UNIVERSITA' DEGLI STUDI DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

DIPARTIMENTO DI SCIENZE CHIMICHE

SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE MOLECOLARI
INDIRIZZO SCIENZE CHIMICHE
CICLO XX

Unconventional approaches to the solution of the crystallographic phase problem

DIRETTORE DELLA SCUOLA: CH.MO PROF. MAURIZIO CASARIN

SUPERVISORE: CH.MO PROF. GIUSEPPE ZANOTTI

DOTTORANDO: ANTON THUMIGER

31 GENNAIO 2008

Index

Summary

The phase problem represents one of the major challenges along the way of structure solution by x-ray crystallography. This is especially true when the object under study are macromolecular crystals, which have many atoms in the unit cell and often diffract poorly, so not allowing to measure the higher resolution shells of the diffraction pattern.

For these reasons, no general method exists nowadays to solve the macromolecular phase problem, but rather a variety of approaches, mostly experimental, and requiring an additional amount of work to be carried out in order to collect several different datasets from chemically modified crystals. *Ab initio* methods, that is, knowledge-independent methods working on a single dataset, are rarely used, because of the lack of high-resolution data. However, some arguments support the thesis that *ab initio* phasing should be possible even in the absence of atomic resolution data or specific additional information. In fact, some general constraints on the resulting electron density distribution exist that can in principle lead to an overdetermined problem; finding an effective way to impose such constraints would result in a general phasing method able to solve the majority of macromolecular structures starting from a single, medium resolution dataset.

The work described in this thesis was addressed to experiment alternative methods for macromolecular *ab initio* solution, and can be divided in three sections. First, neural networks were investigated as potential tools for encoding unknown relationships between phases and magnitudes; then the power of some real-space constraints was studied, jointly with iterative algorithms to impose them; and finally, an analysis of Patterson maps was carried out in the hope of identifying autocorrelation features that could be related to the presence and orientation into the unit cell of known secondary structure elements, like α-helices and β-strands and sheets. This last topic, while not aiming at direct solution of the phase problem, but rather at extracting some raw structural information from the measured data, is among the three different approaches the one that gave the most promising results.

The neural network section (chapter 2) describes one-dimensional tests that were carried out in order to assess the network ability in learning unknown relationships between diffracted magnitudes and the corresponding phases. A simple, atomic case was chosen to evaluate the network behaviour in conditions that are known to be favourable. The result of this investigation is mainly that neural networks are not the right tool for phasing, at least

not with the approach described here.

The work about iterative methods (described in chapter 3) has been motivated by some recent results (Lunin *et al*., 2002) showing that macromolecular phasing at low resolution can be accomplished if a binary mask, instead of a continuously valued electron density, is searched for. Some attempts are described here in building iterative phasing algorithms that impose the binary constraint on the electron density; the implemented methods have shown to work in some simple binary 2D cases, but it is doubtful that they can be applied to find binary approximations to continuous densities. A subsection of this work has been conducted on modifications of an existing algorithm in order to accommodate topological restraints, with interesting but not conclusive results about phase extension.

The final part of this thesis (chapter 4) outlines a new approach to Patterson map analysis, aiming at elucidating its connection with the secondary structure content of the unit cell. It is shown that, in favourable cases, information on the presence and orientation of α-helices and β-sheets can be easily extracted from the Patterson map. There is no need for high resolution data since the concept of atom is not used in the derivation of the method. The approach needs further refinement to be turned into a reliable tool for macromolecular crystallography; in perspective, it could provide phase estimates, to be used as a starting point for and extension and refinement procedure.

At the end of this thesis, a short experimental work on TpF-1 protein from the pathogenic bacterium *Treponema Pallidum* has been reported. The structure of this immunogenic protein was solved at the beginning of the Ph.D. Project.

Sommario

Il problema della fase costituisce uno dei maggiori ostacoli nel processo di determinazione strutturale per via cristallografica. Ciò è particolarmente vero nel caso dei cristalli macromolecolari, a causa dell'elevato numero di atomi nella cella e del disordine intrinseco, che limita la risoluzione massima delle intensità misurabili.

Per questi motivi, non esiste al momento un metodo generale per risolvere il problema della fase in campo macromolecolare, bensì una grande varietà di tecniche (per lo più sperimentali), che richiedono una mole di lavoro supplementare al fine di misurare diversi set di dati da cristalli opportunamente trattati. Per via della limitata risoluzione a cui i dati vengono normalmente raccolti, i metodi *ab initio*, che non necessitano informazioni aggiuntive e sono in grado di ricostruire la struttura da un singolo set di dati, possono venire utilizzati solo raramente. D'altra parte, diverse argomentazioni fanno supporre che la risoluzione strutturale *ab initi*o di macromolecole dovrebbe essere possibile anche in assenza di dati a risoluzione atomica e di informazioni *a priori* sulla struttura. Infatti, vi sono vincoli di natura generale in grado di rendere il problema sovradeterminato; se si trovasse un modo efficiente di imporre tali vincoli la maggior parte delle macromolecole potrebbe essere risolta da un singolo set di dati di diffrazione a media risoluzione (1.5-3 Å).

Il lavoro di ricerca esposto nella presente tesi era volto a sperimentare metodi alternativi per la risoluzione *ab initio* di macromolecole, e si articola in tre parti. In un primo momento, si è indagato sulle potenzialità delle reti neurali nell'apprendere relazioni esistenti tra fasi e moduli diffratti; in seguito, sono stati studiati algoritmi iterativi in grado di imporre specifici vincoli sulla densità; infine, si è tentato di stabilire quale relazione sussista tra la mappa di Patterson e gli elementi di struttura secondaria presenti nella cella elementare (α-eliche, β-strands e β-sheets), con lo scopo di individuare un modo per desumerne la presenza e l'orientazione. In quest'ultimo ambito, che non mira alla risoluzione del problema della fase bensì ad ottenere direttamente dai dati sperimentali alcune informazioni strutturali di base, si sono ottenuti i risultati più interessanti.

Nella parte relativa alle reti neurali (capitolo 2) vengono descritti esperimenti basati su dati 1-D artificiali, ideati allo scopo di verificare la capacità di apprendimento di una rete neurale riguardo a relazioni non-lineari esistenti tra moduli e fasi. Per valutare il comportamento della rete, si è scelto un caso semplice e dotato di atomicità, per il quale è effettivamente

possibile ricavare relazioni probabilistiche. Questo studio ha portato a concludere che le reti neurali non sono in grado di svolgere questo compito; ciò non esclude che non possano essere sfruttate in altro modo all'interno di una procedura di *phase retrieval*.

La parte di lavoro relativa ai metodi iterativi (descritta nel capitolo 3) è stata ispirata da studi condotti in questi ultimi anni (Lunin *et al*., 2002), nel corso dei quali si è dimostrato che il problema della fase per cristalli macromolecolari può essere risolto a bassa risoluzione se si approssima la densità elettronica (dotata di una distribuzione continua di valori) a una funzione binaria. In una prima fase del lavoro sono stati studiati algoritmi iterativi in grado di imporre il vincolo binario, nel tentativo di trovare un metodo più rapido ed efficace di quello originale. Gli algoritmi implementati si sono dimostrati in grado di ricostruire semplici densità binarie bidimensionali, ma la loro applicazione per approssimare densità continue è risultata difficile. In una seconda fase sono state introdotte modifiche ad algoritmi esistenti in modo da imporre vincoli topologici; questo ha portato a risultati interessanti, ma non conclusivi, per quanto riguarda l'estensione a partire da fasi esistenti.

L'ultima parte della tesi (capitolo 4) descrive un nuovo metodo di analisi della mappa di Patterson, che è stato sviluppato nel tentativo di individuare le relazioni tra questa mappa e il tipo di struttura secondaria presente nella cella elementare. In casi favorevoli è possibile estrarre facilmente informazioni sulla presenza e l'orientazione di α-eliche e β-sheets; il metodo non richiede l'uso di dati a risoluzione atomica perché si basa su caratteristiche a media risoluzione della mappa di Patterson. Questo tipo di approccio necessita di essere ulteriormente affinato in vista di applicazioni reali nel campo della cristallografia di proteine; in prospettiva, esso potrebbe fornire stime iniziali delle fasi a partire dalle intensità diffratte, utilizzabili come punto di partenza per un processo di estensione.

In calce alla tesi è stato riportato un articolo relativo alla determinazione strutturale della proteina immunogenica TpF-1, codificata nel genoma del batterio patogeno *Treponema Pallidum*. Questo lavoro sperimentale è stato portato a termine durante il primo anno di dottorato.

Chapter 1

The Phase Problem

Introduction

The aim of protein crystallography is to reconstruct, from a diffraction spectrum, a molecular electron density distribution, which can be interpreted in terms of an atomic model. Such a model can provide insight into the protein structural properties, mechanisms of enzymatic catalysis, protein-substrate or protein-protein interactions and offers a rational approach to drug design.

Key steps in the process of structural determination via x-ray crystallography are protein production or extraction from living cells or tissues, protein purification, crystallization, diffraction data collection, the solution of the phase problem, and model building and refinement. In many stages of this pipeline intrinsic difficulties can arise, making the success of a project rather unpredictable.

Solving the phase problem is one of the bottlenecks of the process. It consists in reconstructing the electron density distribution from diffraction data. These latter are in principle complex quantities, describing the electric field of diffracted waves, and furnishing an alternative, equivalent description of the crystal. The diffracted waves contain in fact the full spatial information about the unit cell content; experimentally, however, it is not possible to measure the phase differences of the diffracted beams, which can be known only in absolute value (real quantities). This substantially incomplete knowledge of the diffraction image causes an infinite number of possible object to be compatible with the measured data.

Different techniques are available for solving the macromolecular phase problem, most of them being experimental ones (Zanotti, 2002). They rely on some kind of chemical modification of the crystal, like introducing heavy (metal) atoms or anomalous scatterers (typically replacing the sulfur in methionines with selenium). This allows to measure an independent set of diffracted intensities, which can be used to solve a system of equations involving phase values. These techniques are nowadays widely used, although being time-consuming and not guaranteed to succeed. Often, they represent the only way to solve structures lacking relevant sequence similarity with other proteins of known structure. When a significant similarity with previously solved structures is found, at least in some domain, a frequent option is represented by *molecular replacement* (*MR*) methods, in which one tries to optimally reproduce the observed data by orienting and translating a known fragment in the unit cell (Rossmann and Blow, 1962).

Both experimental phasing methods and *MR* involve the introduction of some kind of independent knowledge specific to the particular case to be solved.

Instead, by the name of '*ab initio* methods' are commonly indicated phasing strategies that do not involve any prior knowledge about specific structural features. These methods rely instead on general properties of the density to be reconstructed, and require only a single set of measured intensities as input data. When they can be applied, such techniques are likely to represent the best option; in fact, they find wide application in the field of small molecule crystallography. Unfortunately, and despite their constant improvement, they are still seldom useful for proteins, mainly because of the low data to parameters ratio of the problem.

This latter is function of the resolution limit d_{min}, or minimum distance between adjacent crystal planes giving rise to a measurable diffracted intensity. A small value for d (high resolution) corresponds to a wide angle between incoming and scattered rays, and is related to fine details of the electron density (such as atomic peaks). Conversely, reflections with high values of d (low resolution) carry the information about more global parameters of the image (like molecule position and general shape).

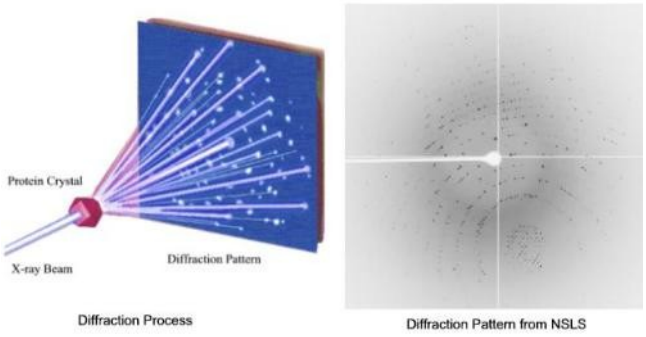


Fig. 1 Schematic representation of the diffraction process from a crystal. The incoming x-ray beam interacts with the sample and diffracted intensities are recorded on the detector plane as isolated spots.

The nature of the phase problem

The *phase problem* is a very important topic in many fields of optics, affecting all those techniques known as *diffractive* (or *lensless*) *imaging*. In all these methods, experiments can be set to measure the diffraction pattern arising from some object under study, with the aim of reconstructing the distribution of scattering matter in the object. So diffractive imaging exploits the wave properties of radiation, differently from refractive techniques which rely over the laws of geometric optics and appear to be *diffraction-limited*. In optical microscopy, for example, diffraction is an unavoidable phenomenon that sets a limit for resolution: it is not possible to image an object which is smaller than the wavelenght of the illuminating radiation. The resolution is given in fact by the expression $d = \lambda / 2 A_N$, where the *numerical aperture* A_N is always lower than 1 (for lenses operating in air). In lensless imaging, conversely, it is not possible to obtain images at a level of detail much smaller than the wavelenght of the radiation used. This sets the useful wavelenght range for crystallographic studies to the hard x-ray region and precisely around 1 Å (0.1 nm), which is the order of magnitude of atomic bonding distances. Another difference is that in diffractive imaging the quality of the data is strongly correlated to the coherence of the light beam, since the measured quantities arise from interference phenomena. Temporal and spatial incoherence degrade the quality of the diffraction data, limiting the resolution of the reconstructed images and in extreme cases preventing image reconstruction (Thibault, 2007).

Restricting the analysis to the far-field (Fraunhofer) diffraction regime, the diffracted wave $f(\boldsymbol{s})$ is given by the Fourier transform ($\mathcal{FT}$) of the object density $\rho(\boldsymbol{x})$, so that a complete knowledge of the former would allow the straighforward calculation of $\rho(\boldsymbol{x})$ by means of an inverse transform:

$$f(\boldsymbol{s}) = \mathcal{FT}[\rho(\boldsymbol{x})], \quad \rho(\boldsymbol{x}) = \mathcal{FT}^{-1}[f(\boldsymbol{s})]. \tag{1.1}$$

The space where $f(\boldsymbol{s})$ is defined is called *frequency space* [1], since the function *f* describes the object density in terms of its spatial frequencies.

The quantity $f(\boldsymbol{s}) = |f(\boldsymbol{s})| e^{i\phi(\boldsymbol{s})}$ is in general a complex function describing the electric field of the diffracted wave in amplitude $|f(\boldsymbol{s})|$ and phase $\phi(\boldsymbol{s})$; because of the difficulty of

1 It is also called *Fourier space* or *reciprocal space* (this last name is of common use in crystallography).

realizing interference experiments with a reference beam, often the relative phases cannot be measured, and the only quantities available from most diffraction experiments are the intensities $I=|f(\boldsymbol{s})|^2$ (that is, only half of the information needed for object reconstruction). It turns out that some *a priori* knowledge about the object is needed in order to make the problem overdetermined. Stated in mathematical terms, solving the phase problem consists in retrieving the unknown phases of a complex function $f(\boldsymbol{s})$, given its modulus $|f(\boldsymbol{s})|$ and a sufficient amount of independent information on the nature of its inverse Fourier transform $\rho(\boldsymbol{x})$. In position space (also called *direct* or *real* space), where $\rho(\boldsymbol{x})$ is defined, the solution of an ideal phase problem is represented by the intersection of two subsets: the first one is the subset of all the possible densities which are consistent with the observed moduli, the other is the subset of all the densities satisfying the *a priori* constraints. An analogous representation can be given in the *phase space* (the subspace of frequency domain spanned by phase values), since for fixed moduli there is a bijective correspondence between the functions $\rho(\boldsymbol{x})$ and $\phi(\boldsymbol{s})$. It should be observed that the intersection between the two subsets is not generally a single point, because the choice for the origin and the handedness of the axes (*enantiomorph*) in real space is arbitrary and not constrained in any way by the diffracted intensities. This means that the equivalence

$$\rho(\boldsymbol{x}) \sim \rho(\sigma \boldsymbol{x}+\boldsymbol{t}) \tag{1.2}$$

holds, where $\sigma=\pm 1$ gives the axes handedness and $\boldsymbol{t}$ is an arbitrary translation. Consequently, in Fourier space one has

$$\phi(\boldsymbol{s}) \sim \sigma[\phi(\boldsymbol{s})-2\pi \boldsymbol{s}\cdot\boldsymbol{t}]\,. \tag{1.3}$$

In the most general case, when the object is specified by a complex, non-periodic $\rho(\boldsymbol{x})$, the corresponding Fourier transform is non-centrosymmetric and continuous. In three dimensions, it can be written as

$$F(\boldsymbol{s}) = \int \rho(\boldsymbol{x}) e^{2\pi i(\boldsymbol{s}\cdot\boldsymbol{x})} d\boldsymbol{x}\,. \tag{1.4}$$

Moreover, assuming a given sampling in both spaces, one has:

$$F(h_r,k_s,l_t)=\sum_{u=1}^{N(x)}\sum_{v=1}^{N(y)}\sum_{w=1}^{N(z)}\rho(x_u,y_v,z_w)\ \exp[\,2\pi i(h_r x_u+k_s y_v+l_t z_w)] \\ =\sum_{u=1}^{N(x)}\sum_{v=1}^{N(y)}\sum_{w=1}^{N(z)}\rho(x_u,y_v,z_w)\ q_h^x q_k^y q_l^z . \tag{1.5}$$

If F can be factorized, then the solution of the problem is not unique. In fact, $F=F_1F_2\ldots F_n$ and any function like $H=F_1F_2^*\ldots F_n$, where conjugation is applied to one or more factors, have the same squared modulus. This is always the case for one-dimensional functions, because the fundamental theorem of algebra states that any polynomial in a single variable can be written as a product of first-order terms. However, almost all cases of practical interest occur in 2- or 3-dimensional spaces, and since almost all polynomials in two or more variables are irreducible, the uniqueness of the solution is guaranteed in practice (Millane, 1990).

Given that there is no intrinsic degeneracy of solutions, it is necessary to set the problem so that is is overdetermined, by specifying the minimum amount of *a priori* information needed to make the ratio equations/unknowns favourable. In *d* dimensions, once introduced a sampling in direct and reciprocal space, we have

$$F(\boldsymbol{s}_j)=\sum_{i=0}^{N-1}\rho(\boldsymbol{x_i})\exp(2\pi i\,\boldsymbol{s}_j\cdot\boldsymbol{x_i})\quad,\qquad N=\prod_{i=1}^{d}n_i\quad,\qquad j=1,-,N \tag{1.6}$$

where N is the total number of pixels and n_i is the number of sampling intervals (pixels) along the i-th dimension. Since F is known only in modulus, we need to solve a system of N non linear equations:

$$|F(\boldsymbol{s}_j)|=\left|\sum_{i=0}^{N-1}\rho(\boldsymbol{x_i})\exp(2\pi i\,\boldsymbol{s}_j\cdot\boldsymbol{x_i})\right| \tag{1.7}$$

If the density is complex valued, then the number of unknowns will be *2N*, since for each $\boldsymbol{x}_i$ one needs to know the values of the real and the imaginary part of $\rho(\boldsymbol{x_i})$. The problem is doubly underdetermined since we dispose of N equations only. The same

equations/unknowns ratio is valid for a real valued density; in that case the number of unknowns would drop to *N*, but since the Fourier transform of a real function has to be even ($F(\boldsymbol{s})=F(-\boldsymbol{s})$), the number of equations would also reduce to *N/2*. The conclusion is that the problem is always underdetermined by a factor of 2, regardless of the dimensionality *d* and of the nature of $\rho(\boldsymbol{x})$ (Miao, 1998).

To make the problem solvable, it is necessary to increase the number of equations, by introducing some knowledge about the values of the function $\rho(\boldsymbol{x})$. It is clear that knowing a priori the ρ values for *M* pixels provides *M* new equations, so that for *M>N/2* the problem becomes overdetermined. An alternative, once the size of the object is known[2], is to choose a fine enough sampling in reciprocal space; this corresponds in direct space to surrounding the object domain with a zero-valued jacket. This method (known as *oversampling*) finds wide application in optics, where the unknown ρ is non periodic and the quantity $F(\boldsymbol{s})$, being continuous, can be sampled as finely as needed. So, any non-periodic object can be reconstructed from its diffraction pattern, given that the latter is known at a sufficient level of detail. In these last years, this consideration has brought interesting developments in the field of lensless imaging, where many successful experiments have been carried out. Among the reconstructed objects with synchrotron x-ray radiation are living cells (Shapiro *et al.*, 2005); a similar technique based on electron diffraction allowed imaging of single carbon nanotubes down to a resolution of ~1 Å (Zuo *et al.*, 2003). With the most powerful existing x-ray sources, the free electron lasers (FELs), it has been shown that images can be obtained down to a resolution of some tenths of nanometers, by recording a femtosecond diffraction pattern from the object before that this latter is turned into a plasma (Chapman *et al.*, 2006). The birth of this new field in x-ray science has led to the development of very efficient phasing algorithms, capable of working in absence of any knowledge about the object under study (Marchesini *et al.*, 2003; Wu *et al*., 2004; He, 2006; Marchesini, 2007). A difficult task remains that of bringing the FELs to work at atomic (~1 Å) wavelenghts; this technological achievement would in principle allow imaging of single molecules, and some theoretical studies have already been carried out to address some of the most challenging aspects of the project, such as the creation of a molecular beam of proteins (Wu and Spence, 2005; Spence *et al*., 2005), the control over molecular orientation, the lifetime of the strongly irradiated sample and the reconstruction

2 The size can be estimated from the autocorrelation function $A(\boldsymbol{u})=\mathcal{FT}^{-1}[|f(\boldsymbol{s})|^2]$

of a three-dimensional image from multiple dataset recorded from many copies of the molecule of interest (Miao *et al.*, 2001). The single-molecule thechnique would overcome one of the most difficult and unpredictable steps in protein crystallography, that is, crystallization; this in turn would allow structural investigation of very unsoluble species, like membrane proteins (which constitute less than 0.1% of the coordinates files deposited in the PDB[3] database).

The crystallographic phase problem

While for non periodic objects the phase problem is in principle always solvable, a very different situation occurs in crystallography. Here the object has a periodicity given by the crystal lattice, whose axes $\boldsymbol{a}$,$\boldsymbol{b}$,$\boldsymbol{c}$ coincide in lenght and direction with the edges of the unit cell. This latter is the fundamental 'building block' of the crystal, which can be seen as the collection of identical copies of the cell translated by lattice vectors $u\,\boldsymbol{a}+v\,\boldsymbol{b}+w\,\boldsymbol{c}$. The repetition of unit cells in the crystal has the effect of superposing the diffraction fringes of the crystal lattice to the diffraction pattern of the single unit cell, giving rise to a 'natural sampling' that diffraction intensities to be non zero only for angles satisfying Bragg's law (Bragg, 1913):

$$n\lambda = 2d\sin\theta \tag{1.8}$$

where d is the lattice spacing between two adjacent crystallographic planes, θ is the angle of incidence of the incoming beam on that family of planes, λ is the x-ray wavelenght and n is the order of diffraction. Since the same crystallographic planes give rise to many diffracted beams of different order, the Bragg formula can be simplified to $\lambda = 2d_{\boldsymbol{h}}\sin\theta$ assuming that the higher orders of diffraction (n>1) are due to imaginary planes of spacing $d_{\boldsymbol{h}} = d/n$. Here $\boldsymbol{h} \equiv (h,k,l)$ is a triple of Miller indexes specifying a family of crystallographic planes that intercepts the three points $\boldsymbol{a}/h, \boldsymbol{b}/k, \boldsymbol{c}/l$, or some multiple thereof. Diffracted spots (also somewhat improperly called *reflections*) can be recorded by a single-beam detector or by a two-dimensional one. Usually, the smallest distance d_{min} for which measurable intensities arise (and related to the widest angle of diffraction θ_{max} by $d_{min} = \lambda/2\sin\theta_{max}$) is called

3 The PDB database is accessible on web at the address: http://www.rcsb.org/pdb/home/home.do

resolution of an experimental dataset. The name is due to the level of detail expected in the reconstructed electron density: a dataset with *high resolution* is one comprising intensities recorded at wide angles, that is, with small d_{min} values.

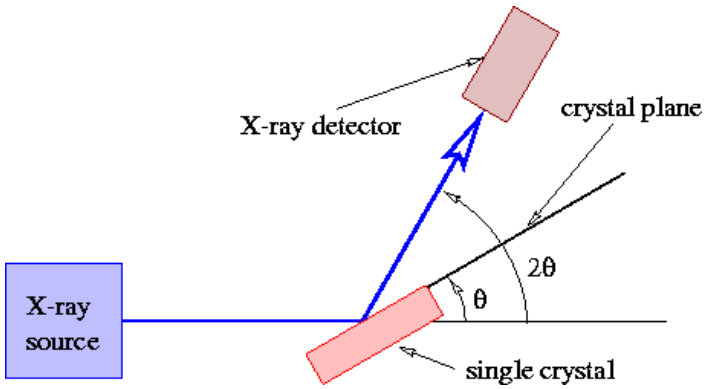


Fig. 2 Diffraction geometry: the incoming x-rays interact with the crystal and give rise to a diffracted beam at angle 2θ, which can be thought as being 'reflected' by a crystallographic plane. This has nothing to do with true reflection, however, since it only occurs for the angles satisfying the Bragg law, that is, $\sin\theta = n\lambda/(2d)$

This means that the diffraction pattern $F(\boldsymbol{S})$ is non-zero only for those values of $\boldsymbol{S}$ which satisfy the *Laue equations* (Laue, 1912):

$$\begin{cases} \boldsymbol{a}\cdot\boldsymbol{S} = h \\ \boldsymbol{b}\cdot\boldsymbol{S} = k \\ \boldsymbol{c}\cdot\boldsymbol{S} = l \end{cases} \tag{1.9}$$

where the *scattering vector* $\boldsymbol{S}$ is given by $\boldsymbol{S} = (\boldsymbol{s} - \boldsymbol{s_0})/\lambda$ ($\boldsymbol{s_0}$ and $\boldsymbol{s}$ are unit vectors specifying the direction of incoming and diffracted x-rays). This can be put in a more condensed form by observing that the values of $\boldsymbol{S}$ for which a diffracted intensity can be observed lie at the nodes of an imaginary lattice $\boldsymbol{S} = h\boldsymbol{a}^* + k\boldsymbol{b}^* + l\boldsymbol{c}^*$. This abstract object is called *reciprocal lattice* and its axes $\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*$ are defined by the nine scalar products $\boldsymbol{a}_i \cdot \boldsymbol{a}_j^* = \delta_{ij}$[4]. The diffracted wave observed at the nodes of the reciprocal lattice is proportional to the unit cell Fourier transform or *structure factor*, $F(\boldsymbol{h})$:

4 The generic vector $\boldsymbol{a}_i$ can take the values $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$.

$$F(\boldsymbol{h}) = \int_V \rho(\boldsymbol{r}) e^{2\pi i \boldsymbol{h}\cdot\boldsymbol{r}} dV \,. \tag{1.10}$$

The value of $F(\boldsymbol{h})$ in any point in between reciprocal nodes (that is, for fractional **h** indices) cannot be known. Early observations about the connection between phase problem and the unknown diffracted intensities between reciprocal lattice nodes were made by Sayre (Sayre, 1952b) on the basis of the Shannon sampling theorem (Shannon, 1949). The nature of crystal diffraction patterns is consistent with Fourier transform theory, which states that a periodic function must have a discrete spectrum[5]; in fact, the inverse relationship writes

$$\rho(\boldsymbol{r}) = \frac{1}{V}\sum_{\boldsymbol{h}} |F(\boldsymbol{h})| e^{i(\phi(\boldsymbol{h}) - 2\pi \boldsymbol{h}\cdot\boldsymbol{r})} \tag{1.11}$$

The physical meaning in position space is that the unit cell cannot be surrounded by a zero contour by simply enlarging the bounds of the real domain, since other identical unit cells are encountered. An obvious conclusion is that, when diffraction occurs from a periodic object, oversampling is no longer feasible, and the phase problem is always underdetermined by a factor of 2. Additional *a priori* constraints are then needed to identify the correct set of phases.

Constraints in direct and reciprocal space

The most powerful constraint in crystallography is represented by the *atomicity* property; that is, the solution of the phase problem must correspond to an electron density made of well resolved atomic peaks, emerging from a sea of near-zero values. Two general approaches have been followed to put this constraint in action, the *real space* and the *reciprocal space* approach. Both classes of methods have co-existed since their birth in the 1930s: real space methods came in usage first, once it was clear that the vectorial properties of the Patterson function[6] could be used to unravel simple structures. First observations were made by Patterson (Patterson, 1934), who proposed a method to restrain the number of

5 The Fourier transform reduces then to a *Fourier series*.

6 In crystallography, the Patterson function is an aliased autocorrelation function of the unit cell, that can be obtained by Fourier transforming the experimental intensities with zero phases.

possible atomic coordinates based on the Patterson function. In the subsequent years, vector approaches exploiting Patterson map superposition were developed and applied. Starting in the early 1950s, the development of *direct methods*, working entirely in reciprocal space, ended by establishing a new standard in *ab initio* crystallographic phasing. During the last 5 decades, direct methods have grown in power (fig. 3), allowing solution of structures with hundreds of atoms in the asymmetric unit, and also of small macromolecules. Patterson-based approaches continued to be improved but could not reach such a broad use. However, recent developments have shown that Patterson deconvolution methods can outperform direct methods, allowing to solve macromolecular structures with as much as 2000 non-H atoms in the asymmetric unit, provided that data with a resolution higher than 1.5 Å are available.

Direct methods: theory

Despite of a long period of skepticism, in the early 1950s it became clear that the crystallographic phase problem could be solved from measured intensities alone. The discovery can be ascribed to a change in the point of view, which had shifted from the unknown electron density to the atomic coordinates. It turned out that stating the problem in terms of atomic coordinates led to an overdetermined system of equations; so the solution was expected to be unique and only an appropriate method to find it was needed. In fact, since the electron density of the unit cell can be to a good approximation assumed equal to the sum of the individual atomic densities (that is, neglecting the deformation of the electron clouds caused by chemical bonding), the structure factor can be expressed as

$$F(\boldsymbol{h}) = \sum_{j=1}^{N} f_j(r^*)\, e^{2\pi i \boldsymbol{h}\cdot\boldsymbol{r}} \tag{1.12}$$

where the sum extends over all j atoms in the cell. The $f_j(r^*)$ are the *atomic scattering factors*, defined as Fourier transforms of the atomic electron densities, which are assumed to be spherically symmetric:

$$f_j(r^*)=\int_0^\infty 4\pi r^2 \rho(r)\frac{\sin(2\pi r r^*)}{2\pi r r^*}\,dr\,, \quad r^* = \frac{2\sin\theta}{\lambda}\,. \tag{1.13}$$

[mettere grafico f(r*) - didascalia: the decrease of fj with angle reflects in a similar behaviour for the total scattering intensity of a crystal]

Most of the direct methods theory has been developed from a description of the unit cell content in terms of point atoms (that is, Dirac δ(**r**) distributions of electron density) of constant scattering factor. The appropriate quantities in reciprocal space are the *normalized structure factors* $E(\boldsymbol{h})$, giving the Fourier transform for the point-atom structure. These can be obtained, at least approximately, from the observed structure factors, taking into account the symmetry, the unit cell content and the thermal motion:

$$E(\boldsymbol{h})=\frac{F(\boldsymbol{h})\exp[B(\sin^2\theta/\lambda^2)]}{\sqrt{\langle|F^{obs}(\boldsymbol{h})|^2\rangle_{r^*(\boldsymbol{h})}}} \tag{1.14}$$

The oscillation of the atoms about their equilibrium position has the effect of speeding up the angular decay of the diffracted wave, and can be accounted for in the scaling procedure introducing an average *temperature factor* (*B*).

The observed moduli are a function of the 3*N* atomic coordinates $\{x_j, y_j, z_j\}$, which represent the unknowns of the problem. Each modulus contributes with and independent equation

$$|E(\boldsymbol{h})| = \sigma^{-1/2}\left|\sum_{j=1}^{N} Z_j \exp(2\pi i\,\boldsymbol{h}\cdot\boldsymbol{r}_j)\right|\,. \tag{1.15}$$

When the data are measured up to atomic resolution, the number *M* of observed moduli largely exceeds 3*N*, so that the problem is overdetermined. A typical value of the observables/parameters ratio is *O/P* = ~ 8 at a resolution of 1 Å; this proves that with atomic resolution data the solution must be unique (the ratio *O/P* decreases to a value of ~ 3 at 1.4 Å, where is it assumed the breakdown of atomicity occurs). In direct methods, statistical relationships between phases and moduli are exploited to obtain the phases for strong reflections first. These relationships are derived without any assumption about the relative

coordinates of the peaks, which are assumed to be uncorrelated, as for atoms randomly distributed in the unit cell. This is obviously not true, since in any crystal structure the atoms are chemically bonded, but the correlations arising from chemical constraints are very difficult to introduce in the probabilistic framework of direct methods. Working entirely in reciprocal space, classical direct methods avoid to calculate many Fourier transforms and result to be quite fast; only in the final stage an electron density map is calculated and searched for atomic peaks. The probabilistic method for deriving moduli/phases statistical relationships (Hauptman and Karle, 1953) can be described as follows:

- fix the set $\boldsymbol{S}$ of reflections $E(\boldsymbol{h})$ to be used for phase determination;
- calculate the *characteristic function*

$$C(\theta_1, -, \theta_m, \rho_1, -, \rho_m)\,, \tag{1.16}$$

  where θ_i, ρ_i are carrying variables associated with the phases $\phi(\boldsymbol{h}_i)$ and the moduli $|E(\boldsymbol{h}_i)|$ ($\boldsymbol{h}_i \in \boldsymbol{S}$);
- calculate the *joint probability distribution*

$$P_J(\phi(\boldsymbol{h}_1), -, \phi(\boldsymbol{h}_m), |E|(\boldsymbol{h}_1), -, |E|(\boldsymbol{h}_m)) \tag{1.17}$$

  by Fourier transforming the characteristic function. From the joint distribution P_J the *conditional probabilities*

$$P_C(\phi(\boldsymbol{h}_i) | \{\phi, |E|\}) \tag{1.18}$$

  may be obtained. These provide the probability of the phase of indices $\boldsymbol{h}_i$ to take the value $\phi(\boldsymbol{h}_i)$, given the set of known phases and moduli $\{\phi, |E|\}$.

The most interesting quantities provided by this approach are of the form $P_C(\psi_n | \{|E|\})$, giving the probability distribution for a linear combination of phases

$$\psi_n = \sum_{i=1}^{n} \phi_i \tag{1.19}$$

once the set of known moduli $\{|E|\}$ has been fixed. In fact, since the phases depend on the choice of the origin, they are not uniquely defined; direct methods do not allow to make predictions about the values of individual phases, but rather on some linear combinations of them which are origin-independent. Those combinations are called *structure invariants* (s.i.) and take the general form

$$\psi(\boldsymbol{h}_1, -, \boldsymbol{h}_n) = \sum_{j=1}^{n} \phi(\boldsymbol{h}_j) \quad \textit{with} \quad \sum_{j=1}^{n} \boldsymbol{h}_j = 0\,. \tag{1.20}$$

A less general (but useful) kind of linear combination of phases are the *structure semi-invariants* (s.s.): these quantities are not independent from *any* origin shift, but they do not change when the origin moves between cell positions possessing the same point symmetry (*permissible origins*). Obviously the definition of s.s. is space-group dependent.

In general, a given structure invariant ψ_n will depend primarily on a small number of structure factor magnitudes |*E*|; the *neighborhood principle* (Hauptman, 1975) and the more general *representation theory* (Giacovazzo, 1977; Giacovazzo, 1980) allow to class the |*E*| magnitudes in order of their decreasing effectiveness for phase estimation. The vectors belonging to the set $\{\boldsymbol{h}_i\}_{i=1,-,n}$, whose phases appear in the expression of ψ_n, are called the *basis vectors* of ψ_n. If the crystal symmetry is higher than triclinic then one or more additional structure invariants of the form

$$\psi_n^{(k)} = \phi(\boldsymbol{h}_1 \boldsymbol{R}_s) + ... + \phi(\boldsymbol{h}_n \boldsymbol{R}_m) \tag{1.21}$$

where $\boldsymbol{R}_s, -, \boldsymbol{R}_m$ vary over the set of rotation matrices and the condition

$$\boldsymbol{h}_1 \boldsymbol{R}_s + ... + \boldsymbol{h}_n \boldsymbol{R}_m = \boldsymbol{0} \tag{1.22}$$

is satisfied. The *first representation* of ψ_n is defined as ψ_n itself plus the set of invariants $\{\psi_n^{(k)}\}$, all differing by a constant phase angle which depends on the symmetry operators

only. A set of cross vectors is defined by the linear combination of basis vectors

$$m_1 \boldsymbol{h}_1 \boldsymbol{R}_s + ... + m_n \boldsymbol{h}_n \boldsymbol{R}_m \quad (m_i = 0,1) \,. \tag{1.23}$$

The moduli corresponding to all the basis and cross vectors appearing in the first representation constitute the *first phasing shell* of ψ_n.

The most important s.i., extensively used in direct methods, are the *triplets*

$$\psi_{\boldsymbol{hk}} = \phi_{-\boldsymbol{h}} + \phi_{\boldsymbol{k}} + \phi_{\boldsymbol{h}-\boldsymbol{k}} \;; \tag{1.24}$$

triplet values are expected to follow the Cochran distribution (Cochran, 1955):

$$\begin{aligned} P(\psi_{\boldsymbol{hk}}) &= [2\pi I_0(G_{\boldsymbol{h},\boldsymbol{k}})]^{-1} \exp(G_{\boldsymbol{hk}} \cos\psi_{\boldsymbol{hk}}) \\ G_{\boldsymbol{hk}} &= 2\sigma_3 \sigma_2^{-3/2} |E_{\boldsymbol{h}} E_{\boldsymbol{k}} E_{\boldsymbol{h}-\boldsymbol{k}}| \end{aligned} \tag{1.25}$$

[7]

which constitutes a fundamental result of direct methods theory and can be derived applying the *central limit theorem* under the assumption of atoms randomly distributed in the cell with uniform probability. The expression for $P(\psi_{\boldsymbol{hk}})$ is an example of *von Mises distribution*, the generalization of a normal distribution for cyclic variables.

The distribution (1.25) peaks around zero, and gets narrower as the |*E*| values of the three involved reflections increase; this means that a triplet with big *Es* is very likely to have a value close to zero. More sophisticated expressions can be derived for quartets (s.i. linking 4 reflections), quintets and so on, but the usefulness of these higher-order invariants is very limited.

If *r* pairs of phases $\{\phi_{\boldsymbol{k}_j}, \phi_{\boldsymbol{h}-\boldsymbol{k}_j}\}_{j=1,-,r}$ are known, together with the correspondent moduli $\{|E_{\boldsymbol{k}_j}|, |E_{\boldsymbol{h}-\boldsymbol{k}_j}|\}_{j=1,-,r}$, the total probability distribution for the phase $\phi_{\boldsymbol{h}}$ is given by the product of the corresponding distributions (assuming that thay are independent):

$$P(\phi_{\boldsymbol{h}}) = \prod_{j=1}^{r} P_j(\phi_{\boldsymbol{h}}) = A \exp\left[\sum_{j=1}^{r} G_{\boldsymbol{h}\boldsymbol{k}_j} \cos(\phi_{\boldsymbol{h}} - \phi_{\boldsymbol{k}_j} - \phi_{\boldsymbol{h}-\boldsymbol{k}_j})\right] \tag{1.26}$$

7 I_0 is the modified Bessel function and the σ parameters depend on the atomic numbers of the atoms involved.

and can be put in the form

$$P(\phi_{\boldsymbol{h}})=A\exp[\alpha_{\boldsymbol{h}}(\phi_{\boldsymbol{h}}-\beta_{\boldsymbol{h}})]\,. \tag{1.27}$$

with the following definitions:

$$\alpha_{\boldsymbol{h}}=\sqrt{S_{\boldsymbol{h}}^{\,2}+C_{\boldsymbol{h}}^{\,2}}\,,\quad \tan\beta_{\boldsymbol{h}}=\frac{S_{\boldsymbol{h}}}{C_{\boldsymbol{h}}} \tag{1.28}$$

and

$$C_{\boldsymbol{h}}=\sum_{j=1}^{r}G_{\boldsymbol{h}\boldsymbol{k}_j}\cos(\phi_{\boldsymbol{k}_j}+\phi_{\boldsymbol{h}-\boldsymbol{k}_j})\,,\qquad S_{\boldsymbol{h}}=\sum_{j=1}^{r}G_{\boldsymbol{h}\boldsymbol{k}_j}\sin(\phi_{\boldsymbol{k}_j}+\phi_{\boldsymbol{h}-\boldsymbol{k}_j})\,. \tag{1.29}$$

The expression obtained has still the form of a von Mises distribution; the maximum is attained for $\phi_{\boldsymbol{h}}=\beta_{\boldsymbol{h}}$ and the variance depends on $\alpha_{\boldsymbol{h}}$. The formula giving the value for $\tan\beta_{\boldsymbol{h}}$ is known as *tangent formula* and represents one of the cornerstones of direct methods phasing; the phase estimates obtained through it have an uncertainty that gets smaller as the quantity $\alpha_{\boldsymbol{h}}$ increases.

An exact relationship for a structure made of identical, resolved atoms can be derived (Sayre, 1952a). The electron density in the unit cell can be written

$$\rho(\boldsymbol{x})=\sum_{j=1}^{N}\rho_j(\boldsymbol{x})=\sum_{j=1}^{N}\rho_{atom}(|\boldsymbol{x}-\boldsymbol{x}_j|) \tag{1.30}$$

Given that $f(\boldsymbol{h})$ is the Fourier transform of the electron density $\rho_{atom}(\boldsymbol{x})$ of a single atom, the transform of the unit cell is simply the sum of *N* atomic contributions:

$$F(\boldsymbol{h})=F[\rho(\boldsymbol{x})]=f(\boldsymbol{h})\sum_{j=1}^{N}\exp(2\pi i\,\boldsymbol{h}\cdot\boldsymbol{r}_j)\;,\quad f(\boldsymbol{h})=F[\rho(\boldsymbol{x})] \tag{1.31}$$

Since the atoms are assumed to be well resolved, the squared density can be written as a sum of individual 'squared atoms':

$$\rho^2(\boldsymbol{x})=\sum_{j=1}^{N}\rho_j^2(\boldsymbol{x})=\sum_{j=1}^{N}\rho_{atom}^2(\boldsymbol{x}-\boldsymbol{x}_j) \tag{1.32}$$

and the corresponding Fourier transform will also be a sum:

$$G(\boldsymbol{h})=V\,F\,[\rho^2(\boldsymbol{x})]=g(\boldsymbol{h})\sum_{j=1}^{N}\exp(2\pi\, i\,\boldsymbol{h}\cdot\boldsymbol{r}_j)\quad,\quad g(\boldsymbol{h})=F[\rho^2(\boldsymbol{x})]\,. \tag{1.33}$$

We see that

$$F(\boldsymbol{h})=\frac{f(\boldsymbol{h})}{g(\boldsymbol{h})}G(\boldsymbol{h})=\theta_{\boldsymbol{h}}G(\boldsymbol{h})\,. \tag{1.34}$$

Observing that squaring in direct space means self-convolution in the Fourier space, one has

$$V^{-1}G(\boldsymbol{h})=F\,[\rho^2(\boldsymbol{x})]=F\,[\rho(\boldsymbol{x})]*F\,[\rho(\boldsymbol{x})]=V^{-2}F(\boldsymbol{h})*F(\boldsymbol{h}) \tag{1.35}$$

$$G(\boldsymbol{h})=V^{-1}F(\boldsymbol{h})*F(\boldsymbol{h})=V^{-1}\sum_{\boldsymbol{k}}F(\boldsymbol{k})F(\boldsymbol{h}-\boldsymbol{k}) \tag{1.36}$$

and we are led to the Sayre equation:

$$F(\boldsymbol{h})=\theta_{\boldsymbol{h}}V^{-1}\sum_{\boldsymbol{k}}F(\boldsymbol{k})F(\boldsymbol{h}-\boldsymbol{k})\text{ , or} \tag{1.37}$$

$$|F(\boldsymbol{h})|\exp(i\,\phi_{\boldsymbol{h}})=\theta_{\boldsymbol{h}}V^{-1}\sum_{\boldsymbol{k}}|F(\boldsymbol{k})\,F(\boldsymbol{h}-\boldsymbol{k})|\exp[i\,(\phi_{\boldsymbol{k}}+\phi_{\boldsymbol{h}-\boldsymbol{k}})]\,. \tag{1.38}$$

The equation (1.37) links together all the structure factors amplitudes and phases, showing that applying self-convolution to the structure factors is equivalent to scaling them by a function $V/\theta_{\boldsymbol{h}}$, where $\theta_{\boldsymbol{h}}$ is the ratio between the scattering factors of 'normal' and 'squared' atoms. Although strictly valid for equal atom structures only, the equation has been succesfully applied even to organic structures containing a few heavier atoms. The error introduced by the different chemical identities of the atoms leads mainly to an overweighting of heavier atoms, since in that case the Sayre equation tends to give the phases for the squared density.

The Sayre (complex) equation can be partitioned in two (real) equalities, considering that

both sides must be equal in phase and in modulus. The phase equality resembles much the tangent formula:

$$\tan\phi_{\boldsymbol{h}}=\frac{\sum_{\boldsymbol{k}} F_{\boldsymbol{k}}F_{\boldsymbol{h}-\boldsymbol{k}}\sin(\phi_{\boldsymbol{k}}+\phi_{\boldsymbol{h}-\boldsymbol{k}})}{\sum_{\boldsymbol{k}} F_{\boldsymbol{k}}F_{\boldsymbol{h}-\boldsymbol{k}}\cos(\phi_{\boldsymbol{k}}+\phi_{\boldsymbol{h}-\boldsymbol{k}})} \tag{1.39}$$

However, it must be pointed out that the two equations have a very different meaning. The eq. (1.39) one is an exact equation, provided that the structure contains only equal atoms and that the sum is carried over the whole reciprocal space (that is, including *all* the possible indices $\boldsymbol{k}$). The tangent formula (1.28) gives instead the maximum of a statistical distribution, which is the best estimate for the phase $\phi(\boldsymbol{h})$ compatible with the terms included in the summation. Moreover, no proof exists that the various estimates combined in the tangent formula are truly independent as required.

A more effective phase equality can be derived from the whole Sayre equation by considering that a good set of phases should satisfy a system of equalities of the form

$$E_{\boldsymbol{h}}=\frac{K}{g_{\boldsymbol{h}}}\sum_{\boldsymbol{k}} E_{\boldsymbol{k}}E_{\boldsymbol{h}-\boldsymbol{k}} \tag{1.40}$$

where $g_{\boldsymbol{h}}$ is the scattering factor for 'squared' atoms and K is an overall scaling constant. By minimizing the residual

$$R=\sum_{\boldsymbol{h}}\left|g_{\boldsymbol{h}}E_{\boldsymbol{h}}-K\sum_{\boldsymbol{k}} E_{\boldsymbol{k}}E_{\boldsymbol{h}-\boldsymbol{k}}\right|^2 \tag{1.41}$$

one obtains the *Sayre-equation tangent formula* (Debaerdemaeker *et al*., 1988b):

$$\phi_{\boldsymbol{h}}=\textit{phase of}\left[\sum_{\boldsymbol{l}}(g_{\boldsymbol{h}}+g_{\boldsymbol{l}}+g_{\boldsymbol{h}-\boldsymbol{l}})E_{\boldsymbol{l}}E_{\boldsymbol{h}-\boldsymbol{l}}-2K\sum_{\boldsymbol{l}}\sum_{\boldsymbol{k}} E_{\boldsymbol{h}-\boldsymbol{l}}E_{\boldsymbol{k}}E_{\boldsymbol{l}-\boldsymbol{k}}\right] \tag{1.42}$$

This expression contains both triplet and quartet terms and takes into account the information from small-valued structure factors. It has been used in the program *SAYTAN*.

The first implementations of direct methods programs were based upon *symbolic addition*: a small starting set of strong reflections is selected, while some origin-defining

phases are chosen and given explicit values. Symbols are assigned to the phases in the starting set, which then is expanded by means of a network of triplet relationships. At the end of this procedure, different numeric values are given to each symbolic phase to generate different solutions, which can be ranked according to specific figures of merit (*FOM*s) measuring the deviation of phase values from their expected statistical behaviour. Since an initial error propagates through the process, unreliable starting triplets can give rise to wrong solutions. For this reason specific procedures have been devised to select a very good starting set. A drawback in symbolic procedures is that, because of the cyclic nature of the variables, it is not possible to combine together multiple estimates relating to the same phase but involving different symbols. To overcome this difficulty, the multisolution approach has been introduced: several sets of values are given to the initial phases, each set is expanded by means of a weighted tangent formula, and *FOM* values are calculated for each solution. An example of such an approach is offered by the program *MULTAN* (Debaerdemaeker *et al*., 1988a). The software SHELX (Usón, 1999) makes use of a simulated annealing procedure into phase refinement. This latter is still carried out by means of a modification of the well-known tangent formula, but phases are subjected to 'thermal fluctuations' around their predicted values, allowing a random walk in phase-space that avoids local minima.

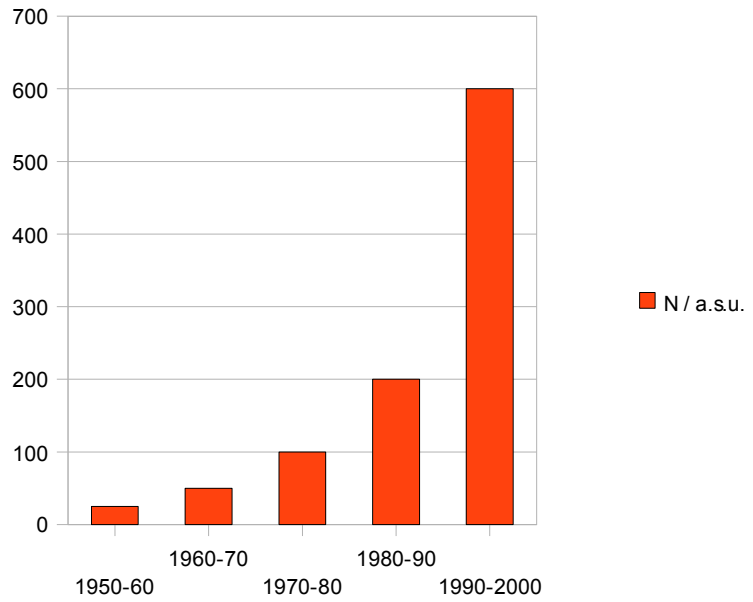


Fig. 3 The growth in power of Direct Methods, expressed as the maximum number of atoms in the asymmetric unit of solvable structures.

Direct methods have been an unvaluable tool in crystallography, allowing for automated structure solution up to hundreds of atoms in the asymmetric unit; they are still the preferred choice in solving the structure of small molecules. Since they operate only in reciprocal space, their request for computing resources is limited. However, they suffer from many limitations that have prevented until now their extensive use in macromolecular crystallography. These are:

- the probabilistic relationships become weaker as the number of atoms N in the cell increases. Structures with more than ~200 atoms in the asymmetric unit are difficult to solve;
- they cannot deal with data at a resolution less than 1.2 Å, because of the breakdown of the atomicity assumption (the 1.2 Å limit has been somewhat relaxed in these last years);
- the atoms are assumed to be uniformly and randomly distributed in the cell, an unrealistic approximation since atom positions are correlated and the distribution is non-uniform (this is especially true for macromolecular crystals). Thus a lot of available chemical information, which would considerably strenghten the method, remains unexploited.
- the probability distributions are derived through the use of questionable mathematical approximations (Bricogne, 1997a).

Classical direct methods only compute the conditional probability distributions for many small sets of phases (mainly triplets and quartets), and try then to put them together to get estimates on larger sets. This approach avoids many mathematical difficulties, but it relies on questionable approximations and ends up in weakening considerably the predictive power implicit in the random atom model.

A completely general approach (Bricogne, 1997a) calls for bayesian statistics, whose framework allows the computation of joint probability distributions for large sets of structure factors. In that case, the mathematical treatment is much more complex than in classical direct methods, and no general analytical expression for the probabilities can be derived; one is forced instead to formulate different phase hypotheses, which can be then evaluated through a likelihood criterion to be accepted or rejected. A phasing procedure must then be strongly hierarchical, starting from a small set of phased structure factors and

extending it along a 'multisolution tree'. At each node of the tree, two quantities, *entropy* and *likelihood*, are used to estimate the best phases of the current reflection set. While entropy measures the strenght of phase estimates in relation to the basis set, the likelihood is the probability of having the observed values for reflections out of the basis set given the phases and constitutes a 'look-ahead' tool. Being more general and flexible, the bayesian approach would accommodate stereochemical constraints (Bricogne, 1997b) into a *random fragment model* (an extension of the random atom model) and, if mathematical difficulties are overcome, allow *ab initio* phasing for macromolecules even when the diffraction data do not extend to atomic resolution.

The Shake-and-bake approach

A powerful dual-space method, capable of solving large structures, is the *Shake-and-Bake* algorithm, which has been implemented in the *SnB* computer program (Weeks *et al*., 1994), and also inspired the *Half-baked* procedure (Sheldrick and Gould, 1995) within the SHELX-97 package. SnB applies a real-space filtering by selecting atomic peaks in the electron density maps, while in reciprocal space a phase constraint arising from direct methods theory is applied. This consists in finding the minimum for the 'minimal function' (Hauptman, 1991):

$$m(\{\psi\}) = \left(\sum_{h,k} G_{h,k}\right)^{-1} \sum_{h,k} G_{h,k}\left[\cos(\psi_{h,k}) - \frac{I_1(G_{h,k})}{I_0(G_{h,k})}\right]^2 \qquad (1.43)$$

which is a measure of the deviation of all the considered triplet values $\{\psi\}$ from their expected value, and allows to state the phase problem in terms of global minimization. The dual space approach forces the same atomicity (and positivity) constraint in both spaces, resulting in a more powerful approach than the purely reciprocal-space one of traditional direct methods. In fact, one of the drawbacks in the minimization of $m(\{\psi\})$ is the presence of false minima; in this procedure, the problem is overcome by the real space peak search. Moreover, atomicity is imposed from the beginning, since the procedure directly starts with a random distribution of atoms in the asymmetric unit. This method is effective in solving structures containing up to 1000 independent non-hydrogen atoms, provided that atomic

resolution (d<1.1 Å) data are available. It is frequently used to solve the heavy atom substructures in SAS (*Single Anomalous Scattering*) and SIR (*Single Isomorphous Replacement*) applications, requiring only 3 Å resolution data.

The charge flipping algorithm

An interesting dual space method has been developed recently, the *Charge Flipping* (*CF*) algorithm (Oszlányi and Sütő, 2004), which is inspired by the Fienup algorithms used in optics. The algorithm forces atomicity in a strange, indirect way, by flipping the values of every density pixel under a given (positive) threshold δ. This flip preserves the norm of electron density while inducing a phase perturbation. In reciprocal space, no value is assumed for the zero-frequency term F_{000}, which is initially set to zero and then allowed to change freely during the iterations, while the other known moduli are simply imposed at each cycle. The evolution in phase space is chaotic, showing a strong dependence on initial conditions. Since the true, atomic solution represents a limiting cycle for the algorithm, a succession of iterations leads to structure reconstruction. When the solution is found, an abrupt change in total charge and R-factor occurs. No symmetry information is used and the reconstruction is carried out in a P1 cell; avoiding to fix the origin and the enantiomorph has the advantage that the structure can appear everywhere, so the efficiency of the algorithm is higher (the solution is not a point in phase space but rather a set of points, each of which corresponding to a different origin/enantiomorph choice).

$$\rho_{n+1}(\boldsymbol{r}) = \begin{cases} \rho_n(\boldsymbol{r}) & if \quad \rho_n(\boldsymbol{r}) > \delta \\ -\rho_n(\boldsymbol{r}) & if \quad \rho_n(\boldsymbol{r}) \leq \delta \end{cases} \tag{1.44}$$

This algorithm belongs to the *output-output* class, the less powerful among Fienup iterative schemes (Fienup, 1978). As a result, the iterations suffer from stagnation. Many modifications of the algorithm have been proposed to overcome stagnation problems, including a separated treatment for weak reflections (Oszlányi, 2005) and the introduction of the tangent formula into the algorithm (Coelho, 2007). The algorithm has been succesfully applied to incommensurately modulated structures (Palatinus, 2004). A modified

form including histogram matching has also proven to be effective in solving difficult structures from powder diffraction data (Baerlocher, 2007).
Due to the dimensionality of the phase space to be explored, even the best modifications of the method cannot work for large structures (more than ~300 non-hydrogen atoms in the asymmetric unit), even if atomic resolution data are available.

Patterson methods

Only in the last few years this class of methods has raised new attention, due to their unsuspected power in solving macromolecular structures. Patterson deconvolution relies over an atomistic interpretation of the Patterson function. This latter is the Fourier transform of the squared structure factors:

$$P(\boldsymbol{u}) = \frac{1}{V}\sum_{\boldsymbol{h}} |F(\boldsymbol{h})|^2 \exp(-2\pi \boldsymbol{h}\cdot\boldsymbol{u}) \,. \tag{1.45}$$

According to the Wiener-Khinchin theorem, $P(\boldsymbol{u})$ coincides with the autocorrelation of electron density:

$$P(\boldsymbol{u}) = \int_V \rho(\boldsymbol{r})\,\rho(\boldsymbol{r}+\boldsymbol{u})\,d\boldsymbol{r} \,. \tag{1.46}$$

This function $P(\boldsymbol{u})$ differs from the autocorrelation of any localized function in being a cyclic autocorrelation, that is, it has the same translational periodicity (unit cell) of the electron density. The squared structure factors are proportional to the observable intensities; their expression as function of the atomic coordinates writes

$$\begin{aligned} |F(\boldsymbol{h})|^2 = F(\boldsymbol{h})F^*(\boldsymbol{h}) &= \left(f_j \exp(2\pi i\, \boldsymbol{h}\cdot\boldsymbol{r}_j)\right)\left(f_k \exp(-2\pi i\, \boldsymbol{h}\cdot\boldsymbol{r}_k)\right) \\ &= \sum_j \sum_k f_j f_k \exp[2\pi i\, \boldsymbol{h}\cdot(\boldsymbol{r}_j - \boldsymbol{r}_k)] \end{aligned} \tag{1.47}$$

where f_j and $\boldsymbol{r}_j$ represent the atomic scattering factor and the positional vector of the *j*-th atom. Substituting this expression in the Fourier series for $P(\boldsymbol{u})$ gives:

$$P(\boldsymbol{u}) = \frac{1}{V}\sum_{\boldsymbol{h}}\left\{\sum_j\sum_k f_j f_k \exp[2\pi i\,\boldsymbol{h}\cdot(\boldsymbol{r}_j-\boldsymbol{r}_k)]\right\}\exp(-2\pi\,\boldsymbol{h}\cdot\boldsymbol{u})$$

$$= \sum_j\sum_k \frac{1}{V}\sum_{\boldsymbol{h}} f_j f_k \exp\{-2\pi i\,\boldsymbol{h}\cdot[\boldsymbol{u}-(\boldsymbol{r}_j-\boldsymbol{r}_k)]\} \qquad (1.48)$$

$$= \sum_j\sum_k P_{jk}[\boldsymbol{u}-(\boldsymbol{r}_j-\boldsymbol{r}_k)].$$

This shows that the Patterson function can be written as a sum of individual contributions or peak functions P_{jk}, each taking its maximum value when $\boldsymbol{u}=\boldsymbol{r}_j-\boldsymbol{r}_k$, that is, for a given interatomic vector. Provided that the resolution is high enough, the individual peaks of which $P(\boldsymbol{u})$ is made will be resolved from each other, and the set of maxima of the Patterson function will coincide with the set of interatomic vectors $\{\boldsymbol{r}_j-\boldsymbol{r}_k\}$ of the structure. Assuming that the Friedel law $|F(\boldsymbol{h})|=|F(-\boldsymbol{h})|$ is valid (that is, that electron density and so the f_j are real functions - no anomalous scatterers are present), the function will be symmetric with respect to the origin: $P(\boldsymbol{u})=P(-\boldsymbol{u})$. This is consistent with the simultaneous presence of a vector $\boldsymbol{r}_j-\boldsymbol{r}_k$ and its opposite $\boldsymbol{r}_k-\boldsymbol{r}_j$.

The complete set of interatomic vectors can be obtained by joining many sets of atomic coordinates of the same structure, each translated of a vector which is taken each time equal to one of the atomic coordinates; that is,

$$\{\boldsymbol{r}_j-\boldsymbol{r}_k\} = \{\boldsymbol{r}_j-\boldsymbol{r}_1\}\cup\{\boldsymbol{r}_j-\boldsymbol{r}_2\}\cup\ldots\cup\{\boldsymbol{r}_j-\boldsymbol{r}_N\}\,. \qquad (1.50)$$

Considering the distribution of peaks of the Patterson function, there is in principle one simple way to unravel the structure: it consists in a multiple superposition of the peak map with itself. If the map is translated by one of the interatomic vectors, that is, by setting the new origin on an arbitrary pivot peak $\boldsymbol{u}_p = \boldsymbol{r}_a-\boldsymbol{r}_b$, a set of peaks

$$\{\boldsymbol{r}_j-\boldsymbol{r}_k-(\boldsymbol{r}_a-\boldsymbol{r}_b)\} = \{\boldsymbol{r}_j-\boldsymbol{r}_1-(\boldsymbol{r}_a-\boldsymbol{r}_b)\}\cup\ldots\cup\{\boldsymbol{r}_j-\boldsymbol{r}_N-(\boldsymbol{r}_a-\boldsymbol{r}_b)\} \qquad (1.51)$$

will be obtained. The intersection between the original set and the new one will correspond to the vectors

$$\{\boldsymbol{r}_j-\boldsymbol{r}_a\}\cup\{-(\boldsymbol{r}_j-\boldsymbol{r}_b)\}\,, \qquad (1.52)$$

that is, to a single coordinate set with origin $\boldsymbol{r}_a$ plus its enantiomorph with origin $\boldsymbol{r}_b$. If the peak in $\boldsymbol{u}_p$ is a multiple one, arising from the superposition of m different peaks, then the intersection operation leads to a union of m different sets of the form (1.52). In any case, many interatomic vectors will be ruled out; multiple intersections, using each time a different pivot peak, can be performed, until the image has been extracted from the map. Unfortunately, this simple procedure is not applicable in practice, since real peaks will always have a finite width, giving rise to superpositions; only multiple peaks will emerge from the background of the map. For this reason, Patterson superposition methods have been used in the last decades mostly to exploit the knowledge represented by a partial model. To estimate the coincidence of the peaks some *image-seeking functions* have been proposed (Buerger, 1959):

$$\begin{aligned} \Pi(\boldsymbol{r}) &= P(\boldsymbol{r})\,P(\boldsymbol{r}-\boldsymbol{u}) \\ \Sigma(\boldsymbol{r}) &= P(\boldsymbol{r})+P(\boldsymbol{r}-\boldsymbol{u}) \\ M(\boldsymbol{r}) &= min\{P(\boldsymbol{r}),\,P(\boldsymbol{r}-\boldsymbol{u})\} \end{aligned} \tag{1.53}$$

These three functions (product, sum and minimum) are expected to take great values when a part of the map superposes exactly with itself.

Recently, the Patterson deconvolution approach has proven to be a very powerful one in solving macromolecular structures. A new approach combining Patterson vector methods and real space refinement has been devised and implemented first in the SIR2002 package (Burla *et al.*, 2002); with some improvements it has been included in the crystallographic package *IL MILIONE* (Burla *et al*., 2007). In this multisolution procedure, different superpositions are calculated, by choosing each time one of the highest peaks as pivot point; the maps arising from the vector process are used as starting point for a series of density modification cycles. The starting point are *implication transformations*, which allow to take into account the crystallographic symmetry. They are defined as

$$I_s(\boldsymbol{r}) = P(\boldsymbol{r}-\boldsymbol{C}_s\boldsymbol{r})/n_s \tag{1.54}$$

where $(\boldsymbol{r}-\boldsymbol{C}_s\boldsymbol{r})$ is an Harker vector, and $\boldsymbol{C}_s$ is the s-th symmetry operator with multiplicity n_s. When the space-group symmetry has more than two primitive operators, all the implication transformations can be combined into the *multiple implication function*:

$$SMF(\boldsymbol{r}) = \min_{s=1}^{m} I_s(\boldsymbol{r}) \tag{1.55}$$

which acts by selecting at each point the lowest value between all the set of functions $I_s(\boldsymbol{r})$. The *SMF* map will present peaks for atomic positions compatible with any of the permissible origins for the space group and also for their enantiomorphs. The map is then cleaned by computing a *minimum superposition function*

$$S(\boldsymbol{r}) = ISF[P(\boldsymbol{r}-\boldsymbol{r}_p), SMF(\boldsymbol{r})] \tag{1.56}$$

where $\boldsymbol{r}_p$ is a pivot peak selected from the highest peaks in the *SMF* map and *ISF* is one of the image-seeking functions (product, sum or minimum). The $S(\boldsymbol{r})$ functions obtained by each chosen pivot peak are then used as starting points for a real space refinement. In the first cycles of real space refinement the electron density map can be further cleaned by means of the *PFF* map, defined as

$$PFF(\boldsymbol{r}) = min[P(\boldsymbol{r}-\boldsymbol{r}_p), FF(\boldsymbol{r}+\boldsymbol{r}_p)] \tag{1.57}$$

where the *FF* map is a map showing peaks for the sum of atomic vectors and, unlike the Patterson map, depends on the phases:

$$FF(\boldsymbol{u}) = \frac{1}{V}\sum_{\boldsymbol{h}} |F(\boldsymbol{h})|^2 \exp(2i\phi_{\boldsymbol{h}} - 2\pi i\,\boldsymbol{h}\cdot\boldsymbol{u}) \tag{1.58}$$

The overall process can be seen as a repeated 'filtering' of the Patterson function, aiming at eliminating a number of 'multiple images' of the structure as well as the symmetry due to the inversion center. The method is less resolution-sensitive than direct methods; it is also faster with respect to the tangent procedure implemented in the same package, since a lower number of trials are needed. Moreover, the efficiency of the deconvolution procedure does not depend on the number of atoms in the asymmetric unit (a.s.u), as direct method do, allowing the solution of structures with as many as 6000 atoms in the a.s.u., provided that atomic resolution data are available and that at least a moderately heavy atom (Ca) is

present.

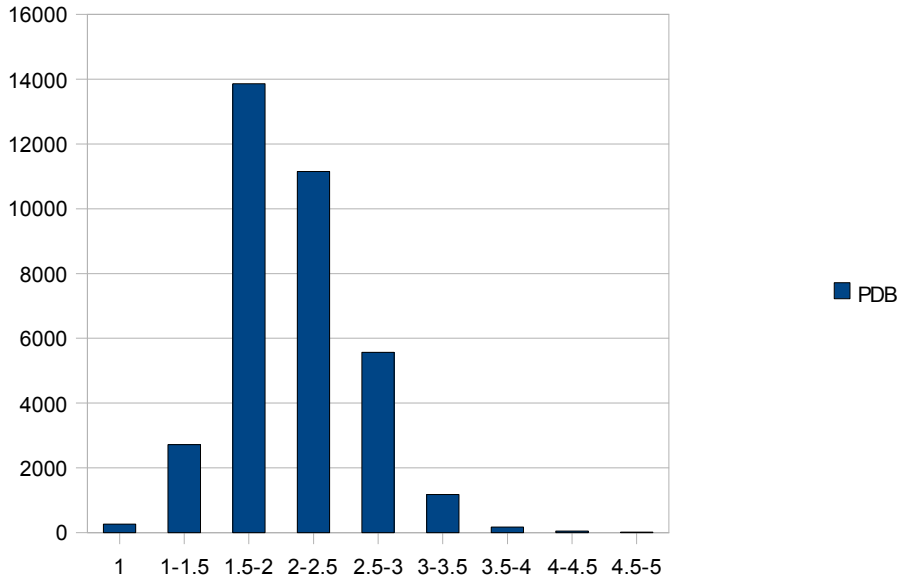


Fig. 4 Number of protein structures in the Protein Data Bank (PDB) ordered by resolution intervals. The majority (87%) of them falls in the resolution range 1.5-3 Å; less than 10% has atomic resolution and can thus be solved in principle by current *ab initio* methods.

Resolution and uniqueness of the phase problem for proteins

The real obstacle to apply *any* of the conventional *ab initio* methods, regardless of their real- or reciprocal-space nature, to protein structure solution, is the limited resolution at which most macromolecular crystals diffract. As reported in (fig. 4), the majority of protein datasets has a resolution limit falling in the range 1.5-3 Å. In such cases, since the atomic resolution reflections are missing, the atomicity restraint is not strong enough to guarantee the uniqueness of solution. This degeneracy can be probed by a simple experiment (Baker *et al.*, 1993a): the atoms of a protein are randomly placed in the unit cell, and after minimization against less-than-atomic resolution data (d_{min}>1.5 Å), it is observed that low *R*-factor values are always attained within little atomic displacements (rmsd 1.5-2.0 Å). The crystallographic *R*-factor is the residual defined by

$$R = \frac{\sum_{\boldsymbol{h}} \left| |F_{calc}(\boldsymbol{h})| - K|F_{obs}(\boldsymbol{h})| \right|}{\sum_{\boldsymbol{h}} |F_{obs}(\boldsymbol{h})|} \quad , \tag{1.59}$$

where K is a scale factor, required to bring the $|F_{obs}(\boldsymbol{h})|$ and the $|F_{calc}(\boldsymbol{h})|$ on the same scale:

$$K = \frac{\sum_{\boldsymbol{h}} |F_{calc}(\boldsymbol{h})|}{\sum_{\boldsymbol{h}} |F_{obs}(\boldsymbol{h})|} \quad . \tag{1.60}$$

The quantity R is a measure of the deviation between the structure factor magnitudes $|F_{calc}(\boldsymbol{h})|$ calculated from the model and the experimental ones ($|F_{obs}(\boldsymbol{h})|$). When the observations/parameters ratio is high, low R values are indicative of a good (physically meaningful) model; however, when this condition is not fulfilled, arbitrarily low R values can be obtained even from models that are physically meaningless. Indeed, the phases corresponding to the atomic arrangements obtained do not show any correlation with the true ones (mean phase error *MPE* = 84-89 Å, very close to the 90° expected for complete uncorrelation); moreover, the phases resulting from different runs are also completely uncorrelated with each other. This shows that the $R = R(\boldsymbol{r}_1, .., \boldsymbol{r}_N)$ hypersurface has many local minima, the vast majority of them bearing no resemblance with the true structure. Those minima are so many that any random set of coordinates has one in its close neighborhood. Restricting the atomic positions inside the true molecular envelope does not change the results: at low resolution (14 Å) strong phase correlations with the true solution are obtained, reflecting the *a priori* knowledge about the protein/solvent boundary, while the higher resolution shells are still randomly phased.

Besides atomicity, another restraint named *connectivity* can be defined on the basis of electron density topology, by selecting all the points of the density map above a given threshold and joining them to form a skeleton; a high connectivity corresponds to a skeleton made of a few long segments, and the corresponding density is likely to be close to the true one. Connectivity is strongly correlated to phase error, decreasing smoothly as *MPE* increases; indeed, false solutions obtained from minimization of random atomic coordinates show a very low connectivity (<0.1, compared to 0.97 of the true map). A very interesting property is that this correlation persists until very low resolutions (d_{min}>12 Å), so that it could be exploited in phasing even in absence of atomic resolution data. Although a phasing process through a direct optimization of the connectivity is not possible, because no analytical expression exists for that quantity and so no derivatives can be computed,

connectivity can be used as a figure of merit to judge the quality of a set of phases. An *iterative skeletonization* procedure, based on this idea, has been proposed for phase improvement and implemented in the *PRISM* program (Baker *et al*., 1993b).

Conditional optimization

Stereochemical data, like known ideal bond lenghts and angles, are a powerful constraint for macromolecular refinement, allowing the quality of the atomic model to be improved by maximizing the fit with experimental data. In absence of bond and angle constraints, refinement would diverge, because of the low observables/parameters ratio. Unfortunately, the usefulness of stereochemical constraints in *ab initio* phasing is limited by the difficulty of imposing them on electron density or phases without explicitly building an atomic model. Starting a conventional refinement from a random conformation of a pre-built model does not lead to the correcty structure because of the many local minima of the problem and the intrinsic slow nature of the search.

An attempt to introduce stereochemistry into a phasing process is the method of *conditional optimization* (Scheres and Gros, 2004), in which a kind of refinement is carried out on loose atoms. The protocol starts from random coordinates to which a chemical identity is assigned each time according to their neighborhood and are refined under a force field based upon ideal geometry. The method has been tested for model building from experimental phases, but also for *ab initio* solution with experimental data truncated at 2 Å resolution. All the helices of a four-helical bundle could be reconstructed at the end of 1000 optimization steps, although the directionality of the resulting chain is not always correct and loops were missing in the final model due to intrinsic limitations of the force field.

Ab initio phasing starting from low resolution

The major problem in macromolecular phasing is the limited resolution of the measurable diffraction data. Alternative approaches to direct methods, which try to estimate first the phases of the strongest reflections, are an ensemble of techniques whose starting point is to build phase estimates for the lowest resolution shell. The basic idea is to choose a few reflections at very low resolution (usually less than 15 Å) and assign them phases according

to real-space criteria (that is, requiring the corresponding electron density map to match some expected properties). This small set of phased reflections can be used in principle as a starting point for an extension procedure for phases extrapolation to higher resolution. Since some very-low-resolution reflections are always missing in data sets recorded with a straightforward procedure, these methods require the diffraction experiment to be carried out with an experimental setting a little different from usual. Lunin and co-workers (Lunin *et al*., 2000a) investigated a general phasing process in which:

- a small number of very low resolution reflections are chosen (e.g. 39 independent reflections at 16 Å resolution), defining a starting set **S**.
- a great number of phase sets $\{\phi(\boldsymbol{h})\}_{\boldsymbol{h}\in\boldsymbol{S}}$ for these reflections are generated at random;
- for each set of phases the corresponding electron density (Fourier synthesis) is computed;
- each of these low-resolution maps is evaluated according to a selection criterion;
- the phase sets giving rise to the more likely maps are retained;
- if necessary, cluster analysis of the phase sets is performed, in order to group them in classes of similar solutions. To calculate the closeness of two phase sets an origin alignment is performed by maximizing the map correlation coefficient;
- aligned phase sets belonging to the same cluster are averaged, giving rise to one of the final solutions.

The alignment procedure is required because of the origin and enantiomorph ambiguity in defining the phases. Two phase sets that look completely different can actually give rise to a pair of similar electron density maps that are related by an origin shift (and/or enantiomorph inversion). An origin-shift is applied to the phases until the real space map correlation C_ϕ (Lunin and Woolfson, 1993) is maximized:

$$C_\phi = \frac{\int [\rho_1(\boldsymbol{r})-\langle\rho_1\rangle][\rho_2(\boldsymbol{r})-\langle\rho_2\rangle]d\,\boldsymbol{r}}{\int [\rho_1(\boldsymbol{r})-\langle\rho_1\rangle]^2 d\,\boldsymbol{r}\int [\rho_2(\boldsymbol{r})-\langle\rho_2\rangle]^2 d\,\boldsymbol{r}} = \frac{\sum_{\boldsymbol{h}} F^{obs}(\boldsymbol{h})^2\cos[\phi_1(\boldsymbol{h})-\phi_2(\boldsymbol{h})]}{\sum_{\boldsymbol{h}} F^{obs}(\boldsymbol{h})^2} \tag{1.61}$$

Some selection criteria which have been tested are:

- the closeness of map histogram to the expected one at the working resolution;
- the connectivity properties of the maps (number of connected regions above a given threshold);
- the probability of obtaining the observed structure factor magnitudes by placing the atoms at random inside the connected regions.

Each of these criteria is rather weak in discriminating between the sets which are close to the true solution from those that are far apart. Many individual phase sets can be wrong, and nevertheless obtain a high score according to some criterion, while better sets can be incorrectly classified as bad. Nevertheless, there is a statistical correlation between the closeness of the selected phase sets to the correct one and the measure of goodness offered by each criterion. This means that selecting out of the random ensemble the phase sets with higher score leads to a population statistically enriched in variants which are closer to the true solution. Averaging over this smaller ensemble cancels out the random differences between its elements, while common features are enhanced; the final averaged set of phases is much closer to the true solution than the majority of the selected variants. As by-product of the averaging procedure over M individual phases $\phi_j(\boldsymbol{h})$, a figure of merit $m(\boldsymbol{h})$ is obtained:

$$m(\boldsymbol{h})\exp[i\phi^{best}(\boldsymbol{h})] = \frac{1}{M}\sum_{j=1}^{M}\exp[i\phi_j(\boldsymbol{h})] \tag{1.62}$$

Often, simply averaging the selected variants leads to a reasonable solution of the phase problem. In some cases, however, the variants tend to concentrate around more than one center. A better averaging strategy is then to group the selected phase sets in increasingly large clusters according to their 'distance', defined by

$$\text{dist}\,[\{\phi_1(\boldsymbol{h})\},\{\phi_2(\boldsymbol{h})\}] = [2(1-C_\phi)]^{1/2} \tag{1.63}$$

A convenient level of clustering is then chosen, and averaging is carried out for each of the resulting clusters. Low resolution maps obtained with different selection criteria show a

correlation of about 60-70 % with the exact one.

Among all the tested criteria, the most effective has proven to be connectivity (Lunin *et al.*, 2000b). In the low-resolution context, this word does not imply the construction of a skeleton, but rather the segmentation of the electron density in a number of isolated regions. To analise the connectivity, a Fourier synthesis is computed on a grid from the trial phases and the observed moduli, and a mask Ω_κ is then obtained by selecting the points in the synthesis which have values above a cutoff level κ:

$$\Omega_\kappa = \{\boldsymbol{r} : \rho(\boldsymbol{r}) > \kappa\} \tag{1.64}$$

In order to make the connectivity independent of the scale of the synthesis it is convenient to define the cutoff level κ as a function of the specific volume α:

$$\alpha = \frac{Volume(\Omega_k)}{number\ of\ residues\ per\ unit\ cell}\ . \tag{1.65}$$

Once α has been fixed, κ becomes a function of the scale, allowing the region Ω_κ to be scale-independent. An optimal choice when working with ~15 reflections is $\alpha = 25\ A^3/residue$; for this specific volume each molecule in the cell is expected to give rise to a single separated region in the mask Ω_κ. The mask is then separated into its *connected components*; each c.c. is formed by grid points that can be joined by a continuous chain of neighboring points belonging to the same component. The selection of phase sets is then performed on the basis of the number of connected components in the unit cell, which should be equal to the number of expected molecules. Additional criteria, like the volume of connected regions, can be applied.

Connectivity-based phasing is a valuable tool when the standard crystallographic methods fail, or when only low resolution data are available; for example, successful attempts to phase low-density lipoprotein to a resolution of 27 Å (Lunin *et al*., 2001) and lectin SML-2 to 16 Å (Müller *et al.*, 2006) have been reported. Some interesting results were obtained in a test case on the small, 61 residues long protein G, whose structure had been previously solved (Derrick and Wigley, 1994). Only one molecule per asymmetric unit is present, and since there are four asymmetric units (space group $P2_12_12_1$) the low resolution map was

expected to show four separated regions of approximately the same volume. This criterion, with some additional restraints at higher resolution, allowed phasing to be performed from a very low resolution starting set (16 Å) up to an effective resolution of about 4.5 Å (Lunina *et al*., 2003). It must be observed that at the stage of variant clustering human intervention is still of great usefulness. Instead, for determining the most crucial parameters of the method, like the correct choice for the space group and the number of molecules per unit cell, automated procedures have been proposed (Urzhumtseva *et al.*, 2004).

Weak constraints and the density modification scheme

Although they cannot be generally considered tools for *ab initio* phasing, the so-called *density modification* methods, widely used in macromolecular crystallography, offer some examples of successful application of weak constraints for improving or extending existing phase sets (Cowtan and Zhang, 1999). The use of weak constraints is justified by the fact that the initial estimates for the phases, usually arising from experiment, are in a close neighborhood of the correct solution. From an algorithmic point of view, density modification follows a simple alternating scheme, like the classical Gerchberg-Saxton algorithm (Gerchberg and Saxton, 1972). The procedure goes back and forth between real and reciprocal space (fig. 5), restoring in turn the real constraints (which can be of many kinds) and the Fourier constraints (experimental moduli and initial phase estimates).

At each *i*-th cycle, proper weights must be calculated in order to combine the new phases $\{\phi(\boldsymbol{h})\}_i$ with the initial ones $\{\phi(\boldsymbol{h})\}_0$. This is done by multiplying their respective probability distributions:

$$P_{new}[\phi(\boldsymbol{h})] = P_{init}[\phi(\boldsymbol{h})]\, P_{mod}[\phi(\boldsymbol{h})] \tag{1.66}$$

The recombination of the modified phases with the initial ones is mandatory, because in almost all applications underdetermined constraints are used; however, the recombination process assumes independence between the two phase sets, which is not true and leads to strong bias with respect to the initial phases. The weights for the modified phases are given according to the agreement of the calculated factors with the observed ones. Unfortunately, due to underdeterminacy of the constraints, a large enough number of cycles will lead to an

arbitrarily good agreement between model magnitudes and their observed value; this agreement is not necessarily correlated with phase improvement. To limit this possibility, a small number of cycles is usually performed when weakly phased reflections are included.

Solvent flattening

Most biological molecules have a roughly globular shape, so that the crystal packing shows many gaps that are filled with disordered solvent from the crystallization solution. With the exception of the hydration shell in the proximity of protein molecules, the solvent regions can be assumed to a good approximation as having a flat electron density, which is an average value over a great number of unit cells. If the solvent regions have been identified, then the phases can be improved by imposing a constant value to the electron density in these regions. The most common method (Wang, 1985) for selecting the solvent regions is the following:

- the electron density map is truncated:

$$\rho_{tr}(\boldsymbol{x})=\begin{cases}\rho(\boldsymbol{x}), & \rho(\boldsymbol{x})>\rho_{solv}\\ 0, & \rho(\boldsymbol{x})<\rho_{solv}\end{cases} \tag{1.67}$$

- the truncated map is smoothed by convoluting it with a smearing function (this operation is readily accomplished in reciprocal space):

$$\rho_{av}(\boldsymbol{x})=\rho_{tr}(\boldsymbol{x})*g(\boldsymbol{x}) \tag{1.68}$$

- the solvent region Ω_{solv} is obtained as

$$\Omega_{solv}=\{\boldsymbol{x}\,|\,\rho_{av}(\boldsymbol{x})<\rho_{cut}\} \tag{1.69}$$

The real space operation is then simply

$$\rho_{mod}(\boldsymbol{x})=\begin{cases}\rho(\boldsymbol{x}) & \boldsymbol{x}\in\Omega_{solv}\\ \rho_{solv} & \boldsymbol{x}\notin\Omega_{solv}\end{cases} \tag{1.70}$$

Histogram matching

Histogram matching is the most conservative among density modification constraints, since it does not change the ordering of values of the electron density. The histogram of electron density is the distribution of its values; it can be constructed by discarding all the positional information present in an electron density map and by ordering all its values in increasing order. The number of values falling in given ranges is then counted to obtain an histogram that can be approximated by a continuous frequency distribution $P(\rho)$. At a resolution better than 6 Å, the histogram calculated over protein regions only is a function of the resolution and the temperature factor, whose effect can be removed. So, once the protein region is known, the point falling inside it should have a predictable density distribution that depends only on their resolution. According to a standard technique in image processing, a given density can be modified in order to have the theoretical histogram; since an infinite number of possible densities exist for a given histogram, the convention is not to alter the order of values in the map.

A variant of the method is the two-dimensional histogram matching (Nieh and Zhang, 1999). This technique exploits the joint information of the distribution electron density and its gradient to decouple the ordering of the density values between different cycles of refinement. In fact, the histogram constraint is a general but very weak constraint in that the locations of maxima and minima in the density is left unchanged. The 2-D histogram performs slightly better than the simple one.

NCS averaging

Non-crystallographic symmetry (NCS) arises in crystals when two or more molecules are related to each other by a symmetry operation that is not common to the whole crystal (that is, a local symmetry). The information about NCS can be used to average portions of the map corresponding to copies of the same molecule; moreover, the NCS operators cause the reflections being linked together by supplementary equations, providing powerful phase constraints. To be exploited, this property needs first to identify the masks for the single molecules and the NCS matrices relating them. This kind of constraint can be very strong, so that NCS-averaging can be considered the most powerful density modification technique; for very symmetrical molecules (or molecular assemblies like viruses), a*b initio* phasing is also possible.

Iterative skeletonisation

Connectivity can be imposed on electron density when an initial set of phases is available (from experimental methods or molecular replacement); the procedure, known as *iterative skeletonisation*, has been first implemented in the *PRISM* crystallographic software (*Baker 1993b*) and is available into the *CCP4* package as part of the *DM* program. At each cycle of the procedure, a *skeleton* is traced into the density by joining all the neighbouring points that lie above a given threshold, and the density is modified by setting to zero the points belonging to small disconnected elements of the resulting graph. After Fourier transforming the map, new phases are obtained which can be combined with the starting ones. Effectiveness varies much from one case to another.

Motivation of the project

From the exposition outlined above, it is clear that x-ray structural investigation in the macromolecular field is significantly hindered by the lack of high resolution data and the consequent underdeterminacy of the phase problem. Alternative approaches, which could be based for example on the topological properties of the electron density, have begun to develop in the last 20 years; however, they are still of very limited use, and often require human intervention, as well as careful setting of their basic parameters. Taking inspiration by these recent techniques, the present work is aimed at exploring some possible alternative approaches, that do not rely directly on the existence of atoms, but rather on some properties of the electron density that are likely to arise in (or persist until) a medium resolution regime (1.5-4.5 Å).

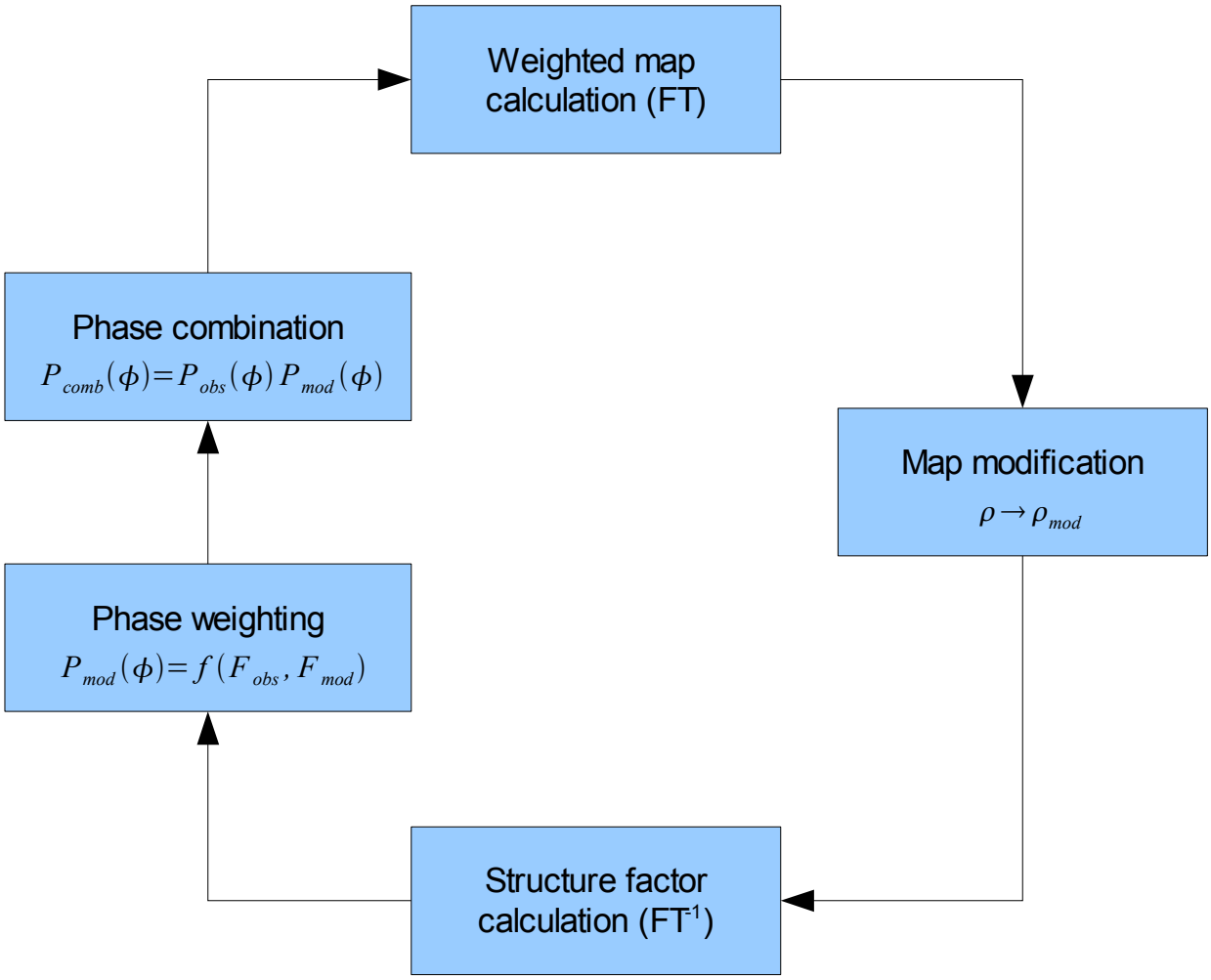


Fig. 5 The schematic representation of the density modification procedure

Chapter 2

A neural-network based approach

Introduction

In search of a new approach to phase problem for macromolecular structures, we addressed *artificial neural networks* (ANNs) as a potential tool to extract phase relationships from real examples. ANNs are parallel processors whose basic architecture is inspired to biological neural networks (Haykin, 1999): in fact, they are built of several interconnected layers of simple processing units, named *neurons*. Each neuron is an implementation of a mathematical function $\mathbb{R}^n \rightarrow \mathbb{R}$, which integrates many input signals and transforms them through a *transfer function,* to give a single output signal which can be redistributed to other neurons connected to it. The ANN as a whole is a mathematical model defining a function $f : A \rightarrow B$. While the general form of this function depends on the specific network architecture and on the nature of the transfer functions, its precise operation is defined by the connection parameters.

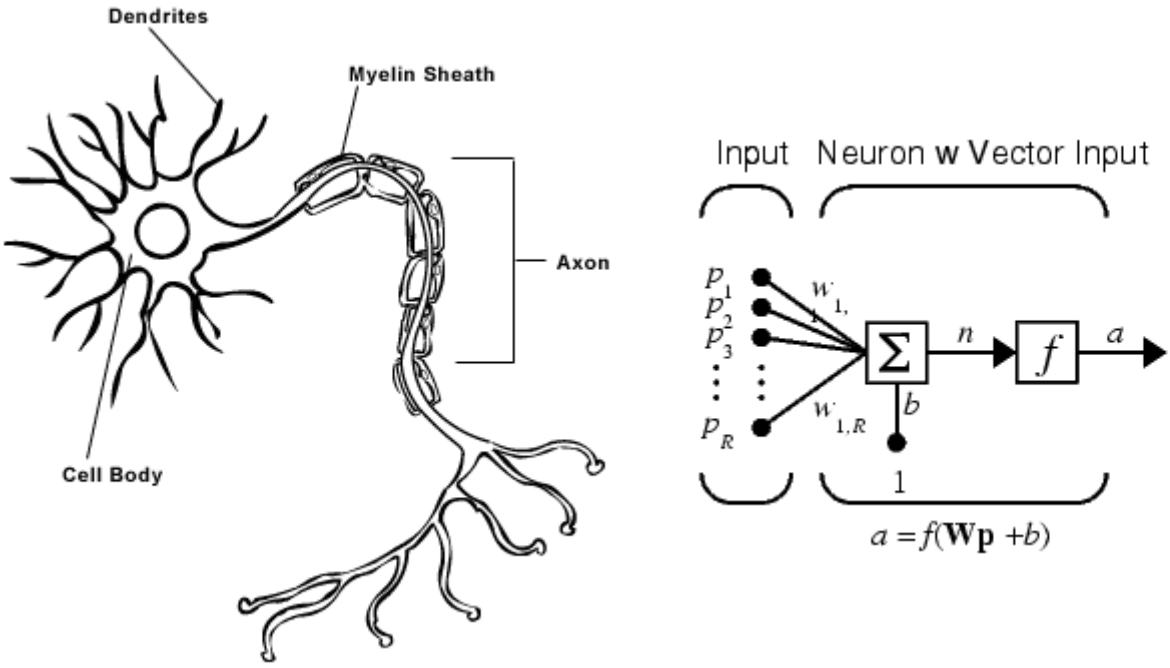


Fig. 1 A biological neuron and the 'neuron' processor as defined in ANNs. Both are able to integrate many inputs to give a single output, which will be sent in turn to many similar processing units. The 'body' of an artificial neuron is a weighted sum (the scalar product **Wp** of input vector and weight vector) upon which a *transfer function f* will act. (*Adapted from: left, http://www.drugabuse.gov/MOM/TG/momtg-introbg.html; right, Matlab package documentation (The Mathworks)*)

Two principal classes of architectures can be defined, depending if cycles are present (*recurrent ANNs)* or not (*feedforward ANNs*) in the connection graph. Cycles can lead to a

time-dependent behaviour; moreover they are present in Hopfield networks, a well-known class of networks that can act as associative memories. In feedforward networks the neurons are organized in layers (fig. 2), the output of one layer becoming the input for the next.

As in biological neural networks, the strenght of the connections between neurons can vary, leading to different behaviours of the whole network. Each connection between neurons is given a *synaptic weight*, a number by which the quantities sent through it are being scaled: for example, in order to become an input for the neuron *j,* the output from a neuron *i* must be multiplied by the weight wij. Specific *learning algorithms* allow a network to modify its weights according to a set of examples (*training set*), in order to reproduce input/output relationships or to classify the inputs. If only input data are used, allowing the network to find by itself their characteristic features, we speak of *unsupervised learning*; weight modification is carried out by minimizing a cost function which depends on the data and on the output values and can have any form. On the other hand, in *supervised learning* both input and output are used, and the cost function to be minimized is a measure of the difference between calculated and expected outputs. If transfer functions are differentiable the learning process can be accomplished using a *backpropagation* algorithm, in which the corrections to be applied to the weights are computed backwards from the output layer up to the input layer.

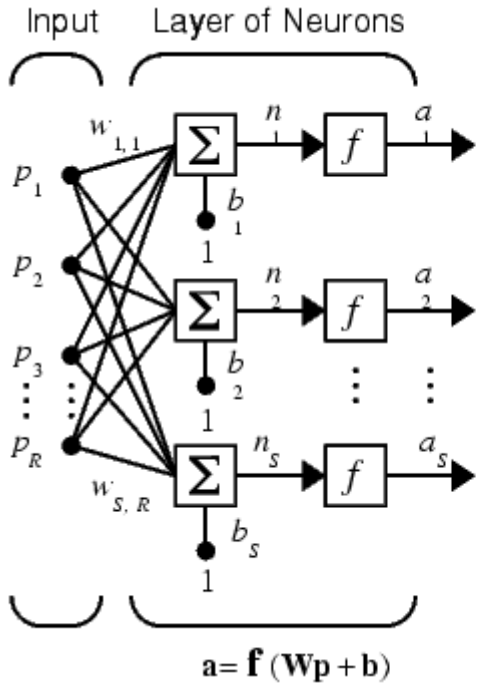


Fig. 2 A single layer of neurons, receiving an input vector **p** and returning the output vector **a**. This last can become in turn the input vector for the next layer of the network. The dimensions of input (R) and output (S) can be different. (**W**, weight matrix; b, bias vector; **f,** vector of the transfer functions relative to each neuron). (*Adapted from: Matlab package documentation, The Mathworks*)

In a so-called *feedforward ANN* several layers of neurons (fig. 2) are interconnected; each neuron in the i-th layer collects its inputs from all the neurons in the (*i-1*)-th layer and processes them, originating an output which is sent to all the neurons in the (*i+1*)-th layer. The first layer of the network (*input layer*) is fed with the input vector, while the last layer (*output layer*) generates the output vector. In between,

a number of inner (*hidden layers*) can be present (fig. 3).

Neural networks possess an intrinsic learning capability, which arises from having many adjustable parameters; a feedforward network can be used to model a complex function from a series of known points. During the learning process, input/output pairs constituting the *training set* are presented in succession to the network, and a learning algorithm modifies the weights in order to achieve the minimum difference between each calculated output and its target value found into the training set. At the end of this process, the network is able to reproduce at its best the output values corresponding to training set inputs. This means that, if some general relationship $a_i \sim b_i$ exists between corresponding elements of two sets A and B, and we train a suitable network with a training set made of elements of the subsets $A_{tr} \in A$ and $B_{tr} \in B$, the network will learn to predict b values from a values also for those elements of A and B not belonging to the training set. It is clear that a good generalization is attained only with a wisely chosen training set; this latter should not be too small and its elements have to be evenly distributed across the whole sets A and B.

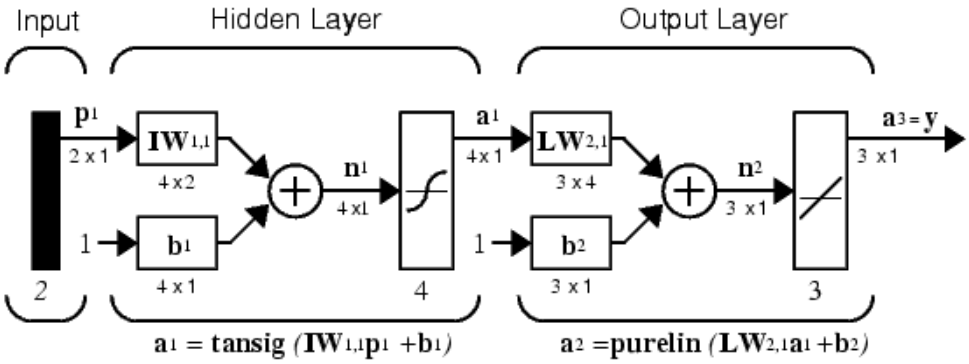


Fig. 3 An example of feedforward network made of two layers with different transfer functions, a sigmoid and a linear one. (Adapted from: Matlab package documentation, The Mathworks)

Network dimensions (number of neurons per layer and number of layers) have to be tailored for a specific application. While the input and output layer must have a number of elements respectively equal to input and output vectors, the dimensions of any other layer can vary. It is clear that a network is too small to accomplish a given task when the number of neurons (and as a consequence the number of adjustable parameters) is not sufficient to generalize the information contained in the training set. In that case no learning is observed.

A network is conversely too big when it has so many parameters that it always learns to reproduce well a given training set, even in absence of general relationships; such a network won't be useful, because in presence of a significant relationship in the data it will specialize in reproducing the specific features of the training set, losing any predicting capability on the rest. For real, noisy data, this means that the network learns to model exactly the training set noise, leading to a degraded mapping of the true relationship.

Neural networks and the phase problem

The phase problem is in principle solvable when appropriate *a priori* information is provided to constrain the solution. Since atomicity implies strong moduli/phases relationships, when atomic resolution data are available sufficient *a priori* information is always present. If resolution is less than 1.2 Å, however, more detailed stuctural information is needed, which, in the case of proteins, ranges from protein/solvent boundary to stereochemical data (ideal bond distances and angles). Although this kind of information is currently exploited at the stage of model refinement, when an atomic model has already been fitted into the density, no efficient way has been devised yet to incorporate it into an *ab initio* phasing procedure. The main reasons for this are the mathematical difficulty and the high dimensionality of the search space. The hypothesis made in this work is that some general but unknown moduli/phases relationships should exists for protein structures at non-atomic resolution. On this basis, one can guess that it could be possible to generalize them from examples; in this perspective neural networks are a very useful tool.

One-dimensional tests

As first approach, the problem was greatly simplified by restricting the study to one-dimensional atomic models. The choice of one dimension only was dictated by simplicity, while atomicity was required to test if the ANN approach could give results at least in a well known solvable case. The purpose of these over-simplified experiments was in fact to explore ANN capabilities in a case where the solution is known to be unique and mathematical relationships between moduli and phases have already been derived in the context of direct methods. In the more complex case of macromolecular phasing in absence of atomic data, direct methods relationships like Sayre equation have lost their strenght, although other relationships can exist, owing to general protein features. The training sets

were obtained by generating random coordinates for a given number of atoms along a line (fig. 5). The coordinates were passed to the software Shelx for structure factor computation, and the output file from Shelx was splitted in moduli and phases in order to create input and output training data.

When working with a neural network it is necessary to properly define the data to be processed. In particular, to a given input must correspond a single output, while it is known not to be the case when speaking about moduli and phases. These latter not only depend on the choice of origin and handedness of real-space axes, but in addition (being cyclic variables) are only defined modulo 2π. A given set of phases $\{\phi_h\}$ is equivalent to any other set of the form $\{\sigma(\phi_h - 2\pi h t) + 2k_h\pi\}$, where $\sigma = \pm 1$ expresses the enantiomorph choice, *t* is a real space origin shift and the k_h are arbitrary integer numbers. The cyclic ambiguity can in principle be avoided by transporting every phase value into the interval $[-\pi, \pi]$, but this approach can generate strong discontinuities in the function to be mapped since two extreme phase values (like 0.99π and -0.99π), although very different numerically, relate in fact to very similar phase choices. For this reason, one should not use directly the phase values, but rather their sines and cosines, which do not suffer from this ambiguity. The multiplicity of phase sets due to origin definition can be addressed in two ways:

- the origin can be fixed by choosing an index $\tilde{h}$ and applying to every phase set the required origin shift in order to have $\phi(\tilde{h}) = a$ (where *a* is the same for every set). Taking $\tilde{h} = 1$ and $a = 0$, the proper shift *t* can be obtained from $\phi'_1 = \phi_1 - 2\pi t = 0$ and one finds that each phase value of the set needs to be modified according to

$$\phi'_h = \phi_h - h\phi_1 = 0 \,. \tag{2.1}$$

- the network can be trained to predict only phases of *structure invariants* (s.i.). These are linear combinations of phases:

$$\psi(k_1, -, k_n) = \sum_{i=1}^{n} \phi(k_i) \quad , \tag{2.2}$$

which are independent from the origin choice and are defined by

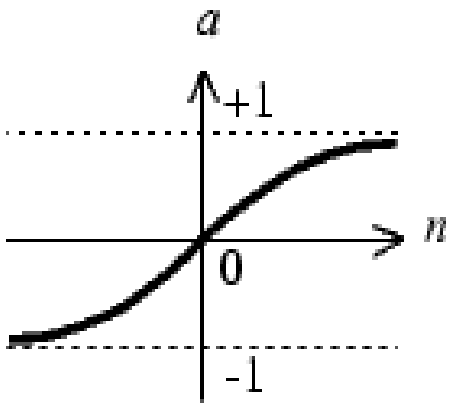


Fig. 4 A sigmoid transfer function used in neural networks: the hyperbolic tangent (Matlab *tansig* function) $\tanh(n)=2(1+e^{-2n})^{-1}-1$. Another one is the *logsig:* $logsig(n)=(1+e^{-n})^{-1}$. *(Adapted from: Matlab package documentation - The Mathworks)*

$$\sum_{i=1}^{n} k_i = 0 \,. \tag{2.3}$$

The enantiomorph choice affects only the sign of the phases, so that the sign of sines will be affected in the same way, while the cosines will be independent from the handedness. In principle we can deal with the enantiomorph choice in a similar way to that for origin definition: we choose an enantiomorph-fixing reflection with index $\bar{h}$ and we multiply all the phase values in a set by (+1) or (-1) in order to make the sign of $\phi_{\bar{h}}$ always positive (or negative) for every set of phases. The same idea can be applied to s.i., which being linear combinations of phases depend on the enantiomorph in the same way. The neural networks used in this work were created with the Matlab Neural Network Toolbox, which contains many predefined functions and allows the creation and training of a neural network object with a few lines of code. Structure factors were always normalized with the Matlab function *premnmx*, to get scaled moduli in the interval $[0,1]$. The output of the network did not need any post-processing since the output layer neurons in all the tested cases had a tangent sigmoid (*tansig*) transfer function (fig. 4), whose output values lie in the interval $[-1,1]$ (as is expected for sine and cosine).

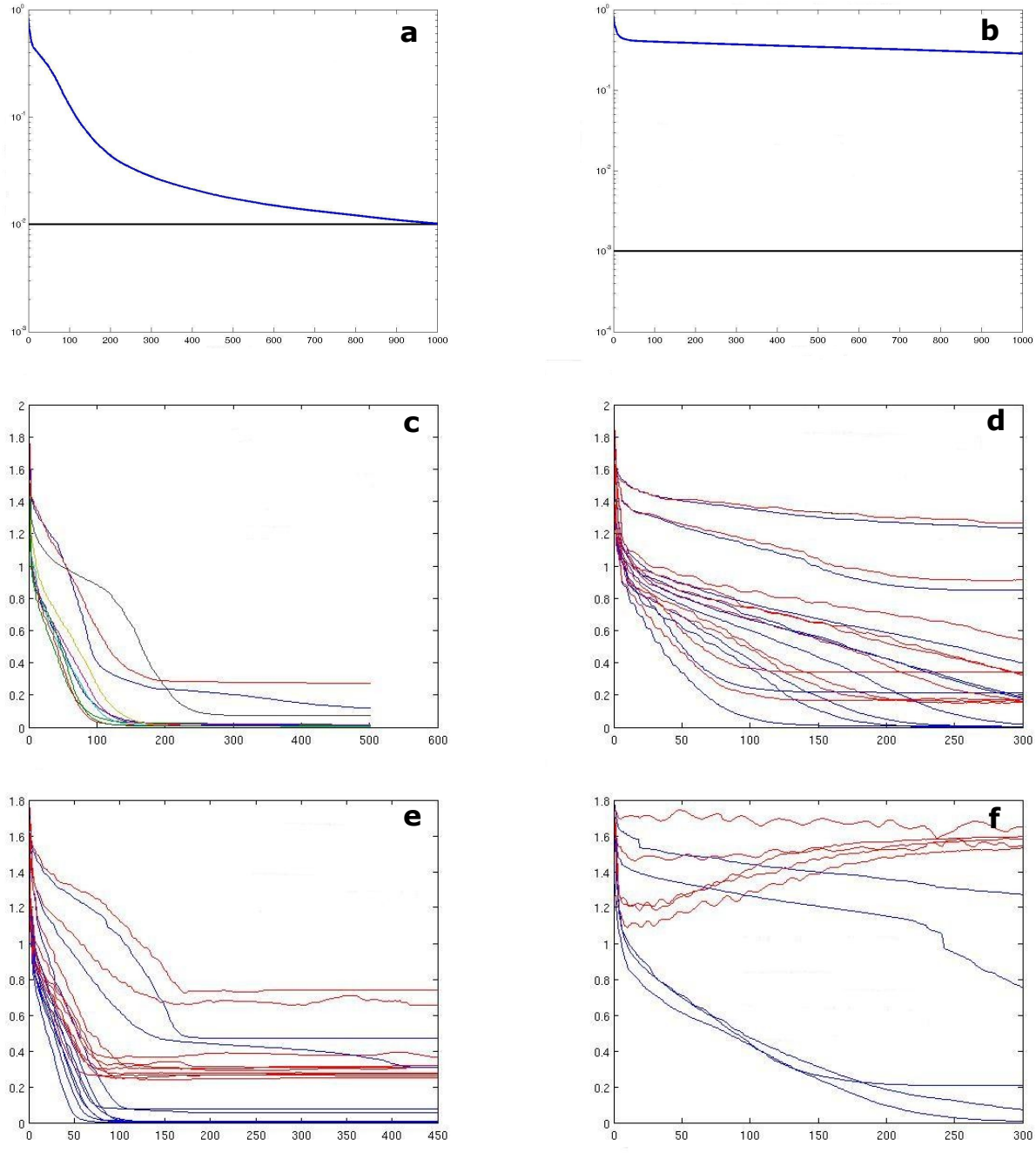


Fig. 6 Plots of mean squared error versus number of epochs for different neural networks and training sets. Error on training set is blue, error on test set is red. Training and test sets size always is of 500 elements each. Number of neurons per layer is preceded by the transfer function specification, T (tansig) or L (logsig). (a) T150-T10 network trained on moduli (h=1-20) and cosines (h=1-10) (4-atom centrosymmetric structures), single run - (b) L50-T100-T10 network trained on moduli (h=1-20) and sines (h=1-10) (4-atom non-c.s.), single run - (c) T15-T100-T10 network trained with moduli (1-10) and cosines (1-10) (4-atom c.s.) - (d) L60-L100-T10 network, moduli (1-10) / cosines (1-10). (4-atom c.s.) - (e) T60-T100-T10 network, moduli (1-10) / cosines (1-10) (4-atom non-c.s.) - (f) L60-T100-T10 network, moduli (1-10) / cosines (1-10). training set: 4-atom c.s., test set: 10-atom c.s.

The chosen architecture was unidirectional (*feedforward*) and the *backpropagation* training algorithm was applied. The synaptic weights are corrected in a backwards process, in order to minimize the mean square error (*mse*) between calculated and expected outputs. The only requirement for the training algorithm is to use differentiable transfer functions; since the phase problem is highly non-linear we used non-linear functions as the tangent sigmoid (which is used very often for is ability to compress any input value in $[-\infty,+\infty]$ into the output domain $[-1,1]$). The learning process was performed in *batch training* mode. In each training cycle (an *epoch*) all the data in the training set are presented to the network, and the corresponding output errors are recorded; weight updating takes place only at the end of each epoch, correcting for the deviations of all the training set elements in a single step process.

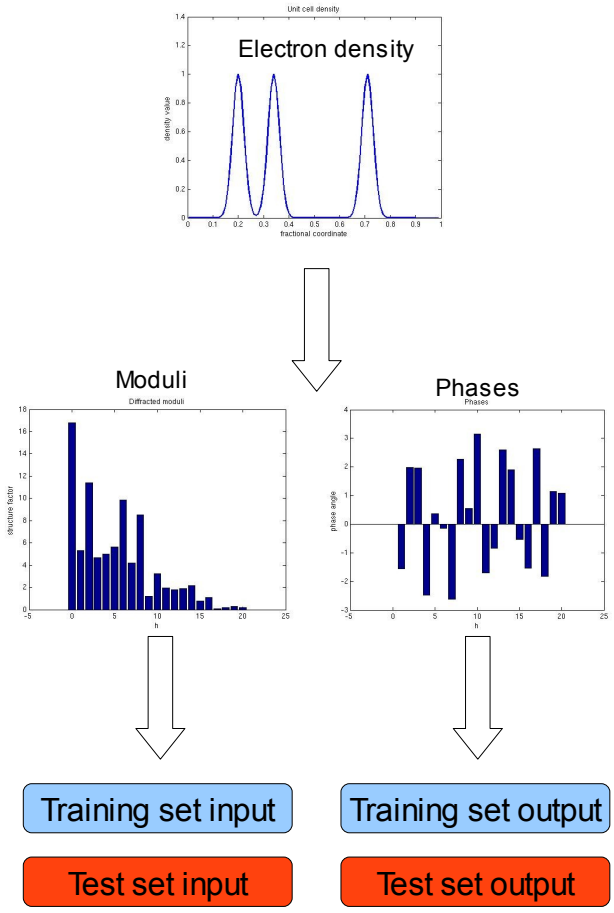


Fig. 5 Scheme of training and test set generation: from a random set of atomic coordinates moduli and phases are calculated and used as input and output vectors.

Two cases were first considered: the centrosymmetric (c.s) one, with 4 atoms per cell (i.e. 2 atoms per asymmetric unit), and the non-centrosymmetric, with 3 atoms per cell. In both cases, training sets composed by 500 elements were used. 1-D structure factors were computed with Shelx (ref.) as the *h00* reflections of a fictitious structure obtained by placing C atoms in the specified random fractional positions (x_i,0,0) in a P1 cell with edges a=20 Å, b=c=10 Å. Initially, no restriction was imposed on the atomic positions, which were randomly generated with uniform probability along the segment. This means that two atoms could happen to be very close, the corresponding peaks in the electron density being completely merged. This obviously violates the atomicity assumption which implies well resolved peaks; we can speak of this relaxed condition on the model as pseudo-atomicity.

Results and discussion

Many tests have been carried out with cosine values (h=1-10) as output. These values were calculated after aligning the phase angles according to (eq. 1.1). The input consisted in the scaled moduli with h=1-10 or h=1-20; the training process was followed by monitoring the mean squared error (m.s.d.) for the training set and for an independent test set of the same size. For the smallest structures (3 atoms non-c.s. and 4 atoms c.s.) the network correctly learns to predict cosine values (m.s.e. after 1000 epochs is as low as 0.01 – fig. 6a). Two layers of neurons (i.e. no hidden layers) are enough to give the network the learning ability; adding a hidden layer does not change much in performance while adding two hidden layers only results in a learning process which is slower and prone to stagnation. In a similar way, increasing the number of neurons in the layers over a certain limit has small or negative effect (when expanding the input layer), as can be predicted on the basis of the overfitting phenomenon. Good results were obtained with a two layer network with respectively 150 and 10 sigmoid neurons in the input and output layers. In analogous experiments carried out with the sines as output no learning was observed, and this is not surprising since the enantiomorph had not been defined (fig. 6b).

Unfortunately, when the number of atoms is increased, the learning is dramatically reduced, even in going just from 4 to 6 atoms (in both c.s. and non-c.s. cases, although in the former only half of of the positions are independent). This phenomenon cannot be due to the increased complexity of the problem, since expanding the network in terms of number of layers and neurons per layer does not modify the situation, nor does the choice of bigger training sets (up to 5000 elements). The observed behaviour can be explained if we assume that increasing the number of atoms the moduli/phases relationships cease to be one-to-one, i.e. different sets of phases can correspond to very similar sets of moduli and generalization is no more possible. In other words, the pseudo-atomicity criterion guarantees a unique solution only for structures made of very few atoms. The results learned for such small structures are not valid for larger ones, as seen from the plot in fig. 6f, where it is evident that a decreasing error for the 4-atom training set does not imply the same for a 10-atom test set.

Furthermore, to prove that lack of learning has nothing to do with convergence problems, a different kind of neural networks was exploited: the *Generalized Regression Neural Networks* (GRNN), a class of architectures commonly used for function approximation.

GRNNs are conceived to perform interpolation tasks: the input layer has as many neurons as are the elements of the target set, and each of these neurons uses a gaussian transfer function centered on one input element of the target set. With GRNNs, it was again observed that actual predictive ability only shows up for 3-4 atoms structures, corroborating the hypothesis of unicity breakdown for larger structures.

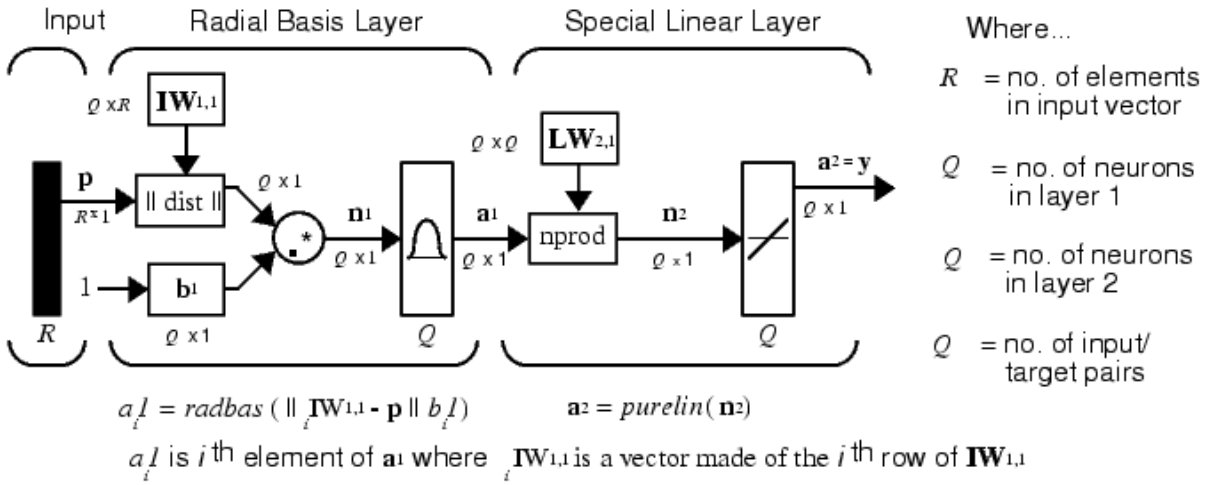


Fig. 7 A *generalized regression neural network* (GRNN). The radial basis layer uses neurons with a gaussian firing function, each reaching the maximum of activation for a given target vector. (*Adapted from: Matlab program documentation, The Mathworks*)

Some additional trials for 2D problems, as well for some chosen phase invariant cosines, did not give positive results. In particular, there is no easy explanation for cosine invariants failure in the same cases where simple (origin-aligned) cosines work.

The results outlined here do not appear to support this simple approach of 'moduli-phases learning'; in addition, the quantity of data involved for protein structures makes it clearly intractable through neural networks. In fact, following the simple moduli/phase cosines scheme outlined here, the input and output vectors should have as many as 10^4-10^5 elements, requiring an impracticable network size. Moreover, taking into account the symmetry and 3D structure of the data are not trivial aspects of the task.

Nevertheless, the idea of exploiting neural networks for protein phasing still remains an

interesting one. The key is to find a suitable representation of the problem and to devise an approach in which the ANNs would do only a part of the work; that is, they could be used to model unknown functions into a well-defined mathematical framework, as well as tools for analising the protein electron density shape and topology. In this perspective, some recent advances in the ANN field allowing dimensionality reduction of multidimensional images (Hinton, 2006) could be useful for the real-space processing of electron density.

Chapter 3

Iterative methods

Chapter 3

Introduction

Many iterative methods exist for non-periodic object reconstruction; from a general point of view, all these methods operate by creating some succession of points in phase (or density) space, that is, in the space where possible solutions are defined. Each point represents a set of phases $\{\phi_h\}$, or, equivalently, the corresponding density function $\rho(\boldsymbol{x})$. Usually, a starting point is chosen at random; the succession is constructed in such a way that, almost for an appreciable percentage of starting points, convergence occurs to the solution. This latter, satisfying all the constraints simultaneously, must lie at the intersection between two constraint subsets: one defined by the experimental moduli and the other determined by a priori constraints (which are often easier to express in real space). The generator of the succession is a map

$$\Gamma : \rho_n \rightarrow \rho_{n+1}\ ^{1}, \tag{3.1}$$

usually devised in such a way that the solution $\hat{\rho}$ is a *fixed point attractor* for the iterations:

$$\Gamma(\hat{\rho}) = \hat{\rho} \tag{3.2}$$

(some cases will be presented in which the attractor is a *limiting cycle* $\Gamma^n(\hat{\rho}) = \hat{\rho}$). A fixed point is left unchanged by an application of the map, so that once the iterations have converged to it, no further evolution occurs. Nevertheless, the existence of fixed points does not suffice *per se* to ensure convergence, so it is not possible to set an upper bound to the number of iterations needed to reach the solution. In this sense, a completely satisfactory phase retrieval algorithm has not been proposed yet.

Given an *N*-point sampling, a generic density is represented by a vector in $\mathbb{R}^N$. If we call C_R and C_{MOD} the two subsets corresponding to the densities consistent respectively with real-space constraints and observed moduli, the solution must belong to their intersection $C^* = C_R \cap C_{MOD}$. In absence of supplementary data, the starting point is a randomly chosen element in C_{MOD}, which is generated simply by fourier transforming the known moduli with

1 Iterating a map gives rise to a memory-less trajectory or *Markov chain*, because only the last point determines the next. It can be argued that, in this way, some useful information from the whole past trajectory remains unexploited.

random phases. A repeated application of the map Γ generates a trajectory in phase space, which in favourable conditions is likely to end in the intersection. When the origin is not fixed from the beginning (for example, by specifying some region in which the object density has known values) the intersection is not represented by a point, but rather by a continuous set of points (filament), since all the possible choices for origin and enantiomorph are equally valid. The trajectory can be thought to evolve in real space (object density) as well in phase space, since for a given set of moduli there is a one-to-one correspondence between points in the two spaces.

Usually, the map used in iterative phasing can be constructed by composing elementary operations known as *vectorial subset projections*. The projection of an element $x \in U$ on a subset of U $Y \subset U$ is written as $\Pi_Y : x \to \{\tilde{y}\}$ and associates to x the set $\{\tilde{y}\}$ of its nearest elements in Y :

$$\Pi_Y(x) = \{\tilde{y} \in Y \ : \ \|\tilde{y} - x\| = \inf_{y \in Y} \|y - x\|\} \tag{3.3}$$

The set $\{\tilde{y}\}$ always contains a single element when the subset Y is convex. A set Y is said to be convex when, for every arbitrary pair of points $x_1, x_2 \in A$, all the points x_μ in the segment

$$\{x_\mu = (1-\mu)x_2 + \mu x_1 \ , \ 0 \le \mu \le 1\} \tag{3.4}$$

also belong to Y. For subsets of the euclidean plane $\mathbb{R}^2$ the meaning is intuitive (see fig. 1). It is easy to show that the subset C_{MOD} is not convex. In fact, given two densities ρ_1, ρ_2 corresponding to the observed moduli $\{F(\boldsymbol{h})\}$ with the phase sets $\{\phi_1\}, \{\phi_2\}$, the densities

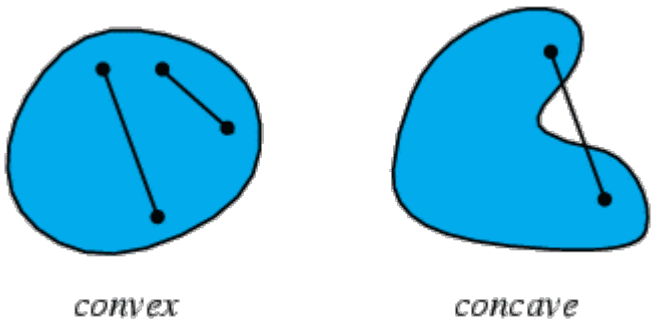


Fig. 1 Convex and concave sets in euclidean plane. Every point lying on the segment drawn between any two points of a convex set belongs to the set itself.

on the segment $\rho_\mu=(1-\mu)\rho_2+\mu\rho_1$ in general will not belong to C_{MOD}, since they will not correspond to the moduli $\{F(\boldsymbol{h})\}$ unless a very special choice for $\{\phi_1\},\{\phi_2\}$ is made. As a consequence, the projection on C_{MOD} is not uniquely defined, since a zero-valued $F(\boldsymbol{h})$ is projected onto the set of points lying on the circle of radius $F^{obs}(\boldsymbol{h})$ (fig. 2). In Fourier space the projection of a generic element of $\{F(\boldsymbol{h})\}$ on C_{MOD} can be written:

$$\Pi_{MOD}\ :\ F(\boldsymbol{h})\rightarrow\begin{cases}F^{obs}(\boldsymbol{h})\dfrac{F(\boldsymbol{h})}{|F(\boldsymbol{h})|} & \text{if } F(\boldsymbol{h})\neq 0\\ F^{obs}(\boldsymbol{h})\ e^{i\psi_{\boldsymbol{h}}} & \text{otherwise}\end{cases} \tag{3.5}$$

where the function $\psi_{\boldsymbol{h}}$ is an arbitrary one. It is common to selects among the many possibilities the projection with $\psi_{\boldsymbol{h}}=0$, which in the following will be called Π_{MOD}.

Another drawback due to non-convexity of the C_{MOD} subset is the presence of *traps* in a sequence of iterated projections (Stark, 1998). Traps are fixed points which do not correspond to an intersection between the subsets. When the map is a simple alternation of projections, $\Gamma=\Pi_1\Pi_2$, and the constraints are non-convex, traps can represent a serious problem. If the trajectory of the representative point gets to a trap, in each successive iteration the density will oscillate between $\rho_1\in C_1$ and $\rho_2\in C_2$, each being the projection of the other, i.e. $\Pi_1(\rho_2)=\rho_1$ and $\Pi_2(\rho_1)=\rho_2$ (fig. 3). This can be viewed as a consequence of the two subset attaining a local minimum of distance; if their boundaries are continuous, the surface of the subset C_1 in ρ_1 and that of subset C_2 in ρ_2 will be parallel. In cases of nearly parallel surfaces the evolution is not completely blocked but becomes very slow; in that case we say the algorithm has entered a *tunnel*. These undesirable phenomena are known as *stagnation*.

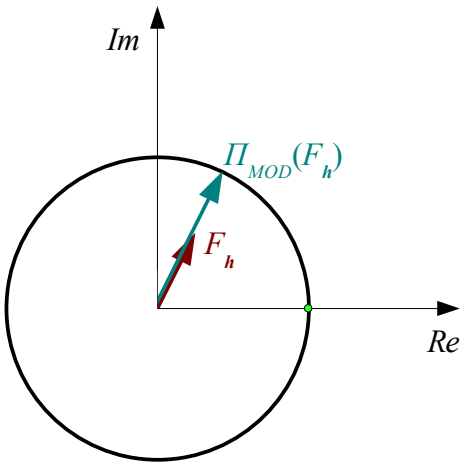


Fig. 2 The Fourier modulus projection represented on the Argand plane. The correct modulus subset is a circle of radius $|F_{\boldsymbol{h}}{}^{obs}|$; a generic $F_{\boldsymbol{h}}$ is projected on it by leaving the phase angle unchanged and substituting the modulus with the correct one. A null vector $F_{\boldsymbol{h}}$=0 would lie at the same distance from any point of the circle, and the arbitrary choice made in defining Π_{MOD} is to project it with zero phase (green point).

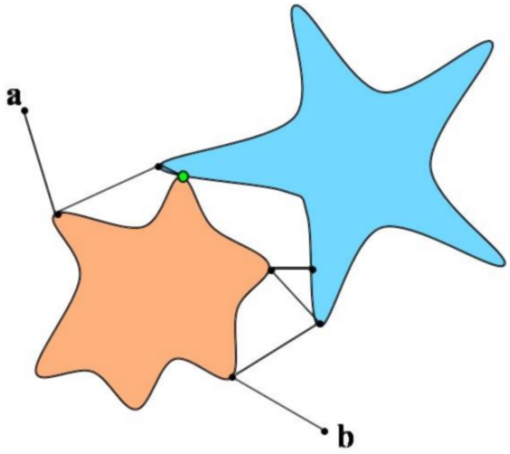


Fig. 3 Two trajectories constructed by alternated projections on non-convex subsets. The succession of points starting from a converges to the intersection, while the one beginning in b ends in a trap. This means that the iterates are being projected back and forth between two points lying at a local minimum of distance between the sets.

An overview of some existing phasing algorithms

Traps and tunnels potentially occur in phasing when using the Gerchberg-Saxton (*GS*) algorithm (Gerchberg and Saxton, 1972), which constitutes the first phase retrieval algorithm ever proposed. The real space constraint allowing image reconstruction is represented by the knowledge of the *object support S,* defined as the region in which the density is expected to be non zero. The *GS* map is simply the repeated projection on support and moduli subsets:

$$\Gamma_{GS} = \Pi_S \Pi_{MOD} \,. \tag{3.6}$$

The projection onto the correct support subset is obtained by simply setting to zero the density values outside the region *S:*

$$\Pi_S : \rho_x \rightarrow \begin{cases} \rho_x & if\ \boldsymbol{x} \in S \\ 0 & if\ \boldsymbol{x} \notin S \end{cases} \tag{3.7}$$

The success of the reconstruction obviously relies on some knowledge about the object size and shape. An upper bound for the support can be inferred from its autocorrelation

function, directly computable from Fourier moduli. In general, for a *N*-point sampling, a necessary (but not sufficient) condition for solution uniqueness is that the sum of the dimensions of the two subspaces must not exceed the dimension of the search space, that is, $dim\, C_S + dim\, C_{MOD} \leq N$, otherwise the intersection cannot be empty. For this to be true, since the subset defined by known moduli has a dimension of N, we must have $dim\, C_S < N/2$, that is, the object must be smaller than half of the image for the problem to be well posed..

The progress of the iterations can be followed by monitoring the summed distance error *J*, which corresponds to the sum of the distances between the current density and its projections on the two subsets:

$$J(\rho) = \left\| \Pi_{MOD}(\rho) - \rho \right\| + \left\| \Pi_{R}(\rho) - \rho \right\| \tag{3.8}$$

Since this quantity can vanish only at the intersection of the subsets, a trap is characterized by the fact that J stabilizes on a non-zero value. A powerful alternative to GS map was introduced by Fienup algorithms (Fienup, 1978), the most effective being the so-called *Hybrid Input-Output* (*HiO*):

$$HiO \;:\; \rho_{\boldsymbol{x}} \rightarrow \begin{cases} \Pi_{MOD}(\rho_{\boldsymbol{x}}) & if\ \boldsymbol{x} \in S \\ \rho_{\boldsymbol{x}} - \beta \Pi_{MOD}(\rho_{\boldsymbol{x}}) & if\ \boldsymbol{x} \notin S \end{cases} \tag{3.9}$$

Density within the support is modified by imposing the observed moduli like in GS algorithm; the difference lies in the outside region, where the density is no more set to zero but rather is to its previous value diminished by the feedback term $\beta \Pi_{MOD}(\rho_{\boldsymbol{x}})$, which increases with the difference between the projected density outside the support and its expected value of zero. When the intersection has been found, the resulting density $\hat{\rho}$ is consistent with the observed moduli and is also zero outside the support, so that no further evolution is observed:

$$\Pi_{MOD}(\hat{\rho}_{\boldsymbol{x}}) = \hat{\rho}_{\boldsymbol{x}} = 0 \quad \forall\, \boldsymbol{x} \notin S \tag{3.10}$$

Compared to GS, the *HiO* algorithm does not suffer from traps, and the convergence is faster. In terms of projections, the *HiO* map can be written as

$$\Gamma_{HiO}=\Pi_S\Pi_{MOD}+(1-\Pi_S)(1-\Pi_{MOD}) \tag{3.11}$$

Recently a general form of map has been proposed (Elser, 2003), the *difference map (DM)*, which avoids stagnation and can be applied to any kind of non-convex constraints. The *HiO* algorithm turns out to be a particular case of *DM* in which the support constraint is used and a given choice of the parameters is made. The *DM* operator is defined by

$$\Gamma_{DM}=1+\beta\Delta\,, \tag{3.12}$$

$$\Delta=\Pi_1 f_2-\Pi_2 f_1\,. \tag{3.13}$$

The operator Γ_{DM} adds to the density a quantity Δ proportional to the difference of two composed maps. Each of these two maps results from the successive application of a map f_i and a projection Π_j on one of the two constraint subsets.

A fixed point $\tilde{\rho}$ of the difference map is characterized by $\Delta=0$, so that

$$\Pi_1 f_2(\tilde{\rho})=\Pi_2 f_1(\tilde{\rho})=\rho_{1\cap 2} \tag{3.14}$$

where the element $\rho_{1\cap 2}$, lying at the intersection between the subsets C_1 and C_2, represents the solution to the phase problem. It should be pointed out that here the solution does not coincide with the fixed point $\tilde{\rho}$. Since in a fixed point Δ must vanish, its norm

$$\varepsilon_i=\|\Delta(\rho_i)\| \tag{3.15}$$

can be used to follow the progress of the iterations.

While the global behaviour of the algorithm is not dependent on the nature of the f_i, a careful choice of them is necessary to allow convergence. Setting for instance $f_1=f_2=1$ (the identity map) does not give attractive fixed points. A possible choice is to construct f_i in a way that its operation on ρ produces a point on the line joining ρ to $\Pi_i(\rho)$:

$$f_i(\rho)=(1+\gamma_i)\Pi_i(\rho)-\gamma_i\rho \tag{3.16}$$

The optimal parameter values are $\gamma_1=-\beta^{-1}$, $\gamma_2=\beta^{-1}$, as found by considering the local

behaviour in the proximity of a fixed point. It can be shown that the difference map can escape traps; these cannot behave like fixed points because they do not allow the quantity Δ to vanish.

The binary approximation

A possibility for restraining the number of solutions is to approximate the electron density in the unit cell to a binary function. This approximation is motivated by the physical reality of separated solvent and protein regions (fig. 4a). The densities of the two zones differ in average value and in variance, both quantities being greater in the protein region. The solvent density can be assumed to be flat to a good approximation, while in the protein region the density can deviate much from its average value (fig. 4b). Numerical tests show that approximating an image with a binary one leads, in Fourier space, to essentially correct phases, while the moduli are more seriously affected. In terms of constraint subsets, the binary densities subset is not expected to intersect the moduli subset, so that an approximate solution would lie between the closest points of the two sets. Moreover, the (euclidean) distance between these two elements of the two sets should be appreciable.

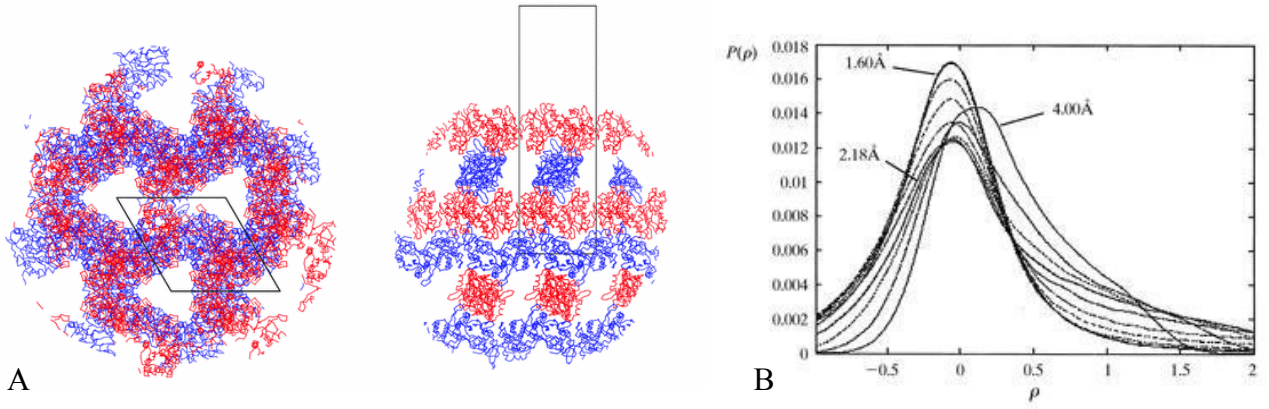


Fig. 4 A) An example of molecular packing in a protein crystal: the trigonal form of rhamnogalacturonan acetylesterase, with a solvent content of 60% (*adapted from: Mølgaard, 2003*). The structure is viewed along two of the crystallographic axes and the unit cell edges are shown in black. The molecules are displayed as Ca traces; the red and blue colors distinguish between the two independent molecules within the asymmetric unit. The solvent channels, accounting for a relevant fraction of the cell volume, are clearly visible as empty spaces. B) Electron density histograms for the protein at different resolutions (*adapted from: Goldstein, 1998*).

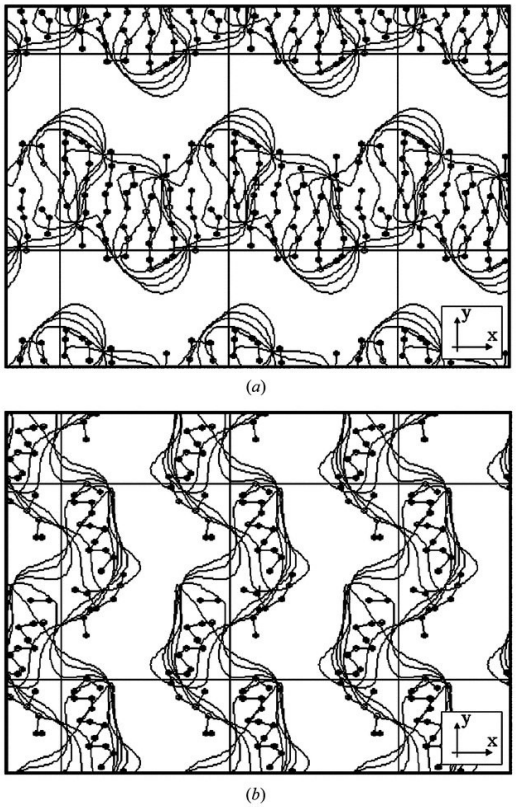


Fig. 5 The BIP phased electron density for Protein G (effective resolution of 12 Å) superimposed with the atomic model *(adapted from: Lunin, 2002)*

Nevertheless, the two-value approximation can be justified to some extent if the resolution is low (> 4 Å). A search for a binary mask (Lunin, 2002) has proven to be successful in reconstructing the density at a resolution of about 12 Å (fig. 5). A Binary Integer Programming (BIP) approach was used in that case, whose main drawback is that the computing time grows exponentially with the complexity of the problem (i.e. with the number of grid points chosen to sample the electron density). In this perspective a more efficient search method, as an iterative one, could perhaps help in extending the resolution limit (at least in the range where the binary approximation is justified). A two-valued function can be scaled to a binary one (having only 0 and 1 as possible values), by shifting and scaling its values. To operate this scaling in Fourier space one needs to know the expected fraction of ones in the unit cell, that is, the volume defined by the molecular envelope which is to be searched for.

Two binary algorithms

The subset of binary densities $C_{01}=\{\rho(\boldsymbol{x})\in\{0,1\}\,\forall\,\boldsymbol{x}\}$ is formed by disjoint points (the corners of an hypercube) and so it is not convex. The projection of ρ on C_{01} is element $\tilde{\rho}_{01}\in C_{01}$ which minimizes the distance

$$\|\rho-\rho_{01}\|=\sum_k(\rho(\boldsymbol{x}_k)-\rho_{01}(\boldsymbol{x}_k))^2 \tag{3.17}$$

and this means that the quantities $|\rho(\boldsymbol{x}_k)-\rho_{01}(\boldsymbol{x}_k)|$ must be minimum for every pixel *k*. This leads to the simple expression for the binary projector:

$$\Pi_{01} : \rho(\boldsymbol{x}) \rightarrow \begin{cases} 0 & : \rho(\boldsymbol{x}) < 1/2 \\ \{0,1\} & : \rho(\boldsymbol{x}) = 1/2 \\ 1 & : \rho(\boldsymbol{x}) > 1/2 \end{cases} \tag{3.18}$$

This projector is not single-valued and some arbitrary choice has to be made about the treatment of densities with value ½ , since they can be indifferently set to 0 or 1.

Here both subsets are non-convex, so that alternate projections will fail. In fact, iteration of a map $\Pi_{01}\Pi_{MOD}$ rapidly gets to a trap, because many different $\rho(\boldsymbol{x})$ possess the same projection. Once $\Pi_{MOD}(\rho_{n+1})$ becomes too close to $\Pi_{MOD}(\rho_n)$ the evolution stops, since Π_{01} projects both of them on the same point of C_{01} .

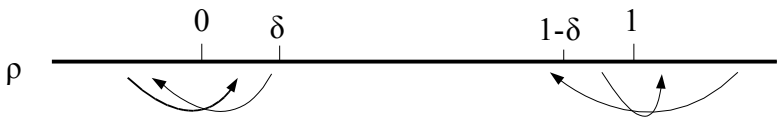
To find a solution to the binary phase problem is thus necessary to avoid that any iteration ρ_n exactly belongs to the subset C_{01} . For this reason in the present work an heuristic algorithm inspired to the *HiO* map, and in particular to the feedback concept, was conceived. It is based on a map Γ_B , consisting in the alternate application of the two operations $\Xi_{MOD}^{(\gamma)}$ and $\Xi_{01}^{(\beta,\delta)}$, each one flipping the density or the moduli about their 'expected values':

$$\Gamma_B = \Xi_{MOD}^{(\gamma)} \, \Xi_{01}^{(\beta,\delta)}. \qquad \beta, \gamma, \delta > 0 \tag{4.19}$$

$$\Xi_{01} : \rho \rightarrow \begin{cases} -\beta\rho & : \rho < \delta \\ 1+\beta(1-\rho) & : \rho > 1-\delta \end{cases} \tag{4.20}$$

$$\begin{aligned} &\Xi_{MOD} = F^{-1} \, \tilde{\Xi}_{MOD} \, F \\ &\tilde{\Xi}_{MOD} : |F_{\boldsymbol{h}}| e^{i\phi_{\boldsymbol{h}}} \rightarrow \left[|F_{\boldsymbol{h}}^{O}| + \gamma(|F_{\boldsymbol{h}}^{O}| - |F_{\boldsymbol{h}}|) \right] e^{i\phi_{\boldsymbol{h}}} \end{aligned} \tag{4.21}$$

The $\Xi_{01}^{(\beta,\delta)}$ operator leaves unchanged the density values falling into the interval $[\delta, 1-\delta]$, while the remaining are flipped about the nearest expected value (0 or 1):



The extent by which each pixel value is flipped is proportional to the parameter β .

A similar operation is carried out in reciprocal space on the values of the moduli by the operator $\Xi_{MOD}^{(\gamma)}$; in that case the expected value of each fourier modulus $|F_{\boldsymbol{h}}|$ is simply the known quantity $|F_{\boldsymbol{h}}^{O}|$ and every modulus is flipped of a quantity proportional to γ. It must be noted that both flipping operations, in real and reciprocal space, are needed for the iterations to converge. Moreover, the previous knowledge of the zero-frequency term $F_{\mathbf{0}}$ (which usually is experimentally unmeasurable) is also necessary, and a special flipping parameter γ_0 was introduced for it. The progress of the iterations can be followed by means of a kind of summed distance error (*SDE*):

$$SDE = N^{-1}\left[\sum_{k=1}^{M}|\Pi_{MOD}(\rho)(\boldsymbol{x}_k)-\rho(\boldsymbol{x}_k)|+\sum_{k=1}^{M}|\Pi_{01}(\rho)(\boldsymbol{x}_k)-\rho(\boldsymbol{x}_k)|\right] \quad [k=pixels] \qquad (3.22)$$

The algorithm was implemented in Fortran 90 for the two-dimensional case, using the static libraries GFT (Chergui, 2002) for FFT computation. Its behaviour has been studied for different values of $\beta, \gamma, \gamma_0, \delta$, in order to identify the set of parameters giving the quickest convergence. Some test results are reported with a 2D trial density (20×20 pixels). In fig. 6 the SDE plots are shown for 20 independent runs of the algorithm (each relates to a different starting set of random phases). In each run, 1000 iterations were performed. Three cases can be identified:

- (a): convergence to the true solution. It occurs suddenly, once the algorithm enters the basin of attraction of the solution after a chaotic trajectory. Very low values of *SDE* are attained (~ 0.01).

- (b): stagnation. At some moment the figure of merit begin to decrease, but slowly sets to a non-zero value (~0.1) because some kind of trap has been entered.

- (c): the trajectory extends over the performed 1000 iterations without entering any basin of attraction.

The dependence of the behaviour on the different parameters can be rationalized as:

- δ affects mostly the speed of convergence, which increases with δ until it rapidly goes to zero above $\delta \approx 0.4$, probably because the basins of attraction of the fixed points become very small.

- β and γ, since they determine the flipping magnitude, influence the ability of the algorithm to 'jump over' local minima (traps). Setting these parameters to small values leads to stagnation, while, at the other extreme, too high values prevent convergence. Since these two quantities play a similar role, they cannot be optimized independently; in fact, for each β value there exists a given range of γ in which convergence is possible.

The situation after choosing the optimal parameters can be seen in fig. 7. Traps are avoided, and at the same time the basin of attraction of the true solution has been enlarged, so that the two unwanted situations (b) and (c) of fig. 6 are both much less probable.

The algorithm does not need knowledge about the support, but only about the fraction κ_1 of non-zero pixels in the solution (which relates to the zero frequency term through $\kappa_1 = F_0 / N$, where N is the number of pixels); the object can appear anywhere in the cell and obviously the two possible enantiomorph choices are equally probable. Since the origin cannot be fixed *a priori*, such a kind of algorithm will always work with a P1 cell, independently from crystallographic symmetry, which cannot be taken into account. Symmetry can only emerge by itself and for this reason it could be used to test the correctness of the solution. For other phase retrieval algorithms without support it has been shown that any attempt to fix the origin reduces very much the algorithm power, probably because of the solution space collapsing to a single point.

An alternative algorithm can be derived as a special case of the difference map $D = 1 + \beta \Delta$ with

$$\Delta = \Pi_{01}[(1+\beta^{-1})\Pi_{MOD} - \beta^{-1}] - \Pi_{MOD}[(1-\beta^{-1})\Pi_{01} + \beta^{-1}] \qquad (3.23)$$

where the binary projector Π_{01} has been defined according to one of the two possible choices in eq. (4.18). An advantage over the binary flipping algorithm is that the zero-frequency term F_0 can be unknown, as it will be found automatically by the algorithm itself; moreover, there is one single parameter to be optimized.

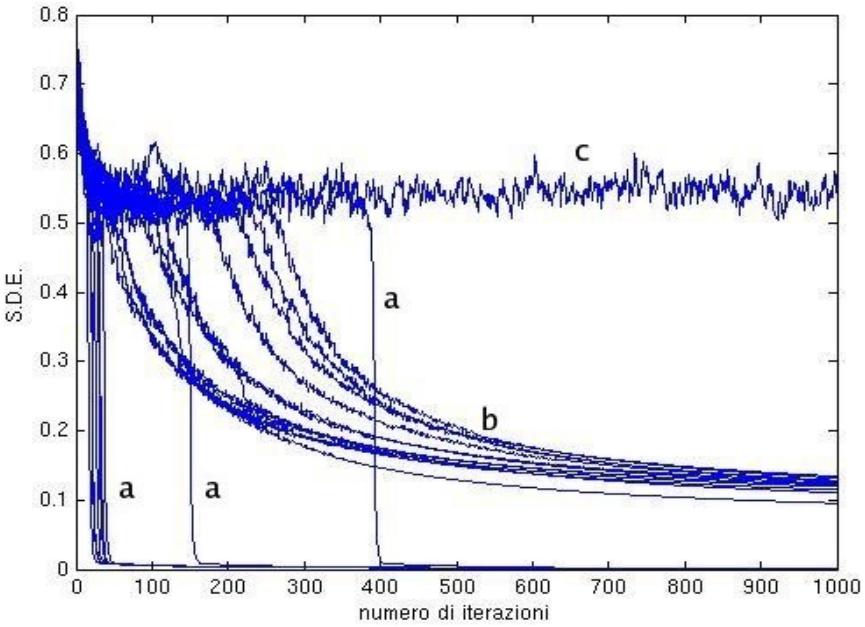


Fig. 6 SDE versus iteration number for non optimized parameters

(β= 0.5, δ=0.2, γ_0=1.2, γ= 1.3) – 20 runs

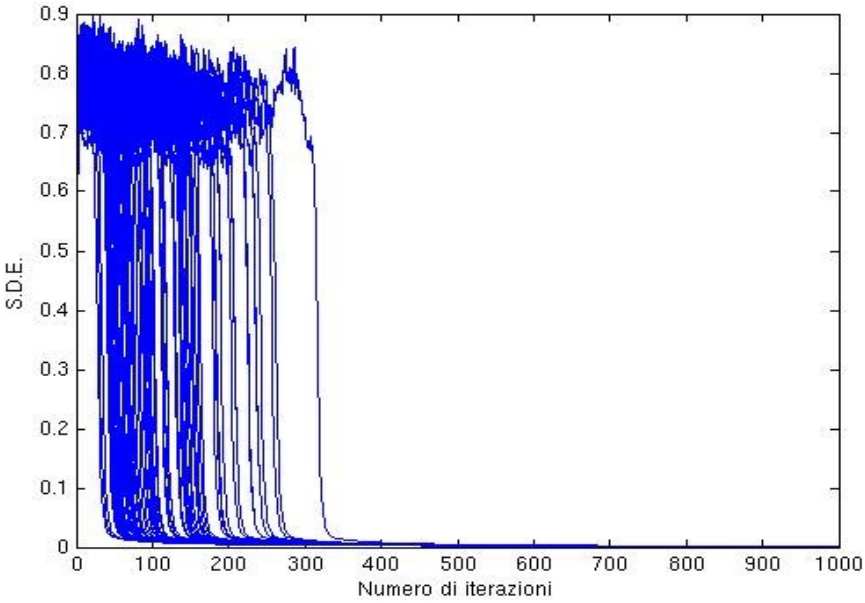


Fig. 7 SDE for optimized parameters (β= 0.5, δ= 0.2, γ_0= 1.2, γ= 1.6) – 100 runs

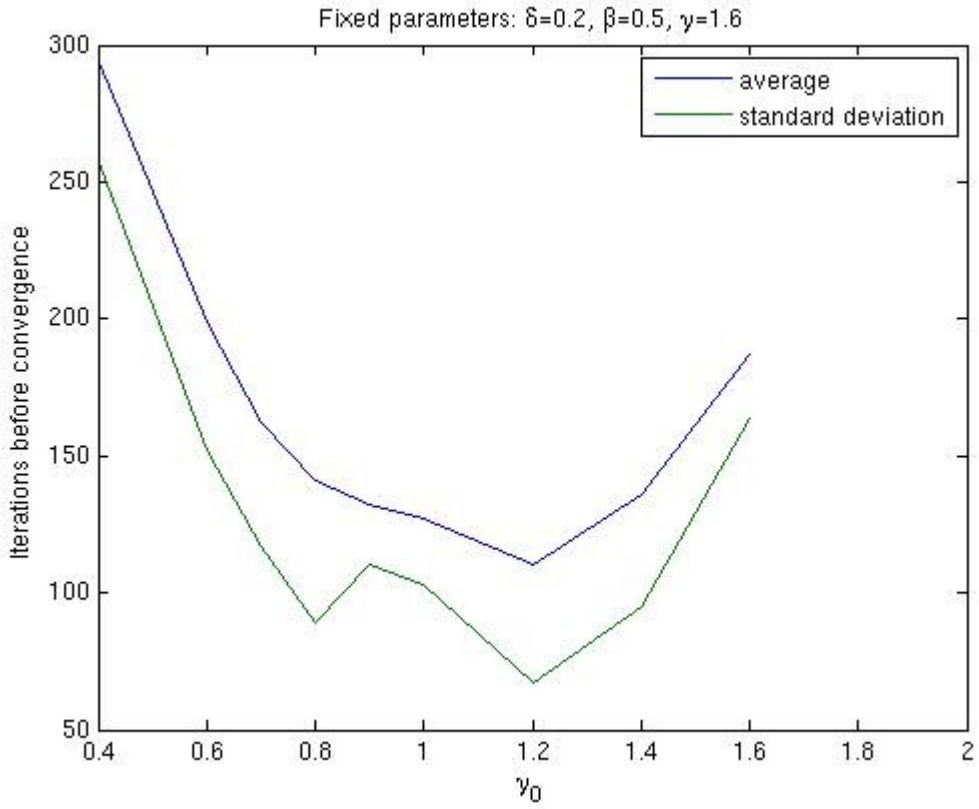


Fig. 8 Optimization plot for the parameter γ_0 for fixed values of the other three parameters

Various experiments have been conducted with different trial densities to determine the influence of β on the speed of convergence and to compare the behaviour of the two algorithms. Two different optimal ranges of β have been found, one centered about -1 and the other about 0.8 (fig. 11). This is in agreement with the literature (Elser, 2003), where the optimum values for the β parameter are found to be close to ± 1 . The comparison between binary flipping and difference map shows that their efficience depends greatly on the features of the object to be reconstructed, but the dependence differs from one algorithm to the other. The two methods are, at some degree, complementary; putting aside very simple cases, often one of the two appears to perform well in those situations where the other exhibits a very slow convergence (fig. 12).

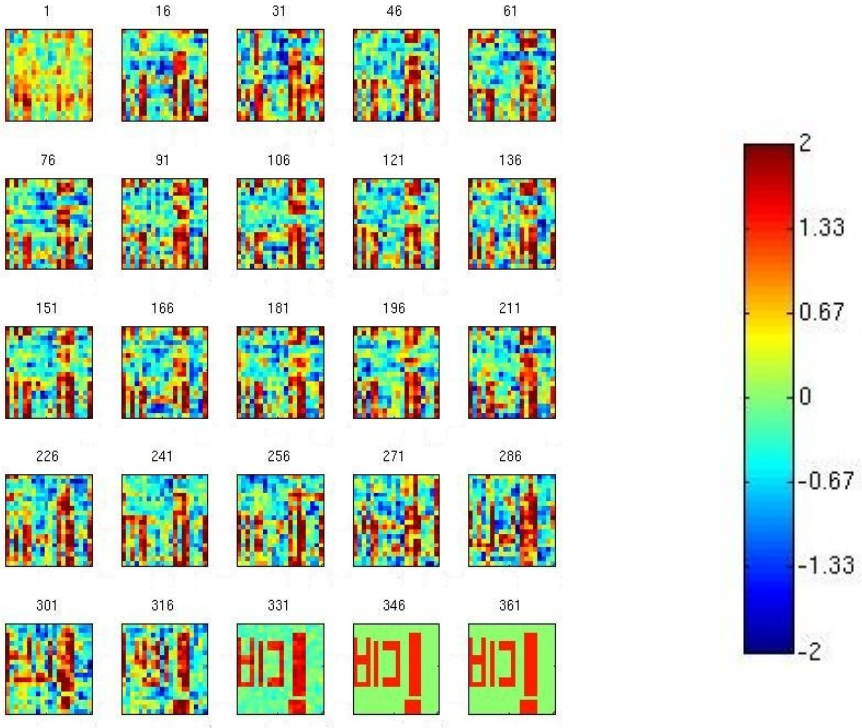


Fig. 9 Snapshots of the density during its evolution, taken every 15 iterations. The abrupt change (fig. 6, case (a)) in the figure of merit (SDE) occurs near cycle 320, when the density suddenly begins to converge to the correct (binary) one.

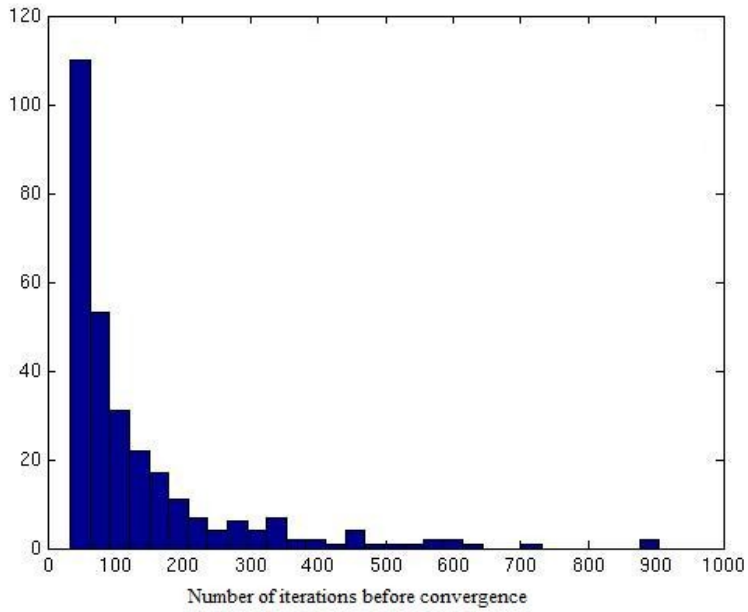


Fig. 10 An histogram showing the distribution of number of iterations needed for convergence (a sort of trajectory lenght) for the binary flipping algorithm. The distribution has an approximately exponential decay, indicative of a memory-less process.

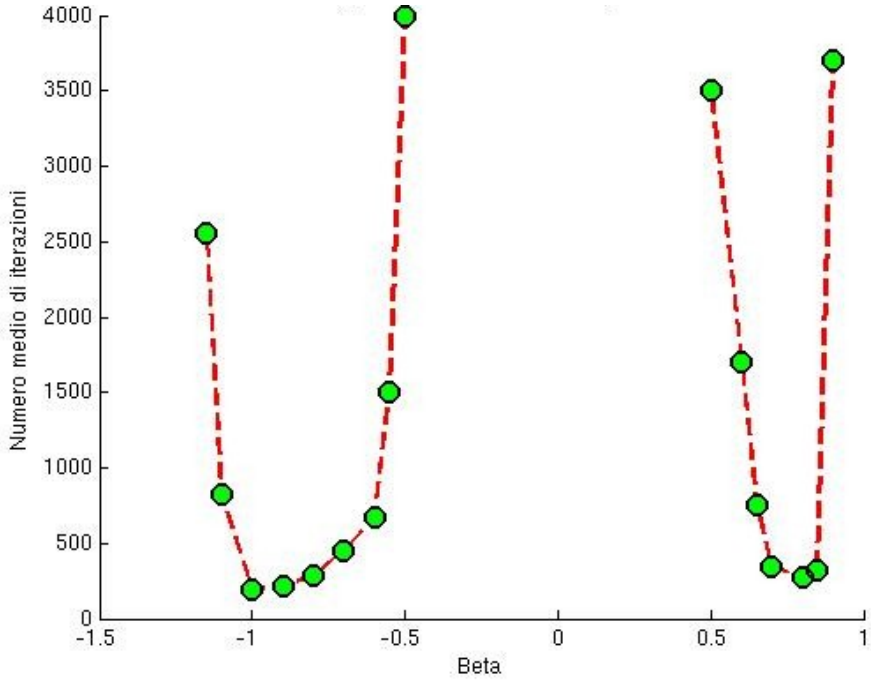


Fig. 11 Optimization plot for the binary difference map. The average number of iterations needed for convergence is shown as function of the single parameter β. Two optimal ranges are found, the first (centered on β=-1, the global minimum) being larger and deeper.

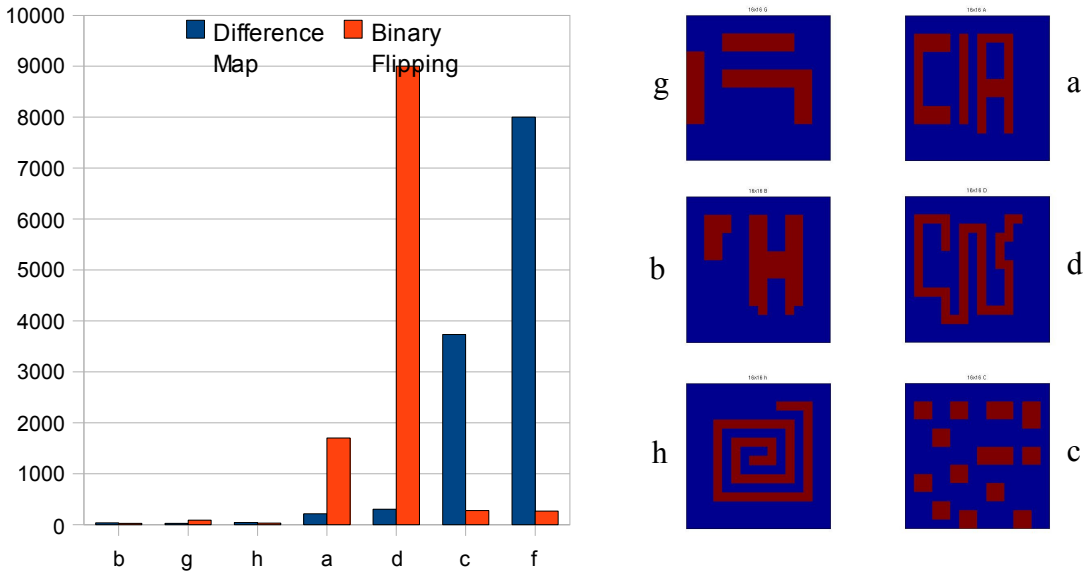


Fig. 12 Comparison between the performance of the two algorithms, expressed as average number of iterations before convergence, on different two-dimensional binary masks (20×20 pixels). Putting aside very simple cases (b,g,h), which are solved in less than 100 iterations, there is a considerable difference in convergence rate.

Binary approximations and real cases

Once established that a method existed to solve the binary problem, a more realistic case was considered, consisting in pseudo-molecular data in two dimensions. The moduli were obtained by fourier transforming the density of benzene molecules projected onto the molecular plane. The cell was a square of 10 Å edge in which one, two or four benzene molecules had been placed. Data were used up to a resolution of 2 Å.

A binary approximation to the real density can be constructed by scaling the density and then setting a threshold *z*. The points with values higher than *z* are given the new value of 1 and the others of 0.

$$\rho_{01}(\boldsymbol{x})=\begin{cases}1 & :\rho(\boldsymbol{x})\geq z\\ 0 & :\rho(\boldsymbol{x})<z\end{cases} \qquad F\,(\rho_{01}(\boldsymbol{x}))=F_{01}(\boldsymbol{h}) \tag{3.24}$$

The structure factors corresponding to the binary density, $F_{01}(\boldsymbol{h})$, can be assumed proportional to the true ones, as in (Lunin, 2002):

$$F_{01}(\boldsymbol{h})\approx kF(\boldsymbol{h}) \tag{3.25}$$

where the constant *k* can be calculated from the knowledge of the fraction κ_1 of non-zero pixels in ρ_{01} :

$$k=\left[\frac{\kappa_1-\kappa_1^2}{\sum_{\boldsymbol{h}\neq 0}|F(\boldsymbol{h})|^2}\right]^{1/2}, \qquad \kappa_1=\frac{(\sum_i \rho_{01}(\boldsymbol{x}_i))}{N} \tag{3.26}$$

The data from molecular structures were scaled in this way and then given as input to the binary flipping algorithm. No convergence was observed, for none of the β, γ, δ parameter sets that had worked better for the ideal binary cases. This can be explained assuming that there is no intersection between the two constraint subsets, that is, no binary density exists that could reproduce the non-binary moduli. In fact, binarization of a density not only will affect the moduli in the chosen resolution sphere (in 2D, a circle), but it will create non-zero frequency components outside the sphere (where the original moduli had been set to be zero). To allow the two subsets to intersect in some point, out-of-sphere moduli should be

allowed to deviate to some extent from their expected value of zero; it is not clear, however, if any physically meaningful solution could be found in this way.

A simplified Sayre equation for binary images

Another possibility for phasing diffraction data from a binary object can be derived outside the iterative methods context, taking inspiration from the Sayre equation (Lunin, 1985). While this relationship has been derived to exploit the atomicity property, it can be shown that it holds, in a simplified form, for binary densities too. In fact, the Sayre equation presupposes that density and squared density are related by convolution with a spread function *g*:

$$\rho = g * \rho^2 \tag{3.27}$$

This is true for a density made of identical, well resolved, spherical peaks (equal atom structure); nevertheless, it is also consistent with a binary function, in which case *g* reduces to a constant. Assuming the density can take only the values 0 or *a*, we have

$$a\,\rho = \rho^2 \tag{3.28}$$

which in reciprocal space is equivalent to:

$$\boldsymbol{F}_h = (aV)^{-1} \sum_k \boldsymbol{F}_k \boldsymbol{F}_{h-k} \tag{3.29}$$

where *V* is the unit cell volume (in the 3D case). This convolution relationship would allow the solution search to be carried out entirely in reciprocal space, borrowing a variety of existing algorithms from the field of direct methods. Moreover, a binary approximation to a non-binary object can be found by minimizing the deviation between the two sides of the equation, while iterative algorithms fail in this task. As seen before, from the lack of intersection between the constraint subsets follows that only a global minimum of the distance between the subsets can be searched. But this minimum is not qualitatively different from those non-meaningful local minima (traps) that a good algorithm is expected to avoid.

Modifications of the charge flipping algorithm

A possible criticism to the application of the binary flipping approach to non-binary density is that, while the lowest density region (corresponding to solvent in protein structures and to vacuum in small molecule structures) can be effectively assumed to be sharply distributed around zero, the object (molecular) density has a broader distribution. The behaviour of the algorithm becomes more interesting after suppression of the flipping about the upper value of 1, we let β tend to 1 and γ to 0, and give F_0 the freedom to vary during the iterations: the density of a single benzene ring in the cell could be slowly reconstructed. With these modifications, the algorithm reduces to the known method of *charge flipping* (Oszlànyi, 2003), which alternates moduli projection to a change in sign of low-valued density:

$$\Gamma_{CF} = \Pi_{MOD}\ \Xi_0^{\delta}, \tag{3.30}$$

$$\Xi_0^{\delta} : \rho \rightarrow \begin{cases} \rho & : \rho \geq \delta \\ -\rho & : \rho < \delta \end{cases} \qquad (\delta > 0) \tag{3.31}$$

In term of projections, the flip operator can be written

$$\Xi_0^{\delta} = 2\Pi_{S(\delta)} - 1 \tag{3.32}$$

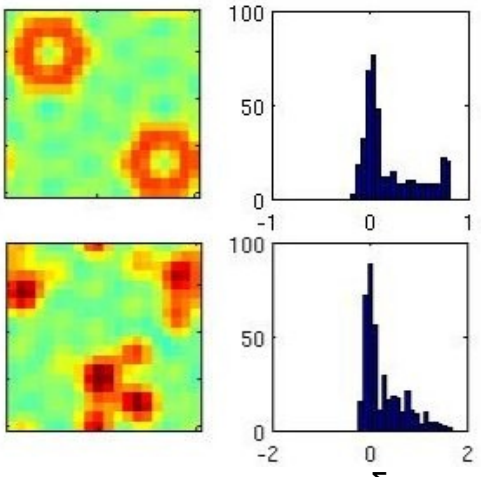


Fig. 13 Test of the *CF* algorithm (δ=0.2) on a single benzene molecule projection (2 Å resolution). Upper plots: density and its histogram for the true map. Lower plots: same for the reconstructed map

where $\Pi_{S(\delta)}$ stands for support projection. The important thing is that the support $S(\delta)$ is a dynamic one, being updated at each iteration by selecting the points with $\rho \geq \delta$. The *CF* algorithm has been proposed in crystallography for reconstructing atomic (<1.2 Å) resolution structures, but it has been shown to be also applicable to the phase retrieval of non-periodic objects that lack atomicity. In both cases, however, the uniqueness of solution is

guaranteed by the presence of extended regions of density with near-zero values and by (not strict) positivity. For non-atomic objects the algorithm tends more to stagnation, so that it has been used in conjunction with the *HiO* map: *CF* provides support evolution, while *HiO* drives to convergence because it is insensitive to traps.

The 2D benzene ring at 2 Å resolution does not display atomicity, but the presence of a vast majority of pixels with small absolute values of density still causes the solution to be unique. Because of the lack of atomicity sudden convergence is never observed; what happens is instead a slow, gradual approach to the solution. This good behaviour is compromised in going from one molecule to two and four molecules per cell, because the ratio of null pixels to the total number of pixels decreases. With two molecules, although the null pixels still occupy more than half of the cell, the algorithm fails to reconstruct the rings, whose density is rather flat, and shows a preference for 'peaky' solutions with higher variance (fig. 13).

The only way to find a solution with the required characteristics is to introduce new restraints; for example, an upper limit to density values can be used to force density flatness. A choice that proved to be effective is to set a proportionality constant α between average density (calculated with the values above the flipping threshold δ) and the maximum allowed density *s*; at each *CF* cycle, the density values are modified by inversion about the expected maximum (*plateau*) value.

$$\rho \to s - \eta(\rho - s)\,, \qquad s = \alpha \langle \rho \rangle_{\rho > \delta} \tag{3.33}$$

The value of *s* is calculated at each cycle. With this additional restraint, correct solutions could be found for the cases of 2 and 4 molecules per cell (fig. 14). The best values for the parameters were $\alpha \approx 1.3$, $\eta \approx 2$; the first one depends on the expected maximum value for the density and can be varied only in a very narrow range if wrong solutions are to be avoided.

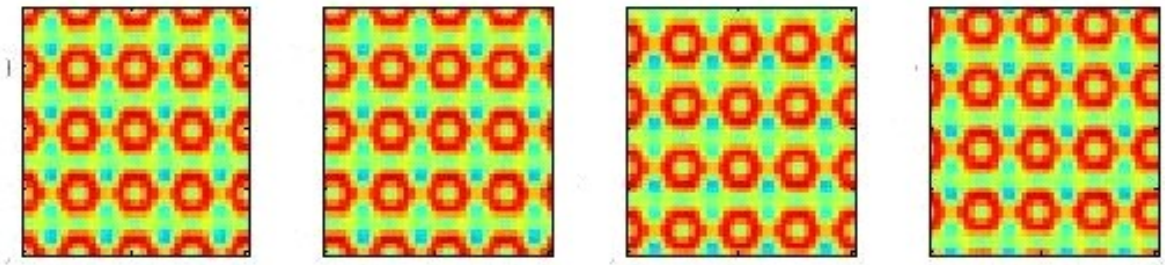


Fig. 14 Densities reconstructed in four different runs of the upper-bounded *CF* algorithm. Four molecules per cell are present. Four unit cells are shown for clarity – note the different origin positions, which depend on the (random) starting point.

A 3D case was then considered, to test if this modified *CF* algorithm with upper bound restraint could phase bigger structures. Synthetic trial data were calculated with the software SHELX (Sheldrick, 1996) from the PDB coordinates of one molecule of *Fatty Acid Binding Protein* (*FABP,* PDB code 2HMB - Zanotti, 1992). This protein consists in 131 aminoacids, organized in a β structure which defines an internal cavity. The reflections were computed from a single molecule positioned in a P1 cell (for simplicity, a=b=c, α=β=γ=90° were chosen). Since zero density zones (which can be identified here with the solvent regions) define the degree of determinacy of the problem, different tests have been carried out varying the lenght of the cell edge, that is, the unit cell volume. The effect of data resolution was also investigated, across the range 20÷2.5 Å.

It has been found that setting an upper bound for the density has no or little effect on converging to the correct solution, which could be retrieved in a small percentage of runs only when the solvent content is very high (at least 85% of the unit cell volume, far too high to be found in any real crystal). This probably means that, under a given fraction of null pixels, the correct solution ceases to be a strong attractor for the *CF* algorithm, and this happens well before the problem become underdetermined.

In fact it was noted that, even starting from the correct phases, there is a tendency to escape from the correct solution; the rate of this process increases with the flipping threshold δ. This can be seen in fig. 15, where the correlation between final and starting (true) density (computed with 4 Å resolution data) has been reported as a function of δ, for a fixed number of iterations. Each line relates to a different solvent content. It can be noticed that there is a change in trend between 77% and 66% of solvent: the 66% solvent density gets worse as δ increases.

Another modification of the *CF* algorithm was tested in the perspective of phase extension of protein diffraction data. It consists in imposing on the electron density a topological restraint motivated by very general features of protein structures. A key process consist in dividing the image in the connected components, i.e. separated features appearing in density when the isosurface for a given cutoff value is constructed. For a given threshold κ a mask Ω_κ is defined as

$$\Omega_\kappa=\{\boldsymbol{x} \,:\, \rho(\boldsymbol{x})\geq\kappa\}\,; \tag{3.34}$$

the set of points Ω_κ can be decomposed in a number *M* of *connected components* $\omega_\kappa^{(i)}$, each with a given volume $v_k^{(i)}$. A subset of points $\omega\subset\Omega$ is said to be a connected component when every pair of points $\{\boldsymbol{x}_1,\boldsymbol{x}_2\}\in\omega$ can be joined by some curve entirely contained in ω.

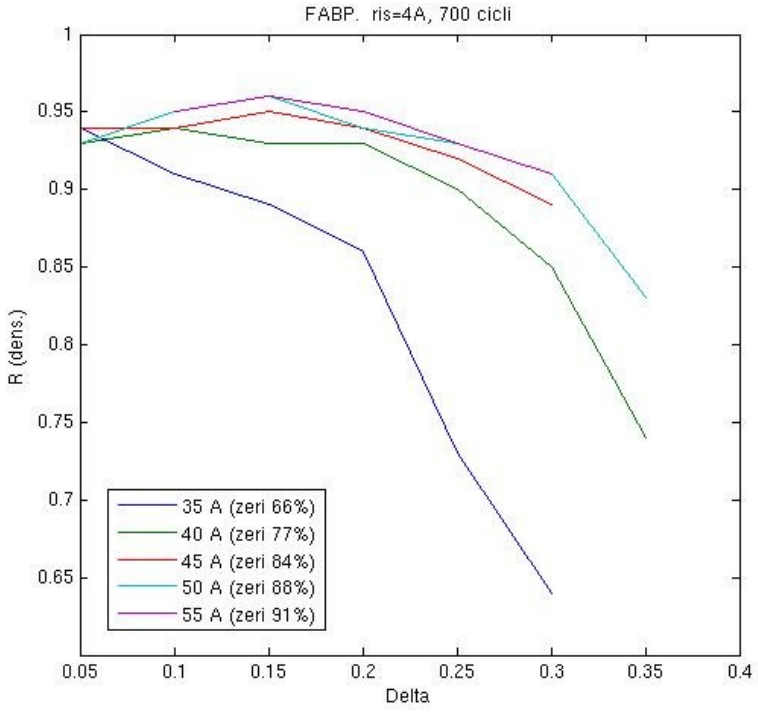


Fig. 15 Correlation between final and starting density maps for different values of solvent content and δ parameter. In each case the true density at a resolution of 4 Å was used as starting point for 700 cycles of CF algorithm.

While connected component analysis identifies volume segments, without saying nothing about their shape, we can define a useful quantity for estimating the linear lenght of density pieces. This topological property, named *connectivity*, is computed by tracing the *skeleton*,

that is, the set of lines joining all neighboring points above a given threshold (Baker, 1993a). With density defined on a grid, the procedure is to select grid points having density greater than 1.4 standard deviations above the mean, and connect by edges the points which are nearest neighbors. Two grid points belong to the same graph if they are connected by a continuous set of edges.

By means of the skeleton we define the connectivity as:

$$\text{Connectivity} = \frac{\text{number of points in longest graph}}{\text{total number of points in graphs}} . \tag{3.35}$$

(An alternative definition for connectivity is the total number of graphs).
This quantity is obviously a function of the threshold and the phase set. If the threshold is appropriately chosen, the global maximum for connectivity should coincide with the true phases, for which the electron density shows a single continuous polypeptide chain. Connectivity values relative to random phase sets are smaller than 0.1 while for correct phases a value above 0.9 is expected. It has been shown that the addition of an increasing phase error to the correct phase set always decreases the connectivity in a gradual way.

The connectivity restraint could be exploited into an iterative algorithm by selectively eliminating the densities that belong to the shortest graphs. Although it is impossible to know if those small segments would result to be correctly placed in the final density, one surely knows that correct density should not show small, isolated blobs. So the idea is to force the density to evolve by growth of the longest fragments rather than by fusion of many small segments. An encouraging observation is that connectivity only depends on strong reflections and it is preserved even if a consistent fraction of moduli are given completely random phases (up to 80% of the weakest ones – test carried out at 4 Å resolution).

An implementation was tried in this work using a weaker topological constraint, based on the segment volume rather than graph lenght. The volume constraint is expected to be weaker than connectivity (as defined by Baker *et al.*) because it involves no restriction on the shape of the density; there is no reason to think that a general relationship between the volume of a connected component and its skeleton lenght should exist. However, for a densities in a neighborhood of the solution (so that the phase error is acceptable and connectivity is not too low) some kind of local relationship should arise, since the longest

elements will also be the largest ones. For that reason, one expects that the requirement for the density to display a minimum number of volume elements $\omega_{\kappa}^{(i)}$ could be used to improve or extend a set of known phases. Thus, a modified *CF* algorithm was devised, introducing supplementary real space operations:

- a binary mask is created to distinguish between points above and below a fixed threshold.
- a segmentation algorithm is used to identify the connected components into the density;
- a sorted list of segments is created on the basis of their volume (number of pixels);
- segments with volume below a certain minimum value v_{min} are set to zero in the density map.

The segmentation method used here was essentially the 'burning grass' algorithm described by (Lunina, 2003) and consisting in the following steps (fig. 16):

- *Initialization*: the points above the threshold are given a value 1, the others 0. No found components are present.

- *Search for a new component*: the nodes of the grid are scanned until a node with value '1' is found. The number of found components is increased by one. A 'current front' is defined as a set consisting of this node only. The new found component is marked with a consecutive number *m*. If no more '1' nodes are present the algorithm stops.

- *Isolation of a connected component*: the 'future front' is defined as the set of the nodes with value 1 that are neighbouring to one of the nodes of the 'current front';

- *Propagation of the front*: the nodes of the current front are marked as belonging to the *m*-th component. The 'future front' becomes the 'current front' and the algorithm goes back to the preceding point. This loop is repeated until the 'future front' is empty, then a new component search is performed.

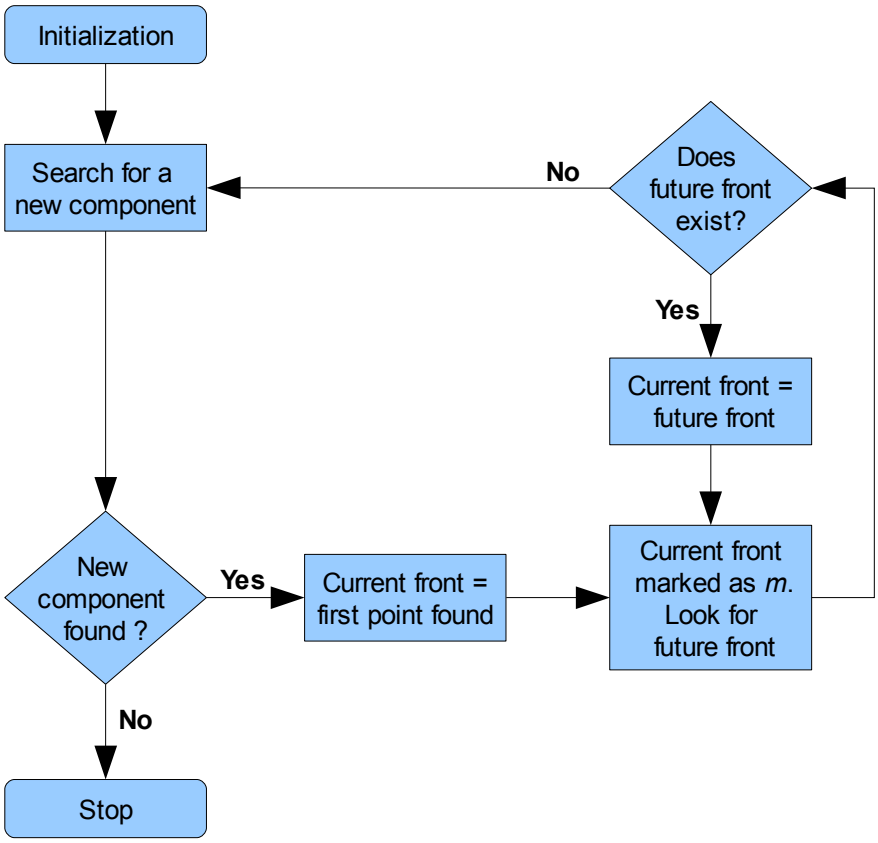


Fig. 16 Flow chart for the 'burning grass' segmentation algorithm. Each *m*-th time a new initial point is found, the propagation loop is entered. The loop defines a 'future front' as the list of those points which are nearest-neighbors to points of the 'current front'; these latter are then marked as belonging to the *m*-th segment and the procedure is repeated until no more nearest neighbors are found and all the *m*-th connected component has been isolated.

The modified *CF* algorithm has shown some phase extension power in a series of error-free tests conducted with a starting set of exact phases (fig. 17) that were extended to cover a larger sphere of reflections. Phases not belonging to the starting set were initially given random values, while known phases were kept constant at each run. It must be noted, however, that the algorithm is not able to improve a set of error-affected phases if these are given the freedom to vary from one cycle to the other. In fact, a divergent behaviour was always observed in that case, probably due to overdeterminacy.

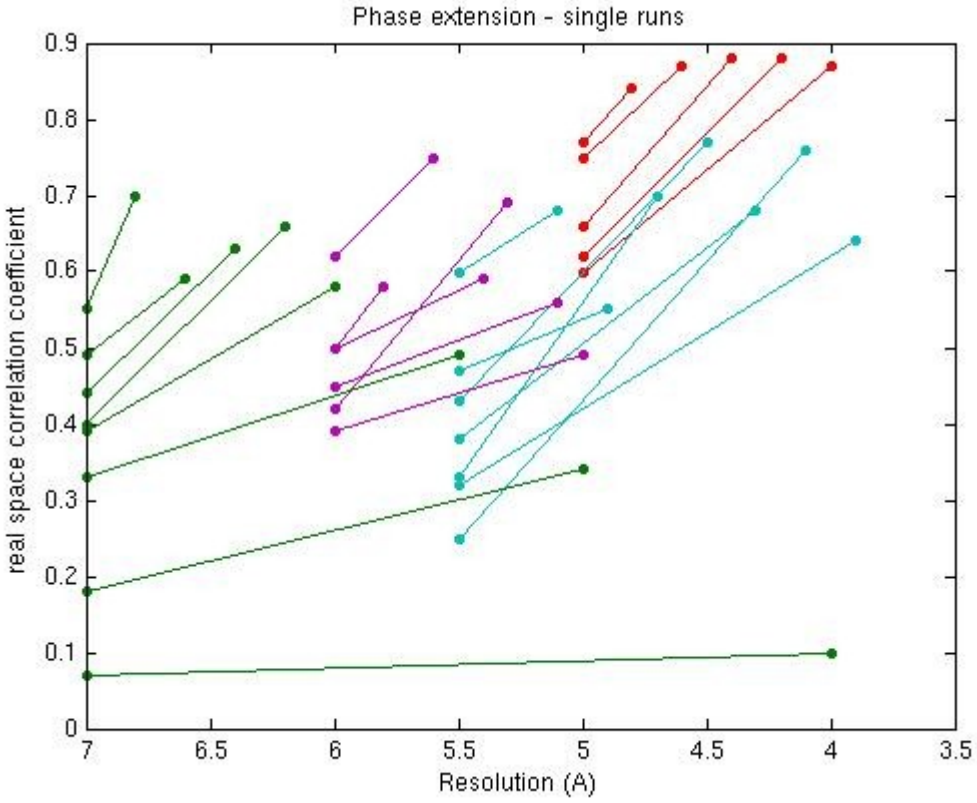


Fig. 17 Starting and final correlation coefficient for some runs of the connectivity-restrained *CF* performed on ideal data from the protein FABP (reference!) with an exact starting set. The map correlation coefficient has been reported as function of starting and final resolution. Several hundreds of cycles were carried out, but in many cases the density ceased to evolve after only 50÷100 iterations.

Chapter 4

Patterson function and secondary structure

Introduction

The last part of this work focuses on the correlations between secondary structure and Patterson map. While not offering a strategy for the *ab initio* solution of the phase problem, analysis of these correlations could provide a useful crystallographic tool, allowing detection and orientation of secondary structure elements in the cell without a prior knowledge of phases.

In these last years, the relationship between the angular distribution of the diffracted intensities and the percentage of alpha and beta structure in the crystal has been investigated (Morris, 2004); the origin of this connection must be found in the characteristic interatomic distances and periodic arrangements occurring in protein structures and closely related to secondary structure elements. The diffracted intensities provide information about the spatial frequencies of electron density; in angular intensity distribution this information is averaged over all orientations, so that on this basis only global information, as the fraction of α and β structure, can be extracted. In this work, an attempt was made to explicitly analyse the frequency content in each spatial direction, and to put it in relation with the presence of alpha helices, beta strands and beta sheets.

The reciprocal space point of view has been abandoned and the study has been carried out on the Fourier transform of the intensities, the well known Patterson function. This represents the self-correlation of electron density and can be more easily interpreted in terms of real space properties. Analysis of the Patterson map has shown that alpha helices give rise to strong, recognizable features in the direction of their axis. More difficult has proven to detect single beta strands, while it is possible to derive some indications about whole beta sheets.

Secondary structure and diffracted intensities

Recently it has been pointed out how the different layers of structural organization in protein crystals do reflect in the intensity distribution (Morris 2004). The quantity investigated was the square of the normalized structure factor

$$|E(\boldsymbol{h})|^2 = \frac{|F_{obs}(\boldsymbol{h})|^2}{\langle |F(\boldsymbol{h})|^2 \rangle} \tag{4.1}$$

where $\langle|F(\boldsymbol{h})|^2\rangle$ is the expectation value of the computed structure factor, which is estimated on the basis of the knowledge about the structure. In the absence of any structural information, the atomic positions are assumed to be random variables, uniformly and independently distributed. The normalized structure factors are dimensionless quantities and are suitable to characterize the atomic arrangements into the crystal since they are as independent as possible from the chemical identity of the atoms. In fact, if the assumption of random atomic positions was true, $|E(\boldsymbol{h})|$ would coincide with the Fourier modulus of a point-atom structure, and its square would have a constant expectation value: $\langle|E(\boldsymbol{h})|^2\rangle=1$.

In real crystals, atomic positions are correlated because of chemical bonding, and their distribution can deviate very much from a uniform one. This is especially true for protein crystals, which show different levels of atomic organization: aminoacid structure, secondary structure elements, protein fold and molecular packing. As a consequence, the normalized intensity distribution $\langle|E|^2\rangle(d^*)$ shows a series of peaks which are connected with the relative abundance of some interatomic distances in the structure. In fact, for an equal atom structure, the following relationship holds:

$$I(d^*)=N\,f(d^*)^2\left[1+\frac{1}{N}\int p(r)\,\mathrm{sinc}(2\pi d^* r)\,dr\right] \tag{4.2}$$

where $f(d^*)$ is the atomic scattering factor, N is the number of atoms and $p(r)$ is the radial pair distribution function. This formula can be derived by rotationally averaging the expression for $I(\boldsymbol{h})$ in terms of atomic positions; the quantity $p(r)\,dr$ gives the number of atoms with a separation distance within (r, $r + dr$), which contributes to the intensity through a sinc function[1]. In particular, secondary structure is responsible for characteristic atomic distances in the range 4.5-7 Å, appearing like a series of peaks in the pair distribution function (located at 4.5, 4.9, 5.4, 6.2, 7.3 Å for alpha structure, at 4.8, 6.1, 6.6, 6.9, 7.6 Å for beta).

In reciprocal space, a peak around 4-5 Å resolution[2] (fig. 1) in the $\langle|E|^2\rangle(d^*)$ distribution is always present in the secondary structure. Such a peak could not be reproduced by

1 The sinc function is defined as $\mathrm{sinc}(x)=\frac{\sin x}{x}$.

2 The resolution $d=(d^*)^{-1}$ is used rather than d^* because it represents real space distances.

artificial structures unless the typical Ramachandran angles for α and β structure are forced.

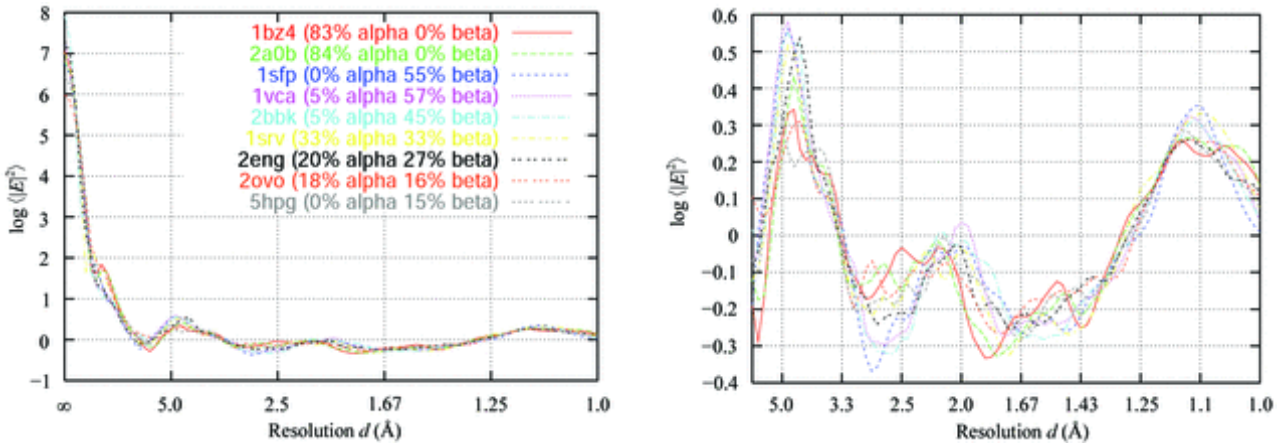


Fig. 1 $\log\langle|E|^2\rangle(d)$ plots for some protein structures. (the intensity has been reported in function of the resolution $d=(d^*)^{-1}$ for a better visualization). A) The $\infty \div 1$ resolution range. Curves from different proteins look very similar. B) details of the same curves in the resolution range $1 \le d_{min} \le 6$.

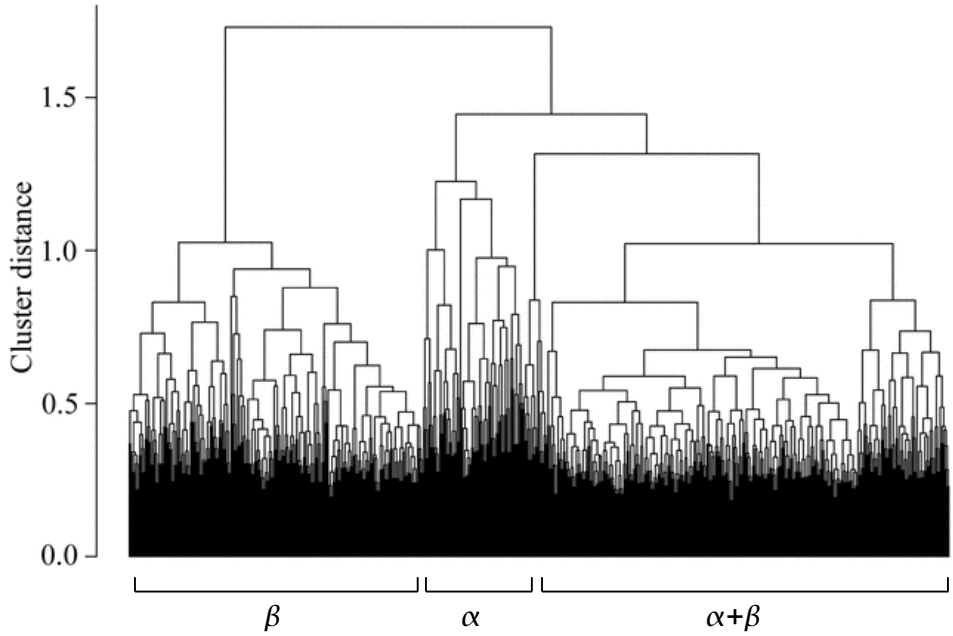


Fig. 2 Hierarchical clustering tree relative to $\langle|E|^2\rangle(d^*)$ profiles from 600 different proteins. At a clustering distance of about 1.3 three main clusters can be identified, corresponding to alpha, beta and mixed structures. [*Both fig. 1 and 2 are adapted from (Morris, 2004)*]

Cluster analysis of $\langle |E|^2 \rangle (d^*)$ distribution for a great number of protein crystals has shown that plots with similar content of alpha and beta structure tend to group together (fig. 2). The clustering process has been accomplished starting from single curves, which were grouped according to their euclidean distance in increasingly larger partitions (Morris, 2004).

Secondary structure and Patterson maps

The Patterson function can be obtained as Fourier transform of the diffracted intensities and coincides with the self-correlation of the electron density:

$$P(\boldsymbol{u}) = \rho(\boldsymbol{r}) * \rho(\boldsymbol{r}) = \int_V \rho(\boldsymbol{r})\rho(\boldsymbol{r}+\boldsymbol{u})\,d\boldsymbol{r} = V^{-1}\sum_{\boldsymbol{h}} |F(\boldsymbol{h})|^2 \cos(2\pi\,\boldsymbol{h}\cdot\boldsymbol{u}) \quad [3] \qquad (4.3)$$

The two functions $I(\boldsymbol{h})$ and $P(\boldsymbol{u})$ provide two alternative descriptions of the same object. The intensity represents the frequency spectrum of the Patterson function and the choice to analise the latter was made because it bears a more direct relationship with electron density (both are defined in real space). From the definition of self-correlation it is possible to interpret $P(\boldsymbol{u})$ as a measure of the global superposition (a scalar product of square integrable functions) between the electron density of the unit cell and the same density translated of a vector $\boldsymbol{u}$. At atomic resolution (d<1.2 Å) the peaks arising in the Patterson function correspond to interatomic vectors and the map can be used for *ab initio* phasing; recently it has been shown that these approaches, which were known since long time, can be succesfully applied to macromolecular structures (Burla, 2006).

In the medium resolution range (1.5-4.5 Å) the peaks due to interatomic vectors cannot be isolated and the map should have lost any *ab initio* phasing power; this reflects the gradual merging of atomic peaks occurring in electron density as the resolution decreases. Nevertheless, in a protein electron density the polypeptide chain still displays good connectivity (that breaks down above ~4.5 Å) and one can easily recognize the secondary structure. So it is likely that the Patterson map will show some characteristic patterns due to correlations between secondary structure elements.

3 The star ($*$) denotes function correlation, related to convolution ($\times$) by $f(x)*g(x) = f(x)\times g(-x)$

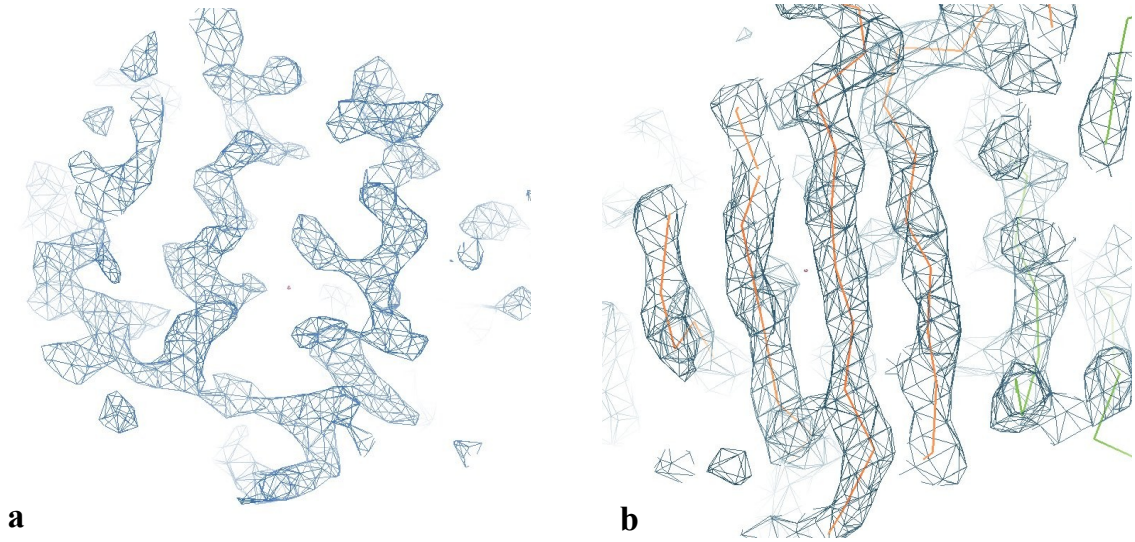


Fig 3 Portions of electron density maps for (a) alpha helices (myoglobin) and (b) beta strands (*FABP*) at 4 Å resolution.

Thinking about the unit cell content as a collection of α-helices, β-strands and segments without secondary structure (*loops*), the density $\rho(\boldsymbol{r})$ can be written as a sum of contributions $\rho_i(\boldsymbol{r})$ of the different elements. The Patterson function then results by summing the self-correlations of individual fragments $P_{ii}(\boldsymbol{u})$ and the cross-correlation terms ($P_{ij}(\boldsymbol{u})$) between fragment pairs:

$$\begin{aligned} P(\boldsymbol{u}) &= \rho(\boldsymbol{r})*\rho(\boldsymbol{r}) = \sum_i \sum_j \rho_i(\boldsymbol{r})*\rho_j(\boldsymbol{r}) = \sum_i P_{ii}(\boldsymbol{u}) + 2\sum_i \sum_{j>i} P_{ij}(\boldsymbol{u}) \\ P_{ij}(\boldsymbol{u}) &= \rho_i(\boldsymbol{r})*\rho_j(\boldsymbol{r}) \end{aligned} \tag{4.4}$$

Assuming that atomic models are known for each fragment, from which the ideal electron densities $\rho_i^0(\boldsymbol{r})$ can be costructed, and that the orientations and positions in the unit cell (specified by rotation matrices $\boldsymbol{R}_i$ and translation vectors $\boldsymbol{t}_i$) are unknown, one can write

$$P_{ij}(\boldsymbol{u}) = \rho_i^0(\boldsymbol{R}_i\boldsymbol{r}-\boldsymbol{t}_i) * \rho_j^0(\boldsymbol{R}_j\boldsymbol{r}-\boldsymbol{t}_j) = \rho_i^0(\boldsymbol{R}_i\boldsymbol{r}) * \rho_j^0(\boldsymbol{R}_j\boldsymbol{r}-\boldsymbol{t}_{ij}) \ , \quad \boldsymbol{t}_{ij} = \boldsymbol{t}_i - \boldsymbol{t}_j \ . \tag{4.5}$$

For $i = j$ this reduces to

$$P_{ii}(\boldsymbol{u}) = \rho_i^0(\boldsymbol{R}_i\boldsymbol{r}) * \rho_i^0(\boldsymbol{R}_i\boldsymbol{r}) = P_{ii}^0(\boldsymbol{R}_i\boldsymbol{u}) \tag{4.6}$$

where $P_{ii}^{0}(\boldsymbol{u})$ stands for the *i*-th fragment self-correlation expressed in the model reference frame. From this it is clear that the self-correlation functions $P_{ii}(\boldsymbol{u})$ depend only on the orientation of the structure elements, while cross-terms also on the relative positions $\boldsymbol{t}_{ij}$.

If the element *k* gives rise to a characteristic distribution of peaks in the correlation space, one can think to detect its contribution $P_{kk}(\boldsymbol{u})$ in the Patterson function $P(\boldsymbol{u})$, obtaining at the same time its orientation. This principle is the basis of *molecular replacement* (MR) methods, which orient a known fragment in the unit cell by seeking the maximum for the *rotation function* $R(\boldsymbol{C})$[4]:

$$\begin{aligned} R(\boldsymbol{C}) &= \int_{\Omega} P(\boldsymbol{u})\, P_{kk}^{0}(\boldsymbol{C}\boldsymbol{u})\, d\boldsymbol{u}\ , \quad \boldsymbol{C} = \boldsymbol{C}(\theta_1,\theta_2,\theta_3) \\ & \qquad\qquad P_{kk}^{0}(\boldsymbol{u}) = P_{kk}(\boldsymbol{R}_k^{-1}\boldsymbol{u}) \end{aligned} \tag{4.7}$$

where the partial Patterson function $P_{kk}^{0}(\boldsymbol{u})$ can be computed from the known fragment k arbitrarily positioned in a arbitrary unit cell, and the rotation matrix is expressed as function of a set of three rotation angles (usually, the Euler angles). Again, expressing the function as sum contributions from the fragments *1,--,k,--,N*, one has:

$$\begin{aligned} R(\boldsymbol{C}) &= \sum_i \sum_j \int_{\Omega} P_{ij}(\boldsymbol{u}) P_{kk}^{0}(\boldsymbol{C}\boldsymbol{u})\, d\boldsymbol{u} \\ &= \sum_i \sum_{j}^{ij\neq kk} \int_{\Omega} P_{ij}(\boldsymbol{u})\, P_{kk}^{0}(\boldsymbol{C}\boldsymbol{u})\, d\boldsymbol{u} + \int_{\Omega} P_{kk}(\boldsymbol{u}) P_{kk}^{0}(\boldsymbol{C}\boldsymbol{u})\, d\boldsymbol{u} \end{aligned} \tag{4.8}$$

where the terms $\int_{\Omega} P_{ij}(\boldsymbol{u})\, P_{kk}^{0}(\boldsymbol{C}\boldsymbol{u}) d\boldsymbol{u}\,, \quad ij \neq kk$ can be thought as 'noise', while the signal of interest is represented by the function

$$\int_{\Omega} P_{kk}(\boldsymbol{u})\, P_{kk}^{0}(\boldsymbol{C}\boldsymbol{u}) d\boldsymbol{u}\ , \tag{4.9}$$

which exhibits a pronounced peak of height $\int_{\Omega} [P_{kk}(\boldsymbol{u})]^2 d\boldsymbol{u}$ when $\boldsymbol{C} = \boldsymbol{R}_k^{-1}$, giving the correct orientation for the fragment. To be identified in the rotation function, the solution peak must be great when compared with the standard deviation $\sigma[R(\boldsymbol{C})]$. This means that the model fragment *k* cannot be too small with respect to the asymmetric unit content,

4. The integration domain Ω is a spherical shell chosen in such a way that the origin peak and the intermolecular vectors are excluded.

otherwise its signal will be indistinguishable among the many local maxima arising from the other terms. In that case the problem will be undetermined, since many acceptable ways will exist to superpose the fragment self-correlation with the Patterson map. Even then, however, a way could exist to orient the fragment, if attention was paid to match only some special features of the model with the map; this hypothesis was the starting point of the present work.

The self-correlation of an alpha helix

To compute the self-correlation of an isolated alpha helix one can imagine to translate its electron density by every possible vector ***u,*** each time evaluating the integral of the product (original density)×(translated density). Since the helix has a periodicity in the axis direction, when the translation vector ***u*** has the same direction as the axis a periodic trend in self-correlation can be predicted. Maxima are expected for translations by one or more helical pitches, since these transport a part of the helix on itself: $\pm n\, l\, \hat{\boldsymbol{a}}$, where l is the pitch of the helix ($l_\alpha \approx 5.4$ Å) and $\hat{\boldsymbol{a}}$ is a unit vector specifying axis direction. Furthermore, the height of the peaks should decrease with n since the portion of the helix which superposes is each time shorter; for an helix that is k turns long, the last (and smallest) peaks occur at $\pm(k-1)\, l\, \hat{\boldsymbol{a}}$. These reasonings are strictly valid only for a very idealized helix, since they do not take into account that in a α-helix there is not an integer number of residues per turn, so that after a translation of l the C_α positions would not coincide.

In the same way, the model does not consider that each turn of the helix will present different side chains, nor that the helix can be bent (which gives the notion of 'helical axis' only an approximate or average meaning). Nevertheless, in going from this idealized helix to a real one only minor changes are expected to take place in the Patterson map, since the effect of side chains can be neglected if compared to the strong main chain contribution. In fact, the computed Patterson map from a single alpha helix shows a periodic series of peaks in the same direction of the helical axis (fig. 4a,b). The periodicity can be seen in the linear profile of the Patterson map, which for generic direction $\hat{\boldsymbol{n}}$ is given by

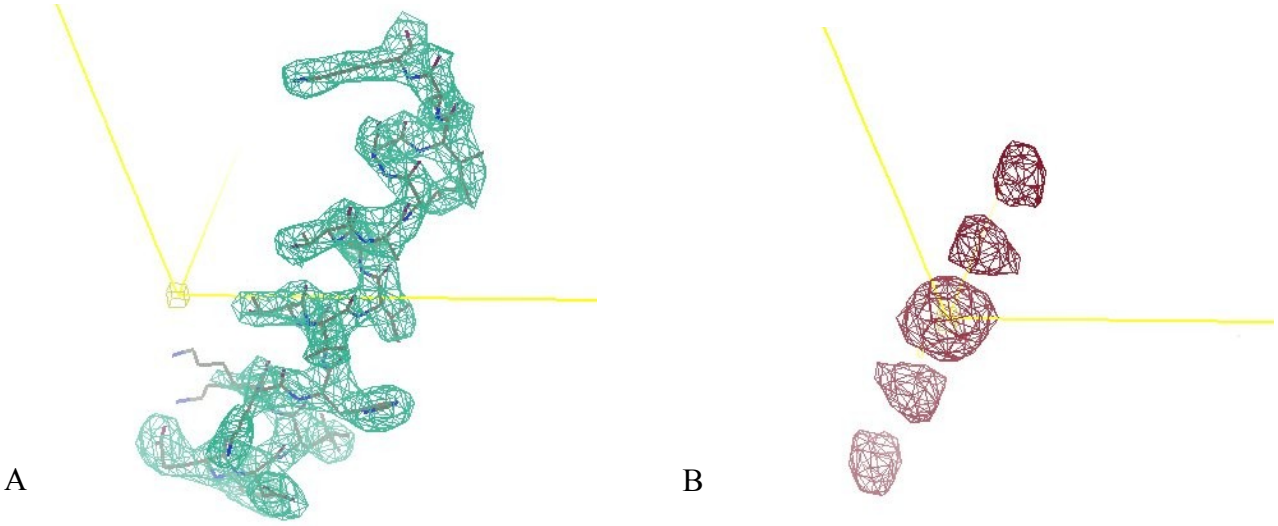


Fig. 4 A. Atomic model and electron density of an alpha helix B. The corresponding Patterson map (resolution is 3 Å and the maps are contoured at 3 sigma) showing a series of peaks with a spacing of about 5.4 Å

$$P_{\hat{\boldsymbol{n}}}(t)=P(\hat{\boldsymbol{n}}\,t)=V^{-1}\sum_{\boldsymbol{h}}|F(\boldsymbol{h})|^2\cos(2\pi\,\boldsymbol{h}\cdot\hat{\boldsymbol{n}}\,t) \tag{4.10}$$

The linear periodicity can be analysed with the Fourier transform[5] :

$$\begin{aligned} G_{\hat{\boldsymbol{n}},L}(s)&=FT\,[P_{\hat{\boldsymbol{n}}}(r)]=V^{-1}\sum_{\boldsymbol{h}}|F(\boldsymbol{h})|^2\int_{-L}^{+L}\exp[\,2\pi\,i(s-\boldsymbol{h}\cdot\hat{\boldsymbol{n}})\,r\,]dr \\ &=2\,L\,V^{-1}\sum_{\boldsymbol{h}}|F(\boldsymbol{h})|^2\;sinc\;[\,2\pi\,L(s-\boldsymbol{h}\cdot\hat{\boldsymbol{n}})\,] \end{aligned} \tag{4.11}$$

where the unit vector $\hat{\boldsymbol{n}}$ can be expressed as function of the two spherical angles θ,ϕ. Both quantities $P_{\theta,\phi}(r)$ and $|G_{\theta,\phi}(s)|^2$ can be seen in fig. 5, where the plots for different spatial directions are compared.

It turned out to be simpler to compute the Patterson map on a grid of crystallographic coordinates using an existing software; the one-dimensional profiles were successively obtained by interpolation. The whole procedure consisted in the following steps:

5. The unit vector $\hat{\boldsymbol{n}}$ is expressed in fractional crystallographic coordinates. The transformation to a polar spherical system must follow a conversion in cartesian coordinates through a matrix $\boldsymbol{A}$ depending on the cell parameters: $\hat{\boldsymbol{n}}=\boldsymbol{A}\hat{\boldsymbol{n}}_c$ $\hat{\boldsymbol{n}}_c(\theta,\phi)=(\sin\phi\cos\theta,\sin\phi\,sen\,\theta,\cos\phi)^T$

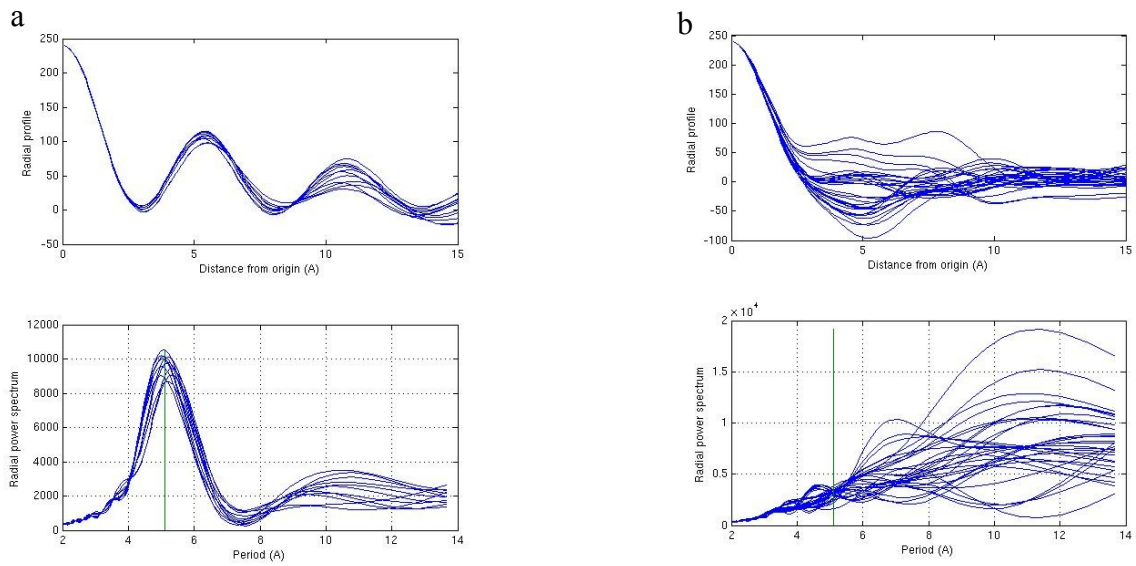


Fig. 5 a) Profiles $P_{\theta,\phi}(r)$ (upper plots) and power spectra $|G_{\theta,\phi}(s)|^2$ (lower plots) for directions near to the helical axis. The strong peak in correspondence of the period *d*=5.4 Å indicates the presence of an helix. b) Same plots for a sparse set of directions being very different from the helix orientation.

- The Patterson function is computed at the desired resolution with the CCP4 package (CCP4, 1994). The output map covers the whole unit cell.
- The map in CCP4 format is read in by a Fortran routine which performs spline interpolation in all the points of a polar grid covering half a sphere. In this way the conversion $P(n_x\Delta x, n_y\Delta y, n_z\Delta z) \rightarrow P(n_r\Delta r, n_\theta\Delta\theta, n_\phi\Delta\phi)$ is carried out.
- The power spectrum $|G_{\theta,\phi}(s)|^2$ is computed along the *r* coordinate by Fourier transforming $P(r,\theta,\phi)$ in the range $a \le r \le L$, where *a* is chosen to exclude the origin peak and the choice of *L* should limit the calculation to a single unit cell.

Since the Fourier transform is linear, one can write $G_{\theta,\phi}(s)$ as a sum of simple and mixed contributions from the different partial structures in which the model has been divided:

$$G_{\theta,\phi}(s) = \sum_i G^{(ii)}_{\theta,\phi}(s) + 2\sum_i \sum_{j>i} G^{(ij)}_{\theta,\phi}(s) \qquad G^{(ij)}_{\theta,\phi}(s) = \int_{-L}^{+L} P^{(ij)}_{\theta,\phi}(r)\, e^{-2\pi i r s}\, dr \qquad (4.12)$$

A structural element can be identified by means of the power spectrum $|G_{\theta,\phi}(s)|^2$ if it has

a characteristic spatial period d since a peak for $s=d^{-1}$ will appear. Thus in principle the study of $|G_{\theta,\phi}(s)|^2$ can allow detection of substructures (as α-helices) which are too small to be oriented by a rotation function search.

Detecting helices in real structures

The best way to represent the spectrum is to plot the various sections $|G_d(\theta,\phi)|^2$. Each of these sections represents the content of the cell in period d as a function of the spatial orientation. As a first step, ideal maps were derived from atomic models to assess the validity of the assumption for single helices. As expected, in each case a strong peak was observed in the section corresponding to a period of 5.4 Å for a pair of angles (θ,ϕ) giving the direction of the helical axis. The first tests have been carried out using a P1 cell. Obviously when symmetry is present more peaks appear in the power spectrum, being related by the symmetry of the Patterson space group.

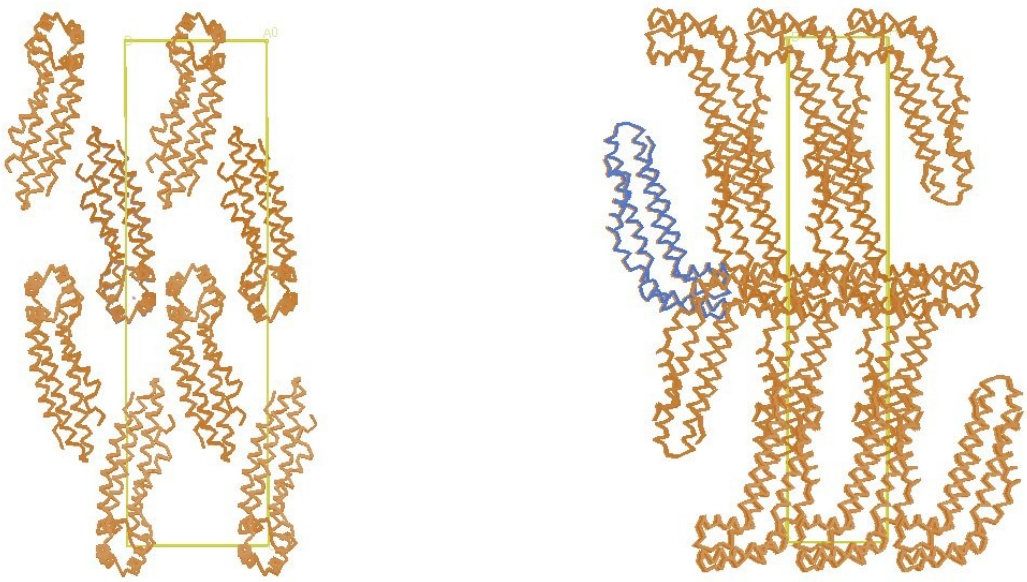


Fig. 6 Crystal packing of CagZ, as seen along two of the crystallographic axes. The unit cell edges have been traced. Four different orientations are possible, corresponding the fourfold symmetry observed in the angular plots.

In a subsequent step, real diffraction data relative to already solved structures were considered. Here the results can vary much, probably because in some cases the signal is

partially buried and appears for a period slightly different from the expected. The results were more interesting for structures with long helices (20-30 residues); in the case of CagZ (Cendron, 2004), an entirely α protein from *Helicobacter pylori*, the presence of three long, nearly parallel helices (fig. 7a) gives rise to a strong signal in the power spectrum, while the five much shorter helices are responsible for a group of weaker signals (fig. 7b). Here the space group was $P2_12_12_1$, corresponding to four asymmetric units, each one containing one single molecule (fig. 6), so that only one fourth of the power spectrum section is independent (two planes of symmetry are present). It is interesting to notice that the elongated shape of the strong peak does not arise from a difference in the orientation between the three long helices but is mostly due to the bent shape of the longest one.

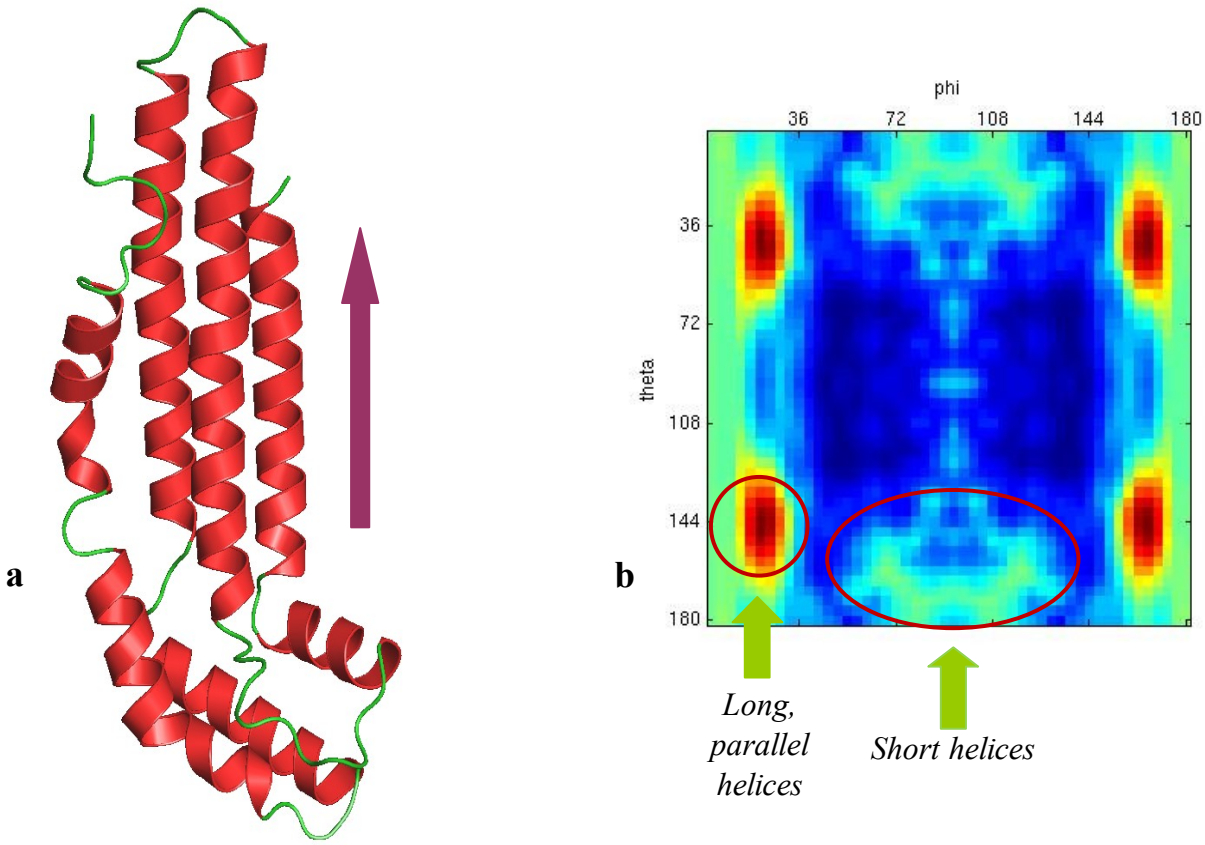


Figura 7. (a): Ribbon representation of the protein CagZ (Cendron, 2004). The arrow indicates the direction of the three longest helices. (b): Section of the power spectrum $|G_d(\theta,\phi)|^2$ corresponding to the 5.4 Å period. Data up to 4 Å of resolution have been included in the calculation.

Considerations about β structure

An analogous approach was tried in order to identify the beta strands from the Patterson map. Observing a nearly linear strand (fig. 8) one can notice that some periodicity exists, although it must be very weak compared to that of an helix. Instead, translating the strand along its axis, something different happens. Considering the main chain atoms, there will always be an appreciable superposition, which should decrease in a roughly linear way with the translation. This can be seen in a plot of $P_{\theta,\phi}(r)$ (fig. 8).

A possible way to identify this trend in the Patterson function is to compute the integral

$$H_{\hat{\boldsymbol{n}}} = \int_{l_1}^{l_2} P_{\hat{\boldsymbol{n}}}(t)\, dt = \int_{l_1}^{l_2} P(\hat{\boldsymbol{n}}\, t)\, dt\,. \tag{4.13}$$

Since the contribution of the strand to $P_{\theta,\phi}(r)$ slowly decreases, the integral over a carefully chosen range $[l_1, l_2]$ should have a great value in the direction of a strand axis. As for the helices, the integration limits are to be chosen in order to get the maximum signal to noise ratio. Signal is relative to the single element contributions while the 'noise' arises from the mixed terms P_{ij}.

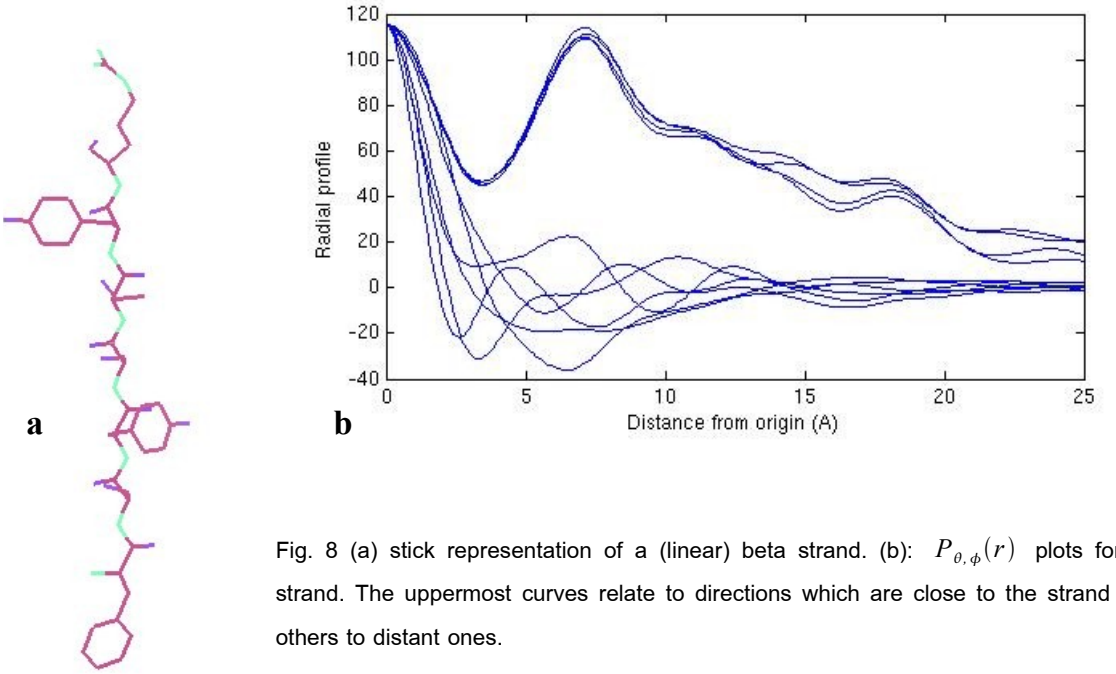


Fig. 8 (a) stick representation of a (linear) beta strand. (b): $P_{\theta,\phi}(r)$ plots for a single beta strand. The uppermost curves relate to directions which are close to the strand orientation, the others to distant ones.

The lower limit should exclude the origin peak, while the upper one must be kept smaller than the maximum lenght of a beta strand in order to avoid as much as possible those radial contributions that do not arise from beta strands.

The idea was tested on data from the protein HiUase (Zanotti, 2006), whose structure is almost entirely beta, with rather long and linear strands. For this purpose, a $H(\theta,\phi)$ plot was constructed for the whole structure and compared with the single strand contributions, obtained by positioning only one strand at a time into the unit cell (fig. 9).

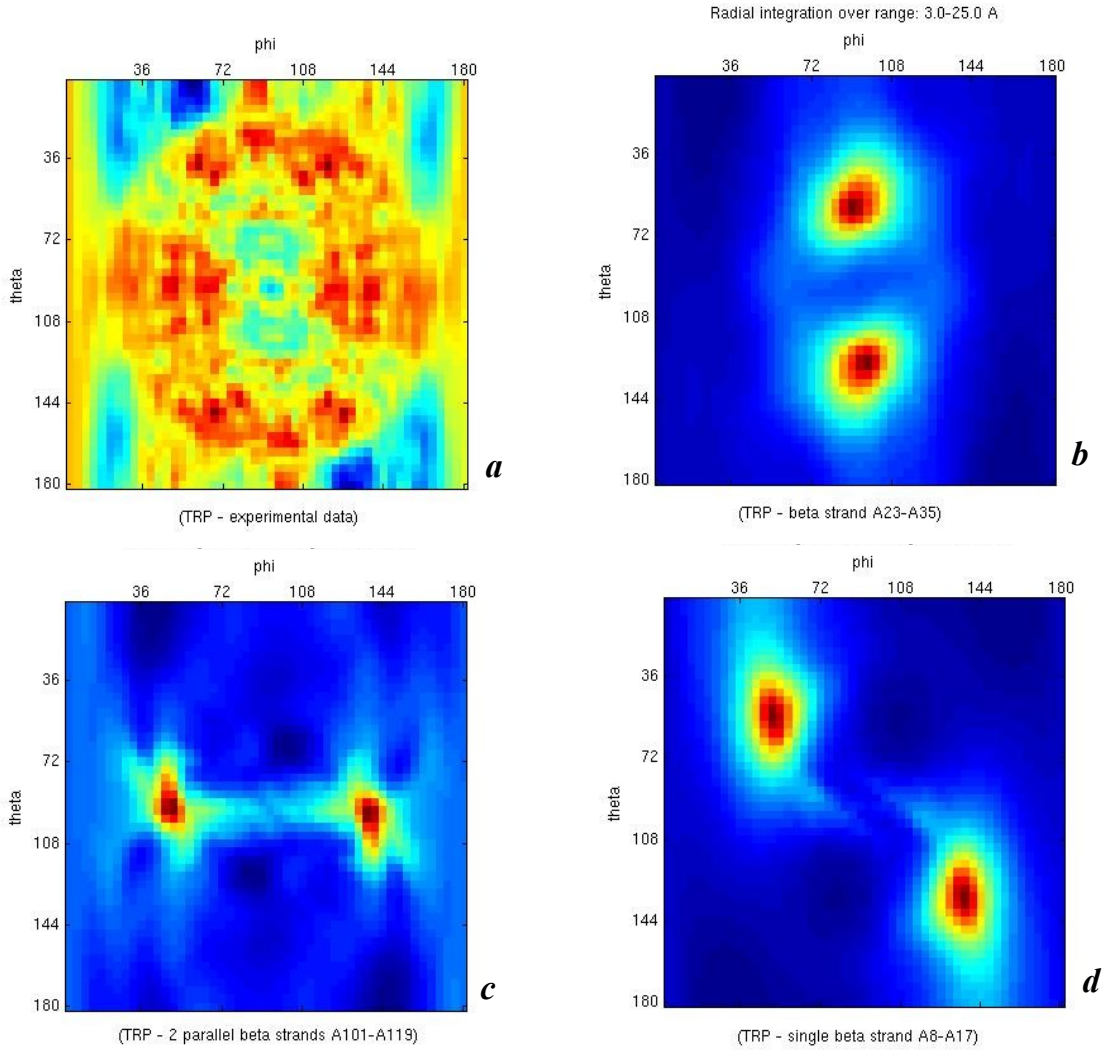


Fig.9 Integral plots for the protein HiUase. (a) is the plot relative to the whole structure (computed from experimental data). The three plots (b), (c), (d) are obtained from individual beta strands correctly positioned in a cell with the same dimensions and same space-group symmetry as the real structure. One can notice that only the peaks from (c) are clearly visible in the plot (a) relating to the whole structure.

Although the peaks arising from the isolated beta strands can be identified in the integral plot, many others strong signals cannot be ascribed to strands, so that this criterion alone is not strong enough to identify the strands. Other approaches were tried by taking into account also the frequency content, but no simple criterion could be found, and a statistical study of the profiles is probably needed to get a reliable way (if any exists) to identify the strands.

While the single strands are difficult to identify, the association of many strands to give a so-called beta sheet structure introduces a periodic repetition which should be detectable in the power spectrum in the direction orthogonal to the strand axes. In principle the period expected (~ 4.8 Å) is shorter than the alpha helix period, but the corresponding peak can be so broad to be well visible also in the 5.4 Å section. An example can be seen in the power spectrum relative protein G (Derrick, 1994), where the short helix gives rise to a weak signal, while the beta sheet is responsible for a big, broad peak (fig. 10b).

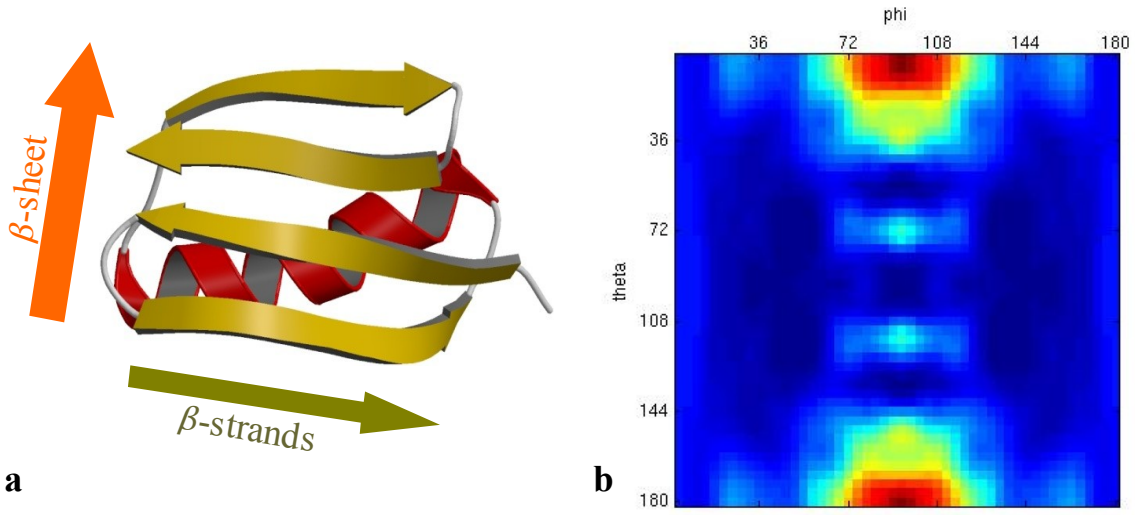


Fig. 10 a) Ribbon representation of the Protein G structure. The two arrows show the two orthogonal directions of elongation of the strands and of beta sheet periodicity. b) 5.4 Å section of the power spectrum $|G_d(\theta,\phi)|^2$. The strong, broad peak arises from the beta sheet, while small peaks in the middle are due to the short alpha helix.

Conclusions

The work carried on was aimed at finding new *ab initio* phasing methods to be used in the field of macromolecular crystallography. Probably, more efficient phasing strategies, with a reduced sensitivity to resolution, can only be achieved through a radical change in strategy. Since now, *ab initio* phasing has been based mainly on the existence of atoms; only in the last years this classical approach has been complemented with methods specific for low resolution phasing. These latter often start from a small shell of very low resolution, and allow to retrieve rough structural information, such as position and shape of the individual molecules present in the cell.

Difficulties arise when trying to extend the low resolution phases; it can be argued that the amount of general information about electron density at low resolution (d_{min} > 5 Å) is very limited and strong constraints are difficult to identify. However, is one takes into account the data up to ~4 Å resolution the secondary structure should appear in the correct electron density and the polypeptide chain could be traced into the 'worm-like' density. Probably, this is the resolution region where potential *ab initio* methods could be effective, since many topological constraints arise but the number of reflections to be phased is not so high (of the order of ~10^3). An interesting observation is that the connectivity depends mainly on the strongest reflections, so that it is preserved even when as much as 80% of the weakest structure factors are randomly phased (provided that the remaining ones are given the exact phases).

In the course of this project very different approaches were tried. While neural networks seem not a promising option in phasing (at least, if used in the simple way described here), iterative methods could be useful if constraints strong enough are defined. The *charge flipping* algorithm, whose modifications are discussed in chapter 3, does not represent an efficient way of exploring the phase space, owing to its strong tendency to stagnation; moreover, it does not allow to exploit arbitrary constraints (as the topological ones) in a simple, understandable way. A natural development of this strategy would make use of the more powerful *difference map* algorithm, which shows a clever behaviour with respect to false minima, and has been proposed to solve a variety of non-convex optimization problems aside from phase retrieval (Elser *et al*., 2007).

There is no doubt that the most interesting results are those obtained in the context of

Patterson map analysis. It has been shown that the presence of alpha helices and beta sheets reflects on significant frequency contributions to the Patterson map, allowing for detection and orientation of these elements. The generality of this approach is reduced by the wide variety of effects that can be observed across a class of structural elements; for instance, only helices long enough would give rise to detectable signals, and bent helices result in deformed peaks. Single beta strands can deviate much from linearity; even the most linear ones are difficult to locate in the Patterson function on the simple basis of radial integral values.

However, the full potential of the method could be assessed only by a more extended feasibility study. Probably, statistical analysis of Patterson profiles, together with viable methods to increase the signal-to-noise ratio (as, for example, an appropriate integration over Patterson peak width in a direction orthogonal to the profile axis), would allow for more sensitive secondary structure detection. The approach described here has an intrinsic interest in putting aside any atomistic interpretation, which of course is no more expected to hold at the resolutions chosen for this study. Further developments of the method would hopefully allow to assign initial phases on the basis of some identified fragment, or at least provide an useful tool to predict some of the protein structural features directly from the diffracted intensities.

References

Baerlocher C., McCusker L.B. and Palatinus L. (2007). Charge flipping combined with histogram matching to solve complex crystal structures from powder diffraction data. *Z. Kristallogr.* **222**, 47-53.

Baker D., Krukowski A.E. and Agard D.A. (1993 a). Uniqueness and the *ab initio* phase problem in macromolecular crystallography. *Acta Cryst.* D**49**, 186–192

Baker D., Bystroff C., Fletterick R., Agard D. (1993 b)**.** PRISM: Topologically Constrained Phase Refinement for Macromolecular Crystallography**.** *Acta Cryst.* D**49**, 429-439

Bragg W.L. (1913). The diffraction of short electromagnetic waves by a crystal. *Proc. Cambridge Phil. Soc.* **17**, 43-57

Bricogne G. (1997 a). The Bayesian Statistical Viewpoint on Structure Determination: Basic Concepts and Examples. In: *Methods in Enzymology*, **276A**, 361-423.

Bricogne, G. (1997 b)**.** *Ab Initio* Macromolecular Phasing: A Blueprint for an Expert System Based on Structure Factor Statistics with Built-In Stereochemistry. In: *Methods in Enzymology*, **277A**, 14-18.

Buerger, M.J. (1959). *Vector space and its applications in crystal structure investigation.* Wiley, New York

Burla M.C., Caliandro R., Carrozzini B., Cascarano G.L., De Caro L., Giacovazzo C. and Polidori G. (2004). Ab initio protein phasing: the Patterson deconvolution method in *SIR2002*. *J. Appl. Cryst.* **37**, 258-264

Burla M.C., Caliandro R., Carrozzini B., Cascarano G.L., De Caro L., Giacovazzo C., Polidori G, Siliqi D. (2006). The revenge of Patterson methods. I. Protein *ab initio* phasing – J. Appl. Cryst. . **39,** 527-535

References

Burla M.C., Caliandro R., Camalli M., Carrozzini B., Cascarano G.L, De Caro L., Giacovazzo C., Polidori G., Siliqi D. and Spagna R. (2007). *IL MILIONE*: a suite of computer programs for crystal structure solution of proteins. *J. Appl. Cryst*. **40**, 609-613

Cendron L., Seydel A., Angelini A., Battistutta R., Zanotti G. (2004). Crystal structure of CagZ, a protein from the Helicobacter pylori pathogenicity island that encodes for a type IV secretion system. - *J. Mol. Biol.* **340**, 881-889.

Chapman H.N., Barty A., Bogan M.J., Boutet S., Frank M., Hau-Riege S.P., Marchesini S., Woods B.W., Bajt S., Benner W.H., London R.A., Plönjes E., Kuhlmann M., Treusch R., Düsterer S., Tschentscher T., Schneider J.R., Spiller E., Möller T., Bostedt C., Hoener M., Shapiro D.A., Hodgson K.O., van der Spoel D., Burmeister F., Bergh M., Caleman C., Huldt G., Seibert M.M., Maia F.R.N.C., Lee R.W., Szöke A., Timneanu N. and Hajdu J. (2006). Femtosecond diffractive imaging with a soft-X-ray free-electron laser. *Nature Physics* **2**, 839-843

Chergui J. (2002). GFT, Generic Fourier Transform, copyright CNRS/IDRIS, France. http://www.idris.fr/data/publications/GFT/

Cochran W. (1955). Relations between the phases of structure factors. *Acta Cryst.* **8**, 473-478

Coelho A.A. (2007). A charge-flipping algorithm incorporating the tangent formula for solving difficult structures. *Acta Cryst.* A**63**, 400-406

Cowtan K.D., Zhang K.Y.J. (1999). Density modification for macromolecular phase improvement. *Prog. Biophys. Mol. Biol.* **72**, 245-270

Debaerdemaeker T., Germain G., Main P., Refaat L.S., Tate C. and Woolfson M.M. (1988a). *MULTAN88: computer programs for the automatic solution of crystal structures from X-ray diffraction data***.** University of York, U.K.

Debaerdemaeker T., Tate C. and Woolfson, M.M. (1988b). On the application of phase relationships to complex structures. Acta Cryst. B**24**, 91-96

Derrick J.P. and Wigley D.B. (1994). The Third IgG-Binding Domain from Streptococcal Protein G: an Analysis by X-ray Crystallography of the Structure Alone and in a Complex with Fab. *J. Mol. Biol.* **243**, 906-918.

Elser V. (2003). Phase retrieval by iterated projections. *J. Opt. Soc. Am*. A**20**, 40-55

Elser V., Rankenburg I. and Thibault P. (2007). Searching with iterated maps. *Proc. Natl. Acad. Sci.* **104**, 418-423

Fienup, J.R. (1978). Reconstruction of an object from the modulus of its Fourier transform. Opt. Lett., **3**, 27-29.

Gerchberg R.W., Saxton W.O. (1972). A practical algorithm for the determination of phase from image and diffraction plane pictures. Optik, **35**, 237-246.

Giacovazzo C. (1977). A general approach to phase relationships: the method of representations. Acta Cryst. A**33**, 933-944

Giacovazzo C. (1980). The method of representations of structure semiinvariants. II. New theoretical and practical aspects. Acta Cryst. A**36**, 362-372

Goldstein A., Zhang K.Y.J (1998). The two-dimensional histogram as a constraint for protein phase Improvement. *Acta Cryst*., D**54**, 1230-1244

Hauptman H.A. and Karle J. (1953). The solution of the Phase Problem I. The centrosymmetric Crystal. ACA Monograph no. 3, Polycrystal Book Service, New York

Hauptman H.A. (1975). A new method in the probabilistic theory of structure invariants.

Acta Cryst. A**31**, 680-687

Hauptman H.A. (1991). A minimal principle in the phase problem. In: Moras D., Podjarny A.D. and Thierry J.C. (eds.), Crystallographic Computing 5: from Chemistry to Biology, IUCr Oxford Univ. Press, 324-332

Haykin S. Neural Networks, a comprehensive foundation (2nd Edition – 1999). Prentice Hall International, London

He, H. (2006). Simple constraint for phase retrieval with high efficiency. J. Opt. Soc. Am. A**23**, 550-556

Hinton G.E, Salakhutdinov R.R. Reducing the dimensionality of data with neural networks. Science (2006). **313**, 504-507

Laue, M. (1912). Eine quantitative Prüfung der Theorie für die Interferenz-Erscheinungen bei Röntgenstrahlen. Sitzungberichte der Kgl. Bayer Akad. Der Wiss, 363-373; reprinted in Ann. Phys. (1913), **41**, 989-1002

Lunin V.Y. (1985). Use of the fast differentiation algorithm for phase refinement in protein crystallography. *Acta Cryst*., A**41**, 551-556

Lunin V.Y. and Woolfson M.M. (1993). Mean phase error and the map correlation coefficient. Acta Cryst. D**49**, 530-533

Lunin V.Y., Lunina N.L., Petrova T.E., Skovoroda T.P., Urzhumtsev A.G. and Podjarny A.D. (2000a). Low-resolution ab initio phasing: problems and advances. Acta Cryst. D**56**, 1223-1232

Lunin V.Y., Lunina N.L. and Urzhumtsev A. (2000b). Connectivity properties of high-density regions and ab initio phasing at low resolution. Acta Cryst. A**56**, 375-382

Lunin V.Y., Lunina N.L., Ritter S., Frey I., Berg A., Diederichs K., Podjarny A.D., Urzhumtsev A. and Baumstark M.W. (2001). Low-resolution data analysis for low-density lipoprotein particle. Acta Cryst. D**57**, 108-121

Lunin V.Y., Urzhumtsev A., Bockmayr A. (2002). Direct phasing by binary integer programming. *Acta Cryst.,* A**58**, 283-291

Lunina N., Lunin V. and Urzhumtsev A. (2003). Connectivity-based ab initio phasing: from low resolution to a secondary structure. Acta Cryst. D59, 1702-15

Marchesini S., He H., Chapman H.N., Hau-Riege S.P., Noy A., Howells M.R., Weierstall U. and Spence J.C.H. (2003). X-ray image reconstruction from a diffraction pattern alone. *Phys. Rev.* B**68**, 140101(R). [arXiv:physics/0306174v2]

Marchesini S. (2007). Benchmarking iterative projection algorithms for phase retrieval. [arXiv:physics/0404091v1]

Miao J., Sayre D. and Chapman H.N. (1998). Phase retrieval from the magnitude of the Fourier transforms of nonperiodic objects. *J. Opt. Soc. Am.* A**15**, 1662-1669

Miao J., Hodgson K.O. and Sayre D. (2001). An approach to three-dimensional structures of biomolecules by using single-molecule diffraction images. *Proc. Natl. Acad. Sci.* **98**, 6641-6645

Millane R.D. (1990) Phase retrieval in crystallography and optics. *J. Opt. Soc. Am.* A**7**, 394-411

Mølgaard A., Larsen S. (2003). Crystal packing in two pH-dependent crystal forms of rhamnogalacturonan acetylesterase. *Acta Cryst*., D**60**, 472-478

Morris R.J., Blanc E., Bricogne G. (2004). On the interpretation and use of the $<|E|^2>(d^*)$ profiles. *Acta Cryst.* **D60**, 227-240.

References

Müller J.J, Lunina N.L, Urzhumtsev A., Weckert E., Heinemann U. and Lunin V. (2006). Low-resolution ab initio phasing of Sarcocystis muris lectin SML-2. *Acta Cryst.* D**62**, 533-540

Nieh Y.P., Zhang K.Y. (1999). A two-dimensional histogram-matching method for protein phase refinement and extension. *Acta Cryst.,* D**55**, 1893-1900

Oszlányi G. and Sütő A. (2004). *Ab initio* structure solution by charge flipping. *Acta Cryst.* A**60**, 134-41

Oszlányi G. and Sütő A. (2005). *Ab initio* structure solution by charge flipping. II. Use of weak reflections. *Acta Cryst.* A**61**, 147-152

Palatinus, L. (2004)**.** *Ab initio* determination of incommensurately modulated structures by charge flipping in superspace. *Acta Cryst.* A**60**, 604-610.

Rossmann M.G. and Blow D.M. (1962). The detection of sub-units within the crystallographi asymmetric unit. *Acta Cryst*. **15**, 24-31

Sayre, D. (1952a). The squaring method: a new method for phase determination. *Acta Cryst.* **5**, 60-65

Sayre, D. (1952b). Some implications of a theorem due to Shannon – *Acta Cryst.* **5**, 843

Scheres S.H.W. and Gros P. (2004). The potentials of conditional optimization in phasing and model building of protein-structures**,** *Acta Cryst.* D**60,** 2202-2209

Shannon C.E. (1949). Communication in the presence of noise. *Proceeding of the IRE*, **37**(1), 10-21, January 1949

Shapiro D., Thibault P., Beetz T., Elser V., Howells M., Jacobsen C., Kirz J., Lima E., Miao

H., Neiman A.M. and Sayre D. (2005). Biological imaging by soft x-ray diffraction microscopy. *Proc. Natl. Acad. Sci.* **102**, 15343-15346

Sheldrick G.M. and Gould R.O. (1995). Structure solution by iterative peaklist optimization and tangent expansion in space group P1. *Acta Cryst.* B**51**, 423-431

Sheldrick G.M. (1996). *SHELX-96, Programs for X-Ray Crystallography*. University of Göttingen

Spence J.C.H., Schmidt K., Wu J.S., Hembree G., Weierstall U., Doak B. and Fromme P. (2005). Diffraction and imaging from a beam of laser-aligned proteins: resolution limits. *Acta Cryst*. A61, 237-245

Stark H., Yang Y. (1998). *Vector Space Projections*. John Wiley & Sons, New York

Thibault P. (2007). *Algorithmic methods in diffraction microscopy.* Ph.D. Thesis, Cornell University, pp. 82-90

Urzhumtseva L., Lunina N., Fokine A., Samama J.-P., Lunin V.Y. and Urzhumtsev A. (2004). *Ab initio* phasing based on topological restraints: automated determination of the space group and the number of molecules in the unit cell. *Acta Cryst.* D**60**, 1519-1526

Usón I. and Sheldrick G.M. (1999). Advances in direct methods for protein crystallography**.** *Curr. Opin. Struct. Biol.* **9** 643-648.

Wang B.C. (1985). Resolution of phase ambiguity in macromolecular crystallography. In: Wyckoff, H.W., Hirs, C.H.W., Timasheff S.N. (Eds.), Diffraction methods for Biological Macromolecules, vol. 115. Academic Press, Orlando, pp. 90-113

Weeks C.M., DeTitta G.T., Hauptman H.A., Thuman P. and Miller R. (1994). Structure solution by minimal function phase refinement and Fourier filtering: II. implementation and applications. *Acta Cryst.* A**50**, 210-220.

References

Wu J.S., Weierstall U. and Spence J.C.H. (2004). Iterative phase retrieval without support. *Opt. Lett.* **29**, 2737-2739

Wu J.S. and Spence J.C.H. (2005). Phasing diffraction data from a stream of hydrated proteins. *J. Opt. Soc. Am.* A**22**, 1453-1459

Zanotti G., Scapin G., Spadon P., Veerkamp J.H., Sacchettini J.C. (1992). Three-dimensional structure of recombinant human muscle fatty acid-binding protein. *J. Biol. Chem.,* **267,** 18541-18550.

Zanotti G. (2002). Protein Crystallography. In: Fundamentals of Crystallography. Giacovazzo C. (ed.). 2nd ed., Oxford University Press

Zanotti G., Cendron L., Ramazzina I., Folli C., Percudani R., Berni R. (2006). Structure of zebra fish HIUase: insights into evolution of an enzyme to a hormone transporter. - *J. Mol. Biol.* **363**, 1-9.

Zhang K.Y.J. and Main P. (1990). Histogram matching as a new density modification technique for phase refinement and extension of protein molecules. *Acta Cryst.* A**46**, 41-46

Zuo J.M., Vartanyants I, Gao M., Zhang R. and Nagahara L.A. (2003). Atomic Resolution Imaging of a Carbon Nanotube from Diffraction Intensities. *Science*, **300**, 1419-1421

STRUCTURE NOTE

Crystal Structure of Antigen TpF1 from *Treponema pallidum*

Anton Thumiger,[1] Alessandra Polenghi,[2,3] Elena Papinutto,[1,3] Roberto Battistutta,[1,3] Cesare Montecucco,[2,3] and Giuseppe Zanotti[1,3]

[1]*Dipartimento di Scienze Chimiche e ICB-CNR, Università di Padova, Padova, Italy*
[2]*Centro CNR Biomembrane e Dipartimento di Scienze Biomediche, Università di Padova, Padova, Italy*
[3]*Venetian Institute of Molecular Medicine (VIMM), Padova, Italy*

Introduction. Several bacterial genomes code for proteins belonging to the Dps family, which includes dodecamers, made up of 12 identical subunits, each of them with a four-helix bundle folding similar to that of ferritins. The crystal structure of several members of the family have been determined: Dps from *Escherichia coli* (1DPS, 1F30, 1F33, 1JRE, 1JTS, 1L8H, 1L8I),[1] *Listeria Innocua* Dps (1QGH, 2BJY, 2BK6, 2BKC),[2,3] HP-NAP from *Helicobacter pylori* (1JI4),[4] Dlp1 and Dlp2 from *Bacillus anthracis* (1JI5, 1JIG),[5] archaeal Dps-homolog from *Halobacterium salinarum* (1MOJ, 1TJO),[6] Dps protein from *Bacillus brevis* (1N1Q),[7] *Agrobacterium tumefaciens* Dps (1O9R),[8] Dps-like peroxide resistance protein from *Streptococcus suis* (1UMN),[9] Dps from *Mycobacterium smegmatis* (1VEI, 1VEQ, 1UVH).[10]

Despite their structural similarity and the fact that most of these proteins are capable of incorporating iron in vitro, their biological function appear to differ among family members. The *E. coli* and the *B. subtilis* proteins protect DNA from oxidative damage (Dps, *D*NA *p*rotecting protein under *s*tarved conditions),[11–13] whereas the *L. innocua* protein (Flp) is a true dodecameric ferritin functioning in iron storage.[2] The FtpA protein from *H. ducreyi* is a structural protein of fine tangled pili.[14] At variance from these Dps proteins, the *H. pylori* homolog HP-NAP appears to display different activities. It induces migration and activation of human neutrophils and monocytes,[15] adhesion of neutrophils to endothelial cells,[16] and it causes mast cell degranulation.[17] HP-NAP binds to neutrophil glycosphingolipids and to mucin, a component of the stomach mucus layer.[18,19] A major property of HP-NAP is that of being highly immunogenic in humans.[20,21] This property is shared by a Dps-like protein, named TpF1, produced by *Treponema pallidum*,[22–25] and therefore, we decided to undertake the determination of the crystal structure of this protein, which is presented in this report.

Methods. The TpF1 gene, amplified by PCR starting from a preparation of *Treponema pallidum* genome, was cloned and expressed in *E. coliXl1blue. E. coli* containing the plasmid pSM214G-TpF1 was grown for 15 h in Luria Bertani medium supplemented with chloramphenicol 15 μg/mL. Details of expression and purification are better described in supplementary material. Briefly, cells were suspended in 10 mL of Tris-HCl 30 mM, pH 7.8, and subjected to three passages through a French press. After fractionated precipitation with ammomium sulphate, the pellet containing the protein was suspended in NaCl 0,1 M, Tris 30 mM, DTT 5 mM, pH 8.4 and purified using an ion-exchange chromatography (MonoQ column, Amersham Biosciences) and a by gel filtration chromatography (superdex 200 HR 10/30 column, Amersham Biosciences).

Crystals were obtained using the vapor diffusion technique with hanging or sitting drops at 20°C, using as precipitant a solution containing 0.1 M Tris buffer, pH 7.5, 10% PEG 6000 or 8000, 8% ethylene glycol. They belong to the trigonal P321 space group. The V_M value of 2.48 is compatible with the presence of one dodecamer and one tetramer in the asymmetric unit, corresponding to a solvent content of about 50%.

Diffraction data were measured at the X-ray diffraction beam-line of the ELETTRA synchrotron in Trieste (Italy). Data were processed with the software MOSFLM[26] and merged with SCALA.[27]

The structure of TpF1 was solved using the molecular replacement method with the program AMoRe,[28] using as a template the model of HP-NAP from *Helicobacter pylori* (PDB code 1JI4).[4] Two different sets of solutions were found, one of them corresponding to a dodecameric molecule in a general position, the other one with the crystallographic threefold axis running through the dodecamer.

Grant sponsor: the Italian Ministero per l'Università e la Ricerca Scientifica (MURST); Grant number: PRIN 2003; Grant sponsor: the Italian National Research Council (CNR); Grant sponsor: the University of Padova.

*Correspondence to: Giuseppe Zanotti, Dipartimento di Scienze Chimiche e ICB-CNR, Università di Padova, Via Marzolo 1, 35131 Padova, Italy. E-mail: Giuseppe.zanotti@unipd.it

Received 2 September 2005; Revised 27 September 2005; Accepted 28 September 2005

Published online 00 Month 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20828

tapraid5/z7e-protein/z7e-protein/z7e00206/z7e2529d06g | heckt | S=6 | 11/17/05 | 16:07 | Art: 20828

TABLE I. Statistics on Data Collection and Refinement

X-ray data	
Space group	P 321
Cell parameters, a, c [Å]	184.92, 154.88
Resolution (Å)	60–2.45 (2.55–2.45)
Independent reflections	111,030 (11,886)
Multiplicity	5.1 (3.9)
Completeness (%)	95.7 (70.8)
$I/\sigma(I)$	8.5 (1.5)
R_{merge}	0.072 (0.37)
Refinement	
Number of residues included	2416
Total number of atoms, including ligands and solvent	19,892
$R_{cryst.}/R_{free}$ (%)	22.5/25.3
Ramachandran plot (%)	
Most favored	97.2
Additionally allowed	2.0
Generously allowed	0.7
Disallowed	0.0
RMS on bonds length (Å), angles (°)	0.007/1.2

Crystals were frozen at 100 K under a nitrogen gas cold stream without the need of any cryoprotectant solution. A wavelength of 1.2 Å was used. A CCD detector was positioned at a distance of 150 mm from the sample. Rotations of 0.5° were performed.

Consequently, the refinement was carried out including one dodecamer and one tetramer in the asymmetric unit, using the software package CNS.[29] Cycles of simulated annealing and energy minimization, followed by manual adjustments, reduced the crystallographic R factor to the final value of 0.225 (R_{free} = 0.253). Statistics and data processing and refinement are reported in Table I.

Results and Discussion. TpF1 is a dodecamer, about 90 Å in diameter, displaying 32 symmetry [Fig. 1(A)]. Each TpF1 subunit folds in a way very similar to the other miniferritins: a four-helix bundle, with helices B and C connected through a long stretch that includes a short helix [Fig. 1(B)]. An alignment search using the TpF1 amino acid sequence(BLAST[30]) shows a high similarity with Dlp2 miniferritin from *B. anthracis* (score 99, 38% identity), and with other proteins of the same family with known structure: HP-NAP from *H. pylori* (score 97), Dlp1 from *B. anthracis* (94), *B. brevis* (82), *S. suis* (79), *L. innocua* (72), *A. Tumefaciens* (54), *E. coli* (45). The comparison of corresponding Cα atoms of TpF1 model with *H. pylori* HP-NAP and Dlp-1 and Dlp-2 from *B. anthracis* gives root-mean-square deviation values of 0.8 and of 0.9 Å, respectively. Major differences are observed in the long connection between helix B and helix C, and minor but significant differences are present in the other loops connecting the helices. Moreover, TpF1 is about 30 residues longer at the N-terminus with respect to most of the other Dps family members. An SDS-PAGE electrophoresis and the N-terminal sequence analysis performed on dissolved crystals (data not shown) shows that the protein present in the crystals has undergone a proteolitic cleavage at residue Ser 22. A partial degradation took place during the purification process, despite the use of protease inhibitors, and proteolysis was completed during the few days necessary for crystal growth. This finding strongly

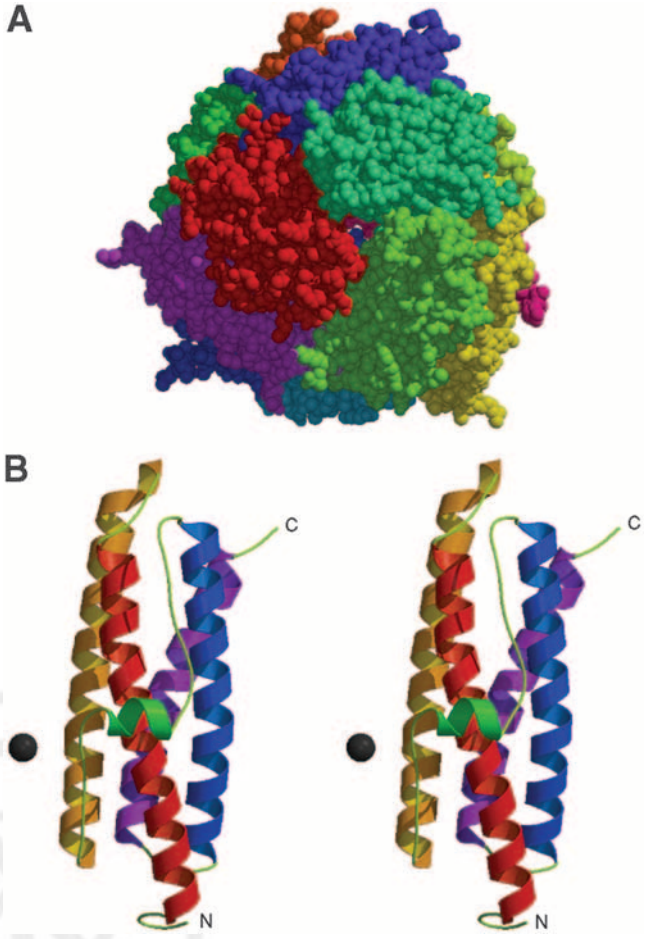


Fig. 1. (**A**) Van der Waals representation of the dodecamer of TPF1. Each subunit is colored differently. One of the threefold axis is running approximately perpendicular to the plane of the paper in the center of the image, through one of the putative tunnel for the iron entrance. (**B**) Stereoview of the ribbon representation of TPF1 monomer. Each α-helix is shown in different colors, Fe(II) ion as a black sphere. N- and C-terminus are labeled N amd C, respectively.

COLOR

suggests that the N-terminal 21 residues are quite flexible. Moreover, the electron density in the crystal is clearly visible only from residue 27 or 28, with the exception of monomer B that starts at residue 22. In the latter, the first six residues are organized as two short β-strands, running antiparallel, connected by a tight β-turn in correspondence of Gly25 and Pro24. The ordering of this chain in monomer B is possibly favored by intermolecular contacts in the crystals; while in the other monomers it is more flexible and not clearly visible in the crystal, at least at this resolution.

The arrangement of the 12 monomers of TpF1 results in the nearly spherical shell typical of miniferritins, with an internal cavity where the iron is stored. The dodecamer possesses four threefold axes, each of them passing through the shell in two different threefold environments that arrange as pores. One of the two threefold pores possibly corresponds to the postulated iron entry channel, because it presents a strongly hydrophilic, negatively charged

environment, contributed by Glu 147 at the entrance, Asp-154 inside the tunnel, and Asp 159 at the end of it, and their symmetry mates. A similar situation is present in other members of the family, like HP-NAP, Dlp1 and 2 and Flp, but the negatively charged residues are not conserved in the amino acid sequence position. The second of the two threefold pores is smaller, but it is not hydrophobic like in others Dps-like proteins, owing to the presence of Glu64, Gln 67, and Lys 66 and symmetry mates. The internal surface of TpF1 also differs from that of other members of the family, because it presents six negatives and two positive charges per monomers pointing towards the interior of the cavity.

Dps-like proteins, like ferritins, bind one Fe ion per monomer. This ion possibly represents the iron oxidation site, but it has also some structural relevance, because it strengthens the interaction among monomers. This Fe(II) atom presents a coordination similar to that of the iron of the other members of the family: the environment of the cation roughly corresponds to a tetrahedral coordination, where three corners of the tetrahedron are occupied by protein atoms (two oxygen, one of Asp 84, and the other of Glu 84, from one monomer and nitrogen of His 57 from another monomer), whereas the fourth coordination position is apparently occupied by a solvent molecule. Atomic details of the metal coordination cannot be described, as the resolution of the model is not high enough. TpF1 displays *in vitro* ferroxidase activity and the presence in the iron coordination of carboxylate and histidine residues suggests that this site functions as a ferroxidase center, where histidines may play a role in the redox process.[31]

Immunogenic properties of the protein are more difficult to rationalize on the basis of the crystal structure: they largely depend from the flexible N-terminal portion, which protrudes from the surface of the spherical shell.

Acknowledgments. We thank the staff of the X-ray diffraction beam-line of ELETTRA, Trieste, Italy, for technical assistance during data measurements.

REFERENCES

1. Grant RA, Filman DJ, Finkel SE, Kolter R, Hogle JM. The crystal structure of Dps, a ferritin homologue that binds and protect DNA. Nat Struct Biol 1998;5:294–303.
2. Ilari A, Stefanini S, Chiancone E, Tsernoglou D. The dodecameric ferritin from listeria innocua contains a novel intersubunit iron-binding site. Nat Struct Biol 2000; 7:38–43.
3. Ilari A, Latella MC, Ceci P, Ribacchi F, Su M, Giangiacomo L, Stefanini S, Chasteen ND, Chiancone E. The unusual intersubunit ferroxidase center of listeria innocua dps is required for hydrogen peroxide detoxification but not for iron uptake. A study with site-specific mutants. Biochemistry 2005;44:5579–5587.
4. Zanotti G, Papinutto E, Dundon WG, Battistutta R, Seveso M, Del Giudice G, Rappuoli R, Montecucco C. Structure of the neutrophil-activating protein from *Helicobacter Pylori*. J Mol Biol 2002;323: 125–130.
5. Papinutto E, Dundon WG, Pitulis N, Battistutta R, Montecucco C, Zanotti G. Structure of two iron-binding proteins from *Bacillus anthracis*. J Biol Chem 2002;277:15093–15098.
6. Zeth K, Offermann S, Essen LO, Oesterhelt D. Iron-oxo clusters biomineralizing on protein surfaces: structural analysis of halobacterium salinarum Dpsa in its low- and high-iron states. Proc Natl Acad Sci USA 2004;101:13780–13785.
7. Ren B, Tibbelin G, Kajino T, Asami O, Ladenstein R. The multi-layered structure of dps with a novel di-nuclear ferroxidase center. J Mol Biol 2003;329:467–477.
8. Ceci P, Ilari A, Falvo E, Chiancone E. The Dps protein of *Agrobacterium tumefaciens* does not bind to DNA but protects it toward oxidative cleavage: X-ray crystal structure, iron binding, and hydroxyl-radical scavenging properties. J Biol Chem 2003;278: 20319–20326.
9. Kauko A, Haataja S, Pulliainen A, Finne J, Papageorgiou A. Crystal structure of streptococcus suis dps-like peroxide resistance protein dpr: implications for iron incorporation. J Mol Biol 2004;338:547–558.
10. Roy S, Gupta S, Das S, Sekar K, Chatterji D, Vijayan M. X-Ray Analysis of *Mycobacterium smegmatis* Dps and a comparative study involving other Dps and Dps-like molecules. J Mol Biol 2004;339:1103–1113.
11. Almiron M, Link AJ, Furlong D, Kolter R. A novel DNA-binding protein with regulatory and protective roles in starved *Escherichia coli*. Genes Dev 1992;6:2646–2654.
12. Chen L, Helmann JD. R. Bacillus subtilis MrgA is a Dps(PexB) homologue: evidence for metalloregulation of an oxidative-stress gene. Mol Microbiol 1995;18:295–300.
13. Antelmann H, Engelmann S, Schmid R, Sorokin A, Lapidus A, Hecker M. Expression of a stress- and starvation-induced dps/pexB-homologous gene is controlled by the alternative sigma factor sigmaB in Bacillus subtilis. J Bacteriol 1997;179:7251–7256.
14. Brentjens RJ, Ketterer M, Apicella M A, Spinola SM. Fine tangled pili expressed by *Haemophilus ducreyi* are a novel class of pili. J Bacteriol 1996;178:808–816.
15. Satin B, Del Giudice G, Della Bianca VG, Dusi S, Laudanna C, Tonello F, Kelleher D, Rappuoli R, Monetcucco C, Rossi F. The neutrophil activating protein (HP-NAP) of *Helicobacter pylori* is a protective antigen and a major virulence factor. J Exp Med 2000;191:1467–1476.
16. Evans DJ Jr, Evans DG, Takemura T, Lampert HC, Nakano H. Identification of four new prokaryotic bacterioferritins, from *Helicobacter pylori*, *Anabaena variabilis*, *Bacillus subtilis* and *Treponema pallidum*, by analysis of gene sequences. Gene 1995;153:123–127.
17. Montemurro P, Nishioka H, Dundon WG, De Bernard M, Del Giudice G, Rappuoli R, Montecucco C. The Neutrophil activating protein (HP-NAP) of *Helicobacter pylori* is a potent stimulant of mast cells. Eur J Immunol 2002;32:671–676.
18. Teneberg S, Miller-Podraza H, Lampert HC, Evans DJ Jr, Evans DG, Danielsson D, Karlsson KA. Carbohydrate binding specificity of the neutrophil-activating protein of *Helicobacter pylori*. J Biol Chem 1997;272:19067–19071.
19. Namavar F, Sparrius M, Veeman EC, Appelmelk BJ, Vandenbroucke-Grauls CM. Neutrophil-activating protein mediates adhesion of *Helicobacter pylori* to sulfated carbohydrates on high-molecular-weight salivary mucin. Infect Immun. 1998;66:444–447.
20. Del Giudice G, Covacci A, Telford JL, Montecucco C, Rappuoli R. The design of vaccines against *Helicobacter pylori* and their development. Annu Rev Immunol 2001;19:523–563.
21. Tomb, JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Gill S, Dougherty BA, Nelson K, Quackenbush J, Zhou L, Kirkness EF, Peterson S, Loftus B, Richardson D, Dodson R, Khalak HG, Glodek A, McKenney K, Fitzegerald LM, Lee N, Adams MD, Hickey EK, Berg DE, Gocayne JD, Utterback TR, Peterson JD, Kelley JM, Cotton MD, Weidman JM, Fujii C, Bowman C, Watthey L, Wallin E, Hayes WS, Borodovsky M, Karp PD, Smith HO, Fraser CM, Venter JC. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 1997;388:539–547.
22. Fehniger TE, Walfield AM, Cunningham TM, Radolf DJ, Miller JN, Lovett MA. Purification and characterization of a cloned protease-resistant *Treponema pallidum*-specific antigen. Infect Immun 1984;46:598–607.
23. Coates SR, Sheridan PJ, Hansen DS, Laird WJ, Erlich HA. Serospecificity of a cloned protease-resistant *Treponema pallidum*-specific antigen expressed in *Escherichia coli*. J Clin Microbiol 1986;23:460–464.
24. Bellini AV, Galli G, Fascetti E, Frascotti G, Branduzzi P, Lucchese G, Grandi G. Production processes of recombinant IL-1 beta from

Bacillus subtilis: comparison between intracellular and exocellular expression. J Biotechnol 1991;18:177–192.
25. Noordhoek GT, Hermans PW, Paul AN, Schouls LM, van der Sluis JJ, van Embden JD. *Teponema pallidum* subspecies pallidum (Nichols) and *Treponema pallidum* subspecies pertenue (CDC 2575) differ in at least one nucleotide: comparison of two homologous antigens. Microb Pathog 1989;6:29–42.
26. Leslie AGW. In: Moras D, Podjarny AD, Thierry JP, editors. Crystallographic computing V. Oxford: Oxford University Press; 1991. p 27–38.
27. Collaborative Computational Project Number 4. The CCP4 suite. Acta Crystallogr 1994;D50:760–763.
28. Navaza, J. On the computation of the fast rotation function. Acta Crystallogr 1994;A50:157–163.
29. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang J-S, Kuszewski J, Nilges N, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. Acta Crystallogr 1998;D54:905–921.
30. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
31. Wade VJ, Levi S, Arosio P, Treffry A, Harrison PM, Mann S. Influence of site-directed modifications on the formation of iron cores in ferritin. J Mol Biol 1991;221:1443–1452.

AQ: 1 AQ: 2

SUPPLEMENTARY MATERIAL

CLONING, EXPRESSION AND PURIFICATION

TpF1 was cloned and expressed in *E. coliXl1blue*. *TpF1* gene was amplified by PCR starting from a preparation of *Treponema pallidum* genome. The PCR reaction was carried out using standard method. The thermal cycling parameters were as follows: 5 min at 94°C, 30 cycles of 1 min at 94°C, 1 min at 50°C, and 45 s at 72°C, and a final extension cycle of 10 min at 72°C. Primers used were TP1: 5'-ccggaattcacgatgaacatgtgtaca-3' and TP2: 5'-cccaagcttctaggctttcagggtagc-3', containing restriction site for *Eco*RI and *Hin*dIII, respectively. The amplified fragment was excised by digestion with *Eco*RI and *Hin*dIII and ligated into *Eco*RI and *Hin*dIII sites of the expression vector pSM214G. pSM214G contains an artificial constitutive promoter, a chloramphenicol resistance cassette, and two origins of replication that allow expression of cloned genes both in *E. coli* and in *Bacillus subtilis*.

E. coli containing the plasmid pSM214G-TpF1 was grown for 15 hours in Luria Bertani medium supplemented with chloramphenicol 15 µg/ml. The cells were pelleted by centrifugation at 6000 x g and suspended in 10 ml of Tris-HCl 30 mM pH7.8 plus protease inhibitors (Roche) for 500 ml of culture. After three passages through a French press and removal of debris by centrifugation at 32000 X g, a solution of saturated ammonium sulfate was added to a final concentration of 12,5% w/v, at 4°C. At this percentage of ammonium sulfate most of the protein remained in solution. After 3 h at 4°C at slow stirring the sample was centrifuged at 32000 x g for 30 min, the supernatant was recovered and ammonium sulfate was added to a final concentration of 22,5% w/v. The sample was kept for 3 h at at 4°C at slow stirring and then centrifuge at 32000xg for 30 min. The pellet containing the protein TpF1 was suspended in NaCl 0,1 M, Tris 30 mM, DTT 5 mM, pH 8,4 (buffer A) and dialyzed overnight in buffer A. The sample was fractionated by ion-exchange chromatography using a MonoQ column (Amersham Biosciences) equilibrated with buffer A. After the sample was applied, the column was eluted with a linear NaCl gradient in Tris 30mM, DTT 5 mM, pH 8,4. Fractions were analysed by SDS-PAGE and TpF1-containing fractions were pooled. TpF1 was further purified by gel filtration chromatography using a superdex 200 HR 10/30 column (Amersham Biosciences) equilibrated with phosphate buffer saline, pH 7,8.

Crystals were obtained using the vapor diffusion technique with hanging or sitting drops at 20 °C, using as precipitant a solution containing 0.1 M Tris buffer, pH 7.5, 10% PEG 6000 or 8000, 8% ethylene glycole. They belong to the trigonal P321 space group. The V_M value of 2.48 is compatible with the presence of one dodecamer and one tetramer in the asymmetric unit, corresponding to a solvent content of about 50%.

DIFFRACTION DATA COLLECTION

Diffraction data were measured at the x-ray diffraction beam-line of the ELETTRA synchrotron in Trieste (Italy). A crystal was frozen at 100 °K under a nitrogen gas cold stream without the need of any cryoprotectant solution. For the measurements a wavelength of 1.2 Å was selected. A CCD detector (MAR Research, …) was positioned at a distance of 150 mm from the sample, corresponding to a maximum resolution of 2.45 Å. Rotations of 0.5 were performed. Data were processed with the software MOSFLM (Leslie, 1991) and merged with SCALA (CCP4).

STRUCTURE SOLUTION AND REFINEMENT

The structure of TpF1 was solved using the molecular replacement method with the program AMoRe (Navaza, 1994). The search for the rotation and translation function was performed at 4 Å resolution using as templates the models of HP-NAP from *Helicobacter pylori* (PDB code 1JI4, Zanotti et al., 2003). Two different sets of solutions were found, one of them corresponding to a dodecameric molecule in a general position, the other one with the crystallographic 3-fold axis running through the dodecamer. Consequently, the refinement was carried out using one dodecamer and one tetramer in the asymmetric unit. In the initial stages of refinement the strict non-crystallographic symmetry, as implemented in the software package CNS (Brünger et al., 1998), was used, whilst in the final stages restraints were imposed. Cycles of simulated annealing and energy minimization, followed by manual adjustments, reduced the crystallographic R factor to the final value of 0.225. (Rfree = 0.253). Water molecules were introduced in peaks of electron density close to hydrophilic residues and forming possible hydrogen bonds.

Ringraziamenti

Desidero ringraziare il prof. Giuseppe Zanotti per avermi coinvolto in questo progetto, per avermi accordato una grande fiducia e libertà, e sostenuto con il suo inguaribile ottimismo. Mi sento decisamente fortunato ad esser stato accolto nel suo gruppo e ad aver potuto lavorare con lui.

Vorrei ringraziare Diego Frezzato, i prof. Giorgio Moro e Alberta Ferrarini per le utili discussioni che abbiamo avuto modo di fare, e per l'interesse che hanno dimostrato per il mio lavoro. E i ragazzi di Chimica Fisica, in particolare Mirko, Mirco e Fabio, che mi hanno in più occasioni dato una mano e dedicato del tempo.

Un grazie particolare a Jacopo, come amico e scienziato, per avermi sempre incoraggiato e stimolato a proseguire su questa strada, con discussioni notturne e interminabili, e per tutto il resto. (Ti sono ancora debitore di un seminario!)

Un pensiero riconoscente a tutte le persone con cui mi sono trovato a lavorare accanto in questi anni e che ho avuto modo di apprezzare umanamente e scientificamente: Nicola “il Pasqui”, Laura Cendron, Nicola Barison, Ale Angelini, Zulia, Laura Fonso, Anke, Elisa, Tommaso T.; Tommaso S. con il quale ho lavorato nei primi mesi; Enrico, Giorgia, Ivan; Elena Papinutto, con cui ho svolto l'unica parte sperimentale del lavoro; Marco, i prof. Paola Spadon e Roberto Battistutta, Elena Vaki, Lorenza, Patrick, Roberto Spricigo, e svariati altri che mi sto dimenticando. Sono contento di avervi conosciuti e che, in un modo o nell'altro, le nostre strade si siano incrociate qui!

Grazie alle persone che hanno reso questi anni e questa città speciali e importanti per me: la mia “terza sorella” Sophi, Claudia e Ina della “casa dell'acqua”; il mio paziente coinquilino Nicola (the projectionist) con cui ho vissuto gli ultimi 4 anni; le persone con cui ho condiviso la passione per la musica: Pieze, Andrea, Marco Q., Laura e i miei insegnanti Gastone e Alfonso. Grazie a Carla per tutto ciò che abbiamo vissuto insieme e per avermi convinto a riprendere gli studi.

Un ringraziamento a tutti i miei compagni di fede buddista, con cui ho condiviso le lotte di questi anni; non vi nomino uno a uno perché siete tanti ma vi sento profondamente vicini. Grazie a “sensei” Daisaku Ikeda, da cui ho deciso di imparare ogni giorno a trasformare le difficoltà in una fonte di gioia e cambiamento e ad essere sinceramente ciò che già sono.

Grazie, infine, ai miei genitori Gianni e Flaminia, che da quando sono venuto al mondo mi hanno sempre aiutato a seguire le mie inclinazioni e stranezze; alle mie sorelle Candida e Martha a cui voglio dedicare questo mio lavoro e auguro una vita piena, felice e realizzata.