Expert system for protein engineering: its application in the study of chloramphenicol acetyltransferase and avian pancreatic polypeptide

Barry Robson, Eric Platt, Robert V Fishleigh, Alan Marsden and Peter Millard

Epsitron Peptide and Protein Engineering Research Unit, Department of Biochemistry and Molecular Biology, The University, Oxford Road, Manchester M13 9PT, UK

Lucifer is a suite of programs for the conformational study of drugs, proteins and other biomolecules. The overall architecture and operation of the suite is outlined with worked examples. New procedures are also described for the identification of secondary structure template similarity based on theorem-proof algebra and for the rapid evaluation of the hydrophobic packing of a protein conformation. These are discussed in relation to the molecular-graphics modelling of chloramphenicol acetyltransferase. Although Lucifer is of proven worth for the study of biological peptides and for protein modelling against homologues, its applicability to de novo *protein structure prediction is untested. A preliminary study of avian pancreatic polypeptide is of this character and is described here. The results of the above attempts are both informative and promising.*

Keywords: chemical structure + prediction, protein structure, chloramphenicol acetyltransferase, avian pancreatic polypeptide, homology, molecular graphics, template pattern matching, hydrophobic packing, expert systems

Received 13 October 1986
Accepted 21 October 1986

It is widely recognized that the ability to design and redesign proteins rationally is essential for the fullest development of biotechnology (see, for example, Reference 1). It may be that the development of the mathematical and computational technologies required to permit this will be seen, in hindsight, as representing one of the greatest scientific challenges confronted in the last quarter of the Twentieth Century.

Essentially, this challenge is one of facing, answering and acting on three questions. First, is the problem of predicting the 3D structure of proteins and other complex biomolecules *de novo* one of NP completeness? That is, is there an intractable combinatorial explosion in the computation involved? Second, can the intelligent analysis of the data provided by molecular biology, genetic engineering and experimental conformational studies be routinely used to overcome fundamental difficulties? Third, if our capabilities in both of the above areas are restricted at present, can the human intervenor be more objectively involved, through media such as molecular graphics, to give us an edge on the problem? That is, can the human brain contribute anything more to the problem than simply designing the hardware and software used?

DEVELOPMENT OF THE LUCIFER SUITE

Based on a recognition of these problems over a decade ago, the Lucifer suite has evolved from the early investigations that this group made into the field of the computer-aided design of complex biomolecules. Lucifer stands for: Logical Use of Conformational Information and Fast Energy Routines. As the name suggests, the suite explicitly seeks to carry out global energy minimization as rapidly as possible[2] but puts great emphasis on the rational use of external data to overcome any deficiencies. This external data is of three basic types. The first comes from the scanning of databases to identify primary, secondary, supersecondary and tertiary structural homologies. Second, experimental data from physical chemists and pharmacologists is analysed and exploited. Such data is exemplified by intergroup distances from NMR spectroscopy, infrared spectral data, characteristic ratios of polymers, net dipole moments and the pharmacological potencies of structural analogues. Third, the Immam graphics system developed by this group provides a vehicle for more qualitative judgements. The value of this system has been illustrated in the modelling of immunoglobulins against electron microscopic and other experimental data[3-5].

Although a considerable portion of the program is dedicated to interfacing Lucifer with the more specialized peripheral programs, the core routines are primarily concerned with optimization problems — the search for the lowest energy conformation, or the optimization of a conformational model's fit to a homologous protein, in whole or in part. The use of the core routines is interesting. Some 50 'primitive' commands cover the fundamental operations required for the manipulation of molecules and the data relating to them. While there is great flexibility in the operation of the program, the software checks that the overall protocol of command use is sensible. Complex protocols composed of a number of these primitive commands can be developed for performing a series of operations. While some such protocols are 'experimental' and may prove to be inefficient, others are found to be of general use and can be called as 'master' commands by a given name. The availability of established master commands enables the inexperienced user to operate the program productively by calling tried and trusted protocols such as *BUILD (sequence) and *MINIMISE, both of which invoke a fairly complex series of operations. At the same time, the flexible, language-like command structure, which includes IF and GOTO type constructs, permits the practised user to investigate novel protocols freely.

Procedures of the automatic program that are new or otherwise not described by Robson and Platt[2] include novel methods for quadratic extrapolation, the extension of Euclidean space into a fourth dimension in order to tunnel through energy barriers, and 'node coupled trajectory programming', as reviewed by Robson and Garnier[5]. Of particular interest here is the Globex method, which rests on an upgraded version of the Nelder–Mead Simplex method[6]. Whereas Simplex will readily cross frictional barriers, which are often transparent to it, it may still become entrapped in very deep minima. The Globex routine uses a stack of conformations and energies corresponding to deep minima found so far. The action on entrapment is to generate a new simplex of $n + 1$ points from the last and lowest energy minima (up to n) in this stack, with the remaining conformers generated at random. The subsequent Simplex operations naturally provide the best chance of locating a new deep minimum, implicitly in the most promising direction. This turns out to be very valuable and is now a Lucifer default.

An equally important option used in the present study is to initialize the empty stack at the beginning of the simulation with conformers combinatorially assembled from the conformations of homologous segments in the database of proteins of known structure. It is important that the conformations are locally minimized before they are placed on the stack, otherwise the conformational information content may be masked by bad clashes. The specific and particular mode of use in the case of avian pancreatic polypeptide is described later.

The apparent preference for Simplex-type methods is simply due to the fact that these, in the past, have been found to perform significantly better than the alternatives. On the other hand, in-house advances in molecular dynamics have made this a very fast and extremely interesting support option. Such advances include the damping out of the fast modes in such a way as to permit longer time steps without interfering with evaluated thermodynamic properties. Although these all form part of our protocol, the automatic methods used here are, for simplicity, confined to Globex (see below), a gradient method to refine deep minima, and the zero minimum and r'-space calculation[2]. Other methods were also used in pilot studies. The less automatic method, described first for comparison, combined the use of molecular graphics and database interrogation. The merits and difficulties of these two Lucifer options are discussed. The results are interim, in that the molecular dynamics procedures will ultimately be applied to both to produce a final result. In this respect, it must also be noted that the molecular dynamics routine used by us is fast and robust and, from previous experience, is known to lead to improved conclusions, although dependent on the above for the first study phase.

MOLECULAR GRAPHICS MODELLING OF CHLORAMPHENICOL ACETYLTRANSFERASE

Background

Chloramphenicol acetyltransferase (CAT) (EC 2.3.1.28) catalyses the O-acetylation of the antibiotic chloramphenicol (CM), leading to its inactivation. CAT uses acetyl coenzyme A (acetyl CoA) as the acyl donor in this reaction, which occurs in both Gram-negative and Gram-positive bacteria.

The entire amino-acid sequences of five CAT forms have been determined — the plasmid-borne CAT I[7–9] and CAT III[10] variants from *Escherichia coli*, two *Staphylococcus aureus* forms from plasmids pC194[11] and pC221[12], and the chromosomal CAT from *Bacillus pumilis*[13].

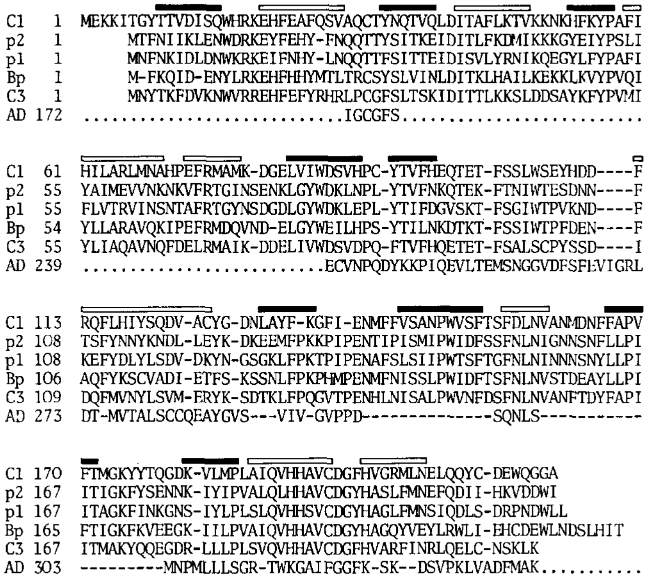
Studies on the kinetics of the type III variant have indicated that the reaction proceeds by a ternary-complex mechanism[14], with a central role in the catalytic mechanism of CAT implicated for His-189[15]. A high level of sequence conservation across the five variants is noted in the region of this histidine, with site-directed mutagenesis of Asp-199 to asparagine in the type I variant having a greatly adverse effect on enzyme activity[16], suggesting a possible catalytic role for Asp-199.

Evidence has also been obtained to suggest that Cys-31 of the type I form may be at or near the active site, but not necessarily involved in the catalytic activity of CAT[10]. It can be noted from Figure 1 that this cysteine is not conserved in the pC194 and pC221 forms, although there is another well conserved cysteine shortly after the putative catalytic histidine.

Lysine-136 of CAT I and Lys-38 of CAT III were found to be unreactive to amidination with methyl acetimidate, and it was suggested that these residues could be buried and perhaps involved in salt-bridge formation with residues from other subunits[17]. Until recently it was believed that CAT formed a tetramer of identical subunits, but X-ray crystallographic studies have indicated a trimeric quaternary structure[18]. The multimer has been shown to be highly stable[17,19], and heteromeric hybrids of the type I and III forms have been produced *in vitro*[19].

Using *in vitro* DNA methods, a 13 amino-acid insertion has been made between residues 71 and 72 of CAT I, with retention of up to 15% of the wild-type enzyme activity[20].

In the present study, an attempt has been made to predict the structure of the trimer using a combination



```
C1   1  MEKKITGYTTVDISQWHRKEHFEAFQSVAQCTYNQTVQLDITAFLKTVKKNKHFKYPAFI
p2   1       MTFNIIKLENWDRKEYFEHY-FNQQTTYSITKEIDITLFKDMIKKKGYEIYPSLI
p1   1       MNFNKIDLDNWKRKEIFNHY-LNQQTTFSITTEIDISVLYRNIKQEGYLFYPAFI
Bp   1       M-FKQID-ENYLRKEHFHHYMTLTRCSYSLVINLDITKLHAILKEKKLKVYPVQI
C3   1       MNYTKFDVKNWVRREHFEFYRHRLPCGFSLTSKIDITTLKKSLDDSAYKFYPVMI
AD 172  ............................IGCGFS..........................

C1  61  HILARLMNAHPEFRMAMK-DGELVIWDSVHPC-YTVFHEQTET-FSSLWSEYHDD----F
p2  55  YAIMEVVNKNKVFRTGINSENKLGYWDKLNPL-YTVFNKQTEK-FTNIWTESDNN----F
p1  55  FLVTRVINSNTAFRTGYNSDGDLGYWDKLEPL-YTIFDGVSKT-FSGIWTPVKND----F
Bp  54  YLLARAVQKIPEFRMDQVND-ELGYWEILHPS-YTILNKDTKT-FSSIWTPFDEN----F
C3  55  YLIAQAVNQFDELRMAIK-DDELIVWDSVDPQ-FTVFHQETET-FSALSCPYSSD----I
AD 239  ..................ECVNPQDYKKPIQEVLTEMSNGGVDFSFEVIGRL

C1 113  RQFLHIYSQDV-ACYG-DNLAYF-KGFI-ENMFFVSANPWVSFTSFDLNVANMDNFFAPV
p2 108  TSFYNNYKNDL-LEYK-DKEEMFPKKPIPENTIPISMIPWIDFSSFNLNIGNNSNFLLPI
p1 108  KEFYDLYLSDV-DKYN-GSGKLFPKTPIPENAFSLSIIPWTSFTGFNLNINNNSNYLLPI
Bp 106  AQFYKSCVADI-ETFS-KSSNLFPKPHMPENMFNISSLPWIDFTSFNLNVSTDEAYLLPI
C3 109  DQFMVNYLSVM-ERYK-SDTKLFPQGVTPENHLNISALPWVNFDSFNLNVANFTDYFAPI
AD 273  DT-MVTALSCCQEAYGVS---VIV-GVPPD--------------SQNLS-----------

C1 170  FTMGKYYTQGDK-VLMPLAIQVHHAVCDGFHVGRMLNELQQYC-DEWQGGA
p2 167  ITIGKFYSENNK-IYIPVALQLHHAVCDGYHASLFMNEFQDII-HKVDDWI
p1 167  ITAGKFINKGNS-IYLPLSLQVHHSVCDGYHAGLFMNSIQDLS-DRPNDWLL
Bp 165  FTIGKFKVEEGK-IILPVAIQVHHAVCDGYHAGQYVEYLRWLI-EHCDEWLNDSLHIT
C3 167  ITMAKYQQEGDR-LLLPLSVQVHHAVCDGFHVARFINRLQELC-NSKLK
AD 303  ---------MNPMLLLSGR-TWKGAIFGGFK-SK--DSVPKLVADFMAK..........
```

Figure 1. Alignment of five CAT sequences (type I, C1; pC221, p2; pC194, p1; Bacillus pumilis, *Bp; type III, C3) with horse E alcohol dehydrogenase (AD). The predicted secondary structures of CAT are indicated (hollow box = helix, solid box = extended chain, blank = aperiodic structure). Deletions are indicated by a dash and regions of alcohol dehydrogenase unalignable with CAT are shown by a full stop*

of techniques for secondary structure prediction, homology detection and molecular-graphics modelling. Consideration has also been given to the structural information suggested by the experimental studies summarized above. The work reported here was first presented at the SERC Protein Prediction Meeting at Daresbury, January 1986, and more fully at the 5th International Meeting of the Molecular Graphics Society, Cap d'Agde, April 1986. Since then, the preliminary results from a crystallographic study of CAT III have become available[18]. The predicted structure is compared with the low-resolution crystal data, and the molecular-graphics approach to the modelling of larger systems is critically evaluated.

Protein structure database scan

The first stage in the analysis of the CAT sequences was to scan the database of proteins of known structure for members bearing sequence homology to CAT, in order to identify a protein with close sequence homology which could be used as the starting point for the molecular-graphics modelling.

The bit-pattern method of homology detection[21] was used in which a search is made for chain sections with similar patterns of hydrophobic and hydrophilic residue character. Using this method, a number of proteins bearing weak homology to CAT were identified, as exemplified by horse E alcohol dehydrogenase (ADH) (EC 1.1.1.1.)[22] and dihydrofolate reductases (EC 1.5.1.3.) from various sources. As can be seen from Figure 1, the sequence homology between ADH and CAT III is rather limited. In view of the length of the insertion regions and the poorer sequence homology between the other CAT variants and ADH, the observed structure of ADH was not selected as the starting point for molecular-graphics modelling.

Secondary structure prediction and template matching

The Garnier *et al.* method[23] and a computerized version of the Lim method[24] were used for the secondary-structure prediction. Predictions were made separately on each of the five sequences, and a procedure was then applied to derive a consensus prediction of potentially enhanced reliability[21] (Figure 1).

The secondary-structure template obtained in this manner was used to probe a database of templates derived from the crystal structures*. A close but not perfect agreement was found between the templates of CAT and cat muscle pyruvate kinase (PYK) (EC 2.7.1.40) domains A and B[25]. These domains correspond to the first half of the β-barrel domain and the entire Greek-key domain respectively, which are consecutive in the polypeptide chain.

It is important to allow for the imperfections of the predictive method when attempting to identify template pattern similarity. Although the availability of a number of homologous sequences can increase the reliability of the template predicted, there is still a significant chance that a whole strand of helix or β-strand might be missed or incorrectly predicted. For this reason, an optimal match, taking into account the strengths of predictions in various regions, must be sought, rather than simply an apparent 'perfect' match.

If a close match is not forthcoming, it may be necessary to search for more weakly homologous proteins, in which case it is important to allow for the possibility that the protein of interest contains a novel combination of supersecondary-structure motifs. In this case, a partial template match with one or more proteins is the best that can be expected, and so the template-matching procedure used must be able to analyse the supersecondary-structure composition of larger proteins and determine which elements may be consistent with the predicted template.

Such a method for template matching should operate on three premises:

- that some points of agreement or disagreement between the templates of the protein of interest and of those in the database are more important than others,
- that, while the supersecondary-structure motifs of two proteins may not be identical, they may be related to the motif of a common ancestral protein,
- that, even if the similarity is not obvious, the rules under which evolution operates mean that the motifs may still be related.

The principles of theorem-proof algebra can be used as the basis for a template-matching procedure that can take account of these points in its operation.

Briefly, the method involves writing a general secondary-structure formula describing the supersecondary-structure arrangement, including reference to interacting component elements. If helices are represented by *a*, β strands by *b*, and *x* corresponds to a strand of aperiodic structure, then a β hairpin can be written as *bb*, two parallel strands as *bxb*, and a Rossman fold element as *bab*. A more complex example is immu-

*Robson, B, Fishleigh, R V and Garnier, J, unpublished noglobulin V, in which the seven β-strands comprising the Greek key unit can be expressed as:

$$b(+5)/b^5/b(-5)$$

It is implied that, within the unit b^5, the adjacent strands are in contact with one another in an antiparallel arrangement. The use of a '/' symbol explicitly indicates that such an arrangement is not found between the adjacent strands, as between the sixth and seventh strands in this example. Longer range interactions may be specified by the numbers in brackets. In this example, '$b(-5)$' means that the strand is hydrogen bonded to the strand five strands away in the N-terminal direction.

Once the supersecondary-structure formulas have been determined, they can then be manipulated according to the principles of theorem-proof algebra. Each formula is a 'theorem' which can be related to another by the sequential application of transforming rules in a manner analogous to that by which one theorem is proven by another. That is, there is a finite number of valid transforming rules that prove formula D from formula A, because they convert A to B, then B to C and finally C to D. In the case of protein supersecondary structure, the transforming rules consistent with evolution are obtained by analysing the structural relationships of proteins of known conformation.

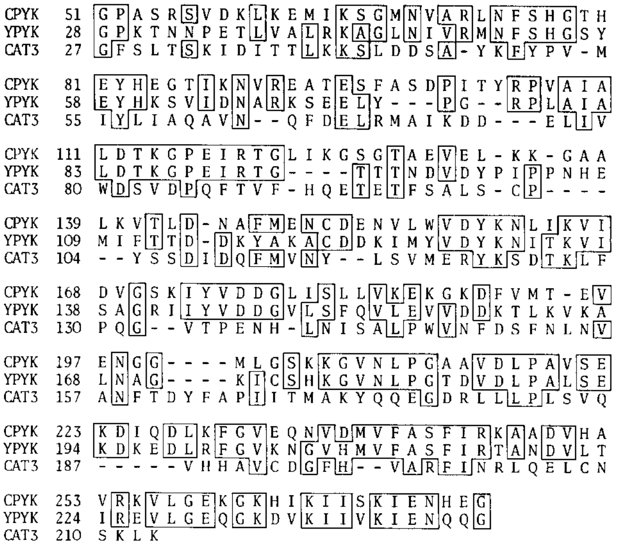
The rules can be illustrated in the following example. The jelly roll Greek key of tomato bushy stunt virus protein domain 3 with the formula:

$$b(+5)/b(+3)/b^3/b(-3)/b/(-5)$$

can be derived from the seven-stranded Greek key of immunoglobulin V in one step. This is because the sequence b^5 of five antiparallel β strands has entrance and exit strand directions which are the same as in the $b(+3)/b^3/b(-3)$ case. In this sense the topography of the two structures is consistent.

In making the transformations, it is important to appreciate the significance of the changes being made. A step that changes the connectivity of the strands introduces a relatively major topographical distortion into the structure. If two algebraic formulas can be related only by making such a change, it implies a fundamental difference in the supersecondary structures of the two proteins concerned.

Less distortion in the overall motif is introduced by a step that changes the number of strands in the structure, providing that the change in numbers is not too great. For example, if $b(+n)b^{n-1}b(-n)$ represents a hypothetical 'ideal' antiparallel β barrel, removing strands such that n is below a critical number would force a Greek key pleated sheet $b(+n)b^{n-1}/b(-n)$. This would imply a major difference between the spatial organization of the motifs. In contrast, the two barrels of chymotrypsin are considered related because the formula for one can be derived simply from the other by substituting bxb with b. This preserves the direction and connectivity of the strands and does not reduce n below a critical number of strands required for a barrel. It is possible to apply a sequence of allowable transformations to one complex motif and hence to relate it to another, despite the fact that there is no other apparent relationship. By the rules, the complex intertwining of two β sheets generates two separate sheets by transformation rules, and the two sheets can then be shown algebraically to be related by further simple rules[5].



```
CPYK  51 G P A S R S V D K L K E M I K S G M N V A R L N F S H G T H
YPYK  28 G P K T N N P E T L V A L R K A G L N I V R M N F S H G S Y
CAT3  27 G F S L T S K I D I T T L K K S L D D S A - Y K F Y P V - M

CPYK  81 E Y H E G T I K N V R E A T E S F A S D P I T Y R P V A I A
YPYK  58 E Y H K S V I D N A R K S E E L Y - - - P G - - R P L A I A
CAT3  55 I Y L I A Q A V N - - Q F D E L R M A I K D D - - - E L I V

CPYK 111 L D T K G P E I R T G L I K G S G T A E V E L - K K - G A A
YPYK  83 L D T K G P E I R T G - - - - T T T N D V D Y P I P P N H E
CAT3  80 W D S V D P Q F T V F - H Q E T E T F S A L S - C P - - - -

CPYK 139 L K V T L D - N A F M E N C D E N V L W V D Y K N L I K V I
YPYK 109 M I F T T D - D K Y A K A C D D K I M Y V D Y K N I T K V I
CAT3 104 - - Y S S D I D Q F M V N Y - - L S V M E R Y K S D T K L F

CPYK 168 D V G S K I Y V D D G L I S L L V K E K G K D F V M T - E V
YPYK 138 S A G R I I Y V D D G V L S F Q V L E V V D D K T L K V K A
CAT3 130 P Q G - - V T P E N H - L N I S A L P W V N F D S F N L N V

CPYK 197 E N G G - - - - M L G S K K G V N L P G A A V D L P A V S E
YPYK 168 L N A G - - - - K I C S H K G V N L P G T D V D L P A L S E
CAT3 157 A N F T D Y F A P I I T M A K Y Q Q E G D R L L L P L S V Q

CPYK 223 K D I Q D L K F G V E Q N V D M V F A S F I R K A A D V H A
YPYK 194 K D K E D L R F G V K N G V H M V F A S F I R T A N D V L T
CAT3 187 - - - - - V H H A V C D G F H - - V A R F I N R L Q E L C N

CPYK 253 V R K V L G E K G K H I K I I S K I E N H E G
YPYK 224 I R E V L G E Q G K D V K I I V K I E N Q Q G
CAT3 210 S K L K
```

Figure 2. Alignment of CAT III (CAT3) with yeast and chicken pyruvate kinases (YPYK and CPYK). Deletions are indicated by a dash. Identical residues are boxed

The finding of structural relationships between two proteins in this way is followed by a scan for sequence homology. The template pattern similarity between CAT and cat muscle PYK was investigated further by comparing the sequences of the CAT variants with the sequences of PYK from chicken[26] and yeast[27]. (PYK sequences were taken from the Protein Sequence Database of the Protein Investigation Resource[28].) A direct comparison of the cat muscle PYK with CAT was not possible, since the amino-acid sequence could not be determined at the 2.6Å resolution of the crystal structure[25].

An alignment of the PYK sequences with CAT III is shown in Figure 2. Sequence conservation across the three sequences is not high, and the homology is undoubtedly weak. However, the alignment is more convincing in view of the relatively small number of insertions and deletions that have to be introduced, and the close agreement between the predicted secondary structures of the CAT and PYK sequences.

Molecular graphics modelling

For the purposes of the molecular-graphics modelling of CAT III described here, it was assumed that there was a tertiary structure relationship between CAT and PYK, and that this was reflected in the similarity in the secondary-structure templates of the two proteins. A major aim of the study was to investigate the strengths and weaknesses of interactive molecular graphics for the modelling of larger systems and to evaluate the potential for integration of this computer-based system with the energy minimization routines of Lucifer, which is the system generally used in this laboratory for conformational studies of all molecules, irrespective of size.

The C_α Cartesian coordinates of cat muscle PYK were taken from the Brookhaven Data Bank[29]. The absence of a primary structure from this low-resolution structure[25] was not considered to be a handicap, since the precise locations of the various secondary-structure elements were indicated.

Table 1. Lengths of corresponding elements of secondary structure in cat muscle pyruvate kinase and the model for CAT III

Secondary structure identifier	Length of element	
	CAT III	PYK crystal
AE1	7	5
AH1	9	7
AE2	6	5
AH2	8	15
AE3	5	4
AH3	11	—
AH4	6	—
BE1	8	5
BE2	5	7
BE3	—	6
BH1	14	5
BE4	6	4
BE5	9	7
BH2	5	—
BE6	6	11
BE7	5	6
BH3	9	—
BH4	7	—

The CAT III assignments are based on the consensus secondary structure prediction. Units of secondary structure with 'A' as the first identifier are in Domain A, and 'B' indicates elements contained within the Greek key, Domain B. 'H' and 'E' in the second position of the identifier refer to α helix and β sheet respectively. Elements are uniquely identified by the number in the third position. A dash indicates that the element is missing.

The first step in the modelling process was the separation of the two domains of interest from the remainder of the PYK structure. The first half of the β barrel domain A, and the entire Greek key domain B were then independently modified to be consistent with the predicted secondary structure of CAT. In a number of cases there were differences in the lengths of corresponding helices and the extended chains in CAT and PYK. Length was reduced by removing residues from the C-terminal end of the structural element, while extensions were made by adding more residues to the C-terminal end. The lengths of the equivalent elements of periodic secondary structures are summarized in Table 1, together with the designated codes for these regions, which will be used below.

Certain helices predicted in CAT did not have equivalents in PYK, and so had to be modelled on homologous helices from elsewhere in the protein-structure database. These additional helices were then fitted into the structure and positioned such that their polar faces were oriented towards the surface of the protein.

The greatest variation between CAT and PYK was in the loop and turn regions, and so these were modelled last. Where available, homologous loops from the protein-structure database were used as templates, but in other cases a plausible loop was simply interactively modelled.

The final juxtapositioning of the two CAT domains relative to each other, and the adjustment of their internal structure, took into account the available biochemical information summarized earlier. The relative distances between conserved or putative active residues were obtained by the introduction of the substrate 'measuring stick' concept in the interactive modelling process. That is, the substrates CM and acetyl CoA were used as measuring sticks. The coordinates of CM and acetyl CoA were obtained from the Cambridge Crystal Structure Data Base[30], the latter being constructed from the component parts (L-lysine D-pantothenate, an adenosine moiety and a β-mercaptoethylamine moiety).

Side chains were approximately positioned on the CAT III model for the polar residues Arg/Lys-13, -14, -68 and -171, and Asp-35. Similarly placed were the side chains for His-189, Gln-173 and Cys-26. It was found that the 3′ and 5′ phosphate moieties could bind to conserved cationic residues at 13 and 14 on the same surface loop of domain A and were restricted to that position by Asp-35. The third phosphate could bind to residue 171 by placing the β sheet surface of each domain together, forming a relatively hydrophobic cleft (see Colour Plates 1 and 2). The fully extended pantothenic acid backbone of acetyl CoA lies along the first (AE1) extended strand, placing the thioester moiety close to Cys-26. Generation of the van der Waals' dot surface on the coenzyme aided the separation of the two domains (see Colour Plate 3).

Helix BH3 lies naturally along the lower edge of the cleft with the active His in close proximity to the coenzyme acyl donor. The substrate fits into the cleft with the phenyl ring lying alongside Tyr-172 and the chloride bound to Arg-68.

As a final step in the modelling, the three monomer units were fitted together to form the trimer. A relatively hydrophobic face had been created by the side of the Greek key distant to the cleft, and it was through the hydrophobic interactions of these regions that the trimer formation was envisaged to occur (see Colour Plate 4).

Evaluation of the model — the hydrophobic radius of gyration

The molecular-graphics modelling described above was neither time consuming nor computationally expensive, and the structure produced could have provided a good starting point for refinement by energy minimization and molecular dynamics, both of which would be relatively long processes. An attractive feature of the structure was that its hydrophobic radius of gyration was very close to the value that would have been expected for a protein of this size. The hydrophobic radius of gyration of a protein is a criterion by which nonnative-like conformations may be rapidly identified. Until this study, it appeared to be a powerful general criterion by which an incorrect fold could be identified and rejected.

For a protein of NRES amino-acid residues length containing n hydrophobic residues, where residue types Leu, Ile, Val, Cys, Phe, Met, Trp and Tyr are classified as hydrophobic, the hydrophobic radius of gyration of the molecule, R, can be calculated from the equation:

$$R = \left[\frac{\sum_{i=1}^{n} \sum_{j=2}^{n} d_{ij}^{2}}{n \times n} \right]^{\frac{1}{2}}$$

where d is the distance between the centroids of the side chains in the hydrophobic residues.

The hydrophobic radii of gyration of 34 proteins in

Table 2. Proteins used in the investigation of the hydrophobic radius of gyration

Brookhaven Data Bank code	Protein name
1CRN	Crambin
3RXN	Rubredoxin
1FDX	Ferredoxin
3PTI	Pancreatic trypsin inhibitor
1HIP	High potential iron protein
2B5C	Cytochrome b_5 oxidized
3FXC	Ferredoxin
1PCY	Plastocyanin
4CYT	Cytochrome C — reduced
2CPV	Calcium-binding parvalbumin B
1C2C	Ferricytochrome C_2
3RSA	Ribonuclease A
7LYZ	Lysozyme
3FXN	Flavodoxin — oxidized
1HHB	Human haemoglobin — chain 1
1HHB	Human haemoglobin — chain 2
1SNS	Staphylococcal nuclease
3MBN	Myoglobin
1DFR	Dihydrofolate reductase
1LZM	T_4 lysozyme
2ADK	Adenylate kinase
8PAP	Papain
2ACT	Actinidin hydrolase
1PTN	β trypsin
3CNA	Concanavalin A
2GCH	Gamma chymotrypsin A
1CAC	Carbonic anhydrase — form C
2SBT	Subtilisin
1RHD	Rhodanese
1ABP	L-arabinose binding protein
1APP	Acid proteinase, penicillinopepsin
1APR	Acid proteinase
1LDX	Lactate dehydrogenase
4ADH	Apo liver alcohol dehydrogenase
2GRS	Glutathione reductase

the range of 40 to 500 residues length have been calculated. These proteins are listed in Table 2, and their values of $<R^2>$ are plotted against NRES in Figure 3. As can be seen from the Figure, the points fall around a straight line of the formula:

$$<R^2> = 6.9 + 0.967\text{NRES}$$

It should be noted that this formula holds only for proteins of the size given in the range above. For much larger proteins, a greater degree of divergence from that predicted by the formula is observed, while small proteins have values markedly above that expected for their size. Monomeric APP, for example, has a value of $<R^2>$ of 104Å^2.

Despite these reservations, it would appear that, for a molecule of around 215 residues length, such as CAT, the hydrophobic radius of gyration can be used as a quick and easy way to gain an impression of the hydrophobic packing of the modelled structure. Indeed, we have elsewhere invoked it during energy minimization to direct folding to a rational result on the basis of the hydrophobic packing. The value of $<R^2>$ obtained for CAT was 232Å^2, but this was based on the C_α positions of the hydrophobic residues. A correction factor of -8Å^2 (obtained from an analysis of protein crystal structure) must be added to this value to gain an estimate of the hydrophobic side chain centroid value. The corrected value of 224Å^2 for CAT III is very close to the expected figure of 222Å^2, and this is satisfying, especially since the distribution of the hydrophobic residues in the pyruvate kinase template was unknown.

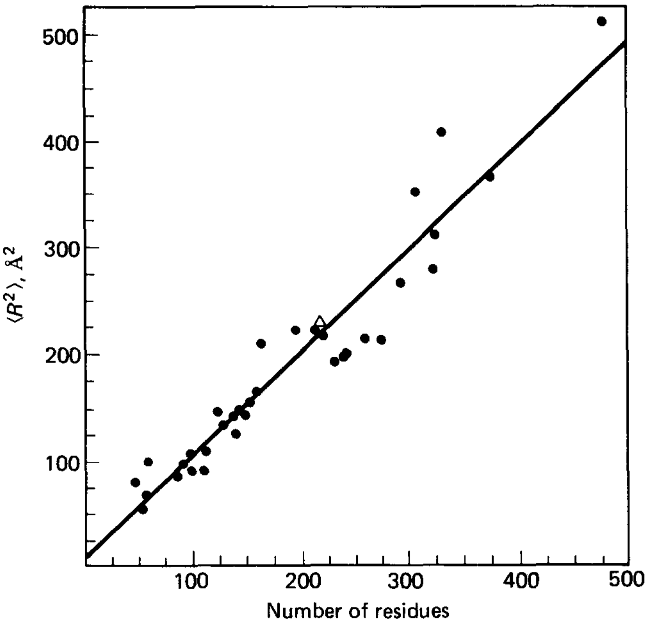


Figure 3. Graph of the square of the hydrophobic radius of gyration, $<R^2>$, against the number of residues in the protein (see Table 2 for the proteins plotted). The straight line was determined using a least-squares fit. The value determined for the predicted structure of CAT III is indicated by △

Comparison with the native

The following analysis is necessarily tentative since contact with the details in advance of this publication has been purposely avoided. Experimental features considered below and explicitly noted are done so retrospectively only by independent assessment or through 'accidental exposure' to discussion of CAT.

In the light of the finding that the modelled structure was plausible on the grounds of the hydrophobic radius of gyration, a continuation of the study using energy minimization and molecular dynamics to refine the structure was anticipated. However, this was pre-empted by the preliminary results of the crystallographic study[18]. Although the current resolution is only at the 2.7Å level, clearly there are significant differences between the predicted conformation and cartoons describing the crystal structure. The major fold motif observed in the crystal structure is a six-stranded antiparallel β sheet with flanking α helices. In addition, a further strand appears to hydrogen bond with the first β-sheet strand of the adjacent subunit of the trimer. This interaction may represent an important factor in trimer formation. A detailed comparison of the predicted and observed structures must wait until the crystal structure is refined further, but it is appropriate to summarize some of the attractive features of the modelled structure.

The model was based on a common supersecondary-structural feature of proteins, a Greek key, with an additional section based on a Rossman fold motif. These two domains were positioned such that they formed a long hydrophobic cleft suitable for binding acetyl CoA, while the overall hydrophobic radius of gyration for the structure was very close to the expected value, despite the fact that the primary structure of cat muscle PYK was unknown. The predicted α-helical content was similar to the content of around 30% indicated by circular dichroism studies*. In addition, residues that had been implicated as being in the active site or as having substrate-binding roles were located in appropriate places in the model. The site of the engineered insertion in the chain[20] was in a peripheral helix where no major disruption of the overall fold would be made. Nevertheless, the structure predicted is significantly different from that observed.

The authors are informed that the content and relative location of predicted periodic elements of secondary structure agree fairly well with those in the crystal structure, but the tertiary-structure template used, the Greek key, was incorrect. It is interesting to note that some of the weakly homologous protein sequences identified by a protein database scan, such as alcohol dehydrogenase and dihydrofolate reductase, do contain β-sheet supersecondary structure motifs. In the absence of the stronger homology identified with the pyruvate kinase sequences, this kind of motif might have been taken as the tertiary template. Even so, it is unlikely that a different model would have included one important and unusual feature suggested by preliminary analysis of the X-ray data — that the active and/or binding site region comprises sections of the chain from two or more of the three subunits. It is assumed that this preliminary statement is valid in the following. Contribution of residues from more than one subunit to the active site is unusual but not unprecedented (e.g. glutamine synthetase[31]). However, there had been no indication from the various experimental studies summarized earlier that this was the case in CAT, and the experimental conformational data had been thoroughly considered when constructing the model. That is, assurances as to the value of this data were taken too literally. Because of the above considerations, it is tempting to draw the far-reaching conclusion that the model presented here describes an ancestral form of CAT. That is, the formation of the di- or trimeric active/binding site may have been a relatively recent event.

Some important lessons can be learned from the molecular-graphics modelling of CAT. On the positive side, the use of a consensus secondary structure prediction can enhance the reliability of the template, and if a matching template can be found, it is possible to derive very rapidly a fairly detailed model for the tertiary structure. Furthermore, the structure predicted can be consistent with at least one criterion of native-like folds, in having an appropriate hydrophobic radius of gyration. It is also of value to find for the first time a case where this criterion is not sufficient *per se* to identify and reject improbable protein structures. On the negative side, clearly one must not assume that a protein with weak sequence homology has the same tertiary structure, especially if the detected sequence homology is indirect, as in the present study. Here, evolution to such a strongly integrated trimer might cause supersecondary rearrangement. The detection of weak homologies between the protein of interest and a number of other proteins with different fold motifs, as noted in this study, can alert one to this danger. Lastly, one must not rely exclusively on the data from experimental conformational studies, but should consider other interpretations of the results.

Molecular graphics undoubtedly has an important place in protein structure analysis, but, on the basis of the present findings, the subjectivity of the homology-molecular graphics approach would appear to limit its usefulness in the *de novo* modelling of large proteins, unless a very closely homologous protein of known structure can be identified. In the context of the Lucifer suite, the main role of molecular graphics is in the rapid generation of plausible structures that can then be studied objectively and in detail by robust energy minimization and molecular-dynamics routines.

CONFORMATIONAL STUDY OF AVIAN PANCREATIC POLYPEPTIDE

Background

The crystal structure of avian pancreatic polypeptide (APP) was determined by Blundell *et al.*[32]. APP is 36 residues long and as such is one of the smallest proteins to have been studied crystallographically. It is strictly a 'biologically active peptide', and, for this class of molecule, studies using nuclear magnetic resonance spectroscopy in solution and X-ray crystallography often lead to different conclusions about the conformational preferences. It is notable that the crystal structure of monomeric APP lacks the compact hydrophobic core characteristic of larger proteins; as has been mentioned above, the hydrophobic radius of gyration of monomeric APP is well above that expected from an analysis of the trend for other proteins. However, dimeric APP has a value of $<R^2>$ less than 50% above the expected figure. It must therefore be said from the outset that this special integrity of the dimer, the small size of the 'protein' and the importance of bound ions in the lattice make crystalline APP a weak test. The structure of the isolated monomer favoured in solution may differ drastically, as in the case of mellitin.

There is nevertheless some hope of calculating a dominant solution conformer of APP as a 'biological peptide', since Lucifer has proved to be quick and efficient in the folding simulations of, for example, neurotensin[33,34] luteinizing hormone-releasing hormone[35] and dynorphins[36], although these are all less than half the size of APP.

The observed tertiary structure of APP is fairly simple. The first six residues are in extended conformation, with φ/ψ angles in the range characteristic of a polyproline helix. A loop region of seven residues follows and leads into a long α helix which packs against the initial section of extended chain. The final five residues are loop-like but interact with the helix to leave the carboxyl terminus fairly close to the N terminus (see Colour Plate 5).

*Pain, R H, unpublished

	APP	1ALP	1APR	1SGR	3CPA
	G	X	R	X	L
	P	V	L	X	Q
	S	F	G	X	I
	Q	X	G	X	G
	P	X	G	X	R
	T	X	G	V	S
Loop	Y	X	F	C	Y
	P	P	P	P	E
	G	G	G	G	G
	D	N	D	D	R
	D	D	N	S	P
	A	R	D	G	I
	P	A	G	G	Y
	V	W	L	S	V
	E	V	L	L	L

Figure 4. Homologues of the loop region of APP: an alignment of loops from α lytic protease (1ALP), acid proteinase A (1APR), proteinase A from Streptomyces giseus *(1SGR), and carboxypeptidase A (3CPA), with the predicted loop region (residues 7 to 13) of APP. X indicates that the residue type is unknown*

Use of the Globex stack

It should be recalled that when a deep new minimum is identified, its conformation is added to the new Simplex, together with the conformations of previous minima, and the remainder (if any) of the Simplex is filled with new points semirandomly generated around the new minimum. The storage and use of previous minima and the action of the Simplex procedures imply that the best direction to look for the global minimum is indicated by the previous local minima identified. This, then, is the normal Lucifer operation of the Globex stack.

The procedure for initialization of the stack so as to exploit the conformational information in protein structure databases is being developed in a series of studies, including the present one. At the start of the simulation of APP, the stack contained the conformations of metastable minima identified from $n + 1$ local, independent minimizations from $n + 1$ different starting conformations. In the case of APP, 137 such metastable conformers were introduced into the Globex stack. In general, such conformers represent justified unprejudiced 'best guesses' concerning plausible conformations. In this study, the starting conformations for these 137 short minimizations runs were objectively decided in the following manner.

The first step was a scan of the protein-structure database for regions homologous in sequence to regions in APP. There is a natural homology between the N terminus of APP and the proline-rich regions of collagen. However, this factor was neglected to provide 'a greater challenge'. Four segments bearing limited homology to the middle of the APP sequence were found in α lytic protease, acid proteinase, carboxypeptidase A and proteinase A from *Streptomyces giseus* (see Figure 4). These segments are loops which could provide starting points for four minimization runs on the corresponding APP segment. As it turns out, the corresponding APP segment is also a loop, but none of these four loops has a structure particularly similar to that of crystalline APP.

A secondary structure prediction of APP using the Garnier *et al.* method[23] supported the idea (assuming no prior knowledge of APP conformation) that residues 7 to 13 were in a loop-like conformation, while the C-terminal portion was weakly predicted as helical. From this, one might reasonably deduce that there were three sections of polyproline helix, loop and α helix. Instead, in view of the weakness of the predictions, and for the purposes of objectivity, it was merely inferred that there were three regions, the middle one of which is apparently a loop.

An objective procedure must be found for regions lacking an obvious homology. Six starting points were taken for the two remaining sections. In six short runs, both of these sections were minimized from the following: α helix, extended chain, polyproline helix, left-handed helix, 2_7 axial, and 2_7 equatorial. Once the separate sections had been independently minimized, they were combinatorially assembled to give 144 (6 × 4 × 6) different structures. These were then relaxed by local minimization, and the lowest-energy 137 were used to fill the Globex stack for the minimization proper.

Hydrophobic compaction factor

Minimization from the Globex stack was performed over longer runs and with various strengths of hydrophobic compaction factor (HCF). The HCF is an extra variable term added to the energy calculation to penalize open structures. It was reported by Robson and Platt[2] that, to obtain a 'native-like' structure for APP using the rigid geometry model, a suitable solvent model is required. Here the extra energy term E_{HCF} is used, which is related to the hydrophobic radius of gyration.

The E_{HCF} term can be defined as:

$$E_{\mathrm{HCF}} = H_c \times \mathrm{HCF}$$

where H_c is a constant and HCF is given by the expression:

$$\mathrm{HCF} = \frac{\sum_{i=1}^{n-1} \sum_{j=2}^{n} h_i \times h_j \times d_{ij}^2}{[\sum_{i=1}^{n} h_i]^2}$$

where d_{ij} is the distance between atoms i and j and h_i is the energy of interaction of the hydrophobic atom type i. Thus, the term represents the energy of interaction between hydrophobic groups and is equivalent to an entropic term of the free energy.

The solvent effects were further represented by displacable 'dummy'-water molecules positioned on potentially hydrogen-bonding groups, as has been used elsewhere (see, for example, Reference 33).

Results and discussion

The results of the minimization runs carried out under various HCFs can be assessed by three criteria:

- the root mean squared fit, RMS, of the predicted structure to the observed structure[2], which gives an indication of the similarity of the two,
- the potential energy of the predicted structures, which is compared with that of the structure RMS-fitted to the observed, crystal structure,
- the HCF of the predicted structures as compared with that of the observed structure.

The results of this study, summarized in Table 3, indicate a trend for the conformer energy to increase with increas-

Table 3. Minimization run results for APP conformers obtained with various values of H_c

H_c	Potential energy, kcal/mol	HCF	RMS for all nonH atoms, Å	RMS for C_α atoms, Å
0	33.04	259.3	11.25	12.13
1	70.91	164.0	6.00	6.05
2	93.82	110.1	3.78	2.80
3	100.47	110.0	3.85	2.60
4	119.67	110.0	3.70	2.60
5	148.60	101.0	4.40	3.60
RMS fitted to the observed	71.71	115.5	1.07	0.54
Observed	—	118.0	0.00	0.00

ing H_c, while the HCF value decreases. The structures obtained with an H_c of less than 2 are of lower energy than the best RMS-fitted-energy minimized structure, although their HCF value is markedly greater than that of the observed structure. The most compact structures (i.e. those with the smallest HCF values) were obtained using values of H_c in the range 2–4. All three conformers identified under these conditions had HCF values of around 110Å. The lowest energy structure obtained with the inclusion of the E_{HCF} term is that using an H_c value of 2. This has an RMS of 3.78Å for all nonhydrogen atoms and an RMS of 2.80Å for C_α atoms only.

There is evidence that the last four residues have flexibility in solution[37]. Thus if the RMS calculations are restricted to the first 32 residues, then the RMS values of the above structure become 2.98Å for all nonhydrogen atoms and 2.4Å for C_α atoms. The energy of the APP monomer when using an H_c of 3 is only 7 kcal/mol higher than that of the latter conformer and has a similar HCF value. The RMS values for this structure are: (a) 3.89Å (all nonhydrogen atoms) and 2.6Å (C_α atoms) for the whole molecule and (b) 2.96Å and 2.15Å respectively for the first 32 residues. The dimeric structure could easily be lower for this structure than for that with an H_c of 2, and so both structures could be considered as equally plausible. Colour Plate 6 is the stereo diagram of the structure calculated with an H_c of 3, and Colour Plate 5 shows a similar diagram for the observed structure. Figure 5 (a) is a stereo plot of the structure with an H_c of 2. The conformation obtained after a 10.5ps molecular dynamics run on the lowest energy structure is shown in Figure 5 (b).

Several factors could contribute to the apparent difference between the predicted and observed structures. The crystal structure may be significantly different from the form dominant in solution, and it is the preferred solution conformations that Lucifer seeks to identify. The crystal environment, in terms of the intermolecular forces and the presence of the Zn^{2+} ions, may serve to stabilize a form less favoured in solution. More technically, the discrepancies between the two structures may be due in part to the representation used, and in particular to the absence of an explicit treatment of hydrogen atoms. In this study, all χ^1 angles were initially set to $-60°$, with the exception of valine for which a value of 180° was chosen. It would probably be preferable to consider further the initial prediction of the backbone conformation before predicting starting values for side-chain angles.

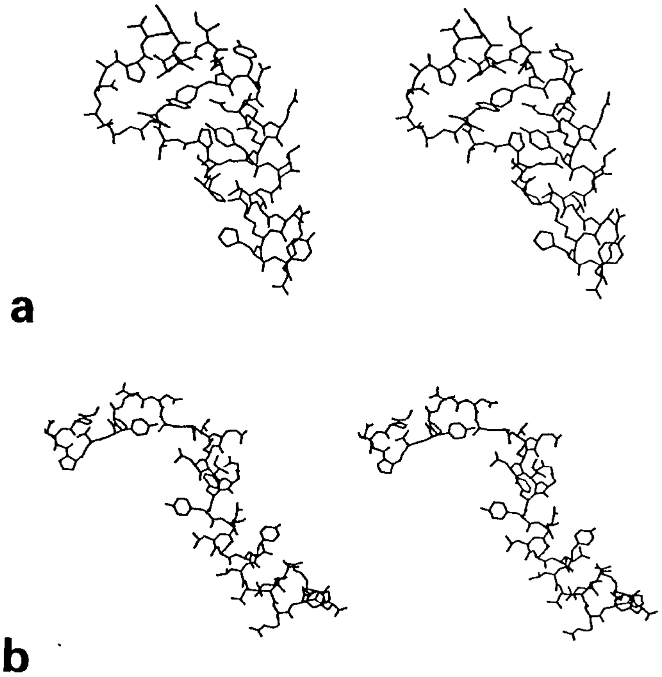


Figure 5. Stereo plots of APP conformations obtained: (a) with an H_c *of 2 and (b) after a 10.5ps molecular dynamics run on the lowest energy structure with an* H_c *of 0*

Polypeptide hormones often have different conformers in solution and in the crystal, and the multimeric assembly of such small proteins is important in determining conformation (e.g. mellitin, where four separate molecules make up a 'globin fold'). The crystal structure may not be a good test for small systems where intermolecular forces dominate intramolecular forces. In this sense, the above results are presented prior to more detailed studies in the solution state, as an objective and unprejudiced test of the procedures. The possible need for refinement of methodology is of course recognized, and disagreement with experiment is of as much value as predictive success.

ACKNOWLEDGEMENTS

The authors are grateful to Tracy Wootton for her assistance in the preparation of the manuscript, and would like to thank the University of Manchester Regional Computer Centre and the University of Manchester Computer Graphics Unit for use of their computer facilities. Eric Platt is supported by the Science and Engineering Research Council (SERC), Alan Marsden is funded jointly by the SERC and Unilever Research (CASE), and Robert Fishleigh is supported by the Medical Research Council.

REFERENCES

1 'Proposal for a research action programme in biotechnology' *EC COM* (1984) Number 230

2 **Robson, B and Platt, E** 'Refined models for computer calculations in protein engineering: calibration and

testing of atomic potential functions compatible with more efficient calculations' *J. Mol. Biol.* Vol 181 (1986) pp 259–281

3 **Pumphrey, R S H** 'Computer models of the human immunoglobulins I. Shape and segmental flexibility' *Immunol. Today* Vol 7 (1986) pp 174–178

4 **Pumphrey, R S H** 'Computer models of the human immunoglobulins II. Binding sites and molecular interactions' *Immunol. Today* Vol 7 (1986) pp 206–211

5 **Robson, B and Garnier, J** *Introduction to proteins and protein engineering* Elsevier, The Netherlands (1986)

6 **Nelder, J A and Mead, R** 'A simplex method for function minimization' *Comput. J.* Vol 7 (1965) pp 308–313

7 **Alton, N K and Vapneck, D** 'Nucleotide sequence analysis of the chloramphenicol resistance transposon Tn 9' *Nature* Vol 282 (1979) pp 864–869

8 **Shaw, W V et al.** 'Primary structure of a chloramphenicol acetyltransferase specified by R plasmids' *Nature* Vol 282 (1979) pp 870–872

9 **Marcoli, R et al.** 'The DNA sequence of an IS1-flanked transposon coding for resistance to chloramphenicol and fusidic acid' *FEBS Lett.* Vol 110 (1980) pp 11–14

10 **Shaw, W V** 'Chloramphenicol acetyltransferase: enzymology and molecular biology' *CRC Crit. Rev. Biochem.* Vol 14 (1983) pp 1–46

11 **Horinouchi, S and Weisblum, B** 'Nucleotide sequence and functional map of pC194, a plasmid that specifies inducible chloramphenicol resistance' *J. Bact.* Vol 150 (1982) pp 815–825

12 **Shaw, W V et al.** 'Chloramphenicol acetyltransferase gene of staphylococcal plasmid pC221. Nucleotide sequence analysis and expression studies' *FEBS Lett.* Vol 179 (1985) pp 101–106

13 **Harwood, C R et al.** *Gene* 'Nucleotide sequence of a *Bacillus pumilus* gene specifying chloramphenicol acetyltransferase' Vol 24 (1983) pp 163–169

14 **Kleanthous, C and Shaw, W V** 'Analysis of the mechanism of chloramphenicol transferase by steady state kinetics. Evidence for a ternary complex mechanism' *Biochem. J.* Vol 223 (1984) pp 211–220

15 **Kleanthous, C et al.** '3-(bromoacetyl)chloramphenicol, an active site directed inhibitor for chloramphenicol acetyltransferase' *Biochem.* Vol 24 (1985) pp 5307-5313

16 **Murray, I A et al.** 'Catalytic mechanism of chloramphenicol acetyltransferase investigated by site directed mutagenesis' *Biochem. Soc. Trans.* Vol 14 (1986) pp 1227–1228

17 **Packman, L C and Shaw, W V** 'Identification of "buried" lysine residues in two variants of chloramphenicol acetyltransferase specified by R factors' *Biochem. J.* Vol 193 (1981) pp 525–539

18 **Leslie, A G W and Shaw, W V** 'Structural studies of chloramphenicol acetyltransferase' *Biochem. Soc. Trans.* Vol 14 (1986) pp 1224–1225

19 **Packman, L C and Shaw, W V** 'The use of natural occurring hybrid variants of chloramphenicol acetyltransferase to investigate subunit contacts' *Biochem. J.* Vol 193 (1981) p 541

20 **Betz, J L and Sadler, J R** 'Variants of a cloned synthetic lactose operator II. Chloramphenicol-resistant revertants retaining a lactose operator in the CAT gene plasmid pBR325' *Gene* Vol 15 (1981) p 187

21 **Fishleigh, R V et al.** 'Studies on rationales for an expert system approach to the interpretation of protein sequence data: preliminary analysis of the human epidermal growth receptor' *FEBS Lett.* submitted

22 **Jornvall, H** 'Horse liver alcohol dehydrogenase. On the primary structures of the isoenzymes' *Europ. J. Biochem.* Vol 16 (1970) pp 41–49

23 **Garnier, J et al.** 'Analysis of the accuracy and implications of simple methods for predicting the secondary structures of globular proteins' *J. Mol. Biol.* Vol 120 (1978) pp 97–120

24 **Lim, V I** 'Algorithms for prediction of α-helical and β-structural regions in globular proteins' *J. Mol. Biol.* Vol 88 (1974) pp 873–894

25 **Stuart, D I et al.** 'Crystal structure of cat muscle pyruvate kinase at a resolution of 2.6Å' *J. Mol. Biol.* Vol 134 (1979) pp 109–142

26 **Lonberg, N and Gilbert, W** 'Primary structure of chicken muscle pyruvate kinase mRNA' *Proc. Natl. Acad. Sci. USA* Vol 80 (1983) pp 3661–3665

27 **Burke, R L et al.** 'The isolation, characterization and sequence of the pyruvate kinase gene of Saccharomyces cerevisiae' *J. Biol. Chem.* Vol 258 (1983) pp 2193–2201

28 *Protein investigation resource* National Biomedical Research Foundation, Georgetown University Medical Centre, Washington, USA (1986)

29 **Bernstein, F C et al.** 'The protein data bank: a computer-based archival file for macromolecular structures' *J. Mol. Biol.* Vol 112 (1977) pp 535–542

30 **Allen, F H et al.** 'The Cambridge Crystallographic Data Centre: computer based search retrieval, analysis and display of information' *Acta Cryst.* B35 (1979) pp 2331–2339

31 **Almassy, R J et al.** 'Novel subunit — subunit interactions in the structure of glutamine synthetase' *Nature* Vol 323 (1986) pp 304–309

32 **Blundell, T L et al.** 'X-ray analysis (1.4Å resolution) of avian pancreatic polyp polypeptide: small globular protein hormone' *Proc. Natl. Acad. Sci. USA* Vol 78 (1981) pp 4175–4179

33 **Fishleigh, R V et al.** 'Conformational study of neurotensin and some of its analogues' *Biochem. Soc. Trans.* Vol 14 (1986) pp 1259–1260

34 **Ward, D J et al.** 'Prediction of preferred solution conformers of analogues and fragments of neurotensin' *Regul. Pept.* Vol 15 (1986) p 197

35 **Fishleigh, R V et al.** 'Comparative conformational analogues of luteinising hormone releasing hormone' *Biol. Chem. Hoppe-Seyler* Vol 367 (*Suppl.*) (1986) p 266

36 **Griffiths, E C et al.** 'Conformational analysis of dynorphins[1–17] and [1–8]' *Brit. J. Pharmacol.* Vol 88 (*Suppl.*) (1986) p 361

37 **Strasburger, W et al.** 'Calculated tyrosyl circular dichroism of proteins. Absence of tryptophan and cystine interferences in avian pancreatic polypeptide' *FEBS Lett.* Vol 139 (1982) pp 295–299