Phil. Trans. R. Soc. Lond. B **324**, 447–460 (1989)
Printed in Great Britain

Protein engineering and design

By T. L. Blundell, F.R.S.[1], G. Elliott[1], S. P. Gardner[1], T. Hubbard[1], S. Islam[1], M. Johnson[1], D. Mantafounis[1], P. Murray-Rust[2], J. Overington[1], J. E. Pitts[1], A. Sali[1], B. L. Sibanda[1], J. Singh[1], M. J. E. Sternberg[1], M. J. Sutcliffe[1], J. M. Thornton[1] and P. Travers[1]

[1] *Laboratory of Molecular Biology, Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, U.K.*

[2] *Glaxo Research Laboratories, Greenford Road, Greenford, Middlesex UB6 0ME, U.K.*

Rapid advances in site-directed mutagenesis and total gene synthesis combined with new expression systems in prokaryotic and eukaryotic cells have provided the molecular biologist with tools for modification of existing proteins to improve catalytic activity, stability and selectivity, for construction of chimeric molecules and for synthesis of completely novel molecules that may be endowed with some useful activity. Such protein engineering can be seen as a cycle in which the structures of engineered molecules are studied by X-ray analysis and two-dimensional nuclear magnetic resonance. The results are used in the improvement of the design by using knowledge-based procedures that exploit facts, rules and observations about proteins of known three-dimensional structure.

1. The engineering and design cycle

Protein engineering is a multidisciplinary technology for the design and construction of proteins. It offers new possibilities for modification of natural proteins for specific industrial, clinical or agricultural purposes, for synthesis of chimeric proteins that combine the properties of two differing natural proteins and for construction of totally novel proteins that have properties as yet unexplored by natural evolution in living organisims.

In many ways, nature is the most successful protein engineer. However, random mutagenesis followed by selection at the level of the whole organism is a slow and laborious process. Developments in protein synthesis in the 1960s and 1970s gave hope of more effective approaches but the complexities of the chemistry made it useful only for smaller peptides or for semisynthesis of small proteins such as insulin. Two developments that were largely ignored by the Spinks Committee in 1980 transformed the scene. First, site-directed mutagenesis (Zoller & Smith 1982) allowed specific modifications to be made to DNA; it was used by Winter *et al.* (1982) to introduce changes at known sites into structural genes, thereby modifying the function of the protein in a predetermined way. Secondly, computational and graphics tools became available for display and manipulation of three-dimensional structures of proteins defined by X-ray analysis; these allowed new designs to be explored before the recombinant DNA steps were undertaken. The widespread use of interactive computer graphics by biological chemists and biotechnologists has stimulated the development of knowledge-based approaches to design; these exploit our understanding of protein structure and function in a more systematic way.

Protein engineering is now seen as a cycle of interdependent steps, illustrated in figure 1. The

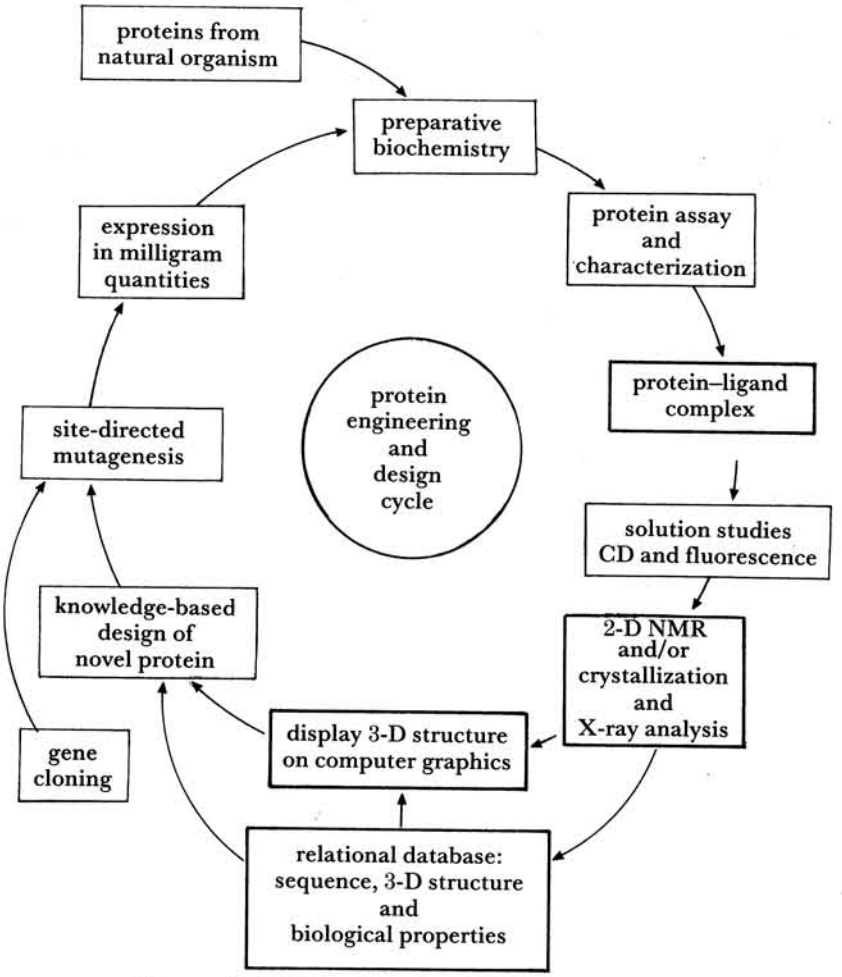

FIGURE 1. The protein engineering and design cycle.

point of entry into the cycle might be the preparation of a protein from a natural organism and its biochemical characterization with a ligand, i.e. as an enzyme–substrate, receptor–hormone or antibody–antigen interaction. The three-dimensional structure of the protein–ligand complex is defined by high-resolution X-ray analysis or two-dimensional nuclear magnetic resonance (NMR) and displayed with interactive computer graphics. Knowledge-based procedures are then used to suggest a novel design and the requisite DNA sequence is synthesized either by total synthesis of the gene or directed mutagenesis of an existing cloned gene. Finally, the novel gene is introduced into a suitable expression system and the gene product is purified and biochemically characterized, so completing the cycle. Usually several cycles are required to reach an optimally designed protein in sufficient quantities. This is because our understanding of protein structure and function is still unsatisfactory and the protein design will always need experimental testing. Thus the cycles can be seen not only as a series of steps leading to an improved protein-engineered product, but also as steps designed to test or falsify the hypothesis generated earlier in the cycle. As each step involves a different methodology – biochemistry, biophysics, biocomputing recombinant DNA, microbiology and cell biology – a closely integrated multidisciplinary research organization is required.

In this paper we briefly review, using work from our own laboratories, the stages of the

protein-engineering cycle and identify the advances made since the Spinks Committee reported in 1980. Our emphasis is on those parts of the cycle that concern protein three-dimensional structure and design. In this respect it is worth noting that the protein-engineering cycle shares steps with the design and synthesis cycles for drugs, insecticides, herbicides and peptide vaccines, illustrated in figure 2, where rational approaches also require knowledge of a protein–ligand complex, i.e. receptor–drug or antibody–antigen. The differences rest in the fact that drug design concerns only the ligand, which can usually be modified more effectively by chemical synthesis than by recombinant DNA methodology.

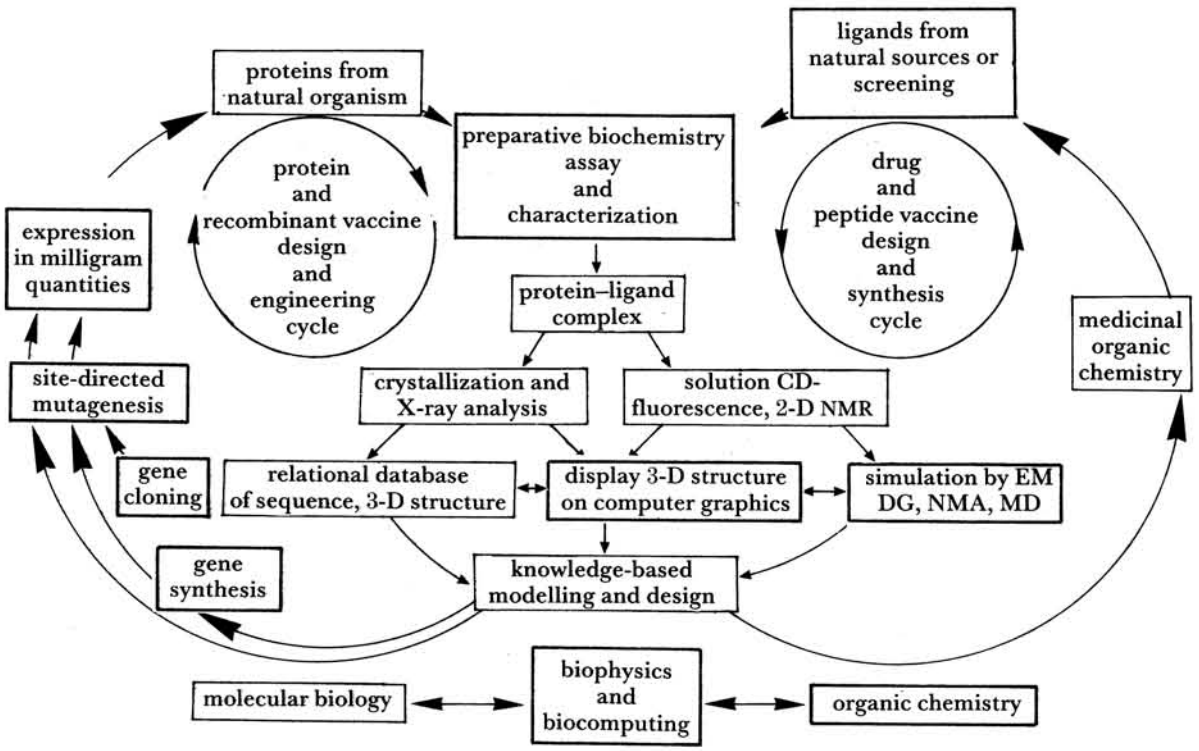

FIGURE 2. A bi-cycle illustrating the common structure analysis and design steps in the engineering of proteins, vaccines and drugs. (MD, molecular dynamics; NMA, normal mode analysis; DG, distance geometry; EM, electron microscopy.)

2. KNOWLEDGE-BASED DESIGN

The results of X-ray crystallographic studies of several hundred proteins have provided a detailed knowledge base for guiding directed mutagenesis and the design of chimeric and novel proteins. This will be enhanced by results now becoming available from two-dimensional NMR for small proteins in solution. However, three-dimensional structures will be available for few of the proteins of interest to the protein engineer, and so in many cases a procedure is required for generation of the structure of the proteins of interest. This may be achieved by *de novo* procedures or, more realistically, by modelling from homologous or analogous proteins if they are available. Such a modelling procedure has much in common with the techniques required for modelling replacements, insertions and deletions in proteins of known structure during site-directed mutagenesis. For these reasons we discuss the general requirements for knowledge-based modelling and design and then describe their application to various specific protein engineering problems.

Figure 3 illustrates the software components – computer graphics, computer simulations and databases – required in an integrated knowledge-based system. A truly integrated system involving all these components does not yet exist, although many academic and commercial packages have several components.

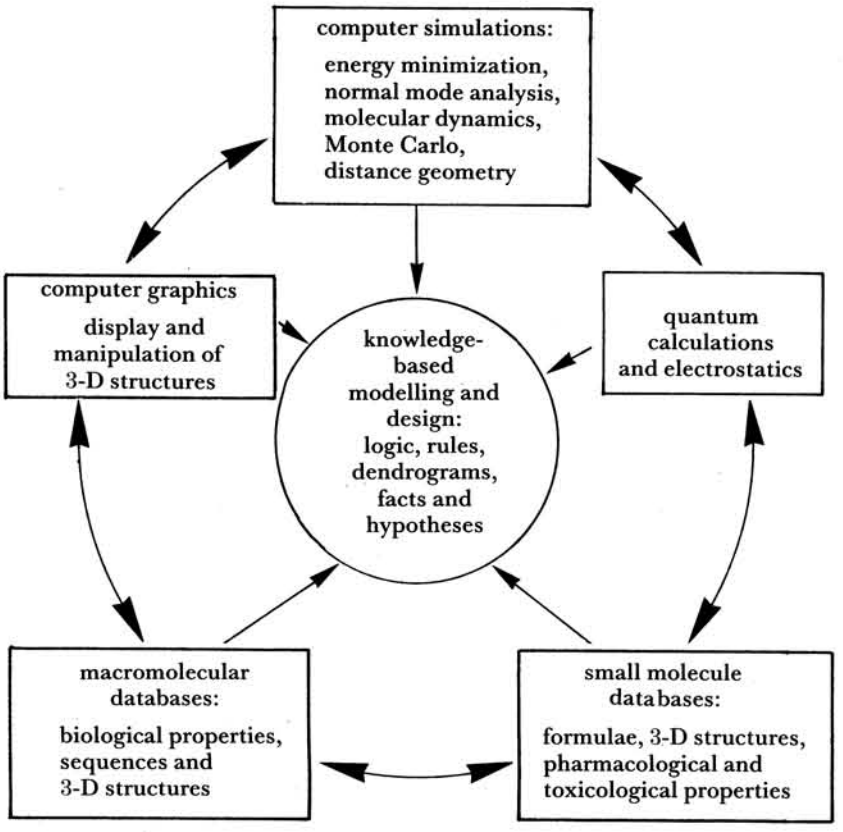

FIGURE 3. Computational aspects of knowledge-based approaches to drug, protein and vaccine design.

2.1. *Computer Graphics*

The interactive computer graphics workstation provides the man–machine interface. In the past few years the technology has developed fast, so that raster graphics systems using high-resolution commercial television technology have overtaken previously popular vector graphics or calligraphics technology based on the oscilloscope. Raster graphics can now give images of proteins either as 'ball and stick' models or as 'dotted' or solid surfaces. The computational power is increasing so that even complex solid surfaces can be rotated and otherwise manipulated smoothly in real time through a series of dials or by using a 'mouse'.

2.2. *Databases*

Figure 3 also emphasizes the importance of the knowledge resource, not only for small molecules that might be ligands but also for the sequences and structures of the macromolecules. In the first area the Cambridge Crystallographic Database and its associated software provides a versatile and well-maintained source for the three-dimensional structures of small molecules defined by X-ray analysis. For the sequences and structures of proteins and nucleic acids the

position is more complex and less satisfactory. However, our groups at Birkbeck College, London University, in collaboration with Leeds University and funded by the SERC Protein Engineering Club have now developed an integrated database of protein sequence and structure (ISIS) that incorporates derived data relating to protein properties and relations (Akrigg *et al.* 1988). The OWL sequence database is an integration of six publicly available databases together with a translation of GENBANK. The BIPED relational database of three-dimensional structure is based on coordinates provided by Brookhaven. Data derived from the coordinates include the secondary structure, ϕ, ψ, ω and χ angles, solvent accessibilities, hydrogen bonds, salt bridges, disulphide bridges and neighbours. In addition, crystallographic data such as *R*-factors and *B*-values are stored so that selections can be dependent on the quality of the data.

The data are organized in a hierarchical order with information at the PROTEIN, CHAIN, RESIDUE and ATOM levels. To facilitate easy and flexible access we have used the Oracle Relational Database Management System with the simple query language SQL. The queries take the form of 'SELECT column..., FROM Table..., WHERE conditions... are met.' A simple enquiry to extract the ϕ, ψ angles of all prolines in the fourth position of an α-helix took 14 seconds of μVax II c.p.u. time to locate the information from more than 80000 residues in 296 proteins.

Integration of the OWL and BIPED databases is achieved through a simple system of protein codes and residue identifiers. A powerful extension to the storage of basic data comes from explicit incorporation of features of protein sequence and structure. The term 'feature' is deliberately unspecific and refers to any structure or substructure of proteins that can be defined by a set of aligned sequences or superposed three-dimensional structures.

The structural features are defined by structural superpositions in three dimensions, rather than just sequence alignments. There will be three major components. First, tables of homologous protein families of known structure will be aligned on the basis of superposition of their three-dimensional structures. For example, the table for the aspartic proteinases includes three proteins (penicillopepsin, endothiapepsin and rhizopuspepsin) whose known structures have been aligned in three dimensions. Related sequences of unknown structure are also aligned on the basis of sequence. These tables will allow easy extraction of conserved residues in a family, or for example all cases where an alanine is replaced by a proline. Secondly, tables of structural features (e.g. β-strands, β-hairpins, β-meanders, Greek keys) are being established. This will allow easy extraction of, for example, data on helices containing proline or β-turns with a glycine at position 2. The third type of structural feature will be related to biochemical function and will include nucleotide binding pockets, DNA binding structures, sites binding ions, metals and other ligands, sites of covalent modification, and many others.

2.3. *Computer simulations*

A fully integrated knowledge-based system will require the full range of computer simulations for energy minimizations, normal-mode analyses, molecular dynamics and Monte Carlo calculations. Distance geometry routines and various procedures for calculating electron distributions and electrostatics are also required. In many cases the protein structures are too complex for rigorous calculations and so we proceed by using knowledge-based procedures. However, simulations will be useful where the knowledge base is insufficient, for example where no fragment exists that has appropriate geometry or sequence. Thus sampling of torsion angles

and energy calculations may be used for systematic conformational search for variable 'loop' regions (Bruccoleri *et al.* 1988). Simulations such as energy minimization and molecular dynamics are also useful for finding local minima once a rough model has been produced by knowledge-based procedures.

For very small changes in structure introduced by protein engineering, the perturbation method may give the most reliable predictions (Tembe & McCammon 1984). This procedure depends on growing the structural change incrementally in the folded and unfolded states while the protein-solvent system is simulated by molecular dynamics. This gives estimates of free-energy changes for site-directed mutations.

2.4. *Knowledge-based modelling*

Any knowledge-based approach requires the establishment of a series of rules, logic, facts and hypotheses that can be used in the modelling. In our approach we have made a systematic study of evolutionary relationships in divergent protein families to establish the rules that define structural changes consequent upon limited sequence change (Blundell *et al.* 1987, 1988). This has been achieved at three levels for the conserved framework, for the variable regions and for side chains.

At the level of the framework we have developed the ideas of Eventoff & Rossmann (1975) and established procedures for the independent construction of phylogenetic trees from sequences and three-dimensional structures of homologous proteins (Johnson *et al.* 1989). These trees, examples of which are shown in figure 4, can be used to classify related structures and for selection of appropriate molecules for modelling. The conserved or 'framework' region for the unknown is then constructed from a weighted average of the superposed structures of the homologous or analogous subgroup (Sutcliffe *et al.* 1987*a*). In this case the rules simply define the weights in the averaging procedures and these depend on defined relations between root mean square differences of three-dimensional structures and differences in sequences. For the establishment of frameworks for more distantly related structures, direct superposition is not useful and we have developed software for comparison of structures based on local properties (hydrophobicity, secondary structure, etc.) and relations (hydrogen bonding, nearest neighbours) at different levels in the hierarchy of protein structure (A. Sali & T. L. Blundell, unpublished results). These comparison algorithms lead to more complex rules for relating families of analogous structures.

The rules for modelling variable regions that often occur between secondary structural elements of the framework are equally challenging. The most straightforward approach involves the classification of particular regions, such as β-hairpins so that conformation can be related to length and residue identities (Sibanda & Thornton 1985; Milner-White & Poet 1986). These rules can be used systematically in modelling (Sibanda *et al.* 1989) but unfortunately account for a relatively small percentage of the variable regions. An alternative approach selects conformers on the basis of key residues that define the conformation of the region. Such key residues may be inaccessible to solvent (i.e. being buried in the protein core), or have an unusual ϕ angle, etc. This approach derives from the idea of canonical structures defined by Chothia *et al.* (1986) but has been usefully coded for general modelling (Sutcliffe 1988).

The third set of rules defines the conformations of side chains replacing a side chain (perhaps by site-directed mutagenesis) in a protein of known structure. Thus Sutcliffe *et al.* (1987*b*)

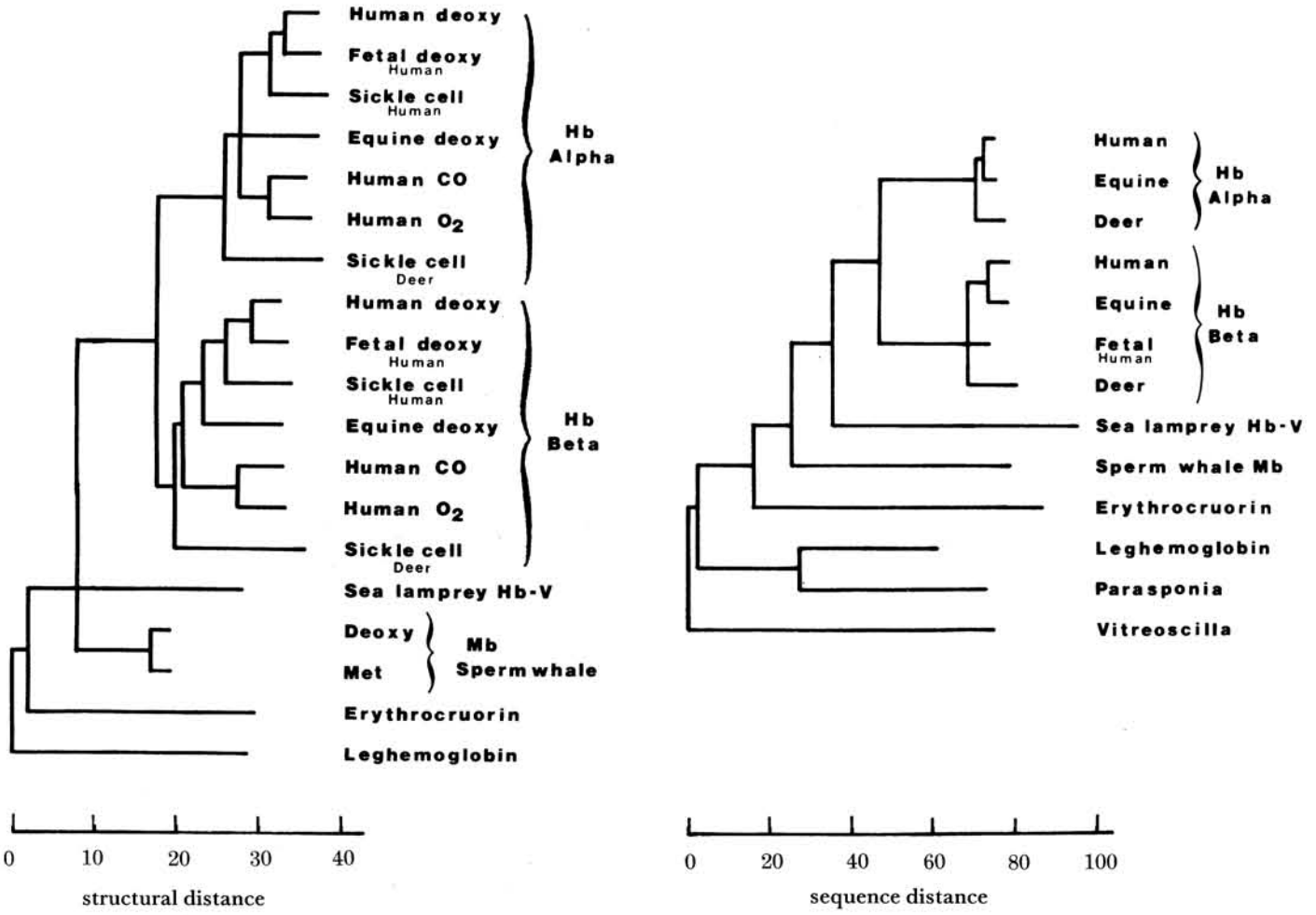

FIGURE 4. Phylogenetic trees derived from comparisons of sequences and three-dimensional structures for globins.

analysed the relation between side chains at topologically equivalent positions in homologous families and developed rules for spacial relations for most of the 20×20 possible replacements for α-helical, β-sheet and non-regular structural regions. Similar studies have been reported by Summers *et al.* (1987). When there is no useful rule the most probable structure is chosen (McGregor *et al.* 1987).

Data on preferred side-chain interaction geometries are also usefully included in modelling (Reid & Thornton 1989). However, this approach is limited by the size of the database for any pair of residues with the required C_α separation; it is useful only for the more common interactions, such as adjacent residues on an α-helix. Alternatively the interactions between side chains can be analysed, independent of backbone geometries, by constructing geometric distributions. Studies on aromatic (Singh & Thornton 1985) and arginine–carboxylate interactions (Singh *et al.* 1987) have both revealed unexpected preferred geometires. Currently a database of the 20×20 side-chain interactions is being established (Thornton *et al.* 1988).

Thus we have a series of rule-based procedures that can be used for building a structure by homology or analogy, or for introducing replacements, insertions or deletions. Much of the software is presently incorporated into a computer program, COMPOSER (Sutcliffe *et al.* 1987*a*, *b*;

Blundell *et al.* 1988; Sutcliffe 1988). Figure 3 shows that the knowledge-based modelling procedure needs to be closely integrated with databases, computer simulations and computer graphics. It also needs to be developed into a knowledge-engineering environment in which other information and hypotheses can be used to guide the modelling and design procedures.

3. Protein engineering

In this section we shall discuss the application of the knowledge-based modelling and design to three areas of current interest at Birkbeck. These are modelling of tissue-type plasminogen activator and its interactions with inhibitors, the site-directed mutagenesis of an enzyme (chymosin) to vary activity and pH optimum, and *ab initio* design of a novel protein.

3.1. *Modelling by homology*

For many proteins of biotechnological interest, the three-dimensional structure has not been determined by X-ray analysis or NMR. One such example is tissue-type plasminogen activator (tPA). This is a protein where four other domains are linked to a serine proteinase domain (Pennica *et al.* 1983). The additional domains target the site of action by specifically binding to the major substrate, fibrin, to give proteolytic activation *in situ* (Gething *et al.* 1988). tPA has great potential use as a selective thrombolytic agent. More recently, protein-engineered chimeric molecules have been designed (Harris 1989). The regulation of enzyme interactions with the potent endogenous inhibitor PAI-1 (Ny *et al.* 1986) may also be of clinical value. For these reasons the enzyme–inhibitor complex has been studied by model building with COMPOSER (Blundell *et al.* 1988; Overington *et al.* 1989).

Seven distinct serine proteinase structures were available from the Brookhaven Databank (Bernstein *et al.* 1977). Phylogenetic analysis of the sequences of these and the serine proteinase domain of tPA suggests the use of only the mammalian-derived enzymes (elastase, chymotrypsin, kallikrien-A, rat mast-cell proteinase-II and trypsin), in the model building of tPA. Determination of the structurally conserved core from this set defines 145 residues as being in the framework (54% of the tPA sequence being modelled). The peptide backbone of the core regions are built by least-squares fitting to the framework of the most similar corresponding fragment from the family. Experience in model-building known structures reveals that the error in the positioning of the mainchain atoms at this stage will be of the order of 0.5 Å†.

The interconnecting variable regions were modelled by an analysis of the corresponding variable regions from the family. If possible equivalent variable regions from the family are used, then side-chain orientation and main-chain conformation are already optimized to the constraints of the global serine proteinase fold.

The main criterion for the selection of these variable-region fragments is the conservation of residues between the sequence of the unknown and the sequence of the known fragments, at positions determined to be important in defining the conformation. Of the 14 variable regions that required modelling, ten can be taken directly from the family, and a further three can be satisfactorily built by local extension of the framework within the bounds defined by the family followed by insertion of a fragment identified by a distance-based search of a database. Only one fragment required manual modelling; for this section no fragments obeying the necessary distance constraints could be found in a search of the database.

† 1 Å = 10^{-10} m = 10^{-1} nm.

The active site of PAI-1 was modelled on the observed structure of several peptide inhibitor – serine proteinase complexes. Although the PAI-1 and structurally known inhibitor sequences display no homology, it is likely that the conformation about the cleaved bond will be similar in all peptide inhibitors owing to the required close-packed nature of the complex.

Examination of the model and comparison with the structurally known family reveals that there are important differences in the regions surrounding the active site that will probably affect ligand binding. Such predictions are being probed by site-directed mutagenesis. The preference for a positively charged P1 and a non-polar P1′ residue in the inhibitor is also apparent from analysis of the corresponding predicted specificity pockets in the tPA model.

3.2. *Site-directed mutagenesis*

Specific alterations of the sequence of an enzyme may lead to a protein of altered stability, specificity or other characteristics desirable in a useful commercial product (Smith 1985). We have explored the use of such approaches with chymosin, an aspartic proteinase used in cheese production. Similar studies have been made by others (Harris *et al.* 1982; Beppu *et al.* 1982; J. Sedlacek & P. Strop, unpublished results). A cloned gene fragment of chymosin is transferred to the filamentous phage M13 and the single-stranded form isolated. This template can be primed with a synthetic oligonucleotide either for mutagenesis or for DNA sequencing.

In our experiments we decided to attempt to influence the optimal pH by introducing a change of a conserved aspartate (Asp 304) to an alanine in the vicinity of the active site. Similarly the natural mutation of Gly 244 to aspartate has been made (chymosin B → A). To effect this, an oligonucleotide of 24 base pairs was synthesized for the change and site-directed mutagenesis experiments were done by using the method of Eckstein (Taylor *et al.* 1985). The resulting phages were isolated, grown up and samples were applied to a dot blot manifold to bind them to a nitrocellulose filter. The oligonucleotide, which had been radioactively labelled with ^{32}P by using T4 polynucleotide kinase, was then used to select mutant isolates by hybridization and autoradiography (figure 5). In such experiments all the dots bind probe initially, but upon warming and washing the filter at high temperature only the perfectly matched mutant sequences retain radioactivity. The high efficiency of mutagenesis is obtained because the protocol allows the wild-type strand to be selectively destroyed *in vitro* and the

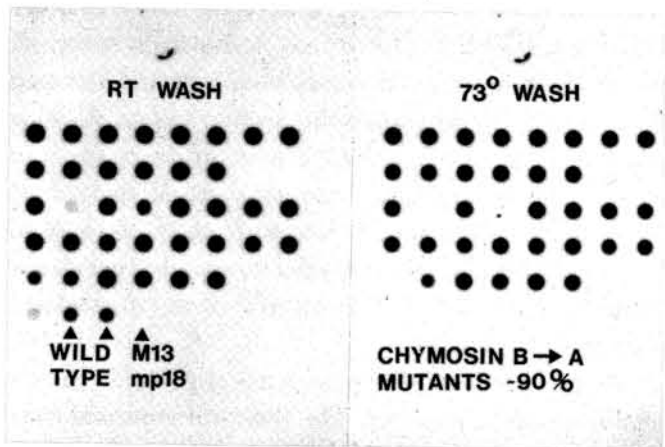

FIGURE 5. A dot blot used in the selection of mutant clones.

resynthesis occurs using the altered strand as template to give a pure homoduplex. The selected phage can be grown from the original stocks and single-stranded template prepared.

A DNA sequencing gel using the dideoxy chain-termination method of Sanger *et al.* (1977) is shown in figure 6. The wild-type sequence shown on the left is compared to mutant from the above dot blot. The antisense sequence CCA (≡ GGT for Gly) has been changed to the desired sequence CTA (≡ GAT for Asp). The result of this mutation is a change in the optimal pH towards neutrality in its substrate cleavage.

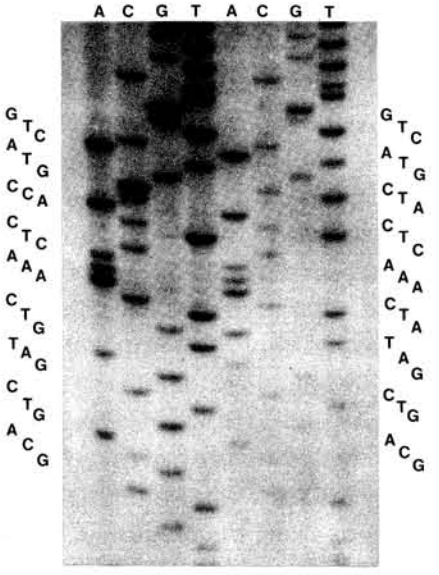

FIGURE 6. A DNA sequencing gel for native and mutant G244D.
A restriction site for *EcoRV* has also been introduced.

Although experiments on chymosin are of little value to cheese making, they may lead the way for the design for more general applications in food and other protein processing.

3.3. *Ab initio protein design*

Families of proteins with homologous tertiary structures that have diverged from a common ancestor show amino acid sequence differences that derive from two processes: the first is the selection for divergence of function between family members and the second is a consequence of neutral drift. In time neutral drift will allow the protein to sample all amino acid sequences compatible with the overall fold and function. These processes can obscure an underlying simplicity of a sequence that might constitute a sequence paradigm: the simplest sequence that will fold to create a given structure (Ponder & Richards 1987). It may prove possible for any given structure to determine the sequence paradigm, and by varying that sequence or by overlapping paradigmatic sequences of different structures, to construct proteins of novel function yet based on known structures.

As a first step in this process we have designed a novel protein, CRYSTANOVA, based on the structure of the γ-crystallin eye-lens proteins. The three-dimensional structure of two of these proteins is known (Blundell *et al.* 1981; White *et al.* 1989) and each consists of two very similar domains that have clearly arisen through gene duplication. Each domain again consists of two

similar motifs. Very few residues are completely conserved in all motifs (figure 7*a*) and the structure is extremely stable making it a good framework for a protein design. CRYSTANOVA is an attempt to design a single-domain protein where the two motifs are more similar to each other than those found naturally and thus may reflect the sequence of an ancestor of the family.

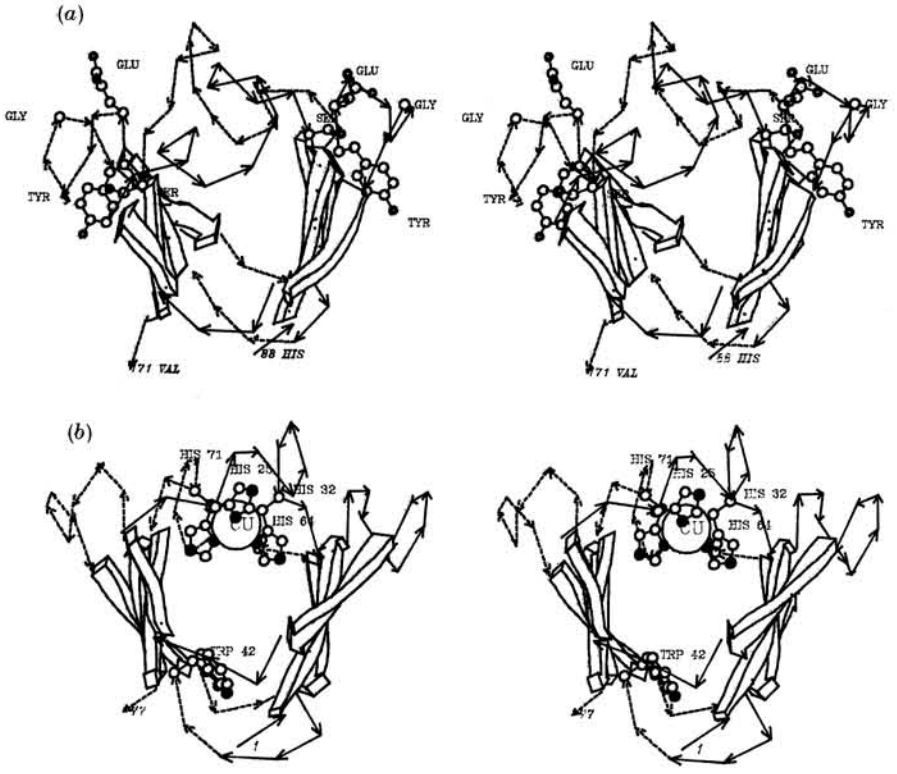

FIGURE 7. Stereo diagram of α-carbon atoms of (*a*) a single crystallin domain (γ-IV-2nd-domain) with four conserved residue side chains labelled; (*b*) CRYSTANOVA with four histidine copper ligand side chains and non-symmetrical tryptophan labelled. Dotted lines connect α-carbon atoms in the second motif of each structure. β-sheets formed of four strands in each motif are shown as solid strands with arrows. Picture produced by a computer program written by Lesk & Hardman (1982, 1985)

Natural crystallins have no catalytic activity or other feature that can provide a sensitive assay of the integrity of the tertiary structure, so an attempt was made to introduce a symmetrical metal binding site (figure 7*b*) from the known structure of superoxide dismutase. Copper was chosen because it has excellent spectral properties which can be used to assay binding. The program COMPOSER was used to build CRYSTANOVA, based on the structure of the most symmetrical natural domain (figure 7*a*). The sequence was evolved as a result of careful analysis using computer graphics and analytical tests for the integrity of the model as well as molecular simulation. A synthetic gene was designed for expression of the final sequence in bacteria as a fusion protein (figure 8). The gene has been cloned and the expressed protein now awaits structural investigation.

FIGURE 8. The oligonucleotides used in the synthesis of the *Crystanova* gene. 255 base pairs; base usage A = 59, C = 83, G = 58, T = 55. Bold amino acid sequences 4–36 and 43–75 are identical. Capital His residues are active site residues. Italic amino acid residues are translations of extra nucleotides included to give restriction sites at ends of synthetic gene for cloning purposes. CN1–CN16 are oligonucleotides used to synthesize the complete gene sequence. Bold nucleotides are cleaved from gene when it is cut with *EcoRI*/*Hin*DIII.

REFERENCES

Akrigg, D., Bleasby, A. J., Dix, N. I. M., Findlay, J. B. C., North, A. C. T., Parry-Smith, D., Wootton, J. C., Blundell, T. L., Gardner, S. P., Hayes, F., Islam, S., Sternberg, M. J. E., Thornton, J. M. & Tickle, I. J. 1988 A protein sequence/structure database. *Nature, Lond.* **335**, 745–746.

Beppu, T., Nishimori, Y., Kawaguchi, Y., Hidaka, M. & Vozumi, T. 1982 Cloning of cDNA for prochymosin and its expression in E. coli. In *Genetics of industrial microorganisms*, (ed. T. Beppu), pp. 195–202. Tokyo: Kodansha.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. 1977 The protein data bank: a computer-based archival file for macromolecular structures, *J. molec. Biol.* **112**, 535–542.

Blundell, T. L., Carney, D., Gardner, S., Hayes, F., Howlin, B., Hubbard, T., Overington, J., Singh, D. A., Sibanda, B. L. & Sutcliffe, M. 1988 Knowledge-based protein modelling and design. *Eur. J. Biochem.* **172**, 513–520.

Blundell, T. L., Lindley, P., Miller, L., Moss, D., Slingsby, C., Tickle, I., Turnell, B. & Wistow, G. 1981 The molecular structure and stability of the eye lens: X-ray analysis of γ-crystallin II. *Nature, Lond.* **289**, 771–777.

Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E. & Thornton, J. M. 1987 Knowledge-based prediction of protein structures and the design of novel molecules. *Nature, Lond.* **326**, 347–352.

Bruccoleri, R. E., Naber, E. & Novotny, J. 1988 Structure of antibody hypervariable loops reproduced by a conformational search algorithm. *Nature, Lond.* **335**, 564–568.

Chothia, C., Lesk, A. M., Levitt, M., Amit, A. G., Mamurra, R. A., Phillips, S. E. & Poljak, R. J. 1986 The predicted structure of immunoglobulin D1.3 and its comparison with the crystal structure. *Science, Wash.* **233**, 755–758.

Efimov, A. V. 1986 Standard conformations of a polypeptide chain in irregular regions of proteins. *Molec. Biol., Moscow* **20**, 250–260.

Eventoff, W. & Rossmann, M. G. 1975 The evolution of dehydrogenases and kinases. *Crit. Rev. Biochem.* **3**, 111–140.

Gething, M. J., Adler, B., Boose, J.-A., Gerard, R. D., Madison, E. L., McGookey, D., Meidel, R. S., Roman, L. M. & Sambrook, J. 1988 Variants of human tissue-type plasminogen activator that lack specific structural domains of the heavy chain. *EMBO J.* **7**, 2731–2740.

Harris, T. J. R. 1989 Second-generation plasminogen activators. *Protein Engng.* (In the press.)

Harris, T. J. R., Lowe, P. A., Lyons, A., Thomas, P. G., Eaton, M. A. W., Millican, T. A., Patel, T. P., Bose, C. C., Carey, N. H. & Doel, M. T. 1982 Molecular cloning and nucleotide sequence of cDNA coding for calf preprochymosin. *Nucl. Acids Res.* **10**, 2177–2187.

Johnson, M. S., Sutcliffe, M. J. & Blundell, T. L. 1989 Molecular anatomy: phylogenetic relationships derived from three-dimensional structures of proteins. *J. molec. Evol.* (In the press.)

Lesk, A. M. & Hardman, K. D. 1982 Computer-generated schematic diagrams of protein structure. *Science, Wash.* **216**, 539–540.

Lesk, A. M. & Hardman, K. D. 1985 Computer-generated pictures of proteins. *Methods Enzymol.* **115**, 381–390.

McGregor, M. J., Islam, S. A. & Sternberg, M. J. E. 1987 Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. *J. molec. Biol.* **198**, 295–310.

Milner-White, J. & Poet, R. 1986 Four classes of β-hairpins in proteins. *Biochem. J.* **240**, 289–292.

Ny, T., Sawdey, M., Lawrence, D., Millan, J. L. & Loskutoff, D. J. 1986 Cloning and sequence of a cDNA coding for the human β-migrating endothelial-cell-type plasminogen activator inhibitor. *Proc. natn. Acad. Sci. U.S.A.* **83**, 6776–6780.

Overington, J., Sutcliffe, M., Watson, F., Campbell, S., James, K. & Blundell, T. 1989 The modelling of the serine proteinase domain of tissue-type plasminogen activator and its interaction with inhibitors. *Proc. Int. Biotech. Symp.* (ed. G. Durand), pp. 279–304. Société Francaise de Biologie.

Pennica, D., Holmes, W. E., Kohr, W. J., Harkins, R. N., Vehar, G. A., Ward, C. A., Bennet, W. F., Yelverton, E., Seeburg, P., Heynecker, H. L., Goeddel, D. V. & Collen, D. 1983 Cloning and expression of human tissue-type plasminogen activator cDNA in *E. coli. Nature Lond.* **301**, 214–221.

Ponder, J. W. & Richards, F. M. 1987 Tertiary templates for proteins. *J. molec. Biol.* **193**, 775–791.

Reid, L. & Thornton, J. M. 1989 Rebuilding flavodoxin from C_α coordinates. *Proteins.* (In the press.)

Sanger, F., Nicklen, S. & Coulson, A. R. 1977 DNA sequencing with chain terminating inhibitors. *Proc. natn. Acad. Sci. U.S.A.* **74**, 5463–5467.

Sibanda, B. L., Blundell, T. L. & Thornton, J. M. 1989 The conformation of β-hairpins in protein structures. *J. molec. Biol.* **207**. (In the press.)

Sibanda, B. L. & Thornton, J. M. 1985 β-hairpin families in globular proteins. *Nature, Lond.* **316**, 170–174.

Singh, J. & Thornton, J. M. 1985 The interaction between phenylalanine rings in proteins. *FEBS Lett.* **191**, 1–6.

Singh, J., Thornton, J. M., Snarey, M. & Campbell, S. F. 1987 The geometries of interacting arginine-carboxyls in proteins. *FEBS Lett.* **224**, 161–171.

Smith, M. 1985 *In vitro* mutagenesis. *A. Rev. Genet.* **19**, 423–462.

Summers, N. L., Carlson, W. D. & Karplus, M. 1987 Analysis of side-chain orientations in homologous proteins. *J. molec. Biol.* **196**, 175–198.

Sutcliffe, M. J. 1988 An automated approach to the systematic modelling of homologous proteins. Ph.D. thesis, University of London.

Sutcliffe, M. J., Haneef, I., Carney, D. & Blundell, T. L. 1987*a* Knowledge-based modelling of homologous proteins, part I: three-dimensional frameworks derived from simultaneous superposition of multiple structures. *Protein Engng* **1**, 377–384.

Sutcliffe, M. J., Hayes, F. R. F. & Blundell, T. L. 1987*b* Knowledge-based modelling of homologous proteins part II: rules for the conformations of substituted sidechains. *Protein Engng* **1**, 385–392.

Taylor, J. W., Ott, J. & Eckstein, F. 1985 The rapid generation of oligonucleotide-directed mutations at high frequency using phosphorothioate-modified DNA. *Nucl. Acids Res.* **13**, 8765–8785.

Tembe, B. L. & McCammon, J. A. 1984 Ligand–receptor interactions. *Computers Chem.* **8**, 281.
Thornton, J. M., Singh, J., Campbell, S. F. & Blundell, T. L. 1988 Protein–protein recognition via side-chain interactions. *Biochem. Soc. Trans.* **16**, 927–930.
White, H. E., Driessen, H. P. C., Slingsby, C., Moss, D. S. & Lindley, P. F. 1989 Structure analysis of bovine lens γIVa–crystallin symmetry and lattice contacts. (In preparation.)
Winter, G. A. R., Fersht, A. J., Wilkinson, M., Zoller, M. & Smith, M. 1982 Redesigning enzyme structure by site-directed mutagenesis. *Nature, Lond.* **299**, 756–758.
Zoller, M. J. & Smith, M. 1982 Oligonucleotide-directed mutagenesis using M13 derived vectors. *Nucl. Acids Res.* **10**, 6487–6500.