Eur J Cancer, Vol. 26, No. 11/12, pp. 1163–1166, 1990.
Printed in Great Britain

0277-5379/90 $3.00 + 0.00
© 1990 Pergamon Press plc

Feature Articles

Prediction of Protein Structure from Sequence

Michael J.E. Sternberg and Markéta J.J.M. Zvelebil

Methods to predict the three-dimensional structure of a protein from its sequence are reviewed. The approaches to derive information about the local conformation from the local sequence include hydrophobicity plots, secondary structure prediction and the identification of short, functional sequence motifs. The most reliable method of tertiary structure prediction is model building from the experimentally determined coordinates of a protein with an homologous sequence. This approach is illustrated by a prediction of the three-dimensional structure of human cytochrome P450-IA1. If no known homologous structure is available, then the only approach is to suggest models for the tertiary fold of proteins by packing together predicted secondary structures. A three-dimensional model for the dimerisation of the transmembrane α-helices of *neu*, a tyrosine kinase growth factor receptor, is described. In general, structure prediction can suggest approaches for regulating protein activity that may lead to new pharmaceutical therapies for cancer.
Eur J Cancer, Vol. 26, No. 11/12, pp. 1163–1166, 1990.

INTRODUCTION

THE GENES of many proteins of importance for cancer research are today being cloned, sequenced and expressed. A major problem is to relate the translated aminoacid sequence into information about the structure and function of the protein to guide the systematic design of experiments, for instance in mutagenesis and antibody mapping, and the development of novel pharmaceutical regulators of activity. We describe computer methods to predict protein structure from sequence [1, 2]. Structure prediction is vital since the conformations of only 400 proteins have been experimentally determined, as X-ray crystallography and two-dimensional nuclear magnetic resonance remain difficult and time-consuming. In contrast, the sequences of more than 15 000 proteins are known and this number will increase rapidly with the impetus of the genome projects.

ENERGY CALCULATIONS

In principle, it should be possible to locate the folded structure of a protein by searching for a conformation of minimum free energy. However, this approach is not feasible since there is a myriad of conformations to examine and the energy terms cannot be represented with sufficient accuracy. Thus all methods of prediction use rules developed from analysis of the experimentally known structures. Energy calculations can, however, be used to explore small variations of conformation given a rough starting model. One particularly powerful use of such calculations is when there is an experimentally-determined structure and one wishes to evaluate free energy differences between two closely related molecules, such as the effects of point mutations or the binding of different ligands to a protein [3].

ANALYSIS OF LOCAL STRUCTURE AND FUNCTION

Hydrophobicity profile

The simplest form of analysing a local sequence is the hydrophobicity plot in which a moving average of the hydrophobicity of the residues is evaluated along the sequence. The most commonly used scales of hydrophobicity are those of Kyte and Doolittle [4] and of Hopp and Woods [5]. When an average over 7 residues is calculated, the peaks of hydrophilicity (troughs of hydrophobicity) are likely to be surface loops and the sequences of these regions are good candidates for the synthesis of linear peptides against which antibodies can be raised that could recognise the native protein. The location of membrane spanning regions can be identified from the peaks of hydrophobicity in a plot averaged over 15 residues.

Secondary structure prediction

The location of regular secondary structure, α-helices and β-sheet strands, can also be predicted from local sequence. Some algorithms, such as that of Chou and Fasman [6] and of Robson [7], are based on empirically derived statistical propensities Others, such as Lim's [8], recognise the periodicities of hydrophobic residues on the surfaces of the secondary structures. Sophisticated analyses based on neural networks [9, 10] or artificial intelligence [11] have been used. However, the accuracy for a general prediction of α-helices, β-sheet or coil remains at about 65% compared with a random value of about 33%. Improvement of up to 10% can be obtained if the sequences of a family of related proteins are used both to average the statistical propensities and to include the observation that sequence variations tend to occur in the loop regions [12].

Local sequence motifs for function

A common local sequence motif often occurs in several proteins due to a functional requirement. These motifs generally are quite simple and can be represented by a string of one or more residue types that occur in a short region of the protein sequence. The motifs are derived by examination of a family of

Correspondence to M.J.E. Sternberg.
M.J.E. Sternberg is at the Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, 44 Lincoln's Inn Fields, London WC2A 3PX, U.K.; M.J.J.M. Zvelebil was previously at the Imperial Cancer Research Fund and is now at the Biomolecular Structure and Modelling Unit, University College, London, U.K.

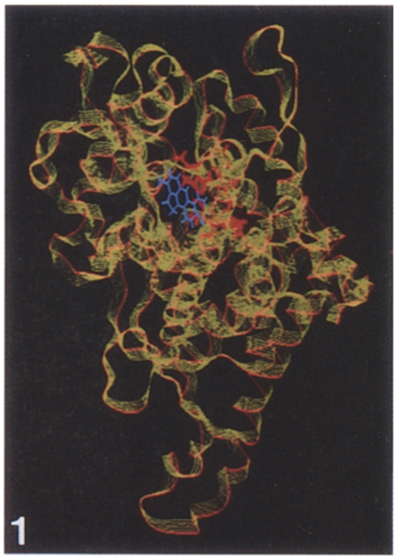

Fig. 1. Predicted three-dimensional model of human cytochrome P450-IA1. Fold of chain is illustrated by a ribbon, with α-helices forming spirals. Haem group is in red and the proposed position of 3-methylcholanthrene is in blue. Graphics software, QUANTA/ CHARMm, was from Polygen (Reading, Berks, UK).

proteins with a common function and identifying a local region of high sequence conservation. There are several types of such functional motifs that can specify post-translational modification (e.g. glycosylation), the localisation of a protein (e.g. endoplasmic reticulum targeting), a specific biological function (e.g. ATP/GTP binding) or a protein active site (e.g. aspartyl proteinase). For example, the sequence motif for a ATP/GTP binding site is (A or G)-X-X-X-X-G-K-(S or T), where X is any aminoacid. A list of over 300 such patterns is available from Dr A. Bairoch via the EMBL Data Library at Heidelberg. We have developed a program (PROMOT) that will scan a new sequence against the Bairoch data base to locate pattern matches.

When a functional motif common to two proteins is found, it does not always imply that there is a structural similarity. Some patterns such as glycosylation are dictated by the linear sequence and even locally the protein can adopt a different conformation. Others, such as the ATP/GTP binding site, occur due to a common local structure for the function but the tertiary fold of several proteins with this motif has been shown by crystallography to be different. However, a common pattern for an enzyme active site strongly suggests that the protein domains will have a similar three-dimensional fold.

PREDICTION OF TERTIARY STRUCTURE

Homology searches

Often all, or a domain, of a newly determined protein is found to have sequence similarity to another molecule because of divergent evolution of a protein family. The standard procedure to detect this relation is a computer scan against a data base of all protein sequences. For these scans, a matrix that scores the similarity between residues is required. The simplest of such scores is to flag identities but greater sensitivity occurs if the matrix evaluates the likelihood of one residue mutating to another [13]. The most sensitive searches now identify local [14] rather than global [15] sequence similarities and thus can detect the presence of a common domain within two larger proteins. Protein sequence similarity suggests that the related regions will have a common structure and function.

Model-building by homology

When a homology search finds a relation between a new sequence and a protein of experimentally known structure, it is possible to predict accurately a three-dimensional structure for the new protein. Model-building by homology is based on the observation from X-ray structures of homologous proteins that the protein core of α-helices and/or β-sheets tends to be conserved structurally during evolution with the major conformation changes occurring with insertions and deletions in the connecting loop regions.

The approach for model building will be illustrated by our recent prediction [16] of the structure of human cytochrome P450 from the IA1 gene family (Fig. 1). The primary role of the P450 family of enzymes is the detoxification of a wide range of substrates by oxidation and other chemical reactions. However, these reactions also result in the activation of many procarcinogens. Indeed susceptibilities to certain forms of cancer have been linked to the presence or absence of different members of the P450 gene family [17]. Since there is no experimental structure for any mammalian P450, to guide experimental work we have predicted models from the crystal structure [18] of the related P450 that metabolises camphour (P450-cam).

The first step was to establish an accurate sequence alignment of P450-IA1 with distantly-related P450-cam. The sequences of several mammalian P450s were automatically aligned [19] and then a secondary structure prediction was done on the multiply-aligned sequences [12]. Predicted secondary structures were matched against the known α-helices and β-sheets in P450-cam. Then the precise sequence equivalence was based on the alignment of functionally conserved residues, the matching of non-polar sidechains of P450-IA1 with hydrophobic residues buried in P450-cam, and ensuring that sequence insertions primarily occurred in the loop regions.

Based on this alignment, the core regions that, to a first approximation, will be conserved in structure between P450-IA1 and cam were identified. The coordinates of the core region in P450-cam was used to model the P450-IA1 main-chain conformation. The next step was to model the connecting loops where there will be structural differences. The approach was to scan a data-base of the experimentally-determined structures of proteins to extract a suitable fragment that could make the necessary connection [20]. This knowledge-based approach provided a series of possible connections. Further inspection on a molecular graphics package suggested which connection was compatible with the model of P450-IA1. Generally, loop selection remains uncertain and simply finds a likely conformation. With the main chain built for P450-IA1, the appropriate sidechains were placed, based on their probable conformation. Energy minimisation was used to make minor adjustments to the model and remove any steric clashes that have been introduced during model building. Finally a possible model for the interaction of P450-IA1 with its substrate, 3-methylcholanthene, was obtained.

The predicted structure of P450-IA1 provides information for several experimental studies. The exposed loops, which are likely epitopes, have been located and thus peptides with the sequences of these loops could be used to raise antibodies. From the sequence alignment between P450-IA1 and P450-cam, possible active site residues of P450-IA1 have been suggested and these residues are targets for mutagenesis.

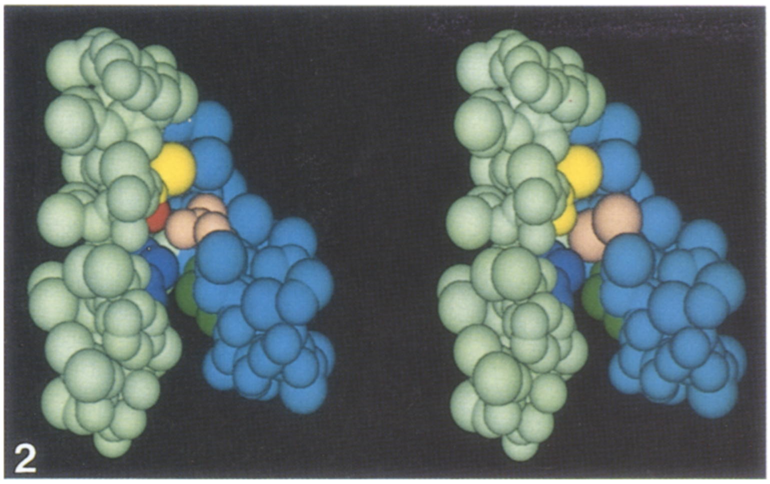

Fig. 2. Space-filling models for packing of part of the transmembrane α-helices in *neu* [26]. Left helix is mainly in light green, right in light blue. On left, *neu* with Glu 664 in right helix in pink forming proposed hydrogen bond between carbonyl oxygen (red) of Ala 661 of left helix. Exact conformation for side-chains cannot be modelled and figure just shows that hydrogen bond could be formed. On right *neu* with Val 664 in right helix in pink packing against Gly 665 in left helix.

A similar approach was followed to predict a three-dimensional model for cytochrome P450-17α from the coordinates of P450-cam [21]. P450-17α catalyses two key steps in the biosynthesis of the androgens from pregnanes. Since 80% of prostatic cancers are sensitive to the levels of androgens, especially testosterone, there is a search for selective inhibitors of this enzyme. The modelling suggested that there may be two modes of binding steroid substrates at the enzyme active site.

Improvements in the above strategy can be made if crystal structures are known for several proteins in the family. The parent structure for modelling is chosen based on careful analysis of sequence similarities between the protein to be predicted and the known structures. When a loop is being modelled, often an appropriate model is an analogous loop from one member of the known structures. This approach is particularly useful in predicting the hypervariable regions of antibodies, because for many of the complementary determining regions there are defined families of loop conformations that can be readily identified from sequence [22]. Thus predictions of high accuracies are now being obtained for the hypervariable loops of antibodies.

Tertiary packing of α-helices and β-sheets

When no experimental structure is available for prediction by homology, the only approach is to pack together predicted α-helices and β-sheets. The rules governing the packing geometry of these secondary structures were obtained by examination of the X-ray structures of proteins. The most amenable system for modelling is the packing of a few α-helices and this approach has led to predictions for the three-dimensional fold of α-interferon [23] and of interleukin-2 [24]. Subsequently, the model for interleukin-2 was found to be partly correct when compared with the X-ray structure [25].

This approach of packing α-helices has been applied to model the transmembrane region of the *neu* receptor, a tyrosine kinase growth receptor [26, 27]. The experimental observation [28] that promoted this study was that c-*neu* can be activated into an oncogene (onc-*neu*), with a marked increase in tyrosine kinase activity, by a single point mutation Val 664 → Glu in the transmembrane region. In general, the intracellular tyrosine kinase activity of this family of receptors is considered to be activated by dimerisation promoted by extracellular binding of the growth factor. It was suggested [26, 27] that the Glu side-chain in *neu* was protonated in the hydrophobic environment of the membrane and thus able to form a hydrogen bond. A detailed model (Fig. 2) for a dimerisation of a pair of transmembrane α-helices was obtained in which the helix association was stabilised in onc-*neu* by the Glu forming a hydrogen bond. In c-*neu*, the same helix association would occur but would have a lower free energy of association since the Val side-chain would interact simply by packing forces rather than a hydrogen bond. Analysis of the sequences of transmembrane regions of the entire family of tyrosine kinase growth factor receptors suggested that the dimerisation of the transmembrane α-helices may be a general phenomenon. If this is correct, then the dimerisation of a specific receptor might be inhibited by a peptide with the sequence of the transmembrane region because the peptide would compete with one molecule of the receptor during dimerisation (Fig. 3). Such a peptide could have a therapeutic role for those cancers associated with over-expression or mutation of growth factor receptors [29].

CONCLUSION

The strategy for predicting structure from a newly-determined sequence is shown in Fig. 4. Of course, the model obtained by prediction is less reliable than a protein crystal structure and it is important, therefore, to evaluate critically its likely accuracy. Models obtained from a close homology with an experimental structure will have the correct overall fold of the α-helices and

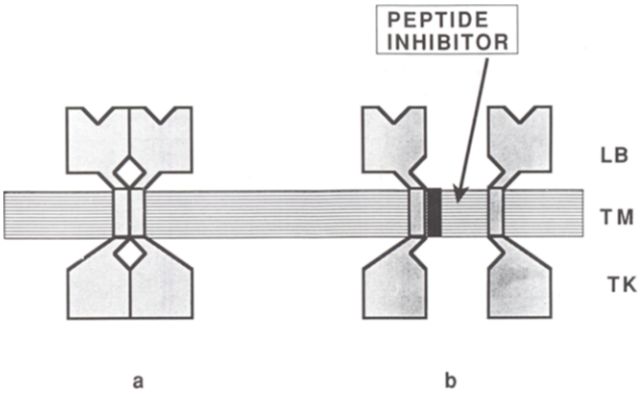

Fig. 3. Use of peptide to inhibit dimerisation of tyrosine kinase growth factor receptors. LB, TM and TK denote ligand-binding, transmembrane and tyrosine kinase domains of receptor. (a) Modelling of *neu* suggests that transmembrane regions of receptors pack together during activation of tyrosine kinase activity. (b) Peptide with sequence of transmembrane region might be able to prevent dimerisation.

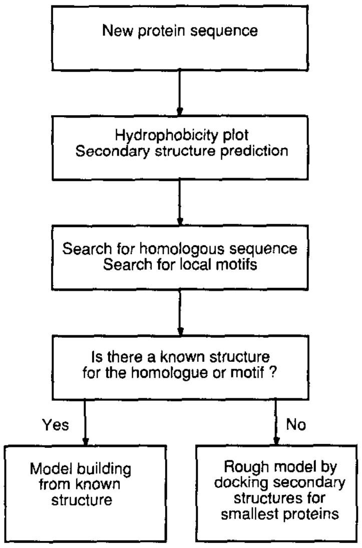
New protein sequence
Hydrophobicity plot
Secondary structure prediction
Search for homologous sequence
Search for local motifs
Is there a known structure for the homologue or motif ?
Yes
No
Model building from known structure
Rough model by docking secondary structures for smallest proteins

Fig. 4. Approach to analysis of protein sequence for structure prediction.

β-sheets but the conformation of loops cannot generally be accurately modelled. As the level of sequence homology decreases, the model becomes less reliable with even some of the α- and β-regions being subject to error. Predictions based on identification of a common local sequence motif tend to be less reliable than those obtained from a global sequence homology. Finally, models obtained by docking secondary structures should be regarded simply as an hypothesis that suggests further experimental work. However, despite these limitations, protein structure prediction provides a powerful method to obtain rapidly three-dimensional information from the ever-increasing number of gene sequences. This information suggests approaches to regulate the activity of proteins and over the next decade predictions might well lead to the development of novel treatments for cancer.

1. Blundell TL, Sibanda BL, Sternberg MJE, Thornton JM. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* 1987, **326**, 347–352.
2. Fasman GD. *Prediction of Protein Structure and the Principles of Protein Conformation.* New York, Plenum Press, 1989.
3. Merz KM, Kollman PA. Free energy perturbation simulations of the inhibition of thermolysin: prediction of the free energy of binding of a new inhibitor. *J Am Chem Soc* 1989, **111**, 5649–5658.
4. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982, **157**, 105–132.
5. Hopp TP, Woods KR. Prediction of protein antigenic determinants from amino acid sequences. *Proc Natl Acad Sci* USA 1981, **78**, 3824–3828.
6. Chou PY, Fasman GD. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol* 1978, **47**, 45–148.
7. Garnier J, Osguthorpe DJ, Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol* 1978, **120**, 97–120.
8. Lim VI. Algorithms for prediction of alpha-helical and beta-structural regions in globular proteins. *J Mol Biol* 1974, **88**, 873–894.
9. McGregor MJ, Flores TP, Sternberg MJE. Prediction of beta-turns in proteins using neural networks. *Protein Eng* 1989, **2**, 521–526.
10. Kneller DG, Cohen FE, Langridge R. Improvements in protein secondary structure prediction by an enhanced neural network. *J Mol Biol* 1990, **214**, 171–182.
11. King RD, Sternberg MJE. A machine learning approach for the prediction of protein secondary structure. *J Mol Biol* 1990 (in press).
12. Zvelebil MJJM, Barton GJ, Taylor WR, Sternberg MJE. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J Mol Biol* 1987, **195**, 957–961.
13. Dayhoff M. *Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation, Silver Spring, MD, 1978.
14. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981, **147**, 195–197.
15. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 1970, 443–453.
16. Zvelebil MJJM, Wolf CR, Sternberg MJE. A predicted three-dimensional structure of human cytochrome P450: implications for substrate specificity. *Protein Eng* 1990 (in press).
17. Gough AC, Miles JS, Spurr NK *et al.* Identification of the primary gene defect at the cytochrome P450 CYP2D, debrisoquine hydroxylase, locus and association with susceptibility to lung and bladder cancer. *Nature* 1990, **347**, 773–776.
18. Poulos TL, Finzel BC, Howard AJ. High-resolution crystal structure of cytochrome P450cam. *J Mol Biol* 1987, **195**, 687–700.
19. Barton GJ, Sternberg MJE. A strategy for the rapid multiple alignment of protein sequences—Confidence levels from tertiary structure comparisons. *J Mol Biol* 1987, **198**, 327–337.
20. Bates PA, McGregor MJ, Islam SA, Sattentau QJ, Sternberg MJE. A predicted three-dimensional structure for the human immunodeficiency virus binding domains of CD4 antigen. *Protein Eng* 1989, **3**, 13–21.
21. Laughton CA, Neidle S, Zvelebil MJJM, Sternberg MJE. A molecular model for the enzyme cytochrome P450-17α, a major target for the chemotherapy of prostatic cancer. *Biochem Biophys Res Commun* (in press).
22. Chothia C, Lesk AM, Tramontano A, *et al.* Conformations of immunoglobulin hypervariable regions. *Nature* 1989, **342**, 877–883.
23. Sternberg MJE, Cohen FE. The prediction of the secondary and tertiary structures of interferon from four homologous amino acid sequences. *Int J Biol Macromol* 1982, **4**, 137–144.
24. Cohen FE, Kosen PA, Kuntz ID, Epstein LB, Ciardelli TL, Smith KA. Structure-activity studies of interleukin-2. *Science* 1986, **234**, 349–352.
25. Brandhuber BJ, Boone T, Kenney WC, McKay DB. Three-dimensional structure of interleukin-2. *Science* 1987, **238**, 1707–1709.
26. Sternberg MJE, Gullick WJ. Neu receptor dimerization. *Nature* 1989, **339**, 587–
27. Sternberg MJE, Gullick WJ. A sequence motif in the transmembrane region of growth factor receptors with tyrosine kinase activity mediates dimerization. *Protein Eng* 1990, **3**, 245–248.
28. Bargmann CI, Weinberg RA. Ocogenic activation of the neu-encoded receptor protein by point mutation and deletion. *EMBO J* 1988, **7**, 2043–2052.
29. Sainsbury J, Farndon JR, Needham GK, Malcolm AJ, Harris AL. Epidermal-growth-factor receptor status as a predictor of early recurrence of and death from breast cancer. *Lancet* 1987, **i**, 1398–1402.

Acknowledgements—The model-building of human cytochrome P450-IA1 was a collaboration with Dr R. Wolf (ICRF Molecular Pharmacology, Edinburgh) and the modelling of cytochrome P450-17α with Dr C. Laughton and Dr S. Neidle (Institute of Cancer Research, Sutton); the dimerisation of the *neu* receptor was investigated with Dr W. Gullick (ICRF Oncology Unit, Hammersmith Hospital, London). We thank our colleagues at the ICRF, Paul Bates, Suhail Islam, Mansoor Saqi and Peter Walls, for helpful discussions. We also thank Polygen for contributing to the cost of colour printing.