Structure-based drug design

Tom L. Blundell

The three-dimensional structures of more than 4,000 macromolecules have already been solved, and the number will continue to increase steadily. Many of these macromolecules are important drug targets and it is now possible to use the knowledge of their three-dimensional structure as a good basis for drug design.

THE revolution in biology over the past two decades has resulted in radical new opportunities for drug discovery. Most importantly it has defined major drug targets in the form of molecular components of disease processes, which are now being developed for use in automated assays. Once lead compounds have been identified by screening natural compounds, chemical databanks or combinatorial libraries, the target macromolecules can provide a starting point for structure-based approaches (see ref. 1 for a review). These involve definition of the topographies of the complementary surfaces of ligands and their macromolecular targets. Here I describe recent progress in using such knowledge of the three-dimensional structures of receptor or target proteins as a basis for drug design.

Three-dimensional structures and their accuracy

Although the number of macromolecular three-dimensional structures increased linearly for about 30 years[2], more powerful synchrotrons, better X-ray detectors, faster computers and graphics and new multi-dimensional nuclear magnetic resonance (NMR) methods have changed the position radically in the past decade[3,4]. X-ray and NMR approaches have both taken advantage of the expression and purification of stable domains, substructures and mutants of often complex proteins such as receptor tyrosyl kinases, oncogenes and repressors (Fig. 1). As a consequence, the number of protein structures has begun to rise exponentially, like that of sequences over the previous decade. Now there are more than 4,000 macromolecular three-dimensional structures in the Protein Data Bank (refs 2, 5) and many of these are key drug targets (Figs 2 and 3).

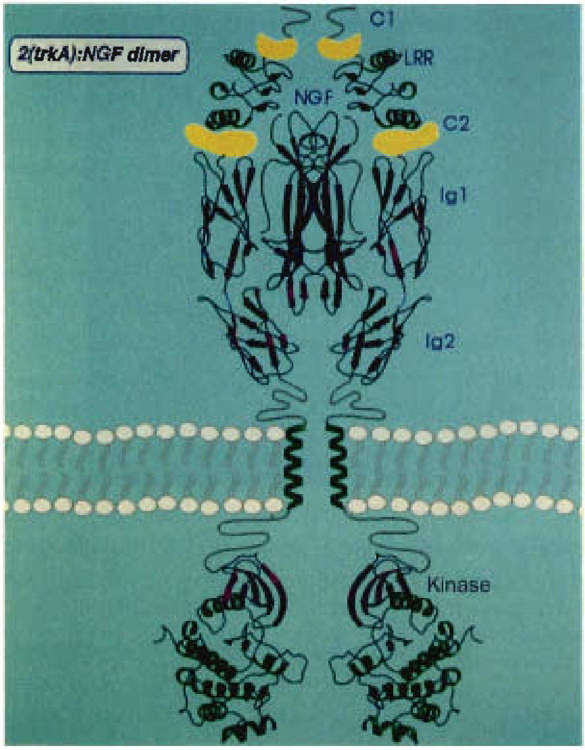

FIG. 1 A schematic representation of the nerve growth factor (NGF) complex with its receptor tyrosyl kinase, TrkA. Figure devised by Judith Murray Rust on the basis of the crystal structure of NGF and comparative models of the receptor domains. C1 and C2, cystine-rich regions; Ig1 and Ig2, immunoglobulin-like domains; LLR, leucine-rich region.

The accuracy required of a macromolecular structure reflects the use to which it will be put. If the design is predicated on the assumption that a lead molecule will complement a known binding site precisely, an accurate model will be required at the highest resolution possible, although designers must remember that proteins are flexible and can easily accommodate small changes. However, if the designer wishes only to know the general availability of space, essential hydrogen bonds, key electrostatic interactions or where to cyclize ligand groups, a rough model may be adequate.

Of course, the accuracy of a three-dimensional structure depends on the refinement, the resolution and the restraints introduced in the structure analysis[3,6]. However, much structure-based design appears to assume that the structure is correct, precise and rigid. Modelling software should perhaps oblige the user to know more about the experimental approach, the statistical parameters indicating the agreement between model and data and the thermal parameters giving clues about disorder, which are available in the original Protein Data Bank files.

Interactive graphics and lead development

Once the three-dimensional structure of a target protein has been defined, then computational procedures are required to suggest ligands that will bind at the active site. This can be approached either by elaborating a known ligand, preferably where the protein–ligand complex has been defined by X-ray analysis, or by searching for ligands (database approach) or molecular fragments (construction approach) that complement the receptor topography[7].

Structure-based design begins with the graphical display of hydrogen bonds, molecular surfaces[8,9] and electrostatic fields[10,11]. Traditionally, key interactions have been identified visually from three-dimensional structures of macromolecular ligand complexes, as illustrated in Fig. 2. New ligand designs are then explored that optimize a transition-state isostere, modify groups to improve complementarity and cyclize side groups to increase the rigidity of the ligand. Other important objectives include modification of bonds susceptible to hydrolysis, such as peptide bonds, decrease in size of ligand to assist nasal or oral absorption and identification of sites for modulation of physical–chemical properties to improve bioavailability.

Making the new molecules

Although some attempts have been made at interactive docking of a putative ligand molecule into a receptor site[12–14], it is more effective to evaluate the electrostatic, steric or more complex energy terms during a systematic search of rotational and translational space for the two molecules[15,16]. Computational time can be reduced by precalculating terms for each point on a grid using electrostatic terms for probes[17] or by using pseudo-energies calculated from pairwise distributions of atoms in protein complexes or crystals of small molecules[18,19]. Probe molecules are fitted to these potentials and ranked according to energy.

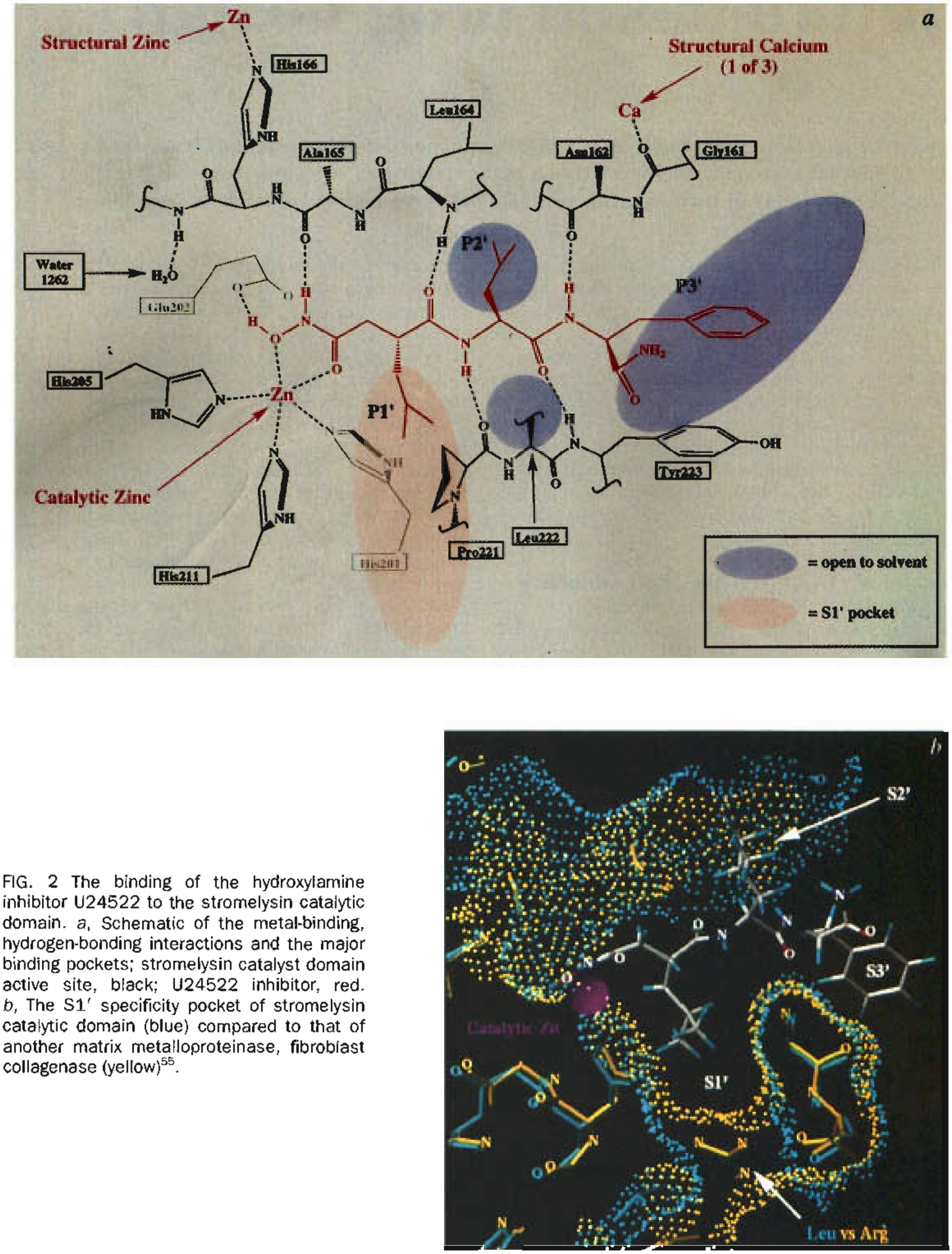

FIG. 2 The binding of the hydroxylamine inhibitor U24522 to the stromelysin catalytic domain. *a*, Schematic of the metal-binding, hydrogen-bonding interactions and the major binding pockets; stromelysin catalyst domain active site, black; U24522 inhibitor, red. *b*, The S1′ specificity pocket of stromelysin catalytic domain (blue) compared to that of another matrix metalloproteinase, fibroblast collagenase (yellow)[55].

For example, the program DOCK[20] creates a negative image of the target site and selects and ranks putative ligands on the basis of a comparison of internal distances. Up to 100,000 compounds can be examined in a week. However, difficulties arise in finding the proper conformation and in discriminating between putative interaction modes of similar energy. Procedures for matching involving genetic algorithms[21] and graph theory[22] can also be used to generate molecular structures within constraints of an enzyme active site or a receptor binding site.

Alternatively, fragments can be positioned in the binding cleft of macromolecules and then 'grown' to fill the space available, exploring the electrostatic, van der Waals or hydrogen-bonding interactions involved in molecular recognition[7]. For example, GROWMOL[23] gives multiple highly diverse structures complementary to active sites, GenStar[24] generates chemically reasonable structures from sp^3 carbons to fill the binding site, whereas the multiple-copy simultaneous search (MCSS) method[25] maps out the structure by determining energetically favourable positions and orientations of functional groups on the receptor surface. LUDI[26] positions molecules or new substituents into clefts so that hydrogen bonds are formed and hydrophobic pockets are filled with hydrocarbon groups. A standard library of 1,000 fragments is used to fit the interaction sites and a further library of 1,200 link fragments is used to connect these into a single molecule. Such methods depend on the existence of large databases of small molecule structures such as the Cambridge Structure

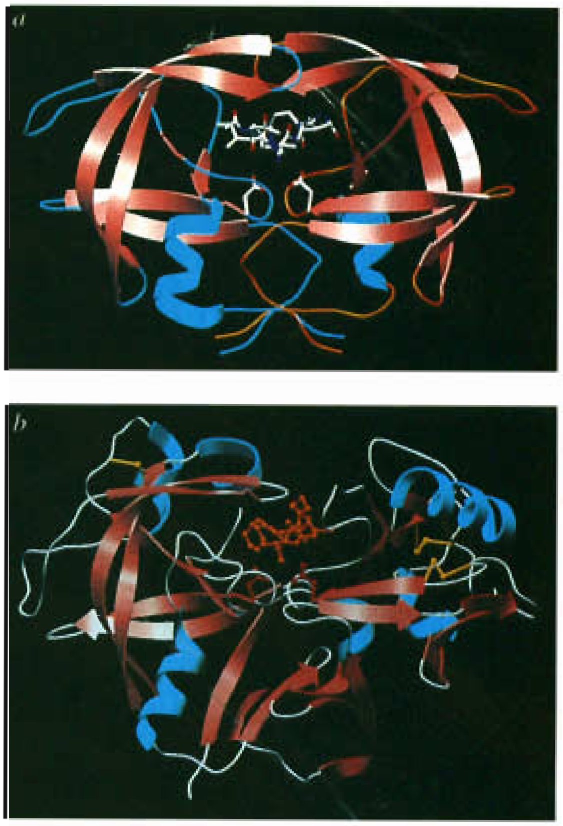

FIG. 3 Structure of *a*, HIV proteinase and *b*, human renin complexed with inhibitors viewed along the active site. Figures derived from coordinates of *a*, 4hvp (ref. 52) and *b*, human renin CP85339 (ref. 56).

Data Base, which contains 100,000 crystal structures[27], or the Fine Chemicals Directory where molecular formulae can be automatically processed into a useful three-dimensional representation by CONCORD[28].

Comparative modelling on a common fold

Where there is no three-dimensional structure of the target, a protein with a similar fold can provide the basis for constructing a useful model[29–31]. For homologous proteins with sequence identities >30%, the common fold can be recognized by sequence searches. For more distantly related proteins, profiles or templates are useful in the search for the common fold and alignment of the sequences[32–38]. Once a related fold is identified, this can be used to model the three-dimensional structure. Most methods depend on the assembly of rigid fragments[39–41], which are used in programs such as COMPOSER to define first the framework, second, the structurally variable, mainly loop regions and, third, the side chains[29,42–44]. An alternative approach, encoded in MODELLER[45], seeks to satisfy structural restraints derived from homologues and other proteins and expressed as probability density functions. These modelling procedures are most successful where the percentage sequence identity to the unknown is high (greater than 40%)[46]. In the absence of a common fold, combinatorial approaches[47], which bring together prediction of secondary and supersecondary structures with procedures for docking these together, can be used to predict the tertiary structure.

Successes, contributions and problems

Structure-based approaches have already played a role in the discovery of several drugs now in clinical use. One of the most notable examples has been the development of HIV inhibitors as AIDS antivirals[48]. An early report[49] that retroviruses code a proteinase that is related to the aspartic proteinases was reminiscent of our earlier suggestion[50] that aspartic proteinases are evolved from a symmetrical dimer. We suggested that HIV proteinase might be an analogous symmetrical dimer, and this was supported by several approximate three-dimensional models[40,51] and later by X-ray structures[48,52,53]. This gave important clues about inhibitors. Structures of several hundred inhibitor complexes have been experimentally defined, providing a previously unparalleled structural database for design (see, for example, Fig. 3*a*). These complexes have exploited a range of different structural features, including 2-fold symmetry in the ligand, replacement of a bound water molecule, cyclization and replacement of scissile peptide bonds; see, for example, the cyclic, symmetrical inhibitor of the Dupont–Merck team[54]. Several studies have involved the use of programs such as DOCK or fragment searching to identify non-peptidic structures. Useful molecules are now exploited as cocktails in the treatment of AIDS, with encouraging results, although it is evident that mutation in HIV allows the virus to escape quickly if challenged with a single antiviral agent.

Similar approaches have been used to design inhibitors of a number of other drug targets. These include antihypertensives that inhibit human renin (Fig. 3*b*), anticancer and antiarthritis inhibitors of matrix metalloproteinases, such as collagenase and stromelysin (Fig. 2), selective immunosuppressants that target purine nucleotide phosphorylase, agents for treatment of the common cold that bind the rhinovirus canyon, and antiproliferative agents that inhibit thymidylate synthase (see ref. 1 for review). Most of these have been carried out in pharmaceutical companies where thousands of crystal structure analyses have been carried out in recent years. New compounds have been developed that would never have arisen from conventional drug discovery techniques.

One major challenge for drug discovery is a consequence of the very large surfaces that characterize many of the protein complexes involved in receptor recognition and signal transduction. This is illustrated by the diagram of the nerve growth factor interaction with its receptor tyrosyl kinase, Trk (Fig. 1). Not only would it appear difficult to bind a small molecule to the large, relatively flat surfaces of many proteins involved in protein interactions, in contrast to the deep clefts of many

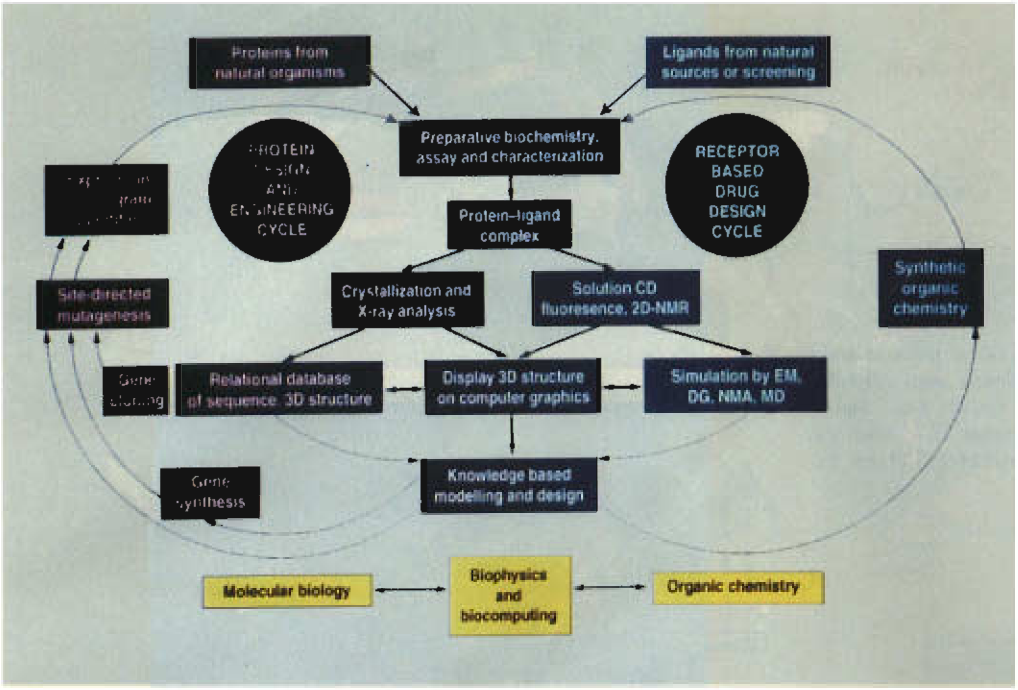
Proteins from natural organisms
Ligands from natural sources or screening
Preparative biochemistry, assay and characterization
PROTEIN DESIGN AND ENGINEERING CYCLE
RECEPTOR BASED DRUG DESIGN CYCLE
Protein–ligand complex
Site-directed mutagenesis
Crystallization and X-ray analysis
Solution CD fluoresence, 2D-NMR
Synthetic organic chemistry
Gene cloning
Relational database of sequence 3D structure
Display 3D structure on computer graphics
Simulation by EM, DG, NMA, MD
Gene synthesis
Knowledge based modelling and design
Molecular biology
Biophysics and biocomputing
Organic chemistry

FIG. 4 A multidisciplinary design bi-cycle. CD, Circular dichroism; NMR, nuclear magnetic resonance; EM, energy minimization; DG, distance geometry; NMA, normal mode analysis; MD, molecular dynamics.

enzymes, but it would also be difficult to disrupt the interaction entirely even if one did.

Future of structure-based design

Iterative optimization of lead compounds usually involves multidisciplinary design cycles (Fig. 4) starting from the cloning, expression, characterization and definition of the three-dimensional structure of the protein or nucleic acid, preferably as a complex with a ligand or a pseudo-substrate. This structure is the basis for suggesting modifications either to the ligand (to be introduced by the chemists) or to the macromolecule (to be introduced by the genetic engineer). The latter cycle will be of value when the molecule of interest is itself a protein, for example for 'humanizing' monoclonal antibodies, for engineering enzymes and for modifying polypeptide hormones, growth factors or cytokines. It is increasingly used to test hypotheses about drug–receptor interactions, by mutating key residues on the receptor topography. In a perfect world a single cycle should produce an improved molecule. In practice, designs are developed iteratively using several cycles, each providing small improvements.

Structure-based approaches will undoubtedly be important in the design of new proteins, drugs and vaccines. The structure-based approaches have provided many surprises and the subsequent experimental stages have been a constant reminder of the fragility of our predictive ability. Indeed the sceptical chemist will rightly still want to test some of the structurally less-likely options and the imaginative chemist will no doubt have his own intuition as well. Furthermore, there is one major shortcoming of a rational approach to design. If one company can arrive at a more effective drug through a rational approach, then so can a competitor. Most drug companies will feel happier with a lead that they have chanced upon randomly in a screen or by combinatorial chemistry as others are less likely to have found the same molecule. However, this may be just the start of a rational process in which the interactions of the lead molecule with its target receptor are defined as I have described and an improved molecule is designed using a more structure-based approach. □

Tom L. Blundell is in the Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1QW, UK, and the Imperial Cancer Research Fund Unit of Structural Molecular Biology, Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, UK.

1. Whittle, P. J. & Blundell, T. L. *A. Rev. Biophys. biomolec. Struct.* **23**, 349–375 (1994).
2. Bernstein, F. C. *et al. J. molec. Biol.* **112**, 535–542 (1977).
3. Wuthrich, K. *Les Cahiers Fondation Louis Jeantet* **8**, 1–16 (1993).
4. Clore, M. & Gronenborn, A. *Prog. Biophys. molec. Biol.* **62**, 153–184 (1995).
5. Sussman, J. *Prot. Databank Q. Newslett.* **76** (April 1996).
6. Blundell, T. L. & Johnson, M. S. *Protein Crystallography* (Academic, London, 1976).
7. Cohen, N. C. & Tschinke, N. *Progr. Drug Res.* **45**, 205–235 (1995).
8. Langridge, R. *et al. Science* **211**, 661–666 (1981).
9. Connolly, M. L. *Science* **221**, 709–713 (1983).
10. Gilson, M. K., Sharp, K. A. & Honig, B. H. *J. comput. Chem.* **9**, 327–333 (1988).
11. Goodsell, D. S., Mian, I. S. & Olson, A. J. *J. molec. Graphics* **7**, 41–47 (1989).
12. Busetta, B., Tickle, I. J. & Blundell, T. L. *J. appl. Crystallogr.* **16**, 432–438 (1983).
13. Pattabiraman, N. *et al. J. comput. Chem.* **6**, 432–439 (1985).
14. Tomioka, N., Itai, A. & Iitaka, Y. *J. Comput. Aided molec. Design* **1**, 197–203 (1987).
15. Wodak, S. J. & Janin, J. *J. molec. Biol.* **124**, 323–329 (1983).
16. Kuntz, T. *et al. J. molec. Biol.* **161**, 269–278 (1982).
17. Goodford, P. J. *J. med. Chem.* **28**, 849–857 (1985).
18. Cruciani, G. & Goodford, P. J. *J. molec. Graphics* **12**, 116–129 (1994).
19. Pastor, M. & Cruciani, G. *J. med. Chem.* **38**, 4637–4647 (1995).
20. Leach, A. R. & Kuntz, I. D. *J. comput. Chem.* **13**, 730–748 (1992).
21. Payne, A. W. R. & Glen, R. C. *J. molec. Graphics* **11**, 76–83 (1993).
22. Lewis, R. A. *J. molec. Graphics* **10**, 31–38 (1993).
23. Bohacek, R. S. & McMartin, C. *J. Am. chem. Soc.* **116**, 5560–5565 (1994).
24. Rotstein, S. H. & Murcko, M. A. *J. Comput. Aided molec. Design* **7**, 23–43 (1993).
25. Miranker, A. & Karplus, M. *Proteins* **11**, 29–34 (1991).
26. Bohm, H. J. *J. Comput. Aided molec. Design* **6**, 61–78; 593–606 (1992).
27. Allen, F. H. *et al. Acta crystallogr.* **835**, 2331–2339 (1979).
28. Rusinko, A. *et al. J. chem. Inf. Comput. Sci.* **29**, 327–333 (1989).
29. Blundell, T. L. *et al. Nature* **326**, 347–352 (1987).
30. Šali, A. *et al. Trends biochem. Sci.* **15**, 235–240 (1990).
31. Johnson, M. S. *et al. Crit. Rev. biol. Chem. molec. Biol.* **29**, 1–70 (1994).
32. Taylor, W. R. *J. molec. Biol.* **188**, 233–258 (1986).
33. Gribskov, M. *et al. Proc. natn. Acad. Sci. U.S.A.* **84**, 4355–4358 (1987).
34. Ponder, J. W. & Richards, F. M. *J. molec. Biol.* **193**, 775–791 (1987).
35. Sippl, M. *J. molec. Biol.* **213**, 859–883 (1990).
36. Jones, D. T., Taylor, W. R. & Thornton, J. M. *Nature* **358**, 86–89 (1992).
37. Johnson, M. S., Overington, J. P. & Blundell, T. L. *J. molec. Biol.* **231**, 735–752 (1993).
38. Bowie, J. U., Luthy, R. & Eisenberg, D. *Science* **253**, 164–170 (1991).
39. Jones, T. H. & Thirup, S. *EMBO J.* **5**, 819–822 (1986).
40. Blundell, T. L. *et al. Eur. J. Biochem.* **172**, 513–520 (1988).
41. Claessens, M. *et al. Prot. Engng* **2**, 335–345 (1989).
42. Sutcliffe, M. J., Hayes, F. R. F. & Blundell, T. L. *Prot. Engng* **1**, 385–392 (1987).
43. Summers, N. L., Carlson, W. D. & Karplus, M. *J. molec. Biol.* **196**, 175–198 (1987).
44. Topham, C. *et al. J. molec. Biol.* **229**, 194–220 (1993).
45. Šali, A. & Blundell, T. L. *J. molec. Biol.* **234**, 779–815 (1993).
46. Srinivasan, N. & Blundell, T. L. *Prot. Engng* **6**, 501–512 (1993).
47. Presnell, S. R., Cohen, B. I. & Cohen, F. E. *Biochemistry* **31**, 983–988 (1992).
48. Wlodawer, A. & Erickson, J. *A. Rev. Biochem.* **62**, 543–585 (1993).
49. Toh, H., Ono, M., Saigo, K. & Miyata, T. *Nature* **315**, 691 (1985).
50. Tang, J., James, J., Jenkins, J. A. & Blundell, T. L. *Nature* **271**, 618–621 (1978).
51. Pearl, L. H. & Taylor, W. R. *Nature* **329**, 351–354 (1987).
52. Miller, M. *et al. Science* **246**, 1149–1152 (1989).
53. Lapatto, R. *et al. Nature* **342**, 299–302 (1989).
54. Lam, P. Y. S. *et al. Science* **263**, 380–384 (1994).
55. Dhanaraj, V. *et al. Structure* **4**, 375–386 (1996).
56. Dhanaraj, V. *et al. Nature* **357**, 466–472 (1992).

ACKNOWLEDGEMENTS. I thank J. Murray Rust, V. Dhanaraj and K. Guruprasad for preparation of figures and for helpful comments on the manuscript. I am grateful to J. Overington for providing an 'industry' view.