Modelling and Computer
Methods in Molecular
Biology and Genetics

ISBN 1-56072-077-8

FROM THE COMPARATIVE ANALYSIS OF PROTEINS TO SIMILARITY-BASED MODELLING

Mark S. Johnson, J. Overington, A. Sali and T.L. Blundell

Imperial Cancer Research Fund Unit of Structural Molecular Biology
Birkbeck College, Malet Street, London WC1E 7HX, England

The comparative analysis of homologous protein structures has led to a number of general observations that have a bearing on similarity-based protein modelling. First, in structures related through a common ancestor the elements of secondary structure, the α-helices and β-strands, are arranged to yield similar three-dimensional topologies (for review see [1]). Indeed, sequence similarity can all but vanish while the three-dimensional structures remain clearly recognizable! Second, the replacement of residues in the solvent inaccessible hydrophobic "core" is usually accompanied by relatively small shifts in the orientation of the elements of secondary structure [2-3]. A comparison of homologous tertiary structures and their corresponding sequences has also shown that residues of the hydrophobic core are more conserved (both in sequence and position in space) than surface residues that are exposed to solvent [4-5]. The bulk of amino acid replacements, insertions and deletions that do occur are located primarily on the protein surface and are many times the "loops" which connect elements of secondary structure. All told, it is not too surprising that one can use the coordinates of homologous three-dimensional structures to obtain an estimate of the coordinates for the structure of a related protein but with structure unknown.

Browne et al. [6] were the first to describe a model based on the coordinates of a homologous structure: They modelled α-lactalbumin using the structure of lysozyme as a starting point. Other models, mainly for serine proteinases, soon followed: The α-lytic proteinase from elastase [7], the trypsin-like proteinase from *S. griseus* modelled on the coordinates of bovine trypsin [8] and the insulin-like growth factors and relaxins using the coordinates of an insulin structure [9-10]. In general, these models were constructed through the intensive use of interactive graphics devices in which alterations were made to the known structure itself. Indeed, it was not until Greer published his account of modelling serine proteinases [11] that a more systematic approach to modelling was considered.

The approach suggested by Greer [11] is the following:

1. The superposition of pairs of structure coordinates as rigid bodies in order to locate the most similar and therefore the most conserved portions of the structures. These regions are most likely to remain spacially conserved in the unknown structure.
2. The alignment of the unknown sequence with the aligned sequences from the known structures.
3. The use of the existing coordinates from the known structures to build the structurally conserved core and therefore maintain the correct main chain geometry.
4. For the structurally variable loop regions, the use of the corresponding loop positions from homologous structures if possible.
5. The construction of the side chains from the existing side chain geometries in the known structures as much as is possible.

In our attempts to produce both a general and automated approach to protein modelling we have developed two different procedures. The first, COMPOSER [12-14], relies upon the use of fragments of real structures, either homologous or analogous pieces. This I refer to as Frankenstein-monster modelling. A second more elegant approach that is at an early stage of development, concerns itself not with discrete pieces of structure to be glued together in the formation of a finished hybrid model, but instead establishes constraints on the position of each atom in the model and obtains a result that best agrees with these constraints [14].

COMPOSER

COMPOSER is an automated procedure, developed at Birkbeck College, that facilitates the construction of a protein model based on the comparison of known homologous protein structures. This procedure depends upon the analysis of both tertiary and primary structures and the selection of those structures that have the closest inferred similarity to the protein to be modelled, the determination of the conserved protein "core" based on a "family framework," the search and selection of variable "loop" regions and the construction of side chain geometries. COMPOSER can be used to derive an estimate of the relative positions of all nonhydrogen main chain and side chain atoms. The basic steps incorporated into this procedure are:

1. **The identification and selection of homologous protein structures.** Phyletic trees are derived from both the alignment of tertiary structures and the multiple alignment of amino acid sequences. The amino acid sequence of the protein to be modelled, the "unknown", is also included. The resulting two trees are then mapped onto one another and the proteins which bracket the the position of the unknown are used in the construction of the model [15-17].
2. **Identification and construction of the three-dimensional framework representing the unknown.** The selected structures are simultaneously aligned with a procedure that treats the structures as rigid bodies and seeks to provide the best global superposition [12]. Equivalent positions over each of the structures are identified: Those aligned positions from these structures, which lie within a specified distance of each other, are considered part of the structurally conserved core, which most often consists of a set of discontinuous fragments.

3. **Alignment of the unknown with the conserved core fragments.** The sequence of the unknown is aligned with the sequences from the proteins of known structure. This is one of the most critical steps in the entire procedure. For if the alignment is wrong at this stage, then the model will certainly be incorrect when completed. To help insure that an accurate alignment is obtained, we have developed a method for the multiple alignment of proteins based on structural criteria [18,16], the extensive analysis of the influence of structural environments on amino acid substitution patterns (based on a large number of aligned families of proteins of known structure) [19] and the production of a structural profiling alignment procedure which incorporates the results of these analyses [M.S. Johnson, unpublished results].
4. **Building of the main chain for the structurally conserved regions.** The main chain coordinates, for the discontinuous conserved core fragments, are constructed from those corresponding portions of the actual structures having the lowest RMS deviations from the average of the superposed structures; This insures that the chain has proper geometry. These rigid-body fragments are then fitted to the average framework for the family and thereby secure their likely position relative to one another.
5. **Construction of the regions that connect the fragments of the conserved core.** Loops are selected based on a search for substructures that meet endpoint-to-endpoint distance criteria to sequential fragments, as well as features thought to play a key role in a particular loop structure. The selected structural fragments are then annealed to the main chain pieces that comprise the structurally-conserved core using procedures developed by Eisenmenger and Sklenar (unpublished results). This leads to a single contiguous set of main chain coordinates for the model, which extends from the amino-terminus to the carboxyl-terminus.
6. **Side chain coordinates.** Side chain coordinates are built using information obtained from the topologically equivalent side chains across the family of known structures or, when necessary, the most probable conformations are used [13]. A set of 1200 rules describe how much information from the side chains of the known structures can be used in the construction of the side chain in the unknown [13, M.S. Johnson, unpublished results].
7. **Manual Modelling.** Models are then inspected on an interactive graphics display for any atom-atom clashes and any obvious corruptions of the model structure that might inhibit energy refinement.
8. **Energy refinement.** For energy refinement we have made use of the programs that form part of the SYBYL graphics package (TRIPOS Associates).

If we are given several structures with $\geq$ 35-40 percent sequence identity to the protein to be modelled, then based on our experience in constructing models with COMPOSER and making comparisons to the "real" X-ray defined structure, we can make the following general statements with regard to the accuracy of the model:

1. The core α-carbon coordinates that were constructed directly from fragments of homologous structures will have root mean square (RMS) deviations in the range of 0.6 - 0.8 Å.
2. The loop α-carbon coordinates will be more variable and have RMS deviations on the order of 1 Å if the loops can be modelled using homologous structures or the loops are not too large.

3. The model will be close to the average of the contributing structures.
4. For the constructed main chain coordinates, the core will be more certain than short loops ($\leq$ 10 residues), whereas long loops can have large errors and may be worth excluding from the model.
5. For the side chain coordinates, those which lie within the core of the protein will have less uncertainties in their coordinates than those that reside at loop and solvent accessible positions. (It is well known in X-ray defined structures that the positions of some accessible residues may be poorly defined due to side chain mobility). If the residue in the known structure(s) is identical or very similar to that in the unknown protein, then the prediction will be better than for dissimilar residues.

PROTEIN MODELLER

One of the difficulties with an approach such as COMPOSER reveals itself when the sequence similarity between the unknown and the known structures drops below ~35 percent identity. The number of residues that makeup the conserved core falls off sharply with increasing dissimilarity: Although the topology of the compared proteins is still similar, rigid body movements of elements of secondary structure can disguise this fact and lead to few equivalences after superposition. The resulting larger loop regions will in turn be more difficult to predict with reasonable accuracy and the final model will have large errors. In order to escape this dilemma a new structural comparison procedure, COMPARER [18, 16], and a novel approach to modelling, PROTEIN MODELLER [14], were developed.

COMPARER [18, 16] considers features of both sequence and structure in order to obtain the optimal alignment of two or more proteins. Features of individual residues or segments of residues (identity, physical properties, torsion angles, solvent accessibility, direction of chain, etc.) and relationships between residues (hydrogen bonding, van der Walls interactions, etc.) can be taken into account. The result is an alignment of all residues and including gap positions and one which is not degraded by low to insignificant levels of sequence similarity.

The PROTEIN MODELLER takes the structural alignment produced by COMPARER and in turn aligns it with the sequence of the of the protein of unknown structure. From this alignment, features of aligned positions in the structures are extrapolated to the corresponding residue in the unknown. For example, if two atoms are hydrogen bonded in all structures, then we can assume this hydrogen bond will be maintained in the unknown (provided that this residue has the potential to form the required hydrogen bond). This represents a distance constraint on the atoms involved in the hydrogen bond. As one can see there will be a number of constraints placed on every atom in the unknown: The predicted distances are the constraints implied for the atoms due to main chain and side chain dihedral angles, standard bond lengths, cystine bonds, $C\alpha$-$C\alpha$ distances, van der Waals restrictions, etc.

Each of these predicted distances are conveniently expressed as Gaussian probability density functions, all of which can be combined together. The object is to obtain a three-

dimensional model that is in best agreement with these constraints and the most probable structure is the one that maximizes the product of all the individual feature probability density functions. The optimization procedure is performed in Cartesian coordinate space using the variable target function approach of Braun and Go [20] and a combination of conjugate gradients and simulated annealing.

A preliminary model [14] has been constructed for a domain of an aspartic proteinase and based only on Cα-Cα distance constraints from two other aspartic proteinase structures. A comparison of this initial model with the corresponding X-ray determined structure exhibits the potential usefulness of the PROTEIN MODELLER:

1. The overall RMS deviation for a comparison of the model with the known structure (icluding both the core and loops) was 0.8 Å.
2. In contrast to COMPOSER, the model is closer to the crystal structure than to either the average or to the two contributing structures.

CONCLUDING REMARKS

It is safe to say that given a protein to be modelled that has on the order of 35-40 percent sequence identity with one or more proteins of known structure, that a model on the order of a medium resolution X-ray structure can be obtained using building-block methods like COMPOSER. Below ~35 percent sequence identity, other methods such as the PROTEIN MODELLER may be required due to the uncertainties in both the sequence and structural alignments, as well as the relative movements of strands and helices that are likely to occur in the contributing structures and, more importantly, also in the model.

AKNOWLEDGEMENTS

The program COMPOSER was initially produced by Mike Sutcliffe and Tom Blundell. Major modifications have been made by John Overington, Pam Thomas, Frank Eisenmenger and myself. The PROTEIN MODELLER is a concept of Andrej Šali and Tom Blundell with assistance from Dan Donnelly. I would like to thank the American Cancer Society and the Ramsay Memorial Fellowships Trust (University College London) for a postdoctoral fellowship and an honorary fellowship respectively. Current funding is now being provided by the Imperial Cancer Research Fund.

REFERENCES

[1] M. Bajaj and T.L. Blundell, Ann. Rev. Biophys. Bioeng. 13, (1984) 453.
[2] C. Chothia and A.M. Lesk, J. Mol. Biol. 160, (1982) 309.
[3] A.M. Lesk and C. Chothia, J. Mol. Biol. 160, (1982) 325.
[4] A.M. Lesk and C. Chothia, Phil. Trans. Roy. Soc. Lond. A317, (1986) 345.
[5] T.J.P. Hubbard and T.L. Blundell, Prot. Engineer. 1, (1987) 159.
[6] W.J. Browne, A.C.T North, D.C. Phillips, K. Brew, T.C. Vanaman and R.L. Hill, J. Mol. Biol. 42, (1969) 65.

[7] A.D. McLachlan and D.M. Shotton, Nature New Biol. 229, (1971) 202.
[8] L. Jurasek, R.W. Olafson, P. Johnson and L.B. Smillie, In Proteolysis and Physiological Regulation, Vol. 11 (Eds. D.W. Ribbons and K. Brewer), Miami Winter Symposium, Academic Press, N.Y., (1976) p.93.
[9] T.L. Blundell, S. Bedarker, E. Rinderknecht and R.E. Humbel, Proc. Natl. Acad. Sci. U.S.A. 75, (1978) 180.
[10] T.L. Blundell and R.E. Humbel, Nature 287, (1980) 781.
[11] J. Greer, J. Mol. Biol. 153, (1981) 1027.
[12] M.J. Sutcliffe, I. Haneef, D. Carney and T.L. Blundell, Prot. Engineer. 1, (1987) 377.
[13] M.J. Sutcliffe, F.R.F. Hayes and T.L. Blundell, Prot. Engineer. 1, (1987) 385.
[14] A. Šali, J.P. Overington, M.S. Johnson and T.L. Blundell, TIBS 15, (1990) 235.
[15] M.S. Johnson, M.J. Sutcliffe and T.L. Blundell, J. Mol. Evol. 30, (1990) 43.
[16] M.S. Johnson, A. Šali and T.L. Blundell, Methods Enzymol. 183, (1990) 670.
[17] M.S. Johnson, J.P. Overington and A. Šali, In Current Research in Protein Chemistry: Techniques, Structure, and Function, (Ed. J.J. Villifranca), Academic Press, San Diego, (1990) p. 567.
[18] A. Šali and T.L. Blundell, J. Mol. Biol. 212, (1990) 403.
[19] J.P. Overington, M.S. Johnson, A. Šali and T.L. Blundell, (1990) Proc. Roy. Soc. B241, (1990) 132.
[20] W. Braun and N. Go, J. Mol. Biol. 186, (1985) 611.