Phosphofructokinase as a target for rational drug design

Daniel Rigden

Ph.D. Thesis
University of Edinburgh, 1993.

Declaration

The work presented in this thesis is the result of my own research unless otherwise stated, and has not been presented for any other academic degree elsewhere.

Daniel Rigden.
Edinburgh, Dec 1993.

All intellectual improvement arises from leisure.

Samuel Johnson

One never knows, do one?

Thomas 'Fats' Waller

Acknowledgements

I first wish to thank my supervisors Linda Gilmore and Paul Taylor for their help and support throughout the course of this work.

Thanks are due to Dr Mike Blackburn at Sheffield and Professor Perie in Toulouse for various useful chemistry discussions. Thanks too to Dr David Apps for his comments on some of this thesis. I particularly wish to thank Dr Phil Evans at Cambridge for much help with all things PFK - plasmids, strains and coordinates.

None of this would have been any fun without the folk of Lab 327. Thanks then to Fiona, Jacqui, Lluis, Malcolm, Nantana, Rebecca and Rich.

I'm particularly indebted to Thales who has helped preserve my sanity on several occasions. I couldn't have wished for a more understanding friend.

Special mention too for Simon whose wit and *joie de vivre* enlivened the whole department and more. He and Carol are much missed. As is Barbara, though differently.

Finally, thanks to the family, both Crewe and Edinburgh contingents, for their constant and unconditional support.

Abstract

There is a pressing need for new anti-parasitic drugs. Parasitic infection is an enormous, and growing, worldwide problem. Current treatments are generally ineffective and cause side-effects. Organisms resistant to some of the more effective drugs are appearing.

A good target for drug intervention is glycolysis on which many parasites rely for their energy supply. Phosphofructokinase (PFK) is a good choice of glycolytic enzyme for targeting since it is believed to catalyse the major regulatory step of glycolysis. However, inhibition of parasite PFK must take place with as little inhibition of the host PFK as possible.

The aim of this work was to tackle this problem of species selectivity. In the absence of much information about parasite PFKs, a model system was used which comprised two of the best-studied PFKs. The goal was inhibition of *Escherichia coli* PFK by compounds little affecting rabbit muscle PFK.

Models of the structure for rabbit muscle PFK were built based on the justified assumption of a shared protein fold with *Escherichia coli* PFK. Several ligand design approaches were used. Some used protein structure information like the program LUDI. Database searches used information from the structures of known ligands. The question of species selectivity was addressed during the design process. Several classes of putative ligands were produced. These ranged from disparate organic compounds suggested by LUDI to substrate analogs modelled using molecular mechanics procedures. Testing of putative ligands was by a variety of techniques such as enzyme assay, fluorescence and protection against protein denaturation. Finally, the results were summarised and future prospects considered.

Contents

List of Figures

List of Tables

Abbreviations

a) The Amino Acids

One letter code	Three letter code	Name
A	Ala	Alanine
C	Cys	Cysteine
D	Asp	Aspartic acid
E	Glu	Glutamic acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
Y	Tyr	Tyrosine

b) Chemicals

AMP	adenosine 5'-monophosphate
AMPPCP	β,γ-methylene adenosine 5'-triphosphate
ADP	adenosine 5'-diphosphate
ATP	adenosine 5'-triphosphate
DMF	dimethylformamide
DMSO	dimethylsulphoxide
DTT	dithiothreitol
EDTA	ethylenediaminetetraacetic acid
F-2,6-BP	fructose-2,6-bisphosphate

F6P	fructose-6-phosphate
FBP	fructose-1,6-bisphosphate
GdnHCl	guanidinium hydrochloride
GDP	guanosine 5'-diphosphate
GTP	guanosine 5'-triphosphate
IPTG	isopropylthiogalactoside
MTT	3-[4,5-dimethylthiazol-2-yl]-2,5-diphenyltetrazolium bromide
NAD^+	nicotinamide adenine dinucleotide (oxidised form)
NADH	nicotinamide adenine dinucleotide (reduced form)
$NADP^+$	nicotinamide adenine dinucleotide phosphate (oxidised form)
NADPH	nicotinamide adenine dinucleotide phosphate (reduced form)
NAT	N-acetyltryptophanamide
PC	phosphocholine
PDS	phospho-D-serine
PE	phosphoethanolamine
PEP	phospho-*enol*-pyruvate
PLS	phospho-L-serine
PP_i	inorganic pyrophosphate
PS	phosphoserine
Tris	tris (hydroxymethyl) aminomethane

c) Others

2D	two dimensional
3D	three dimensional
A_x	absorbance at x nm
BSA	bovine serum albumin
CD	circular dichroism
DHFR	dihydrofolate reductase
F-1,6-BPase	fructose-1,6-bisphosphatase
F-2,6-BPase	fructose-2,6-bisphosphatase
FPLC	fast protein liquid chromatography
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
H-bond	hydrogen bond
K_R	binding constant for R state
K_T	binding constant for T state
L	allosteric constant
M	molar (mol l^{-1})
n_H	Hill coefficient

NMR	nuclear magnetic resonance
PFK	phosphofructokinase
RM	rabbit muscle
SDS-PAGE	sodium dodecyl sulphate - polyacrylamide gel electrophoresis

Chapter 1

Introduction

Parasitic infection is an enormous worldwide problem. The figures are shocking. For example, nearly half the world's population, some 2100 million people, are at risk from malaria (Fullick and Fullick, 1993). The situation is worsening due to social factors. One of these is migration of the population from the country into the cities. Frequently these people are too poor to take account of the potential health advantages; better drinking water and sewerage facilities, for example. The obvious health disadvantages of city dwelling like overcrowding, however, remain. Ironically, potentially favourable developments such as reservoirs and irrigation ditches can provide new environments in which parasites and their vectors may breed.

Parasitic infections are hugely more prevalent in tropical, generally poorer, nations than they are in the developed world. The most common intestinal parasite in the U.S. causes around 4000 hospital admissions per year. Cryptosporidiosis infects 250-500 million people in Asia, Africa and Latin America annually (Warhust and Smith, 1992). The great weighting of parasitic diseases on poorer nations is the main reason why efforts towards pharmaceutical intervention trail so far behind those directed at cancer, heart disease etc. Drug companies find the potential profits in anti-parasitic compounds to be too small to justify research and development costs. The onus has therefore fallen much more on public bodies like the World Health Organisation (Patel, 1993).

There are two main ways to attack parasitic diseases; vaccines and drugs. Vaccines have been effective against diseases such as smallpox (now virtually eradicated), measles and polio. However there are intrinsic limits on their usefulness in preventing parasitic diseases. Those parasites that live inside their host have generally evolved ways of evading the host's immune system. For example, the parasite may live inside host cells or may change its external coat so that host antibodies no longer recognise it. Drugs are not subject to these limitations but the relative lack of research in this field means that those currently available are generally of limited effectiveness and cause significant side-effects. Some of the better drugs, such as chloroquine for malaria treatment, are now suffering from the evolution of resistant target organisms.

In this chapter I shall first discuss what makes glycolytic enzymes, and phosphofructokinase

(PFK) in particular, such good targets for anti-parasitic drug design. I will then explain the choice of *E.coli* and rabbit muscle PFKs as a model system for species-selective drug design.

1.1 Why target parasite PFKs ?

Various considerations influence the choice of pathway for drug targeting. Of course blockage of the pathway should have as severe as possible consequences for the parasite. Less obviously, the pathway chosen should be one about which basic biochemical knowledge is available or easily obtained. Ideally, the enzymes of the pathway should be readily purified. This is significant since one of the lessons of drug design is that the number of available methods and the chance of success increase with increasing knowledge of the system. In practise this often leads to avoidance of metabolic pathways unique to the parasite and concentration instead on pathways that have been well studied in other organisms.

It is generally thought that most parasites use glycolysis exclusively for their energy production. This is certainly the case in the helminths *Ascaris suum* and *Fasciola hepatica*.(Köhler, 1985). Glycolysis is therefore a good target for anti-parasitic drug design. The glycolytic pathway and its enzymes are shown in Fig 1.1.

The best choice of enzyme for targeting is that which has the greatest influence on the flux through the pathway - the most 'rate-controlling' enzyme or that with the highest flux control coefficient. This arises from the recognition that complete inhibition of an enzyme is unlikely to be achieved. Incomplete inhibition will have more effect on pathway flux if the enzyme is of the above type rather than one with a lower flux control coefficient.

The activity of PFK has long been assumed to have an important effect on the rate of glycolysis as a whole. This belief originated in the discovery that most PFKs are highly regulated enzymes: why would they be so if glycolysis as a whole was not dependent, to some extent, on PFK rate? Various modelling studies of erythrocytes (Rapoport *et al.*, 1974), *E.coli* (Torres and Babul, 1991), muscle tissue (Brooks and Storey, 1991) and ascites tumour cells (Kohn and Lemieux, 1991) tend to confirm the importance of PFK in controlling glycolytic

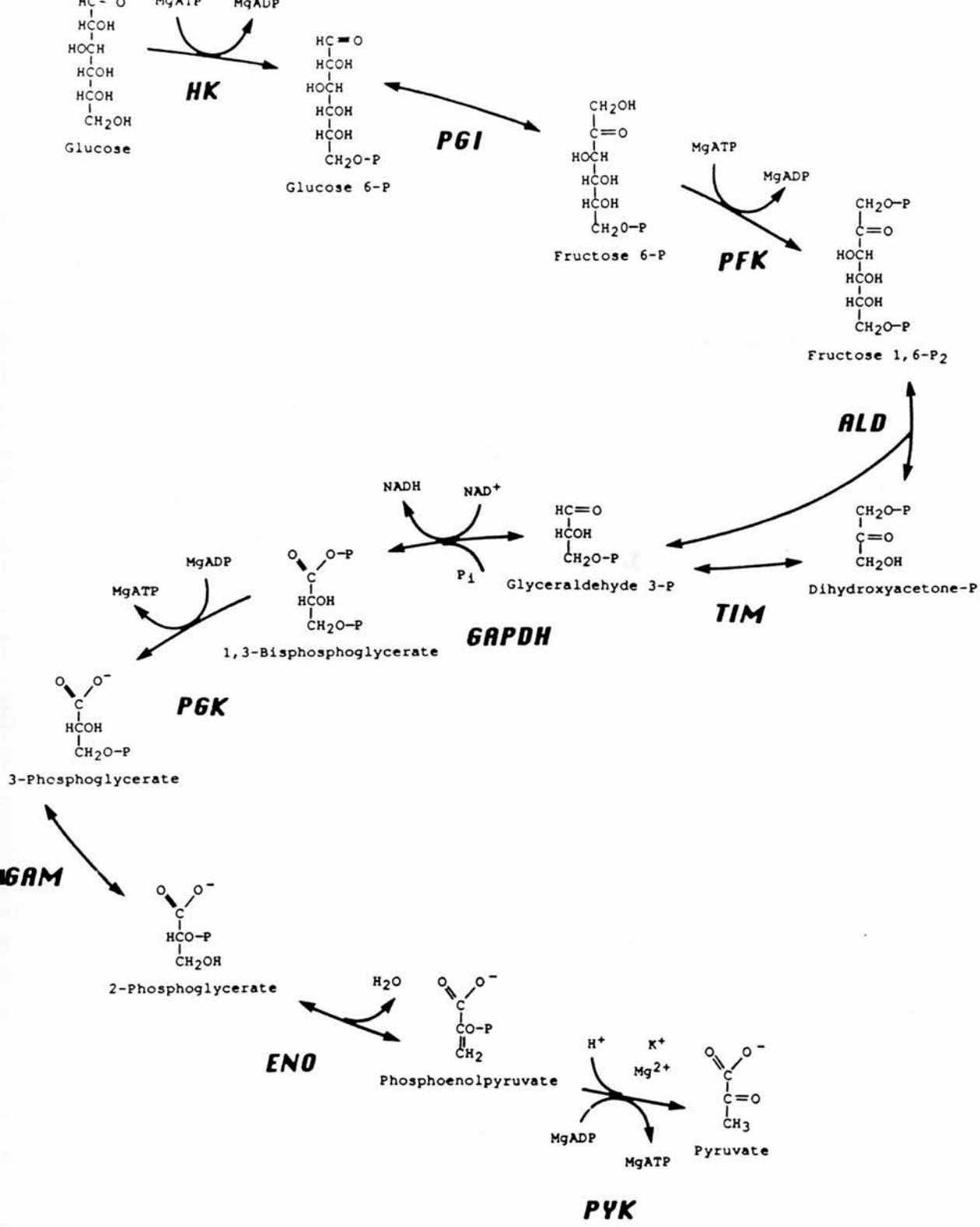


Figure 1.1 Glycolysis. The abbreviations for the enzymes are as follows: HK, hexokinase; PGI, glucosephosphate isomerase; PFK, phosphofructokinase; ALD, aldolase; TIM, triosephosphate isomerase; GAPDH, glyceraldehyde-phosphate dehydrogenase; PGK, phosphoglycerate kinase; PGAM, phosphoglycerate mutase; ENO, enolase; PYK, pyruvate kinase. The letter 'P' in the chemical structures represents a phospho group. Taken from Fothergill-Gilmore and Michels, 1992.

rate. In addition, the drastically increased rate of glycolysis seen in erythrocytes infected with the malarial parasite *Plasmodium berghei* can be explained using a kinetic model of the parasite-specific PFK (Buckwitz *et al.*, 1990). Also suggestive of an important regulatory role for PFK is the finding, in another parasite - *Ascaris suum*, that PFK is operating under physiological conditions at only 2% of its maximal activity (Hofer *et al.*, 1982).

The regulatory importance of an enzyme to a pathway can be determined experimentally by modulating its activity, for example using inhibitors or over-expression. PFK has been over-expressed five-fold in *Saccharomyces cerevisiae* and the consequences monitored using n.m.r. spectroscopy and assays for enzymes and metabolites (Davies and Brindle, 1992). Glycolytic flux measurements showed only small increases in the over-expressing cells suggesting a low PFK flux control coefficient for glycolysis. However, these results could be explained by an observed, compensatory decrease in 6-phosphofructo-2-kinase activity and hence fructose-2,6-bisphosphate (F26BP) concentration. F26BP is a potent activator of yeast PFK so that the decline in its concentration could explain the absence of a large increase in glycolytic flux. Thus, frustratingly, no conclusion could be drawn about the importance of PFK in regulating glycolytic flux.

It can be seen that there is little evidence to contradict the intuitive notion that PFK is an important control enzyme of glycolysis. Amongst the glycolytic enzymes it is therefore a good target for drug intervention. Indeed, some of the earliest anti-parasitic drugs, antimonial compounds, acted on PFK (Bueding and Mansour, 1957).

1.2 Variety of PFKs

There are a variety of 'phosphofructokinases' - enzymes capable of phosphorylating phosphofructose. They may be categorised according to their substrates.

The first division that can be made is between those PFKs whose substrate is fructose-1-phosphate and those acting on fructose-6-phosphate. The former class catalyses the reaction :-

$$\text{F-1-P} + \text{ATP} \rightleftharpoons \text{F-1,6-BP} + \text{ADP}$$

This reaction is part of the fructose uptake pathway in bacteria such as *E.coli* (Buschmeier *et al.*, 1985).

The enzymes whose substrate is fructose-6-phosphate can be further divided according to the position at which they attach the new phospho group. Some phosphorylate the 2 position of F6P :-

$$\text{F6P} + \text{ATP} \rightleftharpoons \text{F-2,6-BP} + \text{ADP.}$$

These 'PFK-2' enzymes are located at the N-terminal portion of a bifunctional polypeptide chain which has at its C-terminal portion a domain with F-2,6-BPase activity. F26BP is a quite recent discovery (Furuya and Uyeda, 1980; Furuya and Uyeda, 1981). It is an important regulatory molecule in liver, for example, since it both activates glycolytic PFK and inhibits F-1,6-BPase (for a review see Pilkis *et al.*, 1988). It therefore has a large effect on the balance between glycolysis and gluconeogenesis.

The remaining PFKs form part of the glycolytic pathway. They can again be divided into two groups - those using ATP as phospho donor :-

$$\text{F6P} + \text{ATP} \rightleftharpoons \text{F-1,6-BP} + \text{ADP}$$

and those using inorganic pyrophosphate (PP_i) :-

$$\text{F6P} + PP_i \rightleftharpoons \text{F-1,6-BP} + P_i.$$

The ATP-dependent PFK is much more commonly found. The PP_i-dependent PFK is present in some prokaryotes and eukaryotes but not in archaebacteria. Two different situations have been noted (Mertens, 1991). In higher plants and *Euglena gracilis* (a photosynthetic protist) PP_i-dependent PFK is allosterically regulated and coexists with ATP-dependent PFK. In all

other organisms the PP_i-dependent PFK appears not to be allosterically regulated, and ATP-dependent PFK levels are very low.

1.2.1 Relationships between PFKs

With enzymes catalysing the same reaction or similar reactions it is interesting to speculate about their possible evolutionary relationships; do they arise from a common ancestor? The most fundamental tool here is sequence comparison. If two enzymes share a sufficiently high degree of sequence identity then it is reasonable to assume that they have diverged from a common ancestor; the probability of independent convergent evolution to similar sequences becomes small. What then constitutes a 'sufficiently high degree of sequence identity'? Gapped alignments of random sequences produce 10-20% sequence identity (Fothergill-Gilmore and Michels, 1992) so observed values in this range must be considered doubtful. Above these values the percentage identity required to be confident of common ancestry decreases with the length of alignment. This has been studied by Sander and Schneider (1991). Figure 2.2 illustrates their findings. For alignments of greater than 80 residues they suggest common ancestry may be inferred from observed sequence identity of greater than 25-30%.

By these criteria, all the ATP-dependent glycolytic pathway PFKs share a common ancestor. Interestingly, in the case of yeast PFK (both α and β subunits) and mammalian PFKs, separate homology can be seen between the N- and C-terminal halves and the bacterial enzymes. From this it has been inferred that they arose through duplication of a precursor PFK gene with subsequent gene fusion. Sequence comparisons show that the C-terminal repeats of both yeast and mammalian PFKs can no longer be catalytically active. In both cases the essential catalytic aspartate residue has been replaced by a serine. Instead, it seems that the C-terminal repeats are responsible for binding the much broader range of effector molecules active on eukaryotic PFKs.

The situation is more complicated when considering the possible relatedness of ATP- and PP_i-

dependent glycolytic pathway PFKs. Unfortunately, there are still only three PP_i-dependent PFK sequences - the α and β subunits of the potato tuber enzyme (Carlisle *et al.*, 1990) and the enzyme from *Propionibacterium freudenreichii* (Ladror *et al.*, 1991). The potato tuber α and β subunits are clearly related to each other since they have about 40% identical sequences. However, comparison with the *E.coli* ATP-dependent PFK produced only 16 and 22% identity with the α and β subunits respectively. This alone is insufficient to infer common ancestry, but there is suggestive circumstantial evidence. For example, the predicted secondary structure of the PP_i-dependent enzyme closely matches that of the *E.coli* enzyme. In addition, the alignments with the *E.coli* enzyme can explain the observation that the β, but not the α, subunit is catalytically active. A 'low but significant' level of sequence identity between the *Propionibacterium* enzyme and *E.coli* PFK has been asserted (Green *et al.*, 1993). This value is 23%, not in the region in which common ancestry may be confidently deduced. The effects of mutations in two catalytic aspartates separated by one residue (Green *et al.*, 1993) are certainly consistent with homology. However, a D-X-D sequence is readily explained by convergent evolution and the possible relationship of the *Propionibacterium* enzyme and *E.coli* PFK must remain doubtful. Results obtained here using the PROFILESEARCH program (see Section 2.3.1) strongly support the idea that the potato enzymes, but not the *Propionibacterium* PFK, are evolutionarily divergently related to the ATP-dependent PFKs. These results are consistent with the highly dispersed distribution of PP_i-dependent PFKs among the prokaryotes and eukaryotes (Mertens, 1991) that suggested that PP_i-dependent PFK has evolved independently, from ATP-dependent PFK or not, several times. Fig 1.2 illustrates the distribution of PP_i-dependent PFKs among prokaryotes and eukaryotes on a multi-kingdom tree derived from rRNA sequences (taken from Mertens, 1991).

The ATP-dependent PFKs may share structural similarity and common ancestry with the PFK-2s althought the evidence is not unequivocal (Bazan *et al.*, 1989). A likely nucleotide-binding fold can be identified from characteristic sequence motifs in the PFK-2 sequence. The match between known secondary structure patterns of various ATP-binding proteins and the predicted secondary structure of the PFK-2 sequence was best for bacterial PFK. Using this information an alignment was produced which suggestively showed greatest sequence similarity in active

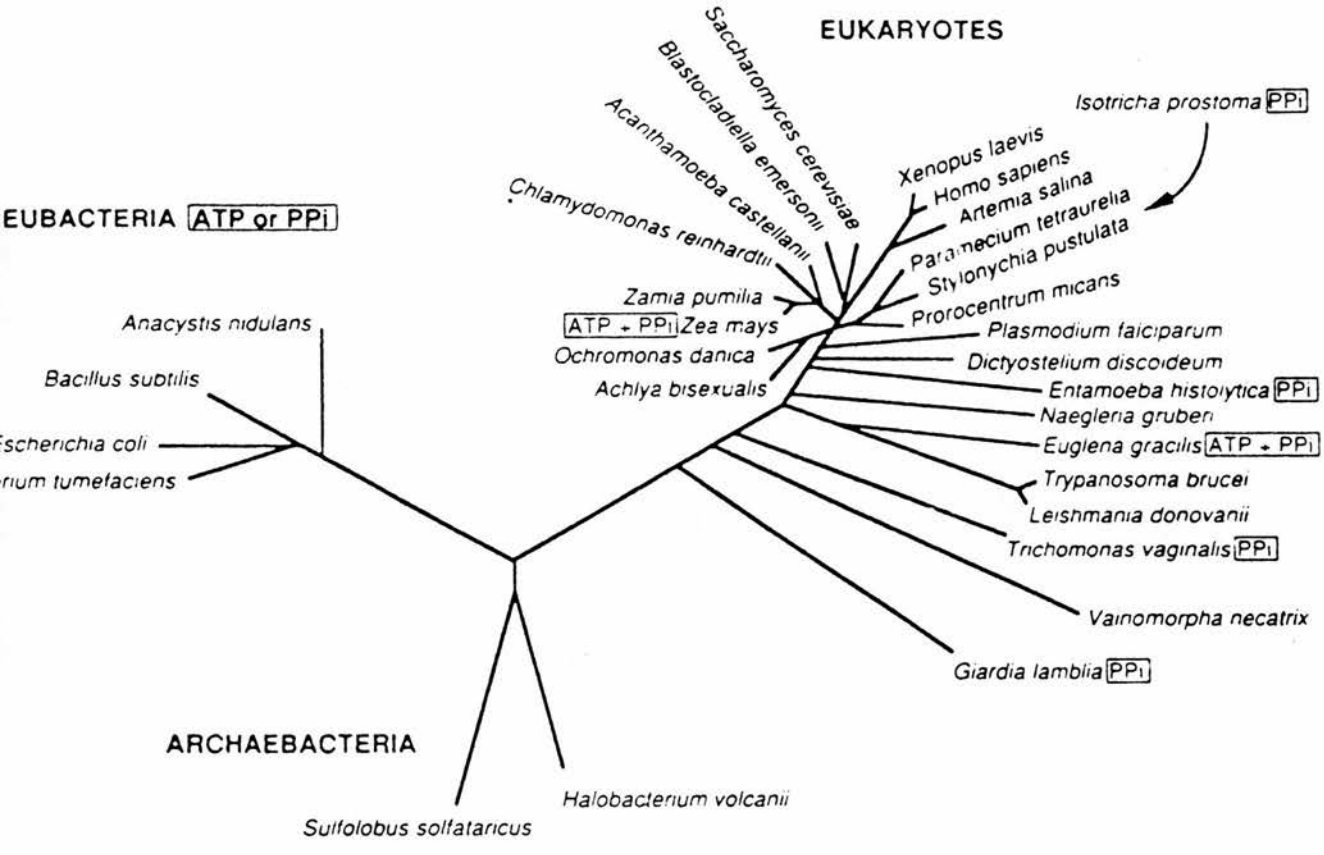


Fig 1.2 Distribution of PP_i-dependent phosphofructokinases among prokaryotes and eukaryotes. Taken from Mertens, 1991.

site and subunit interface regions. The latter was important since dimerisation of PFK-2 domains is known to be required for activity. Nevertheless, proof of structural similarity must await X-ray crystallographic studies.

Fructose-1-phosphate kinases from *E.coli* (Orchard and Kornberg, 1990) and *Rhodobacter capsulatus* (Wu *et al.*, 1991) show clear homology to other sugar kinases like phosphotagatokinase and ribokinase and to Rubisco large subunits. A superfamily including all these enzymes plus the ATP-dependent PFKs has been suggested (Wu *et al.*, 1991). However, the evidence for this is very weak: subunit size, similarity of function, ambiguous levels of sequence identity, etc.

1.2.2 Parasite PFKs

Both ATP- and PP_i-dependent PFKs have been found individually, but not together, in parasites. There seems to be a relationship between the environment of the parasite and the presence of either ATP- or PP_i-dependent PFK (Mertens, 1993). The obvious advantage of a PP_i-dependent PFK is greater efficiency; PP_i, a by-product of other reactions that would otherwise go to waste, is usefully harnessed instead of ATP. Consistent with this is the high biomass yield of propionic bacteria which use this mechanism. This energy advantage would be of greater importance to some parasites than others. Blood-dwelling parasites like trypanosomes will always have a plentiful supply of glucose, and indeed PP_i-linked glycolysis is not found in these organisms. On the other hand, many parasites with less plentiful food supplies depend on fermentation for energy since they lack mitochondria. More efficient glycolysis is clearly highly desirable in these organisms.

There are as yet no extensive sequences of parasite ATP-dependent PFKs. The only sequences available are those surrounding the phosphorylation sites of PFK from *Ascaris suum* (Kulkarni *et al.*, 1987) and *Fasciola hepatica* (Mahrenholz *et al.*, 1991). These sequences are only 11 and 12 residues long respectively, and neither shows any similarity to known PFK sequences. Indirect evidence from the effect of proteolysis on *Ascaris suum* PFK has been interpreted as indicating similarity between this parasite PFK and mammalian PFKs (Ahanotu

et al., 1991). There are some sequences of peptide fragments from *Trichomonas vaginalis* PP_i-dependent PFK (Mertens, personal communication). These indicate 40-60% identity with ATP-dependent PFKs; enough to suggest common ancestry.

Phosphorylation of parasite PFKs causes stimulation of their activity. This has been shown for *Dirofilaria immitis* (Srinivasan *et al.*, 1988), *Ascaris suum* (Kulkarni *et al.*, 1987) and *Fasciola hepatica* (Mahrenholz *et al.*, 1991). Stimulation on phosphorylation has also been seen for mollusc muscle PFK (Biethinger *et al.*, 1991), but not for yeast PFK (Huse *et al.*, 1988) or mammalian enzymes such as rabbit muscle (Kemp and Foe, 1983). A possible alternative, indirect, way to inhibit parasite PFK would therefore be to effect dephosphorylation of the enzyme. In *Ascaris suum* at least, knowledge is accumulating about the kinases (Thalhofer *et al.*, 1988) and phosphatases (Daum *et al.*, 1992) that act on PFK. However, since fully dephosphorylated enzyme is still active to some extent it remains unclear whether this would be very detrimental to the organism.

1.3 A model system for species selective drug design

From the above discussion it will be apparent that rational drug design may not yet be attempted for parasite PFKs. Rational drug design ideally uses 3D structural information on the target enzyme (see Section 3.1.1.1). This is not available and neither is the sequence information that could allow model building by homology with other PFKs. Also there is no large body of information on parasite PFK ligands that could allow ligand-based drug design efforts (see Section 3.1.1.2).

One particular problem that arises from targeting parasite glycolysis is species-selectivity. It would clearly be undesirable to inhibit host glycolysis if, for example, drugs were being administered to infected humans. Rational species-specific drug design has been the focus of little attention although a recent attempt has been made to selectively target trypanosomal triose phosphate isomerase (Verlinde *et al.*, 1993). It was therefore decided to use a model PFK

system to study methods of species-specific drug design. Any effective compounds discovered could well be useful against other PFK species or act as lead compounds.

It was eventually decided to search for compounds that might inhibit *E.coli* PFK but affect rabbit muscle PFK to a lesser degree. The choice of *E.coli* PFK as target was made for two important reasons. First, the only 3D PFK structures available are those for *E.coli* PFK and the homologous enzyme from *B.stearothermophilus*. A 3D structure is required for many of the currently available rational drug design methods. Secondly a system enabling the great overexpression of *E.coli* PFK has been established (Lau *et al.*, 1987) and was kindly made available by Dr. P.R. Evans (University of Cambridge). The ready availability of target enzyme enables the testing of putative ligands by even methods requiring relatively large amounts of protein such as C.D. Rabbit muscle enzyme was chosen also for two reasons. First, the degree of sequence identity to *E.coli* PFK (39%) is sufficiently high for protein modelling by homology to be attempted while being low enough for there to exist significant exploitable differences in structure. Secondly, after the bacterial enzymes, RMPFK is probably the best studied PFK. Knowledge of the properties of an enzyme is crucial if testing of putative inhibitors is to be carried out.

A stepwise design process was envisaged. First the *E.coli* PFK structure would be used to rationally design inhibitors and these would then tested on both the *E.coli* and the rabbit muscle enzymes. Any differential inhibition would be explained using the *E.coli* PFK structure and a model of RMPFK built by homology. The two structures would then be used to suggest modifications that might lead to enhanced selective inhibition of the *E.coli* enzyme.

I shall now introduce the two chosen enzymes, concentrating more on the *E.coli* PFK about which more is known and on which most of the drug design and testing efforts focussed.

1.3.1 *E.coli* PFK

1.3.1.1 Structure

The determined X-ray crystal structures of bacterial PFKs reveal a tetramer of identical subunits each 319-320 amino acids in length. Each subunit consists of two domains, both of

which contain central β-sheet surrounded by α-helices. The tetramer is a dimer-of-dimers with 222 symmetry. Each subunit has three binding sites. The allosteric site which binds inhibitor PEP and activator ADP is situated at the interface between subunits in each dimer. The ATP substrate site is composed solely of residues from the large domain whereas the F6P substrate site consists of small domain residues and two arginines from the other dimer. Fig 1.3 shows a ribbon diagram of the *E.coli* PFK dimer with effector ADP molecules bound between the subunits. Figure 1.4 shows the *E.coli* PFK active site with products bound (Shirakihara and Evans, 1988). Inferences drawn from the crystal structures about possible roles of residues at binding sites have been confirmed by site-directed mutagenesis experiments (Evans, 1992 and references therein).

Crystal structures determined for unligated PFK (Rypniewski and Evans, 1989) or PFK in the presence of subtrates (Evans and Hudson, 1979; Evans *et al.*, 1981) or products (Shirakihara and Evans, 1988) can be identified with the R-state of the Monod-Wyman-Changeux allostery model (see below). These and a T-state structure, determined in the presence of allosteric inhibitor (Schirmer and Evans, 1990), have enabled a molecular description of the allosteric transition of bacterial PFK (Schirmer and Evans, 1990).

1.3.1.2 Kinetics and Regulation

The Monod-Wyman-Changeux (MWC) Model

This allostery model (Monod *et al.*, 1965), also known as the 'concerted' or 'two-state' model, postulates two forms of an oligomeric protein. The T (tense) state and the R (relaxed) state are in equilibrium with the T state having lower affinity for substrates. Each type of site has binding constants for the T and R states - K_T and K_R. Allosteric activators operate by binding to, and stabilisation of, the active R state. Conversely, allosteric inhibitors bind to and stabilise the inactive T state. The third constant defined is L, the allosteric constant, which is the ratio [T]/[R] for the unligated states.

E.coli PFK shows hyperbolic kinetics with respect to ATP but strongly sigmoidal kinetics for F6P. The kinetic behaviour of *E.coli* PFK has been studied in detail and found to be consistent

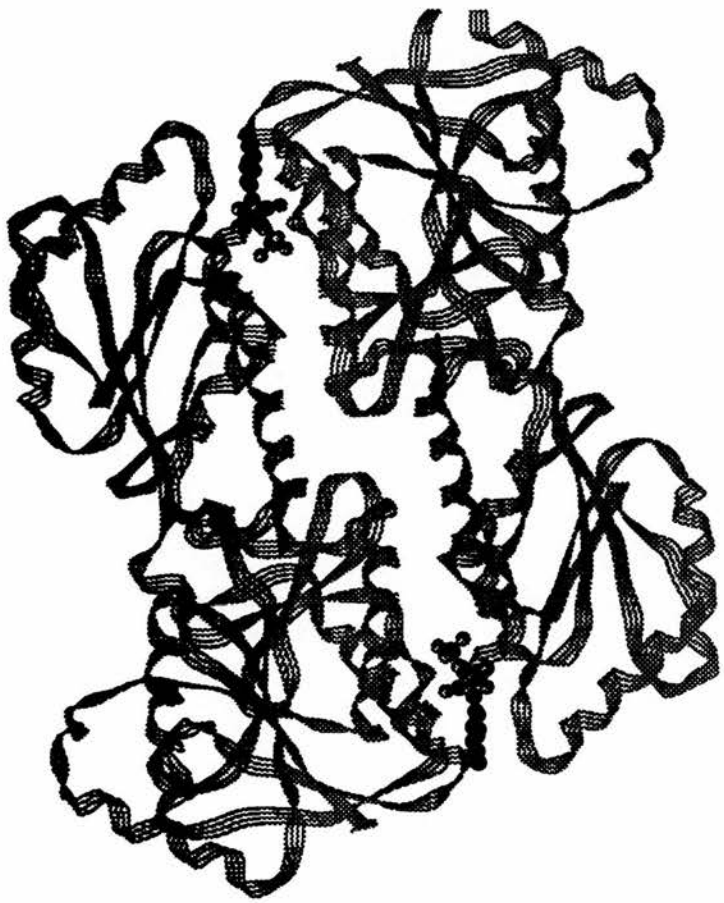


Fig 1.3 Ribbon diagram of the *E.coli* PFK dimer. The ADP molecules bound to the allosteric sites between subunits are shown in ball-and-stick representation.

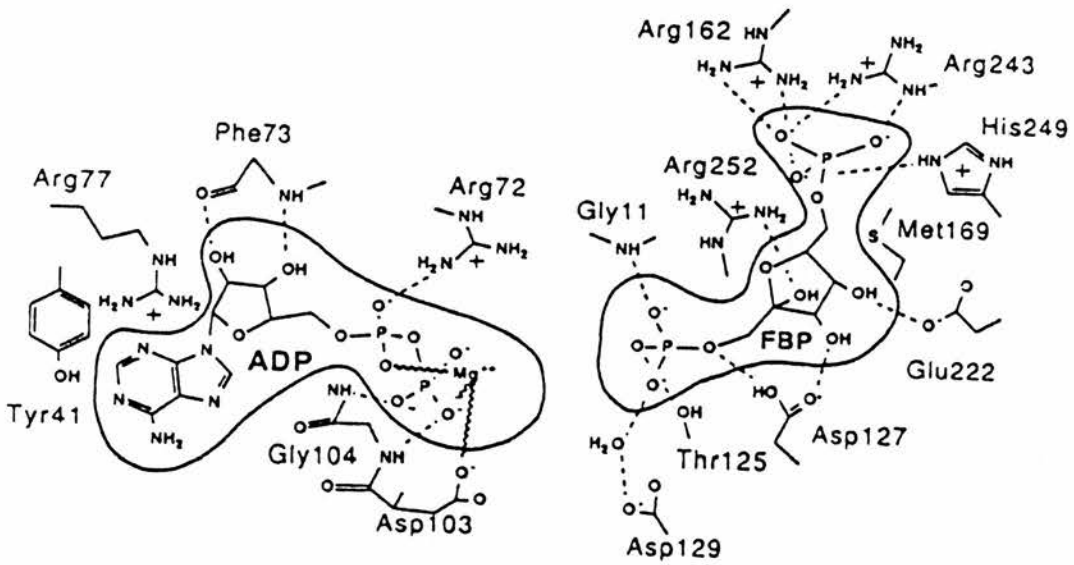


Fig 1.4 The active site of *E.coli* PFK with bound products. Taken from Shirakihara and Evans, 1988.

with the MWC model (Blangy *et al.*, 1968) shown in Fig 1.5a. That model was further supported by the observation of two states by X-ray crystallography (see above). However, recent work, mostly using fluorimetry, has demonstrated features of the behaviour of *E.coli* PFK that are inconsistent with the MWC model. These data are now briefly discussed.

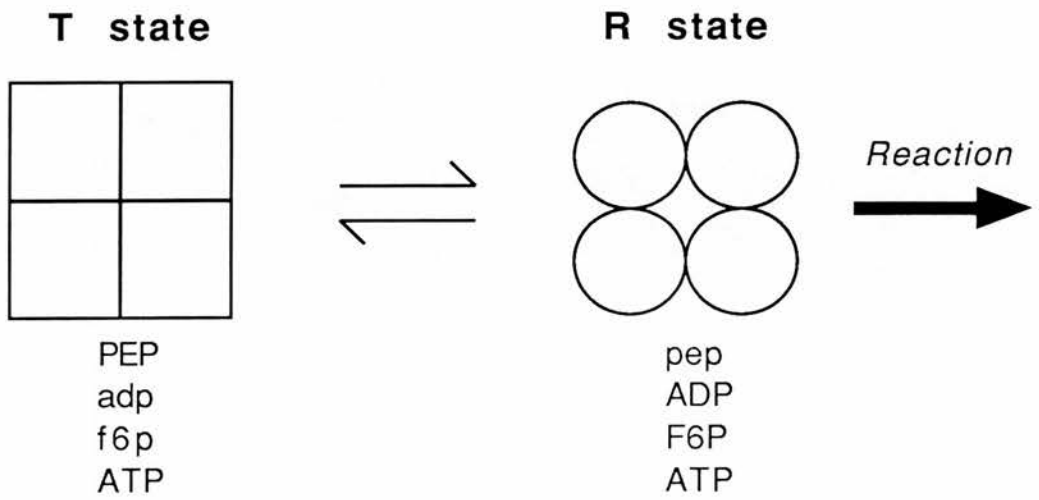
Complications

The fluorescence behaviour of *E.coli* PFK has been studied in some detail (Berger and Evans, 1991; Deville-Bonne and Garel, 1992). The later results provide a consistent framework for understanding ligand binding, although they differ in some details from the earlier data, and are now summarised.

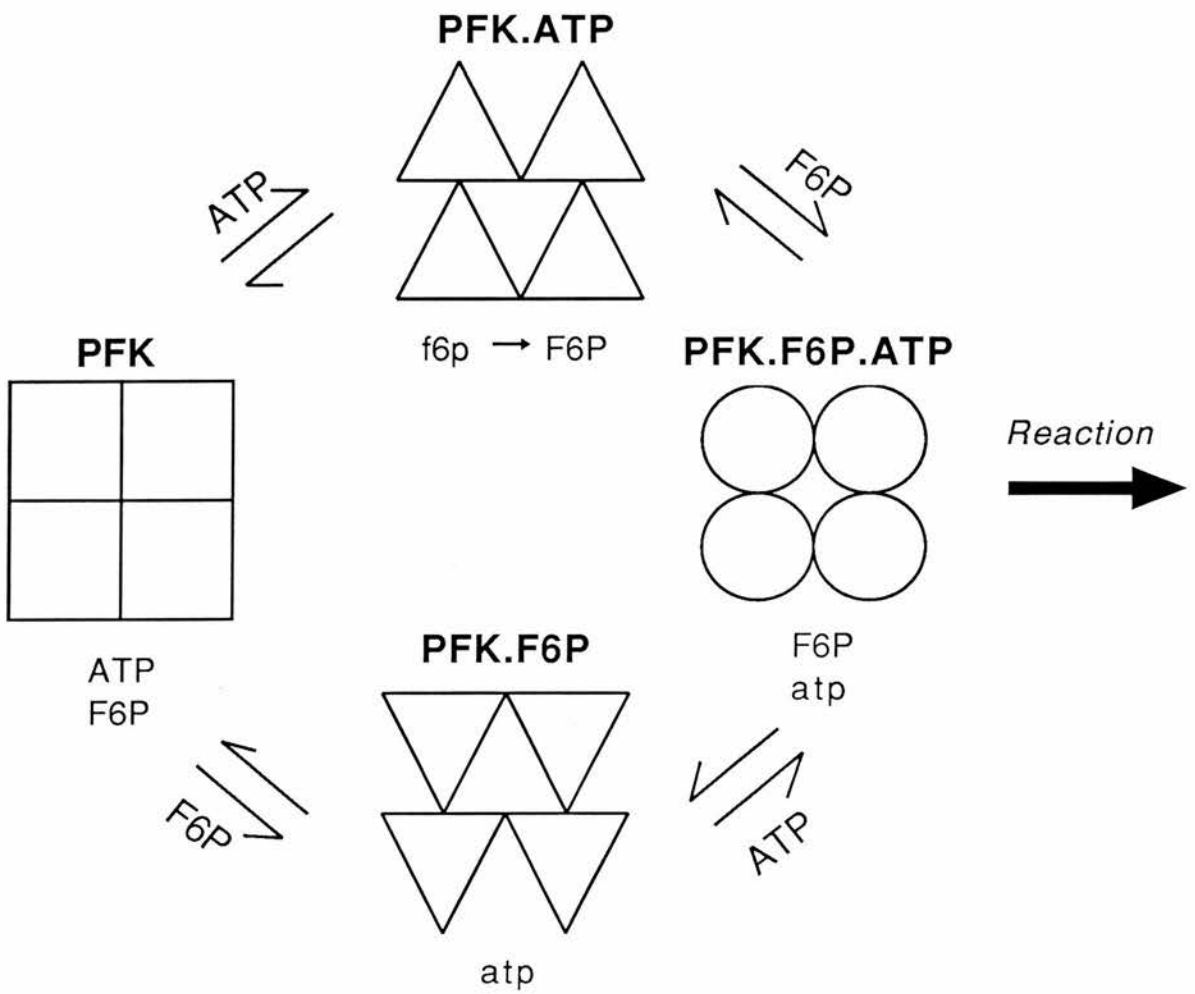
Deville-Bonne and Garel studied the individual effects of substrates, F6P and ATP (and its analog AMPPCP), and effectors, ADP and PEP. They were able to divide these into two classes. F6P and ADP produced a similar, around 20% fluorescence decrease relative to unligated PFK. ATP (or AMPPCP) and PEP gave an 8-10% fluorescence increase. From this they postulate the existence of two states, high-fluorescence and low-fluorescence with the intermediate overall fluorescence of the unligated enzyme arising from a mixture of the two. They assign the high-fluorescence form to the T-state of the concerted model (Blangy *et al.*, 1968) and the low-fluorescence form to the R-state. From this an allosteric constant, L, of around two is inferred, in good agreement with equilibrium dialysis measurements of GDP binding (Blangy, 1971) but very different from the 4×10^6 calculated from steady state measurements of activity (Blangy *et al.*, 1968). An L value near unity provides the explanation for the near hyperbolicity of binding of all ligands to the unligated state. However, when the enzyme is put predominantly in one state by the presence of a ligand from one class, binding of ligands from the other class becomes sigmoidal. For example, in the presence of AMPPCP or PEP, F6P shows n_H values of around two. This is in agreement with the equilibrium dialysis results (Deville-Bonne and Garel, 1992) that show two F6P binding sites per PFK tetramer. More surprisingly, AMPPCP and PEP binding in the presence of F6P also produces n_H value around 2 despite the demonstration by equilibrium dialysis of four binding sites for ATP per tetramer.

Fig 1.5 Models of *E.coli* PFK allostery. Lower case ligand indicates low affinity binding. Upper case indicates high affinity.

a) the two-state, Monod-Wyman-Changeux model.



b) the four-state model of Johnson and Reinhart, 1992.



Many of the fluorescence data are in agreement with the concerted allosteric model of Monod *et al* . (1965), particularly the identification of two states and the effects on affinities of interacting ligands. However, other data are not. In particular, n_H values of only 2 are observed for F6P binding whereas steady-state activity measurements give n_H=3.8-4.0. This suggests a kinetic, as well as a binding, aspect to F6P cooperativity. The lack of a simple relationship between n_H and number of binding sites had already been suggested by n_H values as large as 5.5-6.0 seen under some conditions (Deville-Bonne *et al.*, 1991a).

Another recent unexpected finding is the substrate antagonism of ATP and F6P; the apparent K_m for one is decreased about 20-fold on binding of the other (Deville-Bonne *et al.*, 1991b). In parallel, a regulatory role was suggested for ATP after the finding, by fluorimetry, that F6P binds with much greater affinity in the absence of ATP than in the presence of saturating ATP (Berger and Evans, 1991). A detailed fluorescence study of ATP and F6P interaction has been carried out (Johnson and Reinhart, 1992). The authors explained their results using a four state model (free enzyme, enzyme saturated with F6P, enzyme saturated with ATP and the catalytically competent ternary complex) rather than the two state MWC model. The four states have different affinities for the two substrates and probably for effectors too. Their four-state model is shown in Fig 1.5b. The PFK.ATP state tetramer binds its first F6P molecule weakly but later ones increasingly tightly.

In conclusion, the MWC model provides a good approximation to many basic characteristics of *E.coli* PFK but there is much evidence that it is incapable of encompassing a detailed description of the enzyme.

1.3.2 Rabbit Muscle PFK

RMPFK has been the best studied of eukaryotic PFKs but the attention focussed on this enzyme, and consequently our level of understanding, lag well behind *E.coli* PFK. One reason for this is the paucity of available biochemical methods for its study (see Section 4.1.3). The enzyme is also less stable than *E.coli* PFK and has proved impossible, so far, to crystallise.

The regulatory properties of RMPFK have been reviewed (Kemp and Foe, 1983). RMPFK

has six binding sites. Two of these are the catalytic ATP and F6P binding sites. There are activating allosteric sites for adenine nucleotides and sugar bisphosphates and inhibitory allosteric sites for citrate and ATP. A three state model for allosteric transitions in RMPFK has been proposed (Kemp and Foe, 1983). This includes two tetrameric states, one active and the other inhibited, and an inactive dimeric state. The tertiary and quaternary structures of RMPFK and potential locations of binding sites have been a matter for speculation (Poorman *et al.*, 1984). This is discussed in Section 2.1.4.

Chapter 2

Protein Modelling

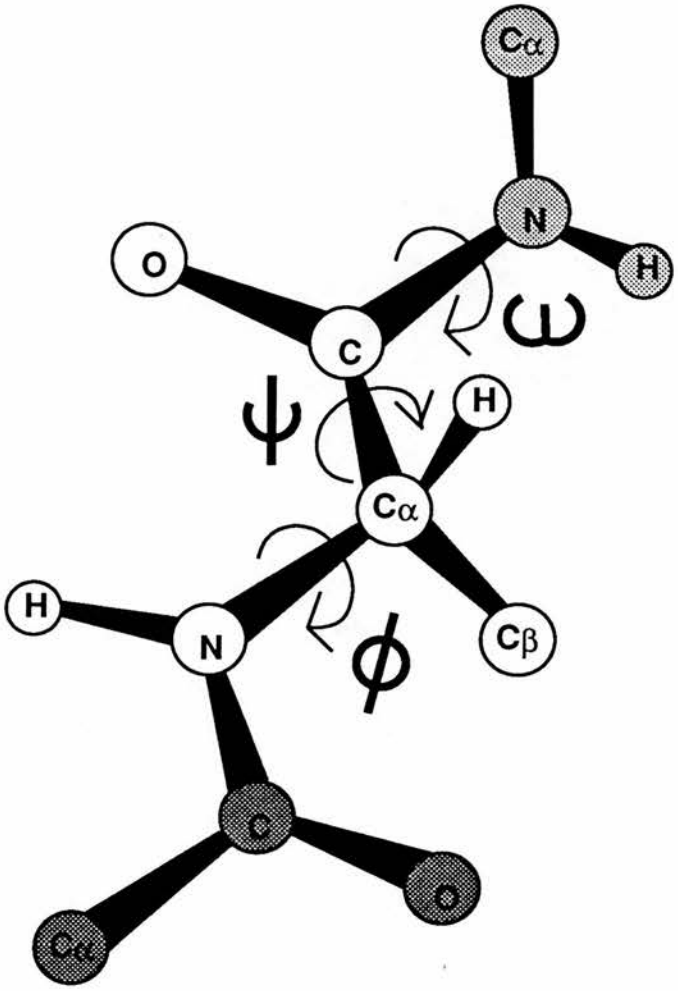
2.1 Introduction

Recent advances in DNA technology have made possible the facile determination of nucleotide sequences. Consequently the number of protein sequences known has risen rapidly. There are over 40,000 sequences deposited in the GenBank sequence database. Detailed understanding of the action of proteins requires the determination of their three-dimensional atomic structure. There are presently two ways to determine this *de novo*. The n.m.r. method is still confined to relatively small proteins of fewer than 150 residues. X-ray crystallography can be used for larger proteins but, despite recent advances, remains a slow and laborious process. Thus despite an exponential rise in the number of 3D protein structures deposited (Protein Databank Quarterly Newsletter, April 1992), the number of known structures lags an increasing distance behind the number of known sequences. This has focussed attention on prediction of 3D structure from protein sequence. Discussion of these efforts requires a brief introduction to protein structure.

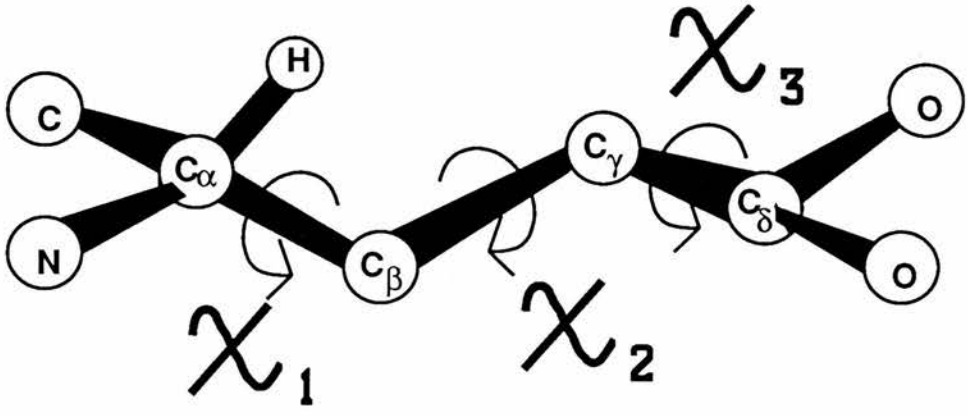
The database of known protein 3D structures has revealed the hierarchical nature of protein structure. At the simplest level, the primary structure defines the sequence of amino acids. The secondary structure is then defined by backbone conformation. The three backbone torsion angles ϕ, ψ and ω define rotation about the N-Cα, Cα-C and C-N bonds respectively (see Fig 2.1a). The ω angle is fixed at approximately 180° by the planarity of the peptide bond. Some of the ϕ,ψ combinations are disallowed by steric hindrance for non-glycine residues (Ramachandran, 1963). Of the remainder, some are compatible with repeating structures such as helices (α, 3_{10} and polyproline-type) or β-sheet. Such features are called secondary structure elements. When the amino acid does not belong to one of these structures, it is said to have a random coil conformation. The tertiary structure is defined by the packing of the secondary structure elements. For soluble proteins, this usually produces a compact, globular shape. The tertiary structure of a protein is sometimes called its fold. Finally,

Figure 2.1 : Definition of torsion angles

a) main chain angles



b) side chain angles



quaternary structure refers to the arrangement of polypeptide protomers in an oligomeric structure. Side chains are defined by a series of torsion angles $\chi_1 \cdots \chi_n$ referring to rotation about bonds Cα-Cβ, Cβ-Cγ etc. respectively. Figure 2.1b shows the definition of these angles for glutamate. The 'virtual' torsion angle ζ (Cα-N-C-Cβ) is sometimes used to measure Cα chirality.

For most proteins the three dimensional structure is thought to depend only on the amino acid sequence. This assumption arises from the observation of unfolding and refolding of isolated proteins on exposure to, and removal of, denaturants such as GdnHCl (Kim and Baldwin, 1990). This is thought to hold true despite the discovery of chaperonin proteins (Gething and Sambrook, 1992) which are required for folding or assembly of some, generally large and oligomeric, proteins. Theoretically, therefore, protein dynamics simulation should be able to determine protein structures. However, the huge computational demands render this still impossible for proteins of even moderate size (Abagyan, 1993).

Attempts can be made to use empirical guidelines to predict protein structure, generally with decreasing success as the level of complexity increases. Secondary structure can be predicted to some extent from primary structure. However, despite recent advances such as the use of neural networks, the accuracy of assignment of alpha helix, beta sheet or random coil remains stuck at just under 70%. The prediction of tertiary structure is even more problematic. Recent work has demonstrated the feasibility of determining overall protein folds if sufficient distance constraints between secondary structure elements are known (Smith-Brown *et al.*, 1993). Such constraints, for example from n.m.r. or fluorescence experiments were incorporated into the energy function used. Also included were torsion constraints from known or predicted secondary structure, a Lennard-Jones contribution and a dihedral angle term. Monte Carlo calculations on a polyglycine chain were used to sequentially satisfy the distance constraints. The best final structures showed r.m.s. Cα differences of 3-5Å to crystal structures; of the order of difference between, for example, globins sharing the same fold. However, modelling side chains, an obvious necessity for understanding of

protein function, increases the computational demands enormously.

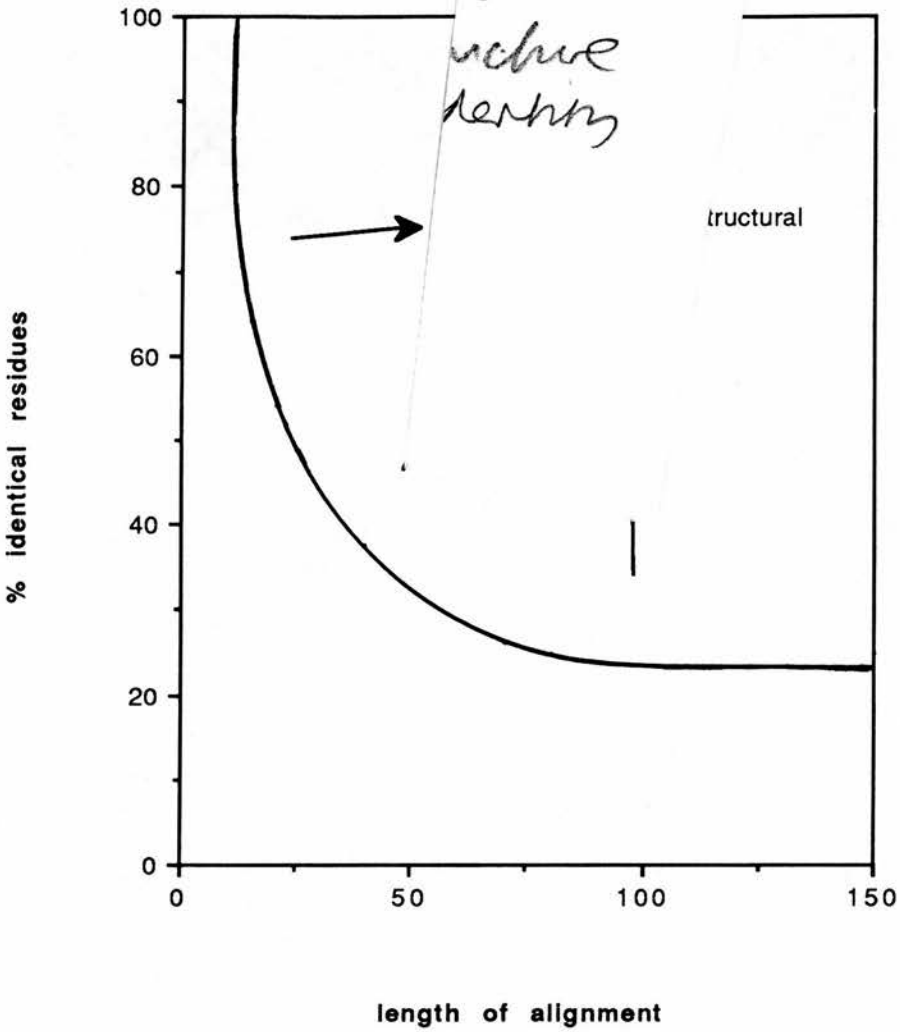
Hence, it is clear that at the present time all methods of prediction of protein 3D structure *de novo* , even with low-resolution structural information from, for example, CD and fluorescence, remain unsatisfactory.

2.1.1 The basis of modelling by homology

As the number of protein 3D structures determined has increased it has become clear that rather than each protein having its own fold, there is a finite number of folds which proteins can adopt. This has opened the way to modelling unknown protein sequences into known folds. Where possible, this approach can be rapid and quite accurate, but care must be taken in attributing to a new sequence a particular fold. The most obvious indicator of a particular fold is sequence similarity with a protein adopting that fold. The reliability of inferring shared fold from sequence similarity has been investigated by Sander and Schneider (1991). They produced standard sequence alignments for all proteins in a large structural database. Putting each alignment on a plot of length of alignment against % identical residues enabled the authors to identify two zones (see Fig 2.2). The region of safe structural homology contains essentially only sequence alignments corresponding to good structural similarity. Short alignment lengths require very high sequence identity whereas only >25% identity is required for alignments of longer than 80 residues to be identified as structurally similar.

It has become clear that sequence alone is not an ideal guide to structural relatedness. For example, distantly related globins, sharing the same fold, can show only 15% identity, well into the 'grey zone' of 10-20% identity seen for gapped alignments of random sequences. Function in this case would provide further evidence for structural similarity although this is not always the case. Different classes of serine proteases provide an example of similar function without structural similarity. Conversely, actin and a bovine heat shock protein show very good structural agreement at low sequence similarity (Kabsch *et al.*, 1990). Efforts have therefore been made to provide more

Figure 2.2 : Safe[illegible] 2 > [illegible]omology



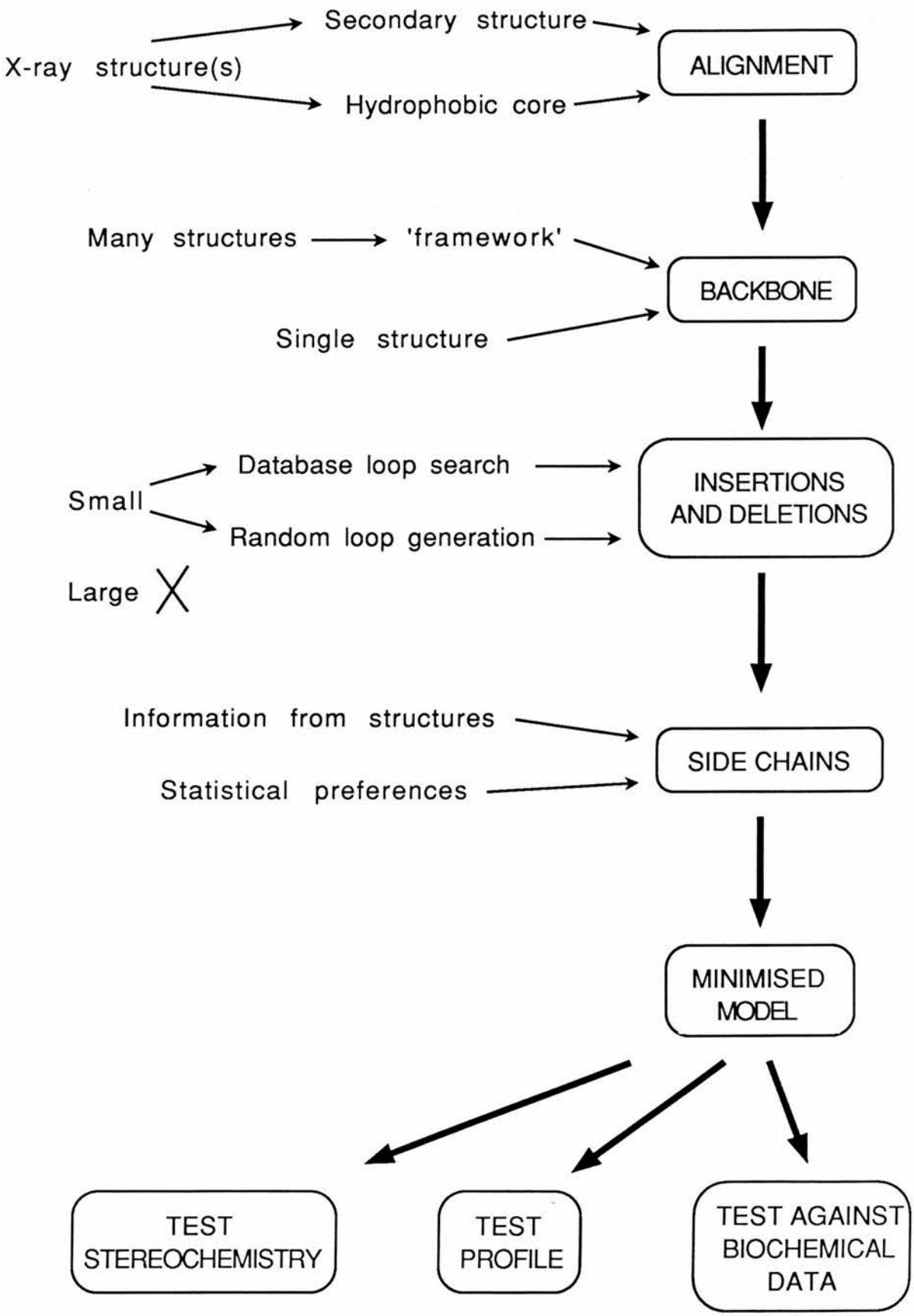
reliable indicators of structural relatedness. Three approaches have been able to predict structural similarity in the absence of significant sequence similarity for various test cases. Bowie *et al.* (1991) first calculate environments (including consideration of secondary structure, buried side chain area, % polar buried area) for residues in all known protein structures. A table is also compiled indicating the relative probabilities of finding each residue in each environment. Alignment of the unknown with all known 3D protein structures is then carried out to maximise a score relating to the favourability of finding the unknown's residues in the environments of the database protein. High scores indicate structural similarity. Jones *et al.* (1992) calculate energies for various 'threadings' of the unknown sequence onto known structures in three dimensions. This energy has two components, interaction potentials for each residue pair and a solvation potential. Expressions for these potentials are derived from database analysis. In most cases, threadings of sequences onto structures known to be related gave energy scores at the extreme lower end of the range observed for the whole database. Godzik *et al.* (1992) generate a 'structural fingerprint' for each protein in a structural library. This contains information on solvent exposure, pairwise interactions and triplet interactions. Again the energy parameters derive from database analysis. Low energy threadings from an initial approximate assessment are subject to a more detailed modelling. A protein lattice model is generated and a Monte Carlo rearrangement regime applied. The two important consequences of this procedure are a more accurate energy measurement and an indication of the stability of the model.

If structural relatedness is convincingly suggested by the above analysis then modelling by homology can be confidently attempted. The methods available for this are discussed in the following section.

2.1.2 Approaches to modelling by homology

A consensus approach has emerged from the work of a many groups. Here modelling by homology is a sequential process as shown in Fig 2.3. Most of the process can be

Figure 2.3 Modelling by homology



automated, the best known of these programs being COMPOSER (Blundell *et al.*, 1988). A different approach has been taken by Havel and Snow (1991) which will be discussed later.

I shall now consider the various steps of the consensus method in turn.

2.1.2.1 Alignment

The method of alignment to use depends on whether or not multiple homologous structures are available.

When only a single structure is available then it is used to modify a standard (multiple) alignment produced by a program such as CLUSTAL (Higgins and Sharp, 1988). Alterations are made to place insertions and deletions preferentially at the surface between secondary structure elements (Barton and Sternberg, 1987) and in order to avoid disturbance of the hydrophobic core (Kanaoka *et al.*, 1989). This has been shown to improve the quality of alignments.

When multiple structures are available then the first step is a structural alignment carried out to minimise inter-Cα distances between residues that correspond in the different structures. From this alignment Structurally Conserved Regions (SCRs), defined as stretches of topologically equivalent residues having inter-Cα distance of less than 3Å, can be identified. A sequence alignment with the unknown protein is then carried out incorporating this information to avoid placing insertions and deletions in SCRs.

2.1.2.2 Framework

A framework can only be determined when multiple homologous structures are available. Where only one structure has been determined then the main chain of this is used as the basis of the model.

A framework derives from the multiple structural alignment determined above (Sutcliffe *et al.*, 1987a). SCRs for the unknown are constructed using SCRs of homologous proteins. Previously this has been by either choosing the SCR with most

sequence identity to the unknown or by using a mean Cα position at each topologically equivalent position. Recently, superior results have been obtained using a weighted average Cα position (Srinivasan and Blundell, 1993). The contribution of Cα atoms in different SCRs is weighted either by sequence similarity to the unknown or the square of the sequence similarity. Particularly large improvements in results are seen where several distinct homologous structures are known that are equally related to the unknown.

2.1.2.3 Structurally Variable Regions (SVRs), Insertions and Deletions

SVRs are the regions between the SCRs in the structural alignment described above. They are the sites of any insertions and deletions but do not necessarily contain either. Where no insertion or deletion is present then the SVR used in the model of the unknown is chosen from the homologous set on the basis of best fit to the end of the SCRs and steric interactions (Topham *et al.*, 1990).

Various approaches have been developed for modelling insertions and deletions.

The most popular method for shorter loops is the database search. Polypeptide segments from the database of known protein structures, both homologous and non-homologous, are scanned to find those that can closely match 'anchor' atoms on either side of the insertion or deletion. This approach was first suggested by Jones and Thirup (1986). The most significant advance since then has been the development of a protocol to include energy assessment of the loop candidates (Summers and Karplus, 1990). Sequence similarity between the database loop and the unknown region of the model has sometimes been used in loop selection although the work of Sternberg and Islam (1990) shows that similar sequences can adopt very different conformations. More valid is the use of environmental amino acid substitution tables to rank loop candidates (Topham *et al.*, 1993). An analysis of 33 homologous protein families was used to assess the probability of substitution of one residue type with another in various environments. The environment definition included main chain conformation, side chain accessibility and side chain involvement in H-bonds. This analysis takes

account of the special conformational characteristics of glycines and prolines. When such analysis is unavailable these residues require special treatment. Glycines in candidate loops should be examined carefully since they can adopt backbone conformations not accessible to other amino acids. Prolines in the sequence to be modelled require special consideration since they can only adopt certain backbone conformations and influence the conformation of preceding residues (MacArthur and Thornton, 1991).

Database searches can fail to produce good loop candidates, particularly if the region to be modelled is large or contains residues that adopt unusual conformations. A more successful approach in these cases can be generation of loops with random backbone torsion angles which may then be assessed sterically and energetically. This approach was first used for the prediction of antibody hypervariable loop conformations (Fine *et al.*, 1986). Care must be taken with such results since commonly available implementations (such as present in SYBYL and INSIGHT) choose ϕ and ψ angles completely at random. In contrast, non-glycine amino acids are known to show strong preferences due to steric clashes between side chains and main chain in some conformations.

A third early approach to loop modelling has failed to gain favour. Moult and James (1986) used a systematic search procedure, intended to sample all possible main chain and side chain loop conformations, to model five-residue segments of a protease. Filters, such as avoidance of clashes, low electrostatic energy and low exposed hydrophobic surface area, were used to eliminate as many as possible of the huge number of candidate loops. Reasonable results were obtained but this approach is very demanding of computer time. Equally good results can be obtained with the quicker database loop search method for loops of this size.

2.1.2.4 Side Chains

There are two situations that may apply to a residue of the model whose side chain is to be constructed. In the first, side chains at equivalent positions in homologous

structures are of similar or greater size to the side chain to be modelled. In this case rules determined for substitution of homologous side chains (see below) can be applied. In the other case, less information is available. For example, the residue may be part of a non-homologous loop or may be the result of a mutation from Gly in homologous structures. In such instances, typical side chain conformations are modelled.

Analyses of amino acid substitutions in homologous proteins (Sutcliffe *et al.*, 1987b; Summers *et al.*, 1987) have discovered the general rule that the replacement side chain adopts a conformation such that it occupies as much of the space occupied by the original side chain as possible. Chemical nature is also considered so that a Thr would be modelled onto a Ser by overlay of the hydroxyls. In most cases this is implemented by transfer of corresponding side chain torsion angles. Several special cases have been discovered (Summers *et al.*, 1987). An example is the case of aromatic χ_2 angles which are most commonly ±90°. These are not equivalent for Trp so that both possible conformations should be considered.

The most commonly observed side chain conformations can be used when no information is available from homologous structures. These have been analysed by Sutcliffe *et al.* (1987b) and in more detail by Dunbrack and Karplus (1993).

Methods have been developed for the completely automatic addition of side chains to a given main chain without reference to homologous proteins. A simulated annealing method to optimise packing has been produced (Lee and Subbiah, 1991). Two groups have used rotamer libraries for initial placement followed by iterative energy-minimising refinement to model side chains (Wilson *et al.*, 1993; Dunbrack and Karplus, 1993). These methods are much more computer-intensive than rule-based modelling and it remains to be seen whether they can produce superior results.

2.1.2.5 The distance geometry approach

A very different approach to modelling by homology to that described above has been described (Havel and Snow, 1991). Their first step is the multiple alignment of

homologous sequences. A subset of atoms is then defined consisting of Cα and Cγ atoms of residues that are identical in all sequences. Distance constraints applying to the unknown structure are deduced from interatomic distances within this subset. Additional distance constraints are deduced from conserved H-bonds and disulphide bridges. Volume constraints from consecutive Cα positions are also determined in order to ensure the appearance of conserved secondary structure elements in the models of the unknown protein. Chirality constraints are also applicable from the knowledge that all amino acids are of the L chirality. All of these constraints are used as input to the DISGEO program (Havel and Wüthrich, 1984). This produces sets of structures that satisfy the constraints in a way directly analogous to structure determination by n.m.r. Advantages of this method are first its efficiency in terms of requirement for human input and computer time. Secondly, multiple structures are usually produced with the differences between them providing some indication of the reliability and precision of the predictions. Failure of the method is informative too since it indicates contradictions in the assumptions made in generating the constraints.

Performance seems to be worse than that of COMPOSER. In a test COMPOSER was used to build a phospholipase structure from three homologous structures of 40-50% identity (Srinivasan and Blundell, 1993). A comparable task was set for the DISGEO method - build a trypsin inhibitor from two homologous structures, again of 40-50% identity (Havel and Snow, 1991). Comparison of model with known structure gives an r.m.s. deviation between Cα atoms of 1.56Å for the COMPOSER model and 3.88Å for the DISGEO model. While remembering that these are distinct tasks, such a large difference in overall result suggests that the DISGEO method performs significantly worse.

There are many obvious improvements that could be made to the distance geometry method such as incorporation of side chain rotamer preferences, but it remains to be seen whether the method can perform as well as the established COMPOSER approach.

2.1.3. Testing of modelling results

Many methods have been devised for testing the validity of protein models built by homology and other methods. They can be divided into two classes. The first compares stereochemical parameters such as bond lengths and angles with typical values obtained from structures determined by crystallography or nuclear magnetic resonance spectroscopy. The second class looks at the environments of amino acids, particularly in relation to solvent or the hydrophobic protein core. Again expected distributions are determined from the database of known structures. Where available, biochemical information such as susceptibilty to proteolysis or chemical modification can also be checked against the model.

2.1.3.1 Stereochemical checks

A detailed study of the stereochemistry of main and side chains of proteins from the Brookhaven Protein Structure Databank (Bernstein *et al.*, 1977) has recently been published (Morris *et al.*, 1992). The most important indicators of good structure were found to be the main chain torsion angles ϕ and ψ, the main chain hydrogen bond energies and the first side chain torsion angle χ_1. A clear correlation was seen between these properties and the resolution of the X-ray structures. Other properties such as peptide bond planarity and disulphide bond angles were also studied.

The restrictions placed on the main chain dihedral angles of a protein, ϕ and ψ, by the presence of side chains, were first recognised by Ramachandran *et al.* (1963). Plots of ϕ against ψ for the whole protein database show pronounced clustering around values for regular secondary structures - alpha helix, beta sheet and to a lesser extent left-handed alpha helix. These plots can be digitised into zones by the degree to which they are populated. Good protein structures may then be expected to densely populate the most favoured regions of ϕ,ψ space with few non-glycine residues in the sparsely-populated 'disallowed' regions.

Values for the side chain dihedral angle χ_1 were found to cluster around three preferred conformers g^- (+60°), t (+180°) and g^+ (-60°). The extent of clustering provides an indication of the quality of the structure and outlying points should be treated as suspicious.

The distribution of hydrogen bond energies calculated for backbone atoms was also studied. It was found that the better, higher-resolution structures showed a tighter distribution of energies around the mean so that this distribution too can be used as a protein quality indicator.

A suite of programs (PROCHECK) has been written to check protein stereochemical quality using such indicators (Laskowski *et al.*, 1993).

2.1.3.2 PROFILE

The PROFILE program tests protein structures by assessing how well each amino acid is suited to the environment in which it is placed.

Eighteen environment classes were defined on the basis of side chain polarity and area buried (six classes) and secondary structure - alpha helix, beta sheet or other. (Bowie *et al.*, 1991). A database of sixteen protein structures and sets of homologous sequences were used to determine the probabilities of finding each amino acid type in a particular environment. A table was then produced containing entries for each residue in each environment. Each entry was a score calculated from the probability of finding a residue in a particular environment and the probability of finding that residue in any environment. Charged residues in exposed environments, for example, consequently score highly.

Used as a way of testing protein models (Lüthy *et al.*, 1992) PROFILE produces two results. An overall score for a model can be calculated by adding the residue scores. Typical values for proteins of different sizes have been determined with which the model value can be compared. Mean scores can also be calculated over a running window of a given length. Profile window plots can be used to localise probable

defects in the protein model.

The program was tested on proteins from the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977) and also on a series of deliberately misfolded structures and early models later shown to be incorrect. Most incorrect models were identified by low overall score. An incorrect *ras* p21 structure was subjected to profile window analysis and the program shown to be capable of localising the known errors in the structure.

2.1.3.3 Other methods

Solvent Free Energy (S.F.E.) of folding was one of the earliest discriminators of correct and incorrect protein models. Atomic solvation parameters for atom types found in side chains were calculated by least squares fit to observed distribution data relating to transfer of residues from octanol to water (Eisenberg and McLachlan, 1986). These can be used, along with solvent accessible atom areas, to calculate S.F.E.s for folded, misfolded and hypothetical unfolded states of a protein. The S.F.E.s of folding of correct and incorrect protein structures enabled their easy discrimination. Later work with a large database demonstrated a strong linear relationship between S.F.E. of folding and protein length (Chiche *et al.*, 1990).

Potentials of mean force (P.M.F.s) have been widely used to assess protein model validity. P.M.F.s are calculated from the relative probabilities of finding atoms (Cβ atoms are commonly used) at a certain separation, by the inverse Boltzmann law. A large dataset is used to derive the P.M.F.s which may then be used to calculate energies for protein models. One such study (Hendlich *et al.*, 1990) calculated potentials able to discriminate consistently between correct and incorrect models, excepting only proteins with prosthetic groups and non-globular proteins. Another study concentrating solely on hydrophobic interactions was also capable of good discrimination (Casari and Sippl, 1992).

Better results have been produced by an approach in which contact potentials are derived to discriminate best between a correct structure and a set of incorrect ones (Maiorov and Crippen, 1992). Here three types of contacts were considered - side-

chain-side chain, side chain-main chain and main chain-main chain, in contrast with previous emphasis only on side chains (Hendlich *et al.*, 1990). A linear relationship was seen between contact energies of native structures and their length.

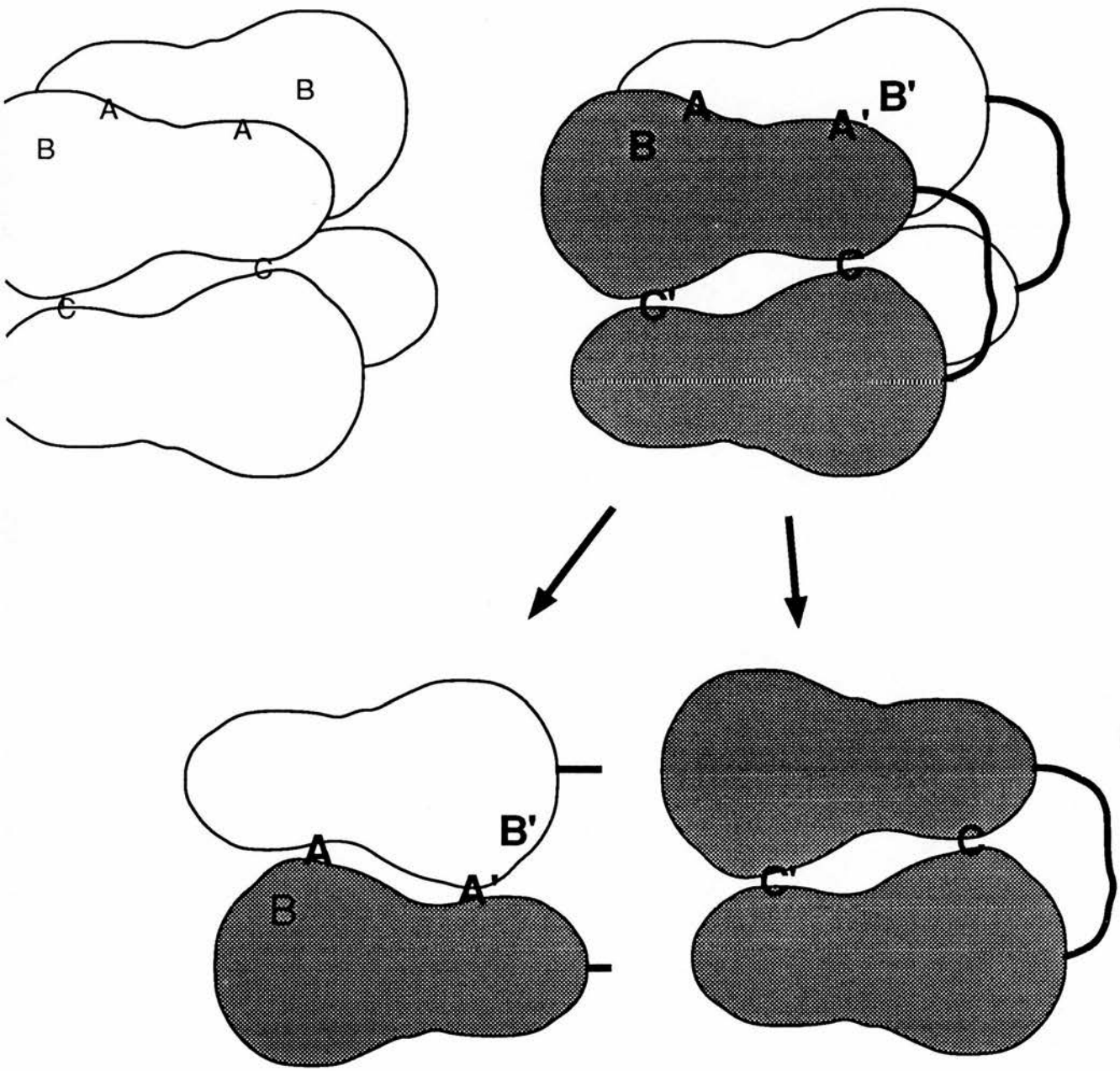
Holm and Sander (1992) use an approach similar to that used in PROFILE, focussing on the favourability of the potential protein-solvent interactions present in a protein structure. The differences are the consideration of different atom types within side chains and the ignoring of secondary structure. Like PROFILE, their approach was able to discriminate between correct and incorrect models by overall score and to give an indication of the positions of poor regions of the model.

2.1.4 The structure of Rabbit Muscle PFK

The only 3D structural information available on RMPFK has come from sedimentation studies and electron microscopy (Hesterberg *et al.*, 1981). This established that the active form of RMPFK was a tetramer with the individual subunits of 4 x 6 x 6 nm in size arranged with D_2 symmetry. With this knowledge and the recognition of internal duplication in the RMPFK sequence Poorman *et al.* (1984) were able to speculate first on the spatial arrangement of RMPFK protomers in the tetramer and then about possible effector sites.

The N- and C-termini of adjacent subunits in the bacterial PFK structure are close enough to be joined by the connecting peptide of about 30 residues that joins the N- and C-terminal halves of the rabbit muscle enzyme. Given the dimer-of-dimers structure of the bacterial enzyme there are two ways in which this could happen. The arrangement preferred on packing grounds puts the larger interface containing the C sites between N- and C-terminal RMPFK halves. This arrangement also puts the sites for the cooperative substrate F6P between subunits. Two of these rabbit muscle protomers are then equivalent to the bacterial tetramer. The quaternary structures of bacterial tetramer and proposed rabbit muscle dimer are shown in Fig 2.4. In the rabbit enzyme one protomer is shaded and one unshaded with the thick lines representing the connecting

cterial PFK tetramer Rabbit Muscle PFK dimer



re 2.4 : Quaternary structure of bacterial PFK tetramer and predicted quaternary structure of rabbit muscle PFK dimer
See text for details.

peptides. Beneath are two views of the rabbit muscle dimer showing A and A' sites between protomers and C and C' sites within protomers. Generating a RMPFK tetramer with D_2 symmetry from two such dimers could be done in two ways, placing all or none of the connecting peptides at the inter-dimer interface (see Fig 2.5). Poorman *et al.* favoured an interface position for the connecting peptides since they are susceptible to few proteases where as limit digestion with subtilisin degrades the N-termini which are exposed in this arrangement.

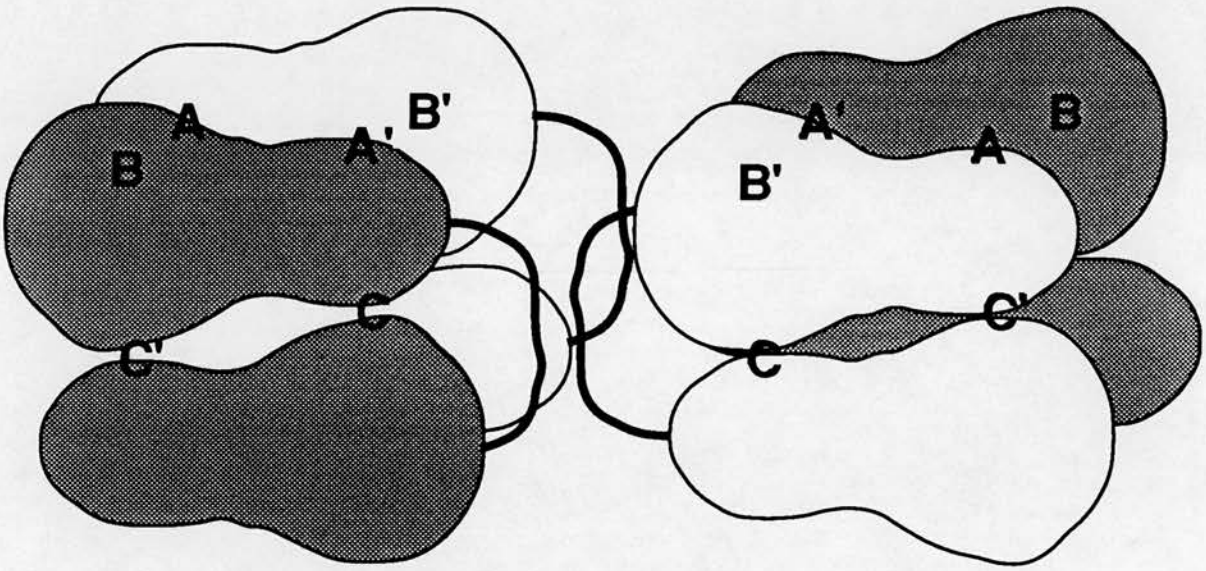
From sequence comparisons Poorman *et al.* propose ligands for the mutated binding sites of RMPFK. These they label A, B, C (in the N-terminal half) and A', B', C' (C-terminal half) by analogy with the A, B, C sites defined for the bacterial enzyme. Based on the conservation of the catalytic Ser127 in site A and its replacement by Ser in site A' they suggest A is the F6P substrate binding site. They propose A' as the sugar bisphosphate activating site since the Asp to Ser change would cause less repulsion of the second phosphate group. They suggest C is the ADP activation site and C' the ATP inhibitory site, rather than the reverse. This is because limit subtilisin digestion, which removes some N-terminal residues abolishes inhibitory ATP binding but not activating ADP binding. In the hypothetical RMPFK tetramer arrangement N-terminal residues contribute to the C' site whereas the C site is composed of residues central in the sequence. The assumption is made that site B remains the substrate ATP site but no speculation is made about the role of site B'.

2.2 Methods

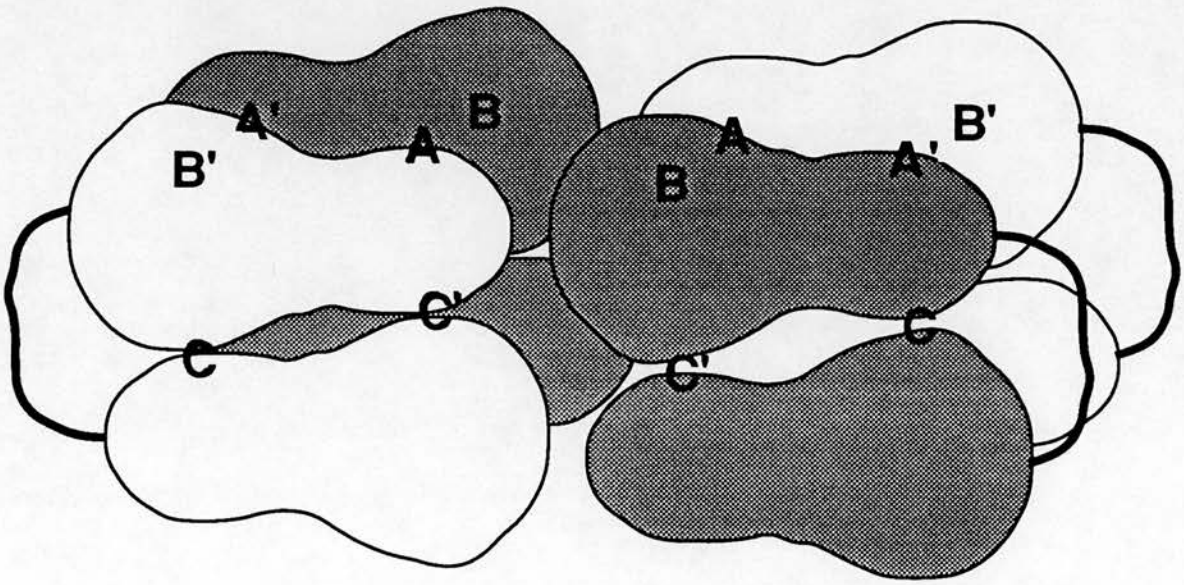
2.2.1 Modelling

Two models of rabbit muscle phosphofructokinase were produced. In each case the N-terminal and C-terminal halves (which both show homology to bacterial PFKs - see introduction) were modelled separately. They were then combined to produce RMPFK

Figure 2.5 : The two possible RMPFK tetramers with D2 symmetry



a) all connecting peptides together



b) connecting peptides apart

dimers of the kind postulated by Poorman *et al.* (1984).

The earlier model was produced largely by manual modelling. Later the program INSIGHT (INSIGHT II User Guide, 1992) became available including facilities for more automatic modelling. The second model was produced with as little intervention as possible so that a comparison of the two approaches could be made. The two models will be referred to as the manual model and the INSIGHT model.

The X-ray crystallographic structures of *B.stearothermophilus* PFK (Evans *et al.*, 1981) and *E.coli* PFK (Shirakihara and Evans, 1988) were used as the basis for the modelling. Both of these structures show PFK in the R-state, the former bound to phosphate, the latter bound to allosteric activator ADP and reaction products ADP and F16BP. These structures are very similar (Shirakihara and Evans, 1988) and show equal sequence similarities with rabbit muscle and other mammalian PFKs.

A model of the C' site was built from T-state coordinates of *B.stearothermophilus* PFK bound to inhibitor phosphoglycollate (Schirmer and Evans, 1990).

2.2.1.1 Validation of modelling by homology

The programs of Bowie *et al.* (1991) were used to check that the RMPFK N- and C-terminal halves were likely to adopt the fold of the bacterial PFKs. ENVIRONMENT calculated environments for residues of the *E.coli* PFK crystal structure bound to products using the secondary structure assignment produced by DSSP (Kabsch and Sander, 1988). PROFILESEARCH then scanned the supplied sequence database, to which PFK sequences had been added, for sequences likely to adopt the *E.coli* PFK fold. A gap opening penalty of 5 and a gap extension penalty of 0.35 were used.

2.2.1.2 Alignment

A published alignment (Lee *et al.*, 1987) was used as a basis for the manual model alignments. An alignment produced using CLUSTAL differed in positioning of some insertions and deletions. The bacterial crystal structure was used to guide the choice. The alternative which placed the insertion or deletion on the protein surface between

secondary structure elements was chosen. This has been shown to improve the quality of alignments when tested on homologous proteins the structures for which are known (Barton and Sternberg, 1987). Alignments can also be improved by avoiding disruption of the protein core (Kanaoka *et al.*, 1989). In this case the hydrophobic core was found to coincide with secondary structure elements and so provided no extra information. The alignments used for manual modelling are shown in Figs 2.6 and 2.7.

The INSIGHT model alignments were produced using commands in the HOMOLOGY module. A structural comparison of the *B.stearothermophilus* and *E.coli* PFK structures produced an alignment containing eight structurally conserved regions (SCRs). The N- and C- terminal halves of the rabbit muscle enzyme were then aligned using a standard algorithm, but including SCR information. The RM N-terminal alignment used is shown in Fig 2.6. The automatic alignment produced for the RM C-terminal half was extremely poor. The alignment from the manual model was therefore used modified slightly to take into account SCR information. This is shown in Fig 2.7.

2.2.1.3 Backbone

Given the close similarity of the *B.stearothermophilus* and *E.coli* PFK structures, the choice of backbone to use for modelling was arbitrary. The *B.stearothermophilus* structure was used.

The option to use an average structure over SCR regions for the INSIGHT model was unavailable so the *B.stearothermophilus* structure was again arbitrarily chosen.

2.2.1.4 Side Chains

For the manual models, the starting *B.stearothermophilus* side chains were mutated to the RM side chains using SYBYL. They were then adjusted manually according to the rules of Summers and Karplus (1989) using the homologous conformations in both *B.stearothermophilus* and *E.coli* structures. Where no guidance could be obtained

Figure 2.6 : Alignments used in modelling RMPFK N-terminal half. **Numbering is for B.stearothermophilus PFK.**

```
                                      10        20        30        40        50
E.coli          ----------MIKKIGVLTSGGDAPGMNAAI--RGVVRSALTEGLEVMGIYDGYLGLYED--RMVQL
B.stearo.       ----------MKRIGVLTSGGDSPGMNAAI--RSVVRKAIYHGVEVYGVYHGYAGLIAG--NIKKL
Manual model    MTHEEHHAARTLGVGKAIAVLTSGGDAQGMNAAV--RAVVRVGIFTGARVFFVHEGYQGLVDGGDHIREA
INSIGHT model   MTHEEHHAARTLGVGKAIAVLTSGGDAQGMNAAVRAVVRVGIFTGARVFFVHEGYQGLVDGGD--HIREA

                    60        70        80                                  90
E.coli          DRYSVSDMINRGGTFLGSARFPEFRDENIRA----------------------VAIENLKKRGIDALVV
B.stearo.       EVGDVGDIIHRGGTILYTARCPEFKTEEGQK----------------------KGIEQLKKHGIEGLVV
Manual model    TWESVSMMLQLGGTVIGSARCKDFREREGRL----------------------RAAHNLVKRGITNLCV
INSIGHT model   TWESVSMMLQLGGTVIGSARCKDFREREGRLRAAHNLVKRGITNLCVIGGDGSLTGADTFRSEWSDLLSD

                  110                                   120       130       140
E.coli          IGGDGSYMGAMRLTEMG----------------------FPCIGLPGTIDNDIKGTDYTIGFFTALSTV
B.stearo.       IGGDGSYQGAKKLTEHG----------------------FPCVGVPGTIDNDIPGTDFTIGFDTALNTV
Manual model    IGGDGSLTGADTFRSEWSDLLSDLQKAGKITAEEATRSSYLNIVGLVGSIDNDFCGTDMTIGTDSALHRI
INSIGHT model   LQKAGKITAEEATRSSY----------------------LNIVGLVGSIDNDFCGTDMTIGTDSALHRI

                150       160       170       180       190           200           210
E.coli          VEAIDRLRDTSSSHQRISVVEVMGRYCGDLTLAAAIAGGCEFVVVPE---VEFSREDLV---NEIKAGIA
B.stearo.       IDAIDKIRDTATSHERTYVIEVMGRHAGDIALWSGLAGGAETILIPE---ADYDMNDVI---ARLKRGHE
Manual model    TEIVDAITTTAQSHQRTFVLEVMGRHCGYLALVTSLSCGADWVFIPECPPDDNWEDHLC---RRLSETRT
INSIGHT model   TEIVDAITTTAQSHQRTFVLEVMGRHCGYLALVTSLSCGADWVFIPE---CPPDDNWEDHLCRRLSETRT

                  220                   230       240       250       260
E.coli          KGKKHAIVAITE-------H-------MCDVDELAHFIEKETGRETRATVLGHIQRGGSPVPYDRILASR
B.stearo.       RGKKHSIIIVAE-------G-------VGSGVDFGRQIQEATGFETRVTVLGHVQRGGSPTAFDRVLASR
Manual model    RGSRLNIIIVAE-------GAIDRNGKPITSEGVKDLVVRRLGYDTRVTVLGHVQRGGTPSAFDRILGSR
INSIGHT model   RGSRLNIIIVAEGAIDRNGK-------PITSEGVKDLVVRRLGYDTRVTVLGHVQRGGTPSAFDRILGSR

                270       280           290       300       310
E.coli          MGAYAIDLLLAGYGG---RCVGIQNEQLVHHDIIDAIENMKRPFKGDWLDCAKKLY
B.stearo.       LGARAVELLLEGKGG---RCVGIQNNQLVDHDIAEALANKHTIDQRMYALSKELSI
Manual model    MGVEAVMALLEGTPDTPACVVSLSGNQAVRLPLMECVQVTKDVTKAMDEKRFDEAMKLRGRSFMNNWEVY
INSIGHT model   MGVEAVMALLEGTPD---TPACVVSLSGNQAVRLPLMECVQVTKDVTKAMDEKRFDEAMKLRGRSFMNNW

E.coli
B.stearo.
Manual model    KLLAHIRPPAPKSGSYTV
INSIGHT model   EVYKLLAHIRPPAPKSGSYTV
```

Figure 2.7 : Alignments used in modelling RMPFK C-terminal half. **Numbering is for B.stearothermophilus PFK.**

```
                        10        20        30        40        50        60        70
E.coli          MIKKIGVLTSGGDAPGMNAAIRGVVRSALTEGLEVMGIYDGYLGLYEDRMVQLDRYSVSDMINRGGTFLGS
B.stearo.        MKRIGVLTSGGDSPGMNAAIRSVVRKAIYHGVEVYGVYHGYAGLIAGNIKKLEVGDVGDIIHRGGTILYT
Manual model         AVMNVGAPAAGMNAAVRSTVRIGLIQGNRVLVVHDGFEGPAKGQIEEAGWSYVGGWTGQGGSKLGS
INSIGHT model        AVMNVGAPAAGMNAAVRSTVRIGLIQGNRVLVVHDGFEGPAKGQIEEAGWSYVGGWTGQGGSKLGS

                       80        90       100               110       120       130
E.coli          ARFPEFRDENIRAVAIENLKKRGIDALVVIGG-------DGSYMGAMRLTEMGFPCIGLPGTIDNDIKGT
B.stearo.       ARCPEFKTEEGQKKGIEQLKKHGIEGLVVIGG-------DGSYQGAKKLTEHGFPCVGVPGTIDNDIPGT
Manual model    KRTL---PKKSFEQISANITKFNIQGLVIIGGFEAYTGGLEIMEGRKQFDELCIPFVVIPATVSNNVPGS
INSIGHT model   KRTLPKKSFEQISA---NITKFNIQGLVIIGGFEAYTGGLEIMEGRKQFDELCIPFVVIPATVSNNVPGS

                   140       150       160       170       180       190
E.coli          DYTIGFFTALSTVVEAIDRLRDTSS-SHQ-RISVVEVMGRYCGDLTLAAAIAGGCEFVVVPE---VEFSR
B.stearo.       DFTIGFDTALNTVIDAIDKIRDTAT-SHE-RTYVIEVMGRHAGDIALWSGLAGGAETILIPE---ADYDM
Manual model    DFSVGADTALNTICTTCDRIKQSAA-GTKRRVFIIETMGGYCGYLATMAGLAAGADAAYIFEEPFTIRDL
INSIGHT model   DFSVGADTALNTICTTCDRIKQSAAGTKR-RVFIIETMGGYCGYLATMAGLAAGADAAYIFE---EPFTI

                          210       220       230       240       250       260
E.coli          EDLVN---EIKAGIAKGKKHAIVAITEH----MCDVDELAHFIEKETG-RETRATVLGHIQRGGSPVPYD
B.stearo.       NDVIA---RLKRGHERGKKHSIIIVAEG----VGSGVDFGRQIQEATG-FETRVTVLGHVQRGGSPTAFD
Manual model    QANVE---HLVQKMKTTVKRGLVLRNEKCNENYTTDFIFNLYSEEGKGIFDSRKNVLGHMQQGGSPTPFD
INSIGHT model   RDLQANVEHLVQKMKTTVKRGLVLRNEKCNENYTTDFIFNLYSEEGKGIFDSRKNVLGHMQQGGSPTPFD

                     270       280                                           290       300
E.coli          RILASRMGAYAIDLLLAGYG---------------G----------------RCVGIQNEQLVHHDIIDAI
B.stearo.       RVLASRLGARAVELLLEGKG---------------G----------------RCVGIQNNQLVDHDIAEAL
Manual model    RNFATKMGAKAMNWMAGKIK---------------ESYRNGRIFANTPDSGCVLGMRKRALVFQPVTELQ
INSIGHT model   RNFATKMGAKAMNWMAGKIKESYRNGRIFANTPDSG---------------CVLGMRKRALVFQPVTELQ

                     310
E.coli          ENMKRPFKGDWLDCAKKLY
B.stearo.       ANKHTIDQRMYALSKELSI
Manual model    NQ-TDFEHRIPKEQWWLKLRPILKILAKYEIDLDTSEHAHLEHISRKRSGEATV
INSIGHT model   NQ-TDFEHRIPKEQWWLKLRPILKILAKYEIDLDTSEHAHLEHISRKRSGEATV
```

from these structures, the most common side chain conformation for the relevant secondary structure (Sutcliffe, 1987b) was generated. A SYBYL macro written for this purpose is listed in Appendix 1. Further manual adjustments were made to alleviate any clashes with main chain atoms or previously-positioned side chains.

The side chain conformations of the INSIGHT models were generated completely automatically. They were transferred from the *B.stearothermophilus* structure when identical. Where a mutation was to be made the program transfers as many dihedral angles as are available and produces an extended conformation for the remainder. No account was taken of the side chain conformations of homologous residues in the *E.coli* structure.

2.2.1.5 Loops

Insertions and deletions in the manual models were generated by two different methods - a database loop search algorithm and random generation. The basic loop search algorithm of SYBYL was greatly extended using three macro programs. These aimed to reproduce as far as possible the approach of Summers and Karplus (1990) and are listed in Appendix 2.

The first stage of this method involves a standard loop search using SYBYL. Both N- and C-terminal anchors are three residues in length and only the Cα atoms are used in the search. The loops showing the best fit to the anchors were then considered in three further tests (Summers and Karplus, 1990). The first of these tests whether the loop can be deformed, under constrained energy minimisation, to closely fit the anchors. Loops passing this test were then subject to two further tests. After further energy minimisation in the field of the protein the energy per residue of the loop and the energy consequence of loop insertion were measured. Loops passing these tests were inspected visually to check for favourable disposition of side chains in relation to solvent. If no suitable candidates were found for a given anchor pair then the anchors were iteratively shifted and the procedure repeated (Summers and Karplus, 1990).

If this loop search method failed then a random generation method was employed.

The SYBYL TWEAK command was used to generate a large number of loops with randomly chosen ϕ and ψ angles which fitted anchors on either side of the insertion/deletion site. Loops with main chain torsion angles mostly in favourable areas of the Ramachandran plot were considered. In cases where the loop contained one or more Pro residues then favoured Pro torsion angles were required giving greater selectivity. When a particular type of turn was strongly predicted (Wilmot and Thornton, 1988) then loops containing corresponding torsion angles were favoured.

For the INSIGHT models, all insertions and deletions were modelled using randomly generated loops. Loops closely fitting the anchors were judged visually for solvent exposure of side chains.

2.2.1.6 Energy minimisation of models

Each of the four models was subjected to 200 rounds of energy minimisation *in vacuo* using XPLOR. A simulated annealing procedure was then carried out consisting of 1ps of molecular dynamics simulation at 600K followed by cooling to 300K over 0.6ps. A further 200 rounds of energy minimisation was then applied to each.

2.2.1.7 Modelling dimers

The RM PFK dimers were constructed by fitting models for the N- and C-terminal halves over the monomers of the tetrameric bacterial enzyme. Atoms used in this fit were the Cα atoms of residues defined as active site or allosteric site in the bacterial enzyme crystal structure (Shirakihara and Evans, 1988) and the Cα atoms of homologous residues in RM N- and C-terminal halves. No attempt was made to model the interfaces in detail since the uncertainty in relative positioning was so great. For the same reason a model of the RM PFK tetramer was not produced.

2.2.1.8 Modelling the T-state C' site with citrate bound.

This model was constructed by mutating side chains of the C site into the corresponding RMPFK residues (Shirakihara and Evans, 1988). The rules for side

chain conformations were as above. Low energy conformations for citrate were investigated using SYBYL's SEARCH command with electrostatics included using charges assigned by the method of Gasteiger and Marsali (1980). Among the 12 lowest energy conformations there were two groups of three, the members of each differing only in rotation of the hydroxyl group. Visual inspection showed that these two groups had carboxylates in similar positions. One of these conformations was used for subsequent docking. The citrate was initially positioned in the model so two of its carboxylates were in favourable GRID carboxylate energy contours of about -5 kcal/mol. The GRID program (Goodford, 1985; Boobbyer *et al.*, 1989) was used extensively for ligand modelling (Chapter 3) and will be discussed there. The DOCK command of SYBYL was then used (again with Gasteiger and Marsali charges) to search for lower energy positions. This command provides a continual approximation of the docking energy which was minimised by manual rotation and translation of the citrate. This involved movement of the citrate further into the pocket towards Arg21 and enabled relief of a bad contact between one carboxylate and Arg25. The structure was then energy minimised using SYBYL's ANNEAL command centred on citrate and with 'Hot' and 'Interesting' radii of 4 and 6Å respectively. Inspection of the minimised structure showed that relocation of Arg154 would place it favourably close to the citrate carboxylates. After this adjustment the structure was again energy minimised during which Arg154 and the citrate moved closer together. The final result showed Arg154 salt bridged to the citrate.

2.2.2 Testing

2.2.2.1 PROCHECK

The PROCHECK program was run on each of the four models and for comparison on the *B.stearothermophilus* PFK structure upon which the models were built. The resolution parameter required by the program was set to 2.4Å as the resolution to which the *B.stearothermophilus* PFK structure was determined.

2.2.2.2 PROFILE

The PROFILE program was run on each of the four models. The models were present in the context of the other RMPFK halves which they contact in the RMPFK dimer model. This was shown to improve their scores compared to the isolated model presumably by taking into account burial of hydrophobic residues at the interfaces between RMPFK halves. PROFILE requires a secondary structure definition which was generated for each model using DSSP.

2.2.2.3 Others

The RMPFK dimer model was checked visually for consistency with biochemical data from chemical modification with iodoacetate (Latshaw *et al.* 1987), pyridoxal phosphate (Kemp *et al.*, 1987) and *p*-fluorosulphonylbenzoyl-5'-adenosine (Weng *et al.*, 1980) and proteolysis by V8 protease (Valaitis *et al.*, 1987).

2.3 Results and Discussion

2.3.1 Construction of the models

The 3D profile programs of Bowie *et al.* (1991) showed that the N- and C-terminal halfs of RMPFK are very likely to adopt the fold of the bacterial PFKs. The output of the PROFILESEARCH program is shown in Fig 2.8. A line has been added between those sequences very likely to adopt the fold (those with a Zscore over 8 according to the authors) and those unlikely to adopt the fold. The bacterial enzymes score very highly as expected, but in addition, both halfs of RMPFK score over 8, the N-terminal half 18 and the C-terminal half 8.5. These results justify the use of the bacterial PFK fold for modelling of RMPFK by homology. The sequences of the known

Fig2.8 Result of PROFILESEARCH using E.coli PFK

Name		ZScore	Score	Length		Documentation ..
X0005	+	62.50	134.85	320.	!	ESCHERICHIA COLI PFK
X0004	+	52.56	115.33	319.	!	B.STEAROTHERMOPHILUS PFK
X0006	+	18.17	48.49	403.	!	RABBIT MUSCLE PFK N-TERMINAL HALF
X0002	+	10.19	32.88	615.	!	PPi-DEPENDENT PFK FROM POTATO TUBER - ISOZYME A
X0003	+	10.04	32.51	512.	!	PPi-DEPENDENT PFK FROM POTATO TUBER - ISOZYME B
X0007	+	8.57	29.34	377.	!	RABBIT MUSCLE PFK C-TERMINAL HALF
A02506	+	3.17	16.45	153.	!	Sperm Whale myoglobin
C8HUB	+	3.04	18.60	1565.	!	Complement component C8 beta chain precursor - Huma
MYWHC	+	2.64	15.52	153.	!	Myoglobin - California gray whale
MYWHL	+	2.25	14.83	153.	!	Myoglobin - Killer whale
HUMMHSP70F_1	+	2.21	11.12	65.	!	MHC heat shock protein HSP70-2 (AA at 1) 0/0
JWEBT	+	2.19	15.46	190.	!	DNA-invertase - Salmonella sp.
FEAH	+	1.90	12.36	96.	!	Ferredoxin - Aphanothece sacrum
GNNYH2	+	1.90	16.31	1895.	!	Genome polyprotein - Rhinovirus 2
ODUTMB	+	1.87	16.25	804.	!	Cytochrome-c oxidase (EC 1.9.3.1) polypeptide I - T
HUMFIXH_1	+	1.85	12.54	102.	!	Hemophilic factor IX, exon 8 (AA at 1) /nomgen="F9"
HUMFMSB_1	+	1.81	9.73	52.	!	Fms protein, exon 2 (AA 277 at 58) 0/0
KIHUG	+	1.78	15.97	417.	!	Phosphoglycerate kinase (EC 2.7.2.3) - Human and ho
MYBO	+	1.72	13.91	153.	!	Myoglobin - Bovine
MYHH	+	1.63	13.75	153.	!	Myoglobin - Western European hedgehog
MNIV2	+	1.49	12.66	121.	!	Nonstructural protein NS2 - Influenza A virus (stra
WJFFBX	+	1.46	10.65	74.	!	Ultrabithorax homeotic protein - Fruit fly (fragmen
SYECVT	+	1.46	15.42	696.	!	Valine--tRNA ligase (EC 6.1.1.9) - Escherichia coli
DROEGFRD_1	+	1.44	15.40	1071.	!	Pre-epidermal growth factor receptor, exon 3 (AA at
B26147	+	1.39	14.49	232.	!	Egg-laying hormone-2 precursor - Little sea hare (f
TARHBG_1	+	1.36	13.17	148.	!	Gamma-globin, exon 1 gamma-globin, exon 2 gamma-glo
FENM2M	+	1.36	11.62	98.	!	Ferredoxin II - Nostoc muscorum
HAMQR	+	1.35	12.99	141.	!	Hemoglobin alpha chain - Rhesus macaque and Japanes
X0001	+	1.33	15.05	403.	!	PPi-DEPENDENT PFK FROM PROPRIONIBACTERIUM
HUMBHRPA_1	+	1.32	9.07	51.	!	Basic histidine-rich protein precursor /nomgen="HRG
HJECD2	+	1.31	15.13	788.	!	Helicase II (EC 3.6.1.-) - Escherichia coli
A26147	+	1.30	14.54	263.	!	Egg-laying hormone-1 precursor - Little sea hare

pyrophosphate-dependent PFKs from potato tuber and *Propionibacterium freudenreichii* were also included. There has been debate over whether these enzymes are evolutionarily related to the ATP-dependent PFKs. The results in Fig 2.8 strongly suggest that the potato enzymes share the same fold as the ATP-dependent PFKs and are therefore likely the result of divergent evolution. The *Propionibacterium* enzyme, in contrast, scores very poorly suggesting that it has a different fold and is not evolutionarily related.

Few problems were encountered in applying the side chain subsitution rules of Summers *et al.* (1987) during construction of the manual models. Small manual alterations were sufficient to avoid clashes in the cores of the models. Figure 2.9 shows the backbones of the manual models with the *B.stearothermophilus* PFK structure on which they were based. The loops are labelled, with numbering as in Table 2.1, and coloured red (modelled) and magenta (unmodelled). N- and C-termini are also labelled and are coloured blue.

Application of the database loop search algorithm was more problematic in some cases than in others. Table 2.1 shows the modelled and unmodelled insertions. The length values are positive for insertions and negative for deletions. The source of the fragment used is shown as the Brookhaven code in brackets. The number of iterations of the search algorithm required before a loop was selected varied from 3 to 11, with a mean around 5. Loops that passed the tests describe in Methods were assessed visually for general packing, exposure of charged side chains and burial of hydrophobic side chains. In the case of C-terminal half insertion 3 there was a special requirement that the salt bridge to the 6-phosphate of bound sugar bisphosphate be preserved. This is provided by Arg162 in the bacterial enzyme. A successful loop search provided a subtilisin segment that placed RMC Arg162 in a suitable position to form the salt bridge. This insertion is shown in Fig 2.10. The failure of the loop search algorithm to provide suitable candidates for the seven-residue insertions was expected since these are at the upper length limit of the method's capabilities. Modelling of the 15- and 23-residue insertions was not attempted. The algorithm's failure with the four-residue

Table 2.1: Modelling of insertions and deletions

Manual Model :N-terminal half

No.	Residue preceding	Length	Modelled?	Method
1	Gly61	2	yes	Database loop search (2CAB)
2	Trp132	23	no	
3	Glu232	3	yes	Random generation
4	Gly265	7	no	
5	Asp330	3	yes	Database loop search (2APR)

Manual Model :C-terminal half

No.	Residue preceding	Length	Modelled?	Method
1	Leu70	-3	yes	Database loop search (2SBT)
2	Gly95	7	no	
3	Lys161	1	yes	Database loop search (2SNI)
4	Glu194	3	yes	Database loop search (2CTS)
5	Lys227	4	no	
6	Gly247	1	yes	Database loop search (2APR)
7	Glu290	15	no	
8	Gln326	-1	no	

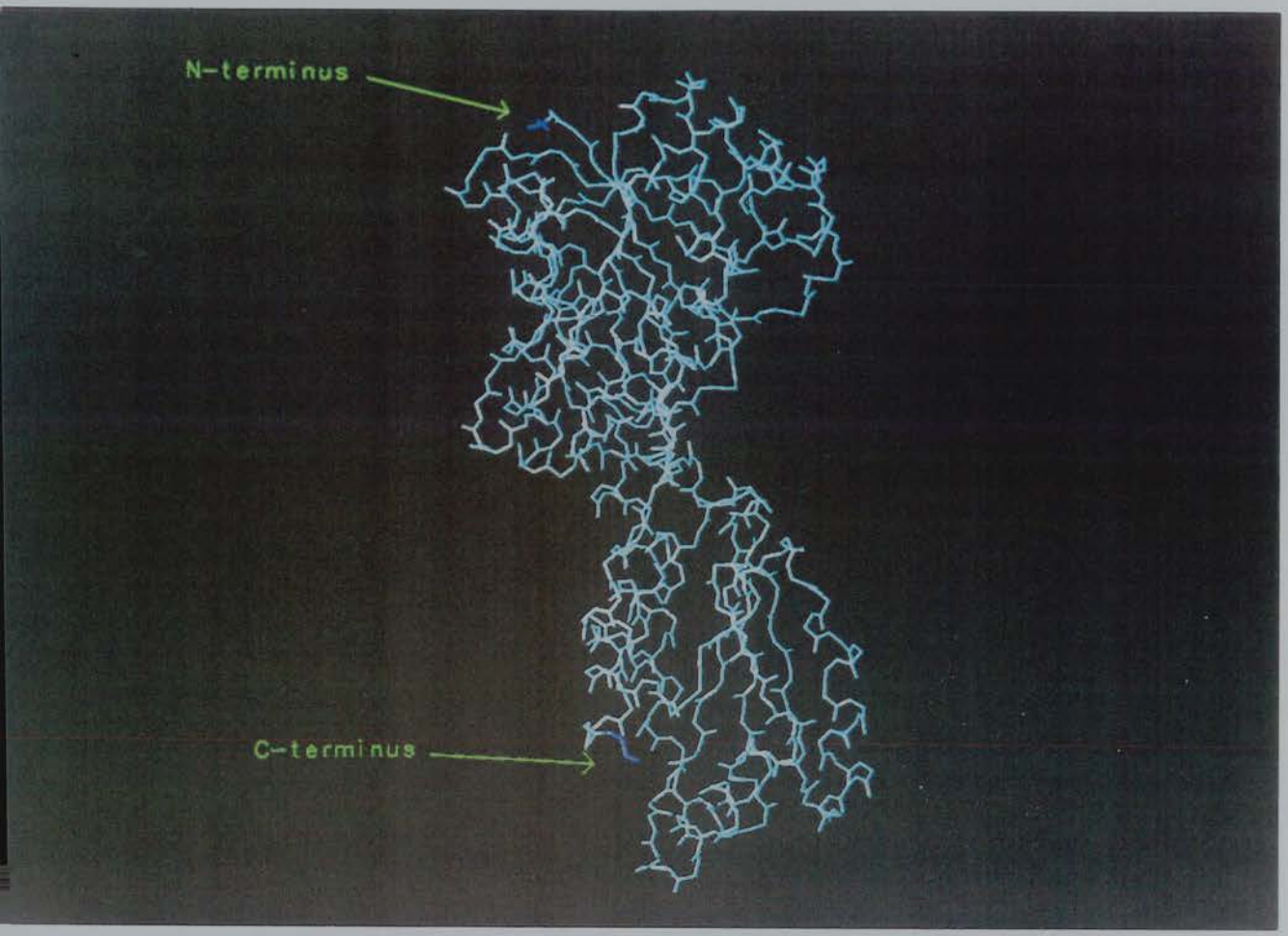
NSIGHT Model :N-terminal half

No.	Residue preceding	Length	Modelled?	Method
1	Gly97	25	no	
2	Asp241	3	yes	Random generation
3	Glu264	7	yes	Random generation
4	Met349	-1	yes	Random generation
5	Lys356	-4	yes	End Joining

NSIGHT Model :C-terminal half

No.	Residue preceding	Length	Modelled?	Method
	Ala80	-3	yes	Random generation
	Gly95	7	yes	Random generation
	Ala158	1	yes	Random generation
	Ala204	3	yes	Random generation
	Lys227	4	yes	Random generation
	Lys289	15	no	

Fig 2.9 Main chain atoms of *E.coli* PFK crystal structure,



manual N-terminal half model and C-terminal half model. (see text)

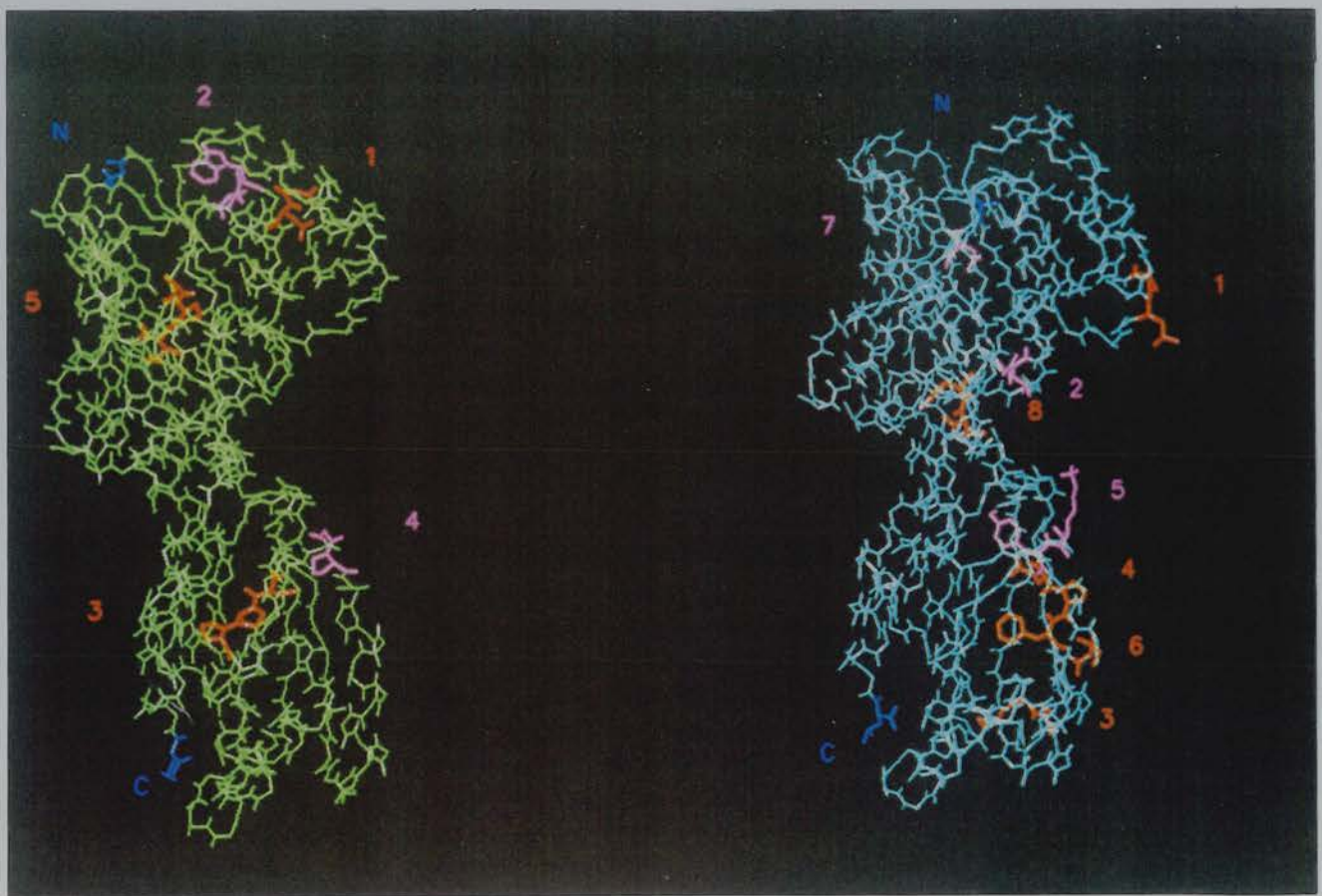


Fig 2.10 Manual C-terminal half model - insertion 3

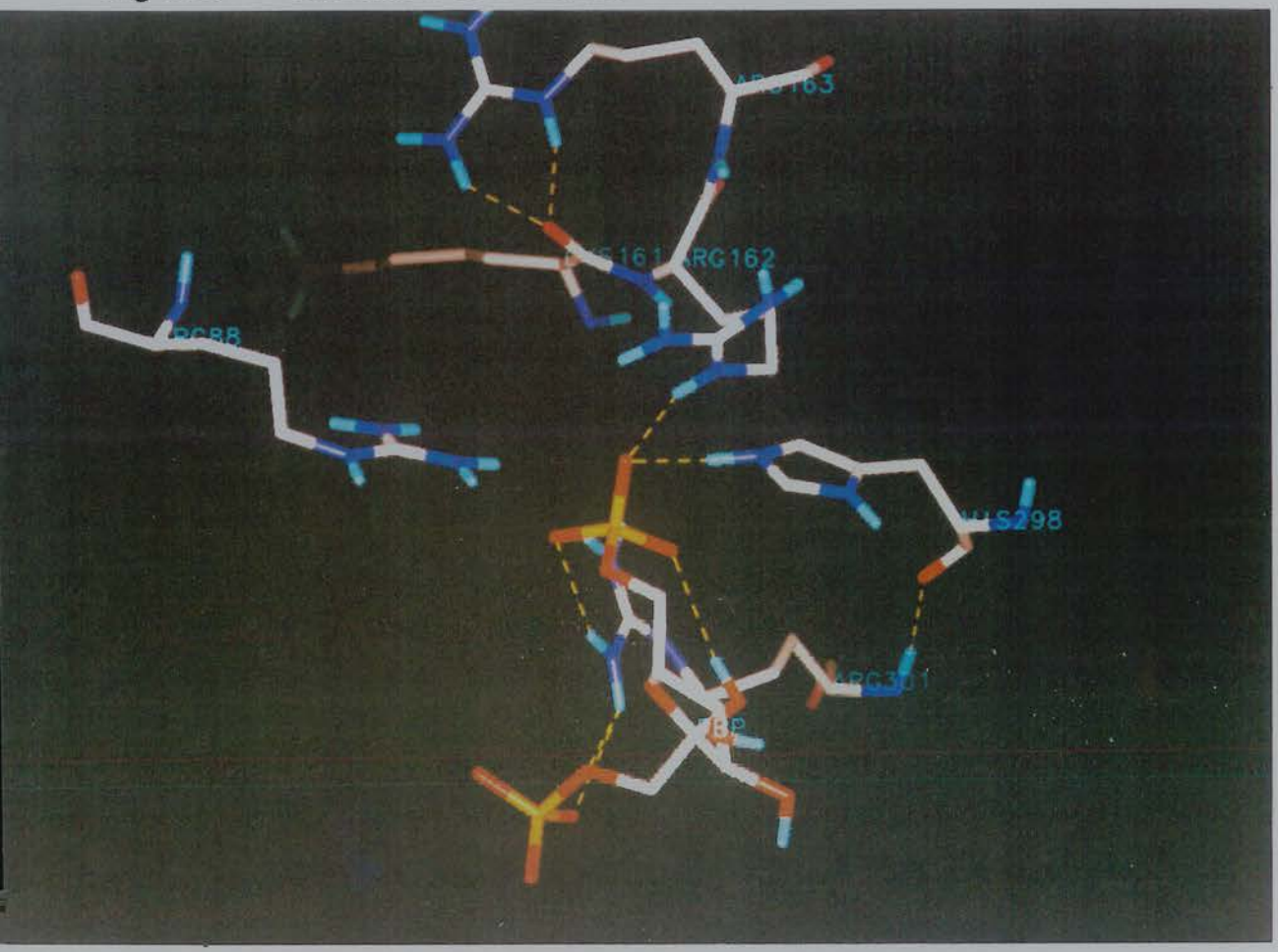
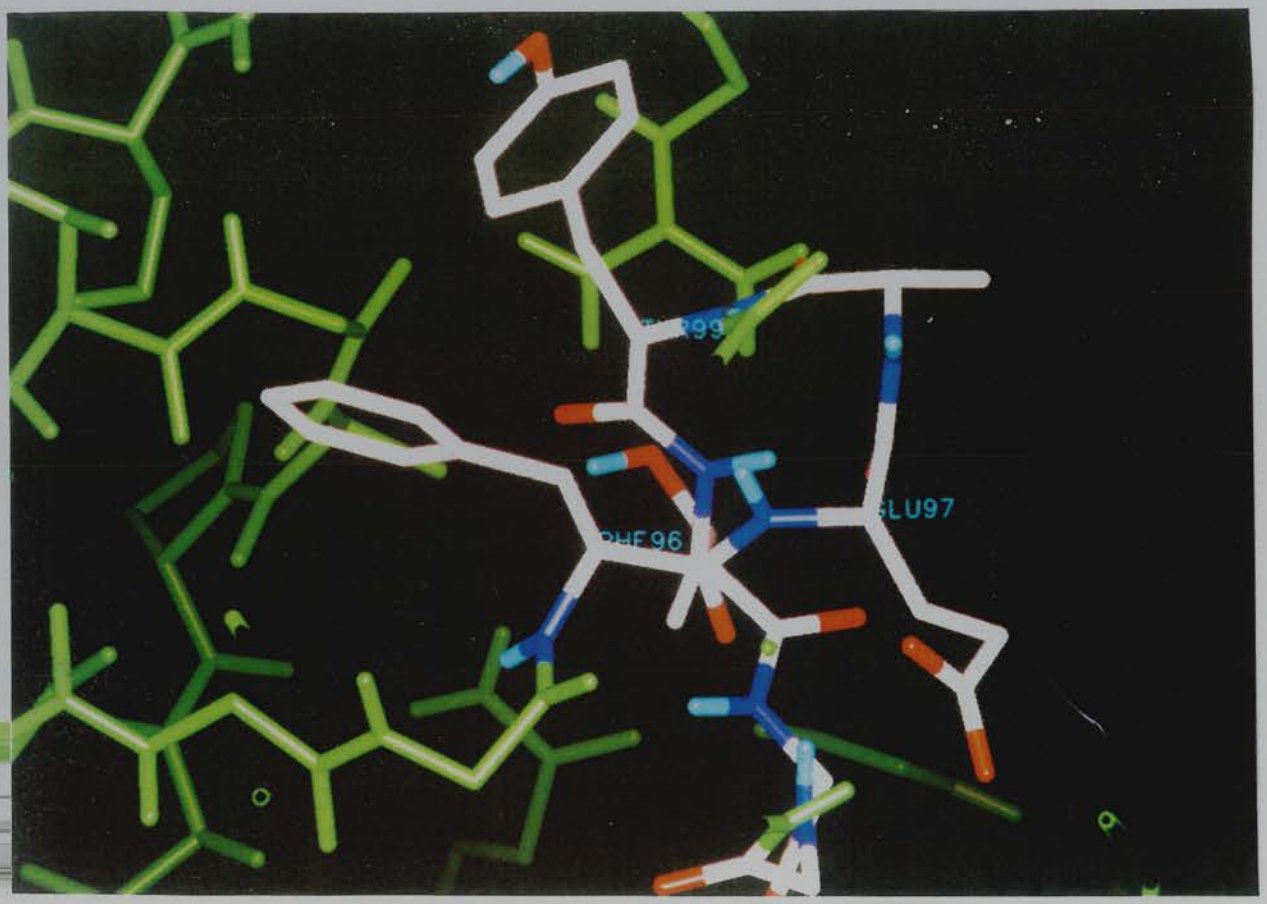


Fig 2.11 INSIGHT C-terminal model - insertion 2



insertion of the C-terminal half was unexpected. INSIGHT provided a randomly generated loop candidate for this position in the INSIGHT model, but this performed poorly in tests (see below). Insertion 3 of N-terminal half was also not satisfactorily modelled by the loop search algorithm. This contains the sequence PPDDN. Application of the B turn analysis of Wilmot and Thornton (1988) suggested that a type I turn was likely here, more probably at the sequence PPDD. The conformations adopted by prolines and their effect on preceding residues (MacArthur and Thornton, 1991) provided backbone torsion angle constraints which were applied to a set of 100 randomly generated loops from SYBYL's TWEAK command. Candidates with approximately correct backbone torsion angles were assessed visually as above and the best candidate chosen.

Randomly generated loops for the INSIGHT models were again chosen for good packing and solvent exposure. In addition a high percentage of residues with favourable backbone torsion angles was looked for. This method was able to produce models for seven-residue insertions. That in the C-terminal half appears well modelled (see Fig 2.11), burying both Phe96 and Tyr99 and exposing Glu97 to solvent.

2.3.2 Testing

The models were checked for stereochemistry and good solvent exposure using PROCHECK and PROFILE respectively. They were then checked for consistency with biochemical knowledge of the enzyme, both of its ligands and of its susceptibility to chemical modification and proteolysis.

2.3.2.1 PROCHECK

The PROCHECK program was used to compare main chain and side chain parameters of the models with those expected for good protein structures.

Overall quality

The Ramachandran plots for the manual and INSIGHT models of the RMPFK N- and C-terminal halfs are shown in Figs 2.12 - 2.15. For comparison Fig 2.16 shows the Ramachandran plot for the *B.stearothermophilus* X-ray crystal structure on which the models were based. For consistency the crystal structure was subject to the same energy minimisation and molecular dynamics regime as the models.

On the Ramachandran plots the regions are defined by the analysis of Morris *et al.* (1992). The darkest regions correspond to core alpha, beta and left-handed alpha regions. These are labelled A, B and L. Lighter regions labelled a, b, l and p correspond to allowed alpha, beta, left-handed alpha and epsilon. Lighter still are the generously allowed alpha, beta, left-handed alpha and epsilon (~a, ~b, ~l, ~p). The lightest coloured regions are disallowed. The percentage of residues in the core regions was found to be a good indicator of stereochemical quality with good models at resolution <2.0Å, R factor <20% expected to have at least 90% of residues here.

Table 2.2 summarises the distribution of residues in the models and crystal structure between these regions. This shows that the manual models, by this criterion, are of better stereochemical quality than those generated by INSIGHT. The manual models are comparable to the crystal structure after the molecular dynamics procedure.

Other main chain parameters are summarised by PROFILE in terms of number of band widths from the mean expected for a particular resolution of structure. Where the particular value is within one band width of the mean a structure is rated as Better or Worse. Outside one band width the ratings BETTER or WORSE are given.

Table 2.3 shows these results for the model structures and the crystal structure. The parameters for bad contacts, zeta angles and H-bond energies are good for all models. The ratings for % residues in core regions confirm the superiority of the manual models over the INSIGHT models in this respect. The omega angles parameter is poor for all models and also, unexpectedly for the crystal structure after minimisation and molecular dynamics. When PROCHECK was run on the original crystal structure, a rating for omega angles of BETTER was obtained. This suggests that the poor omega rating after

Fig 2.12

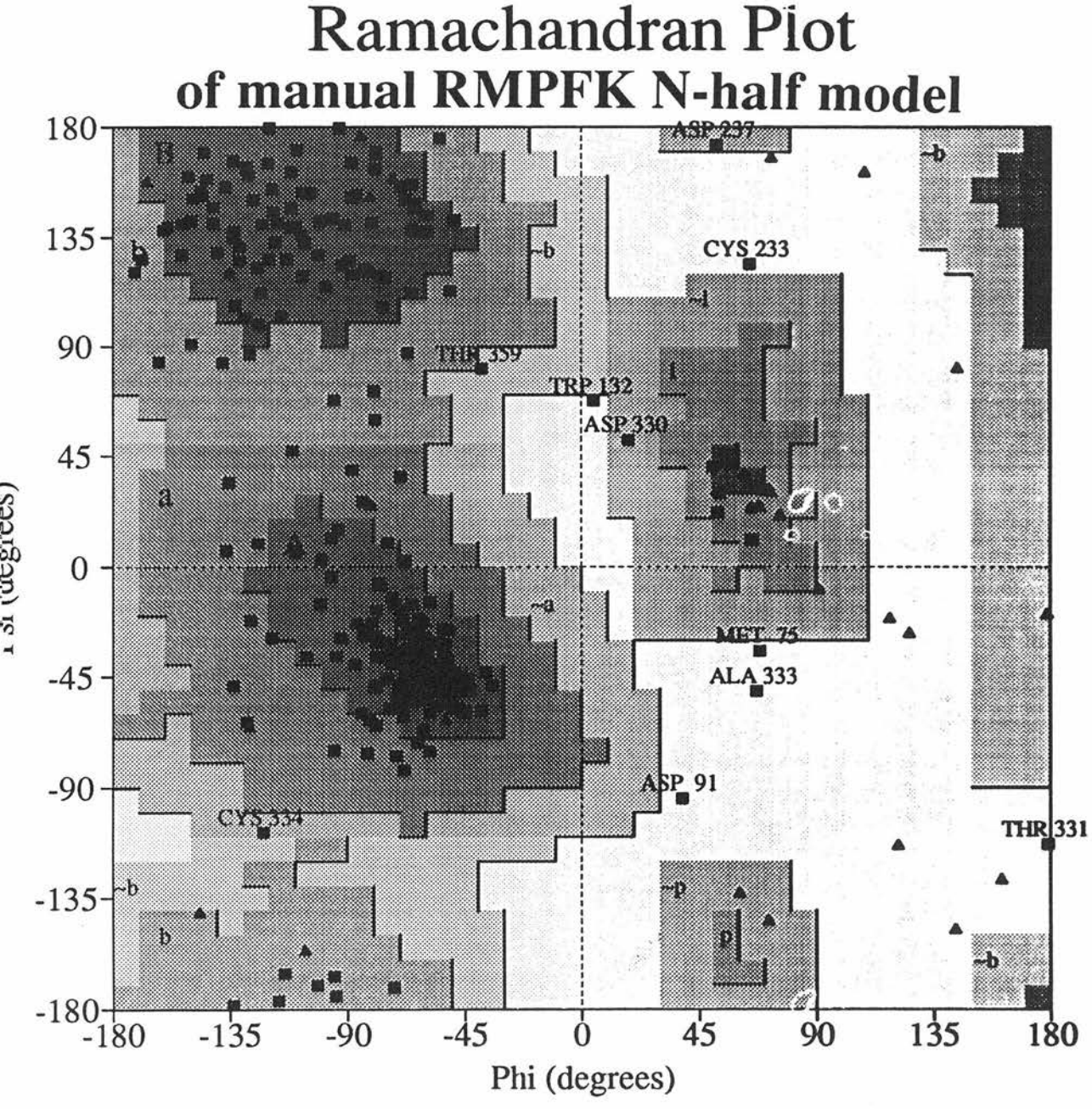

Ramachandran Plot
of manual RMPFK N-half model
Psi (degrees)
Phi (degrees)
180
135
90
45
0
-45
-90
-135
-180
-180
-135
-90
-45
0
45
90
135
180
B
b
~b
a
~a
l
~l
p
~p
ASP 237
CYS 233
THR 359
TRP 132
ASP 330
MET 75
ALA 333
ASP 91
CYS 334
THR 331

Fig 2.13

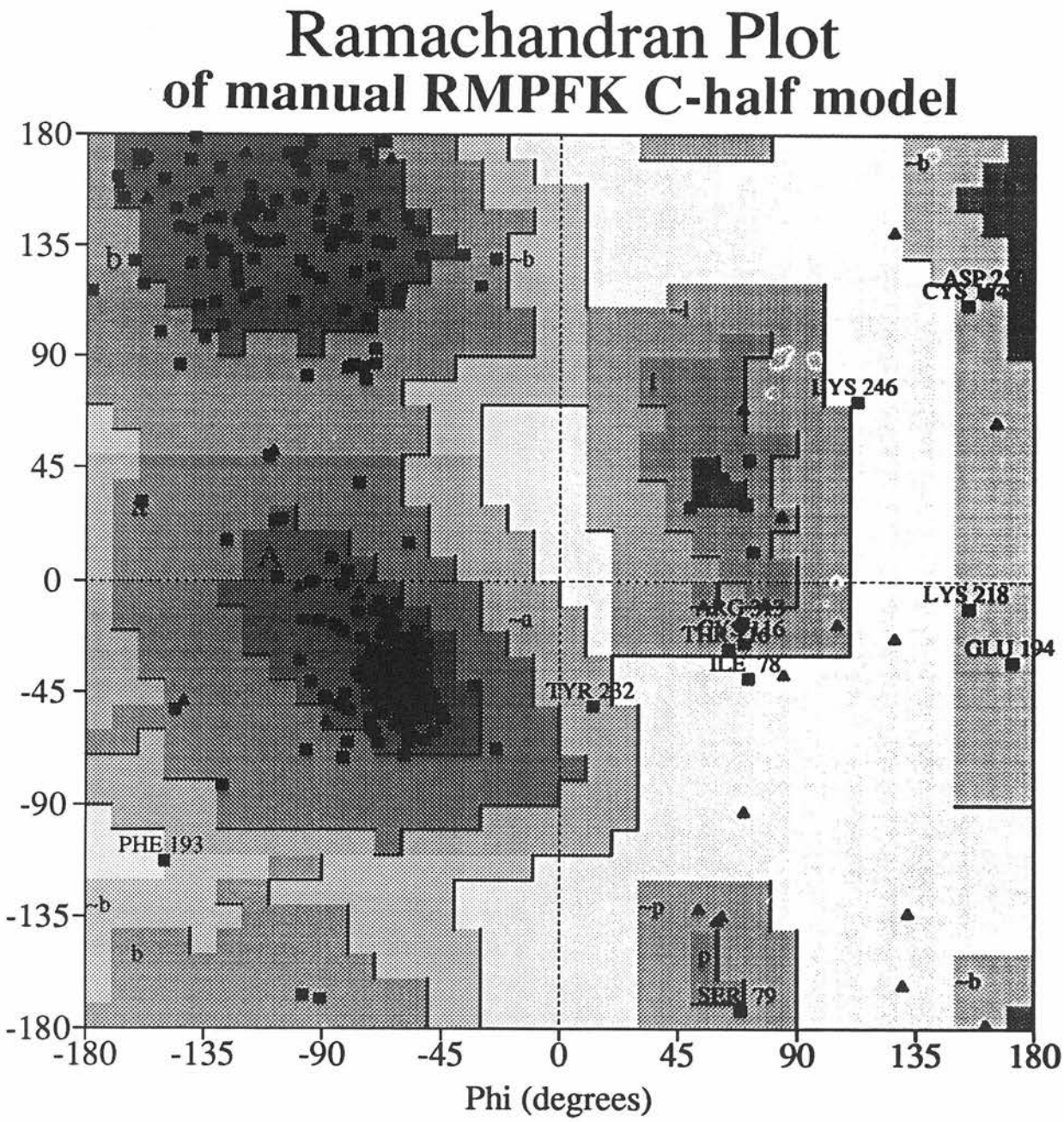

Ramachandran Plot
of manual RMPFK C-half model
180
135
90
45
0
-45
-90
-135
-180
-180
-135
-90
-45
0
45
90
135
180
Phi (degrees)
b
~b
~l
l
~b
LYS 246
LYS 218
GLU 194
ILE 78
TYR 232
PHE 193
~a
~p
p
~b
b

Fig 2.14

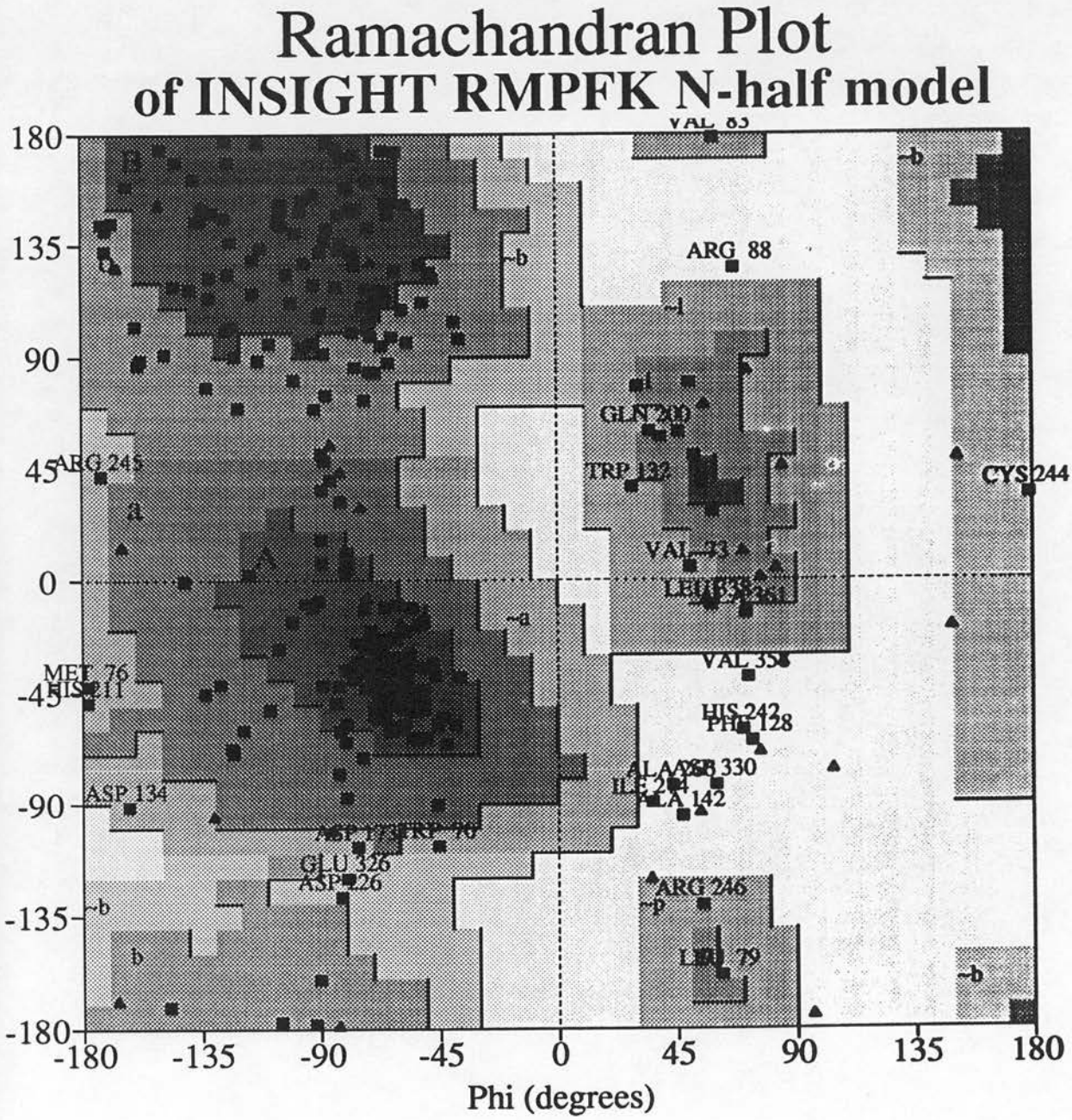

Ramachandran Plot
of INSIGHT RMPFK N-half model
180
135
90
45
0
-45
-90
-135
-180
-180
-135
-90
-45
0
45
90
135
180
Phi (degrees)
ARG 88
GLN 200
ARG 245
TRP 132
CYS 244
VAL 33
VAL 35
HIS 242
PHE 128
ASP 134
ARG 246
GLU 326
ASP 226
MET 76
LEU 79

Fig 2.15

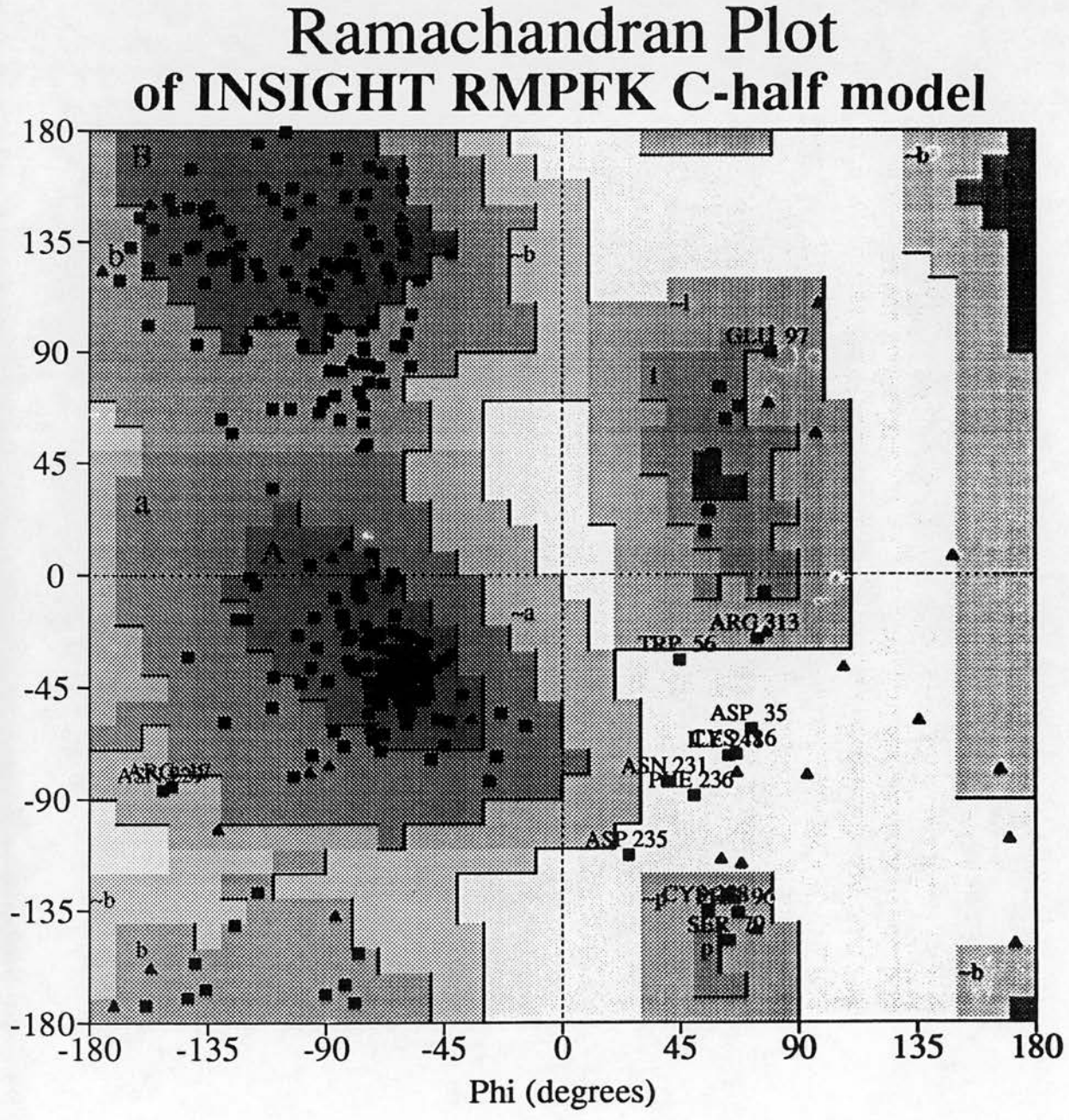

Ramachandran Plot
of INSIGHT RMPFK C-half model
180
135
90
45
0
-45
-90
-135
-180
-180
-135
-90
-45
0
45
90
135
180
Phi (degrees)
B
b
~b
l
GLU 97
i
a
A
~a
ARG 313
TRP 56
ASP 35
ASN 231
PHE 236
ASP 235
SER 70
p

Fig 2.16

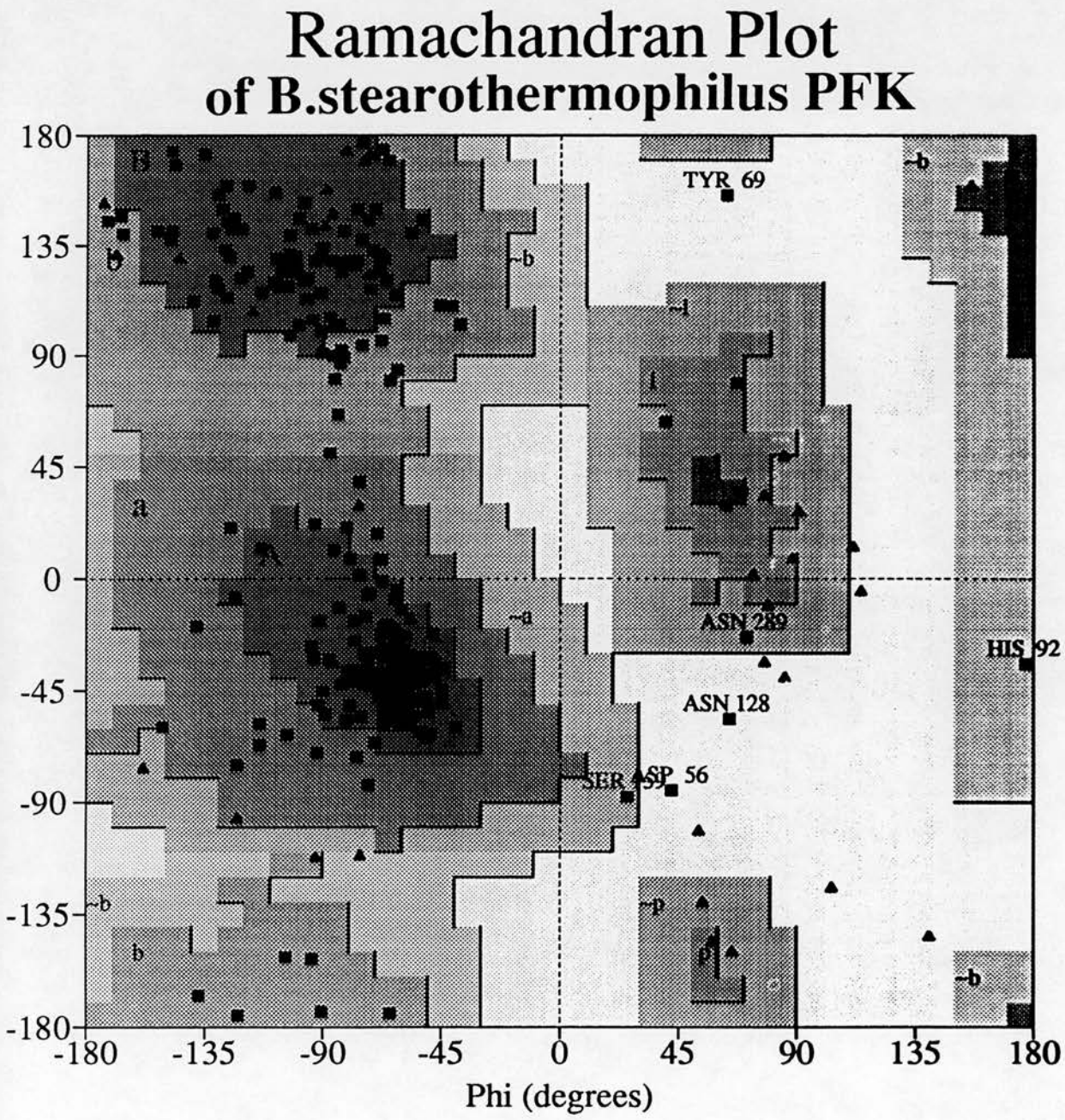

Ramachandran Plot
of B.stearothermophilus PFK
180
135
90
45
0
-45
-90
-135
-180
-180
-135
-90
-45
0
45
90
135
180
Phi (degrees)
TYR 69
ASN 289
HIS 92
ASN 128
SER
ASP 56
a
b
~a
~b
~l
l
~p
p
A
B

Table 2.2 Percentage distribution of residues in models

	Manual Model N	Manual Model C	Insight Model N	Insight Model C	Crystal
Residues in most favoured regions (A,B,L)	81.9	82.9	67.4	69.8	81.1
Residues in allowed regions (a,b,l,p)	14.6	12.6	23.6	24.8	16.7
Residues in generously allowed regions (~a,~b,~l,~p)	1.4	3.7	6.2	2.7	1.1
Residues in disallowed regions	2.1	0.7	2.9	2.7	1.1

Table 2.3 Main chain parameters for model and crystal structures.

	Manual Model N	Manual Model C	Insight Model N	Insight ModelC	Crystal
% residues in A,B,L regions	Better	Better	WORSE	Worse	Better
Omega angle st.dev.	WORSE	Worse	WORSE	WORSE	Worse
Bad contacts/ 100 residues	Better	Better	Better	Better	Better
Zeta angle st.dev.	Better	Better	Better	Better	Better
H-bond energy st.dev.	Better	BETTER	BETTER	Better	Better

Table 2.4 Side chain parameters for model and crystal structures.

	Manual Model N	Manual Model C	Insight Model N	Insight ModelC	Crystal
hi-1 gauche inus st. dev.	Better	Better	Better	Better	BETTER
hi-1 trans . dev.	Better	Better	Better	Better	BETTER
hi-1 gauche us st. dev.	Better	BETTER	BETTER	BETTER	Better
hi-1 pooled . dev.	Better	Better	Better	Better	BETTER
hi-2 trans dev.	BETTER	Better	Better	Better	Better

minimisation and dynamics could be a consequence of the force field used by XPLOR. If this is the case then the poor omega ratings given to the model do not necessarily reflect any intrinsic problems. Although XPLOR worsens the omega angle rating, the deviations produced are small and do not invalidate the results produced by XPLOR.

A similar rating scheme is given by PROCHECK to various side chain parameters. The results obtained for the model and crystal structure are shown in Table 2.4. These values are in the ranges expected for good protein structures and show no overall difference between the models produced manually and those produced by INSIGHT.

Of all these indicators the most useful are ϕ,ψ distribution, χ_1 distribution and H-bond energies. In summary all models showed good values for the latter two with better performance by the manual models in the ϕ,ψ distribution test.

Detailed Examination

The detailed output of PROCHECK was used to conduct a residue-by-residue check on the models to look for especially poor regions. In particular it was of interest to see if they contained, or were adjacent to, unmodelled insertions or deletions or modelled insertions and deletions, perhaps of poor quality. Situation at intra- or intermolecular interfaces could be another explanation for stereochemically poor regions since the influence of adjacent subunit halfs was not considered in detail during the modelling process. The distribution of stereochemically poor residues is summarised in Table 2.5.

Ten non-glycine residues of the manual model of RMPFK N-terminal half fall within the generously allowed or disallowed regions of the Ramachandran plot and are labelled in Fig 2.12. D330, T331, A333 and C334 are near the modelled three residue insertion at D330. C233 and D237 are near the modelled three residue insertion at E232. W132 is near to the unmodelled 23 residue insertion. M75 is at the interface with the C-terminal half of this subunit in the modelled RMPFK dimer. There are no obvious explanations for D91 and T359.

The manual model of RMPFK C-terminal half contains 12 residues in the generously allowed or disallowed areas (Fig 2.13). C174 and Y232 are near the unmodelled four

Table 2.5 Distribution of stereochemically poor residues

	In/near modelled insertions/deletions or unmodelled regions	At interfaces	Others	Total
manual N	7	1	2	10
manual C	5	1	6	12
INSIGHT N	11	5	9	25
INSIGHT C	6	2	5	13
	----	----	----	----
Total	29	9	22	60

residue insertion at Lys227. F193 and E194 are by the modelled three residue insertion at E194. D250 is by the modelled one residue insertion at Lys161. T216 is at the interface with the N-terminal half of the same subunit in the RMPFK dimer model. There are no obvious explanations for I78, S79, C116, K218, K246 and R313.

The INSIGHT model of RMPFK N-terminal half contains 25 poorly placed residues (Fig 2.14). F128 is by the unmodelled 25 residue insertion at G97 while A26 and I274 are near the modelled insertion at E264. H242, C244, R245 and R246 are all near the modelled insertion at D241 while V358 is near the modelled deletion at K356. W132 and D134 are near the unmodelled N-terminal portion while A361 is near the unmodelled connecting peptide. V73, M76 and L338 are at the interface with the C-terminal half of the same monomer whereas L79 and L338 contact the other monomer in the RMPFK dimer model. The remaining W70, V83, R88, A142, D173, H211, D226, E326 and D330 have no obvious explanation.

There are 13 residues in poor ϕ,ψ positions in the INSIGHT model of RMPFK C-terminal half (Fig 2.15). S79 is near the modelled deletion at A80 while C116 is by the unmodelled insertion at K289. F96 and E97 are part of the modelled insertion at G95 while C228 and N229 are part of the modelled insertion at K227. R17 and R313 are at the interface with the N-terminal half of the same subunit in the RMPFK dimer model leaving D35, W56, D235, F236 and I248 with no obvious explanation.

Overall 21 out of 60 poorly placed residues are near modelled insertions or deletions. This casts doubt on the accuracy of these loops, in keeping with the general finding that loops are modelled less accurately in homology-based models than structurally-conserved regions. Specifically, doubt is cast on the accuracy of loops 3 and 5 in the manual N-half model and on loops 4 and 6 in the manual C-half model. Loops 2, 3 and 5 of INSIGHT N-half model all contain or are near poorly placed residues. These too are therefore suspect as are loops 1, 2 and 5 of the INSIGHT C-half model. The INSIGHT N-half model contains much the largest number of poorly placed residues for which there is no obvious explanation, an indication of overall poorer quality. The PROCHECK results suggest that database loop search and random generation methods

are similarly successful. Three of the five loops from database searches contain or are near residues with poor main chain stereochemistry. With randomly generated loops the figure is six suspect out of nine. The one example where a deletion was modelled simply by excision and end-joining, deletion 4 of the INSIGHT N-terminal model, is near stereochemically poor residues and is therefore suspect.

2.3.2.2 PROFILE

PROFILE was run on all the model halfs and the *B.stearothermophilus* structure in the context of dimer models or tetrameric structure respectively.

Overall Scores

The total scores are given in Table 2.6. The predicted scores are derived from the linear relationship seen between protein length and profile score observed for proteins of the Brookhaven Protein Databank (Lüthy *et al.*, 1992). The relationship is not quantitated by the authors but can be estimated from their data as

$$\text{PROFILE Score} = 0.422 \times \text{protein length} + 0.556$$

The crystal structure scores higher than predicted for its length suggesting, like the PROCHECK results, that it is a very good structure. The manual model for the N-terminal half of RMPFK comes impressively close to a typical score for a protein this size. The N-terminal INSIGHT model in contrast gives a poor overall score suggesting major errors in this model. Conversely, the INSIGHT model of the C-terminal half, which is slightly shorter, comes closer to its expected score than the manual model.

Detailed Analysis

The profiles of the models and crystal structure generated with a window length of 21 are shown in Figs 2.17 - 2.19. The residue id does not necessarily correspond exactly with the numbering of the models due to the incompleteness of the models. The

Table 2.6 Total Profile scores for RMPFK half models and crystal structures

Structure	Score	Length	Predicted score
B.stearothermophilus crystal structure	155	320	136
Manual model of N-terminal half	122	329	138
Manual model of C-terminal half	106	318	135
INSIGHT model of N-terminal half	85	320	136
INSIGHT model of C-terminal half	111	309	131

Fig 2.17 Comparison of N–half model profiles

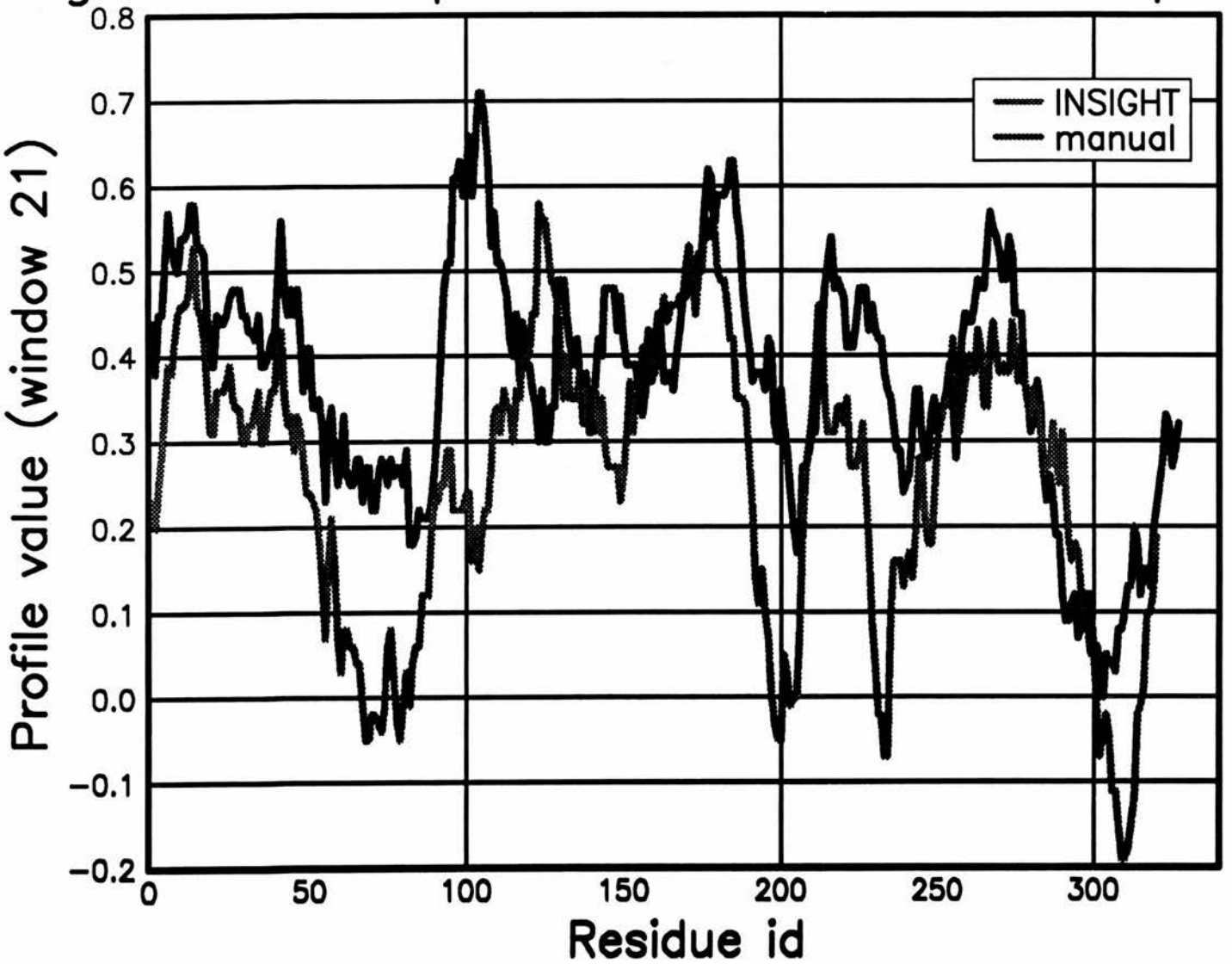


Fig 2.18 Comparison of C–half model profiles

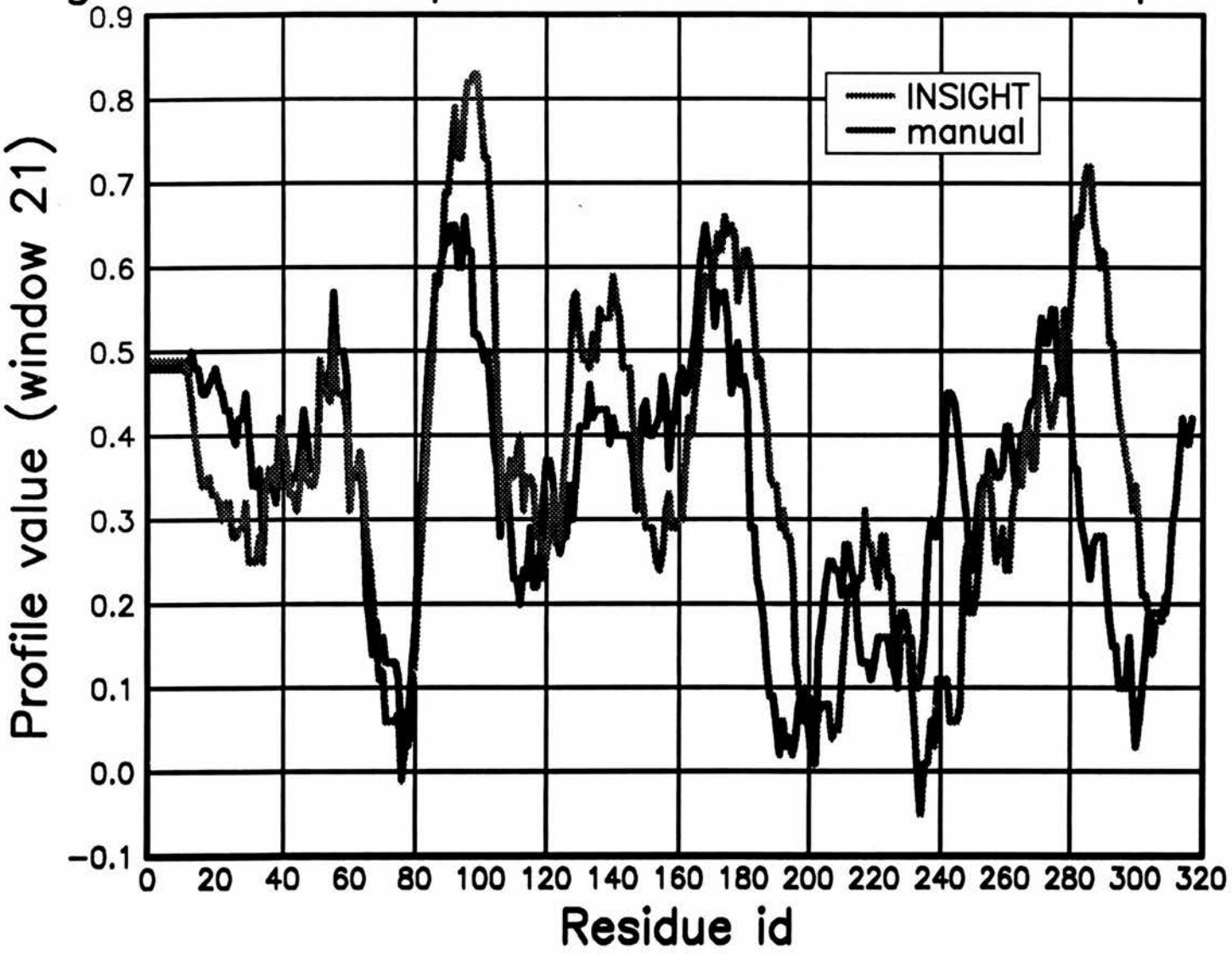
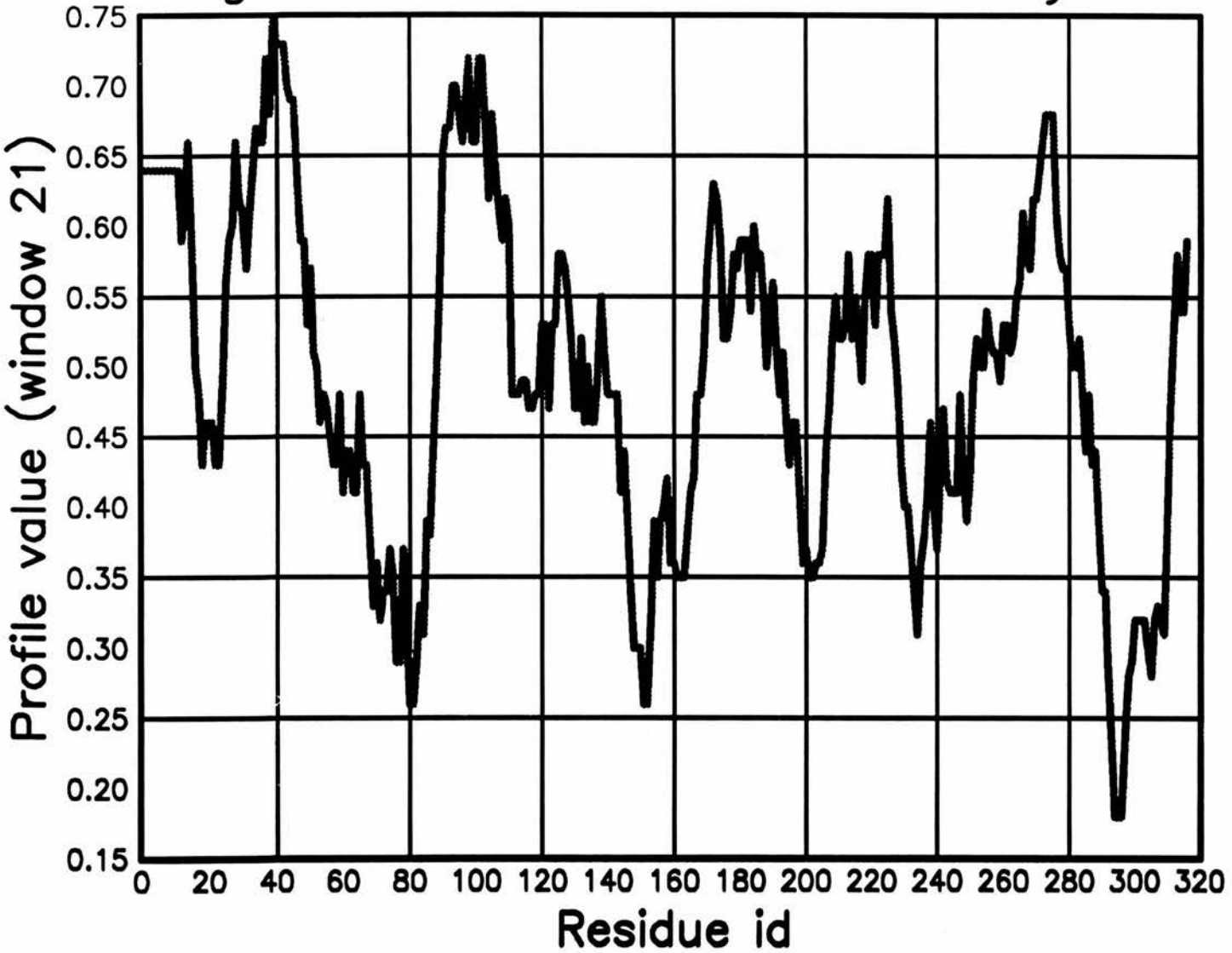


Fig 2.19 Profile of B.stearo crystal



different models of the N- and C-terminal portions do not correspond exactly in residue ids due to differences in alignment. Regions scoring an average of less than 0.2 are rare in good protein sructures. Such regions in the models were analysed to try to find reasons for their probable poor structure.

The manual model of the N-terminal half contains only three such regions, two of which are very small - E96-R98 and D241-L243. These seem to arise from partial burial of an Arg and an Asp respectively. There is a dip in the crystal profile to 0.25 at the corresponding position to the first of these suggesting a probable source of the low score of the model here. Neither of these regions is close to an insertion or deletion, modelled or unmodelled, or to an interface. The third low scoring region extends from P332 to D363. This includes part of the modelled insertion at D330. There is very little similarity to the bacterial enzymes from R345 onwards. This is the C-terminal portion which precedes the connecting peptide which is not present in the bacterial enzymes. The profile suggests that this part of the chain adopts a different conformation in the RM enzyme and is therefore modelled incorrectly using the bacterial structure.

There are four regions scoring below 0.2 in the manual C-half model. S66-A80 includes the modelled deletion at L70. The rest of this region shows little similarity to bacterial enzymes suggesting that the alignment may be incorrect or that this region adopts a very different structure in the RM enzyme. There is a similar situation with F193-Q211. This contains a three residue modelled insertion in a region of low similarity to the bacterial sequences. The third poor region, V222-K246 contains an unmodelled insertion again in a region of little similarity. The final poor region is Q318-P335. There are two possible reasons for its poor score. First the large unmodelled insertion at E290 is adjacent. Secondly this region again shows little similarity to the bacterial enzymes and precedes the unmodelled C-terminal tail.

The first of five poor regions in the INSIGHT N-terminal model is very large, covering E67-G97 and after a large unmodelled insertion T127-E131. The explanation for this lies almost certainly in the alignment produced by INSIGHT. The program

produced an alignment in this region that differed from the manual model alignment, failing, for example to align the sequence IGGDGS with the identical sequence in the bacterial enzymes. The second poor region I145-A147 is also by this large unmodelled insertion. The third region includes the modelled insertion at D241 in a region of weak similarity and is also adjacent to the probable position of the unmodelled C-terminal tail. The fourth region, G271-R285, starts after the modelled insertion at E264 and continues into a region of low similarity. The final region extends from V335 to the end of the model at D363. As stated above sequence similarity is very poor here so that the alignment, which differs from the manual one, may well be incorrect.

The first poor region of the INSIGHT C-terminal half model is R68-A80. This is by the modelled deletion at A80 and once again shows little sequence similarity. P196-L221 includes a modelled insertion at A204 in a region of little sequence similarity. R224-K246 includes the modelled insertion at K227 in a weakly similar region. The score dips below 0.2 finally at F302 and remains there until the final modelled residue Q324. This region of poor homology precedes the unmodelled C-terminal tail.

The most obvious conclusion from this analysis is that low scoring regions are nearly always found in stretches of chain showing weak similarity with the bacterial enzymes used to generate the alignments and build the models. Most probably the alignments used are incorrect or the RM enzyme shows a very different conformation in these regions. If this were true then attempts to model insertions or deletions in these regions would be doomed to failure. The loops indicated as having possible poor stereochemistry by PROCHECK were compared with loops in low-scoring PROFILE regions and a remarkable correlation found. Of the eleven loops containing or near stereochemically poor residues, nine are in the above low scoring regions of low sequence similarity. Of the seven loops built into high scoring, highly similar regions, only two were indicated as stereochemically doubtful by PROCHECK. Of the better loops, those generated by database loop search and random generation are equally represented suggesting, within this small sample size, that both methods can be used successfully.

2.3.2.3 Chemical Modification and Proteolysis

Evidence for assignment of classes of effectors (sugar bisphosphates and AMP/ADP activating, ATP and citrate inhibiting) to putative effector sites C, A', B' and C' has been obtained by chemical modification and proteolysis.

Sheep heart PFK has been reacted with the affinity labelling adenosine analog *p*-fluorosulphonylbenzoyl-5'-adenosine. This modification occurred at a site specific for allosteric activators ADP, AMP and cAMP (Mansour *et al.*, 1978) and was shown to be due to modification at Lys680 (Weng *et al.*, 1980). Another Lys nearby was shown to be highly reactive to pyridoxal phosphate suggesting solvent exposure and proximity to multiple positively-charged residues, as would be expected of a phosphocompound binding site (Kemp *et al.*, 1987). Both of the models of RMPFK place both these lysines (numbers 275 and 279 in my RMC numbering) at the C site, consistent with the identification of this site with AMP/ADP activation by Poorman *et al.* (1984).

Citrate has been shown to protect Lys529 against otherwise rapid modification by pyridoxal phosphate (Kemp *et al.*, 1987). This residue is at the C' site in both manual and INSIGHT RMPFK models contradicting the suggestion of Poorman *et al.* (1984) that this site is the ATP inhibitory site.

Exposure of RMPFK to V8 protease results in the appearance of a form lacking 17 C-terminal residues which shows reduced inhibition by ATP and citrate (Valaitis *et al.*, 1987). Equilibrium binding studies showed lower affinity of the proteolysed form for both. Attention was focussed on ATP however since the synergism of ATP and citrate binding suggests that lower citrate binding was only a consequence of lower binding of ATP which was also present. Removal of the C-terminal seven or eight residues by trypsin did not affect ATP binding (Gottschalk *et al.*, 1983) highlighting the importance of the sequence His-Ala-His-Leu-Glu-His-Ile-Ser-Arg. The histidines had already been implicated in ATP binding in diethylpyrocarbonate modification studies of sheep heart PFK (Setlow and Mansour, 1970). These residues are in the C-terminal tail

which is not present in the bacterial enzyme and is therefore not built in the models. The INSIGHT model stops well short of this sequence. The manual model stops 19 residues short but the previous region shows only low homology with the bacterial enzyme and the model in this region is of doubtful quality (see above). The end of the manual model is certainly close enough to the C' site for the His residues to be contributing to ATP binding there. However the uncertainties are too large for this to be considered strong evidence for identification of the C' site with ATP inhibition.

Several predictions about the environments of methionines and cysteines in RMPFK have been made from consideration of their rates of reaction with iodoacetate (Latshaw *et al.*, 1987).

Cys88 reacted most rapidly suggesting a relationship with a neighbouring nucleophile that lowers its pKa. In the manual model there is a good candidate in Asp91 whose side chain is within H-bonding distance to the thiol. In the INSIGHT model Asp91 is not adjacent and there is no likely nucleophile, indeed the closest residue is Arg95. Cysteines 169 and 232 were found to have high reactivity which was lowered in the presence of F6P. Cys169 is close to modelled substrate FBP in both models - 12Å measured from S to P_1. Cys 232 is further away (21-22Å) consistent with the smaller effect on activity seen on modification compared with Cys169.

Unusually, two methionines in the N-terminal half were also highly reactive to iodoacetate suggesting location in highly positively-charged regions of the enzyme. Both models show concentrations of basic residues around these methionines. The distances are shown in Table 2.7 measured S-Cζ(Arg), S-Nζ(Lys) and S-Cδ(Asp). There is particular uncertainty over distances to residues in the C-terminal half of the model since the relative orientation of N- and C-terminal halfs is not accurately known. Due to differences in alignment Met 74 is in different positions in the two models of the N-terminal half. In the INSIGHT model Met74 is buried which is inconsistent with its high reactivity to iodoacetate.

Modification of Met 74 was found to have no effect on ATP inhibition. This is despite the fact that the manual model and the model of Poorman *et al.* (1984) both place Met74

Table 2.7 Distances of Met74 and Met174 from charged residues

Manual Model

Residue	N or C-terminal half	Distance from Met 74/Å
Lys154	C	7
Arg219	C	8
Lys161	C	14
Arg163	C	16
Lys218	C	16

Residue	N or C-terminal half	Distance from Met 173/Å
Asp 173	N	4
Lys356	N	4
Arg313	C	6
Lys275	C	12
Lys312	C	14

INSIGHT model

Residue	N or C-terminal half	Distance from Met 74/Å
Arg310	N	4
Lys154	C	5
Arg152	C	9
Arg219	C	11

Residue	N or C-terminal half	Distance from Met 173/Å
Asp 173	N	4
Lys356	N	10
Arg313	C	5
Lys275	C	9
Lys312	C	12

at the C' site which Poorman *et al.* allocate to ATP inhibition. However the manual model does not rule out this assignment since the side chain is solvent exposed and therefore presumably quite mobile. Certainly the manual model does not suggest that the Met side chain would contact a putative bound ATP.

2.3.2.4 Phosphorylation

Two serines were identified as phosphorylation sites when RMPFK was incubated with protein kinase A in the presence of calmodulin or troponin C (Zhao *et al.*, 1991). These were Ser774 near the C-terminus and Ser376 in the connecting peptide. Since neither of these regions were present in the RMPFK models suitable exposure of these residues cannot be used as a criterion by which to judge the models.

2.3.2.5 Binding Sites

The putative binding sites of the models were then examined in more detail to see if they were consistent with binding any of the known RMPFK ligands. In particular consistency with the hypotheses of Poorman *et al.* (see Introduction - Section 2.1.4) was tested. Only the manual model of the N-terminal half was studied since the INSIGHT model is clearly of poor quality (see above). Both models of the C-terminal half were studied.

The A site

The alternate subunit of bacterial PFK contributes two arginines to the phosphate-6 binding part of this site. These two, Arg 163 and Arg252 in both alignments, are modelled much better in the manual rather than the INSIGHT model of the C-terminal half. In the manual model Arg163 is next to the modelled insertion 3 which comprises Arg162 (see Fig 2.7). It is Arg162 in this model which salt bridges phosphate-6 rather than Arg163, while Arg252 too is in good position to bind phosphate-6. In the INSIGHT model, by contrast, Arg163 can salt bridge phosphate-6 but Arg252 is too far away. In addition, the modelled insertion 3, comprising Gly159, places Thr160 by

phosphate-6. Clearly, the manual model is superior to the INSIGHT model in this region.

The A' site

Inspection of the model of the N-terminal half is fully consistent with the idea that this site is the sugar bisphosphate activation site. The catalytic Asp127 of the bacterial enzyme which is responsible for repulsion of the product F16BP is replaced by a Ser. The two arginines contributed by the N-terminal half are modelled well in the manual model and no modelled or unmodelled insertions are near this site.

The B site

The manual model of the N-terminal region which wholly contains this site is very similar to the B site of the bacterial enzyme, as would be expected from the high degree of sequence similarity here. No modelled or unmodelled insertions or deletions are near here. Thus, this site is the ATP substrate binding site.

The B' site

This site is wholly contained within the C-terminal half. In both manual and INSIGHT models the alignment places a seven residue insertion after Gly95. This is modelled by random loop generation in the INSIGHT model but is unmodelled in the manual model. The loop in the INSIGHT model completely blocks the entrance to the pocket ruling out binding of ligand. Figure 2.20 shows the ADP-bound B site of *E.coli* PFK. The blocking loop at the B' site in the INSIGHT model of RMPFK C-terminal half is shown in red with the dots indicating the van der Waal's radii of the ADP atoms. There is some doubt in the alignment so that the insertion could be positioned up to four residues later. However the models show that an insertion of this size anywhere in this region would have to be accommodated in the pocket thus ruling out ligand binding here.

Fig 2.20 An insertion (red) blocks the RMPFK B' site

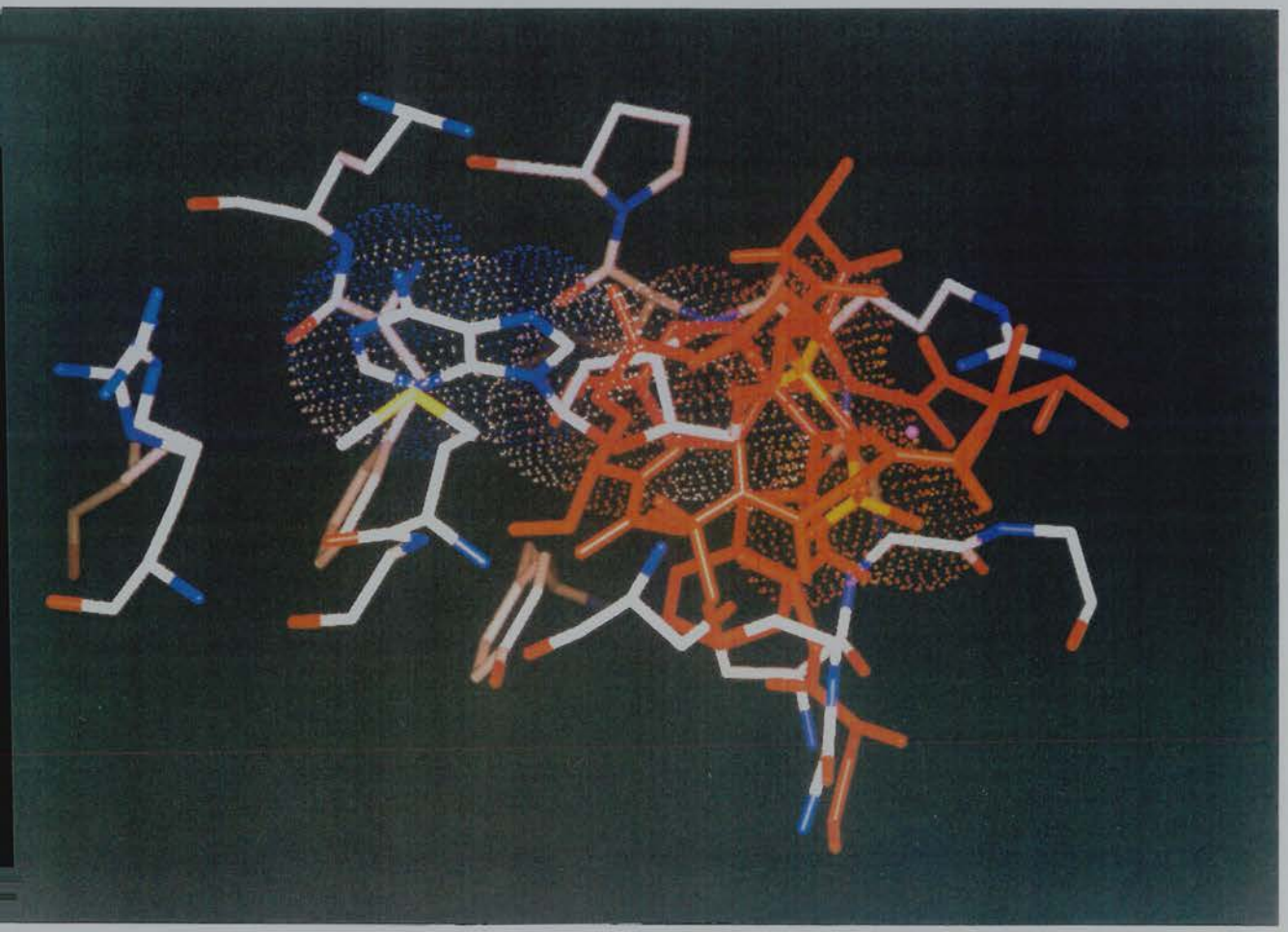
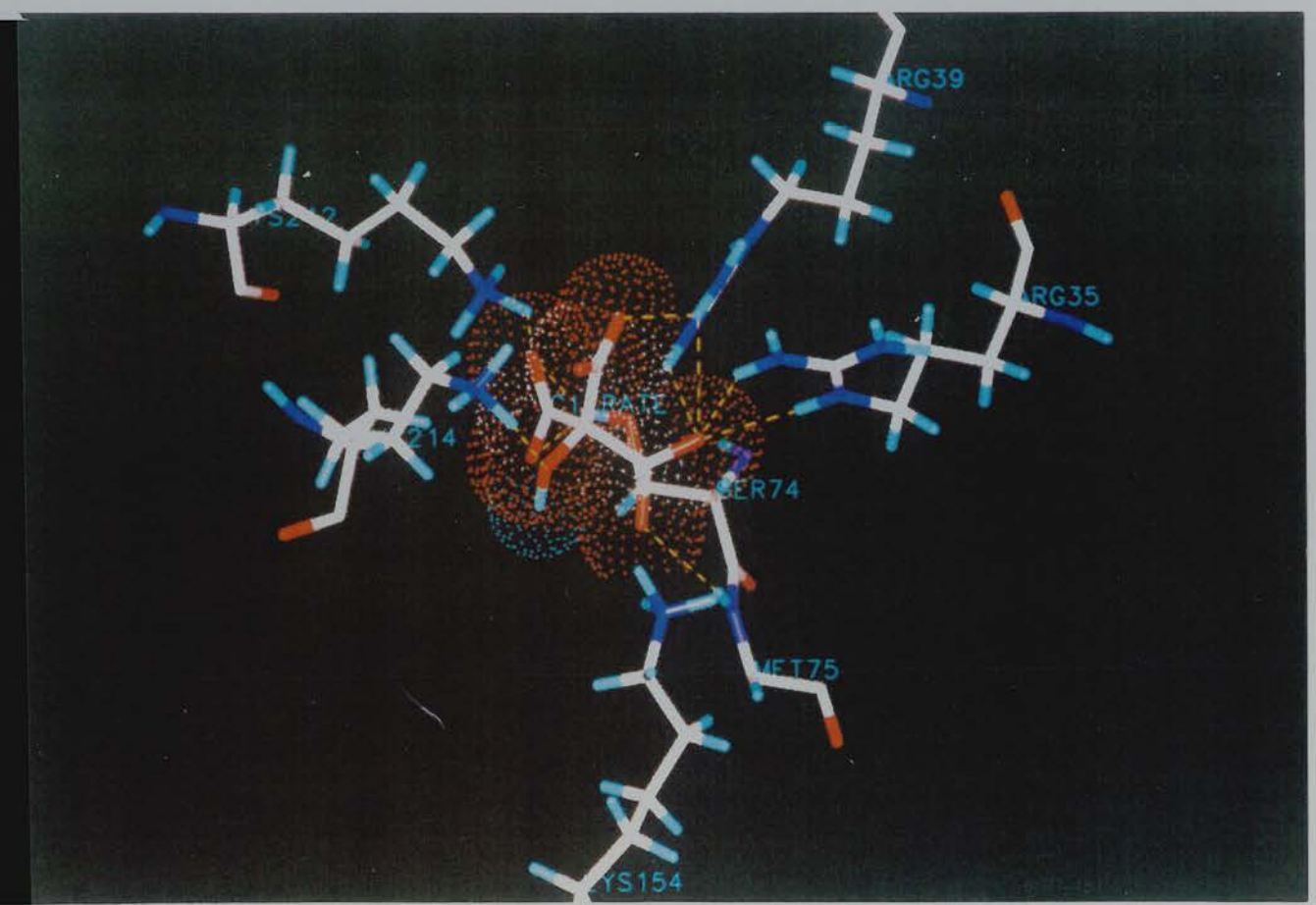


Fig 2.21 Citrate bound to a T-state RMPFK C' site model



The C site

The residues defining this site in the bacterial enzyme are well conserved in the rabbit enzyme and the models are fully consistent with ADP and AMP binding here. There remains doubt however about the possible role of the connecting peptide which cannot be modelled. In the *E.coli* enzyme the C-terminal residue Tyr319 contributes to this site and it is possible that residues of the connecting peptide are close to this site in RMPFK.

The C' site

Again the residues comprising this site are similar to those of the C site of the bacterial enzyme. As discussed above (section 2.3.2.3) there is biochemical evidence linking both ATP inhibition and citrate inhibition to this site.

The gamma phospho group of ATP putatively bound here is positioned where residues of the unmodelled C-terminal extension may well lie. This is consistent with evidence linking histidines of the C-terminal extension to ATP inhibition (Valaitis *et al.*, 1987).

The alternative idea, that the PEP inhibition site of bacterial PFK, has become the citrate inhibition site of RMPFK (which also binds PEP, though with lower affinity - Colombo *et al.*, 1975), seems to be less favoured despite the evidence of chemical modification work (Kemp *et al.*, 1987). This possibility was investigated by construction of a model of the C' site using the structure of T-state, phosphoglycollate bound *B.stearothermophilus* PFK as a basis (see Section 2.2.1.8). Figure 2.21 shows citrate positioned in this model with H-bonds appearing as dashed lines. This appears to be a favourable conformation with all three citrate carboxylates salt bridged to at least one basic residue and the citrate hydroxyl H-bonded to Lys214. The model shows the lysine implicated in citrate binding by chemical modification studies, Lys 529, salt bridged to the citrate. A large uncertainty in this model is the position of Trp70 which was mutated from the *B.stearothermophilus* Val54. As shown, it is unfavourably near to one of the citrate carboxylates. Alternative positions could allow Arg25 to adopt a

conformation in which it salt bridges additional carboxylates.

2.4 Conclusions

The RMPFK models built by homology well illustrate the strengths and weaknesses, both of the general approach, and of the manual and automatic methods.

There are good reasons to think that the overall quality of the manual models is good. Their overall profile scores, particularly the N-terminal model, approach closely the mean scores for proteins of their size in the protain database and are certainly within the spread of values seen for crystallographically-determined structures. Both models exceed 80% of residues on core regions of the Ramachandran plot (Table 2.2) which is similar to the crystal structure subject to the molecular dynamics regime. For good structures determined to 2.0Å 90% of residues would be expected to be in these regions so that this performance for models built on a 2.4Å resolution structure is creditable. In addition, all but one of the diagnostic main chain and side chain parameters of PROCHECK indicate the models to be better than is averagely found for crystal structures. The sole exception is omega angle standard deviation which poor results appear to be a consequence of molecular mechanics using XPLOR's force field.

The INSIGHT models seem less good. The PROFILE score of the C-terminal model is good but that of the N-terminal model is poor. Both models perform more poorly in putting only 65-70% of their residues in core regions of the Ramachandran plot, but again all other PROCHECK parameters are better than average. One factor contributing to the poor Ramachandran plots could be the reliance on random generation for loop modelling in contrast to the emphasis on database loop searches in the manual models. The ϕ and ψ backbone torsion angles generated by the random generation algorithm are not biased towards those typically seen in proteins whereas by using a segment from another protein reasonable protein torsion angles are likely to be obtained. However, with at most 10% of total residues of INSIGHT models modelled by random loop

generation, this is unlikely to be the whole explanation. It is likely that the lack of supervision during side chain substitution of the INSIGHT models lead to some serious clashes with main chain being present in the original model. During minimisation and dynamics these would have caused distortion of the main chain. Where the manual and INSIGHT models differ, the manual models are more consistent with the biochemical evidence.

The decision to produce the INSIGHT models as automatically as possible was made in order to make the comparison between this very rapid procedure (1-2 days) and the manual approach which took weeks. A large factor in the relative slowness of the manual modelling was the complicated loop modelling procedure which aimed to improve on the results obtainable from standard database loop modelling protocols. This more involved approach typically took days for each insertion. For the INSIGHT models, production of ten randomly generated loops typically required less than one hour and a reasonable loop was usually present in the group. Considering the disparity in time requirement, the superiority of the sophisticated loop search over random generation seems quite small although the number of loops modelled is too small to draw firm conclusions. Of the three loops modelled in regions of good homology by each method, none modelled by loop search was inaccurate as judged by PROCHECK whereas one of the randomly generated loops contained poorly positioned residues.

The large misalignment present in the automatic INSIGHT structure alignment is certainly the reason for many of the poor performance indicators of the INSIGHT N-half model. The alignment produced for the C-half was so poor as to make its use futile and a slightly modified manual alignment was substituted. The difference in quality of the two INSIGHT models emphasises the importance of good alignment.

Overall, comparison of the manual and INSIGHT methods suggests that a middle way might be the best future approach. Alignment is probably best handled manually unless a large number of homologous structures are available. Mutation of side chains can be left to the computer, although a check for bad contacts, and their rational removal seems to be valuable. Random generation seems to model loops well given its modest time

demands although lack of bias towards typical torsion angles is a weakness. The convoluted loop search algorithm used in the manual models seems not to be worth the extra time outlay. A recent review of secondary structure prediction, subtitled 'Man versus machine', similarly concluded that both manual and automatic methods have their advantages and to combine both seems the best policy (Benner and Gerloff, 1993).

The models shed some light on the possible binding sites of the extra RMPFK effectors, but are frustratingly incomplete in some of the most interesting regions. There seems to be no doubt about the assignment of sites A and B to substrate F6P and ATP, nor of sites A' and C' to sugar bisphosphate and adenine nucleotide activating sites. Site B' seems to be blocked by an insertion in the mouth of the pocket. That leaves ATP and citrate inhibitory sites to be accounted for by site C' and/or novel sites. Biochemical evidence comes down equally on the sides of ATP or citrate binding to site C'. Given the uncertainty at the C' site due to the probable proximity of the unmodelled C-terminal extension binding of ATP seems possible. The C' site model derived from the T-state bacterial PFK, on the other hand, suggests that binding of citrate here is a possibility. The possibility that inhibitory ATP and citrate share a binding site seems to be ruled out by equilibrium binding measurements (Colombo *et al.*, 1975). ATP is required before significant citrate binding is observed but in the presence of ATP up to 1mol citrate/mol protomer can be bound. The ATP seems to be binding to the ATP inhibitory site rather than to the ATP substrate site since ITP cannot substitute. The two possibilities are therefore that site C (ADP activating; PEP inhibiting) has become site C' (ATP inhibiting) or site C' (citrate inhibiting). The former possibility would require the incorrect modelling of Lys529 despite its being in a region of relatively good homology. The latter possibility would imply the separate evolution of another nucleotide binding site which may be unlikely. At present, therefore, the question remains open.

Chapter 3

Ligand Modelling

3.1 Introduction

Rational drug design is the holy grail of medicinal chemistry. Despite recent advances, traditional random sceening of compounds remains slow and expensive. Advocates have long held that rational drug design holds the key to more efficient drug discovery. Past scepticism has been justified to some extent in view of the relatively small number of rationally-designed drugs which have passed clinical trials to reach prescription. However, the methodological groundwork has now been laid and an acceleration in success rate may now be expected as computers become powerful and cheap enough to run computationally-intensive programs. The potential benefits, both humanitarian, and, for drug companies financial, are enormous.

It is much easier to design a ligand than a drug. Essential properties of a drug include selectivity, chemical stability, ease of synthesis and ready formulation. The pharmacokinetic profile is also important. It may be desirable that the drug is persistent or, conversely, that it is rapidly broken down or excreted to avoid toxic side-effects. It is because of these potential problems that the rational ligand designer should aim to produce many and varied ideas.

The methods available for rational i.e. knowledge-based, ligand design depend very sensitively on the amount and quality of the knowledge accumulated. Rational design can be attempted at levels ranging from an identified substrate to protein-ligand complex X-ray crystal structures. However, the more knowledge is available, the more sophisticated are the available approaches and the greater the chance of success. For a review of available methods see Martin (1991).

Of the many approaches developed most are divisible into two stages. In the first, potential ligand interaction sites are identified for the target region, usually the catalytic

site of an enzyme or the agonist binding site of a receptor. These can be located directly from knowledge of protein structure or inferred from information relating to known ligands. In the second stage known ligands are modified to occupy more interaction sites or new ligands sought to occupy several interaction sites.

Many of the most successful examples of ligand design employ an iterative scheme (Appelt *et al.*, 1991). Ligands designed as above are tested *in vivo* and *in vitro* and the most interesting studied crystallographically in complex with the target. The information so gained is then used to guide further design work aimed at improving affinity.

I shall first discuss methods of analysing the system of interest to obtain information useful to the drug design process. Then I shall discuss the exploitation of this knowledge in attempting to discover new lead compounds.

3.1.1 Finding potential ligand interaction sites

3.1.1.1 From protein structure

GRID - the program

The program GRID (Goodford, 1985; Boobbyer *et al.*, 1989) computes an energy of interaction at points on a three-dimensional grid between a probe and a target. The probe consists of a single atom e.g. carbonyl oxygen atom - O(=C), a small group of atoms e.g. protonated amino group $^{+}NH_3$ or a small molecule e.g. water. Most multi-atom probes are represented as 'extended' atoms. By replacing several atoms with a single 'extended' atom approximating their properties computing demands are much reduced. The target represents the portion of the macromolecule of interest.

The original formulation of the energy function was

$$E = E_{lj} + E_{el} + E_{hb}$$

where E_{lj} is the Lennard-Jones potential, E_{el} is the electrostatic potential and E_{hb} the hydrogen-bond term. For the Lennard-Jones interaction a standard (12,6) potential was used. The electrostatic potential calculation took into account the different dielectric behaviour of the protein interior and solvent compartments. An empirical modification to the standard formula for electrostatic interaction was made using the number of protein atoms within a 4Å radius of the probe. The original hydrogen-bond function has been replaced by a much more sophisticated one (Boobbyer *et al.*, 1989). This considers

$$E_{hb} = E_r \times E_t \times E_p$$

where E_r, E_t and E_p are factors depending respectively on the distance apart of the atoms, the angle made by the H-bond at the target and the angle made by the H-bond at the probe. The angular functions vary for the different atom types. Other important advances were the consideration of possible H-bond partner rotation and the assessment of both possible His tautomers during calculations, with adjustment to the electrostatic term depending on likely tautomeric form. The energy calculated for the grid positions is easily converted into contour maps to be displayed using standard molecular modelling software. A range of different probe-target interaction energy levels can be shown simultaneously as contours of different colours. As figures later in this chapter show, the contours often reveal specific regions of particularly favourable interaction.

The most significant limitation of GRID is its inability to account for changes in both target and ligand that may occur on binding. Protein conformational changes on ligand binding are quite common e.g. the change in hexokinase structure on substrate binding. Changes in electronic distribution of ligands on binding are also possible.

GRID testing

Both papers describing the program contain examples of GRID use. For example GRID accurately located water molecules in the crystal structure of phospholipase A2 (Goodford, 1985). In a protein of pharmacological interest - cytochrome P450-CAM, GRID was able to locate the observed binding position of the hydrophobic substrate within a -2.0 kcal mol^{-1} methyl group probe contour. The single H-bond between camphor and the enzyme was also clearly located.

The program has since been widely used. For example GRID was used to suggest potential glucose analogs that may inhibit glycogen phosphorylase as a treatment for diabetes (Martin *et al.*, 1991). In this case GRID predictions were not confirmed by kinetic and crystallographic experiments due to unexpected changes in protein conformation on binding. GRID has also been used to locate likely positions for substrates on an unligated phospholipase C crystal structure (Byberg *et al.*, 1992). From these a probable mechanism was deduced which can now be tested.

Most excitingly, GRID has recently been instrumental in progress towards influenza sialidase inhibitors (von Itzstein *et al.*, 1993). GRID contours suggested that replacement of a hydroxyl group of a known inhibitor by an amino group would improve binding affinity. Indeed, testing showed improved inhibition both of sialidase *in vitro* and virus replication *in vivo*. In addition the new compounds showed improved specificity for the influenza enzyme over mammalian ones.

Multiple Copy Simultaneous Search

A more recent appoach to finding favourable binding sites for molecular fragments is the multiple copy simultaneous search method (Miranker and Karplus, 1991).

Several thousand copies of the fragment are randomly placed in the region of interest and then energy minimised simultaneously. During minimisation the protein is fixed and interactions between fragments are ignored. Clusters of fragments in similar positions are identified, less favoured clusters discarded and representative examples of the remaining clusters studied by molecular graphics.

Despite the efficiency of simultaneous multiple minimisation this approach is more computer-intensive than GRID. However there are significant potential advantages to this method. First, larger fragments are used such as dimethyl ether and acetonitrile. Secondly the method should be readily extended to take account of protein flexibility. This would be by allowing the protein to move during the multiple minimisation although at the cost of increasing computational demands.

Rule-based methods

Many analyses of H-bond geometry in molecules of crystallographic small molecule databases have been published. Several groups have used this work to identify possible interaction sites.

The program HSITE (Danzinger and Dean 1989a) infers potential ligand H-bond sites from consideration of potential H-bond partners on a protein. Allowance is made for rotation of groups on the protein so that each potential H-bond partner of the protein produces a locus of potential interaction points.

The program LUDI (Böhm, 1992a,b) takes no account of rotation of potential H-bond partners. However this program also locates potential hydrophobic contacts, distinguishing between aliphatic and aromatic contacts.

3.1.1.2 From Ligand Structure

One of the earliest developments in the rational drug design field was Quantitative Structure Activity Relationships (QSAR) (Fujita, 1990). In QSAR the measured potencies of a range of ligands are correlated with the values of physical properties such as pKa, size and hydrophobicity. QSAR properties can be calculated for different positions of the analogs enabling ligand features to be ranked by importance, hence guiding further design. Where the 3D structure of the receptor has been determined, good agreement has been found with the important binding features of the ligands determined by QSAR.

QSAR methods may be extended to three dimensions (Loew and Burt, 1990). Here

the first step is the superposition of molecules to maximise similarity. Ideally the bioactive conformations are experimentally determined, for example by nuclear magnetic resonance spectroscopy. Reasonably low energy conformations can be generated by programs such as CONCORD (Rusinko *et al.*, 1988) although ligands may not always bind in such conformations. Where a molecule has relatively few rotatable bonds then a systematic search for the lowest energy conformation may be carried out. Once the 3D structures have been obtained, from whatever method, the molecules are superimposed. Steric, electrostatic and H-bonding interaction energies are calculated for the overlaid molecules at a range of grid points round the molecules. The energies then replace pKa, hydrophobicity etc. in the QSAR positional analysis. Contour plots may be generated indicating the contributions of different parts of the molecule to binding affinity. These are equivalent to, and can be used in the same way as, ligand binding sites inferred from protein structure by methods described above.

Programs are becoming available for the automatic deduction of structure-activity relationships from experimental data (e.g. Catalyst from BioCAD Corp., 1390 Shorebird Way, Mountain View CA 994043). Such programs can also search 2D and 3D databases for novel molecules that share features with the known active compounds.

3.1.2 Fitting ligands to potential ligand interaction sites

3.1.2.1 Modification of existing ligands

This approach is particularly applicable to systems where a receptor-ligand complex structure is available. Modifications are made to the ligand to occupy favourable interaction sites.

Recent work on purine nucleoside phosphorylase (Montgomery *et al.*, 1993) and thymidylate synthase (Appelt *et al.*, 1991; Varney *et al.*, 1992) has illustrated the power of an iterative approach. From an original receptor-ligand complex structure rational

modifications are proposed and analysed. Promising compounds are then synthesised and tested with the most interesting co-crystallised with receptor. The new complex structure is then used to suggest further ideas, and so on. In the best case only three iterations were required for the transformation of μM-range inhibitors into nM-range inhibitors.

Selective inhibition of a trypanosomal GAPDH has recently been demonstrated using a rationally-modified ligand analog. An added hydrophobic substituent occupied a hydrophobic pocket present in the trypanosomal enzyme but not in the human enzyme (Verlinde *et al.*, 1993).

Another important type of ligand modification is well illustrated by the recent structure-based design of a cyclosporin A analog (Alberg and Schreiber, 1993). By introducing rings, well away from the binding part of the molecule they were able to alter the conformational properties of the molecule, as shown by n.m.r. spectrosopy. They were thus able to ascribe the threefold greater binding affinity of the analog to more favourable entropic properties. This illustrates the importance of entropy in drug design. Unfortunately, entropy is still hard to take account of computationally. This is because entropy relates to degrees of freedom in a system. Assessment of degrees of freedom would require study of large numbers of conformations which would be a highly computationally intensive process.

3.1.2.2 Discovery of novel ligands

One of the results of random screening for inhibitors has been the discovery of inhibitors bearing no chemical resemblance to previous ligands. This has demonstrated the feasibility of rationally designing new classes of inhibitors. Such classes may have important benefits such as ease of synthesis, desirable pharmacokinetics, potential for development etc.

Previous dependence on human ingenuity and skill in suggesting potential new inhibitors is now being gradually superseded by computer database searches.

Novel ligands by manual modelling

Ground-breaking work was carried out on haemoglobin. In 1976, the 2,3-diphosphoglycerate binding site was targeted (Beddell *et al.*, 1976). Dibenzyl-4-4'-dialdehyde modelled into the site showed good steric fit, with the aldehyde groups in good positions to form Schiff's bases with Lys residues. More hydrophilic compounds were also tested, both for increased solubility and for potential salt bridges for higher affinity. Right-shifting of the O_2 association curves was seen for all the compounds with potency in the order predicted from modelling.

Further work on haemoglobin illustrated the potential of sites not exploited by nature for rational drug design. A series of aldehydes designed to occupy the alpha-chain terminal amino region were tested (Beddell *et al.*, 1984). These showed both left-shifting of the O_2-association curves and anti-sickling action in sickle cells.

Manual fitting of compounds into pockets has not yet been totally superseded. In recent work on thymidylate synthase, two series of leads were developed from manually positioned ring systems (Appelt *et al.*, 1991). Initial positioning and modification made extensive use of the GRID program.

Novel ligands by database searches

Fitting ligands to proteins

Several programs have been developed to scan databases of small molecules to search for good fits into a user-defined site. The most successful has been DOCK. Even in its original simple formulation judging fit solely on steric grounds, this was able to locate a sub-millimolar inhibitor of HIV protease from the Cambridge Structural Database of small molecules (DesJarlais *et al.*, 1990). The program has recently been refined to rank molecules by energy of interaction and found several sub-millimolar inhibitors of thymidylate synthase. This work used a larger database of 3D chemical structures, originating not from crystallography, but from programs generating 3D

structures from 2D formulae. The current position with DOCK has been reviewed recently (Kuntz, 1992). Refinements being contemplated include searching ligand conformational space, correction for desolvation on binding and consideration of the subtle differences between closely-related enzymes.

A more recent preogram CLIX (Lawrence and Davis, 1992) obtains initial fits of molecules from a database on steric grounds while ensuring at least three highly favourable contacts. It then ranks them by binding energy. This it approximates as the sum of the energies obtained from precalculated GRID maps for several probes together comprising the molecule. The program performed well in test cases but has not yet been applied to new problems.

Sudarsanam *et al.* (1992) have had success with a much simpler approach. They approximate the shape of each molecule in a database to an ellipsoid and compare the dimensions of each to the ellipsoid obtained from a cast of the binding pocket. This rapid, purely steric approach was able to reproduce some of the results of DOCK and to discover sub-micromolar inhibitors of cysteine proteases.

Other programs dock small fragments from a database and then try to link them together to form a viable inhibitor. The advantages of this fragment-based approach have been pointed out in a series of papers by Chau and Dean (1992a,b,c). In particular, the use of fragments from known structures automatically removes the potential problems with impossible or unstable atomic arrangements that an automatic structure generator might otherwise encounter. Also important is the greater predictability of atomic properties found for a particular atom type in a certain atomic arrangement compared to the same atom type in its full range of possible atomic arrangements. The authors first generated a database of the most commonly occurring molecular fragments (Chau and Dean, 1992a) then determined values and transferability for their bond lengths (Chau and Dean, 1992b) and atomic residual charges (Chau and Dean, 1992c).

LUDI (Böhm, 1992a) begins by locating potential sites for H-bonds to, and hydrophobic contacts with, a user-defined site. Fragments, typically of 5-20 atoms,

excluding hydrogens, are then fitted to these potential interaction sites. They are scored according to goodness of fit to these sites, contact with protein and size. Further searches are then made for fragments that can be linked to the first and that have additional favourable interactions with the protein. The program has been refined by modification of the interaction site location method and inclusion of a facility to add fragments to known inhibitors (Böhm, 1992b). Suggested improvements to known inhibitors were found to be favourable experimentally where data were available.

GroupBuild (Rotstein and Murcko, 1993a) works similarly. An initial fragment is fitted to seed atoms defined by the user, guided for example by GRID. Further fragments are 'grown' from available hydrogens of the first fragment. These are scored by energy using a precalculated potentials grid with correction for changes in solvent-accessible surface areas. GroupBuild was able to reproduce key features of known inhibitors.

Fitting ligands to other ligands

Searching databases for molecules similar to known ligands can involve either 2D or 3D search methods. The former use only measures of atomic connectivity whereas the latter consider the 3D arrangement of atoms and groups. Interest in 3D similarity searches for ligands has been stimulated by the availability of large databases of 3D molecule structures. These are derived from 2D databases using sophisticated molecular mechanics programs such as CONCORD (Rusinko *et al.*, 1988). A comparison of four different 3D similarity search methods and a 2D similarity search has been made (Pepperrell and Willett, 1991). Narrowly the best of the 3D similarity search methods was Atom Mapping. For two molecules A and B, this involves individually locating the atoms in B most like those in A. The overall similarity between A and B is then the sum of the individual atomic similarities. Interestingly, the method was not superior to the 2D similarity search althought the output was different. Subsequent attempts to combine the 2D and 3D methods failed to produce a better result than either method individually. Attempts have been made to take account of the flexibility of molecules in

the search database (Clark *et al.*, 1992; Clark *et al.*,1993). This work has shown that the simplest index of flexibility, the number of rotatable bonds in a structure, produces the same ranking of molecules as more complex indices. Ranking of molecules in a database, so that the least flexible are considered first, has been shown to significantly speed up the search process (Clark *et al.*, 1993).

2D similarity searches can be carried out using various topological indices as search criteria (Rouvray, 1993). The usual approach is to calculate several or many indices for each molecule. Each molecule is then 'plotted' as a point in multidimensional space, with each axis representing an index, and its similarity to the reference molecule calculated as a distance by Pythagorus' theorem. Typical indices are derived from atomic connectivity, with or without consideration of atomic types, or molecular fragments present in each molecule. Topological similarity searching has recently found reverse transcriptase inhibitors (Romero *et al.*, 1991) which have been developed into clinical candidates (Romero *et al.*, 1993).

The most common type of 3D ligand search used to date firstly involves identification of a pharmacophore. A pharmacophore is identified from a superposition of known ligands. It consists of a series of features common to potent ligands (and therefore assumed to be important for binding) such as acidic groups or aromatic systems. The geometric arrangement of these features is defined in terms of distances and angles. In a series of papers the efficiency of a two-pass search technique for scanning molecular databases has been demonstrated (Jakes and Willett, 1986; Jakes *et al.*, 1987; Clark *et al.*, 1991). In a rapid first pass most molecules are discarded on the basis simply of intramolecular distances. A second pass then considers geometric arrangement.

3D similarity searching has recently been used successfully in combination with the DOCK program mentioned above for the design of thymidylate synthase inhibitors (Stoichet *et al.*, 1993). DOCK produced a lead compound with a binding constant of 900μM. Similarity searches produced other compounds which were scored by docking and the highest-ranking tested. This scheme ultimately produced a compound of binding constant 3μM.

Novel ligands by computer generation

Computer generation of putative inhibitors has at least two important advantages over the database search approach. The first of these is the ability to cover a wider range of possible structures, novel molecule types as well as those previously studied. Secondly, there is the potential to produce molecules of superior complementarity to those present in databases.

In a series of papers (Danzinger and Dean, 1989a; Danzinger and Dean, 1989b; Lewis and Dean, 1989a; Lewis and Dean, 1989b) Dean and colleagues have investigated automated site-directed drug design. The first step is the location of potential H-bond partners. Planar, polycyclic molecular skeletons are then fitted to the protein to optimise the fit of skeleton vertices to site points. After pruning of the skeleton to remove parts that sterically clash with the protein, the skeleton can be converted into putative ligands, taking into account chemical considerations. Using observed interaction sites in dihydrofolate reductase (DHFR) and trypsin this approach was able to produce molecular templates strongly resembling known inhibitors.

This approach has recently been combined with the program DOCK (Lewis *et al.*, 1992). Compounds found by DOCK that are suitably relatively positioned are combined to produce molecular skeletons. These require no pruning since the DOCK compounds were selected for steric fit.

Another program, LEGEND, by Nishibata and Itai (1993), builds ligands by adding atoms one at a time. The initial starting structure can be part of a known inhibitor or a computer-generated fragment designed to interact with a functionally important region of the target molecule. During molecule generation, required numbers of rings and H-bonding atoms can be specified. Only carbon atoms are added, but these are changed to heteroatoms if the electrostatic potential shows this to be favourable. Selection from among the generated molecules is by energetic assessment, both inter- and intra-molecularly. The program proved capable of generating the skeletons of known inhibitors of DHFR and provided many other suggestions, although these have yet to

be synthesised and tested.

Very similar to LEGEND is Genstar (Rotstein and Murcko, 1993b). However, Genstar considers intermolecular interactions during molecule generation and allows a realistic range of bond angles and lengths. Once again the program could produce molecules similar to known inhibitors when provided with the protein structure.

An alternative automatic design approach is the construction of a putative ligand from small fragments. The GROW program (Moon and Howe, 1991) automatically produces ideas for peptide and peptide-like ligands by iteratively adding residues from a library of conformations to an initial seed. The seed position is user-controlled so that any of the previously mentioned methods for identification of potential ligand binding sites can be used. The addition of library residues is guided by an energetic assessment which includes steric and electrostatic interaction with the receptor, internal strain energy of the putative ligand and consideration of solvent effects. GROW was able to reproduce crystallographically-determined conformations of known peptide-like inhibitors. More impressively, using a renin structural model built by homology, the program found a novel heptapeptide that proved to inhibit with K_i=30μM.

3.2 Methods

3.2.1 General Molecular Modelling

For general purposes such as molecule building, movement and display the program SYBYL was used (SYBYL 5.5 Manual). Energy minimisation may be carried out using the manufacturer's own force field (Clark *et al.*, 1989). This was used for some small jobs for convenience but all major molecular mechanics was done with the program XPLOR (Brünger, 1990). Like most programs, SYBYL was run on an Evans and Sutherland ESV 20 graphics workstation.

The program INSIGHT was also available (INSIGHT II User Guide) with a similar

range of functions to SYBYL. This was used occasionally for manipulation of LUDI results. INSIGHT (including LUDI) ran on a Silicon Graphics Indigo workstation.

3.2.2 GRID

The GRID programs were kindly supplied by Dr. Peter Goodford (Molecular Design Ltd., Oxford). They were run, unless otherwise indicated, with default parameters and with appropriate probes chosen by menu. The crystal structure of *E.coli* PFK bound to ADP and FBP (Shirakihara and Evans, 1988) was used as the target. Of the slightly different 'open' and 'closed' conformations seen in this structure, the 'open' one was arbitrarily chosen. GRID was run in a box with coordinates $x_{min} = 0$, $x_{max} = 16$; $y_{min} = -25$, $y_{max} = -8$; $z_{min} = -15$, $z_{max} = 0$ (coordinates in Å). This box generously enclosed the FBP binding site. The GSYB program was used to convert the GRID output into contours that could be displayed using SYBYL. Three or four contour levels were generally displayed covering energies of interaction in the range minimum to around half-minimum.

3.2.3 LUDI

The program LUDI was available in the Ligand_Design module of the INSIGHT program. It was run with default parameters except for the Max RMS value. This defines how well the geometry of a particular fragment must fit that of the protein in order to be considered. Experiment showed that with the default value of 0.3Å very few fits were produced. Previous work had also suggested that 0.3Å might be unnecessarily prohibitive (Verlinde *et al.*, 1993). For this reason a value of 0.5Å was used.

The whole of the active site of the product-bound *E.coli* PFK crystal structure (Shirakihara and Evans, 1988), both ADP and FBP sites, was used as receptor for attempted fragment fitting. This was accomplished using three runs; one for a sphere

radius 8Å around Asp127 Cγ, one for a sphere 10Å around O4 of ADP and the third for a sphere 7Å around Ile130 Cδ. Two further runs, intended to focus on regions of potential difference between *E.coli* and RM PFKs were carried out with a Max RMS value of 0.4Å. These were spheres of 10Å around His223 Nε2 and 9Å around Arg305 Cζ.

In the second stage of searching for new fragments to link to original finds, spheres of 8Å centred on the original find and a Max RMS value of 0.3Å were used.

3.2.4 2D similarity searches

The 2D similarity search of the Cambridge Crystallographic Database (Allen *et al.*, 1979) was carried out using the resident Quest88 software on a Vax computer running VMS.

The 2D similarity search for naphtho compounds used a proprietory algorithm based on the distribution of significant chemical groupings (P.Taylor, personal communication). The database used was the Fine Chemical Directory of about 50,000 compounds.

3.2.5 Molecular mechanics with XPLOR

XPLOR was used for both energy minimisation and molecular dynamics simulations of PFK bound to putative inhibitors. The starting protein structure was again the 'open' conformation of the active site seen in the product-bound *E.coli* PFK crystal structure (Shirakihara and Evans, 1988). The following standard procedure was used.

In the large majority of cases water was included in the system. The coordinates of water were generated using SYBYL's SOLVENT command and default parameters. A subset of these water molecules, around 200-250 in an approximately spherical region around the putative inhibitor, was extracted using SYBYL. For both energy minimisation and molecular dynamics simulation, three zones were defined. These

were a zone completely free to move, a zone fixed and an intermediate 'buffer zone' in which the atoms are tied by harmonic constraints to their starting coordinates. The free zone included the putative inhibitor, ADP, Mg^{2+} and all water molecules. Also included were residues defined as contacting FBP and ADP in the complex crystal structure (Shirakihara and Evans, 1988). These were G11, Y41, R72, F73, R77, D103, G104, S107, G108, T125, D127, D129, M169, G170, R171, E222, H249, R252. Excluded, due to limitations in the molecular modelling software, were the residues that contribute to the P6 binding site from the alternate subunit - R162 and R243. The P6 atom of FBP modelled analogs was fixed in an admittedly crude attempt to compensate for their exclusion. The buffer zone comprised residues S9, G10, D12, G40, L42, Y45, A71, P74, E75, F76, D78, R82, G102, Y106, A109, R111, G124, I126, N128, I130, T136, V168, Y172, T221, H223, G248, I250, Q251 and G253.

Charges for protein and water and ADP were as defined in standard XPLOR parameter files. Charges for FBP analogs and LUDI compounds were calculated using the SYBYL command QCPE MOPAC with default parameters. This provides an interface to the MOPAC program which performs a molecular orbital calculation on a given molecule, one of the results of which is a charge assignment for the atoms. This method was shown in tests to produce very similar charges, for example for amino acids, to those present in XPLOR's supplied parameter files. Charges for united atoms were obtained by adding the charges of contributing atoms; a CH_2 united atom charge from the carbon and two hydrogens, for example. The only other modification made to defined charges was averaging of chemically equivalent groups such as carboxylate oxygens. The MOPAC command sometimes produced different charges for these, presumably because of the conformation-dependent component in its calculation.

Energy minimisation consisted of 200 rounds of Powell-method conjugate gradient minimisation. Before molecular dynamics simulation the atoms were given Maxwellian velocities corresponding to a temperature of 600K. One picosecond simulation at 600K (1000 femtosecond time steps) was then carried out. Finally, 0.6 ps simulation was carried out during which the temperature was decreased in 25K steps from 600K to

300K. A further round of energy minimisation as above was then carried out.

3.3 Results and Discussion

3.3.1 Small phosphocompounds

A two-dimensional database search was carried out of the Cambridge Crystallographic Database containing approximately 90,000 compounds. The search template used is shown in Fig 3.1. This was designed to produce compounds with tetrahedral sulphur or phosphorus moieties which could bind into the 6-phosphate binding pocket of *E.coli* PFK. The results can be summarised as follows.

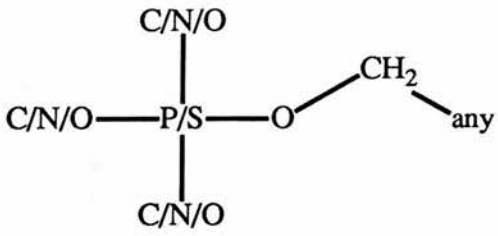
Many nucleotides were suggested, but their effects on PFKs are well known. Pyridoxal compounds were also present in the output. Several naturally-occurring small phospho-compounds were suggested - the phosphates of choline, ethanolamine, serine, glycerol and glycerate. The sulphates of choline and ethanolamine were also present. Other suggestions were propargyl phosphate and the 1,2-bismethylsulphonates of ethane, propane and butane. Some of these suggestions are shown in Fig 3.1.

The compounds were then modelled into the FBP-bound *E.coli* PFK crystal structure so as to place the tetrahedral sulphur or phosphorus part of the molecule in the position occupied by the 6-phosphate of FBP. At this stage it was apparent that the bismethylsulphonates were not interesting since there was not room at the 6-phosphate site for the methoxy group. Modelling of propargyl phosphate, the pyridoxal compounds and phosphoglycerate all produced serious steric clashes with the protein which ruled out these compounds as potential inhibitors.

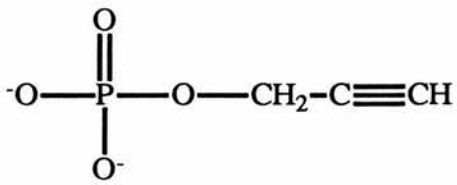
Modelling of phosphoethanolamine (PE) produced a more interesting result. Placing the phosphate in the 6-phosphate binding pocket put the protonated amino group in good position to salt bridge Glu222. Modelling of phospho-D-serine (PDS) from this

Fig 3.1 2-D database search

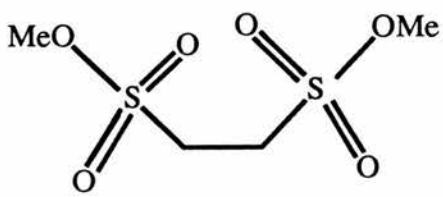
Search template



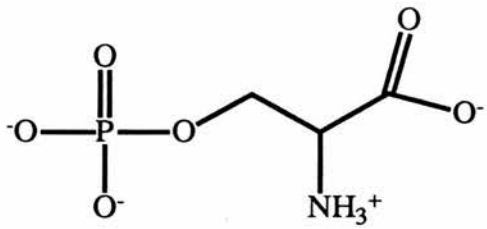
propargyl phosphate



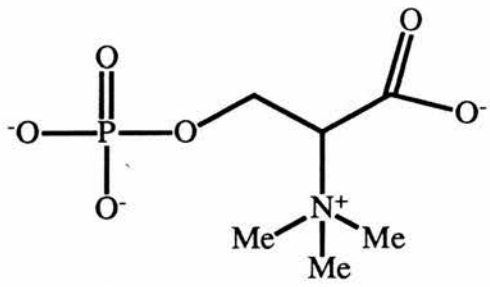
ethane-1,2-bismethylsulphonate



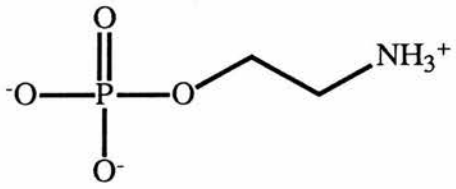
phosphoserine



phosphocholine



phosphoethanolamine



PE conformation placed the carboxylate, favourably, pointing out into solution. The carboxylate of modelled phospho-L-serine (PLS), however, clashed with the protein. Phosphocholine (PC) was also modelled similarly but resulted in unfavourable positioning of the $-NMe_3^+$ group against Glu222. At this stage PE and PDS looked promising and the modelled conformations were subjected to a standard energy minimisation/molecular dynamics simulation in the presence of about 300 water molecules around the small phosphocompound. After this procedure, the salt bridges between phospho group and 6-phosphate binding pocket, and amino group and Glu222 were strengthened. In addition, movement of the phosphocompound and Met169 towards each other produced more favourable hydrophobic contacts. The resultant conformation of PDS is shown in Fig 3.2.

Since these small phosphocompounds bear some resemblance to PEP and phosphoglycollate, known allosteric inhibitors, the attempt was also made to model them into the effector site. The phosphoglycollate-bound, T-state *B.stearothermophilus* PFK crystal structure was used (coordinates kindly supplied by Dr Phil Evans, Cambridge). The phosphate and adjacent carbon of PE were overlaid on the respective atoms of phosphoglycollate. Substituting PE for phosphoglycollate involves putting $CH_2NH_3^+$ in place of COO^-. In the crystal structure the phosphate of phosphoglycollate is completely buried and the pocket wraps around the phosphoglycollate with the carboxylate group at the surface 'stapling the two rims of the cleft together' (Schirmer and Evans, 1990). It was therefore not surprising to find that PE, which is larger than phosphoglycollate, could not be positioned without avoiding serious steric clashes with the protein. In addition the carboxylate of phosphoglycollate is near to several basic residues which are thus unfavourably close to the NH_3^+ of modelled PE. PC and the phosphoserines are larger than PE and therefore cause even greater steric clashes.

The effector site pocket can expand to accommodate the much larger ADP (Schirmer and Evans, 1990). Hence, binding of small phosphocompounds to another, more open, conformation of the pocket cannot be ruled out. Modelling of such a putative

Fig 3.2 PDS after molecular mechanics

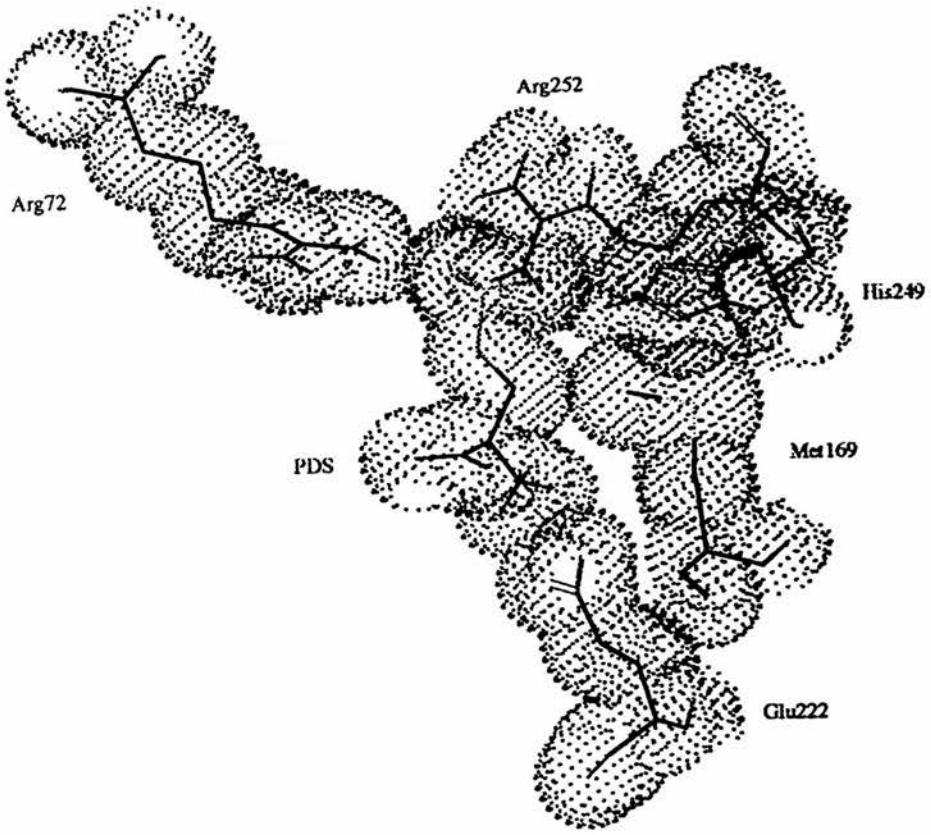
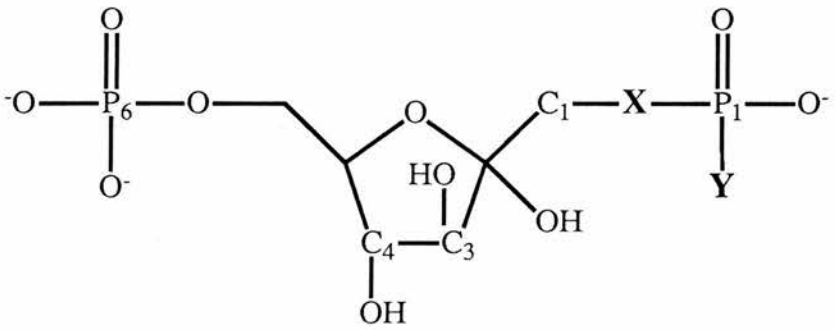


Fig 3.3 Format of FBP analogs



conformation would involve too many uncertainties to be justified.

Thus, the modelling studies suggested that small phosphocompounds, PDS and PE in particular, might bind to PFK. It was decided to test PC and PLS as well. Differences in the behaviour of these four compounds could then be explained using their different structural features.

3.3.2 Bisphosphate compounds

3.3.2.1 Design of fructose-1,6-bisphosphate analogs

Several classes of FBP analogs were modelled. Many were of the formula shown in Fig 3.3 with the P1 region of the molecule modified at bridging atom X and/or at atom Y. The rationale behind modification here was twofold. First, substitution of oxygens at X and Y would prevent the hydrolysis off of the 'P6' moiety by the 'backward' reaction of PFK. This is clearly essential for any useful inhibitor. Secondly, modifications here would reduce the negative charge on this portion of the molecule from the -2 carried by FBP. This would be predicted to improve binding affinity since P1 in the crystal structure is very close to, and indeed may H-bond the catalytic Asp127 and residue Asp129 (via a water molecule). This charge repulsion is probably responsible for the expulsion of product FBP during 'forward' enzyme action so that its reduction would be favourable.

An alternative way to reduce charge here is the substitution of pentavalent phosphorus by hexavalent sulphur. Even with X and Y both remaining oxygens making a sulphate, this gives a charge of -1 compared to the -2 of FBP.

Several attenuated FBP derivatives were also modelled where there was no tetrahedral group attached to C1. These more resemble F6P than FBP. F6P is known to bind with around three orders of magnitude greater affinity.

Also considered were some simple modifications of ring substituents. These alone would not produce non-hydrolysable compounds but could be easily combined with other modifications.

GRID was used in two ways during the design process. In the first the program was used to suggest starting conformations for the molecular mechanics procedure. For example, the Y=Me modification was studied. There are three oxygen atoms of the 1-phosphate group which are chemically equivalent, but differently positioned in the crystal structure. 'Mutating' these to methyl groups would give three distinct possible starting conformations for this analog. GRID enabled a rational choice to be made between these possibilities. A methyl probe was used with GRID and produced a 'tube' of favourable contour to which one of the three oxygens was much closer than the others. That oxygen was therefore the obvious choice for conversion to Me before molecular mechanics. The other way in which GRID was used was to suggest entirely new possibilities. For example, contours produced with a protonated primary amino group probe revealed several highly advantageous positions near active site acidic residues (Fig 3.4). Amino groups placed in these energy minima could be linked to the tetrahedral centre by carbon chains. In Fig 3.4 the contours coloured green, yellow, orange and red indicate, respectively, probe-target interaction energies of -9, -12, -15 and -18 kcal mol^{-1}. The colour scheme for different elements that is used here is standard throughout this thesis. White atoms represent carbon and red, blue, cyan, orange and yellow are used for oxygen, nitrogen, hydrogen, phosphorus and sulphur respectively.

Proper assessment of the results of molecular mechanics on putative inhibitors required that the intrinsic variability of the procedure be ascertained. The molecular dynamics procedure has a large random element so that somewhat different results would be expected from the same starting position after different runs. This was indeed seen on comparision of three applications of the procedure to the hydrated crystal structure. Particular variability was noted in the positions of active site arginine residues 72 and 171, and to a lesser extent 252. This is consistent with their disorder in determined crystal structures. Also seen was a significant variability in the position of His249. A large movement of FBP out of the pocket is seen; the P1 atom moves

Fig 3.4 GRID NH_3^+ contours at the FBP binding site (see text)

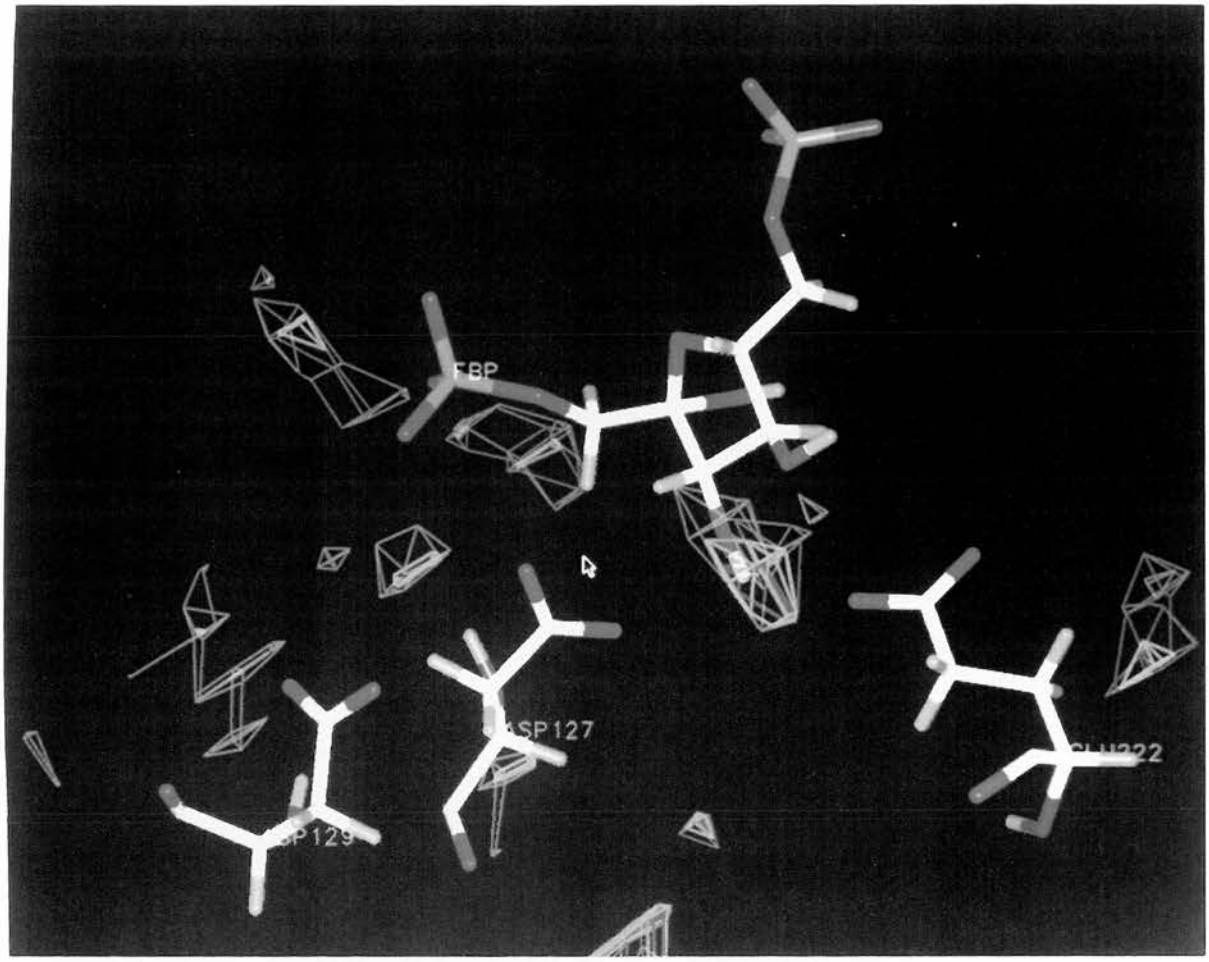
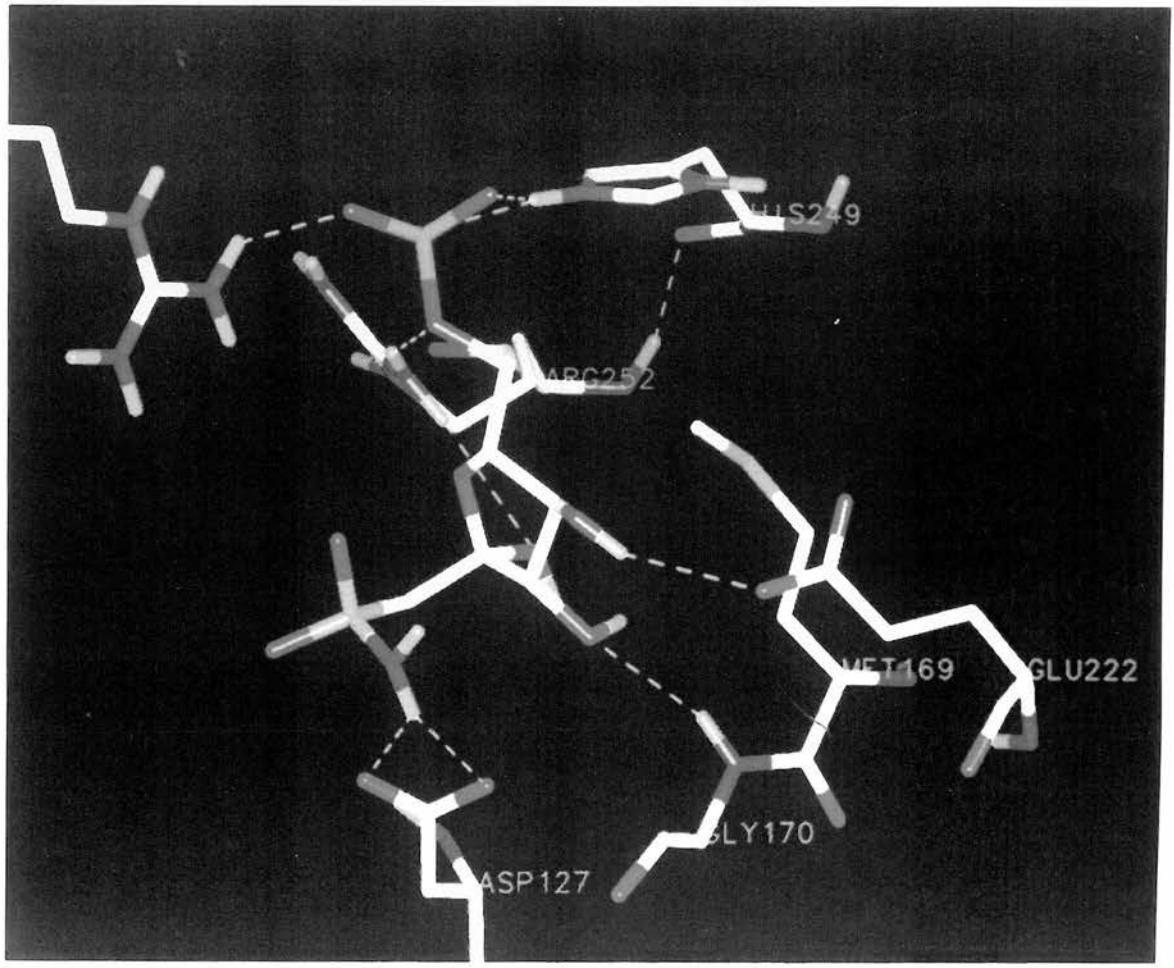


Fig 3.5 A sulphonamide derivative H-bonds Asp127



2.7, 3.1 and 3.5Å in the three simulations. In all three cases Met169 moves to maintain contact with the FBP as it is pushed out of the pocket and water partially fills the gap left behind.

The observation of such variability in the protein after runs from identical starting situations means that differences observed in these regions for runs with analogs are not necessarily analog-specific.

It was originally envisaged that the compounds would be ranked by energy measurement. However, the energy measured was sensitively dependent on the number of water molecules included in the calculation. Arbitrary choices of included water molecules - 50, 100 etc., lead to different rankings of compounds, diminishing confidence in this method of assessment. The measurements would in any case have been of limited utility since they include no entropic component, a factor which would be expected to vary quite widely among the variously sized molecules studied.

Any reliable method of assessment of compounds should suggest that F6P binds significantly more tightly than FBP as seen from fluorescence measurement (see Fig. 4.6). The potential objection to this argument is that studies were carried out on a PFK structure bound to FBP and therefore possibly specifically conformationally changed by its binding. However, this appears invalid in view of the fact that the various PFK structures determined appear to vary surprisingly little in the region of the active site (Schirmer and Evans, 1990). Another method of ranking the molecules, by their buried surface area, was unable to predict F6P binding favoured over FBP.

Ultimately then, only fairly qualitative methods were available to judge the likely binding affinity of compounds. Clearly, H-bonds and salt bridges would contribute to improve binding affinity. An obvious difference between the molecular mechanics results for weak binding FBP and tight binding F6P was the complementarity of enzyme and compound. Thus, a marked movement out of the pocket, presumably driven by P1 - Asp127 repulsion, was seen for FBP, resulting in the buried face of FBP pulling away from the enzyme. This also resulted in the loss of the strong H-

bond to Glu222. In contrast, F6P remains in good contact with the enzyme, its 1-hydroxyl H-bonding Asp127. Movement out of the pocket was hence used as an, admittedly crude, indication of poor binding affinity.

Experiment confirmed that reducing the charge on the P1 region, by bonding the phosphorus to carbon atoms rather than oxygen atoms, seems to produce tighter binding analogs. The modelled derivative in which X=CH_2, and an overall charge here therefore of -1, moves less far out of the pocket than FBP itself. After molecular mechanics the phosphorus atoms of duplicate runs were 2.0 and 1.9Å from the position of P1 in the crystal structure. This compares with 2.7-3.5Å for FBP itself. In addition this derivative retains the H-bonds observed in the crystal structure. The derivative in which X is deleted and P directly bonded to C1 (called here the null derivative) also has a charge of -1 on the P1 region. However, the loss of one bond places the P1 portion rather closer to Asp127. This is probably the reason for the greater movement out of the pocket seen with this compound. Final P atom positions in two null derivative structures are 2.7 and 2.1Å from the P1 position in the crystal. In addition, H-bonds present in the crystal between ligand and protein are broken with for this derivative. The charge on the P1 region is zero for a third studied derivative in which X is again absent and Y=CH_3. This moves very little during the molecular mechanics procedure, if anything deeper into the pocket, and retains H-bonds. Thus the prediction would be that X=CH_2 and X=-, Y=CH_3 more so, would bind with higher affinity than FBP.

Several sulphur derivatives were modelled. Some of the more interesting were null derivatives in which X is absent. The sulphonate with Y=O^- moved very little during the molecular mechanics and retained the H-bonds found in the crystal. A sulphonyl fluoride derivative X=-, Y=F was modelled to see if suitable nucleophiles were near the sulphur in the final structure. Sulphonyl fluorides are a well known class of irreversible enzyme inhibitors. However, no suitable nucleophiles were seen and the molecule moved quite far out of the pocket so that the F finally H-bonded Arg171. One of the most interesting compounds of all was the sulphonamide X=-, Y=NH_2 which

has a charge of zero in this region. Not only does this compound retain the H-bonds of the crystal structure and pack very well against the enzyme, but the new amide group forms H-bonds with Asp127 (Fig 3.5). In Fig 3.5, as elsewhere in this thesis, inferred H-bonds, as defined by SYBYL, are shown as yellow dotted lines. Another sulphonamide - X=O, Y=NH_2, was less interesting; it retained the original H-bonds but with the new amide group forming no new ones.

Amongst the group of compounds suggested by GRID maps was X=O, Y=CH_2OH. Two different favourable positions were obtainable for the hydroxyl group given the crystal position of the P1 moiety. In both cases molecular mechanics produced good final structures which retained the H-bonds found in the crystal structure. In one case the new hydroxyl was H-bonded to Thr125 and Asp127, in the other to Gly11 backbone alone. Several compounds were modelled as a result of the observation of highly favourable GRID contours for an NH_3^+ probe (Fig 3.4). The compound X=CH_2, Y=$CH_2CH_2NH_3^+$ produced a structure in which the new amino group salt bridges both Asp127 and Asp129 and H-bonds Thr125. This is achieved without breaking the H-bonds present in the crystal structure and with the new phosphorus only 1.4Å away from the P1 crystal position. On the debit side is the addition of two rotatable bonds whose freezing out of rotation on binding would be unfavourable entropically. The shorter compound X=CH_2, Y=$CH_2NH_3^+$ also looked interesting, again retaining the crystal's H-bonds and with the new amino group salt bridging Asp127 and H-bonding Thr125. Of the attenuated compounds lacking a tetrahedral centre attached to C1, only the molecule with a chain $CH_2NH_3^+$ looked very interesting. This becomes buried deeper into the pocket during the molecular mechanics procedure enabling the placement of the amino group between Asp127 and Glu222.

Possible favourable modification of FBP ring substituents focussed on the replacement of either C3 or C4 hydroxyl with a primary amino group (NH_3^+). GRID contours suggested that either substitution would be favourable but favoured C3, whereas there appeared to be more space for the new NH_3^+ at the C4 position.

Modelling and molecular mechanics of the C3-substituted analog produced a highly attractive structure with the amino group salt bridging both Asp127 and Glu222 (Fig 3.6). Displacement of the P1 group was comparable to FBP itself at 2.6Å. Substitution at C4 yielded a structure in which the new amino group salt bridged only Glu222 but this appeared to be sufficient constraint to reduce P1 movement to 1.6Å. In both cases substitution of OH with $CH_2NH_3^+$ produced poorer structures. Thus OH to NH_3^+ substitutions at C3 and C4 of the fructose ring of FBP both appear very favourable, particularly since they could be combined with favourable P1 modifications.

3.3.2.2 Modelling of available bisphosphates

Two epoxy derivatives of FBP, α and β anomers were available (kind donations of Pr. Jacque Perié, Toulouse). Their structures are shown in Fig 3.7. In addition to having an epoxide group in place of C3 and C4 hydroxyls, these derivatives were methylated at the C6 hydroxyl in order to preserve anomericity. Epoxy compounds are of interest as potential irreversible enzyme inhibitors since they can react with nucleophilic groups on the enzyme such as aspartates and glutamates. A recent example is the covalent modification at a glutamate and consequent inhibition of carboxypeptidase A by a designed substrate analog (Kim and Kim, 1991; Yun *et al.*, 1992).

In the FBP-bound PFK complex crystal structure the FBP is bound as the β anomer. The β epoxy derivative was modelled from the crystal structure of FBP and energy-minimised to correct the geometry of the new epoxide group. The FIT command of SYBYL, using both phosphates and ring atoms, was then used to optimally overlay the β epoxy derivative onto the FBP crystal position. In this position neither of the potential nucleophiles for reaction with the epoxide group, the side chains of Asp127 and Glu222, were well placed. They were both too far away and in poor geometric relation to the epoxide group.

The α epoxy derivative was modelled by alteration of the β structure. This produced a relatively symmetric structure so both P1 and P6 of the derivative were tested in the P6

Fig 3.6 The C3 -> NH_3^+ derivative salt bridges both Asp127 and Glu222 after molecular mechanics regime.

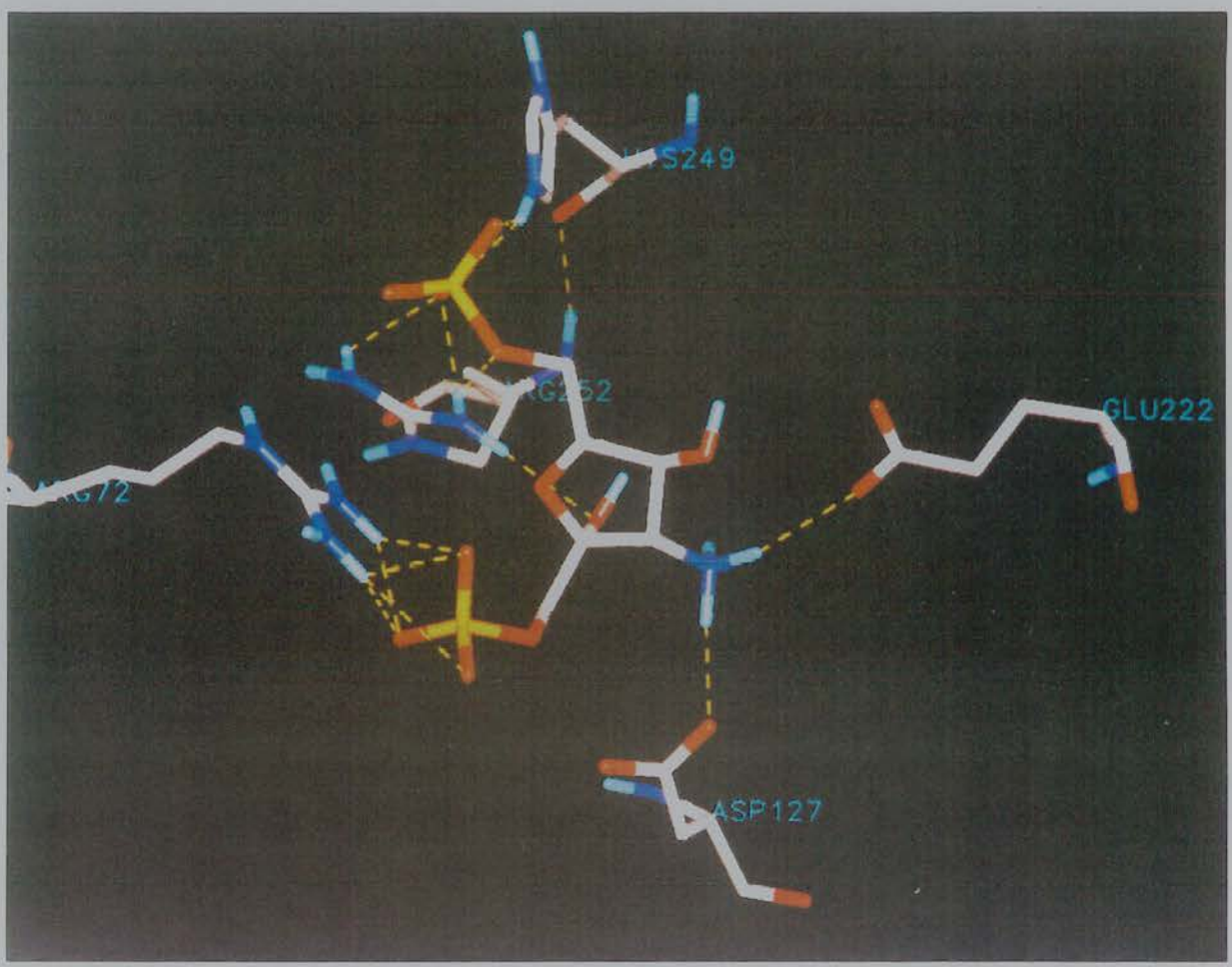
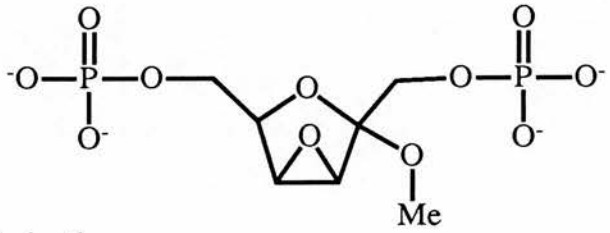
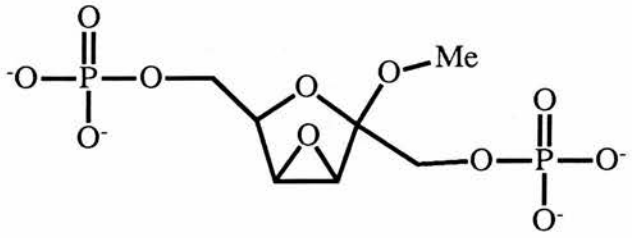


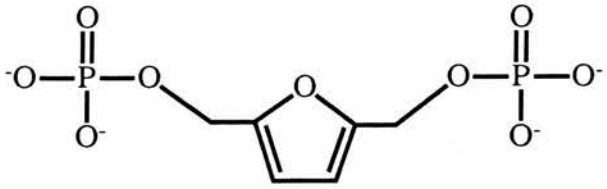
Fig 3.7 Available bisphosphate compounds



alpha epoxy derivative



beta epoxy derivative



Compound 233

binding pocket. With the derivative P6 placed in this pocket bond rotations were made to place the ring as close as possible to the FBP crystal structure. However the resultant structure was poor with P1 clashing with Met169. Placing of P1 in the P6 binding pocket produced a more promising result. Bond rotation was able to overlay four out of five ring atoms onto crystal positions, thus placing the hydrophobic epoxy group favourably close to Met169. In this position the derivative P6 could be rotated to only 1.5Å away from the crystal P1 position. The methoxy group clashes with Arg171 but this residue, along with other active site arginines is known to be flexible from its disorder in crystal structures and from molecular dynamics simulations. However, in this conformation too, neither Asp127 or Glu222 possesses the correct geometry and the required proximity to the epoxide group for reaction to be likely. In addition, in neither this position or in the fitted β analog conformation is there a suitable proton donor for the nascent $-O^-$ that would be generated during reaction.

Thus the modelling studies were pessimistic about possible covalent reaction of either epoxy compound.

Also available was the phosphorylated furan molecule 'Compound 233' shown in Fig 3.7 (another kind gift of Pr. Perié). This was modelled from scratch using the SKETCH command of SYBYL. The furan ring has a very different shape to the ribose ring, being planar with coplanar bonds to the phosphate-bonded carbons. Nonetheless, with one phosphate placed in the 6-phosphate binding pocket, bond rotation enabled the other to be placed close to the position occupied by the 1-phosphate of FBP in the crystal structure. However, the planar furan ring fits very much more poorly than the puckered ribose ring. This and the loss of several H-bonds due to the absence of hydroxyls of Compound 233 suggested that it would bind PFK weakly, if at all.

3.3.3 LUDI compounds

3.3.3.1 Selection criteria

The LUDI runs typically produced several hundred fitted fragments, so that analysis of each was not feasible. The main selection criterion used was LUDI score. LUDI produces a score for each fragment fitted, including components relating to goodness of fit, number of hydrogen bonds and size of fragment. The cut off value was set at 400 or 450 and only those fragments scoring better than this further considered. All of these were inspected visually. This enabled more to be discarded since some fragments contained unprotonated amines which were sometimes unrealistically shown H-bonded to arginines, for example. Another selection criterion was used later after experiment showed little inhibition by tested compounds. Since the size component in the LUDI score is relatively weak the larger fitted fragments (those having more than 24 non-hydrogen atoms) were separately extracted and examined. This produced a few new ideas. Those fragments approximating whole molecules were then checked for commercial availability.

3.3.3.2 Further modelling

For most of the interesting fitted fragments, further LUDI runs were carried out to look for new fragments that might be linked to the original in order to produce a larger, tighter binding inhibitor. LUDI has options to search for new fragments that can link to the original at either one or two atoms. Both searches were carried out for each fragment. The resulting large amount of data will not be discussed in detail, particularly since experimental testing failed to produce evidence of inhibition by any LUDI compounds. Instead some representative examples will be discussed to illustrate the strengths and limitations of this approach, which will undoubtedly be important in future drug design efforts.

It was immediately obvious that linking new fragments would be of more use in some cases than others. The original fragments differed widely in the number of potential

sites for new linkages. Fig 3.8 shows the high scoring fragments from the major LUDI run of the FBP binding site. Sites to which LUDI found extra fragments could be linked are shown as starred atoms (one atom linked to new fragments) and bold bonds (two atoms linked). The number of new fitted fragments suggested varied widely for these fragments; from zero to 41.

Fig 3.9 illustrates the variety of suggestions for new linked fragments produced for just one original fit. The original fitted fragment, 4-nitrodiphenylamine (A18 in LUDI's nomenclature), is shown in Fig 3.9a H-bonding both Gly11 and Asp129 and with hydrophobic contacts to Tyr172. As elsewhere in this thesis, dots are plotted at the van der Waal's radii of the atoms. The dots are used to give some indication of molecular volume. In Figs 3.9b-e the new linked fragment is shown according to atom type and the original A18 fit shown in green. Amongst the suggestions for new fragments doubly linked to the original are those shown in Figs 3.9b and 3.9c. In Fig 3.9b LUDI has suggested adding a phenylene group, making the phenyl group of A18 into a naphthyl group, producing further hydrophobic contacts with Tyr172. In Fig 3.9c an ethylene group has been added to A18 bridging the two rings. This produces no further ligand-enzyme contacts but is a favourable modification since it prevents rotation of the bonds between the two aromatic groups. This removes the unfavourable entropic effect of 'freezing out' their rotation on binding that would otherwise occur. Figs 3.9d and 3.9e illustrate two typical new fragments linked singly to A18. In Fig 3.9d an oxocyclohexyl group has been added, producing further hydrophobic contacts with Tyr172 and a new H-bond with Arg 305. Fig 3.9e illustrates the small addition of a hydroxymethyl group which produces two new H-bonds, with Asp129 and Thr125.

In a few cases more detailed modelling was carried out. An illustrative example is the compound phenylene-1,4-diacetate (BA4), whose original fit appeared to visual examination to be sub-optimal, which was analysed as follows.

Charges were calculated and the improper torsion angle definitions required to maintain ring planarity defined by analogy with tyrosine. A selection of calculated

Fig 3.8 Linkage potential of fitted fragments

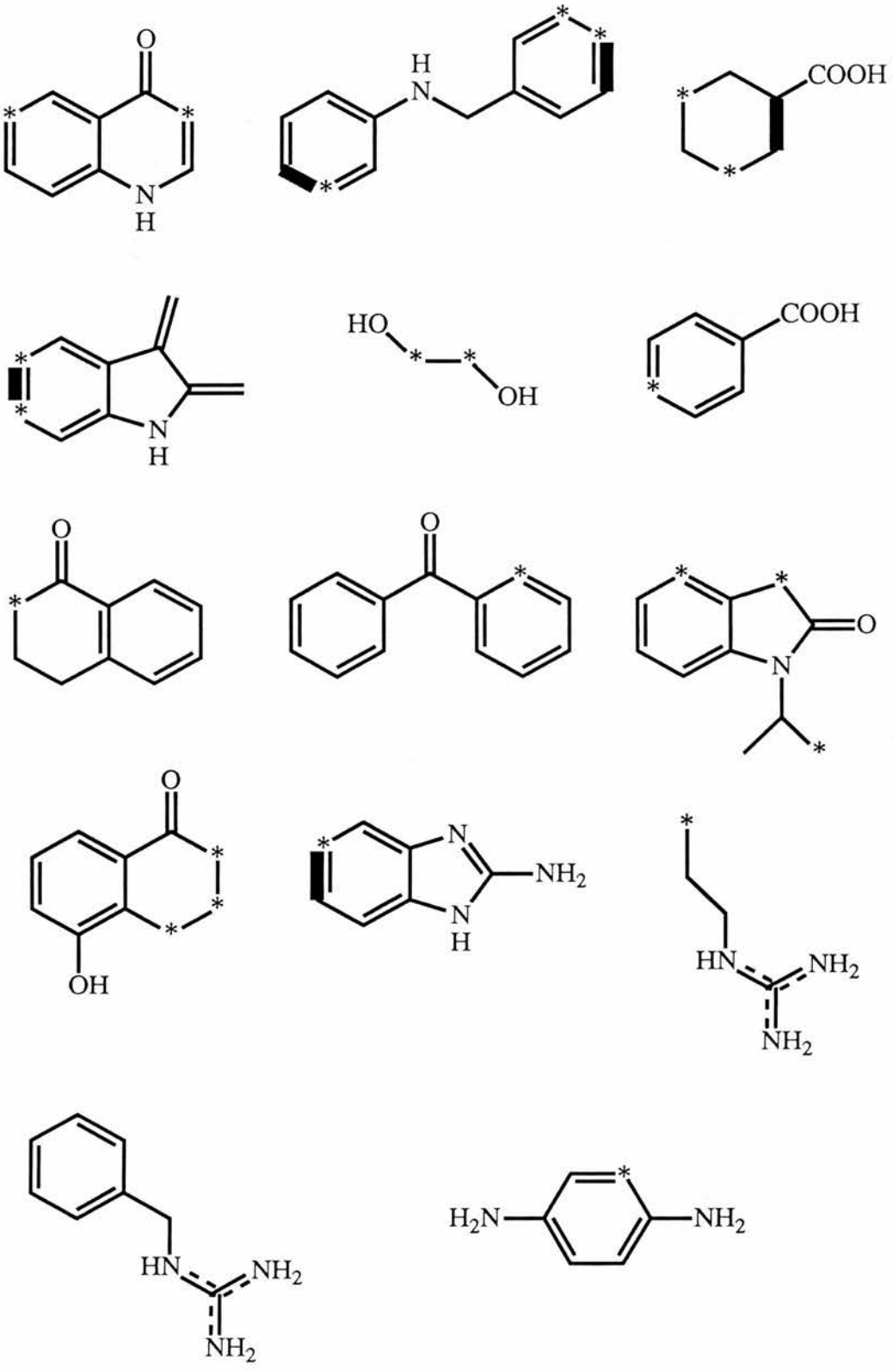


Fig 3.9a Original LUDI A18 fit showing contacts with enzyme

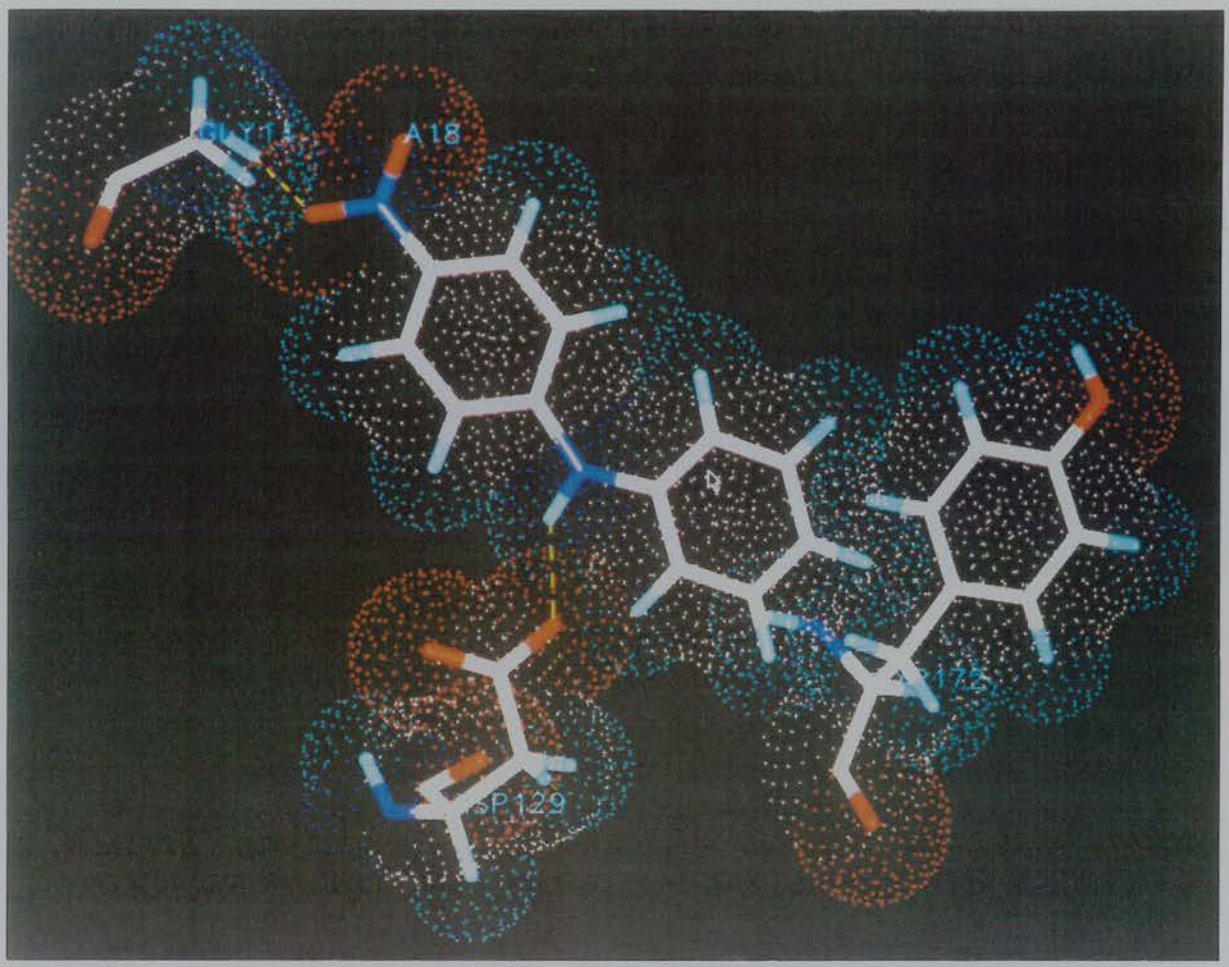


Fig 3.9b Added ring

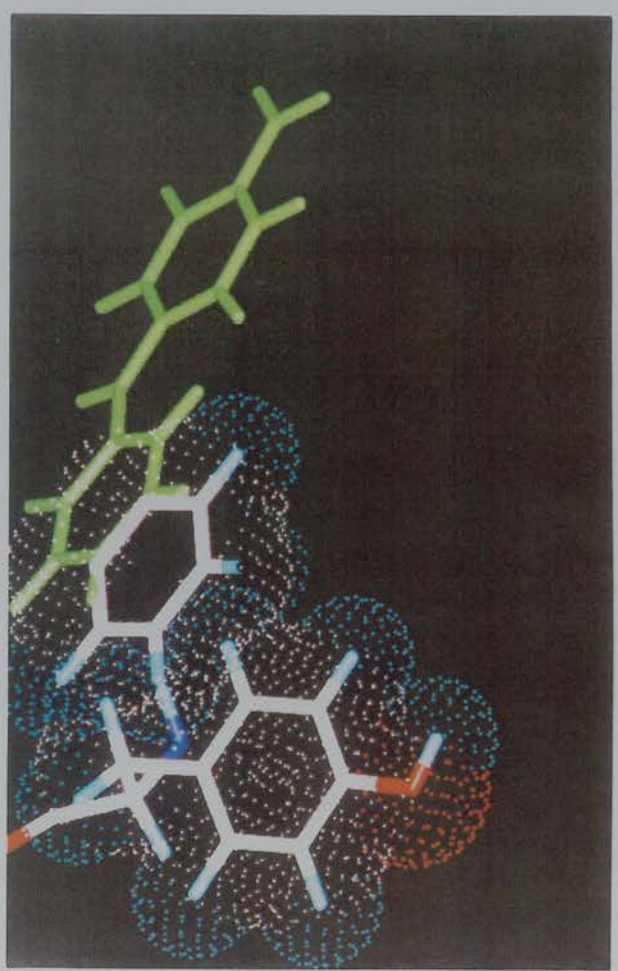


Fig3.9c Intramolecular addition

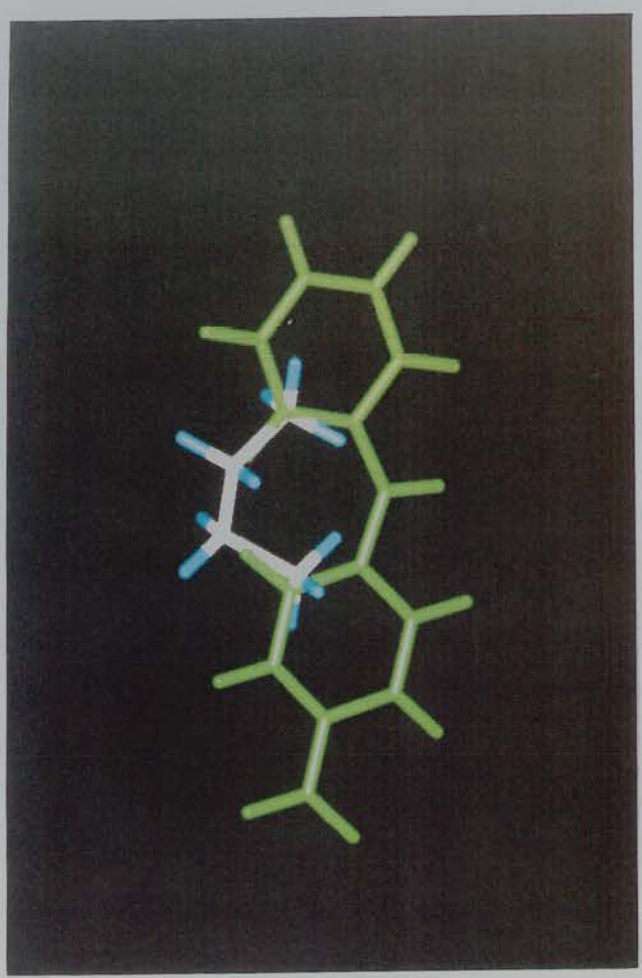


Fig 3.9d Another added ring

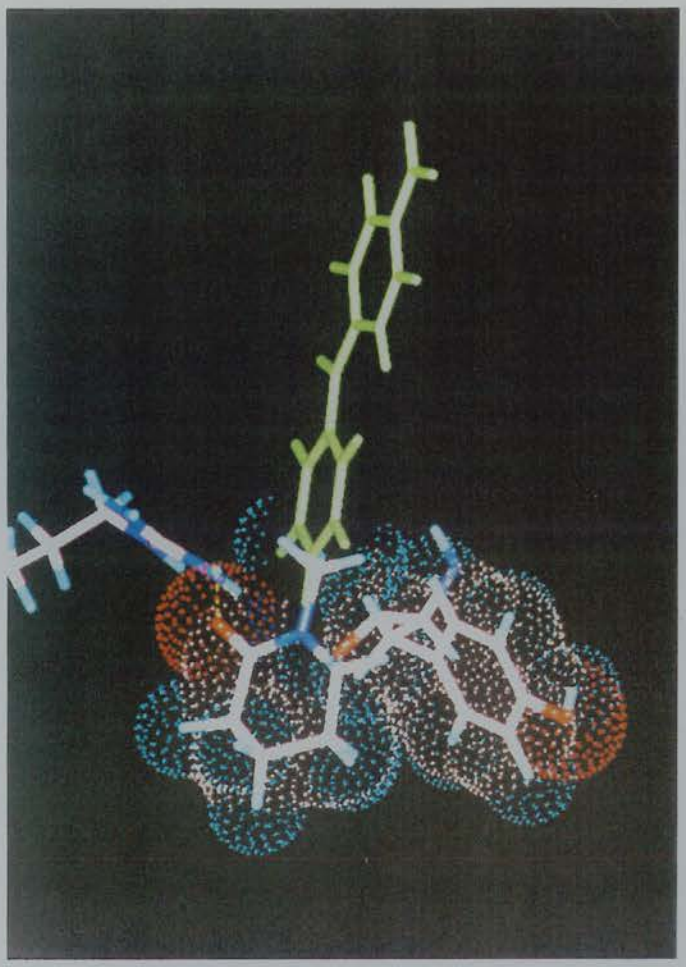


Fig3.9e Added hydroxymethyl

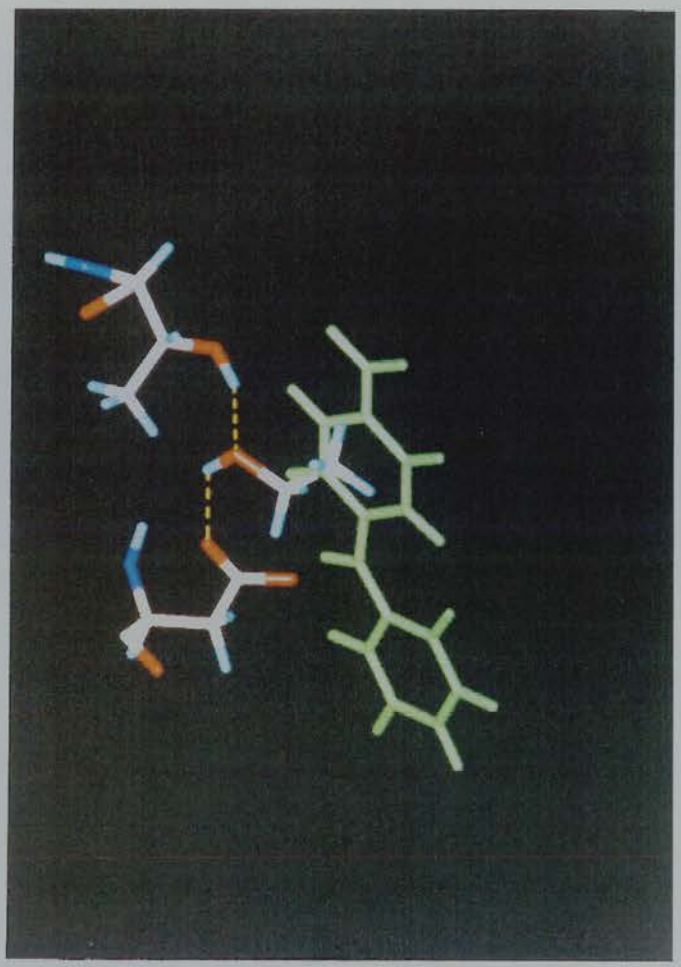
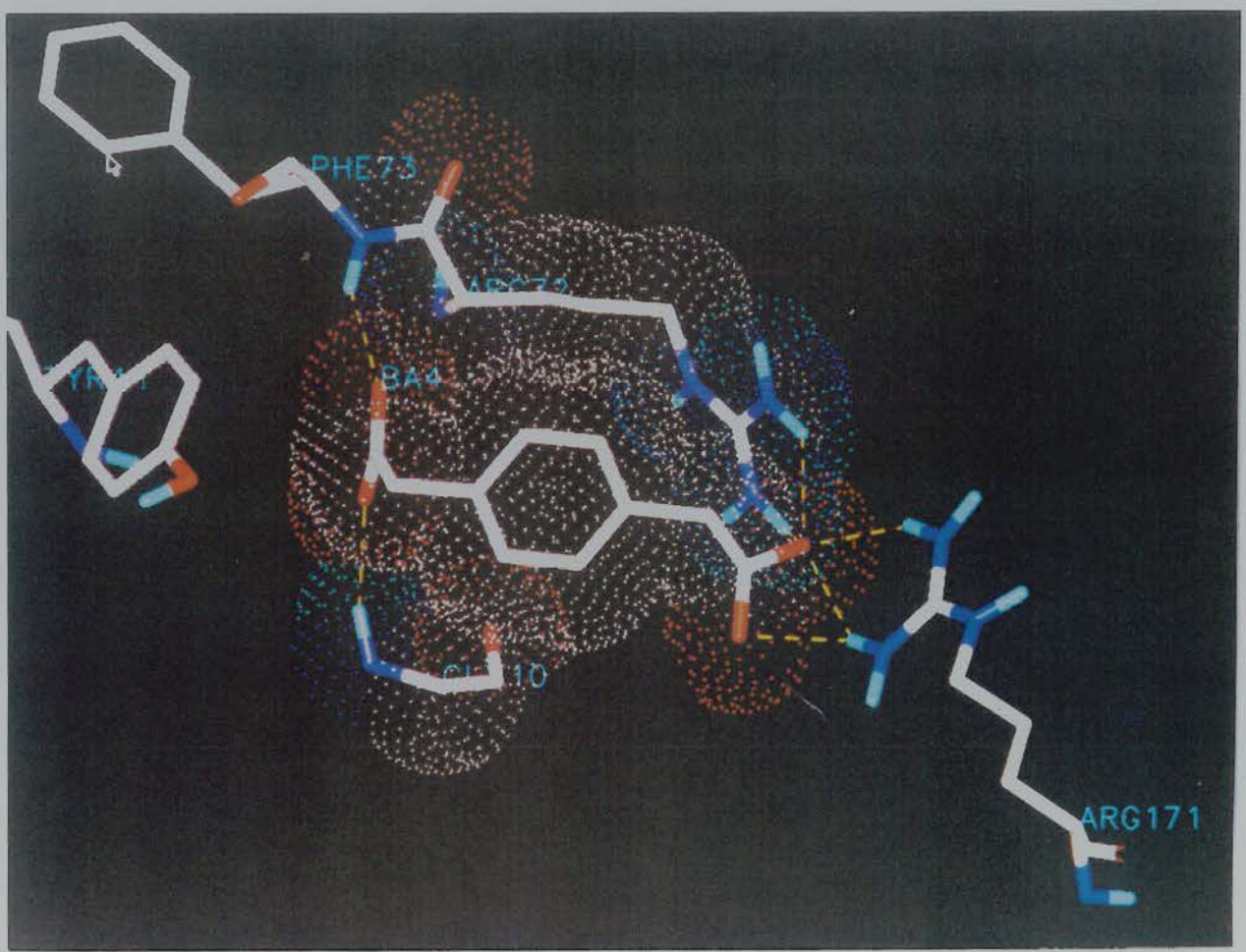


Fig 3.10 BA4 complex after molecular mechanics regime (see text)



water molecules within 10Å of BA4 yielded 288 molecules. The energy minimisation and molecular dynamics protocol described in the Methods was then carried out. Residues of the protein within 5Å of BA4 were mobile during this procedure with a 'buffer zone' defined as residues within 8Å of BA4. The result is shown in Fig 3.10. H-bonds are indicated and the dots indicate the atomic radii of BA4 and of Gly10 and Arg72 with which BA4 forms hydrophobic contacts. This final conformation is the product of large movements of both BA4 and protein. The atoms of BA4 have moved 2-3Å. The largest protein change is the swing of Tyr41 towards the hydrophobic BA4. The hydroxyl oxygen of Tyr41 moved more than 5Å and a new H-bond with Gly104 backbone oxygen was formed. This conformation looks much more favourable than that found by LUDI, particularly in respect of H-bonds. The BA4 carboxylate salt-bridging Arg72 in the LUDI conformation now additionally salt-bridges Arg171. The other has moved so that two H-bonds, to Phe93 and Gly10 backbone NH groups, are present compared to one in the LUDI structure. The improvement in conformation seen after the molecular mechanics procedure illustrates the value of visual examination of potential ligand structures. As elsewhere in this thesis, is is apparent that there remains an important role for the human operator; a combination of computer and human seems to produce the best results.

3.3.3.3 Potential species selectivity

With its approach of linking new, sometimes large, fragments to an original fitted fragment, LUDI was considering regions of the protein quite distant from the catalytic site. A check was therefore made to see if any of the *E.coli* PFK residues contacted by fragments were different in RMPFK. Radical differences could form the basis of binding specificity between the two enzymes and hence possible species selectivity.

Of the residues in the vicinity of the FBP binding site the most interesting were *E.coli* PFK His223 which is replaced by Gly at the RMPFK active site and Arg305 which is Asp in RMPFK. Lys131 in *E.coli* PFK is replaced by Cys in RMPFK. Also different, but less so, were Tyr172 (His in RMPFK) and Ile130 (Phe in RMPFK).

A scan of previous search results showed that fragments contacting His223 fell into two categories. The first was fragments of a guanidino type shown H-bonding His223. Binding of this is made unlikely by charge considerations and is an indication of one LUDI limitation: its ignoring of charge. The second class showed only hydrophobic contacts to His223 whereas exploitation of the His residue's H-bonding potential would appear to be a better route to enzyme inhibition specificity. A *de novo* search was therefore carried out in the region around His223. Among the most promising fits was 1,2 benzene dimethanol (H17 in LUDI's nomenclature). This fragment can be fitted to H-bond Glu222 and Arg171 as well as His223, and also to form hydrophobic contacts to His223. Further runs were carried out to find fragments that could be advantageously linked. Amongst the candidates were a tetrahydrofuranyl group that H-bonded Arg171 as well as providing two further hydrophobic contacts to Tyr172, and a hydroxyethyl group that H-bonded His249. The original H17 fit and the linked tetrahydrofuranyl group are shown in Fig 3.11. Interestingly, a P6 site filling motif - an ethanoylamino group H-bonding His249 and Arg252, that had been previously seen, reappeared. Other extra fragments were found that further contacted His223.

A *de novo* search around Arg305 produced several ideas, the most interesting of which was isatin (AM7 in LUDI's nomenclature). Both carbonyl groups are H-bonded to Arg305 while the NH can H-bond Asp103. In addition this fragment seems close enough to the ATP site to hinder binding. A particularly favourable extra fragment to link to this was a δ-valerolactam ring which produced two extra H-bonds.

A search in the vicinity of Ile130 suggested that a 2,5-diaminopyrimidine (DAP) moiety could be a favourable component of a viable inhibitor . This overall positively-charged fragment can be placed within H-bonding distance of three aspartates - residues 103, 127 and 129. Interestingly, extra fragments can be linked to this base to H-bond Arg305. Simultaneous extra fragments can H-bond Gly11 backbone. Figure 3.12 shows the fitted DAP fragment with a linked N-methyl-2-pyrrolidinone fragment H-bonding Arg305.

Fig 3.11 Original LUDI H17 fit and linked fragment

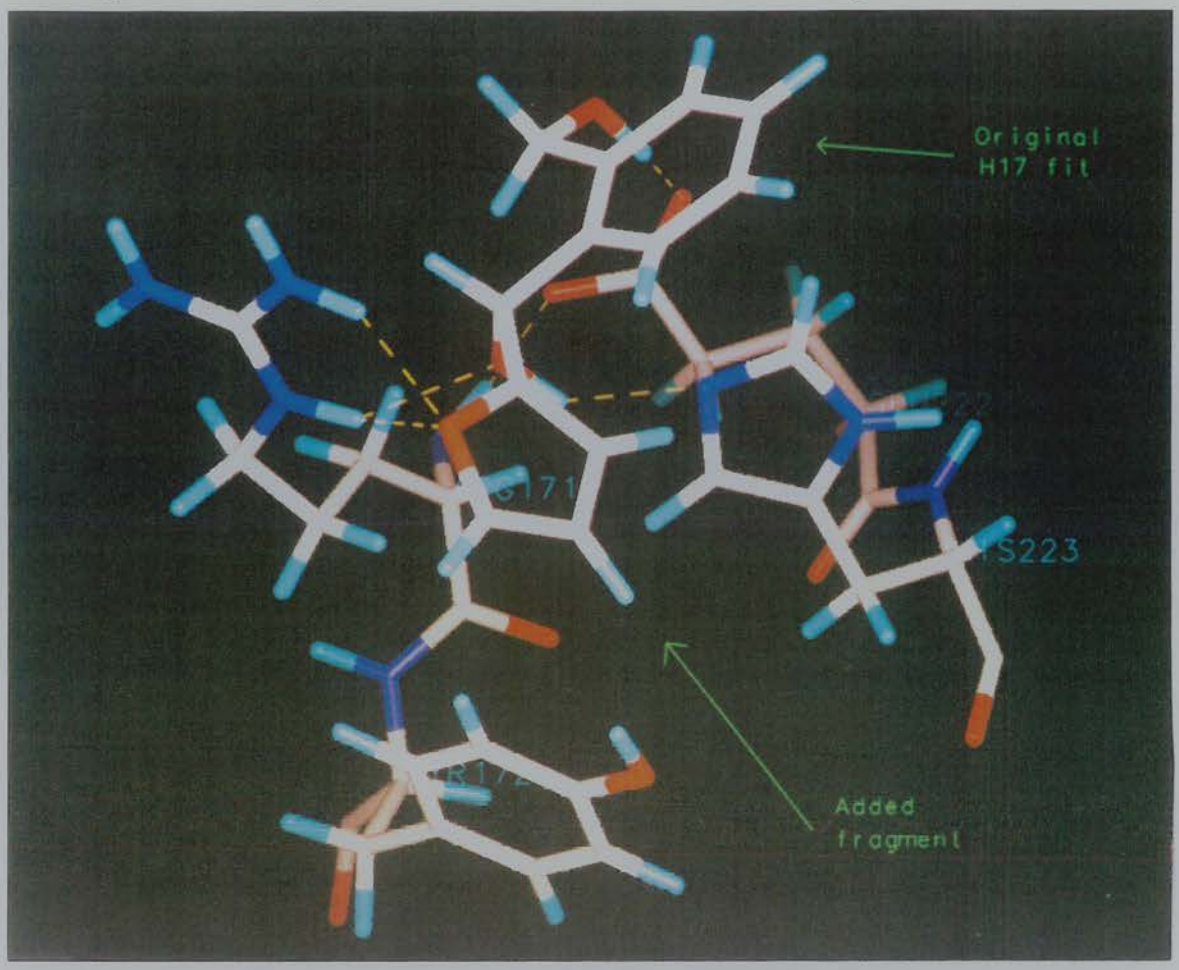
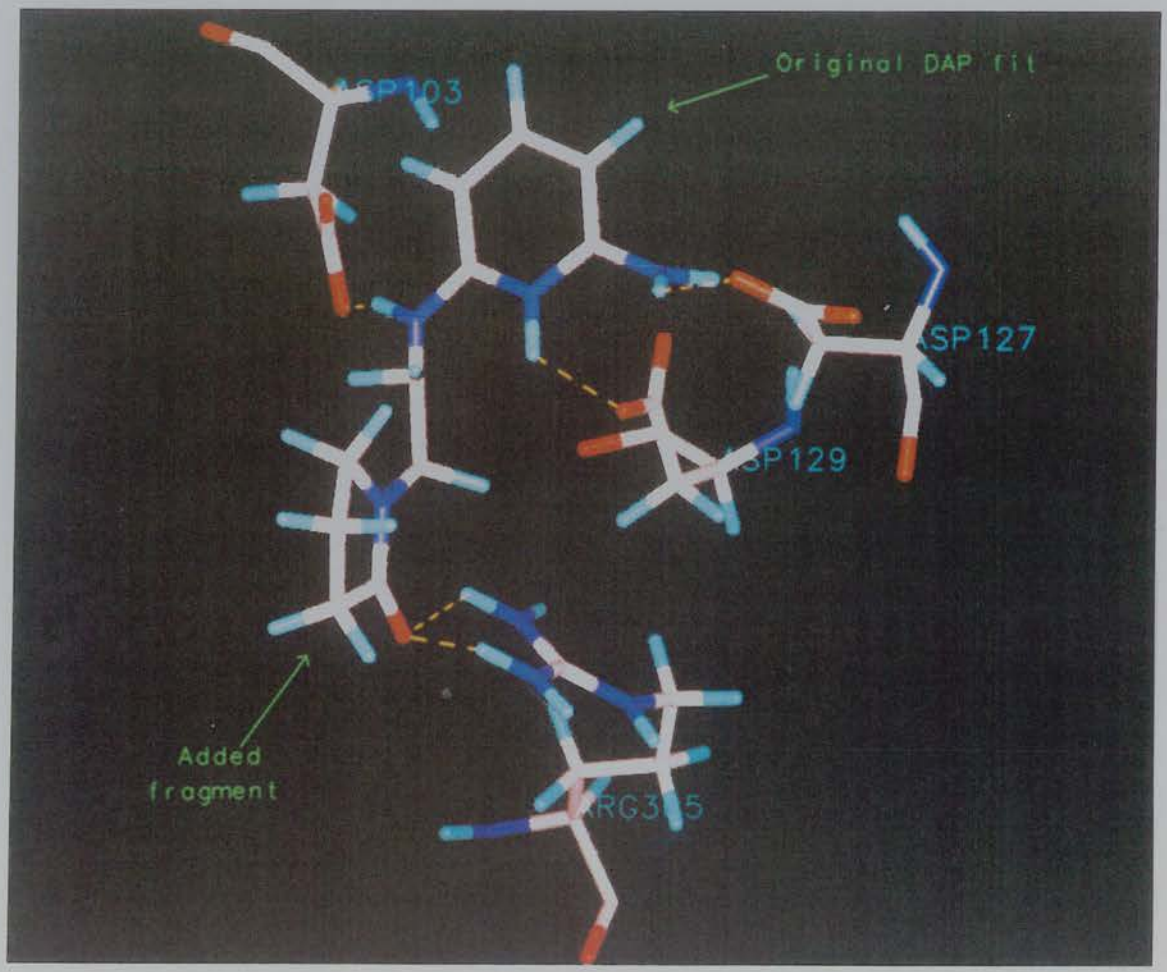


Fig 3.12 DAP at the FBP binding site and linked fragment



3.3.4 Naphtho compounds

3.3.4.1 2D similarity search

The compound naphthoquinone-4-sulphonic acid (NQS) (see Fig 3.14) was tested experimentally (Chapter 4) as a result of a resemblance to a LUDI find pointed out by Pr. Perié. The compound proved interesting and was therefore used as input for a 2D similarity search. A representative selection of the results is shown in Fig 3.13 and their names in Table 3.1. The two significant limitations on the number of compounds that could be tested were commercial availability and the likely insolubility of many of the suggestions. Those selected for further testing are shown in Fig 3.14 and will be referred to as HNNS, AHNS and ANQ. ANQ was not suggested by the similarity search, probably because of non-inclusion in the database, but was commercially available. One of these, AHNS, was found to be unstable with solutions discolouring over days even in the dark at 4°C. No further testing of this compound was carried out. The three remaining compounds were of such variety as to enable the testing for importance of features of the NQS molecule. Thus, ANQ has the quinone grouping of NQS but not the sulphonic acid group. Similarly, HNNS retains the sulphonic acid group of NQS but is not a quinone.

3.3.4.2 Molecular modelling

After the naphtho compounds were found to affect PFK, the attempt was made to find potential binding sites for them on the PFK structure. Attention was focussed on the F6P binding site by the cooperative character of some of the effects (see Fig 4.26). Also borne in mind was the apparent order of potency ANQ > NQS > HNNS (Fig 4.26).

Several GRID maps were used in the search for potential binding sites. The common factor between the naphtho compound is the large hydrophobic naphthalene carbon skeleton. The GRID program was run on a region encompassing the F6P binding site

Fig 3.13 Some 3D similarity search results

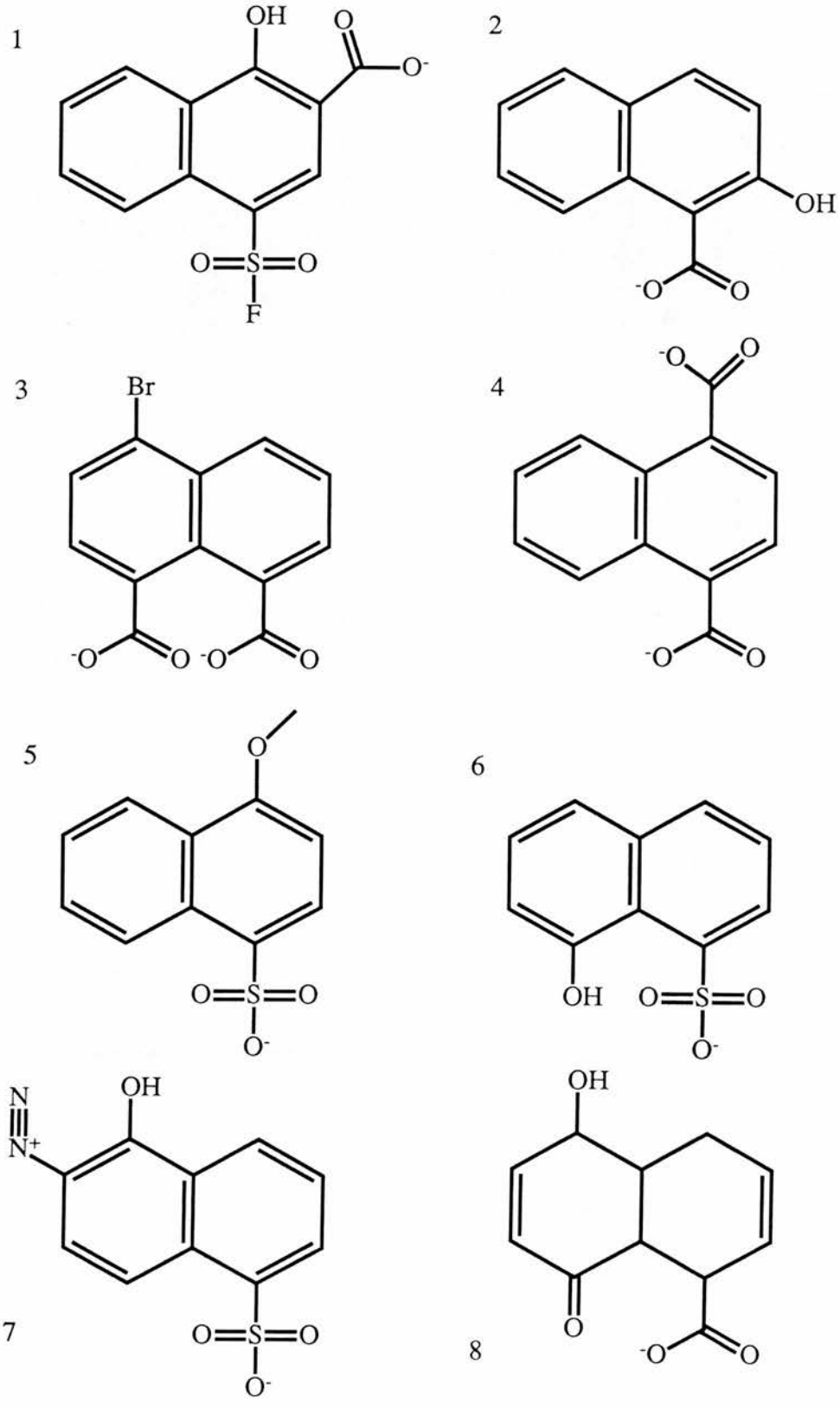
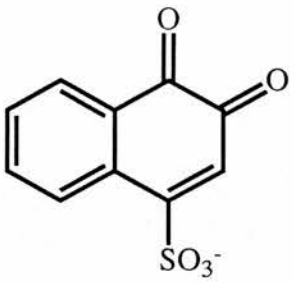


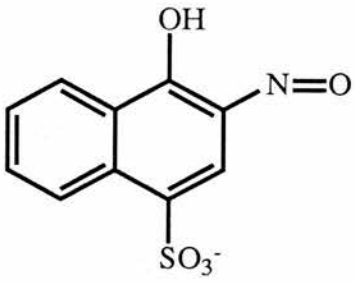
Figure 3.14 Naphtho compounds tested on PFK

NQS



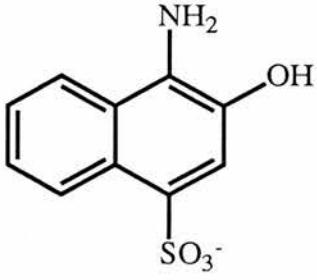
1,2-napthoquinone-4-sulphonic acid

HNNS



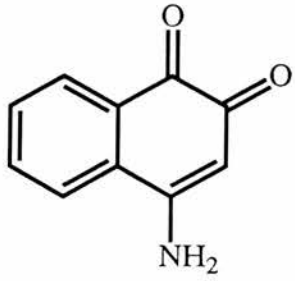
4-hydroxy-3-nitroso-napthalenesulphonic acid

AHNS



3-hydroxy-4-amino napthalenesulphonic acid

ANQ



4-amino-1,2-naphthoquinone

Table 3.1 Some 3D similarity search results

Number in Fig 3.13	Name of compound
1	4-fluorosulphonyl-1-hydroxy-2-naphthoic acid
2	2-hydroxy-1-naphthoic acid
3	4-bromo-1,8-naphthalic acid
4	1,4-naphthalenedicarboxylic acid
5	4-methoxy-1-naphthalenesulphonic acid
6	1-naphthol-8-sulphonic acid
7	2-diazo-1-naphthol-5-sulphonic acid
8	hexahydro-5-hydroxy-8-oxo-1-naphthalenecarboxylic acid

with an aromatic CH probe. Contouring at -4.5, -4, -3.5, -3, -2.5 kcal/mol produces a large relatively favourable region due to a hydrophobic protein patch near Met69 and Ile126. The naphthalene skeleton can be fitted into this region so that most of its carbon atoms are within the favourable contours (Fig 3.15).

The two carbonyl oxygens of NQS were considered next. GRID output for a carbonyl oxygen probe was contoured at -4.5, -4, -3.5 and -3 kcal/mol. With only a small movement of the naphthalene skeleton both NQS oxygen atoms could be placed in favourable contours. The deeper of these contours is due to the nearness of the Ile126 backbone NH to which a carbonyl oxygen could hydrogen bond. In this position the sulphonic acid group of NQS overlaps the end of the Arg171 side chain. The several Arg side chains at the active site are known to be very flexible, both from molecular dynamics simulations and from their disorder in crystal structures. An alteration in Arg171 side chain position would enable a salt bridge to be formed with the NQS sulphonic acid group and so this model seemed to suggest a plausible NQS binding mode. However, this conformation is not consistent with the observed order of potency ANQ > NQS since the sulphonic acid of NQS would be favoured over the amino group of ANQ in the position near Arg171.

An alternative conformation was obtained by rotating the naphthalene skeleton 180° around its long axis. In this position, only one of the carbonyl oxygens of NQS can be placed in the favourable contours while retaining most of the carbon in the CH contours. The other is very close to Arg171, but this is very flexible.

The new positions of NQS sulphonic acid and ANQ amino groups were compared with GRID maps of sulphur double-bonded oxygen (contoured at -5.5, -5, -4.5, -4, -3.5 kcal/mol) and a $:NH_2$ group (contoured at -12, -10.5, -9, -7.5 kcal/mol). Interestingly, patches defined by -9 kcal/mol $:NH_2$ and -5 kcal/mol O= contours were found to coincide between Gly11 and Ile126. By a small movement of ANQ, a compromise position partially satisfying O(=C), $:NH_2$ and CH contours was found. For this conformation, only minor clashes could be seen between the ANQ surface and a solvent-accessible surface calculated for the protein in the vicinity. This conformation

Fig 3.15 Naphthalene skeleton fitted to GRID aryl CH contours

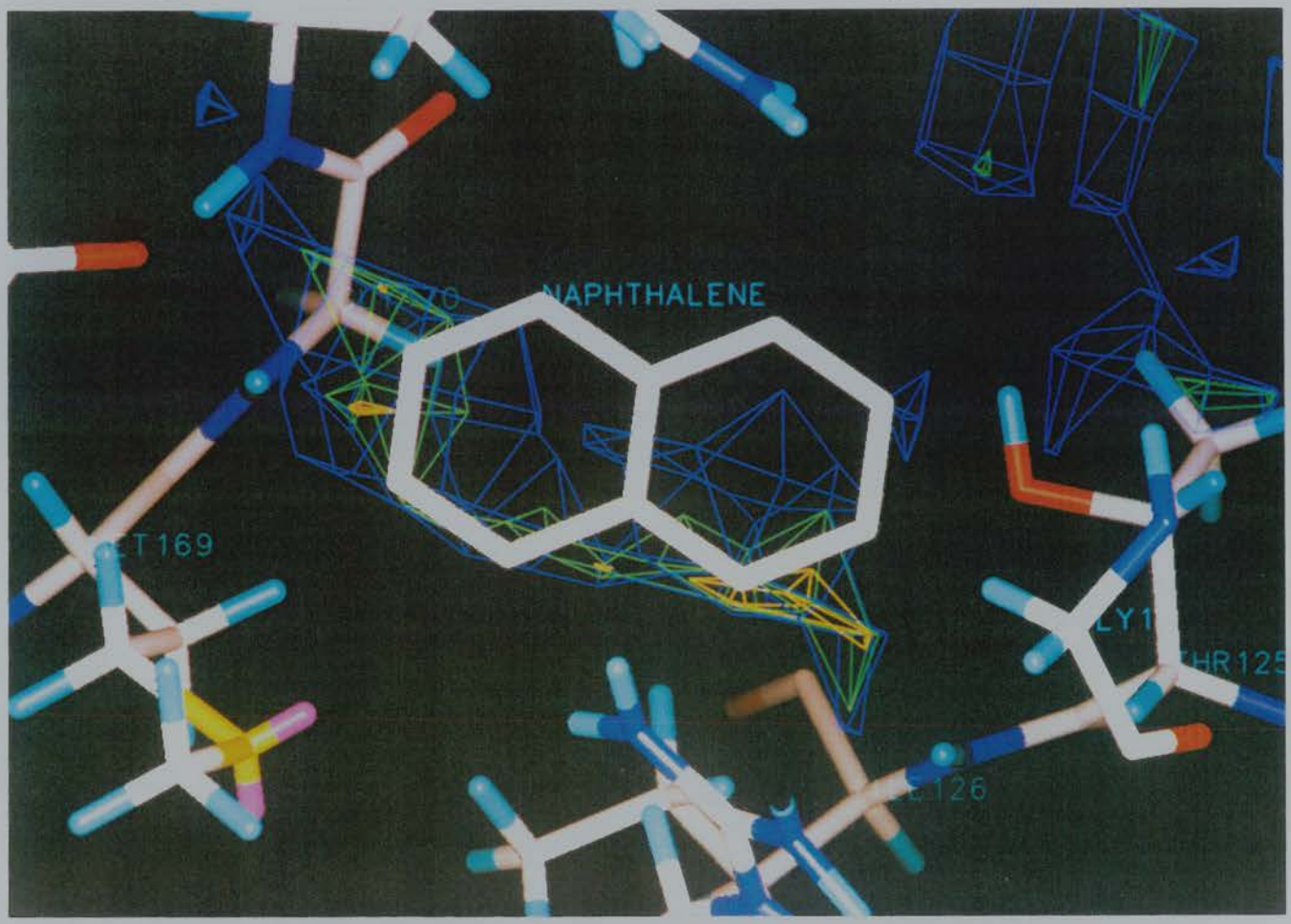
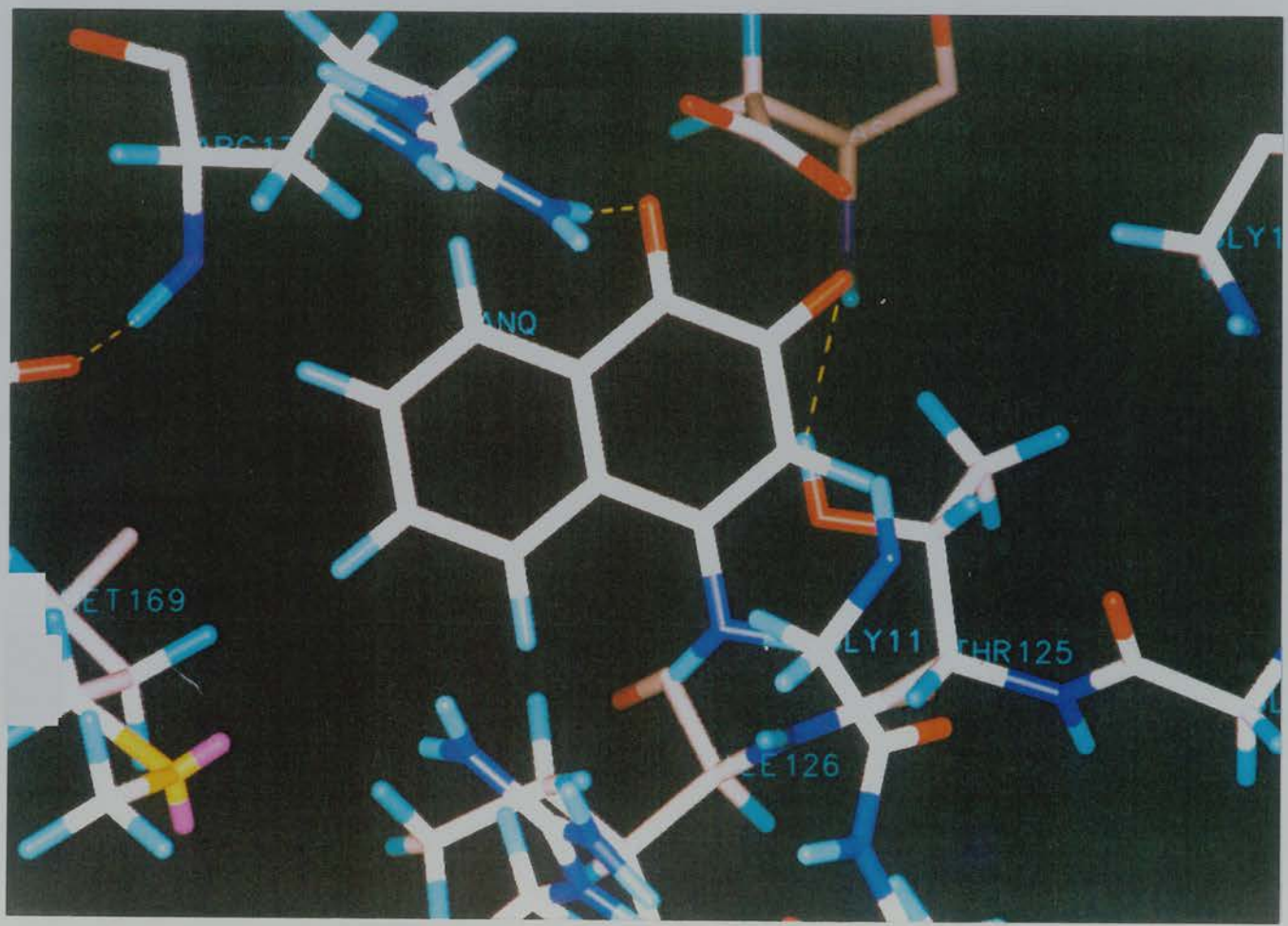


Fig 3.16 Final ANQ fit to FBP binding site (see text)



enables the H-bonding of both carbonyl oxygens to the protein at Thr125 and Arg171 (Fig 3.16). This conformation was also consistent with the order ANQ > NQS since, while there was room for the larger SO_3^- group, none of the three oxygens was even in the -3.5 kcal/mol contours.

The consistency of this position with the order NQS > HNNS was considered next. The two carbonyl oxygens were replaced with hydroxyl and nitroso groups. The GRID hydroxyl contours (at -8, -7, -6 and -5 kcal/mol) encompassed the HNNS OH in the -5 kcal/mol region. The GRID program does not have a nitroso oxygen (O=N) probe and so both O=C and O=S were used. However, no amount of rotation of the nitroso group could place the oxygen in favourable contours for either of these, despite lack of steric problems.

Thus a plausible binding position for the naphtho compounds at the F6P binding site has been found. Semi-quantitative considerations can reasonably explain the potency order ANQ > NQS > HNNS.

Interestingly, the sulphonylfluoride compound (molecule 1 in Fig 3.13) when placed in this conformation has its sulphur atom only 3.2Å from the side chain oxygen of Thr125. The hydroxyl group of this molecule is also favourably placed to H-bond to Arg171. This suggests at least the possibility that this molecule might covalently react with PFK. Reaction at the F6P binding site would of course permanently destroy activity.

3.4 Conclusions

A variety of methods have been applied to the same problem - the design of potential inhibitors of *E.coli* and RM PFKs. The variety of established methods, many unavailable to test here, is a reflection of the intense effort focussed on the field of rational drug design. The different methods produced results of varying quality (see Chapter 4). The relationship between results and method employed will be discussed

in Chapter 5.

Both database searches carried out were relatively rapid, each taking under an hour. The main limitation with each was the size of the database used. This was about 50,000 for the Fine Chemical Directory and about 90,000 (with many duplications) for the Cambridge Crystallographic Database (CCD). The CCD search looked for molecules having a relatively small motif that might fit into the P6 pocket. This, and the size of the database, made it unlikely that high affinity compounds would be found by this method. Nonetheless phosphoethanolamine and phospho-D-serine particularly were worthy of further investigation. The Fine Chemical Directory search produced many ideas. Here the most significant limitations were commercial availability and, to a lesser extent, water solubility. Unfortunately, one of the remaining three compounds appeared to degrade in solution with time. However, the other two were significantly and usefully different to the search molecule. After the observation of activity for these compounds GRID was used to attempt to locate a likely binding site. A plausible site was found which was consistent with characteristics of their activity.

Three donated bisphosphates were modelled. These were designed as potential aldolase inhibitors rather than with PFK in mind. They more resemble FBP, which binds PFKs with low affinity, than substrate F6P. Given this, and the fact that their differences to FBP seemed unfavourable in terms of PFK binding, the likelihoods of their binding with high affinity seemed small.

The designed FBP analogs are a different matter. The fact that they share a large fraction of molecular structure with known PFK ligands suggests that they are highly likely to bind to the enzyme. The fact that generally only small changes were made to the known crystal FBP complex structure also has an important consequence for the validity of the modelling studies. This, and the fact that the modifications that were made were positioned in the light of GRID results, suggests that, after even the relatively short 1.6ps molecular dynamics simulation, an approximation to the global minimum should have been reached. The results themselves were generally in good agreement with GRID results. Thus ligand substituents were found in favourable

GRID contours. Significant exceptions were found in sterically tight positions so that the position adopted by the amino group of $CH_2NH_3^+$ in these places did not coincide perfectly with NH_3^+ contours. However these results overall generally further emphasise the high quality of GRID results. The validity of the notion of reducing the charge on the 'P1' part of the analogs in order to reduce repulsion from Asp127 was amply supported by the results with uncharged analogs particularly favoured. The results also suggest that substitution of ring hydroxyls with primary amino groups would produce tighter binding. Incidentally, this is precisely the change made by von Itzstein *et al.* (1993) in their fruitful work on influenza sialidase inhibitors. The great limitation of these FBP analogs has been their entailing specialised synthesis. Indeed, unfortunately, none of these compounds has been made and so suggestions of their usefulness remain essentially speculative. This was also the slowest design method used; the molecular mechanics procedure alone took several hours for each modelled compound. A particularly good candidate for any later synthetic work would be the sulphonyl chloride (X=-, Y=Cl). This would be readily transformed into a variety of potentially interesting compounds including the sulphonic acid and the sulphonamide.

The LUDI method of drug design had intermediate time requirements; each run typically took under an hour. One apparent disadvantage was the program's ignoring of charge on fragments and protein. This led to its suggesting unrealistic H-bonds between groups almost certain to be similarly charged. Systematic user oversight was therefore essential. This method produced the widest variety of ideas, but with the significant limitation that most of the fragments in the database were quite small (fewer than 20 atoms). This is a reflection of the LUDI approach of linking several fragments to make an inhibitor. The small size, however, does impose a limit on the tightness of binding that may be expected, so that fragment testing for inhibition would be expected to have only limited success. This is important since it hinders the sequential approach of finding one fragment with significant affinity and then building further on that. If fragment combinations must be tested experimentally as a first step then the approach also has chemical synthesis as a limiting factor. An important strength of LUDI over

the substrate analog approach is evident when considering possible enzyme selectivity. Significant amino acid differences between *E.coli* and RM PFKs are only found at a distance from the FBP binding site. They are too far away for FBP modifications such as those described in Section 3.3.2.1 to contact. In contrast LUDI suggests fragments to bind to these different amino acids with other linked fragments able to overlap enough of the active site to hinder substrate binding.

Chapter 4

Ligand Testing

4.1 Introduction

4.1.1 How to study ligand interaction

The most fundamental property of a ligand-receptor system is its binding affinity. Once this has been determined then further studies can be carried out to locate the site of any binding and detailed changes in receptor behaviour arising from it.

Detection of ligand binding usually demands a detectable change in one or more properties of the receptor on binding. From the broad repertoire of detection methods it is likely that only a subset will be appropriate for a particular receptor-ligand system. For example, binding may cause no change in the receptor property or the ligand under investigation may interfere with the method of observation.

Spectroscopic methods are often convenient ways to assess ligand binding. For example, fluorescence and CD spectra are often sensitive to ligand binding. These methods depend on suitably positioned aromatic residues (usually Trp) or prosthetic groups. Nuclear magnetic resonance gives much more detailed information but requires sophisticated instrumentation. All spectroscopic measurements are susceptible to interference by ligand absorption. This may impose an upper limit on the concentration of ligand tested. This can be particularly undesirable during early stages of ligand screening when trying to detect relatively weak binding.

It is usually possible to perturb the quaternary structure (dissociation of subunits) or tertiary and secondary structure (denaturation) using heat or chemicals such as urea or GdnHCl. Ligand interaction often stablises the receptor against such insults thus providing an indirect method to detect binding. The association state of the receptor can be determined by, for example, gel filtration or intra-molecular cross-linking followed by quantitation on SDS-PAGE gels. Receptor denaturation is monitored by loss of activity or changes in fluorescence or optical activity. Protection against chemical modification or proteolysis may also be conferred by ligands and is another way to

assess their binding.

The most direct method of measuring ligand interaction is equilibrium dialysis. This will detect binding which produces no change in receptor properties, but is limited by its requirement for radioactive ligand. Radioactive ligands are more expensive and/or more difficult to prepare than non-radioactive ligands. In addition the method is only applicable to tight binding since the free ligand concentration must be changed by the protein.

4.1.2 Detection methods applied to *E.coli* PFK

4.1.2.1 Kinetics

Enzyme assays for PFK reaction in both directions have been standardised. F16BP synthesis is assayed by conversion to glyceraldehyde-3-phosphate by aldolase and triose phosphate isomerase. This is converted to glycerol-3-phosphate by glycerol-3-phosphate dehydrogenase with the concomitant oxidation of NADH monitored at 340 nm. Creatine kinase and creatine phosphate are included to convert the ADP produced back to ATP (Kotlarz and Buc, 1982). In the reverse direction F6P is acted upon by phosphoglucose isomerase and glucose-6-phosphate dehydrogenase. The latter step is coupled to reduction of $NADP^+$ to NADPH again monitored at 340 nm. Glycerol and glycerol kinase are included to regenerate ADP from ATP (Hellinga and Evans, 1987).

4.1.2.2 Fluorescence

The *E.coli* PFK tetramer contains one Trp residue (Trp 311) per subunit. These tryptophans are situated at the subunit interface within each dimer of the dimer-of-dimers structure (see Section 1.3.1.1). They can be selectively excited by radiation at 295 nm and have a fluorescence emission peak around 337 nm. The magnitude of the fluorescence is sensitive to the binding of ligands to both catalytic and regulatory sites (Berger and Evans, 1991; Deville-Bonne and Garel, 1992). The fluorescence behaviour of *E.coli* PFK provides much information about the nature of its allosteric

behaviour and has already been discussed in Section 1.3.1.2.

4.1.2.3 Protection against denaturation

The dissociation and denaturation of *E.coli* PFK have been studied in some detail. Agents able to induce these processes are urea (Le Bras *et al.*, 1989; Teschner *et al.*, 1990a), GdnHCl (Teschner and Garel, 1989), thiocyanate (Deville-Bonne *et al.*, 1989), high hydrostatic pressure (Deville-Bonne and Else, 1991) and heat (Teschner *et al.*, 1990b). Dissocation and denaturation can then be observed in several ways.

In the absence of ligands the pathway followed by *E.coli* PFK on exposure to denaturant is

$$N <=> 2\ D <=> 4\ M <=> 4\ U$$

where N represents the native tetramer and D, M and U the dimeric, monomeric and unfolded species. As previously stated, the tetramer crystal structure shows two kinds of inter-dimer interfaces. The smaller one of around 1200 $Å^2$ contains the active site (the A interface) and the larger one of 1800 $Å^2$ the regulatory site. With increasing concentrations of denaturant, the A interface dissociates first with simultaneous loss of enzyme activity (Le Bras *et al.*, 1989; Deville-Bonne *et al.*, 1989). The R interface contains the Trp residues so that its later dissociation can be monitored by a large change in fluorescence as the Trp becomes solvent-exposed. At still higher denaturant concentrations the enzyme unfolds. This secondary structure loss can be monitored using ellipticity at 222 nm.

F6P strongly protects the tetramer against dissociation into dimers (Deville-Bonne *et al.*, 1989; Teschner *et al.*, 1990a; Deville-Bonne and Else, 1991). This is consistent with its situation in the crystal structures between subunits. In the presence of saturating F6P the A interface becomes at least as stable as the R interface. No dimers can then be detected during the denaturation with the enzyme displaying all-or-nothing dissociation.

The regulatory sites of the dimeric species can still bind PEP but have lost their ability, seen in the tetrameric species, to bind GDP. PEP, but not GDP, is seen to stabilise the dimeric form against further dissociation (Deville-Bonne *et al.*, 1989).

ATP is seen to cause little or no protection against denaturation (Deville-Bonne and Else, 1989; Le Bras *et al.*, 1989). This is presumably because ATP binds to residues within a single subunit, rather than binding between two subunits as F6P does (see Section 1.3.1.1)

4.1.2.4 Other methods

The regulatory ligands of *E.coli* PFK, GDP and PEP, have been found to protect the enzyme against subtilisin proteolysis whereas F6P gave little protection (Le Bras and Garel, 1982). The proteolysed form showed no regulation by GDP or PEP. In the light of the crystal structures these results are readily explained. The C-terminal residue contributes to the regulatory site so that its loss abolishes regulation. Similarly, ligand binding to the native enzyme stabilises the C-terminal tail reducing its susceptibility to proteolysis.

Recently, important information has been obtained using equilibrium dialysis with radioactive substrates (Deville-Bonne and Garel, 1992). Surprisingly, only binding of two F6P molecules per tetramer could be detected whereas the expected four ATP binding per tetramer was seen. The source of this 'half-of-sites' reactivity is unknown but may be related to visible signs of asymmetry in the dimer-of-dimers crystal structure. For example, in the PFK.ADP.FBP ternary complex two of the active sites are in an 'open' conformation and two in a different 'closed' conformation.

4.1.3 Detection Methods applied to Rabbit Muscle PFK

RMPFK seems to yield much less information to standard ligand binding studies.

Using tryptophan fluorescence (excitation at 294nm and emission at 335nm) two nucleotide binding sites were identified, one associated with an increase in fluorescence

and the other a decrease (Pettigrew and Frieden, 1979). By the responses of different nucleotides the former was identified with the active site and the latter with the inhibitory site. F6P had only a very small effect on fluorescence.

Fluorescence also showed that the environments of tryptophan residues altered on formation of RMPFK tetramer (Cai *et al.*, 1990). CD was also used in this study and showed no significant changes during formation of RMPFK tetramer. CD also showed that no large changes in secondary or tertiary structure were produced on binding of ligands to RMPFK.

The binding of ATP to RMPFK and its dependence on concentration of inhibitors citrate, PEP and phosphoglycerate has been investigated by equilibrium dialysis (Colombo *et al.*, 1975).

These results, taken together, suggest that only equilibrium dialysis, which requires radioactive ligands, or kinetic studies are useful for studying ligand interaction with RMPFK. RMPFK activity can be monitored using the same coupled enzyme systems as for *E.coli* PFK (see Section 4.1.2.1).

4.2 Materials and Methods

4.2.1 Materials

4.2.1.1 Enzymes

The coupling enzymes aldolase, triose phosphate isomerase, glyceraldehyde-3-phosphate dehydrogenase, phosphoglucose isomerase, glucose-6-phosphate dehydrogenase and glycerol kinase were all from Sigma Chemical Co. (Poole U.K.). Rabbit Muscle phosphofructokinase was also from Sigma as an ammonium sulphate

suspension. Creatine kinase was from Boehringer Mannheim (Lewes, U.K.).

4.2.1.2 Chemicals

Tryptone and yeast extract were from Oxoid Ltd. (Haverhill, U.K.) or Difco Labs. (East Molesly, U.K.). Agar was from Oxoid.

Sodium chloride, calcium chloride, potassium chloride, urea, EDTA and Tris were all from Sigma as were ampicillin, IPTG, Meldola blue and MTT. F6P, FBP, NADH and NAD^+ were also from Sigma. GdnHCl was Sigma Grade 1. Phospho-D-serine, Phospho-L-serine, phosphoethanolamine, phosphocholine and phosphocreatine were also from Sigma. Also from Sigma were the proteolysis inhibitors 1,10-phenanthroline, dichloroisocoumarin and E64.

Glycerol and phosphoric acid were from Aldrich Chemical Co. (Poole, U.K.) as were the LUDI compounds shown in Table 4.4.

Sodium hydroxide and sodium azide were from BDH (Poole, U.K.)

4.2.1.3 Bacteria and Plasmids

The *E.coli* PFK expression plasmid pHL1 (Lau *et al.*, 1987) was the kind gift of Dr. P.R. Evans (University of Cambridge) as was the plasmid carrying the K213A mutant sequence. Dr Evans also generously supplied a culture of the *E.coli* PFK strain DF1020 which lacks a genomic copy of the *E.coli* PFK gene. This was used for preparation of mutant PFK.

4.2.1.4 Hardware

Spectrophotometry was carried out on a Unicam 8700 Series UV/Visible spectrophotometer. Fluorescence measurements were made on a Perkin-Elmer LS-3B fluorescence spectometer. A Beckman J2-21 centrifuge with JA20 rotor was used for cell harvesting and during generation of competent cells. Ultracentrifugation was using a Beckman TL100 machine with TL100.3 rotor. A Pharmacia FLPC system was used for analytical gel filtration.

4.2.2 Methods

4.2.2.1 Bacterial Transformation

Competent cells were prepared by the method of Cohen *et al.* (1977). 0.5 ml of an overnight culture of *E.coli* was innoculated into 50 ml of 2xTY medium (16 g l^{-1} tryptone, 10 g l^{-1} yeast extract and 5 g l^{-1} NaCl) and grown until the OD600 was in the range 0.3-0.5. The cells were recovered by centrifugation in a JA20 rotor at 5000 r.p.m. for 15 mins at 4°C. The pellet was resupended in 10 ml of ice-cold 0.1M $CaCl_2$ and stored on ice for 30 mins. The cells were recovered as before by centrifugation.

For transformation, three 200 µl aliquots were taken into microfuge tubes. 2 µl of pHL1 miniprep were added to tube 1. For a DNA control, 1µl of a 10 ng μl^{-1} preparation of pUC18 was added to tube 2. Nothing was added to tube 3 which served as a blank control. The bacteria were then subjected to a heat shock of 42°C for 90 seconds before chilling on ice for 2 minutes. After the addition of 200 µl 2xTY medium, the bacteria were allowed to recover for 30 minutes at 37°C. 200 µl from each tube was then transferred to a 1.5% agar plate made from 1xTY and including 100 µg ml^{-1} ampicillin. After overnight incubation at 37°C, colonies were seen on plates 1 and 2. A stock preparation of TG1 transformed with pHL1 was made by transferring a single colony, by toothpick, to 10 ml of 2xTY containing 100 µg ml^{-1} ampicillin. This was grown overnight in an orbital shaker at 37°C before storage at -20°C as a 15% glycerol solution.

4.2.2.2 Preparation of *E.coli* PFK

100 µl of the TG1/pHL1 glycerol stock was innoculated into 10 ml of 2xTY containing 100 µg ml^{-1} ampicillin and grown overnight at 37°C with shaking. Three 250 ml flasks each containing 150 ml 2xTY with 100 µg ml^{-1} ampicillin and IPTG at 70 µg ml^{-1} were prepared and 1.5 ml of the overnight culture innoculated into each. The cells were grown for about 24 hours at 37°C with shaking and harvested by

centrifugation at 3000 r.p.m. in a JA20 rotor at 4°C for 20 minutes.

The cells were resuspended in 10 ml of low-salt buffer (LSB) comprising 50mM Tris buffer pH 7.9, 1mM EDTA and 10mM DTT. Protease inhibitors were then added to the following concentrations - 100 μM dichloroisocoumarin, 100 μM 1,10-phenanthroline and 20 μM E64. The cells were then sonicated on ice for three 30 second bursts with two minutes between each burst. The resultant solution was spun for 20 minutes at 100,000 r.p.m. and 4°C in an ultracentrifuge. The supernatant was applied to a 2 ml Cibacron blue 3GA column previously equilibrated with LSB. After washes with 2 ml LSB and 6 ml LSB + 1.5M NaCl, PFK was eluted with 9 ml of LSB + 2mM ATP and 10mM $MgCl_2$. The column was cleaned with 6 ml of 8M urea + 0.5M NaOH and stored in LSB + 0.02% sodium azide.

Fractions containing PFK were purified further by gel filtration using a Superose S200 column. The buffer used was 50mM Tris, 4mM EDTA, 2mM DTT at pH 7.9. PFK eluted in a peak at 10 ml and was stored at 4°C.

4.2.2.3 Enzyme Assays

'Forward Assay'

The assay contained 10mM $MgCl_2$, 10mM NH_4Cl, 40mM KCl and 10mM creatine phosphate in 33mM Tris acetate buffer at pH 8. Around 0.4 μg NADH from a solution made fresh each day was present in each assay. Coupling enzymes were present in the following amounts :- 0.6 U aldolase, 3 U glycerol-3-phosphate dehydrogenase, 26 U triose phosphate isomerase, 4 U creatine kinase. ATP and F6P were added from stock solutions of 100mM and 20mM respectively. The total assay volume was 1 ml. Reaction was started by addition of PFK and the reaction rate measured as the rate of decrease of A_{340} in a spectrophotometer connected to a 25°C circulating water bath. When acidic or basic compounds were being tested, the buffer concentration was increased to around 260mM Tris.

Because of the limited quantities of epoxy compounds available, F6P response curves were determined for single reactions. The concentration of F6P in the cuvette

was sequentially increased by addition of small aliquots from a 20mM solution. After each addition the reaction rate was allowed to stabilise for two minutes before rate measurement over the next minute. Calculations showed that at the reaction rates produced, levels of reactants were not significantly depleted during the course of the experiment.

'Backward Assay'

10mM $MgCl_2$, 10mM NH_4Cl, 40mM KCl and around 0.5 μM $NADP^+$ were present in the 30mM Tris acetate buffer which also contained 10% glycerol. The buffer concentration was again increased when assaying acidic or basic potential inhibitors. Coupling enzymes were present in the following amounts :- 2 U phosphoglucose isomerase, 0.5 U glucose-6-phosphate dehydrogenase, 0.4 U glycerokinase. ADP and F16BP were added from stock solutions of 50mM and 100mM respectively. The total assay volume was 1 ml. The reaction was started by the addition of PFK and measured as the rate of increase of A_{340} at 25°C.

Some of the compounds to be tested absorbed strongly at 340 nm severely limiting the concentration that could be tested. In these cases a modified assay was used. A dye mix was prepared from 250 μl of a frozen stock solution of 1 mg ml^{-1} Meldola Blue and 6 mg MTT made up to 1 ml with water. 50 μl of this mix was included in each assay enabling the rate to be monitored at 540 nm.

Rabbit Muscle phosphofructokinase

Before enzyme assay, RMPFK was dialysed overnight against one litre of 33mM Tris acetate buffer pH8 with 1mM EDTA and 1mM DTT. Particulate matter was removed by filtration through a 0.2μm filter. The preparation was found to be unstable and so was used within the next day. Assay conditions were exactly as for *E.coli* PFK.

4.2.2.4 Fluorescence

Intrinsic *E.coli* PFK fluorescence was measured as the emission at 337 nm on excitation at 295 nm. PFK was present at around 30 μg per experiment, in the region in which PFK fluorescence depends linearly on concentration (Berger and Evans, 1991). The buffer used was 33mM Tris acetate pH8 or 495 mM Tris acetate pH8 for acidic and basic compounds. Aliquots of concentrated test compounds were added and the change in fluorescence recorded at 25°C. Correction was made for changes in volume. When ADP fluorescence response was determined 10mM $MgCl_2$ and 10mM KCl were included.

4.2.2.5 Protein Assay

This was by a modification of the method of Bradford (1976). Dye reagent was prepared by dissolving 20mg Coomassie Blue G-250 in 20ml 85% phosphoric acid and 10 ml 95% ethanol then adding 160 ml water and 10ml 1M NaOH. Protein samples of up to 10 μl and up to 2μg were mixed in cuvettes with 1ml of the reagent. After 5 minutes incubation, A_{590} was measured and the protein concentration determined by comparison with a standard curve constructed using bovine serum albumin.

4.2.2.6 Denaturation by GdnHCl

Duplicate incubations were set up over a range of GdnHCl concentrations. Each Eppendorf tube contained 4 μg PFK in a volume of around 10 μl plus 20 μl of 495mM tris acetate buffer pH 8. The volume was made up to 100 μl with 8M GdnHCl and water to give the desired concentration. For binding testing some of the water was substituted with a solution of the compound to give concentrations that were saturating for known ligands or as high as possible for putative ligands. The tubes were then incubated at room temperature, originally overnight and later for 4 hours. 1mM DTT was included in the later assays. The residual activity was measured using the forward assay or, in the case of F16BP protection assays, the backward assay.

4.2.2.7 Thermal Denaturation

Thermal denaturation of *E.coli* PFK was carried out in the absence or the presence of various concentrations of ligand simultaneously using the cell transport facility of the spectrophotometers. Within each set of experiments the amount of PFK was identical at 20-30 μg. The volume of PFK solution was made up to 1 ml with 33mM Tris acetate buffer pH 8 and/or ligand solution in quartz cuvettes. The cuvettes were then transferred to a spectrophotometer prewarmed by circulating water to 60°C. In later experiments ligand solution was prewarmed to 60°C before addition of PFK.

The absorbance at 280 nm was monitored and for each incubation a maximum rate of change estimated from the trace. In some cases maximum change seen was informative. Both quantities were expressed as a % of the control containing no ligand.

4.2.2.8 Analytical Gel Filtration

The association state of *E.coli* PFK was monitored by FPLC. A Superose 12 column was used and a 50mM Tris buffer pH 7.9 containing 1mM EDTA and 2mM DTT and various concentrations of PDS used throughout. Detection was at 280nm. Peaks corresponding to PFK tetramer and monomer were seen at elution volumes of 9.8ml and 14.8ml respectively.

4.2.2.9 Curve Fitting

Fluorescence and kinetic data were fitted to either the hyperbolic (Michaelis-Menten) equation, which for fluorescence is :-

$$F=F_{max}S/(S+K_s)$$

or the sigmoidal (Hill) equation :-

$$F=F_{max}S^n/(S^n + S_{0.5}{}^n).$$

Here, F is the observed fluorescence change at concentration S of ligand. The calculated parameters are F_{max}, the fluorescence change at saturating ligand, K_s, the ligand's dissociation constant, $S_{0.5}$, the concentration of ligand required for half-maximal fluorescence change and n, the Hill coefficient (sometimes called n_H). The

program Fig.P for Windows Version 1.1 (Fig.P Software Corporation, Durham, North Carolina) was used. This calculates a standard error (S.E.) for each parameter. Values are cited as mean ± S.E.

4.3 Results and Discussion

4.3.1 Developments of enzyme assay

4.3.1.1 Effect of solvents on assay

Many of the potential ligands to be tested were found to be poorly soluble in water. The effect of more hydrophobic solvents on the enzyme assays was then determined to see if the compounds could be added to the assay from stock solutions in these solvents.

Figure 4.1 shows the effects of ethanol, methanol, DMF and DMSO on measured *E.coli* PFK rate in the standard 'forward' assay. Methanol and ethanol show similar behaviour with an immediate decline in rate observed at 5% solvent and further decreases to less than 20% original rate at 20% solvent. In contrast, the less hydrophobic solvents DMF and DMSO appear not to affect rate at 5% solvent, but produce a decline thereafter.

Figure 4.2 shows the results of a similar experiment using RMPFK. Again, ethanol even at 5% decreases the observed rate, and at 20% solvent stops the reaction entirely. However, as with *E.coli* PFK it seems that 5% DMSO can be present without affecting the observed rate.

In the light of these results, water-insoluble compounds were made up as DMSO solutions and added to the assay at up to 5% DMSO.

4.3.1.2 Monitoring at 540nm with dyes

The effect of addition of 65µl of 3.9mM NADH on the spectrum of a solution of

Fig 4.1 Effect of solvents on E.coli PFK activity

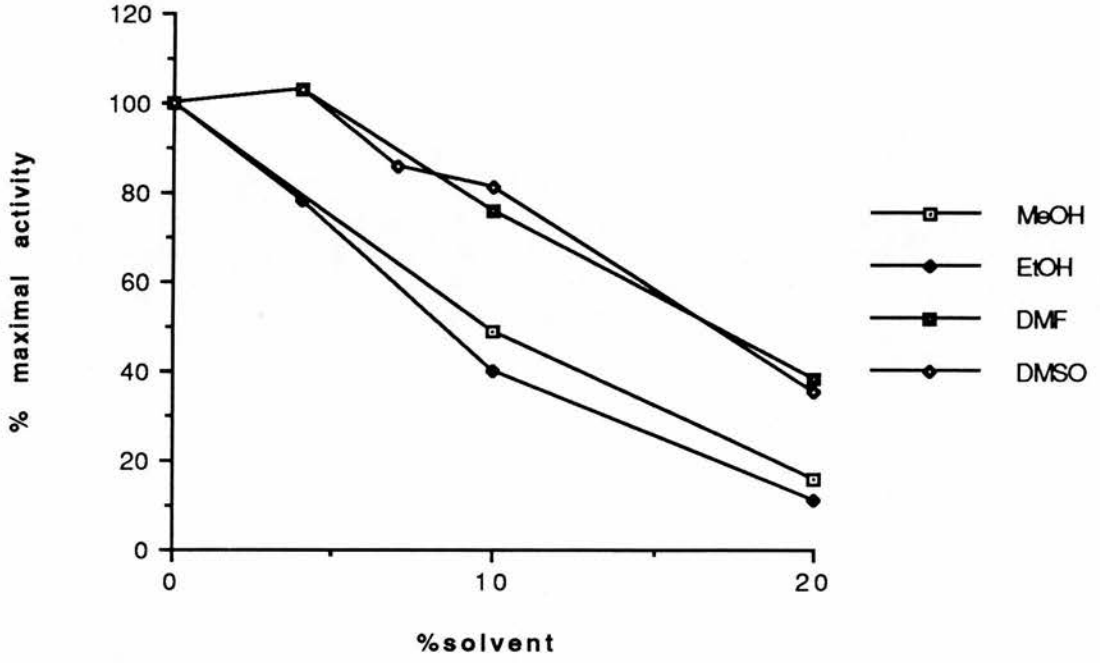
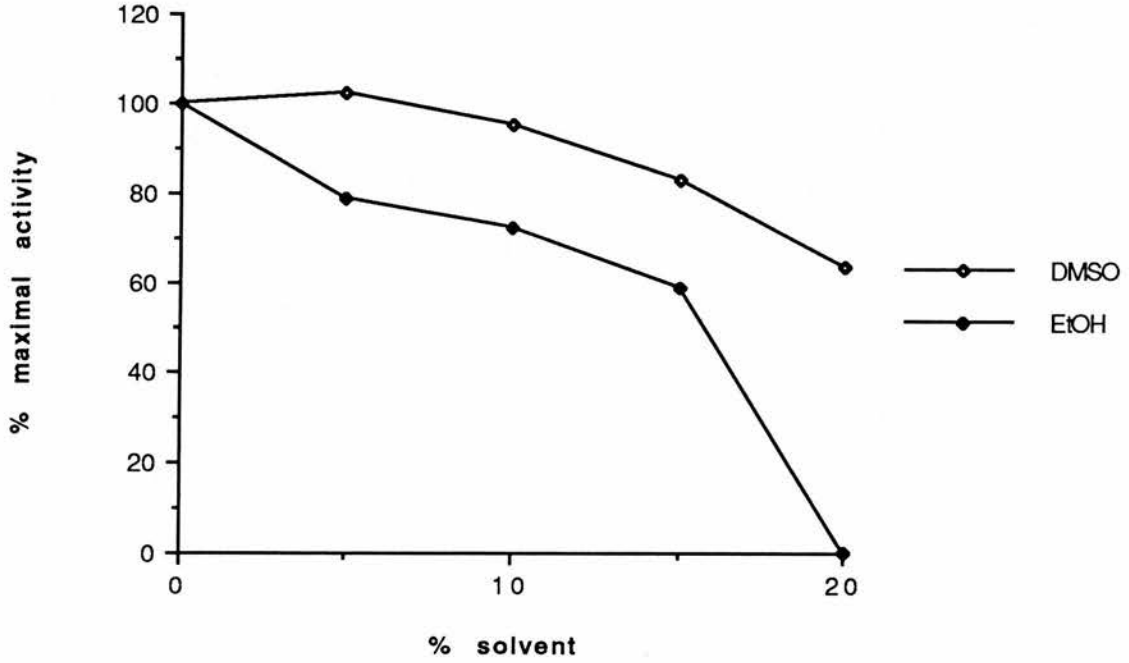


Fig 4.2 Effect of solvents on RMPFK activity



0.00025% Meldola Blue, 0.03% MTT in 33mM Tris acetate buffer pH8 is shown in Fig 4.3. Large increases in absorbance in the region 460nm to 600nm can be seen. The effect of addition of aliquots of the same NADH solution to another dye solution are shown in Fig 4.4. Almost linear changes in absorbance at 540nm and 570nm can be seen as the amount of NADH added increases. The tailing off of absorbance change is only significant after the addition of amounts of NADH in excess of the total NAD^+ present in the standard 'backward' assay.

The rate of change of A540 with respect to enzyme concentration was then determined in a standard 'backward' assay mix containing the dyes. Fig 4.5 shows the linear dependence seen with rate measured between 2 and 4 mins after addition of enzyme. Further experiments showed that DMSO had no effect on the electron transfer from NADH to Meldola Blue and MTT and that these dyes did not affect the rates of PFK or the coupling enzymes of the 'backward' assay.

Taken together these results suggested that the Meldola Blue/MTT system could satisfactorily be used used to test for inhibition by compounds that absorbed strongly at 340nm.

4.3.2 Natural ligands

4.3.2.1 Fluorescence

The intrinsic fluorescence of *E.coli* PFK due to Trp311 is responsive to a variety of natural ligands. Typical traces for F6P, F16BP and ADP are shown in Fig 4.6 with parameters from curve fitting summarised in Table 4.1. Both F6P and F16BP show similar sigmoidicity and maximum fluorescence quench. ADP in contrast gives a purely hyperbolic fluorescence response.

The sigmoid fluorescence response to F6P differs from previous work (Deville-Bonne and Garel, 1992; Berger and Evans, 1991) which found a hyperbolic response. The fluorescence behaviour of PFK does seem to be very sensitive to conditions. For example, Deville-Bonne and Garel report a maximum fluorescence quench of 20% by

Figure 4.3 Effect of NADH on spectrum of dye solution

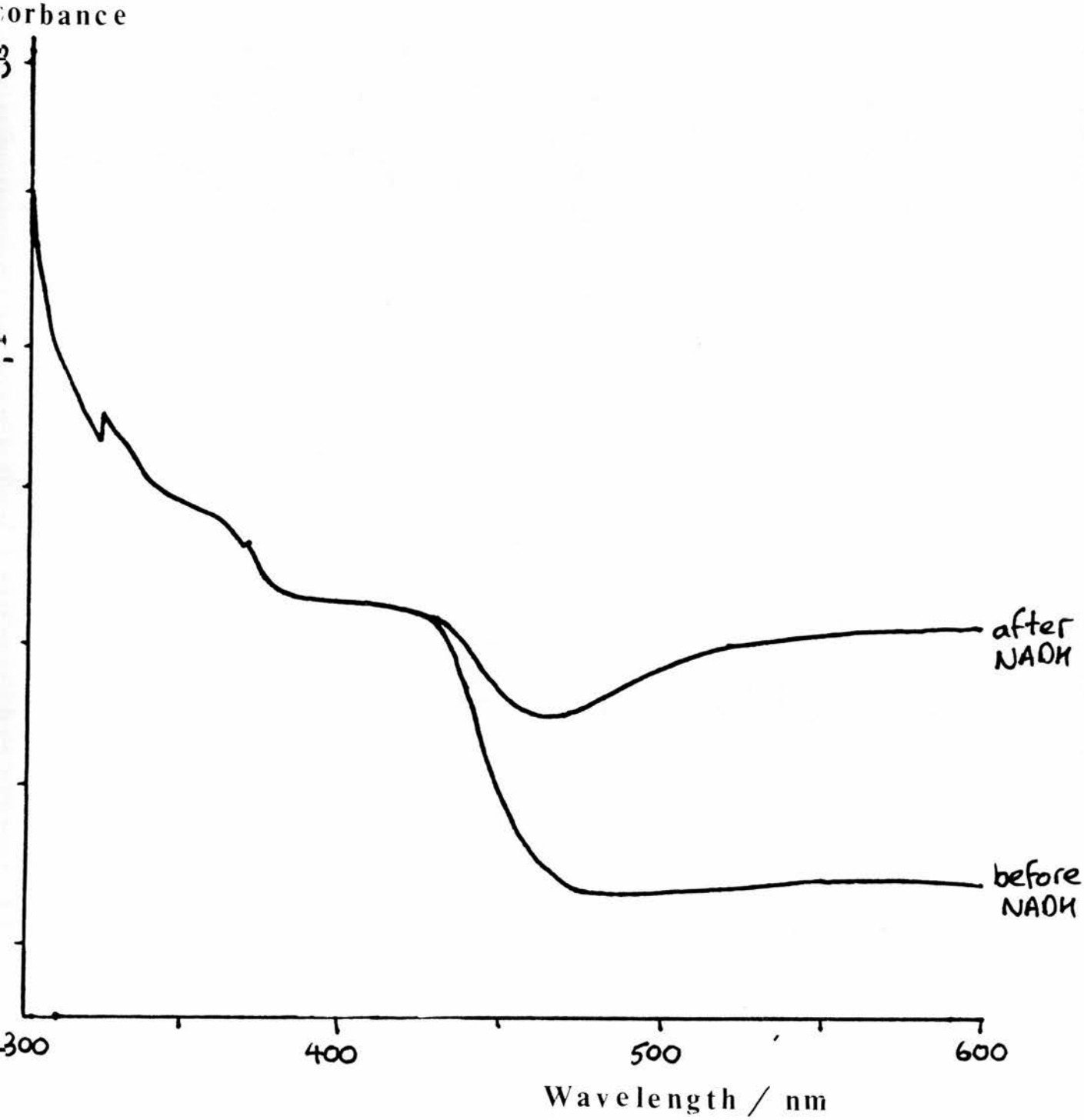


Fig 4.4 Effect of added NADH on dye mix

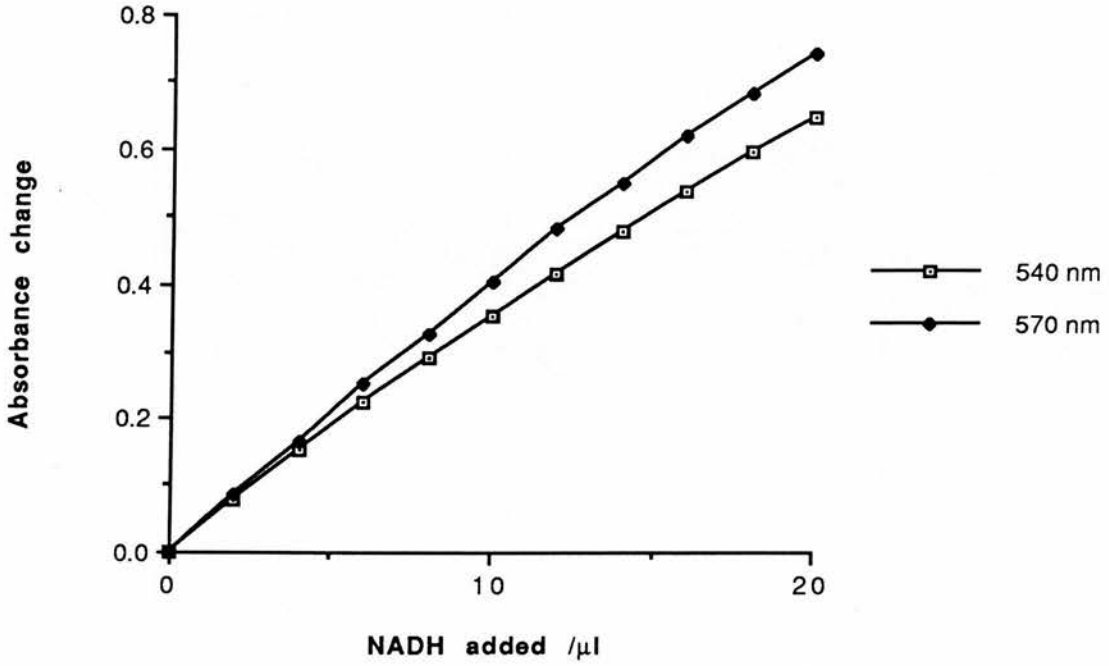


Fig 4.5 Dependence of rate on PFK concentration (Measured at 570nm with dye system)

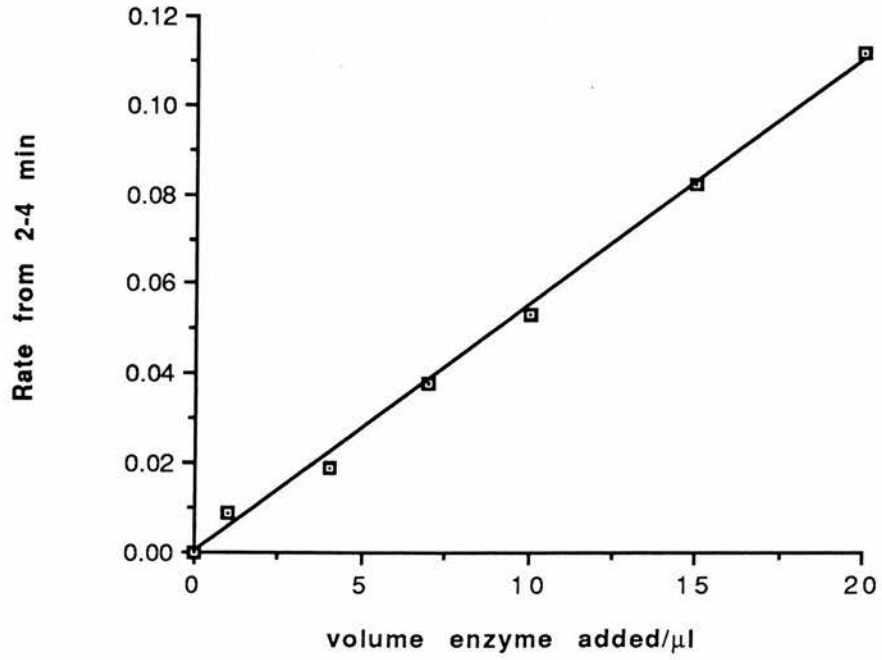


Fig 4.6 Effect of ligands on PFK fluorescence

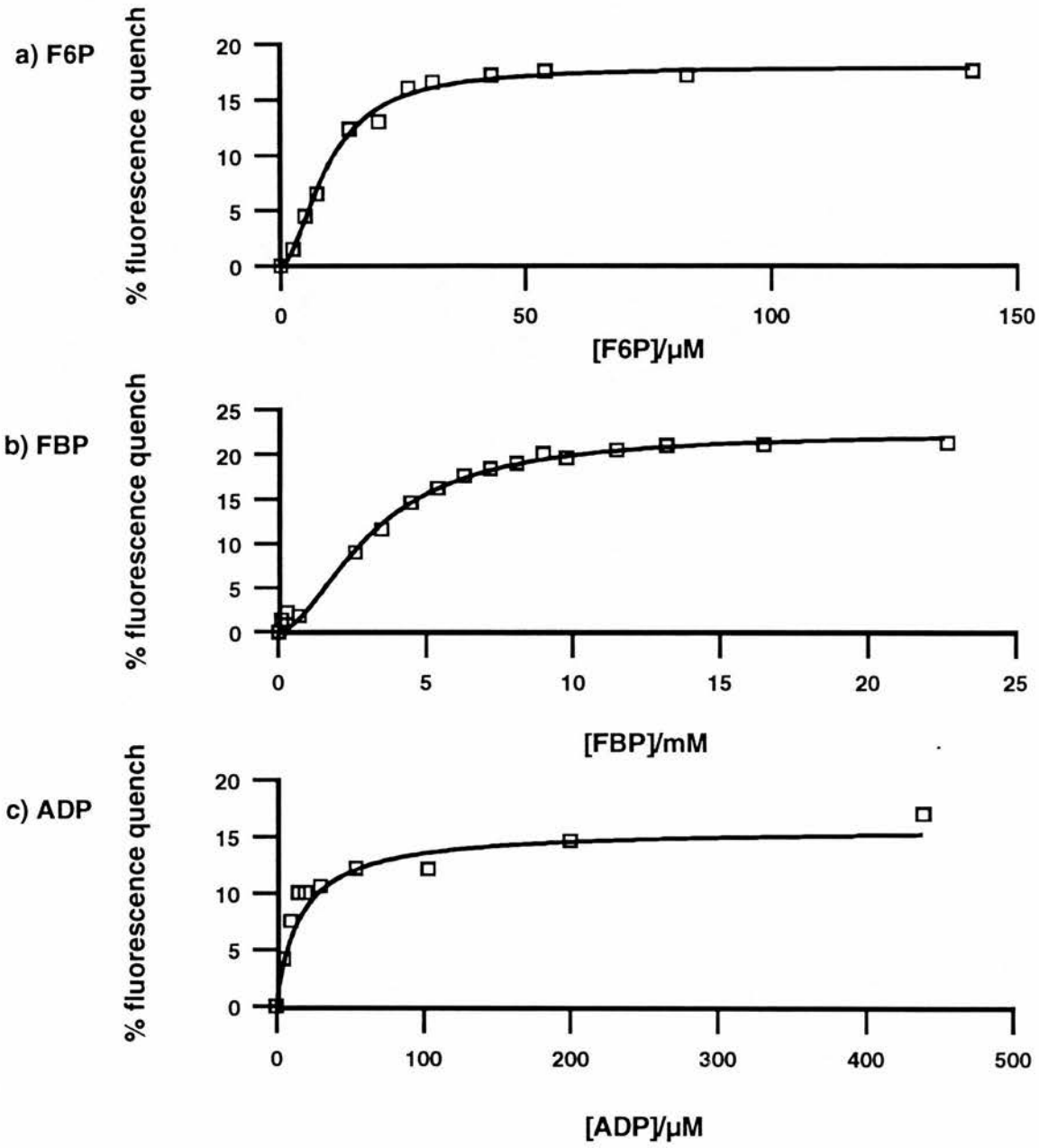


Table 4.1 Fluorescence parameters for natural ligands

	F6P	F16BP	ADP
F_{max} (%)	18±0	24±1	16±1
$S_{0.5}$ /μM	10±1	3200±200	12±2
n_H	1.9±0.2	1.8±0.2	1.0±0.1

F6P whereas Berger and Evans find a 35% quench. However, the finding of a sigmoid fluorescence response in this work could also imply the presence of ATP in the enzyme preparation. This would be surprising after a gel filtration purification step, but the possibility should be borne in mind. Other values in Table 4.1 are in line with previous findings.

4.3.2.2 Thermal denaturation

F6P was found to be able to slow the increase in A_{280} produced by heating PFK at 60°C. Fig 4.7 shows the change in A_{280} with time for PFK solutions in the presence of various concentrations of F6P. Fig 4.8 shows how the maximum rate of change of A_{280} is affected by concentration of F6P. The sigmoid curve fit shown has parameters max inhibition = 103±3%, $S_{0.5}$=180±9μM, n_H=1.9±0.2. Table 4.2 compares parameters for F6P derived from kinetics (Berger and Evans, 1991; no confidence limits given), fluorescence and thermal denaturation studies. The $S_{0.5}$ value from thermal denaturation is much higher than the constants from other studies. This suggests that the PFK species whose binding to F6P slows A_{280} increase may not be the native form, but possibly some intermediate along the denaturation pathway. The fact that sigmoidicity is still observed strongly suggests that this form is still tetrameric and therefore not too dissimilar to the native form. The difference between the n_H value of 4.0 from steady state kinetics and the value of 1.8 for binding alone probably indicates a catalytic as well as a binding component to measured kinetic cooperativity (Berger and Evans, 1991).

ADP and ATP were found to have little effect on PFK denaturation. PEP was mildly destabilising, capable of doubling the rate of change of A_{280} at around 2mM.

4.3.2.3 GdnHCl Denaturation

Both 2mM F6P and 40mM F16BP were able to stabilise *E.coli* PFK against GdnHCl denaturation (Fig 4.9). In the absence of ligand, 50% denaturation of PFK occurs at about 0.4M GdnHCl. F6P is capable of shifting this to 1.7M GdnHCl whereas

Figure 4.7 Effect of F6P on thermal denaturation of *E.coli* PFK

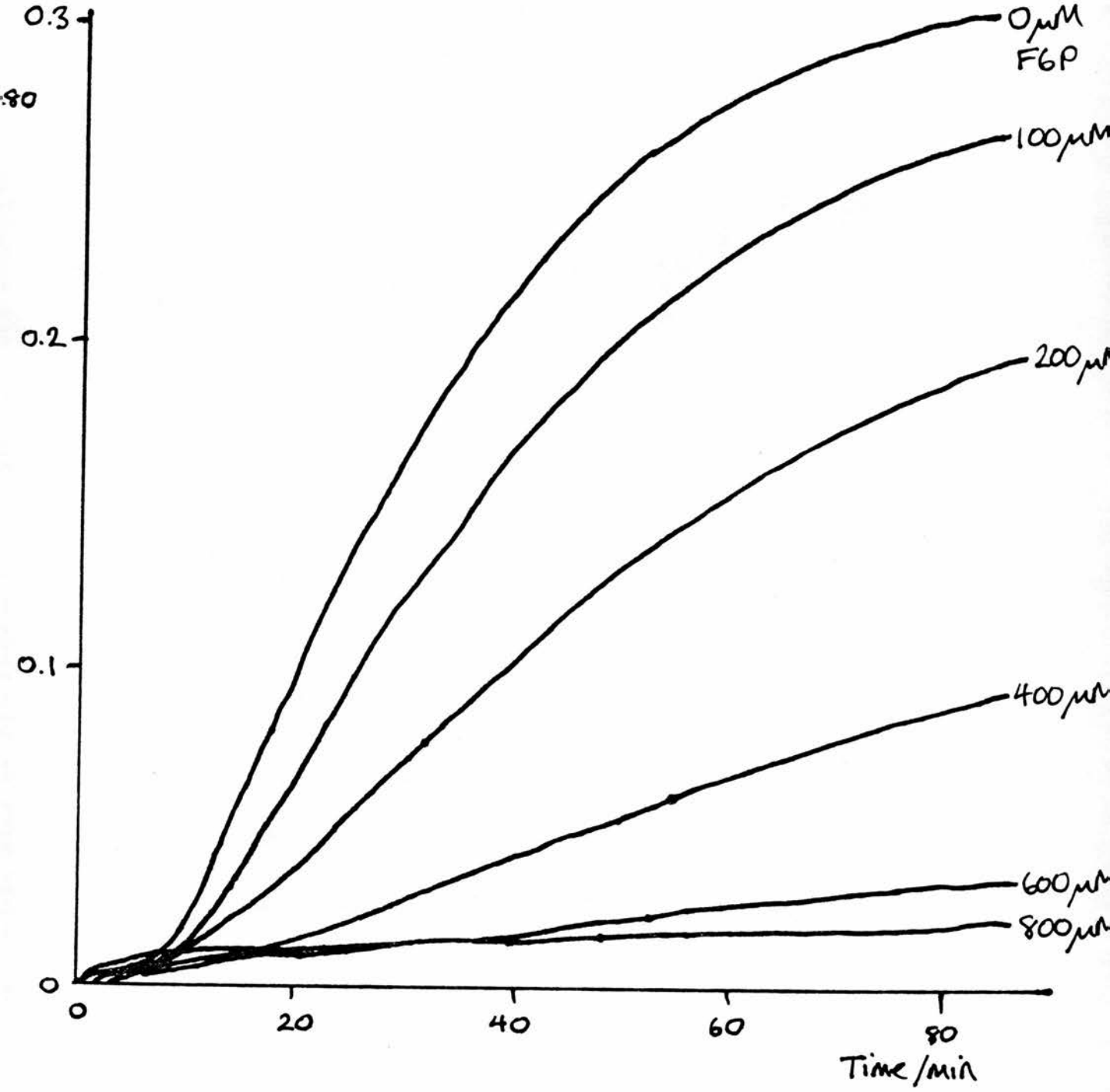


Fig 4.8 Effect of F6P on PFK thermal denaturation

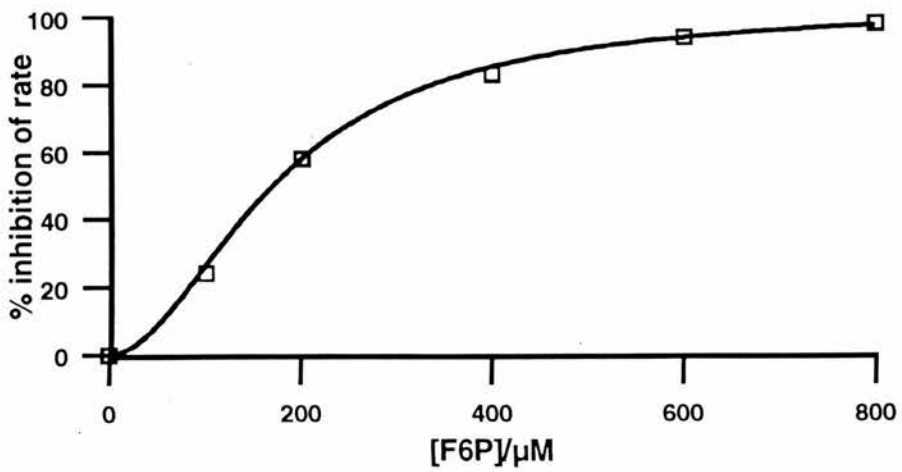


Table 4.2 Comparison of F6P parameters

	Kinetics		Fluorescence	Thermal Denaturation
	no GDP	+ GDP		
'affinity'/μM	$S_{0.5}$=540	K_m=30	K_s=9.7±0.7	$S_{0.5}$=180±9
n_H	4.0	1.0	1.9±0.2	1.9±0.2

Fig 4.9 F6P and FBP affect GdnHCl Denaturation

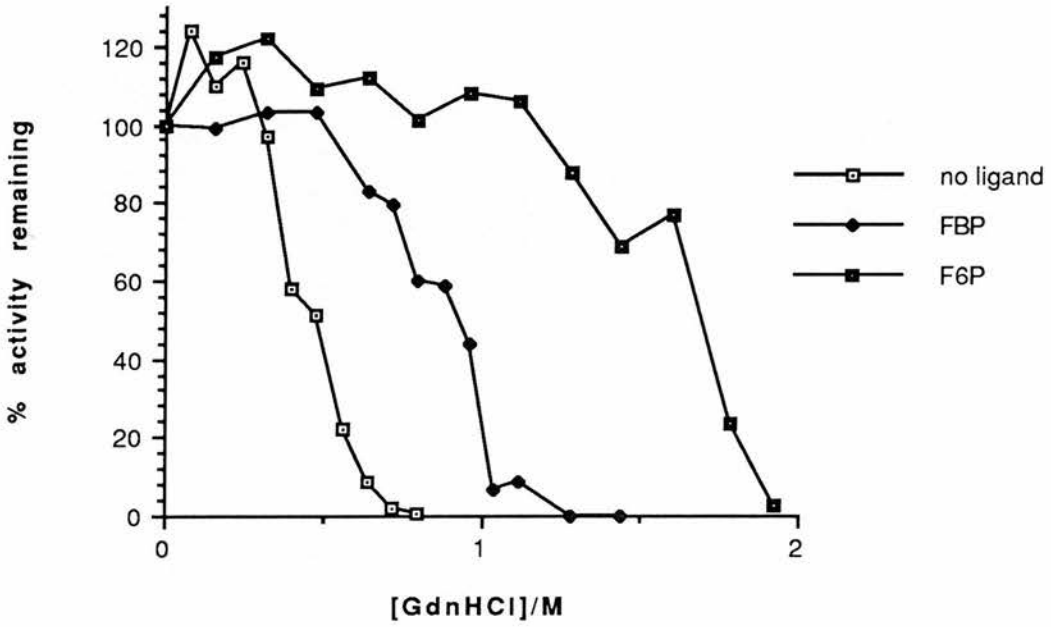
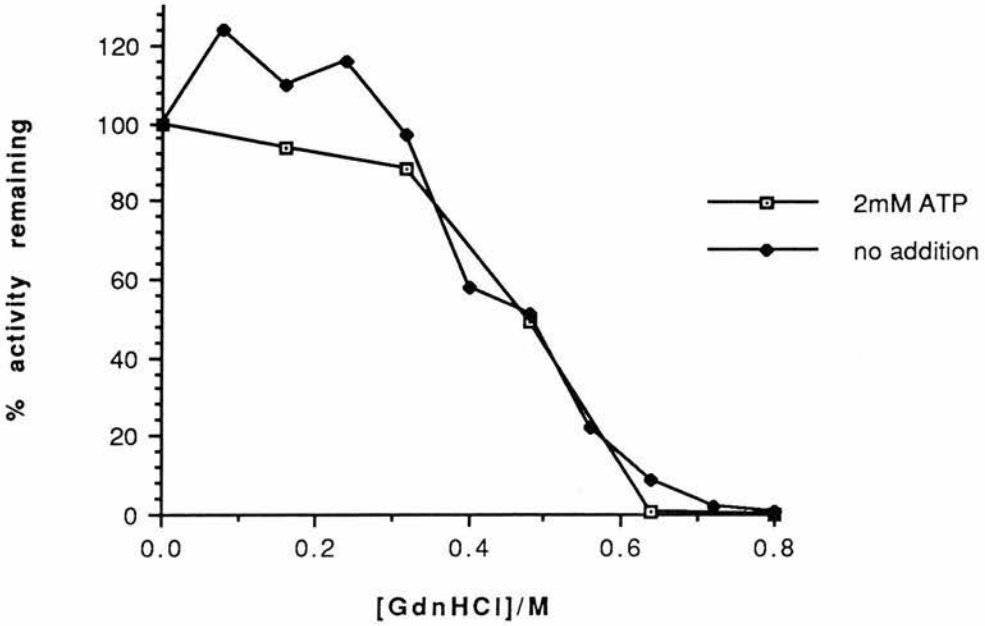


Fig 4.10 Effect of ATP on GdnHCl Denaturation



F16BP shifts the midpoint to about 0.9M GdnHCl. 2mM ATP had no effect on the denaturation (Fig 4.10) and 10mM PEP was found to be destabilising.

RMPFK was found to be very susceptible to GdnHCl. After three hours incubation in the presence of GdnHCl, negligible activity remained even at concentrations as low as 0.16M. In the presence of 2mM F6P a very sharp transition is seen at 0.5-0.6M GdnHCl (Fig 4.11). 2mM concentrations of ADP, ATP, PEP and citrate produced little or no stabilisation.

4.3.3 Small phosphocompounds

4.3.3.1 Enzyme assay

The effect of the small phosphocompounds on PFK enzyme activity proved impossible to determine since they all inhibited at least one coupling enzyme of both 'forward' and 'backward' PFK assays. Both metabolism of added F16BP to glycerol-3-phosphate in the 'forward' assay and metabolism of F6P to 6-phosphogluconic acid in the 'backward' assay were significantly affected by the presence of phosphocholine (PC), phosphoethanolamine (PE), phospho-L-serine (PLS) or phospho-D-serine (PDS).

4.3.3.2 Fluorescence

The effects of PLS and PDS on *E.coli* PFK fluorescence are shown in Fig 4.12 and those of PC and PE in Fig 4.13. These represent typical results from several experiments.

PC produces a maximal fluorescence quench of 9±1% with an approximately hyperbolically-shaped curve. PE gives only a 2±0% fluorescence quench. The two isomers of phosphoserine give indistinguishable fluorescence responses. In both cases the slope of the quench vs. concentration curve increases up to about 23mM phosphoserine before reaching a plateau at about 70% fluorescence quench.

The mutant *E.coli* PFK K213A was also studied in order to try to determine the

Fig 4.11 GdnHCl denaturation of RMPFK

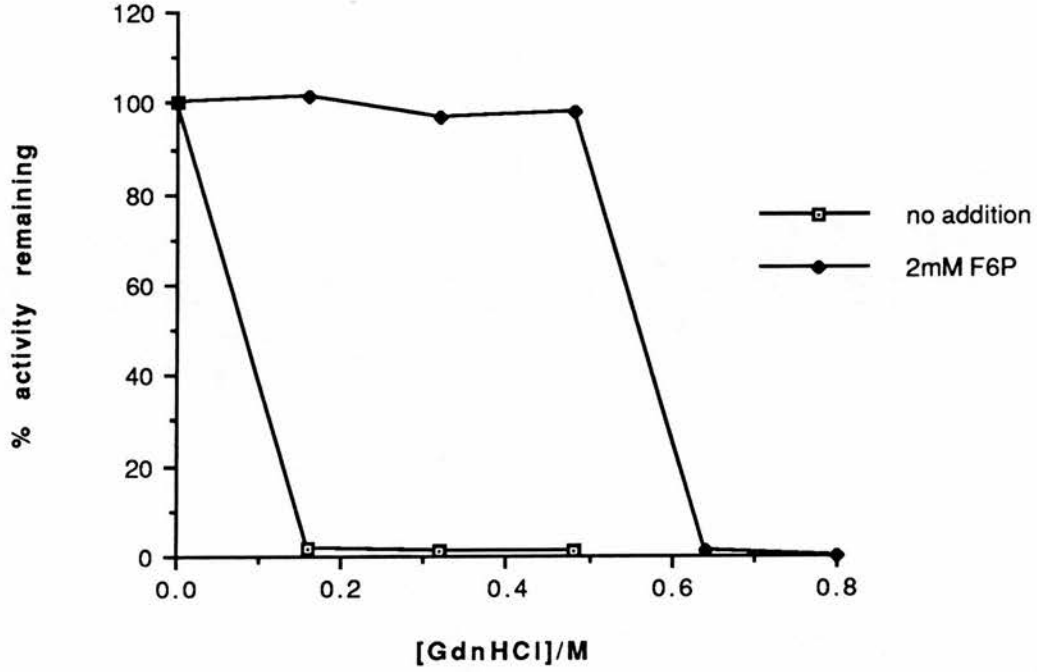


Fig 4.12 Effect of phosphoserines on fluorescence

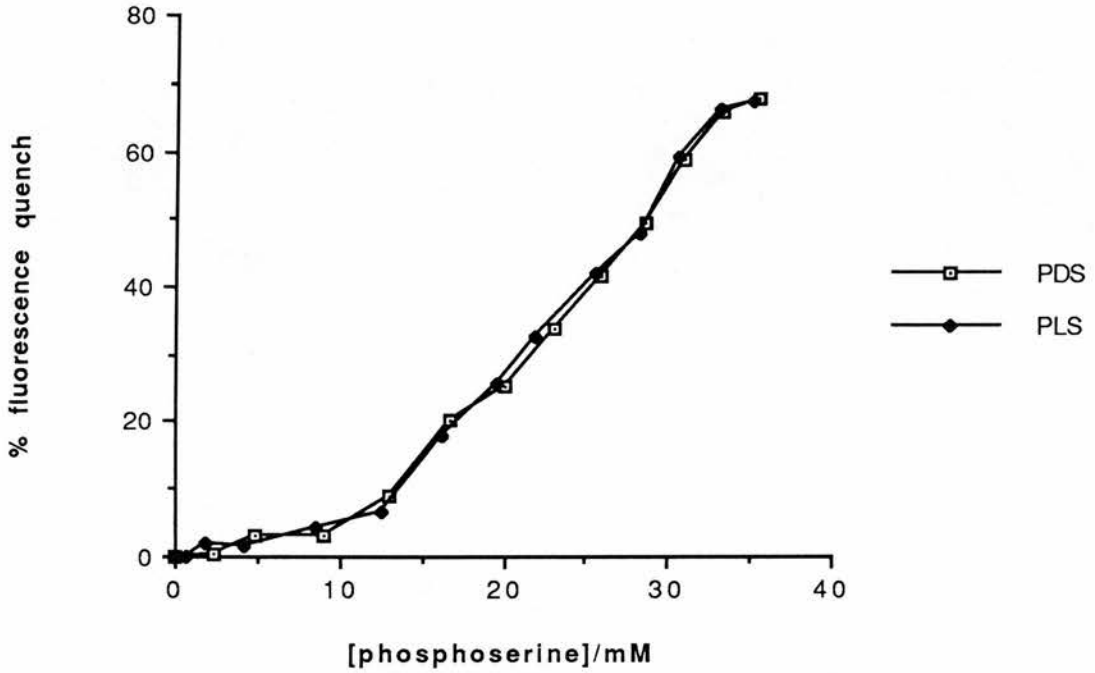
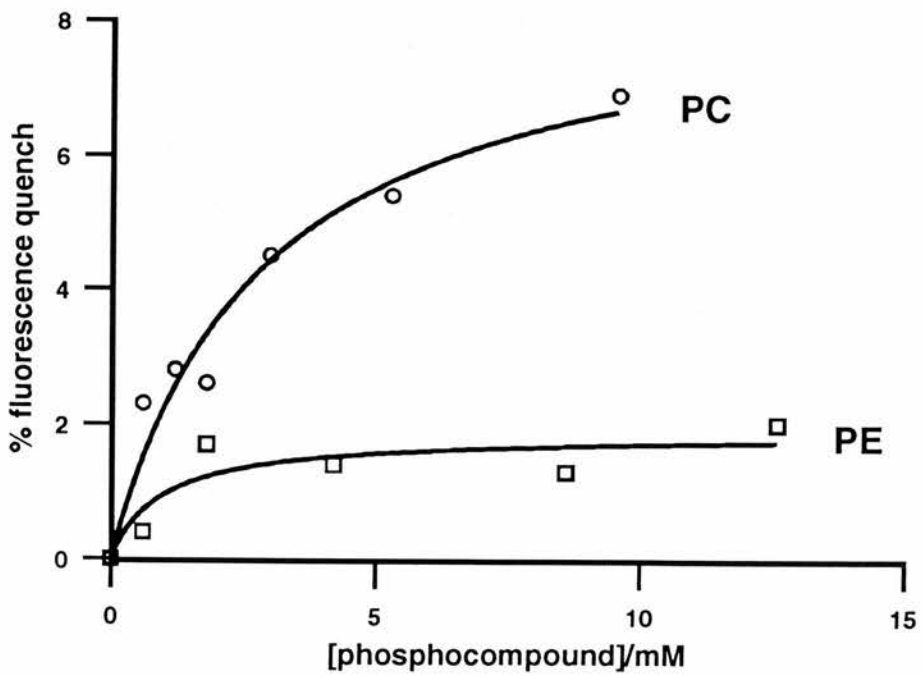


Fig 4.13 Effects of PC and PE on fluorescence



binding site for the small phosphocompounds. This mutant possesses a defective effector site which binds effectors less tightly than the wild-type enzyme. Since the small phosphocompounds under study bore some resemblance to known effectors such as PEP and phosphoglycollate, the effector site was one candidate for the binding site of the phosphoserines and/or PC. The preparation was shown to be the required mutant from its much reduced sensitivity to PEP inhibition compared to the wild-type enzyme. The mutant was also found to show a fluorescence increase in the presence of ADP in contrast to the quench seen for the wild-type PFK. Fig 4.14 shows the fluorescence response of the mutant to the phosphoserines and Fig 4.15 its response to PC and PE. The only change in fluorescence behaviour seen is a raised PC K_S value for the mutant. This is 3±1mM for the wild type enzyme but is increased to 10±2mM for K213A. This suggests that PC may produce its effect on fluorescence by binding to the effector site. Also suggestive is the 9% fluorescence quench seen for saturating PC, matching the 8-10% quench seen for saturating PEP (Deville-Bonne and Garel, 1992) which is known to bind to the effector site. PE and the phosphoserines affect the fluorescence of K213A similarly to the wild-type enzyme suggesting that they do not exert their fluorescence effects by binding to the effector site.

The size of the fluorescence quench produced by the phosphoserines was much larger than that seen for any of the natural ligands. A 70% quench implies a very large change in the environment of the fluorophores, in this cases Trp311. When *E.coli* PFK is dissociated into monomers by 3M urea a fluorescence change of similar magnitude (75%) is observed (Le Bras *et al.*, 1989). This suggested that the phosphoserines were causing a change in the association state of the enzyme.

The observation that saturating F6P shifts the fluorescence quench curve to the right (Fig 4.16) was consistent with this idea since F6P is known to protect *E.coli* PFK against dissociation and denaturation by a variety of agents.

4.3.3.3 Analytical gel filtration

The dependence of *E.coli* PFK association state on PDS concentration was determined

Fig 4.14 Phosphoserines and K213A fluorescence

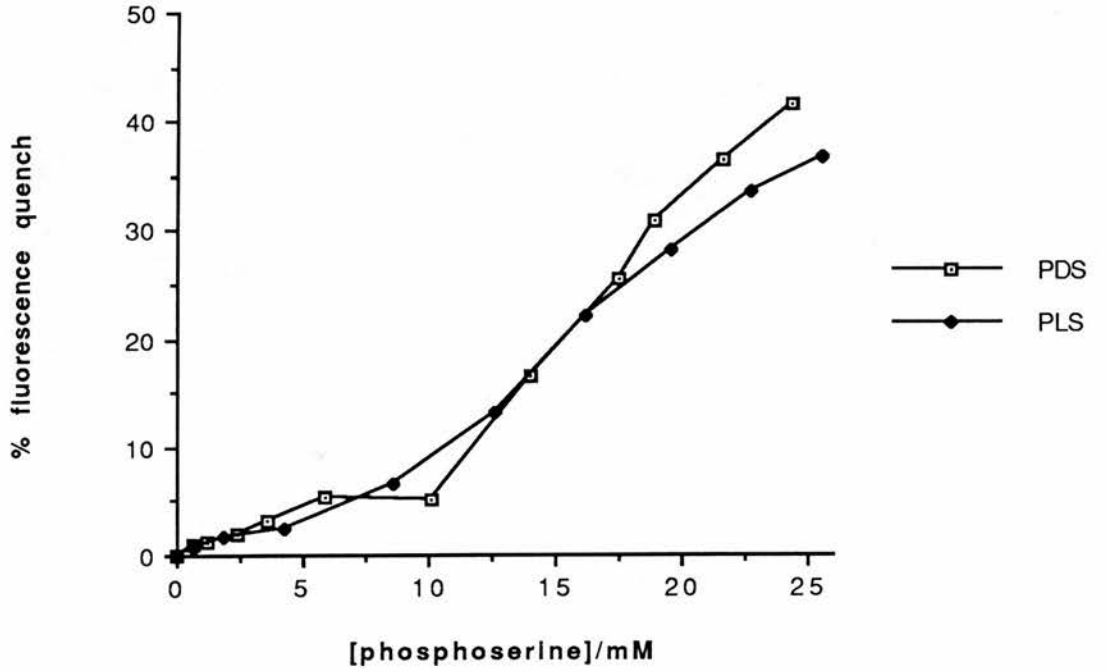


Fig 4.15 Effects of PC and PE on fluorescence of K213A

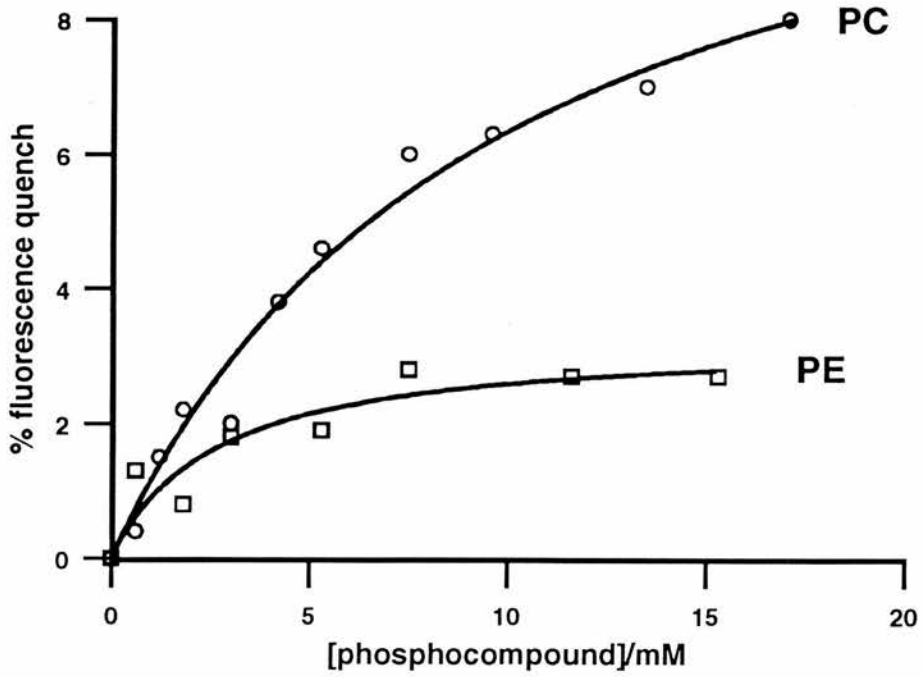


Fig 4.16 Effect of F6P on PDS response

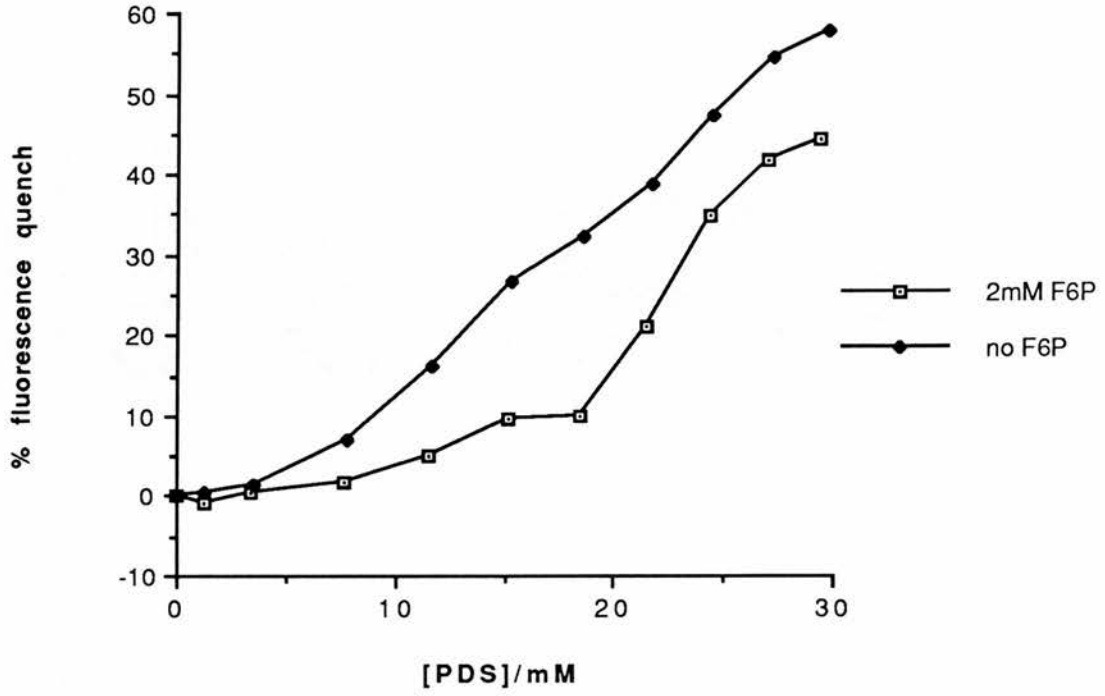
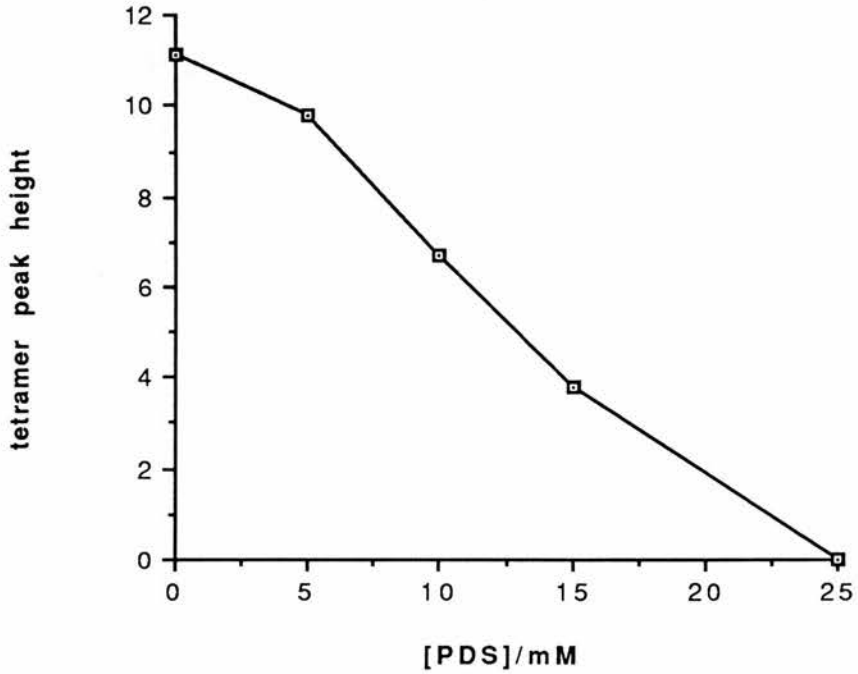


Fig 4.17 PDS and E.coli PFK association state



by analytical gel filtration. The height of the PFK tetramer peak was seen to decrease as the concentration of PDS increased (Fig 4.17). The height of the monomer peak could not be accurately monitored since it was superimposed on another large, unidentified peak.

The negative correlation between fluorescence quench and tetramer peak height confirms that the phosphoserines do indeed effect dissociation of *E.coli* PFK.

4.3.3.4 Thermal denaturation

All of the small phosphocompounds affected the thermal denaturation of *E.coli* PFK. Both the effect on rate of increase in A_{280} and the maximum change in A_{280} were affected. Fig 4.18a shows the effects on rate and Fig 4.18b those on maximum change, both relative to controls containing no small phosphocompound.

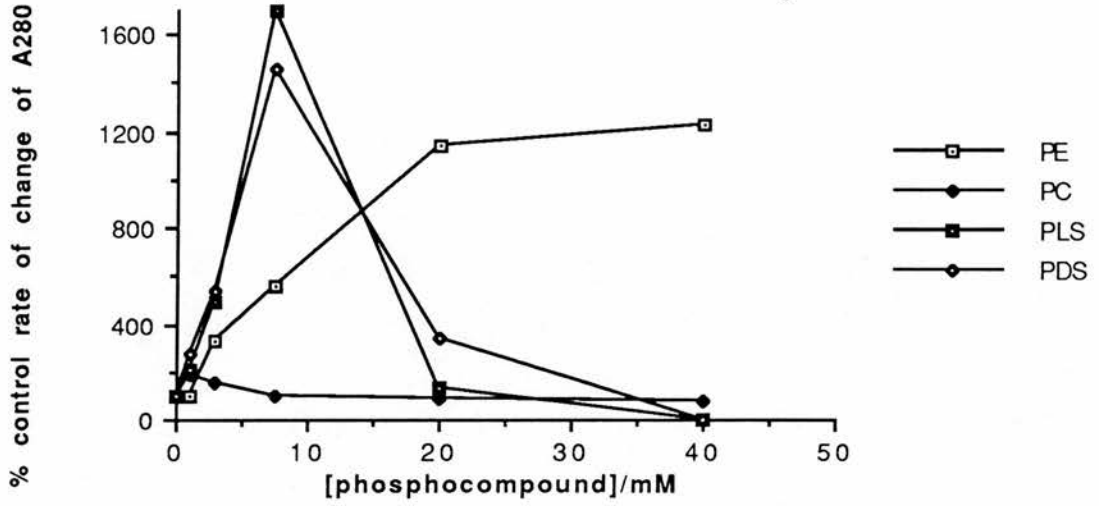
For rate first, different effects are again seen for the three compounds. Phosphoserine shows a biphasic effect on the rate of increase of A_{280}. There is a rapid increase to around 15 times control rate at around 8-12mM PS followed by a decline so that at 40mM PS no increase in A_{280} is observed. Again the effects of PDS and PLS are indistinguishable. The effect of PC is much more modest; a doubling of control rate followed by a slow decline to around 80% control rate at 40mM PC. PE increases the rate of change of A_{280} by up to 12-fold at 20mM followed by a tailing off. Fitting of a saturation curve to the PE data gives a maximum acceleration of rate of 13-fold with half maximal effect seen at 10mM on a slightly sigmoid curve (n_H=1.7±0.4).

Effects on the maximum change in A_{280} seen are more similar between the compounds (Fig 4.18b). With increasing concentration, both phosphoserines and PE are seen to double the maximum change in A_{280}. With the phosphoserines the maximum occurs at around 2mM and is followed by a sharp decline. The PE maximum is between 10 and 15mM and tails off only slowly.

Observation at 280nm means that these results are not easy to explain in detail. Several overlapping processes, each affecting A_{280}, are probably occurring simultaneously. However, the initial increase in A_{280} seems likely to result from

Fig 4.18 Phosphocompounds and thermal denaturation

a) Effect on rate of change of A280



b) Effect on maximum A280 change

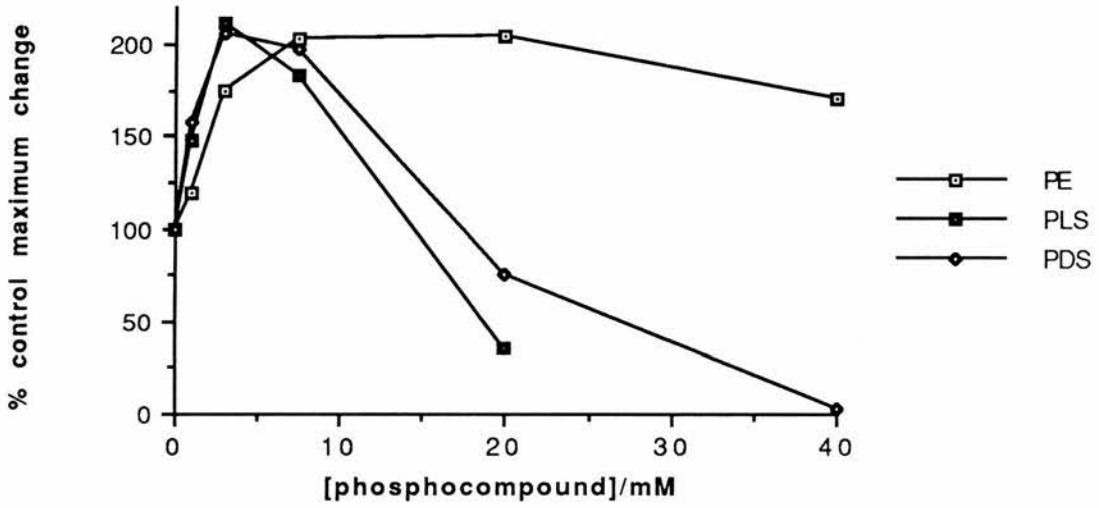
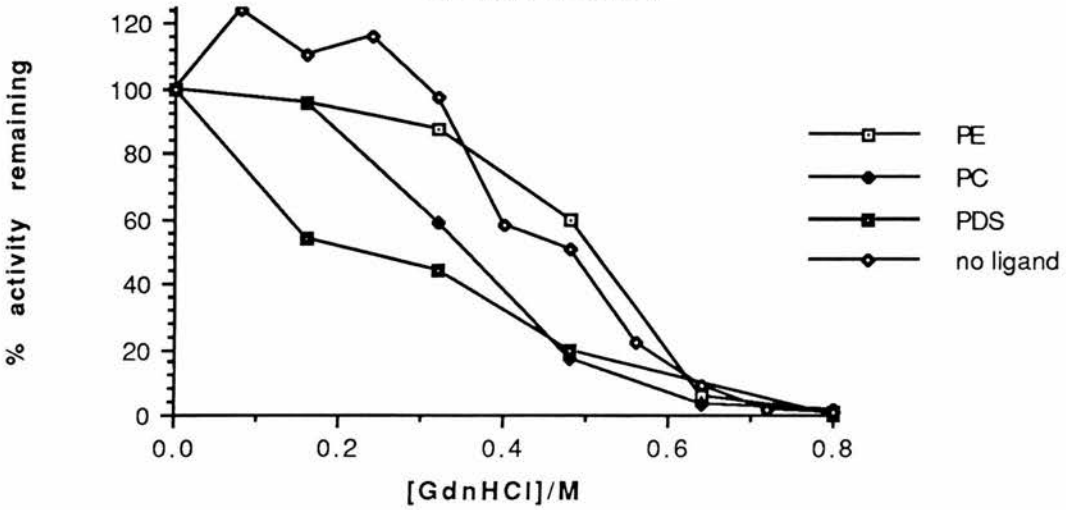


Fig 4.19 Phosphocompounds and GdnHCl denaturation



exposure of aromatic residues to solvent as the tetramer dissociates and denatures. The sharp concentration dependence of the effects of PDS is therefore consistent with the data in sections above: denaturation of monomers will occur more rapidly than denaturation of tetramers. The absence of an A_{280} change in the presence of high concentrations of phosphoserine may indicate an acceleration of transformation of denatured PFK to later low-absorbance species such as large aggregates. Almost as large an acceleration of rate is caused by high concentrations of PE suggesting again destabilisation of the enzyme. This could be a result of binding to either the same site as phosphoserine or a different one. PE differs only from phosphoserine in the absence of one carboxylate. If PE bound to the same site then the absence of a large PE fluorescence effect would mean that that carboxylate was essential for the PDS dissociation effect. The ability of intermediate concentrations of PE and phosphoserine to double the maximum change in A_{280} is most readily explained as being simply a consequence of the accelerated denaturation. Accelerated denaturation followed by relatively slow conversion to low-absorbance species would lead to an accumulation of high-absorbance denatured protein.

4.3.3.5 GdnHCl denaturation

The effects of the small phosphocompounds on the GdnHCl denaturation of PFK after four hours incubation are shown in Fig 4.19. 10mM PDS destabilises PFK to denaturation so that 0.2M GdnHCl causes 50% loss of activity compared to the experiment with no phosphocompound. PC also seems to cause a small destabilisation whereas PE has no effect.

4.3.4 Bisphosphates

4.3.4.1 Enzyme assay

Compound 233 at concentrations of up to 10mM had no effect on the rate of PFK. The effects of 5mM epoxy compounds on the response of PFK to F6P concentration

were determined (Fig 4.20). Both α- and β-epoxy compounds are competitive inhibitors with respect to F6P. Sigmoidal curve fitting gives a $S_{0.5}$ value for F6P in the absence of inhibitor of 460±10μM. In the presence of 5mM α- and β-epoxy compounds the apparent $S_{0.5}$ values were 650±10μM and 610±10μM respectively. Assuming pure competitive inhibition this implies K_i values of 13.7mM and 17.7mM for the α- and β-epoxy compounds respectively.

No evidence of covalent attachment of either epoxy derivative to the enzyme was seen. When reaction was allowed to continue for one hour an absolutely constant reaction rate was seen. A decline in rate would be expected if covalent bonding was occurring.

4.3.4.2 Fluorescence

Compound 233 did not quench PFK fluorescence. Little change was seen up to 25mM concentration. A change would be expected if compound 233 bound in the same way as F16BP so the absence of effect suggests that compound 233 does not bind to *E.coli* PFK.

Both epoxy compounds quench PFK fluorescence. Fig 4.21 compares their fluorescence responses with those produced by F16BP. The parameters derived from these curves by sigmoidal curve fitting are shown in Table 4.3. They show that the epoxy compounds bind more weakly than F16BP itself. The α-epoxy compound is the more tightly binding of the two, in agreement with its greater inhibitory effects (see above). The three compounds bind with a similar degree of sigmoidicity. The greater F_{max} values for the epoxy compounds were unexpected. This suggests that the presence of the epoxide group, in either α or β conformation, effects an additional conformational change altering the environment of Trp311.

4.3.4.3 GdnHCl denaturation

Compound 233 at a concentration of 10mM had, if anything, a slightly destabilising effect on *E.coli* PFK incubated with GdnHCl (Fig 4.22). Insufficient quantities of the epoxy compounds were available for testing by this method.

Fig 4.20 Effect of epoxy compounds on PFK kinetics

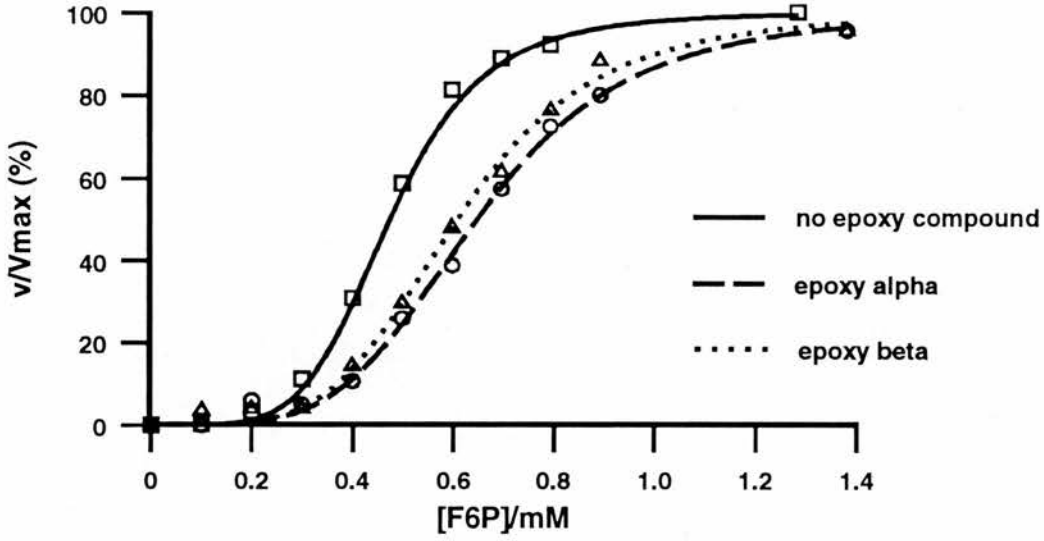


Fig 4.21 Effects of epoxy compounds on PFK fluorescence

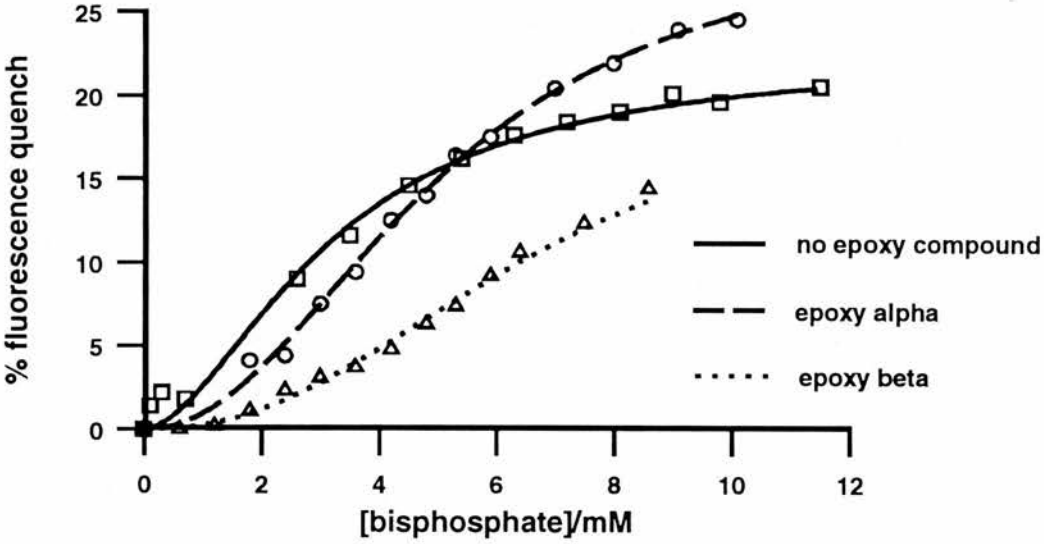


Table 4.3 Effects of FBP and epoxy compounds on PFK fluorescence

	F_{max}/%	$S_{0.5}$/mM	n_H
F16BP	25±2	3.6±0.5	1.5±0.2
epoxy α	31±2	5.1±0.3	2.1±0.1
epoxy β	30±6	8.7±1.5	2.1±0.2

Fig 4.22 10mM '233' and GdnHCl denaturation

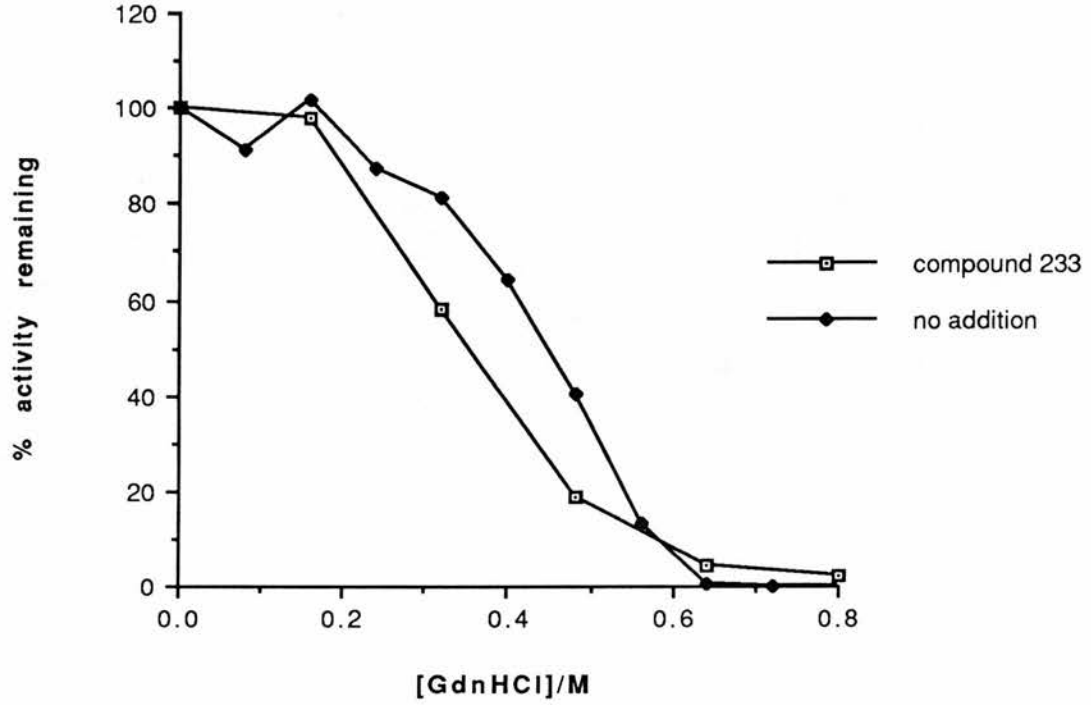
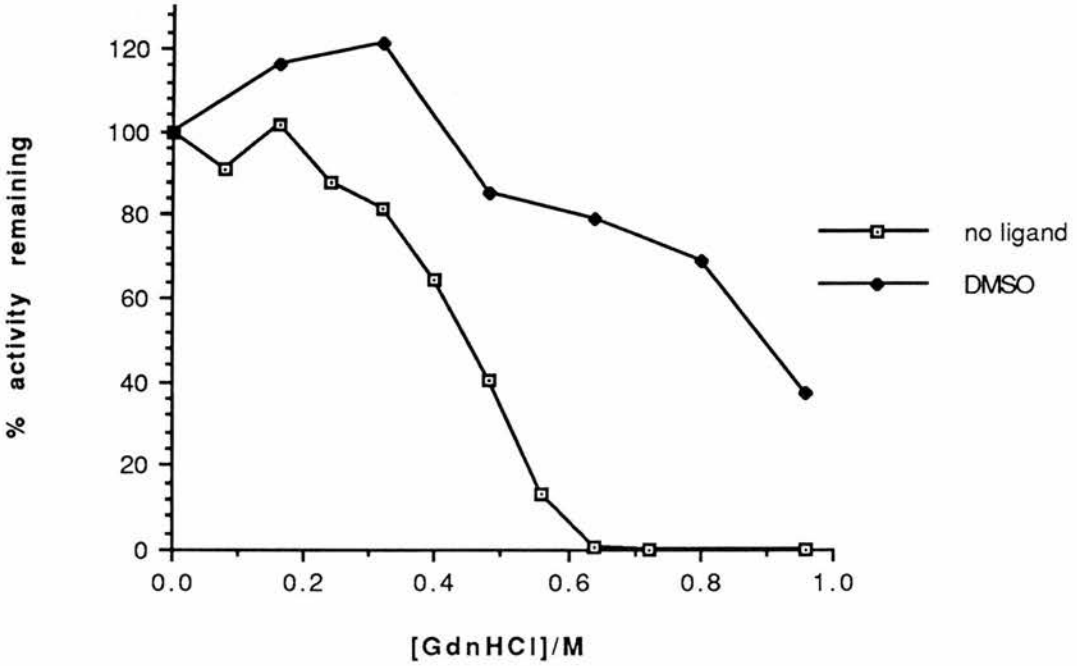


Fig 4.23 Effect of DMSO on GdnHCl denaturation



4.3.5 LUDI compounds

The LUDI compounds will be referred to by abbreviated names, mostly corresponding to LUDI's own nomenclature. The abbreviations and the full compound names are shown in Table 4.4.

4.3.5.1 Enzyme assay

A variety of compounds suggested as possible PFK ligands by the LUDI program were tested in PFK assays for inhibition. In many cases the maximum concentration that could be tested was limited by absorption of the LUDI compound at the wavelength being used to monitor reaction. In cases where the compound was dissolved in DMSO, the effect of DMSO alone on PFK activity (Fig 4.1) limited the concentration that could be tested. A reproducible inhibition of greater than 10% was seen for none of these compounds with either *E.coli* or RM PFK. Table 4.5 summarises some data relating to inhibition testing and includes the maximum concentration tested for each of the compounds.

4.3.5.2 GdnHCl denaturation

Concentrations in the range 1-5mM of AQDS, AM7, DAS, AF5, GU1, CDO, BA4 and DAP were tested to see if they could stabilise *E.coli* PFK against GdnHCl denaturation. In all cases presence of the compound either had no effect on or destabilised the enzyme. This is despite the fact that several of these were dissolved in DMSO which alone was shown to stabilise the enzyme (Fig 4.23).

In the presence of 0.16M GdnHCl RMPFK shows negligible activity after three hours incubation. The following compounds at 2mM were tested to see if they could maintain significant RMPFK activity under these conditions :- PAA,GU1, CDA, AQDS, GBA, A66, BA6, BA4, DAS, AF5, BA2, A18, DAP. None of these proved capable of protecting RMPFK against denaturation.

Table 4.4 LUDI compound names and abbreviations

Abbreviation	Full name
A63	4-nitrodiphenylamine
A66	N-phenylbenzylamine
BA4	phenylene-1,4-diacetic acid
AF5	4-hydroxyquinoline-2-carboxylic acid
GU1	2-aminobenzimidazole
PAA	phenylacetic acid
DAP	2,4-diaminopyridine
BA2	benzene-1,3-dicarboxylic acid
Y41	methylbenzoate
BA6	pyridine-3,5-dicarboxylic acid
CDA	chelidamic acid
DAS	diphenylamine-4-sulphonic acid
AM7	isatin
AQDS	anthraquinone-1,5-disulphonic acid
CDO	chelidonic acid
GBA	4-guanidinobenzoic acid
A18	3-nitrodiphenylamine
BA6	pyridine-3,5-dicarboxylic acid

Table 4.5 Testing of LUDI compounds for inhibition

	Soluble in buffer?	Soluble in DMSO?	540nm testing?	Max conc. tested /mM
A63	no	poorly	no	0.1
A66	no	poorly	no	0.5
BA4	no	yes	no	0.3
AF5	no	yes	yes	1
GU1	yes		no	3
PAA	yes		no	3
DAP	yes		yes	5
BA2	no	no		
Y41	no	no		
BA6	no	yes	no	5
CDA	yes		no	6
DAS	yes		yes	2.5
AM7	inhibits a coupling enzyme			
AQDS	inhibits a coupling enzyme			

4.3.5.3 Thermal denaturation

The following concentrations of LUDI compounds were tested for their effect on the thermal denaturation of *E.coli* PFK :- 5mM BA4, 0.5mM GBA, 0.2mM CDA, 1mM BA6, 0.3mM A18, 1mM BA2, 0.25mM A66, 0.1mM CDO, 0.2mM AQDS. In many of these cases, an upper limit on the concentration of compound to test was imposed by the absorbance of the compound at 280nm. None of these compounds affected the thermal denaturation. However, an effect was seen with A66. This compound reduces the maximum change in A_{280} produced in a concentration-dependent fashion (Fig 4.24) implying an effect on one of the reactions of PFK thermal denaturation.

4.3.6 Naphtho compounds

4.3.6.1 Enzyme assay

Testing of the naphtho compounds was hampered by two factors. First, testing of ANQ proved to be impossible since the compound absorbed strongly, not only at 340nm, the wavelength used in the standard assay, but also at 540nm where the modified assay operates. HNNS and NQS did not absorb so strongly at 540nm but were found to inhibit the coupling enzymes of the modified 'forward' assay. 4mM HNNS was found to inhibit the metabolism of added F16BP by about 70% and 4mM NQS by about 85%. It does seem likely that HNNS at least inhibits PFK - 4mM HNNS in the PFK assay gave an apparent inhibition of 95%. However, the multiple effects prevent accurate measurement of PFK inhibition.

4.3.6.2 GdnHCl denaturation

All three naphtho compounds were found to destabilise *E.coli* PFK towards GdnHCl denaturation at a concentation of 2mM (Fig 4.25).

4.3.6.3 Thermal denaturation

These studies provided clear evidence of binding of the naphtho compounds to *E.coli*

Fig 4.24 Effect of A66 on PFK denaturation

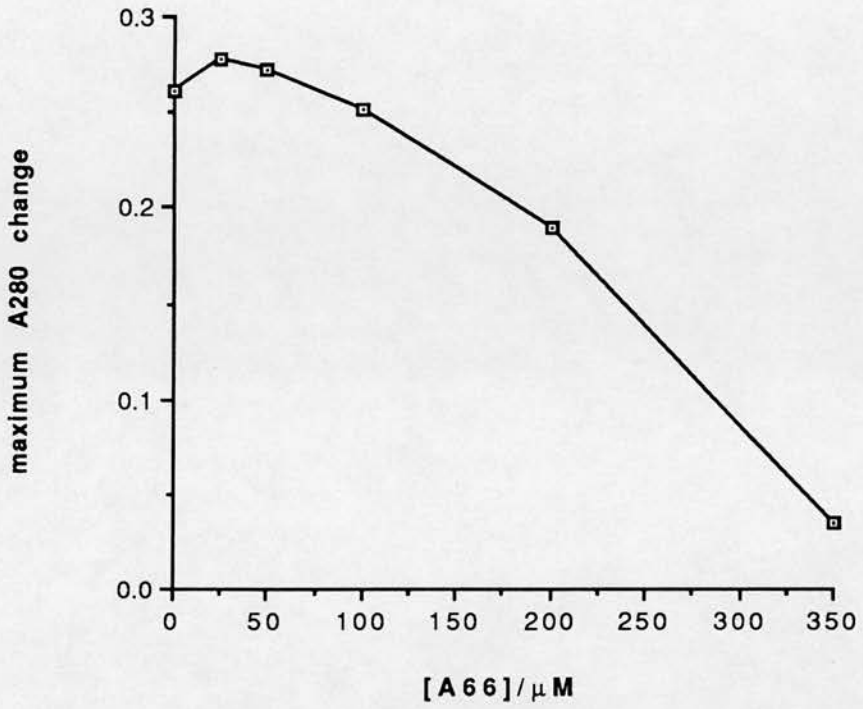
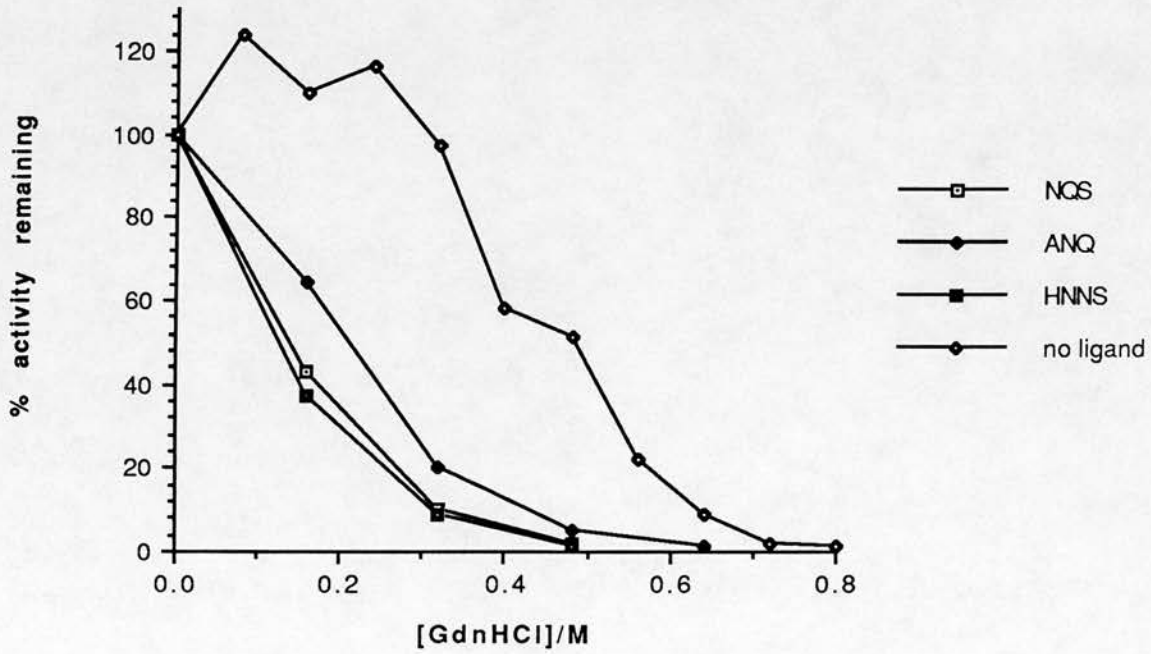


Fig 4.25 Naphtho compounds and GdnHCl denaturation



PFK. The effect of compound concentration on the rate of increase of A_{280} is shown in Fig 4.26. Inhibition in a concentration-dependent manner is seen for all three compounds. Only the NQS data, however, are of sufficiently high quality for reliable curve fitting. Fitting of a sigmoid curve yielded the parameters:- Max inhibition 101±5%, $S_{0.5}$=43±4μM, n_H=1.9±0.2. Inspection of the other traces suggests that ANQ is more potent than NQS with half inhibition seen at around 15μM. Conversely, HNNS is less potent with half inhibition around 75μM. In both these cases some sigmoidicity is suggested.

The clear sigmoidal character of the NQS results is particularly interesting since two inferences can be made from it. First, the F6P/F16BP binding site is the only one on *E.coli* PFK capable of cooperative ligand binding. NQS, and almost certainly HNNS and ANQ too, therefore seem to bind to this site. Secondly, cooperativity would only be seen for oligomeric forms of PFK. Hence the naphtho compounds seem to exert their effect on PFK denaturation by binding to an oligomeric species. This does not necessarily mean that they exert their effect by the binding to the native state of the enzyme; they could be binding to an oligomeric intermediate of the denaturation pathway. There seems to be a precedent for this in the protective effect of F6P (see section 4.3.2.2). There a similar degree of sigmoidicity was observed but the $S_{0.5}$ value did not match any of those determined by other techniques for the native state.

4.3.6.4 Fluorescence

The effects of naphtho compounds on PFK fluorescence were determined. Since these compounds absorb significantly at the wavelengths used for excitation and detection, a correction was made using the compound N-acetyltryptophanamide (NAT). Fluorescence reponses to naphthocompounds of PFK and NAT were determined and the 'true' fluorescence change (that due to change in Trp311 environment) obtained by subtraction. Fig 4.27 shows the raw data, determined in duplicate, and Fig 4.28 the 'true' fluorescence changes.

Fig 4.28 suggests that NQS, HNNS and ANQ are all capable of binding to PFK and

Fig 4.26 Effect of naphtho compounds on thermal denaturation

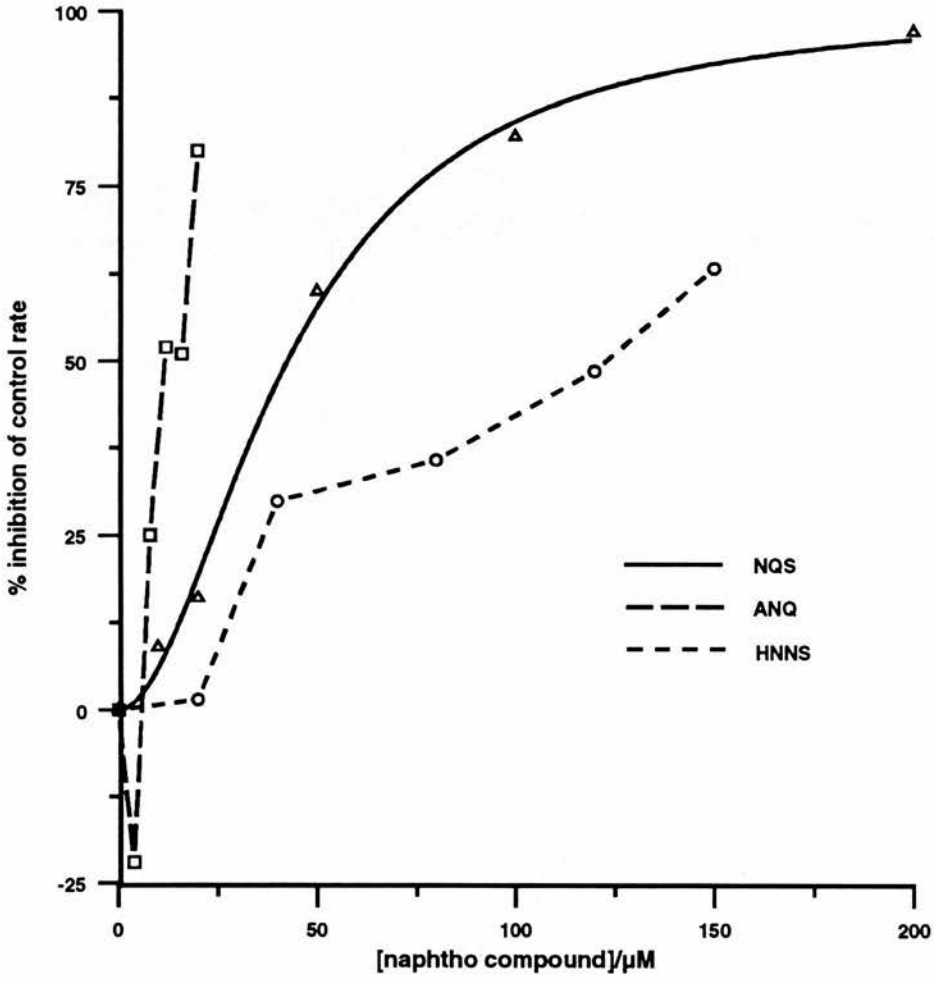
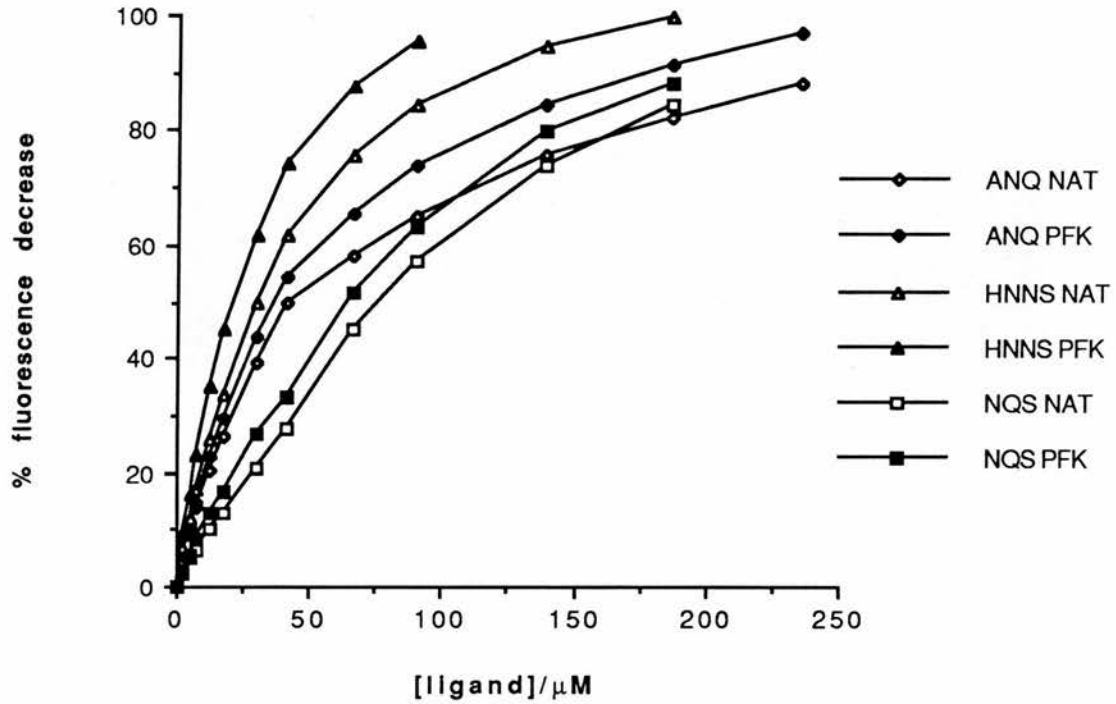
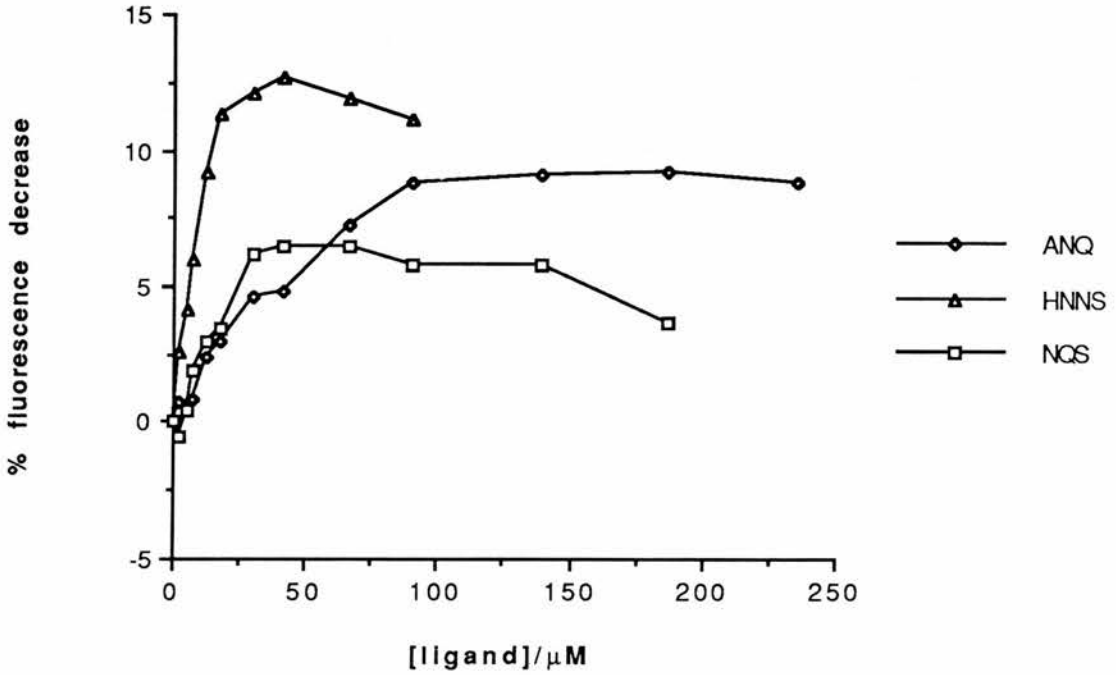


Fig 4.27 Effect of naphtho compounds on fluorescence



4.28 Corrected naphtho compound fluorescence data



thus altering its fluorescence. Curve fitting was not performed since the curves do not conform to standard binding models; in particular the fluorescence decrease observed declines at higher concentrations of naphtho compound. The reason for this is not known but may be a result of deficiencies in the correction method.

These compounds can be modelled into the F6P binding site of *E.coli* PFK (see Section 3.3.4.2) yet give maximum changes in fluorescence smaller than that seen for F6P or F16BP (about 20%). The failure of the naphtho compounds to reach this level could have the same cause as the decline in fluorescence decrease at higher concentrations. Alternatively, the naphtho compounds may not have all the structural necessities for the full fluorescence change.

Table 4.6 presents a summary of the testing of the different classes of compound by the different methods available. 'Yes' indicates that at least some of the compounds in a particular class produced changes in PFK behaviour observable by a particular method. 'No' indicates an absence of response. 'N.D.' means that a particular combination of compound class and analytical technique was not tried.

4.3.7 Combinations of compounds

4.3.7.1 Thermal denaturation

PDS and NQS

Experiments were carried out to see if the protective effect of NQS towards thermal denaturation of PFK could counteract the destabilising effect of PDS. With 7.5mM PDS present in all cuvettes except the control, NQS was present at 0, 20, 50, 100 and 200μM. The curves showing A_{280} against time for these last five incubations were essentially superimposed. They all showed an acceleration of control rate of increase of A_{280} of around 13-fold, in agreement with previous results. Thus NQS cannot protect

Table 4.6 Summary of ligand testing

	Enzyme Assay	GdnHCl denaturation	Thermal denaturation	Fluorescence
Natural Ligands	yes	yes	yes	yes
Small Phospho-compounds	N.D.	yes	yes	yes
Bis-phosphates	yes/no	N.D./no	N.D.	yes/no
LUDI compounds	no	yes	yes	N.D.
naphtho compounds	yes	yes	yes	yes

against the effects of PDS.

PDS and F6P

In the same way the effect of combining PDS and F6P was determined. F6P was present at 0, 100, 200, 400 and 800μM with 7.5mM PDS. A very different result was seen (Fig 4.29). When inhibition of rate is plotted against F6P concentration the data fit well to a sigmoid curve (Fig 4.30). The parameters derived from this fitting - Max inhibition 101±2%, $S_{0.5}$=114±5μM, n_H=2.1±0.2, agree very well with those obtained for F6P in the absence of PDS (Max inhibition 103±3%, $S_{0.5}$=180±9μM, n_H=1.9±0.2). However there are differences between the traces obtained in the presence and absence of PDS (compare Fig 4.7 with Fig 4.29). In the presence of PDS two phases to the traces may be defined. In the first all non-control traces show an increase in A_{280} of about 0.04. There is no lag before this increase appears so that here the traces resemble more those found for PDS alone than F6P alone. In the second phase the rate is F6P-dependent and the traces resemble more those for F6P alone than PDS alone.

An explanation for these results may lie in the PDS gel filtration results. They suggest that 7.5mM PDS would effect about 30% dissociation of PFK. The constant change seen in the first phase of the above results may represent the denaturation of this fraction of the PFK. The remaining tetrameric fraction would then be protected by the F6P in the observed concentration-dependent manner. If this is the case then the F6P-bound portion would have to be insensitive to PDS dissociation. In contrast the results from combination of NQS and PDS suggest that the NQS bound forms of PFK must be PDS-sensitive. This is interesting since the lack of extra protection from NQS in the presence of F6P (data not shown) suggests that NQS and F6P may compete for the same site.

Figure 4.29 Effect of F6P on thermal denaturation of *E.coli* PFK in the presence of PDS

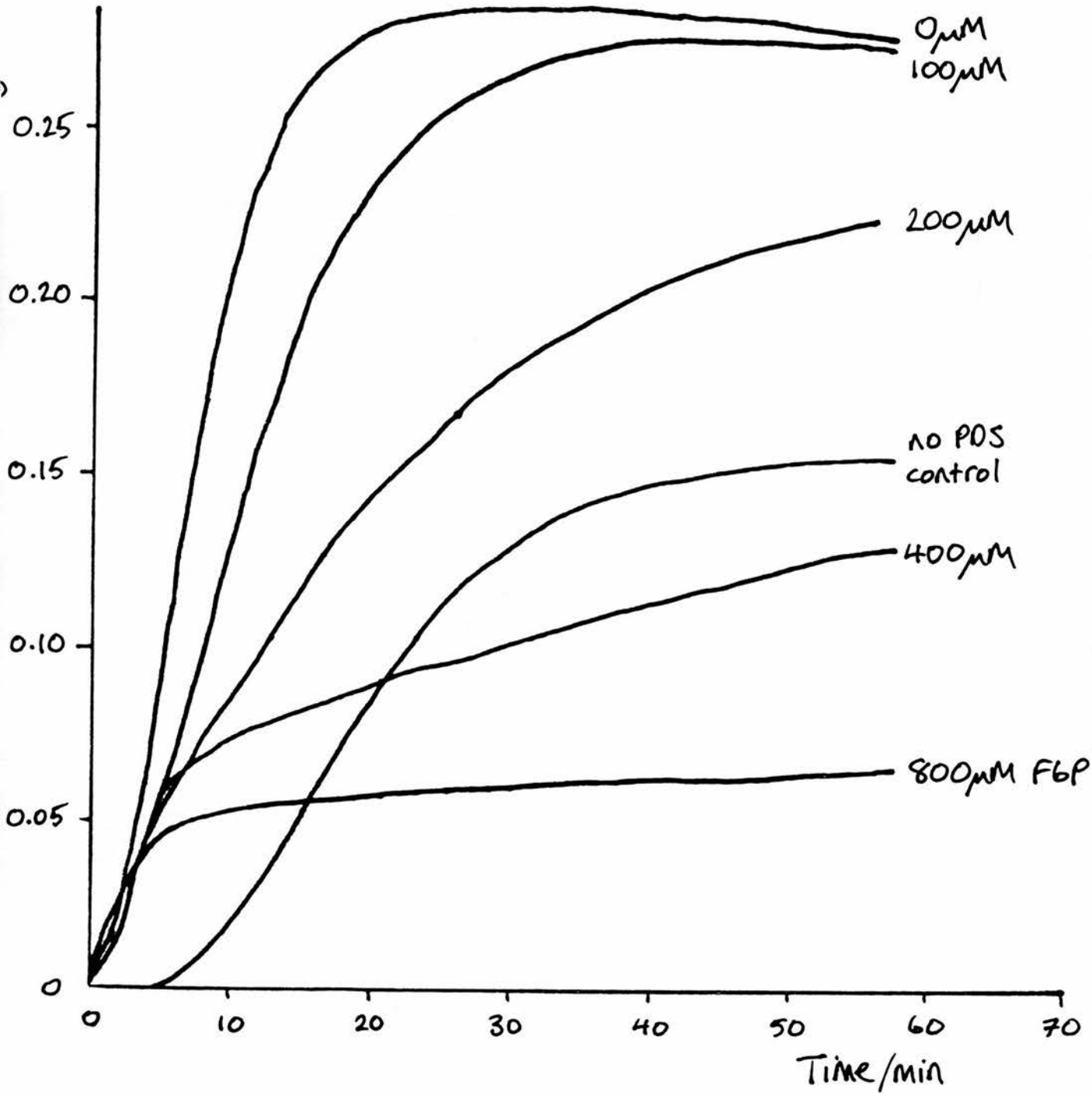
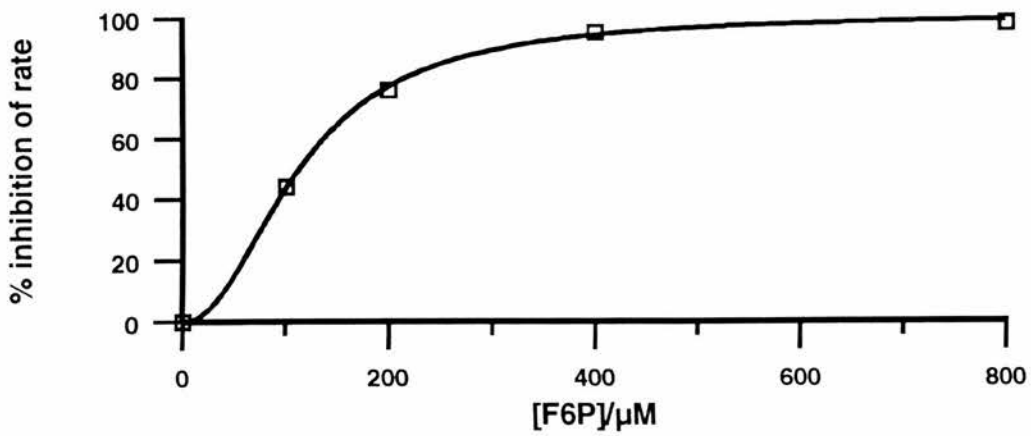


Fig 4.30 Effect of F6P+PDS on thermal denaturation



4.3.7.2 Fluorescence

Small phosphocompounds and natural ligands

The fluorescence responses of PFK to F6P and ADP were determined in the presence of 10mM PE, PC and PDS. The F6P results are shown in Fig 4.31a and the ADP results in Fig 4.31b. The parameters derived from these data are shown in Tables 4.7a and 4.7b. Sigmoid curve fitting was applied to the F6P data. Sigmoid curve fitting of the ADP data gave n_H values of 0.8±0.2 so hyperbolic curve fitting was applied.

In general PE and PC seem to have similar effects with PDS somewhat different. This might be expected from a significant difference in their chemical structures; PDS has an extra charged, bulky carboxylate group that is absent in PC and PE.

Considering the F6P effects first, it is surprising to note that PC and PDS seem capable of increasing the maximum fluorescence quench. This is despite the fact that 10mM PDS itself produces a 5% fluorescence quench and 10mM PC a 7% fluorescence quench. One possible explanation is that binding of PC and PDS loosens the structure (in the case of PDS sufficiently to effect dissociation) so that the conformational changes on binding of F6P produce bigger changes in the environment of Trp311. Both PC and PDS lower the Hill constant to near one and lower the $S_{0.5}$ values for F6P. These results would be explained by conversion of T-state to R-state PFK which could arise from binding to the effector site. If PC or PDS bound in the same way as PEP the opposite effect would be expected. For PDS at least modelling into the T-state structure in place of the PEP analog phosphoglycollate produces serious steric clashes. The difference in PC fluorescence response between wild-type and K213A enzymes (Section 4.3.3.2) suggests that PC does bind to the effector site. However, if PC (or PDS) do bind to the effector site, then they must do so in such a way as to emulate ADP rather than PEP. It is possible that PDS is binding to more than one site with an altered fluorescence quench resulting from the K213A mutation swamped by another larger contribution. In contrast to PC and PDS, PE produces a lower $S_{0.5}$ value for F6P without lowering n_H. This is less easy to explain.

Fig 4.31 Effect of phosphocompounds on fluorescence responses

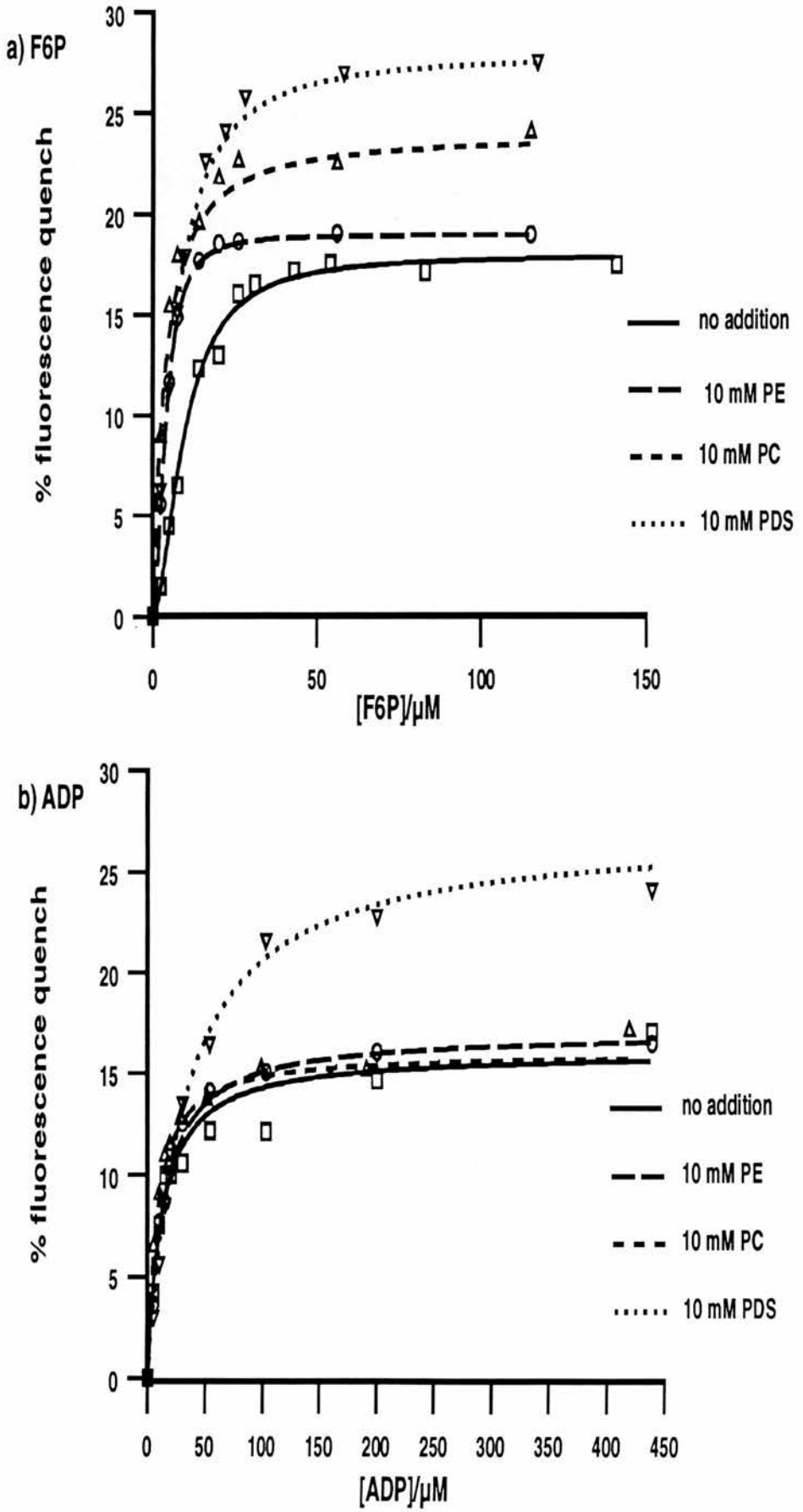


Table 4.7 Effect of small phosphocompounds on PFK fluorescence responses

a) F6P fluorescence response

	no addition	PE	PC	PDS
F_{max} (%)	18±0	19±0	24±1	28±0
$S_{0.5}$/μM	9.6±0.5	4.0±0.0	3.7±0.2	6.5±0.2
n_H	1.8±0.2	2.0±0.0	1.1±0.2	1.4±0.1

b) ADP fluorescence response

	no addition	PE	PC	PDS
F_{max} (%)	16±1	17±0	16±0	27±1
K_s /μM	12±2	13±1	8±1	31±2

PE and PC have little effect on the ADP fluorescence response. The raising of K_s by 10mM PDS would be most easily explained by competition for the same (effector) site. However, this is not consistent with the similar PDS fluorescence response of the K213A mutant unless we again suppose that it binds in a different way to the natural ligands ADP and PEP. The maximum fluorescence quench produced by ADP is increased by PDS. Some synergism between PDS and ADP is suggested since after a 7% fluorescence change due to 10mM PDS, ADP can produce a further 27% quench. ADP alone produces only a 16% fluorescence quench.

NQS and F6P

The F6P fluorescence response of *E.coli* PFK was determined in the presence of 25, 50, 75 and 100μM NQS. Because of the absorbance by NQS at excitation and emission wavelengths, volume changes were kept very small, less than 3%, during each experiment. This enabled changes in absorbance due to NQS to be ignored. The data, means of two determinations, are shown in Fig 4.32 fitted to sigmoid curves. The percentage quench is related to the fluorescence value after addition of NQS which caused an initial non-specific fluorescence decrease. The parameters derived from this curve fitting are shown in Table 4.8.

NQS has a definite effect on the F6P fluorescence reponse, most obviously seen in the decline in F_{max} seen as NQS concentration increases. Also seen is a slight fall in $S_{0.5}$, although this may be within experimental error. The apparent increase in n_H is not convincing. This parameter is determined less accurately than the other two. In particular, a real n_H value of 2.3 seems unlikely given that F6P can only bind to two sites on the tetramer (Deville-Bonne and Garel, 1992).

The thermal denaturation data strongly suggest that F6P and NQS compete for the same site or overlapping sites. These fluorescence data are not consistent with that idea. If the two compounds were to compete then an increase in $S_{0.5}$ for F6P would be expected without a change in F_{max}. In fact, no increase in $S_{0.5}$ is seen and F_{max} is decreased.

Fig 4.32 Effect of NQS on F6P fluorescence response

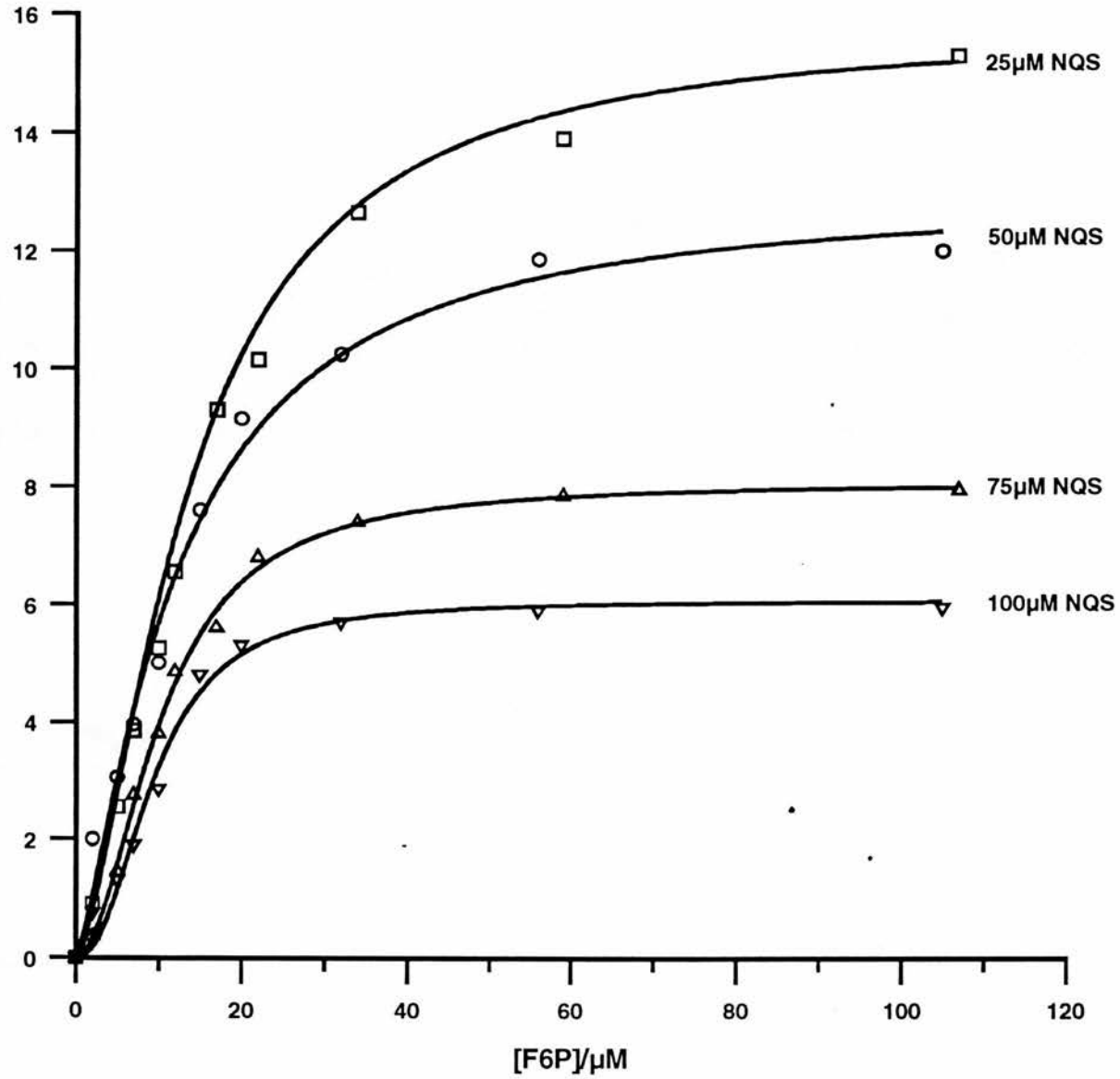
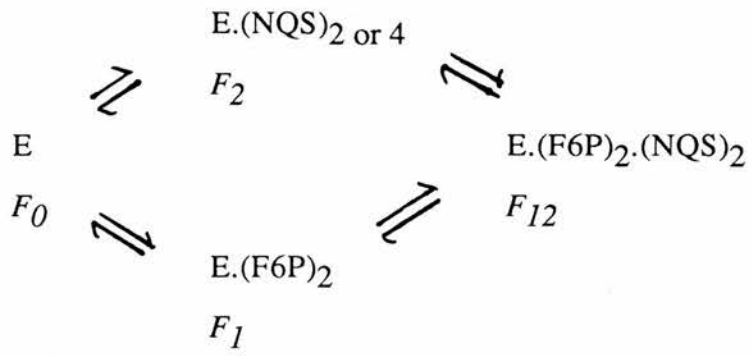


Table 4.8 Effect of NQS on F6P fluorescence response

[NQS]/μM	F_{max}(%)	$S_{0.5}$/μM	n_H
0	18.1±0.2	9.6±0.5	1.8±0.2
25	15.8±0.3	14.7±0.6	1.6±0.1
50	13.0±0.7	12.1±1.4	1.4±0.2
75	8.1±0.1	10.3±0.3	2.0±0.1
100	6.1±0.2	9.4±0.6	2.3±0.3

One interesting explanation could involve the following scheme :-



In this scheme E represents the tetrameric enzyme. The stoichiometry of two F6P molecules per tetramer shown in F6P-bound species is known from equilibrium dialysis experiments (Deville-Bonne and Garel, 1992). The stoichiometry of the NQS-bound species is unknown. The important feature of this model is the postulated $E.(F6P)_2.(NQS)_2$ species. This would arise by binding of NQS to the two F6P binding sites left unoccupied in the F6P-saturated enzyme. The four enzyme species in the scheme are shown with different fluorescence properties. This is consistent with fluorescence data in this thesis. F6P alone causes an 18% fluorescence quench (Fig 4.6) and NQS alone a 7% fluorescence quench (Fig4.28). NQS alters the F6P fluorescence response in a concentration-dependent way (Fig 4.32). The above scheme is also consistent with thermal denaturation data. NQS in the above scheme is binding to the F6P binding sites so that the observed sigmoidicity in the protection against denaturation vs. [NQS] curve could be explained by cooperative NQS binding. All of the ligand-bound species would have the same thermal stability in the above scheme. This seems inconsistent with the observation that F6P, but not NQS, can stabilise *E.coli* PFK against dissociation by PDS. Therefore it must be argued either that thermal- and PDS-induced changes proceed by different pathways, or that PDS binding, to an unknown site, interacts differently with F6P and NQS binding.

4.4 Conclusions

A variety of classes of compound deriving from a variety of design methods have been tested for binding and interaction to both *E.coli* and Rabbit Muscle PFK. The goal of discovery of new potent PFK inhibitors has not been achieved. This was probably an unrealistic target given the time and facilities available. However much information has been accumulated relating to the enzymes and the methods that may be applicable for the design of inhibitors.

Three main problems were encountered in the testing of LUDI compounds for PFK inhibition. First, a majority of the LUDI compounds tested absorbed strongly at 340nm, imposing a severe upper limit on the concentration of the compound that could be tested in the assay. Here the extension of the enzyme assay using MTT and Meldola Blue dyes to enable rate measurement at 540nm proved to be a useful innovation, although some compounds absorb at 540nm as well. Another possibility would be the linking of NADPH oxidation to reduction of O_2 to H_2O_2 via an intermediate electron carrier. An oxygen electrode could then be used to measure the rate of reaction so that absorbance by putative inhibitors would not interfere with the assay. Secondly, coupled enzyme assays are not ideal for inhibition measurement since the effect of potential inhibitors on the coupling enzymes must be determined. If inhibition of coupling enzymes is discovered then the assay must be discounted as a reliable indicator of primary enzyme activity. Direct measurement of the conversion of, for example, radioactive ATP to ADP may be one solution to the problem. Substrate and product would be separated by, for example, ion exchange chromatography and their radioactivity measured by scintillation counter. However, this method would undoubtedly be very laborious and time-consuming. Another possibility would be the assay of ATP produced by the 'backwards' reaction, either in denatured reaction mix or

in some purified form, by a luciferin/luciferase assay. This is more attractive though initial trials ran into reproducibility problems and were not followed up. The third major problem encountered during testing was the relative insolubility of some compounds. The finding that 5% DMSO could safely be included in both *E.coli* and Rabbit Muscle PFK assays provided a partial solution to the problem, but solubility was still occasionally the limiting factor in concentration of compound able to be tested.

The small phosphocompounds suggested by the 2D connectivity search proved interesting though it should be immediately emphasised that high concentrations were required for effects: in no way are these drug leads. The case of the small phosphocompounds was one in which the problems associated with coupled enzyme assays were apparent. Thus their effect on PFK activity could not be determined. It seems likely that ambiguities arising from the results of other methods could be resolved by determination of the effects of the compounds on enzyme rate. The phosphocompounds affected the *E.coli* PFK behaviour seen by all the other methods applied, showing that they bind to the enzyme. To which site(s) they bind seems less clear.

The ability of both phosphoserines to effect dissociation of PFK, as shown by fluorescence and analytical gel filtration, was unexpected. The fact that both isomers were equally effective suggests first that the effect arises not from binding to the F6P binding site; PDS but not PLS can be satisfactorily modelled here. Secondly, it was suggestive of some merely physical effect in an analogous way to urea, thiocyanate etc. However the lack of effect of phosphoserines on control proteins phosphoglucose isomerase, BSA and lysozyme (data not shown) seems to rule this out. Also the concentrations required are relatively small. This leaves us with phosphoserines binding specifically to PFK most likely at or near the subunit interfaces. The effector site is the most obvious candidate but the identical effects of phosphoserines on the K213A mutant argue against this. Alternatively it is possible that phosphoserines bind

to some novel site. Crystallographic or nuclear magnetic resonance spectroscopic study of PFK in the presence of phosphoserine would distinguish between these possibilities. PC is not modelled satisfactorily into the active site and affects the K213A mutant differently from the wild type enzyme. The similar maximal fluorescence quenches seen for PC and PEP further support the idea of PC binding to the effector site. However, the effect of PC on the F6P fluorescence response means that if this was the case, then PC would have to bind in such a way as to emulate ADP rather than PEP. Fluorescence data fail to suggest an obvious binding site for PE.

The destabilising effect of phosphoserine is again evident from the denaturation studies. The rate of increase of A_{280} during thermal denaturation is increased 15-fold by 10mM PDS and the midpoint of transition in the GdnHCl studies shifted from around 0.5M GdnHCl to 0.2M GdnHCl. PE can cause a similar destabilisation of PFK, assessed by thermal denaturation, but at higher concentration. It is therefore tempting to conclude that PE and phosphoserine share a destabilising binding site, and that the presence of the extra carboxylate group on the phosphoserines causes additional destabilisation sufficient to dissociate the enzyme into monomers.

The ability of phosphoserines to dissociate *E.coli* PFK may not be of just academic interest. If nothing else, it illustrates another possible method of action for rationally designed inhibitors. The vast majority of PFKs known are oligomeric and are inactive as monomers so that specific dissociative agents could be as effective as specific catalytic site-directed compounds. In a situation where an enzyme must be oligomeric for activity, non-conservative mutations at subunit interfaces are unlikely to be tolerated so that interface residues would diverge more slowly than surface ones during evolution. This argument is supported by the model of the Rabbit Muscle PFK; no insertions or deletions, for example, were located at presumed interfaces. If this is the case then suitable sites for rationally designed dissociative agents probably exist on RMPFK and other larger PFKs. In addition, RMPFK seems less stable overall than the *E.coli* enzyme. The GdnHCl denaturation results (compare Figs 4.9 and 4.11) illustrate this point. Thus, an *E.coli* PFK structure in the presence of phosphoserine

would enable the identification of a site whose homologue in higher PFKs could be a useful, novel target for rational drug design.

Compound 233 was not expected to bind to PFK. It lacks all three hydroxyl groups of F16BP, each of which forms H-bonds to the enzyme in the FBP-bound crystal structure. Modelling compound 233 into the structure in place of FBP puts the hydrophobic furan ring unfavourably into a largely hydrophilic pocket. Testing confirmed that the compound does not bind.

Lack of material prevented comprehensive testing of the epoxy derivatives. However, the fluorescence and enzyme activity studies provide convincing evidence of their binding to *E.coli* PFK. The results show that both derivatives bind more weakly than FBP itself, presumably due to steric hindrance from the epoxy groups. The lack of covalent modification by either epoxy compound is consistent with the pessimistic modelling results (see Section 3.3.2.2)

Disappointingly, none of the LUDI compounds showed any evidence of PFK inhibition or protection against GdnHCl denaturation. In some cases, solubility or other factors limited the concentration of compound that could be tested. Despite this most were tested at concentrations of greater than 1mM so that even had the compounds bound only weakly, their effects would have been evident. Many of the compounds destabilised PFK against GdnHCl implying some sort of interaction. However it seems likely that this may be non-specific; several of the compounds, for example, have distinct hydrophobic and hydrophilic regions perhaps conferring detergent-like properties on them.

A single compound, A66, influenced the denaturation of *E.coli* PFK. A concentration-dependent decrease in maximum change was seen. However, the compound did not affect the rate of increase of A_{280}. Since this is related to the first step of denaturation i.e. the process involving native tetramer, any ligand interacting

with native protein would be expected to influence this quantity. For this reason it seems likely that A66 may be affecting some later step of the denaturation pathway and is therefore not of interest as an inhibitor.

Since all of the LUDI compounds can be fitted into the PFK structure to form multiple H-bonds and/or hydrophobic contacts, the failure of any to inhibit is disappointing. This is particularly so since detailed modelling can produce convincingly complementary complexes (see Section 3.3.2.2). However, it must be remembered that the original LUDI concept envisaged fitting of multiple fragments and their linking together to produce larger viable inhibitors. Some of the compounds tested were quite small so that exceptional complementarity would have been required for tight binding.

The naphtho compounds certainly bind to *E.coli* PFK as shown by protection against thermal denaturation protection and effects on fluorescence. The two techniques show opposite orders of effectiveness with protection against denaturation in the order ANQ > NQS > HNNS and $S_{0.5}$ values for fluorescence effect in the order HNNS > NQS > ANQ. However, the fluorescence data are not ideal; the apparent tailing off in fluorescence quenching at high concentration of compound is more likely to be the result of somehow deficient data correction than a genuine finding. That said, the reverse order may be the result of the two methods measuring binding to different species. Fluorescence quenching results from effects on a native, 'ground-state' form of the protein whereas the protection against thermal denaturation may arise from effects on an intermediate of the denaturation pathway. This seems to be the case for denaturation protection by F6P where the $S_{0.5}$ value obtained fails to match values derived by other techniques relating to native enzyme. In both that case and protection by naphtho compounds, the clear cooperativity seen means that the form of the enzyme acted upon is still oligomeric. As stated before, this also identifies naphtho compound binding with the F6P binding site. A model of the type proposed in Section 4.3.7.2 is consistent with most of the experimental data. However, ligands tend to stabilise PFK similarly against dissociation and denaturation by various agents. Therefore the similar

stabilisation by NQS and F6P to heat, contrasting with their different stabilisation to PDS, was unexpected and requires explanation.

These data and the apparent evidence of enzyme inhibition, despite measurement difficulties, raise the question of whether these compounds could form the starting point for a rational drug design project - are they lead compounds? Of all the data the most potent effect is half protection from thermal denaturation by ANQ at around 15μM. This compares with typically sub-micromolar values for half inhibition of target shown by industrial drug leads. However, the naphtho compounds are small compared to most drug leads and would therefore be expected to have lower affinities. There is plenty of scope for adding other favourable functionalities, perhaps suggested by LUDI, to the naphtho compounds to improve affinity. The F6P pocket is quite large and adjacent to the even larger nucleotide binding site. In favour of these compounds is their naphthalene carbon skeleton whose chemistry is well known. Addition of other groups to such skeletons is by standard, relatively straightforward reactions.

Much less effort was expended in testing compounds for possible effects on Rabbit Muscle PFK. Only two techniques were used - enzyme assay and GdnHCl denaturation, with neither producing effects for other than natural ligands. The lack of variety of techniques available to study RMPFK hinders the assessment of putative effectors.

Chapter 5

Conclusions

Accurate modelling of proteins, ligands and protein-ligand complexes is an important part of the rational drug design process. In many cases, particularly for parasite enzymes, there is no crystal structure available for the target enzyme. Several factors may cause this including difficulties in obtaining and purifying material, and the intrinsic slowness of the X-ray crystallography method for structure determination. Much more likely to be available is the amino acid sequence of the target enabling, in favourable cases, a model of the structure to be built by homology to a known 3D protein structure (see Section 2.1.2). Accurate protein models are essential for protein-structure based ligand design methods since ligand binding is known to be highly sensitive to changes in target structure. This is readily explained at an atomic level by consideration of, for example, the sensitive angular- and distance-dependence of H-bonds (Fersht, 1985) which frequently contribute to binding affinity. Indeed this sensitivity to the target structure is the premise on which species-selective drug design is based.

Accurate modelling of ligands is important for the same reason although database searches based only on 2D ligand structure can be successful (Pepperrell and Willett, 1991). A major step towards accurate ligand modelling is the development of programs like CONCORD (Rusinko *et al.*, 1988) which are capable of determination of reasonably low energy structures for small molecules. However, probably more than with proteins, ligand flexibility is a complicating factor; ligands do not always bind to proteins at, or even near, their minimum global energy conformation. Attempts to deal with ligand flexibility are therefore important (Clark *et al.*, 1993).

Results in this thesis shed some light on the accuracy with which proteins can be modelled by various methods. Several protein- and ligand-based ligand design procedures, using different levels of information, can be compared for the quality of their results. An overall conclusion of the following analysis is that human intelligence and intuition still have an important part to play alongside brute force computational techniques. Interestingly, artificial intelligence techniques have recently been applied to some of the ligand design operations currently carried out by humans (Cohen and

Shatzmiller, 1993), although this work is at an early stage.

Objective methods have been used to compare models of RMPFK built either mostly manually or mostly by computer. In general, the much slower manual design produced the better model although one of the slowest components, the complex loop modelling, seemed not to be worth the time outlay. Human intervention in the automatic design procedure would seem likely to improve the rather poor results. Inadequacies of the computer-only modelling method were particularly obvious at the alignment stage and during substitution of side chains, when manual relief of steric clashes would seem advantageous.

A variety of ligand design methods were tried. These do not span the full range of procedures under investigation since ligand design programs have great commercial value and are frequently not freely available. Two 2D ligand-based database searches were carried out. Fitting of molecular fragments to protein with LUDI was tried, but fitting of whole molecules from databases, as by the program DOCK, was not an approach that could be tested. Programs for the automatic generation of putative ligand structures, another important field into which energy is being directed, were also unavailable. The protein-ligand complex structure-based approach of ligand modification was, however, attempted.

Both 2D database searches were successful in producing interesting compounds. The small phosphocompounds tested as a result of the very rapid CCD database search proved to bind to PFK, though not with high affinity. The discovery that phospho-D-serine produced dissociation of *E.coli* PFK was of particular interest. It reveals an alternative available means of enzyme inhibition to the usual targeting of the active site. It remains to be seen whether suitable sites for rationally designed dissociative agents exist on other PFKs. A search of a larger database could be carried out. This would provide more ideas and/or enable the use of a more specific search structure.

The database search using NQS provided many ideas. Commercial availability limited

the number of ideas that could be tested to two but both these were active, one more so than NQS. It is worth remembering that NQS was suggested for testing as a result of its resemblance, noted by a chemist, to a LUDI compound being tested. This is evidence, depending on viewpoint, for the importance of either people or serendipity in the rational ligand design process.

None of the LUDI compounds tested could be shown to bind to the *E.coli* PFK 'ground state' (A66 probably binds to *E.coli* PFK at a later stage of the thermal denaturation pathway). This is at least partly expained by their small size, and maybe, in some cases, by the limits imposed by compound absorbance or solubility on the concentration that could be tested. In no cases were compounds derived from linked fragments (the original design concept) available for testing. Hence no conclusion can be drawn about the effectiveness of LUDI. What can be said is that the program runs very rapidly; large runs took less than an hour.

It is difficult to judge the effectiveness of the FBP analog design approach since none of the suggested compounds was available for testing. It was however noted that the results of the molecular mechanics procedure agreed with the predictions made by the GRID program. It can also be said that some of the derivatives modelled looked so favourable as to be very likely to be good inhibitors. There is circumstantial support for the quality of molecular modelling from the predictions made about the likely effectiveness of the available bisphosphates - the epoxides and compound 233. The generally pessimistic predictions were confirmed by experiment. This ligand design approach was by far the slowest. This must be taken into account in any future assessment of the method.

The question of species selectivity has been addressed. In the case of this model system of *E.coli* and RM PFKs, significant exploitable differences between the two were found only quite distant from the substrate binding sites. This meant that only one of the ligand design methods used, LUDI fragment fitting, was applicable to the species selectivity problem. This method, however, produced promising results.

Fragments that were highly complementary to regions differing in the two enzymes were found. Further fragments could be plausibly linked to these in order to occupy at least part of the substrate binding sites.

An important question is whether species selectivity between parasite and mammalian host would be easier or harder to obtain than selectivity between *E.coli* and RM PFKs. As noted (Section 1.2.2), parasites possess either an ATP-dependent PFK or a PP_i-dependent PFK. Species-selective drugs against the ATP-dependent enzyme would probably be harder to obtain than *E.coli* PFK inhibitors in the model system. Although no parasite ATP-dependent PFK sequences are known it is highly likely that they will more closely resemble mammalian PFKs than *E.coli* PFK does.

The situation with the PP_i-dependent parasite PFKs is more promising. The limited amount of sequence data available so far suggests that they are 40-60% identical to the bacterial PFKs. This means that a structural model would be easily built by homology with those enzymes. This would be expected to be of greater overall quality than the RMPFK structures which were produced here on the basis of 30-40% sequence identity. The F6P binding site of the PP_i-dependent PFK model could be targeted as in this thesis. Another attractive possibility is to target the PP_i binding site. PP_i analogs would be unlikely to harm the host since PP_i metabolism is not thought to be significant in mammals. Also, the PP_i binding site is likely to be significantly different to the ATP binding site of host PFKs. This is advantageous for species selective drug design, although it should be borne in mind that this fact simultaneously increases the likelihood of inaccuracies at the PP_i binding site in the homology model of the PP_i-dependent parasite PFK. PP_i analogs are the focus of current synthetic efforts as antiviral agents (Kluger and Huang, 1991). This chemical expertise will be useful if targeting of the PP_i binding site proves promising.

In conclusion, the question of species-specific rational drug design has been addressed using a PFK model system. Protein and ligand modelling studies have illustrated some of the possibilities and problems involved. It seems likely that studies

such as these will soon be able to expedite the discovery of effective anti-parasitic drugs. Such agents will be of great humanitarian value.

Chapter 6

Bibliography

Abagyan, R.A. (1993) *FEBS Letters* **325** 17-22.

Ahanotu, P.A., Ahanotu, E., Srinivasan, N.G. and Harris, B.G. (1991) *Molecular and Biochemical Parasitology* **45** 131-136.

Alberg, D.G. and Schreiber, S.L. (1993) *Science* **262** 248-250.

Allen, F.H., Bellard, S., Brice, M.D., Cartwright, B.A., Doubleday, A., Higgs, H., Hummelink, T.W.A., Hummelinkpeters, B.G., Kennard, O. and Motherwell, W.D. (1979) *Acta Crystallographica* **B35** 2331-2339.

Appelt, K., Bacquet, R.J., Bartlett, C.A., Booth, C.L.J., Freer, S.T., Fuhry, M.A.M., Gehring, M.R., Herrmann, S.M., Howland, E.F., Janson, C.A., Jones, T.R., Kan, C.-C., Kathardekar, V., Lewis, K.K., Marzoni, G.P., Matthews, D.A., Mohr, C., Moomaw, E.W., Morse, C.A., Oatley, S.J., Ogden, R.C., Reddy, M.R., Reich, S.H., Schoettlin, W.S., Smith, W.W., Varney, M.D., Villafranca, J.E., Ward, R.W., Webber, S., Webber, S.E., Welsh, K.M. and White, J. (1991) *Journal of Medicinal Chemistry* **34** 1925-1934.

Barton, G.J. and Sternberg, M.J.E. (1987) *Protein Engineering* **1** 89-94.

Bazan, J.F., Fletterick, R.J. and Pilkis, S.J. (1989) *Proceedings of the National Academy of Sciences* **86** 9642-9646.

Beddell, C.R., Goodford, P.J., Norrington, F.E., Wilkinson, S. and Wootton, R. (1976) *British Journal of Pharmacology* **57** 201-209.

Beddell, C.R., Goodford, P.J., Kneen, G., White, R.D., Wilkinson, S. and Wootton, R. (1984) *British Journal of Pharmacology* **82** 397-407.

Benner, S.A. and Gerloff, D.L. (1993) *FEBS Letters* **325** 29-33.

Berger, S.A. and Evans, P.R. (1991) *Biochemistry* **30** 8477-8480.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T and Tasumi, M. (1977) *Journal of Molecular Biology* **122** 535-542.

Biethinger, M., Hoffman, R. and Hofer, H.W. (1991) *Archives of Biochemistry and Biophysics* **287** 263-267.

Blangy, D. (1971) *Biochimie* **53** 135-149.

Blangy, D., Buc, H. and Monod, J. (1968) *Journal of Molecular Biology* **31** 13-35.

Blundell, T.L., Carney, D., Gardner, S., Hayes, F., Howlin, B., Hubbard, T., Overington, J., Singh, D.A., Sibanda, B.L. and Sutcliffe, M. (1988) *European Journal of Biochemistry* **172** 513-520.

Böhm, H.-J. (1992a) *Journal of Computer-Aided Molecular Design* **6** 61-78.

Böhm, H.-J. (1992b) *Journal of Computer-Aided Molecular Design* **6** 593-606.

Boobbyer, D.N.A., Goodford, P.J., McWhinnie, P.M. and Wade, R.C. (1989) *Journal of Medicinal Chemistry* **32** 1083-1094.

Bowie, J.U., Lüthy, R. and Eisenberg, D. (1991) *Science* **253** 164-170.

Brooks, S.P.J. and Storey, K.B. (1991) *Journal of Theoretical Biology* **149** 361-375.

Brünger, A.T. (1990) XPLOR Version 2.1 Manual.

Buckwitz, D., Jacobasch, G. and Gerth, C. (1990) *Molecular and Biochemical Parasitology* **40** 225-232.

Bueding, E. and Mansour, J.M. (1957) *British Journal of Pharmacology* **12** 159-165.

Buschmeier, B., Hengstenberg, W. and Deutscher, J. (1985) *FEMS Microbiology Letters* **29** 231-235.

Byberg, J.R., Jorgenson, F.S., Hansen, S. and Hough, E. (1992) *Proteins: Structure, Function and Genetics* **12** 331-338.

Cai, G.Z., Lee, L.L., Luther, M.A. and Lee, J.C. (1990) *Biophysical Chemistry* **37** 97-106.

Carlisle, S.M., Blakeley, S.D., Hemmingsen, S.M., Trevanion, S.J., Hiyoshi, T., Kruger, N.J. and Dennis, D.T. (1990) *Journal of Biological Chemistry* **265** 18366-

18371.

Casari, G. and Sippl, M.J. (1992) *Journal of Molecular Biology* **224** 725-732.

Chau, P.-L. and Dean, P.M. (1992a) *Journal of Computer-Aided Molecular Design* **6** 385-396.

Chau, P.-L. and Dean, P.M. (1992b) *Journal of Computer-Aided Molecular Design* **6** 397-406.

Chau, P.-L. and Dean, P.M. (1992c) *Journal of Computer-Aided Molecular Design* **6** 407-426.

Chiche, L., Gregoret, L.M., Cohen, F.E. and Kollman, P.A. (1990) *Proceedings of the National Academy of Sciences* **87** 3240-3243.

Clark, D.E., Willett, P. and Kenny, P.W. (1991) *Journal of Molecular Graphics* **9** 157-160.

Clark, D.E., Willett, P. and Kenny, P.W. (1992) *Journal of Molecular Graphics* **10** 194-204.

Clark, D.E., Willett, P. and Kenny, P.W. (1993) *Journal of Molecular Graphics* **11** 146-156.

Clark, M., Cramer, R.D. and Vanopdenbosch, N. (1989) *Journal of Computational Chemistry* **10** 982-1012.

Cohen, A.A. and Shatzmiller, S.E. (1993) *Journal of Molecular Graphics* **11** 166-173.

Colombo, G., Tate, P.W., Girotti, A.W. and Kemp, R.G. (1975) *Journal of Biological Chemistry* **250** 9404-9412.

Danzinger, D.J. and Dean, P.M. (1989a) *Proceedings of the Royal Society of London* **B 236** 101-113.

Danzinger, D.J. and Dean, P.M. (1989b) *Proceedings of the Royal Society of London* **B 236** 115-124.

Daum, G., Schmid, B., Mackintosh, C., Cohen, P. and Hofer, H.W. (1992) *Biochimica et Biophysica Acta* **1122** 23-32.

Davies, S.E.C. and Brindle, K.M. (1992) *Biochemistry* **31** 4729-4735.

DesJarlais, R.L., Seibel, G.L., Kuntz, I.D., Furth, P.S., Alvarez, J.C., Ortiz de Monellano, P.R., DeCamp, D.L., Babé, L.M. and Craik, C.S. (1990) *Proceedings of the National Academy of Sciences* **87** 6644-6648.

Deville-Bonne, D. and Else, A.J. (1991) *European Journal of Biochemistry* **200** 747-750.

Deville-Bonne, D. and Garel, J.-R. (1992) *Biochemistry* **31** 1695-1700.

Deville-Bonne, D., Le Bras, G., Teschner, W. and Garel, J.-R. (1989) *Biochemistry* **28** 1917-1922.

Deville-Bonne, D., Bourgain, F. and Garel, J.-R. (1991a) *Biochemistry* **30** 5750-5754.

Deville-Bonne, D., Laine, R. and Garel, J.-R. (1991b) *FEBS Letters* **290** 173-176.

Dunbrack, R.L. and Karplus, M. (1993) *Journal of Molecular Biology* **230** 543-574.

Eisenberg, D. and McLachlan, A.D. (1986) *Nature* **319** 199-203.

Evans, P.R. and Hudson, P.J. (1979) *Nature* **279** 500-504.

Evans, P.R., Farrants, G.W. and Hudson, P.J. (1981) *Philosophical Transactions of the Royal Society series B* **293** 53-62.

Evans, P.R. (1992) *Proceedings of the Robert A. Welch Foundation Conference on Chemical Research XXXVI Regulation of Proteins by Ligands.* 39-54.

Fersht, A. (1985) "Enzyme Structure and Mechanism" Chapter 11. W.H. Freeman and Co., New York.

Fine, R.M., Wang, H., Shenkin, P.S., Yarmush, D.L. and Levinthal, C. (1986) *Proteins: Structure, Function and Genetics* **1** 342-362.

Fothergill-Gilmore, L.A. and Michels, P.A.M. (1992) *Progress in Biophysics and Molecular Biology* **59** 105-235.

Fujita, T. (1990) *in* "Comprehensive Medicinal Chemistry" (Hansch, C., Sammes, P.G. and Taylor, J.B., series eds., Ramsden, C.A., vol. ed.), p497. Pergamon, Oxford and New York.

Fullick, A and Fullick, P. (1993) *New Scientist supplement* **137** 1-4 (20th Feb. 1993).

Furuya, E. and Uyeda, K. (1980) *Journal of Biological Chemistry* **255** 11656-11659.

Furuya, E. and Uyeda, K. (1981) *Journal of Biological Chemistry* **256** 7109-7112.

Gasteiger, J. and Marsali, M. (1980) *Tetrahedron* **36** 3219-3228.

Gething, M.-J. and Sambrook, J. (1992) *Nature* **355** 33-45.

Godzik, A., Kolinski, A. and Skolnick, J. (1992) *Journal of Molecular Biology* **227** 227-238.

Goodford, P.J. (1985) *Journal of Medicinal Chemistry* **28** 849-857.

Gottschalk, M.E., Latshaw, S.P. and Kemp, R.G. (1983) *Biochemistry* **22** 1082-1087.

Green, P.C., Tripathi, R.L. and Kemp, R.G. (1993) *Journal of Biological Chemistry* **268** 5085-5088.

Havel, T.F. and Snow, M.E. (1991) *Journal of Molecular Biology* **217** 1-7.

Havel, T.F. and Wüthrich, K. (1984) *Bulletin of Mathematical Biology* **46** 673-698.

Hellinga, H.W. and Evans, P.R. (1987) *Nature* **327** 437-439.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froscauer, R., Gottsbacher, K., Casari, G. and Sippl, M.J. (1990) *Journal of Molecular Biology* **216** 167-180.

Hesterberg, L.K., Lee, J.C. and Erickson, H.P. (1981) *Journal of Biological Chemistry* **256** 9724-9730.

Higgins, D.G. and Sharp, P.M. (1988) *Gene* **73** 237-244.

Hofer, H.W., Allen, B.L., Kaeini, M.R., Pette, D. and Harris, B.G. (1982) *Journal of Biological Chemistry* **257** 3801-3806.

Holm, L. and Sander, C. (1992) *Journal of Molecular Biology* **225** 93-105.

Huse, K., Jergil, B., Schwidop, W.-D. and Kopperschläger, G. (1988) *FEBS Letters* **234** 185-188.

INSIGHT II User Guide, version 2.1.0. San Diego: Biosym Technologies (1992).

Jakes, S.E. and Willett, P. (1986) *Journal of Molecular Graphics* **4** 12-20.

Jakes, S.E., Watts, N., Willett, P., Bawden, D. and Fisher, J.D. (1987) *Journal of Molecular Graphics* **5** 41-48.

Johnson, J.L. and Reinhart, G.D. (1992) *Biochemistry* **31** 11510-11518.

Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992) *Nature* **358** 86-89.

Jones, T.H. and Thirup, S. (1986) *EMBO Journal* **5** 819-822.

Kabsch, W. and Sander, C. (1983) *Biopolymers* **22** 2577-2637.

Kabsch, W., Mannherz, H.G., Suck, D., Pai, E.F. and Holmes, K.C. (1990) *Nature* **347** 37-44.

Kanaoka, M., Kishimoto, F., Ueki, Y. and Umeyama, H. (1989) *Protein Engineering* **2** 347-351.

Kemp, R.G., Fox, R.W. and Latshaw, S.P. (1987) *Biochemistry* **26** 3443-3446.

Kemp, R.G. and Foe, L.G. (1983) *Molecular and Cellular Biochemistry* **57** 147-154.

Kim, D.H. and Kim, K.B. (1991) *Journal of the American Chemical Society* **113**

3200-3202.

Kim, P.S. and Baldwin, R.L. (1990) *Annual Review of Biochemistry* **59** 631-666.

Kluger, R. and Huang, Z. (1991) *Journal of the American Chemical Society* **113** 5124-5125.

Köhler, P. (1985) *Molecular and Biochemical Parasitology* **17** 1-18.

Kohn, M.C. and Lemieux, D.R. (1991) *Journal of Theoretical Biology* **150** 3-25.

Kotlarz, D. and Buc, H. (1982) *Biochimica et Biophysica Acta* **484** 35-48.

Kulkarni, G., Rao, G.S.J., Srinivasan, N.G., Hofer, H.W., Yuan, P.M. and Harris, B.G. (1987) *Journal of Biological Chemistry* **262** 32-34.

Kuntz, I.D. (1992) *Science* **257** 1078-1082.

Ladror, U.S., Gollapudi, L., Tripathi, R.L., Latshaw, S.P. and Kemp, R.G. (1991) *Journal of Biological Chemistry* **266** 16550-16555.

Laskowski, R.A. (1993) *Journal of Applied Crystallography* **26** 283-291.

Latshaw, S.P., Bazaes, S., Randolph, A., Poorman, R.A., Heirikson, R.L. and Kemp, R.G. (1987) *Journal of Biological Chemistry* **262** 10672-10677.

Lau, F.T.-K., Fersht, A.R., Hellinga, H.W. and Evans, P.R. (1987) *Nature* **326** 811-812.

Lawrence, M.C. and Davis, P.C. (1992) *Proteins: Structure, Function and Genetics* **12** 31-41.

Le Bras, G. and Garel, J.-R. (1982) *Biochemistry* **21** 6656-6660.

Le Bras, G., Teschner, W., Deville-Bonne, D. and Garel, J.-R. (1989) *Biochemistry* **28** 6836-6841.

Lee, C.-P., Kao, M.C., French, B.A., Putney, S.D., Chang, S.H. (1987) *Journal of Biological Chemistry* **262** 4195-4197.

Lewis, R.A. and Dean, P.M. (1989a) *Proceedings of the Royal Society of London* **B 236** 125-140.

Lewis, R.A. and Dean, P.M. (1989b) *Proceedings of the Royal Society of London* **B 236** 141-162.

Lewis, R.A., Roe, D.C., Huang, C., Ferrin, T.E., Langridge, R. and Kuntz, I.D. (1992) *Journal of Molecular Graphics* **10** 66-78.

Loew, G. and Burt, S.K. (1990) *in* "Comprehensive Medicinal Chemistry" (Hansch, C., Sammes, P.G. and Taylor, J.B., series eds., Ramsden, C.A., vol. ed.), p105. Pergamon, Oxford and New York.

Lüthy, R., Bowie, J.U. and Eisenberg, D. (1992) *Nature* **356** 83-85.

MacArthur, M.W. and Thornton, J.M. (1991) *Journal of Molecular Biology* **218** 397-412.

Mahrenholz, A.M., Hefta, S.A. and Mansour, T.E. (1991) *Archives of Biochemistry and Biophysics* **288** 463-467.

Maiorov, V.N. and Crippen, G.M. (1992) *Journal of Molecular Biology* **227** 876-888.

Mansour, T.E. and Colman, R.F. (1978) *Biochemical and Biophysical Research Communications* **81** 1370-1376.

Martin, J.L., Veluraja, K., Ross, K., Johnson, L.J., Fleet, G.W.J., Ramsden, N.G., Bruce, I., Orchard, M.G., Oikonomakos, N.G., Papageorgiou, A.C., Leonidas, D.D. and Tsitoura, H.S. (1991) *Biochemistry* **30** 10101-10116.

Martin, Y.C. (1991) *Methods in Enzymology* **203** 587-613.

Mertens, E. (1991) *FEBS Letters* **285** 1-5.

Mertens, E. (1993) *Parasitology Today* **9** 122-126.

Miranker, A. and Karplus, M. (1991) *Proteins: Structure, Function and Genetics* **11** 29-34.

Moon, J.B. and Howe, W.J. (1991) *Proteins: Structure, Function and Genetics* **11** 314-328.

Monod, J., Wyman, J. and Changeux, J.P. (1965) *Journal of Molecular Biology* **12** 88-118.

Montgomery, J.A., Niwas, S., Rose, J.D. Secrist, J.A., Babu, Y.S., Bugg, C.E., Erion, M.D., Guida, W.C. and Ealick, S.E. (1993) *Journal of Medicinal Chemistry* **36** 55-69.

Morris, A.L., MacArthur, M.W., Hutchinson, E.G. and Thornton, J.M. (1992) *Proteins: Structure, Function and Genetics* **12** 345-364.

Moult, J. and James, M.N.G. (1986) *Proteins: Structure, Function and Genetics* **1** 146-163.

Nishibata, Y. and Itai, A. (1993) *Journal of Medicinal Chemistry* **36** 2921-2928.

Orchard, L.M.D. and Kornberg, H.L. (1990) *Proceedings of the Royal Society of London* **B 242** 87-90.

Patel, T. (1993) *New Scientist* **138** 11.

Pepperrell, C.A. and Willett, P. (1991) *Journal of Computer-Aided Molecular Design* **5** 455-474.

Pettigrew, D.W. and Frieden, C. (1979) *Journal of Biological Chemistry* **254** 1887-1895.

Pilkis, S.J., El-Maghrabi, M.R. and Claus, T.H. (1988) *Annual Review of Biochemistry* **57** 755-783.

Poorman, R.A., Randolph, A., Kemp, R.G. and Heinrikson, R.L. (1984) *Nature* **309** 467-469.

Ramachandran, G.N., Ramakris, C. and Sasisekh, V. (1963) *Journal of Molecular Biology* **7** 95-99.

Rapoport, T.A., Heinrich, R., Jacobasch, G. and Rapoport, S. (1974) *European Journal of Biochemistry* **42** 107-120

Romero, D.L., Busso, M., Tan, C.-T., Reusser, F., Palmer, J.R., Poppe, S.M., Aristoff, P.A., Downey, K.M., So, A.G., Resnick, L. and Tarpley, W.G. (1991) *Proceedings of the National Academy of Sciences* **88** 8806-8810.

Romero, D.L., Morge, R.A., Genin, M.J., Biles, C., Busso, M., Resnick, L., Althaus, I.W., Reusser, F., Thomas, R.C. and Tarpley, W.G. (1993) *Journal of Medicinal Chemistry* **36** 1505-1508.

Rotstein, S.H. and Murcko, M.A. (1993a) *Journal of Medicinal Chemistry* **36** 1700-1710.

Rotstein, S.H. and Murcko, M.A. (1993b) *Journal of Computer-Aided Molecular Design* **7** 22-43.

Rouvray, D. (1993) *New Scientist* **138** 35-38.

Rusinko, A., Skell, J.M., Balducci, R., McGarity, C.M. and Pearlman, R.S. (1988) "CONCORD, A Program for the Rapid Generation of High Quality Approximate 3-Dimensional Molecular Structures." Tripos Associates, St. Louis, Missouri.

Rypniewski, W.R. and Evans, P.R. (1989) *Journal of Molecular Biology* **207** 805-821.

Sander, C. and Schneider, R. (1991) *Proteins: Structure, Function and Genetics* **9** 56-68.

Schirmer, T. and Evans, P.E. (1990) *Nature* **343** 140-145.

Setlow, B. and Mansour, T.E. (1970) *Journal of Biological Chemistry* **245** 5524-5533.

Shirakihara, Y. and Evans, P.R. (1988) *Journal of Molecular Biology* **204** 973-994.

Smith-Brown, M.J., Kominos, D. and Levy, R.M. (1993) *Protein Engineering* **6** 605-614.

Srinivasan, N. and Blundell, T.L. (1993) *Protein Engineering* **6** 501-512.

Srinivasan, N.G., Wariso, B.A., Kulkarni, G., Rao, G.S.J. and Harris, B.G. (1988)

Journal of Biological Chemistry **263** 3482-3485.

Sternberg, M.J.E. and Islam, S.A. (1990) *Protein Engineering* **4** 125-131.

Stoichet, B.K., Stroud, R.M., Santi, D.V., Kuntz, I.D. and Perry, K.M. (1993) *Science* **259** 1445-1450.

Sudarsanam, S, Virca, G.D., March, C.J. and Srinivasan, S. (1992) *Journal of Computer-Aided Molecular Design* **6** 223-233.

Summers, N.L., Carlson, W.D. and Karplus, M. (1987) *Journal of Molecular Biology* **196** 175-198.

Summers, N.L. and Karplus, M. (1989) *Journal of Molecular Biology* **210** 785-811.

Summers, N.L. and Karplus, M. (1990) *Journal of Molecular Biology* **216** 991-1016.

Sutcliffe, M.J., Haneef, I., Carney, D. and Blundell, T.L. (1987a) *Protein Engineering* **1** 377-384.

Sutcliffe, M.J., Hayes, F.R.F. and Blundell, T.L. (1987b) *Protein Engineering* **1** 385-392.

SYBYL 5.5 Manual (1992). Tripos Associates, St. Louis, Missouri.

Teschner, W., Deville-Bonne, D. and Garel, J.-R. (1990a) *FEBS Letters* **2647** 96-98.

Teschner, W., Serre, M.C. and Garel, J.-R. (1990b) *Biochimie* **72** 403-406.

Teschner, W. and Garel, J.-R. (1989) *Biochemistry* **28** 1912-1916.

Thalhofer, H.P., Daum, G., Harris, B.G. and Hofer, H.W. (1988) *Journal of Biological Chemistry* **263** 952-957.

Topham, C.M., Thomas, P., Overington, J.P., Johnson, M.S., Eisenmenger, F. and Blundell, T.L. (1990) *Biochemical Society Symposia* **57** 1-9.

Topham, C.M., McLeod, A., Eisenmenger, F., Overington, J.P., Johnson, M.S. and Blundell, T.L. (1993) *Journal of Molecular Biology* **229** 194-220.

Torres, J.C. and Babul, J. (1991) *European Journal of Biochemistry* **200** 471-476.

Valaitis, A.P., Foe, L.G. and Kemp, R.G. (1987) *Journal of Biological Chemistry* **262** 5044-5048.

Varney, M.D., Marzoni, G.P., Palmer, C.L., Deal, J.G., Webber, S., Welsh, K.M., Bacquet, R.J., Bartlett, C.A., Morese, C.A., Booth, C.L.J., Herrmann, S.M., Howland, E.F., Ward, R.W. and White, J. (1992) *Journal of Medicinal Chemistry* **35** 663-676.

Verlinde, C.L.M.J., Pijining, T., Kalk, K.H., van Callenburgh, S., Van Aerschot, A., Herdewijn, P., Callens, M., Michels, P., Opperdoes, F.R., Wierenga, R.K. and Hol, W. (1994) Article in preparation.

von Itzstein, M., Wu, W.-Y., Kok, G.B., Pegg, M.S., Dyason, J.C., Jin, B., Van Phan, T., Smythe, M.L., White, H.F., Oliver, S.W., Colman, P.M., Varghese, J.N., Ryan, D.M., Woods, J.M., Bethell, R.C., Hotham, V.J., Cameron, J.M. and Penn, C.R. (1993) *Nature* **363** 418-423.

Warhurst, D.C. and Smith, H. (1992) *Parasitology Today* **8** 292-293.

Weng, L., Heinrikson, R.L. and Mansour, T.E. (1980) *Journal of Biological Chemistry* **255** 1492-1496.

Wilmot, C.M. and Thornton, J.M. (1988) *Journal of Molecular Biology* **203** 221-232.

Wilson, C., Gregoret, L.M. and Agard, D.A. (1993) *Journal of Molecular Biology* **229** 996-1006.

Wu, L.-F., Reizer, A., Reizer, J., Cai, B., Tomich, J.M. and Saier, M.H. (1991) *Journal of Bacteriology* **173** 3117-3127.

Yun, M.Y., Park, C.Y., Kim, S.S., Nam, D.Y., Kim, S.C. and Kim, D.H. (1992) *Journal of the American Chemical Society* **114** 2281-2282.

Zhao, Z., Malencik, D.A. and Anderson, S.R. (1991) *Biochemistry* **30** 2204-2216.

Appendix 1

A SYBYL macro to automatically set preferred side chain torsion angles. A sample of the data file pref.data follows the program to illustrate its format.

```
# A macro to add preferred side-chain angles to a selected range of a.a.s
# Values taken from Sutcliffe et al. Prot.Eng. 1987 1 385-392
# N.B. Modified to see if each residue is closer to alpha or beta
#
@MACRO
#
Prefauto sybylbasic
#
setvar start \
    %promptif("$1" "int" "" "Start" "Residue number of start")
setvar length \
    %promptif("$2" "int" "" "Length" "Number of residues to set")
uims v off
#
# Open data file
#
setvar f %open(pref.data "r")
#
# Ideal values for alpha and beta structure
#
setvar PHIalpha -57
setvar PSIalpha -47
setvar PHIbeta -129
setvar PSIbeta 124
setvar alphabeta "alpha" "beta"
#
# Enter loop to go through range
#
for res in %range("$start" %math($start + $length -1))
#
# See if closer to alpha or beta
#
setvar atom1 %atom_info(%cat(%subst_info(%math($res -1) name) ".C") id)
setvar atom2 %atom_info(%cat(%subst_info($res name) ".N") id)
setvar atom3 %atom_info(%cat(%subst_info($res name) ".Ca") id)
setvar atom4 %atom_info(%cat(%subst_info($res name) ".C") id)
setvar atom5 %atom_info(%cat(%subst_info(%math($res +1) name) ".N") id)
measure torsion $atom1 $atom2 $atom3 $atom4 >dummy |
setvar g %open(dummy "r")
```

```
setvar dummy %read($g)
setvar report %read($g)
setvar numstart %math(%pos("=" "$report") +2)
setvar numend  %math(%pos("Deg" "$report") -2)
setvar PHIres %substr("$report" $numstart %math($numend - $numstart +1))
%close($g)
measure torsion $atom2 $atom3 $atom4 $atom5 >dummy  |
setvar g %open(dummy "r")
setvar dummy %read($g)
setvar report %read($g)
setvar numstart %math(%pos("=" "$report") +2)
setvar numend  %math(%pos("Deg" "$report") -2)
setvar PSIres %substr("$report" $numstart %math($numend - $numstart +1))
%close($g)
setvar deltaalpha %sqrt(%math(%pow(%math($PHIres - $PHIalpha) 2) + \
                    %pow(%math($PSIres - $PSIalpha) 2)))
setvar deltabeta %sqrt(%math(%pow(%math($PHIres - $PHIbeta) 2) + \
                   %pow(%math($PSIres - $PSIbeta) 2)))
if %lt($deltaalpha $deltabeta)
  setvar ab "a"
else
  setvar ab "b"
endif
#
# What to look for
#
setvar lookfor %cat(%substr(%subst_info(%cat("{" "$res" "}") name) 1 3) $ab)
#
# Enter loop to look for it
#
%rewind($f)
while %not(%streql($lookfor %substr(%read($f) 1 4)))
endwhile
#
# Read number of angles to set,skipping if 0 i.e. ALA,GLY.
#
setvar noang %read($f)
if %not(%eq(0 $noang))
#
# Enter loop to alter angles
#
echo "Setting" %subst_info($res name) "to preferred"\
    %arg(%pos($ab "ab") $alphabeta)  "angles"
for setloop in %range(1 $noang)
BIO SET CONFORMATION %subst_info($res name) %cat("chi" "$setloop" "="
"%read($f)")
endfor
#
```

```
endif
#
endfor
%close($f)


ALAa
0
ALAb
0
ARGa
4
-69
175
174
-75
ARGb
4
-168
176
-177
-176
ASNa
2
-70
-49
ASNb
2
-70
-49
ASPa
2
-69
-3
ASPb
2
-69
-3
CYSa
1
176
CYSb
1
176
```

Appendix 2

Three SYBYL macros which, when run sequentially, emulate the loop search technique of Summers and Karplus, 1990.

```
@MACRO
#
danloop1 sybylbasic
#
setvar runname \
%promptif("$1" "string" "" "Loop run name"  \
"Run name e.g. loop1 to find loop1_1,loop1_13 etc.")
#
setvar gapstart \
%promptif("$2" "int" "" "Gap start" "Residue number of start of gap")
#
setvar gapend \
%promptif("$3" "int" "" "Gap end" "Residue number of end of gap")
#
setvar database \
%promptif("$4" "string" "newloop.mdb" "Database" \
"Name of the database containing the loops and model")
#
setvar model \
%promptif("$5" "string" "model_n_20_3" "Name of model" \
"Name of model protein including region to replace")
#
setvar nanchlength \
%promptif("$6" "int" "3" "N_anchor length" \
"Length of the N-terminal anchor region")
#
setvar canchlength \
%promptif("$7" "int" "3" "C_anchor length" \
"Length of the C-terminal anchor region")
#
setvar loopseq \
%promptif("$8" "string" "val=asp=gly=gly=asp=his" "Loop sequence" \
"Sequence of loop INCLUDING anchors in format XXX=YYY=...")
#
# Open output file
#
setvar f %open(%cat("danloop_" $runname ".out") "w+")
#
data open $database update
```

```
# Exit immediately if no matches
setvar startdb %database(%cat($runname "*"))
setvar noloops %count($startdb)
if %eq(0 $noloops)
 %write($f "No matches found to given loop run name")
 %close($f)
 return
endif
zap *
#
# Set minimisation conditions
#
tailor set maximin2 min_energy_change 0 ||
tailor set maximin2 minimization_method bfgs | |
tailor set force_field parameter_set koll_all ||
tailor set force_field one .5 ||
#
# Get model
#
data get $model m1
default m1
setvar gapminus %subst_info(%math($gapstart -1) name)
setvar gapplus %subst_info(%math($gapend +1) name)
#
# Get the six constraints on minimisation. Use just end-to-end in first
# round BUT all six in second round
#
setvar sourceatom "%cat($gapminus ".C")" "%cat($gapminus ".N")" \
          "%cat($gapminus ".CA")" "%cat($gapminus ".CA")" \
          "%cat($gapminus ".N")" "%cat($gapplus ".CA")"
setvar destatom  "%cat($gapplus ".N")" "%cat($gapplus ".CA")" \
          "%cat($gapplus ".CA")" "%cat($gapplus ".C")" \
          "%cat($gapminus ".CA")" "%cat($gapplus ".C")"
setvar distances %open("dummydist" "w+")
for finddistance in %range(1 6)
setvar atom1 %arg($finddistance $sourceatom)
setvar atom2 %arg($finddistance $destatom)
setvar deltax %math(%atom_info($atom1 x) - %atom_info($atom2 x))
setvar deltay %math(%atom_info($atom1 y) - %atom_info($atom2 y))
setvar deltaz %math(%atom_info($atom1 z) - %atom_info($atom2 z))
setvar diff %sqrt(%math(%pow($deltax 2) + %pow($deltay 2) + %pow($deltaz 2)))
%write($distances $diff)
endfor
%rewind($distances)
setvar endtoend %read($distances)
setvar further1 %read($distances)
setvar further2 %read($distances)
setvar further3 %read($distances)
```

```
setvar further4 %read($distances)
setvar further5 %read($distances)
%close($distances)
#
# Enter outside loop to cycle through all matches to loop run name
#
setvar nopassed 0
for loopcount in %range(1 $noloops)
setvar currloop %arg($loopcount $startdb)
data get $currloop m2
default m2
setvar loopstart %subst_info(%arg(1 %substs(*)) name)
setvar loopend   %subst_info(%arg(%count(%substs(*)) %substs(*)) name)
remove atom {sidechain}-*.cb
bio addh * essential_only
#
# Do first minimisation round
#
default m2
define constraint distance %cat($loopstart ".C") %cat($loopend ".N") \
                $endtoend 600
tailor set maximin2 maximum_iterations 250 ||
tailor set anneal hot 0 int 0 ||
anneal m2 m2(*) continue done interactive
#
# Enter second round of minimisation with extra five constraints to fix
# geometry of ends of loops
#
define constraint distance %cat($loopstart ".N") %cat($loopend ".CA") \
                $further1 250
define constraint distance %cat($loopstart ".CA") %cat($loopend ".CA") \
                $further2 250
define constraint distance %cat($loopstart ".CA") %cat($loopend ".C") \
                $further3 250
define constraint distance %cat($loopstart ".N") %cat($loopstart ".CA") \
                $further4 250
define constraint distance %cat($loopend ".CA") %cat($loopend ".C") \
                $further5 250
tailor set maximin2 maximum_iterations 200 ||
anneal m2 m2(*) continue done interactive
#
# Now do fit and get RMSins
#
fit %cat("m1(" $gapminus ".CA")") %cat($loopstart ".CA") \
            %cat("m1(" $gapminus ".C")")  %cat($loopstart ".C") \
            %cat("m1(" $gapplus ".N)") %cat($loopend ".N") \
            %cat("m1(" $gapplus ".CA)") %cat($loopend ".CA") \
| solve exit >dummy
```

```
setvar g %open(dummy "r")
for dummy in %range(1 20)
setvar junk %read($g)
endfor
setvar int %read($g)
%close($g)
setvar nstart %math(%pos(":" "$int") + 1)
setvar rms %substr("$int" $nstart 9)
#
# Report to output file on rms and add minimised, repositioned loop
# to the database if rms<=.25 A
#
if %lteq($rms 0.25)
 %write($f "")
 %write($f "Loop : " $currloop " passes RMSins test with value" $rms)
 setvar suffix "_pass"
 setvar nopassed %math($nopassed +1)
else
 %write($f "")
 %write($f "Loop : " $currloop " fails RMSins test with value" $rms)
 setvar suffix "_fail"
endif
setvar newname %cat($currloop "_rmsins" $suffix)
%write($f "New database entry is : " $newname)
remove atom *.cb
remove atom %cat($loopstart ".H")
remove atom  %cat($loopstart ".N")
remove atom %cat($loopstart ".CA")
remove atom %cat($loopend ".O")
remove atom %cat($loopend ".C")
remove atom %cat($loopend ".CA")
modify molecule name m2 $newname
data add m2
#
# Create parameter file for this loop to run danloop2
#
setvar pars %open(%cat("danloop_" $currloop ".pars") "w")
%write($pars $newname)
%write($pars $gapstart)
%write($pars $gapend)
%write($pars $database)
%write($pars $model)
%write($pars $nanchlength)
%write($pars $canchlength)
%write($pars $loopseq)
%write($pars $endtoend)
%close($pars)
endfor
```

```
#
%close()
echo
echo "Now run danloop2 ..."
.


@macro
danloop2 sybylbasic
#
# A macro to do Emin and Esub tests given output of danloop1
#
setvar parsource \
    %promptif("$2" "string" "danloop.pars" "File containing danloop1 parameters"
"Filename in format danloop.parsX where X is the loop number")
setvar pars %open($parsource "r")
setvar currloop %read($pars)
setvar gapstart %read($pars)
setvar gapend %read($pars)
setvar database %read($pars)
setvar model %read($pars)
setvar nanchlength %read($pars)
setvar canchlength %read($pars)
setvar loopseq %read($pars)
setvar endtoend %read($pars)
%close($pars)
setvar gapplus %math($gapend +1)
setvar gapminus %math($gapstart -1)
tailor set merge consider_uniq no ||
data open $database update
#
# First get model and remove gap
# first gap subs only
#
data get $model m2
default m2
remove substructure %cat("{" %subst_info($gapstart name) ":" \
                %subst_info($gapend name) "}")
modify molecule name m2 tempminus
data add m2 replace
#
# Then - further atoms (see paper)
#
setvar gapminusname %subst_info($gapminus name)
setvar gapplusname %subst_info($gapplus name)
remove atom %cat($gapminusname ".O")
remove atom %cat($gapminusname ".C")
remove atom %cat($gapplusname ".N")
```

```
modify molecule name m2 tempminusminus
data add m2 replace
#
uims load prefauto.macro
#
#
# Set minimisation conditions
#
tailor set maximin2 min_energy_change 0 ||
tailor set maximin2 minim bfgs ||
tailor set force_field parameter_set koll_all ||
tailor set anneal hot 0 int 12 ||
#
# Get loop into m2
#
data get $currloop m2
default m2
setvar loopstart %subst_info(%arg(1 %substs(*)) name)
setvar loopend   %subst_info(%arg(%count(%substs(*)) %substs(*)) name)
setvar lsno %substr($loopstart 4 3)
setvar leno %substr($loopend 4 3)
setvar lsplusno %math($lsno +1)
setvar leminusno %math($leno -1)
setvar lsid %subst_info($loopstart id)
setvar leid %subst_info($loopend id)
setvar lscid %atom_info(%cat($loopstart ".C") id)
setvar lenid %atom_info(%cat($loopend ".N") id)
#
# Change loop residues to correct sequence - this adds side-chains too
#
bio change %cat(%subst_info($lsplusno name) ":" %subst_info($leminusno name))
\
       $loopseq
#
# Fix any prolines
#
setvar fixstart %substr("$lsid" 2 4)
setvar fixend %substr("$leid" 2 4)
for fixloop in %range(%math($fixstart +1) %math($fixend -1))
setvar fixname %subst_info(%cat("#" $fixloop) name)
if %pos("PRO" $fixname)
 bio fix_proline $fixname
endif
endfor
#
#  - for anchors take conformation from intact model got into m3
#
data get $model m3
```

```
for nanchloop in %range(1 $nanchlength)
default m3
setvar copyuse %cat("m3(" %subst_info(%math($gapminus + $nanchloop) name)
")")
default m2
 bio copy $copyuse \
       %cat("m2(" %subst_info(%math($lsno + $nanchloop) name) ")") \
       chi1,chi2,chi3,chi4
endfor
for canchloop in %range(1 $canchlength)
default m3
setvar copyuse %cat("m3(" %subst_info(%math($gapplus - $canchloop) name) ")")
default m2
 bio copy $copyuse \
       %cat("m2(" %subst_info(%math($leno - $canchloop) name) ")") \
   chi1,chi2,chi3,chi4
endfor
#
#  - for others use prefauto
#
uims v off
prefauto %math($lsno + $nanchlength +1) %math($leminusno - $lsplusno -2)
uims v on
#
# Make dummy copy number +100 in m3 for merging in danloop3
#
copy m2 m3
default m3
for makedummy in %range(1 %math($leno - $lsno +1))
modify substructure name %subst_info(%cat("#" $makedummy) name)
%cat("dummy" \
                %math($lsno + $makedummy +99))
endfor
setvar noatoms %mol_info(m3 natoms)
#
# Now merge in model - correct atoms (from tempminusminus) so that we are
# minimising the loop in the field of the molecule
#
data get tempminusminus m1
merge m1(*) m2
default m2
scan %cat("{sidechain}&{" $loopstart ":" $loopend "}")
#
# Save parameters for danloop3
#
%close()
setvar pars  %open("temp.pars" "w")
%write($pars $noatoms)
```