X-RAY CRYSTALLOGRAPHIC INVESTIGATIONS OF THE STRUCTURES OF ENZYMES OF MEDICAL AND BIOTECHNOLOGICAL IMPORTANCE

by

Richard Lawrence Kingston

A dissertation submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

in the

Department of Biochemistry

at

MASSEY UNIVERSITY, NEW ZEALAND

November, 1996

ABSTRACT

This thesis is broadly in three parts. In the first, the problem of identifying conditions under which a protein will crystallize is considered. Then structural studies on two enzymes are reported, glucose-fructose oxidoreductase from the bacterium *Zymomonas mobilis*, and the human bile salt dependent lipase (carboxyl ester hydrolase).

The ability of protein crystals to diffract X-rays provides the experimental data required to determine their three dimensional structures at atomic resolution. However the crystallization of proteins is not always straightforward. A systematic procedure to search for protein crystallization conditions has been developed. This procedure is based on the use of orthogonal arrays (matrices whose columns possess certain balancing properties). The theoretical and practical background to the problem is discussed, and the relationship of the presented procedure to other published search methods is considered.

The anaerobic Gram-negative bacterium *Zymomonas mobilis* occurs naturally in sugar-rich growth media, and has attracted much interest because of its potential for industrial ethanol production. In this organism the periplasmic enzyme glucose-fructose oxidoreductase (GFOR) is involved in a protective mechanism to counter osmotic stress. The enzyme is unusual in that it contains tightly associated NADP which is not released during its catalytic cycle. The crystal structure of *Z. mobilis* GFOR has been determined by the method of multiple isomorphous replacement, and refined by restrained least squares methods using data extending to an effective resolution of 2.7 Å. The structure determination reveals that each subunit of the tetrameric protein is folded into two domains, one of which is the classical dinucleotide binding domain, or Rossmann fold. The C-terminal domain is a nine-stranded predominantly antiparallel β-sheet around which the tetramer is constructed. Preceding the Rossmann fold there is a 30 amino acid proline rich 'arm' which wraps around an adjacent subunit in the tetramer. The N-terminal arm buries the adenine ring of the NADP, and may also be involved in stabilization of the quaternary structure of the enzyme. The tight association of NADP is accounted for by the structure. An unsuspected structural relationship has been discovered between GFOR and the cytoplasmic enzyme glucose-6-phosphate dehydrogenase (G6PD). It is proposed that GFOR and G6PD derive from an common ancestral gene, and GFOR has evolved to allow it to function in the bacterial periplasm where it is required.

The human bile salt dependent lipase (BSDL) is secreted by the pancreas into the digestive tract, and by the lactating mammary gland into human milk, and is integral to the effective absorption of dietary lipids. It is markedly non-specific, and as its name implies is only active against water-insoluble substrates in the presence of primary bile salts. This differentiates the

enzyme from conventional lipases. Diffraction data has been collected from crystals of native BSDL (isolated from human milk), and from crystals of recombinant BSDL (including a truncated variant which lacks a C-terminal heavily glycosylated tandem repeat region found in the native enzyme). The structure of the truncated variant has been partially determined at 3.5 Å resolution, by the method of molecular replacement. The recent collection of a higher resolution (2.8 Å) data set should allow the completion of the structure. The current status of the crystallographic investigations of the human bile salt dependent lipase are reported.

ACKNOWLEDGMENTS

I thank my principal supervisor, Professor Ted Baker for his friendship, his enthusiasm for science, and for allowing me much freedom to follow my ideas. I also thank Mrs. Heather Baker for her help and constant encouragement.

I thank my assistant supervisors, Dr. Bryan Anderson (for many discussions about crystallographic computing), and Professor Sylvia Rumball (who initiated a collaboration with Umeå University on bile salt dependent lipase).

I would like to specifically acknowledge the scientific contributions of Dr. Rick Faber, and Dr. Stanley Moore to the work presented in this thesis (when they weren't fly fishing, that is).

I would like to thank all the other friends I have made during my time at Massey. Andrew, Rosemary, Isobel, Shaun, Treena, Ross, Catherine, Alain and Anne-Gael, Mark, Michelle, Jakki and Paul, Neil and Liz, Phil, and Maria to name only some. Thanks.

The work on bile salt dependent lipase was carried out in collaboration with Professor Olle Hernell and Dr. Lars Bläckberg (Umeå University, Sweden); and Dr. Kerry Loomes (Auckland University, New Zealand). The work on glucose-fructose oxidoreductase was carried out in collaboration with Professor Robert Scopes (La Trobe University, Australia). I thank these people for their scientific contributions, and especially Kerry, for believing we could solve the structure and working so hard to overcome difficult technical problems.

For financial assistance while this work was completed I thank Massey University (through the award of a doctoral scholarship), and latterly Professor Ted Baker.

Thanks to the members of my family for encouragement and support. I owe much to my parents for supporting what I do, and helping to finance my study.

Finally I would like to thank Wendy, a very special friend, and someone whom it has been difficult to be separated from while this thesis was written

TABLE OF CONTENTS

Chapter 2
GFOR: STRUCTURE DETERMINATION

Chapter 3
GFOR: STRUCTURE AND FUNCTION

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

ABBREVIATIONS

AChE Acetylcholinesterase
AMPSO 3-[(1,1-Dimethyl-2-hydroxyethyl)amino]2-hydroxypropanesulfonic acid
BIS-TRIS PROPANE 1,3-bis[tris(Hydroxymethyl)-methylamino]propane
BSDL Bile salt dependent lipase
BSSL Bile salt stimulated lipase
CDL Colipase-dependent lipase
CHAPS 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate
CRL *Candida rugosa* lipase
DHPR Dihydrodipicolinate reductase
DNA Deoxyribonucleic acid
EPPS N-[2-Hydroxyethyl]piperazine-N′[3-propanesulfonic acid]
FAD Flavin-adenine dinucleotide
FMN Flavin mononucleotide
G6PD Glucose-6-phosphate dehydrogenase
GAPDH Glyceraldehyde-3-phosphate dehydrogenase
GCL *Geotrichum candidum* lipase
GFOR Glucose-fructose oxidoreductase
HEPES N-[2-Hydroxyethyl]piperazine-N'-[2-ethanesulfonic acid]
IEF Isoelectric focussing
LDH Lactate dehydrogenase
MDH Malate dehydrogenase
MES 2-[N-Morpholino]ethanesulfonic acid
MIR Multiple isomorphous replacement
MOPS 3-[N-Morpholino]propanesulfonic acid
NAD Oxidized or reduced form of nicotinamide adenine dinucleotide
NADP Oxidized or reduced form of nicotinamide adenine dinucleotide phosphate
$NADP^+$ Oxidized form of nicotinamide adenine dinucleotide phosphate
NADPH Reduced form of nicotinamide adenine dinucleotide phosphate
NAD(P) NAD or NADP
NCBI National Center for Biotechnology Information
NCS Non-crystallographic symmetry
NIST National Institute of Standards and Technology
PCR Polymerase chain reaction
PEG Polyethylene glycol
PEG-mme Polyethylene glycol monomethyl ether

PIPES 1,4-Piperazinediethanesulfonic acid
PQQ Pyrrolo-quinoline quinone
RMS Root mean square
SEL Sequential elimination of levels
SIR Single isomorphous replacement
TAPS N-tris[Hydroxymethyl]methyl-3-aminopropanesulfonic acid
TcAChE *Torpedo californica* acetylcholinesterase
TRIS Tris(hydroxymethyl)aminomethane

RELATED PUBLICATIONS

Some of the material presented in this thesis has already been published, or has been accepted for publication.

Kingston, R.L., Baker, H.M. & Baker, E.N. (1994) Search designs for protein crystallization based on orthogonal arrays. **Acta Crystallographica. D50**, 429-440.

Kingston, R.L., Scopes, R.K. & Baker, E.N. (1996) The structure of glucose fructose oxidoreductase from *Zymomonas mobilis*: an osmoprotective periplasmic enzyme containing non-dissociable NADP. **Structure**. in Press.

Chapter 1

PROTEIN CRYSTALLIZATION

1.1 INTRODUCTION

1.1.1 Historical background

The ability of protein crystals to diffract X-rays provides the experimental data required to determine their three-dimensional structures at atomic resolution. However, protein crystallization was being used as a technique for the isolation and purification of proteins well before the advent of X-ray diffraction, and has a long and interesting history [1].

By the turn of the nineteenth century the crystallization of hemoglobin, and of the plant seed globulins, had already been extensively studied. Early in this century, the American biochemist James Sumner set himself the task of isolating and purifying an enzyme. At the time little was known about the chemistry of enzymes, and it had not even been established that they were proteins. Sumner spent nine discouraging years attempting to crystallize the plant enzyme urease before succeeding and publishing his findings [2]. This work and the crystallization of the enzyme pepsin, by John Northrop at the Rockefeller Institute [3] established the chemical nature of enzymes. During the 1930's a number of other proteins and viruses were isolated and crystallized. By 1948, it was able to be reported that some forty enzymes 'have been isolated and crystallized and have been found to be proteins' [4].

Across the Atlantic, the modern physical significance of protein crystallization was being gradually realized (see [5, 6, 7] for historical reviews). In 1934 at Cambridge, John Bernal and Dorothy Crowfoot had obtained X-ray diffraction photographs from crystalline pepsin, the first recorded diffraction pattern from a crystalline protein. They concluded that the means now existed for 'arriving at far more detailed conclusions about protein structure than previous physical or chemical methods have been able to give' [8]. Max Perutz, who joined Bernal as a graduate student in 1936, later noted 'his visionary faith in the power of X-ray diffraction to solve the structures of molecules as large and complex as enzymes or viruses at a time when the structure of ordinary sugar was still unsolved' [7]. The work of the Cambridge group culminated in the elucidation of the first protein crystal structures, those of the oxygen carrying proteins myoglobin and hemoglobin [9, 10].

1.1.2 The experimental problem today

The structural information provided by X-ray crystallographic techniques has revolutionized biochemistry. Recent years have seen a massive increase in the number of protein crystal structures determined. However, a primary difficulty with the application of this technique is that it depends on obtaining large, well ordered protein crystals. As the results of the early biochemists suggested, some proteins are readily crystallized. Regrettably, this is not always the case, and protein crystallization can be both time-consuming and frustrating. The importance of this problem is reflected in the large number of recent reviews on this topic [11, 12, 13, 14, 15, 16, 17, 18, 19]).

For a water-soluble protein, a typical crystallization trial will involve at least three solution components; the protein, a buffer to control the solution pH, and a precipitant to exclude the protein from solution. What must be experimentally determined are an appropriate buffer and precipitant together with the ranges of temperature, pH and concentration of the solution components that will support nucleation and crystal growth. Given the small amounts of protein often available, the need for efficient and economical search experiments is evident. A number of papers have described experimental approaches to the problem of searching for protein crystallization conditions [20, 21, 22, 23, 24, 25, 26, 27]. Some of these approaches are based on formal ideas of experimental design, others are simply collections of solution conditions that have worked previously. The purpose of this chapter is to suggest a systematic search method based on the use of orthogonal arrays (a special class of matrices), and to discuss the methods currently used to search for crystallization conditions with reference to the statistical literature.

It should be emphasized that even if only small amounts of protein are available, this will still be enough to perform a systematic search using several classes of protein precipitants. Using the hanging drop vapour diffusion technique, it is possible to carry out a single crystallization trial with 1 μL of protein solution. If there is, for example, only 2 mg of purified protein available, and the protein is concentrated to 10 mg/ml (which is fairly typical), this is enough material to perform some 200 trials, which allows for a relatively systematic search.

1.1.3 Physical background

Protein crystals form in supersaturated solutions in which the protein concentration exceeds its equilibrium solubility. Hence all the physical techniques for crystallizing proteins involve bringing a protein solution into the supersaturated state by alteration of some property of the

system. Typically this is accomplished by the gradual introduction of substances which serve to reduce protein solubility (protein precipitants), via some diffusive process. Salts, simple organic compounds and long chain synthetic polymers have all been used for this purpose.

From a supersaturated solution, equilibrium can be restored by phase separation. A solid phase can result from the formation of disordered protein aggregates, leading to an amorphous precipitate or flocculate. Alternatively, the formation of ordered aggregates leads (once a critical size is reached) to the nucleation and growth of protein crystals [28]. Protein crystals will only form given an appropriate degree of supersaturation, and a suitable physical and chemical environment. The initial experimental problem is to establish the environmental conditions favouring the formation of protein crystals rather than an amorphous solid or flocculate. This is essentially a search problem. Subsequent to this is the problem of producing the large single crystals required for X-ray diffraction. For most laboratories this experimental program is carried out using one of the available micro-methods for protein crystallization (see [29]), and examining the results with an optical microscope.

The experimental difficulties that are often associated with protein crystallization arise from a number of sources.

Proteins must be isolated and purified from a complex biological medium. The difficulties associated with doing this reproducibly and consistently are often considerable.

There are also properties that are intrinsic to the structure of proteins that make them much more difficult than small organic and inorganic molecules to crystallize. In comparison with small molecules, they are much larger, have a lower degree of symmetry, and typically make a relatively small number of contacts within the crystal lattice [30]. The nature of the interactions which direct the association of proteins into ordered and repetitive arrays are still poorly understood. It has been proposed that when associating, proteins have a large number of potential attachment sites that are energetically almost as favourable as the small number of specific sites through which a crystal is formed [28]. Consequently amorphous rather than crystalline aggregates are frequently formed. The complexity of the problem is further increased as the physical properties of proteins themselves, such as charge distribution, conformation, state of association and stability may depend on the physical environment, with obvious implications for crystallization.

When crystallization does occur, terminal crystal size will depend both on the rate of crystal nucleation and of crystal growth. Both these rate processes have been shown to have a functional dependency on the degree of supersaturation in a variety of systems [18]. The rate of

nucleation is typically an exponential function of the degree of supersaturation, which accounts in part for the frequently observed sensitivity of crystallization experiments to changes in the environmental variables. In addition, many of the experimentally controllable factors will influence the crystallization process in a complex fashion. For example, a change in temperature will affect not only the protein solubility (and hence the degree of supersaturation in a given system), but also the rate of the diffusion processes often used to bring the protein into the supersaturated state.

Finally, using a light microscope to inspect the results of crystallization trials, the amount of information that can be obtained is limited. In the case of solid phase formation, it is often difficult to distinguish between an amorphous and a micro-crystalline solid phase. When a solid phase does not result then there is no way to determine, by sight, how close the system is to supersaturation or whether an amorphous or crystalline solid is likely to result should the system be driven further towards supersaturation. In this situation there is no effective method of measuring an experimental response. Even in the favourable cases where protein crystals are clearly distinguishable, it is difficult to assess the results in a quantitative fashion (although use can be made of some observable physical property such as crystal size, regularity of growth, or number of crystals per unit volume). This situation contrasts with that of a typical controlled experiment where a result can be observed and quantified for every applied treatment.

1.2 SEARCH DESIGNS FOR PROTEIN CRYSTALLIZATION

It is with the initial problem of searching for protein crystallization conditions that this chapter is concerned. This can be a formidable experimental problem. The principal difficulties are the size of the experimental region (over much of which crystal nucleation and growth may not be observed at all), the physical complexity of the situation, and the limited information available from the crystallization experiments themselves. All of this is coupled with the small amounts of protein often available.

1.2.1 Terms associated with experimental design

Before proceeding further, it is necessary to clearly define some general terms associated with experimental design. The experimental factors are the variables which influence (or are believed to influence) the attribute of interest in the experiment (for protein crystallization this will include factors such as the solution pH, the nature of the protein precipitant, the concentration of the protein and the protein precipitant, etc.). Factors can be continuous (that is, having a numerical value, e.g. temperature, pH) or discrete (that is, having a non-numerical

value from a finite set of values, e.g. precipitant type). In any experiment the effects of a factor will be evaluated at a number of levels (in the case of continuous factors), or over a number of classes or categories (in the case of discrete factors). However the term 'level' is often applied to both continuous and discrete factors. A factorial experiment simply consists of all possible level combinations of all the factors included in the investigation.

1.2.2 Current approaches to searching for protein crystallization conditions

There are a vast number of publications whose principal concern is to report crystallization conditions and preliminary crystallographic investigations of proteins (much of this information can be accessed through the NIST Biological Macromolecule Crystallization Database). However the majority of these do not discuss how crystallization conditions were arrived at. The general approach of most laboratories to the identification and optimization of crystallization conditions can be described as follows. Firstly, an initial search experiment is executed to try and determine a physical and chemical environment which will support nucleation and crystal growth. In addition to the experiments referenced previously (Section 1.1.2), in many laboratories search experiments seem to have been loosely based on a factorial structure (such experiments are often described as being based on multivariate arrays, matrices or grids). Following this initial experiment (assuming solution conditions supporting crystal growth can be identified) these solution conditions are 'optimized'. Typically this is done in iterative fashion, by simple grid searching, in which the levels of the experimental factors are centred around the best levels identified in the previous experiment, with the levels of the continuous factors spaced increasingly finely in each subsequent experiment.

In an aside, iterative optimization methods like this, which are not based on any formal model of the response, have not been extensively studied in the statistical literature. Underlying much of the statistical research work in experimental design has been the assumption that the experiment can be adequately described by a statistical model [31, 32, 33]. Recently Wu and coworkers [34] considered some physical problems for which this assumption might not hold, and proposed a quite general class of model free optimization methods, which they termed SEL (sequential elimination of levels). The procedure given above is closely related to these methods.

The nature of the initial search experiment is critical, since without the identification of conditions supporting crystal growth, no subsequent optimization can be undertaken. The problem of optimizing protein crystallization conditions is not considered in detail here. It should be noted that even given seemingly appropriate starting points, it may not be possible to grow

crystals of a suitable quality for X-ray diffraction by simple diffusive techniques coupled with manipulation of the environmental variables. The potential importance of seeding procedures [26] or of techniques for protein modification (for example proteolytic cleavage, or enzymatic deglycosylation [35]) should not be overlooked.

1.2.3 General criteria for initial search experiments

Several criteria can be proposed which an initial search experiment should realize.

Firstly, the experiment needs to be of practicable size. The amount of purified protein available for crystallization studies is often quite limiting. Very large initial search experiments, involving many trials conducted in parallel, will typically be inefficient. This inefficiency arises from the failure to exploit information which could have been obtained if a more sequential strategy had been adopted. Sequential experimental procedures can have an important impact on the required experimental size [36]. Since the time in which results from crystallization trials can be expected is relatively short (typically days or weeks), use of sequential experimentation is realistic.

Secondly, the experiment should clearly define the experimental region, and explore it comprehensively. The requirements to limit the experimental size and to perform a comprehensive search are exclusive of one another, and a suitable compromise will always need to be found. Search experiments having a factorial structure clearly define the experimental region, with the probability of locating regions which will support protein crystallization (if they exist) related to the spacing of the levels of the continuous factor(s) (e.g. pH, precipitant concentration). This is the reasoning behind the successive automated grid searches described by Weber [27]. However as the number of factors increases, the size of factorial experiments can quickly become very large (the size being the multiple of the number of levels of each factor).

Thirdly, an initial search experiment should employ the simplest possible physical system in attempts to crystallize the protein. Some approaches to the identification of protein crystallization conditions employ three or four component physical systems, which in addition to a buffer and protein precipitant contain small organic molecules or inorganic salts (see e.g. [23, 25]). The inclusion of additional solution components in this fashion serves to unneccesarily complicate the experiment. The impact of adding such compounds can usually be assessed at a later stage, when crystallization conditions have been identified (see e.g. [37]). There are examples in which additives are found to be absolutely required for the crystallization process (e.g. Cd^{2+} in the crystallization of ferritin [38]). Generally however such

examples seem to be atypical, and if initial experiments based on simple systems fail, experiments employing more complicated physical systems can always be executed. The addition of additives on a 'trial-and-error' basis is distinguished from the situation in which a protein is known to have, for example, a metal ion dependency; in this case the implications for crystallization need to be carefully considered when the initial search experiments are planned.

Finally, the experiment should be flexible. Proteins may vary markedly in their stability and ability to maintain biological activity as a function of temperature, pH and solvent composition. As a consequence of this dissimilar physical behaviour, it is difficult to design initial search experiments which will have a widespread applicability. However there are practical advantages associated with the use of standard initial search designs; it is time consuming to develop a completely new experiment for every protein crystallization problem. Evidently a standard initial search design should be sufficiently flexible to accommodate a likely diversity of physical behaviour. One way to achieve this flexibility is to ensure that the overall design can be readily partitioned into a number of smaller designs (i.e. is modular). For example a search experiment might be constructed so that it can be seen as being comprised of a set of smaller experiments, each covering a limited pH range. This enables a quite general initial search experiment to be adapted to a specific protein crystallization problem.

1.2.4 Orthogonal arrays

Full factorial experiments (multivariate arrays or matrices) are attractive because they clearly define and explore the experimental region. However, the consideration of even a small number of possibilities for the solution composition will result in factorial experiments that are very large. There may also be some redundancy within the factorial structure (in the sense that crystal nucleation and growth may not be critically dependent on some of the factors considered). For example, the crystallization of a protein may not be pH dependent. This idea has been long recognized in the statistical literature where it is sometimes termed effect sparsity.

Therefore it is logical to consider whether a subset of a factorial experiment might be selected which is in some way representative of the full factorial arrangement. This subset can be used as the basis of an initial search experiment. If this initial experiment is successful, no further searching may be required. Even if this is not the case then the information gained (principally concerning the protein solubility) serves to illuminate the problem, and can be used in designing better subsequent experiments. This relates directly to the notion of limiting the experimental size (and of sequential experimentation) which was discussed above.

Here, we consider the selection of subsets using a special class of matrices, termed orthogonal arrays.

The concept of an orthogonal array is best introduced by a simple example. Consider a 2 x 2 x 2 factorial (this involves 3 factors, each at two levels). There are a total of 8 possible level combinations. Following convention, the levels of each factor are denoted with the integers 0 and 1. Then the factorial structure can be represented by the following 8 x 3 matrix.

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \textit{Eq. 1.1}$$

The columns of this matrix correspond to the experimental factors, the entries in the columns correspond to the levels of the factors, and the rows correspond to the level combinations. This matrix is represented geometrically in Figure 1.1. Now consider the following 4 x 3 matrices

$$B = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \qquad \textit{Eq. 1.2}$$

$$C = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix} \qquad \textit{Eq. 1.3}$$

$$D = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \qquad \textit{Eq. 1.4}$$

The rows of each of these matrices are a subset of those in the preceding 8 x 3 matrix (Equation 1.1). Again they are geometrically represented in Figure 1.1. Intuitively matrix (D) corresponds to the most 'uniform' distribution of points, and is the most representative of the full factorial arrangement. This matrix is termed an orthogonal array, an idea which is now developed more formally.

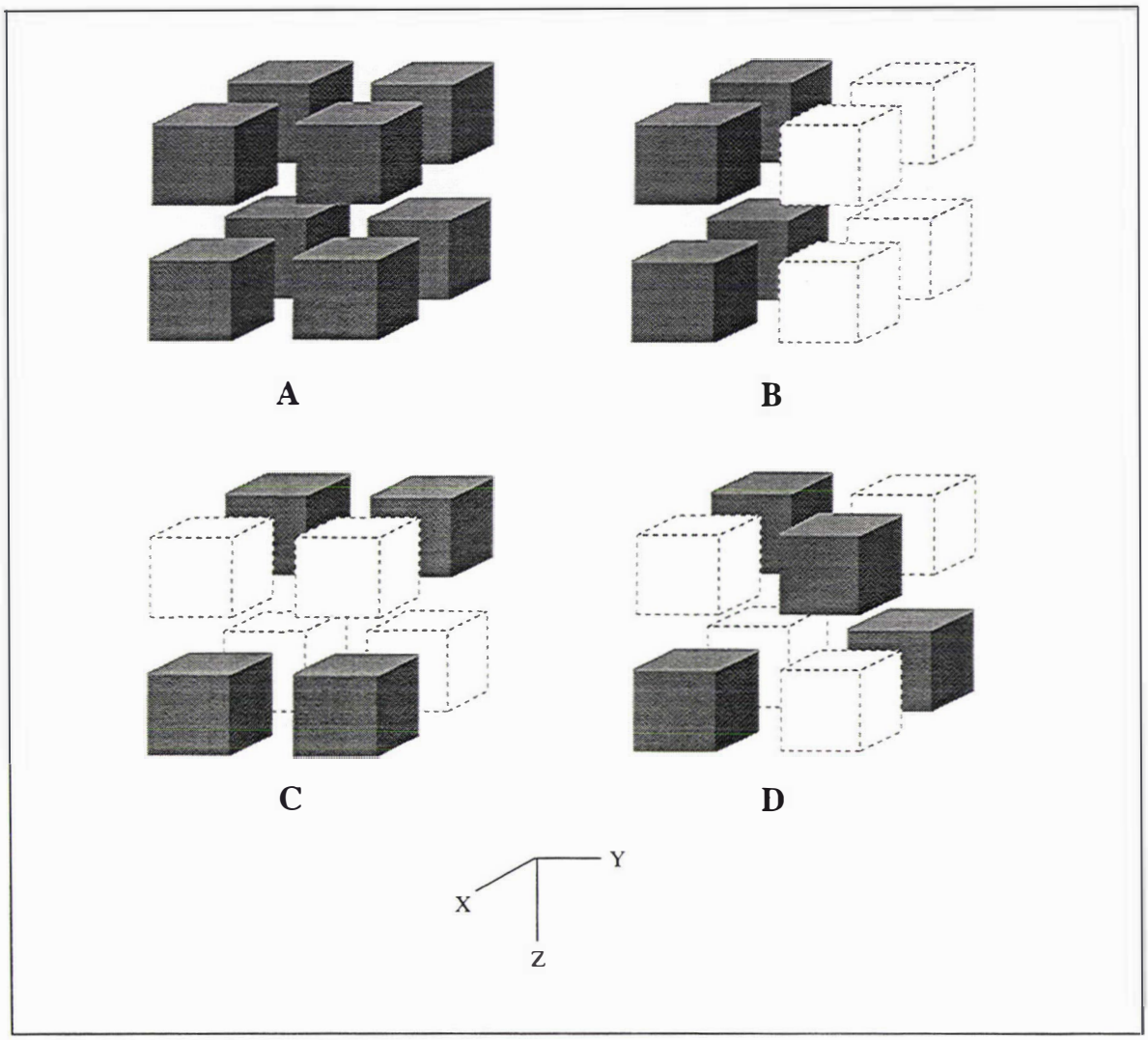


Figure 1.1 Geometric representation of the 2 x 2 x 2 factorial and some possible subsets.

Geometric representation of Equations 1.1 - 1.4. (A) represents the full 2 x 2 x 2 factorial (size = 8) (Equation 1.1). (B), (C) and (D) represent three possible subsets (size = 4) (Equation 1.2, Equation 1.3, and Equation 1.4 respectively). The matrices have been 'plotted' on standard orthogonal axes *X*, *Y* and *Z*. Each darkened box represents a matrix row, the lightened boxes represent the rows 'missing' from each subset relative to the full factorial arrangement. Subset (D) is an orthogonal array (Section 1.2.4.1). Note that when (D) is projected onto the axial planes *XY*, *YZ* and *XZ*, it forms a saturated projection (a complete copy of a 2 x 2 factorial). This is not true of subsets (B) (saturated projection on *XZ*, but not *XY* or *YZ*) and (C) (saturated projection on *XY* and *YZ*, but not *XZ*), which are not orthogonal arrays. The geometric projection properties of orthogonal arrays arise as a consequence of the formal definition given in the following section.

1.2.4.1 Definition of an orthogonal array

For a situation in which the experimental factors are all continuous, the problem of spacing the available experimental points 'uniformly' throughout the experimental region was considered by Kennard and Stone [39]. They implemented an algorithm to achieve this by sequentially choosing that point furthermost from the current design points. The problem of systematic search in high dimensional spaces was considered more formally by Aird [40] and Sobol [41], who present algorithms to generate designs in rectangular domains, based on some measure of dispersion in the set of selected points. However, for a situation in which some of the experimental factors are discrete (qualitative) a more general approach is required.

Rao [42] identified a special class of subsets with a great deal of symmetry. These subsets (or fractions) are known as orthogonal arrays. The formal definition of an orthogonal array follows:

An orthogonal array of strength d, with k_i columns with entries from a set of s_i symbols ($i = 1, ..., r$), is an N x m matrix ($m = k_1 + ... + k_r$) in which all possible combinations of the symbols in any d columns appear with equal frequency [42, 43]. Such an array can be denoted OA(N, m, $s_1^{k1} \times s_2^{k2} \times ... \times s_r^{kr}$, d).

N, m, s_i and d are said to be the parameters of the array. N refers to the size of the array, m to the number of constraints or factors, and d to the strength of the array.

Some authors prefer to use the term orthogonal array exclusively for the special (symmetric) case $s_1 = s_2 = ... = s_r = s$, and refer to the more general (asymmetric) case as an orthogonal array with variable symbols. No such distinction is made here. Note also that an orthogonal array is sometimes defined in a transpose fashion (i.e. with rows and columns interchanged).

With respect to the simple example given above, note that the 4 x 3 matrix specified by Equation 1.4 is an OA(4, 3, 2 x 2 x 2, 2). Consistent with the definition of an orthogonal array of strength 2, in each possible pair of columns the possible level combinations ((0,0), (0,1), (1,0), (1,1)) occur the same number of times (once). Orthogonal arrays are matrices whose columns possess certain balancing properties. In contrast the 4 x 3 matrices specified by Equations 1.2 and 1.3 do not satisfy the definition, and are not orthogonal arrays.

More complex examples, used later in the construction of search experiments for protein crystallization, are given in Table 1.1 and Table 1.2.

Table 1.1 OA(64, 6, 2^4x4x8, 3) [44]

	A	B	C	D	E	F
1	0	1	0	1	0	0
2	0	0	0	0	1	0
3	0	1	1	0	2	0
4	0	0	1	1	3	0
5	1	0	1	0	0	0
6	1	1	1	1	1	0
7	1	0	0	1	2	0
8	1	1	0	0	3	0
9	0	0	1	0	0	1
10	0	1	1	1	1	1
11	0	0	0	1	2	1
12	0	1	0	0	3	1
13	1	1	0	1	0	1
14	1	0	0	0	1	1
15	1	1	1	0	2	1
16	1	0	1	1	3	1
17	0	1	1	0	0	2
18	0	0	1	1	1	2
19	0	1	0	1	2	2
20	0	0	0	0	3	2
21	1	0	0	1	0	2
22	1	1	0	0	1	2
23	1	0	1	0	2	2
24	1	1	1	1	3	2
25	0	0	0	1	0	3
26	0	1	0	0	1	3
27	0	0	1	0	2	3
28	0	1	1	1	3	3
29	1	1	1	0	0	3
30	1	0	1	1	1	3
31	1	1	0	1	2	3
32	1	0	0	0	3	3
33	0	1	0	1	0	4
34	0	0	0	0	1	4
35	0	1	1	0	2	4
36	0	0	1	1	3	4
37	1	0	1	0	0	4
38	1	1	1	1	1	4
39	1	0	0	1	2	4
40	1	1	0	0	3	4
41	0	0	1	0	0	5
42	0	1	1	1	1	5
43	0	0	0	1	2	5
44	0	1	0	0	3	5
45	1	1	0	1	0	5
46	1	0	0	0	1	5
47	1	1	1	0	2	5
48	1	0	1	1	3	5
49	0	1	1	0	0	6
50	0	0	1	1	1	6
51	0	1	0	1	2	6
52	0	0	0	0	3	6
53	1	0	0	1	0	6
54	1	1	0	0	1	6
55	1	0	1	0	2	6
56	1	1	1	1	3	6
57	0	0	0	1	0	7
58	0	1	0	0	1	7
59	0	0	1	0	2	7
60	0	1	1	1	3	7
61	1	1	1	0	0	7
62	1	0	1	1	1	7
63	1	1	0	1	2	7
64	1	0	0	0	3	7

1.2.4.2 Construction and properties of orthogonal arrays

Known orthogonal arrays and their methods of construction were reviewed and catalogued by Dey [45], and research in this area continues (see e.g. [46, 47, 48]. Note that because of the strict conditions relating the parameters of an orthogonal array, such an array cannot exist for all possible values of these parameters.

The following properties of orthogonal arrays follow from the definition [49], and are useful in the manipulation of such arrays.

(I) Any array obtained from an orthogonal array by permuting columns, rows, or symbols in one or more columns will again be an orthogonal array with the same parameters.

(II) Consider an orthogonal array of size N, strength d, and having m factors. Any N x m' sub-matrix formed by deleting some columns of this array is also an orthogonal array with strength $d' = \min(m', d)$

(III) Any orthogonal array of strength d is an orthogonal array of strength d', with $d' < d$.

(IV) Combining rows of OA(N_i, m, s_1 x s_2 x ... x s_r, d), i=1,2 leads to an OA(N, m, s_1 x s_2 x ... x s_r, d) where $N = N_1 + N_2$

Table 1.2 OA(32, 8, 4^8x8, 2) [45]

	A	B	C	D	E	F	G	H	I
1	0	0	0	0	0	0	0	0	0
2	0	1	1	1	1	1	1	1	1
3	0	2	2	2	2	2	2	2	2
4	0	3	3	3	3	3	3	3	3
5	1	3	2	1	0	3	2	1	0
6	1	2	3	0	1	2	3	0	1
7	1	1	0	3	2	1	0	3	2
8	1	0	1	2	3	0	1	2	3
9	2	3	1	3	1	2	0	2	0
10	2	2	0	2	0	3	1	3	1
11	2	1	3	1	3	0	2	0	2
12	2	0	2	0	2	1	3	1	3
13	3	0	3	2	1	1	2	3	0
14	3	1	2	3	0	0	3	2	1
15	3	2	1	0	3	3	0	1	2
16	3	3	0	1	2	2	1	0	3
17	4	2	2	3	3	1	1	0	0
18	4	3	3	2	2	0	0	1	1
19	4	0	0	1	1	3	3	2	2
20	4	1	1	0	0	2	2	3	3
21	5	1	0	2	3	2	3	1	0
22	5	0	1	3	2	3	2	0	1
23	5	3	2	0	1	0	1	3	2
24	5	2	3	1	0	1	0	2	3
25	6	1	3	0	2	3	1	2	0
26	6	0	2	1	3	2	0	3	1
27	6	3	1	2	0	1	3	0	2
28	6	2	0	3	1	0	2	1	3
29	7	2	1	1	2	0	3	3	0
30	7	3	0	0	3	1	2	2	1
31	7	0	3	3	0	2	1	1	2
32	7	1	2	2	1	3	0	0	3

1.2.4.3 Uses of orthogonal arrays

The principal use of orthogonal arrays (and the original justification for their construction) has been in the planning of comparative experiments. Their properties of balance result in orthogonal (uncorrelated) estimates of effects for the linear models commonly used in the interpretation of such experiments. In this context they are often referred to as orthogonal

fractional factorial designs. They have also found application in computer experiments and numerical integration procedures [50, 51, 52].

A property of orthogonal arrays is that the points in an orthogonal array are usually spread regularly throughout the factor space. This suggests that appropriate orthogonal arrays might be used as the basis of initial search experiments in protein crystallization. The idea of using orthogonal arrays as the basis of a search procedure has been suggested in a more general context by Wu and coworkers [34].

1.2.5 Underlying factorial structure for search experiments

In the construction of a protein crystallization search experiment based on a suitable orthogonal array, the preliminary step is the specification of the underlying factorial structure describing the experimental region. This involves consideration of both the factors to be included in the initial study, and the specification of appropriate levels for these factors.

The selection of a suitable physical method to bring the protein solution into the supersaturated state is assumed (see [29]). The choice of physical procedure together with the volume and geometry of the experimental arrangement will influence the crystallization process, principally through the kinetics of equilibration processes. However such considerations seem likely to be of secondary importance to the definition of the temperature and composition of the system.

Even in the simplest system there will typically be three components, the protein, a hydrogen ion buffer to maintain the solution pH, and a protein precipitant. For membrane proteins a detergent to solubilize the protein is also required because of their characteristic hydrophobic surface features [53, 54]. Detergents may also be useful or necessary for the crystallization of water soluble proteins [55]. In this chapter the simpler three component systems are considered. The experimental problem is then to determine an appropriate precipitant and buffer, together with the ranges of temperature, pH and concentration of the solution components that will support nucleation and crystal growth. Experimentally it is usually most convenient to work at a fixed protein concentration and govern the degree of supersaturation in the system by manipulating the concentration of the other solution components. Considering here the situation in which the buffer concentration is also fixed, the experimental factors to be considered are Precipitant type, Buffer type, Precipitant concentration, pH and Temperature. Experimentally, because it does not involve variation in the solution composition, temperature can be treated differently from other experimental factors.

With regard to the protein precipitant, three classes of compounds (inorganic salts, synthetic polymers and alcohols) have found widespread use in the growth of protein crystals. While in all cases, the reduction in protein solubility is associated with the exclusion of the precipitant from the immediate domain of the protein, the principal sources of this effect differ for the differing classes of precipitant [56]. Consequently, it is physically sensible to consider each class of precipitant within the framework of a separate search experiment. Because the number of potentially useful compounds within each class is very large, a representative selection will have to be made.

The control of pH during crystallization requires the presence of a suitable buffering system. However, for a given buffer, effective buffering capacity is limited to pH values close to its pK_a, so it is unlikely to be useful over the entire pH range of interest in an initial search experiment. The consequence of this for an experiment having a factorial structure is that the factor 'buffer type' will need to be nested within the factor 'pH' (i.e.the levels of the factor 'buffer type' will be dependent upon the level of the factor 'pH').

For the continuous variables (precipitant concentration and pH), the spacing of the levels is clearly very important. Recent work suggests that in some cases, protein crystallization may occur over reasonably large bounded ranges of these variables [27]. Once the number of levels has been decided upon it would seem reasonable to space the levels evenly with neither the highest nor lowest level being set at the extreme of the feasible range.

The concentration of the precipitant will govern the degree of protein supersaturation in the system. A principal difficulty arises in relating the commonly used expressions of concentration (molarity, molality and volume percentage) to the relative effectiveness of compounds in reducing protein solubility, a problem related to the lack of adequate physical models of this phenomenon.

1.2.6 Practical implementation of orthogonal array-based search designs

1.2.6.1 A Search design based on the use of polymeric protein precipitants.

A description is now given of an initial search design for protein crystallization, based on an orthogonal array, which illustrates the general approach which has been adopted. Polyethylene glycols are used as protein precipitants. The experiment is based on a factorial structure, with variation in polymer type, polymer concentration, pH, and buffer type. By necessity the buffer type is nested within the pH. The factorial structure is given in Table 1.3.

Table 1.3 Factorial structure of a search experiment for protein crystallization employing polyethylene glycols as protein precipitants

Abbreviations employed: MES, 2-[N-Morpholino]ethanesulfonic acid; PIPES, 1,4-Piperazinediethanesulfonic acid; Bis-tris propane, 1,3-bis[tris(Hydroxymethyl)-methylamino]propane; HEPES, N-[Hydroxyethyl]piperazine-N'-[2-ethanesulfonic acid]; MOPS, 3-[N-Morpholino]propanesulfonic acid; EPPS, N-[2-Hydroxyethyl]piperazine-N'[3-propanesulfonic acid]; Tris, Tris(hydroxymethyl)aminomethane; TAPS, N-tris[Hydroxymethyl]methyl-3-aminopropanesulfonic acid; AMPSO, (3-[(1,1-Dimethyl-2- hydroxyethyl) amino]2-hydroxypropanesulfonic acid.

Factor	Levels		
Polymer type	Polyethylene glycol MW 6000 Polyethylene glycol monomethyl ether MW 5000		
Polymer Concentration %(w/v)	7 14 21 28		
pH (Buffer type)	4.9 5.5 6.1 6.7 7.3 7.9 8.5 9.1	[Acetate [Succinate [MES [PIPES [HEPES [EPPS [TAPS [AMPSO	Citrate] Malate] Cacodylate] Bis-tris propane] MOPS] Tris Bis-tris propane] Borate]

The two polymers employed are polyethylene glycol (PEG) MW 6000 and polyethylene glycol monomethyl ether (PEG-mme) MW 5000. The use of PEG-mme was suggested by recent work at the University of York. [57]. Because these polymers have a similar average molecular weight, the volume exclusion effect should be similar for each [58], so neglecting differences in polymer binding to the protein [59], similar concentrations should be required to effect the same decrease in protein solubility. Polymer concentration is varied over the range 7 - 28%(w/v), employing 4 equally spaced levels.

The pH range of the experiment is 4.9 - 9.1, employing 8 equally spaced levels. At each pH two buffers are employed. The buffers, being obtained as free acids or bases, were titrated to the appropriate pH with KOH or HCl respectively (KOH is used in preference to NaOH because of solubility considerations). While many of the buffers used in the experiment are chemically quite dissimilar, most of those which share the sulfonic acid moiety (MES, PIPES, HEPES, MOPS, EPPS, TAPS, AMPSO) have been assigned to the same class of the nested factor 'buffer type'. The buffer concentration is fixed at 0.2 M (Note that the addition

of aqueous PEG solutions at high concentration to buffered solutions can cause a change in the measured pH [60]).

The experiment as it stands would involve 2x2x4x8 = 128 runs. Using Columns A, B, E and F of the OA(64, 6, 2^4x4x8, 3) (Table 1.1), a subset of 64 runs is selected, forming the basis of an initial protein crystallization search experiment. (Column A has been assigned to polymer type, Column B to buffer type, Column E to Polymer concentration, and Column F to pH). The resulting experiment is given in Table 1.4.

Table 1.4 Search experiment for protein crystallization employing polyethylene glycols as protein precipitants

The experiment is based on the OA(64, 6, 2^4x4x8, 3) (see Table 1.1). The unallocated two-level columns of this array can be assigned to temperature, or to other experimental factors.

Run	Precipitant	Buffer	pH	Unassigned factor	Unassigned factor
1	7.00% PEG 6000	0.20 M Citric acid/KOH	pH 4.90	0	1
2	14.00% PEG 6000	0.20 M Acetic acid/KOH	pH 4.90	0	0
3	21.00% PEG 6000	0.20 M Citric acid/KOH	pH 4.90	1	0
4	28.00% PEG 6000	0.20 M Acetic acid/KOH	pH 4.90	1	1
5	7.00% PEG-mme 5000	0.20 M Acetic acid/KOH	pH 4.90	1	0
6	14.00% PEG-mme 5000	0.20 M Citric acid/KOH	pH 4.90	1	1
7	21.00% PEG-mme 5000	0.20 M Acetic acid/KOH	pH 4.90	0	1
8	28.00% PEG-mme 5000	0.20 M Citric acid/KOH	pH 4.90	0	0
9	7.00% PEG 6000	0.20 M Succinic acid/KOH	pH 5.50	1	0
10	14.00% PEG 6000	0.20 M Malic acid/KOH	pH 5.50	1	1
11	21.00% PEG 6000	0.20 M Succinic acid/KOH	pH 5.50	0	1
12	28.00% PEG 6000	0.20 M Malic acid/KOH	pH 5.50	0	0
13	7.00% PEG-mme 5000	0.20 M Malic acid/KOH	pH 5.50	0	1
14	14.00% PEG-mme 5000	0.20 M Succinic acid/KOH	pH 5.50	0	0
15	21.00% PEG-mme 5000	0.20 M Malic acid/KOH	pH 5.50	1	0
16	28.00% PEG-mme 5000	0.20 M Succinic acid/KOH	pH 5.50	1	1
17	7.00% PEG 6000	0.20 M Cacodylic acid/KOH	pH 6.10	1	0
18	14.00% PEG 6000	0.20 M MES /KOH	pH 6.10	1	1
19	21.00% PEG 6000	0.20 M Cacodylic acid/KOH	pH 6.10	0	1
20	28.00% PEG 6000	0.20 M MES /KOH	pH 6.10	0	0

Run	Precipitant	Buffer	pH	Unassigned factor	Unassigned factor
21	7.00% PEG-mme 5000	0.20 M MES /KOH	pH 6.10	0	1
22	14.00% PEG-mme 5000	0.20 M Cacodylic acid/KOH	pH 6.10	0	0
23	21.00% PEG-mme 5000	0.20 M MES /KOH	pH 6.10	1	0
24	28.00% PEG-mme 5000	0.20 M Cacodylic acid/KOH	pH 6.10	1	1
25	7.00% PEG 6000	0.20 M PIPES /KOH	pH 6.70	0	1
26	14.00% PEG 6000	0.20 M Bis-tris propane/HCl	pH 6.70	0	0
27	21.00% PEG 6000	0.20 M PIPES /KOH	pH 6.70	1	0
28	28.00% PEG 6000	0.20 M Bis-tris propane/HCl	pH 6.70	1	1
29	7.00% PEG-mme 5000	0.20 M Bis-tris propane/HCl	pH 6.70	1	0
30	14.00% PEG-mme 5000	0.20 M PIPES /KOH	pH 6.70	1	1
31	21.00% PEG-mme 5000	0.20 M Bis-tris propane/HCl	pH 6.70	0	1
32	28.00% PEG-mme 5000	0.20 M PIPES /KOH	pH 6.70	0	0
33	7.00% PEG 6000	0.20 M MOPS /KOH	pH 7.30	0	1
34	14.00% PEG 6000	0.20 M HEPES /KOH	pH 7.30	0	0
35	21.00% PEG 6000	0.20 M MOPS /KOH	pH 7.30	1	0
36	28.00% PEG 6000	0.20 M HEPES /KOH	pH 7.30	1	1
37	7.00% PEG-mme 5000	0.20 M HEPES /KOH	pH 7.30	1	0
38	14.00% PEG-mme 5000	0.20 M MOPS /KOH	pH 7.30	1	1
39	21.00% PEG-mme 5000	0.20 M HEPES /KOH	pH 7.30	0	1
40	28.00% PEG-mme 5000	0.20 M MOPS /KOH	pH 7.30	0	0
41	7.00% PEG 6000	0.20 M EPPS /KOH	pH 7.90	1	0
42	14.00% PEG 6000	0.20 M Tris /HCl	pH 7.90	1	1
43	21.00% PEG 6000	0.20 M EPPS /KOH	pH 7.90	0	1
44	28.00% PEG 6000	0.20 M Tris /HCl	pH 7.90	0	0
45	7.00% PEG-mme 5000	0.20 M Tris /HCl	pH 7.90	0	1
46	14.00% PEG-mme 5000	0.20 M EPPS /KOH	pH 7.90	0	0
47	21.00% PEG-mme 5000	0.20 M Tris /HCl	pH 7.90	1	0
48	28.00% PEG-mme 5000	0.20 M EPPS /KOH	pH 7.90	1	1
49	7.00% PEG 6000	0.20 M Bis-tris propane/HCl	pH 8.50	1	0
50	14.00% PEG 6000	0.20 M TAPS /KOH	pH 8.50	1	1
51	21.00% PEG 6000	0.20 M Bis-tris propane/HCl	pH 8.50	0	1
52	28.00% PEG 6000	0.20 M TAPS /KOH	pH 8.50	0	0
53	7.00% PEG-mme 5000	0.20 M TAPS /KOH	pH 8.50	0	1

Run	Precipitant	Buffer	pH	Unassigned factor	Unassigned factor
54	14.00% PEG-mme 5000	0.20 M Bis-tris propane/HCl	pH 8.50	0	0
55	21.00% PEG-mme 5000	0.20 M TAPS /KOH	pH 8.50	1	0
56	28.00% PEG-mme 5000	0.20 M Bis-tris propane/HCl	pH 8.50	1	1
57	7.00% PEG 6000	0.20 M AMPSO /KOH	pH 9.10	0	1
58	14.00% PEG 6000	0.20 M Boric acid/KOH	pH 9.10	0	0
59	21.00% PEG 6000	0.20 M AMPSO /KOH	pH 9.10	1	0
60	28.00% PEG 6000	0.20 M Boric acid/KOH	pH 9.10	1	1
61	7.00% PEG-mme 5000	0.20 M Boric acid/KOH	pH 9.10	1	0
62	14.00% PEG-mme 5000	0.20 M AMPSO /KOH	pH 9.10	1	1
63	21.00% PEG-mme 5000	0.20 M Boric acid/KOH	pH 9.10	0	1
64	28.00% PEG-mme 5000	0.20 M AMPSO /KOH	pH 9.10	0	0

1.2.6.2 Extension to other classes of protein precipitants

A similar factorial structure can also be used for search designs based on other classes of protein precipitants, such as normal salts that do not possess an appreciable buffering capacity (e.g. $(NH_4)_2SO_4$), or alcohols and poly-hydroxy compounds (e.g. ethanol, 2-methyl-2,4-pentanediol). The difficulty in assessing the relative effectiveness of such compounds in reducing protein solubility makes the setting of appropriate levels for the precipitant concentration difficult.

In the case of salts, the relative effectiveness in altering protein solubility is reflected in the lyotropic or Hofmeister series [61]. When ion binding to the protein is negligible, the predominant protein-salt interaction is exclusion of the salt from the immediate domain of the protein, leading to greatly reduced protein solubility at high salt concentrations (salting-out behaviour) [62]. The likely physical basis of this exclusion is the increase in surface tension of water caused by the addition of salts. Accordingly Melander and Horváth [63] proposed that the differing relative effects of salts on protein solubility could be directly related to their effect on the surface tension of water (i.e. that surface tension might be used as the physical basis for a quantitative lyotropic scale). This suggestion has been adopted, and surface tension is used as the basis for assigning relative salt concentrations in our experiments. The change in surface tension can be calculated from the surface tension/molality and density/molality data available in the literature [64, 65]. However, complication of this simple phys-

ical picture will arise when there is significant salt binding to the protein as for example with the divalent cation salts such as $MgCl_2$, $BaCl_2$ and $CaCl_2$ which are known to be largely ineffective at salting out proteins despite having large molal surface tension increments [62].

We experimented with a number of neutral salts as protein precipitants (e.g. LiCl, NH_4NO_3, $(NH_4)_2SO_4$, LiBr). Considering results from a number of proteins, it was found that while all the salts were effective at precipitating proteins, crystallization occurred far more frequently from ammonium sulfate (results not shown). Why this should be so is not clear. Accordingly, an experiment was designed based on the use of ammonium sulfate alone.

The factors varied are the ammonium sulfate concentration, the pH, and the buffer type (nested in the pH as before). The four levels of Ammonium sulfate concentration were set at 0.87, 1.65, 2.33 and 2.94 mol/l which should effect a change in the surface tension of water of 2, 4, 6 and 8 x 10^{-3} Nm^{-1} respectively (based on available literature data). The factorial structure is given in Table 1.5. This experiment would involve 8 x 4 x 2 = 64 runs in total. Using the orthogonal array OA(32, 9, 4^8 x 8, 2), shown in Table 1.2 (for the construction method, see [45], section 3.4.2), a subset of 32 is selected. Note that this is an orthogonal array of strength 2

Table 1.5 Factorial structure of a search experiment for protein crystallization employing ammonium sulfate as a protein precipitant

Factor	Levels		
Salt Concentration (mol/l)	0.87		
	1.65		
	2.33		
	2.94		
pH (Buffer type)	4.9	[Acetate	Citrate]
	5.5	[Succinate	Malate]
	6.1	[MES	Cacodylate]
	6.7	[PIPES	Bis-tris propane]
	7.3	[HEPES	MOPS]
	7.9	[EPPS	Tris
	8.5	[TAPS	Bis-tris propane]
	9.1	[AMPSO	Borate]

With reference to the orthogonal array given in Table 1.2, Column A has been assigned to pH, column B to Ammonium sulfate concentration, and Column I to Buffer type (after being collapsed to three two-level factors by the following correspondence scheme [66]).

Levels of 4-level factor		Levels of two-level factors
0	→	0 0 0
1	→	0 1 1
2	→	1 0 1
3	→	1 1 0

This results in the experiment presented in Table 1.6. For convenience, several of the unassigned 4-level columns are also presented.

Table 1.6 Search experiment for protein crystallization employing ammonium sulfate as a protein precipitant

The experiment is based on the OA(32, 9, 4^8x8, 2) (see Table 1.2). There are a number of unallocated columns of this array (see text), for convenience, two (C and D) are presented in this table.

Run	Precipitant	Buffer	pH	Unassigned factor	Unassigned factor
1	0.87 M Ammonium sulfate	0.20 M Acetic acid/KOH	pH 4.90	0	0
2	1.65 M Ammonium sulfate	0.20 M Citric acid/KOH	pH 4.90	1	1
3	2.33 M Ammonium sulfate	0.20 M Acetic acid/KOH	pH 4.90	2	2
4	2.94 M Ammonium sulfate	0.20 M Citric acid/KOH	pH 4.90	3	3
5	2.94 M Ammonium sulfate	0.20 M Succinic acid/KOH	pH 5.50	2	1
6	2.33 M Ammonium sulfate	0.20 M Malic acid/KOH	pH 5.50	3	0
7	1.65 M Ammonium sulfate	0.20 M Succinic acid/KOH	pH 5.50	0	3
8	0.87 M Ammonium sulfate	0.20 M Malic acid/KOH	pH 5.50	1	2
9	2.94 M Ammonium sulfate	0.20 M MES/KOH	pH 6.10	1	3
10	2.33 M Ammonium sulfate	0.20 M Cacodylic acid/KOH	pH 6.10	0	2
11	1.65 M Ammonium sulfate	0.20 M MES/KOH	pH 6.10	3	1
12	0.87 M Ammonium sulfate	0.20 M Cacodylic acid/KOH	pH 6.10	2	0
13	0.87 M Ammonium sulfate	0.20 M PIPES/KOH	pH 6.60	3	2
14	1.65 M Ammonium sulfate	0.20 M Bis-tris propane/HCl	pH 6.60	2	3
15	2.33 M Ammonium sulfate	0.20 M PIPES/KOH	pH 6.60	1	0

Run	Precipitant	Buffer	pH	Unassigned factor	Unassigned factor
16	2.94 M Ammonium sulfate	0.20 M Bis-tris propane/HCl	pH 6.60	0	1
17	2.33 M Ammonium sulfate	0.20 M HEPES/KOH	pH 7.30	2	3
18	2.94 M Ammonium sulfate	0.20 M MOPS/KOH	pH 7.30	3	2
19	0.87 M Ammonium sulfate	0.20 M HEPES/KOH	pH 7.30	0	1
20	1.65 M Ammonium sulfate	0.20 M MOPS/KOH	pH 7.30	1	0
21	1.65 M Ammonium sulfate	0.20 M EPPS/KOH	pH 7.90	0	2
22	0.87 M Ammonium sulfate	0.20 M Tris/HCl	pH 7.90	1	3
23	2.94 M Ammonium sulfate	0.20 M EPPS/KOH	pH 7.90	2	0
24	2.33 M Ammonium sulfate	0.20 M Tris/HCl	pH 7.90	3	1
25	1.65 M Ammonium sulfate	0.20 M TAPS/KOH	pH 8.50	3	0
26	0.87 M Ammonium sulfate	0.20 M Bis-tris propane/HCl	pH 8.50	2	1
27	2.94 M Ammonium sulfate	0.20 M TAPS/KOH	pH 8.50	1	2
28	2.33 M Ammonium sulfate	0.20 M Bis-tris propane/HCl	pH 8.50	0	3
29	2.33 M Ammonium sulfate	0.20 M AMPSO/KOH	pH 9.10	1	1
30	2.94 M Ammonium sulfate	0.20 M Boric acid/KOH	pH 9.10	0	0
31	0.87 M Ammonium sulfate	0.20 M AMPSO/KOH	pH 9.10	3	3
32	1.65 M Ammonium sulfate	0.20 M Boric acid/KOH	pH 9.10	2	2

For alcohols and polyhydroxy compounds, the lack of adequate physical theory means that we have no choice at present but to set levels completely empirically. For example, a useful experiment has been constructed using ethanol and 2-methyl-2,4-pentanediol as protein precipitants. For this experiment, the factorial structure is the same as for the experiment based on polyethylene glycols (Table 1.3) with the two alcohols replacing the two polymers. Concentrations are set at 10, 20, 40 and 60%(v/v) for both compounds.

Finally, salts of a number of acids (e.g. Citric, Tartaric, Phosphoric and Acetic acids) have been used with considerable success in protein crystallization. Since all of these salts have an appreciable buffering capacity, especially at the high concentrations used to crystallize proteins, trying to control the pH of such solutions using another buffer at much lower concentration is futile. If a genuine variation in pH is desired, the best strategy would seem to be to titrate each acid to the required pH using a strong base. An experiment based on the use of these acid salts remains to be implemented.

1.2.7 Experimental considerations

We have normally executed such experiments using vapour diffusion techniques, in which a small volume of the buffer/precipitant solution (typically 1 - 10μL) is mixed with an equal volume of protein solution (having a typical concentration of 5 - 50 mg/ml). This drop is then equilibrated in a sealed system against a much larger volume of the buffer/precipitant solution. The experiments are not usually replicated because of constraints on the amounts of purified protein available. The protein itself is suspended in water, unless this is not possible because of constraints due to stability, solubility or activity. Where control of the pH is required, low concentrations (10 - 20 mM) of an appropriate buffer are employed. In specific cases, the potential of many buffers as ligands for free metal ions in solution should not be overlooked, nor their ability to serve as enzymatic substrates.

No mention has been made of the temperature at which the experiments are conducted. The orthogonal arrays on which the experiments described in Tables 1.4 and 1.6 are based, have two or more unassigned columns. One of the remaining columns could be assigned to temperature if this was desired. It should be recognized that for some of the buffers included in these experiments, the pK_a values are temperature dependent, which may result in quite large pH changes with temperature.

An important consideration is the direction of further search should the initial experiment(s) fail to identify suitable solution conditions. This is a difficult question to address, principally because negative results give so little information on the most promising neighbourhood for further search. A well planned initial search experiment will at least serve to clearly define the protein solubility (i.e. to restrict the likely range of precipitant concentration to be investigated in further experiments).

If we consider, specifically, the search experiment based on polyethylene glycols (Table 1.4), then extending the search might involve (i) executing the remaining treatment combinations required to complete the factorial arrangement (or at least those that are appropriate based on the knowledge of protein solubility), (ii) extending the pH range of the search, (iii) conducting similar experiments using related compounds (other non-ionic polymers, see e.g. [67]). What strategy is generally appropriate depends on the strength of the orthogonal array used in the search experiment (and the spacing and number of the levels of the continuous factors in the underlying factorial structure). This particular experiment employs an orthogonal array of strength 3, with fairly close spacing of the levels of the continuous factors, and consequently covers the experimental region in a reasonably compre-

hensive fashion. If the results from this experiment are all negative, then it might be more worthwhile considering experiments based on a different class of precipitant (e.g. salts) rather than pursuing the search further with polymeric precipitants. In other cases, for example the search experiment based on ammonium sulfate (Table 1.6), a orthogonal array of strength two is employed, and it might be worthwhile executing the remaining treatment combinations required to complete the factorial arrangement. Reduction of the strength of the orthogonal array used in the initial search procedure substantially reduces the initial experimental size, but also reduces the completeness of the coverage of the experimental region, and increases the probability that solution conditions supporting crystal growth will not be located in the initial search.

The use of orthogonal arrays in the fashion described in this chapter can be viewed as a way to sequentially execute factorial experiments. Where the quantity of protein available is very limited, it may be possible to reduce the experimental size further by adopting an even more sequential approach and executing only part of an orthogonal array. This might also be necessary if it is desired to avoid some part of the experimental region for physical reasons. For example, the pH range explored in the initial search experiment might be reduced. Subsequent to this initial experiment, the search could be extended over a wider pH range if necessary, with some levels of precipitant concentration having been eliminated from further consideration. It should be noted that it is often possible to decompose orthogonal arrays into sub-arrays which are also orthogonal arrays (in fact, working in the reverse fashion, this is a common construction method for orthogonal arrays). For example, consider the experiment based on polyethylene glycols (Table 1.4). Ignoring the unassigned columns of the array, at each level of pH, the 8 experimental points constitute an OA(8, 3, 2x2x4, 2).

It should be apparent from this discussion that searching for crystallization conditions remains an essentially empirical process. Even with an organized experimental approach to the problem such as the one presented here, many difficult questions remain relating to the coverage of the experimental region (i.e. the number and spacing of levels in the underlying factorial experiment, and the strength of the orthogonal array used in the search procedure), and the direction of further search when the initial experiments fail. However, a principal advantage of the search experiments based on orthogonal arrays described in this chapter over less formal approaches, is that they provide (at least in a qualitative sense) a way of assessing the degree of coverage of the experimental region (through the concept of the strength of the array). They also provide a way of organizing a sequential experimental program for proteins which prove difficult to crystallize.

1.3 PRACTICAL APPLICATION TO SEVERAL PROBLEMS

Critical evaluation of experiments to search for crystallization conditions is very difficult. In the preceding section, criteria that a search design should attempt to satisfy were given, and some orthogonal array-based search designs were described which attempt to meet these criteria. Since these ideas were developed they have been used in attempts to crystallize upwards of twenty proteins. This work has been carried out by Heather Baker, and other members of the protein crystallography laboratory at Massey university

We consider in detail only the search design based on the use of polyethylene glycols described in Table 1.4. It is evident that this systematic search method works better than more ad hoc methods (for example the experiment described by Jancarik and Kim [23]), in identifying crystallization conditions using this class of precipitant. In a number of cases it has identified crystallization conditions where the latter experiment has failed. This is not a surprising result. Orthogonal array-based designs also provide a much better basis for comparative analysis of the effects of the experimental factors on protein solubility and protein crystallization. This follows directly from the well established properties of orthogonal arrays when associated with the analysis of linear models (see Section 1.2.4.3). Upwards of 70% of the proteins studied have been crystallized using the experiment described in Table 1.4, which is based on a single class of precipitant (although not all of these crystals have been suitable for X-ray diffraction studies). This goes some way to justifying the premise that search designs should employ simple physical systems. It is apparent that there is little difference between the properties of polyethylene glycol and polyethylene glycol monomethyl ether with respect to the crystallization of proteins. This is perhaps not suprising given that non-ionic polymers are believed to reduce protein solubility by simple steric exclusion effects (see [58, 68]). At some stage in the future the experiment should perhaps be redesigned to reflect this finding. With a little more effort it should be possible to effectively extend these ideas to all classes of protein precipitant, giving rise to a complete and relatively systematic search procedure.

The application of the ideas presented in this chapter is now briefly discussed with reference to three specific crystallization problems, the human bile salt dependent lipase, the glucose-fructose oxidoreductase from *Zymomonas mobilis*, and the human $\alpha_2\varepsilon_2$ embryonic hemoglobin.

1.3.1 Bile salt dependent lipase

Human bile salt dependent lipase (BSDL) (discussed in Chapter 4) is a protein of 722 amino acids. The N-terminal catalytic domain of the protein comprising the first 530 amino acids, is followed by a glycosylated C-terminal proline-rich repeat region of unknown function. This tandem repetitive region cannot have a compact globular structure, due to the presence of a high number of proline residues and the bulky glycan chains. It is known that the glycosylation is heterogeneous. Consequently, this presents a difficult crystallization problem, and it was during work on this protein that most of the work described in this chapter was developed.

The protein was purified from human milk using standard procedures [69]. This work was carried out at the University of Umeå by a collaborating group (Professor Olle Hernell and Dr. Lars Bläckberg), and the protein sent as a freeze-dried powder. In all of the experiments described below, the lyophilized protein (which had been extensively dialyzed against water prior to freeze drying) was redissolved in water, and concentrated to 10 - 30 mg/ml (based on absorption at 280 nm).

Some preliminary attempts to crystallize this protein had not been successful (results not shown). In a first attempt at a systematic crystallization experiment, polyethylene glycols were used to precipitate the protein, the experiment having the factorial structure shown in Table 1.7. This factorial experiment involves 4^4 x 8 = 2048 runs. An orthogonal array of strength 2 was used to select 128 of these (the orthogonal array used was based on the designs of Box and Hunter for two level fractional factorial experiments [70, 71], but was sub-optimal and will not be described further).

Table 1.7 Factorial structure of the search experiment used in initial crystallization attempts for BSDL

Factor	**Levels**							
pH (Buffer)	5 (Citric acid)		6 (Citric acid)		7 (MOPS)		8 (HEPES)	
Buffer counter ion	K^+	Na^+	NH_4^+	$N(CH_3)_4^+$				
Buffer concentration (mol/l)	0.2	0.3	0.4	0.5				
PEG MW	400	1500	4000	8000				
PEG Concentration %(w/v)	9.0	12.0	15.0	18.0	21.0	24.0	27.0	30.0

The results of this experiment were extremely encouraging. For the first time large single crystals of native BSDL were obtained. A number of clear features of the crystallization were apparent. The crystals grew most readily at pH 7 in the presence of MOPS buffer. Small crystals also grew at pH 8 in HEPES buffer, but not nearly as readily. No crystals grew at the other two levels of the pH/buffer. The buffer counter-ion and the buffer concentration were not critical to crystallization. The molecular weight of the PEG employed was also not critical, except that the concentrations employed in the experiment were generally not high enough to precipitate the protein for the two PEG's of lower molecular weight.

In a subsequent factorial experiment to establish if the buffer type was a critical determinant in the growth of these crystals, eight different buffers (HEPES/KOH, PIPES/KOH, MOPSO/KOH, Phosphate/KOH, Ethylenediamine/HCl, Imidazole/HCl, Bis-tris propane/HCL and 2,4,6-Collidone/HCl) all at pH 7 and a concentration of 0.3M, were used together with PEG 8000 as the protein precipitant. Surprisingly crystals grew only from solutions buffered with MOPSO (chemically very similar to MOPS). The presence of the other buffers resulted in precipitation of the protein.

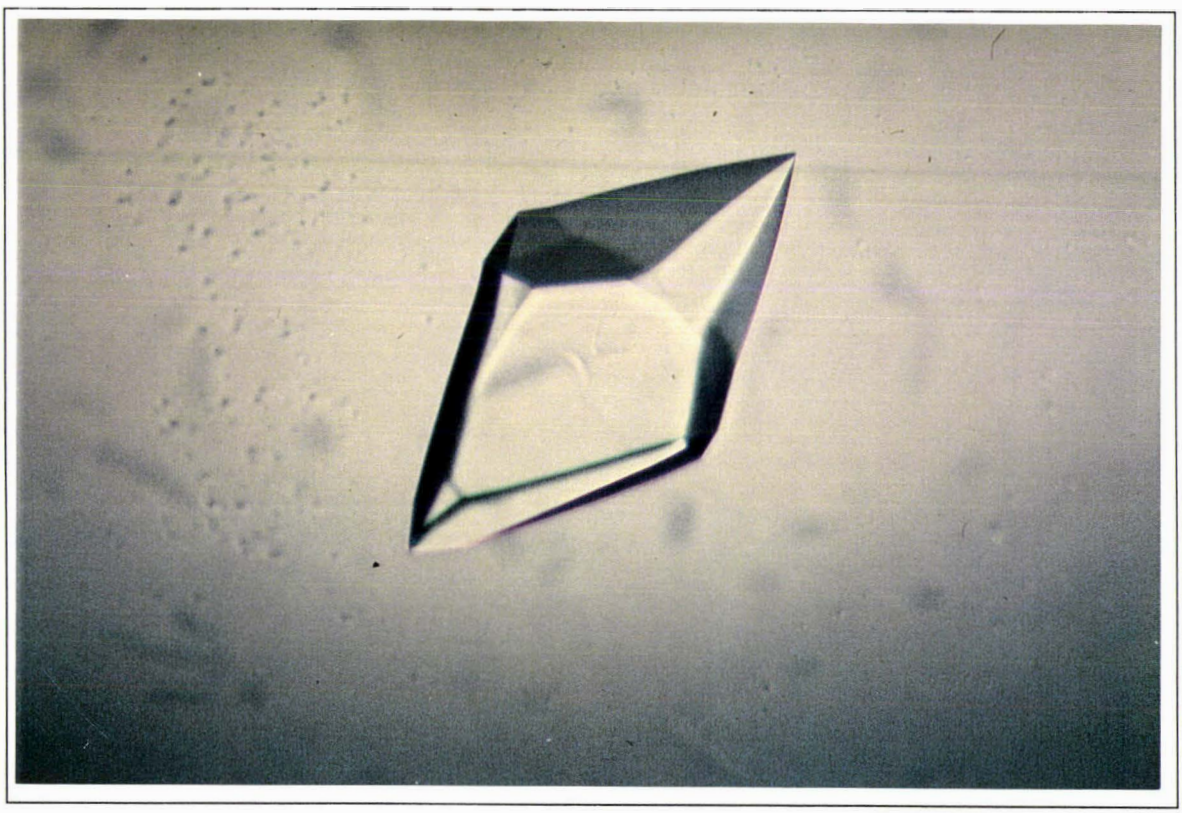


Figure 1.2 Crystal of Native BSDL

Crystals can be grown up to 0.8 mm in length in their longest dimension

Subsequent to the identification of the crystallization conditions, careful optimization of the solution composition led to the growth of large single crystals (see Figure 1.2). The best crystals were grown by sitting-drop vapour diffusion from solutions containing 0.40 M MOPS/ NH_4OH buffer at pH 6.8, and 15 - 20% (w/v) PEG 8000. Unfortunately these crystals diffract only poorly (see Chapter 4). Subsequent work has involved modification of the protein by partial enzymatic deglycosylation, and the crystallization of a recombinant truncated variant, which lacks the C-terminal repeat region altogether.

1.3.2 Glucose-fructose oxidoreductase

Glucose-fructose oxidoreductase from *Zymomonas mobilis* (discussed in Chapters 2 and 3), is a protein of 381 amino acids. It is not glycosylated, and exists as a stable tetramer at low pH. In contrast to native BSSL, it can be crystallized extremely readily. The protein was concentrated to 20 mg/ml in water (by absorbance at 280 nm), and search experiments were executed using the hanging drop vapour diffusion technique. A search experiment was conducted using polyethylene glycols as protein precipitants, which was very similar to that described in Table 1.4. It covered the pH range 5 - 8, with four equally spaced levels, and four buffers at each pH. Again, an orthogonal array of strength 3 was used as the basis of the search experiment. It immediately became apparent that the solubility of GFOR in PEG solutions was not high, consequently only the runs at the two lowest polymer concentrations were executed. Crystals grew across the entire pH range virtually independent of the buffer type, but were best defined at pH 5 - 6. The crystals appeared as masses of extremely thin stacked plates, growing in some cases within minutes at the higher (14% (w/v)) polymer concentration. Subsequent factorial experiments centred on the best conditions identified in this search experiment, and involved variation in the solution pH, and the concentration of the buffer and polymer. By careful control of the degree of protein supersaturation it was possible to grow single crystals (see Figure 1.3), although these were usually still relatively thin and fragile. These crystals were used to successfully determine the structure (Chapter 2). Crystals used in the structure determination by multiple isomorphous replacement were grown at ambient temperature from 5 -15% (w/v) PEG 6000, in the presence of 0.2M Succinic acid/ KOH or 0.2M Citric acid/KOH buffer at pH 5.5. It is also possible to grow very small crystals of GFOR using salts as protein precipitants, but this has not been pursued further.

In the case of GFOR, since it crystallized so readily, the nature of the search procedure was not critical to the successful resolution of the problem. However the use of a balanced experiment allowed us to draw useful informal inferences about the crystallization process, for

example that its behaviour appears essentially independent of the buffer type or the pH over the range 5 - 8.

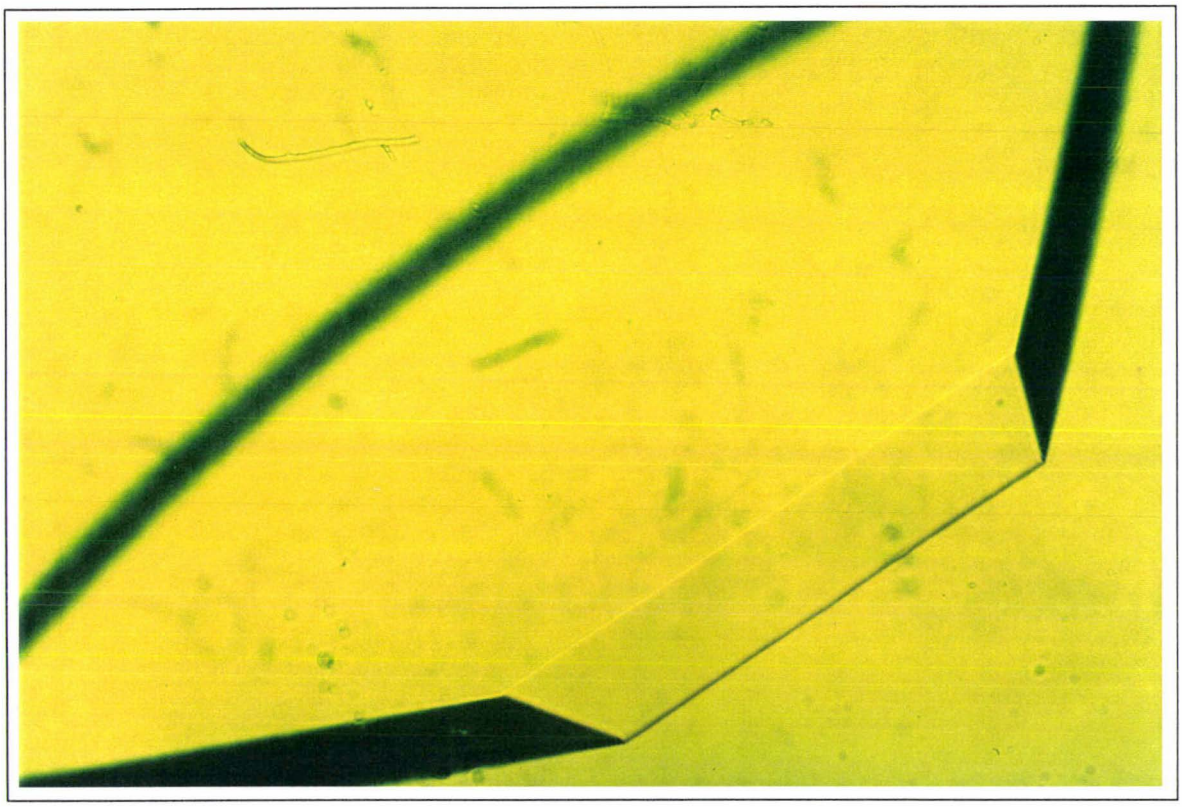


Figure 1.3 Crystal of *Zymomonas mobilis* GFOR

The crystal shown is about 1.5 mm in length and 0.2 -0.3 mm in thickness. Typical crystals are much smaller than this, and are usually quite thin and fragile.

1.3.3 $\alpha_2\varepsilon_2$ *embryonic hemoglobin*

The work in this section was carried out by Mrs. Heather Baker, as part of an ongoing research project investigating the structure of the hemoglobins of early human development. Since the ε-globin chain has substantial sequence homology with the β-globin chain [72] several initial crystallization experiments were conducted based around solution conditions which support crystal growth of the human $\alpha_2\beta_2$ adult hemoglobin [73, 74]. These factorial experiments involved variation in both pH and precipitant concentration, but resulted only in amorphous precipitation of the embryonic hemoglobin. Subsequently, search experiments based on orthogonal arrays were employed to try to identify suitable crystallization conditions. An initial search experiment used polyethylene glycols as protein precipitants, as given in Table 1.4, but covering a reduced pH range (6.1 - 8.5) (40 runs from the proposed 64). The embryonic hemoglobin was concentrated to 40 mg/ml in water, and the experiment conducted as described above. Crystals of the embryonic hemoglobin form in solutions buffered

at pH 8.5, in the presence of high concentrations of polymers (>21% (w/v)). At a lower pH, amorphous precipitation of the protein resulted. These crystals were enlarged by repeat seeding procedures and have subsequently been used to solve the structure at 2.9 Å resolution [A. J. Sutherland-Smith, unpublished work]. Subsequent experiments using salts as protein precipitants have not identified any further crystallization conditions. By way of comparison, a search experiment based on the solution conditions specified by Jancarik & Kim [23] did not identify any conditions under which the protein would crystallize.

1.4 RELATIONSHIP TO PUBLISHED SEARCH PROCEDURES

The justification for the use of orthogonal arrays as search experiments in protein crystallization was that the points in such an array were likely to be evenly or uniformly distributed throughout the experimental region. Consequently, it is hoped that the response among these points will be indicative of the response over the entire experimental region, and hence regions supporting nucleation and crystal growth can be identified with a smaller number of runs than would otherwise be possible.

By far the most widely used experiments to search for protein crystallization conditions appear to be the 'Crystal Screen' [23] and 'Crystal Screen II' experiments sold commercially (Hampton Research, California, USA). These experiments have no formal underlying structure (they are not factorial experiments). They are collections of solution conditions which have been used to crystallize proteins in the past. While these experiments are often successful, they are completely unsystematic. As a result, they will often fail to identify crystallization conditions where a more systematic approach will succeed, and they cannot be used to make useful inferences about how protein solubility and protein crystallization depend on the experimental variables.

The problem of designing systematic search experiments for protein crystallization has been previously addressed by Carter & Carter [75], and discussed further in subsequent papers [76, 20]. In these papers an alternative procedure for constructing a subset of a factorial experimental design is proposed. In this procedure, the points comprising the subset are essentially chosen by simple random selection without replacement. Two restrictions govern this selection; firstly, that each factor is represented in the subset 'a nearly identical' number of times at each level; and secondly, that each possible pairwise combination of levels occurs at least once in the subset [20]. Experiments resulting from this procedure have been termed incomplete factorial designs. It was proposed to analyze the results from such experiments using multiple linear regression.

This procedure has an interesting precedent in the statistical literature. In the late 1950's it was proposed [77, 78, 79] that useful experiments with a small number of runs could be derived from full factorial experiments by selecting a suitably sized subset at random from the treatment combinations comprising the full factorial experiment. Such experiments were known as random balance or random allocation [80] experiments. One variant of this idea was that the random sampling be conditional on each factor being represented in the sample a prearranged number of times at each level. It is clearly possible to extend this technique to balancing with regard to the combinations of factors. Such an approach would then appear essentially equivalent to the incomplete factorial method of Carter and Carter [75]. Dempster [80] suggested the name random allocation with partial balance for a procedure such as this.

In connection with model-based inference, there has been extensive discussion of the efficiency of such designs and of the potential difficulties associated with their interpretation [81, 82, 83, 84]. In terms of design properties, necessary and sufficient conditions for orthogonality of linear model effect estimates were given by Addelman [66], which may be useful in assessing the 'goodness' of designs generated by a random allocation or incomplete factorial procedure.

Here interest is centred on the (related) properties of arrays generated by random allocation as search designs in protein crystallization. For this purpose we require in some sense the even or uniform distribution of points throughout the experimental region. It is important to recognize that while random allocation designs appeal because of their inherent simplicity, random selection alone will not necessarily ensure uniform distribution of the experimental points throughout the experimental region. A random sample is not a representative sample 'in the sense that the sample is like the population or is a typical cross section of the population' [85]. Hence the importance of random selection conditional on certain balancing properties in the final sample (stratification of the sampling procedure), if uniform coverage of the experimental region is required.

Stratified sampling procedures have been studied in connection with computer experiments. Latin hypercube sampling was introduced by McKay and coworkers [86]. This is essentially random sampling subject to univariate stratification. Here the constraint is that for each factor, each level occurs with a fixed frequency in the final sample. A Latin hypercube sample is essentially an orthogonal array of strength 1. It has been shown that Latin hypercube sampling could be generalized using orthogonal arrays of any strength d, resulting in sampling plans that stratify on all d-variate margins simultaneously {50, 52]. In this paper, we have described search experiments for protein crystallization, based on the use of orthogonal

arrays. The procedure of Carter & Carter [75] will result in subsets which approximate the conditions of balance which orthogonal arrays of low strength fulfil exactly. Relaxing the mathematically restrictive requirement for orthogonality can result in a considerable reduction in experimental size. However, by relaxing these conditions, the experimental region is necessarily covered less completely. Search experiments based on orthogonal arrays provide a more balanced coverage of the experimental region, but are more expensive in terms of the required number of runs. The work described in this paper is essentially a compromise between search experiments based on large full factorial arrays (exemplified by the approach of Weber [27]), and search experiments based on the small partially balanced arrays generated by the procedure of Carter & Carter [75]. It was motivated by our need for a systematic search procedure for proteins which might not crystallize readily (i.e. for proteins which crystallize over small bounded ranges of continuous variables such as pH and precipitant concentration, or proteins whose crystallization shows a marked dependence on the chemical nature of the solution components (e.g buffer type)).

One advantage of random allocation designs is that for any given factorial experiment, a random allocation subset can always be quickly constructed. Notwithstanding this, useful orthogonal designs are increasingly accessible in the literature and elsewhere. Dey [45] comprehensively catalogues construction methods for orthogonal arrays. A collection of C programs for the construction and manipulation of symmetrical orthogonal arrays has been deposited at statlib (lib.stat.cmu.edu).

1.5 DISCUSSION AND CONCLUSION

1.5.1 Analysis using Linear models

The use of empirical models to analyze the results of crystallization experiments has been proposed by a number of authors (see e.g. [20, 25]). Note, however that these studies either ignore (or treat improperly) the correlation of linear model effect estimates caused by the non-orthogonality of their designs. In the construction of empirical models such as these, the lack of a quantitative and measurable response in protein crystallization is a considerable problem. It is always possible to rank the results of the experiments on an arbitrary numerical scale, and then proceed in the standard fashion, but what meaning is then associated with the coefficients of the linear model? Does the linear model explain the observations any better than a more qualitative description? Does it help in the design of further crystallization experiments? Any thinking about a complex phenomenon like protein crystallization is likely to involve elements of simplification and idealization, and hence rely on a model of some sort, even if this is extremely informal. Models of physical phenomena do not always have to be mathematical. Thus a statement such as 'the crystallization is dependent on the

pH', is a statement which implies a model. Properly designed experiments can support such inferences, whether they are made formally or informally; badly designed experiments can not. The use of multiple linear regression to analyze the results of experiments such as those described in this chapter is only useful if accompanied by careful thinking about the problem.

1.5.2 Distribution properties of orthogonal arrays

The use of orthogonal arrays as the basis for search experiments was justified on the basis that the points in an orthogonal array are usually spread regularly throughout the factor space. However it is clear that not all orthogonal arrays of a given strength are equal in this respect. Instead of a general formulation, this idea is illustrated by simple example. Consider two orthogonal arrays OA(8, 3, 2x2x4, 2):

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 2 \\ 0 & 0 & 2 \\ 0 & 1 & 3 \\ 1 & 0 & 3 \end{bmatrix} \qquad \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 2 \\ 1 & 0 & 2 \\ 1 & 1 & 3 \\ 0 & 0 & 3 \end{bmatrix}$$

Both are orthogonal arrays of strength two (represented geometrically in Figure 1.4), yet intuitively the second has a much more regular or even distribution of points (note also that one can be obtained from the other by permutation of the symbols in column 3). Formally we see that, without row rearrangement, the first can divided into two sub-arrays OA(4, 3, 2x2x2, 2)

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 1 & 1 & 2 \\ 0 & 0 & 2 \\ 0 & 1 & 3 \\ 1 & 0 & 3 \end{bmatrix}$$

While the second can be partitioned into three

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 2 \\ 1 & 0 & 2 \end{bmatrix} \qquad \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 2 \\ 1 & 1 & 3 \\ 0 & 0 & 3 \end{bmatrix}$$

Thus the intuitive idea of uniform distribution can be related to the sub-divisibility of an orthogonal array. This property needs to be investigated further. Analogously, Hedayat [49] has considered the subdivision of orthogonal arrays by column, which leads to useful insights about estimable effects in the analysis of experiments using linear models.

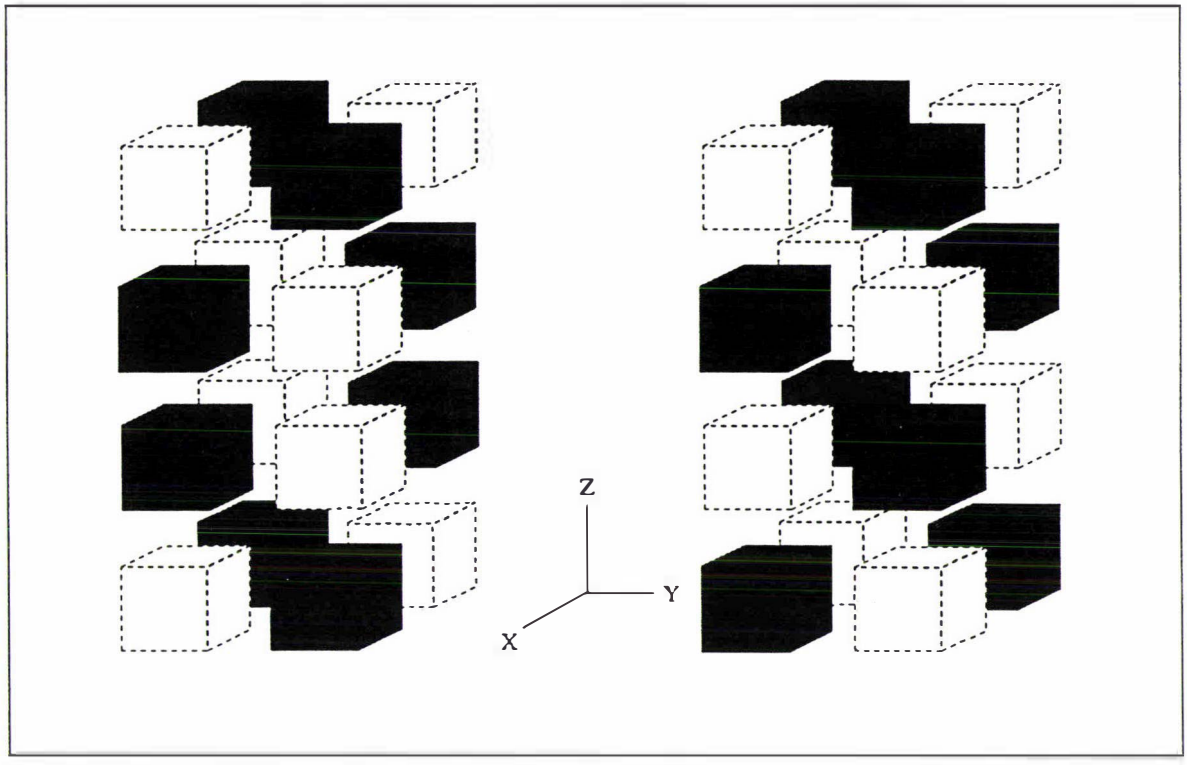


Figure 1.4 Geometric representation of two orthogonal arrays, OA(8, 3, 2x2x4, 2)

Orthogonal arrays are geometrically represented as in Figure 1.1.

1.5.3 Dynamic light scattering

Given the limitations of visual inspection, there is a clear need for a more efficient assay in the search for crystal nucleation conditions. Some recent studies of protein crystallization have used dynamic light scattering to monitor particle size distributions in solution, resulting in the identification of several necessary (but not sufficient) conditions for protein crystalli-

zation to occur (see e.g. [87, 88]). Its future use as a diagnostic technique for protein crystallization seems likely to depend on the extension of these studies from model systems to real experimental problems, and on the degree of miniaturization that can be achieved. Such physical techniques will still need to be associated with efficient experimental procedures for searching potential crystallization conditions.

1.5.4 Crystallization of other biological macromolecules

This chapter has concerned the construction of protein crystallization search designs for water soluble proteins, based on orthogonal arrays. In the examples given, the size of the underlying factorial experiments was fairly modest (for example, the experiment using polyethylene glycols would require 128 runs if conducted at a single temperature). For other biological molecules, the experimental situation is more complicated. In the case of membrane protein crystallization, detergents are required [53, 54]. For the crystallization of nucleic acids, cationic additives (polyamines and divalent cations) are required [89]. These additional solution components introduce further factors into the experimental structure (e.g. detergent type and detergent concentration in the case of membrane protein crystallization). Even if each of these factors have only two levels, this will result in very large factorial experiments. Orthogonal arrays might be particularly useful in such experimental situations.

1.5.4.1 Limitations of factorial experimental structure

Attempting to design search experiments for protein crystallization quickly highlights one disadvantage of describing the experimental region with a factorial structure. That is that certain sub-regions of the proposed experimental region may be physically unrealizable, as a result of phase separation or precipitation of the solution components. If the experimental region is still to be described by a factorial structure, the ranges of some of the variables must be reduced. However, this approach can be unsatisfactory, because interesting parts of the original region may be excluded from the experiment. Further consideration needs to be given to the design of search experiments where the experimental region is irregular (i.e. non-rectangular)

1.5.5 Conclusion

In conclusion, the use of experiments having an explicit factorial structure as search designs in protein crystallization is well established. The attraction of such experiments is that they provide a wide 'inductive basis for our conclusions' [90]. The difficulty with such experiments is that they rapidly become extremely large. In practice, this has led to the widespread

adoption of many unsystematic experimental approaches to searching for crystallization conditions, many of which lack any factorial structure. Here, what is in essence a sequential method of execution for factorial experiments has been proposed, based on the use of orthogonal arrays. These arrays provide a unified framework for considering the problem of searching for the physical and chemical conditions which will support protein crystal growth. The concept of the strength of such arrays provides some measure of the degree of coverage of the experimental region. Such an approach allows for a systematic exploration of the experimental region while keeping the experimental size within reasonable limits, and may be particularly important as the complexity of the experimental problem increases, as it does, for example, with membrane protein and nucleic acid crystallization.

Chapter 2

GFOR: STRUCTURE DETERMINATION

2.1 INTRODUCTION

The anaerobic Gram-negative bacterium *Zymomonas mobilis* occurs naturally in sugar-rich growth media [94, 95]. The bacterium ferments glucose, fructose and sucrose, utilizing the Entner-Doudoroff pathway, with ethanol and carbon dioxide as the principal products. Much of the work on *Z. mobilis* has been motivated by the technological interest in an efficient microbial process for ethanol production [96, 97]. An exceptional property of the bacterium is its tolerance of high concentrations of sugars and ethanol in the growth medium. In the presence of high concentrations of sugars, *Z. mobilis* produces substantial quantities of sorbitol [98, 99]. The sorbitol results from the reduction of fructose, a reaction which is coupled with the oxidation of glucose to gluconolactone [100].

Both reactions are catalyzed by a single enzyme, glucose-fructose oxidoreductase (GFOR) [101]. GFOR is an tetrameric enzyme, comprising four identical subunits. Each subunit contains one tightly, but non-covalently bound NADP molecule which is not released during the catalytic cycle. The enzyme operates by a ping pong mechanism, catalyzing the reaction of one of its substrates to yield a product that dissociates before the other substrate binds. Hence the overall reaction consists of two half reactions with alternate reduction of the bound $NADP^+$ (as glucose is oxidized to gluconolactone) and oxidation of NADPH (as fructose is reduced to sorbitol) [102]. The gluconolactone is subsequently converted to ethanol [101, 103], however sorbitol is not further metabolized by the cell.

GFOR is located in the periplasm of the bacterial cells [104, 105]. Here its proposed biological function is to protect the bacterium against osmotic stress caused by high external sugar concentrations [106]. The protective mechanism arises from the conversion of fructose into sorbitol, which is a compatible solute for the bacterium and can be accumulated in the cell. Steady state kinetic studies have shown that significant sorbitol formation will only occur in the presence of high concentrations of glucose and fructose (this is implied by the relatively large Michaelis constants for the two substrates; K_m for glucose 10.8 ± 0.8 mM, K_m for fructose 400 ± 30 mM [102]). This restricts the formation of sorbitol to conditions of hyperosmotic stress.

The export of GFOR into the periplasmic region is in accordance with the gene sequence for the enzyme [107]. There is a signal sequence of 52 amino acids preceding the N-terminal

sequence of the mature enzyme, a general feature of proteins destined for transport across cellular membranes [108]. The cleavage of the signal peptide at an Ala-Ala peptide linkage, gives rise to a protein of 381 amino acids (43 kDa) which corresponds to the observed subunit size in purified preparations of GFOR [101]. During the course of the X-ray structure determination it became apparent that there were conflicts with the published sequence [13] in several regions which could be explained by short frameshift errors in the original sequence determination. This has now been confirmed by resequencing of the gene [T. Wiegert, H. Sahm & G. A. Sprenger, personal communication]. The structural model reported in this thesis incorporates the corrected sequence.

The similar nature of the polypeptide chain organization in the dinucleotide binding domains of many NAD(P) and FAD binding enzymes is well documented [109]. The way in which this domain associates with NAD(P) places certain restrictions on the amino acid sequence. In particular the domain is commonly associated with the sequence GXGXXG or GXGXXA (where X is any amino acid), which forms a tight turn at the beginning of the dinucleotide binding helix [110]. GFOR contains such a fingerprint sequence (GLGKYA, corresponding to amino acids 38 - 43). This suggested prior to the structure determination the likely presence of the now familiar Rossmann fold.

GFOR is one of a small number of enzymes now known which use NAD(P) as an endogenous redox carrier (i.e. one which is not released during the catalytic cycle of the enzyme; see [111] and [112] for discussion). The enzymes which share this property are both structurally and functionally diverse. Some other examples include a formaldehyde dismutase (E.C. 1.2.99.4) from *Pseudomonas putida* [113], lactate-malate oxidoreductase (E.C. 1.1.99.7) from *Micrococcus lactilyticus* [114] and UDP-galactose 4-epimerase (E.C. 5.1.3.2) from *Escherichia coli* [115]. The only one for which the three-dimensional structure is known is UDP-galactose 4-epimerase [115]. In this thesis, the crystal structure of glucose-fructose oxidoreductase is reported, determined at 2.7 Å resolution. This has enabled us to account for the tight binding of NADP, and reveals an unsuspected structural and probable evolutionary relationship with the enzyme glucose-6-phosphate dehydrogenase.

In this chapter, a full description of the methods used in the structure determination of GFOR is given. Attention is focussed on the difficulties encountered during this study. Where appropriate, there is discussion of the theoretical background. A description of the structure itself, and discussion of its biological implications, is given in the following chapter.

2.2 OVERVIEW OF THE STRUCTURE DETERMINATION

The crystal structure of GFOR has been determined by the method of multiple isomorphous replacement (MIR) at an effective resolution of 2.7 Å. The structure determination was complicated by the occurrence of two closely related crystal forms, and difficulties in obtaining truly isomorphous derivatives. Two poor derivatives were obtained for one of the crystal forms. Real space electron density modification procedures were employed to improve the MIR map, and allow the building of an initial model. Subsequently, iterative combination of phase information from the partial model and the heavy atom derivatives, accompanied by model rebuilding and restrained least squares refinement against the intensity data allowed the determination of the missing parts of the structure. The model was fully refined using diffraction data from the second crystal form, for which it had not been possible to obtain good isomorphous derivatives, but for which the data set had a greater degree of completeness and a higher multiplicity.

2.3 PROTEIN PURIFICATION AND CRYSTALLIZATION

2.3.1 Cell growth and protein purification

Cell growth and protein purification were carried out by Professor Robert Scopes (La Trobe University, Australia). *Z. mobilis* cells were grown with 15% (w/v) glucose as substrate, harvested after fermentation had ceased, and lysed (see [116] for details). After centrifugation, the crude cell extract was passed through both negative and positive dye adsorbent columns as described previously [101]. GFOR was then further purified by cation exchange chromatography, using a Sepharose-S column, and the active enzyme precipitated with ammonium sulfate prior to crystallization.

2.3.2 Protein crystallization

Crystallization conditions for GFOR were identified using search experiments based on orthogonal arrays (see Section 1.3.2). Crystals were grown using hanging-drop vapour diffusion methods, and appear as thin, fragile plates from polyethylene glycol solutions buffered between pH 5 and 8. Crystals used in the MIR structure determination were grown at ambient temperature from 5 - 15% (w/v) PEG 6000, in the presence of 0.2M Succinic acid/KOH or 0.2M Citric acid/KOH buffer at pH 5.5. Protein concentration was between 10 and 30 mg/ml (by absorbance at 280 nm). Similar crystallization conditions have been reported by others [117]. Crystals of the same space group and morphology can also be grown in the presence of various additives, including high concentrations of sorbitol (a product of the

reaction catalyzed by the enzyme). The data used for the final refinement of the structure was collected from crystals grown as above, but with the inclusion of 0.8M sorbitol.

The crystals were not stable over time and would visibly deteriorate 2-3 weeks after their appearance (this involved a brown discolouration of the formerly clear crystals, visible fracturing, and disappearance of clean exterior faces). These degraded crystals could not be redissolved by dilution of the drop. The cause of this behaviour has not been identified.

2.4 X-RAY DATA COLLECTION

2.4.1 Characterization of the crystals

Two different (but morphologically indistinguishable) crystal forms grew under apparently identical solution conditions. Examination of the measured intensities showed that for both crystal forms the Laue class was *mmm*, and there were systematic absences (for all odd indices) along two of the axial directions. Consequently the space group for both crystal forms was established as $P2_12_12$. The first (Form I) had unit cell dimensions a = 84.82, b = 93.86, c= 117.02 Å, while the second (Form II) had cell dimensions a= 84.49, b= 283.69, c= 116.99 Å. They are related by a tripling of the *y* axis. Inspection of the diffraction pattern from the Form II crystals revealed the presence of an approximate sublattice, with every third reflection in the direction of the *y* axis being relatively strong. This effect is particularly pronounced in the low resolution terms. The Form I crystals seemed likely to contain 2 molecules in the asymmetric unit, and the Form II crystals 6. This corresponds to a Matthews coefficient [118] of 2.7 $Å^3Da^{-1}$, and a solvent content of ~ 54% in each case.

2.4.1.1 The Patterson and self-rotation functions

Tetrameric proteins are commonly associated with approximate or exact *222* point group symmetry [119]. The self-rotation function [120] of the Form I crystals was consistent with molecular *222* point group symmetry, revealing the presence of two-fold non-crystallographic symmetry axes in the *xy* plane, perpendicular to the crystallographic two-fold axis along *z* (Figure 2.1). The self-rotation function of the Form II crystals appears essentially identical to that of the Form I crystals (Figure 2.1). Inspection of the Patterson function calculated from the Form II data revealed two very large non-origin peaks at (0, 1/3, 0) and (0, 2/3, 0) (Figure 2.2). This, and the presence of an approximate sublattice in the diffraction data, indicated that the two crystal forms were related by a very slight packing rearrangement.

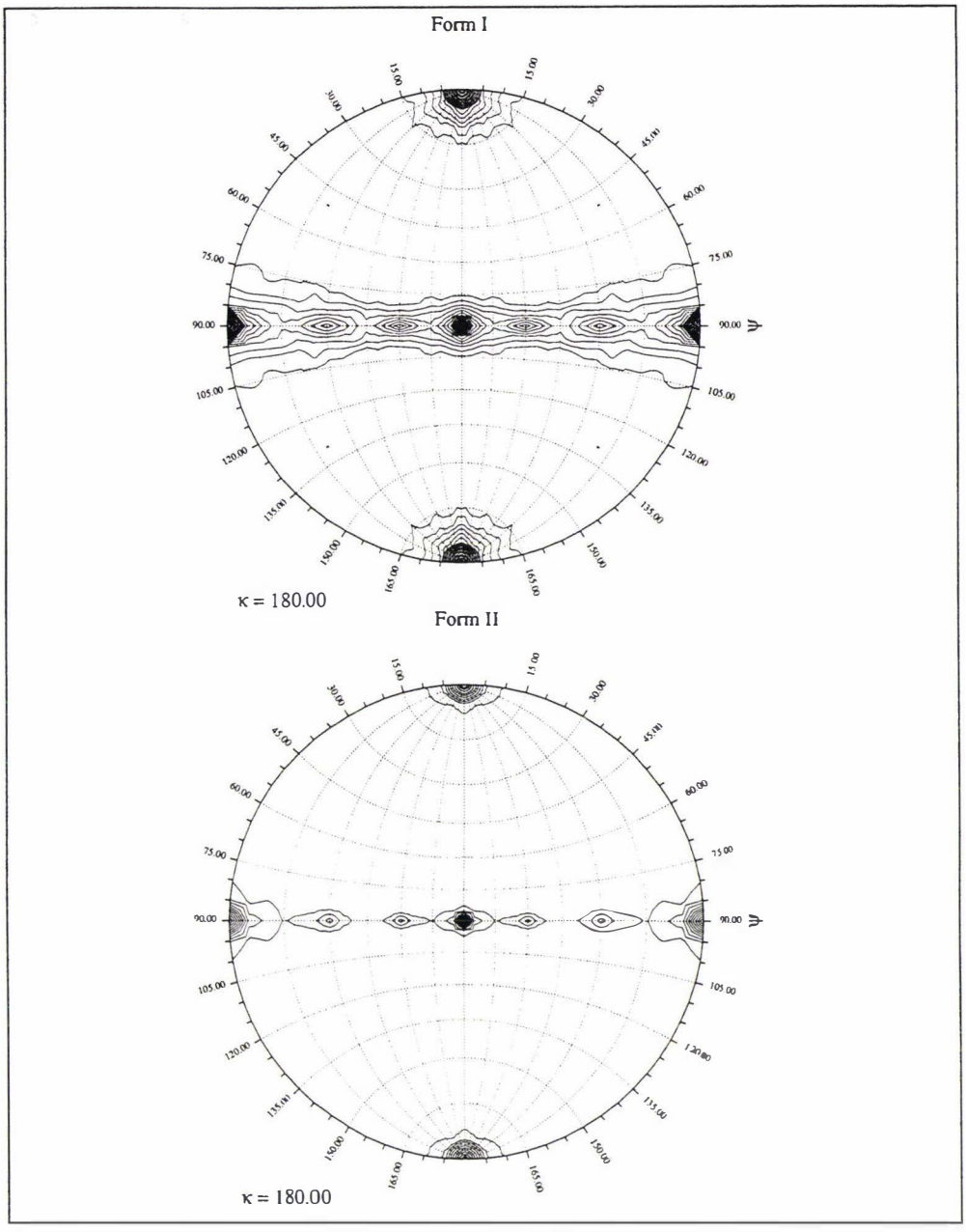


Figure 2.1 Stereographic projections of the self-rotation functions of the two crystal forms of GFOR.

The stereographic projections show the self-rotation function for all rotations with κ=180° (spherical polar angles) calculated from (top) the Form I crystal data; (bottom) the Form II crystal data (see Section 2.4.2 for details of data collection). In both cases the function is contoured starting at 3σ above the mean, with contour intervals of 2σ (where the mean and standard deviation have been calculated over the κ=180° section, not the entire rotation function). The self-rotation function was calculated by the reciprocal space method of Rossmann and Blow [120], using the program GLRF [121]. The calculation used terms from 20.0 - 2.7 Å resolution for the Form II data, and 20.0 - 2.5 Å for the Form I data. The following polar angle definition is used; φ is the angle from the cartesian x axis, and ψ the angle from the cartesian z axis. Only the largest terms (approximately 10% of the data) were used in the calculation, as discussed by Tollin and Rossmann [122]. When calculating the mean and standard deviation of the rotation function, the correction suggested by Yeates was applied [123] to compensate for the non-Euclidean nature of rotation space.

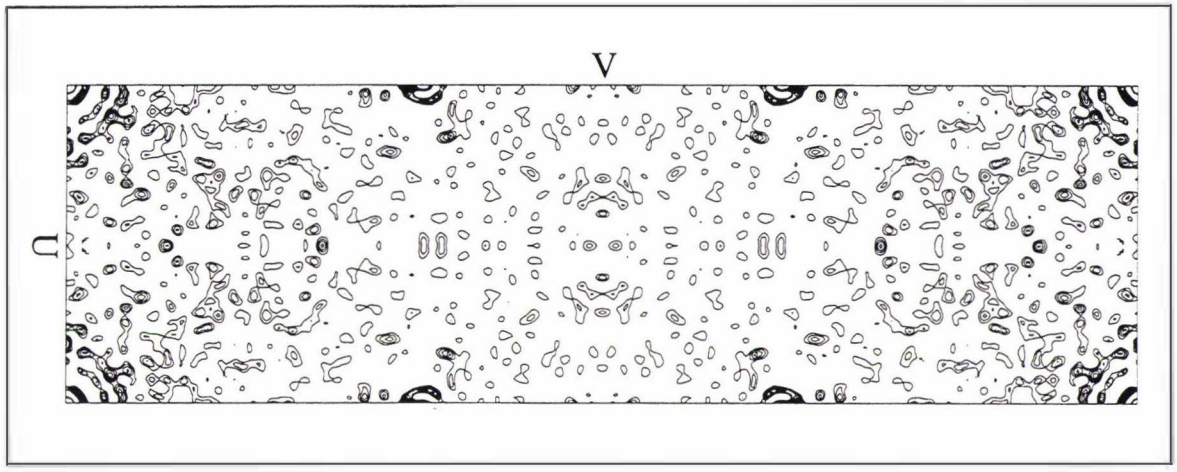


Figure 2.2 Patterson function calculated from the Form II data

The Patterson function was calculated using all observed data (see Section 2.4.2 for details of data collection). The section *(U, V, 0)* is shown (with 1σ contour intervals, starting at 1σ above the mean). The peaks at (0, 1/3, 0) and (0, 2/3, 0) are 25% of the origin peak height.

The X-ray structure determination confirms that the Form I crystals contain 2 monomers in the asymmetric unit. These belong to a tetramer with *222* point group symmetry with one of the symmetry axes of the molecule being coincident with a crystallographic two-fold axis. The other symmetry axes of the molecule lie in the *xy* plane, oriented approximately 30° from the crystallographic axes. The molecules pack in such a fashion that there are discrete layers of tetramers perpendicular to the *y*-axis. In the Form II crystals each of these layers is interleaved with two layers of tetramers with a slightly differing orientation (in which the molecular two-fold axes are no longer parallel to the *z*-axis) (Figure 2.3). This triples the apparent cell dimension along *y* and results in 6 monomers in the asymmetric unit. This accounts for the marked sublattice in the diffraction pattern of the Form II crystals, the indistinguishable self-rotation functions of the two crystal forms, and the presence of large non-origin peaks in the Patterson synthesis of the Form II crystals (which correspond to the simple translations between the almost identically oriented tetramers). The pattern of intermolecular contacts which leads to this slight rearrangement, has not yet been analyzed in detail.

2.4.2 Data collection and processing

2.4.2.1 Methods

Diffraction data from crystals of GFOR were collected by the oscillation method on a R-axis IIC system [124] utilizing the Fuji imaging plate as an X-ray detector and CuK_α radiation from a Rigaku RU-200 rotating anode generator operated at 50kV and 100mA. Profile-fitted relative intensities were obtained using the program DENZO [125]. Relative scaling of the

intensities was performed using the algorithm of Fox and Holmes [126]. Structure factor amplitudes were obtained from the intensity measurements employing the Bayesian treatment of French and Wilson [127]. This procedure also serves to put the data on an approximately absolute scale via a conventional Wilson plot [128]. Post-refinement of data was performed to check the estimated mosaicity of the crystals and to obtain reliable cell dimensions [129]. The scaling and merging of data was carried out using the CCP4 program suite [130].

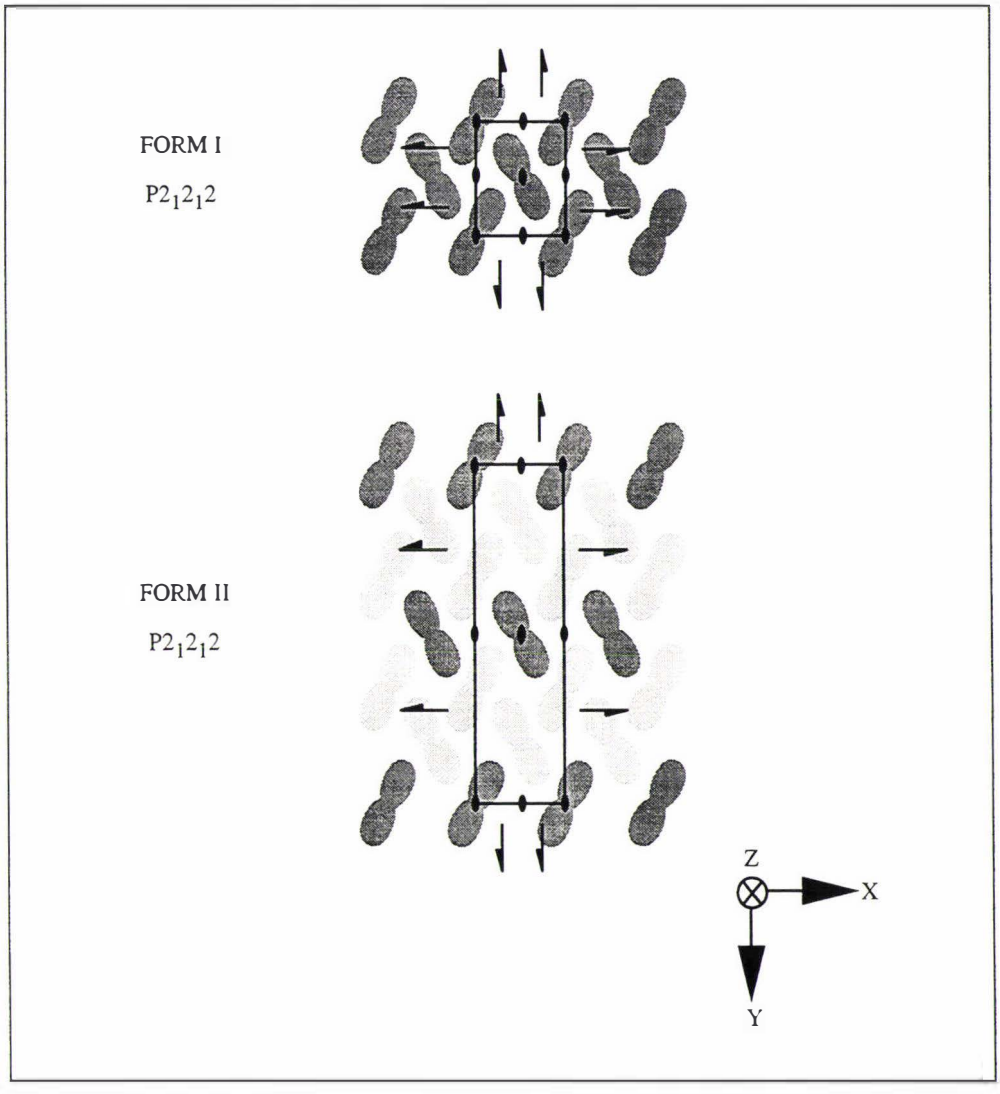


Figure 2.3 The relationship between the two crystal forms

Schematic diagram illustrating the relationship between the two crystal forms of GFOR. The tetrameric molecules (represented as pinched ellipsoids) are projected onto the *xz* plane of the crystal. Tetramers with a molecular two-fold axis parallel to *z* are shown in dark grey, tetramers having a slightly different orientation (in which a molecular two-fold axis is no longer perfectly parallel to *z*) are shown in light grey. The symmetry operations of the space group are shown according to the International Tables for X-ray Crystallography.

2.4.2.2 Experimental

Data collection was complicated by the existence of the two morphologically indistinguishable crystal forms. Form II crystals are markedly predominant, and only two native crystals of Form I have been observed. Heavy atom soaking experiments (Section 2.5) produced several derivatized Form I crystals, but whether these resulted from a conversion of Form II crystals is not known. On several occasions interconversion of the crystal forms (Form II → Form I) was observed during X-ray data collection. During the transition, the Bragg diffraction peaks became very diffuse, but became sharp again once the transition was complete. The diffraction limit of the crystals fell as a result of the transition, indicating that they were partially disordered by this process. The process could be monitored by the number of effectively observed reflections, which fell by two-thirds over the time course of the transition. The factors which promote this interconversion are not known, and it has been observed very infrequently.

A data set was collected at room temperature from one of the Form I crystals. This crystal suffered quite severe radiation damage, with an accompanying increase in crystal mosaicity, and decrease in diffraction intensity (see [131]). At the time, no equipment was available to cool crystals during data collection (which will often slow radiation damage [132]). A complete data set could not be collected from this single crystal (the data set was 79% complete to 2.5 Å resolution, see Table 2.1). It should be noted that in almost all respects this is a poor data set. The overall redundancy is low, the data is very weak beyond 3 Å resolution, and the data set is incomplete across the entire resolution range.

Several native data sets were collected on Form II crystals. The long *y*-axis cell dimension made data collection difficult, as large crystal to detector distances were needed to achieve effective spot resolution. A helium box was employed to help prevent X-ray attenuation by air. The most complete, and most reliably estimated data were collected at 4 °C from a total of 4 Form II crystals grown in the presence of 0.8M sorbitol. Because of the fragility of the GFOR crystals, and the potential for a solid state phase transition between the two crystal forms, a special crystal mounting procedure was adopted for this data collection, used recently in the structure determination of bovine F_1 ATPase [133]. In this procedure crystals are mounted in capillaries filled with mother liquor (see also [134]). With respect to GFOR, this mounting technique offers several advantages. Firstly it prevents mechanical damage to the crystals. Using conventional mounting techniques it was very easy to shatter the thin plate-like crystals as the liquid was withdrawn from around the crystal. Secondly it maintains the crystals in a controlled physical state (as discussed in Section 2.4.3 below, solid state

phase transformations in protein crystals are commonly linked to hydration or dehydration of the crystal). This allowed data to be collected and merged from several crystals without encountering problems due to non-isomorphism. The principal disadvantage of the technique is the increased background scattering due to the solvent which surrounds the crystal.

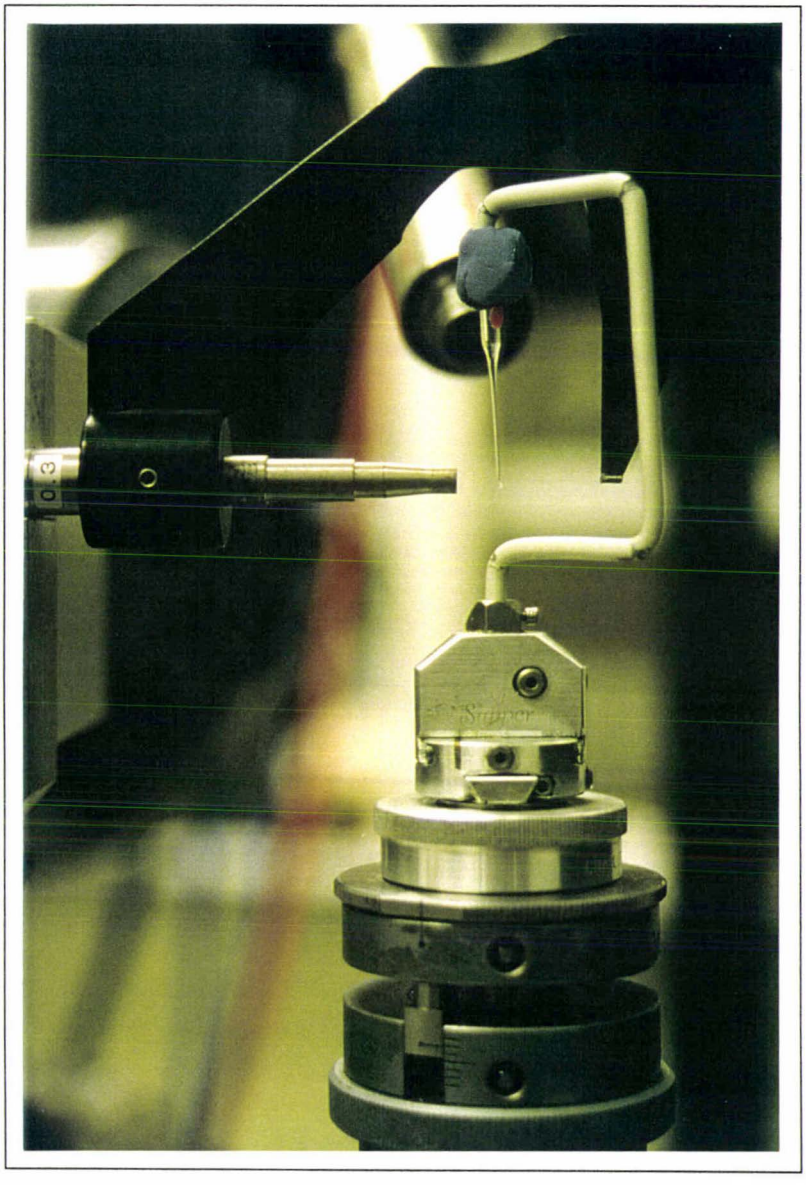


Figure 2.4 Data collection using crystals mounted in liquid-filled capillaries

The photograph shows a liquid-filled capillary, suspended from a special jig used in data collection. The crystal is wedged at the fused end of the capillary, and held in place by gravity.

In practice, a thin-walled glass capillary (0.5 mm diameter) was drawn out in a flame, fusing the end of the capillary. The capillary was filled with liquid, and a crystal introduced at the liquid surface. The crystal drops down under gravity until it wedges itself at the fused end of

the capillary, the top of which can then be sealed with wax. For data collection the capillary can be suspended from a special jig (see Figure 2.4). It was found that crystal slippage was not a significant problem.

The Form II data is much more reliably estimated than the Form I data. The redundancy of the data set is high (mean redundancy for each observation 4.8), and it is essentially complete to 2.9 Å resolution (loss in completeness beyond this limit is largely due to the finite detector size, and the long crystal to detector distance). In the outer resolution shell (2.87 -2.70 Å) the mean value for each intensity divided by its estimated standard deviation ($\langle I/\sigma(I)\rangle$) is 2.7, and the completeness 60%. The overall R_{merge} is 9%.

Table 2.1 Native data processing statistics. Form I crystals

30 oscillation images (2° oscillations) collected from a single crystal were used in data processing. On each image, data were integrated to a resolution limit at which the mean $I/\sigma(I)$ in a thin isotropic resolution shell fell below 2.

	Upper resolution limit (Å)								
	5.00	3.97	3.47	3.15	2.92	2.75	2.61	2.50	All
No. of measured reflections*	9702	9486	9180	8628	7993	6534	4477	2853	58853
No. of rejected reflections†	246	271	118	35	14	2	1	0	687
No. of unique reflections	3484	3515	3513	3499	3454	3249	2866	2400	25980
Completeness (%)	80	84	85	86	85	80	71	59	79
$\langle I/\sigma(I)\rangle$	11.7	8.9	6.3	4.2	2.5	2.1	1.9	1.9	5.7
R-merge (%)§	4.9	6.7	9.9	15.0	24.3	30.3	34.2	36.5	10.3

* The total number of integrated observations used in data processing.

† The number of observations rejected as outliers during data processing. The fall-off with increasing resolution reflects the decreasing multiplicity of the data set, making it more difficult to detect aberrant measurements.

§ $R_{Merge} = \sum_{hkl}\sum_{j}\left| I_j(hkl) - \langle I(hkl)\rangle\right| / \sum_{hkl}\sum_{j}\left| I_j(hkl)\right|$ where $I_j(hkl)$ are the symmetry-equivalent intensity measurements for a reflection and $\langle I(hkl)\rangle$ is their weighted mean value.

All attempts to freeze the crystals at liquid-nitrogen temperatures have failed, seeming to result in a partial lattice transformation between the two crystal forms. The data from the Form I crystals (Table 2.1) were used in the structure solution by MIR, the data from the Form II crystals (Table 2.2) were used in subsequent refinement of the structural model. The Form II crystals diffract to well beyond 2.7 Å resolution, however a high resolution data set has not

yet been collected due to the practical problems described above.

Table 2.2 Native data processing statistics. Form II crystals

115 oscillation images (1.5° oscillations) collected from 4 different crystals were used in data processing. On each image, data were integrated to a resolution limit at which the mean I/σ(I) in a thin isotropic resolution shell fell below 2. Definition of terms as for Table 2.1.

	Upper resolution limit (Å)						
	4.91	3.89	3.40	3.09	2.87	2.70	All
No. of measured reflections	85588	80982	81540	57672	31909	15509	353200
No. of rejected reflections	2491	2921	1391	357	52	7	7219
No. of unique reflections	13399	13053	12971	12856	11520	7701	71500
Completeness (%)	99	100	100	100	89	60	91
<I/σ(I)>	8.7	7.9	6.0	3.4	3.3	2.7	6.3
R-merge (%)	5.9	7.4	10.8	15.7	21.1	26.5	9.0

2.4.3 Space group transitions

It is not uncommon for proteins to crystallize in more than one form. Some proteins (for example bovine pancreatic ribonuclease [135, 136]) have a very large number of known crystal forms, and in some cases changes to the crystal space group can result from seemingly minor changes in solution composition [137]. Phase transitions between protein crystal forms are more rarely observed. Examples exist of proteins with crystal forms related by a doubling of one of the cell axes (see Table 2.3). In two of these cases [92, 93] a temperature-dependent phase transition between the two forms was observed. Several other examples of phase transitions in protein crystals have now been reported [138, 139, 140, 141]. In all cases the transitions have been brought about by changes in temperature, or in the composition of the solution from which the crystals were grown.

For GFOR, it is not known what factors are responsible for the crystal polymorphism, or under what conditions the rarely-observed phase transition between the two crystal forms takes place. Some preliminary experiments were carried out in order to see if the transition was temperature-dependent. GFOR crystals were grown as described, but with the addition of small quantities (5%(v/v)) of ethylene glycol, glycerol or PEG400 to the buffer-precipitant solution. These small organic molecules serve to depress the solution freezing point. Crystals grown under such conditions (and conventionally mounted) were then cooled from room temperature until the freezing point of the crystals was reached (typically at -20 to -30 °C), monitoring the X-ray diffraction. A nitrogen-gas-stream cooling device [142] (Oxford Cry-

osystems) was used to control the temperature. There was no evidence of a phase transition in the crystals (all Form II) over this temperature range (20 to -30 °C). It may be that hydration or dehydration of the crystals promotes the phase transition, however this remains to be verified by experiment. A proper evaluation of the physical behaviour of these crystals (establishing a phase diagram) would require a great deal of experimental effort.

Table 2.3 Protein crystal forms related by doubling of cell axes.

Protein	Spacegroup	Cell Dimensions						Reference
		a (Å)	b (Å)	c (Å)	α (°)	β (°)	γ (°)	
Inorganic	R32	110.6	110.6	154.7	90.0	90.0	120.0	[91]
pyrophosphatase	R32	111.4	111.4	76.6	90.0	90.0	120.0	
δ-crystallin	$P2_12_12_1$	100.6	133.4	140.2	90.0	90.0	90.0	[92]
	$P2_12_12$	99.9	133.6	69.2	90.0	90.0	90.0	
Pyruvate	P1	81.0	82.4	233.2	69.5	72.6	62.4	[93]
decarboxylase	P1	81.0	82.4	116.6	69.5	72.6	62.4	

2.5 MULTIPLE ISOMORPHOUS REPLACEMENT

While the search for heavy atom derivatives remains an essentially empirical process, several authors have discussed the chemistry of heavy-atom attachment to proteins [143, 144, 145], and the experimental approach adopted reflects the biochemical information available on GFOR at the time. In all of the experiments, pre-existing crystals were soaked in solutions containing the heavy atom compound (as opposed to reacting the protein with the compound in solution prior to crystallization). Only two derivatives were obtained, both for the Form I crystals (details are given in Table 2.4). Efforts to obtain further derivatives were consistently frustrated by the predominance of Form II crystals. Derivatives obtained for the Form II crystals were all poorly isomorphous with the native crystals, and were not ultimately used in the structure determination.

The nucleotide sequence of GFOR indicated that the protein contains five cysteine residues. While it was not known if these amino acids were involved in disulfide linkages, it seemed that there would be at least one free sulfhydryl group in the protein (in fact the subsequent revision of the sequence shows that the protein contains only four cysteine residues, the structure determination that none are involved in disulfide linkages). Consequently the first reaction attempted was a mercuration of accessible cysteine residues. Crystals soaked in ethyl mercury phosphate showed large intensity differences from the native crystals.

Table 2.4 Derivative data collection and phasing statistics: Form I crystals

	Ethylmercury Phosphate	Chloro (2,2′:6′,2″-terpyridine)platinum(II) chloride
Number of crystals	1	1
Temperature	Ambient	Ambient
Maximum resolution (Å)	3.30	3.30
No. of measured reflections	43478	25446
No. of unique reflections	13601	11583
Completeness (%)	93	79
R_{merge} (%)*	10.6	7.4
Soaking concentration (mM)	1	1
Soaking time (hours)	22	72
No. of binding sites	4	2
Binding Locations	Cys54, Cys158	His308
R_{Cullis}†: acentric (centric)	0.82 (0.71)	0.93 (0.88)
Phasing power§: acentric (centric)	1.16 (1.12)	0.63 (0.50)

* R_{Merge} as previously defined (Table 2.1)

† $R_{Cullis} = \sum_{hkl}|F_H(obs) - F_H(calc)| / \sum_{hkl}|F_H(obs)|$ where F_H is the heavy-atom structure factor.

§ Phasing power = $\sum_{hkl}|F_H(calc)| / \sum_{hkl}|F_H(obs) - F_H(calc)|$

In order to determine the heavy atom positions, a Patterson synthesis was computed with coefficients $(|F_{PH}| - |F_P|)^2$, where $|F_{PH}|$ is the derivative structure factor amplitude, and $|F_P|$ the native structure factor amplitude (see [146] for discussion). Scaling of the native to derivative data was by the method of Kraut *et al* [147]. By examination of the difference Patterson function it was possible to identify a set of four heavy atom positions that accounted for all the principal peaks on the Harker sections. The four heavy atom sites could be divided into two sets, which are related to one another by a two-fold rotation in the *xy* plane (consistent with the results of the self-rotation function). From this it was evident that the mercury compound bound to equivalent positions in the two crystallographically independent molecules. The presence of non-crystallographic symmetry could have been used to directly aid the Patterson map interpretation [148], but this did not prove necessary. An

anomalous difference Patterson function, calculated with coefficients $(\Delta|F|_{ano})^2$ [149], was noisy, but consistent with the assignment of the heavy-atom positions. The Form I data are not of sufficient accuracy to obtain reliable estimates of the small effects due to anomalous dispersion. Refinement of the heavy atom positions and occupancies, and computation of phase estimates was done with the program MLPHARE [150].

Subsequently, a second derivative was obtained using the compound chloro (2,2′:6′,2′′-terpyridine)platinum(II) chloride [151]. Heavy atom sites were determined by inspection of a difference Fourier synthesis calculated with coefficients $(|F_{PH}| - |F_P|)$, and SIR phases computed from the mercury derivative. The atomic positions derived in this fashion were checked for consistency with the difference Patterson synthesis as described above. The platinum compound attached to each molecule at a single site.

As previously mentioned, both the mercury and platinum compounds could also be used to prepare derivatives of the Form II crystals. Unfortunately, these were only poorly isomorphous with the native crystals (results not shown). Trimethyllead acetate [152] was also used to prepare a poorly isomorphous derivative of the Form II crystals, but in this case a Form I analogue could not be prepared.

2.6 DENSITY MODIFICATION

Unsurprisingly, the initial electron density map was largely uninterpretable (the mean figure of merit to 3.3 Å resolution was 0.36, although the version of MLPHARE used to perform this calculation slightly underestimates the true figure of merit [E. J. Dodson, personal communication]). The initial MIR phases (to 3.3 Å resolution) were improved and extended (to 2.5 Å resolution) using real space density modification procedures (2-fold averaging, histogram matching, and solvent flattening), as implemented in the DEMON program suite [153]. At later stages of the structure determination, phase information from the partial model and the heavy atom derivatives was combined (at 3.3 Å resolution), and then essentially the same density modification protocol used to improve and extend the phases.

2.6.1 Envelope definition

By definition, non-crystallographic symmetry (ncs) extends only a finite distance through the crystal. Application of electron density averaging procedures requires not only that the ncs operations are known, but also that the volume to which they apply is defined. In order to do this, the assumption was made that tetrameric GFOR possessed exact *222* point group

symmetry (proper non-crystallographic symmetry, in the terminology of Bricogne [154]). As has been discussed, there was no indication from either the self-rotation function or the results of isomorphous replacement of large deviations from *222* symmetry. This assumption simplifies the definition of the molecular envelope since it is not necessary to define envelopes corresponding to the individual subunits, only an envelope corresponding to the oligomer (in other words, it is not important to know which part of the molecular envelope corresponds to which subunit). This point has been discussed in more detail elsewhere [155, 156, 157].

2.6.1.1 Refinement of the non-crystallographic symmetry operation.

Before proceeding, it was necessary to accurately define the non-crystallographic symmetry operation. There are several ways of approaching this problem. Using the refined heavy atom positions, it is possible to calculate the orientation and position of the non-crystallographic symmetry element relating these sites by standard vector superposition procedures. However this takes no account of any variation in the accuracy with which the heavy atom positions have been determined. An elegant procedure for determining the position of a non-crystallographic rotation or screw-rotation axis was suggested by Blow *et al* [158]. Given the orientation of a symmetry axis and the translational elements along that axis, the location of the axis in the unit cell can be determined by correlating the electron density of the molecules related by the symmetry operation. This is in essence a real space translation function, and can either be computed directly (see [159]), or expressed as a Fourier summation [158, 160] (this latter approach has been implemented in the program GLRF [121]).

Direct computation of the correlation between electron density was employed. The program CCMAX2 in the DEMON/ANGEL software suite [153] was used to search for the maximum correlation between regions related by the ncs operator, while modifying the rotational and translational parameters. The constraint of exact *222* point group symmetry means that the rotation must be exactly 180°, and that the rotation axis must pass through the point group centre (at the special position (*0.5, 0.5, z*)), and lie in the xy plane. Therefore there are only two degrees of freedom associated with the definition of this symmetry operator. The program CCMAX2 was modified to allow constrained optimization of the rotational and translational parameters. The correlation coefficient was evaluated over a sphere (radius = 15 Å), centered on the best current estimate of the point group centre (since at this stage the true molecular envelope was unknown). The starting estimates for the axial orientation and position were obtained from the self-rotation function, and the real space translation function as implemented in the program GLRF [121]. The optimization was done by simple grid searching on successively finer grids.

2.6.1.2 Envelope definition

Once the ncs operation had been refined, an envelope enclosing the protein region was calculated from a local correlation map [153]. This procedure was first suggested by Rees and coworkers [161]. The function evaluated is

$$r(x) = \frac{\sum_{sphere} [\rho(x_1) - \langle \rho(x_1) \rangle] \times [\rho(x_2) - \langle \rho(x_2) \rangle]}{\left[\sum_{sphere} [\rho(x_1) - \langle \rho(x_1) \rangle]^2 \times \sum_{sphere} [\rho(x_2) - \langle \rho(x_2) \rangle]^2 \right]^{1/2}} \qquad \textit{Eq. 2.1}$$

the correlation coefficient between spheres of electron density (ρ), centered on the points x_1 and x_2, which are related by the ncs operation. The radius of the sphere over which the correlation coefficient is calculated is critical to the appearance of the function. A sphere of radius 6 Å seemed to give satisfactory results. An envelope was created from the local correlation map by assigning all points in the map above a fixed threshold to the molecular envelope. The envelope volume was chosen so that the protein would have a partial specific volume that was typical of globular proteins (see [162]). The envelope obtained in this fashion was smoothed and modified to erase isolated islands and fill enclosed voids [163]. Interestingly, tetrameric GFOR proves to have a large solvent-filled cavity at its centre. This cavity was clearly evident in the initial molecular envelope.

An alternative procedure when the molecular symmetry operators are known, and are proper, is to average the entire isomorphous replacement map using these operators, without regard to the molecular or asymmetric unit boundary (envelope-free averaging). Such a procedure should enhance the electron-density features within the molecular envelope, and simultaneously suppress features outside as a result of 'mis-averaging' their electron densities [156]. However this procedure has been found in practice to be less effective than envelope definition based on a correlation map [R.J. Read, personal communication].

At later stages, when the structural model had been largely completed, envelope generation was carried out starting from the coordinates of the molecular model. Grid points within a fixed distance of any atomic position were assigned to the envelope [153].

2.6.2 Phase improvement and extension

2.6.2.1 Iterative density modification

The general philosophy of iterative density modification has been well described in a number of recent reviews [156, 155, 164, 165]. Here some specific features of the density modification procedures applied to GFOR are described.

2.6.2.2 Real space averaging

The operation of electron density averaging was carried out using an input map computed on a fine grid (ca 1/5 of the high resolution limit as suggested by Bricogne [159]). Electron density values at sub-grid intervals are computed by eight point linear interpolation. Rossmann and coworkers [166] have suggested that Bricogne's estimate of the required grid sampling is unduly pessimistic, and that a coarser grid can be used, which would result in a substantial saving in computer time.

2.6.2.3 Solvent flattening

The electron density in the region outside the molecular envelope was set equal to its mean value at each cycle of density modification (solvent flattening [167]).

2.6.2.4 Histogram matching

Histogram matching (or histogram specification), was introduced as method for phase refinement and extension in the late 1980's [168, 169, 170, 171]. In this technique, a histogram of density values is computed from the current image of the structure (the electron density). A mapping is found which transforms this histogram into an ideal one. This is essentially a way of incorporating prior information about the electron density distribution. Typically, the histogram matching procedure is applied to points in the molecular region only. When combined with solvent flattening this provides constraints on the electron density over the entire volume of the crystal. Histogram matching within the DEMON program suite is based on the implementation of Harrison [168]

To use the information inherent in an electron density histogram for phase refinement or extension, it is necessary to know what the true (unknown) histogram should look like. In practice it has been found that if calculations are restricted to the molecular region, the form of the electron density histogram is independent of the structure itself. This means that the electron density histogram of a known structure can be used to predict the histogram of an

unknown structure. However, there are two important caveats: (1) The histogram is dependent on the resolution of the electron density map; (2) the histogram is dependent on the mean atomic displacement parameter of the structure used to calculate the map.

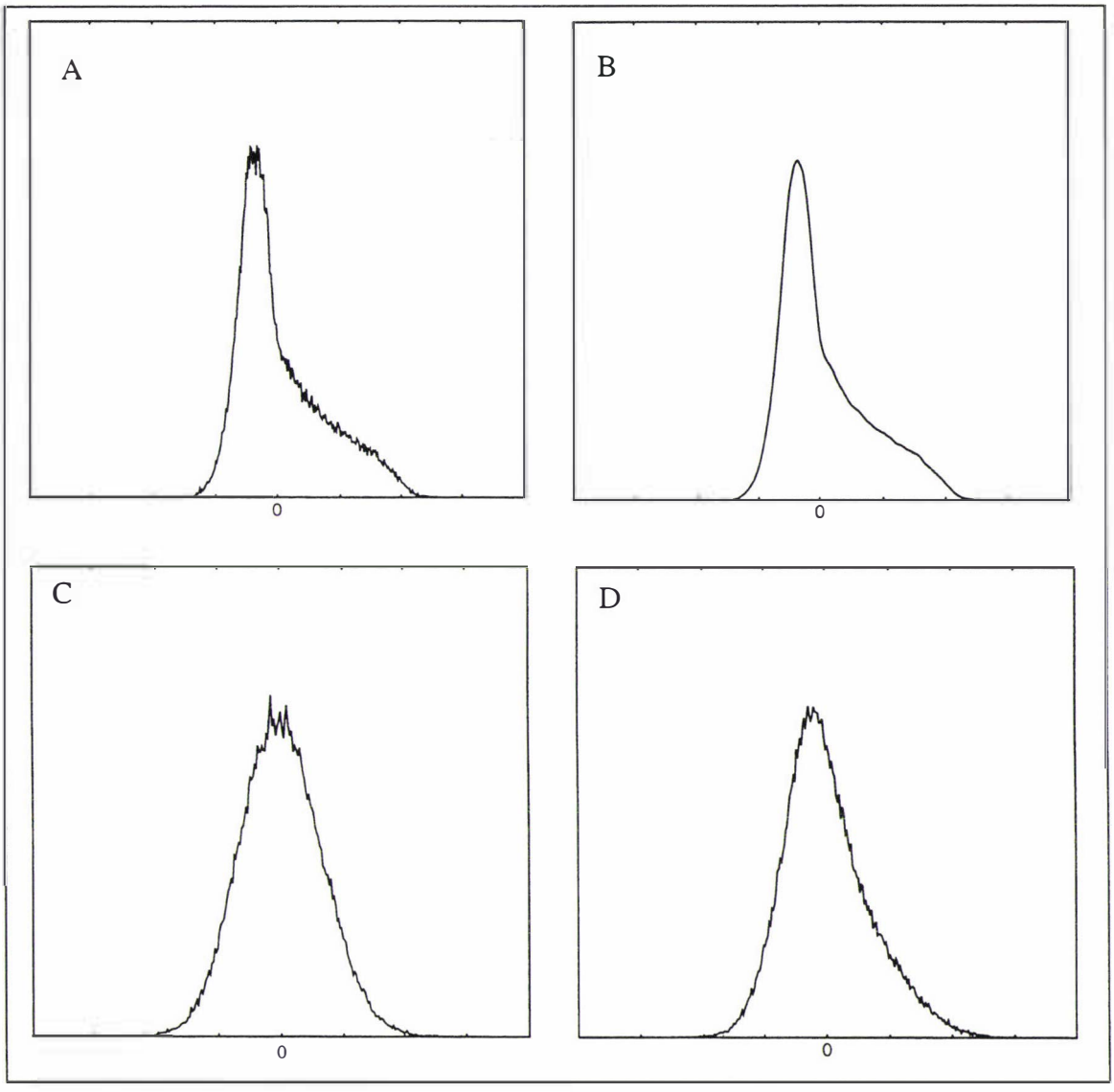


Figure 2.5 Calculated and observed electron density histograms

Electron density histograms, (A) calculated directly from the structure of *B. sterolicum* cholesterol oxidase at 3.3 Å resolution, (B) calculated from the structure of *B. sterolicum* cholesterol oxidase at 3.3 Å resolution, and smoothed, (C) calculated from the MIR phased map at 3.3 Å resolution, (D) calculated from the MIR phased map at 3.3 Å resolution after phase improvement by real space density modification (2-fold averaging, histogram matching and solvent flattening). Since the scales are arbitrary (only the form of the histogram is critical) they have been omitted (electron density is plotted on the horizontal axis, relative frequency on the vertical axis). In each case 250 bins were used to compute the histogram.

Electron density histograms were calculated from the crystal structure of *Brevibacterium sterolicum* cholesterol oxidase, refined at 1.8 Å resolution [172]. This was an arbitrary choice. The atomic displacement parameters of this structure were incremented so that their mean value was 40 $Å^2$, consistent with the expected value for GFOR (based on a conventional Wilson plot). Since phase extension was being performed, and the shape of the electron density histogram depends on the resolution of the image, histograms were calculated from a series of maps with a upper resolution limit between 3.3 and 2.5 Å (with a 0.1 Å interval). The molecular region was defined by an envelope computed from the structural model of cholesterol oxidase.

Electron density histograms calculated in this fashion show small random fluctuations superimposed on the overall form of the curve (see Figure 2.5 (a)). To eliminate this, the curves were smoothed (using locally weighted regression [173]) before being used in density modification procedures (see Figure 2.5 (b)). Also shown is the electron density histogram calculated from the initial MIR phased map (see Figure 2.5 (c)), demonstrating the potential for phase improvement by this method.

Several authors have reported an extension of this method which takes into account the local environment of each grid point in the map [174, 175]. This development, while offering a clear improvement over density modification procedures based on conventional electron density histograms, was too recent to be implemented in this work.

2.6.2.5 Phase extension

A concise formulation of the theory of phase extension has been given by Lawrence [156], and this will not be repeated. Unphased observations were introduced during the phase extension process by adding data in a thin isotropic resolution shell to the current set of phased observations (the shells were of width 0.01 Å in the extension from 3.3 to 2.5 Å resolution). In practice this resulted in the introduction of between 100 and 230 reflections at each extension step (12200 reflections were initially phased by MIR). More properly the shell width at each extension step should be governed by the volume of the added shell (which would avoid adding progressively more reflections each step at higher resolution). The density modification procedure was iterated three times at a fixed resolution, before the introduction of the next shell of reflections. Before phase extension was initiated, a number of cycles of density modification were performed at fixed resolution, iterating until convergence had been achieved.

2.6.2.6 Amplitude completion

In the computation of electron density maps during each density modification cycle, missing structure factor amplitudes were substituted with calculated structure factor amplitudes (from backtransformation of the previous electron density map). Systematic omission of these unobserved data from map calculation slows the rate of convergence. This assumes some importance because of the incompleteness of the Form I data. The theoretical reasons for this have been discussed by Rossmann [155]

2.6.2.7 Weights associated with the modified phases

A Sim weighting procedure [176, 177] was initially employed in the density modification protocol, to estimate the weight associated with the modified phases at each iteration. As discussed by Cowtan and Main [178], such a weighting scheme has little theoretical justification. The conventional Sim weighting procedure, applied in iterated density modification calculations, results in overestimation of the phase accuracy, and a spurious inflation of the weights. Consequently, at a later stage, the 'free-Sim' weighting method of Cowtan and Main was implemented in the DEMON program suite.

In calculating the Sim weight for the modified phase, it is required to estimate Σ_Q, the expected square structure factor amplitude of the difference structure between the modified and true maps. Various empirical estimates of Σ_Q are in use, of which that of Bricogne [159] was used in this density modification protocol

$$\Sigma_Q = \langle |F_{obs}|^2 - |F_{mod}|^2 \rangle \qquad \textit{Eq. 2.2}$$

where $|F_{obs}|$ is the observed structure factor amplitude, and $|F_{mod}|$ is the structure factor amplitude calculated by backtransformation of the density-modified map. In the 'Free-Sim' weighting method a small fraction of the data is omitted from the map calculation at each density modification cycle. This data is used in the estimation of Σ_Q by Equation 2.2. This estimate is then used in the calculation of weights for the rest of the reflections. In order to avoid systematically biasing the phase estimates a different (randomly selected) set of data is omitted each cycle. This is obviously an empirical solution to the weighting problem (as all weighting schemes currently used in density modification are), but in practice it results in much more realistic estimation of the weights (largely uninterpretable maps do not now have figures-of-merit which uniformly approach 1).

2.6.2.8 Phase combination

In the initial application of the density modification procedures, the calculated phases (from backtransformation of the modified map) were combined with the experimental phases by multiplication of their respective probability distributions (see Section 2.7.4 for discussion). Thus, combined weights and phases were obtained for the calculation of the map for the next cycle. At later stages, the experimental phase probability distribution was ignored (phase combination was not carried out), and the calculated phases (and associated weights) were used directly.

2.6.3 Results of Phase improvement and extension

At this stage no formal *a posteriori* analysis of the density modification protocol has been carried out. The general effectiveness of the techniques of solvent flattening, histogram matching and electron density averaging in phase improvement and extension is now well established. That it provided striking improvements in this particular case can be seen by inspection of Figure 2.6, which shows the electron density at the core of the tetramer at varying stages of the structure determination.

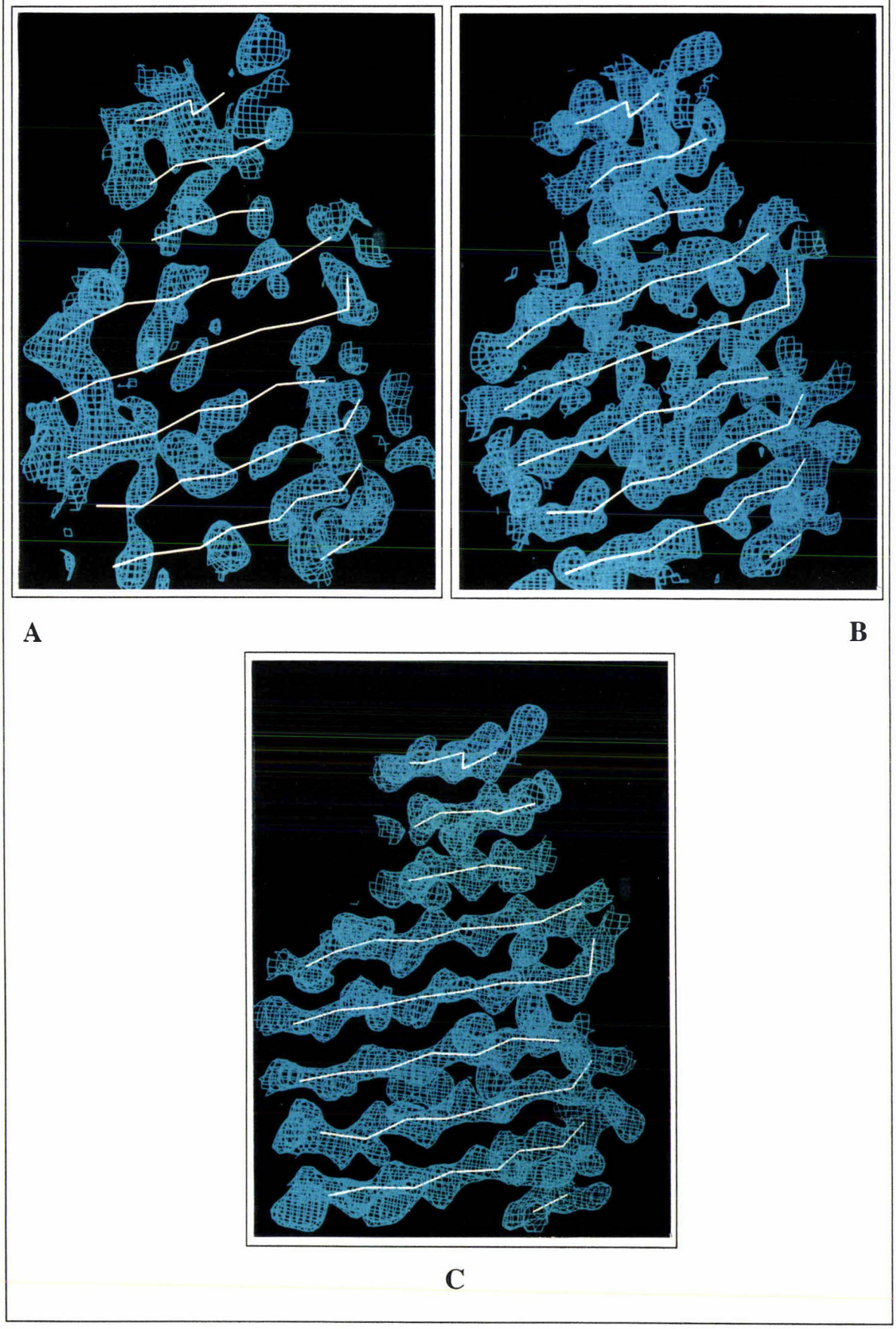
2.7 MODEL BUILDING AND REFINEMENT

2.7.1 Building the initial model

Inspection of the electron density map obtained after density modification (calculated at 2.5 Å resolution, using the Form I data), revealed regular structural features such as α-helices and β-strands. The helices were right-handed, indicating that the correct hand had been chosen for the heavy atom structure in the isomorphous replacement procedure [179]. It was also clear that the map was not going to be of sufficient quality to allow unambiguous tracing of the entire polypeptide chain of the molecule. The initial objective then, was to build the polypeptide backbone in those regions of the map where regular structural features could be recognized. Since the map had been averaged, only the structure of a single subunit was being considered.

Figure 2.6 (following page) Electron density maps for GFOR

Electron density maps calculated from, (A) the initial MIR phases at 3.3 Å resolution (Form I data), (B) the phases after iterative phase improvement and extension to 2.5 Å, using real space density modification (Form I data), (C) the phases calculated from the final refined model at 2.7 Å (Form II data). In all cases the electron density maps are contoured at 1.5σ above the mean. Electron density corresponding to the central β-sheet of the molecule is shown (for clarity the Cα trace of the strands of the sheet is superimposed).


A
B
C

The model building was carried out on SGI workstations using at first the program TOM/FRODO [180], and at a later stage the program TURBO-FRODO [C. Cambillau, A. Roussel, A.G. Inisan and E. Koups-Mouthuy]. Polyalanine fragments, corresponding to regular α-helices and β-strands were generated and placed in the map. Small adjustments to their regular geometry usually allowed adequate fitting. It should be emphasized that the quality of the initial map was generally poor, but that knowledge of the regular structures found in proteins meant that a model with reasonable conformation could be fitted to such electron density. This point has been emphasized by others [181]. During this process it became apparent that the tetramer must be constructed around a stacked β-interface [182]. There were clearly two β-sheets in the map, stacked on top of one another, and related by the two-fold non-crystallographic symmetry axis.

In several of the strands in the β-sheet area, the density was broken along the polypeptide backbone, sometimes at more than one point (see Figure 2.6 (b)). In addition, the strands were frequently connected along the hydrogen-bonding direction in the sheet. This is a common feature of poorly phased experimental electron density maps calculated at this resolution [183]. It is also a frequent source of error in chain tracing (for example, in the incorrect structure of 6-phosphogluconate dehydrogenase, the model had a U-turn across a hydrogen bond in a β-sheet converting what should have been 3 parallel strands into an anti-parallel structure [184]). Hence, the map was interpreted with special caution in this region. The connections between the strands in the sheet, and to adjacent helices, were in general poorly defined and could not be confidently determined. As a consequence of this, the direction of the strands could not always be unambiguously assigned (since the electron density corresponding to the carbonyl oxygen atoms was often unclear). No attempt was made at this stage to interpret the partial structure in terms of the protein sequence. The quality of the map would have made this a difficult undertaking at best.

2.7.2 Recovery of the missing structure

The initial model comprised 16 polyalanine fragments, containing a total of 207 amino acids (out of a total of 381). This constitutes only 1/3 of the number of atoms included in the final model. A significant proportion of the initial electron density map could not be interpreted. It was hoped that this situation might be improved by calculating an electron density map whose phases resulted from combination of the original MIR phase information and that obtained from the partial structure [185]. Phase improvement and extension could be carried out starting from this map using the real-space density modification procedures previously described. Prior to performing such phase combination, refinement of the initial model was carried out, in order to improve the agreement with the experimental observations.

2.7.3 Model refinement

Refinement of a very incomplete model in reciprocal space is problematic because it is unclear how much of each observed amplitude is due to the missing part of the model. This problem has been considered by Bhat and Blow [186], and Terwilliger and Berendzen [187] but neither of these procedures were used in this study (this problem is also implicitly addressed in the development of maximum likelihood refinement methods, see Section 2.7.3.3 for discussion). The problem does not occur in real space where it is possible to refine the model against the relevant electron density independent of the effects of uninterpreted regions of the map. Consequently real space refinement methods were employed for the initial refinement. The limitation of this approach is that the phases were fixed during this process, and still contained large errors, so the potential for improving the partial model was restricted. In real space refinement, the function minimized is

$$f(x) = \sum_{Space} (\rho_o(r) - \rho_c(r,x))^2 \qquad \textit{Eq. 2.3}$$

where ρ_o is the observed electron density, ρ_c the electron density calculated from the model, x an atomic model vector, and r a general vector. This refinement was carried out using the program TNT [188, 189].

2.7.3.1 Least squares refinement and the difference Fourier synthesis

The relationship between least squares minimization and the difference Fourier synthesis (computed with coefficients $\left((F_o - F_c)\, e^{i\alpha_c} \right)$ was established some time ago [190, 191, 192, 193]. In fact 'difference Fourier refinement' was widely used in the early stages of protein crystallography (see [194, 195, 196]). Later it was shown by Agarwal that the gradient vector for the least squares calculation could be computed from a difference Fourier synthesis by a relatively simple transformation [197]. Thus while the argument is qualitative, effective refinement in reciprocal space by least squares methods would seem to require that difference Fourier syntheses be interpretable. Unsurprisingly, difference maps calculated during the initial stages of model building were almost completely uninterpretable (as would be expected with only 30 - 40% of the structure modelled, and this with relatively large coordinate errors). Consequently, in the initial stages of the structure determination, when the model was still very incomplete, real space (phase-invariant) refinement was employed, switching to reciprocal space refinement when difference Fourier syntheses became inter-

pretable. Reciprocal space least squares refinement was also carried out using the program TNT [188, 189].

2.7.3.2 Treatment of non-crystallographic symmetry.

At all stages during the initial refinement, the two independent copies of the monomer in the asymmetric unit of the Form I crystals were constrained to be identical. The justification for this is straightforward; recent analysis shows that many of the reported differences between ncs-related molecules in structures determined at moderate resolution are likely to be artifactual [198]. It was felt that any genuine differences between the molecules could be accounted for at a later stage in the refinement. The program TNT handles ncs constraints quite elegantly. During refinement a single prototype molecule is expanded by the ncs operators, then the calculated shift vectors for each molecule are transformed by the ncs-operations, averaged, and applied to the prototype.

2.7.3.3 Maximum likelihood Refinement

Protein structure refinement has traditionally been carried out by minimization of a conventional least-squares residual. It has been pointed out that such a minimization is not strictly appropriate, since the conditions necessary for the use of the least-squares residual are not met in protein structure refinement. Refinement procedures based on maximum likelihood analysis have many advantages over the traditional approach, and a number of practical implementations have very recently been reported [199, 200, 201]. In particular, Bricogne [199] has emphasized the potential of this approach in the refinement of incomplete models, and in the provision of information to help complete the partial structure. With respect to the refinement of GFOR, it is clear that a maximum likelihood approach, by virtue of its ability to deal more correctly with the phase uncertainties introduced by model inaccuracies and incompleteness, would have been superior, and may have considerably alleviated the difficulties encountered during the structure determination. However, given that the initial model was essentially only 30% complete, its seems likely that many of the procedures described in this chapter would still have been necessary, although the process of refinement and rebuilding would have been considerably aided by the availability of the new refinement methods.

2.7.4 Combination of phase information

Following real space refinement of the model, phases calculated from the partial structure were combined with the MIR phases at 3.3 Å resolution, using the program SIGMAA.

The problem of combining two sources of phase information, for example phases from a partial model and phases from isomorphous replacement, is not trivial. No rigorous theoretical treatment of the problem is available. Rossmann [202] proposed that the combination of independent sources of phase information could be achieved by multiplication of the corresponding phase probability distributions. Thus

$$P(\alpha) = P_{iso}(\alpha)\, P_{par}(\alpha) \qquad \textit{Eq. 2.4}$$

where $P(\alpha)$ is the combined phase probability distribution, and $P_{iso}(\alpha)$ and $P_{par}(\alpha)$ are the isomorphous and partial structure phase probability distributions respectively. A computationally convenient representation of the phase probability distributions was suggested by Hendrickson and Lattmann [203].

This approach was used by Rice in his work on the combination of partial structure and isomorphous phase information [185, 204]. Phase probability distributions from isomorphous replacement were derived by the method of Blow and Crick [205]. For the partial structure, the phase probability distribution was based on the analysis of Sim [176, 177], using the procedure of Bricogne [159] to evaluate the contribution of the missing structure (see Section 2.6.2.7). The probability distributions were multiplied as described above to give the combined phase estimate.

While this procedure proved effective, it has a number of clear limitations. Firstly as discussed elsewhere [159], it is often not justified to consider the phase information from isomorphous replacement and the partial structure as independent. Additionally, the probability analysis of Sim neglects the effect of coordinate errors in the partial structure. Alternative procedures for combining these two sources of phase information have been proposed by Stuart and Artymiuk [206] and Read [207]. While they differ in detail, the basic idea behind both of these approaches is the same; the expressions for the map coefficients vary according to the extent to which the model determines the phase. The approach of Read (program SIGMAA) has been adopted here.

2.7.5 Iterative cycles of rebuilding, refinement and phase combination

Following combination of the phase information from the partial model and isomorphous replacement, the density modification protocol was repeated and an improved map calculated. This procedure was repeated iteratively, and slowly allowed the missing structure to be built. At each stage the model was refined by restrained least squares, using the program

TNT [189]. Non-crystallographic symmetry constraints were employed in refinemnet, at all stages.

2.7.5.1 Sequence assignment

Following several rounds of phase combination, the Rossmann fold of the dinucleotide binding domain could be clearly identified. This also allowed the association of the sequence with the partial structure.

The direction of several strands in the central β-sheet of the C-terminal domain had to be reversed as the connections with other secondary structural elements became clear. Starting from such a poor initial map, we were very cautious in determining the connectivity and assigning the sequence. Supporting the sequence assignment, the derivatives were found to have bound in chemically reasonable positions (Table 2.4). At this stage, it also became clear that there were three regions in which the structure could not be reconciled with the published gene sequence. These regions are discussed further in Section 2.7.6.3.

It was unfortunate that one of the errors in the published sequence occurred a short distance from the GXGXXG motif. After tracing the sequence through the first helix of the dinucleotide binding domain and into the second β-strand, it became apparent that the sequence and the structure conflicted. Since the connectivity of the secondary structural elements in this domain was not unequivocally established at this stage (and had been partly inferred from the assumption that the topology of the Rossmann fold would be conserved), this held up progress for some time, until it was clearly established that there were no other possibilities for the connectivity. At this stage, it had not been recognized that the conflicts with the sequence could be explained by frameshift errors. In fact this was not unequivocally established until refinement was begun using the Form II data (Section 2.7.6 below), for which the maps were much less ambiguous.

2.7.5.2 The use of a dummy-atom procedure to build the N terminal region

One difficult region to build was the N-terminal region from each subunit, which in the structure forms an 'arm' which wraps around an adjacent subunit in the tetramer. This was poorly connected, and we could not confidently fit an atomic model. A dummy atom procedure was employed to model this region at first, placing 30 dummy atoms along the presumed backbone, and using the globic scattering factors suggested by Guo *et al* [208] in structure factor calculations to 3.3 Å resolution. In this approximation, the X-ray scattering of amino acids is approximated by their spherically-averaged Fourier transform.

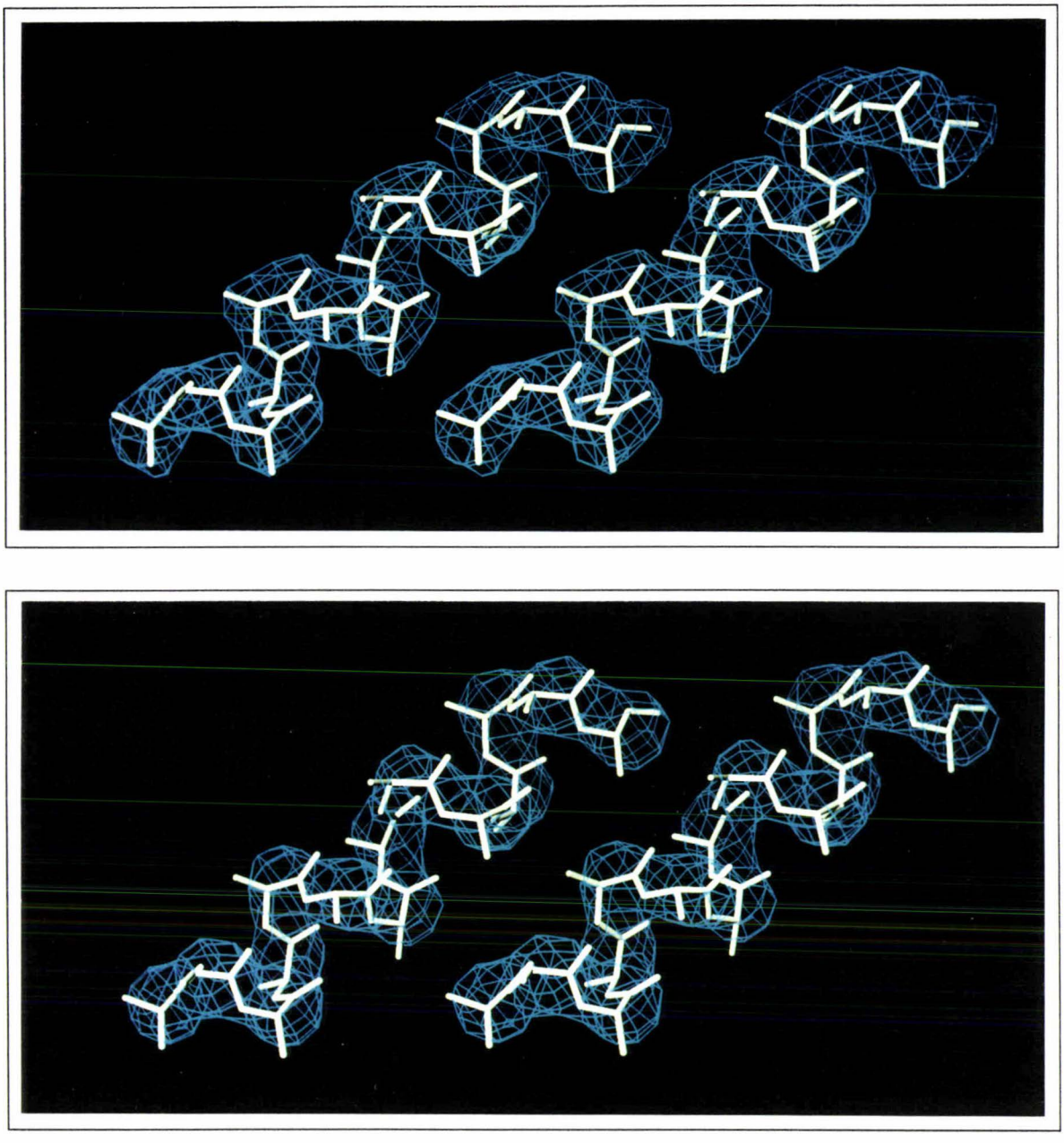


Figure 2.7 Stereoview of electron density maps calculated from an atomic and a 'globic' representation of an α-helix at 3.0 Å resolution.

Electron density was calculated from, (A) a conventional atomic representation of an ideal α-helix (all alanine), (B) a 'globic' representation of the same helix. For the first calculation individual atomic scattering factors were employed; for the second the globic scattering factors suggested by Guo *et al* [208]. Individual isotropic atomic displacement parameters (B=20 $Å^2$) were used in both cases. The structure factor calculation was carried out in space group P1 (cell dimensions a=b=c= 40 Å, α=β=γ=90°). Both electron density maps are contoured at 0.5 electrons/$Å^3$.For all terms to **3.3** Å resolution, the correlation coefficient between the two sets of structure factors amplitudes is 0.95, the conventional crystallographic R-factor (with no relative scale factor) 0.27 (0.42 for reflections between 3.42 and 3.30 Å resolution), and the mean phase difference 31° (59° for reflections between 3.42 and 3.30 Å resolution).

Positions of the dummy atoms were refined (by reciprocal space least-squares methods) prior to the phase calculation. Crude contact restraints were enforced between the dummy atoms during this refinement procedure (the exact distance between the centre of mass of each peptide unit is dependent on the backbone torsion angles). That the 'globic' scattering factor approximation is realistic at low resolution can be seen by inspection of Figure 2.7.

This procedure improved the phase-combined maps to the extent that unambiguous building of this region was then possible. At this stage the R-factor for data to 2.5 Å resolution was 33%, and the free R-factor (for 1300 reflections omitted from all refinement procedures) 37%. Once the model was largely complete, refinement was begun using the data from the Form II crystals. The improvement in electron density maps calculated using this data was striking.

2.7.6 Final Refinement of the model

The molecules were roughly positioned in the Form II cell by expansion of the Form I model using the crystallographic symmetry operators. Rigid body refinement of the individual subunits (of which there are 6 in the asymmetric unit of the form II crystals) quickly converged. The much improved quality of the Form II maps allowed the definitive identification of the sequence conflicts as frameshift errors in the sequence determination. It also allowed correction of a number of other minor errors in the model. Final refinement of the model was by restrained least squares using the program TNT [189], with the geometry library of Engh and Huber [209]. Refinement was against all Form II data to 2.7 Å, employing a bulk solvent correction for the low resolution terms (Section 2.7.6.5 below).

2.7.6.1 NCS constraints

The strict ncs constraints employed at the start of the refinement have been maintained throughout, thus the determined structure represents an average of the six copies in the asymmetric unit of the Form II crystals. At several stages rigid body positional refinement of the molecules was employed to check and improve the ncs symmetry operators. Deviations from exact *222* point group symmetry for the GFOR tetramers are very small. However, correlation coefficients calculated during map averaging procedures suggest that there are genuine differences between subunits. These differences are between those subunits related by the non-crystallographic two-fold rotation axes which lie (approximately) in the xy plane. It seems that these symmetry operations are not exact. However given the small magnitude of the differences (six-fold averaged maps are everywhere interpretable), we do not feel justified in modelling them at 2.7 Å resolution.

A question which has not yet been properly addressed is the effect of errors in the noncrystallographic symmetry operations on the refined structure. Clearly this will depend on the nature of the non-crystallographic symmetry, and the problem may be less acute in the case of high non-crystallographic redundancy, where averaging over many different orientations should prevent systematic distortions of the model.

2.7.6.2 Cross-validation

We monitored the free R-factor during the course of refinement against the Form (II) data. Predictably, because of the reciprocal space relationships between the structure factors created by the non-crystallographic symmetry (some of which are almost purely translational), it simply mirrored the conventional R-factor. For example, in the final stages of refinement, when the conventional R-factor was 21%, the free R-factor, for 1032 randomly-selected reflections omitted from all refinement procedures, was 22%. It is clear that random selection of the reflections used for the free R-factor calculation is not sufficient in cases such as this. All measured data were used in the final refinement cycles.

2.7.6.3 Identification of frameshift errors in the published nucleotide sequence

The conflicts between the X-ray structural results and the derived sequence for GFOR could be explained by short frameshift errors in the published nucleotide sequence [107]. There appeared to be a total of three such errors. Once it had been established that frameshift errors had occurred, the inferred sequence was quickly fitted, with the only uncertainties arising in the determination of the frameshift boundaries. The electron density in all three regions unambiguously supported these conclusions (see Figure 2.8), although it was difficult to be certain of the sequence assignment at the frameshift boundaries.

An independent redetermination of the gene sequence has confirmed these results [T. Wiegert, H. Sahm & G.A. Sprenger, personal communication; Genbank accession number Z80356]. This also acted as a validation of the correctness of the structure (with the exception of several residues at the frameshift boundaries, the errors in the sequence had been correctly identified). The frameshift errors affect the following regions of the published sequence (numbering according to that sequence); Met 59 - Thr 66, Leu 194 - Ser 204, Val 377 - Gly 387. One result of the frameshift errors is that the mature enzyme is now seen to be 381 amino acids in length, in contrast to the 387 amino acids previously reported [107].

2.7.6.4 Modelling the ordered water structure

In the crystal almost all the space not occupied by protein atoms is filled with solvent. Both X-ray and neutron diffraction studies have shown that some water molecules are sufficiently well ordered for their electron density to be resolved. Water sites detected in this fashion reveal favoured water positions, or equivalently, measure the extent to which a given site is occupied by water (for recent reviews see [210, 211, 212, 213]).

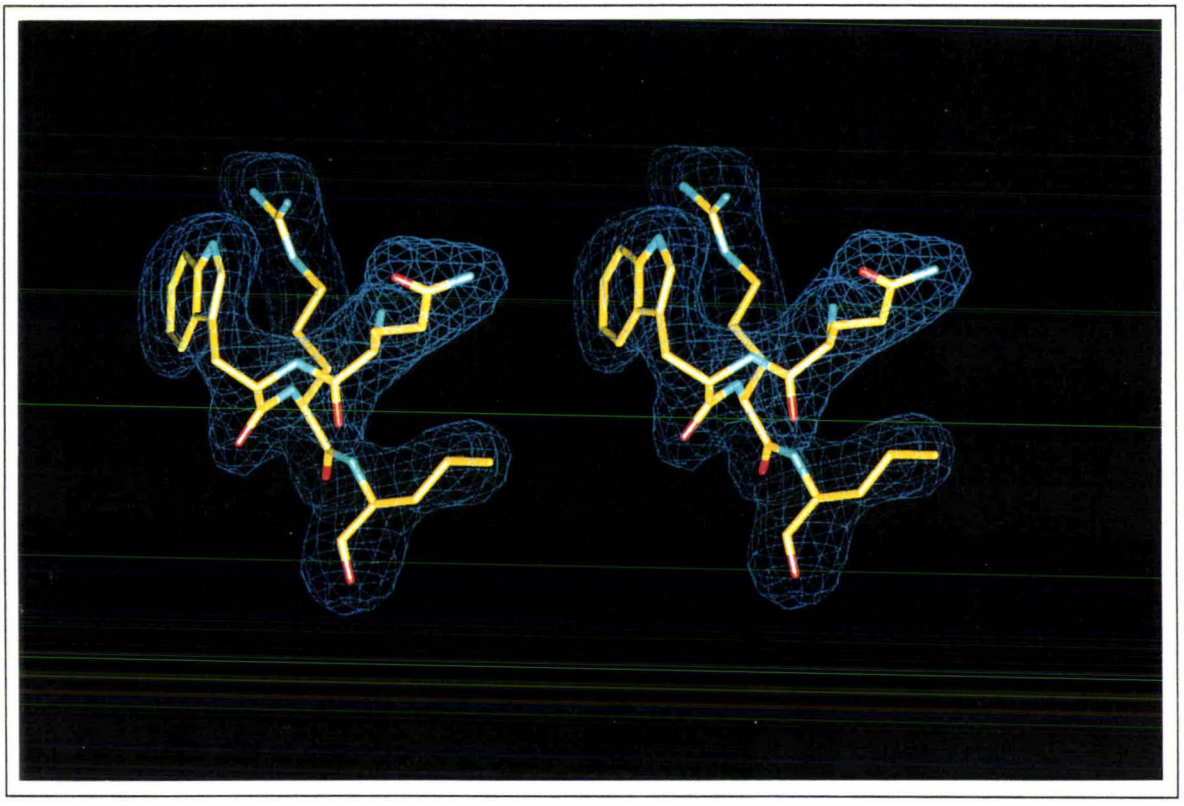


Figure 2.8 Stereoview of a difference Fourier synthesis with residues in the region conflicting with the published sequence omitted.

The difference Fourier synthesis was calculated with residues 198 -201 (Gln Trp Arg Leu) omitted from the model. The omitted residues and the corresponding density are shown. These four residues are in a region which conflicted with the published nucleotide sequence (the corresponding residues in the published sequence are Gly Val Cys Val). The map was calculated with all data to 2.7 Å resolution and contoured at 3.0σ. Fourier coefficients employed in the map calculation were of the form $(m|F_o| - D|F_c|)$ (SIGMAA weighting) where $|F_o|$ is the native structure factor amplitude, $|F_c|$ is the calculated structure factor amplitude, and m and D have been defined by Read [207]. The figure was prepared using Turbo-Frodo [C. Cambillau, A. Roussel, A.G. Inisan and E. Koups-Mouthuy]

Modelling of the ordered solvent structure was not begun until the late in the refinement, by which stage a number of potential sites (inside or at the surface of the protein) were clearly

resolved. Water molecules were assigned to peaks in difference Fourier syntheses (typically contoured at 3σ above the mean), subject to the constraint that hydrogen bonds were formed with protein atoms or other water molecules. Care was taken not to include water molecules in regions were the protein structure was still ambiguous. During refinement, no restraints were placed on the positional or atomic displacement parameters of the water molecules.

The constraint that the subunits be identical (applied throughout the refinement) extends to the water structure associated with each subunit. Clearly this will not be a good approximation in some surface regions, since the water structure will reflect the differing local environment of the subunits. Some of the modelled water molecules may also represent other solvent species. These limitations in the modelled water structure should be recognized.

The ordered water molecules seen in protein structures determined by X-ray crystallography may in fact have quite short residence times [214], and be in rapid exchange with the bulk solvent. This makes refinement in terms of a conventional occupancy and atomic displacement parameter problematic (this point has been the subject of some debate [215, 216, 217]). Especially at moderate resolution, this problem appears intractable. The procedure adopted during this refinement was to fix the occupancy of the water molecules at a value of 1.0, and refine for each an isotropic displacement parameter. For those water molecules for which the displacement parameter refined to high values (>70 $Å^2$), the occupancy was set to 0.5. Any of these water molecules for which the displacement parameter subsequently refined to high values were removed from the model.

2.7.6.5 Correction for the contribution of the bulk solvent.

The ordered water structure seen adjacent to the protein surface, and accounted for explicitly in the structural model, represents only a fraction of the total solvent in the crystal. The effective scattering from the disordered solvent regions cannot be neglected at low resolution, where its effects cause large discrepancies between the observed and calculated structure factors [218]. This clearly has important implications for structural refinement. Because of this effect, it has been common in the past to omit the low resolution terms during refinement. However, systematic omission of the low resolution terms will cause series termination errors in Fourier syntheses (see [219]).

A number of methods have been proposed to address this problem. Empirical weighting schemes have been suggested, which in effect employ a special scale factor for the low resolution terms (see e.g. [220, 221]). In an approach with more clear reference to the physical basis of the problem, the solvent region is modelled by a grid, with each grid point assigned

a uniform value of electron density. Structure factors for the solvent region can be calculated by Fourier transformation, and modified by application of a scale factor, and an artificial temperature factor to smooth the discontinuity that would otherwise occur at the protein-water interface. These structure factors can then be combined with the structure factors calculated from the model of the protein and the ordered water, to give the total calculated structure factors for the crystal [222, 223]. This is the method implemented in the program X-PLOR [224]. A similar model has been used to account for the bulk solvent scattering contribution in the context of neutron diffraction [225], and this model extended by dividing the solvent volume into shells which extend outward from the surface of the protein [226].

An alternative approach is based on the application of Babinet's principle (see [227, 228]), which relates the diffraction patterns produced by two complementary masks. Although this provides only a qualitative description of the effect of the disordered solvent, it can be expressed in a fashion which is computationally very convenient. This approach was adopted in the refinement of GFOR. For completeness, the theoretical background is given, since this is not made particularly clear in the literature.

Two masks are said to be complementary if the opaque regions in one mask are replaced by transparent regions in the other, and vice versa (i.e. the transparent and opaque regions are interchanged). In its most general form Babinet's principle states that the vector amplitude produced at a given point by one mask, when added to that produced by the complementary mask, gives the amplitude due to the unscreened wave. Hence we may write the vector equation

$$A_1 + A_2 = A_0 \qquad \textit{Eq. 2.5}$$

where the subscripts 1 and 2 refer to the complementary masks, and 0 to the absence of any mask. When applied to Fraunhofer diffraction (of which Bragg diffraction is a special case), this has an interesting consequence. Here the unscreened wave yields zero intensity everywhere, except at the image of the source itself. Hence $A_0 = 0$, and $A_1 = -A_2$. When these amplitudes are squared to obtain intensities, we have the result that the diffraction patterns due to complementary masks are identical (the diffracted radiation from the complementary masks at each given point has the same amplitude, but opposite phase). It should be noted that Babinet's principle is not perfectly rigorous but involves approximations [229].

At very low resolution the protein and bulk solvent regions of a protein crystal can be considered to be smooth and featureless. By application of Babinet's principle, the scattering due

to the bulk solvent has the same amplitude, but with opposite phase, as the scattering that would be produced by water if it occupied the molecular volume of the protein. So we can write

$$F_s = -K_s F_m \qquad \textit{Eq. 2.6}$$

where F_s and F_m are the structure factors due to the bulk solvent and the protein respectively, and K_s is a scale factor which accounts for the differing electron density levels in the protein and solvent regions. A correction must be applied to account for the static or dynamic disorder in the bulk solvent region, so Equation 2.6 can be written

$$F_s = -K_s F_m \exp\left(-B_s \frac{\sin^2\theta}{\lambda^2}\right) \qquad \textit{Eq. 2.7}$$

Now the total calculated structure factor scattering from the crystal is

$$\begin{aligned} F_c &= F_m + F_s \qquad \textit{Eq. 2.8} \\ &= F_m - K_s F_m \exp\left(-B_s \frac{\sin^2\theta}{\lambda^2}\right) \\ &= F_m\left(1 - K_s \exp\left(-B_s \frac{\sin^2\theta}{\lambda^2}\right)\right) \end{aligned}$$

This is the expression presented by Moews and Kretsinger [228]. Thus the problem of modelling the bulk solvent has been reduced to a scaling problem. The scaling function used in the program TNT [188] is derived from this; with the addition of a scale factor K to bring F_o onto an absolute scale, and an overall atomic displacement parameter B, F_o can be scaled to F_m using

$$F_o = \frac{1}{K}\exp\left(-B \frac{\sin^2\theta}{\lambda^2}\right) F_m\left(1 - K_s \exp\left(-B_s \frac{\sin^2\theta}{\lambda^2}\right)\right) \qquad \textit{Eq. 2.9}$$

The problem is then to determine appropriate values for the parameters of this function. In fact at moderate resolution, this problem is not trivial, and the parameters of this scaling function cannot be independently refined. Therefore the following procedure was adopted. All four scaling parameters (K, B, K_s, B_s) were systematically varied, and the conventional crystallographic R-factor between the observed and calculated structure factors evaluated for

all observed data. In practice it was found that a number of different values for these parameters gave equivalent agreement.

The data had already been put on an approximately absolute scale using a conventional Wilson plot [128]. From this procedure, the estimated mean atomic displacement parameter for the structure was 44 $Å^2$ (using diffraction terms between 4.5 and 2.7 Å resolution). Analysis based on the Patterson origin peak [230], gave results consistent with this. In the absence of any better procedure K_s and B_s were chosen so that agreement was obtained with the prior estimates of the scale factor (K), and overall mean atomic displacement parameter (which relates to B), based on Wilson statistics or analysis of the Patterson origin peak.

The effect of the correction can be seen in Table 2.5, where the structure factor amplitudes of several low angle reflections, both with and without the bulk solvent correction specified by Equation 2.8, are compared with the observed data.

Table 2.5 Structure factor amplitudes for typical low resolution reflections

Calculations were made using the final refined model for GFOR. Observed and calculated structure factor amplitudes were scaled according to Equation 2.9. Scaling parameters were {K= 1.068, B=0.0 $Å^2$, K_s= 0.0, B_s=0.0 $Å^2$} (no bulk solvent correction, R-factor for all data 0.351) or {K= 1.068, B=0.0 $Å^2$, K_s=0.88 B_s=140 $Å^2$} (bulk solvent correction, R-factor for all data 0.203). The standard deviations for the observed data derive from profile fitting only, and considerably underestimate the true variance (See [231, 232]).

h	k	l	Resn. (Å)	|F|(observed) (σ|F|(observed))	|F|(calculated) Solvent correction	|F|(calculated) No solvent correction
0	0	7	16.71	1897 (11)	1326	5929
0	1	5	23.32	158 (5)	436	2494
0	12	13	9.70	299 (3)	117	251
3	13	6	12.92	435 (3)	590	2060
6	0	15	6.82	581 (5)	463	792
7	2	19	5.48	700 (4)	851	1173
10	14	25	4.01	64 (28)	70	78

Evidently this model of the solvent structure is oversimplified, and there is experimental evidence to suggest that the solvent distribution around the protein surface may be non-uniform [233].

2.7.6.6 B-factor refinement

B-factor modelling was begun when the crystallographic R-factor was below 25%. At first two B-factors per residue were refined (for the side and main chain atoms respectively), then individual isotropic B-factors were employed in the final stages of refinement, using the restraints suggested by Tronrud [234]. Several side chains show evidence for more than one discrete conformation (notably Gln 256 and Met 314) but these have not yet been modelled. A number of side-chains, principally lysine and arginine residues on the surface of the protein are clearly disordered, and are not included in the model.

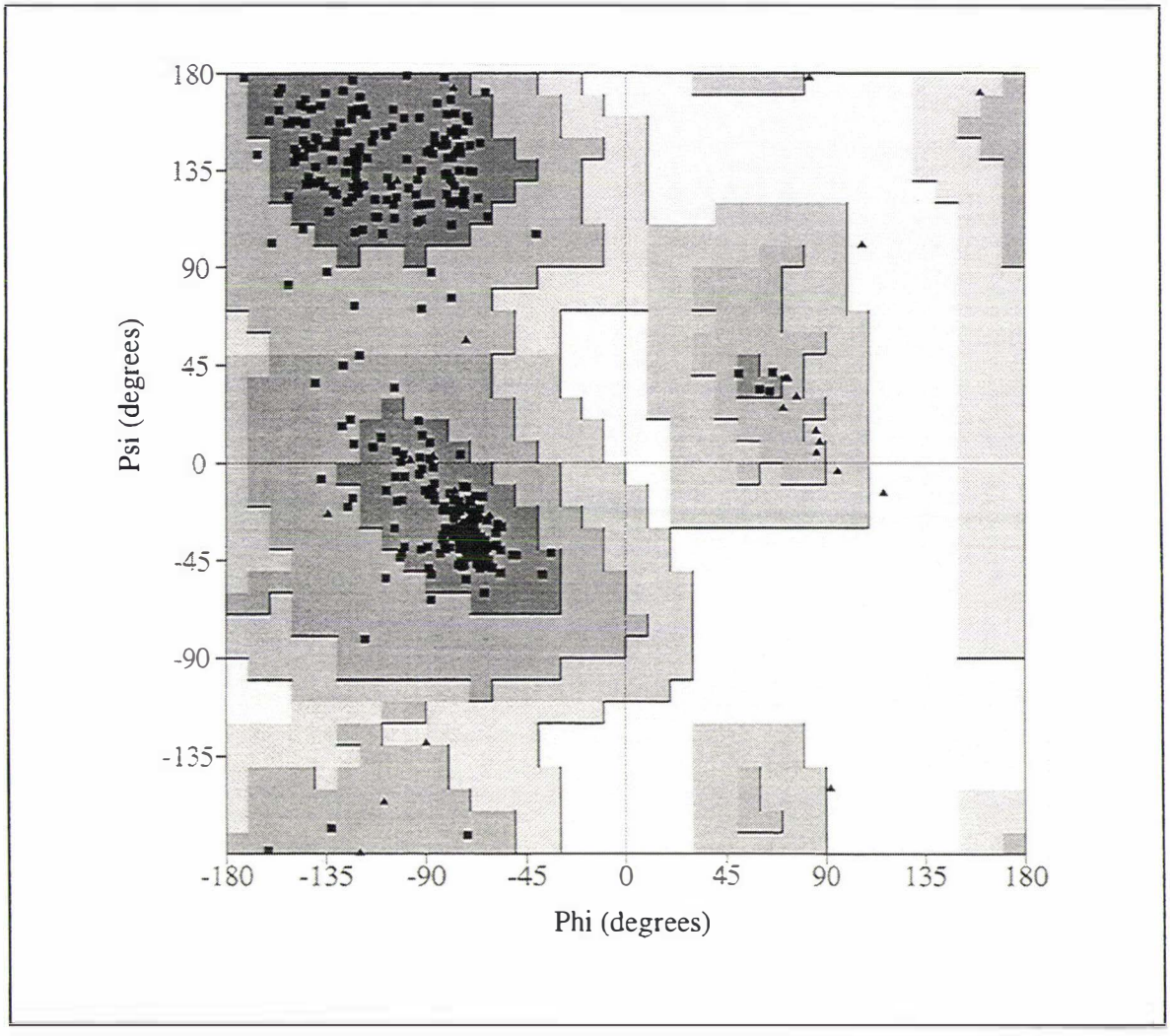


Figure 2.9 Ramachandran Plot for the refined GFOR monomer

Non-glycine residues are plotted as filled squares, glycine residues as filled triangles. Figure produced with the program PROCHECK [235].

2.7.6.7 Structure validation

The final model for the monomer contains a total of 3081 non-hydrogen atoms, which includes 141 water molecules, and 48 atoms for the NADP. The entire protein chain has been modelled. The current crystallographic R-factor for all data to 2.7 Å resolution is 20.2% (30% for terms between 2.75 and 2.70 Å resolution). The mean B-factor for all atoms is 44 $Å^2$. The final model is tightly restrained (the root mean square (rms) deviation from standard bond lengths 0.013 Å; angles 1.60°).

A general validation of the structure comes from its agreement with chemical data. The heavy atom compounds are found to have bound at chemically reasonable positions. The frameshift errors in the published sequence were predicted prior to the corrected sequence becoming available.

Perhaps the most useful geometric validation comes from examination of the backbone torsion angles, which were not restrained during refinement (however, information about these angles is incorporated during interactive rebuilding of the model). The Ramachandran plot for the monomer is shown in Figure 2.9. 91% of the residues are in the core regions of the Ramachandran plot, with no residues in the disallowed regions (as defined by the program PROCHECK [235]).

2.8 SUMMARY

Considering the structure determination of GFOR in retrospect, it is clear that the best course of action (had this been proved possible) would have been to collect a better native data set from the Form I crystals, and to obtain further isomorphous derivatives. This was unfortunately prevented by the occurrence of two closely related crystal forms, and the predominance of the second crystal form, for which derivatives were poorly isomorphous. It was however possible to solve the structure starting from a very poor initial electron density map, by using real-space density modification procedures and repeatedly incorporating information from the structural interpretation of the map. This process was quite laborious The general quality of the electron density map calculated from the Form II data is excellent (see Figure 2.10). This contrasts sharply with the map calculated from the Form I data, which using phases calculated from the final refined model, still shows discontinuities in the electron density of the polypeptide backbone in several places. This highlights the importance of well estimated data. The poor quality of the Form I data probably accounts for many of the difficulties encountered in the structure determination.

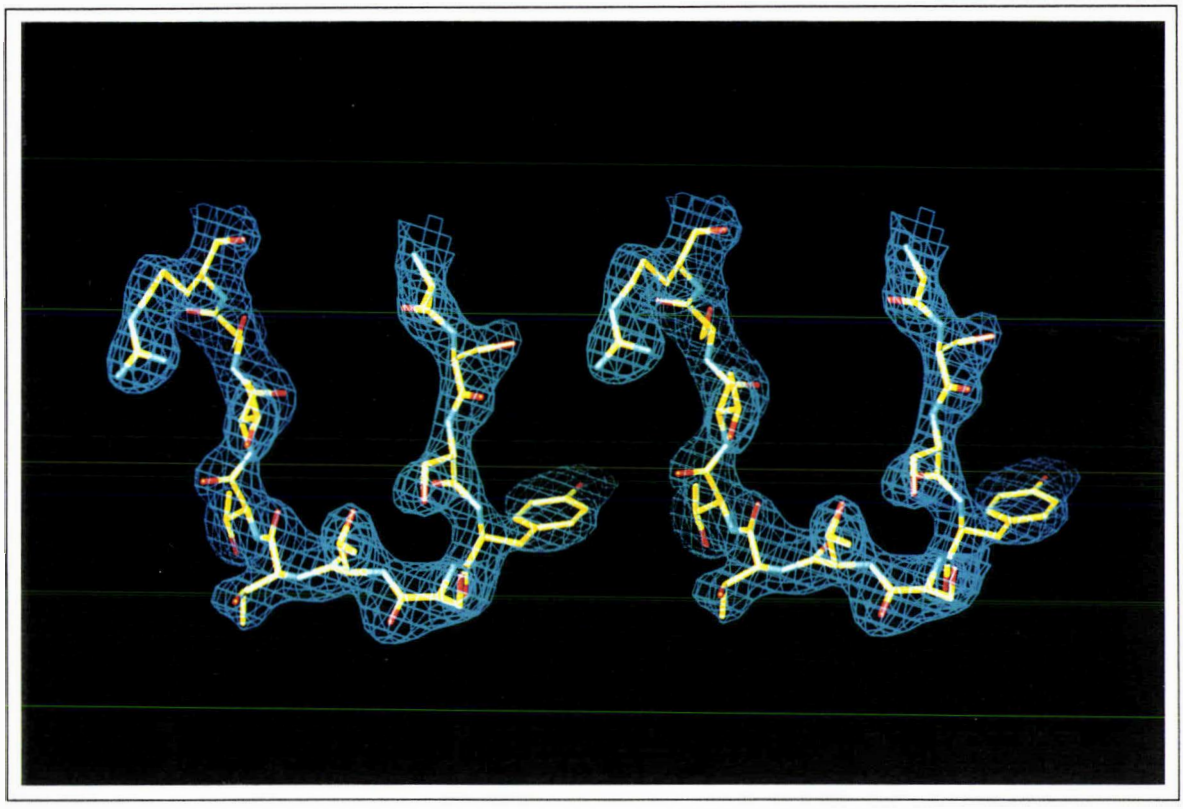


Figure 2.10 Electron density map calculated using the Form II data

Electron density corresponding to residues Ser 270 - Arg 279 which form a rarely observed ψ loop in the central β-sheet of the GFOR monomer. The map was calculated with all data to 2.7 Å resolution and contoured at 1.5σ. Fourier coefficients employed in the map calculation were of the form $(2m|F_o| - D|F_c|)$ (SIGMAA weighting) where $|F_o|$ is the native structure factor amplitude, $|F_c|$ is the calculated structure factor amplitude, and m and D have been defined by Read [207]. The figure was prepared using Turbo-Frodo [C. Cambillau, A. Roussel, A.G. Inisan and E. Koups-Mouthuy]

Chapter 3

GFOR: STRUCTURE AND FUNCTION

3.1 STRUCTURE OF THE MONOMER

3.1.1 N-terminal domain

In common with a number of other enzymes utilizing NAD(P) as a cofactor in oxidation-reduction reactions, the structure of GFOR is comprised of two well defined domains. These domains are not loosely associated as is sometimes seen (e.g as in dihydrodipicolinate reductase [236]), but are packed tightly together (Figure 3.1).

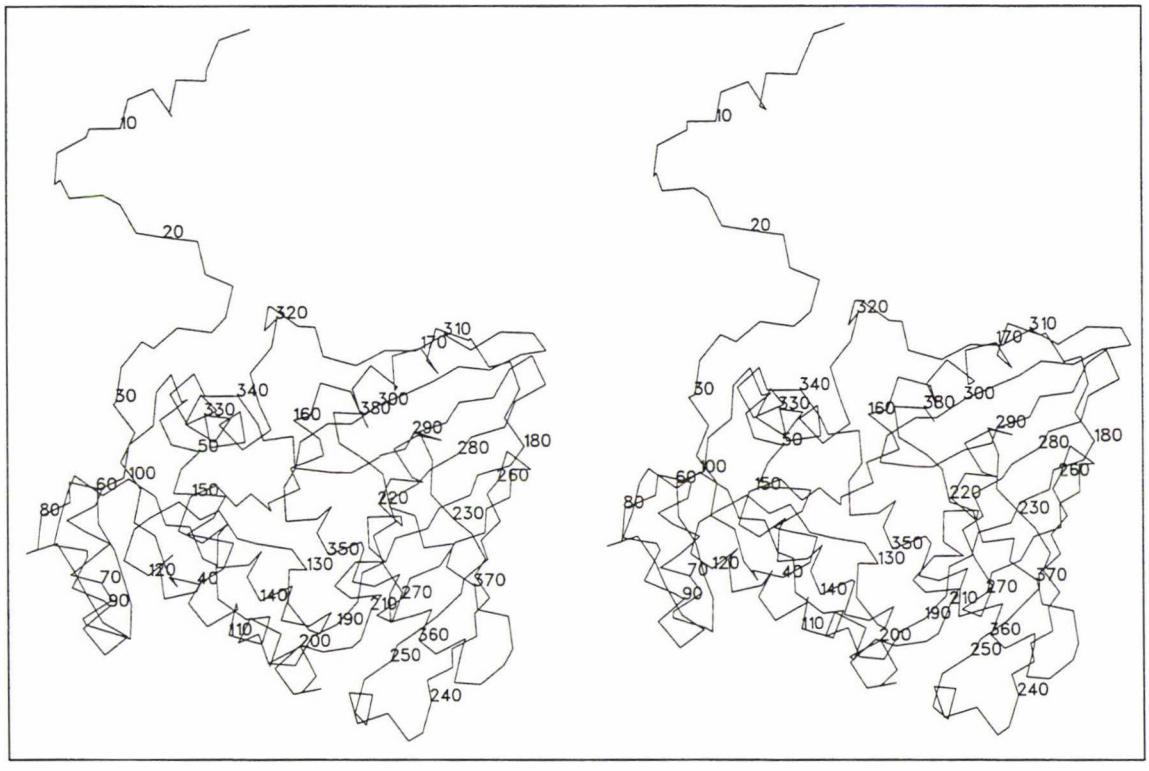


Figure 3.1 Cα plot of GFOR

Stereoview of the Cα plot of the GFOR monomer. Every tenth residue in the sequence is numbered.

The N-terminal domain has the classical dinucleotide binding fold containing the canonical 6 β-strands [237, 109, 238]. That is, it is comprised of two βαβαβ motifs which form a single parallel β-sheet, flanked on either side by alpha helices (Figure 3.2 (a)). The strands, labelled in the order they appear in the sequence, occur in the sheet as βF βE βD βA βB βC (i.e the sheet topology [239] is 1x 1x -3x -1x -1x). The crossover connection between strands βC and

βD, known to be among the least conserved structural elements of the fold [109], incorporates a relatively long and regular 3_{10} helix. The NADP, which is very clearly defined in the electron density map, is bound in conventional fashion with the pyrophosphate group located at the N-terminus of helix αa. Details of the NADP conformation, and the interactions between the protein and the dinucleotide are described later.

In comparison with other NAD(P) binding domains, the domain seen in GFOR is relatively small (comprising residues 32 - 154). In all classical dinucleotide binding domains, the strand βA terminates before the two adjacent strands in the sheet (βB and βD), creating a cleft where the adenine ribose is positioned [240]. GFOR is no exception to this, however the cleft is not at all marked, as the helices which pack on either side of the sheet do not extend much beyond the sheet boundary, and the connecting loops are in general short. Using the structure comparison algorithm of Holm and Sander [241], we compared the dinucleotide binding domain of GFOR with structures deposited in the protein data bank. The greatest degree of overall structural similarity was with the coenzyme A binding domain of succinyl-CoA synthetase [242] (rms difference for 107 equivalenced Cα positions 2.2 Å), and the NADP binding domain of dihydrodipicolinate reductase (DHPR) [236] (rms difference for 111 equivalenced Cα positions 2.2 Å).

3.1.2 C-terminal Domain

The C-terminal domain is based around a mixed β-sheet, the strands of which are linked by a number of helices and surface loops (Figure 3.2 (b)). In this domain, the sheet topology [239] is (3x, -1, -1, -2, -1, -1, -1, 8). The central β-sheet has a very pronounced right-handed twist. The domain is 'open-faced' [239] in that the helices and loops cover only one side of the β-sheet. In this sense GFOR resembles members of the glyceraldehyde-3-phosphate dehydrogenase (GAPDH) family [245], and also DHPR [236]. The sheet is entirely antiparallel, with the exception of the first strand in the sequence (βG) which is found at the center of the sheet and is involved in a ψ loop formed between the antiparallel strands βJ and βK. Psi loops are rarely observed in protein structures, and are characterized by two sequentially adjacent antiparallel strands in a β-sheet, connected by a '+2' hairpin turn (i.e. with one strand in-between, hydrogen-bonded to both of them) [244]. There are four possibilities for the topology of these loops. Interestingly, despite differing overall sheet topology, psi loops of the same kind are found in GAPDH and DHPR (Figure 3.2 (c)). In these oxidoreductases, which are both tetrameric, the open-faced sheet of the C-terminal domain is involved in the formation of a subunit interface. This is also the case in GFOR.

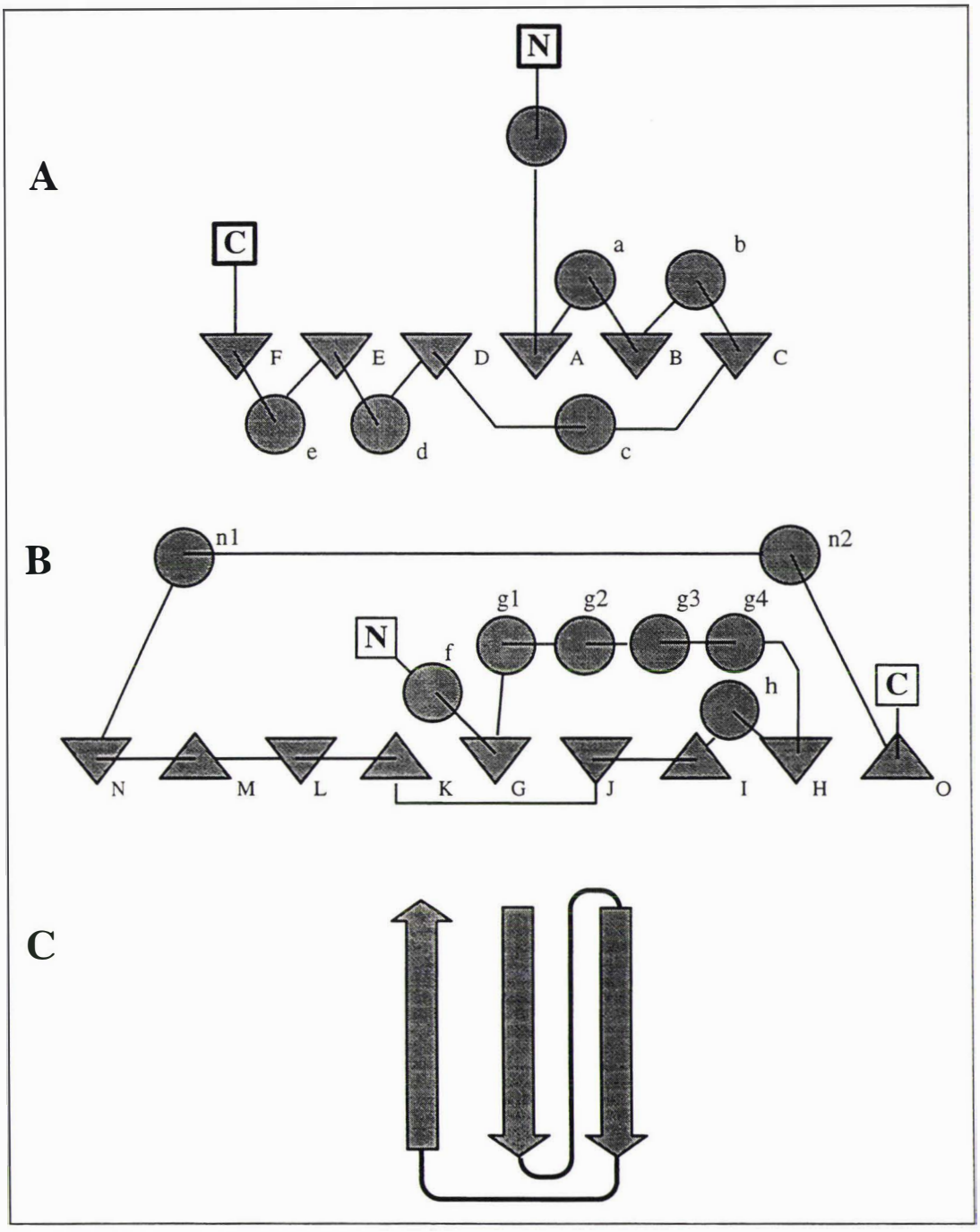


Figure 3.2 Topology of GFOR

(a) A topology diagram of the N-terminal domain. Definition of the principal secondary structural elements is as follows. Helices **a** 41-50; **b** 67-77; **c** 90-95; **d** 108-120; **e** 135-148. β-Strands **A** 32-37; **B** 57-63; **C** 84-85; **D** 101-104; **E** 124-127; **F** 152-154. Secondary structural elements were defined using PROMOTIF [243].

(b) A topology diagram of the C-terminal domain. Definition of the principal secondary structural elements is as follows. Helices **f** 162-172; **g1** 195-198; **g2** 203-206; **g3** 210-213; **g4** 215-226; **h** 244-246; **n1** 324-337; **n2** 346-365. β-Strands **G** 181-187; **H** 230-238; **I** 252-

259; **J** 264-270; **K** 276-282; **L** 287-290; **M** 301-305; **N** 308-312; **O** 369-370. Secondary structural elements were defined using PROMOTIF [243].

(c) The topology of the psi-loop common to the C-terminal domains of GFOR, G6PD, GAPDH, and DHPR. The crossover connection between the two right-most strands varies in complexity, and incorporates other strands of the sheet in all but DHPR. The loop is Type 1X´ according to the nomenclature of Hutchinson and Thornton [244]

3.1.3 N-terminal arm

Preceding the dinucleotide binding domain there is an extended N-terminal 'arm', which is found wrapped around an adjacent subunit in the tetramer. The N-terminal sequence contains a high number of proline residues (7 in the first 31 amino acids), which is reflected in its extended conformation in the structure. The only element of regular secondary structure in this region is a short α-helix (residues 5-8). Interestingly, an N-terminal arm (though not proline rich) precedes the dinucleotide binding domain in some tetrameric lactate dehydrogenases (LDH) (e.g. dogfish muscle LDH [246]). Here it is responsible for stabilization of the quaternary structure, as proteolytic cleavage of this region results in the formation of stable dimers [247]. Correspondingly, members of the closely related malate dehydrogenase (MDH) family lack the N-terminal arm and are typically dimeric [238]. Note however, that the N-terminal arm is not required for stabilization of the quaternary structure in all species, as many bacterial LDHs are tetrameric yet lack this feature [248].

3.2 COMPARISON WITH GLUCOSE 6-PHOSPHATE DEHYDROGENASE

3.2.1 Structure comparison

For GFOR, the most striking structural similarity is not with GAPDH or DHPR, mentioned above, but with the recently determined structure of glucose-6-phosphate dehydrogenase (G6PD) from *Leuconostoc mesenteroides* [249] (Figure 3.3). The sheet topology of the C-terminal domains of these two structures is almost identical (the only difference being a reversal in the direction of the last short β-strand of the sheet). This relationship had not been anticipated as there is no detectable sequence homology between these two proteins. Significantly, G6PD catalyzes a reaction essentially equivalent to one of the half-reactions of GFOR (the oxidation of glucose to gluconolactone), differing only in the requirement that the substrate is phosphorylated.

While GFOR and G6PD are virtually identical in a topological sense, G6PD is a substantially larger protein. Corresponding elements of secondary structure often differ in both length and relative orientation, and the connecting loops between them are often elaborated in G6PD (Fig. 3). For example in the C-terminal domain of GFOR, the last four strands of the central

β-sheet (βK - βN) are markedly shorter than their counterparts in G6PD. There are also additional structural elements present in G6PD, but not in GFOR. In the connection between strands βG and βH, there is a large extension in G6PD, which contains several helices which have no counterpart in GFOR.

Figure 3.3 Ribbon diagrams of GFOR and G6PD

Ribbon diagrams of the monomers of (top) GFOR and (bottom) G6PD. For both proteins the N-terminal domain is in blue and the C-terminal domain in purple, with connecting loops in yellow. The figure was produced with Ribbons 2.0 [250].

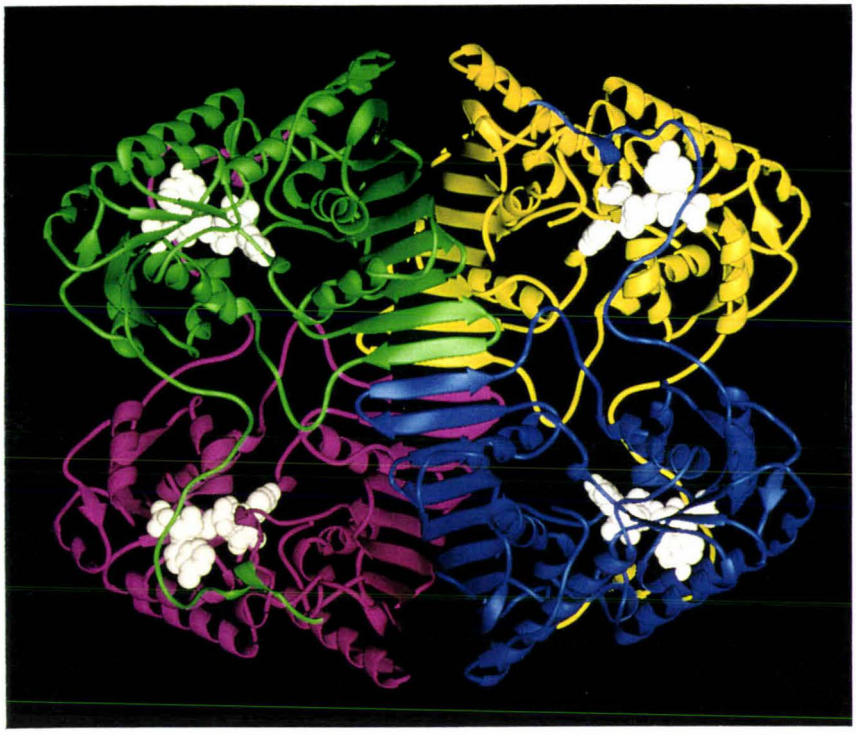
Another difference occurs in the C-terminal region where an additional helix occurs in G6PD, associated with the otherwise open-faced side of the β-sheet. This is of significance for the quaternary structure of both enzymes, as the subunits in tetrameric GFOR associate through an aligned face to face packing of the central β-sheets.

As noted above the structure of the dinucleotide binding domain of GFOR, as a whole, resembles domains in other oxidoreductases (succinyl-CoA synthetase, DHPR) more closely than it does G6PD. In particular the βB-βC, βC-βD and βD-βE connections all differ substantially between GFOR and G6PD (Figure 3.3). There are however several intriguing similarities between the domains. In GFOR the characteristic hydrogen bonding pattern of the first helix of the domain (αa) is disrupted by the incorporation of a proline (Pro 49) within the helix. In G6PD, a proline is observed at an identical position, and is conserved between all known G6PD sequences. The effect in both proteins is to distort the helix so that an extra residue is accommodated in the second turn. Another striking similarity is in the loop following strand βE, which is found adjacent to the nicotinamide ring. Residues 127 - 129 in GFOR (Glu Lys Pro) correspond to residues 147 -149 in G6PD (again Glu Lys Pro), and are found in a very similar conformation in both proteins.

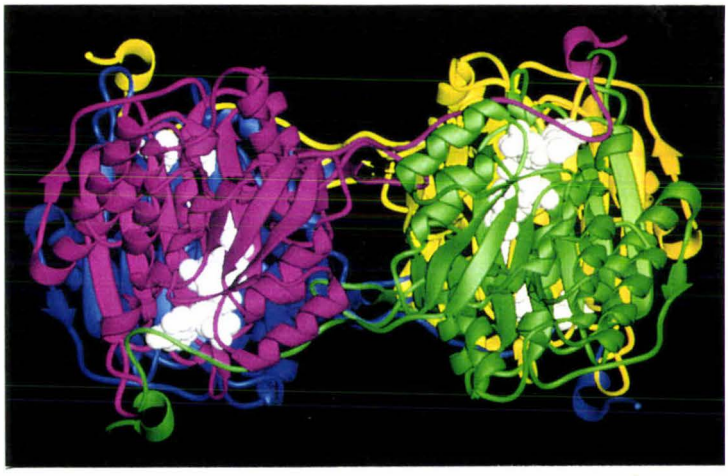
3.2.2 Evolutionary implications

Overall the topological equivalence of the C-terminal domains of GFOR and G6PD, the conservation of several key features in the dinucleotide binding domains, and the similar functions of the two enzymes amount to a persuasive argument for a common evolutionary origin. Given that GFOR has not been positively identified in other organisms, and is apparently responsible for the tolerance of *Z. mobilis* to the high sugar concentrations found in its natural growth media, the idea that it has been 'recruited' from a cytoplasmic enzyme involved in glucose metabolism has considerable attraction. G6PD is a major metabolic enzyme in *Z. mobilis,* utilizing NAD in mainstream catabolism of sugars, but also capable of using NADP to generate anabolic reducing equivalents. In this respect it is functionally similar to the enzyme from *L. mesenteroides*, with which it has clear sequence homology [251] (sequence identity ~ 33% using standard pairwise alignment procedures). Interestingly, the *Z. mobilis* enzyme is tetrameric [116], in contrast to the dimeric *L. mesenteroides* G6PD. We propose that GFOR and G6PD have both evolved from a common 'glucose-oxidizing' ancestral gene, but note that both are structurally distinct from the glucose dehydrogenase from the archaeon *Thermoplasma acidophilum* [252].

A



B



C

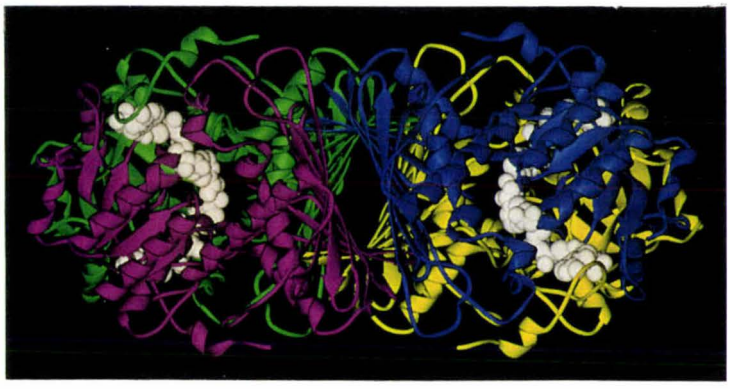


Figure 3.4 (Preceding page) Quaternary structure of GFOR

A ribbon diagram of the GFOR tetramer, with a space-filling representation of the bound NADP in white. Three orthogonal views (A, B, C) of the tetramer are shown, looking down the molecular two-fold axes. Each subunit is in a different colour. The figure was produced with Ribbons 2.0 [250].

3.3 STRUCTURE OF THE TETRAMER

The GFOR tetramer possesses almost perfect *222* point group symmetry (Figure 3.4). Measured along the internal symmetry axes of the molecule, the tetramer extends approximately 85 x 100 x 43 Å, thus having a slightly flattened overall appearance. The principal inter-subunit contacts involve the central β-sheet of each C-terminal domain. In two of the subunits, the contacts are largely between the last strands in each domain, which are hydrogen bonded to each other in a typical antiparallel fashion. Consequently a continuous 18-stranded β-sheet is formed. As a result of a marked right-handed twist of the β-sheet of each subunit, the sheet resulting from the interaction between the subunits turns by almost 180 degrees over the length of the molecule. In the tetramer, the extended sheets formed by two such pairs of subunits stack against one another, forming an extensive interface. Hence tetrameric GFOR contains both the stacked and extended β-interfaces described by Jones and Thornton [182] in their analysis of protein dimers. A very similar association of subunits is found in the all β protein concanavalin A [253], and in DHPR [236].

The stacked β-sheets in GFOR are oriented so that the strand direction in one sheet is at an angle of about 30° to the strand direction in the other (in correspondence with arguments based on packing considerations [254, 255]). However, the two sheets do not pack tightly along the entire length of the interface, but spread apart in the middle creating a cavity at the centre of the tetramer. Electron density maps indicate the presence of a number of ordered water molecules in this region, hence the cavity is solvent filled.

The other subunit-subunit interaction in GFOR is mediated by the N-terminal arm. In Figure 3.4 (a) it can be seen that the association between the two left-most (or two right-most) subunits is principally due to contacts between the NADP-binding domain of each subunit and the N-terminal arm from the other subunit of the pair. This is illustrated by calculation of the molecular surface area buried by subunit association (employing an analytic surface calculation method [256]). When the two leftmost subunits in Figure 3.4 (a) are associated, the total buried surface area is 6780 $Å^2$. Repeating the calculation with truncation

of the N-terminal arm (residues 1-31), the total buried surface area is 1510 $Å^2$, or 22% of the original value. Hence almost all the contact between these two subunits involves the N-terminal arm.

When the structure of GFOR was first determined, the proximity of the N-terminal arm to the NADP binding site, and the implications of this for the tight association of NADP (discussed below), was immediately striking. However, in light of the structural studies on LDH and MDH, where the N-terminal arm has been shown to be important for subunit association, it may be implicated in such events as well. Further experiments will be required to determine the exact role of the N-terminal arm in cofactor binding and oligomerization.

3.4 DINUCLEOTIDE BINDING

3.4.1 NADP Conformation

The crystals of GFOR used in the structure determination were grown in the presence of 800 mM sorbitol, a product of one of the reactions catalyzed by the enzyme. Based on the equilibrium constant for the fructose/sorbitol half reaction [102] it is expected that the cofactor is in its reduced state. The conformation of the bound NADP together with the corresponding electron density is displayed in Figure 3.5. In the structure the adenine ring is found in a *syn* conformation with respect to the glycosidic bond, while the nicotinamide ring is found in an *anti* conformation. The pucker of the sugar groups is C(3´) - endo for the adenine ribose and C(2´) - endo for the nicotinamide ribose [257]. The oxygen atoms of the two phosphate groups in the pyrophosphate bridge are perfectly staggered. The *syn* conformation of the adenine ring, with the bulky adenine group positioned 'above' the ribose sugar, is unusual. However it is not without parallel in protein-nucleotide complexes (see e.g. [258, 259]). As a consequence of its *syn* conformation, the adenine ring points away from the dinucleotide binding domain and interacts with the N-terminal arm of an adjacent subunit. In fact the only hydrogen-bonding interactions by atoms of the adenine ring involve main chain carbonyl oxygen atoms from the N-terminal arm (residues Pro 11, Thr 13 and Ala 15). These interactions must help stabilize the energetically disfavoured *syn* conformation [257].

3.4.2 Interactions with GFOR

The hydrogen-bonding interactions between NADP and the protein are shown schematically in Figure 3.6. In GFOR there are 21 potential hydrogen bonds between NADP and protein atoms, and at least a further six hydrogen-bonding interactions with ordered water molecules (these water molecules were well defined in difference Fourier syntheses and have B-factors comparable to those of the protein atoms). The hydrogen-bonding interactions between the

protein and the 2′-phosphate group are particularly extensive. In all the NADP-specific enzymes whose structures have been solved to date, the 2′-phosphate group interacts with one or more basic amino acids, which provide a favorable electrostatic interaction with the negative charge(s) carried by the 2′-phosphate group. In GFOR Lys 69 fulfils this role. In a sense the question of specificity for NAD or NADP is irrelevant for GFOR, because the bound cofactor is never released from the enzyme. However, specific association of GFOR with NADP is consistent with the presence of a positively charged binding pocket for the 2′-phosphate group. In comparison with some other NADP-dependent dehydrogenases, the number of direct hydrogen-bonding interactions between the dinucleotide and the protein is quite large. For example in the complex of NADP with G6PD, there are only 9 potential hydrogen bonds to protein atoms [M.J. Adams, personal communication]; and in glutathione reductase there are 11 [260]. One other interesting interaction is the stacking of one face of the nicotinamide ring against Tyr 42 (see Figure 3.7). Similar stacking interactions have now been seen in a number of enzymes (e.g. aldose reductase [261] and glutathione reductase [260]).

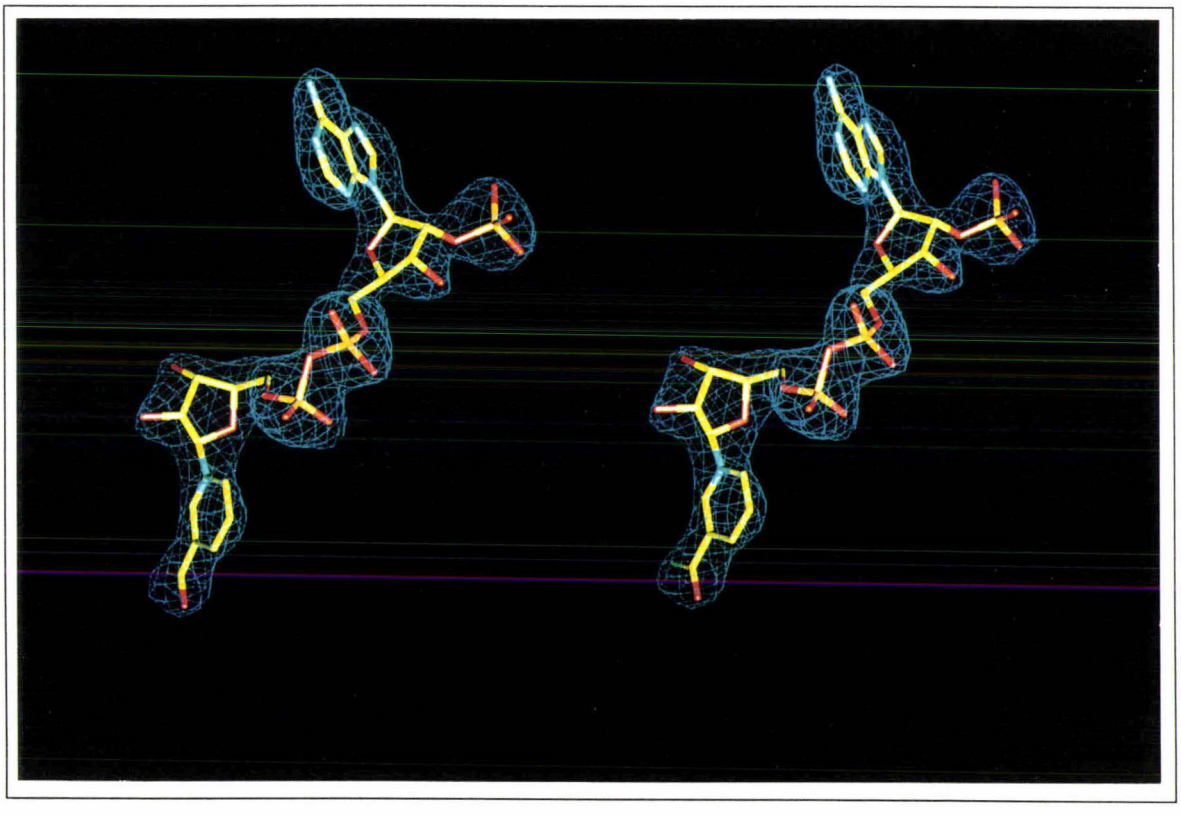


Figure 3.5 Conformation of the enzyme-bound NADP

Stereoview of the electron density corresponding to NADP. The map was calculated with all data to 2.7 Å resolution and contoured at 1.5 σ. Fourier coefficients employed in the map calculation were of the form $(2m|F_o| - D|F_c|)$ (SIGMAA weighting) where $|F_o|$ is the native structure factor amplitude, $|F_c|$ is the calculated structure factor amplitude, and m and D have been defined by Read [207]. The figure was prepared using Turbo-Frodo [C. Cambillau, A. Roussel, A.G. Inisan and E. Koups-Mouthuy]

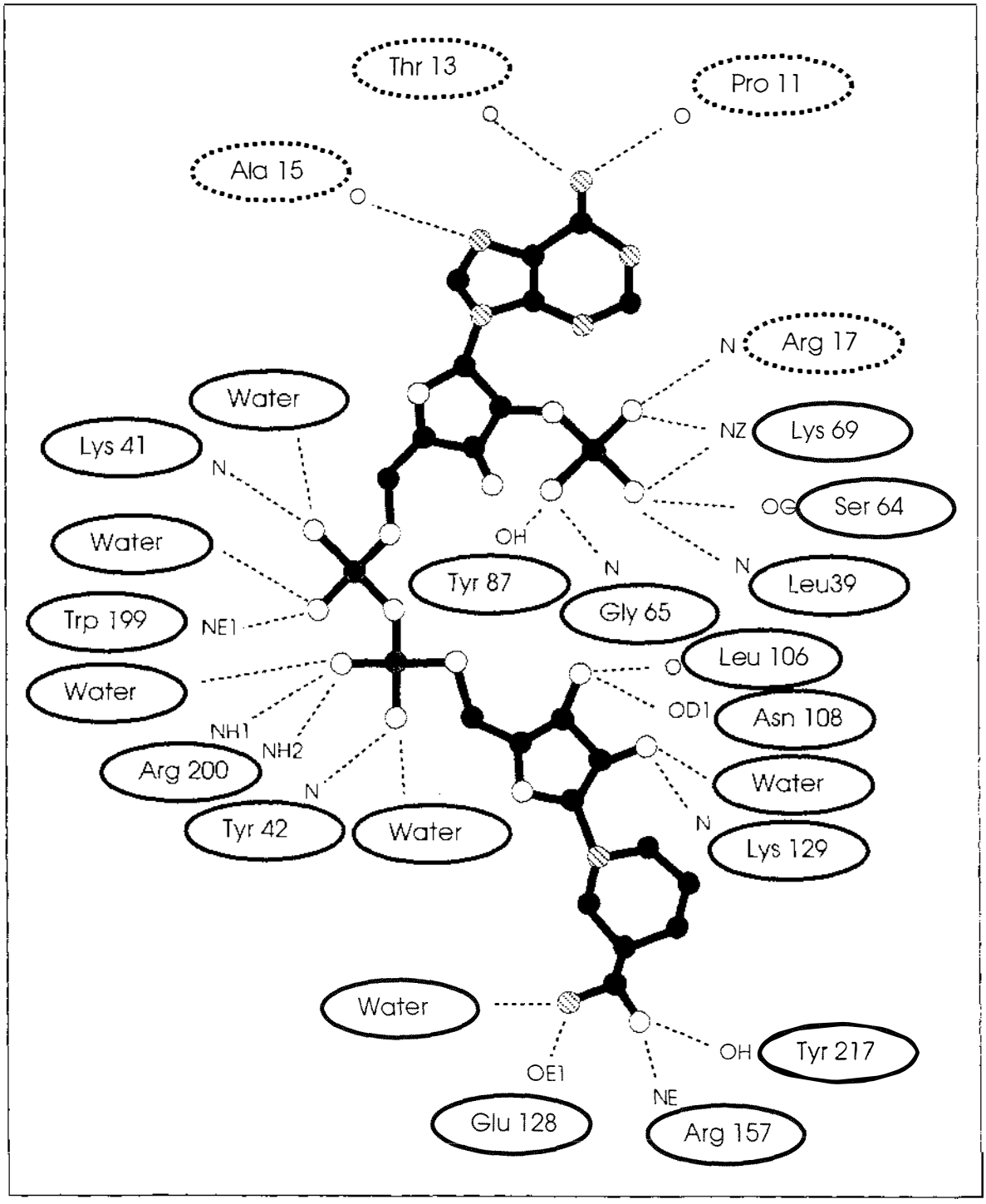


Figure 3.6 Hydrogen-bonding interactions between GFOR and NADP

Schematic representation of the hydrogen bonding between NADP and GFOR. Dashed lines around residue names indicate that the residues are from the N-terminal arm of an adjacent subunit in the tetramer. Distances of potential hydrogen bonds indicated in the diagram range from 2.7 - 3.1 Å.

3.4.3 Tight association with GFOR

The bound NADP is almost entirely buried in the interior of the protein, with 97% of its molecular surface area buried in the complex. Only atoms of the nicotinamide ring (C4 and C5) are at all exposed. This effective burial of the NADP is in large part due to interactions with a short helix from the second domain of the protein (αg1; residues 195 -200) and also with the N-terminal arm of an adjacent subunit in the tetramer, which covers the adenine ring. Extensive burial of NAD(P) in the protein interior has been seen in a number of oxidoreductases; enzymes which must release NAD(P) during the catalytic cycle accomplish this by rigid body domain motions (e.g. horse liver alcohol dehydrogenase) or by more local conformational changes such as loop movements (e.g. lactate dehydrogenase) [240].

It is clear that NADP cannot be released from GFOR without a concerted displacement of both the N-terminal arm and the loop containing helix αg1. Both of these regions are well defined in electron density maps, and the atomic displacement parameters of the constituent atoms are comparable to those in the dinucleotide binding domain itself. Hence there is no indication that these regions are inherently more mobile than contiguous parts of the structure. In addition, the high number of proline residues in the N-terminal arm places a number of conformational constraints on the polypeptide backbone. These findings are all consistent with the non-dissociable nature of NADP in GFOR.

These observations are mirrored in the structure of UDP-galactose 4-epimerase, which contains tightly associated NAD [115]. In this protein, 96% of the molecular surface area of the bound NAD is buried, and again almost all of the exposed surface area is associated with atoms of the nicotinamide ring. As with GFOR, the protein is involved in a large number of hydrogen-bonding interactions with the bound dinucleotide. However the way in which the tight association of NAD(P) is achieved differs in an important respect. In UDP-galactose 4-epimerase, residues which interact with NAD are almost exclusively within the dinucleotide binding domain, which is much larger than its counterpart in GFOR (comprising the first 180 residues of the protein). Strands βB and βD and the loops that follow form a marked cleft which encloses the adenine ribose and the adenine ring. In contrast, in GFOR the cleft is barely noticeable, as the strands βB and βD do not extend much further than the central strand βA. Instead, structural elements external to the dinucleotide binding domain itself seem to be important for the tight association of NADP, consistent with the idea that GFOR has evolved from a cytoplasmic precursor which would release NADP during its catalytic cycle. Consequently, it may be that the way in which the tight association of NADP has been achieved in GFOR reflects the evolutionary history of the enzyme rather than any structural

necessity.

3.4.4 Evolutionary implications of the N-terminal arm

The way in which GFOR prevents dissociation of NADP, employing several structural elements external to the dinucleotide binding domain, seems economical from an evolutionary perspective. The classical dinucleotide binding fold is conserved, and the association of the cofactor with this domain appears fundamentally similar to that reported for many other NAD(P) dependent oxidoreductases. The role of the N-terminal arm is particularly interesting. In other proteins, the presence of N or C-terminal extensions to a core domain has been implicated in protein-protein or protein-membrane association. In addition to LDH and MDH (discussed previously), some other examples include a N-terminal arm which precedes the $(\beta/\alpha)_8$ TIM barrel domain in *Propionibacterium shermanii* methylmalonyl-CoA mutase, and is involved in subunit interactions [262]; a N-terminal arm which is involved in the association of the β-crystallins of the eye lens [263, 264]; a C-terminal arm which participates in subunit interactions in the decameric muconolactone isomerase from *Pseudomonas putida* [265]; a C-terminal sequence implicated in aggregation of spinach chloroplast GAPDH; and a short N-terminal sequence preceding the catalytic domain of a cyclic AMP phosphodiesterase which confers membrane association on an essentially soluble protein [266]. These examples, toghether with the apparent evolutionary relationship between GFOR and 6PGD, lead us to speculate that the addition of N or C-terminal extensions to pre-existing structural domains may be a general evolutionary mechanism for controlling domain association and other binding events, and regulating protein function.

3.5 IMPLICATIONS FOR CATALYSIS

3.5.1 Background

The reactions catalyzed by GFOR involve a carbonyl/alcohol interconversion, coupled with the oxidation and reduction of NADP. Such reactions are known to proceed by hydride transfer to or from the C4 carbon of the nicotinamide ring, catalyzed by the polarization of the reacting group in the substrate [267]. In the absence of a metal ion, the polarization can be achieved through an amino acid which acts as an acid-base catalyst, by hydrogen-bonding with the carbonyl or alcohol group of the substrate. For example, in lactate and malate dehydrogenase, a histidine residue acts as a general acid/base catalyst [268]. Reduction of the carbonyl group of the substrate involves the concerted donation of a proton (H^+) from the active site histidine to the carbonyl oxygen atom, and of a hydride ion (H^-) from the nicotinamide ring of NADPH to the carbonyl carbon atom. In the reverse reaction, the histidine accepts a

proton from the hydroxyl group, coupled with hydride transfer from the carbonyl carbon to the nicotinamide ring.

3.5.2 The active site of GFOR

Identification of catalytically important residues in GFOR is complicated by the fact that the overall reaction catalyzed by the enzyme is comprised of two half reactions (the oxidation of glucose to gluconolactone, and the reduction of fructose to sorbitol), and the question of whether the two substrates bind to the enzyme in an analogous fashion. An anomeric specificity of GFOR for β-D-glucose has been demonstrated [102] which implies that this sugar binds in its pyranose ring form, but it is not clear if fructose binds in a ring or open-chain form. If it binds in a ring form then the reduction must be accompanied by ring opening, since sorbitol is an acyclic molecule. It is not necessarily the case that the two substrates bind in the same way, or share the same proton donor/acceptor.

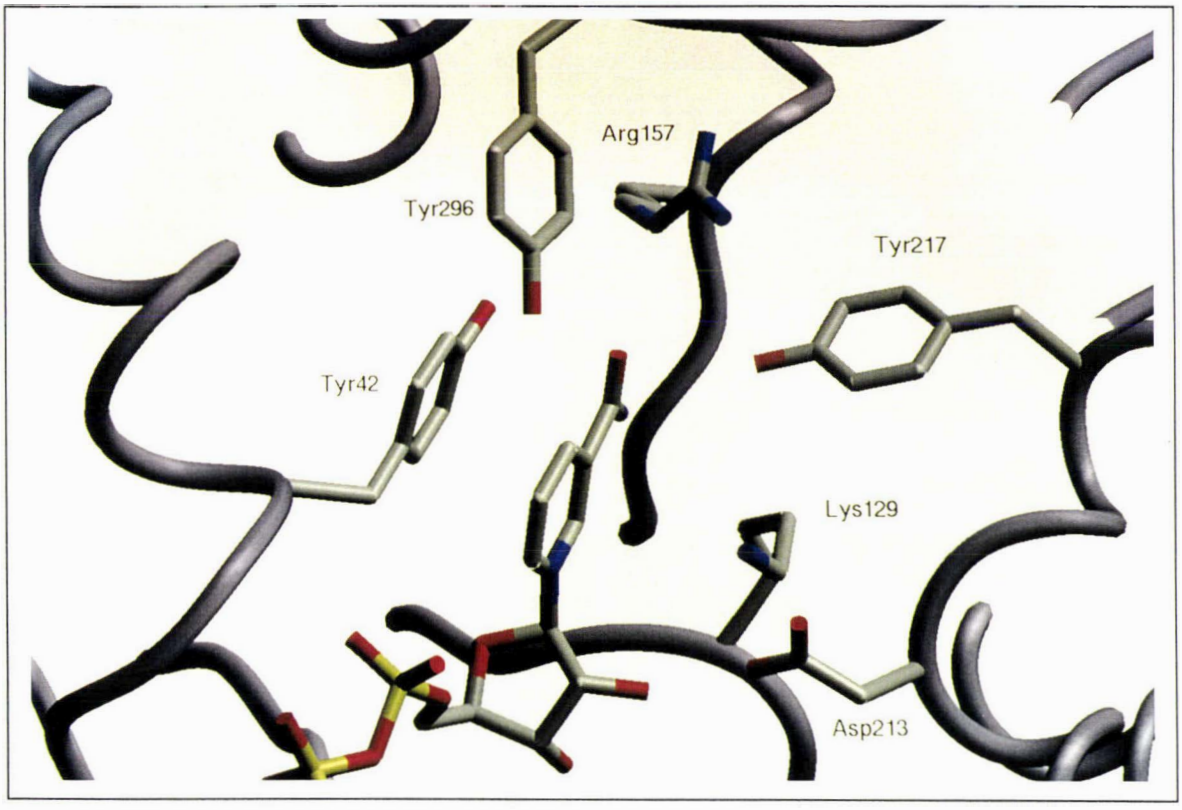


Figure 3.7 The active site of GFOR

Figure prepared with SETOR [272].

In GFOR the orientation of the nicotinamide ring and the position of the C4 carbon are unequivocal. The solvent-exposed atoms of the nicotinamide ring are found in a deep cavity formed at the interface between the two domains. There are a number of potential proton donors/acceptors at reasonable distances from the C4 carbon (Figure 3.7). These include two tyrosine residues, Tyr 217 and Tyr 296, together with Lys 129 and Asp 213. There are no histidine residues adjacent to the nicotinamide ring. In both the aldo-keto reductase and short-chain dehydrogenase/reductase protein families, a tyrosine residue has been implicated as the acid-base catalyst [269, 270, 271], in each case with a hydrogen-bonded lysine which is proposed to depress the pK_a, thereby facilitating proton transfer. Thus it seems possible that either Tyr 217 or Tyr 296 could act as a proton/donor acceptor in GFOR. Of these, Tyr 217 seems the more likely. It is hydrogen-bonded to the carboxamide group of the nicotinamide ring; it is adjacent (but not hydrogen bonded) to Lys 129 (discussed below); and this tyrosine and the proposed acid-base catalyst in G6PD (His 240) [249] originate from a topologically equivalent helix in the two structures (αg4 in GFOR), and indeed occupy a roughly equivalent position.

The lysine residue, Lys 129, is intriguing. Its sidechain has an unusual rotamer conformation {g+,g+} which is rarely observed in protein structures [273]. The side chain electron density is weak, and the side chain atoms have relatively high B-factors, yet we are confident that it is correctly modelled. It was not included in the model until late in the refinement, when Fourier difference maps could be unambiguously interpreted. In the present structure the amino group of the side chain is adjacent to the face of the nicotinamide ring (~ 3.8 Å distant) and makes a hydrogen bond to the carbonyl oxygen of Asp 213. It is suggested that this conformation may well be influenced by the oxidation state of the cofactor; thus it is possible that Lys 129 might rearrange during catalysis and participate in the fashion that has been suggested for the aldo-keto reductase and short-chain dehydrogenase/reductase protein families (by hydrogen bonding to the adjacent Tyr 217 and facilitating proton transfer).

3.5.3 Sequence and structural similarities

Regardless of its exact role, several lines of evidence point to the potential importance of Lys 129. The first is that it is involved in a structural motif which is conserved between GFOR and G6PD. Residues 128 - 130 in GFOR (Glu Lys Pro) correspond to residues 147 -149 in G6PD (also Glu Lys Pro). This is striking in view of the almost complete lack of sequence identity between the two proteins. These 3 residues occur at the end of strand βE in the dinucleotide binding domain. In GFOR, Glu 128 is hydrogen bonded to the carboxamide group of the nicotinamide ring and Pro 130 has a *cis* peptide bond and is integral to the turn at the end of strand βE. In G6PD the situation is a little more complicated. There are two indepen-

dent subunits in the asymmetric unit of the crystals of the holo-enzyme. In the first of these the conserved proline residue has a *cis* peptide and the overall conformation is very similar to that found in GFOR. In the second subunit the proline has a *trans* peptide bond, and the conformation of the preceding lysine also differs [249]. However, in a complex of G6PD with NADP, both subunits of the dimer contain Pro 149 in a *cis* conformation [M.J. Adams, personal communication]. This raises the possibility that there may be two conformational states for the motif which interconvert, as there are examples of *cis/trans* proline isomerization occurring in folded proteins [274, 275].

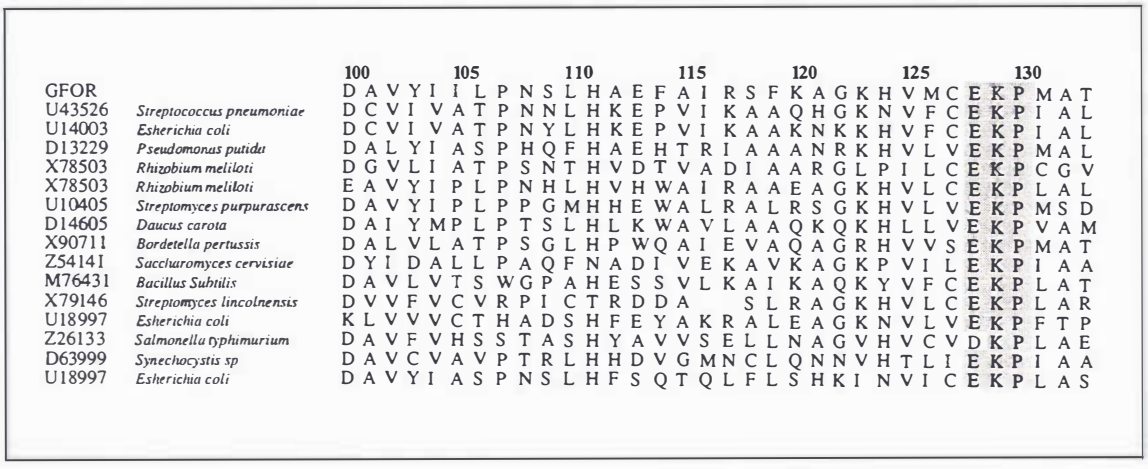


Figure 3.8 Alignment of sequences with homology to GFOR.

Multiple sequence alignment of deduced gene products having sequence homology to the dinucleotide binding domain of GFOR. Genbank accession numbers and the source organism are reported. The conserved motif EKP is highlighted (numbering corresponds to the sequence of GFOR). The figure was produced with ALSCRIPT [281].

A search of the NCBI non-redundant protein sequence database (May 1996) using the BLASTP algorithm [276] revealed 15 sequences with clear homology to the dinucleotide binding domain of GFOR. Unfortunately only two of these deduced gene products have an assigned function, an inositol 2-dehydrogenase ([277], Genbank accession number M76431) and a dihydro-4,5-dihydroxyphthalate dehydrogenase ([278], Genbank accession number D13229). Multiple sequence alignment [279] revealed that the motif EKP was almost completely conserved in all of the sequences (Figure 3.8). The only other completely conserved residue in the alignment is the first glycine in the dinucleotide binding loop motif GXGXXG. Significantly this seems to be the only absolutely required glycine in the motif, the other two glycine residues representing preferences not requirements [280]. This supports the idea that these sequences represent dinucleotide binding domains in which the conserved motif EKP has some well defined structural or functional role. The similarity of the sequences following the N-terminal domain is less marked. However one interesting feature is that the amino acid

corresponding to Tyr 217 in GFOR is a histidine or a tyrosine in all of the sequences, strengthening the view that this residue plays a role in catalysis.

3.5.4 General discussion

The early work on NAD(P) dependent oxidoreductases highlighted a number of common structural features [282]. The enzymes comprised two distinct domains, a dinucleotide binding domain which had a similar structure in all the enzymes studied, and a catalytic domain, which differed markedly between the proteins. The first domain carried most of the residues critical to dinucleotide binding, while the second bound the substrate and facilitated catalysis. More recently however, a number of structures have been determined in which only a single domain is present [271]. While these enzymes still possess the canonical dinucleotide binding fold, connecting loops have been elaborated in order to facilitate substrate binding and catalysis within a single domain [109]. Thus the original distinction between the catalytic and dinucleotide binding domains has become blurred, and it seems entirely possible that a dinucleotide binding domain in an enzyme with several domains might contribute residues essential for catalysis.

Concerning catalysis, an interesting discovery was made in a recent biotechnological application of the GFOR. The enzyme, which is stable in the presence or absence of any of its substrates or products, is irreversibly inactivated during the time course of its own catalytic action [283] (this inactivation is probably not of biological significance, since it occurs over a time period greater than the life-time of the bacterial cell). The inactivation was shown to be linked to the oxidation of cysteine residues. There are four cysteine residues in GFOR, none of which are involved in disulfide bond formation. Three of these (Cys 127 , Cys 138, and Cys 158) are in the vicinity of the active site. Cys 127 directly precedes the conserved EKP motif, Cys 138 is at the beginning of the following helix (αE); and Cys 158 is in a short 3_{10} helical turn which links the two domains. Of these three amino acids, only Cys 158 reacted with ethyl mercury phosphate during crystal soaking experiments. Reaction of accessible thiol groups with the aldehyde group of the open chain form of glucose is likely to occur at room temperature. If catalysis was linked to a perturbation of the structure in the active site, this might allow one of these cysteine residues to react and subsequently inactivate the enzyme. This is a possible explanation for the strict link between catalysis and irreversible inactivation.

Obviously a great many questions concerning catalysis by GFOR remain to be answered. One of the more interesting concerns how the substrates bind and are released. Although GFOR shows a strong preference for the sugars which are its natural substrates (i.e. glucose

and fructose) [101], appreciable product formation only occurs in the presence of very high concentrations of these sugars [102]. Despite the fact that the crystals used in this structural study were grown in the presence of high concentrations of sorbitol, there is no convincing crystallographic evidence for the presence of either sorbitol or fructose in the active site of the enzyme. It should be noted that the conclusions in this respect are restricted by the moderate resolution of the structure determination.

3.6 GFOR AS A PERIPLASMIC ENZYME

Many of the enzymes found in the periplasm of Gram negative bacteria are involved in the degradation of molecules destined for import into the cell; the biosynthesis of the cell wall and other structural elements of the periplasmic region; and the modification of cytotoxic compounds [284]. The involvement of GFOR in a mechanism to protect the cell against osmotic stress makes sense of its periplasmic location. Here, both of its substrates are simultaneously available at saturating concentrations [285].

The existence of free NAD(P) in the periplasm seems unlikely for a number of reasons (although there is no direct experimental evidence regarding this matter). Firstly, such small hydrophilic molecules should readily diffuse through the solvent channels in the outer membrane [286]. Secondly, several phosphatase genes from *Z. mobilis* have been characterized; the cellular location of such enzymes in Gram negative bacteria is usually the periplasm [287]. Finally, NAD(P) would require an active transport system to cross the cell inner-membrane, yet has no assigned function in the periplasm. Hence GFOR, consistent with its periplasmic location, seems to have evolved a mechanism to retain NADP as an endogenous cofactor; the redox cycle is completed while NADP remains attached to the same enzyme.

This invites the question of why GFOR employs NADP as a cofactor, and not a group that more normally functions as an endogenous redox carrier. Most oxidoreductases that do not use NAD(P) as a cofactor employ riboflavin derivatives (FMN or FAD), which are covalently attached to the protein. Many enzymes in the periplasm of Gram negative bacteria employ the cofactor pyrrolo-quinoline quinone (PQQ) [288]. In this case PQQ, while not covalently attached to the protein, is bound sufficiently tightly to allow the entire redox cycle to occur on a single enzyme molecule. In fact, many aerobic Gram negative bacteria contain a PQQ-dependent glucose dehydrogenase in the periplasm, or associated with the inner cell membrane, which catalyses a reaction which corresponds to one of the half reactions of GFOR [288]. This enzyme is also found in anaerobic *Z. mobilis* [103] supporting the suggestion that this organism may have originated from aerobic ancestors [95]. However reduction of fructose to sorbitol by a PQQ-dependent enzyme seems thermodynamically improbable,

due to the relatively high redox potential of the PQQ/$PQQH_2$ couple [288]. Ultimately, the choice of a pyridine-nucleotide-linked or flavin-linked enzyme for the biological role fulfilled by GFOR may have been an evolutionary one.

A final question concerns the transport of GFOR across the cytoplasmic membrane and its assembly into an active tetramer. It has been shown that in GFOR-recombinant strains of *Z. mobilis,* exhibiting 5-6 fold increased GFOR enzyme activity, a precursor form of GFOR accumulates in the cytoplasm [289]. The N-terminal sequence of this precursor matches the leader sequence in the coding region of the gene. The precursor is enzymatically active and contains the cofactor NADP. From this it was suggested that NADP is bound by the precursor GFOR before it is processed and exported to the periplasm. More recent results support the idea that GFOR is exported by the conventional secretory pathway [285]. However it is generally believed that proteins are translocated across membranes in a partially unfolded state [290, 108, 291]. It has been shown that NADP binds to GFOR in a conventional fashion, and that the tight association of NADP may in part be linked to the quaternary structure of the enzyme. NADP is released on denaturation of the protein [117], and the crystallographic results also indicate that the cofactor is not covalently bound. It is not clear if NADP could remain associated with the protein during transport if the quaternary structure is disrupted and the protein partially unfolded. The problem of cofactor acquisition by periplasmic enzymes is not restricted to GFOR, and must also be faced by a number of other proteins (for example the haem containing cytochrome family) [284].

3.7 CONCLUSION

There is much general interest in the anaerobic Gram-negative bacterium Zymomonas mobilis because of its potential application as a biocatalyst in industrial ethanol production. This microorganism will tolerate high concentrations of sugars in its growth medium. In order to overcome the associated osmotic stress, it appears that the periplasmic enzyme glucose-fructose oxidoreductase (GFOR) produces the compatible solute sorbitol from fructose (coupled with the oxidation of glucose to gluconolactone). GFOR is of interest because it has not been positively identified in other organisms, and because in contrast to many oxidoreductases, NADP is very tightly associated with the protein.

The structure determination by X-ray crystallography reveals that each subunit of the tetrameric protein is folded into two domains, one of which is the classical dinucleotide binding domain, or Rossmann fold. The second domain is a nine-stranded predominantly antiparallel β-sheet around which the tetramer is constructed. Preceding the Rossmann fold there is a 30 amino acid proline rich 'arm' which wraps around an adjacent subunit in the tetramer. The

N-terminal arm buries the adenine ring of the NADP, and may also be involved in stabilization of the quaternary structure of the enzyme, as is the case for mammalian lactate dehydrogenases. An unsuspected structural relationship has been discovered between GFOR and the cytoplasmic enzyme glucose-6-phosphate dehydrogenase (G6PD), and a strong argument can be made for the existence of a corresponding evolutionary relationship between them. It is suggested that GFOR and G6PD derive from an common ancestral gene, and GFOR has evolved to allow it to function in the periplasm where it is required. Thus GFOR would seem to provide a clear example of how bacteria adapt preexisting structural domains for new roles in the cell.

The enzymes that use NAD(P) as an endogenous redox carrier (i.e. one which is not released during their catalytic cycle) appear to be structurally and functionally diverse. Few have been extensively characterized. The structure of GFOR reveals that the NADP is bound in a conventional fashion but cannot dissociate from the enzyme because it is effectively buried by several structural elements, including the extended N-terminal arm. Hence, it seems likely that proteins which tightly bind NAD(P) will employ conventional dinucleotide binding motifs, with suitable modifications to prevent cofactor dissociation. Several intriguing questions still remain unanswered. One of these concerns how GFOR is transported across the cell inner membrane into the periplasmic region and is assembled into an active tetramer. Another concerns whether the two substrates of the enzyme bind in an analogous fashion, and how the oxidative and reductive half-reactions are catalyzed.

Chapter 4

BILE SALT DEPENDENT LIPASE

4.1 INTRODUCTION

4.1.1 General background

Long chain triacylglycerols cannot be absorbed directly by the epithelial cells which line the walls of the small intestine. In order for absorption to occur, dietary triacylglycerols must be converted to more polar free fatty acids and monoacylglcerols. Other dietary lipids such as phospholipids and the esters of cholesterol and the fat soluble vitamins must also be hydrolyzed before absorption can take place (see [293] for general reviews of dietary lipid absorption). The enzymes responsible for facilitating this conversion are the lipases active in the gastrointestinal tract.

4.1.2 Lipases

Lipases catalyze the hydrolytic cleavage of ester bonds in lipids. Hence lipases are a special class of esterases which act on water-insoluble substrates. Sarda and Desnuelle [294] proposed that lipases could hydrolyze aggregated (but not dispersed) lipids, or in other words that lipases function at a water-lipid interface. Consequently when substrates are presented to these enzymes in a properly fashioned second phase, a micelle or a lipid droplet, a dramatic increase in enzymatic activity is observed. This phenomenon was termed interfacial activation. In order to reach the interface the enzyme must be soluble, and it was proposed by Desnuelle, Sarda and Ailhard [295] that binding to a water-lipid interface might be associated with a conformational change in the enzyme.

The recent determination of the three dimensional structures of a number of lipases by X-ray crystallography has begun to establish the structural basis for lipase behaviour (for reviews see [296, 297, 298, 299, 300]). Several common structural features shared by the lipases have emerged from these studies.

Virtually all lipases and esterases sequenced contain the Gly-X-Ser-X-Gly motif (where X is any amino acid, and the Ser is essential for catalysis) [301]. In all the lipase structures solved to date, the serine is located in a sharp turn between a β-strand and a buried α-helix. The catalytic serine is involved in a hydrogen bonded network of Ser...His...Asp (or Glu), which will serve to make the serine hydroxyl strongly nucleophilic. By analogy with other serine hydrolases [302, 303, 304] the hydrolysis of the ester bond will proceed by an acyl

transfer mechanism, which involves a nucleophilic attack of the serine hydroxyl on the scissile carbonyl carbon of the substrate. Proton donation by the histidine results in release of the alcohol moiety of the original substrate (e.g. a diacylglycerol in triacylgycerol hydrolysis), and the concomitant formation of an acylenzyme intermediate. The second half of the reaction is essentially a reversal of the first, with a water molecule as the attacking nucleophile, resulting in the release of a free fatty acid and regeneration of the enzyme.

In all conventional lipases which exhibit interfacial activation, the active site serine is buried beneath a flexible 'lid' region, comprising one or more surface loops (see Figure 4.1, Figure 4.2). Displacements of the lid regions are associated with the exposure of a hydrophobic surface around the active site. However the topology of the lid regions and the mechanism of lid opening are very different for the lipases for which this process has been characterized. In essence this would seem to be a mechanism for maintaining high solubility of the closed form while maintaining the ability of the active (open) form to associate with the hydrophobic water-lipid interface. Strong supporting evidence for the involvement of the lid region in interfacial activation comes from the study of natural and recombinant lipase mutants (see e.g. [305, 306, 307]).

While many advances have been made in the understanding of lipolytic enzymes, much remains to be determined. What is at present poorly understood is the regulation of lid opening in the lipase family (see [308, 309] for discussion). The structure of cutinase [310] shows that there are some soluble lipases which do not have a lid structure and which do not exhibit interfacial activation. Additional questions are posed by the phospholipases, which lack the lid structures seen in the conventional lipases, yet still exhibit interfacial activation, mediated perhaps by more subtle structural changes (see [311, 312].

The two principal lipases active in the intestinal tract of humans are the pancreatic colipase-dependent lipase (CDL) and bile salt dependent lipase (BSDL). It is this second, intriguing, enzyme which is the subject of this chapter. Bile salt dependent lipase is synthesized and secreted by the pancreas, but is also found in appreciable quantities in human milk. In contrast to the fungal and microbial lipases (which form the bulk of the lipases structurally characterized to date), CDL and BSDL act in the intestinal tract in the presence of high concentrations of the amphiphilic bile acids (which act to emulsify the lipid substrates of the enzymes). This additional *in vivo* complexity is reflected in their properties, and the elegant investigations of pancreatic CDL by X-ray crystallography [313, 314, 315, 316, 317] have begun to explain these at a structural level. BSDL, an enzyme which does not exhibit interfacial activation in the classical sense, but is absolutely dependent on bile salts for activity

against water insoluble substrates, is at this stage less well characterized. In this chapter, progress toward the structure determination of BSDL is reported.

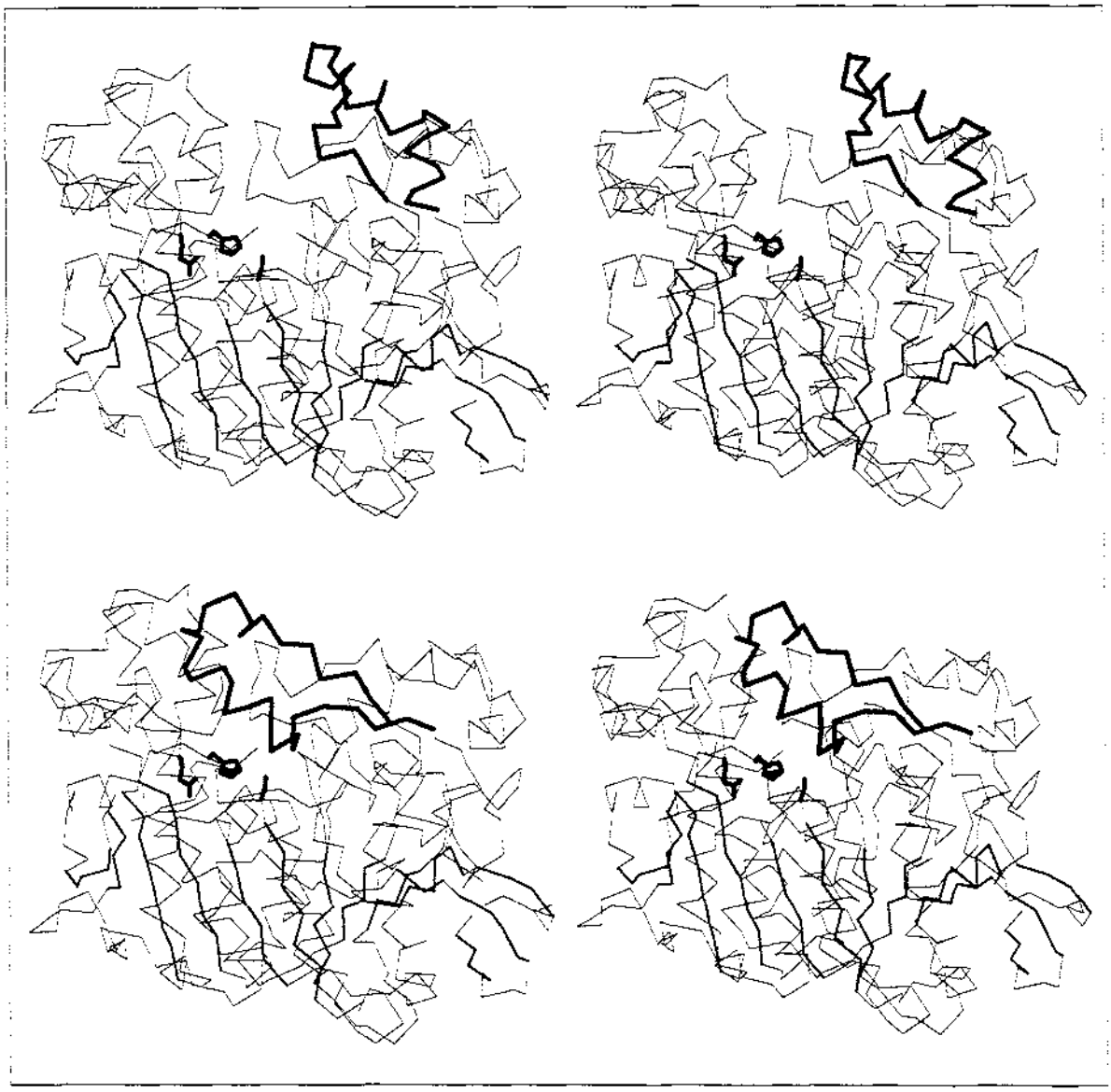


Figure 4.1 Diagram showing the conformational change associated with activation in *Candida rugosa* lipase

Cα plots of the structure of *Candida rugosa* lipase in (top) the 'open' or active state. (bottom) the 'closed' or inactive state [318, 319]. Ser 209, Glu 341 and His 449, the 3 residues of the catalytic triad are shown in full. β-strands are shown in medium lines, the active site 'lid' in thick lines. Conformational variability in loop(s) covering the active site has been observed by X-ray crystallography for lipases from the fungi *Rhizomucor miehei* [320, 321], *Candida rugosa* [319], *Humicola lanuginosa* [322] and *Rhizopus delemar* [323]; and also for mammalian pancreatic colipase-dependent lipases [315, 317, 324]. Other lipase structures (e.g. the lipases from *Pseudomonas glumae* [325], and *Candida Antarctica* [326]) also reveal lid regions adjacent to the active site, and it is reasonable to infer that these undergo conformational change during activation. The figure was produced using the program Molscript [327].

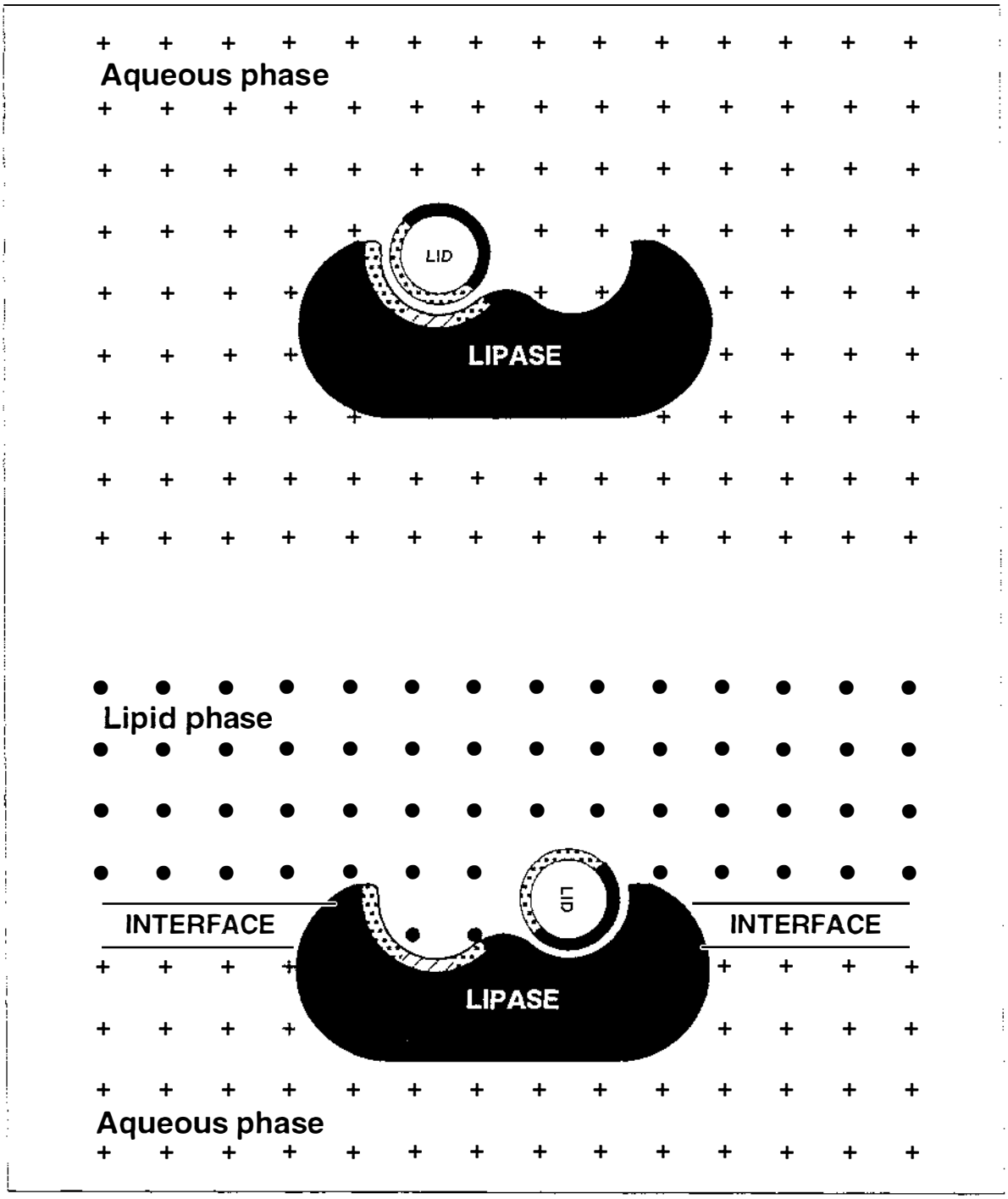


Figure 4.2 Schematic diagram showing the conformational change associated with interfacial activation in the fungal lipases.

Hydrophilic surfaces are coloured black, hydrophobic surfaces are stippled, and the active site is hatched. Diagram after Dodson *et al* [296].

4.1.3 Bile salt dependent lipase

4.1.3.1 Historical background

Lipolytic activity in human milk was discovered early this century [328], and studied extensively in the following years by Freudenberg [329]. Subsequently it was demonstrated that human milk contained at least two lipases, a lipoprotein lipase at a relatively low concentration, and a second lipolytic enzyme present at much higher concentrations. This latter enzyme was termed bile salt stimulated lipase (BSSL) because of an apparent activatory effect of bile salts [330, 331]. It is stable enough to survive passage through the stomach [332] and acts in the intestinal lumen of the newborn where it encounters activatory concentrations of bile salts. In contrast, the lipoprotein lipase is rapidly inactivated in the gut, and its physiological function remains unclear [333]. Because of the potential importance of BSDL in infant nutrition (see [334]), it has been extensively studied.

The same protein is synthesized and secreted into the intestine by the acinar cells of the human pancreas. It was designated carboxylic ester hydrolase by Lombardo, Guy and Figarella [335] who purified the human enzyme to homogeneity. This enzyme also attracted much interest because of its role in dietary lipid absorption, and in particular for its ability to catalyze the hydrolysis of cholesterol esters. While it was recognized early that the enzymes secreted by the pancreas and by the lactating mammary gland might be identical [336], this was confirmed only relatively recently by isolation and sequencing of cDNA from these two tissues [337, 338, 339, 340]. It is clear that both are products of the same gene, which has now been characterized [341, 342]. Thus historically the enzyme has been studied from two different perspectives, and it is only recently that this knowledge has been effectively unified.

As discussed below the enzyme has an extremely broad substrate specificity. An unfortunate consequence of this is that the enzyme is known by a plethora of names in the literature (catalogued by Wang and Hartsuck [343]). Here we use the name bile salt dependent lipase (BSDL).

4.1.3.2 Enzyme characteristics

BSDL has a broad specificity with respect to the chemical structure of the substrate (see [344, 345] and references therein). It hydrolyses water insoluble long-chain fatty acid esters of glycerol and cholesterol, as well as esters of the fat soluble vitamins A, D and E and long chain phospholipids, in the presence of primary bile salts (i.e. bile salts containing the 7α

hydroxyl group, see Section 4.1.3.12). BSDL also hydrolyses water-soluble esters such as short chain fatty acid esters of glycerol, 4-nitrophenyl acetate and methyl butyrate, but in this case activation by bile salts is not essential. A lipoamidase activity (which is not bile-salt-dependent) has also been demonstrated [346]. The specificity of the enzyme with respect to the physical state of the substrate (e.g. emulsified, micellar, vesicular) is less well defined. The general difficulties associated with studying lipolytic reactions are compounded, because the addition of bile salts required to activate the enzyme against lipid substrates alters the physical state of the lipids themselves. There is still much to be learned, as recent studies have shown that the solubilization and transport of lipids and lipolytic products *in vivo* is more complex than originally assumed [347, 348].

4.1.3.3 BSDL in milk

For a long time it was believed that BSDL was a constituent of milk of only the highest primates. The milk from many other mammals showed no apparent BSDL activity [349, 350]. More recently however, BSDL activity was demonstrated in the milk of several species of monkey [333], in the milk of cats and dogs [351] and ferrets [352]. At the DNA level, expression of the BSDL gene in the lactating mammary gland has been detected in mouse, cow and goat [353]. Hence it seems likely that the gene is expressed in the lactating mammary gland of many mammalian species. Earlier failure to detect BSDL activity may have been due to variation in expression levels during lactation (see e.g. [351]), or to low overall levels of expression.

The importance of BSDL in the utilization of milk lipids is supported by *in vivo* experimentation. Several studies have shown that pasteurization of human milk under conditions which inactivate BSDL (see [354]) leads to a significant reduction in fat absorption in pre-term infants [355, 356, 357]. More recently a study of the effect of BSDL on the weight gain of kittens [358] has provided further direct evidence for the importance of the enzyme for lipid digestion in the neonate. It is of interest that in the newborn infant, exocrine pancreatic function is not fully developed, and the intraluminal levels of the pancreatic digestive enzymes are considerably lower than in adults, which emphasizes the compensatory effect of the BSDL in human milk [359].

The digestion of milk lipids must be considered to be the result of the concerted action of several lipases (in addition to BSDL, gastric lipase, and the pancreatic CDL), which have overlapping but not identical functions [360, 361]. BSDL will not catalyze the hydrolysis of long chain triacylglycerols in milk, but will become activated when, after passage through the stomach into the intestinal lumen, the partially digested milk mixes with bile salts. Con-

sidering the hydrolysis of triacylglycerols, BSDL is unique in that it shows no positional or stereospecificity, and will hydrolyze all three ester bonds. A recent study has suggested a role for BSDL in the efficient use of long chain polyunsaturated fatty acids in milk [362].

4.1.3.4 Pancreatic BSDL

BSDL has been found to be excreted from the pancreas in all species studied [363]. In some non-mammalian species, BSDL has been found to be the only lipase excreted by the pancreas [364], which underlines its potential to mediate complete intestinal fat absorption because of its marked non-specificity.

Because BSDL is the only digestive enzyme capable of catalyzing hydrolysis of cholesterol and vitamin esters, this has been considered its most important function in adults [344]. Here again it has been shown that lipid digestion is the result of the concerted action of a number of enzymes [365, 366]. The general conclusion from these studies was that the actions of one lipolytic enzyme, by changing the physical and chemical state of the lipids, could make substrates for other lipases available for hydrolysis, resulting in a cooperative lipolytic process. The involvement of BSDL in cholesterol absorption has been controversial. It was proposed that BSDL, in addition to its ability to hydrolyze cholesterol esters might have a cholesterol transport function [367, 368]. However recent work, including studies on BSDL-knockout mice, has demonstrated that BSDL is responsible for mediating intestinal absorption of cholesterol esters but does not play a primary role in free cholesterol absorption [369, 370].

4.1.3.5 Genomic organization and regulation

In humans, little is known about the regulation of the gene, other than that BSDL is synthesized in the pancreas, and in the mammary gland during lactation. In the pancreas, recent evidence supports the idea that BSDL synthesis may be regulated by transcriptional and translational mechanisms in response to diet [371, 372, 373]. Evidence for synthesis in some other tissues is more equivocal. There have been conflicting reports of BSDL synthesis and activity in the liver of various species [374, 375, 376, 377, 378, 340, 379, 380]. However, using PCR based techniques, BSDL transcripts were shown to be absent from total RNA from the liver, in both humans [381] and mice [382]. In the rat, a hepatic cholesterol ester hydrolase has now been cloned which is homologous to (but clearly distinct from) BSDL [383]. This may explain some of the confusion which has arisen. It appears likely that BSDL, or a BSDL-like protein, is produced by the human eosinophils (leukocytes of the myeloid lineage) [384]. BSDL activity has been reported in the plasma of humans and other mammals, but again evidence based on DNA techniques is lacking [385, 386]. Finally BSDL gene

transcripts have been detected in many fetal tissues, which is in contrast to the marked tissue-specificity of transcription found in adults [387]. The physiological significance of this result has not yet been explained.

While the exact tissue distribution and other regulatory aspects of BSDL remain to be determined, current evidence suggests that the enzyme serves several different functions, and that it is not simply utilized in lipid digestion. In eosinophils for example, it is proposed that the enzyme might play a protective role, by preventing the cell from being lysed by the lysophospholipid present in the plasma membrane of parasites [384].

The BSDL gene itself contains 11 exons and 10 introns and has been mapped to the most distal part of the long arm of human chromosome 9 (9q34-qter) [342, 388, 341]. The BSDL locus exhibits polymorphism [381] which seems certain to be associated with a tandem repeat region found in exon 11. Tandem repetitive regions in vertebrate DNA often show substantial allelic variation in the number of repeating units [389, 390]. The presence of an apparent pseudogene, closely related to BSDL has also been reported. This product of this pseudogene, which differs principally from BSDL in the deletion of exons 2-7 (the active site serine is in exon 5), is unlikely to have any catalytic activity, and seems to be expressed ubiquitously at a much lower level than BSDL [381, 387].

The secretory pathway of BSDL in the pancreas has been studied [391, 392], leading to the suggestion that biosynthesis may involve association with intracellular membranes, and that the folding of the enzyme may be assisted by molecular chaperones.

4.1.3.6 Protein sequence

The nucleotide sequence for human BSDL specifies an open reading frame of 742 amino acids. The mature protein contains 722 amino acids, and is preceded by a 20 amino acid signal sequence. The N-terminal catalytic domain of the protein (comprising residues 1 - 535) is homologous with a number of other lipases and esterases (discussed in the following section).

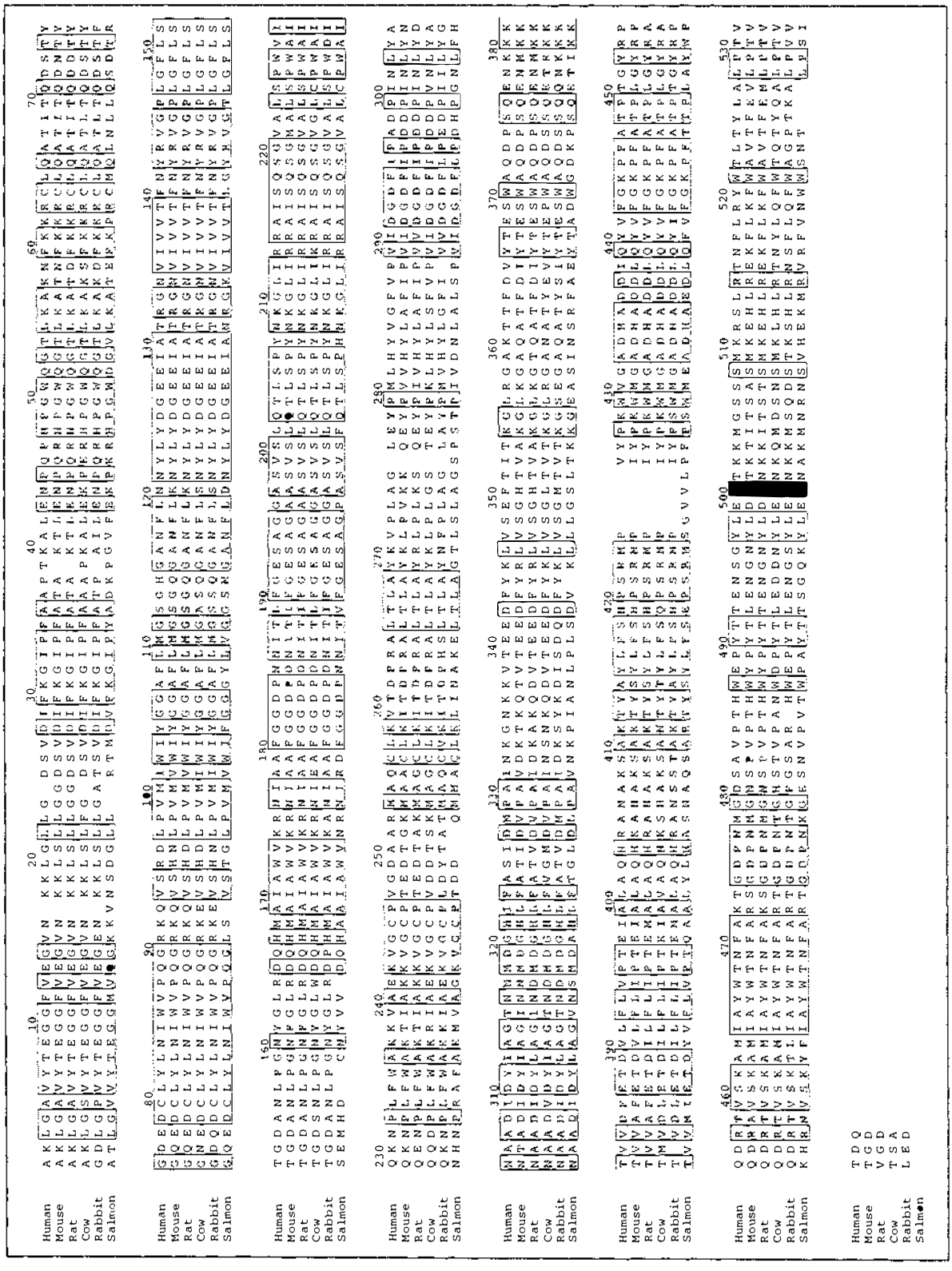


Figure 4.3 Alignment of known BSDL sequences.

A multiple sequence alignment of the N-terminal domains of BSDL (the C-terminal tandem repetitive region is not shown). Conserved amino acids are boxed. Multiple sequence alignment was by the method of Barton and Sternberg [279]; the figure was produced with the program ALSCRIPT [281].

Following this domain is the C-terminal tandem repeat sequence, comprising 16 almost identical repeats of 11 residues each. The repeats have the consensus sequence (Gly Ala Pro Pro Val Pro Pro Thr Gly Asp Ser). In addition to the human enzyme, the sequences of the BSDL from mouse [382, 353], rat [393, 378, 394], cow [395], rabbit [396] and salmon [397] have also been determined, by characterization either of cDNA or of the BSDL gene itself. A multiple sequence alignment (excluding the C-terminal tandem repeat region) is shown in Figure 4.3. The tandem repeats differ both in composition and number between species [353], and are missing altogether in the salmon enzyme. The N-terminal domain shows a high degree of overall sequence conservation, with the salmon BSDL being clearly demarcated as a result of its evolutionary distance from the other (mammalian) sequences.

4.1.3.7 The esterase/lipase protein family

BSDL is a member of a functionally diverse group of proteins which share what is now termed the α/β hydrolase fold [398, 399]. All of these proteins have at their core a topologically similar β-sheet, which is packed on either side with α-helices. Based on structural and sequence homology, this superfamily can be further subdivided. BSDL is a member of a lipase/esterase subfamily which includes the choline esterases, a number of conventional lipases and esterases (including some which, through gene mutations, confer organophosphorous insecticide resistance in insect species; see e.g [400, 401]), and also domains of proteins devoid of known catalytic function (e.g. the thyroid hormone precursor, thyroglobulin) [402, 403]. The structures of three proteins from this subfamily have been determined by X-ray crystallography, the acetylcholinesterase from *Torpedo californica* (TcAChE) [404], and the lipases from the fungi *Geotrichum candidum* (GCL) [405, 406] and *Candida rugosa* (CRL) [318]. These proteins have pairwise amino acid identities of 30 - 40%, and have in common a core structure which is remarkably conserved. A topology diagram for the family is shown in Figure 4.4. A ribbon diagram of *T. californica* acetylcholinesterase, a prototypical member of the family, is shown in Figure 4.5 (see also the Cα plots of *C. rugosa* lipase in Figure 4.1). The α/β hydrolase fold appears to be a stable scaffold, with the functional determinant of the specific enzymes being the surface structural elements found 'inserted' between elements of the core structure.

4.1.3.8 The catalytic triad

Prior to the availability of nucleotide sequence data for the human BSDL, the involvement of Asp, Ser and His in the mechanism of substrate hydrolysis had been inferred [407]. Sequence comparison techniques coupled with site directed mutagenesis of the rat pancreatic BSDL (which has a very high level of sequence identity with the human enzymes) led to the

definitive identification of the active site serine (Ser 194), histidine (His 435) and aspartic acid (Asp 320) residues [408, 409, 410]. The structure determination of several enzymes with a clear relationship to BSDL has confirmed this.

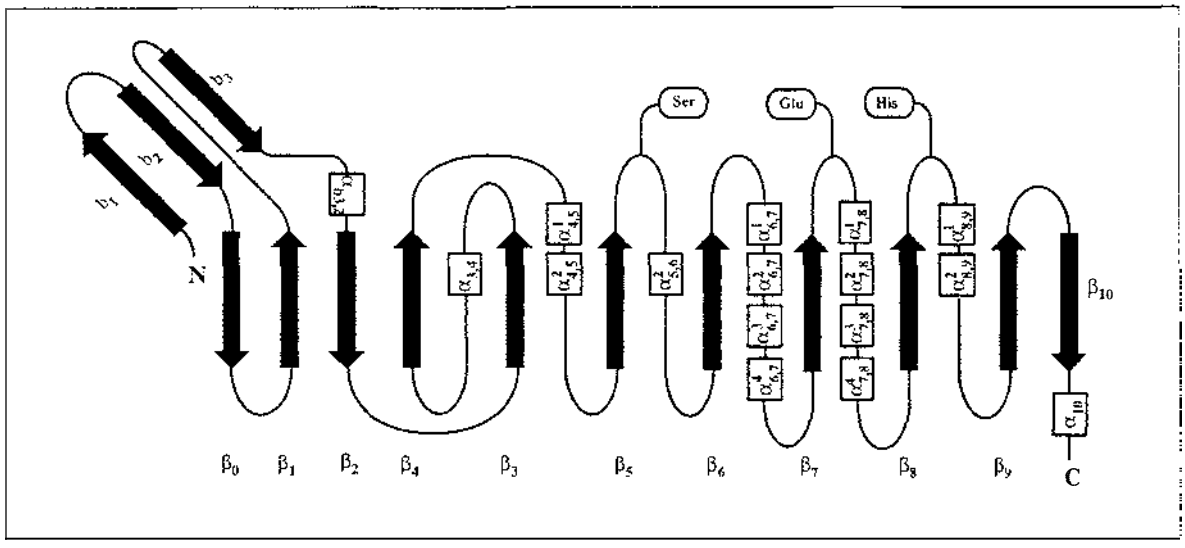


Figure 4.4 Topology diagram of the lipase/esterase family fold.

The strands in the large central β-sheet are labelled β_i (i = 0 to 10) with the numbering starting at 0 for consistency with the original definition of the α/β hydrolase fold [398]. The strands in the smaller β-sheet are labelled b_i (i = 1 to 3). α–helices are designated $\alpha^k_{i,j}$, where the subscripts i and j refer to the β-strands between which the helix is found, and the superscript k refers to the sequential number of the helix within this connection. Not all of the helices are conserved between the three structures of the family that have been determined to date. The topological position of the residues of the catalytic triad (Section 4.1.3.8) is also indicated. After Cygler *et al* [403], Grochulski *et al* [318].

4.1.3.9 Glycosylation

Both the human milk BSDL and its pancreatic counterpart are heavily glycosylated [411]. Recent reviews of glycobiology [412, 413, 414] emphasize that there is no single unifying function for the oligosaccharides attached to proteins, and with respect to BSDL, their role is not fully understood. There is a single potential N-glycosylation site in the human enzyme (Asn 187), which appears to be conserved across species (see Figure 4.3). Human BSDL expressed in both the mammary gland and the pancreas is glycosylated at this site, and the N-linked oligosaccharide structure of the pancreatic BSDL has been characterized and shown to be relatively heterogeneous [415, 416]. N-glycosylation does not affect enzymatic function [417, 418]. Some studies have suggested that the N-linked oligosaccharide structure is important for enzyme secretion, but other studies have not supported this idea [419, 420, 421, 422].

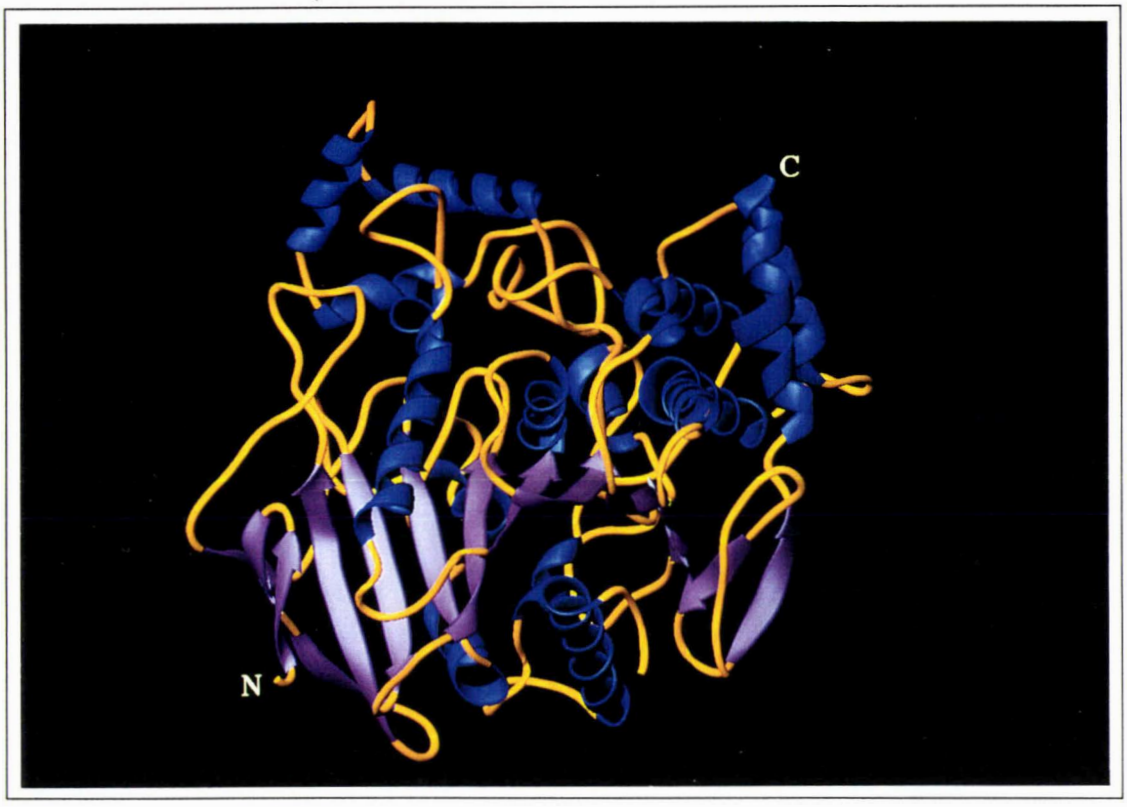


Figure 4.5 Ribbon diagram of *T. californica* acetylcholinesterase

β-strands are shown in purple, and helices in blue, with connecting loops in yellow. The figure was prepared using the program Ribbons 2.0 [250].

Most of the carbohydrate in the protein is associated with the C-terminal repeat region which contains many potential O-glycosylation sites. [339, 417]. Recently it was shown that the repeat region in milk BSDL contains threonine-linked oligosaccharides, the structures of which were partially characterized [423]. Reported differences in electrophoretic mobility between the BSDL expressed in the mammary gland and the pancreas [69] are most likely due to differences in glycosylation.

4.1.3.10 The C-terminal repeat region

The expression of recombinant BSDL variants which lack the C-terminal tandem repeat region has shown that this domain is not essential for catalytic activity or bile salt activation of the enzyme [417, 418, 422, 424]. In some respects the C-terminal region of BSDL might be termed a mucin-like domain. The mucous glycoproteins have a protein backbone consisting of repetitive amino acid sequences [413, 425, 426, 427]. The repeats are rich in serine and threonine residues, the potential O-glycosylation sites, and usually also contain many proline

residues. These structures are believed to adopt an extended conformation with the addition of numerous, bulky, O-linked glycans.

The biological significance of the C-terminal domain remains a challenging question. One hypothesized role is stabilization of the enzyme, as BSDL expressed in the mammary gland must survive passage through the stomach and into the intestinal tract. This hypothesis is consistent with the report of increased proteolytic susceptibility of a truncated variant lacking the C-terminal repeats [424]. It is also consistent with the absence of the repeat region in the salmon BSDL. It has also been suggested that the C-terminal region might contribute an adhesive activity to BSDL, through the O-glycan structure [423]. Finally, it has been suggested [378] that the C-terminal repeat region, being relatively enriched in Pro, Asp, Glu, Ser and Thr residues, may serve as a signal for selective proteolysis of the protein, in accordance with the PEST region hypothesis of Rogers, Wells and Rechsteiner [428]. However, with respect to these latter two hypotheses, there is no direct evidence for a physiological role of the C-terminal region in adhesive events, or that BSDL has a fast turnover rate *in vivo*.

4.1.3.11 Heparin binding

Heparin-Sepharose affinity chromatography is used in some purification procedures for human milk BSDL [69, 429], effectively showing that BSDL binds the glycosaminoglycan heparin reversibly (glycosaminoglycans are linear, sulphate-substituted polymers composed of alternating hexouronic acid and hexosamine units [430, 431]). Glycosaminoglycans are a component of the glycocalyx which surrounds the microvilli on the epithelial cell surface [430] (see [432] for a review of intestinal epithelial cell structure). Consequently it was suggested that *in vivo*, this could lead to localization of the enzyme at the luminal surface (the site for membrane transport of the enzymatic reaction products) [429]. More recently it has been shown that pancreatic BSDL does bind to membrane-associated heparin of intestinal epithelial cells, and that binding was reversed by the addition of soluble heparin [433]. This supports the idea that the enzyme is localized at the cell surface to facilitate the uptake of hydrolyzed dietary lipids. Other studies have also indicated that BSDL may be membrane associated [434]. Of related interest may be the identification of relatively high concentrations of glycosaminoglycans on the surface of the human milk fat globule [435].

Cyanogen bromide cleavage of the enzyme showed that the N-terminal fragment (residues 1-101) retained heparin binding properties [339]. Since glycosaminoglycan binding domains in proteins are known to be associated with clusters of basic amino acids [430, 431, 436] several groups have proposed that a highly basic region in the BSDL sequence (Lys Ala Lys Asn Phe Lys Lys Arg, between residues 55 and 64) may be associated with the heparin binding

activity [339, 378].

4.1.3.12 Bile acid activation of the enzyme

The bile acids are a family of steroidal compounds based (in mammals and other higher vertebrates) on the C_{24} compound cholanoic acid (Figure 4.6) [437, 438]. Cholanoic acid possesses a four-ring steroidal nucleus, and a five carbon side chain terminating in a carboxylic acid. The steroidal nucleus can be substituted with hydroxyl groups at positions 3, 6, 7 and 12. Since each of these hydroxyl groups may be positioned above (β) or below (α) the plane of the molecule, a large number of permutations are possible, although some are much more common in nature than others. Because they possess both substantial hydrophilic and hydrophobic regions, bile salts are amphiphilic molecules (see Figure 4.7).

Bile acid synthesis occurs in the liver; these primary bile acids are then amidated with glycine or taurine and secreted into the small intestine, where they are involved in the solubilization of lipids. The enterohepatic circulation of bile acids is quite complex (see [437] for details). The primary bile acids in man are cholic ($3\alpha,7\alpha,12\alpha$) and chenodeoxycholic ($3\alpha,7\alpha$) acids. As a result of bacterial transformation reactions in the gut substantial quantities of deoxycholic ($3\alpha,12\alpha$) and lithocholic (3α) acids are also formed. These bile acids (which lack the 7α hydroxyl group) are sometimes termed 'secondary bile acids'.

The mechanism of activation of BSDL by bile acids appears complex and is not understood in structural terms. A tentative model for the bile salt interaction with the enzyme has been proposed, consistent with experimental results to date [421]. Under this hypothesis, bile salts can interact with two sites on the protein. The first site is less specific, and can bind both primary and secondary bile salts. Binding of bile salt to this site is a prerequisite for binding the enzyme to the aggregated-lipid surface. However simple association with the lipid-water interface in this fashion does not cause activation (if it did there would be no obligatory requirement for primary bile salts). Consequently it is proposed that a second site must be present which is specific for primary bile salts (containing the 7α hydroxyl group), and is associated with enzyme activation. While this is a plausible hypothesis, direct evidence for multiple bile salt-binding sites is lacking. Chemical modification studies suggest that arginine residues are involved in bile salt binding [439, 421]. Several groups have used analysis of circular dichroism in an attempt to monitor conformational changes associated with bile salt binding, but their results have been contradictory [440, 441].

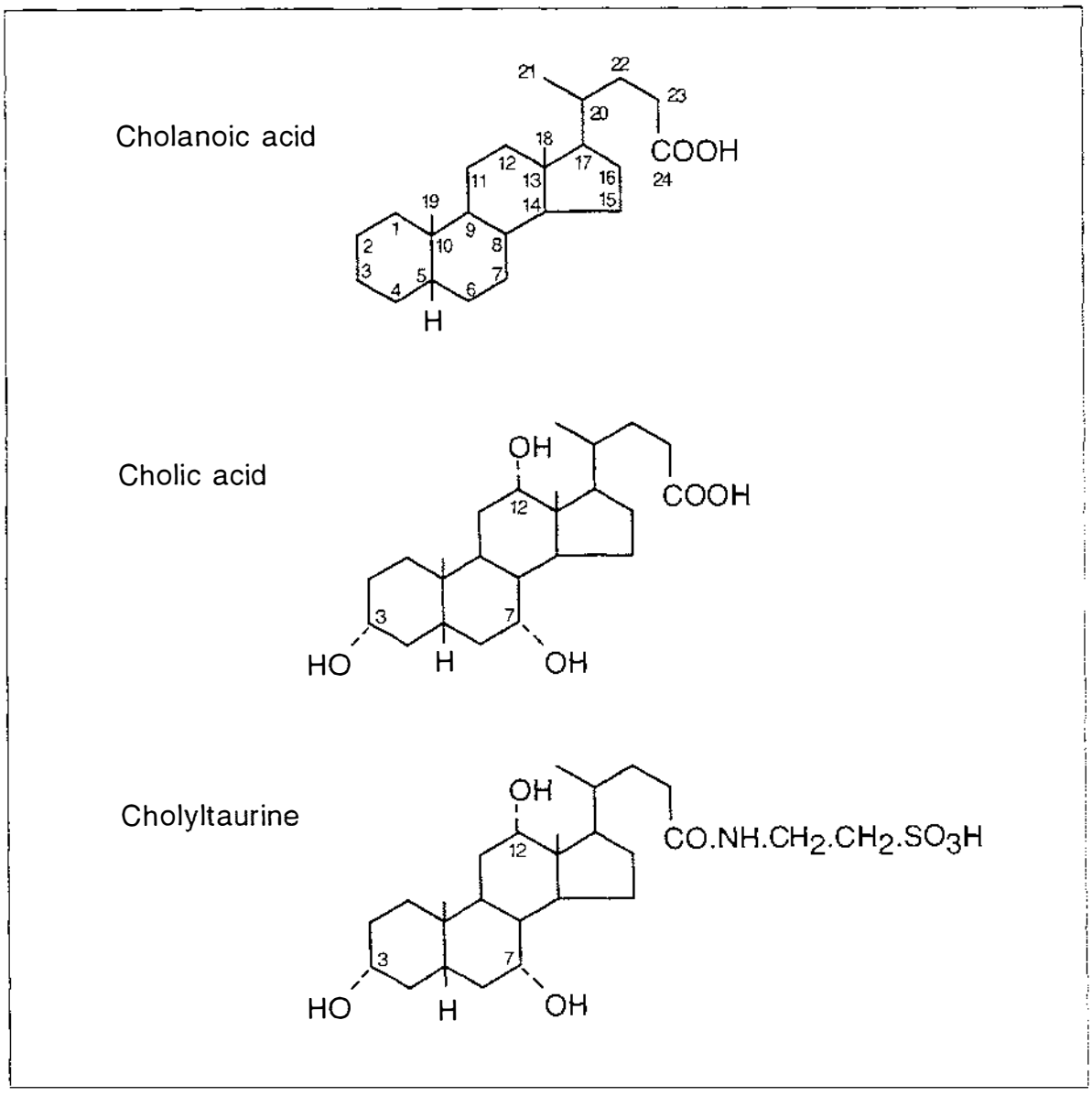


Figure 4.6 Bile acid structure

For a full description of bile acid nomenclature see Hofmann *et al* [442].

BSDL is readily inactivated by classical serine hydrolase inhibitors, such as diisopropylfluorophosphate [335, 69], whereas other lipases appear much more resistant to inhibition (e.g. pancreatic colipase-dependent lipase [443], and the lipase from *Rhizomucor miehei* [320]). These latter two observations are consistent with the presence of a lid which covers the active site serine. The ready inhibition of BSDL, considered in conjunction with its known esterase activity, implies that the structural basis for the bile salt-mediated activation is different from the activation by 'lid' movements seen in the lipase structures solved to date. In fact the region which contains the active site lid in the related fungal lipases GCL and CRL (the con-

nection between strands b_2 and β_3 (Figure 4.4, see also Figure 4.1) contains a conserved deletion in the BSDL family.

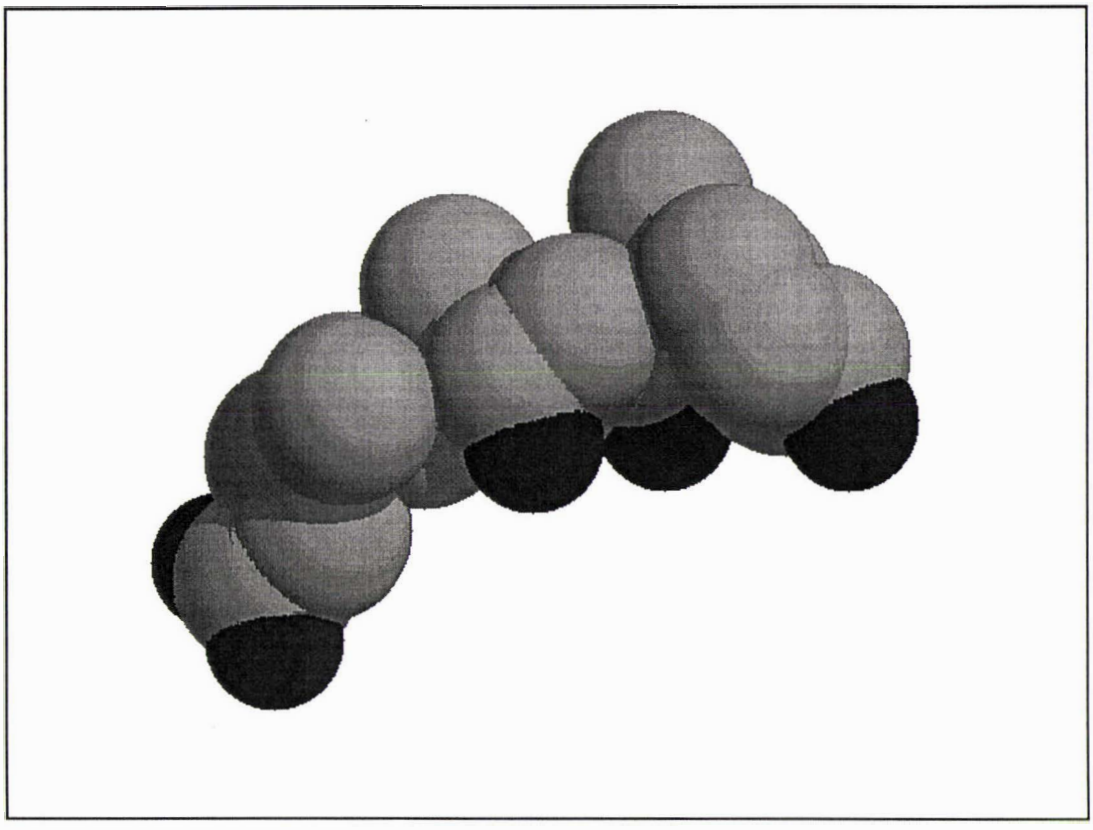


Figure 4.7 Space filling model of cholic acid

Carbon atoms are in grey, and oxygen atoms in black. Atomic coordinates for cholic acid were taken from a crystallographic structure determination [444]. The figure was prepared with the program RasMol 2.4 [R. Sayle].

4.1.3.13 Pancreatic colipase-dependent lipase

The interaction of BSDL with bile salts suggests interesting parallels with the pancreatic colipase-dependent lipase, the other lipolytic enzyme active in the intestinal tract. This enzyme is strongly inhibited by bile salts [445]. The physical basis of this inhibition is probably the reduction in the hydrophobicity of the lipid-water interface caused by the introduction of the amphiphilic bile salts, which can form mixed micelles with the lipid substrates. In order to overcome this inhibition the pancreas secretes a small activatory protein, colipase, which anchors the enzyme to the bile salt-coated water-lipid interface [446]. Crystallographic studies of the pancreatic lipase-colipase complex [314, 315, 317, 324] show that colipase facilitates the adsorption of pancreatic lipase to the lipid surface by binding to the C-terminal non-catalytic domain of the enzyme on the same face of the molecule as the

active site (which is covered by a helical lid). The colipase molecule is comprised of three 'fingers', the tips of which are hydrophobic and face outwards from the complex. The structure of a lipase-colipase complex with a bound phospholipid reveals that the lid is displaced and undergoes a marked change in secondary structure, exposing a solvent-accessible hydrophobic groove, at the bottom of which is located the active site serine. In contrast, for BSDL it appears that adsorption to the lipid-water interface is facilitated by binding to bile salts themselves (however activation of the enzyme is mediated only by primary bile salts which contain the 7α hydroxyl group). While the events that surround the binding and activation of these enzymes are still unclear (particularly for BSDL) it seems that both proteins have evolved mechanisms which reflect the presence of the amphiphilic bile acids found at high concentrations in the intestinal milieu.

4.1.4 The role of structural studies

Structural studies were initiated on BSDL for several reasons. Firstly the work will contribute directly to understanding of lipolytic enzymes at a fundamental level. BSDL, in its activation by bile acids, is unique among lipases. Understanding how bile acids mediate the activation of the enzyme, its association with the lipid phase, and the connection between these events, will be greatly advanced by knowledge of the structure. Secondly, BSDL is a protein of considerable biological importance. It is directly implicated in fat digestion in the newborn, and in the absorption of esterified cholesterol in adults. Attempts to modify the activity of the protein, or to use recombinant proteins in therapeutic or nutritional roles (see [447]) require a much clearer understanding of both the N-terminal catalytic domain, and the C-terminal tandem repeat region. Full understanding of the marked non-specificity of the enzyme can only come from a structural characterization of the substrate binding site, and comparison with other lipase structures (see [448]).

In this chapter progress is reported toward the structure determination of BSDL. This has involved crystallographic work on native and recombinant proteins, including a truncated variant which lacks the C-terminal repeat region altogether. This last approach has provided crystals which diffract to a useful resolution. At the time of writing, a partial model of the N-terminal domain has been built, by interpretation of a map with effective resolution 3.5 Å (the structure was solved by the method of molecular replacement, using *T. californica* acetylcholinesterase as a search model). Very recently, a much better data set has been collected with effective resolution 2.8 Å. Completion of the structure now seems imminent. Here, preliminary crystallographic results are reported for the native and recombinant proteins, and a brief description given of the current status of the structure determination of the N-terminal catalytic domain.

4.2 NATIVE BSDL

4.2.1 Protein purification and crystallization

Purification of the protein from human milk was carried out by Dr. Lars Bläckberg (University of Umeå, Sweden), following published procedures [69]. Following purification the enzyme was extensively dialyzed against water, freeze-dried, and sent as a lyophilized powder. The crystallization of the enzyme has been fully described in Chapter 1 (Section 1.3.1). Large single crystals of native BSDL were grown by hanging or sitting drop vapour diffusion, from solutions containing 0.40 M MOPS/NH_4OH buffer at pH 6.8, and 15 - 20% (w/v) PEG 8000, with a protein concentration of 10 - 30 mg/ml (determined by absorbance at 280 nm) (see Figure 4.8, also Figure 1.2).

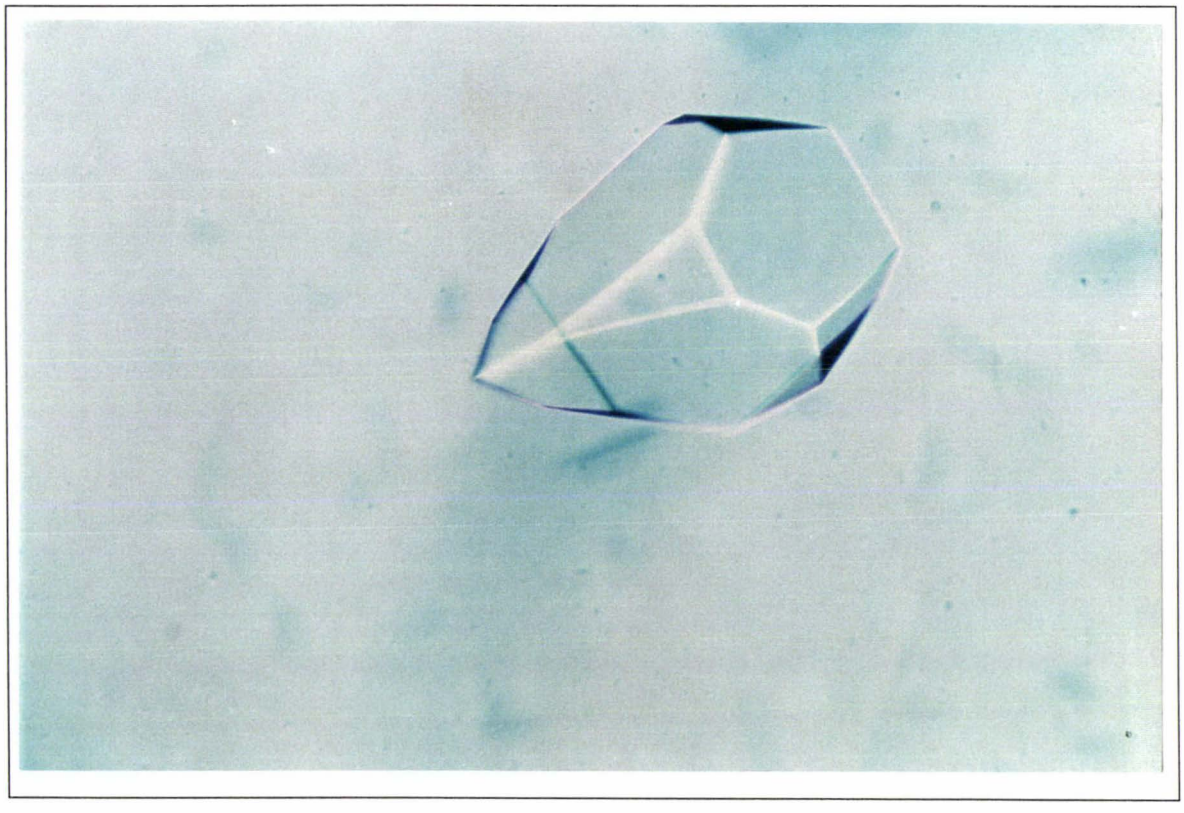


Figure 4.8 Crystal of Native BSDL

Crystals can be grown up to 0.8 mm in length in their longest dimension

4.2.2 Characterization of the crystals

X-ray diffraction patterns were characterized from numerous crystals of native BSDL. Crystals were conventionally mounted in thin glass capillaries, and data collected on the Raxis IIC system, with the crystals at ambient temperature. Unfortunately, in all cases the crystals

were poorly ordered, and diffracted anisotropically (see Section 4.2.3 below) to low resolution.

Autoindexing algorithms [449, 450] determined a lattice consistent with trigonal or hexagonal crystal symmetry. Examination of the measured intensities showed that the Laue class of the diffraction pattern was $\bar{3}m$. The orientation of the Laue symmetry with respect to the crystal lattice, and the systematic absence of every third reflection *00l,* indicates that the space group of the native BSDL crystals is $P3_121$ or its enantiomorph $P3_221$. Cell parameters are a = b = 90.0 Å, c = 156.1 Å, $\alpha = \beta = 90°$, $\gamma = 120°$.

Crystals typically diffracted usefully to 3.5 - 4.5 Å along the unique lattice direction (c*), but to only 5 - 7 Å in orthogonal directions. Nonetheless, given that the structure of the intact protein is of considerable biological interest, a data set was collected at the synchrotron radiation source at the Photon factory, Tsukuba, Japan.

4.2.2.1 Data collection at the Photon Factory

At the Photon factory a synchrotron radiation source is coupled with Weissenberg camera geometry [292] (see Section 4.4.3 for discussion). Data were collected from a single crystal mounted with the unique axis perpendicular to the spindle axis of the camera. A total of twelve diffraction images were collected. Total time of data collection (from first exposure of the crystal) was 80 minutes, with the crystal maintained at 4° C throughout this time. An oscillation range of 8° and a coupling constant of 2.0 °/mm were employed for the first 9 images (with oscillation range 6° and coupling constant 1.4°/mm for the last 3 images). Data processing was carried out using the program WEIS [451, 452]. Relative scaling of profile-fitted intensities from each image was by the method of Fox and Holmes [126]. The relative temperature factors obtained from the scaling procedure are consistent with significant radiation damage occurring over the time course of the experiment (results not shown).

The presence of significant anisotropy in the diffraction pattern introduces difficulties during data reduction. Usually, when processing film or image plate data, diffraction is assumed to be isotropic, and observations are measured out to a fixed resolution limit. This limit is usually determined by monitoring the mean signal to noise ratio, or the internal consistency of the data (conventionally assessed by the merging R-factor

$$R_{Merge} = \sum_{hkl}\sum_{j} \left| I_j(hkl) - \langle I(hkl) \rangle \right| / \sum_{hkl}\sum_{j} \left| I_j(hkl) \right| \qquad \textit{Eq. 4.1}$$

where $I_j(hkl)$ are the symmetry-equivalent intensity measurements for a reflection and $\langle I(hkl) \rangle$ is the weighted mean value for this reflection).

Even in the case of a crystal which diffracts isotropically, the determination of the effective resolution limit, and the related problem of the treatment of weak reflections, is problematic. In the anisotropic case it is even less clear how to proceed. There are problems in evaluating structure factor amplitudes $|F_o|$ from the observed intensities $|F_o|^2$ for the weak reflections (for general discussion of this problem see [232]). The Bayesian statistical approach of French and Wilson [127] cannot be used without modification, since it makes use of Wilson expectation values for intensities [128], with the implicit assumption that the diffraction is isotropic (see [453]). Of course a Bayesian approach could still be used, with an alternative probability distribution (one which reflects the anisotropy in the diffraction, and does not rely on Wilson statistics), but this remains to be implemented.

Table 4.1 Native BSDL data processing statistics

As discussed in the text, the statistics include only those reflections for which the profile-fitted intensity I > 0. A total of 73 observations were rejected as outliers and not included in the statistics.

	Upper resolution limit (Å)										
	7.54	5.98	5.23	4.75	4.40	4.14	3.94	3.77	3.63	3.50	ALL
No. of measured reflections	4986	4687	4298	3758	3194	2693	2062	1410	1285	1132	29505
No. of unique reflections	1039	974	950	911	849	857	732	543	528	505	7888
No. of reflections with I > 5σ(I)	1005	855	704	598	463	299	161	102	77	51	4315
Total possible observations	1055	986	972	969	950	973	941	950	941	945	9682
R-merge (%)	5.3	7.8	12.0	15.8	17.2	26.6	33.8	40.4	44.7	50.0	9.6

No attempt has been made to deal with this problem systematically at this stage. The data reduction statistics in Table 4.1 include all those reflections for which the profile-fitted intensity I > 0 (the program WEIS, in its current implementation, does not output negative intensities). Structure factor amplitudes have been obtained from the observed intensities by simply taking the square root of these quantities (an inadequate treatment for the weak reflections [232]). At some future stage the data needs to be reprocessed, without rejection of any measured reflections, and with the inherent anisotropy in diffraction properly treated. With

respect to Table 4.1, the very high R_{Merge} at high resolution reflects the fact that most of the reflections at these scattering angles are weak due to the anisotropic diffraction (only 51 observations between 3.63 and 3.50 Å resolution have $I > 5\sigma(I)$, which constitutes 5% of the possible observations in this resolution range).

4.2.3 Anisotropic diffraction

The degree of anisotropy in the measured intensities can be evaluated by calculating an overall ('cell averaged') anisotropic displacement parameter from the diffraction data. This problem has been considered several times in the literature, with the methods proposed being based on the use of Wilson statistics (see e.g. {454, 455]), or on analysis of the Patterson origin peak [230]. In both cases, the overall anisotropic displacement factor which is fitted to the data expands to

$$\exp\Big(-2\pi^2\Big(h^2U_{11}(a^*)^2 + k^2U_{22}(b^*)^2 + l^2U_{33}(c^*)^2 + 2hkU_{12}(a^*)(b^*) + 2hlU_{13}(a^*)(c^*) + 2klU_{23}(b^*)(c^*)\Big)\Big) \qquad \textit{Eq. 4.2}$$

Wilson expectation values for the diffracted intensity are known to deviate from experimental observation for protein data at low resolution (as a result of the violation of Wilsons assumptions). Since the diffraction data from the native BSDL crystals extends at best to 3.5 Å resolution it is not possible to make use of methods based on Wilson statistics, and analysis of the Patterson origin peak was employed.

The estimated coefficients are presented in Table 4.2. For comparison, results are also presented for diffraction data from crystals of *Alcaligenes denitrificans* cytochrome c′ [456]. These hexagonal crystals diffract to relatively high resolution without marked anisotropy. It should be noted that because these are 'cell averaged' values, there are restrictions on the coefficients which reflect the Laue class of the crystal. In addition the overall anisotropic displacement parameter, the equivalent isotropic displacement factor is reported, for which the general expression [457] is

$$U_{eq} = \frac{1}{3}\sum_{i=1}^{3}\sum_{j=1}^{3} U_{ij}a_i^*a_j^*a_ia_j \qquad \textit{Eq. 4.3}$$

For trigonal and hexagonal crystal systems (in the hexagonal setting) this expression simplifies to

$$U_{eq} = \frac{1}{3}\left(U_{33} + \frac{4}{3}(U_{11} + U_{22} - U_{12}) \right) \qquad \textit{Eq. 4.4}$$

For native BSDL the displacement parameters are small along the c* direction, relative to a* and b*, consistent with the anisotropic fall-off in diffraction. The value of U_{eq} (the mean square amplitude of displacement) is extremely large, reflecting the high degree of positional disorder in the native BSDL crystals (although *a priori* estimation of displacement parameters based on analysis of the Patterson origin peak has only recently been implemented, and it is not yet clear how reliable these estimates are). Individual crystals of native BSDL have been found to differ in the overall magnitude of the displacement parameters (results not shown), but qualitatively, all display the same behaviour.

Table 4.2 Overall anisotropic displacement parameters calculated for diffraction data from crystals of two proteins.

Restrictions on the elements of the displacement parameter tensor for trigonal and hexagonal crystal systems are $U_{11} = U_{22}$, $U_{12} = 1/2U_{11}$, $U_{13} = U_{23} = 0$. For the Patterson origin peak analysis, all observed data up to the specified resolution limit were used. For the least-squares calculation based on Wilson statistics (*A. denitrificans* cytochrome c´ only) data between 3.5 Å and the upper resolution limit were used. The *a priori* estimates obtained for U_{eq} for cytochrome c´ are consistent with the isotropic displacement parameters of the refined atomic model [456]. For BSDL, the atomic contents of the asymmetric unit were estimated from the protein sequence alone. Programs LEVY and ROGERS, used to perform the calculations, were provided by Dr. Robert Blessing (Medical Foundation of Buffalo, New York).

	Space-group	D_{min} (Å)		Anisotropic displacement parameters (Å^2)						
				U_{11}	U_{22}	U_{33}	U_{12}	U_{13}	U_{23}	U_{eq}
Human BSDL	$P3_121$ $P3_221$	5.0	Patterson analysis	4.35	4.35	1.81	2.18	-	-	3.50
A. denitrificans	$P6_522$	2.2	Wilson statistics	0.32	0.32	0.31	0.15	-	-	0.31
Cytochrome c´			Patterson analysis	0.39	0.39	0.45	0.19	-	-	0.41

It is not possible at this stage to provide a physical interpretation of these results. Anisotropic diffraction from protein crystals is not uncommon (see e.g. [458, 454, 459, 460, 461]), and in structurally characterized cases is often seen to be related to asymmetry in the molecular contacts which create the crystal lattice (see [454]).

It is clear that the native BSDL crystals characterized here exhibit a high degree of disorder. This is evident from both the anisotropic Bragg diffraction, which extends only to moderate resolution, and the marked diffuse scattering seen in the diffraction patterns (described

below). However, native BSDL is one of the few (if not the only) proteins crystallized with an intact, and heavily glycosylated tandem repetitive region. The structure of this region, and its relationship with the N-terminal catalytic domain are of considerable interest. If the structure of the N-terminal domain can be independently solved, then it may be possible to position this domain within the cell using standard molecular replacement procedures. Even if the electron density corresponding to the C-terminal repeat region cannot be resolved (and this seems likely), consideration of the packing of the N-terminal domain may still give indications as to the relative position and extent of the tandem repeat region, which would help in defining its likely biological role.

4.2.4 Diffuse scattering

Diffraction patterns collected from native BSDL crystals exhibit very marked diffuse scattering. That is, there is a clear background intensity distribution in addition to the discrete Bragg diffraction (see Figure 4.9). These effects are due to the breakdown of strict translational symmetry within the protein crystal, broadly termed *disorder* (see [462, 463] for reviews of diffuse scattering and its application to protein crystallography). This disorder can be both substitutional and positional; it is the latter that is usually of interest in protein crystallography. Diffuse scattering may arise from a large number of processes; intramolecular flexibility at various levels (from the displacement of individual side chains, to relative displacement of domains and subunits) or rigid body displacement of entire molecules. The interest in diffuse scattering arises because it can provide information on the correlations between the atomic displacements in a crystal. It should be noted that these displacements may be either static or dynamic, as both will give rise to the same apparent effects over the time scale of the X-ray diffraction experiment.

As a consequence, the observed diffuse scattering pattern results from the superposition of diffuse features having a number of different physical origins. The interpretation of experimentally observed diffuse scattering from protein crystals is neither straightforward nor routine, and requires an atomic model, which we do not yet have for native BSDL. So at this stage a detailed analysis of the diffuse scattering pattern is not possible.

Figure 4.10 shows one characteristic of the X-ray scattering by the native BSDL crystals, diffuse streaks which run parallel to c*. This is suggestive of a correlated rigid-body displacement of molecules in the direction of c in the crystal lattice [463]. Similar diffuse features arising from correlated rigid-body displacements were observed for the orthorhombic crystal form of hen egg white lysozyme [464]. Such displacements can only occur in directions characterized by strong contacts between neighbouring molecules. Taken together with the

anisotropic fall-off in the Bragg diffraction (which implies that the crystals are most ordered along c), these observations suggest that the crystals are characterized by a chain of strong interactions in this direction.

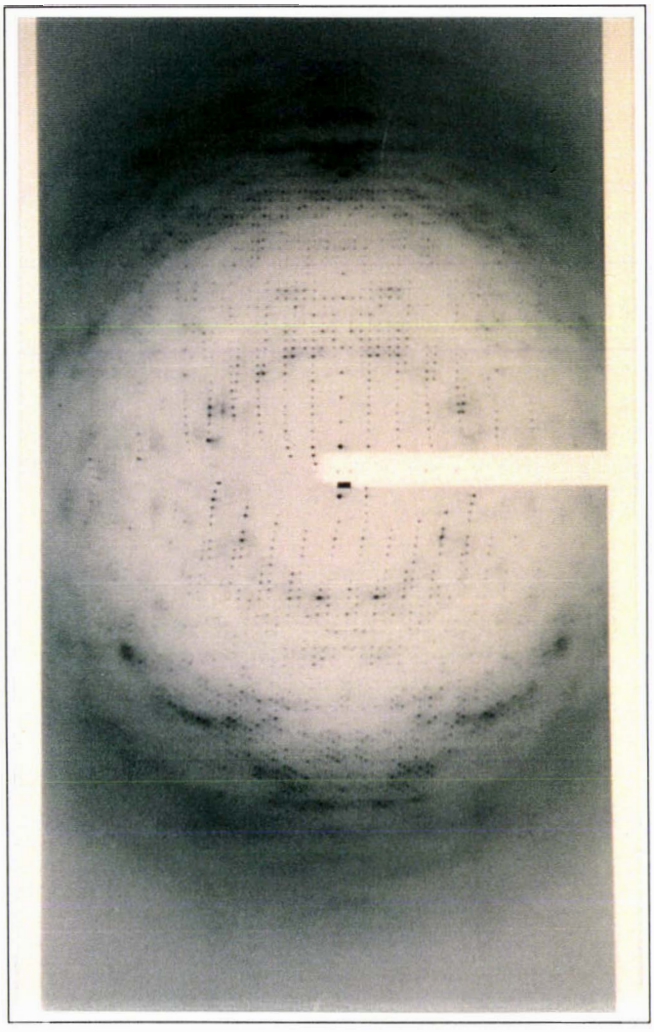


Figure 4.9 Diffuse scattering patterns from native BSDL crystals (I).

Diffraction pattern resulting from a native BSDL crystal, collected on a Weissenberg geometry camera, at the Photon factory, Tsukuba, Japan. The oscillation range was 8°, the coupling constant 2 °/mm, and the camera cassette radius 429 mm. The crystal was mounted with the unique axis c perpendicular to the rotation axis, which lies in the horizontal direction in the figure.

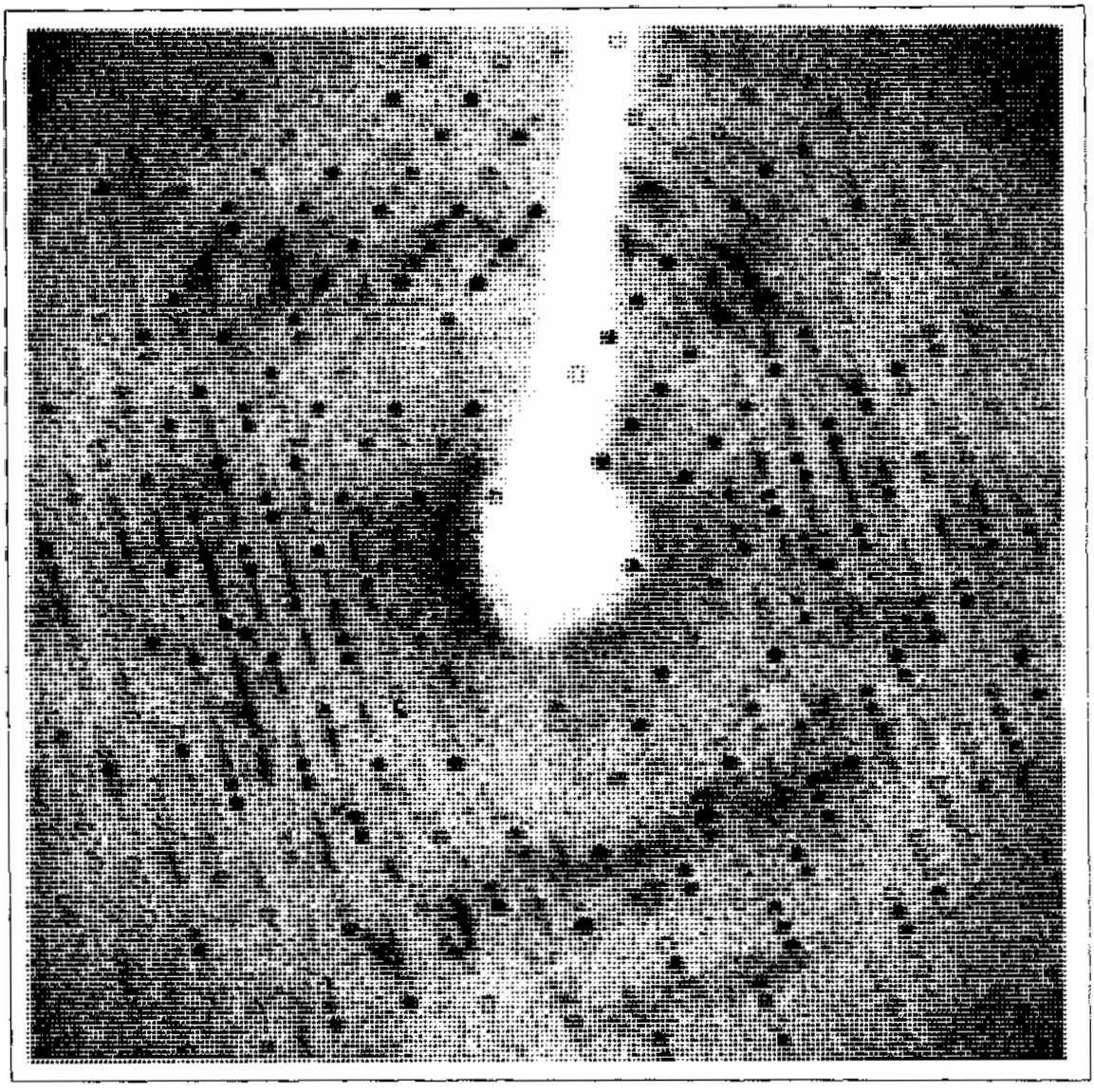


Figure 4.10 Diffuse scattering patterns from native BSDL crystals (II).

Part of the diffraction pattern resulting from a native BSDL crystal. Light grey circles show the predicted Bragg diffraction (calculated using the program DENZO [150], with effective mosaicity 0.3°). Diffuse streaks run parallel to c* (which is in the near vertical direction in the figure). The diffraction pattern results from a 2.5° oscillation of the crystal, and was collected on a Raxis IIC system (using CuKα radiation from a rotating anode generator, and a Fuji imaging plate as a detector).

Finally it should be noted that the diffuse scattering pattern of the native BSDL crystals may contain features due to substitutional as well as positional disorder, since the glycosylation of the C-terminal repeat region is known to be heterogeneous.

4.2.5 Enzymatic deglycosylation

Human BSDL, in common with many secretory proteins, is heavily glycosylated (Section 4.1.3.9). The heterogeneous nature of protein-associated oligosaccharides often hinders crystallization attempts [35]. This problem may contribute to the disorder apparent in the crystals of the native protein. Therefore a strategy for partial enzymatic deglycosylation of BSDL was investigated.

We were interested at first in modification of the O-linked oligosaccharides of the tandem-repetitive region, which comprise the bulk of the carbohydrate in BSDL. Removal of the O-linked oligosaccharides, by sequential treatment with sialidase (to remove terminal sialic acid residues) and O-glycosidase, was used successfully in the crystallization of the T Cell coreceptor CD8 [465]. However in the case of BSDL, it seemed that total removal of the O-linked oligosaccharides would be undesirable, because they are an integral part of the structure of the tandem repeat region (see Section 4.1.3.10 for discussion).

There was already strong evidence that BSDL-associated oligosaccharides contained sialic acid residues. In 1988, the isolation of a lipoamidase from human milk was reported [466], an activity that was subsequently shown to be associated with BSDL [346]. In the first paper it was shown that treatment of lipoamidase with sialidase shifted the isoelectric point of the protein. As an initial step toward obtaining a protein more amenable to crystallization, the removal of any terminating sialic acid residues was therefore attempted.

4.2.5.1 Experimental methods

Sialidase (*Clostridium perfringens*) was purchased from Boehringer Mannheim. Lypohilized human milk BSDL was dissolved in 100 mM Potassium acetate buffer (pH 5.5), containing 10 mM Na_2EDTA and 0.02% (w/v) sodium azide. Final BSDL concentration was 25 mg/ml. Sialidase, dissolved in the same buffer was added (0.04 mg sialidase per 1 mg BSDL), and the proteins incubated for 24 hours. The desialidation procedure was monitored using isoelectric focussing (IEF) [467]. Enzymatic desialidation was followed by gel filtration to separate BSDL from sialidase and sialic acid. After the incubation period, the solution was filtered, and passed down a Superdex 75 column (Pharmacia). The collected BSDL was transferred into water by repeated concentration and dilution using Amicon microconcentrators (30 000 Mw cutoff). The final protein concentration was 8 mg/ml (by absorption at 280 nm); this material was used without further modification in crystallization trials. IEF was performed using a PhastSystem (Pharmacia), and PhastGel IEF media (pH 3-9) (precast

polyacrylamide gels containing carrier ampholytes). The silver staining technique was used to detect proteins in the polyacrylamide gels [468].

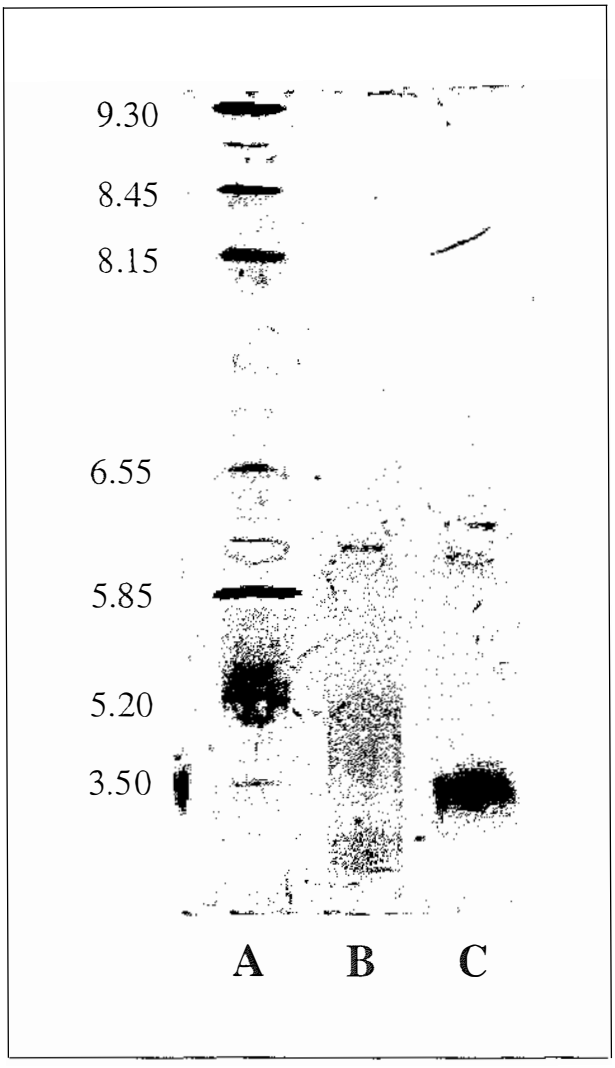


Figure 4.11 Isoelectric focusing of BSDL

Lane (A) contains protein markers of known isoelectric point (as indicated). Lane (B) contains the unmodified native BSDL. Lane (C) contains BSDL after treatment with sialidase. Analytical IEF using the PhastSystem was carried out according to the manufacturer's instructions.

4.2.5.2 Results of desialidation

In Figure 4.11 typical results are shown (as monitored by analytical IEF). It can be seen that before desialidation, the protein does not focus as a single band, indicating charge heterogeneity. Following incubation with sialidase, BSDL focuses as a discrete band, with an isoelectric point (pI) of 3 - 4 (c.f. a pI of 4.6 reported for the desialidated lipoamidase (BSDL) from human milk [466], and a pI of 5.1 calculated from the protein sequence alone [469]). This

provides direct proof that the protein is heterogeneously glycosylated, and that the charge heterogeneity is due to the presence of terminal sialic acid residues. Support for this conclusion comes from later studies which have detected (by immunochemical techniques) the occurrence of sialidation, and suggest heterogeneity among the oligosaccharides of the C-terminal repeat region [423].

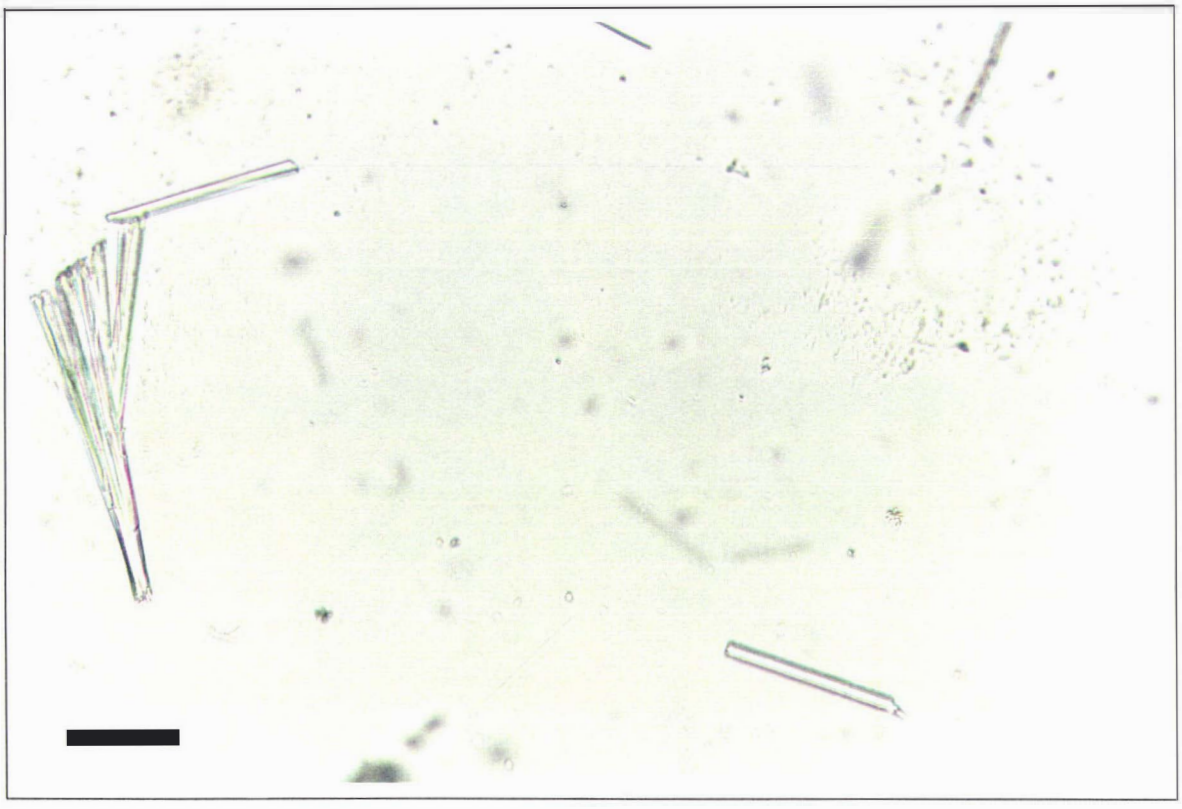


Figure 4.12 Crystals of desialidated BSDL.

The crystals shown were grown by hanging-drop vapour diffusion techniques, with a reservoir solution 14%(w/v) PEG 6000, 0.2M Bis-tris propane/HCl buffer, pH 6.7. The thin black bar represents 0.1 mm

4.2.5.3 Crystallization of the desialidated protein

Crystallization trials on the desialidated protein were promising. Again, the experiment detailed in Table 1.4 was employed (using polyethylene glycols), in combination with the hanging drop vapour diffusion technique. Very thin rod-like crystals grew readily over the pH range 6.1 - 8.5, appearing in some cases after several weeks. These crystals have a completely different morphology to the crystals of the native protein (see Figure 4.12). However, they are very small, and no X-ray diffraction studies have yet been carried out.

As this work was in progress, recombinant forms of BSDL (expressed in a mammalian cell line) became available through a collaboration with Dr Kerry Loomes (University of Auckland). Because of the promising results obtained using these recombinant proteins, the partial enzymatic deglycosylation of native BSDL, and characterization of the resulting crystals, has not been further developed. However this work should not be abandoned. If the crystals of the desialidated protein can be grown larger (and diffract usefully) they have the potential to provide data on the intact protein, with a nearly native pattern of glycosylation in the C-terminal repeat region. This becomes of greater relevance now that real progress has been made toward the structure determination of the N-terminal catalytic domain, described in Section 4.4

4.3 RECOMBINANT FULL LENGTH BSDL

The first recombinant protein to be studied possessed the entire amino acid sequence of the native protein, including the C-terminal repeat region. However because glycosylation is species- and cell-specific [412, 414, 413] it is certain that the oligosaccharide structures associated with the recombinant and native proteins will differ (the recombinant protein is produced by transformed Baby Hamster Kidney (BHK) cells [424]). Therefore, the recombinant full length BSDL presents a different (and potentially more tractable) crystallization problem. However, because of the likely structural role of the O-linked oligosaccharides in the C-terminal repeat region, it is difficult to know if the biological integrity of this structure will be maintained.

4.3.1 Expression, purification and crystallization

The expression and purification of full-length recombinant BSDL was carried out by Dr Kerry Loomes (University of Auckland, N.Z.), as previously described [424]. The protein had been concentrated to between 12 and 15 mg/ml in 20 mM MOPS/NH_4OH buffer (pH 7.0). This material was used without further modification in crystallization trials. Only 60 µL of the protein solution was initially available. Part of the search experiment based on polyethylene glycols (Table 1.4) was executed (36 of the 64 trials), omitting the highest and lowest pH levels and also the lowest PEG concentration in order to minimize the number of trials. Over a time period of several weeks to several months, crystals appeared over the pH range 6.1 - 8.5. These had either a thin 'needle-like' appearance or a more rectangular 'plate-like' shape. The crystallization of full-length recombinant BSDL has subsequently proved to be frustratingly irreproducible. Crystals can be induced to grow more readily if a seed crystal (from a previous experiment) is introduced into the hanging drop prior to equilibration. Usually the seed crystal does not grow further, but presumably the introduction of crystal nuclei during the transfer process initiates the growth of other crystals. Typical crystals

obtained by this method (having the 'plate-like' morphology) are shown in Figure 4.13. Only a single crystal of suitable size for X-ray diffraction has been obtained (again having the 'plate-like' morphology), so it is not yet known whether the two observed morphologies represent distinct packing arrangements of the molecules.

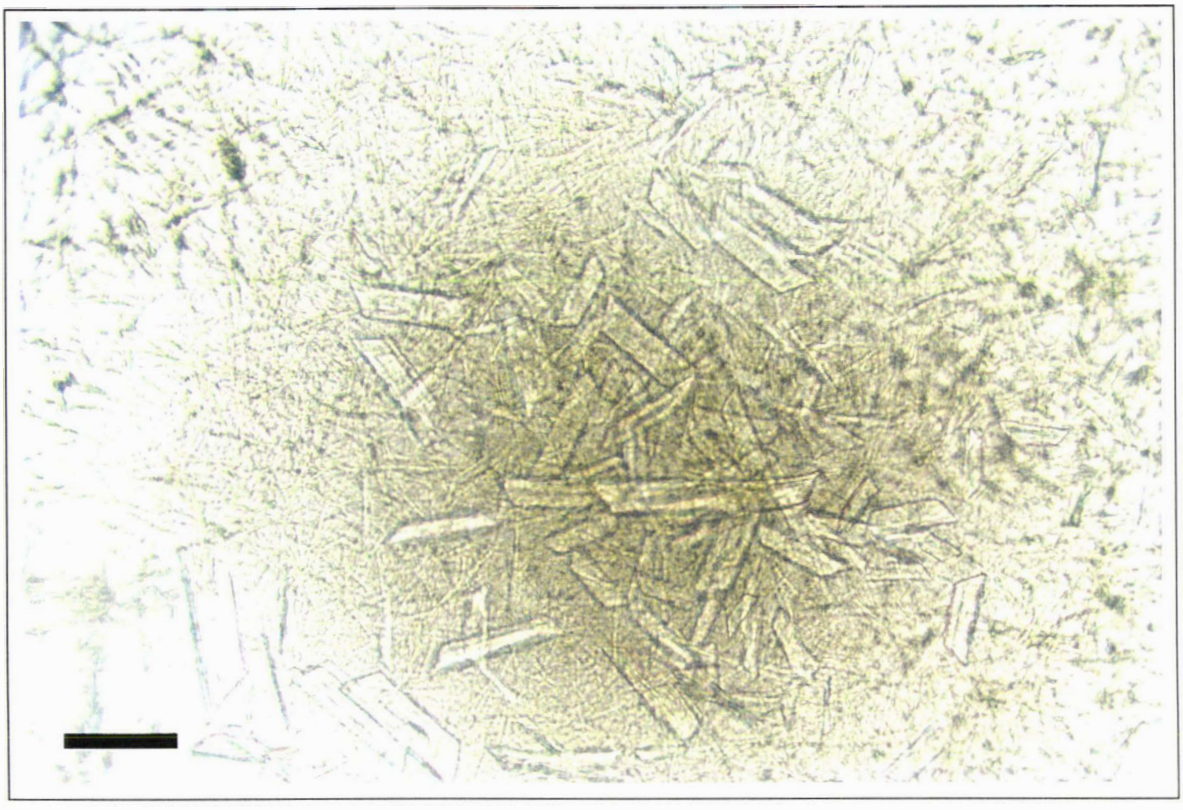


Figure 4.13 Crystals of full-length recombinant BSDL

The crystals shown were grown by hanging drop vapour diffusion, with a reservoir solution containing 16%(w/v) PEG-mme 5000,and 0.20 M HEPES/KOH buffer (pH 7.30). A single seed crystal was included in the hanging drop to initiate crystal growth. The thin black bar represents 0.1 mm

4.3.2 Preliminary crystallographic investigation

The single crystal of useful size for X-ray diffraction was grown directly from a solution containing 14%(w/v) PEG 6000, and 0.20 M MES/KOH buffer (pH 6.1), and resembled in appearance those in Figure 4.13. It was mounted conventionally in a thin-walled glass capillary, and preliminary crystallographic investigations were carried out using the Raxis IIC system. Unfortunately problems with water vapour condensation on the face of the detector (caused by the crystal cooling system at the time) resulted in severe absorption effects, and the data could not be usefully merged. This remains the only crystal of the recombinant full length protein that has been of suitable size for X-ray diffraction.

The diffraction pattern could be indexed on a primitive lattice, with approximate cell dimensions a = 90.5 Å, b = 145.8 Å, c = 115.8 Å, $\alpha = 90.0°$, $\beta = 103.3°$, $\gamma = 90.0°$. The metric symmetry of this lattice suggests the crystals will be monoclinic (Space group P2 or $P2_1$), however further data needs to be collected to confirm this. The crystal appeared to diffract usefully to about 3.3 Å. Unlike the crystals of the native full length protein, the diffraction did not appear to be markedly anisotropic.

4.4 RECOMBINANT TRUNCATED BSDL

The difficulties created by the presence of a glycosylated tandem repeat region following the N-terminal catalytic domain of BSDL, suggested the study of a truncated variant lacking this region. Several laboratories have reported the expression of such recombinant variants in mammalian or bacterial cells [424, 417, 418]

4.4.1 Expression and purification

The expression and purification of the recombinant truncated variant was carried out by Dr Kerry Loomes (University of Auckland, N.Z.). The truncated variant (which comprises residues Ala 1 - Phe 518, plus four additional amino acids at the C-terminus) was expressed in a mammalian cell line as previously described [424]. The purification procedure has now been further developed, and is given here for completeness.

The first step in the purification involves Heparin-sepharose chromatography (Hep-Pac cartridge, Pharmacia). The optimum cell culture medium was centrifuged at 10,000 rpm in a SS34 rotor to remove debris, and diluted with water by a factor of two. This medium was then loaded directly onto a Hep-Pac cartridge, and fractions corresponding to the truncated variant were eluted with a NaCl gradient in 10 mM sodium phosphate buffer, pH 7.6.

This was followed by Gel filtration. At this stage CHAPS (3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate) (0.8 mM final concentration) was added to prevent absorptive losses during concentration. The volume was then reduced using Amicon concentrators (30 000 Mw cut-off), and the sample loaded on to a Superose-12 column (Pharmacia) equilibrated with 20 mM MOPS/NH_4OH, pH 7.0, + 0.1 M NaCl + 0.8 mM CHAPS. This step served to both purify the protein, and to exchange the solution buffer. Salt was included at this stage to prevent the protein adhering to the column, which occurs in its absence.

Finally, the salt was removed by repeated concentration and dilution with 20 mM MOPS/ NH_4OH, pH 7.0, + 0.8 mM CHAPS without NaCl (again using Amicon concentrators). The final protein concentration was between 0.5 - 0.8 mg/ml (determined by the Bradford dye-binding assay [470], using bovine serum albumin as a standard). Considerable care had to be taken in the final concentration steps, as the enzyme would irreversibly precipitate if the volume was reduced by too great an amount.

4.4.2 Crystallization

Because of initial difficulties with expression and purification, the quantity of the truncated variant initially available was very small. Despite its low concentration (0.5 - 0.8 mg/ml), it was known that the protein solution must be nearly saturated, and it was therefore used at this concentration in crystallization experiments. Using the hanging drop vapour diffusion technique, the experiment described in Table 1.4 (using polyethylene glycols as protein precipitants) was executed. Crystals grew over several days from solutions containing PEG 6000 and PEG-mme 5000, buffered at pH 5.5 - 9.1, with no apparent buffer dependency.

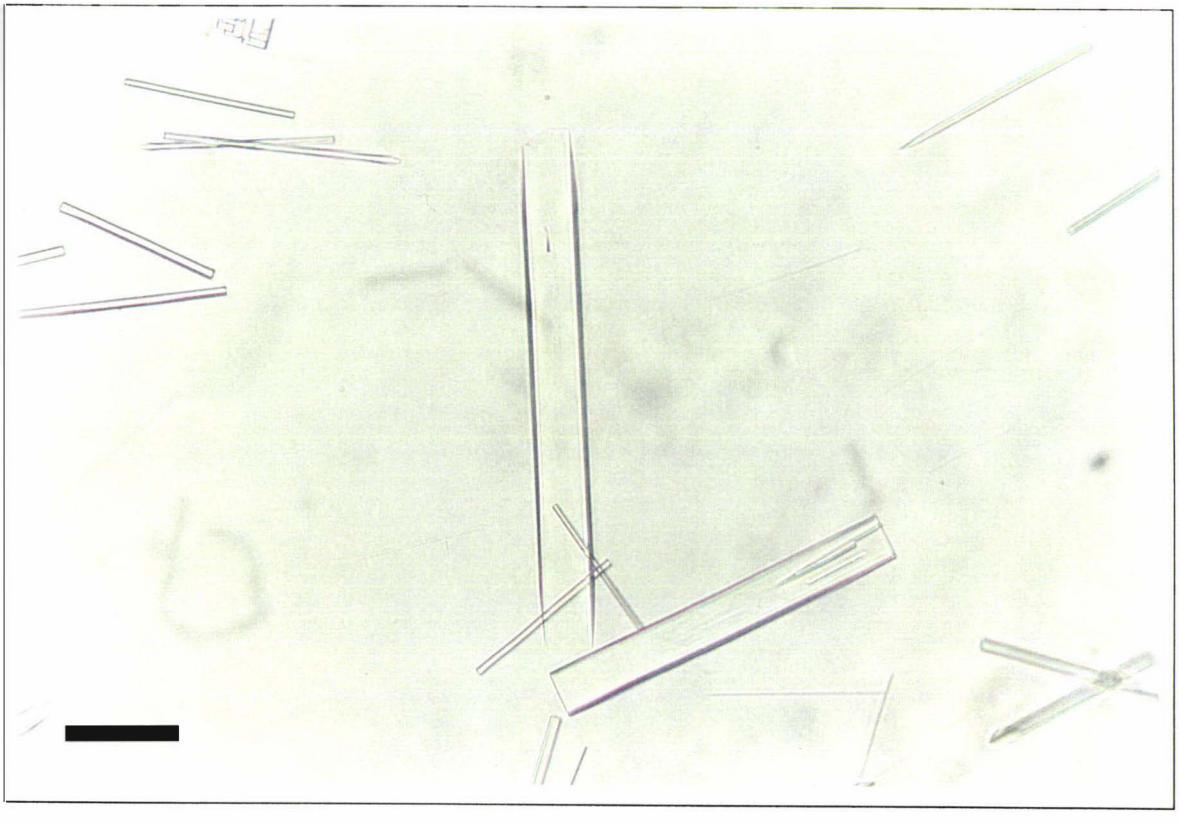
The best crystals have been grown by hanging drop vapour diffusion, from solutions buffered with PIPES/KOH at pH 6.7, with PEG 6000 concentrations of 15 - 25% (depending on protein concentration). It has not proved possible to obtain crystals of suitable size for X-ray diffraction studies without repeat seeding of the crystals [26]. In this procedure, individual crystals were isolated from a hanging drop, transferred into a fresh protein/precipitant solution, and again equilibrated against a precipitant solution with concentration sufficient to push the system into the supersaturated state. Addition of small quantities (2% (v/v)) of glycerol to the precipitant solution improved the outward appearance of the crystals. The seeding procedure was repeated 2 to 3 times in order to get crystals of sufficient size for X-ray diffraction (Figure 4.14). Provided the crystals used in seeding had been recently grown, it did not seem to be necessary to perform a washing step (to partially redissolve the crystals).

Figure 4.14 (following page) Crystals of truncated recombinant BSDL.

(A) Typical crystals obtained directly by hanging drop vapour diffusion. (B) Crystals after repeat seeding procedures. Both photographs are on the same scale. The thin black bar represents 0.1 mm.

A

B



Even after seeding the crystals remain very small (typically less than 0.1 mm in two of their dimensions, and 0.01 - 0.03 mm in their smallest dimension). This borders on the usable limit for X-ray diffraction studies. However the crystals can be seeded only a finite number of times before they begin to incorporate visible growth defects.

The assistance of Mrs. Heather Baker in growing crystals of the recombinant truncated variant during the latter stages of this project is gratefully acknowledged.

4.4.3 Data collection and processing

Three native data sets have been used in the structure determination to date. Each data set represents an improvement over the previous in terms of effective resolution, measurement redundancy and completeness. These improvements reflect both an increasing proficiency at growing and seeding the crystals, and also the use of better data collection methods. The experimental details associated with each data collection are given in Table 4.3, and the data processing statistics in Tables 4.4 - 4.6. What follows is a general discussion of the data collection methods.

Table 4.3 Recombinant truncated BSDL; experimental details of X-ray data collection.

Data set	I	II	III
Data collection method	Oscillation	Weissenberg	Oscillation
X-ray source	CuKα radiation Rotating anode generator	Synchrotron (Tsukuba, Japan)	CuKα radiation Rotating anode generator
X-ray detector	Fuji imaging plate	Fuji imaging plate	Fuji imaging plate
Temperature	279 K (6 °C)	277 K (4 °C)	113 K (-160 °C)
Crystal mounting method	conventional capillary mount	wet mount (Section 2.4.2.2)	solid-surface mount (Section 4.4.3.4)
Oscillation range	2°	3° (coupling constant 2.5 °/mm)	1°
Crystal to detector distance	90 mm (air path)	287 mm (helium-filled path)	110 mm (air path)
X-ray beam collimation	0.3 mm pinhole collimation	see [292] for details	0.1 mm pinhole collimation
Exposure time per image	50 minutes	24 seconds	30 minutes
Number of crystals	1	2	1

4.4.3.1 Data processing methods

Diffraction data from crystals of recombinant truncated BSDL were collected both on the laboratory X-ray data collection system, and at a synchrotron radiation source (refer to Table 4.3). Profile-fitted relative intensities were obtained using the program DENZO [125]. Relative scaling of the intensities was performed using the algorithm of Fox and Holmes [126]. Structure factor amplitudes were obtained from the intensity measurements employing the Bayesian treatment of French and Wilson [127]. This procedure also serves to put the data on an approximately absolute scale via a conventional Wilson plot [128]. The scaling and merging of data was carried out using the CCP4 program suite [130].

4.4.3.2 Space Group determination.

X-ray diffraction data could be indexed on a primitive lattice with approximate cell dimensions, a = 59.3 Å, b = 90.0 Å, c = 107.7 Å, $\alpha = \beta = \gamma = 90°$. Examination of the observed diffraction patterns revealed clear *mmm* symmetry, hence the crystals belong to the orthorhombic crystal system. There were systematic absences (for all odd indices) along two of the axial directions. In the third direction (corresponding cell axis 108 Å), systematic absences could not be determined because of the incompleteness of the data that were initially collected. This restricted the possible space groups to $P2_12_12$ or $P2_12_12_1$. Subsequently, the solution of the structure by molecular replacement, and the collection of further data, have confirmed the space group as $P2_12_12_1$. With respect to the cell dimensions, the z-axis cell dimension shrinks markedly when the crystals are flash frozen at liquid nitrogen temperatures (Section 4.4.3.4) (the other cell dimensions appear relatively invariant).

Table 4.4 Data processing statistics: data set (I)

28 images were used in data processing, collected from a single crystal. Data were obtained by profile fitting of intensities out to a fixed isotropic resolution limit of 3.5 Å on each image. Approximate cell dimensions a = 59.2Å, b = 90.0 Å, c = 107.7 Å, $\alpha = \beta = \gamma = 90.0°$.

	Upper resolution limit (Å)										
	7.54	5.98	5.23	4.75	4.40	4.14	3.94	3.77	3.63	3.50	ALL
No. of measured reflections	1518	1473	1491	1437	1502	1350	1435	1346	1264	1334	14150
No. of unique reflections	603	590	606	591	614	591	610	614	612	615	6036
Completeness (%)	70	75	76	78	78	79	80	80	81	81	78
$\langle I/\sigma(I)\rangle$	8.1	5.9	5.0	4.4	4.5	3.7	3.3	2.9	2.5	2.3	3.9
R-merge (%)	7.1	12.2	14.4	15.4	16.0	18.7	21.1	24.5	28.0	32.3	17.5

Table 4.5 Data processing statistics: data set (II)

29 images were used in data processing, collected from a total of two crystals. Data were obtained by profile fitting of intensities out to a fixed isotropic resolution limit of 3.5 Å on each image. Approximate cell dimensions a = 59.0 Å, b = 89.5 Å, c = 106.7 Å, $\alpha = \beta = \gamma = 90.0°$.

	Upper resolution limit (Å)										
	7.54	5.98	5.23	4.75	4.40	4.14	3.94	3.77	3.63	3.50	ALL
No. of measured reflections	2381	2394	2477	2384	2338	2400	2338	2419	2303	2238	23672
No. of unique reflections	715	689	697	690	683	688	693	708	678	686	6927
Completeness (%)	85	89	91	91	92	92	94	94	93	94	91
$\langle I/\sigma(I)\rangle$	15.2	7.4	6.3	6.3	6.3	4.5	4.1	3.1	2.6	2.2	4.8
R-merge (%)	4.5	10.1	12.0	12.0	11.9	16.8	18.6	24.2	29.1	34.6	15.1

Table 4.6 Data processing statistics: data set (III)

Data were obtained by profile fitting of intensities from 88 images, collected from a single crystal. On each image, data were integrated to a resolution limit at which the mean $I/\sigma(I)$ in a thin isotropic resolution shell fell below 2. Approximate cell dimensions a = 58.8 Å, b = 90.2 Å, c = 103.6 Å, $\alpha = \beta = \gamma = 90.0°$.

	Upper resolution limit (Å)										
	5.60	4.45	3.88	3.53	3.28	3.08	2.93	2.80	2.69	2.60	ALL
No. of measured reflections	6216	6190	6193	6155	5972	4627	4132	3865	3596	3101	50047
No. of unique reflections	1846	1723	1711	1690	1669	1534	1462	1410	1352	1253	15650
Completeness (%)	97	97	98	97	97	89	85	82	79	66	89
$\langle I/\sigma(I)\rangle$	14.2	10.4	8.1	5.5	4.1	3.8	3.3	2.9	2.3	2.0	6.3
R-merge (%)	4.6	6.5	8.2	12.0	16.6	18.5	22.1	25.2	31.6	34.9	10.4

4.4.3.3 Data collection strategy

It became evident that due to the extremely small size of the crystals special care was needed with experimental techniques. In the interests of the maximum peak to background ratio, it is important that the primary X-ray beam diameter is not larger than the crystal [471]. The truncated BSDL crystals are in their smallest dimension typically < 0.05 mm. The Raxis IIC data collection system [124], is equipped with conventional monochromator-collimator X-ray optics. At the time of the initial data collection, the finest pinhole collimator available was 0.3 mm in diameter, far exceeding in size the smallest dimensions of the crystals. Subsequently a 0.1 mm diameter collimator was purchased (Molecular Structure Corporation, Texas, USA), which was used in the collection of data set III.

Another simple consideration is that the signal-to-noise ratio in the oscillation method is inversely proportional to the rotation range (i.e. larger oscillations result in a lower signal to noise ratio) [472]. This is because at any one instant, only a very small area of the detector is receiving useful information (i.e. Bragg diffraction). This must be balanced against the fact that smaller oscillations will result in a larger number of partially recorded Bragg reflections on each image. While improvements in the processing of partially recorded reflections have been made, summation of such reflections is still susceptible to the introduction of systematic errors [see [473, 474] for discussion). In the collection of data set I, 2° oscillations were employed (clearly too large), whereas in the collection of data set III, 1° oscillations were used.

Data set II was not collected using a laboratory X-ray source, but at the Photon factory (Tsukuba, Japan) where a synchrotron radiation source is coupled with a Weissenberg camera [292]. In principle, the high intensity synchrotron X-radiation source should have allowed the collection of a high quality data set, but for a number of reasons this was not effectively realized. The principal advantages of Weissenberg camera geometry are operational; a large angular range can be recorded on each image, which avoids producing a significant number of partially recorded reflections [475]. The wide angular range produces a larger background over an individual diffraction spot than would be the case if it was measured only over its rocking width; in practice this is compensated for by the extremely large crystal to detector distances used (see [292, 476] for discussion).

Because of their very small size, it proved at first to be extremely difficult to conventionally mount the crystals in glass capillaries without damaging them. For this reason, the crystals taken to the Photon factory were mounted in liquid-filled capillaries, as was described for GFOR (Section 2.4.2.2). Unfortunately, because of their small volume, mounting the crystals in this fashion resulted in an extremely high background scatter from the surrounding solvent. It also made the crystals difficult to center and align in the X-ray beam because of the poor optical properties of the liquid-filled capillaries. Retrospectively, mounting the crystals in this fashion was a mistake. Relatively small oscillations (3°) were employed during the data collection in an attempt to minimize the background scatter. Even so, the effective resolution of the data set collected was only 3.5 Å. Subsequently it was found that it was possible to mount the crystals conventionally with reasonable ease, provided that capillaries were not siliconized, and were of very small (0.1 - 0.2 mm) diameter.

4.4.3.4 Cryocrystallography (data set III)

The primary advantage of data collection at liquid nitrogen temperatures, is the slowing or prevention of radiation damage to the sample [477, 132]. Once sufficient crystals became available, a nitrogen-gas-stream cooling device [142] (Oxford Cryosystems) operating at 113 K, was used to flash-freeze the crystals, and to maintain this state during data collection.

Several mounting techniques have been developed in order to realize the effective and rapid freezing of protein crystals. In the most widespread technique, the crystal is suspended in a thin film, formed within a small loop [478]. The loop can be made from a number of natural and synthetic fibers (see e.g. [479]). In another technique, the crystals are suspended on a thin glass or quartz support (see e.g. [480, 481]. In both cases, the mounted crystal is then quickly transferred into the cold (80 - 120 K) nitrogen gas stream. In order to prevent ice nucleation, the crystals must be transferred into a solution containing a cryoprotectant prior to flash-freezing [482] (or other suitable measures must be taken [483]).

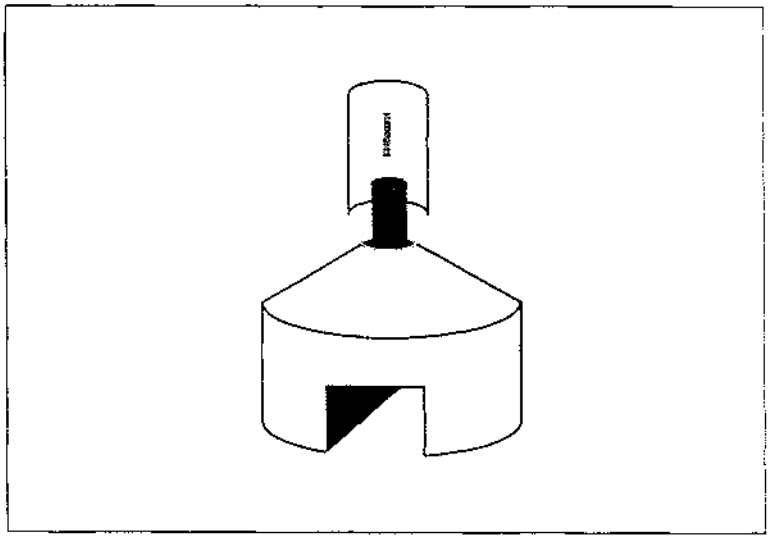


Figure 4.15 Glass-slide mounting device for cryocrystallography

A schematic illustration of the mounting device used in the collection of data set (III) from the truncated recombinant BSDL crystals. The crystal, after soaking in a cryoprotectant solution, can be transferred directly to the glass surface, and the excess liquid removed using a pasteur pipette which has been drawn out in a flame.

When using the loop technique, minimizing the thickness of the free-standing film was an important consideration, because of the very small size of the crystals. It was found in practice that it was very difficult to remove sufficient liquid from the loop prior to flash-freezing, and that it was also difficult to align the very small crystals within the loop. We therefore

found the second technique preferable. In practice, a fragment of a conventional capillary tube provides a suitable rigid surface. This fragment can be glued to a metal pin and attached to a magnetic base capable of being quickly transferred to the goniometer head (Figure 4.15). It is relatively easy to remove excess liquid prior to flash freezing.

In order to prevent ice nucleation, the crystals were introduced to a new mother liquor containing a cryoprotective agent prior to freezing. Crystals were transferred directly to a solution containing 0.2M PIPES/KOH (pH 6.7), 25% PEG 6000 and 25% Glycerol. The soak time in the cryoprotectant solution before flash freezing was typically short (5 minutes). Damage to protein crystals is commonly associated with flash freezing [477, 484] (this is often evidenced by an increase in crystal mosaicity). This damage can often be minimized by careful choice of the cryoprotectant and its concentration. No studies of this type have been carried out, thus the conditions for the flash-freezing of the crystals reported here can almost certainly be improved.

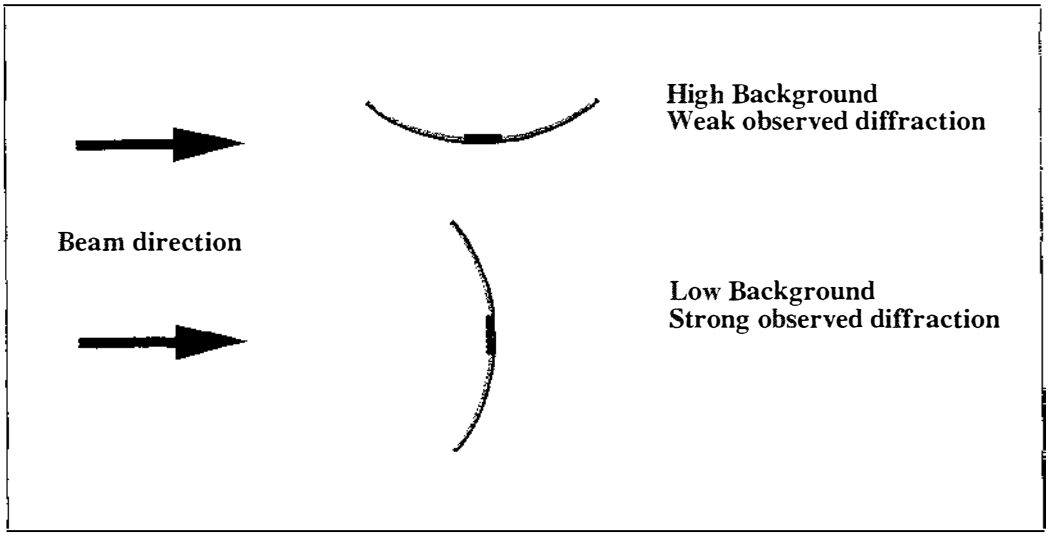


Figure 4.16 Background scatter and absorption due to the solid-surface mount.

Schematic diagram of the favoured and disfavoured directions for cryocrystallographic data collection using a solid-surface mounting device. These effects are only likely to be significant for data collection with very small crystals.

The drawback of the solid-surface mounting technique was that background scatter and absorption from the capillary (and the liquid which coats its surface) was much more marked in some directions than in others (Figure 4.16). These effects are no more severe than are encountered in the mounting of crystals conventionally in sealed capillaries, but assume much greater significance because of the small size of the crystals. In practice this meant that the effective resolution limit in the least favoured direction was ~ 3.3 Å, while that in the

most favoured direction was ~ 2.6 Å. This effect was not due to any inherent anisotropy in diffraction from the crystals (see Section 4.2.3), since it was not consistent with the symmetry of the diffraction pattern. These effects could be minimized by using the smallest possible solid surface as a support for the crystal; in practice this creates difficulties in placing the crystal and removing the surrounding liquid. Overall however, this mounting procedure, coupled with a more finely collimated beam, larger crystals, and data collection at liquid nitrogen temperatures, allowed the collection of an essentially complete data set which extends usefully to 2.7 Å resolution (Table 4.6). This represents a vast improvement over what had been previously achieved. Data collection from such small protein crystals is not unprecedented (see e.g. [485, 486]), however it is apparent that careful attention to experimental technique is needed to collect high resolution data from these crystals.

4.4.3.5 Data collection from bile acid-soaked crystals

Some preliminary experiments have been performed in an attempt to diffuse bile acids into preformed crystals. Inspection of the molecular replacement solution (discussed below) shows that the face of the molecule which contains the active site (and which is expected to be involved in bile acid binding) is not involved in crystal contacts, so such experiments would seem to be worthwhile. Crystals can be soaked in solutions containing bile acids (25% (w/v) PEG 6000, 0.25 M PIPES/KOH (pH 6.7), 25 mM cholic acid) without visible damage. A data set has been collected (at cryogenic temperature) from a crystal which had been soaked in the above solution for 150 minutes (results not shown). While diffusion of small molecules into protein crystals seems to be relatively rapid [487], it has not yet been established whether the bile acid has bound to the protein or if CHAPS (the detergent used in protein solubilization, and a bile acid analogue) will competitively bind to the protein (see [421]).

4.4.4 Structure solution by molecular replacement

Even the first, relatively poor, data set collected (I), was sufficient to demonstrate the feasibility of using molecular replacement to solve the phase problem for recombinant truncated BSDL. This was later verified using the second data set collected (II), and the statistics of the rotation and translation functions reported below relate to these data. It should be emphasized that the third, higher resolution data set (III) was collected very recently. This thesis will describe only the initial stages of the structure determination, carried out using the second data set (and therefore at an effective resolution of 3.5 Å).

The refined atomic coordinates of TcAChE [404] were used as a search model for molecular

replacement, employing the entire polypeptide chain with all side chains truncated to the Cβ position. A real-space cross-rotation function [488] was evaluated over its asymmetric unit [489], with the program X-plor [224] (using native data set II; diffraction terms between 8 and 4 Å resolution; inner and outer integration radii for the Patterson functions of 5 and 30 Å respectively; and the pseudo-orthogonal Eulerian angle system of Lattman [490] to define the search grid). The highest peaks of the rotation function were subject to Patterson correlation refinement [491]. Here the standard linear correlation coefficient

$$PC(\Omega) = \frac{\langle |E_{obs}|^2 |E_m(\Omega)|^2 - \langle |E_{obs}|^2 \rangle \langle |E_m(\Omega)|^2 \rangle \rangle}{\sqrt{\langle |E_{obs}|^4 - \langle |E_{obs}|^2 \rangle^2 \rangle \langle |E_m(\Omega)|^4 - \langle |E_m(\Omega)|^2 \rangle^2 \rangle}} \qquad \textit{Eq. 4.5}$$

is maximized as a function of the molecular orientation Ω. E_{obs} denotes the normalized observed structure factors, and $E_m(\Omega)$ the normalized calculated structure factors of the search model oriented according to Ω, and placed in an arbitrary position in the triclinic spacegroup *P1*, with a cell possessing the same metric symmetry as that of the real crystal. The angle brackets denote an average over the observed reflections expanded to space group *P1*. Fujinaga and Read [492] have discussed the equivalence of Equation 4.5 and the correlation coefficient between Patterson functions. This procedure serves both to discriminate between correct and incorrect orientations of the search model, and to improve the accuracy of the rotational operator [493].

Following Patterson correlation refinement, a single orientation of the search model was clearly distinguished from the other peaks of the rotation function (Figure 4.17), with a maximum value for the correlation coefficient of 0.069 (mean 0.029, standard deviation 0.011, for the top 235 peaks of the rotation function).

Using the oriented molecule, a translational search was performed, computing the conventional correlation coefficient between calculated and observed structure factor amplitudes (normalized), as a function of the position of the molecule in the unit cell. The function evaluated is

$$CC(\Omega, T) = \frac{\langle |E_{obs}|^2 |E_{calc}(\Omega, T)|^2 - \langle |E_{obs}|^2 \rangle \langle |E_{calc}(\Omega, T)|^2 \rangle \rangle}{\sqrt{\langle |E_{obs}|^4 - \langle |E_{obs}|^2 \rangle^2 \rangle \langle |E_{calc}(\Omega, T)|^4 - \langle |E_{calc}(\Omega, T)|^2 \rangle^2 \rangle}} \qquad \textit{Eq. 4.6}$$

This is a very similar function to that used in Patterson correlation refinement (Equation 4.5),

except that the averaging operations are now performed over an asymmetric unit of diffraction data, and $E_m(\Omega)$ has been replaced by $E_{calc}(\Omega, T)$ (a function of both the orientation Ω, and position T, of the search model), which denotes the normalized structure factors calculated from the search model and all its symmetry-related molecules.

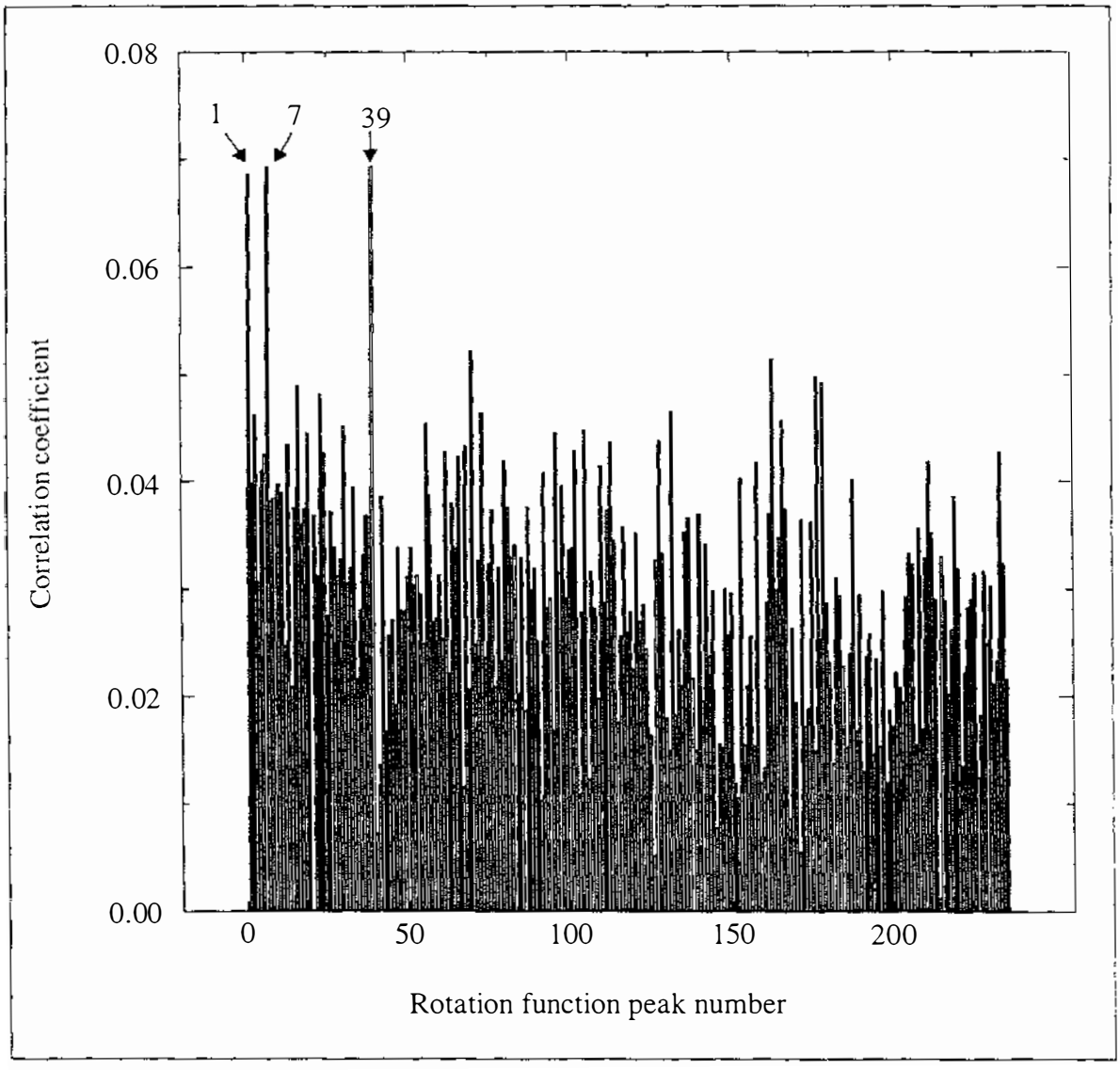


Figure 4.17 Results of Patterson correlation refinement

Utilizing the top 235 peaks of the real-space cross-rotation function, the oriented (but unpositioned) search models were subject to Patterson correlation refinement [491]. Three peaks (labelled 1, 7, and 39) were clearly differentiated from the remainder. These three peaks converge to the same rotation (θ_1=321°, θ_2=55°, θ_3=197°) (Eulerian angles according to the convention of Rossmann and Blow [120]). Their values before Patterson correlation refinement were respectively, (1) (θ_1=320°, θ_2=55°, θ_3=200°), (7) (θ_1=322°, θ_2=53°, θ_3=189°), (39) (θ_1=310°, θ_2=55°, θ_3=212°). Diffraction terms between 8 and 4 Å resolution were used in the calculation.

Because of the space group ambiguity (which had not been resolved at this stage) translational searches were carried out in space groups $P2_12_12$ and $P2_12_12_1$. The region of the unit cell which needs to be searched is defined by the unit cell of the Cheshire group of the space group [494], in this case *Pmmm* with unit cell $\frac{1}{2}a \times \frac{1}{2}b \times \frac{1}{2}c$. A modestly contrasted solution was obtained in space group $P2_12_12_1$, but not in space group $P2_12_12$ (see Table 4.7)

Table 4.7 Results of the translational search

The top four peaks of the translation function are shown (computed in both of the possible spacegroups). The translation function (Equation 4.6) was evaluated using all observed data between 8 and 4 Å resolution.

Space Group	Peak	Fractional translation			Function value	Function mean	Function standard deviation
		a	b	c			
$P2_12_12_1$	1	0.259	0.056	0.407	0.1660	0.071	0.016
	2	0.333	0.426	0.148	0.1428		
	3	0.259	0.056	0.370	0.1394		
	4	0.259	0.056	0.241	0.1383		
$P2_12_12$	1	0.463	0.444	0.148	0.1364	0.070	0.017
	2	0.000	0.296	0.167	0.1305		
	3	0.500	0.296	0.167	0.1305		
	4	0.407	0.148	0.167	0.1302		

Inspection of the molecular replacement solution in $P2_12_12_1$ showed that it was physically reasonable. The molecules were non-overlapping, and there were plausible contacts between neighbouring molecules. After rigid-body positional refinement (program X-plor), the conventional crystallographic R-factor was 53% (for all observed data, 6.0 - 3.5 Å resolution).

4.4.5 Building an initial model

Phases were calculated from the positioned search model, and maps (at 3.5 Å resolution) computed with Fourier coefficients of the form $(2m|F_o| - D|F_c|)$ and $(m|F_o| - D|F_c|)$ (SIGMAA weighting), where $|F_o|$ is the native structure factor amplitude, $|F_c|$ is the calculated structure factor amplitude, and m and D have been defined by Read [495, 496]

The structures of the two fungal lipases GCL and CRL were superimposed on the oriented and translated TcAChE molecular replacement solution. The superposition was based on Cα positions in the first seven strands of the central beta sheet (β1 - β7, see Figure 4.4). Equivalent Cα positions were deduced by inspection of the structures. The initial superposition was by the method of Kabsch [497]; the set of equivalenced positions and the transformation were then refined in iterative fashion by the method of Rao and Rossmann [498] (program SUPPOS in the BIOMOL program suite). The RMS difference between 42 equivalenced Cα positions was 0.55 Å for TcAChE and GCL, and 0.60 Å for TcAChE and CRL.

A structure-based sequence alignment was constructed for the lipase/esterase family by Cygler *et al* [403]. This was retrieved from the 3D_ali database [499]. The alignment is based on the two members of the family for which structures were then available (TcAChE and GCL). The program TURBO-FRODO was used for interactive model building [C. Cambillau, A. Roussel, A.G. Inisan and E. Koups-Mouthuy]. Making use of the structure-based sequence alignment, and with reference to the superimposed structures of TcAChE, GCL and CRL, side chains were added to the model, and the backbone conformation corrected where possible. In some cases side chain positions were clearly indicated in the difference Fourier synthesis. In other cases they were not, but could be safely inferred from structural conservation. In regions where the maps were clearly uninterpretable, the polypeptide backbone was removed (the initial search model comprised the entire TcAChE molecule). As noted by Cygler *et al* [403], the similarity between members of the lipase/esterase family is less evident in the C-terminal part of the molecule. Correspondingly, while the sequence assignment was obvious in most of the initial strands of the central β-sheet, in the latter half of the molecule it became more difficult, and for strands β_9 and β_{10} it was not possible to assign the sequence unambiguously. In regions which contained BSDL-specific insertions relative to the other members of the family (for example the connection between strands β_3 and β_4), there were few indications of the path of the polypeptide chain. Unsuprisingly, it was also apparent that some of the surface loops (e.g. that which connects strands β_0 and β_1) must differ in conformation from the other structures. The initial model constructed in this fashion contained 382 amino acids (c.f. the 522 amino acids in total which comprise the truncated variant). In general this comprised the strands of the large and small β-sheets, and many of the surrounding helices. There were three extensive regions in which the structural conservation was very poor (110-128, 263-286, and 330-383); none of these regions could be initially modeled.

4.4.6 Refinement at low resolution

Refinement of structural models using low resolution data is problematic because at some point there will be insufficient observations to explicitly determine all parameters of the model. The problem of indeterminacy in crystallographic refinement is not fully understood, largely because the standard practice of treating stereochemical and geometric data as 'observations' during refinement [500, 501] obscures the true observation to parameter ratio (recent work based on full matrix least squares analysis has begun to address this problem [502]). Even so, with 4049 non-hydrogen atoms in the complete structure of truncated BSDL, and only 6893 observations, the problem appeared hopelessly underdetermined.

Two refinement procedures were briefly (and unsuccessfully!) investigated. The first was straightforward reciprocal space least squares minimization using the program TNT [188, 189]. The second was based on torsion angle molecular dynamics (coupled with simulated annealing). This recently-described procedure [503] has been successfully applied to some difficult crystallographic refinement problems [504]. Here bond lengths and angles are fixed during the refinement procedure, and only rotations about bonds are considered (which substantially reduces the number of degrees of freedom associated with the model). Qualitatively similar results were obtained using both refinement procedures. Large reductions in the conventional crystallographic R-factor were achieved, but the free R-factor [505, 506], calculated using 259 reflections omitted from refinement, increased. This cross-validation procedure indicates that the accuracy of the model was not improving. Similar observations were reported by Sauer-Eriksson *et al* in their low resolution (3.5 Å) structure determination of a streptococcal protein G-immunoglobulin complex (also in the absence of non-crystallographic symmetry) [507]. There were other indications that the problem was underdetermined. The Ramachandran plot of the 'refined' structure steadily worsened (not shown), indicating local distortions of the model were occurring.

Despite this, it is clear that refinement was not a totally futile exercise. For example there was a concerted shift in the long, kinked, helix $\alpha^{4}_{7,8}$ which was certainly not wholly the result of interactive rebuilding of the model. Consequently it seems that at least some of the parameters of the model are specified by the data, but the overall system is poorly determined, and this results in ready distortions at the local level (implied by the worsening Ramachandran plot and the raised free R-factor). The problem is that the level of detail inherent in an atomic representation of the structure is not supported by the data. It is not clear how the model might be reparameterized to allow for effective refinement at low resolution

In practice then, crystallographic refinement programs were used only to achieve global regularization of the protein structure after rebuilding (by running several refinement cycles with very tight geometrical restraints).

4.4.7 Difficulties in completion of the partial structure

There is an acute problem in phasing by the method of molecular replacement when a search model with low sequence identity to the target structure is used. In such a case there will almost certainly be regions of the structure which differ significantly from the search model (this is the case for BSDL, using TcAChE as a search model). Since the phase determines the appearance of an electron density map much more than the amplitude [207], and the phases are calculated from the search model, regions which differ substantially from this initial model are usually very poorly defined in electron density maps. Difference Fourier methods are only fully effective when most of the structure is determined, and the phases calculated from the partial structure are very accurate [508], Hence the problem of recovering the missing structure is not trivial, especially in the absence of any additional means to improve the phases (e.g. non-crystallographic symmetry).

Real space density modification procedures such as solvent flattening and histogram matching are not particularly useful. Since the phase information is calculated from a model, the electron density is already effectively flattened outside the molecular region. Electron density histograms already closely resemble those of a refined structure, since most of the map has been interpreted in terms of a stereochemically correct atomic model, and this provides a very weak phase constraint. In the particular case of BSDL, there is no non-crystallographic symmetry or multiple crystal forms, which would allow electron density averaging.

A number of dummy-atom procedures have been reported, in which electron density is described by a set of atomic positions with no attention paid to their structural sense (thus they can be any distance from each other) (see [509, 510, 511, 512, 513]). Such an artificial atomic model can be used to describe regions in the crystal for which a stereochemically correct atomic model has not yet been determined. This can be used to improve the phase estimates, which allows the partial atomic model to be corrected and extended. One variant on such procedures was used in the structure determination of GFOR (see Section 2.7.5.2).

Another procedure was recently suggested by Urzhumtsev [514]. In a sphere surrounding the region with the missing structure, a fine grid of dummy atoms is placed. Any real atoms (from the partial, stereochemically-correct, atomic model) within this sphere are set to an occupancy of zero. Then the occupancy and position of the dummy atoms are refined (using

calculated structure factors which result from the known (partial) structure and the dummy atom model). Dummy atoms which refine to low occupancies are removed from the model. This procedure is very similar to the standard crystallographic 'omit' map (see [515]), except that in the omitted region a dummy-atom model is constructed. Again the hope is that the resulting map (with phases calculated from the partial structure and the refined dummy-atom model) will be more readily interpretable in this region, and allow the extension of the partial structure.

In practice it was found that with only low resolution data available, the refinement of the dummy atom model was often poorly determined. Nonetheless, this procedure, coupled with careful inspection of conventional electron density maps, allowed the partial structure to be extended in some regions. In other regions, the situation remained ambiguous, and the missing structure could not be determined with any confidence.

Recently Szöke and coworkers [516] have also considered the problem of recovering missing structure, given a partial structure, which was recast in terms of holographic theory. However, this procedure is still under development.

4.4.8 Current status of the structure determination

At 3.5 Å resolution it appears that the problem is essentially intractable. The core of the structure is well defined as expected, but in the absence of some means to improve the phases (e.g. experimental phase determination) it is impossible to build the missing structure with any confidence. Refinement of a conventional atomic representation of the structure also appears impossible with current methodology. Fortunately the recent collection of better, higher resolution, data (data set III) now opens the way for the completion of the structure.

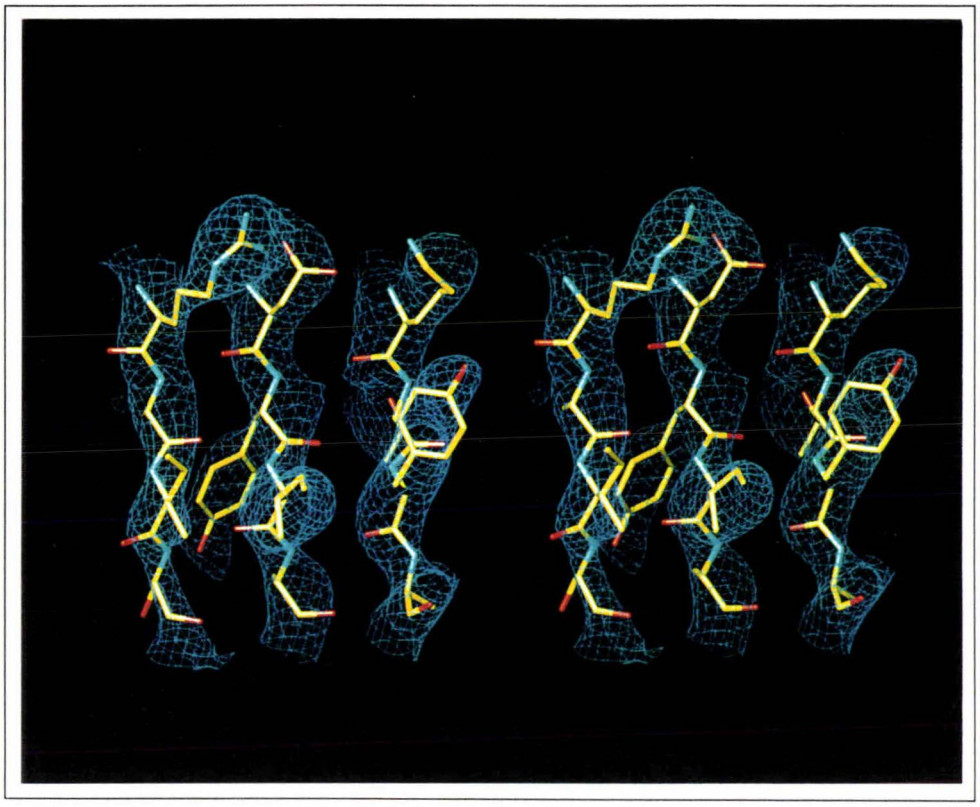
Rebuilding and refinement of the atomic model using this data falls outside the scope of this thesis, but some brief observations can be made. While refinement of protein structures at moderate resolution (2.7 Å) is still not completely straightforward, the problem is not now underdetermined. The conventional crystallographic R-factor cannot be driven down to arbitrary values by extensive refinement. Refinement of the partial model (428 amino acids) using the new data set resulted in an R-factor of 43% for all data used in refinement, and 48% for a test set (590 reflections) excluded from refinement (using the program TNT [188, 189] and a bulk solvent correction (see Section 2.7.6.5) with $K_{sol} = 0.8$ and $B_{sol} = 150$ Å^2). Even at this preliminary stage, the difference in the detail of the electron density maps is striking (see Figure 4.18). Difference Fourier syntheses now clearly indicate parts of the missing

structure, and also errors in the current partial structure. While the completion and refinement of the structure will not be trivial (and will perhaps require the use of computational procedures such as the dummy-atom techniques described above) there are clear indications that the problem is now tractable. Thus it should not be long before the structure of the N-terminal catalytic domain of BSDL is determined, and the structural basis for the bile salt activation of this unique lipase revealed.

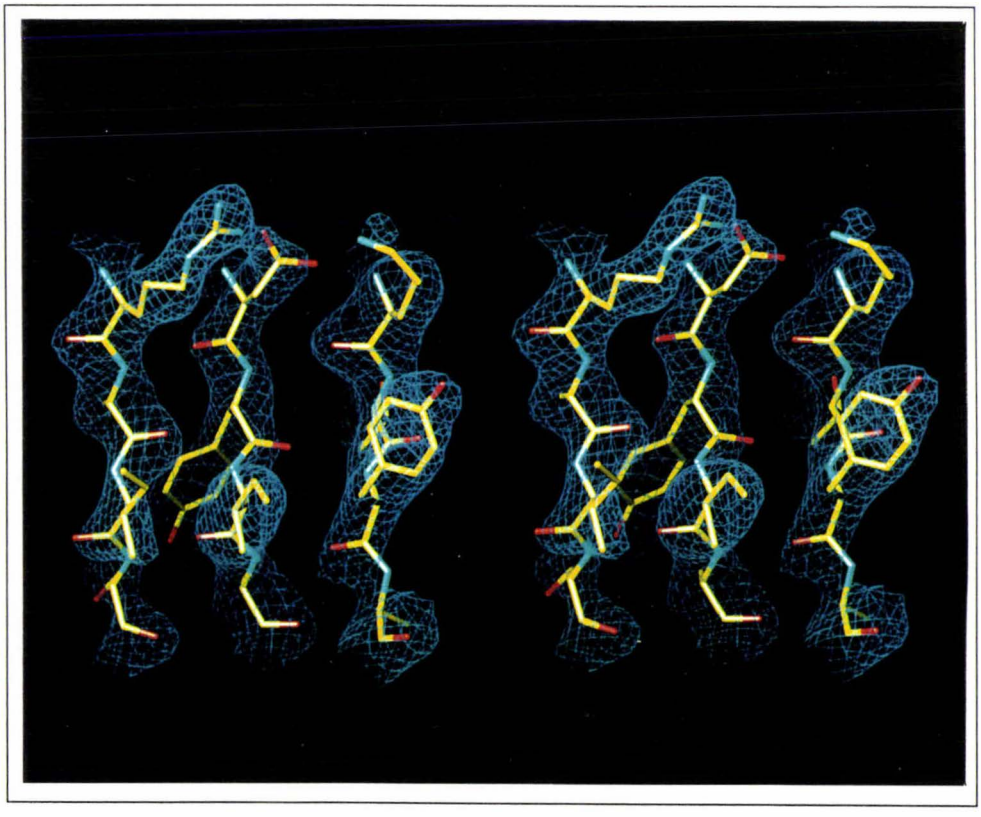
Figure 4.18 (following page) Electron density for truncated recombinant BSDL.

Electron density maps for the same region of BSDL, calculated at (A) 3.5 Å resolution (data set II), and (B) 2.7 Å resolution (data set III). Only preliminary refinement of the model had been carried out using the high resolution data set. Both maps are contoured at 1.5 σ. Displayed are residues from strands β_6, β_7, and β_8. Fourier coefficients employed in the map calculation were of the form $(2m|F_o| - D|F_c|)$ (SIGMAA weighting) where $|F_o|$ is the native structure factor amplitude, $|F_c|$ is the calculated structure factor amplitude, and m and D have been defined by Read [207]. The figure was prepared using Turbo-Frodo [C. Cambillau, A. Roussel, A.G. Inisan and E. Koups-Mouthuy]

A



B



REFERENCES

[1] A. McPherson. A brief history of protein crystal growth. *Journal of Crystal Growth*, 110:1–10, 1991.

[2] J. B. Sumner. The isolation and crystallization of the enzyme urease. *The Journal of Biological Chemistry*, 69:435–441, 1926.

[3] J. H. Northrop. Crystalline pepsin. *Science*, 69:580, 1929.

[4] J. H. Northrop, M. Kunitz, and R. M. Herriot. *Crystalline Enzymes (second edition)*. Columbia University Press: New York, 1948.

[5] L. Bragg. First stages in the X-ray analysis of proteins. *Reports on Progress in Physics*, 28:1–14, 1965.

[6] D. C. Hodgkin. Crystallographic measurements and the structure of protein molecules as they are. *Annals of the New York Academy of Sciences*, 325:121–148, 1979.

[7] M. Perutz. Early days of protein crystallography. *Methods in Enzymology*, 114:3–18, 1985.

[8] J. D. Bernal and D. Crowfoot. X-ray photographs of crystalline pepsin. *Nature*, 133:794–795, 1934.

[9] J. C. Kendrew, R. E. Dickerson, B. E. Strandberg, R. Hart, D. R. Davies, D. C. Phillips, and V. C. Shore. Structure of myoglobin: A three dimensional Fourier synthesis at 2 Å resolution. *Nature*, 185:422–427, 1960.

[10] A. F. Cullis, H. Muirhead, M. F. Perutz, M. G. Rossmann, and A. C. T. North. The structure of haemoglobin IX. A three-dimensional Fourier synthesis at 5.5 Å resolution: Description of the structure. *Proceedings of the Royal Society*, A265:161–187, 1961.

[11] R. Giegé, B. Lorber, and A. Théobald-Dietrich. Crystallogenesis of biological macromolecules: Perspectives and facts. *Acta Crystallographica*, D50:339–350, 1994.

[12] G. L. Gilliland and D. R. Davies. Protein crystallization: The growth of large-scale single crystals. *Methods in Enzymology*, 104:370–381, 1984.

[13] G. L. Gilliland and J. E. Ladner. Crystallization of biological macromolecules for X-ray diffraction studies. *Current Opinion in Structural Biology*, 6:595–603, 1996.

[14] J. A. Littlechild. Protein crystallization: magical or logical: can we establish some general rules. *Journal of Physics D: Applied Physics*, 24:111–118, 1991.

[15] A. McPherson. Current approaches to macromolecular crystallization. *European Journal of Biochemistry*, 189:1–23, 1990.

[16] A. McPherson, A. J. Malkin, and Y. G. Kuznetsov. The science of macromolecular crystallization. *Structure*, 3:759–768, 1995.

[17] D. Ollis and S. White Protein crystallization. *Methods in Enzymology*, 182:646–659, 1990.

[18] P. Weber. Physical principles of protein crystallization. *Advances in Protein Chemistry*, 41:1–36, 1991.

[19] G. Weigand. How do you get large protein crystals? Two familiar laboratory methods improved and simplified. In H. Tschesche, editor, *Modern methods in protein and nucleic acid research: Review articles*. Walter de Gruyter, Berlin, 1990.

[20] C. W. Carter, Jr. Design of crystallization experiments and protocols. In A. Ducruix and R. Giegé, editors, *Crystallization of Nucleic acids and Proteins*, pages 47–71. Oxford University Press, Oxford, 1992.

[21] B. Cudney, S. Patel, K. Weisgraber, Y. Newhouse, and A. McPherson. Screening and optimization strategies for macromolecular crystal growth. *Acta Crystallographica*, D50:414–423, 1994.

[22] A. McPherson. Two approaches to the rapid screening of crystallization conditions. *Journal Of Crystal Growth*, 122:161–167, 1992.

[23] J. Jancarik and S.-H. Kim. Sparse matrix sampling: a screening method for crystallization of proteins. *Journal of Applied Crystallography*, 24:409–411, 1991.

[24] J. Sedzik. DESIGN: A guide to protein crystallization experiments. *Archives of Biochemistry and Biophysics*, 308:342–348, 1994.

[25] H.-S. Shieh, W. C. Stallings, A. M. Stevens, and R. A. Stiegman. Using sampling techniques in protein crystallization. *Acta Crystallographica*, D51:305–310, 1995.

[26] E. A. Stura, G. R. Nemerow, and I. A. Wilson. Strategies in the crystallization of glycoproteins and protein complexes. *Journal of Crystal Growth*, 122:273–285, 1992.

[27] P. C. Weber. A protein crystallization strategy using automated grid searches on successively finer grids. *METHODS: A Companion to Methods in Enzymology*, 1:31–37, 1990.

[28] G. Feher and Z. Kam. Nucleation and growth of protein crystals: General principles and assays. *Methods in Enzymology*, 114:77–112, 1985.

[29] A. Ducruix and R. Giegé. Methods of crystallization. In A. Ducruix and R. Giegé, editors, *Crystallization of Nucleic acids and Proteins*, pages 73–98. Oxford University Press, Oxford, 1992.

[30] F. R. Salemme, L. Genieser, B. C. Finzel, R. M. Hilmer, and J. J. Wendoloski. Molecular factors stabilizing protein crystals. *Journal of Crystal Growth*, 90:273–282, 1988.

[31] D. M. Steinberg and W. G. Hunter. Experimental design: review and comment. *Technometrics*, 26:71–97, 1984.

[32] E. L. Lehmann. Model specification: The views of Fisher and Neyman, and later developments. *Statistical Science*, 5:160–168, 1990.

[33] D. R. Cox. Role of models in statistical analysis. *Statistical Science*, 5:169–174, 1990.

[34] C. F. J. Wu, S. S. Mao, and F. S. Ma. SEL: A search method based on orthogonal arrays. In S. Ghosh, editor, *Statistical Design and Analysis of Industrial Experiments*, pages 279–310. Marcel Dekker, New York, 1990.

[35] H. M. Baker, C. L. Day, G. E. Norris, and E. N. Baker. Enzymatic deglycosylation as a tool for crystallization of mammalian binding proteins. *Acta Crystallographica*, D50:380–384, 1994.

[36] B. K. Ghosh and P. K. Sen, editors, *Handbook of Sequential Analysis*.. Marcel Dekker, New York, 1991.

[37] S. Trakhanov and F. A. Quiocho. Influence of divalent cations in protein crystallization. *Protein Science*, 4:1914–1919, 1995.

[38] D. M. Lawson, P. J. Artymuik, S. J. Yewdall, J. M. A. Smith, J. C. Livingstone, A. Treffry, A. Luzzago, S. Levi, P. Arosio, G. Cesareni, C. D. Thomas, W. V. Shaw, and P. M. Harrison. Solving the structure of human H ferritin by genetically engineering intermolecular crystal contacts. *Nature*, 249:514–544, 1991.

[39] R. W. Kennard and L. A. Stone. Computer aided design of experiments. *Technometrics*, 11:137–148, 1969.

[40] T. J. Aird and J. R. Rice. Systematic search in high dimensional sets. *SIAM Journal of Numerical Analysis*, 14:296–312, 1977.

[41] I. M. Sobol. On the systematic search in a hypercube. *SIAM Journal of Numerical analysis*, 16:790–793, 1979.

[42] C. R. Rao. Factorial experiments derivable from combinatorial arrangements of arrays. *Journal of the Royal Statistical Society Supplement*, 9:118–139, 1947.

[43] C. R. Rao. Some combinatorial problems of arrays and applications to the design of experiments. In J. N. Srivastava, editor, *A Survey of Combinatorial Theory*, pages 349–359. North Holland, Amsterdam, 1973.

[44] A. Dey and V. Agrawal. Orthogonal fractional plans for asymmetrical factorials derivable from orthogonal arrays. *Sankhya: The Indian Journal of Statistics*, B 47:56–66, 1985.

[45] A. Dey. *Orthogonal Fractional Factorial Designs*. Wiley, New Delhi, 1985.

[46] J. P. Mandeli. Construction of asymmetrical orthogonal arrays having factors with a large non-prime power number of levels. *Journal of Statistical Planning and Inference*, 47:377–391, 1995.

[47] K. K. Goswami and S. Pal. On the construction of orthogonal factorial designs of resolution iv. *Communications in Statistical Theory and Methods*, 21:3561–3570, 1992.

[48] J. C. Wang and C. F. J. Wu. An approach to the construction of asymmetrical orthogonal arrays. *Journal of the American Statistical Association*, 86:450–456, 1991.

[49] A. S. Hedayat. New properties of orthogonal arrays and their statistical applications. In S. Ghosh, editor, *Statistical Design and Analysis of Industrial Experiments*, pages 407–422. Marcel Dekker, New York, 1990.

[50] A. B. Owen. Orthogonal arrays for computer experiments, integration and visualization. *Statistica Sinica*, 2:439–452, 1992.

[51] A. B. Owen. Lattice sampling revisited: Monte Carlo variance of means over randomized orthogonal arrays. *The Annals of Statistics*, 22:930–945, 1994.

[52] B. Tang. Orthogonal array-based latin hypercubes. *Journal of the American Statistical Association*, 88:1392–1397, 1993.

[53] R. M. Garavito and D. Picot. Crystallization of membrane proteins: a minireview. *Journal of Crystal Growth*, 110:89–95, 1991.

[54] W. Kühlbrandt. Three-dimensional crystallization of membrane proteins. *Quarterly Reviews of Biophysics*, 21:429–477, 1988.

[55] A. McPherson, S. Koszelak, H. Axelrod, J. Day, R. Williams, L. Robinson, M. McGrath, and D. Cascio. An experiment regarding crystallization of soluble proteins in the presence of β-octyl glucoside. *The Journal of Biological Chemistry*, 261:1969–1975, 1986.

[56] S. N. Timasheff and T. Arakawa. Mechanism of protein precipitation and stabilization. *Journal of Crystal Growth*, 90:39–46, 1988.

[57] A. M. Brzozowski. Crystallization of a *Humicola lanuginosa* lipase-inhibitor complex with the use of polyethylene glycol monomethyl ether. *Acta Crystallographica*, D49:352–354, 1993.

[58] H. Mahadaven and C. K. Hall. Statistical-mechanical model of protein precipitation by nonionic polymer. *AIChE Journal*, 36:1517–1528, 1990.

[59] T. Arakawa and S. N. Timasheff. Mechanism of poly(ethylene glycol) interaction with proteins. *Biochemistry*, 24:6756–6762, 1985.

[60] D. H. Atha and K. C. Ingham. Mechanism of precipitation of proteins by polyethylene glycols. *Journal of Biological Chemistry*, 256:12108–12117, 1981.

[61] P. H. von Hippel and T. Schleich. The effect of neutral salts on the structure and conformational stability of macromolcules in solution. In S. N. Timasheff and G. D. Fasman, editors, *Structure and Stability of Biological Macromolecules*, pages 417–574. Dekker, New York, 1969.

[62] T. Arakawa, R. Bhat, and S. N. Timasheff. Preferential interactions determine protein solubility in three-component solutions: the $MgCl_2$ system. *Biochemistry*, 29:1914–1923, 1990.

[63] W. Melander and C. Horváth. Salt effects on hydrophobic interactions in precipitation and chromatography of proteins: an interpretation of the lyotropic series. *Archives of Biochemistry and Biophysics*, 183:200–215, 1977.

[64] J. Timmermans. *Physico-chemical Constants of Binary Systems*. Interscience, New York, 1960.

[65] V. M. M. Lobo and J. L. Quaesma. *Electrolyte Solutions: Literature Data on Thermodynamic and Transport Properties*. University of Columbia, Portugal, 1981.

[66] S. Addelman. Techniques for constructing fractional replicate plans. *Journal of the American Statistical Association*, 58:45–71, 1963.

[67] S. Patel, B. Cudney, and A. McPherson. Polymeric precipitants for the crystallization of macromolecules. *Biochemical and Biophysical Research Communications*, 207:819–828, 1995.

[68] H. Mahadaven and C. K. Hall. Experimental analysis of protein precipitation by polyethylene glycol and comparison with theory. *Fluid Phase Equilibria*, 78:297–321, 1992.

[69] L. Bläckberg and O. Hernell. The bile-salt-stimulated lipase in human milk. Purification and characterization. *European Journal of Biochemistry*, 116:221–225, 1981.

[70] G. E. P. Box and J. S. Hunter. The 2^{k-p} fractional factorial designs part I. *Technometrics*, 3:311–351, 1961.

[71] G. E. P. Box and J. S. Hunter. The 2^{k-p} fractional factorial designs part II. *Technometrics*, 3:449–458, 1961.

[72] F. E. Baralle, C. C. Shoulders, and N. J. Proudfoot. The primary structure of the human ε-globin gene. *Cell*, 21:621–626, 1980.

[73] M. M. Silva, P. H. Rogers, and A. Arnone. A third quaternary structure of human hemoglobin at 1.7-Å resolution. *The Journal of Biological Chemistry*, 267:17248–17256, 1992.

[74] M. Perutz. Preparation of haemoglobin crystals. *Journal of Crystal Growth*, 2:54–56, 1968.

[75] C.W. Carter, Jr. and C. W. Carter. Protein crystallization using incomplete factorial experiments. *The Journal of Biological Chemistry*, 254(23):12219–12223, 1979.

[76] C. W. Carter, Jr., E. T. Baldwin, and L. Frick. Statistical design of experiments for protein crystal growth and the use of a precrystallization assay. *Journal of Crystal Growth*, 90:60–73, 1988.

[77] F. E. Satterthwaite. Random balance experimentation. *Technometrics*, 1:111–137, 1959.

[78] T. A. Budne. The application of random balance designs. *Technometrics*, 1:139–155, 1959.

[79] F. J. Anscombe. Quick analysis methods for random screening experiments. *Technometrics*, 1:195–209, 1959.

[80] A. P. Dempster. Random allocation designs I: On general classes of estimation methods. *Annals of Mathematical Statistics*, 31:885–905, 1960.

[81] J. W. Tukey, W. J. Youden, O. Kempthorne, G. E. P. Box, and J. S. Hunter. Discussion of 'Random balance experimentation' by F. E. Satterthwaite and 'The application of random balance designs' by T. A. Budne. *Technometrics*, 1:157–193, 1959.

[82] A. Herzberg and D. R. Cox. Recent work on the design of experiments: A bibliography and a review. *Journal of the Royal Statistical Society*, A 132:29–67, 1969.

[83] J. P. C. Kleijnen. *Statistical Techniques in Simulation*. Marcel Dekker, New York, 1975.

[84] D. K. J. Lin. Generating systematic supersaturated designs. *Technometrics*, 37:213–225, 1995.

[85] J. L. Folks. Use of randomization in experimental research. In K. Hinkelmann, editor, *Experimental Design, Statistical Models, and Genetic Statsitcs*, pages 17–32. Marcel Dekker, New York, 1984.

[86] M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21:239–245, 1979.

[87] A. R. Ferré-D'Amaré and S. K. Burley. Use of dynamic light scattering to assess crystallizability of macromolecules and macromolecular assemblies. *Structure*, 2:357–359, 1994.

[88] V. Mikol, P. Vincendon, G. Eriani, E. Hirsch, and R. Giegé. Diagnostic of protein crystallization by dynamic light scattering; an application to an aminoacyl-tRNA synthetase. *Journal of Crystal Growth*, 110:195–200, 1991.

[89] A.-C. Dock-Bregeon and D. Moras. Crystallization of nucleic acids and co-crystallization of proteins and nucleic acids. In A. Ducruix and R. Giegé, editors, *Crystallization of Nucleic acids and Proteins*, pages 145–174. Oxford University Press, Oxford, 1992.

[90] R. A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.

[91] P. Heikinheimo, T. Salminen, R. Lahti, B. Cooperman, and A. Goldman. New crystal forms of *Esherichia coli* and *Saccharomyces cervisiae* soluble inorganic pyrophosphatases. *Acta Crystallographica*, D51:399–401, 1995.

[92] S. Mylvaganam, C. Slingby, P. Lindley, and T. Blundell. Preliminary studies of adult δ-crystallin: evidence of a space group transition. *Acta Crystallographica*, B43:580–582, 1987.

[93] F. Dyda, W. Furey, S. Swaminathan, and M. Sax. Mutiple crystal forms of brewers' yeast pyruvate decarboxylase: Characterization and preliminary crystallographic analysis. In H. Bisswanger and J. Ulllrich, editors, *Biochemistry and Physiology of Thiamin Diphosphate Enzymes*, pages 115–122. VCH Publishers, 1991.

[94] J. Swings and J. De Ley. The biology of *Zymomonas*. *Bacteriological Reviews*, 41:1–46, 1977.

[95] H. Sahm, S. Bringer-Meyer, and G. Sprenger. The genus *Zymomonas*. In A. Balows, H. G. Trüper, M. Dworkin, W. Harder, and K.-H. Schleifer, editors, *The Prokaryotes*, pages 2287–2301. Springer-Verlag, New York, second edition edition, 1992.

[96] L. Viikari. Carbohydrate metabolism in *Zymomonas*. *CRC Critical Review of Biotechnology*, 7(3):237–261, 1988.

[97] M. Zhang, C. Eddy, K. Deanda, M. Finkelstein, and S. Picataggio. Metabolic engineering of a pentose metabolism pathway in ethanologenic *Zymomonas mobilis*. *Science*, 267:240–243, 1995.

[98] L. Viikari. Formation of levan and sorbitol from sucrose by *Zymomonas mobilis*. *Applied Microbiology and Biotechnology*, 19:252–255, 1984.

[99] K. D. Barrow, J. G. Collins, D. A. Leigh, P. L. Rogers, and R. G. Warr. Sorbitol production by *Zymomonas mobilis*. *Applied Microbiology and Biotechnology*, 20:225–232, 1984.

[100] D. Leigh, R. Scopes, and P. Rogers. A proposed pathway for sorbitol production by *Zymomonas mobilis*. *Applied Microbiology and Biotechnology*, 20:413–415, 1984.

[101] M. Zachariou and R. K. Scopes. Glucose-fructose oxidoreductase, a new enzyme isolated from *Zymomonas mobilis* that is responsible for sorbitol production. *Journal of Bacteriology*, 167(3):863–869, 1986.

[102] M. J. Hardman and R. K. Scopes. The kinetics of glucose-fructose oxidoreductase from *Zymomonas mobilis*. *European Journal of Biochemistry*, 173:203–209, 1988.

[103] M. Strohdeicher, B. Schmitz, S. Bringer-Meyer, and H. Sahm. Formation and degradation of gluconate by *Zymomonas mobilis*. *Applied Microbiology and Biotechnology*, 27:278–382, 1988.

[104] H. Loos, M. Völler, B. Rehr, Y.-D. Stierhof, H. Sahm, and G. Sprenger. Localisation of the glucose-fructose oxidoreductase in wild type and overproducing strains of *Zymomonas mobilis*. *FEMS Microbiology Letters*, 84:211–216, 1991.

[105] H. C. Aldrich, L. Mcdowell, F. S. M. de Barbosa, L. P. Yomano and R. K. Scopes, and L. O. Ingram. Immunocytochemical localization of glycolytic and fermentative enzymes in *Zymomonas mobilis*. *Journal of Bacteriology*, 174:4504–4508, 1992.

[106] H. Loos, R. Krämer, H. Sahm, and G. A. Sprenger. Sorbitol promotes growth of *Zymomonas* mobilis in environments with high concentrations of sugar: Evidence for a physiological function of glucose-fructose oxidoreductase in osmoprotection. *Journal of Bacteriology*, 176(24):7688–7693, 1994.

[107] V. Kanagasundaram and R. K. Scopes. Cloning, sequence analysis, and expression of the structural gene encoding glucose-fructose oxidoreductase from *Zymomonas mobilis*. *Journal of Bacteriology*, 174(5):1439–1447, 1992.

[108] G. Schatz and B. Dobberstein. Common principles of protein translocation across membranes. *Science*, 271:1519–1526, 1996.

[109] A. M. Lesk. NAD-binding domains of dehydrogenases. *Current Opinion in Structural Biology*, 5:775–783, 1995.

[110] R. K. Wierenga, M. C. D. Maeyer, and W. Hol. Interaction of pyrophosphate moieties with α-helices in dinucleotide binding proteins. *Biochemistry*, 24:1346–1357, 1985.

[111] P. A. Frey. Complex pyridine nucleotide dependent transformations. In D. Dolphin, O. Avramovic, and R. Poulson, editors, *Pyridine Nucleotide Coenzymes Part B*, pages 461–512. Wiley, 1987.

[112] P. W. van Ophem and J. A. Duine. Microbial alcohol, aldehyde and formate ester oxidoreductases. In H. Weiner, D. W. Crabb, and T. G. Flynn, editors, *Enzymology and Molecular Biology of Carbonyl Metabolism 4*, pages 605–620. Plenum Press, New York, 1993.

[113] N. Kato, T. Yamagami, M. Shimao, and C. Sakazawa. Formaldehyde dismutase, a novel NAD-binding oxidoreductase from psedomonas putida F61. *European Journal of Biochemistry*, 156:59–64, 1986.

[114] S. G. Allen and J. R. Patil. Studies on the structure and mechanism of action of the malate-lactate transhydrogenase. *The Journal of Biological Chemistry*, 247(3):909–916, 1972.

[115] J. B. Thoden, P. A. Frey, and H. M. Holden. Molecular structure of the nadh/udp-glucose abortive complex of UDP-galactose 4-epimerase from *Escherichia coli*: Implications for the catalytic mechanism. *Biochemistry*, 35:5137–5144, 1996.

[116] R. K. Scopes, V. Testolin, A. Stoter, K. Griffiths-Smith, and E. M. Algar. Simultaneous purification and characterization of glucokinase, fructokinase and glucose-6-phosphate dehydrogenase from *Zymomonas mobilis*. *Biochemistry Journal*, 228:627–634, 1985.

[117] H. Loos, U. Ermler, G. A. Sprenger, and H. Sahm. Crystallization and preliminary X-ray analysis of glucose-fructose oxidoreductase from zymomonas mobilis. *Protein Science*, 3:2447–2449, 1994.

[118] B. Matthews. Solvent content of protein crystals. *The Journal of Molecular Biology*, 33:491–497, 1968.

[119] B. W. Matthews and S. Bernhard. Structure and symmetry of oligomeric enzymes. *Annual Review of Biophysics and Bioengineering*, 2:257–317, 1973.

[120] M. G. Rossmann and D. M. Blow. The detection of sub-units within the crystallographic asymmetric unit. *Acta Crystallographica*, 15:24–31, 1962.

[121] L. Tong and M. G. Rossmann. The locked rotation function. *Acta Crystallographica*, A46:783–792, 1990.

[122] P. Tollin and M. G. Rossmann. A description of various rotation function programs. *Acta Crystallographica*, 21:872–876, 1966.

[123] T. O. Yeates. Statistics for rotation functions. *Journal of Applied Crystallography*, 26:448–449, 1993.

[124] M. Sato, M. Yamamoto, K. Imada, Y. Katsube, N. Tanaka, and T. Higashi. A high speed data-collection system for large-unit-cell crystals using an imaging plate as a detector. *Journal of Applied Crystallography*, 25:348–357, 1992.

[125] Z. Otwinowski. Oscillation data reduction program. In L. Sawyer, N. Isaacs, and S. Bailey, editors, *Proceedings of the CCP4 Study Weekend*, pages 56–62. SERC Daresbury Laboratory, Warrington, U.K., 1993.

[126] G. Fox and K. Holmes. An alternative method of solving the layer scaling equations of Hamilton, Rollet and Sparks. *Acta Crystallographica*, 20:886–891, 1966.

[127] S. French and K. Wilson. On the treatment of negative intensity observations. *Acta Crystallographica*, A34:517–525, 1978.

[128] A. J. C. Wilson. Determination of absolute from relative X-ray intensity data. *Nature*, 150:152, 1942.

[129] F. Winkler, C. Schutt, and S. Harrison. The oscillation method for crystals with very large unit cells. *Acta Crystallographica*, A35:901–911, 1979.

[130] Collaborative Computational Project No. 4. The CCP4 suite: Programs for protein crystallography. *Acta Crystallographica*, D50:760–763, 1994.

[131] J. R. Helliwell. Protein crystal perfection and the nature of radiation damage. *Journal of Crystal Growth*, 90:259–272, 1988.

[132] C. Nave. Radiation damage in protein crystallography. *Radiation Physics and Chemistry*, 45:483–490, 1995.

[133] J. Abrahams and A. Leslie. Methods used in the structure determination of bovine mitochondrial F_1 ATPase. *Acta Crystallographica*, D52:30–42, 1996.

[134] S. Narayana, M. S. Weininger, K. L. Heuss, and P. Argos. A method to increase protein-crystal lifetime during x-ray exposure. *Journal of Applied Crystallography*, 15:571–573, 1982.

[135] M. V. King, J. Bello, E. H. Pignataro, and D. Harker. Crystalline forms of bovine pancreatic ribonuclease: Some new modifications. *Acta Crystallographica*, 15:144–147, 1962.

[136] M.-P. Crosio, J. Janin, and M. Jullien. Crystal packing in six crystal forms of pancreatic ribonuclease. *Journal of Molecular Biology*, 228:243–251, 1992.

[137] F. Jurnak. Effect of chemical impurities in polyethylene glycol on macromolecular crystallization. *The Journal of Crystal Growth*, 76:577–582, 1986.

[138] B. Bax and C. Slingsby. Crystallization of a new form of the eye lens protein βB2-crystallin. *Journal of Molecular Biology*, 208:715–717, 1989.

[139] C. Reynolds, B. Stowell, K. Joshi, M. M. Harding, S. Maginn, and G. Dodson. Preliminary study of a phase transformation in insulin crystals using synchrotron radiation Laue diffraction. *Acta Crystallographica*, B44:512–515, 1988.

[140] K. Y. Zhang and D. Eisenberg. Solid-state phase transition in the crystal structure of ribulose 1,5-bisphosphate carboxylase/oxygenase. *Acta Crystallographica*, D50:258–262, 1994.

[141] T. Kawashima, C. Berthet-Colominas, S. Cusack, and R. Leberman. Interconversion of crystals of the escherichia coli EF-Tu.EF-Ts complex between high and low-diffraction forms. *Acta Crystallographica*, D52:799–805, 1996.

[142] J. Cosier and A. M. Glazer. A nitrogen-gas-stream cryostat for general X-ray diffraction studies. *Journal of Applied Crystallography*, 19:105–107, 1986.

[143] C. C. F. Blake. The preparation of isomorphous derivatives. *Advances in Protein Chemistry*, 23:59–120, 1968.

[144] G. A. Petsko. Preparation of isomorphous heavy-atom derivatives. *Methods in Enzymology*, 114:147–156, 1985.

[145] J. Drenth. The chemistry of heavy atom attachment. In W. Wolf, P. R. Evans, and A. G. W. Leslie, editors, *Proceedings of the CCP4 Study Weekend*, pages 1-8. SERC Daresbury Laboratory, Warrington, U.K., 1991.

[146] D. Phillips. Advances in protein crystallography. *Advances in Structural Research by Diffraction Methods*, 2:75–140, 1966.

[147] J. Kraut, L. C. Sieker, D. F. High, and S. T. Freer. Chymotrypsinogen: A three dimensional Fourier synthesis at 5 Å resolution. *Proceedings of the National Academyof Sciences, USA*, 48:1417–1424, 1962.

[148] L. Tong and M. G. Rossmann. Patterson-map interpretation with noncrystallographic symmetry. *Journal of Applied Crystallography*, 26:15–21, 1993.

[149] M. G. Rossmann. The position of anomalous scatterers in protein crystals. *Acta Crystallographica*, 14:383–388, 1961.

[150] Z. Otwinowski. Maximum liklihood refinement of heavy atom parameters. In *Proceedings of the CCP4 Study Weekend*, pages 80-86. SERC Daresbury Laboratory, Warrington, U.K., 1991.

[151] D. M. Lawson. A novel platinum reagent [chloro (2,2':6',2"-terpyridine)platinum(II) chloride] for use in heavy atom derivatization of protein crystals. *Acta Crystallographica*, D50:332–334, 1994.

[152] H. M. Holden and I. Rayment. Trimethyllead acetate: A first choice heavy atom derivative for protein crystallography. *Archives of Biochemistry and Biophysics*, 291:187–194, 1991.

[153] F. Vellieux, J. Hunt, S. Roy, and R. Read. DEMON/ANGEL: a suite of programs to carry out density modification. *Journal of Applied Crystallography*, 28:347–351, 1995.

[154] G. Bricogne. Geometric sources of redundancy in intensity data and their use for phase determination. *Acta Crystallographica*, A30:395–405, 1974.

[155] M. G. Rossmann. The molecular replacement method. *Acta Crystallographica*, A46:73–82, 1990.

[156] M. C. Lawrence. The application of the molecular replacement method to the *de novo* determination of protein structure. *Quarterly Reviews of Biophysics*, 24:399–424, 1991.

[157] M. G. Rossmann. *Ab initio* phase determination and phase extension using non-crystallographic symmetry. *Current Opinion in Structural Biology*, 5:650–655, 1996.

[158] D. Blow, M. G. Rossmann, and B. Jeffery. The arrangement of α-chymotrypsin molecules in the monoclinic crystal form. *Journal of Molecular Biology*, 8:65–78, 1964.

[159] G. Bricogne. Methods and programs for direct-space exploitation of geometric redundancies. *Acta Crystallographica*, A32:832–847, 1976.

[160] L. Tong, H.-K. Choi, W. Minor, and M. G. Rossmann. The structure determination of Sindbis virus core protein using isomorphous replacement averaging between two crystal forms. *Acta Crystallographica*, A48:430–442, 1992.

[161] B. Rees, B. Bilwes, J. P. Samama, and D. Moras. Cardiotoxin V_4^{II} from *Naja mossambica*: The refined crystal structure. *Journal of Molecular Biology*, 214:281–297, 1990.

[162] Y. Harpaz, M. Gernstein, and C. Chothia. Volume changes on protein folding. *Structure*, 2(7):641–649, 1994.

[163] G. J. Kleywegt and T. A. Jones. Halloween ... masks and bones. In S. Bailey, R. Hubbard, and D. Waller, editors, *Proceedings of the CCP4 study weekend*, pages 59–66. EPSRC Daresbury Laboratory, Warrington, U.K., 1994.

[164] R. P. Millane. Phase retrieval in crystallography and optics. *Journal of the Optical Society of America. A*, 7:394–411, 1990.

[165] A. D. Podjarny, T. N. Bhat, and M. Zwick. Improving crystallographic macromolecular images: the real-space approach. *Annual Review of Biophysics and Bioengineering*, 16:351–373, 1987.

[166] M. G. Rossmann, R. McKenna, L. Tong, D. Xia, J. Dai, H. Wu, H. K. Choi, D. Marinescu, and R. E. Lynch. Molecular replacement real-space averaging. *Journal of Applied Crystallography*, 25:166–180, 1992.

[167] B.-C. Wang. Resolution of phase ambiguity in macromolecular crystallography. *Methods in Enzymology*, 115:90–112, 1985.

[168] R. W. Harrison. Histogram specification as a method of density modification. *Journal of Applied Crystallography*, 21:949–952, 1988.

[169] K. Y. J. Zhang and P. Main. Histogram matching as a new density modification technique for phase refinement and extension of protein molecules. *Acta Crystallographica*, A46:41–46, 1990.

[170] V. Y. Lunin and T. P. Skovoroda. Frequency-restrained structure factor-refinement. I. histogram simulation. *Acta crytstallographica*, A47:45–52, 1991.

[171] V. Y. Lunin. Electron-density histograms and the phase problem. *Acta Crystallographica*, D49:90–99, 1993.

[172] A. Vrielink, L. F. Lloyd, and D. M. Blow. Crystal structure of cholesterol oxidase from *Brevibacterium sterolicum* refined at 1.8 Å resolution. *Journal of Molecular Biology*, 219:533–554, 1991.

[173] W. S. Cleveland and S. J. Devlin. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83:596–610, 1988.

[174] L. Refaat, C. Tate, and M. Woolfson. Direct-space methods in phase extension and phase refinement. IV. The double-histogram method. *Acta Crystallographica*, D52:252–256, 1996.

[175] S. Xiang and C. W. Carter, Jr. Representing stereochemical information in macromolecular electron-density distributions by multi-dimensional histograms. *Acta Crystallographica*, D52:49–56, 1996.

[176] G. Sim. The distribution of phase angles for structures containing heavy atoms. ii. A modification of the normal heavy-atom method for non-centrosymmetrical structures. *Acta Crystallographica*, 12:813–815, 1959.

[177] G. Sim. A note on the heavy-atom method. *Acta Crystallographica*, 13:511–512, 1960.

[178] K. D. Cowtan and P. Main. Phase combination and cross validation in iterated density-modification calculations. *Acta Crystallographica*, D52:43–48, 1996.

[179] T. L. Blundell and L. N. Johnson. *Protein Crystallography*. Academic Press, London, 1976.

[180] T. Jones. TOM: A graphics fitting program for macromolecules. In D. Sayre, editor, *Computational crystallography*, pages 303–317. Clarendon Press, Oxford, 1982.

[181] T. A. Jones and S. Thirup. Using known substructures in protein model building and crystallography. *The EMBO journal*, 5:819–822, 1986.

[182] S. Jones and J. M. Thornton. Protein-protein interactions: A review of protein dimer structures. *Progress in Biophysics and Molecular Biology*, 63:31–65, 1995.

[183] J. S. Richardson and D. C. Richardson. Interpretation of electron density maps. *Methods in Enzymology*, 115:189–206, 1985.

[184] M. J. Adams and S. Gover. The right model? : Lessons from building a structure of 6PGDH starting from a wrongly connected model with an incorrect sequence. In S. Bailey, R. Hubbard, and D. Waller, editors, *Proceedings of the CCP4 Study Weekend*, pages 19–30. EPSRC Daresbury Laboratory, Warrington, U.K., 1994.

[185] D. Rice. The use of phase combination in the refinement of phosphoglycerate kinase at 2.5 Å resolution. *Acta Crystallographica*, A37:491–500, 1981.

[186] T. Bhat and D. Blow. A method for refinement of partially interpreted protein structures including a procedure for scaling between a model and an electron-density map. *Acta Crystallographica*, A39:166–170, 1983.

[187] T. C. Terwilliger and J. Berendzen. Bayesian weighting for macromolecular crystallographic refinement. *Acta Crystallographica*, D52:743–748, 1996.

[188] D. E. Tronrud, L. F. Ten Eyck, and B. W. Matthews. An efficient general-purpose least squares refinement program for macromolecular structures. *Acta Crystallographica*, A43:489–501, 1987.

[189] D. E. Tronrud. Conjugate-direction minimization: an improved method for the refinement of macromolecules. *Acta Crystallographica*, A48:912–916, 1992.

[190] D. W. J. Cruickshank. The convergence of the least-squares and Fourier refinement methods. *Acta Crystallographica*, 3:10–13, 1950.

[191] D. W. J. Cruickshank. On the relations between Fourier and least-squares methods of structure determination. *Acta Crystallographica*, 5:511–518, 1952.

[192] W. Cochran. Some properties of the $(F_o\text{-}F_c)$-synthesis. *Acta Crystallographica*, 4:408–411, 1951.

[193] Y. C. Leung, R. E. Marsh, and V. Schomaker. The interpretation of difference maps. *Acta Crystallographica*, 10:650–652, 1957.

[194] L. H. Jensen. Protein model refinement by Fourier and least-squares methods. In F. R. Ahmed, K. Huml, and B. Sedlácek, editors, *Crystallographic computing techniques*, pages 307–316. Munksgaard, Copenhagen, 1976.

[195] S. T. Freer, R. A. Alden, S. A. Levens, and J. Kraut. Refinement of five protein structures by constrained F_o - F_c Fourier methods. In F. R. Ahmed, K. Huml, and B. Sedlacek, editors, *Crystallographic computing techniques*, pages 317–321. Munksgaard, Copenhagen, 1976.

[196] S. T. Freer. Classic $(F_o - F_c)$ fourier refinement. *Methods in Enzymology*, 115:235–237, 1985.

[197] R. C. Agarwal. A new least-squares refinement technique based of the fast Fourier transform algorithm. *Acta Crystallographica*, A34:791–809, 1978.

[198] G. J. Kleywegt. Use of non-crystallographic symmetry in protein structure refinement. *Acta Crystallographica*, D52:842–857, 1996.

[199] G. Bricogne and J. Irwin. Maximum-likelihood refinement of incomplete models with BUSTER + TNT. In E. Dodson, M. Moore, A. Ralph, and S. Bailey, editors, *Proceedings of the CCP4 Study Weekend*, pages 85–92. CCLRC Daresbury Laboratory, Warrington, U.K., 1996.

[200] G. N. Murshudov, E. J. Dodson, and A. A. Vagin. Application of maximum likelihoood methods for macromolecular refinement. In E. Dodson, M. Moore, A. Ralph, and S. Bailey, editors, *Proceedings of the CCP4 Study Weekend*, pages 93–104. CCLRC Daresbury Laboratory, Warrington, U.K., 1996.

[201] N. S. Pannu and R. J. Read. Improved structure refinement through maximum liklihood. *Acta Crystallographica*, A52:659–668, 1996.

[202] M. G. Rossmann. The refinement of structures partially determined by the isomorphous replacement method. *Acta Crystallographica*, 14:641–647, 1961.

[203] W. A. Hendrickson and E. A. Lattman. Representation of phase probability distributions for simplified combination of independent phase information. *Acta Crystallographica*, B26:136–143, 1970.

[204] D. Rice, B. Anderson, and E. Baker. A practical guide to the use of partial structural phase combination. In S. Bailey, E. Dodson, and S. Phillips, editors, *Proceedings of the CCP4 Study Weekend*, pages 113–120. SERC Daresbury Laboratory, Warrington, U.K., 1988.

[205] D. Blow and F. Crick. The treatment of errors in the isomorphous replacement method. *Acta Crystallographica*, 12:794–802, 1959.

[206] D. Stuart and P. Artymiuk. The use of phase combination in crystallographic refinement: the choice of amplitude coefficients in combined syntheses. *Acta Crystallographica*, A40:713–716, 1985.

[207] R. J. Read. Model bias and phase combination. In S. Bailey, R. Hubbbard, and D. Waller, editors, *Proceedings of the CCP4 Study Weekend*, pages 31–40. EPSRC Daresbury Laboratory, Warrington, U.K., 1994.

[208] D. Guo, G. D. Smith, J. F. Griffin, and D. A. Langs. Use of globic scattering factors for protein structures at low resolution. *Acta Crystallographica*, A51:945–947, 1995.

[209] R. A. Engh and R. Huber. Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallographica*, A47:392–400, 1991.

[210] M. Levitt and B. H. Park. Water: now you see it, now you don't. *Structure*, 1:223–226, 1993.

[211] P. A. Karplus and C. Faerman. Ordered water in macromolecular structure. *Current Opinion in Structural Biology*, 4:770–776, 1994.

[212] M. Frey. Water structure associated with proteins and its role in crystallization. *Acta Crystallographica*, D50:663–666, 1994.

[213] M. Billeter. Hydration water molecules seen by NMR and by X-ray crystallography. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 27:635–645, 1995.

[214] G. Otting, E. Liepinsh, and K. Wüthrich. Protein hydration in aqueous solution. *Science*, 254:974–980, 1991.

[215] C. E. Kundrot and F. M. Richards. Use of the occupancy factor in the refinement of solvent molecules in protein crystal structures. *Acta crystallographica*, B43:544–547, 1987.

[216] T. N. Bhat. Correlation between occupancy and temperature factors of solvent molecules in crystal structures of proteins. *Acta Crystallographica*, A45:145–146, 1989.

[217] L. H. Jensen. Solvent model for protein crystals: on occupancy parameters for discrete solvent sites and the solvent continuum. *Acta Crystallographica*, B46:650–653, 1990.

[218] M. Vijayan. Phase evaluation and some aspects of the Fourier refinement of macromolecules. In R. Diamond, S. Ramaseshan, and K. Venkatesan, editors, *Computing in Crystallography*, pages 19.01–19.25. Indian Institute of Science, Bangalore, 1980.

[219] D. E. Tronrud. The limits of interpretation. In E. Dodson, M. Moore, A. Ralph, and S. Bailey, editors, *Proceedings of the CCP4 Study Weekend*, pages 1–10. CCLRC Daresbury Laboratory, Warrington, U.K., 1996.

[220] J. L. Chambers and R. M. Stroud. Difference Fourier refinement of the structure of DIP-trypsin at 1.5 Å with a minicomputer technique. *Acta Crystallographica*, B33:1824–1837, 1977.

[221] G. D. Smith. Weighting diffraction data. In E. Dodson, M. Moore, A. Ralph, and S. Bailey, editors, *Proceedings of the CCP4 study weekend*, pages 193–200. CCLRC Daresbury Laboratory, Warrington, U.K., 1996.

[222] C. C. F. Blake, W. C. A. Pulford, and P. J. Artymuik. X-ray studies of water in crystals of lysozyme. *Journal of Molecular Biology*, 167:693–723, 1983.

[223] S. E. V. Phillips. Structure and refinement of oxymyoglobin at 1.6 Å resolution. *Journal of Molecular Biology*, 142:531–554, 1980.

[224] A. T. Brünger. *X-PLOR Version 3.1. A system for X-ray crystallography and NMR*. Yale University Press, U.S.A, 1991.

[225] B. P. Schoenborn. Solvent effect in protein crystals. A neutron diffraction analysis of solvent and ion density. *Journal of Molecular Biology*, 201:741–749, 1988.

[226] X. Cheng and B. P. Schoenborn. Hydration in protein crystals. A neutron diffraction analysis of carbonmonoxymyoglobin. *Acta Crystallographica*, B46:195–208, 1990.

[227] R. Langridge, D. A. Marvin, W. E. Seeds, H. C. Wilson, C. W. Hooper, M. H. F. Wilkins, and L. Hamilton. The molecular configuration of deoxyribonucleic acid II. Molecular models and their Fourier transforms. *Journal of Molecular Biology*, 2:38–64, 1960.

[228] P. C. Moews and R. H. Kretsinger. Refinement of the structure of carp muscle calcium-binding parvalbumin by model building and difference Fourier analysis. *Journal of Molecular Biology*, 91:201–228, 1975.

[229] E. T. Copson. An integral-equation method of solving plane diffraction problems. *Proceedings of the Royal Society (London)*, 186 100–118, 1946.

[230] R. H. Blessing and D. A. Langs. *A priori* estimation of scale and overall anisotropic temperature factors from the Patterson origin peak. *Acta Crystallographica*, A44:729–735, 1988.

[231] P. R. Evans. Data reduction. In L. Sawyer, N. Isaacs, and S. Bailey, editors, *Proceedings of the CCP4 Study Weekend*, pages 114-122. SERC Daresbury Laboratory, Warrington, U.K., 1993.

[232] R. H. Blessing. Data reduction and error analysis for accurate single crystal diffraction intensities. *Crystallographic Reviews*, 1:3–58, 1987.

[233] F. T. Burling, W. I. Weis, K. M. Flaherty, and A. T. Brünger. Direct observation of protein solvation and discrete disorder with experimental crystallographic phases. *Science*, 271:72–77, 1996.

[234] D. E. Tronrud. Knowledge-based B-factor retraints for the refinement of proteins. *Journal of Applied Crystallography*, 29:100–104, 1996.

[235] R. Laskowski, M. MacArthur, D. Moss, and J. Thornton. PROCHECK: A program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, 26:283–291, 1993.

[236] G. Scapin, J. S. Blanchard, and J. C. Sacchettini. Three dimensional structure of *Escherichia coli* dihydrodipicolinate reductase. *Biochemistry*, 34:3502–3512, 1995.

[237] C. R. Bellamacina. The nicotinamide dinucleotide binding motif: a comparison of nucleotide binding proteins. *FASEB Journal*, 10:1257–1269, 1996.

[238] M. G. Rossmann, A. Liljas, C.-I. Brändén, and L. J. Banaszak. Evolutionary and structural relationships among dehydrogenases. In P. D. Boyer, editor, *The Enzymes*, volume 11, pages 61–102. Academic Press, New York, 3rd edition, 1975.

[239] J. S. Richardson. The anatomy and taxonomy of protein structure. *Advances in Protein Chemistry*, 34:167–339, 1981.

[240] H. Eklund and C.-I. Brändén. Crystal structure, coenzyme conformations, and protein interactions. In D. Dolphin, O. Avramovic, and R. Poulson, editors, *Pyridine Nucleotide Coenzymes Part A*, pages 51–98. Wiley, New York, 1987.

[241] L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology*, 233:123–138, 1993.

[242] W. T. Wolodko, M. E. Fraser, M. N. G. James, and W. A. Bridger. The crystal structure of succinyl-CoA synthetase from *Esherichia coli* at 2.5 Å resolution. *The Journal of Biological Chemistry*, 269:10883–10890, 1994.

[243] E. G. Hutchinson and J. M. Thornton. PROMOTIF - a program to identify and analyze structural motifs in proteins. *Protein Science*, 5:212–220, 1996.

[244] E. G. Hutchinson and J. M. Thornton. HERA - a program to draw schematic diagrams of protein secondary structures. *Proteins*, 8:203–212, 1990.

[245] M. Buehner, G. C. Ford, D. Moras, K. W. Olsen, and M. G. Rossmann. Three-dimensional structure of D-glyceraldehyde-3-phosphate dehydrogenase. *The Journal of Molecular Biology*, 90:25–49, 1974.

[246] C. Abad-Zapatero, J. P. Griffith, J. L. Sussman, and M. G. Rossmann. Refined crystal structure of dogfish M4 apo-lactate dehydogenase. *Journal of Molecular Biology*, 198:445–467, 1987.

[247] U. Opitz, R. Rudolph, R. Jaenicke, L. Ericsson, and H. Neurath. Proteolytic dimers of porcine muscle lactate dehydrogenase: Characterization, folding, and reconstitution of the truncated and nicked polypeptide chain. *Biochemistry*, 26:1399–1406, 1987.

[248] R. M. Jackson, J. E. Gelpi, A. Cortes, D. C. Emery, H. M. Wilks, K. M. Moreton, D. J. Halsall, R. N. Sleigh, M. Behan-Martin, G. R. Jones, A. C. Clarke, and J. J. Holbrook. Construction of a stable dimer of *Bacillus stearothermophilus* lactate dehydrogenase. *Biochemistry*, 31:8307–8314, 1992.

[249] P. Rowland, A. K. Basak, S. Gover, H. R. Levy, and M. J. Adams. The three-dimensional structure of glucose 6-phosphate dehydrogenase from *Leuconostoc mesenteroides* refined at 2.0 Å resolution. *Structure*, 2:1073–1087, 1994.

[250] M. Carson. Ribbons 2.0. *Journal of Applied Crystallography*, 24:958–961, 1991.

[251] W. O. Barnell, K. C. Yi, and T. Conway. Sequence and genetic organization of a *Zymomonas mobilis* gene cluster that encodes several enzymes of glucose metabolism. *Journal of Bacteriology*, 172:7227–7240, 1990.

[252] J. John, S. J. Crennell, D. W. Hough, M. J. Danson, and G. L. Taylor. The crystal structure of glucose dehydrogenase from *Thermoplasma acidophilum*. *Structure*, 2:385–393, 1994.

[253] George N. Reeke, Jr, J. W. Becker, and G. M. Edelman. The covalent and three-dimensional structure of concanavalin A. *The Journal of Biological Chemistry*, 250:1525–1547, 1975.

[254] C. Chothia and J. Janin. Relative orientation of close-packed β-sheets in proteins. *Proceedings of the National Acadamy of Sciences, U.S.A*, 78:4146–4150, 1981.

[255] F. E. Cohen, M. J. Sternberg, and W. R. Taylor. Analysis of the tertiary structures of protein β-sheet sandwiches. *Journal of Molecular Biology*, 148:253–272, 1981.

[256] F. Eisenhaber and P. Argos. Improved strategy in analytic surface calculation for molecular systems: Handling of singularities and computational efficiency. *Journal of Computational Chemistry*, 14(11):1272–1280, 1993.

[257] W. Saenger. Structure and function of nucleosides and nucleotides. *Angewandte Chemie*, 12:591–601, 1973.

[258] L. W. Tari, A. Matte, U. Pugazhenthi, H. Goldie, and L. T. Delbaere. Snapshot of an enzyme reaction intermediate in the structure of the ATP-Mg^{2+}-oxalate ternary complex of *Escherichia coli* PEP carboxykinase. *Nature Structural Biology*, 3:355–363, 1996.

[259] J. M. Thorn, J. D. Barton, N. E. Dixon, D. L. Ollis, and K. J. Edwards. Crystal structure of *Escherichia coli* QOR quinone oxidoreductase complexed with NADPH. *Journal of Molecular Biology*, 249:785–799, 1995.

[260] P. A. Karplus and G. E. Schulz. Substrate binding and catalysis by glutathione reductase as derived from refined enzyme:substrate crystal structures at 2 Å resolution. *The Journal of Molecular Biology*, 210:163–180, 1989.

[261] D. K. Wilson, K. M. Bohren, K. H. Gabbay, and F. A. Quiocho. An unlikely sugar substrate site in the 1.65 Å structure of the human aldose reductase holoenzyme implicated in diabetic complications. *Science*, 257:81–84, 1992.

[262] F. Mancia, N. H. Keep, A. Nakagawa, P. F. Leadlay, S. McSweeney, B. Rasmussen, P. Bösecke, O. Diat, and P. R. Evans. How coenzyme B_{12} radicals are generated: the crystal structure of methylmalonyl-coenzyme A mutase at 2 Å resolution. *Structure*, 4:339–350, 1996.

[263] J. N. Hope, H.-C. Chen, and J. F. Hejtmancik. βA3/A1-crystallin association: Role of the N-terminal arm. *Protein Engineering*, 7(3):445–451, 1994.

[264] V. Nalini, B. Bax, H. Driessen, D. S. Moss, P. F. Lindley, and C. Slingsby. Close packing of an oligomeric eye lens β-crystallin induces loss of symmetry and ordering of sequence extensions. *Journal of Molecular Biology*, 236:1250–1258, 1994.

[265] S. K. Katti, B. A. Katz, and H. W. Wyckoff. Crystal structure of muconolactone isomerase at 3.3 Å resolution. *Journal of Molecular Biology*, 205:557–571, 1989.

[266] G. Scotland and M. D. Houslay. Chimeric constructs show that the unique N-terminal domain of the cyclic AMP phosphodiesterase RD1 (RNPDE4A1A; rPDE-IV_{a1}) can confer membrane association upon the normally cytosolic protein chloramphenicol acetyltransferase. *Biochemistry Journal*, 308:673–681, 1995.

[267] F. Westheimer. Mechanism of action of the pyridine nucleotides. In D. Dolphin, O. Avramovic, and R. Poulson, editors, *Pyridine Nucleotide Coenzymes Part A*, pages 253–322. Wiley, New York, 1987.

[268] J. J. Birktoft and L. J. Banaszak. The presence of a histidine-aspartic acid pair in the active site of 2-hydroxyacid dehydrogenases. *The Journal of Biological Chemistry*, 258:472–482, 1983.

[269] S. S. Hoog, J. E. Pawlowski, P. M. Alzari, T. M. Penning, and M. Lewis. Three-dimensional structure of rat liver 3α-hydroxysteroid/dihydrodiol dehydrogenase: A member of the aldo-keto reductase superfamily. *Proceedings of The National Acadamy of Sciences*, 91:2517–2521, 1994.

[270] K. M. Bohren, C. E. Grimshaw, C.-J. Lai, D. H. Harrison, D. Ringe, G. A. Petsko, and K. H. Gabbay. Tyrosine-48 is the proton donor and histidine-110 directs substrate stereochemical selectivity in the reduction reaction of human aldose reductase: Enzyme kineteics and crystal strcuture of the Y48H mutant enzyme. *Biochemistry*, 33:2021–2032, 1994.

[271] H. Jörnvall, B. Persson, M. Krook, S. Atrian, R. Gonzàlez-Duarte, J. Jeffery, and D. Ghosh. Short-chain dehydrogenases/reductases (sdr). *Biochemistry*, 34(18):6003–6013, 1995.

[272] S. V. Evans. Setor: hardware lighted three-dimensional solid model representations of macromolecules. *Journal of Molecular Graphics*, 11:134–138, 1993.

[273] J. R> L. Dunbrack and M. Karplus. Conformational analysis of the backbone-dependent rotamer preferences of protein side chains. *Nature Structural Biology*, 1:334–340, 1994.

[274] P. A. Evans, C. M. Dobson, R. A. Kautz, G. Hatfull, and R. O. Fox. Proline isomerism in *Staphylococcal nuclease* characterized by NMR and site directed mutagenesis. *Nature*, 329:266–268, 1987.

[275] D. M. Truckses, J. R. Somoza, K. E. Prehoda, S. C. Miller, and J. L. Markley. Coupling between *trans/cis* proline isomerization and protein stability in *Staphlococcal nuclease*. *Protein Science*, 5:1907–1916, 1996.

[276] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[277] Y. Fujita, K. Shindo, Y. Miwa, and K. Yoshida. *Bacillus subtilis* inositol dehydrogenase-encoding gene (idh): sequence and expression in *Escherichia coli*. *Gene*, 108:121–125, 1991.

[278] Y. Nomura, M. Nakagawa, N. Ogawa, S. Harahima, and Y. Oshima. Genes in pht plasmid encoding the initial degradation pathway of pthalate in *Pseudomonas putida*. *Journal of Fermentation and Bioengineering*, 74:333–344, 1992.

[279] G. J. Barton and M. J. E. Sternberg. A strategy for the rapid multiple alignment of protein sequences. *Journal of Molecular Biology*, 198:327–337, 1987.

[280] M. B. Swindells. Classification of doubly wound nucleotide binding topologies using automated loop searches. *Protein Science*, 2:2146–2153, 1993.

[281] G. J. Barton. ALSCRIPT: a tool to format multiple sequence alignments. *Protein Engineering*, 6:37–40, 1993.

[282] C.-I. Brändén and H. Eklund. Structure and mechanism of liver alcohol dehydrogenase, lactate dehydrogenase and glyceraldehyde-3-phosphate dehydrogenase. In J. Jeffery, editor, *Dehydrogenases Requiring Nicotinamide Coenzymes*. Birkhäuser Verlag, Basel, 1980.

[283] D. Gollhofer, B. Nidetzky, M. Fuerlinger, and K. D. Kulbe. Efficient protection of glucose-fructose oxidoreductase from *Zymomonas mobilis* against irreversible inactivation during its catalytic action. *Enzyme and Microbial Technology*, 17:235–240, 1995.

[284] S. Ferguson. The periplasm. In S. Mohan, C. Dow, and J. Coles, editors, *Prokaryotic Structure and Function*, pages 315–339. Cambridge University Press, 1992.

[285] T. Wiegert, H. Sahm, and G. A. Sprenger. Export of the periplasmic NADP-containing glucose-fructose oxidoreductase of zymomonas mobilis. *Archives of Microbiology*, 166:32–41, 1996.

[286] H. Nikaido and M. H. Saier, Jr. Transport proteins in bacteria: Common themes in their design. *Science*, 258:936–942, 1992.

[287] I. R. Beacham. Periplasmic enzymes in gram-negative bacteria. *International Journal of Biochemistry*, 10:877–883, 1979.

[288] C. Anthony. Quinoproteins and energy transduction. In C. Anthony, editor, *Bacterial Energy Transduction*, pages 293–316. Academic Press, London 1988.

[289] H. Loos, H. Sahm, and G. Sprenger. Glucose-fructose oxidoreductase, a periplasmic enzyme of *Zymomonas mobilis*, is active in its precursor form. *FEMS Microbiology Letters*, 107:293–298, 1993.

[290] A. P. Pugsley. The complete general secretory pathway in gram-negative bacteria. *Microbiological Reviews*, 57:50–108, 1993.

[291] T. A. Rapoport, B. Jungnickel, and U. Kutay. Protein translocation across the eukaryotic endoplasmic reticulum and bacterial inner membranes. *Annual Review of Biochemistry*, 65:271–303, 1996.

[292] N. Sakabe. X-ray diffraction data collection system for modern protein crystallography with a Weissenberg camera and an imaging plate using synchrotron radiation. *Nuclear Instruments and Methods in Physics Research*, A303:448–463, 1991.

[293] A. Kuksis, editor. *Fat Absorption*, volume I & II. CRC Press, Boca Raton, Florida, 1987.

[294] L. Sarda and P. Desnuelle. Action de la lipase pancréatique sur les esters en émulsion. *Biochimica et Biophysica Acta*, 30:513–521, 1958.

[295] P. Desnuelle, L. Sarda, and G. Ailhaud. Inhibition de la lipase pancréatique par le diéthyl-p-nitrophényl phosphate en emulsion. *Biochimica et Biophysica Acta*, 37:570–571, 1960.

[296] G. G. Dodoson, D. M. Lawson, and F. K. Winkler. Structural and evolutionary relationships in lipase mechanism and activation. *Faraday Discussions of the Chemical Society*, 93:95–105, 1992.

[297] Z. S. Derewenda and A. M. Sharp. News from the interface: the molecular structures of triglyceride lipases. *Trends in Biochemical Sciences*, 18:20–25, 1993.

[298] C. Cambillau and H. van Tilbeurgh. Structure of hydrolases: lipases and cellulases. *Current Opinion in Structural Biology*, 3:885–895, 1993.

[299] B. Rubin Grease pit chemistry exposed. *Nature Structural Biology*, 1:568–570, 1994.

[300] Z. S. Derewenda. A twist in the tale of lipolytic enzymes. *Nature Structural Biology*, 2:347–349, 1995.

[301] Z. S. Derewenda and U. Derewenda. Relationships among serine hydrolases: Evidence for a common structural motif in triglyceride lipases and esterases. *Biochemistry and Cell Biology*, 69:842–851, 1991.

[302] D. M. Blow. Structure and mechanism of chymotrypsin. *Accounts of Chemical Research*, 9:145–152, 1976.

[303] T. A. Steitz and R. G. Shulman. Crystallographic and NMR studies of the serine proteases. *Annual Review of Biophysics and Bioengineering*, 11:419–444, 1982.

[304] J. J. Perona and C. S. Craik. Structural basis of substrate specificity in the serine proteases. *Protein Science*, 4:337–360, 1995.

[305] A. Hjorth, F. Carrière, C. Cudrey, H. Wöldike, E. Boel, D. M. Lawson, F. Ferrato, C. Cambillau, G. G. Dodson, L. Thim, and R. Verger. A structural domain (the lid) found in pancreatic lipases is absent in the guinea pig (phospho)lipase. *Biochemistry*, 32:4702–4707, 1993.

[306] F. Carrière, K. Thirstrup, E. Boel, R. Verger, and L. Thim. Structure-function relationships in naturally occuring mutants of pancreatic lipase. *Protein Engineering*, 7:563–569, 1994.

[307] M. L. Jennens and M. E. Lowe. A surface loop covering the active site of human pancreatic lipase influences interfacial activation and lipid binding. *Journal of Biological Chenistry*, 269:25470–25474, 1994.

[308] S. H. Northrup. Activation of an enzyme simulated by explicit dynamics of an active site lid. *Biophysical Journal*, 71:3, 1996.

[309] G. H. Peters, O. H. Olson, A. Svendsen, and R. C. Wade. Theoretical investigation of the dynamics of the active site lid in *Rhizomucor miehei* lipase. *Biophysical Journal*, 71:119–129, 1996.

[310] C. Martinez, A. Nicolas, H. van Tilbeurgh, M. P. Egloff, C. Cudrey, R. Verger, and C. Cambillau. Cutinase, a lipolytic enzyme with a preformed oxyanion hole. *Biochemistry*, 33:83–89, 1994.

[311] B. van den Berg, M. Tessari, R. Boelens, R. Dijkman, G. H. de Haas, R. Kaptein, and H. M. Verheij. NMR structures of phospholipase A_2 reveal conformational changes during interfacial activation. *Nature Structural Biology*, 2:402–406, 1995.

[312] M. F. Roberts. Phospholipases: structural and functional motifs for working at an interface. *FASEB journal*, 10:1159–1172, 1996.

[313] F. K. Winkler, A. D'Arcy, and W. Hunziker. Structure of human pancreatic lipase. *Nature*, 343:771–774, 1990.

[314] H. van Tilbeurgh, L. Sarda, R. Verger, and C. Cambillau. Structure of the pancreatic lipase-procolipase complex. *Nature*, 359:159–162, 1992.

[315] H. van Tilbeurgh, M. P. Egloff, C. Martinez, N. Rugani, R. Verger, and C. Cambillau. Interfacial activation of the lipase-procolipase complex by mixed micelles revealed by X-ray crystallography. *Nature*, 362:814–820, 1993.

[316] Y. Bourne, C. Martinez, B. Kerfelec, D. Lombardo, C. Chapus, and C. Cambillau. Horse pancreatic lipase. the crystal structure refined at 2.3 Å resolution. *Journal of Molecular Biology*, 238:709–732, 1994.

[317] M. P. Egloff, F. Marguet, G. Buono, R. Verger, C. Cambillau, and H. van Tilbeurgh. The 2.46 Å structure of the pancreatic lipase-colipase complex inhibited by a C11 alkyl phophonate. *Biochemistry*, 34:2751–2762, 1995.

[318] P. Grochulski, Y. Li, J. D. Schrag, F. Bouthillier, P. Smith, D. Harrison, B. Rubin, and M. Cygler. Insights into interfacial activation from an open structure of *Candida rugosa* lipase. *The Journal of Biological Chemistry*, 268:12843–12847, 1993.

[319] P. Grochulski, Y. Li, J. D. Schrag, and M. Cygler. Two conformational states of *Candida rugosa* lipase. *Protein Science*, 3:82–91, 1994.

[320] A. M. Brzozowski, U. Derewenda, Z. S. Derewenda, G. G. Dodson, D. M. Lawson, J. P. Turkenburg, F. Bjorkling, B. Huge-Jensen, S. A. Patkar, and L. Thim. A model for interfacial activation in lipases from the structure of a fungal lipase-inhibitor complex. *Nature*, 351:491–494, 1991.

[321] U. Derewenda, A. M. Brzozowski, D. M. Lawson, and Z. S. Derewenda. Catalysis at the interface: The anatomy of a conformational change in a triglyceride lipase. *Biochemistry*, 31:1532–1541, 1992.

[322] D. M. Lawson. Probing the nature of substrate binding in *Humicola lanuginosa* lipase through X-ray crystallography and intuitive modelling. *Protein Engineering*, 7:543–550, 1994.

[323] U. Derewenda, L. Swenson, Y. Wei, R. Green, P. M. Kobos, R. Joerger, M. J. Haas, and Z. S. Derewenda. Conformational lability of lipases observed in the absence of an oil water interface: crystallographic studies of enzymes from the fungi *Humicola lanuginosa* and *Rhizopus delemar*. *Journal of Lipid Research*, 35:524–534, 1994.

[324] J. Hermoso, D. Pignol, B. Kerfelec, I. Crenon, C. Chapus, and J. C. Fontecilla-Camps. Lipase activation by non-ionic detergents. *The Journal of Biological Chemistry*, 271:18007–18016, 1996.

[325] M. E. M. Noble, A. Cleasby, L. N. Johnson, M. R. Egmond, and L. G. J. Frenken. Analysis of the structure of *Pseudomonas glumae* lipase. *Protein Engineering*, 7:559–562, 1994.

[326] J. Uppenberg, M. T. Hansen, S. Paktar, and T. A. Jones. The sequence, crystal structure determination and refinement of two crystal forms of lipase B from *Candida antarctica*. *Structure*, 2:293–308, 1994.

[327] P. J. Kraulis. Molscript: A program to produce both detailed and schematic plots of protein structures. *Journal of Applied Crystallography*, 24:946–950, 1991.

[328] A. B. Marfan. Allaitement naturel et allaitement artificiel. *Presse Med.*, 9:13–16, 1901.

[329] E. Freudenberg. *Die Frauenmilchlipase*. J. Karger, Basel, 1953.

[330] O. Hernell and T. Olivecrona. Human milk lipases. I. Serum-stimulated lipase. *Journal of Lipid Research*, 15:367–374, 1974.

[331] O. Hernell and T. Olivecrona. Human milk lipases II. Bile salt-stimulated lipase. *Biochimica et Biophysica Acta*, 369:234–244, 1974.

[332] O. Hernell. Human milk lipases III. Physiological implication of the bile salt-stimulated lipase. *European Journal of Clinical Investigation*, 5:267–272, 1975.

[333] O. Hernell, L. Bläckberg, and S. Bernbäck. Milk lipases and *in vivo* lipolysis. In S. A. Atkinson and B. Lonnerdäl, editors, *Protein and non-protein nitrogen in human milk*, pages 221–236. CRC Press, Boca Raton, 1989.

[334] L. Bläckberg and O. Hernell. Utilization of human milk fat: Biochemical and physical-chemical concepts. In *Mechanisms regulating lactation and infant nutritient utilization*, pages 241–258.Wiley-Liss 1992.

[335] D. Lombardo, O. Guy, and C. Figarella. Purification and characterization of a carboxyl ester hydrolase from human pancreatic juice. *Biochinica et Biophysica Acta*, 527:142–149, 1978.

[336] L. Bläckberg, D. Lombardo, O. Hernell, O. Guy, and T. Olivecrona. Bile salt-stimulated lipase in human milk and carboxyl ester hydrolase in pancreatic juice: Are they identical enzymes? *FEBS Letters*, 136:284–288, 1981.

[337] J. Nilsson, L. Bläckberg, P. Carlsson, S. Enerbäck, O. Hernell, and G. Bjursell. cDNA cloning of human-milk bile-salt-stimulated lipase and evidence for its identity to pancreatic carboxylic ester hydrolase. *European Journal of Biochemistry*, 192:543–550, 1990.

[338] D. Y. Hui and J. A. Kissel. Sequence identity between human pancreatic cholesterol esterase and bile salt-stimulated milk lipase. *FEBS Letters*, 276:131–134, 1990.

[339] T. Baba, D. Downs, K. Jackson, J. Tang, and C. S. Wang. Structure of human milk bile salt activated lipase. *Biochemistry*, 30:500–510, 1991.

[340] K. Reue, J. Zambaux, H. Wong, G. Lee, T. H. Leete, M. Ronk, J. E. Shively, B. Sternby, B. Borgström, D. Ameis, and M. C. Schotz. cDNA cloning of carboxyl ester lipase from human pancreas reveals a unique proline-rich repeat unit. *Journal of Lipid Research*, 32:267–276, 1991.

[341] A. K. Taylor, J. L. Zambaux, I. Klisak, T. Mohandos, R. S. Sparkes, M. C. Schotz, and A. J. Lusis. Carboxyl ester lipase: A highly polymorphic locus on human chromosome 9qter. *Genomics*, 10:425–431, 1991.

[342] U. Lidberg, J. Nilsson, K. Strömberg, G. Stenman, P. Sahlin, S. Enerbåck, and G. Bjursell. Genomic organization, sequence analysis, and chromosomal localization of the human carboxyl ester lipase (CEL) gene and a CEL-like (CELL) gene. *Genomics*, 13:630–640, 1992.

[343] C. S. Wang and J. A. Hartsuck. Bile salt-activated lipase. a multiple function lipolytic enzyme. *Biochimica et Biophysica Acta*, 1166:1–19, 1993.

[344] E. A. Rudd and H. L. Brockman Pancreatic carboxyl ester lipase (cholesterol esterase). In B. Borgström and H. L. Brockman, editors, *Lipases*, pages 185–204. Elsevier, Amsterdam, 1984.

[345] T. Olivecrona and G. Bengtsson. Lipases in milk. In B. Borgström and H. L. Brockman, editors, *Lipases*, pages 205–261. Elsevier, Amsterdam, 1984.

[346] D. Y. Hui, K. Hayakawa, and J. Oizumi. Lipoamidase activity in normal and mutagenized pancreatic cholesterol esterase (bile salt-stimulated lipase). *Biochemical Journal*, 291:65–69, 1993.

[347] J. E. Staggers, O. Hernell, R. J. Stafford, and M. C. Carey. Physical-chemical behaviour of dietary and biliary lipids during intestinal digestion and absorption. 1. Phase behaviour and aggregation states of model lipid systems patterned after aqueous duodenal contents of healthy adult human beings. *Biochemistry*, 29:2028–2040, 1990.

[348] O. Hernell, J. E. Staggers, and M. C. Carey. Physical-chemical behaviour of dietary and biliary lipids during intestinal digestion and absorption. 2. Phase analysis and aggregation states of luminal lipids during duodenal fat digestion in healthy adult human beings. *Biochemistry*, 29:2041–2056, 1990.

[349] E. Freudemberg. A lipase in the milk of the gorilla. *Experentia*, 22:317, 1966.

[350] L. Bläckberg, O. Hernell, T. Olivecrona, L. Domellöf, and R. Malinov. The bile salt-stimulated lipase in human milk is an evolutionary newcomer derived from a non-milk protein. *FEBS letters*, 112:51–54, 1980.

[351] L. M. Freed, C. M. York, M. Hamosh, J. A. Sturman, and P. Hamosh. Bile salt-stimulated lipase in non-primate milk: longitudinal variation and lipase characteristics in cat and dog milk. *Biochimica et Biophysica Acta*, 878:209–215, 1986.

[352] L. A. Ellis and M. Hamosh. Bile salt stimulated lipase: Comparitive studies in ferret milk and lactating mammary gland. *Lipids*, 27:917–922, 1992.

[353] A. S. Lidmer, M. Kannius, L. Lundberg, G. Bjursell, and J. Nilsson. Molecular cloning and characterization of the mouse carboxyl ester lipase gene and evidence for expression in the lactating mammary gland. *Genomics*, 29:115–122, 1995.

[354] S. H. L. Pan, C. W. Dill, E. S. Alford, R. L. Richter, and C. Garza. Heat inactivation of bile salt-stimulated lipase activity in human milk and colostrum. *Journal of Food Protection*, 46:525–527, 1983.

[355] S. Williamson, E. Finucane, H. Ellis, and R. H. Gamsu. Effect of heat treatment of human milk on absorption of nitrogen, fat, sodium, calcium, and phosphorous by preterm infants. *Archives of Disease in Childhood*, 53:555–563, 1978.

[356] B. Alemi, M. Hamosh, J. W. Scanlon, C. Salzman-Mann, and P. Hamosh. Fat digestion and very low-birth-weight infants: Effect of addition of human milk to low-birth-weight formula. *Pediatrics*, 68:484–489, 1981.

[357] S. A. Atkinson, M. H. Bryan, and G. H. Anderson. Human milk feeding in premature infants: protein fat and carbohydrate balances in the first two weeks of life. *Journal of Pediatrics*, 99:617–624, 1981.

[358] C.-S. Wang, M. E. Martindale, M. M. King, and J. Tang. Bile salt-activated lipase: Effect on kitten growth rate. *American Journal of Clinical Nutrition*, 49:457–463, 1989.

[359] V. Sbarra, E. Mas, T. R. Henderson, M. Hamosh, D. Lombardo, and P. Hamosh. Digestive lipases of the newborn ferret: compensatory role of milk bile salt-dependent lipase. *Pediatric Research*, 40:263–268, 1996.

[360] S. Bernbäck, L. Bläckberg, and O. Hernell. The complete digestion of human milk triacylgycerol *in vitro* requires gastric lipase, pancreatic colipase-dependent lipase and bile salt-stimulated lipase. *Journal of Clinical Investigation*, 85:1221–1226, 1990.

[361] S. J. Iverson, C. L. Kirk, M. Hamosh, and J. Newsome. Milk lipid digestion in the neonatal dog: the combined actions of gastric and bile salt stimulated lipases. *Biochimica et Biophysica Acta*, 1083:109–119, 1991.

[362] O. Hernell, L. Bläckberg, Q. Chen, B. Sternby, and A. Nilsson. Does the bile salt-stimulated lipase of human milk have a role in the use of the milk long-chain polyunsaturated fatty acids? *Journal of Pediatric Gastroenterology and Nutrition*, 16:426–431, 1993.

[363] O. Hernell and L. Bläckberg. Digestion and absorption of human milk lipids. In R. Dulbecco, editor, *Encyclopedia of Human Biology, Volume 3*, pages 47–56. Academic Press, San Diego, 1991.

[364] D. R. Gjellesvik, D. Lombardo, and B. Walther. Pancreatic bile salt dependent lipase from cod (*Gadus morhua*): purification and properties. *Biochimica et Biophysica Acta*, 1124:123–134, 1992.

[365] M. B. Lindström, B. Sternby, and B. Borgström. Concerted action of human carboxyl ester lipase and pancreatic lipase during lipid digestion *in vitro*: Importance of the physiochemical state of the substrate. *Biochimica et Biophysica Acta*, 959:178–184, 1988.

[366] M. B. Lindström, J. Persson, L. Thurn, and B. Borgström. Effect of pancreatic phospholipase A_2 and gastric lipase on the action of pancreatic carboxyl ester lipase against lipid substrates *in vitro*. *Biochimica et Biophysica Acta*, 1084:194–197, 1991.

[367] A. Lopez-Candales, M. S. Bosner, C. A. Spilburg, and L. G. Lange. Cholesterol transport function of pancreatic cholesterol esterase: Directed sterol uptake and esterfication in enterocytes. *Biochemistry*, 32:12085–12089, 1993.

[368] S. C. Myers-Payne, D. Y. Hui, H. L. Brockman, and F. Schroeder. Cholesterol esterase: A cholesterol transfer protein. *Biochemistry*, 34:3942–3947, 1995.

[369] R. Shamir, W. J. Johnson, R. Zolfaghari, H. Sook Lee, and E. A. Fisher. Role of bile salt-dependent cholesteryl ester hydrolase in the uptake of micellar cholesterol by intestinal cells. *Biochemistry*, 34:6351–6358, 1995.

[370] P. N. Howles, C. P. Carter, and D. Y. Hui. Dietary free and esterfied cholesterol absorption in cholesterol esterase (bile salt-stimulated lipase) gene-targeted mice. *The Journal of Biological Chemistry*, 271:7196–7202, 1996.

[371] J. Brodt-Eppley and D. Y. Hui. Calcium mobilization and protein kinase Cactivation are required for cholecystokinin stimulation of pancreatic cholesterol esterase secretion. *Biochemistry Journal*, 306:605–608, 1995.

[372] J. Brodt-Eppley and D. Y. Hui. Dietary regulation of cholesterol esterase mRNA level in rat pancreas. *Journal of Lipid Research*, 35:27–35, 1994.

[373] Y. Hunag and D. Y. Hui. Increased cholesterol esterase level by cholesterol loading of rat pancreatoma cells. *Biochimica et Biophysica Acta*, 1214:317–322, 1994.

[374] F. Li, Y. Huang, and D. Y. Hui. Bile salt stimulated cholesterol esterase increases uptake of high density lipoprotein-associated cholesteryl esters by hepg2 cells. *Biochemistry*, 35:6657–6663, 1996.

[375] C. J. Rojas and E. H. Harrison. Bile salt-dependent and bile salt-independent cholesteryl ester hydrolase activities in rat liver cytosol. *Proceedings of the Society for Experimental Medicine and Biology*, 206:60–68, 1994.

[376] K. E. Winkler, E. H. Harrison, J. B. Marsh, J. M. Glick, and A. C. Ross. Characterization of a bile salt-dependent cholesteryl ester hydrolase activity secreted from hepG2 cells. *Biochimica et Biophysica Acta*, 1126:151–158, 1992.

[377] R. Zolfaghari, E. H. Harrison, A. C. Ross, and E. A. Fisher. Expression in *Xenopus oocytes* of rat liver mRNA coding for a bile salt-dependent cholesteryl ester hydrolase. *Proceedings of the National Academy of Sciences USA*, 86:6913–6916, 1989.

[378] J. A. Kissel, R. Fontaine, C. W. Turck, H. L. Brockman, and D. Y. Hui. Molecular cloning and expression of cDNA for rat pancreatic cholesterol esterase. *Biochimica et Biophysica Acta*, 1006:227–236, 1989.

[379] E. H. Harrison. Bile salt-dependent, neutral cholesteryl ester hydrolase of rat liver: Possible relationship with pancreatic cholesteryl ester hydrolase. *Biochimica et Biophysica Acta*, 963:28–34, 1988.

[380] E. D. Camulli, M. J. Linke, H. L. Brockman, and D. Y. Hui. Identity of a cytosolic neutral cholesterol esterase in rat liver with the bile salt stimulated cholesterol esterase in pancreas. *Biochimica et Biophysica Acta*, 1005:177–182, 1989.

[381] J. Nilsson, M. Hellquist, and G. Bjursell. The human carboxyl ester lipase-like (CELL) gene is ubiquitously expressed and contains a hypervariable region. *Genomics*, 17:416–422, 1993.

[382] K. Mackay and R. Lawn. Characterization of the mouse pancreatic/mammary gland cholesterol esterase-encoding cDNA and gene. *Gene*, 165:255–259, 1995.

[383] S. Ghosh, D. H. Malloneee, P. B. Hylemon, and W. M. Grogan Molecular cloning and expression of rat hepatic neutral cholesteryl ester hydrolase. *Biochimica et Biophysica Acta*, 1259:305–312, 1995.

[384] F. W. Holtsberg, L. E. Ozgur, D. E. Garsetti, J. Myers, R. W. Egan, and M. A. Clark. Presence in human eosinophils of a lysophospholipase similar to that found in the pancreas. *Biochemistry Journal*, 309:141–144, 1995.

[385] R. Shamir, W. J. Johnson, K. Morlock-Fitzpatrick, R. Zolfaghari, L. Li, E. Mas, D. Lombardo, D. W. Morel, and E. A. Fisher. Pancreatic carboxyl ester lipase: A circulating enzyme that modifies normal and oxidized lipoproteins *in vitro*. *Journal of Clinical Investigation*, 97:1696–1704, 1996.

[386] D. Lombardo, G. Montalto, S. Roudani, E. Mas, R. Laugier, V. Sbarra, and N. Abouakil. Is bile salt-dependent lipase concentration in serum of any help in pancreatic cancer diagnosis? *Pancreas*, 8:581–588, 1993.

[387] S. Roudani, F. Miralles, A. Margotat, M. J. Escribano, and D. Lombardo. Bile salt-dependent lipase transcripts in human fetal tissues. *Biochimica et Biophysica Acta*, 1264:141–150, 1995.

[388] B. V. Kumar, J. A. Aleman-Gomez, N. Colwell, A. Lopez-Candales, M. S. Bosner, C. A. Spilburg, M. Lowe, and L. G. Lange. Structure of the human pancreatic cholesterol esterase gene. *Biochemistry*, 31:6077–6081, 1992.

[389] A. J. Jeffreys, N. J. Royle, V. Wilson, and Z. Wong. Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature*, 332:278–281, 1988.

[390] Y. Nakamura, M. Leppert, P. O'Connell, R. Wolff, T. Holm, M. Culver, C. Martin, E. Fujimoto, M. Hoff, E. Kumlin, and R. White. Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science*, 235:1616–1622, 1987.

[391] N. Bruneau and D. Lombardo. Chaperone function of a grp 94-related protein for folding and transport of the pancreatic bile salt dependent lipase. *The Journal of Biological Chemistry*, 270:13524–13533, 1995.

[392] N. Bruneau, P. Lechene de la Porte, V. Sbarra, and D. Lombardo. Association of bile-salt-dependent lipase with membranes of human pancreatic microsomes. *European Journal of Biochemistry*, 233:209–218, 1995.

[393] R. N. Fonatine, C. P. Carter, and D. Y. Hui. Structure of the rat pancreatic cholesterol esterase gene. *Biochemistry*, 30:7008–7014, 1991.

[394] J. H. Han, C. Stratowa, and W. J. Rutter. Isolation of full length putative rat lysophospholiapse cDNA using improved methods for mRNA isolation and cDNA cloning. *Biochemistry*, 26:1617–1625, 1987.

[395] E. M. Kyger, R. C. Wiegand, and L. G. Lange. Cloning of the bovine pancreatic cholesterol esterase/ lysophopholipase. *Biochemical and Biophysical Research Communications*, 164:1302–1309, 1989.

[396] N. S. Colwell, J. A. Aleman-Gomez, and B. V. Kumar. Molecular cloning and expression of rabbit pancreatic cholesterol esterase. *Biochimica et Biophysica Acta*, 1172:175–180, 1993.

[397] D. R. Gjellesvik, J. B. Lorens, and R. Male. Pancreatic carboxylester lipase from atlantic salmon (*Salmo salar*): cDNA sequence and computer-assisted modelling of tertiary structure. *European journal of Biochemistry*, 226:603–612, 1994.

[398] D. L. Ollis, E. Cheah, M. Cygler, B. Dijkstra, F. Frolow, S. M. Franken, M. Harel, S. J. Remington, I. Silman, J. Schrag, S. J. L., K. H. G. Vershueren, and A. Goldman. The α/β hydrolase fold. *Protein Engineering*, 5:197–211, 1992.

[399] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: Astructural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 1995.

[400] R. D. Newcomb, P. D. East, R. J. Russell, and J. G. Oakeshott. Isolation of alpha cluster esterase genes associated with organophosphate resistance in *Lucilia cuprina*. *Insect Molecular Biology*, In Press, 1996.

[401] J. Vaughan, V. znd Hemingway. Mosquito carboxylesterase Est$\alpha 2^1$ (A_2). *The Journal of Biological Chemistry*, 270:17044–17049, 1995.

[402] E. Krejci, N. Duval, A. Chatonnet, P. Vincens, and J. Massoulié. Cholinesterase-like domains in enzymes and structural proteins: Functional and evolutionary relationships and identification of a catalytically essential aspartic acid. *Proceedings of the Natural Acadamy of Sciences USA*, 88:6647–6651, 1991.

[403] M. Cygler, J. D. Schrag, J. L. Sussman, M. Harel, I. Silman, M. K. Gentry, and B. P. Doctor. Relationship between sequence conservation and three-dimensional structure in a large family of esterases, lipases and related proteins. *Protein Science*, 2:366–382, 1993.

[404] J. L. Sussman, M. Harel, F. Frolow, C. Oefner, A. Goldman, L. Toker, and I. Silman. Atomic structure of acetlycholineseterase from *Torpedo californica*: A prototypic acetylcholine-binding protein. *Science*, 253:872–878, 1991.

[405] J. D. Schrag, Y. Li, S. Wu, and M. Cygler. Ser-His-Glu triad forms the catalytic site of the lipase from *Geotrichum candidum*. *Nature*, 351:761–764, 1991.

[406] J. D. Schrag and M. Cygler. 1.8 Å refined structure of the lipase from *Geotrichum candidum*. *Journal of Molecular Biology*, 230:575–591, 1993.

[407] N. Abouakil, E. Rogalska, and D. Lombrado. Human milk bile-salt stimulated lipase: further investigations on the amino-acids residues involved in the catalytic site. *Biochimica et Biophysica Acta*, 1002:225–230, 1989.

[408] L. P. DiPersio, R. N. Fontaine, and D. Y. Hui. Identification of the active site serine in pancreatic cholesterol esterase by chemical modification and site-specific mutagenesis. *The Journal of Biological Chemistry*, 265:16801–16806, 1990.

[409] L. P. DiPersio, R. N. Fontaine, and D. Y. Hui. Site-specific mutagenesis of an essential histidine residue in pancreatic cholesterol esterase. *The Journal of Biological Chemistry*, 266:4033–4036, 1991.

[410] L. P. DiPersio and D. Y. Hui. Aspartic acid 320 is required for optimal activity of rat pancreatic cholesterol esterase. *Journal of Biological Chemistry*, 268:300–304, 1993.

[411] N. Abouakil, E. Rogalska, J. Bonicel, and D. Lombardo. Purification of pancreatic carboxylic-ester hydrolase by immunoaffinity and its application to the human bile-salt-stimulated lipase. *Biochimica et Biophysica Acta*, 961:299–308, 1988.

[412] R. A. Dwek. Glycobiology: Toward understanding the function of sugars. *Chemical Reviews*, 96:683–720, 1996.

[413] H. Lis and N. Sharon. Protein glycosylation. Structural and functional aspects. *European Journal of Biochemistry*, 218:1–27, 1993.

[414] G. Opdenakker, P. M. Rudd, C. P. Ponting, and R. A. Dwek. Concepts and principles of glycobiology. *The FASEB journal*, 7:1330–1337, 1993.

[415] T. Sugo, E. Mas, N. Abouakil, T. Endo, M. J. Escribano, A. Kobata, and D. Lombardo. The structure of N-linked oligosaccharides of human pancreatic bile-salt-dependent lipase. *European Journal of Biochemistry*, 216:799–805, 1993.

[416] E. Mas, N. Abouakil, S. Roudani, J. L. Franc, J. Montreuil, and D. Lombardo. Variation of the glycosylation of human pancreatic bile-salt-dependent lipase. *European Journal of Biochemistry*, 216:807–812, 1993.

[417] L. Hansson, L. Bläckberg, M. Edlund, L. Lundberg, M. Strömqvist, and O. Hernell. Recombinant human milk bile salt-stimulated lipase: Catalytic activity is retained in the absence of glycolsylation and the unique proline-rich repeats. *The Journal of Biological Chemistry*, 268:26692–26698, 1993.

[418] D. Downs, Y. Y. Xu, J. Tang, and C. S. Wang. Proline-rich domain and glycosylation are not essential for the enzymic activity of bile salt-activated lipase. Kinetic studies of t-BAL, a truncated form of the enzyme , expressed in *Escherichia coli*. *Biochemistry*, 33:7979–7985, 1994.

[419] N. Abouakil, E. Mas, N. Bruneau, A. Benajiba, and D. Lombardo. Bile salt-dependent lipase biosynthesis in rat pancreatic AR 4-2 J cells. *The Journal of Biological Chemistry*, 268:25755–25763, 1993.

[420] K. R. Morlock-Fitzpatrick and E. A. Fisher. The effects of O- and N-linked glycosylation on the secretion and bile salt-stimulation of pancreatic carboxyl ester lipase activity. *Proceedings of the Society for Experimental Biology and Medicine*, 208:186–190, 1995.

[421] L. Bläckberg and O. Hernell. Bile salt-stimulated lipase in human milk: Evidence that bile salt induces lipid binding and activation via binding to different sites. *FEBS Letters*, 323:207–210, 1993.

[422] L. Bläckberg, M. Strömqvist, M. Edlund, K. Juneblad, L. Lundberg, L. Hansson, and O. Hernell. Recombinant human-milk bile-salt-stimulated lipase: Functional properties are retained in the absence of glycosylation and the unique proline-rich repeats. *European Journal of Biochemistry*, 228:817–821, 1995.

[423] C. S. Wang, A. Dashti, K. Jackson, J. C. Yeh, R. D. Cummings, and J. Tang. Isolation and characterization of human milk bile salt-activated lipase C-tail fragment. *Biochemistry*, 34:10639–10644, 1995.

[424] K. M. Loomes. Structural organization of human bile-salt-activated lipase probed by limited proteolysis and expresion of a recombinant trunctaed variant. *European Journal of Biochemistry*, 230:607–613, 1995.

[425] J. Hilkens, M. J. L. Ligtenberg, H. L. Vos, and S. V. Litvinov. Cell membrane-associated mucins and their adhesion-modulating properties. *Trends in Biochemical Sciences*, 17:359–363, 1992.

[426] N. Jentoft. Why are proteins O-glycosylated. *Trends in Biochemical Sciences*, 15:294, 1990.

[427] S. E. Harding. The macrostructure of mucus glycoproteins in solution. *Advances in Carbohydrate Chemistry and Biochemistry*, 47:345–381, 1989.

[428] M. Reschsteiner. Regulation of enzyme levels by proteolysis: the role of PEST regions. *Advances in Enzyme Regulation*, 27:135–151, 1988.

[429] C. S. Wang and K. Johnson. Purification of human milk bile salt-activated lipase. *Analytical Biochemistry*, 133:457–461, 1983.

[430] R. L. Jackson, S. J. Busch, and A. D. Cardin. Glycosaminoglycans: Molecular properties, protein interactions, and role in physiological processes. *Physiological Reviews*, 71:481–539, 1991.

[431] L. Kjellén and U. Lindahl. Proteoglycans: Structures and interactions. *Annual Review of Biochemistry*, 60:443–475, 1991.

[432] J. L. Madara and J. S. Trier. Functional morphology of the mucosa of the small intestine. In L. R. Johnson, editor, *Physiology of the Gastrointestinal Tract, Second Edition*, pages 1209–1249. Raven Press, New York, 1987.

[433] M. S. Bosner, T. Gulick, D. J. S. Riley, C. A. Spilburg, and L. G. Lange. Receptor-like function of heparin in the binding and uptake of neutral lipids. *Proceedings of the National Academyof Science USA*, 85:7438–7442, 1988.

[434] M. Withiam-Leitch, R. P. Rubin, S. E. Koshlukova, and J. M. Aletta. Identification and characterization of carboxyl ester hydrolase as a phospholipid hydrolyzing enzyme of zymogen granule membranes from rat exocrine pancreas. *The Journal of Biological Chemistry*, 270:3780–3787, 1995.

[435] M. Shimizu, N. Uryu, and K. Yamauchi. Preparation of heparan sulfate in the fat globule membrane of bovine and human milk. *Agricultural and Biological Chemistry*, 45:741–745, 1981.

[436] D. Spillmann and U. Lindahl. Glycosaminoglycan-protein interactions: a question of specificity. *Current Opinion in Structural Biology*, 4:677–682, 1994.

[437] R. Coleman. Bile salts and biliary lipids. *Biochemical Society Transactions*, 15:68S–80S, 1987.

[438] C. J. O'Conner and R. G. Wallace. Physico-chemical behaviour of bile salts. *Advances in Colloid and Interface Science*, 22:1–111, 1985.

[439] D. Lombardo, D. Campese, L. Multigner, H. Lafont, and A. de Caro. On the probable involvement of arginine residues in the bile-salt-binding site of human pancreatic carboxylic ester hydrolase. *European Journal of Biochemistry*, 133:327–333, 1983.

[440] T. Tsujita, N. K. Mizuno, and H. L. Brockman. Nonspecific high affinity binding of bile salts to carboxylester lipases. *Journal of Lipid Research*, 28:1434–1443, 1987.

[441] P. W. Jacobson, P. W. Wiesenfeld, L. L. Gallo, R. L. Tate, and J. C. Osborne, Jr. Sodium cholate-induced changes in the conformation and activity of rat pancreatic cholesterol esterase. *The Journal of Biological chemistry*, 265:515–521, 1990.

[442] A. F. Hofmann, J. Sjövall, G. Kurz, A. Radominska, C. D. Schteingart, G. S. Tint, Z. R. Vlahcevic, and K. D. R. Setchell. A proposed nomenclature for bile acids. *Journal of Lipid Research*, 33:599–604, 1992.

[443] M. F. Maylié, M. Charles, and P. Desnuelle. Action of organophosphates and sulfonyl halides on porcine pancreatic lipase. *Biochimica et Biophysica Acta*, 276:162–175, 1972.

[444] L. Lessinger. Cholic acid monohydrate, $C_{24}H_{40}O_5.H_2O$. *Crystal Structure Communications*, 11:1787–1792, 1982.

[445] B. Borgström and C. Erlanson. Pancreatic lipase and co-lipase. Interactions and effects of bile salts and other detergents. *European Journal of Biochemistry*, 37:60–68, 1973.

[446] C. Erlanson-Albertsson. Pancreatic colipase. Structural and physiological aspects. *Biochimica et Biophysica Acta*, 1125:1–7, 1992.

[447] B. Lonnerdäl. Recombinant human milk proteins - an opportunity and a challenge. *American Journal of Clinical Nutrition*, 63:622S–626S, 1996.

[448] P. Grochulski, F. Bouthillier, R. J. Kazlauskas, A. N. Serrequi, J. D. Schrag, E. Ziomek, and M. Cygler. Analogs of reaction intermediates identify a unique substrate binding site in *Candida rugosa* lipase. *Biochemistry*, 33:3494–3500, 1994.

[449] T. Higashi. Auto-indexing of oscillation images. *Journal of Applied Crystallography*, 23:253–257, 1990.

[450] W. Kabsch. Automatic processing of rotation diffraction data from crystals of initially unkown symmetry and cell constants. *Journal of Applied Crystallography*, 26:795–800, 1993.

[451] T. Higashi. The processing of diffraction data taken on a screenless Weissenberg camera for macromolecular crystallography. *Journal of Applied crystallography*, 22:9–18, 1989.

[452] B. A. Fields, J. M. Guss, M. C. Lawrence, and A. Nakagawa. The Weissenberg method for the collection of X-ray diffraction data from macromolecular crystals: Modifications to the data-processing program WEIS. *Journal of Applied Crystallography*, 25:809–811, 1992.

[453] C. J. Gilmore. Maximum entropy and Bayesian statistics in crystallography: a review of practical applications. *Acta Crystallographica*, A52:561–589, 1996.

[454] S. Sheriff and W. A. Hendrickson. Description of overall anisotropy in diffraction from macromolecular crystals. *Acta crystallographica*, A43:118–121, 1987.

[455] R. H. Blessing, D. Y. Guo, and D. A. Langs. Statistical expectation value of the Debye-Waller factor and E(hkl) values for macromolecular crystals. *Acta Crystallographica*, D52:257–266, 1996.

[456] A. J. Dobbs, B. F. Anderson, H. R. Faber, and E. N. Baker. Three-dimensional structure of cytochrome c' from two *alcaligenes* species and the implications for four-helix bundle structures. *Acta Crystallographica*, D52:356–368, 1996.

[457] R. X. Fischer and E. Tillmanns. The equivalent isotropic displacement factor. *Acta Crystallographica*, C44:775–776, 1988.

[458] J. T. Finch, R. S. Brown, D. Rhodes, T. Richmond, B. Rushton, L. C. Lutter, and A. Klug. X-ray diffraction study of a new crystal form of the nucleosome core showing higher resolution. *Journal of Molecular Biology*, 145:757–769, 1981.

[459] G. F. X. Schertler, H. D. Bartunik, H. Michel, and D. Oesterhelt. Orthorhombic crystal form of bacteriorhodopsin nucleated on benzamide diffracting to 3.6 Å resolution. *Journal of Molecular Biology*, 234:156–164, 1993.

[460] T.-P. Ko, J. D. Ng, J. Day, A. Greenwood, and A. McPherson. Determination of three crystal structures of canavalin by molecular replacement. *Acta Crystallographica*, D49:478–489, 1993.

[461] M. Z. Papiz and S. M. Prince. Group anisotropic thermal parameter refinement of the light-harvesting complex from purple bacteria *Rhodopseudomonas acidophila*. In E. Dodson, M. Moore, A. Ralph, and S. Bailey, editors, *Proceedings of the CCP4 study weekend*, pages 115–123. CCLRC Daresbury Laboratory, Warrington, U.K., 1996.

[462] T. Thüne and J. Badger. Thermal diffuse X-ray scattering and its contribution to understanding protein dynamics. *Progress in Biophysics and Molecular Biology*, 63:251–276, 1995.

[463] J. P. Benoit and J. Doucet. Diffuse scattering in protein crystallography. *Quarterly Reviews of Biophysics*, 28:131–169, 1995.

[464] J. Doucet and J. P. Benoit. Molecular dynamics studied by analysis of the X-ray diffuse scattering from lysozyme crystals. *Nature*, 325:643–646, 1987.

[465] D. J. Leahy, R. Axel, and W. A. Hendrickson. Crystal structure of a soluble form of the human Tcell coreceptor CD8 at 2.6 Å resolution. *Cell*, 68:1145–1162, 1992.

[466] K. Hayakawa and J. Oizumi. Isolation and characterization of human breast milk lipoamidase. *Biochimica et Biophysica Acta*, 957:345–351, 1988.

[467] D. E. Garfin. Isoelectric focussing. *Methods in Enzymology*, 182:459–477, 1990.

[468] J. Heukeshoven and R. Dernick. Improved method for silver staining of proteins in polyacrylamide gels and the mechanism of silver staining. *Electrophoresis*, 6:103–112, 1985.

[469] B. Bjellqvist, G. Hughes, C. Pasquali, N. Paquet, F. Ravier, J. C. Sanchez, S. Frutiger, and D. F. Hochstrasser. The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis*, 14:1023–1031, 1993.

[470] M. M. Bradford. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry*, 72:248–254, 1976.

[471] U. Arndt. X-ray collimation and generation. In L. Sawyer, N. Isaacs, and S. Bailey, editors, *Proceedings of the CCP4 study weekend*, pages 33–43. SERC Daresbury laboratory, Warrington U.K., 1993.

[472] U. W. Arndt and A. J. Wonacott, editors. *The Rotation Method in Crystallography*. Elsevier, Amsterdam, 1977.

[473] T. J. Greenhough and F. L. Suddath. Oscillation camera data processing. 4. Results and recommendations for the processing of synchrotron radiation data in macromolecular crystallography. *Journal of Applied Crystallography*, 19:400–409, 1986.

[474] T. J. Greenhough. Partials and partiality. In J. R. Helliwell, P. A. Machin, and M. Z. Papiz, editors, *Proceedings of the CCP4 study weekend*, pages 51–57. SERC Daresbury laboratory, Warrington, U.K., 1987.

[475] J. R. Helliwell. *Macromolecular Crystallography with Synchrotron Radiation*. Cambridge University Press, Cambridge, U.K., 1992.

[476] D. I. Stuart and E. Y. Jones. Weissenberg data collection for macromolecular crystallography. *Current Opinion in Structural Biology*, 3:737–740, 1993.

[477] D. W. Rodgers. Cryocrystallography. *Structure*, 2:1135–1140, 1994.

[478] T. Y. Teng. Mounting of crystals for macromolecular crystallography in a free standing film. *Journal of Applied Crystallography*, 23:387–391, 1990.

[479] L. Blond, S. Pares, and R. Kahn. An easy technique for making fiber loops for cryocrystallography. *Journal of Applied Crystallography*, 28:653–654, 1995.

[480] H. Hope, F. Frolow, K. von Böhlen, I. Makowski, C. Kratky, Y. Halfon, H. Danz, P. Webster, K. S. Bartels, H. G. Wittmann, and A. Yonath. Cryocrystallography of ribosomal particles. *Acta Crystallographica*, B45:190–199, 1989.

[481] M. Nakasako, T. Ueki, C. Toyoshima, and Y. Umeda. A crystal mounting device made from a capillary tube for cryogenic macromolecular crystallography. *Journal of Applied Crystallography*, 28:856–857, 1995.

[482] G. Petsko. Protein crystallography at sub-zero temperatures: Cryo-protective mother liquors for protein crystals. *Journal of Molecular Biology*, 96:381–392, 1975.

[483] H. Hope. Cryocrystallography of biological macromolecules: a generally applicable approach. *Acta Crystallographica*, B44:22–26, 1988.

[484] E. P. Mitchell and E. F. Garman. Flash freezing of protein crystals: investigation of mosaic spread and diffraction limit with variation of cryoprotectant concentration. *Journal of Applied Crystallography*, 27:1070–1074, 1994.

[485] J. C. Fontecilla-Camps, R. de Llorens, M. H. le Du, and C. M. Cuchillo. Crystal structure of ribonuclease A.d(ApTpApApG) complex. *The Journal of Biological Chemistry*, 269:21526–21531, 1994.

[486] M. R. Haynes, E. A. Stura, D. Hilvert, and E. A. Wilson. Routes to catalysis: Structure of a catalytic antibody and comparison with its natural counterpart. *Science*, 263:646–652, 1994.

[487] P. O'Hara, P. Goodwin, and B. L. Stoddard. Direct measurement of diffusion rates in enzyme crystals by video absorbance spectroscopy. *Journal of Applied Crystallography*, 28:829–833, 1995.

[488] R. Huber. Experience with the application of Patterson search techniques. In P. Machin, editor, *Proceedings of the CCP4 study weekend*, pages 58–61. SERC Daresbury Laboratory, Warrington, U.K., 1985.

[489] S. N. Rao, J. H. Jih, and J. A. Hartsuck. Rotation function space groups. *Acta Crystallographica*, A36:878–884, 1980.

[490] E. A. Lattman. Optimal sampling of the rotation function. *Acta Crystallographica*, B28:1065–1068, 1972.

[491] A. T. Brünger. Extension of molecular replacement: A new search strategy based on Patterson correlation refinement. *Acta Crystallographica*, A46:46–57, 1990.

[492] M. Fuginaga and R. J. Read. Experiences with a new translation function program. *Journal of Applied Crystallography*, 20:517–521, 1987.

[493] A. T. Brünger. Molecular replacement with X-PLOR: Pc-refinement and free R value. In E. Dodson, S. Gover, and W. Wolf, editors, *Proceedings of the CCP4 study weekend*, pages 49–61. SERC Daresbury Laboratory, Warrington, U.K., 1992.

[494] F. L. Hirshfeld. Symmetry in the generation of trial structures. *Acta Crystallographica*, A24:301–311, 1968.

[495] R. J. Read. Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallographica*, A42:140–149, 1986.

[496] R. J. Read. Structure-factor probabilities for related structures. *Acta Crystallographica*, A46:900–912, 1990.

[497] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica*, A32:922–923, 1976.

[498] S. T. Rao and M. G. Rossmann. Comparison of super-secondary structures in proteins. *Journal of Molecular Biology*, 76:241–256, 1973.

[499] S. Pascarella and P. Argos. A data bank merging related protein structures and sequences. *Protein Engineering*, 5:121–137, 1992.

[500] W. A. Hendrickson and J. H. Konnert. Incorporation of stereochemical information into crystallographic refinement. In R. Diamond, S. Ramaseshan, and K. Venkatesan, editors, *Computing in Crystallography*, pages 13.01–13.25. Indian Institute of Science, Bangalore, 1980.

[501] W. A. Hendrickson and J. H. Konnert. Stereochemically restrained crystallographic least-squares refinement of macromolecule structures. In R. Srinivasan, editor, *Biomolecular Structure, Conformation, Function and Evolution*, volume I, pages 43–57. Pergamon Press, Oxford, 1981.

[502] L. F. Ten Eyck. Full matrix least squares. In E. Dodson, M. Moore, A. Ralph, and S. Bailey, editors, *Proceedings of the CCP4 study weekend*, pages 37–45. CCLRC Daresbury laboratory, Warrington, U.K., 1996.

[503] L. M. Rice and A. T. Brünger. Torsion angle dynamics: Reduced variable conformational sampling enhances crystallographic structure refinement. *Proteins*, 19:277–290, 1994.

[504] K. Braig, P. D. Adams, and A. T. Brünger. Conformational variability in the refined crystal strcuture of the chaperonin GroEL at 2.8 Å resolution. *Nature Structural Biology*, 2:1083–1094, 1995.

[505] A. T. Brünger. Free R-value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, 355:472–475, 1992.

[506] G. J. Kleywegt and A. T. Brünger. Checking your imagination: applications of the free R value. *Structure*, 4:897–904, 1996.

[507] A. E. Sauer-Erickson, G. J. Kleywegt, M. Uhlen, and T. A. Jones. Crystal structure of the C2 fragment of streptococcal protein G in complex with the Fc domain of human IgG. *Structure*, 3:265–278, 1995.

[508] M. Vijayan. On the Fourier refinement of protein structures. *Acta Crystallographica*, A36:295–298, 1980.

[509] R. Agarwal and N. W. Isaacs. Method for obtaining a high resolution protein map starting from a low resolution map. *Proceedings of the National Academy of Sciences, USA*, 74:2835–2839, 1977.

[510] V. Y. Lunin and A. G. Urzhumtsev. Improvement of protein phases by coarse model modification. *Acta Crystallographica*, A40:269–277, 1984.

[511] V. Y. Lunin, A. G. Urzhumstev, E. A. Vernoslova, Y. N. Chirgadze, N. A. Nevskaya, and N. P. Fomenkova. Phase improvement in protein crystallography using a mixed electron density model. *Acta Crystallographica*, A41:166–171, 1985.

[512] V. S. Lamzin and K. S. Wilson. Automated refinement of protein models. *Acta Crystallographica*, D49:129–147, 1993.

[513] P. M. D. Fitzgerald. A generalized approach to the fitting of non-peptide electron density. In S. Bailey, R. Hubbard, and D. Waller, editors, *Proceedings of the CCP4 study weekend*, pages 125–132. EPSRC Daresbury Laboratory, Warrington, U.K., 1994.

[514] A. G. Urzhumtsev. Density growing: a method for local improvement of electron density maps. *CCP4 Newsletter on protein crystallography*, (32):37–40, 1996.

[515] A. Hodel, S.-H. Kim, and A. Brünger. Model bias in macromolecular crystal structures. *Acta Crystallographica*, A48:851–858, 1992.

[516] J. R. Somoza, H. Szöke, D. M. Goodman, P. Béran, D. Truckses, S.-H. Kim, and A. Szoke. Holographic methods in X-ray crystallography. IV. A fast algorithm and its application to macromolecular crystallography. *Acta Crystallographica*, A51:691–708, 1995.