Current Research in Protein Chemistry: Techniques, Structure, and Function

Published under the Auspices of
the Protein Society

Edited by
Joseph J. Villafranca
Department of Chemistry
Pennsylvania State University
University Park, Pennsylvania

ACADEMIC PRESS, INC.
Harcourt Brace Jovanovich, Publishers
San Diego New York Boston
London Sydney Tokyo Toronto

53

Knowledge-Based Protein Modelling: Human Plasma Kallikrein and Human Neutrophil Defensin

Mark S. Johnson
John P. Overington
Andrej Šali

Laboratory of Molecular Biology
Department of Crystallography
Birkbeck College
University of London
London WC1E 7HX
United Kingdom

Abstract

COMPOSER is an automated procedure, developed at Birkbeck College, that facilitates the construction of a protein model based on a comparison of known homologous protein structures. This procedure depends upon the analysis of both tertiary and primary structures and selection of those structures that have the closest relationship with the protein to be modelled, determination of the conserved protein core based on a "family framework," the search and selection of variable regions, and the construction of sidechain geometries. Here we have applied this methodology to the construction of two different models. First, a model was derived for human plasma kallikrein and based on the comparison of six mammalian serine proteinases. In the second case, a model for human neutrophil defensin NP-3 was obtained using information derived from published solution-NMR distance constraints for rabbit neutrophil defensin NP-5.

I. Introduction

In the early 1960's, Zuckerkandl and Pauling (1965) suggested that amino acid sequence data could be used to chart evolution among homologous proteins. Given a roughly constant rate of change for the same protein within different species, fewer amino acid changes will be observed in

Current Research in Protein Chemistry

comparisons among proteins from more recently divergent organisms than from those species which shared a common ancestor at a more distant time. A similar trend is seen for tertiary structures: It has been shown that the root mean square deviation (RMS) over equivalenced alpha-carbon coordinates from superposed structures increases as the percentage sequence identity becomes smaller (Chothia and Lesk, 1986; Hubbard and Blundell, 1987).

It may, however, come as a surprise that tertiary structures are even more conserved in evolution than their amino acid sequences would suggest (see, for instance, Bajaj and Blundell, 1984). Indeed, comparisons of tertiary structures can be made when the level of the sequence relationship is statistically insignificant (Johnson *et al.*, 1989a,b; Šali and Blundell, 1989). For structures where the sequence similarity is obvious, the observed alterations generally occur at the solvent exposed surface of a protein rather than within the tightly packed hydrophobic interior. However, if the sequence similarity is low among related structures, many amino acid differences are also found within the hydrophobic core of the proteins and the resultant structural alterations manifest themselves in the rigid-body rotations and translations of secondary structural elements (Lesk and Chothia 1982, 1986; Chothia and Lesk, 1982); there is an overriding tendency to conserve the overall fold within a family of homologous proteins. It is this conservation of the basic protein fold that is exploited in the COMPOSER procedure for protein modelling.

Our basic strategy in the construction of a model for the tertiary structure of a protein includes the knowledge base provided by an examination of sequences and tertiary structures for proteins that are homologous with the protein to be modelled. Two cases will serve as examples. In the first we have modelled human plasma kallikrein, a serine proteinase important for its vasoactive role. This model is derived from information provided from the structures of six mammalian serine proteinases. In the second example, a model was constructed for a 30-residue antimicrobial peptide: human neutrophil defensin NP-3.

II. Methods

The automatic modelling program COMPOSER was used to derive the coordinates of all nonhydrogen mainchain and sidechain atoms for the models of kallikrein and defensin and follows the procedures detailed elsewhere (Sutcliffe *et al.*, 1987a,b; Overington *et al.*, 1988; Blundell *et al.*, 1989). The basic steps are briefly described here.

1. **The identification and selection of homologous protein structures.** Phyletic trees are derived for both the tertiary structures and the amino acid sequences (includes the sequence to be modelled), the trees are mapped onto one another, and the structures which bracket the "unknown" are selected for study (Johnson *et al.*, 1989a,b).
2. **Identification and construction of a three-dimensional framework representing the unknown.** Selected structures are simultaneously aligned with a procedure that treats the structures as rigid bodies and seeks to provide the best global superposition (Sutcliffe *et al.*, 1987a). Equivalent positions over each of the structures are identified: Those aligned positions from these structures, which also lie within a specified distance of each other, are considered a 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 templates derived from the fragments that are deemed structurally equivalent. This alignment allows the delineation of the structurally more variable regions, generically called "loops," that connect these fragments of the conserved core.
4. **Building of the mainchain for the structurally conserved regions.** The mainchain coordinates for the discontinuous conserved-core fragments are constructed from those corresponding portions of the actual structures having the lowest RMS deviation from the average of the superposed structures. These regions are then fitted to the average framework for the family.
5. **Construction of the regions that connect portions of the conserved core.** Loops are selected based on a search for substructures that meet distance criteria, as well as features thought to play a key role in a particular loop structure. The selected structural fragments are melded to the mainchain pieces that comprise the structurally conserved core. This leads to a single continuous set of mainchain coordinates for the model, which extend from the amino-terminus to the carboxyl-terminus.
6. **Sidechain coordinates.** Sidechain coordinates are built using information obtained from topologically equivalent sidechains across the family or, when necessary, the most probable conformations are used (Sutcliffe *et al.*, 1987b).
7. **Manual modelling.** Models are inspected on a graphics display for atom-atom clashes and any obvious corruptions of the model structure that might inhibit energy refinement.
8. **Energy refinement.** Energy refinement is conducted using the programs of the SYBYL graphics package (TRIPOS Associates).

Human plasma kallikrein: The sequence of human plasma kallikrein was obtained from the original source (Chung *et al.*, 1986). The position of the amino-terminus of the serine proteinase domain in the mature protein was determined by the clear homology to other known serine proteinase amino-termini. The carboxyl-terminus shows a ten-residue insertion relative to the structurally known proteins and the sequence was truncated to residue 246 (chymotrypsinogen A numbering). The sequences and tertiary structures of the mammalian serine proteinases used in the construction of this model include bovine α-chymotrypsin (4CHA), bovine trypsin (2PTN), porcine glandular kallikrein (2PKA), rat mast cell proteinase II (3RP2), rat tonin (1TON), and porcine elastase (3EST). All coordinates and sequences were obtained from the April 1989 release of the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977).

Human neutrophil defensin: Distance constraints (set C; Pardi *et al.*, 1988) for the solution-NMR structure of rabbit neutrophil defensin NP-5 (Pardi *et al.*, 1988) was used to generate ten structures with the program DISGEO (Havel *et al.*, 1983). These structures were then input to COMPOSER where the average structure was determined in order to give a framework (the average RMS of the individual structures to the framework was about 3 Å). The distances between each structure and the framework at each position were then examined to aid the disection of the molecule into a number of rigid bodies. These fragments were fitted to the framework to assemble the mainchain of the model. The sidechains were built using an analysis of the observed conformations determined from the NMR constraints. The appropriate disulfide linkages, as determined for the closely related human neutrophil defensin NP-2 (NP-3 differs only in having an amino-terminal aspartic acid residue; Selsted and Harwig, 1989), were incorporated into the model using SYBYL prior to energy refinement.

Analyses were conducted on a IRIS 4D/20 graphics workstation under the UNIX operating system (Silicon Graphics), a microVAX II running VAX-VMS (Digital Equipment Corporation) and with an 80386-based personal computer under the XENIX operating system (SCO version 3.2).

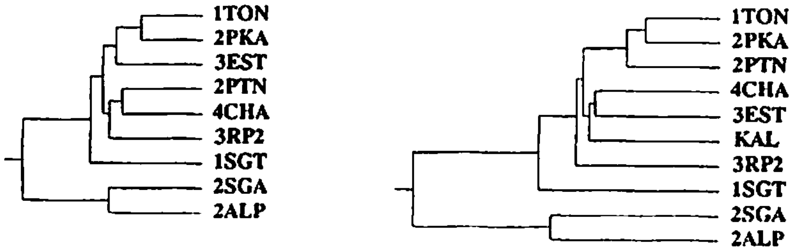

Fig. 1. Phylogenetic trees determined for (left) the serine proteinases structures and (right) the serine proteinase sequences, which also includes the sequence of human plasma kallikrein (KAL). Codes for the proteins are: 1TON (rat tonin); 2PKA (porcine glandular kallikrein A); 2PTN (bovine trypsin); 4CHA (bovine α-chymotrypsin); 3EST (porcine elastase); 3RP2 (rat mast cell proteinase II); 1SGT (*S. griseus* trypsin); 2SGA (*S. griseus* proteinase A); 2ALP (*L. enzymogenes* α-lytic proteinase).

III. Results and Discussion

Human plasma kallikrein: The three-dimensional structures of the serine proteinase family were compared structurally. Besides the six mammalian structures listed above, three microbial structures were also considered: The α-lytic proteinase of *Lysobacter enzymogenes* (2ALP), trypsin from *Streptomyces griseus* (1SGT), and the proteinase A from *Streptomyces griseus* (2SGA). The sequence of plasma kallikrein was then added to the nine serine proteinase sequences to progressively align the whole family according to the procedure of Feng and Doolittle (1987). A comparison of structure-based and sequence-based trees generated from the two data sets (Fig. 1) suggests that all of the mammalian serine proteinases should be included in the construction of a model for plasma kallikrein.

An analysis of the structural data by the three-dimensional comparison program, COMPARER (Šali and Blundell, 1989; Johnson *et al.*, 1989b), leads to the structural alignment displayed in Figure 2. Those positions which were structurally equivalenced by COMPOSER and comprise the structurally conserved core fragments are indicated. One can see from an examination of the conserved and variable regions in Figure 2, Figure 3A, and the conserved-core fragments shown in Figure 3B, that most of the structural variation and all of the insertions/deletions occur between these structurally-conserved core regions.

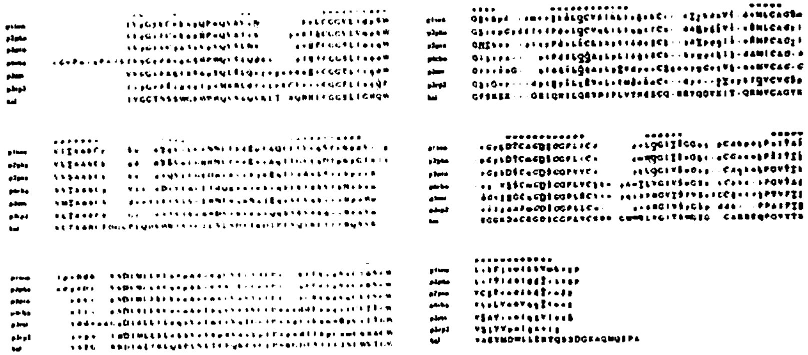

Fig. 2. Structural alignment computed by COMPARER (Šali and Blundell, 1989) for the mammalian serine proteinases used in the modelling of human plasma kallikrein. Structural details were incorporated into the alignment with the program JOY (Overington *et al.*, 1989). See Fig. 1 for identification of the structures. Key to symbols: UPPERCASE, solvent inaccessible, lowercase, solvent accessible (defined as more than 7% relative sidechain accessibility); *italic*, positive phi angle; **bold**, sidechain hydrogen bond to mainchain nitrogen; underline (_), sidechain hydrogen bond to mainchain oxygen; tilde (˜) sidechain-sidechain hydrogen bond. Structural equivalences across all structures as determined by COMPOSER are marked using the equality symbol (=).

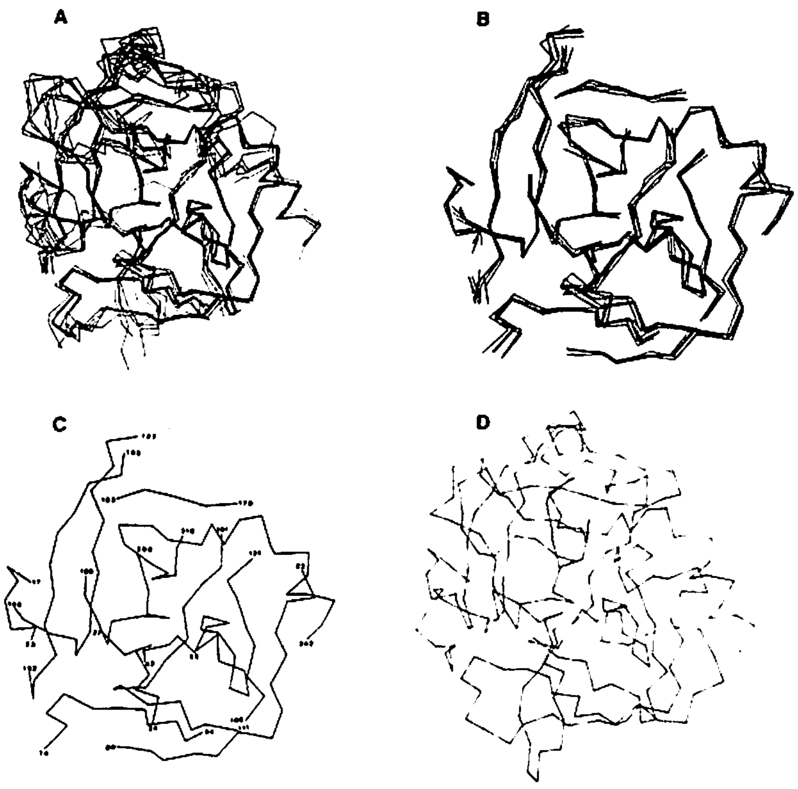

Fig. 3. Alpha-carbon tracing of (A) the entire superposed structures of the serine proteinases used to construct the model, (B) the fragments from the mammalian serine proteinases that form the structurally conserved core regions, (C) the constructed core for plasma kallikrein, and (D) the final model of kallikrein.

The alpha-carbon ribbon for the complete core of the plasma kallikrein model is displayed in Figure 3C. Once the loops had been selected and melded to the core, sidechains were constructed, the model inspected and then subjected to energy refinement. The alpha-carbon trace of the final model is shown in Figure 3D. Although no x-ray structure for plasma kallikrein is yet available, some reasonable estimates on the reliability of the model can be made. The mainchain coordinates for the highly-conserved core regions are likely to have average errors no larger than 0.6 Å, whereas the more variable loop regions, many times less well defined by x-ray analysis due to inherent flexibility along the surface of proteins, may be in error by, on average, 1 Å.

A model for human neutrophil defensin: As a contribution towards the Protein Modelling Workshop, organized by Doug Rees as a part of the third symposium of The Protein Society (Seattle, Washington; July 29-August 2, 1989), a model for human neutrophil defensin NP-3 was predicted. Structural information was available for only one protein in the family and consisted of 2D-NMR distance constraints for rabbit defensin NP-5.

Sequences of nearly a dozen members of the defensin family have been published. In an alignment, which includes both the rabbit and human defensins, no insertions or deletions of residues occur between the half-cystine closest to the amino-termini and the adjacent half-cystines at the carboxyl-termini; the spacing of residues between each of the half-cystines is exactly preserved in each of the aligned defensins (Selsted and Harwig 1989). A loss or gain of residues is only seen at the extreme termini of the sequences.

In the construction of this model, the ten structures calculated by distance geometry for the rabbit defensin were input to COMPOSER. Figure 4A displays the alpha-carbon trace for all of these structures, derived from the NMR constraints and superposed as rigid bodies. In Figures 4B and 4C, the model-built core of human defensin, as well as the the final minimized model, including sidechains, is shown.

The errors in this model are likely to be at least as large as the average RMS observed for the NMR structures input into the program, in this case about 3 Å. It should also be considered that a human defensin x-ray structure will reflect a crystalline environment, whereas this model was built from solution data. This effect will be especially noticeable for sidechain conformations, which in the crystal may be affected by constraining intermolecular contacts.

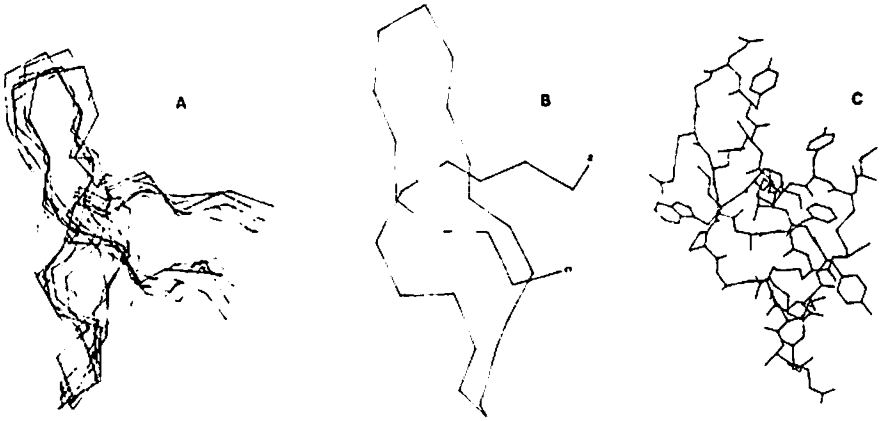

Fig. 4. Alpha-carbon trace for (A) representative structures generated for rabbit defensin NP-5 by DISGEO (fitted over all alpha-carbon atoms) and (B) the alpha-carbon atoms constructed for human defensin NP-3. All atoms (C) of the defensin model after energy minimization.

Acknowledgements

MSJ is a Fellow of the American Cancer Society and a United States Honorary Ramsay Memorial Fellow (University College, London); JPO has a CASE studentship from Pfizer; AŠ acknowledges the support of an ORS Award, Research Council of Slovenia and J. Stefan Institute. We thank Tom L. Blundell for his continuing support and many valuable discussions. The authors would like to acknowledge the sterling efforts of Geri Leahy for such quick and accurate typing.

References

Bajaj, M. and Blundell, T.L. (1984) Ann. Rev. Biophys. Bioeng. 13, 453-492.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Bryce, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) J. Mol. Biol. 112, 535-542.

Blundell, T.L., Johnson, M.S., Overington, J.P. and Šali, A. (1989) Proceedings of the Smith, Kline and Beckmann Symposia, in press.

Chothia, C. and Lesk, A.M. (1982) J. Mol. Biol. 160, 309-323.

Chothia, C. and Lesk, A.M. (1986) EMBO J. 5, 823-826.

Chung, D.W., Fujikawa, K., McMullen, B.A. and Davie, E.W. (1986) Biochemistry, 25, 2410-2417.

Feng, D.-F. and Doolittle, R.F. (1987) J. Mol. Evol. 25, 351-360.

Havel, T.F., Kuntz, I.D. and Crippen, G.M. (1983) Bull. Math. Biol., 45, 665-720.

Hubbard, T.J.P. and Blundell, T.L. (1987) Protein Engineering, 1, 159-171.

Johnson, M.S., Sutcliffe, M.J. and Blundell, T.L. (1989a) J. Mol. Evol., in press.

Johnson, M.S., Šali, A. and Blundell, T.L. (1989b) Methods Enzymol., in press.

Lesk, A.M. and Chothia, C. (1982) J. Mol. Biol., 160, 325-342.

Lesk, A.M. and Chothia, C. (1986) Phil. Trans. Roy. Soc. Lond. A317, 345-356.

Overington, J., Sutcliffe, M., Watson, F., Campbell, S., James, K. and Blundell, T. (1988) Proceedings 8th International Biotechnology Symposium, Paris 1988, 1, 279-304.

Overington, J.P., Johnson, M.S., Šali, A. and Blundell, T.L. (1989), submitted.

Pardi, A., Hare, D.R., Selsted, M.E., Morrison, R.D., Bassolino, D.A. and Bach, A.C. (1988) J. Mol. Biol. 201, 625-636.

Šali, A. and Blundell, T.L., (1989), submitted.

Selsted, M.E. and Harwig, S.S.L. (1989) J. Biol. Chem. 264, 4003-4007.

Sutcliffe, M.J., Haneef, I., Carney, D. and Blundell, T.L. (1987a) Protein Engineering, 1, 377-384.

Sutcliffe, M.J., Hayes, F.R.F. and Blundell, T.L. (1987b) Protein Engineering, 1, 385-392.

Zuckerkandl, E. and Pauling, L. (1965) *In* "Evolving Genes and Proteins" (V. Bryson and H.J. Hogel, eds.), pp. 97-166, Academic Press, New York.