported in Table 2.2.

For PPE, radiation damage from the high photon flux was the experimental parameter exploited to cause lattice changes in the unit cell. The various effects caused by exposing samples to an intense X-ray beam have over the last decade been a widely discussed topic in single-crystal protein crystallography (Garman 2010), as they can compromise structure determination. Among the most common phenomena is an irreversible radiation-induced lattice expansion, although the mechanism by which this occurs is not fully understood (Ravelli et al. 2002). In many instances, however, the changes in the unit-cell parameters are anisotropic. This is also the case for PPE (Besnard et al. 2007), as is quantitatively illustrated in Figure 2.4. In turn, these anisotropic lattice changes lead to differential variations in peak position and can thus be exploited to resolve overlapping reflections in a powder pattern (Besnard et al. 2007, Von Dreele 2007). Figure 2.5 shows such variations for the native sample of PPE.

The variation of the pH of the crystallization buffer-solution and X-ray irradiation are thus appropriate experimental parameters to be used to improve the resolution of accidentally overlapping peaks. However, their efficiency seems to be system dependent. Indeed, no peak shifts as a function of radiation damage were observed for HEWL samples measured on ID31. As a result, radiation damage could not be exploited to deconvolute peaks in the same way it was with PPE.



**Figure 2.4** Radiation-induced anisotropic variation in the lattice parameters of native PPE (top) and of its uranyl derivative (bottom). The absolute values for the lattice parameters at the origin (first data set) are  $a = 51.865 \text{ \AA}$ ,  $b = 57.896 \text{ \AA}$ ,  $c = 75.308 \text{ \AA}$  for the native sample, and  $a = 51.151 \text{ \AA}$ ,  $b = 57.979 \text{ \AA}$ ,  $c = 75.422 \text{ \AA}$  for the uranyl derivative.



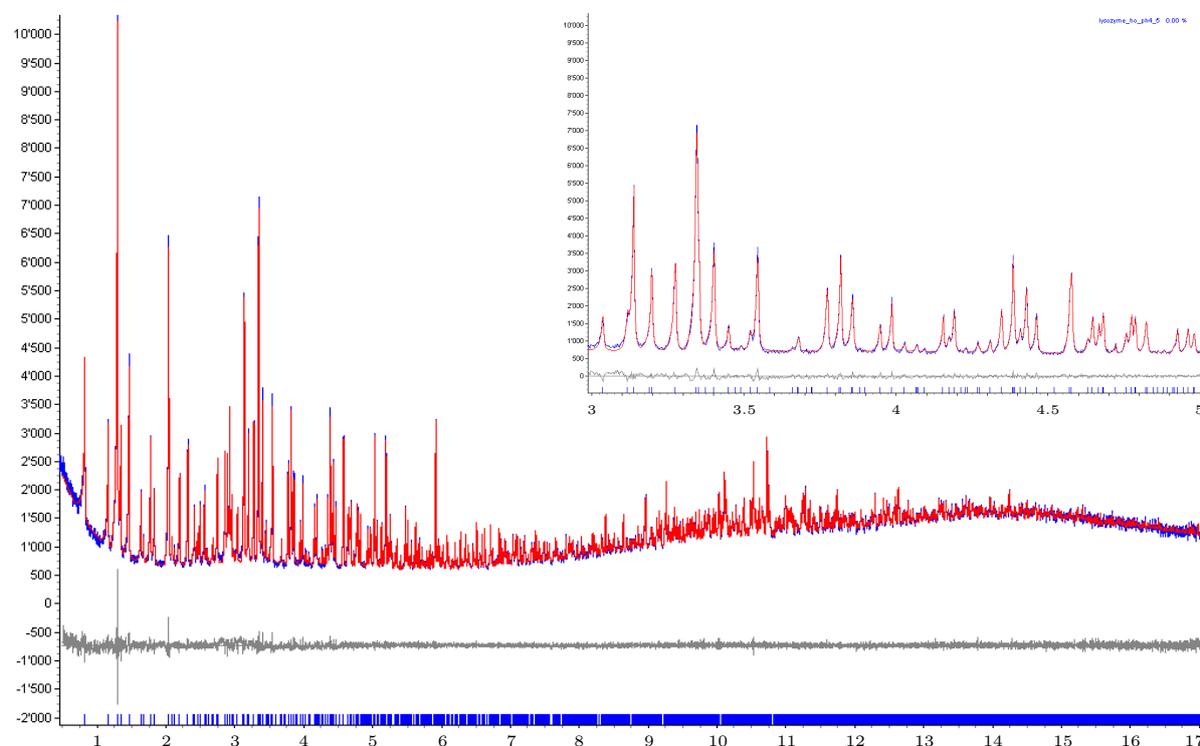
**Figure 2.5** Effect of radiation-induced anisotropic lattice changes on the position of diffraction peaks in the  $2\theta$  region 9.00-9.25 for native PPE. The four different powder patterns correspond to sequential stages of X-ray irradiation.

### 2.3.2 Intensity extraction

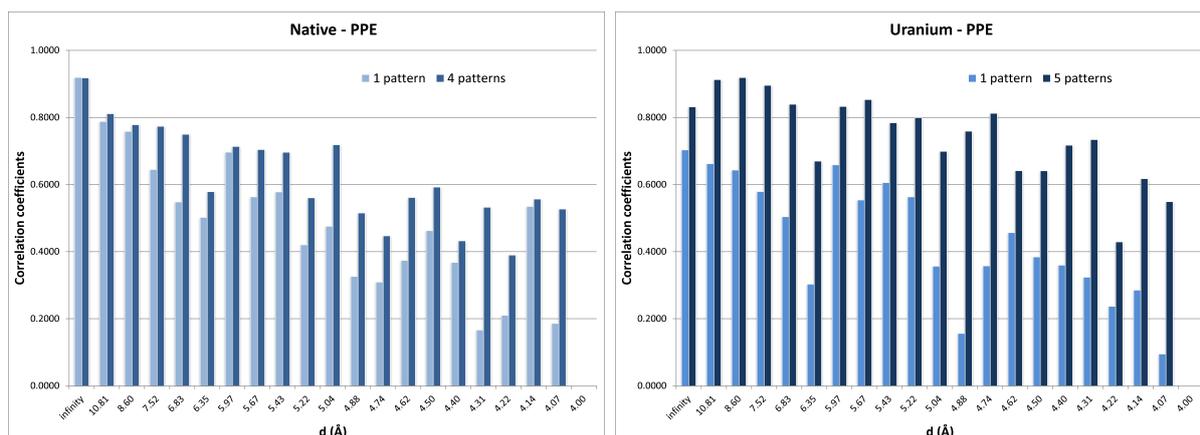
Prior to the obtention of the holmium data, a SIR analysis was performed on both HEWL and PPE using the Gd- and U-derivatives, respectively. In this analysis, the peak intensities in the lysozyme data set were extracted with the program *PRODD* (Wright & Forsyth 2000) while those of elastase were extracted using the software *TOPAS-Academic* (Coelho 2004). Different programs were employed since the HEWL data were processed by a group from the ESRF, as part of the aforementioned collaboration (§ 2.1). Nevertheless, the results of the SIR study reported by Wright et al. (2008) do not critically depend on the methods implementation in the software used for the intensity extraction procedure. Additionally, the data sets remained independent throughout the entire analysis. On the other hand, the MIR investigation conducted on HEWL exploited the combined phasing power of both HEWL derivatives. To ensure consistency in the data treatment, all HEWL data were extracted with one single software, namely *TOPAS-Academic*. Hence, the native and gadolinium data were reprocessed with *TOPAS-Academic* and the following description of the intensity extraction process solely concerns carrying it out with this powder-dedicated software.

Given the well-characterized nature of the two chosen protein systems for this study, both space groups and unit-cell parameters were known and hence no indexation was required. However, prior to refining the intensity of peaks, a LeBail decomposition (LeBail et al. 1988) was performed on each powder pattern (6 for HEWL, 9 for PPE) so as to obtain a well defined peak shape along with precise lattice parameters for each sample. With these, a subsequent Pawley decomposition (Pawley 1981) was carried out in which the peak intensities were variables in a least-squares procedure. A representative example of the result of the above procedure is provided in Figure 2.6. From the flatness of the difference curve (grey line), it is possible to see that the agreement between the observed (blue line) and calculated (red line) profiles is good.

Final intensities were extracted using the method of multi-pattern Pawley intensity extraction (MiPPiE). This method consists in calculating each diffraction pattern as a sum of overlapping reflections, the intensities of which are refined the same way as in a conventional Pawley refinement. Therefore, for each sample type (3 for HEWL, 2 for PPE) a single set of intensities can be fitted to multiple data sets, in this case, recorded for different values of pH of the crystallization solution (for HEWL) or at a different degree of X-ray irradiation (for PPE). In doing so, one is able to benefit from the partial separation of overlapping peaks arising from the anisotropic changes in the lattice dimensions and



**Figure 2.6** Pawley refinement fit for the holmium-HEWL sample at pH 4.5. The blue line corresponds to the observed data, the red line to the calculated pattern and the grey line to the difference between the two. The blue ticks represent the positions of Bragg reflections. A zoomed-in image from  $3^\circ$  -  $5^\circ$  in  $2\theta$  is also included.



**Figure 2.7** Effect of the MiPPIE method for native PPE (left) and its uranyl derivative (right). The graphs show, in shells of resolution, the correlation coefficients between structure factor amplitudes obtained from single-crystal measurements and those obtained from powder diffraction data with and without the use of the MiPPIE method. The correlations reported are for the powder data extracted from one single pattern and for the series of four, respectively five consecutive powder patterns pertaining to different radiation-induced lattice changes.

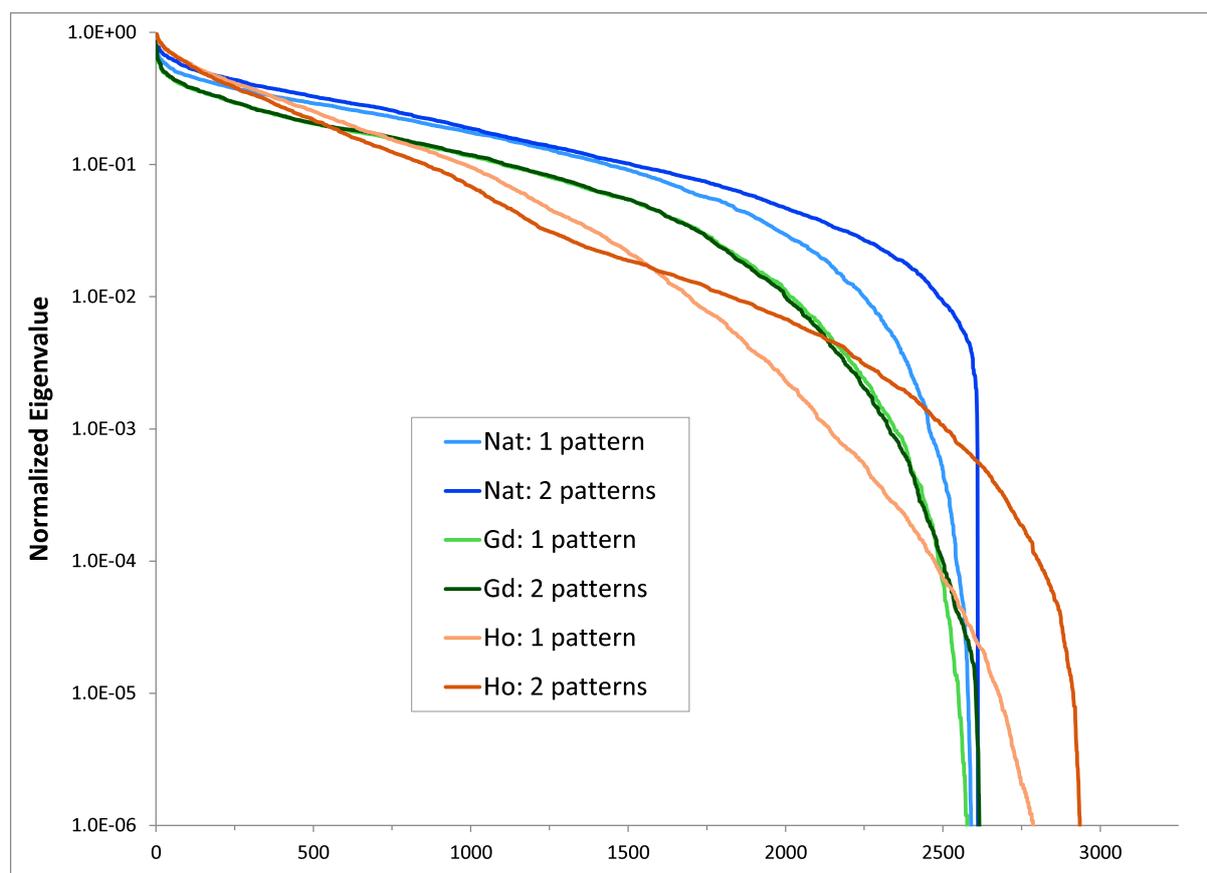
**Table 2.3** Refinement parameters and statistical data of the various multi-pattern Pawley refinements.

Samples	HEWL			PPE	
	Nat	Gd	Ho	Nat	U
$\lambda$ (Å)	1.54999(3)	1.54999(3)	0.8000(1)	1.25085(3)	1.25085(3)
$2\theta$ range (°)	1.40–32.00	1.40–32.00	0.50–17.10	0.36–10.3	0.36–10.3
$d_{min}$ (Å)	2.8	2.8	2.7	3.5	3.5
Nbr. of reflections	3229	3229	3641	2862	3076
$R_{wp}$ (%)	5.73	7.25	3.57	8.34	3.33
$R_{exp}$ (%)	4.26	4.55	2.58	5.78	2.79
$R_p$ (%)	4.72	5.60	3.10	6.17	2.61
$\chi^2$	1.35	1.60	1.39	1.44	1.19

as a result extract data of better quality. This improvement in quality is shown in Figure 2.7 which contains comparisons between PPE powder-data extracted with and without using the MiPPIE method and the structure factor amplitudes obtained from the corresponding single-crystal measurements (§ 2.2.3). The parameters of each multi-pattern Pawley refinement and the resulting statistical data can be found in Table 2.3. The latter reflect the quality-of-fit achieved with the MiPPIE method.

Though the improvement is unequivocal, it could in theory simply be due to increased statistics ensuing the addition of multiple patterns. To determine whether or not the observed improvement is in fact a direct result of the benefits of the MiPPIE method, the least-squares matrix of the Pawley refinement was examined as it provides further information concerning the data quality. Indeed, the information content in powder patterns is given by the eigenvalue spectrum of a reduced matrix constructed from the purely intensity-related elements of the full Pawley normal matrix (Sivia 2000, Wright 2004). Thus, the improvement in the information content upon using the MiPPIE method can be quantified by comparing the eigenvalue spectra generated from the intensities extracted from single patterns to those extracted from multiple patterns. Figure 2.8 shows such a comparison for the three sets of HEWL samples. The number of eigenvalues that have a significant magnitude, which corresponds to those appearing before a discernible dip in the spectrum, determines the quantity of ‘good pieces of intensity information’ in the corresponding set of intensities.

Therefore, from Figure 2.8 it is possible to conclude that the MiPPIE method successfully increased the information content in the intensities extracted for the native and



**Figure 2.8** Effect of the MiPPIE method for the HEWL samples. Each curve represents an eigenvalue spectrum of a reduced matrix constructed from the purely intensity-related elements of the full Pawley normal-matrix. The spectra (normalized by the largest eigenvalue) are given for Pawley intensity extractions from one pattern (at one pH value) and two patterns (at two different pH values) for each sample type: native (Nat), gadolinium derivative (Gd) and holmium derivative (Ho). The quantity of ‘good pieces of intensity information’ in one or a set of powder patterns is given by the number of eigenvalues that have a significant magnitude (Sivia 2000, Wright 2004).

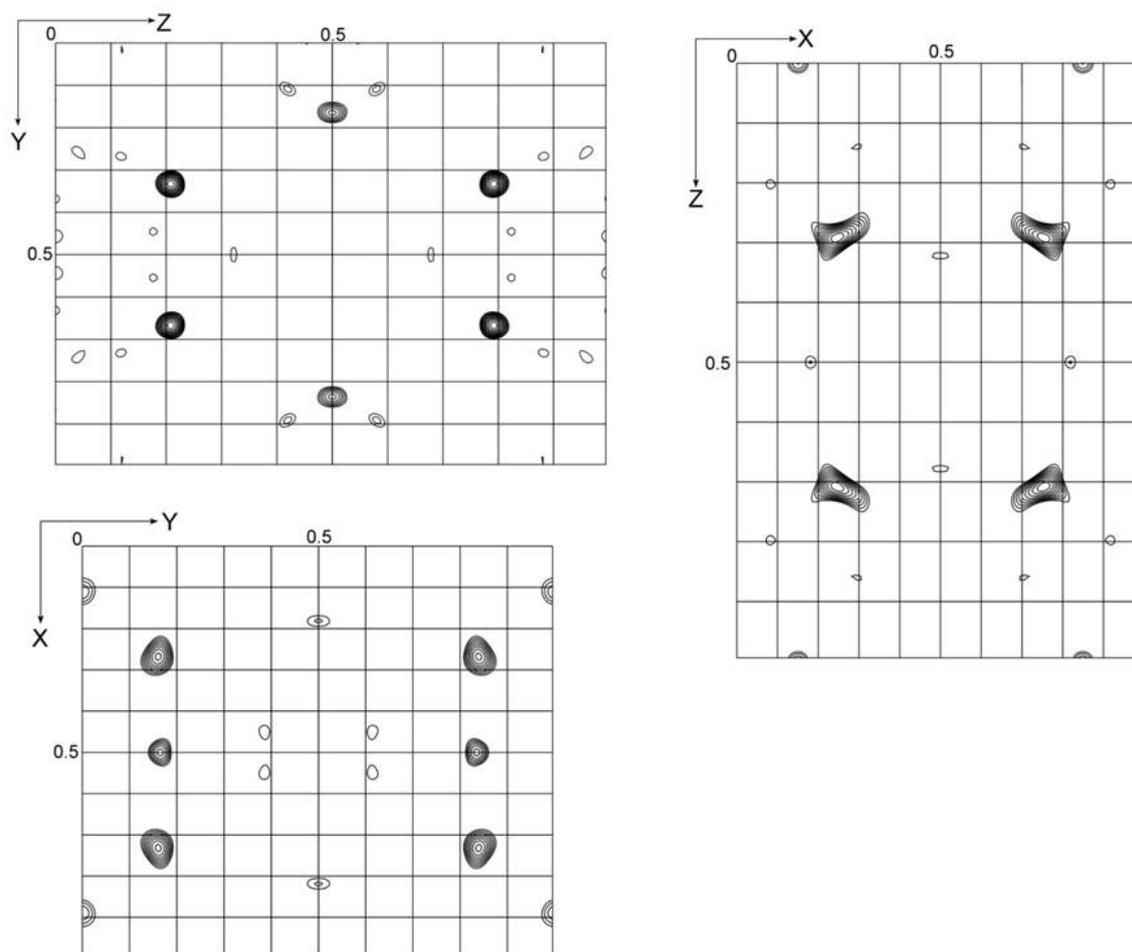
holmium-derivative samples and considerably less so for the gadolinium derivative. The difference in the success rate of the extraction method for the various sample types can be attributed to the percentage change in the lattice constants upon varying the pH of the crystallization solution (Table 2.2). In other words, for the gadolinium derivative, the anisotropic change in lattice parameters is not significant enough to profit from the MiPPiE method. This observation demonstrates that the improvement in the quality of the extracted intensities is a direct consequence of the anisotropic shift in lattice parameters and is not simply due to better statistics achieved through the addition of multiple patterns.

Initially, the standard uncertainties (s.u.) on the extracted intensities were taken from the diagonal elements of the reduced error-covariance matrix obtained after the convergence of the Pawley refinement. However, it was realized to be highly inadequate given that unreasonably disproportionate s.u. values were assigned to peaks which were highly correlated to each other, *i.e.* overlapped. This prevented the subsequent phasing procedure from running properly. As a result, the s.u. were estimated using the bootstrap method (Elfron & Tibshirani 1986) instead. Although the use of this method generated reasonable s.u. values for overlapping peaks, it did not, at first, resolve the issues encountered in the phasing process. Indeed, very weak reflections were assigned s.u. values of zero, or very close to zero, which is not tolerated by the phasing algorithm. To solve this problem, a constant of 0.02 was added to all s.u. values.

Taking no additional measures to further alleviate the problem of overlapping peaks, the final extracted peak-intensities were subsequently treated as though they originated from a single-crystal experiment (*i.e.* a list of  $hkl$ ,  $I_{hkl}$  and  $\sigma(I_{hkl})$ ). This allows for the use of well established biomacromolecular single-crystal software to perform the isomorphous replacement analysis.

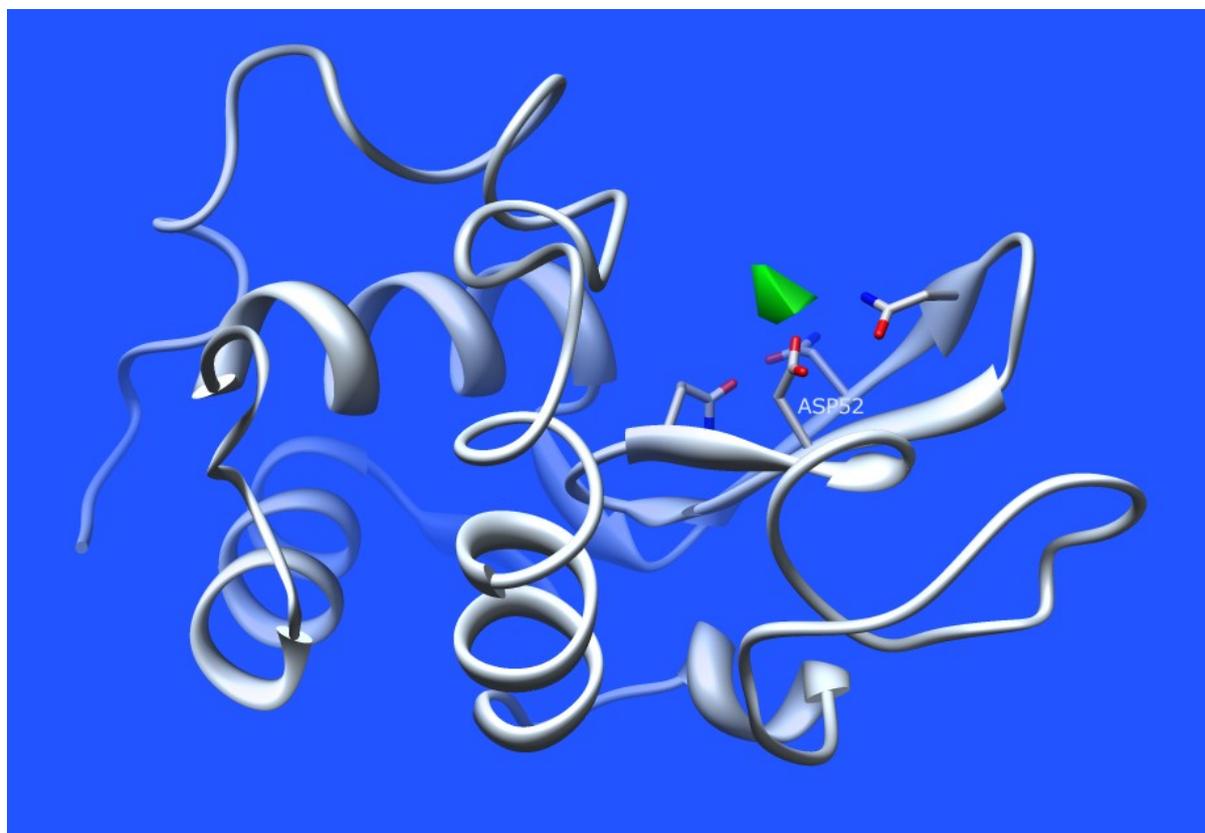
### 2.3.3 Heavy-atom phasing

In an isomorphous replacement experiment, preceding the detection of heavy atoms within the structure of a protein molecule of a heavy-atom derivative, the extracted intensities from the various powder patterns have to be adjusted to a common scale. Each native data set was scaled with its respective derivative intensity data set(s) with software from the *CCP4* package (Collaborative Computational Project, Number 4 1994). All subsequent crystallographic computations were performed using conventional single-crystal programs, most of which are included within the aforementioned package.



**Figure 2.9** Isomorphous difference Patterson map of the uranyl derivative of PPE. The Harker sections  $x = \frac{1}{2}$ ,  $y = \frac{1}{2}$  and  $z = \frac{1}{2}$  are shown. Contours are at intervals of  $0.5\sigma$ , starting  $3\sigma$  above the mean density (where  $\sigma$  is the r.m.s. of the map density). In all sections, Harker peaks which correspond to the interatomic U-U vectors can be seen.

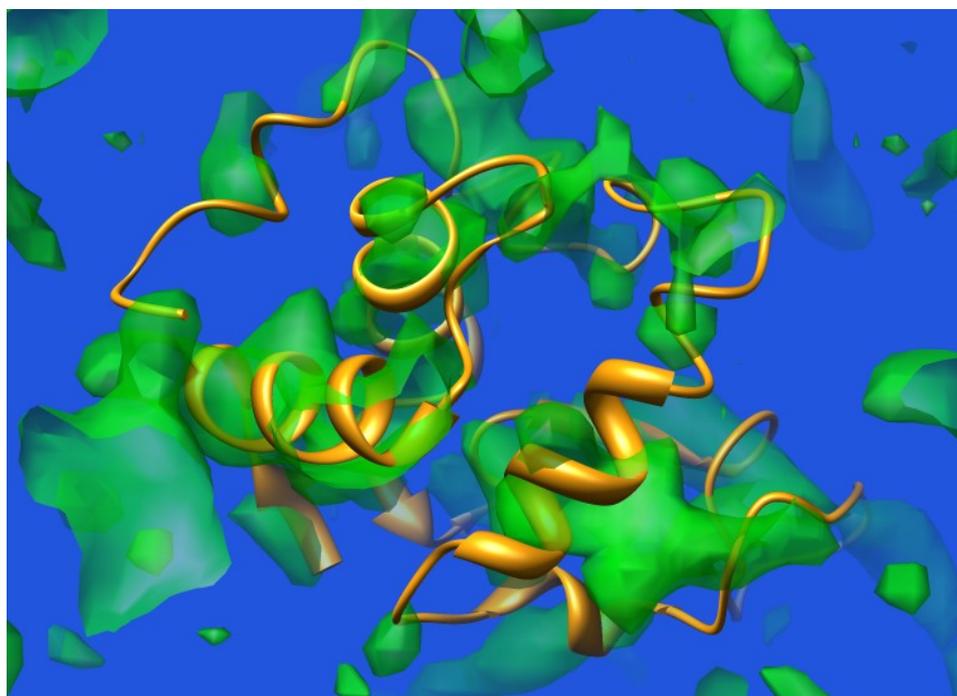
As part of the preliminary SIR analysis, initial heavy-atom positions for both the Gd-HEWL and U-PPE samples were found. In the former, the gadolinium atoms were located by direct methods as implemented in the software SHELXD (Sheldrick 2010). The program delivered a solution for two Gd sites, which were found to correspond to the known positions of these atoms in the reported HEWL-Gd-Hp-Do3A structure (Girard et al. 2002). In the PPE derivative, isomorphous difference Patterson maps revealed clear interatomic peaks in the Harker sections, as shown in Figure 2.9. These peaks are in agreement with a single uranium site in the asymmetric unit, the position of which was subsequently verified using the reference data measured on the equivalent single-crystal sample (§ 2.2.3).



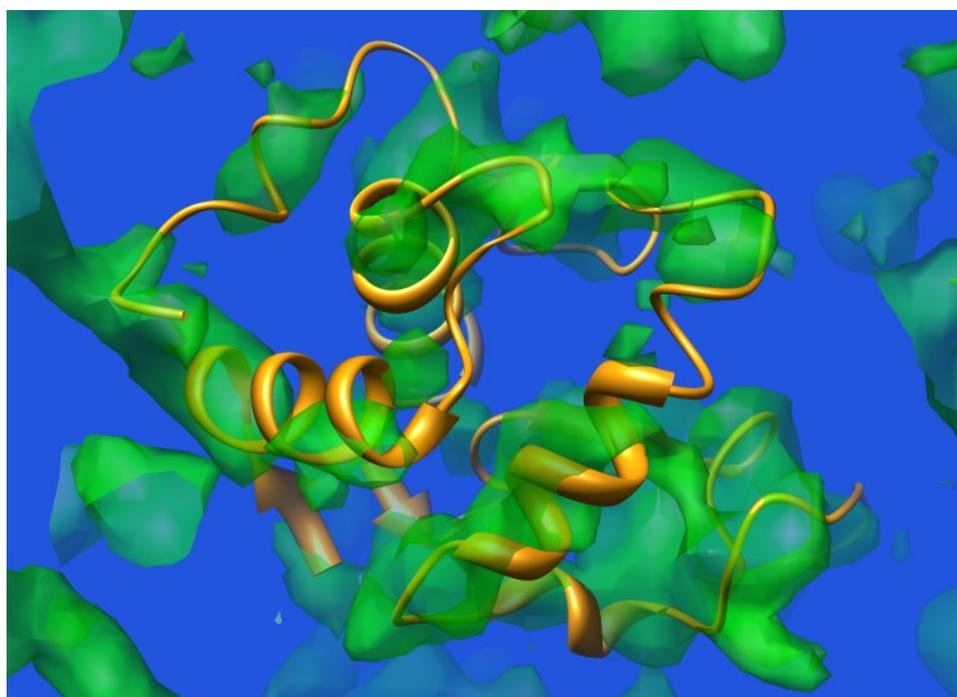
**Figure 2.10** Isomorphous difference Fourier map (green), contoured at the  $5.5\sigma$  level, superimposed onto the known molecular structure (silver) of HEWL (Protein Data Bank identifier 6lyt; Young et al. 1993), revealing the position at which the holmium atoms bind to the protein molecule.

For the succeeding MIR study, the location of holmium atoms in the second HEWL-derivative was found using neither method mentioned above. Instead, an isomorphous difference Fourier map was computed using the phase information obtained from the SIR experiment with the gadolinium sites. This consisted in calculating an electron-density map with the extracted intensities from the native sample and one with those of the holmium derivative, both using the phase information generated with the SIR experiment. The former was then subtracted from the latter to reveal strong peaks in the remaining density which corresponded to the electron density of holmium atoms. In this manner, one site per asymmetric unit was identified. Upon comparison with the known molecular model of HEWL, as shown in Figure 2.10, the site was discovered to be situated in the protein catalytic cleft, close to the Asp52 side chain residue, and in agreement with the second most occupied site reported by Jakoncic et al. (2006).

Heavy-atom refinement and phasing were performed by maximum likelihood techniques as implemented in the program SHARP (Fortelle & Bricogne 1997, Bricogne

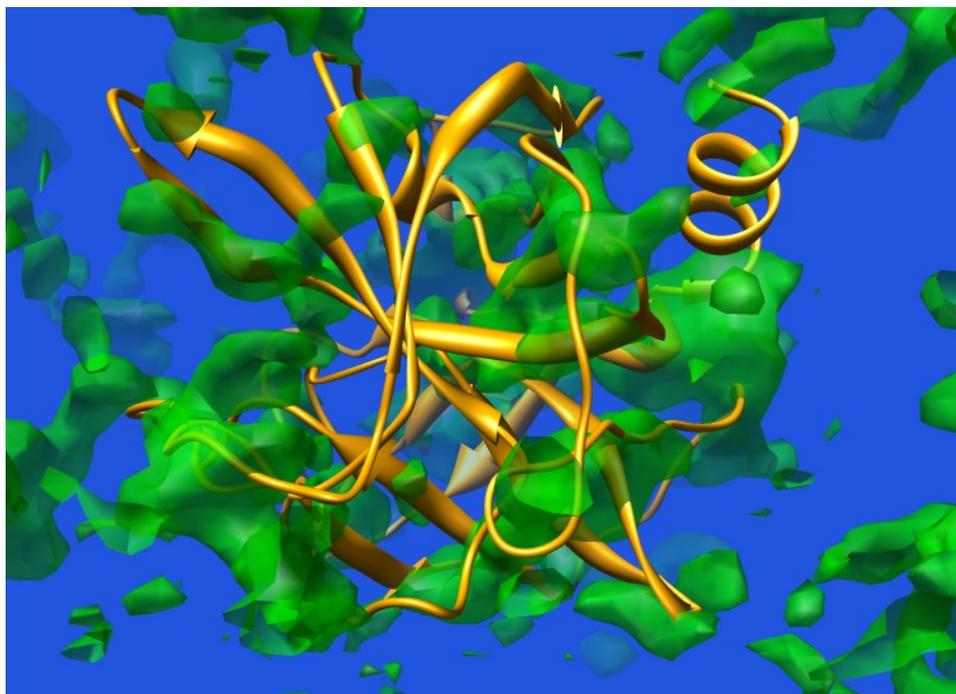


(a)



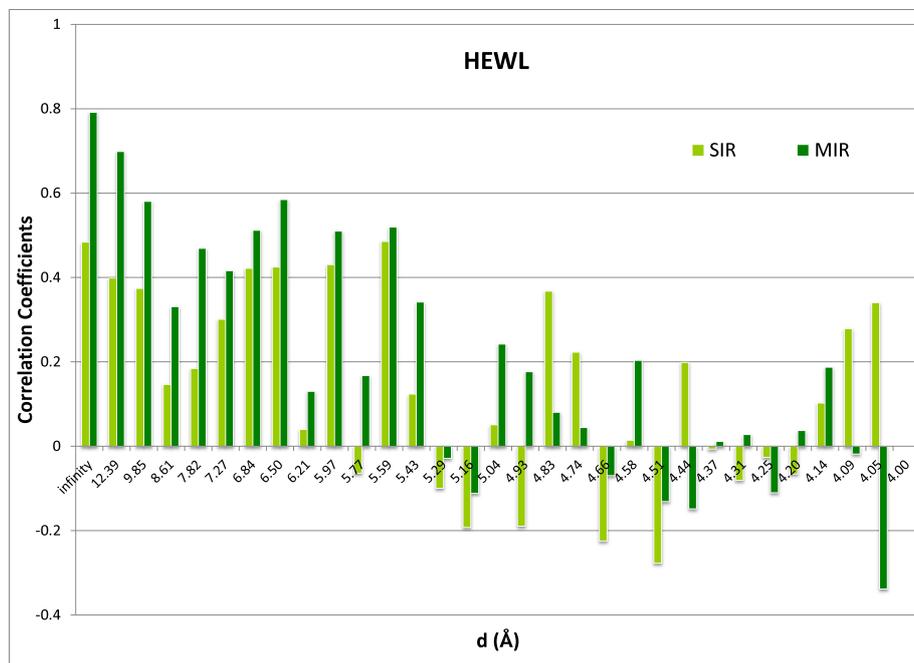
(b)

**Figure 2.11** Electron-density maps of HEWL (green) obtained from SIR data (top) and MIR data (bottom) superimposed onto the known molecular structure (orange) of HEWL (Protein Data Bank identifier 6lyt; Young et al. 1993).

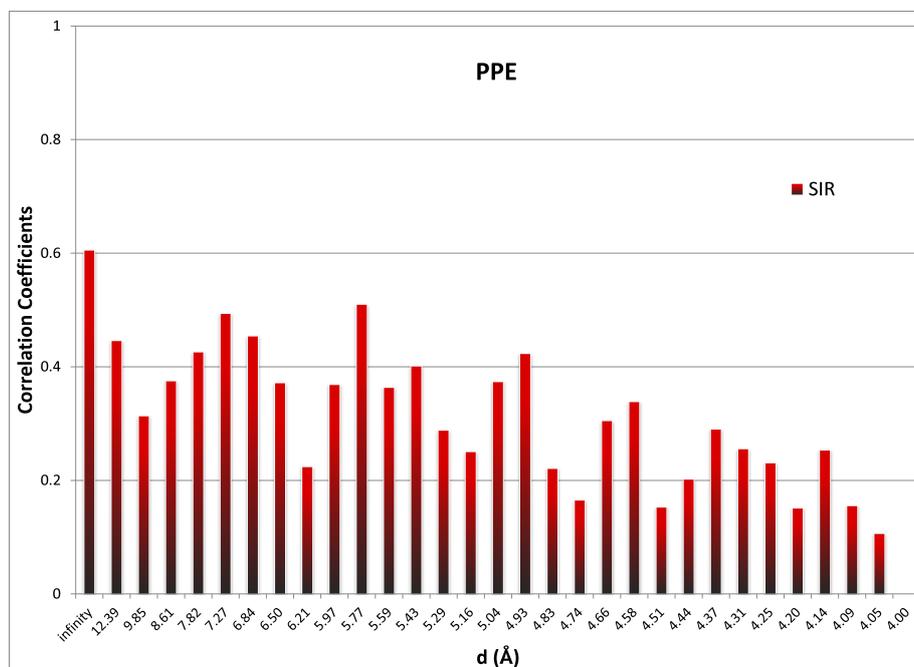


**Figure 2.12** Electron-density map of PPE (green) obtained from SIR data superimposed onto the known molecular structure (orange) of PPE (Protein Data Bank identifier 1lvy; Schiltz et al. 1994).

et al. 2003). Among other parameters, the atomic position and occupancy of each site was refined. However, given the limited resolution of the data, the atomic displacement parameters of the heavy atoms were held fixed. Once refined, the final heavy-atom structural models were used to compute a phase probability distribution for each reflection. Initial electron-density maps (prior to applying density modification methods, see § 2.3.4) were calculated using, as Fourier coefficients, the centroids of the probability distribution of each structure factor afforded by SHARP. The SIR maps (Figure 2.11(a) and 2.12) were particularly noisy mainly due to the bimodality of the phase probability distributions for acentric reflections inherent to the SIR method (§ 2.1.1.2). Not surprisingly, the quality of the MIR map of HEWL is much improved, as can be seen from Figure 2.11, given that it resolves the aforementioned bimodal problem for many reflections. This improvement is also reflected in Figure 2.13 which depicts the correlation coefficients between these experimental maps and maps computed from the known molecular structures. Evidently, for HEWL the SIR method has enabled some phase information to be retrieved up to a resolution limit of roughly 8.5 Å, which was pushed back to about 5.3 Å with the use of the MIR method. Similarly, Figure 2.14 shows that in the case of PPE a certain degree of phase information was recovered to approximately 5 Å.



**Figure 2.13** Correlation coefficients (computed in reciprocal space) between maps computed from the known molecular structure of HEWL (Protein Data Bank identifier 6lyt; Young et al. 1993) and maps computed with SIR and MIR experimental phases.



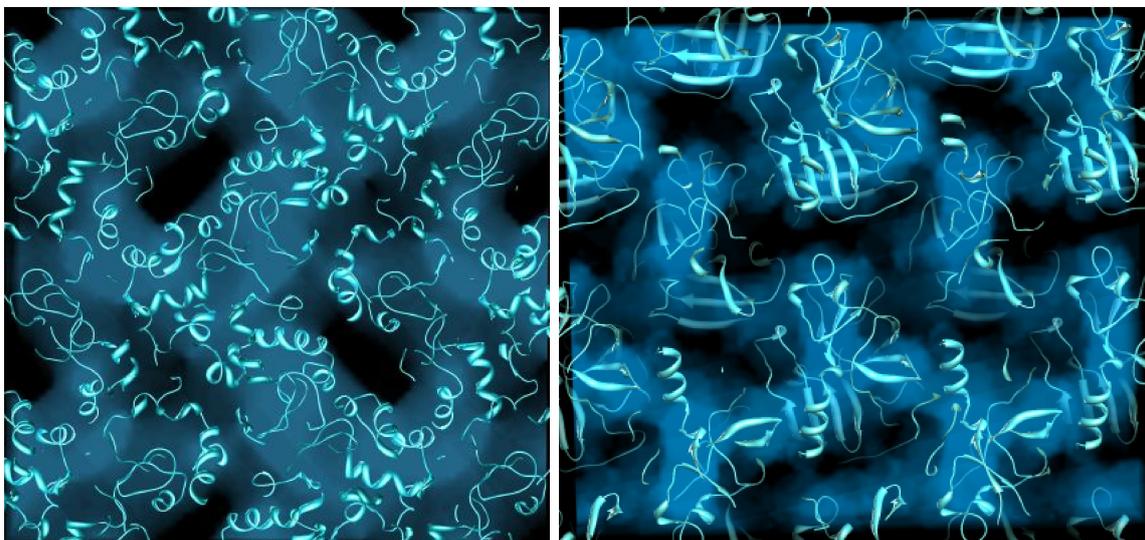
**Figure 2.14** Correlation coefficients (computed in reciprocal space) between maps computed from the known molecular structure of PPE (Protein Data Bank identifier 1lvy; Schiltz et al. 1994) and maps computed with SIR experimental phases.

### 2.3.4 Density modification

The aim of density modification techniques is to improve, and/or generate new, experimental phase estimates through the use of constraints applied to the electron density. This is generally achieved using an iterative approach which is separated into real and reciprocal space. A weighted map is initially calculated from the experimentally obtained structure-factor amplitudes and phases, to which real-space constraints are applied. The resulting modified map is subsequently used in an inverse Fourier transform to generate a modified set of amplitudes and phases. These phases are then weighted, on the basis of a comparison between the initial and modified amplitudes, and combined with the initial ones to produce a new set of phases. With these, and the initial amplitudes, a second weighted map is computed to start the second cycle of the iterative process, which frequently converges to significantly improved phases (Wang 1985).

As part of the present project, all density modification procedures were carried out using the programs *DM* (Cowtan 1994) and *SOLOMON* (Abrahams & Leslie 1996) from the *CCP4* package. Solvent flatness, available in both programs, was one of the real-space constraints used to modify the density of the experimental map. This constraint is employed to identify the protein-solvent boundary based on the following properties of an electron density map of a protein crystal: the local averages and fluctuations of the electron density are markedly higher in the regions occupied by the protein than in those occupied by the solvent. In both programs, applying these properties as constraints roughly translates to performing two steps. In the first step, a smoothed map is computed by calculating either the mean density (Wang 1985), in *DM*, or the standard deviation (Abrahams & Leslie 1996), in *SOLOMON*, at each point in the experimental map over a surrounding sphere of radius  $R$ . A threshold is then applied to this smoothed map so as to produce a binary mask. The second step involves applying the mask to the experimental map, essentially dividing it into two regions: the protein and solvent regions. Subsequently, all density points within the region delineated as solvent by the mask are either set to the expected solvent value, in solvent flattening, or modified according to a  $\gamma$ -correction factor (Abrahams 1997, Cowtan 1999), in solvent flipping.

The second real-space constraint employed was that of histogram matching, available only in *DM*. The aim of histogram matching is to bring the distribution of electron-density values of a map (called the density histogram) as close as possible to those of an ideal map. It was shown that the ideal density histogram of several protein systems is dependent on resolution, the atomic displacement parameters and the errors on the estimated

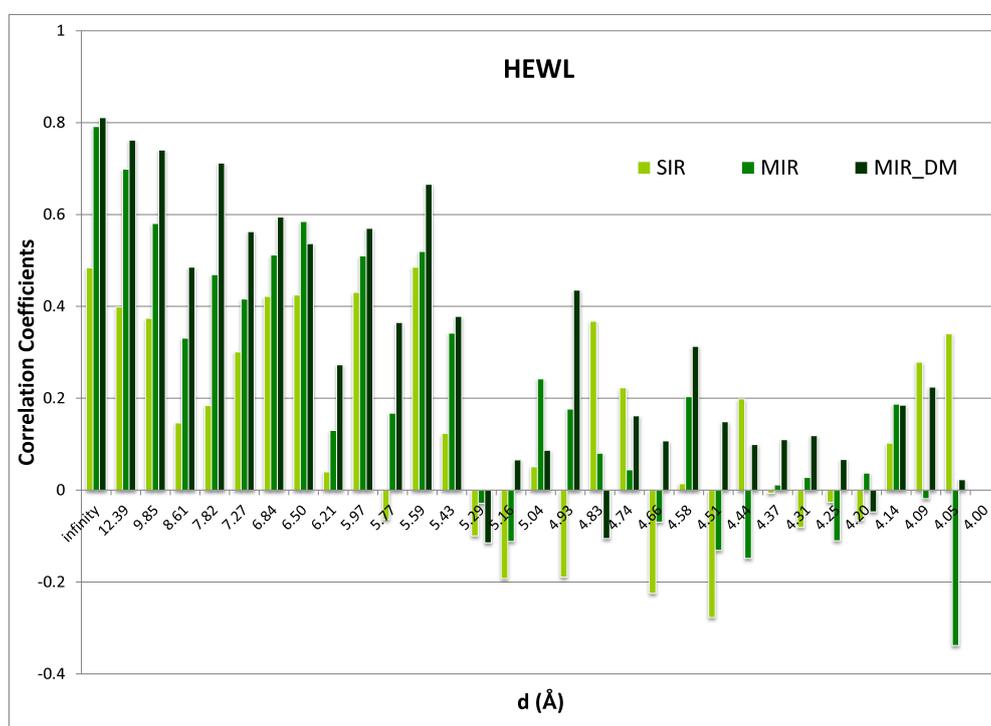


**Figure 2.15** Molecular envelopes derived from the SIR experimental data for HEWL (left) and PPE (right). The envelopes are represented as semi-transparent surfaces superimposed onto the known molecular structures depicted as main-chain ribbon models (Protein Data Bank identifier 6lyt (Young et al. 1993) for HEWL, 1lvy (Schiltz et al. 1994) for PPE). Both are viewed down the crystal  $c$ -axis, with  $a$  vertical in-plane and  $b$  horizontal in-plane for PPE.

phases, while independent of atomic coordinates (Zhang & Main 1990). The sensitivity to the phase errors enables histogram matching to be used as a real-space constraint in density modification algorithms. Furthermore, the lack of correlation between the density histogram and the structural conformation of the protein allows the prediction of the ideal histogram for unknown structures.

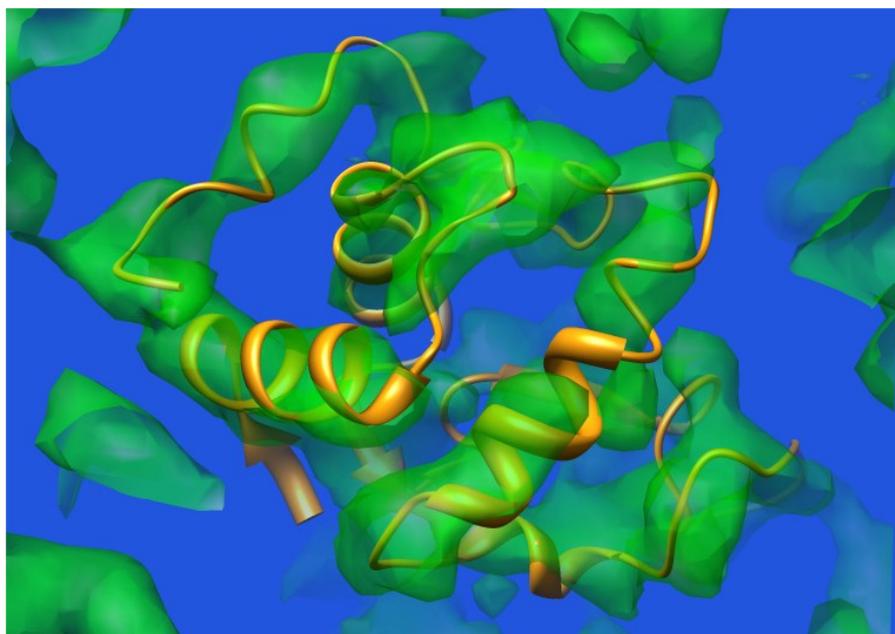
The cyclic procedure of density modification using solvent flatness as a constraint was applied to the SIR data of both HEWL and PPE. However, given the relatively poor quality of the starting estimated phases, no significant improvement was observed. Nevertheless, the solvent masks generated during the application of the real-space constraint proved to be essentially correct. As can be seen in Figure 2.15, the molecular envelopes describe the roughly globular shape of the protein molecules relatively well. Moreover, the HEWL masks allowed for the solvent channels traversing the crystal parallel to the  $c$ -axis to be approximated to 15 Å in diameter. Along the two remaining directions, the packing of protein molecules is such that only narrow and tortuous channels are present. Similarly, in PPE, continuous spiral channels were found along all three crystallographic axes with smallest diameters ranging from 8 to 14 Å.

The application of density modification to the MIR data for HEWL was, in comparison, much more successful. Indeed, running *DM* and *SOLOMON* in solvent flipping mode

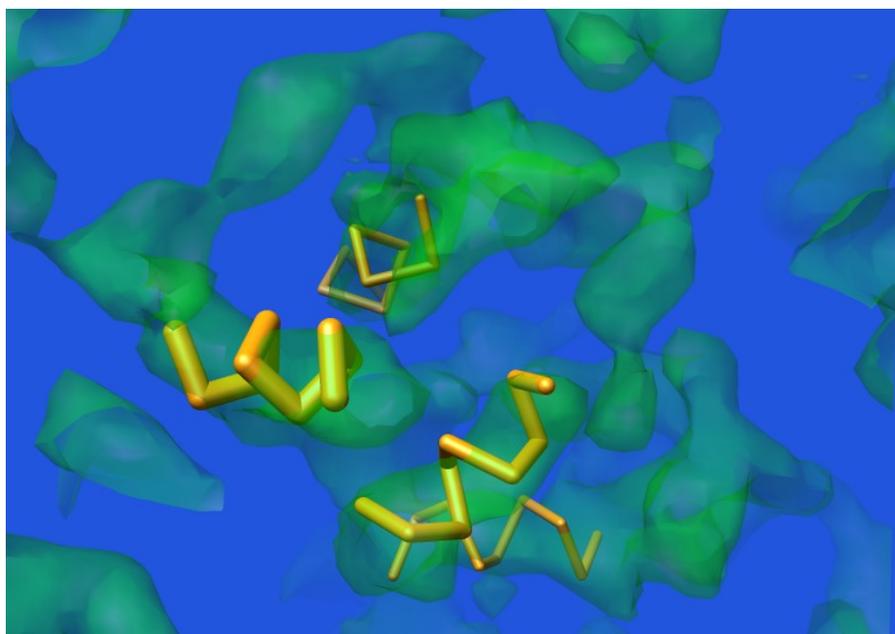


**Figure 2.16** Correlation coefficients (computed in reciprocal space) between maps computed from the known molecular structure of HEWL (Protein Data Bank identifier 6lyt; Young et al. 1993) and maps computed with experimental phases recovered with SIR, MIR and MIR after density modification (MIR\_DM).

both improved the initial estimated phases. As a result, other real-space constraints were tested instead of, or in combination with, solvent flatness, giving rise to variable levels of amelioration of the recovered phases. In the end, however, the algorithm which lead to the greatest improvement was the one in which the two programs were combined: two cycles of *DM*, in histogram-matching and solvent-flipping mode, followed by one cycle of *SOLOMON* in solvent-flipping mode. Although alternative routines running more than one cycle of *SOLOMON* were attempted, the general trend seemed to be that increasing the number of cycles performed somewhat deteriorated the quality of the derived phases. Nevertheless, the effectiveness of the chosen routine can be observed when comparing the columns for MIR and MIR\_DM in Figure 2.16. Although the limit in resolution at which phase information was recovered was not lowered, the quality of phase angles above that limit was greatly improved. This improvement in quality is reflected in the subsequent electron density map, shown in Figure 2.17.



**Figure 2.17** Electron-density map of HEWL (green) obtained from MIR data after density modification superimposed onto the known molecular structure (orange) of HEWL (Protein Data Bank identifier 6lyt; Young et al. 1993).



**Figure 2.18** The four helices (orange) found by the program *FFEAR* superimposed on the electron-density map for HEWL (green) obtained from MIR data after density modification. The thickness of each helix is representative of the accuracy of the fit between the search target and the electron density: the thicker the helix, the lower its score and the more accurate the fit.

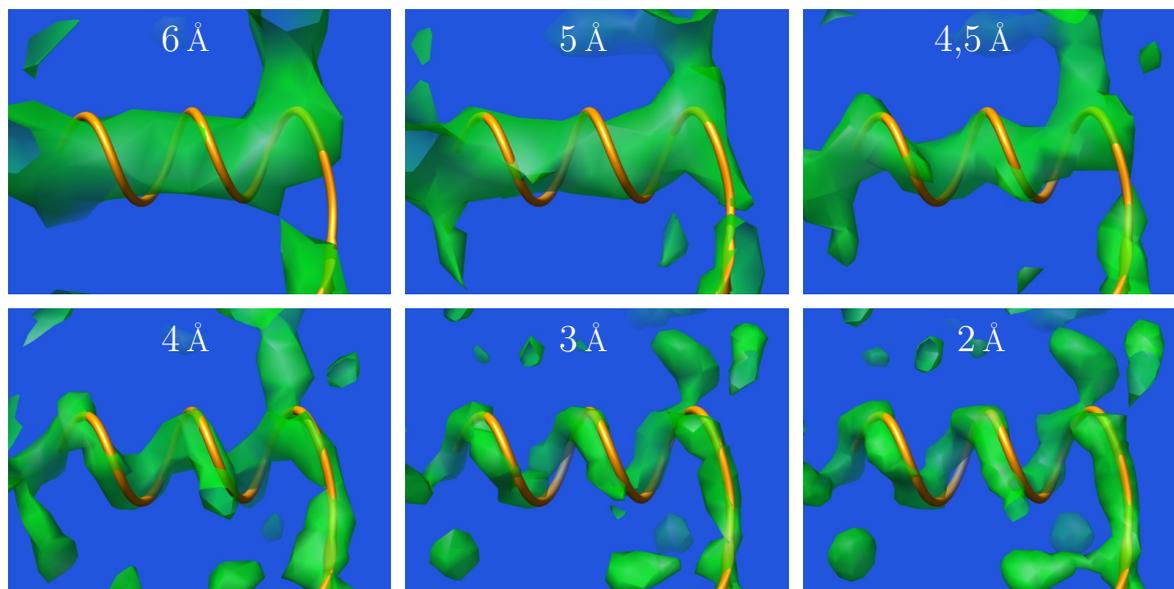
### 2.3.5 Detection of secondary-structure elements

In electron density maps of low to medium resolution, the primary structure of a protein molecule is inaccessible. Hence, in such maps, the focus turns to larger features of a polypeptide chain, namely, elements of the secondary structure. The bulkiest of those elements are the  $\alpha$ -helices, which, using the *CCP4* program *FFFEAR* (Cowtan 1998), can be located at a resolution of 5 Å or higher. Considering that the phase determination was successful to a resolution of roughly 5.3 Å, a search for helices was attempted using the map obtained from MIR after density modification. *FFFEAR* is a program which uses ‘representative’ maximum likelihood search targets to detect molecular fragments in electron density maps. More specifically, the search targets provided are helical fragments of 9 residues at various resolutions. Using a 6 Å resolution helical search target, *FFFEAR* found all four helices present in HEWL (Blake et al. 1967). The results of the search are shown in Figure 2.18 where the thickness of each helix is representative of its score (an agreement function based on the mean squared difference between the model and map over a masked region): the thicker the helix, the lower its score and the more accurate the fit. If compared to Figure 2.17, it is visible that the helices found by the search program are well located.

In light of the poor quality of the density maps obtained using the SIR method, the above search was not attempted on SIR data. Also, the PPE molecule contains only one  $\alpha$ -helix, in a position at which there is very little density in the corresponding electron density map (Figure 2.12).

### 2.3.6 Handedness of protein molecules

The chirality (‘hand’) of protein molecules can be a hurdle in X-ray crystallographic structural analyses which employ the method of isomorphous replacement (Harker 1956) not supplemented by anomalous scattering data. A set of heavy-atom coordinates and its inverse set will yield equally good solutions using Patterson or direct-methods techniques giving rise to a twofold ambiguity. One way to determine which set will ultimately generate the electron density map representing the true structure of the protein molecule is to closely inspect the handedness of  $\alpha$ -helices. In an isocontour representation of the electron density, the direction of helices, as illustrated in Figure 2.19, only starts to be visible to the naked eye at a resolution of approximately 4.5 Å. However, secondary-structure-element search programs profit from the full information contained in an electron density, and therefore they can be employed to discriminate between the two solutions using maps



**Figure 2.19** A set of electron densities computed at various resolution using the structure-factor amplitudes and phases from the known molecular structure of HEWL (Protein Data Bank identifier 6lyt; Young et al. 1993) superimposed with the structural model itself. The model was truncated to show only the first helix in the sequence.

of slightly poorer resolution.

The above was not attempted on the electron density maps obtained using the SIR method, since the resolution gap between 4.5 Å and the achieved resolution was judged too vast. With no way to decide on one of the two possible solutions, the phase information generated as part of both SIR experiments was obtained by voluntarily selecting the known set of coordinates of the gadolinium and uranium substructures in the corresponding HEWL and PPE derivatives. Hence, as part of the MIR analysis on HEWL, the isomorphous difference Fourier map, calculated to determine the positions of heavy atoms in the Ho-derivative, revealed the holmium substructure with the correct handedness. In this case, however, the attained resolution of  $\sim 5.3$  Å was deemed high enough to attempt to use secondary-structure-element search programs to carry out a *de novo* determination of the true set of heavy-atom coordinates.

This was performed by repeating the procedures described from § 2.3.3 to § 2.3.5 using the HEWL-MIR data, while inverting the atomic coordinates of both derivatives and switching to the enantiomorphic space group  $P4_12_12$ . Using the same helical search target, only the three longest helices out of the four present in HEWL were found by *FFFEAR*. The fact that not all helices were found tended to indicate that the search was less successful when using the mirror image of the heavy-atom substructures in the

---

phasing procedure. This was confirmed by the range of scores (reminder: the lower the score, the better the fit) corresponding to each search: 69.51 – 70.91 *vs* 84.50 – 86.68 for the search with the correct and inverted coordinates, respectively. As a result, the level of information content recovered up to approximately 5.3 Å resolution was sufficient to determine which heavy-atom coordinates lead to the electron density representing the HEWL molecule of correct chirality.

## 2.4 Discussion

The results presented in this study have enabled us to evaluate the quality of the information that can be obtained from protein X-ray powder diffraction data. In addition, the information in this type of data proved to be accessible through the use of established single-crystal software applied to a list of intensities extracted from multiple powder-diffraction patterns.

Two different approaches were employed to induce anisotropic lattice changes in order to alleviate the problem of accidentally overlapping peaks. Subsequently, detailed comparisons between the data obtained from single-pattern and multi-pattern Pawley extractions have demonstrated that the latter can lead to a significant improvement in the quality of the intensities. Furthermore, this improvement was shown to be a consequence of the anisotropic changes in the unit-cell parameters as opposed to simply being due to increased statistics as a result of adding patterns to one another. On the other hand, however, the results do not allow for the quantification of a minimum in either the magnitude or the level of anisotropy in the lattice variations required for the multi-pattern Pawley intensity extractions to be successful.

While the experimental data extend to Bragg spacings below 4.0 Å, the counting statistics and especially peak overlap were such that the phasing of the data at relatively high resolution is rather poor. Nonetheless, at lower resolution, the intensity extraction and phasing using SIR data have been successful enough to recover some information to roughly 8.5 Å for HEWL and 5.0 Å for PPE. The discrepancy in resolution between the two can be accounted for by the difference in crystal systems and, more importantly, in the nature of the derivatives. HEWL samples were crystallized in a tetragonal space group ( $P4_32_12$ ) whereas those of PPE were synthesized in an orthorhombic one ( $P2_12_12_1$ ). The two equivalent directions in a tetragonal space group will inevitably lead to exact overlap of some symmetry-independent reflections (§ A.2). This is even the case for tetragonal space groups belonging to the Laue class of high symmetry ( $P4/mmm$ ), for which, de-

pending on the resolution, 10-20% of symmetry-independent reflections overlap exactly (Basso et al. 2005, suppl. mat.). On the other hand, no such overlap occurs in patterns obtained from orthorhombic polycrystallites. Thus, in comparison, the HEWL data set intrinsically contains slightly less measurable information purely due to symmetry effects. Furthermore, as the isomorphous replacement method is dependent on intensity variations, the more scattering power a derivative possesses – *i.e.* the more electrons its bound heavy-atom is comprised of – the more accurately these variations will be measured, and, in turn, the more successful the phasing will be. In light of this, the uranium derivative of PPE was significantly more efficient at phasing than the gadolinium derivative of HEWL, given that uranium atoms ( $Z=92$ ) have twenty-eight additional electrons with respect to gadolinium atoms ( $Z=64$ ). These considerations make the set of PPE data a favourable case for the method of isomorphous replacement.

Despite the above highlighted differences, the SIR method produced a molecular envelope accurately describing the approximately globular shape of both macromolecules. When considering protein crystals as materials possessing chemical and physical properties, rather than a mere required product to be used in an X-ray diffraction experiment, a solvent envelope can prove to be highly informative. Indeed, protein crystals are porous solids with large solvent channels that traverse the entire crystal lattice and through which substrates and products can be transported in and out. The width and topology of the pores can range from 10 Å to occasionally more than 100 Å and can considerably vary between crystal polymorphs as the protein molecules pack differently in one with respect to the other. These properties thus make protein crystals a unique class of molecular sieves which can be considered analogous to zeolites (Vilenchik et al. 1998) and be used in various applications such as catalysis and bioremediation (Margolin & Navia 2001). To characterize these features, the determination at high-resolution of the three-dimensional structure of the macromolecule itself (which usually does not change substantially between different polymorphs) is not necessary. However, it is essential to map out the solvent channels and to determine the packing of the protein molecules within the lattice. As was shown in § 2.3.4, this information is in fact included in molecular envelopes. A further point of interest concerns the inherent chirality of protein molecules. As porous materials, protein crystals have the distinct advantage of containing chiral molecules which can be used in the separation of enantiomers and in enantioselective catalysis (Lalonde et al. 1995, Persichetti et al. 1995, Zelinski & Waldmann 1997, St. Clair et al. 2000, Vuolanto et al. 2003). In spite of this, the chirality of protein molecules does not affect the low resolution features of the structure, since the sizes of pores and

---

solvent channels remain identical in a mirror-inverted image of the structure. In this respect, the fact that the SIR method only enabled low resolution protein structures to be phased up to a two-fold ambiguity concerning the chirality of the molecules, does not preclude the usefulness of the result, *i.e.* the molecular envelopes. Therefore, the method of SIR applied to protein microcrystals may be a well suited technique to deliver the essential structural information at the mesoscopic scale.

While the SIR method enabled the computation of a molecular envelope of HEWL, an electron density map in which elements of its secondary structure could be correctly located was obtained through a MIR analysis. Moreover, with the use of a secondary-structure-element search program, the MIR data proved to be of sufficient quality to distinguish which set of heavy-atom coordinates ultimately lead to an electron density map representing the HEWL molecule of correct chirality. Although promising, these results cannot be claimed to have been entirely obtained using a *de novo* methodological procedure. Indeed, the optimal density modification routine reported in §2.3.4 was in fact determined on the basis of map correlation coefficients between the density-modified experimental map and a map calculated from the known single-crystal structure. However, the analysis itself remains a *de novo* one since in much the same way as for the determination of the electron density that corresponded to the protein molecule of correct chirality, the best density modification protocol could have been identified based on the success of the secondary-structure-element search program. More specifically, the efficiency of a given density modification routine could have been established by judging the quality of the resulting electron density map using, as reference, the score of a helix search performed within that map.

## 2.5 Outlook

In consideration of the above, the improvement on switching from SIR to MIR is unequivocal. In the same line of thought, adding a third derivative to the phasing procedure, provided it could be crystallized, would improve the quality of the recovered phase-angle information. Yet, it is expected that the resolution limit reached in the MIR analysis is primarily due to the underlying fact that single-crystal phasing methods are not entirely adequate for the processing of powder diffraction data rather than the lack of any additional heavy-atom derivatives. Indeed, these methods assume that the errors are uncorrelated between different reflection intensities or amplitudes. However, in protein powder diffraction data, which displays significant peak overlap, this assumption is no longer jus-

tified. While it was shown to be possible to extract individual reflection intensities from a powder pattern, the error distributions of such data reveal a pattern of strong covariances. The information concerning these correlations is actually produced in the Pawley procedure in the form of the inverse of the error covariance matrix (David 2004, Wright 2004). However, this knowledge is lost when passing the data over to single-crystal phasing programs as they use a diagonal approximation. These approximations are, nevertheless, not essential (Bricogne 1991) and expressions for generalized multivariate likelihood functions, capable of accommodating arbitrary patterns of covariances between the various sources of errors of reflection data, have been devised (Bricogne 2000).

There is thus scope for improvement and it is anticipated that the implementation of the above ideas will lead to an increase in the attainable resolution of the resulting electron density maps from *de novo* phasing of protein powder data. This type of study would also benefit from an enhanced performance of the algorithms used for extracting reliable structural information from medium-resolution data (4.5 – 8 Å). As yet, these algorithms remain relatively scarce (Karmali et al. 2009, Brunger et al. 2009), but this is nevertheless expected to change as the gap between electron microscopy and high-resolution X-ray diffraction is bridged.

## Chapter 3

# Implementation of Ramachandran plot restraints in a global optimization algorithm

### 3.1 Introduction

A possible approach to exploit the information in a powder diffraction pattern is to extract single-crystal-like intensities in order to then be able to employ traditional single-crystal structure-solution methods to obtain structural information about the system under study (Chapter 2). However, such a strategy relies significantly on the successful partitioning of accidentally overlapping peaks in powder patterns and ultimately remains an adaptation of one type of data, powder, to methods designed for another type of data, single crystal.

Almost two decades ago, an alternative set of powder-specific structure determination techniques began to emerge, namely global optimization methods. Contrary to conventional single-crystal methods, global optimization strategies are based on algorithms which operate in direct-space and are thus also known as direct-space (or direct-space) methods. More specifically, in these techniques, the conformation, position and orientation of a structural model<sup>1</sup> of the system of interest are continuously varied in direct-space in order to minimize the discrepancies between the calculated and observed

---

<sup>1</sup>In this chapter a model refers to a molecular description based on known chemical quantities rather than a molecular configuration based on a previously determined crystal structure.

powder diffraction pattern. In the same way as in a Rietveld refinement (Rietveld 1969), such an approach entirely bypasses the problem of fortuitous peak overlap since the partitioning of partially overlapped peaks is derived directly from the structural model in a given configuration. Indeed, given a known unit cell and space group, atomic coordinates with corresponding atom types and a set of profile parameters obtained from a pattern decomposition procedure (*e.g.* a LeBail or Pawley decomposition), an intensity for each reflection in the pattern can be calculated irrespective of overlap. Nevertheless, the information ‘lost’ in exact peak overlaps, due to the one dimensionality of powder diffraction data, is unavoidable regardless of the method used. As a result of this reduced amount of information, which is often insufficient with respect to the information required to determine the crystal structure, global optimization strategies frequently incorporate *a priori* chemical knowledge. The most commonly exploited type of supplementary chemical information is that of molecular geometry, which can be included directly in the description of the structural model. One way to do this is to employ a rigid body approach whereby the molecule is expressed in terms of bond lengths, bond angles and torsion angles while setting those that are known (with a certain precision) to their respective values and defining the rest (typically only torsion angles) as variable parameters to be adjusted during the process of structure solution, as is discussed in detail in § 3.3. With the inclusion of such knowledge in the description of the molecular structure, the total number of variable parameters is reduced to a minimum. This aspect has been crucial to the success of global optimization methods in that the fewer the variables, the smaller the search domain to be explored and therefore the more likely it is the solution will be found.

The continuous developments in the various types of global optimization strategies have enabled the tackling of increasingly complex structure-solution problems in the field of structure determination from powder diffraction data (see for example Fernandes et al. 2007). A class of materials which has especially benefited from these methodological advances is that of molecular organic compounds and in particular pharmaceuticals. Indeed, the structures of pharmaceutical compounds have been determined using the direct-space methods of simulated annealing (David & Shankland 2008) and, to a lesser extent, of genetic algorithm (Shankland et al. 1998, Kariuki et al. 1999). Furthermore, the latter has also been successfully employed with another type of organic compounds, namely oligopeptides. More specifically, a total of three peptide structures have been elucidated using genetic algorithms applied to powder diffraction data: Phe–Gly–Gly–Phe (Tedesco et al. 2000), Piv–L-Pro–Gly–NHMe (Tedesco et al. 2001) and Piv–L-Pro– $\gamma$ -Abu–NHMe (Cheung et al. 2002).

---

In spite of their wide success, global optimization techniques cannot yet be considered as being routinely applicable to powder diffraction data for structure-solution purposes. This is especially true in the context of peptide molecules as is demonstrated by the fact that the three aforementioned cases represent, to the best of our knowledge, the only such structures solved with direct-space methods. Further evidence of this can be found in the example of the structure determination of the heptapeptide of sequence Gly–Asn–Asn–Gln–Gln–Asn–Try. Indeed, X-ray powder diffraction data for this peptide, from both laboratory (Balbirnie et al. 2001) and synchrotron (Diaz-Avalos et al. 2003) instruments, had been available prior to the obtention of a crystal of sufficient size for a single-crystal analysis to be performed. Yet, the powder data was used only to produce an approximate  $\beta$ -sheet model of the heptapeptide. The main limitation in exploiting the powder data quoted by the authors (Diaz-Avalos et al. 2003) was that of peak overlap, and it was not until X-ray single-crystal data were collected that a refined structure was established (Nelson et al. 2005). In addition to the importance of structural investigations of oligopeptides, as they can prove vital in gaining insight into the structural properties of polypeptide sequences in protein molecules (Kaul & Balaram 1999), the example of the heptapeptide highlights the need that exists for direct-space methods to be further developed so as to become a tool that can contribute more significantly to the structure determination of relatively short polypeptide chains.

As it has already been mentioned, the reduction in the number of variable parameters, with the inclusion of *a priori* knowledge on molecular geometry, greatly improves the efficiency of global optimization strategies. This knowledge is available for polypeptide chains and the conformational flexibility of such molecules can thus be restricted to its variable torsion angles (§ 3.3.1). Yet, as the number of these angles can rapidly increase as a function of chain length, the structure determination of relatively long oligopeptides with direct-space methods can prove too complex an optimization problem. It is in this respect that the use of additional *a priori* knowledge, namely that contained in the Ramachandran plot (§ 3.4), is herein proposed to be implemented in a simulated annealing protocol as a set of restraints which operate on torsion-angle pairs of amino acids constituting a polypeptide chain expressed as a rigid body.

## 3.2 Global optimization methods

The level of success that global optimization methods have known in the field of structure determination from powder diffraction data (SDPD) can mostly be accounted for by their

ability to incorporate an extensive amount of prior chemical knowledge into the process of structure solution. Furthermore, substantial interest, from a methodological standpoint, has been generated as a result of the notion that any global optimization strategy can be employed in SDPD. This stems from the fact that they all essentially share the same fundamental principles described below.

A structural model of the crystal structure of interest is continuously adjusted in direct-space to generate a theoretically exhaustive set of trial crystal structures independently of the diffraction data. After each adjustment, the resulting structural candidate is assessed based upon the agreement between the calculated and observed powder patterns. The aim is to maximize this agreement, which can be quantified by a cost function based on the full diffraction pattern such as the weighted profile R-factor,  $R_{wp}$ , or the correlated integrated intensity function,  $\chi^2$ . Minimizing this cost function to obtain the best agreement is analogous to finding the deepest well (*i.e.* global minimum) among a multitude of somewhat shallower wells (*i.e.* local minima) of a multi-dimensional hypersurface. The size (dimensionality) and shape of this hypersurface is defined by the number and nature, respectively, of the structural variables employed to characterize the degrees of freedom of the crystal structure. If the set of structural parameters with the lowest corresponding cost-function score (a solution candidate) is located within the well containing the global minimum, the structural model described by these parameters will be of sufficient quality for a Rietveld refinement to determine the ‘true’ structure. Furthermore, the technique of powder diffraction benefits from the fact that the value of the cost-function obtained for a pattern decomposition procedure can be employed as a reference to judge the value of that obtained for the solution candidate and whether or not it is thus worth moving onto the refinement stage with this structural model.

Although a common base exists, there are many different types of global optimization techniques (David & Shankland 2008). In the field of SDPD, the most popular, so far, have been those of a stochastic nature which are based on a sequential or evolutionary algorithm. The former only operates on a single trial structure and includes simulated annealing, which is detailed in the following section (§ 3.2.1). The latter, on the other hand, maintains a randomly generated population of trial structures and is founded on the Darwinian principles of natural evolution (Michalewicz 1996). Each structure in the population is treated as an individual possessing a genetic code defined by the structural variables of the structural model. Through ‘time’, subsequent generations are produced in which the individuals of one generation are created *via* mutation and recombination of the members in the preceding generation. According to their fitness, defined by their

associated cost-function value, the fittest will survive and procreate. In this manner, the overall fitness of the subsequent generations will improve until one or more members of the population ultimately reach the global minimum. Evolutionary techniques which have been successfully applied to SDPD include genetic algorithms (Kariuki et al. 1997, Shankland, David & Csoka 1997, Albesa-Jove et al. 2004, Pan et al. 2006) and differential evolution (Seaton & Tremayne 2002, Tremayne et al. 2002, Chong et al. 2006). These strategies differ in their method of mutation and recombination as well as in their selection process of the fittest members of a generation.

### 3.2.1 Simulated annealing

The method of simulated annealing (SA) was independently described by Kirkpatrick et al. (1983) and Černý (1985), and its name originates from the parallel that exists between its algorithm and the physical process of annealing. This process consists in forming a crystalline solid from a melt by controlled cooling. Provided the cooling is slow enough, at each temperature,  $T$ , the particles in the melt arrange themselves so that the system reaches thermal equilibrium according to a probability,  $P(E_{eq})$ , given by the Boltzmann distribution

$$P(E_{eq}) \propto \exp\left(-\frac{E_{eq}}{k_B T}\right), \quad (3.1)$$

where  $E_{eq}$  is the energy of the system at thermal equilibrium for a given temperature and  $k_B$  is the Boltzmann's constant. As the temperature is decreased down to zero, the Boltzmann distribution favours the lower energy states of the system and ultimately the state of minimum energy. If the temperature is decreased too rapidly, then the resulting solid can be 'frozen' into one of many metastable crystalline or amorphous structures (*i.e.* in a local minima).

In 1953, Metropolis et al. devised an algorithm which simulated the thermal equilibration process at a fixed temperature called the Monte Carlo (MC) method. It is on this method that the SA algorithm is founded, with the additional feature of an annealing schedule giving rise to the above analogy. The MC method requires the definition of a set of configurations, a cost function and a generation mechanism. In the context of SDPD, these correspond to conformations of the molecule of interest, a profile R-factor or the correlated integrated intensity function and a structural model which contains variable parameters, respectively. With this, a sequence of trial structures can be generated as potential structure solutions, where each structure is derived from the preceding one by

applying shifts, random in both size and direction, to the variables of the model. The newest structure in the sequence is then accepted either if its cost-function value,  $CF_i$ , is lower than that of its predecessor,  $CF_{i-1}$ , or with a probability given by

$$\exp\left(-\frac{CF_i - CF_{i-1}}{k_B T}\right) > R, \quad (3.2)$$

where  $R$  is a random number within the range of 0 to 1, if its cost-function value is higher than that of its predecessor. The probabilistic tolerance of an increase in the cost-function value is essential to allow the algorithm to escape from local minima. In this way, the molecular conformation with the lowest corresponding cost-function value, for a given temperature, will be found provided a sufficiently large number of iterations is performed.

Taking the above into consideration, the SA method can be viewed as a series of consecutive MC algorithms evaluated at a sequence of decreasing temperatures as defined by the annealing schedule. Indeed, as the temperature is lowered in steps, the system is allowed to approach equilibrium in each step, until the temperature reaches a value at which virtually no further moves leading to an increase in the cost-function value are accepted. At this point, the algorithm will no longer be able to escape the well in which it finds itself and thus terminate at the corresponding minimum. If the same structure solution is obtained with multiple SA runs, there is a high probability that the global minimum has been reached.

The widespread success of the SA method in the field of SDPD (Deem & Newsam 1989, Andreev et al. 1997, Putz et al. 1999, Pagola & Stephens 2000) is mainly due to its ease of implementation and use. This can be corroborated by the multitude of powder dedicated software in which the method has been implemented (Černý & Favre-Nicolin 2007), such as: *TOPAS-Academic* (Coelho 2004), *DASH* (David et al. 2006) and *FOX* (Favre-Nicolin & Černý 2002). In the context of the work presented herein, the software of choice was *TOPAS-Academic* for the simplicity with which restraints can be incorporated into it as penalty functions.

### 3.3 Rigid bodies

Any given atomic structure can be described by basic building blocks ranging from atoms to entire molecules and passing through polyhedra and molecular fragments. The size

---

of these rigid building units will mainly depend on the availability of *a priori* structural knowledge on the molecule of interest, in terms of molecular geometry, as well as on its intrinsic flexibility.

One approach by which information about the intramolecular connectivity can be incorporated in the description of a molecular model, is to employ a z-matrix construct. This method consists in defining the position of atoms with respect to one another using arbitrary atomic labels and the internal coordinates of a molecule, namely bond distances, bond angles and torsion angles (also called dihedral angles). Such a parameterization is illustrated in Table 3.1 for a hypothetical molecule<sup>2</sup>. Once each atom is defined with internal coordinates, the molecule can be built within a Cartesian coordinate system according to the following three-step convention: (i) the first atom is placed at the origin, (ii) the second atom is positioned at a distance  $b_1$  from the first atom to define the  $z$ -axis and (iii) the third atom is placed at a distance  $b_2$  from either the first or second atom, so as to define the  $xz$ -plane, while forming an angle  $\tau_1$  with them. In the established coordinate system, all subsequent atoms will be set at a distance  $b_n$  to one of the previously described atoms, making an angle  $\tau_n$  with that atom and another, while also forming a torsion angle  $\theta_n$  with these two atoms and a fourth one. In this manner, the number of parameters required to parametrize a molecule containing  $N$ -atoms is  $3N-6$ , as there are  $b(N-1)$ ,  $\tau(N-2)$  and  $\theta(N-3)$  parameters. Finally, a further set of external degrees of freedom is needed to fully describe a molecule, expressed as a z-matrix, in direct space. These are the three rotational and the three translational parameters which enable the atomic matrix to be correctly positioned and oriented within the unit cell. From this point onwards, they will be referred to as the ‘6MR’ parameters, since those are the sought-after-variables in a molecular replacement (MR) algorithm<sup>3</sup>.

In traditional crystallography, the complexity of the structure-solution problem is evaluated based on the number of structural parameters describing the position of atoms

---

<sup>2</sup>The nomenclature used in this section for bond lengths, bond angles and torsion angles, as well as the convention used to compute torsion angles, are the standard nomenclature and conventions for the description of polypeptide conformation proposed by the IUPAC-IUB Commission on Biochemical Nomenclature (1970).

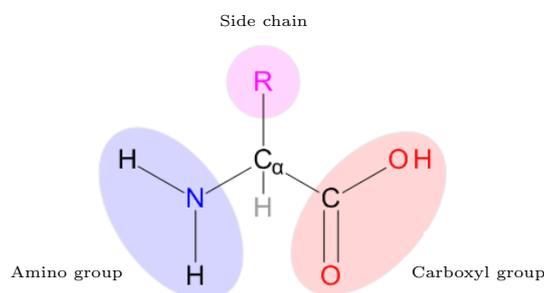
<sup>3</sup>For polar space groups, translational parameters can be fixed in order to define an origin. This is due to the fact that polar space groups are non-centrosymmetric and contain one or several unique polar directions. As a result, systems which crystallize in such space groups can be expressed using only 5 to 3 MR parameters as variables.

**Table 3.1** The general formula of a z-matrix.

Site 1	Site 2	Distance	Site 3	Angle	Site 4	Torsion angle
A						
B	A	$b(A,B)$				
C	B	$b(B,C)$	A	$\tau(A,B,C)$		
D	C	$b(C,D)$	B	$\tau(B,C,D)$	A	$\theta(A,B,C,D)$
E	D	$b(D,E)$	C	$\tau(C,D,E)$	B	$\theta(B,C,D,E)$
...						

within the asymmetric unit. More specifically, for an  $N$ -atom molecule a set of  $x, y, z$  coordinates is assigned to each individual atom giving rise to a problem consisting of  $3N$  variables. However, when using the z-matrix approach with global optimization methods in the context of SDPD, while the total number of parameters will remain  $3N$  ( $(3N-6) + 6MR$ ), the number of variables can potentially be dramatically reduced. Indeed, if, for example, all bond lengths and bond angles of a molecular structure are known, the degree of conformational freedom of that molecule will only be dependent on variables representing the torsion angles which cannot be pre-defined based on prior knowledge. The complexity of the structure-solution problem will then solely depend on these torsion angles and the  $6MR$  parameters.

Despite the ability of the z-matrix construct to efficiently reduce the number of parameters to a minimum, alternative methods have been developed for cases where the inherent rigidity of a z-matrix can sometimes prevent the search domain to be properly explored. One such method was put forth by Favre-Nicolin & Černý (2004), in which atomic positions are defined with a set of  $x, y, z$  coordinates while the conformation of the molecule is statistically imposed through a set of restraints. A more recent alternative was suggested by Zhou et al. (2007), which consists in fragmenting the molecule into multiple z-matrices so as to reduce the number of torsion angles at the expense of an increased total number of variables (each additional fragment is associated with a supplementary set of  $6MR$  parameters). While both approaches successfully improved the efficiency of the structure-solution process for flexible molecules, their benefits are not as significant for intrinsically rigid molecules such as inorganic structures and molecules composed of amino acid building blocks.

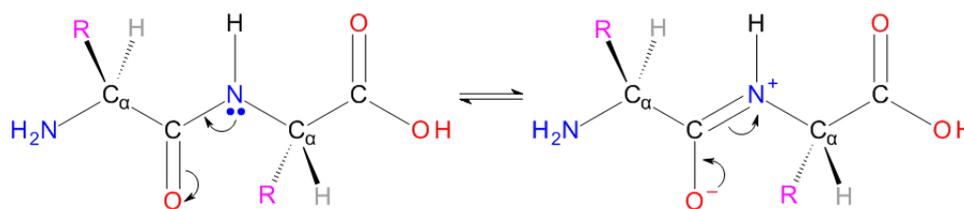


**Figure 3.1** Generic structure of  $\alpha$ -amino acids in their neutral form.

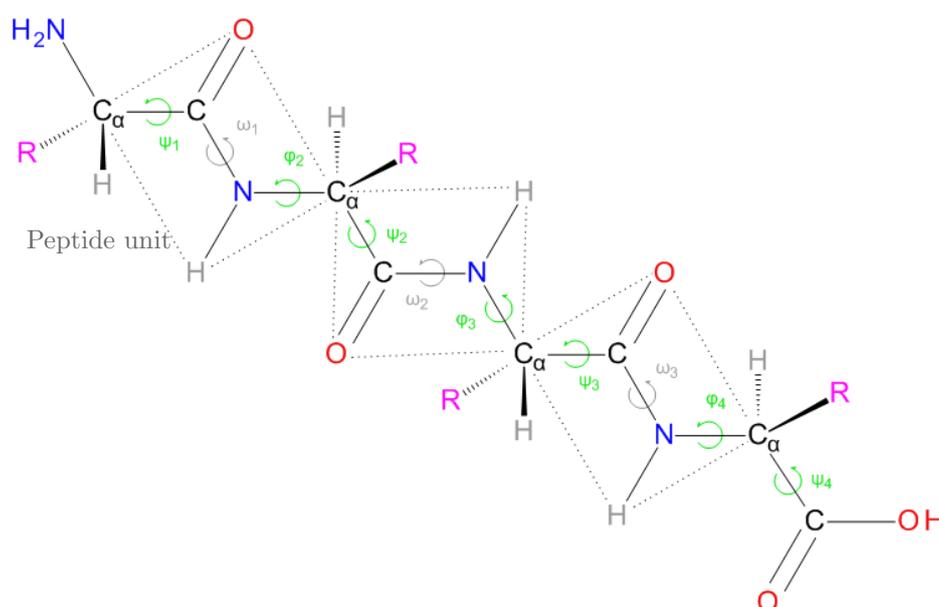
### 3.3.1 Amino acids

All amino acids share the same basic structure (Figure 3.1), each containing an amine group and a carboxylic acid group. In  $\alpha$ -amino acids, the most commonly found in nature, the R-group, referred to as the side chain, is bonded to the  $\alpha$ -carbon atom ( $C_\alpha$ ). The side chain can be varied to generate all 22 proteinogenic amino acids, *i.e.* those naturally incorporated into polypeptide chains. The nomenclature for labeling side-chain carbon atoms is to continue down the Greek alphabet so that the first carbon atom of the R-group bonded to  $C_\alpha$  will be named  $C_\beta$ , the second  $C_\gamma$  and so on. With the exception of glycine, for which the R-group is a hydrogen atom, all proteinogenic amino acids are chiral. Chirality inevitably leads to a pair of enantiomers which are denominated as L- and D-amino acids (IUPAC-IUB Joint Commission on Biochemical Nomenclature 1984), where the L-configuration represents the stereo-configuration of the naturally occurring amino acids. In polypeptide chains, amino acids are linearly bonded to each other through a peptide bond, which, during the polymerisation process, is created *via* a condensation reaction between the carboxyl carbon of one amino acid and the amine nitrogen of another amino acid. As a result of the partial double-bond nature of this bond, accounted for by the resonance structures in Figure 3.2, the peptide unit is planar (Figure 3.3). Furthermore, the *trans*-configuration of the peptide bond is highly favoured over the *cis*-configuration. This is due to the energetically less favourable interactions between  $C_\alpha$  atoms of adjacent amino acids in the latter configuration. In accordance with the previously mentioned set of conventions, the *trans* and *cis* forms of all torsion angles are defined as  $180^\circ$  and  $0^\circ$ , respectively.

As a result of the extensive knowledge available on the intramolecular connectivity of amino acids (Engh & Huber 1991), the z-matrix construct is readily applicable. Consequently, when expressing a polypeptide chain as a z-matrix, its degrees of freedom consist



**Figure 3.2** Resonance structures for a generic dipeptide accounting for the partial double-bond characteristics of the peptide bond.



**Figure 3.3** Representation of a generic polypeptide chain comprised of four amino acids showing the various backbone torsion angles as well as the peptide planes (dotted-lined rectangles).

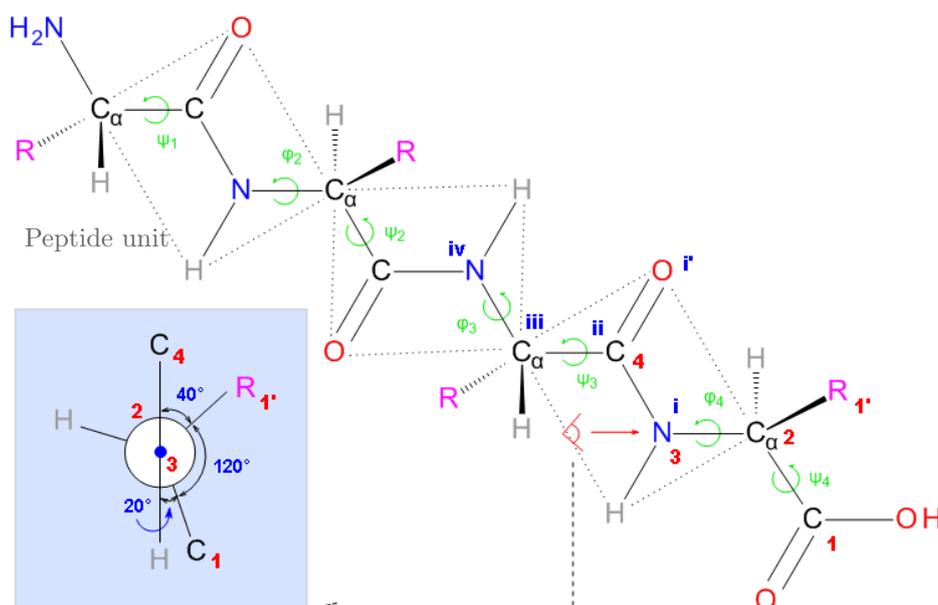
of the torsion angles alone (not taking into account the 6MR parameters). Additionally, assuming planarity and a *trans*-configuration of the peptide bond, the torsion angle around this bond, denoted  $\omega$ , can be fixed at  $180^\circ$ . Therefore, the conformational flexibility of the backbone of a polypeptide described as a *z*-matrix stems exclusively from the torsion angles on either side of the  $\alpha$ -carbon, defined as  $\varphi$  and  $\psi$  (Figure 3.3). Evidently, depending on the nature of the side chain, a variable number of additional degrees of freedom,  $\chi$  torsion angles, would contribute to the overall flexibility of the molecule. Although hydrogen atoms are not considered in the following reflection (for reasons discussed in § 3.6), they are included in all images of polypeptide chains for completeness.

In building the *z*-matrix of polypeptide molecules, its structure gives rise to an unavoidable ‘double’ definition of two out of the three backbone torsion angles. This is

irrespective of whether the molecule is being described from the N-terminus to the C-terminus or *vice versa* (see Appendix B). In the former case, for example, the ‘double’ definition concerns the torsion angles  $\varphi$  and  $\psi$  (the set of atoms only containing backbone atoms was taken as the default definition for both backbone torsion angles, *i.e.*  $\varphi = \theta(\text{C}^1, \text{C}_\alpha, \text{N}, \text{C}^4)$  and  $\psi = \theta(\text{N}^i, \text{C}, \text{C}_\alpha, \text{N}^{iv})$ ). Indeed, in a z-matrix the definition of a carbonyl carbon and that of the first atom of the R-group both define the torsion angle  $\varphi$  (illustrated by the red numbers in Figure 3.4). Similarly, the definition of a backbone nitrogen atom and of a carbonyl oxygen atom both describe the torsion angle  $\psi$  (illustrated by the blue numbers in Figure 3.4). In each case, the overparameterization can be avoided by expressing one definition in terms of the other. This can be done by determining the angular relationship between the planes formed by the sets of three consecutive atoms which are not common to both definitions.

In the case of the bimodal definition of  $\psi$ , the situation is straightforward as it involves two planes which both lie within the peptide plane:  $\tau(\text{N}^i, \text{C}, \text{C}_\alpha)$  and  $\tau(\text{O}, \text{C}, \text{C}_\alpha)$ . Therefore, the overparameterization can be resolved using the following relationship:  $\theta(\text{O}, \text{C}, \text{C}_\alpha, \text{N}^{iv}) = \psi \pm 180^\circ$ . This holds true for the  $\psi$  angle of the C-terminal amino acid ( $\psi_4$  in Figure 3.3) as well, since the carbon atom of the carboxyl group is  $\text{sp}^2$  hybridized and hence planar. The same approach can be utilized to determine the relationship between the two alternative definitions of  $\varphi$ . However, in this case, the two planes of interest –  $\tau(\text{C}^1, \text{C}_\alpha, \text{N})$  and  $\tau(\text{R}, \text{C}_\alpha, \text{N})$  – are not parallel to each other. Consequently, it is useful to draw a Newman projection, seen along the  $\text{N}^3$ - $\text{C}_\alpha^2$  bond, to serve as a visual aid. From such a projection (included in Figure 3.4), in which a hypothetical value of  $160^\circ$  for  $\varphi$  was taken, the following relationship can be deduced:  $\theta(\text{R}, \text{C}_\alpha, \text{N}, \text{C}^4) = \varphi - 120^\circ$ .

Taking the above into account, it is possible to accurately determine the number of conformational degrees of freedom for a polypeptide chain expressed as a z-matrix. In the simplest case, *i.e.* a polypeptide comprised of  $M$  glycine amino-acids, the number of conformational variables is  $2M-1$ . In comparison, the equivalent relationship obtained when simply assigning a set of  $x, y, z$  coordinates to each atom is  $12M+3$ . The factor 12 originates from the multiplication of 4 and 3, where 4 is the number of non-hydrogen atoms per amino acid, with the exception of the C-terminus amino acid which contains five atoms and is accounted for by the addition of the three supplementary degrees of freedom. With non-branched amino acids, other than glycine and alanine, the number of variables increases by 3, with traditional methods, while only by 1, with a z-matrix construct, for every additional atom contained in the side chain. The gap in total number of conformational variables when using a z-matrix rather than three dimensional atomic



**Figure 3.4** Representation of a generic polypeptide chain comprised of four amino acids showing the various backbone torsion angles as well as the peptide planes (dotted-lined rectangles). Superimposed are the two sets of numbers representing the alternatives for defining the torsion angles  $\psi$  (blue numbers) and  $\varphi$  (red numbers). Also included is a Newman projection seen along the  $N^3-C_\alpha^2$  to clearly visualize the relationship between the two different  $\varphi$  definitions.

coordinates, is even more significant for polypeptides consisting of branched amino acids and ones which contain cyclic groups.

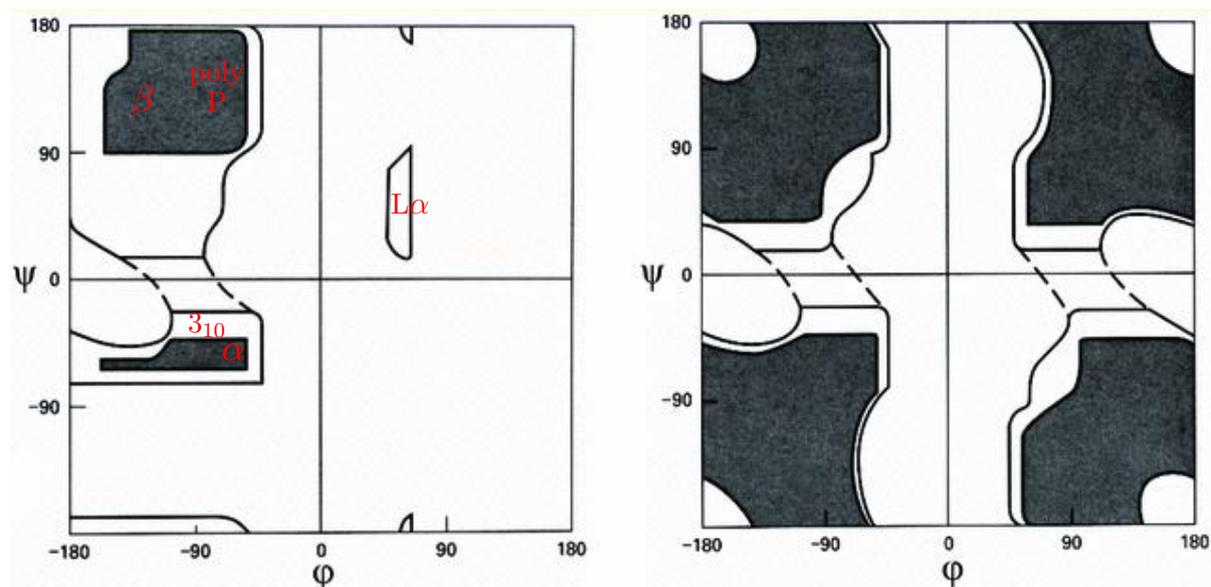
As a result, it is apparent that the z-matrix approach is a very efficient method for describing the structure of polypeptide molecules in a condensed manner. Yet, a pit-fall associated with this approach is the fact that multiple z-matrices exist to represent the same molecule. This can give rise to varying success rates of structure solution depending on the chosen z-matrix, as was empirically demonstrated in a powder diffraction study investigating the effect of algorithmic, crystallographic and molecular factors on the structure-solution process (Shankland et al. 2002). Nevertheless, in this study, the z-matrix parameterization was the method of choice to describe polypeptide structures as it provides an ideal framework to apply Ramachandran plot restraints on  $[\varphi, \psi]$  torsion angle pairs.

## 3.4 The Ramachandran plot

Almost 50 years ago, professor G. N. Ramachandran and co-workers began an investigation into the steric implications of considering a polypeptide chain as a set of consecutive planar peptide units, where the orientation of each unit, relative to the preceding one, is defined by the torsion angles ( $\varphi$  and  $\psi$ ) on either side of the atom ( $C_\alpha$ ) linking the pair of units (Ramachandran et al. 1963). Within the approximation of ideal molecular geometry for peptide units, a *trans*-configuration for the peptide bond and a hard-sphere model, the authors showed that large parts of the geometrically possible conformations ( $[\varphi, \psi]$  combinations) were stereochemically highly unfavourable. This preliminary result was expressed as a diagram, known as the Ramachandran plot (R-plot), of  $\varphi$  vs  $\psi$ .

As part of a succeeding in-depth study (Ramakrishnan & Ramachandran 1965), two separate R-plots were established: one for glycine and one for non-glycine amino acids. These are shown in Figure 3.5, where the two solid-line contours, obtained using slightly different radii of spheres, encapsulate shaded and unshaded regions that correspond to ‘allowed’ and ‘partially allowed’ conformations, respectively. An additional area of ‘possible’ conformations, defined by the dashed-line contours, can be considered when allowing for a slightly increased bond angle between the peptide units, namely  $\tau(NC_\alpha C)$ . Such diagrams were determined on the basis of interatomic contacts involving atoms in adjacent peptide units and the  $C_\beta$  atom. Glycine residues must therefore be treated as a separate case from non-glycine residues, since they have a hydrogen atom at the  $C_\beta$  atomic site which will give rise to substantially less atomic overlaps and, in turn, a comparatively increased number of permitted  $[\varphi, \psi]$  combinations. This is visually apparent from the plots and can be quantified by the fraction of the total area that represents permitted regions (all contoured regions) in each plot: 22.5% and 61.0% for non-glycine and glycine residues, respectively. In the R-plot of non-glycine amino acids, it was also shown that these regions were relatively unaltered when including the  $C_\gamma$  atom in the contacts calculation. As a result, the regions can be regarded as being largely side-chain independent. In addition, the conformation of most types of secondary structures observed in polypeptide chains was found to be located within these favourable areas, as shown in Figure 3.5. At this stage, it is important to note that the above considerations do not apply to  $[\varphi, \psi]$  combinations for residues at chain termini since the N- and C-terminus are not sterically hindered by preceding and succeeding amino acids, respectively.

Since its conception, the R-plot has been utilized as a powerful verification tool in single-crystal macromolecular crystallography to assess the validity of empirical models.

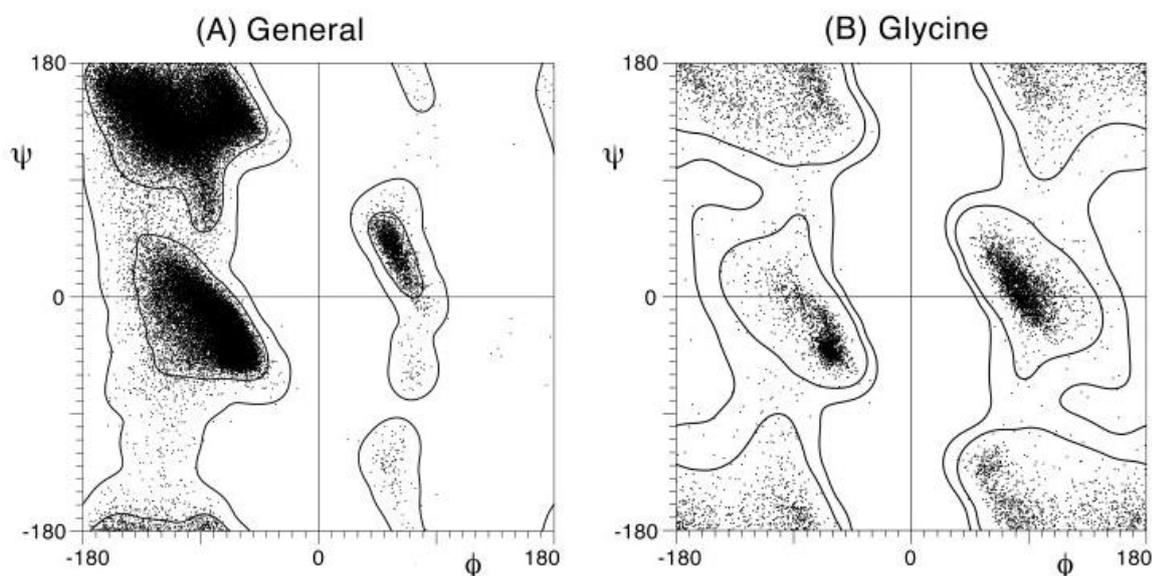


**Figure 3.5** Original Ramachandran plots for (left) non-glycine and (right) glycine residues. The shaded and unshaded regions, contoured with solid lines were obtained using slightly different radii of spheres and correspond to ‘allowed’ and ‘partially allowed’ conformations, respectively. The area, contoured with dashed lines, becomes possible when allowing for a slightly increased bond angle  $\tau(\text{NC}_\alpha\text{C})$ . The red labels show the approximate position of the various secondary-structures elements where:  $\alpha$  = right-handed  $\alpha$ -helix,  $3_{10}$  =  $3_{10}$  helix,  $\beta$  =  $\beta$ -sheet, polyP = polyPro and  $L\alpha$  = left-handed  $\alpha$ -helix.

However, over the years, the increase in both the number of solved protein structures and the resolution at which these are solved have encouraged the use of R-plots derived from experimental data. These R-plots are generated by surveying protein structure databases, such as the Protein Data Bank, for frequencies of occurrence of  $[\varphi, \psi]$  pairs. As a consequence of Boltzmann’s principle, it can be assumed that a  $[\varphi, \psi]$  pair with a high frequency will correspond to a low molecular energy and hence be probable. In this manner, an empirically derived R-plot can be converted into an energy function  $E(\varphi, \psi)$  according to the rearranged form of equation 3.1

$$E(\varphi, \psi) \propto k_B T \left[ -\ln \left( P(\varphi, \psi) \right) \right], \quad (3.3)$$

where  $P(\varphi, \psi)$  is the probability of a given torsion-angle pair. This type of potential function is referred to as a knowledge-based potential and has previously been used to routinely evaluate structure solutions of protein molecules in various programs and servers such as *WHATCHECK* (Hooft et al. 1996) and *PROCHECK* (Laskowski et al. 1993), and *MOLPROBITY* (Chen et al. 2010), respectively.



**Figure 3.6** Ramachandran plots from the Top500 database for (left) the general case and (right) glycine residues. The inner and outer contours correspond to ‘favoured’ and ‘allowed’ regions.

The set of protein structures, on which these potentials are based, is selected according to various filters. In *MOLPROBITY*, for example, the criteria consisted mainly of a cutoff value of 1.8 Å in resolution and of < 30 in backbone atomic displacement parameter (Lovell et al. 2003). Accordingly, a database of 500 structures containing 109799 amino acids was generated. This database is known as the Top500 database, the data set of which is freely available online<sup>4</sup>. With this database, redefined R-plots were produced for the general case (explained below) and for glycine, proline and pre-proline residues. Indeed, given the particularity of proline amino acids of having the nitrogen and the  $\alpha$ -carbon bonded within a single five-membered ring, its conformational freedom, as well as that of pre-proline residues, is greatly reduced in comparison to other non-glycine residues (MacArthur & Thornton 1991). As a result, proline and pre-proline amino acids were treated as separate cases in the Top500 database and the general case is thus representative of non-glycine, non-proline and non-pre-proline amino acids. The redefined R-plot for the general and glycine cases, calculated with approximately 81000 and 7700 residues respectively, are shown in Figure 3.6. In these plots, the inner and outer contours correspond to ‘favoured’ and ‘allowed’ regions which, respectively, envelop 98% and 99.95% of the data points. When comparing the images in Figures 3.5 and

<sup>4</sup><http://kinemage.biochem.duke.edu/databases/top500.php> (valid in May 2011)

3.6, it is possible to note that although the original R-plots contain most of the general features seen in the Top500 R-plots, some discrepancies exist. For the non-glycine case, these mainly consist of the protruding leg below the  $\beta$ -sheet regions and the majority of the areas in the  $[\varphi \geq 0, \psi \leq 0]$  quadrant. On the other hand, in the case of the glycine residues the observed difference is greater since the most ‘favoured’ regions in Top500 R-plot correspond to areas of ‘possible’ conformations in the original glycine R-plot.

In powder diffraction crystallography, the R-plot has not played a major role as a result of the powder technique being, as yet, not ideal for studying relatively large molecules such as polypeptide chains. Nevertheless, the R-plot has previously been exploited in a Rietveld refinement of a protein molecule (Von Dreele 1999). More specifically, the information content of the R-plot was expressed as a set of restraints in order to refine the tertiary structure of Metmyoglobin using powder diffraction data. In a similar way, such restraints could be employed to influence the process of model building in direct space methods. Although doing so could prevent the use of the R-plot as an *a posteriori* validation tool, it could potentially be extremely useful in guiding search algorithms of global optimization methods (§ 3.2).

### 3.4.1 2D Fourier series approximation

In order to include R-plot restraints in a minimization function, the information contained in an R-plot needs to be approximated. One possible approach is to fit the surface of a torsion-angle knowledge-based potential with a given set of functions. Using the knowledge-based potential incorporated into *PROCHECK*, Von Dreele (2005) approximated the resulting  $[\varphi, \psi]$  surface with a sum of two-dimensional Gaussians. However, the use of Gaussian functions results in the loss of the intrinsic periodicity of the R-plot space. In this study, the characteristic of periodicity was retained in the resulting approximation with the use of a two-dimensional Fourier series. The employed knowledge-based potential,  $E_{500}(\varphi, \psi)$ , was derived from the Top500 data set using equation 3.3. The approximation of this potential was thus performed according to

$$E_{500}(\varphi, \psi) = \sum_{h=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} C_{hk} \exp[2\pi i(h\varphi + k\psi)], \quad (3.4)$$

where  $C_{hk}$  are the Fourier coefficients. By applying Euler’s formula and expanding the terms in the summation, equation 3.4 can be re-written as

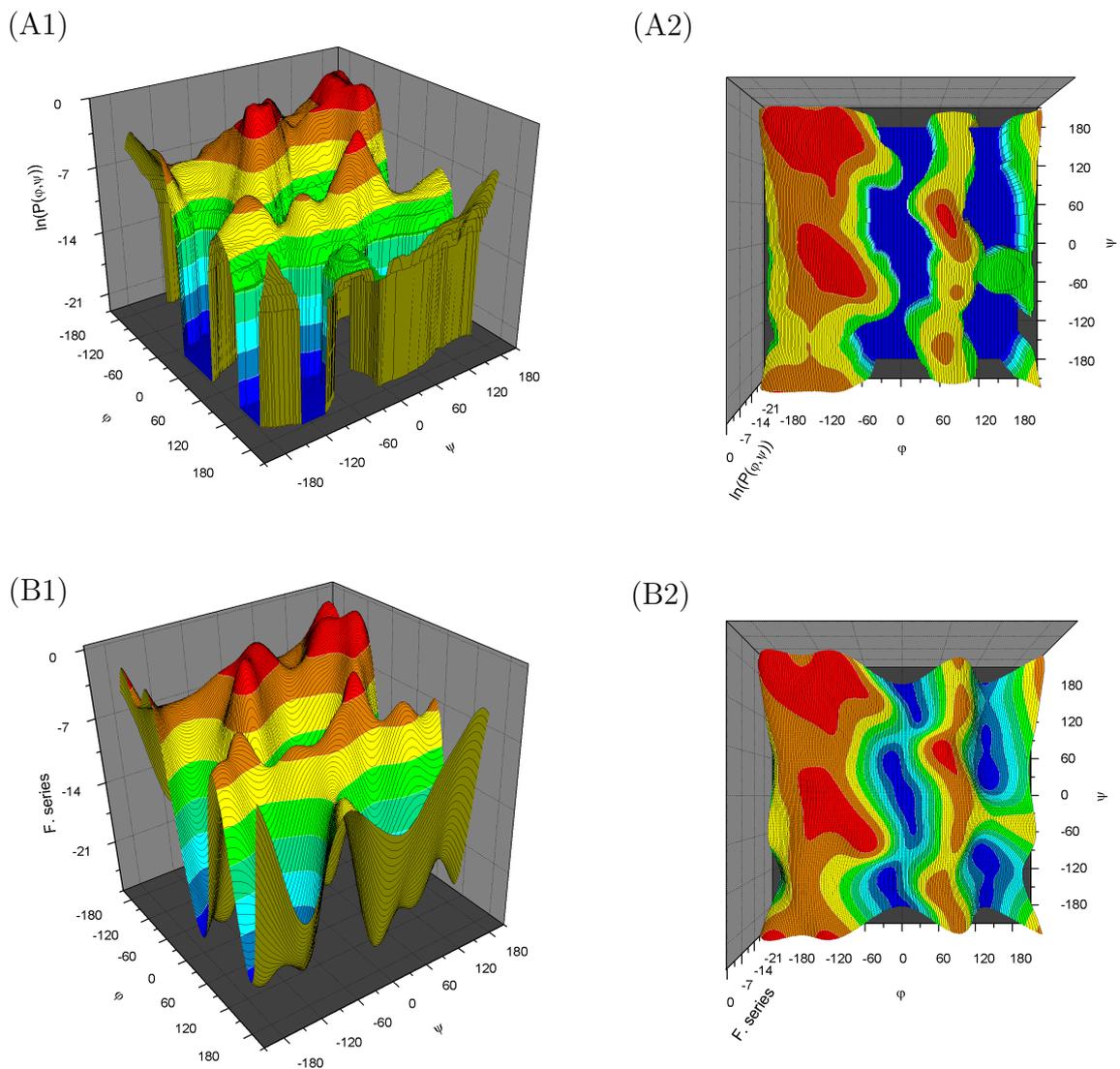
$$\begin{aligned}
E_{500}(\varphi, \psi) = C_{00} + \sum_{h=1}^{\infty} \sum_{k=1}^{\infty} & M_{hk} \cos(h\varphi) \cos(k\psi) \\
& + N_{hk} \sin(h\varphi) \sin(k\psi) \\
& + O_{hk} \cos(h\varphi) \sin(k\psi) \\
& + P_{hk} \sin(h\varphi) \cos(k\psi),
\end{aligned} \tag{3.5}$$

where

$$\begin{aligned}
M_{hk} &= C_{hk} + C_{\bar{h}k} + C_{h\bar{k}} + C_{\bar{h}\bar{k}} \\
N_{hk} &= -C_{hk} + C_{\bar{h}k} + C_{h\bar{k}} - C_{\bar{h}\bar{k}} \\
O_{hk} &= C_{hk} + C_{\bar{h}k} - C_{h\bar{k}} - C_{\bar{h}\bar{k}} \\
P_{hk} &= C_{hk} - C_{\bar{h}k} + C_{h\bar{k}} - C_{\bar{h}\bar{k}}.
\end{aligned} \tag{3.6}$$

From equation 3.5, it is possible to note that the zero order coefficient ( $C_{00}$ ) is a constant and can in turn be considered as a simple scale factor. Therefore, whether it is included in the Fourier series or not, the features of the approximation of the knowledge-based function will remain unchanged. In the present study, when referring to the number of coefficients used in a Fourier series calculation, the quoted number will not contain the zero order coefficient but the calculation will have been performed with it.

In order to approximate a function with a Fourier series, it is necessary to establish the number of coefficients needed to produce a satisfactory approximation and to determine the value of each of these coefficients. As part of a statistical analysis (Pertsemlidis et al. 2005), it was demonstrated that an accurate fit of the non-glycine R-plot (the authors made no mention to whether or not proline and pre-proline amino acids were included) can be achieved using only the first 80 coefficients of a Fourier series. This corresponds to a truncation of the series at order 4. In light of this, a first approximation attempt, using the Top500 data of the R-plot for the general case (general R-plot), was performed with 80 coefficients, the values of which were determined using a Least squares routine (see Appendix C). The obtained coefficients are reported in Table 3.2 while Figure 3.7 depicts two different orientations (B1 and B2) of the resulting approximation as well as the two corresponding orientations (A1 and A2) of the plot of the Top500 original data, for comparison. These four images clearly demonstrate the accuracy of the resulting approximation of the general R-plot with a two-dimensional Fourier series comprised of 80 coefficients. It is important to note that these plots were calculated on



**Figure 3.7** General R-plots, shown in two different orientation (1 and 2), calculated with (A) the Top500 database data and (B) the Fourier series limited to the coefficients of order  $\leq 4$ .

**Table 3.2** Values of the 80 Fourier coefficients used to approximate the general R-plot. In bold, are the values of the 10 highest-absolute-value coefficients. The symbol ‘–’ corresponds to a value of zero originating from a product containing a  $\sin(0)$  term. Although the zero-order coefficient is not included in the quoted numbers of coefficients, it was used in the Fourier series calculations with a value of  $C_{00} = 11.646969$ .

Order	Cumulative sum of the number of $C_{hk}$	h k	$M_{hk}$	$N_{hk}$	$O_{hk}$	$P_{hk}$
1	8	1 0	<b>2.186660</b>	–	–	<b>4.774439</b>
		0 1	-0.489094	–	-0.726836	–
		1 1	-0.389979	0.887410	0.492653	-0.381130
2	24	2 0	<b>4.542012</b>	–	–	<b>-4.639463</b>
		0 2	-0.134110	–	<b>2.223164</b>	–
		2 1	1.033534	-0.514767	0.287876	0.905685
		1 2	<b>2.197744</b>	<b>-1.935359</b>	-1.085766	-0.367026
		2 2	1.537064	<b>-4.759644</b>	-0.029157	0.311634
3	48	3 0	<b>3.589686</b>	–	–	1.480194
		0 3	0.314653	–	0.328429	–
		3 1	-0.187212	0.452896	0.629674	-0.227868
		1 3	-0.322735	0.886116	-0.320815	0.025303
		3 2	0.429013	<b>-2.083962</b>	-0.210571	0.027108
		2 3	0.358759	-0.564862	-0.108413	-0.073484
		3 3	-0.863038	0.371683	0.494425	0.351141
4	80	4 0	-0.371525	–	–	0.634790
		0 4	0.299917	–	0.039048	–
		4 1	0.722906	-0.414505	-0.971600	0.033938
		1 4	0.116726	0.358488	-0.236179	0.477044
		4 2	-0.573654	-1.399517	0.181831	-0.261189
		2 4	0.154788	0.152297	-0.073906	-0.198260
		4 3	-0.132589	0.036222	0.056887	0.047416
		3 4	0.319468	1.187904	-0.027105	0.303679
4 4	-0.514991	0.867724	-0.661937	-0.153835		

**Table 3.3** Values of the 80 Fourier coefficients used to approximate the glycine R-plot. In bold, are the values of the 10 highest-absolute-value coefficients. The symbols ‘-’ and ‘\*’ correspond to a value of zero originating from a product containing a  $\sin(0)$  term and the centrosymmetric nature of the plot (see text), respectively. Although the zero-order coefficient is not included in the quoted numbers of coefficients, it was used in the Fourier series calculations with a value of  $C_{00} = 8.269118$ .

Order	Cumulative sum of the number of $C_{hk}$	h k	$M_{hk}$	$N_{hk}$	$O_{hk}$	$P_{hk}$
1	8	1 0	<b>7.108806</b>	—	—	*
		0 1	0.785656	—	*	—
		1 1	<b>-1.215116</b>	0.497832	*	*
2	24	2 0	<b>6.357630</b>	—	—	*
		0 2	<b>-1.406775</b>	—	*	—
		2 1	<b>0.967572</b>	0.222431	*	*
		1 2	0.679734	<b>-2.811360</b>	*	*
		2 2	<b>0.951812</b>	<b>-4.140176</b>	*	*
3	48	3 0	<b>2.159292</b>	—	—	*
		0 3	-0.025312	—	*	—
		3 1	-0.173884	0.387610	*	*
		1 3	0.527057	0.154763	*	*
		3 2	0.735837	<b>-2.357535</b>	*	*
		2 3	-0.065952	-0.419039	*	*
		3 3	-0.623985	0.439479	*	*
4	80	4 0	0.249873	—	—	*
		0 4	0.304873	—	*	—
		4 1	0.201145	-0.617533	*	*
		1 4	0.885365	0.622105	*	*
		4 2	0.300605	-0.619451	*	*
		2 4	0.330433	0.866410	*	*
		4 3	-0.424458	0.169504	*	*
		3 4	0.190407	0.588112	*	*
4 4	-0.504869	0.370425	*	*		

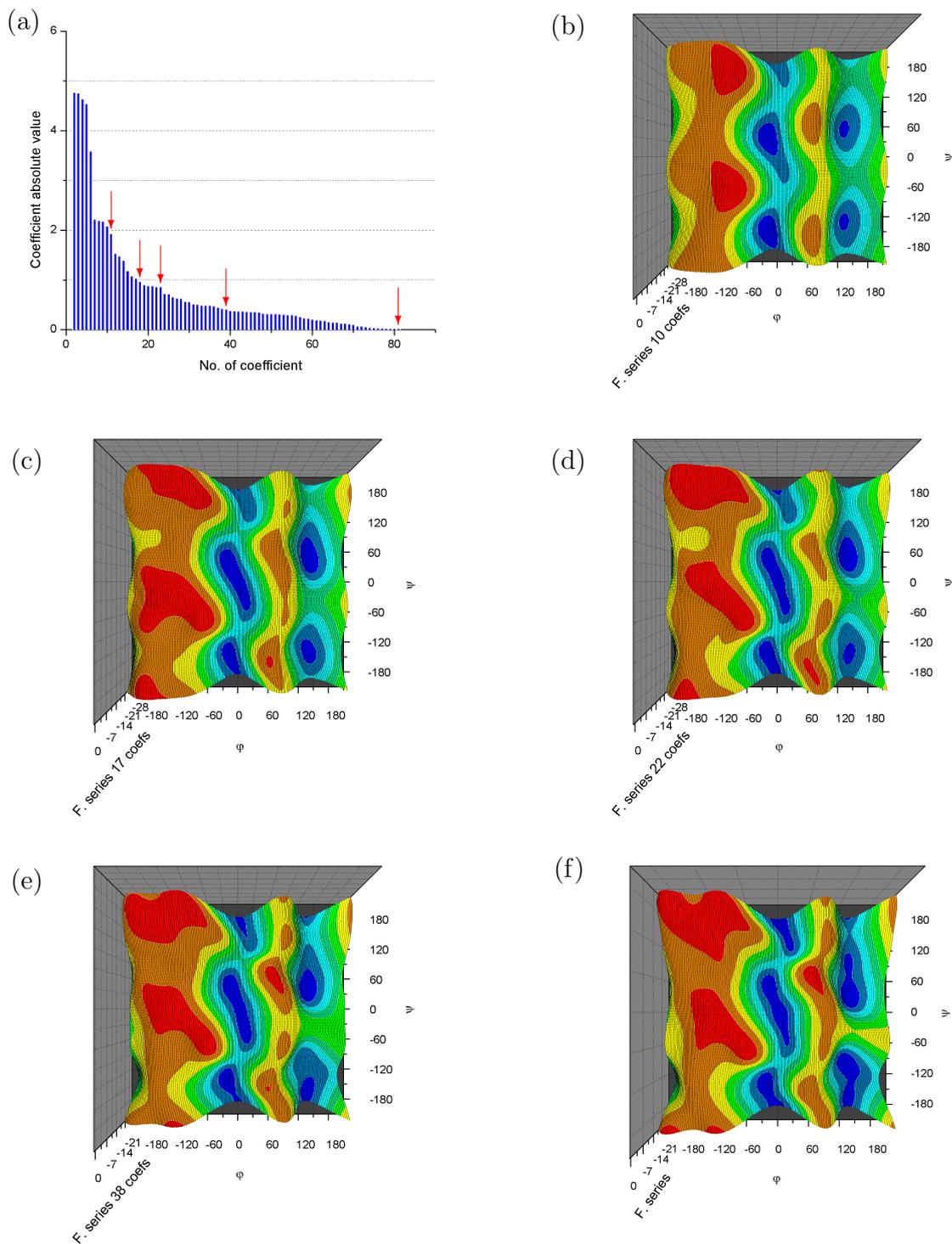
the basis of the opposite of the knowledge-based potential, (*i.e.*  $-E_{500}(\varphi, \psi)$ ). The reason for this is a purely visual one, as it is customary to see peaks for favourable regions of an R-plot rather than troughs. However, in a minimization function, it is  $E_{500}(\varphi, \psi)$  that needs to be employed in order for a low energy value to correspond to a favourable  $[\varphi, \psi]$  combination.

Although the above approximation of the general R-plot is more than satisfactory, including a term comprised of 80 coefficients in a minimization function could be computationally intensive and hence potentially slow down the global optimization process. Consequently, a series of general R-plots were calculated using a decreasing number of low-absolute-value Fourier coefficients, as shown in Figure 3.8. From these images, it is possible to see that a Fourier series composed of the 10 highest-absolute-value coefficients (highlighted in bold in Table 3.2) generates an approximation which accurately represents the general R-plot at a gross level of detail.

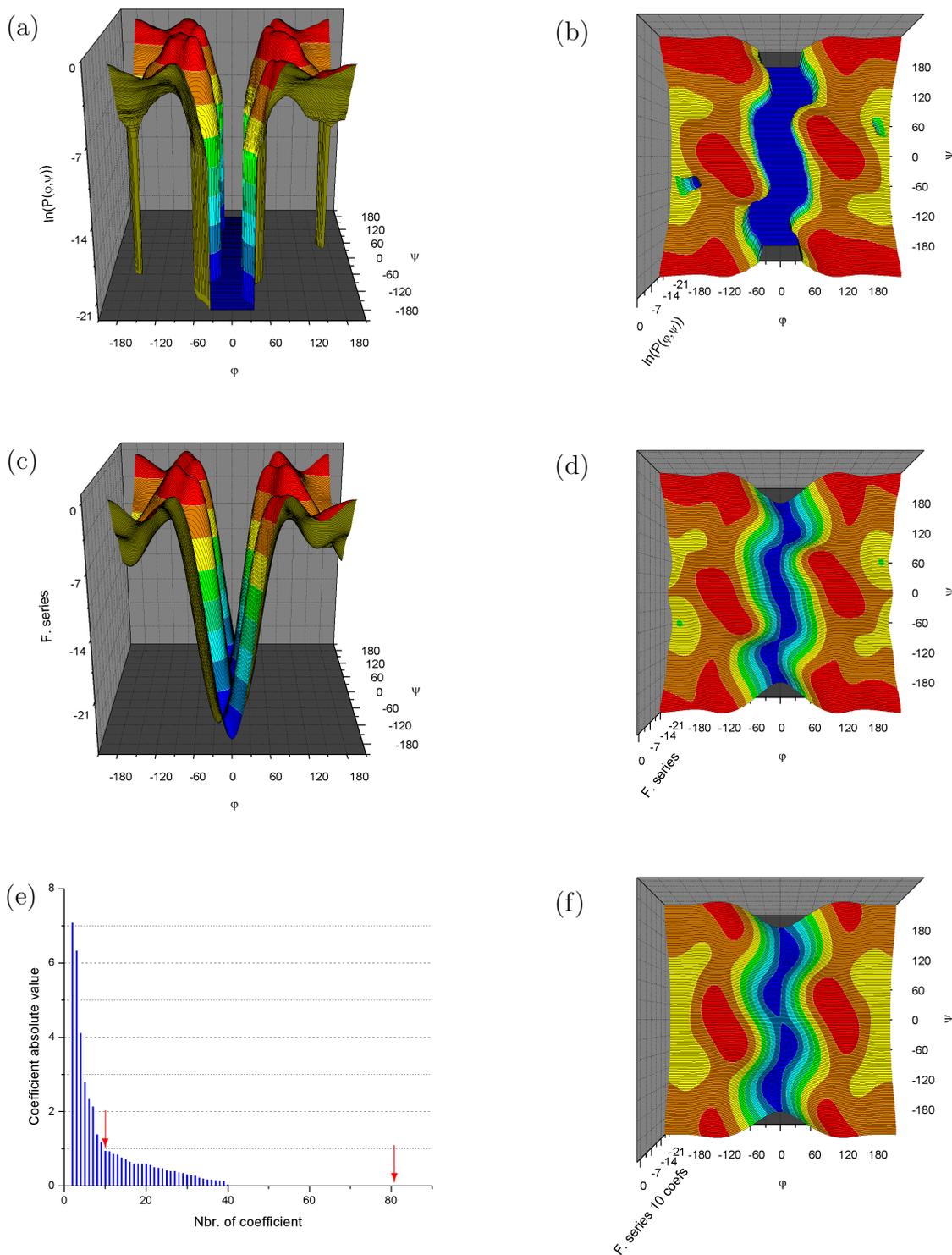
A subsequent approximation of the glycine R-plot of the Top500 database was carried out also starting with a Fourier series of 80 coefficients. The obtained results were similar to those found for the general R-plot approximation, inasmuch as a Fourier series of 80 coefficients (Figure 3.9c and 3.9d) produces a relatively accurate approximation at a detailed level and one of 10 coefficients (Figure 3.9f) generates a broader approximation which contains all the major features of the glycine R-plot. Moreover, the centrosymmetric nature of the glycine R-plot dictates a value of zero for all coefficients containing cosine and sine cross-products, namely  $O_{hk}$  and  $P_{hk}$  (Table 3.3). Consequently, these coefficients were manually set to zero and were not refined during the least squares routine.

In terms of using R-plot potentials as restraints, an effective potential should first and foremost contain all the main features of that R-plot. Additionally, it should be smooth and not contain very sharp maxima so as to allow the value of a restrained variable to be swept throughout the corresponding R-plot space. Hence, the potentials obtained from the Fourier series calculated with 10 coefficients, which contain the aforementioned properties, were chosen to perform preliminary tests described in § 3.6. In this way, R-plot based restraints can be applied to polypeptide chains comprised of non-proline and non-pre-proline residues. Approximations of proline and pre-proline R-plots were not produced as part of this study since the selected polypeptide test-cases, presented in the following section (§ 3.5), did not contain residues of that type. Furthermore, it is important to note that although the above describes the approximation of an R-plot for L-amino acids, the same Fourier series is applicable as a restraint for D-amino acids by simply using  $E_{500}(-\varphi, -\psi)$ .

**Figure 3.8** A series of general R-plots calculated with (b) 10, (c) 17, (d) 22, (e) 38 and (f) 80 coefficients. Also included (a) is the set of coefficients ordered by absolute value from highest to lowest (where  $C_{00}$  has been omitted), with the red arrows pointing at the various points at which the displayed Ramachandran plots were calculated.



**Figure 3.9** Glycine R-plots calculated with the Top500 database data (a & b) and the Fourier series data using 80 coefficients (c & d) and 10 coefficients (f). Also included is the set of coefficients ordered by absolute value from highest to lowest (where  $C_{00}$  has been omitted), with the red arrows pointing at the various points at which the displayed Ramachandran plots were calculated.



## 3.5 Peptide systems

Various short polypeptide chains were used in this project with the aim to assess the potential of incorporating R-plot restraints in the direct-space method of simulated annealing (§ 3.2.1) applied to powder diffraction data. The polypeptides were chosen according to their varying degrees of complexity (*i.e.* number of torsion angles). More specifically, these included a tripeptide (Fmoc–Ala–Ala–Ala), a tetrapeptide (Phe–Gly–Gly–Phe) and a pentapeptide (Tyr–Ala–Gly–Phe–Leu). Of these, the first two were previously studied in our laboratory and diffraction data was thus readily available. For the pentapeptide, however, no experimental data were available as it was found on the Cambridge Structural Database (CSD, Allen 2002). In an attempt to obtain a polypeptide system flexible enough to fully assess the potential of R-plot restraints and to determine the point of complexity at which these restraints become decisive for structure solution, crystallization trials were carried out on a heptapeptide (Adrenocorticotrophic hormone- (4 - 10)).

### 3.5.1 Fmoc–Ala–Ala–Ala

The structure of Fmoc–(Ala)<sub>3</sub> was previously solved in our laboratory as part of an unpublished study. The non-amino acid group, denoted ‘Fmoc’, is an  $\alpha$ -amino acid protecting group extensively used in peptide synthesis for its ease of cleavage under basic conditions. Structure determination of the tripeptide was achieved using a combination of powder and single-crystal synchrotron diffraction data measured at the Swiss-Norwegian beamlines BM01B and BM01A, respectively. The molecules were found to be in an extended conformation and hydrogen bonded to each other in standard parallel  $\beta$ -sheets.

The crystallographic data are presented in Table 3.4, while Figure 3.10 and Table 3.5 illustrate the conformational degrees of freedom of this peptide. To determine the total number of variables, the external degrees of freedom of both the peptide and water molecule have to be considered. Given the polar nature of the  $P2_1$  space group, the former contributes an additional 5MR variables. On the other hand, the water molecule, which consists of a single oxygen atom, can be correctly placed within the unit cell using only translational parameters. Hence, it contributes a further 3MR variables for a total of 14.

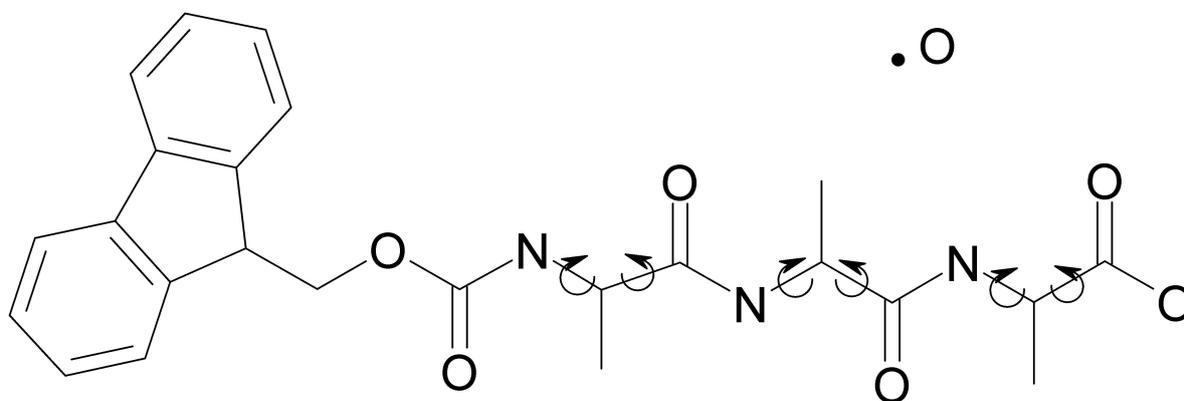
Table 3.5 also indicates the achievable restraining potential using R-plot restraints. In a tripeptide, only the torsion-angle pair of the second residue is theoretically eligible for R-plot restraints, as the other two are at chain termini. However, in this peptide, the presence of the protecting group leads to the existence of a peptide-like plane at the N-terminus. Indeed, the bond between the C-atom of the ‘Fmoc’ group and the N-atom

**Table 3.4** Crystallographic and diffraction data for Fmoc-Ala-Ala-Ala ('expt' = experimental).

Space group	Symmetry				Powder data	
	$a$ (Å)	$b$ (Å)	$c$ (Å)	$\beta$ (°)	Type	$\lambda$ (Å)
$P2_1$	17.198	4.787	14.984	96.023	expt.	0.8000(1)

**Table 3.5** Breakdown of the number of variable torsion angles in Fmoc-Ala-Ala-Ala. Also included are the number of torsion-angle pairs to which R-plot restraints can be applied ( $R(\varphi, \psi)$ ) and, of those, the number of glycine amino acids [Gly].

Total	Torsion angles			R-plot restraints	
	$\varphi$	$\psi$	$\chi$	$R(\varphi, \psi)$	[Gly]
6	3	3	0	1	[0]

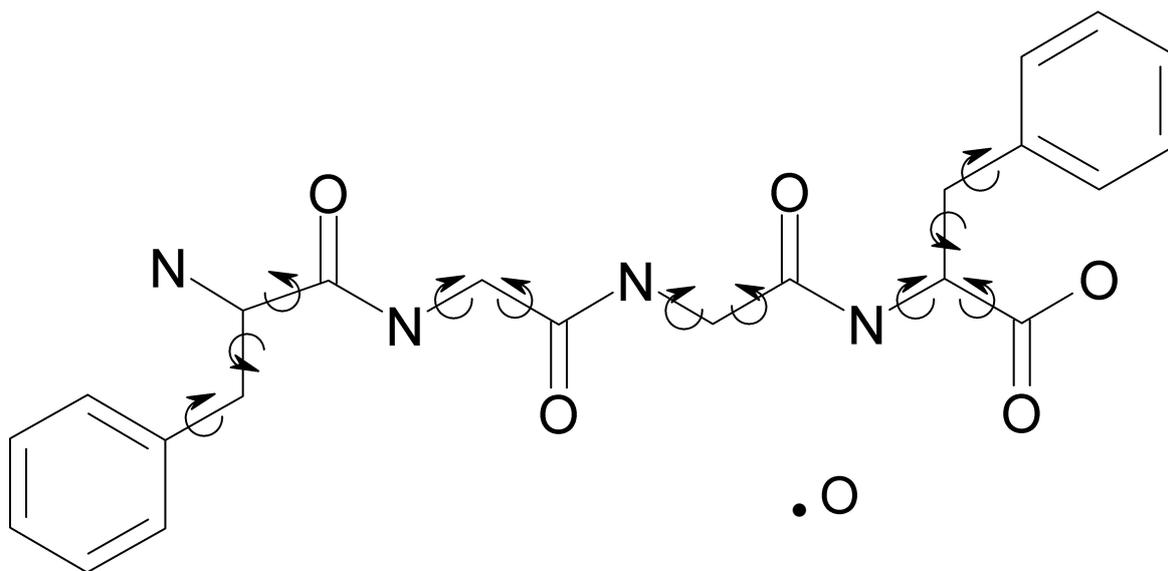
**Figure 3.10** Fmoc-Ala-Ala-Ala  
(All hydrogen atoms are omitted for clarity)

of the first alanine residue is identical to a peptide bond, as is apparent from Figure 3.10. The only difference between this peptide-like plane and a 'normal' peptide plane is that one of the  $C\alpha$ -atoms is replaced by an O-atom. Given the smaller radius of an O-atom compared to that of a C-atom, the corresponding R-plot would be permissive of a few more  $[\varphi, \psi]$  combinations. As a result, an R-plot restraint with a low weight could potentially be employed for the N-terminus alanine residue.

### 3.5.2 Phe–Gly–Gly–Phe

In a structural investigation of Phe–Gly–Gly–Phe performed using a genetic algorithm approach applied to powder diffraction data (Tedesco et al. 2000), the tetrapeptide molecules were found to be in an extended conformation and to contain intermolecular interactions analogous to those present in an antiparallel  $\beta$ -sheet arrangement. A subsequent in-house study revealed an almost identical crystal-structure with the main difference being the added presence of a water molecule. The monohydrate structure was initially determined using single-crystal data collected on the BM01A beamline. It was then also found using simulated annealing in the program *DASH* applied to powder diffraction data measured on the BM01B beamline. The powder diffraction analysis was carried out independently of the single-crystal data as well as of the structural information extracted from it.

The crystallographic data are presented in Table 3.6, while Figure 3.11 and Table 3.7 illustrate the conformational degrees of freedom of this peptide. Given that the space group  $P4_1$  is also polar, the number of external degrees of freedom for this crystal structure is eight (5MR for the peptide and 3MR for the water molecule). Therefore, the complexity of this peptide system can be accounted for by a total number of 19 variables. Table 3.7 also indicates the achievable restraining potential using R-plot restraints.



**Figure 3.11** Phe–Gly–Gly–Phe  
(All hydrogen atoms are omitted for clarity)

**Table 3.6** Crystallographic and diffraction data for Phe–Gly–Gly–Phe ('expt' = experimental).

Space group	Symmetry			Powder data	
	$a$ (Å)	$b$ (Å)	$c$ (Å)	Type	$\lambda$ (Å)
$P4_1$	15.052	15.052	9.716	expt.	0.6504(1)

**Table 3.7** Breakdown of the number of variable torsion angles in Phe–Gly–Gly–Phe. Also included are the number of torsion-angle pairs to which R-plot restraints can be applied ( $R(\varphi,\psi)$ ) and, of those, the number of glycine amino acids [Gly].

Total	Torsion angles			R-plot restraints	
	$\varphi$	$\psi$	$\chi$	$R(\varphi,\psi)$	[Gly]
11	3	4	4	2	[2]

### 3.5.3 L-Tyr–D-Ala–Gly–L-Phe–D-Leu

Enkephalin peptides are pentapeptides involved in neural processes in the peripheral and central nervous system. The sequence of the two existing types of such peptides is: Tyr–Ala–Gly–Phe–X, where 'X' can be amino acids Met or Leu to form the methionine- and leucine-enkephalin, respectively (Noda et al. 1982). The pentapeptide with the titled sequence is a modified leucine-enkephalin where residues 2 and 5 have been replaced by D-ALA and D-LEU, respectively. Its structure was determined in a previously published single-crystal experiment (Deschamps et al. 1996) and was found to be a slightly distorted type I'  $\beta$ -bend conformation stabilised by a single intramolecular hydrogen bond.

The crystallographic data of the pentapeptide are presented in Table 3.8. The powder diffraction data were simulated due to lack of experimental data. The simulation was performed within the program Mercury (Macrae et al. 2006). Figure 3.12 and Table 3.9 illustrate the conformational degrees of freedom of this peptide. Similarly to the tripeptide above, the number of MR parameters for this crystal structure is eight (5MR for the peptide and 3MR for the chlorine atom). As a result, the total number of variables is of 23.

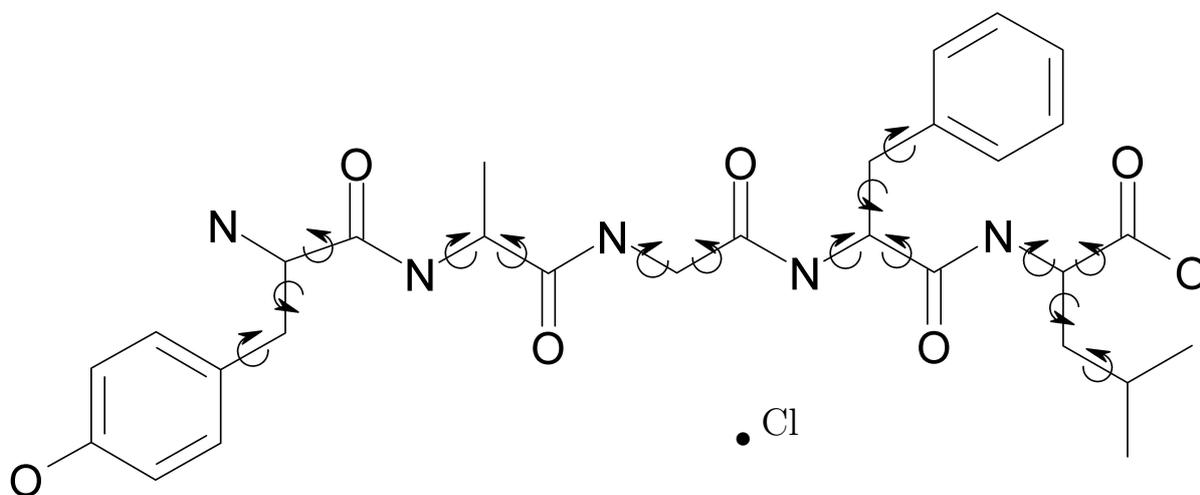
Table 3.9 also indicates the achievable restraining potential using R-plot restraints. Out of the three amino acids for which R-plot restraints are applicable, the alanine one is in the D-configuration. In this case, the general R-plot restraint can be applied to the  $[\varphi,\psi]$  pair by simply calculating the restraint using  $[-\varphi,-\psi]$ .

**Table 3.8** Crystallographic and diffraction data for L-Tyr–D-Ala–Gly–L-Phe–D-Leu ('sim.' = simulated data).

Space group	Symmetry				Powder data	
	$a$ (Å)	$b$ (Å)	$c$ (Å)	$\beta$ (°)	Type	$\lambda$ (Å)
$P2_1$	14.498	9.263	14.703	103.160	sim.	0.8

**Table 3.9** Breakdown of the number of variable torsion angles in L-Tyr–D-Ala–Gly–L-Phe–D-Leu. Also included are the number of torsion-angle pairs to which R-plot restraints can be applied ( $R(\varphi, \psi)$ ) and, of those, the number of glycine amino acids [Gly].

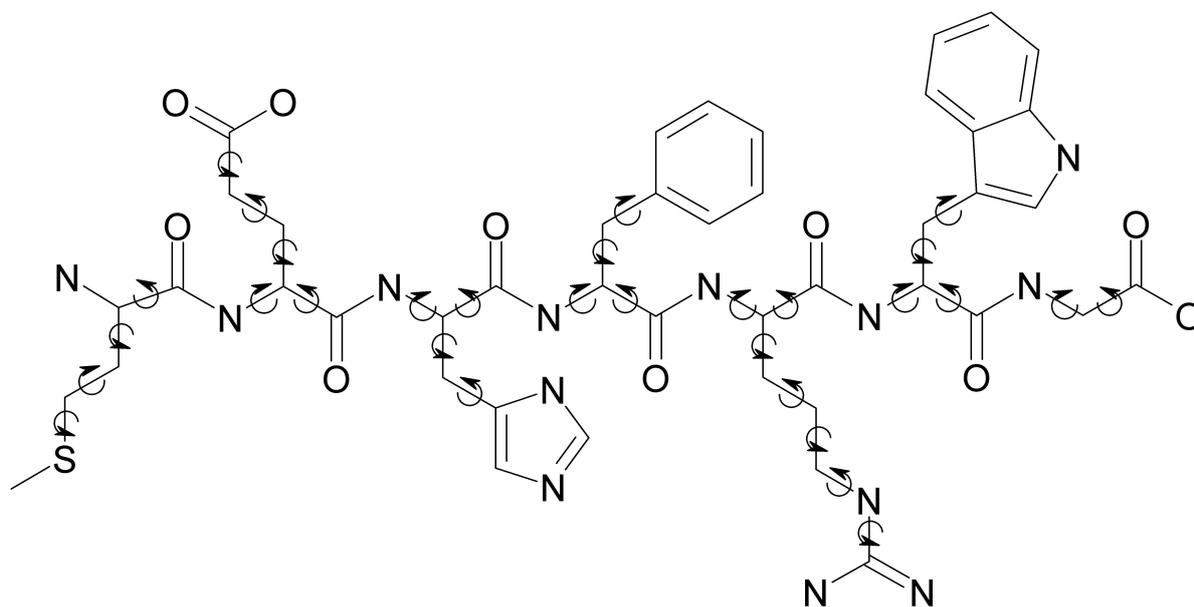
Total	Torsion angles			R-plot restraints	
	$\varphi$	$\psi$	$\chi$	$R(\varphi, \psi)$	[Gly]
15	4	5	6	3	[1]



**Figure 3.12** L-Tyr–D-Ala–Gly–L-Phe–D-Leu  
(All hydrogen atoms are omitted for clarity)

### 3.5.4 Adrenocorticotrophic hormone - (4 - 10)

Pro-opiomelanocortin is a precursor protein which, when cleaved by proteolytic enzymes, produces various hormones and endorphins in the pituitary gland, brain and other tissues. Adrenocorticotrophic hormone (ACTH) is one of such generated hormones. Moreover, fragments of this hormone can be isolated to produce neuropeptides (De Wied 1969). ACTH-(4-10) corresponds to the neuropeptide of amino-acid sequence from residue 4 to 10



**Figure 3.13** ACTH-(4-10)

(All hydrogen atoms are omitted for clarity)

of the hormone, *i.e.* Met–Glu–His–Phe–Arg–Trp–Gly. A review on the neurological role of adrenocorticotrophic hormone and its related peptides was published by Strand (2000). In a preliminary X-ray structural study of ACTH-(4-10), the authors only put forth a unit cell and a model of the secondary structure comprised of antiparallel  $\beta$ -sheets (Admiraal & Vos 1984). Since then, no subsequent structural investigation on the heptapeptide has been reported in the literature.

Figure 3.13 and Table 3.10 illustrate the conformational degrees of freedom of this peptide. Analogously to the three previous peptides, the ACTH-(4-10) crystallizes in a polar monoclinic space group, namely  $C2$ . Consequently, assuming the molecule is anhydrous, 5MR parameters would be required to correctly place the molecule in the unit cell. The total number of variables would then amount to 35. Table 3.10 also indicates the achievable restraining potential using R-plot restraints.

### 3.5.4.1 Crystallization

crystallization experiments were performed using lyophilised ACTH-(4-10) purchased from Bachem. The purity of the product was quoted at 98.9% (HPLC-TFA) and the lyophilised powder was used without further purification. The crystallization trials were based on the conditions previously determined by Admiraal & Verweij (1982). Although the quoted crystal dimensions of up to  $0.2 \times 0.2 \times 0.6 \text{ mm}^3$  were never reproduced, two

**Table 3.10** Breakdown of the number of variable torsion angles in ACTH-(4-10). Also included are the number of torsion-angle pairs to which R-plot restraints can be applied ( $R(\varphi,\psi)$ ) and, of those, the number of glycine amino acids [Gly].

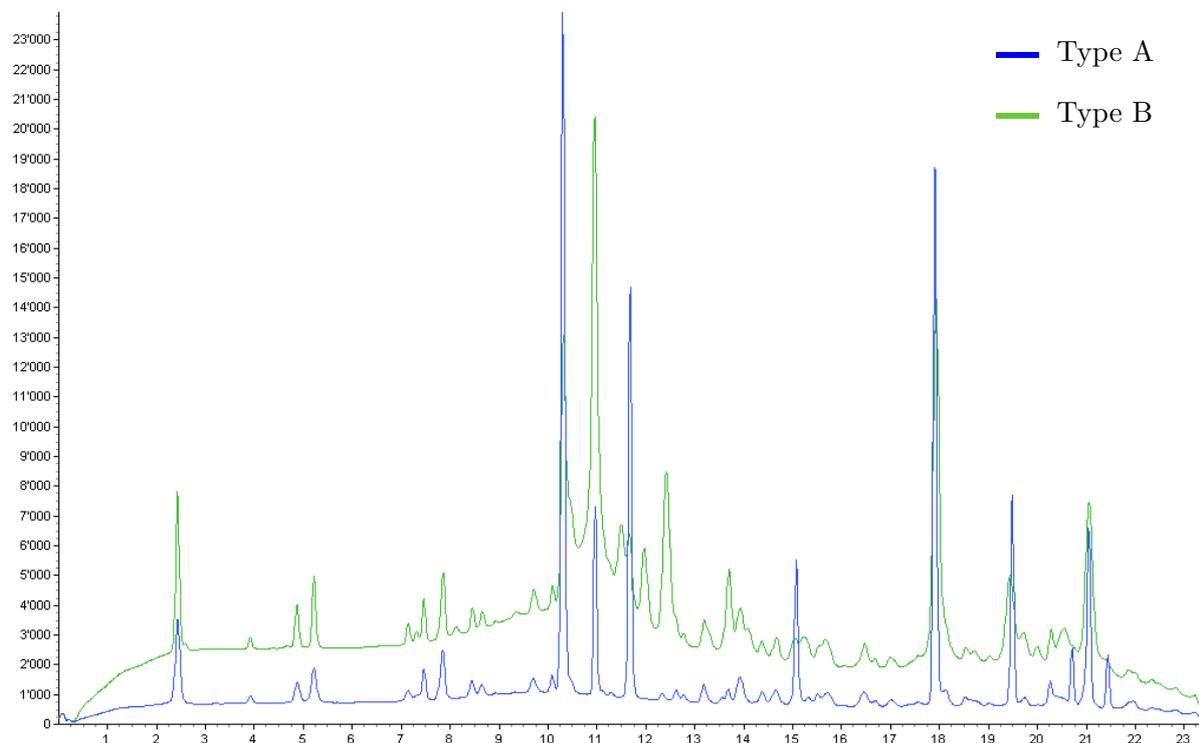
Total	Torsion angles			R-plot restraints	
	$\varphi$	$\psi$	$\chi$	$R(\varphi,\psi)$	[Gly]
30	6	7	17	5	[0]

**Table 3.11** crystallization conditions for the two polycrystalline types of ACTH-(4-10). The provided concentrations for the peptide solution correspond to the initial concentrations (*i.e.* prior to the addition of the precipitant)

	Peptide Solution				Precipitant
	Peptide Concentration	Buffer Agent	pH	CaCl <sub>2</sub>	
Type A	8 mg / ml	0.02M (NH <sub>4</sub> ) <sub>3</sub> citrate	4.5	0.02M	30 $\mu$ L propan-2-ol
Type B	8 mg / ml	0.02M (NH <sub>4</sub> ) <sub>3</sub> citrate	4.5	0.01M	50 $\mu$ L 1-propanol

types of polycrystalline samples were obtained using batch methods (§ 2.2.1). One type of powder sample (Type A) consisted of crystals which formed aggregates of roughly 10  $\mu$ m in diameter and hence it was difficult to describe each crystallite with a particular shape. The second sample type (Type B) was comprised of needle-like microcrystals approximately 5  $\mu$ m in length. These samples were prepared according to the crystallization conditions reported in Table 3.11. Judging from all the crystallization attempts, it was concluded that the concentration of CaCl<sub>2</sub> in the peptide solution was the determining factor in producing one crystal shape or the other. On the other hand, the nature of the alcohol precipitant simply affected the quantity and size of the obtained microcrystallites. More specifically, the use of propan-2-ol gave rise to smaller crystallites and in greater quantity as compared to using 1-propanol.

Preliminary X-ray powder diffraction measurements, performed at the BM01A beamline, revealed that both types of polycrystalline powders diffracted and that they were in all likelihood the same phase. Figure 3.14 shows the powder patterns of the two samples measured using an image-plate area detector. Indexation of these patterns has been unsuccessful, the most probable cause being the relatively extensive broadness of the peaks originating from the type of detector employed to collect the data.



**Figure 3.14** Preliminary X-ray powder diffraction patterns for ACTH-(4-10) samples of Type A (blue) and Type B (green). In the former there seems to be extensive preferred orientation while the latter has a considerable solvent background contribution.

### 3.6 Preliminary results

All simulated annealing runs were carried out using a default protocol incorporated in the software *TOPAS-Academic* which employs the weighted profile R-factor as a cost function. No further measures were taken to tailor the protocol to the problem of solving peptide crystal structures. The structural models of the peptides presented above were described using the z-matrix construct. The z-matrices were built from the N-terminus to the C-terminus without including the hydrogen atoms. These atoms can be excluded from the construct during the structure-solution process, owing to their low scattering power and, in turn, negligible effect on the powder diffraction pattern. The variables to be determined during a simulated annealing run ( $\varphi$ ,  $\psi$ ,  $\chi$  and the MR parameters) were identified and defined as demonstrated in § 3.3.1. These were subsequently set to an initial value of zero in order to generate a starting model of the polypeptide chain in a random configuration. For each peptide system, the scale factor used to adjust the calculated pattern to the observed one was fixed to a value determined from a Rietveld refinement using the single-crystal structure. Moreover, the R-plot restraints were implemented as

simple penalty functions using as a weighting scheme, the weight assigned to the restraints relative to the diffraction data, the default scheme provided within *TOPAS-Academic* (Coelho 2000).

However, prior to incorporating R-plot restraints into the global optimization process, the SA algorithm was tested on the tripeptide Fmoc–Ala–Ala–Ala. The peptide was described with a z-matrix in which the stereochemical information was taken from the International Tables of Crystallography Volume F (Engh & Huber 2001), the non-variable torsion angles were expressed using the relationships established in § 3.3.1 and the number of independent variables was 14 (§ 3.5.1). Using experimental powder diffraction data, a total of 49 SA runs were carried out, each run beginning with the peptide in the same starting configuration and position (*i.e.* with all variables set to zero). The runs were stopped after 100000 iterations, where one iteration corresponds to one random move in the parameter space. In this way, a solution was found 11 times out of the 49 runs<sup>5</sup>. Subsequently, an equivalent test was performed with the addition of general R-plot restraints for the second alanine residue. In this case, a solution was achieved 10 times out of 49 runs. These results would seem to indicate that including R-plots restraints do not improve the process of structure solution. However, a sample of 49 runs is not sufficient to accurately represent the statistics governing the success rate of the SA protocol depending on whether or not restraints were used. Furthermore, with only one torsion-angle pair being eligible for R-plots restraints, only 14.3% of the variables are restrained. Therefore, the effect of the restraints on the search algorithm may not be significant enough to notice a considerable impact on the structure-solution process.

In comparison, the tetrapeptide Phe–Gly–Gly–Phe contains one more amino acid to which R-plot restraints can be applied. Hence, 4 out of the 19 variables for this system are restrainable which corresponds to a restraining potential of 21.1% of the variables. However, both of the torsion-angle pairs belong to glycine residues. Given the relatively low restrictions induced by the glycine R-plot restraints, compared to those of the general R-plot restraints, their effect may not be clearly noticeable either. In light of this, it was judged more productive to move onto the next peptide candidate, namely the pentapeptide Tyr–Ala–Gly–Phe–Leu.

In this pentapeptide, there are 3 restrainable torsion-angle pairs, which corresponds

---

<sup>5</sup>A structural candidate was considered a solution when it gave rise to a cost-function value comparable to that obtained from a pattern decomposition procedure and when it was found to be closely related to the single-crystal structure upon comparison.

---

to 26.1% of the total number of variables. However the increased restraining potential comes at an inevitable cost of overall increased complexity of the structure-solution problem. Indeed, initial SA runs (without restraints) suggested that a system comprised of 23 independent variables was too complex for the employed default SA protocol, as no solution was found. These runs were performed with a z-matrix built as that of the tripeptide and using a simulated powder diffraction pattern (§ 3.5.3). The success of the protocol did not improve when all non-variable internal coordinates were set to the values from the single-crystal structure. Yet, as the aim was to observe the effect of R-plot restraints, all variables representing external degrees of freedom could also be set to those values and held fixed. In this manner, the problem became one of ‘only’ 12 variables, all representing torsion angles, of which 50% were restrainable. Similar tests to those described above were carried out using two sets of 29 runs of SA. A smaller number of runs was performed since the algorithm was allowed to run for longer, namely 450000 iterations, to increase the chances of finding the global minimum. The ensuing results were analogous to those obtained for the tripeptide: the solution was found 28 and 26 times out of 29 runs without and with restraints, respectively. However, as mentioned above, such a small set of runs is not sufficient to accurately represent the statistics involved. No further runs were performed due to lack of time.

### 3.7 Outlook

The implementation of R-plot restraints as penalty functions in a simulated annealing protocol has been shown to be possible. The use of Fourier series was demonstrated to be effective in approximating R-plots. Furthermore, a reduced set of high-absolute-value Fourier coefficients provides a good approximation of the main features of a given R-plot. The periodicity of Fourier series enables a more general implementation of torsion-angle restraints in comparison with previously suggested restraints of a similar type. Moreover, the approach could be extended to rotamer libraries in order to produce restraints for the torsion angles of amino-acid side chains. For most amino acids these angles cluster within narrow ranges of values and the restraining potential of rotamer libraries is thus substantial. Such a library is available in the aforementioned Top500 database.

Despite the successful implementation of R-plot restraints, the preliminary results are inconclusive concerning their effect. In order to draw conclusions from analyses such as those described above, the statistics governing the success rate of the SA protocol with and without the use of restraints need to be dramatically improved. Despite being time

consuming, this can be done without difficulty as it simply involves completing a greater number of runs. Simultaneously, a systematic investigation into the weighting scheme for R-plot restraints is required. Indeed, inasmuch as the default weighting scheme used in the previously described tests does not hamper the structure-solution process, the weight attributed to the restraints could be made heavier to potentially guide the search algorithm more efficiently. Nevertheless, care should be taken not to use too heavy a weight since this will result in the search being restricted to one of the favourable areas of the R-plot.

Additionally, there exists a need to find a peptide system just complex enough to be unsolvable using a non-restrained SA protocol but which becomes solvable once R-plot restraints are incorporated into it. Indeed, the level of conformational complexity at which R-plot restraints become decisive for structure solution will only be determined with such a system. However, for this type of peptide to be found the employed SA protocol needs to perform as efficiently as possible (*i.e.* it should be tailored for structure solution of peptide molecules). This was not the case in this project since the protocol used to produce the described preliminary results was a default SA protocol available in *TOPAS-Academic*. As a result, initial improvements to the protocol could be made by carrying out an investigation on the effect of varying the parameters of the algorithm. These include the temperature at which each new cycle starts within a given run, the rate at which it is decreased and its relationship with the magnitude of the shifts applied to the variables.

Furthermore, the efficiency of the SA protocol is also closely dependent on the quality of the atomic model. Here, the quality of the model is taken to mean how well it strikes the right balance between reducing the total number of variables to a minimum and retaining a certain degree of flexibility so as to allow the molecule to adopt all possible conformations. In this respect, using a z-matrix to describe a polypeptide chain seems to be a suitable choice given the inherently rigid nature of its building blocks. However, the relationships established in § 3.3.1, for the non-variable torsion angles expressed as a function of  $\varphi$  and  $\psi$ , are most probably too rigid. The reason for this stems from the fact that these relationships are based on theoretical assumptions. Indeed, the values of these torsion angles in the structures derived from single-crystal data often deviate (up to  $\sim 10^\circ$ ) from the predicted values using the aforementioned relationships. As a result, for correct values of  $\varphi$  and  $\psi$ , the corresponding atomic positions of the carbonyl oxygen atom and that of the  $C\beta$  atom, respectively, will be incorrect. It is apparent that this is especially damaging to the structure-solution process for peptides made up of residues

---

with long side chains. Similarly, the assumption of planarity made in order to define the torsion angle  $\omega$  as being equal to  $180^\circ$  is in all likelihood also too rigid. In a z-matrix definition which starts at either end of the peptide chain, this rigidity can be especially notorious if the true value of one of the first omega angles deviates from the theoretical value as it will affect the atomic positions of all subsequently defined atoms in the z-matrix. Taking the above into consideration, the results obtained using the method of SA could be improved by making the definition of these non-variable torsion angles less ‘strict’ so as not to lose flexibility at the expense of reducing the number of variables. This could be achieved by expressing these torsion angles as a function of variable torsion angles using highly restrained variables instead of fixed numerical values. Moreover, the effects associated with deviations in the  $\omega$  angle could be made less damaging to the structure-solution process by using a z-matrix description starting closer to the middle of the molecule, as was demonstrated by Shankland et al. (2002).

Another approach to improve the ability with which the solution is found is to vary the starting point within the parameter space at which each SA run begins. All previously reported tests were performed with a molecular conformation described with a starting value of zero for all variable parameters, in the aim to assess the robustness of the SA algorithm. Yet, its success rate could potentially be ameliorated by performing a series of runs starting from different initial configurations. While the various starting points of the external degrees of freedom would be random, those for the torsion-angle variables could be set to favourable regions of the R-plot. The latter suggestion would ensure that the search method initially considers trial structures contained within one of the favourable regions of the R-plot, before either stochastically continuing to explore the R-plot space or explicitly moving onto another one of its favourable regions depending on whether or not R-plot restraints are used.

In its current state, it is unlikely that the R-plot restrained SA protocol could be employed to solve the structure of the heptapeptide ACTH-(4-10) using powder diffraction data. Nevertheless, the implementation of the above changes will certainly increase the likelihood of solving it. In addition, the crystallization trials on the heptapeptide (§ 3.5.4.1) were undertaken with the development of rotamer restraints in mind. Indeed, with more than half of the conformational degrees of freedom of the heptapeptide being side-chain torsion-angles, ACTH-(4-10) would make an ideal candidate for testing such restraints.



# Chapter 4

## Conclusion

In view of the work presented in this manuscript, the technique of powder diffraction is likely to contribute in a non-negligible manner to the field of biomolecular crystallography. The results described in Chapter 2 show that it can be employed to produce relatively detailed structural information for a protein molecule without the use of a molecular model and with methods far from optimal. Simultaneously, these results highlight the potential of powder diffraction to play an especially significant role in the study of relatively small protein molecules (residue count between 0 - 250) which crystallize in space groups with orthorhombic symmetry or lower. Indeed, such crystalline material will give rise to powder diffraction patterns with a manageable density of peaks as well as no exact overlap, thus providing excellent conditions for the MiPPIE method to perform efficiently and thereby reducing the gap in information content between powder and single-crystal data. Inevitably, the requirements on molecular size and crystal symmetry, for optimal exploitation of the powder technique, somewhat limit the number of macromolecular systems which can be structurally investigated using powder diffraction under such favourable conditions. Nevertheless, many such protein molecules exist as is demonstrated by the percentage of structures deposited in the PDB which either fall within the aforementioned residue-count range or crystallize in a space group with symmetry no higher than orthorhombic:  $\sim 36\%$  ( $\approx 26400$  structures) and  $> 60\%$  ( $\approx 42000$  structures), respectively. In spite of this, powder diffraction will most likely remain a complementary tool to single-crystal diffraction in the field of protein structure solution, since, for the foreseeable future, it is unimaginable that it will be employed to generate electron density maps of macromolecules at atomic resolution. However, it is anticipated that it will increasingly be used as a first step in the structural investigation of undetermined protein crystal structures. Moreover, its impact on the structure-solution process of such

structures is expected to intensify as the algorithms to extract structural information from electron density maps of medium-to-low resolution are improved.

On the other hand, protein powder diffraction is foreseen to play a leading role in the field of material science as a result of the recent surge in interest in protein powders as a material which possesses chemical and physical properties exploitable in a series of applications. In the study of such material, given that these properties are mostly governed by crystal packing, and hence polymorphism, it is important to map out the protein-solvent boundaries (*i.e.* determine the solvent envelope). Since these can be identified in a low-to-medium resolution electron density map, knowledge of the high-resolution tertiary structure (which does not vary extensively between polymorphs) of the macromolecule constituting the crystalline powder is not necessary. It is thus apparent that powder diffraction is the ideal tool to study the properties of such materials. Once again, developments in the algorithms mentioned above will allow the exploitation of lower-resolution data obtained, for example, from a powder sample of a relatively large protein which crystallizes in a high symmetry space group.

In terms of molecular size, oligopeptides are situated at the very beginning of the polypeptide spectrum. Nevertheless, the structural complexity of such ‘small molecules’ can escalate relatively quickly depending on the number and nature of the amino acids. Furthermore, like proteins, peptides have a propensity to form polycrystalline powders rather than large single crystals. With the recent progress made in powder-specific structure determination techniques, global optimization methods in particular, powder diffraction could emerge as a viable technique to study such biopolymers. In addition, the successful implementation of future developments in global optimization methods, including those proposed in Chapter 3, will enable powder diffraction to solve the structure of polypeptides with increasingly longer amino-acid sequences and, one day perhaps, those with a sequence long enough to constitute a protein molecule.

# Appendix A

## X-ray diffraction

### A.1 Fundamental principles

**A crystal** is a solid whose microscopic structure is characterized by a periodic repetition of an atomic motif in three dimensions. The symmetry operations of translation can be decomposed as follows:  $\mathbf{t} = u\mathbf{a} + v\mathbf{b} + w\mathbf{c}$ , where  $u, v$  and  $w$  are integers and  $\mathbf{a}$ ,  $\mathbf{b}$  and  $\mathbf{c}$  are the *base vectors*. These vectors generate a unit cell and their norms are referred to as the *unit-cell parameters* (expressed in a unit of length). An arbitrary point in the crystal can be described by  $\mathbf{r} = x\mathbf{a} + y\mathbf{b} + z\mathbf{c}$ , where  $x, y$  and  $z$  are dimensionless coefficients called *fractional coordinates*. The volume of the unit cell is  $V = (\mathbf{a} \cdot \mathbf{b} \wedge \mathbf{c}) > 0$ .

**X-ray diffraction from a crystal** involves an incident beam and a diffracted beam. These can be described by a plane wave characterized by the wave vector  $\mathbf{s}_0$  and a scattered wave characterized by the vector  $\mathbf{s}$ , respectively. We define the *scattering vector* associated with a scattered wave as:  $\mathbf{S} = (\mathbf{s} - \mathbf{s}_0)$ . Since only elastic scattering is considered,  $|\mathbf{s}_0| = |\mathbf{s}| = 1/\lambda$ . Scattered waves only get scattered by a crystal in characteristic directions. More specifically, they have to simultaneously fulfill the three von Laue equations (Friedrich et al. 1912, von Laue 1913):

$$\mathbf{S} \cdot \mathbf{a} = h$$

$$\mathbf{S} \cdot \mathbf{b} = k$$

$$\mathbf{S} \cdot \mathbf{c} = l$$

where  $h, k$  and  $l$  are integers. These are then referred to as *diffracted waves* and the vector  $\mathbf{S}$  is denoted  $\mathbf{h}$ . The *reciprocal lattice* (Ewald 1921) is subsequently defined as the set of endpoints of all these vectors:  $\mathbf{h} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*$ , where  $h, k$  and  $l$  are dimensionless coefficients called *indices*. The reciprocal base vectors are defined as  $\mathbf{a}^* = (\mathbf{b} \wedge \mathbf{c})/V$ ,  $\mathbf{b}^* = (\mathbf{a} \wedge \mathbf{c})/V$  and  $\mathbf{c}^* = (\mathbf{a} \wedge \mathbf{b})/V$ . These vectors have the dimension of the inverse of a length and satisfy the relationships  $\mathbf{a} \cdot \mathbf{a}^* = \mathbf{b} \cdot \mathbf{b}^* = \mathbf{c} \cdot \mathbf{c}^* = 1$ ,  $\mathbf{a} \cdot \mathbf{b}^* = \mathbf{a} \cdot \mathbf{c}^* = \dots = 0$ . It follows that the reciprocal lattice is a geometric representation of the set of scattering vectors which satisfy the von Laue equations. As a result:  $\mathbf{h} \cdot \mathbf{r} = (hx + ky + lz)$ .

**The structure factor** is a complex number the modulus of which is proportional to the intensity of the corresponding diffracted wave and can be expressed by

$$\mathbf{F}(\mathbf{h}) = \sum_j f_j \exp(2\pi i \mathbf{h} \cdot \mathbf{r}_j),$$

where the summation is performed over all atoms  $j$  contained in the unit cell and,  $r_j$  and  $f_j(\mathbf{h})$  represent the positional vector and atomic scattering factor of a given atom. Alternatively, the structure factor can be expressed as a function of the electron density of a unit cell,  $\rho(\mathbf{r})$ , according to

$$\mathbf{F}(\mathbf{h}) = \int_{\text{Unit cell}} \rho(\mathbf{r}) \exp(2\pi i \mathbf{h} \cdot \mathbf{r}) d^3\mathbf{r}.$$

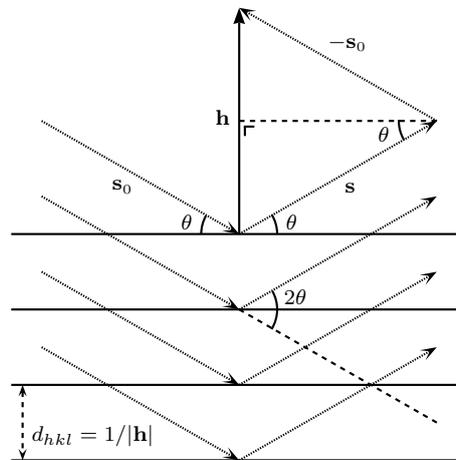
## A.2 Powder diffraction

**Bragg's law** is a scalar interpretation of the von Laue equations. It is ideally suited to describe the diffraction of X-rays from a powder sample given the rotational projection of the three-dimensional reciprocal lattice onto the single  $2\theta$  dimension of a powder pattern. From Figure A.1, Bragg's law can be derived to give

$$2d_{hkl} \sin\theta = \lambda,$$

where  $2\theta$  is the scattering angle of the X-rays and  $d_{hkl}$  is the interatomic distance.

**Peak overlap** in a powder pattern leads to a great information deficit with respect to single-crystal data. The loss of information stems from the fact that only the sum of the



**Figure A.1** Illustration of Bragg's law.

intensities of overlapping peaks can be measured accurately. Nevertheless, not all overlap leads to a loss of information. There are two types of peak overlap: exact and accidental. The former represents the reflections which occur at the exact same angle and can be subdivided into two further categories: symmetry-equivalent reflections and symmetry-independent reflections<sup>1</sup>. Only the latter contributes to the information deficit, since in a single-crystal diffraction measurement the symmetry-equivalent reflections are merged together. While nothing can be done about exact overlap, accidental overlap, which also leads to information loss, can be somewhat circumvented as described in § 2.3.2.

### A.3 Synchrotron radiation

Synchrotron radiation is produced by the acceleration of electrons or positrons traveling at relativistic speeds. In practice, these relativistic particles are circulated within a *storage ring*. Part of the exploited radiation is generated in the bent section of the storage ring by bending magnets. These deviate the trajectory of the charged particles and produce a centripetal acceleration, which is responsible for the emission of X-ray radiation. Synchrotron radiation can also be generated by insertion devices, namely wigglers and undulators, placed in the straight sections of the storage ring. Such devices comprise of a periodic arrangement of small magnets which impose a sinusoidal trajectory on the

<sup>1</sup>Example of symmetry-independent reflections:  $(hkl)$ ,  $(\bar{h}kl)$ ,  $(h'k'l)$  and  $(\bar{h}'k'l)$  for which  $d_{hkl}$  is the same, in space group  $P4/m$ .

charged particles. Such radiation sources exhibit a number of remarkable properties:

- *High brilliance*<sup>2</sup> – Synchrotron radiation possesses a very high degree of collimation. In other words, virtually all of the flux is emitted in an acutely narrow cone whose axis is tangential to the trajectory of the relativistic particles. The great flux is thus concentrated within a narrow region in space. The combination of an intense and fine source with a high degree of collimation results in a radiation of intense brilliance.
- *Choice of wavelength* – The radiation emitted by the bending magnets and the wigglers is polychromatic and spans a wide continuous spectral range, *i.e.* from ultraviolet to hard X-rays. Conversely, the radiation produced by undulators is emitted in a relatively narrow spectral range. The characteristic wavelength can be varied by simply modifying the distance separating the magnets within the undulators.
- *Pulsated structure in time*
- *Polarization characteristics*

The first two properties listed above represent those which are of a particularly significant utility for the exploitation of synchrotron radiation in the field of biomacromolecular crystallography (Dauter et al. 2010). In comparison to standard sources, the great intensity of a synchrotron beam allows for the reduction in exposure time. The high degree of brilliance generates a significantly improved signal to noise ratio in the measured intensities and thus gives rise to data of better quality. Finally, the wide choice in wavelength enables a wide range of experiments to be carried out and provides an ideal framework for multiple anomalous diffraction studies.

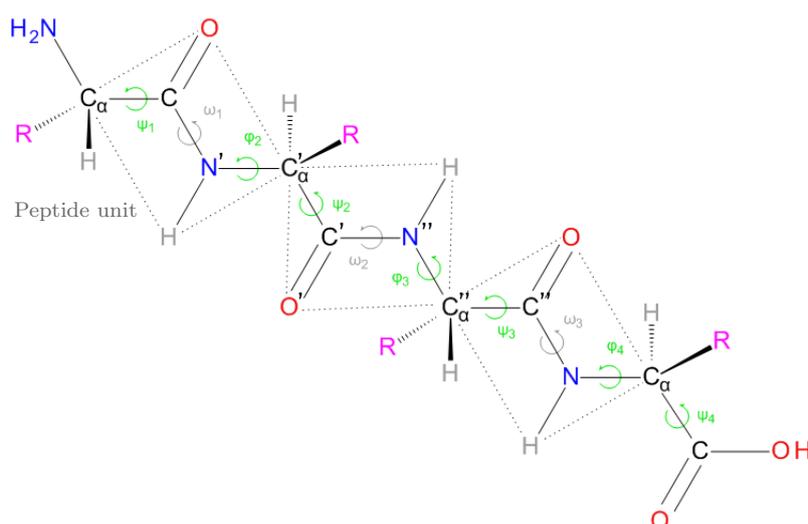
---

<sup>2</sup>Brilliance = photons per s per 0.1%  $\delta\lambda/\lambda$  per mrad<sup>2</sup> per mm<sup>2</sup>

# Appendix B

## Rigid body - definition of $\varphi$ , $\psi$ & $\omega$

$\mathbf{N}_{\text{ter}} \rightarrow \mathbf{C}_{\text{ter}}$	$\mathbf{C}_{\text{ter}} \rightarrow \mathbf{N}_{\text{ter}}$
$\varphi_2 = \theta(\mathbf{C}', \mathbf{C}\alpha', \mathbf{N}', \mathbf{C})$	$\varphi_2 = \theta(\mathbf{C}, \mathbf{N}', \mathbf{C}\alpha', \mathbf{C}')$
$\theta(\mathbf{R}, \mathbf{C}\alpha', \mathbf{N}', \mathbf{C}) = \varphi_2 - 120^\circ$	
$\psi_2 = \theta(\mathbf{N}'', \mathbf{C}', \mathbf{C}\alpha', \mathbf{N}')$	$\psi_2 = \theta(\mathbf{N}', \mathbf{C}\alpha', \mathbf{C}', \mathbf{N}'')$
$\theta(\mathbf{O}', \mathbf{C}', \mathbf{C}\alpha', \mathbf{N}') = \psi_2 \pm 180^\circ$	$\theta(\mathbf{R}, \mathbf{C}\alpha', \mathbf{C}', \mathbf{N}'') = \psi_2 + 120^\circ$
$\omega_2 = \theta(\mathbf{C}\alpha'', \mathbf{N}'', \mathbf{C}', \mathbf{C}\alpha')$	$\omega_2 = \theta(\mathbf{C}\alpha', \mathbf{C}', \mathbf{N}'', \mathbf{C}\alpha'')$
	$\theta(\mathbf{O}', \mathbf{C}', \mathbf{N}'', \mathbf{C}\alpha'') = \omega_2 \pm 180^\circ$



**Figure B.1** Representation of a generic polypeptide chain comprised of four amino acids showing the various backbone torsion angles as well as the peptide planes (dotted-lined rectangles).



# Appendix C

## Least squares approximation

The least squares routine used to approximate Ramachandran plots using data from the Top500 database was coded in Python. The Python code is depicted below. It also contains pieces of code to perform various other tasks, aside from the least squares routine, that were useful during the approximation process. These include ordering Fourier coefficients according to their absolute value, plotting data from the Top500 database and plotting a given Fourier series approximation.

```
1 from numpy import *
2 from pylab import *
3 import scipy
4 from scipy.optimize import leastsq
5 from math import sin,cos,radians,exp
6
7
8 def extract_xy_z(f_in, f_out):
9     xy = []
10    z = []
11    input = open(f_in, 'r')
12    lines = input.readlines()
13    input.close()
14    output = open(f_out, 'w')
15
16    for i in range(6,len(lines)):
17        line = lines[i].split()
18        xy.append((float(line[0]),float(line[1])))
19        if float(line[2]) == 0.0:
20            z.append(log(1.23e-10))
21        else:
22            z.append(log(float(line[2])))
23        output.write(str(float(line[0]))+" "+str(float(line[1]))+" "+str(z[i-6])+'\n')
24    output.close()
25    return xy,z
26
27
28 def F_serie(coefs,xy):
29     p = coefs
30     phi = radians(xy[0])
31     psi = radians(xy[1])
32
33     f = p[0] \
```

# This code uses the scientific packages NumPy and SciPy, and the plotting package PyLab.

# The function converts an empirically derived R-plot of the Top500 database ('f\_in') into a knowledge-based potential (written to a text file 'f\_out'). # 'f\_in' = a \*.data file and contains the columns phi, psi & P(phi,psi). The # function returns x, y & z which corresponds to phi, psi & E(phi,psi).

# The function calculates the Fourier series with 80 coefficients (not # counting the zero order coefficient). 'coefs' = a list of coefficients and # 'xy' = the list of (phi,psi) produced in 'extract\_xy\_z'. The function # returns approximated Fourier series, 'f'.

```

34     + p[1]*cos(phi) + p[2]*sin(phi) \
35     + p[3]*cos(psi) + p[4]*sin(psi) \
36     + p[5]*cos(phi)*cos(psi) + p[6]*sin(phi)*sin(psi) + p[7]*cos(phi)*sin(psi) + p[8]*sin(phi)*cos(psi) \
37     + p[9]*cos(2*phi) + p[10]*sin(2*phi) \
38     + p[11]*cos(2*psi) + p[12]*sin(2*psi) \
39     + p[13]*cos(2*phi)*cos(psi) + p[14]*sin(2*phi)*sin(psi) + p[15]*cos(2*phi)*sin(psi) + p[16]*sin(2*phi)*cos(psi) \
40     + p[17]*cos(phi)*cos(2*psi) + p[18]*sin(phi)*sin(2*psi) + p[19]*cos(phi)*sin(2*psi) + p[20]*sin(phi)*cos(2*psi) \
41     + p[21]*cos(3*phi) + p[22]*sin(3*phi) \
42     + p[23]*cos(3*psi) + p[24]*sin(3*psi) \
43     + p[25]*cos(2*phi)*cos(2*psi) + p[26]*sin(2*phi)*sin(2*psi) + p[27]*cos(2*phi)*sin(2*psi) + p[28]*sin(2*phi)*cos(2*psi) \
44     + p[29]*cos(3*phi)*cos(psi) + p[30]*sin(3*phi)*sin(psi) + p[31]*cos(3*phi)*sin(psi) + p[32]*sin(3*phi)*cos(psi) \
45     + p[33]*cos(phi)*cos(3*psi) + p[34]*sin(phi)*sin(3*psi) + p[35]*cos(phi)*sin(3*psi) + p[36]*sin(phi)*cos(3*psi) \
46     + p[37]*cos(4*phi) + p[38]*sin(4*phi) \
47     + p[39]*cos(4*psi) + p[40]*sin(4*psi) \
48     + p[41]*cos(3*phi)*cos(2*psi) + p[42]*sin(3*phi)*sin(2*psi) + p[43]*cos(3*phi)*sin(2*psi) + p[44]*sin(3*phi)*cos(2*psi) \
49     + p[45]*cos(2*phi)*cos(3*psi) + p[46]*sin(2*phi)*sin(3*psi) + p[47]*cos(2*phi)*sin(3*psi) + p[48]*sin(2*phi)*cos(3*psi) \
50     + p[49]*cos(4*phi)*cos(1*psi) + p[50]*sin(4*phi)*sin(1*psi) + p[51]*cos(4*phi)*sin(1*psi) + p[52]*sin(4*phi)*cos(1*psi) \
51     + p[53]*cos(1*phi)*cos(4*psi) + p[54]*sin(1*phi)*sin(4*psi) + p[55]*cos(1*phi)*sin(4*psi) + p[56]*sin(1*phi)*cos(4*psi) \
52     + p[57]*cos(3*phi)*cos(3*psi) + p[58]*sin(3*phi)*sin(3*psi) + p[59]*cos(3*phi)*sin(3*psi) + p[60]*sin(3*phi)*cos(3*psi) \
53     + p[61]*cos(4*phi)*cos(2*psi) + p[62]*sin(4*phi)*sin(2*psi) + p[63]*cos(4*phi)*sin(2*psi) + p[64]*sin(4*phi)*cos(2*psi) \
54     + p[65]*cos(2*phi)*cos(4*psi) + p[66]*sin(2*phi)*sin(4*psi) + p[67]*cos(2*phi)*sin(4*psi) + p[68]*sin(2*phi)*cos(4*psi) \
55     + p[69]*cos(4*phi)*cos(3*psi) + p[70]*sin(4*phi)*sin(3*psi) + p[71]*cos(4*phi)*sin(3*psi) + p[72]*sin(4*phi)*cos(3*psi) \
56     + p[73]*cos(3*phi)*cos(4*psi) + p[74]*sin(3*phi)*sin(4*psi) + p[75]*cos(3*phi)*sin(4*psi) + p[76]*sin(3*phi)*cos(4*psi) \
57     + p[77]*cos(4*phi)*cos(4*psi) + p[78]*sin(4*phi)*sin(4*psi) + p[79]*cos(4*phi)*sin(4*psi) + p[80]*sin(4*phi)*cos(4*psi)
58     return f
59
60
61 def residuals(coefs, z, xy):                                # The function represents the objective function minimized by 'least_sq'.
62     err = []                                               # 'coefs' = a list of the Fourier coefficients, 'z' = the list of E(phi, psi) and
63                                                         # 'xy' = the list of (phi, psi) both produced in 'extract_xy_z'. The function
64     assert len(xy) == len(z)                               # returns the residue, 'err', of the minimization.
65     for count in range(len(z)):
66         err.append(abs(z[count] - F_serier(coefs, xy[count])))
67     return err
68
69
70 def least_sq(num_c, z, xy):                                # The function carries out the least squares refinement of the Fourier series.
71     c_ini = []                                             # 'num_c' = a number which defines the number of Fourier coefficients, 'z' = the
72                                                         # list of E(phi, psi) and 'xy' = the list of (phi, psi) both produced in
73     for i in range(0, num_c):                               # 'extract_xy'. The function returns a list of coefficients, 'plsq[0]'.
74         c_ini.append(1.)
75     plsq = leastsq(residuals, c_ini, args=(z, xy))
76     print plsq[0],
77     return plsq[0]
78
79
80 def order_coefs(f_in, f_out):                              # The function sorts a list of coefficients in decreasing order of absolute
81     output = open(f_out, 'w')                             # value. 'f_in' = a file containing a list of coefficients and 'f_out' = a file
82     dic_coef = {}                                         # in which the sorted list is written.
83
84     input = open(f_in, 'r')
85     lines = input.readlines()
86     input.close()
87     coefs = []
88
89     for line in lines:
90         coefs.append(float(line))
91
92     for i in range(0, len(coefs)):
93         dic_coef[abs(coefs[i])] = i
94     coef_tup = dic_coef.items()
95     coef_tup.sort()
96     coef_tup.reverse()
97     for i in range(0, len(coef_tup)):
98         output.write(str(coef_tup[i][0])+" "+str(coef_tup[i][1])+"\n")
99     output.close()
100
101
102 def F_serier_red(coefs, xy):                              # The function calculates a reduced Fourier series. 'coefs' = a reduced list of
103     p = coefs                                             # coefficients sorted by absolute value. The function returns the approximation
104     phi = radians(xy[0])                                  # of a reduced Fourier series (with 10 coefficients as it stands), 'f'.

```

```

105     psi = radians(xy[1])
106
107     f = p[0] + p[1]*cos(phi) + p[9]*cos(2*phi) + p[26]*sin(2*phi)*sin(2*psi) + p[18]*sin(phi)*sin(2*psi) \
108         + p[42]*sin(3*phi)*sin(2*psi) + p[21]*cos(3*phi)+ p[11]*cos(2*psi) + p[5]*cos(phi)*cos(psi) + p[13]*cos(2*phi)*cos(psi) \
109         + p[25]*cos(2*phi)*cos(2*psi)
110     return f
111
112
113 def plot_rama(z):                                     # The function plots an empirically derived Ramachandran plot using P(phi,psi),
114     arr = zeros((180,180))                             # 'z', as extracted by 'extract_xy-z'.
115     cpt = 0
116
117     for i in range(0, 180):
118         for j in range(0, 180):
119             arr[i,j] = -z[cpt]
120             cpt +=1
121     figure(1)
122     imshow(arr)
123     show()
124
125
126 def plot_rama_serie(f_in ,xy,f,f_out):                # The function plots an approximated Fourier series. 'f_in' = a file containing a
127     arr = zeros((180,180))                             # list of coefficients, 'xy' = the list of (phi,psi) produced by 'extract_xy-z',
128     cpt = 0                                             # f = a Fourier series (i.e. either 'F_serie' or 'F_serie_red') and 'f_out' = a
129     output = open(f_out, 'w')                          # file containing the coordinates of the plot.
130
131     input = open(f_in, 'r')
132     lines = input.readlines()
133     input.close()
134     coefs = []
135
136     for line in lines:
137         coefs.append(float(line))
138
139     for i in range(0, 180):
140         for j in range(0, 180):
141             arr[i,j] = f(coefs,xy[cpt])
142             output.write(str(xy[cpt][0])+" "+str(xy[cpt][1])+" "+str(arr[i,j])+'\n')
143             cpt +=1
144     output.close()
145     print arr.min(),arr.max()
146     figure(2)
147     imshow(arr)
148     show()

```



# Bibliography

- Abrahams, J. P. (1997), 'Bias Reduction in Phase Refinement by Modified Interference Functions: Introducing the  $\gamma$  Correction', *Acta Cryst.* **D53**, 371–376.
- Abrahams, J. P. & Leslie, A. G. W. (1996), 'Methods used in the structure determination of bovine mitochondrial F<sub>1</sub> ATPase', *Acta Cryst.* **D52**, 30–42.
- Admiraal, G. & Verweij, A. B. (1982), 'Crystallization of ACTH Oligopeptides', *J. Cryst. Growth* **57**, 446–448.
- Admiraal, G. & Vos, A. (1984), 'Antiparallel  $\beta$ -sheets in the crystal structure of the heptapeptide Met-Glu-His-Phe-Arg-Trp-Gly (ACTH 4-10)', *Int. J. Peptide Protein Res.* **23**, 151–157.
- Albesa-Jove, D., Kariuki, B. M., Kitchin, S. J., Cheung, E. Y. & Harris, K. D. M. (2004), 'Challenges in direct-space structure determination from powder diffraction data: A molecular material with four independent molecules in the asymmetric unit', *Chem. Phys. Chem.* **5**(3), 414–418.
- Alcorn, T. & Juers, D. H. (2010), 'Progress in rational methods of cryoprotection in macromolecular crystallography', *Acta Cryst.* **D66**, 366–373.
- Allaire, M., Moiseeva, N., Botez, C. E., Engel, M. A. & Stephens, P. W. (2009), 'On the possibility of using polycrystalline material in the development of structure-based generic assays', *Acta Cryst.* **D65**(379–382).
- Allen, F. H. (2002), 'The Cambridge Structural Database: a quarter of a million crystal structures and rising', *Acta Cryst.* **B58**, 380–388.
- Amos, L. A., Jubb, J. S., Henderson, R. A. & Vigers, G. (1984), 'Arrangement of protofilaments in two forms of tubulin crystal induced by vinblastine', *J. Mol. Biol.* **178**(3), 711–729.

- Andreev, Y. G., Lightfoot, P. & Bruce, P. G. (1997), 'A General Monte Carlo Approach to Structure Solution from Powder-Diffraction Data: Application to Poly(ethylene oxide)<sub>3</sub>:LiN(SO<sub>2</sub>CF<sub>3</sub>)<sub>2</sub>', *J. Appl. Cryst.* **30**, 294–305.
- Astbury, W. T. (1933), *Fundamentals of fiber structure*, Oxford University Press.
- Balbirnie, M., Grothe, R. & Eisenberg, D. (2001), 'An amyloid-forming peptide from the yeast prion Sup35 reveals a dehydrated  $\beta$ -sheet structure for amyloid', *Proc. Natl Acad. Sci. USA* **98**(5), 2375–2380.
- Basso, S., Besnard, C., Wright, J. P., Margiolaki, I., Fitch, A. N., Pattison, P. & Schiltz, M. (2010), 'Features of the secondary structure of a protein molecule from powder diffraction data', *Acta Cryst.* **D66**, 756–761.
- Basso, S., Fitch, A. N., Fox, G. C., Margiolaki, I. & Wright, J. P. (2005), 'High-throughput phase-diagram mapping *via* powder diffraction: a case study of HEWL *versus* pH', *Acta Cryst.* **D61**, 1612–1625.
- Basu, S. K., Govardhan, C. P., Jung, C. W. & Margolin, A. L. (2004), 'Protein crystals for the delivery of biopharmaceuticals', *Expert Opin. Biol. Ther.* **4**(3), 301–317.
- Bernal, J. D. (1939), 'Structure of proteins', *Nature* **143**, 663–667.
- Bernal, J. D. & Crowfoot, D. (1934), 'X-ray photographs of crystalline pepsin', *Nature* **133**, 794–795.
- Bernal, J. D. & Fankuchen, I. (1941), 'X-ray and crystallographic studies of plant virus preparations. III', *J. Gen. Physiol.* **25**(1), 147–165.
- Besnard, C., Camus, F., Fleurant, M., Dahlström, Wright, J. P., Margiolaki, I., Pattison, P. & Schiltz, M. (2007), 'Exploiting X-ray induced anisotropic lattice changes to improve intensity extraction in protein powder diffraction: Application to heavy atom detection', *Z. Kristallogr. Suppl.* **26**, 39–44.
- Blake, C. C. F., Fenn, R. H., North, A. C. T., Phillips, D. C. & Poljak, R. J. (1962), 'Structure Of lysozyme: A fourier map of the electron density at 6 Å resolution obtained by x-ray diffraction', *Nature* **196**(4860), 1173–1176.
- Blake, C. C. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1965), 'Structure of hen egg-white lysozyme. A three-dimensional Fourier synthesis at 2 Å resolution', *Nature* **206**, 757–761.

- Blake, C. C. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1967), 'On the Conformation of the Hen Egg-White Lysozyme Molecule', *Proc. R. Soc. London B* **167**, 365–377.
- Blow, D. M. & Crick, F. H. C. (1959), 'The treatment of errors in the isomorphous replacement method', *Acta Cryst.* **12**, 794–802.
- Blundell, T. L. & Jenkins, J. A. (1977), 'The binding of heavy metals to proteins', *Chem. Soc. Rev.* **6**, 139–171.
- Blundell, T. L. & Johnson, L. N. (1976), *Protein crystallography*, Academic Press, New York.
- Bokhoven, C., Schoone, J. C. & Bijvoet, J. M. (1951), 'The Fourier synthesis of the crystal structure of strychnine sulphate pentahydrate', *Acta Cryst.* **4**, 275–280.
- Boyes-Watson, J., Davidson, E. & Perutz, M. F. (1947), 'An X-ray study of horse methaemoglobin. I', *Proc. R. Soc. London A* **191**(1024), 83–132.
- Bragg, W. L. (1913), 'The determination of some crystals as indicated by their diffraction of X-rays', *Proc. R. Soc. London A* **89**(610), 248–277.
- Bricogne, G. (1991), 'A multiresolution method of phase determination by combined maximization of entropy and likelihood. III. Extension to powder diffraction data', *Acta Cryst.* **47**, 803–829.
- Bricogne, G. (2000), Crystallographic moment-generating and likelihood functions, *in* D. Cocolicchio, G. Dattoli & H. M. Srivastava, eds, 'Proceedings of the Workshop on Advanced Special Functions and Applications, Melfi (PZ), Italy, 9-12 May 1999', Aracne Editrice, Rome, pp. 315–321.
- Bricogne, G., Vonrhein, C., Flensburg, C., Schiltz, M. & Paciorek, W. (2003), 'Generation, representation and flow of phase information in structure determination: recent developments in and around *SHARP 2.0*', *Acta Cryst.* **D59**, 2023–2030.
- Brinkmann, C., Weiss, M. S. & Weckert, E. (2006), 'The structure of the hexagonal crystal form of hen egg-white lysozyme', *Acta Cryst.* **D62**, 349–355.
- Brunelli, M., Wright, J. P., Vaughan, G. B. M., Mora, A. J. & Fitch, A. N. (2003), 'Solving Larger Molecular Crystal Structures from Powder Diffraction Data by Exploiting Anisotropic Thermal Expansion', *Angew. Chem. Int. Ed.* **42**, 2029–2032.

- Brunger, A. T., DeLabBarre, B., Davies, J. M. & Weis, W. I. (2009), 'X-ray structure determination at low resolution', *Acta Cryst.* **D65**, 128–133.
- Budisa, N., Karnbrock, W., Steinbacher, S., Humm, A., Prade, L., Neufeind, T., Moroder, L. & Huber, R. (1997), 'Bioincorporation of telluromethionine into proteins: a promising new approach for X-ray structure analysis of proteins', *J. Mol. Biol.* **271**, 1–8.
- Černý, R. & Favre-Nicolin, V. (2007), 'Direct space methods of structure determination from powder diffraction: principles, guidelines and perspectives', *Z. Kristallogr.* **222**(3–4), 105–113.
- Černý, V. (1985), 'Thermodynamical Approach to the Traveling Salesman Problem: An Efficient Simulation Algorithm', *J. Opt. Theory Appl.* **45**(1), 41–51.
- Chayen, N. E. (1996), 'A novel technique for containerless protein crystallization', *Protein Eng.* **9**, 927–929.
- Chen, V. B., Arendall III, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, D. S. & Richardson, D. C. (2010), 'MolProbity: all-atom structure validation for macromolecular crystallography', *Acta Cryst.* **D66**, 12–21.
- Cheung, E. Y., McCabe, E. E., Harris, K. D. M., Johnston, R. L., Tedesco, E., Raja, K. M. P. & Balaram, P. (2002), 'C–H···O Hydrogen Bond Mediated Chain Reversal in a Peptide Containing a  $\gamma$ -Amino Acid Residue, Determined Directly from Powder X-ray Diffraction Data', *Angew. Chem. Int. Ed.* **41**(3), 494–496.
- Chong, S. Y., Seaton, C. C., Kariuki, B. M. & Tremayne, M. (2006), 'Molecular vs crystal symmetry in tri-substituted triazine, benzene and isocyanurate derivatives', *Acta Cryst.* **B62**, 864–874.
- Cochran, W., Crick, F. H. C. & Vand, V. (1952), 'The structure of synthetic polypeptides. I. The transform of atoms on a helix', *Acta Cryst.* **5**, 581–586.
- Coelho, A. A. (2000), 'Whole-profile structure solution from powder diffraction data using simulated annealing', *J. Appl. Cryst.* **33**, 899–908.
- Coelho, A. A. (2004), 'Topas-Academic', <http://www.topas-academic.net/>.

- Collaborative Computational Project, Number 4 (1994), 'The ccp4 suite: Programs for protein crystallography', *Acta Cryst.* **D50**, 760–763.
- Collings, I., Watier, Y., Giffard, M., Dagogo, S., Kahn, R., Bonneté, F., Wright, J. P., Fitch, A. N. & Margiolaki, I. (2010), 'Polymorphism of microcrystalline urate oxidase from *Aspergillus flavus*', *Acta Cryst.* **D66**, 539–548.
- Corey, R. B. & Wyckoff, R. W. G. (1936), 'Long spacings in macromolecular solids', *J. Biol. Chem.* **114**, 407–414.
- Cowtan, K. (1994), 'dm': An automated procedure for phase improvement by density modification', *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography* **31**, 34–38.
- Cowtan, K. (1998), 'Modified Phased Translation Functions and their Application to Molecular-Fragment Location', *Acta Cryst.* **D54**, 750–756.
- Cowtan, K. (1999), 'Error estimation and bias correction in phase-improvement calculations', *Acta Cryst.* **D55**, 1555–1567.
- Crick, F. H. C. & Magdoff, B. S. (1956), 'The theory of the method of isomorphous replacement for protein crystals. I', *Acta Cryst.* **9**, 901–908.
- Cudney, B., Patel, S. & McPherson, A. (1994), 'Crystallization of macromolecules in silica gels', *Acta Cryst.* **D50**, 479–483.
- Dauter, Z., Jaskolski, M. & Wlodawer, A. (2010), 'Impact of synchrotron radiation on macromolecular crystallography: a personal view', *J. Syn. Rad.* **17**, 433–444.
- David, W. I. F. (2004), 'On the equivalence of the Rietveld method and the correlated integrated intensities method in powder diffraction', *J. Appl. Cryst.* **37**, 621–628.
- David, W. I. F. & Shankland, K. (2008), 'Structure determination from powder diffraction data', *Acta Cryst.* **64**, 52–64.
- David, W. I. F., Shankland, K., van de Streek, J., Pidcock, E., Motherwell, W. D. S. & Cole, J. C. (2006), 'DASH: a program for crystal structure determination from powder diffraction data', *J. Appl. Cryst.* **39**, 910–915.
- De Wied, D. (1969), 'Effects of peptide hormones on behavior', *Front. Neuroendocrinol.* **1**, 97–140.

- Deem, M. W. & Newsam, J. M. (1989), 'Determination of 4-connected framework crystal structures by simulated annealing', *Nature* **342**(6247), 260–262.
- Deschamps, J. R., George, C. & Flippen-Anderson, J. L. (1996), '[D-Ala<sup>2</sup>,D-Leu<sup>5</sup>]-enkephalin hydrochloride', *Acta Cryst.* **52**(6), 1583–1585.
- Diaz-Avalos, R., Long, C., Fontano, E., Balbirnie, M., Grothe, R., Eisenberg, D. & Caspar, D. L. D. (2003), 'Cross-beta Order and Diversity in Nanocrystals of an Amyloid-forming Peptide', *J. Mol. Biol.* **330**(5), 1165–1175.
- Doebbler, J. A. & Von Dreele, R. B. (2009), 'Application of molecular replacement to protein powder data from image plates', *Acta Cryst.* **D65**, 348–355.
- Elfron, B. & Tibshirani, R. (1986), 'Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy', *Statistical Science* **1**, 54–75.
- Engh, R. A. & Huber, R. (1991), 'Accurate bond and angle parameters for X-ray protein structure refinement', *Acta Cryst.* **A47**, 392–400.
- Engh, R. A. & Huber, R. (2001), *International Tables of Crystallography Volume F: Crystallography of biological macromolecules*, The International Union of Crystallography, pp. 382–392.
- Evans, G., Axford, D., Waterman, D. & Owen, R. L. (2011), 'Macromolecular microcrystallography', *Crystallogr. Rev.* **17**(2), 105–142.
- Ewald, P. P. (1921), 'Das "reziproke Gitter" in der Strukturtheorie', *Z. Kristallogr.* **A56**, 129–159.
- Favre-Nicolin, V. & Černý, R. (2002), 'FOX, 'free objects for crystallography': a modular approach to ab initio structure determination from powder diffraction', *J. Appl. Cryst.* **35**, 734–743.
- Favre-Nicolin, V. & Černý, R. (2004), 'A better FOX: Using flexible modelling and maximum likelihood to improve direct-space ab initio structure determination from powder diffraction', *Z. Kristallogr.* **219**(12), 847–856.
- Fernandes, P., Shankland, K., Florence, A. J., Shankland, N. & Johnston, A. (2007), 'Solving molecular crystal structures from X-ray powder diffraction data: The challenges posed by -carbamazepine and chlorothiazide N,N,-dimethylformamide (1/2) solvate', *J. Pharm. Sci.* **96**(5), 1192–1202.

- Fitch, A. N. (2004), 'The High Resolution Powder Diffraction Beam Line at ESRF', *J. Res. Natl Inst. Stand. Technol.* **109**, 133–142.
- Fleming, A. (1922), 'On a Remarkable Bacteriolytic Element Found in Tissues and Secretions', *Proc. Roy. Soc.* **93**(B), 306–317.
- Fortelle, E. & Bricogne, G. (1997), 'Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods', *Methods Enzymol.* **276**, 472–494.
- Franklin, R. & Gosling, R. G. (1953), 'Molecular configuration in sodium thymonucleate', *Nature* **171**(4356), 740–741.
- Friedrich, W., Knipping, P. & von Laue, M. (1912), 'Interferenz-Erscheinungen bei Röntgenstrahlen', *Sitzgsber. Math. Phys. Kl. K. Bayer. Akad. Wiss.* pp. 303–322.
- Garman, E. F. (2010), 'Radiation damage in macromolecular crystallography: what is it and why should we care?', *Acta Cryst.* **D66**, 339–351.
- Girard, E., Chantalat, L., Vicat, J. & Kahn, R. (2002), 'Gd-HPDO3A, a complex to obtain high-phasing-power heavy-atom derivatives for SAD and MAD experiments: results with tetragonal hen egg-white lysozyme', *Acta Cryst.* **D58**, 1–9.
- Girard, E., Stelter, M., Vicat, J. & Kahn, R. (2003), 'A new class of lanthanide complexes to obtain high-phasing-power heavy-atom derivatives for macromolecular crystallography', *Acta Cryst.* **D59**, 1914–1922.
- Green, D. W., Ingram, V. M. & Perutz, M. F. (1954), 'The structure of Haemoglobin. IV. Sign determination by the isomorphous replacement method', *Proc. R. Soc. London A* **225**(1162), 287–307.
- Gruner, S. M. (2010), 'Synchrotron area X-ray detectors, present and future', *AIP Conf. Proc.* **1234**, 69–72.
- Hampel, A., Labanauskas, M., Connors, P. G., Kirkegard, L., Rajbhandary, U. L., Sigler, P. B. & Bock, R. M. (1968), 'Single crystals of transfer RNA from formylmethionine and phenylalanine transfer RNA's ', *Science* **162**, 1384–1386.
- Harata, K. & Akiba, T. (2006), 'Structural phase transition of monoclinic crystals of hen egg-white lysozyme', *Acta Cryst.* **D62**, 375–382.

- Harker, D. (1956), 'The determination of the phases of the structure factors of non-centrosymmetric crystals by the method of double isomorphous replacement', *Acta Cryst.* **9**, 1–9.
- Helliwell, J. R., Helliwell, M. & Jones, R. H. (2005), 'Ab initio structure determination using dispersive differences from multiple-wavelength synchrotron-radiation powder diffraction data', *Acta Cryst.* **A61**, 568–574.
- Hendrickson, W. A. (1985), 'Analysis of protein structure from diffraction measurement at multiple wavelengths', *Trans. Am. Crystallogr. Assoc.* **21**, 11–21.
- Hendrickson, W. A. (1991), 'Determination of macromolecular structures from anomalous diffraction of synchrotron radiation', *Science* **254**(5028), 51–58.
- Hendrickson, W. A. (1999), 'Maturation of MAD phasing for the determination of macromolecular structures', *J. Syn. Rad.* **6**, 845–851.
- Hendrickson, W. A., Horton, J. R. & LeMaster, D. M. (1990), 'Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure', *EMBO J.* **9**, 1665–1672.
- Hendrickson, W. A. & Teeter, M. M. (1981), 'Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulphur', *Nature* **290**(5802), 107–113.
- Hodsdon, J. M., Brown, G. M., Sieker, L. C. & Jensen, L. H. (1990), 'Refinement of triclinic lysozyme: I. Fourier and least-squares methods', *Acta Cryst.* **B46**, 54–62.
- Hooft, R., Vriend, G., Sander, C. & Abola, E. (1996), 'Errors in protein structures.', *Nature* **381**, 272–272.
- Islam, S. A., Carvin, D., Sternberg, M. J. E. & Blundell, T. L. (1998), 'A Databank of Heavy-Atom binding sites in protein crystals: A resource for use in multiple isomorphous replacement and anomalous scattering', *Acta Cryst.* **D54**, 1199–1206.
- IUPAC-IUB Commission on Biochemical Nomenclature (1970), 'Abbreviations and Symbols for the Description of the Conformation of Polypeptide Chains', *J. Mol. Biol.* **52**, 1–17.

- IUPAC-IUB Joint Commission on Biochemical Nomenclature (1984), 'Nomenclature and symbolism for amino acids and peptides (Recommendations 1983)', *Pure Appl. Chem.* **56**(5), 595–624.
- Jakoncic, J., Di Michiel, M., Zhong, Z., Honkimaki, V., Jouanneau, Y. & Stojanoff, V. (2006), 'Anomalous diffraction at ultra-high energy for protein crystallography', *J. Appl. Cryst.* **39**, 831–841.
- Jenner, M. J., Wright, J. P., Margiolaki, I. & Fitch, A. N. (2007), 'Successful protein cryocooling for powder diffraction', *J. Appl. Cryst.* **40**, 121–124.
- Judge, R. A., Jacobs, R. S., Frazier, T., Snell, E. H. & Pusey, M. L. (1999), 'The effect of temperature and solution pH on the nucleation of tetragonal lysozyme crystals', *Biophys. J.* **77**, 1585–1593.
- Kariuki, B. M., Psallidas, K., Harris, K. D. M., Johnston, R. L., Lancaster, R. W., Staniforth, S. E. & Cooper, S. M. (1999), 'Structure determination of a steroid directly from powder diffraction data', *Chem. Commun.* pp. 1677–1678.
- Kariuki, B. M., Serrano-Gonzalez, H., Johnston, R. L. & Harris, K. D. M. (1997), 'The application of a genetic algorithm for solving the crystal structures from powder diffraction data', *Chem. Phys. Lett.* **280**, 189–195.
- Karle, J. (1980), 'Some developments in anomalous dispersion for the structural investigation of macromolecular systems in biology', *Int. J. Quant. Chem.* **S7**, 357–367.
- Karmali, A. M., Blundell, T. L. & Furnham, N. (2009), 'Model-building strategies for low-resolution X-ray crystallographic data', *Acta Cryst.* **D65**, 121–127.
- Kaul, R. & Balaram, P. (1999), 'Stereochemical control of peptide folding', *Bioorg. Med. Chem.* **7**(1), 105–117.
- Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, R. W. G. & Phillips, D. C. (1958), 'A three-dimensional model of the myoglobin molecule obtained by X-ray analysis', *Nature* **181**(4610), 662–666.
- Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C. & Shore, V. C. (1960), 'Structure of myoglobin. A three-dimensional Fourier synthesis at 2 Å resolution', *Nature* **185**, 422–427.

- Kepler, J. (1611), *Strena seu de Nive Sexangula*, Frankfurt: G. Tampach.
- Kirkpatrick, S., Gelatt Jr., C. D. & Vecchi, M. P. (1983), 'Optimization by Simulated Annealing', *Science* **220**(4598), 671–680.
- Lalonde, J. J., Govardhan, C. P., Khalaf, N., Martinez, A. G., Visuri, K. & Margolin, A. L. (1995), 'Cross-linked crystals of *Candida rugosa* lipase: Highly efficient catalysts for the resolution of chiral esters', *J. Am. Chem. Soc.* **117**(26), 6845–6852.
- Larson, A. C. & Von Dreele, R. B. (2004), 'General Structure Analysis System (GSAS)', Technical report, Los Alamos National Laboratory Report LAUR-86-748, Los Alamos, USA.
- 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. Cryst.* **26**, 283–291.
- LeBail, A., Duroy, H. & Fourquet, J. L. (1988), 'Ab-initio structure determination of LiSbWO<sub>6</sub> by X-ray powder diffraction', *Mater. Res. Bull.* **23**(3), 447–452.
- Leslie, A. G. W. (1993), *Proceedings of CCP4 Study Weekend on Data Collection and Processing*, Daresbury: SERC Daresbury Laboratory, pp. 44–51.
- Lipson, H. & Beevers, C. A. (1935), 'The crystal structure of the alums', *Proc. R. Soc. London A* **148**(865), 664–680.
- Lonsdale, K. (1928), 'The structure of the benzene ring', *Nature* **122**(3082), 810.
- Lorber, B. & Giegé, R. (1996), 'Containerless protein crystallization in floating drops: application to crystal growth monitoring under reduced nucleation conditions', *J. Cryst. Growth* **168**, 204–215.
- Lovell, S., Davis, I., Arendall III, W., De Bakker, P., Word, J., Prisant, M., Richardson, J. & Richardson, D. (2003), 'Structure validation by C $\alpha$  geometry:  $\varphi$ ,  $\psi$  and C $\beta$  deviation', *Proteins* **50**, 437–450.
- MacArthur, M. W. & Thornton, J. M. (1991), 'Influence of proline residues on protein conformation', *J. Mol. Biol.* **218**(2), 397–412.

- Macrae, C. F., Edgington, P. R., McCabe, P., Pidcock, E., Shields, G. P., Taylor, R., Towler, M. & van de Streek, J. (2006), 'Mercury: visualization and analysis of crystal structures', *J. Appl. Cryst.* **39**, 453–457.
- Margiolaki, I., Wright, J. P., Fitch, A. N., Fox, G. C., Labrador, A., Von Dreele, R. B., Miura, K., Gozzo, F., Schiltz, M., Besnard, C., Camus, F., Pattison, P., Beckers, D. & Degen, T. (2007), 'Powder diffraction studies on proteins: An overview of data collection approaches', *Z. Kristallogr. Suppl.* **26**, 1–13.
- Margiolaki, I., Wright, J. P., Fitch, A. N., Fox, G. C. & Von Dreele, R. B. (2005), 'Synchrotron X-ray powder diffraction study of hexagonal turkey egg-white lysozyme', *Acta Cryst.* **D61**(423–432).
- Margiolaki, I., Wright, J. P., Wilmanns, M., Fitch, A. N. & Pinotsis, N. (2007), 'Second SH3 Domain of Ponsin Solved from Powder Diffraction', *J. Am. Chem. Soc.* **129**, 11865–11871.
- Margolin, A. L. & Navia, M. A. (2001), 'Protein Crystals as Novel Catalytic Materials', *Angew. Chem. Int. Ed.* **40**(12), 2204–2222.
- Matthews, B. W. (1985), 'Determination of protein molecular weight, hydration, and packing from crystal density', *Methods Enzymol.* **114**, 176–187.
- McPherson, A. (1982), *The preparation and analysis of protein crystals*, John Wiley and Sons, New York.
- Metropolis, N., Rosenblith, A. W. & Freeman, C. M. (1953), 'Equation of state calculations by fast computing machines', *J. Chem. Phys.* **21**, 1087–1092.
- Michalewicz, Z. (1996), *Genetic Algorithms + Data Structures = Evolution Programs*, 3rd edn, Springer-Verlag, Berlin.
- Mitchell, C. M. (1957), 'Phase determination by the two-wavelength method of Okaya & Pepinsky', *Acta Cryst.* **10**, 475–476.
- Nelson, R., Sawaya, M. R., Balbirnie, M., Madsen, A. Ø., Riekel, C., Grothe, R. & Eisenberg, D. (2005), 'Structure of the cross- $\beta$  spine of amyloid-like fibrils', *Nature* **435**(7043), 773–778.

- Noda, M., Teranishi, Y., Takahashi, H., Toyosato, M., Notake, M., Nakanishi, S. & Numa, S. (1982), 'Isolation and structural organization of the human preproenkephalin gene', *Nature* **297**(5865), 431–434.
- Norrman, M., Ståhl, K., Schluckebier, G. & Al-Karadaghi, S. (2006), 'Characterization of insulin microcrystals using powder diffraction and multivariate data analysis', *J. Appl. Cryst.* **39**, 391–400.
- Okaya, Y. & Pepinsky, R. (1956), 'New formulation and solution of the phase problem in X-ray analysis of noncentric crystals containing anomalous scatterers', *Phys. Rev.* **103**(6), 1645–1647.
- Pagola, S. & Stephens, P. W. (2000), 'Towards the solution of organic crystal structures by powder diffraction', *Mater. Sci. Forum* **321–323**, 40–45.
- Pan, Z. G., Xu, M. C., Cheung, E. Y., Harris, K. D. M., Constable, E. C. & Housecroft, C. E. (2006), 'Understanding the Structural Properties of a Dendrimeric Material Directly from Powder X-ray Diffraction Data', *J. Phys. Chem B* **110**(24), 11620–11623.
- Pauling, L. & Corey, R. B. (1951), 'Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets', *Proc. Natl Acad. Sci. USA* **37**, 729–740.
- Pauling, L., Corey, R. B. & Branson, H. R. (1951), 'The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain', *Proc. Natl Acad. Sci. USA* **37**, 205–211.
- Pawley, G. S. (1981), 'Unit-cell refinement from powder diffraction scans', *J. Appl. Cryst.* **14**, 357–361.
- Persichetti, R. A., St. Clair, N. L., Griffith, J. P., Navia, M. A. & Margolin, A. L. (1995), 'Cross-Linked Enzyme Crystals (CLECs) of Thermolysin in the Synthesis of Peptides', *J. Am. Chem. Soc.* **117**(10), 2732–2737.
- Pertsemlidis, A., Zelinka, J., Fondon, J. W., Henderson, R. K. & Otwinowski, Z. (2005), 'Bayesian Statistical Studies of the Ramachandran Distribution', *Stat. Appl. Genet. Mol. Biol.* **4**(1), art. no. 35.

- Perutz, M. F. (1949), 'An X-ray study of horse methaemoglobin. II', *Proc. R. Soc. London A* **195**(1043), 474–499.
- Perutz, M. F. (1951*a*), 'New X-ray evidence on the configuration of polypeptide chains', *Nature* **167**, 1053–1054.
- Perutz, M. F. (1951*b*), 'The 1.5-Å reflexion from proteins and polypeptides', *Nature* **168**, 653–654.
- Perutz, M. F. (1956), 'Isomorphous replacement and phase determination in non-centrosymmetric space groups', *Acta Cryst.* **9**, 867–873.
- Perutz, M. F., Muirhead, H., Cox, J. M. & Goaman, L. C. G. (1968), 'Three-dimensional Fourier synthesis of horse oxyhaemoglobin at 2.8 Å resolution: The atomic model', *Nature* **219**(5150), 131–139.
- Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G. & North, A. C. T. (1960), 'Structure of haemoglobin: A three-dimensional Fourier synthesis at 5.5 Å resolution, obtained by X-ray analysis', *Nature* **185**(4711), 416–422.
- Pflugrath, J. W. (2004), 'Macromolecular cryocrystallography – methods for cooling and mounting protein crystals at cryogenic temperatures', *Methods* **34**(3), 415–423.
- Pompidor, G., Maury, O., Vicat, J. & Kahn, R. (2010), 'A dipicolinate lanthanide complex for solving protein structures using anomalous diffraction', *Acta Cryst.* **D66**, 762–769.
- Putz, H., Schön, J. C. & Jansen, M. (1999), 'Combined method for ab initio structure solution from powder diffraction data', *J. Appl. Cryst.* **32**, 864–870.
- Ramachandran, G., Ramakrishnan, C. & Sasisekharan, V. (1963), 'Stereochemistry of polypeptide chain configurations', *J. Mol. Biol.* **7**, 95–99.
- Ramakrishnan, C. & Ramachandran, G. (1965), 'Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units', *Biophys. J.* **5**(6), 909–933.
- Ravelli, R. B. G., Theveneau, P., McSweeney, S. & Caffrey, M. (2002), 'Unit-cell volume change as a metric of radiation damage in crystals of macromolecules', *J. Syn. Rad.* **9**, 355–360.

- Röntgen, W. C. (1895), 'Ueber eine neue Art von Strahlen (Vorläufige Mittheilung)', *Sitzungsber. Würzburger Phys Med. Gesellschaft*.
- Rietveld, H. M. (1969), 'A profile refinement method for nuclear and magnetic structures', *J. Appl. Cryst.* **2**, 65–71.
- Robert, M. C. & Lefauchaux, F. (1988), 'Crystal growth in gels: principles and applications', *J. Cryst. Growth* **90**, 358–367.
- Robertson, J. M. (1936), 'An X-ray study of the phthalocyanines. Part II. Quantitative structure determination of the metal-free compound', *J. Chem. Soc.* pp. 1195–1209.
- Rodgers, D. W. (1994), 'Cryocrystallography', *Structure* **2**(12), 1135–1140.
- Rupp, B. (2009), *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*, 1<sup>st</sup> edn, Garland Science, New York.
- Rypniewski, W. R., Holden, H. M. & Rayment, I. (1993), 'Structural consequences of reductive methylation of lysine residues in hen egg white lysozyme: An x-ray analysis at 1.8 Å resolution', *Biochemistry* **32**, 9851–9858.
- Salemme, F. R. (1972), 'A free interface diffusion technique for the crystallization of proteins for X-ray crystallography', *Arch. Biochem. Biophys.* **151**, 533–540.
- Sauter, C., Otálora, F., Gavira, J.-A., Vidal, O., Giegé, R. & García-Ruiz, J. M. (2001), 'Structure of tetragonal hen egg-white lysozyme at 0.94 Å from crystals grown by the counter-diffusion method', *Acta Cryst.* **D57**, 1119–1126.
- Schiltz, M., Prangé, T. & Fourme, R. (1994), 'On the preparation and X-ray data collection of isomorphous xenon derivatives', *J. Appl. Cryst.* **27**, 950–960.
- Seaton, C. C. & Tremayne, M. (2002), 'Differential evolution: crystal structure determination of a triclinic polymorph of adipamide from powder diffraction data', *Chem. Commun.* pp. 880–881.
- Shankland, K., David, W. I. F. & Csoka, T. (1997), 'Crystal determination from powder diffraction data by the application of a genetic algorithm', *Z. Kristallogr.* **212**(8), 550–552.

- Shankland, K., David, W. I. F., Csoka, T. & McBride, L. (1998), 'Structure solution of Ibuprofen from powder diffraction data by the application of a genetic algorithm combined with prior conformational analysis', *Int. J. Pharm.* **165**, 117–126.
- Shankland, K., David, W. I. F. & Sivia, D. S. (1997), 'Routine ab initio structure determination of chlorothiazide by X-ray powder diffraction using optimised data collection and analysis strategies', *J. Mater. Chem.* **7**, 569–572.
- Shankland, K., McBride, L., David, W. I. F., Shankland, N. & Steele, G. (2002), 'Molecular, crystallographic and algorithmic factors in structure determination from powder diffraction data by simulated annealing', *J. Appl. Cryst.* **35**, 443–454.
- Sheldrick, G. M. (2010), 'Experimental phasing with *SHELXC/D/E*: combining chain tracing with density modification', *Acta Cryst.* **D66**, 479–485.
- Shotton, D. M. & Watson, H. C. (1970), 'Three-dimensional Structure of Tosyl-elastase', *Nature* **225**(5235), 811–816.
- Sivia, D. S. (2000), 'The number of good reflections in a powder pattern', *J. Appl. Cryst.* **33**, 1295–1301.
- St. Clair, N. L., Wang, Y.-F. & Margolin, A. L. (2000), 'Cofactor-Bound Cross-Linked Enzyme Crystals (CLEC) of Alcohol Dehydrogenase', *Angew. Chem. Int. Ed.* **39**(2), 380–383.
- Strand, F. L. (2000), 'David and Goliath — the slingshot that started the neuropeptide revolution', *Eur. J. Pharmacol.* **405**(1-3), 3–12.
- Sumner, J. B. (1926), 'The isolation and crystallization of the enzyme urease. Preliminary paper', *J. Biol. Chem.* **69**(2), 435–441.
- Tedesco, E., Harris, K. D. M., Johnston, R. L., Turner, G. W., Raja, K. M. P. & Balaram, P. (2001), 'Ab initio structure determination of a peptide  $\beta$ -turn from powder X-ray diffraction data', *Chem. Commun.* pp. 1460–1461.
- Tedesco, E., Turner, G. W., Harris, K. D. M., Johnston, R. L. & Kariuki, B. M. (2000), 'Structure determination of an oligopeptide directly from powder diffraction data', *Angew. Chem. Int. Ed.* **39**(24), 4488–4491.

- Tremayne, M., Seaton, C. C. & Glidewell, C. (2002), 'Structures of three substituted arenesulfonamides from X-ray powder diffraction data using the differential evolution technique', *Acta Cryst.* **B58**, 823–834.
- Vilenchik, L. Z., Griffith, J. P., St. Clair, N. L., Navia, M. A. & Margolin, A. L. (1998), 'Protein Crystals as Novel Microporous Materials', *J. Am. Chem. Soc.* **120**, 4290–4294.
- Von Dreele, R. B. (1999), 'Combined Rietveld and stereochemical restraint refinement of a protein crystal structure', *J. Appl. Cryst.* **32**, 1084–1089.
- Von Dreele, R. B. (2000), 'The first protein crystal structure determined from high-resolution X-ray powder diffraction data: a variant of T<sub>3</sub>R<sub>3</sub> human insulin-zinc complex produced by grinding', *Acta Cryst.* **D56**, 1549–1553.
- Von Dreele, R. B. (2001), 'Binding of N-acetylglucosamine to chicken egg lysozyme: a powder diffraction study', *Acta Cryst.* **D57**, 1836–1842.
- Von Dreele, R. B. (2003), 'Protein crystal structure analysis from high-resolution X-ray powder-diffraction data', *Methods Enzymol.* **368**, 254–267.
- Von Dreele, R. B. (2005), 'Binding of N-acetylglucosamine oligosaccharides to hen egg-white lysozyme: a powder diffraction study', *Acta Cryst.* **D61**, 22–32.
- Von Dreele, R. B. (2006), 'A rapidly filled capillary mount for both dry powder and polycrystalline slurry samples', *J. Appl. Cryst.* **39**, 124–126.
- Von Dreele, R. B. (2007), 'Multipattern Rietveld refinement of protein powder data: an approach to higher resolution', *J. Appl. Cryst.* **40**, 133–143.
- von Laue, M. (1913), 'Röntgenstrahlinterferenzen', *Phys. Z.* **14**, 1075–1079.
- Vuolanto, A., Kiviharju, K., Nevanen, T., Leisola, M. & Jokela, J. (2003), 'Development of Cross-Linked Antibody Fab Fragment Crystals for Enantioselective Separation of a Drug Enantiomer', *Crystal Growth Des.* **3**(5), 777–782.
- Wang, B. C. (1985), 'Resolution of Phase Ambiguity in Macromolecular Crystallography', *Methods Enzymol.* **115**, 90–112.
- Watson, H. C., Shotton, D. M., Cox, J. M. & Muirhead, H. (1970), 'Three-dimensional fourier synthesis of tosyl-elastase at 3.5 Å resolution', *Nature* **225**(5235), 806–811.

- Watson, J. D. & Crick, F. H. C. (1953), 'A structure for deoxyribose nucleic acid', *Nature* **171**(4356), 737–738.
- Wright, J. P. (2004), 'Extraction and use of correlated integrated intensities with powder diffraction data', *Z. Kristallogr.* **219**, 791–802.
- Wright, J. P., Besnard, C., Margiolaki, I., Basso, S., Camus, F., Fitch, A. N., Fox, G. C., Pattison, P. & Schiltz, M. (2008), 'Molecular envelopes derived from protein powder diffraction data', *J. Appl. Cryst.* **41**, 329–339.
- Wright, J. P. & Forsyth, J. B. (2000), *PRODD, Profile Refinement of Diffraction Data using the Cambridge Crystallographic Subroutine Library*, Rutherford Appleton Laboratory Report RAL-TR-2000-012, Didcot, Oxon, UK.
- Würtele, M., Hahn, M., Hilpert, K. & Höhne, W. (2000), 'Atomic resolution structure of native porcine pancreatic elastase at 1.1 Å', *Acta Cryst.* **D56**, 520–523.
- Wyckoff, R. W. G. & Corey, R. B. (1936), 'X-ray diffraction patterns of crystalline tobacco mosaic proteins', *J. Biol. Chem.* **116**, 51–55.
- Young, A. C. M., Dewan, J. C., Nave, C. & Tilton, R. F. (1993), 'Comparison of radiation-induced decay and structure refinement from X-ray data collected from lysozyme crystals at low and ambient temperatures', *J. Appl. Cryst.* **26**, 309–319.
- Zelinski, T. & Waldmann, H. (1997), 'Cross-Linked Enzyme Crystals (CLECs): Efficient and Stable Biocatalysts for Preparative Organic Chemistry', *Angew. Chem. Int. Ed.* **32**(7), 722–724.
- Zhang, K. Y. J. & Main, P. (1990), 'Histogram matching as a new density modification technique for phase refinement and extension of protein molecules', *Acta Cryst.* **A46**, 41–46.
- Zhou, Z., Siegler, V., Cheung, E. Y., Habershon, S., Harris, K. D. M. & Johnston, R. L. (2007), 'Advantages of a redefinition of variable-space in direct-space structure solution from powder X-ray diffraction data', *Chem. Phys. Chem.* **8**(5), 650–653.



# Acknowledgements

First and foremost, I would like to thank Prof. Marc Schiltz for having given me the opportunity to complete a PhD thesis in his research group. I am grateful for his guidance, support and for the freedom he granted me throughout my work. I also greatly appreciated the time and patience he dedicated, as well as the extremely didactic approach he adopted, when explaining concepts unknown to me.

I am indebted to many colleagues within our laboratory for their collaboration and illuminating discussions. I would like to especially thank Dr. Celine Besnard for her collaborative supervision in all aspects of my thesis work and, this, right from day one. Furthermore, I was fortunate enough to share my office with Francesco Gramiccia with whom I had countless discussions and to whom I am grateful for having essentially fulfilled the role of personal tutor in Maths and French grammar as well as for simply being there every step of the way. I hereby thank Dr. Kurt Schenk for having his door constantly open to passionately discuss any question, regardless of its nature. Moreover, I would like to mention Dr. Philip Pattison for his technical help in both obtaining and interpreting part of the data presented in this manuscript.

I would like to acknowledge the financial support from the Swiss National Science Foundation through grant Nbr. 200021-107637/1 and 200021-113339/1. Additionally, I hereby thank the European Synchrotron Radiation Facility and the Swiss-Norwegian Beamline Consortium for provision and allocation of beam time in the framework of the Long-Term Proposal Nbr. CH-1985.

The dependable administrative support of both Francine Bordelais and Anne-Lene Odegaard are gratefully acknowledged, and this in particular in the organization of my private defense. Japhet Bagilishya was in turn instrumental in the smooth proceeding of my thesis work with efficient and amicable IT support.

Many thanks are likewise due to the three examiners of the jury – Prof. Andrew N. Fitch, Dr. Radovan Černý and Prof. Petr Leiman – and to the jury president – Dr. Sandrine Gerber – for taking the time to review this manuscript as well as for attending

the private defense.

Finally, I hereby thank my parents, who provided me with the tools to succeed in this latest endeavor of mine, and Samantha David for her unconditional support and monumental patience on a daily basis and over the course of the past four years.

Thank you all.

# CURRICULUM VITÆ

Sebastian Basso

Born 30 March 1984

seb.basso@gmail.com

## Education

2007 - 2011	<b>Ph.D.</b> – <i>Development of X-ray Powder Diffraction Methods for Biomolecules</i>	Lausanne, Switzerland
	Ecole Polytechnique Fédérale de Lausanne	
2002 - 2006	<b>Master of Chemistry</b>	Bath, United Kingdom
	University of Bath – Graduated with 1 <sup>st</sup> class honors	
1999 - 2002	<b>International Baccalaureate – Bilingual Diploma</b>	New York, United States
	United Nations International School	

## Languages

French - Mother tongue

English - Fluent

Italian - Advanced level (B2 European standard)

German - Basic knowledge (A1 European standard)

## Publications

Basso, S., Fitch, A. N., Fox, G. C., Margiolaki, I. & Wright, J. P. (2005), *Acta Cryst.* **D61**, 1612–1625

Wright, J. P., Besnard, C., Margiolaki, I., Basso, S., Camus, F., Fitch, A., Fox, G. C., Pattison, P. & Schiltz, M. (2008), *J. Appl. Cryst.* **41**, 329–339

Basso, S., Besnard, C., Wright, J. P., Margiolaki, I., Fitch, A. N., Pattison, P. & Schiltz, M. (2010), *Acta Cryst.* **D66**, 756–761