Assessment of Ab Initio Protein Structure Prediction

Arthur M. Lesk
University of Cambridge Clinical School
Cambridge, CB2 2QH, U.K.

Abstract

The assessment of predictions of protein structures from amino acid sequences is related to the question of analysing the differences between similar natural protein structures to study protein evolution and conformational change. However, assessments present certain special features: the alignment is fixed, unlike evolving proteins which are subject to insertions and deletions, and the character of the differences between predicted and target structures is different in kind from the differences between related known proteins.

This report describes the assessment methods used for CASP–2, the recent organized "blind test" of protein structure prediction. This is set in the context of general methods for extraction of similar substructures. It is emphasized that the problem of finding *one* common similar substructure is much simpler than that of guaranteeing the identification of *all* common substructures with similarity better than a specified threshold. Software based on the results presented here will make it possible to extract from two protein structures *all* substructures that have r.m.s. deviations upon optimal superposition no greater than a prespecified value. Applied to a predicted structure and the experimentally-determined coordinates of the target, this will identify any and all three-dimensionally correct features of the prediction.

1 Introduction

The outstanding problem of computational molecular biology is the discovery of an algorithm for predicting the structure and function of the proteins that correspond to the genes emerging from large-scale sequencing projects. We know that such an algorithm exists because experimentally proteins fold spontaneously to unique native three-dimensional structures, both in the living cell and also in isolation *in vitro*. We also know, from studies of protein evolution, that protein structures must be "robust," in the sense that if any mutation would destroy a protein structure, proteins could not evolve by mutation, and could not have achieved their current well-adapted states. Therefore small changes in sequence should in most cases leave the structure intact. This implies that the algorithm for determining structure from sequence should be reasonably well-behaved.

What is the current state of the art of protein structure prediction, and how is it determined? Blind tests are essential: Judging predictions requires experimental structures; but knowing the structure in advance would leave any apparent success suspect. Organized blind tests were initiated by John Moult; the most recent, "Critical Assessment of Structure Prediction–2 (CASP–2)," took place in 1996. Here is the challenge: the amino acid sequences of several proteins, the experimental structure determinations of which are in progress, are made public, and predictors are invited to submit their models before a deadline. After the entries are closed the crystallographic results are revealed ... to the delight of a few and the chagrin of most!

In CASP–2 submissions were invited in three categories: (1) Homology modelling, (2) Fold recognition, and (3) *ab initio* prediction:

Homology modelling

Homology modelling is based on the observation that small changes in sequence often produce relatively small changes in protein structure, so that if a target protein of unknown structure is related to a protein with similar sequence and known structure, the related protein can serve as a basis for predicting the target (See Figure 1.) This can be considered the "differential form" of the protein folding problem: Though the "integral form" – the direct prediction of structure from sequence – is very difficult, the "differential form" – the prediction of *change* in structure from *change* in sequence – is more

RECOMB 98 New York NY USA
Copyright 1998 0-89791-976-9/98/3..$5.00

tractable. The closer the relationship, the better the prediction. The growth of databases will increase the likelihood of successful homology modelling, by making it more likely that a target protein will be closely related to a known structure.

Fold recognition

The goal of fold recognition is to identify the general folding pattern of a protein as similar to one or more known structures in a library of protein folds. The assumption is that the sequence of the target protein is so different from the sequences of all the library structures, that homology modelling methods are not feasible. The methods involve trying different alignments, seeking a good sequence-structure compatibility. There are several approaches to measuring this compatibility.

Ab initio prediction

This category includes submissions of all types in which no *explicit* reference is made to known protein structures. (To be sure the accumulated knowlege of two generations of study of the corpus of solved structures is applied, but no explicit detailed information from any particular known structure.) My work deals with the assessment of predictions in the *ab initio* category. In my role as assessor for CASP–2, I dealt with four types of predictions, that produced: (1) full or partial sets of three-dimensional coordinates, (2) assignments of secondary structure – helices and strands, (3) sets of contacts between residues, and (4) sets of contacts between helices and strands [1].

Here I will discuss only the predictions of three-dimensional coordinates. It might be thought that success and failure would be obvious, but this is not so, because the current state of the art achieves at best only partial success. (Utter failure *is* obvious.) As the completeness and quality of predictions improves, it will be possible to apply methods that have been developed to analyse the structural differences between related proteins. However, given the current state of the art, the challenge is to find the few specks of gold in the murky water – or, less figuratively, to detect small correct features in a background of error.

2 Detection of structural similarity

A general and fundamental problem relevant to assessing predictions of three-dimensional structure with partial success is the following: Given two or more protein structures, find the common similar substructures. We shall discuss the problems associated with the formulation of this problem.

It may be useful to contrast three related problems that arise in computational chemistry (see [2]):
(1) Similarity of two sets of atoms with known correspondences: $p_i \longleftrightarrow q_i, i = 1,\ldots N$. The analog, for sequences, is the Hamming distance: mismatches only.
(2) Similarity of two sets of atoms with unknown correspondences, but for which the molecular structure – specifically the linear order of the residues – restricts the range of the correspondence. In the case of proteins we are restricted to correspondences which retain the order along the chain:

$$p_{i(k)} \longleftrightarrow q_{j(k)}, k = 1,\ldots K \leq N, M$$

with the constraint that: $k_1 > k_2 \Rightarrow i(k_1) > i(k_2)$ and $j(k_1) > j(k_2)$. This can be thought of as corresponding to the Levenshtein distance, or to sequence alignment with gaps.
(3) Similarities between two sets of atoms with unknown correspondence, with no restrictions on the correspondence:

$$p_{i(k)} \longleftrightarrow q_{j(k)}$$

This problem arises in the following important case: Suppose two (or more) molecules have similar biological effects, such as a common pharmacological activity. It is often the case that the structures share a common constellation of a relatively small subset of their atoms that are responsible for the biological activity, called a *pharmacophore*. The problem is to identify them: to do so it is useful to be able to find maximal subsets of atoms from the two molecules that have a similar structure.

The most general approach to these problems is based on a purely geometric closed solution of case (1), the case of fixed correspondence $p_i \longleftrightarrow q_i$. Even if two structures are exactly congruent, the atoms may not be at the same positions in space because the coordinate systems in which they are expressed are independent. The most general motion of a rigid body can be expressed as a rotation plus a translation. Therefore a measure of similarity of two ordered sets of points is the root-mean-square deviation Δ_2 after optimal superposition:

$$\Delta_2^2 = \min_{\mathbf{R},t}\{\sum_{n=1}^{N} \| \mathbf{R}p_i + \mathbf{t} - q_i \|\}$$

where $\mathbf{R}$ is a proper rotation matrix ($\det \mathbf{R} = 1$) and $\mathbf{t}$ is a translation vector. (Later we shall consider alternative measures of structural similarity).

The problem of analysing the similarities in related proteins presents case (2). Solutions of this problem have several applications to the study of natural proteins, including: analysis of the pathways of evolution in protein families, analysis of the mechanism of conformational change, and the classification of protein folding patterns. Several approaches have been developed:
(A) By determining the angles of internal rotation defining the conformation of each residue, the main chain conformation of a protein may be written as a *one-dimensional* sequence of residue conformations. The

algorithms for determining minimal distances and optimal alignments of strings may be applied directly to these sequences of conformations [3-4].
(B) Each residue may be characterized by the distances between a point within the residue to corresponding points in other residues nearby in the sequence. This is another way to reduce the conformation to a one-dimensional sequence of objects. In this case the objects are sets of distances. Nevertheless, it is possible to define a metric on the objects and to apply the dynamic programming algorithms [5-7]. 1985; Taylor & Orengo, 1989).
(C) A common class of methods is, in some senses, the three-dimensional analog of parts of sequence-comparison algorithms that search for matching k-tuples (see, for instance, [8-9]). It is possible to identify all well-fitting subsets of k consecutive residues from the two proteins (typically $k = 6 - 8$), and combine them to determine maximal well-fitting substructures.
(D) A fourth class of methods is based on hash-coding structural features, using techniques from computer vision [10-11].

The case of assessment of structure prediction is however less general than the general problem of finding maximal common substructures of proteins, because the alignment is fixed, so that we have case (1) rather than case (2).

At present the problem of finding maximal common substructures is not well formulated because substructures of different sizes will fit to different accuracies. Figure 2 shows the results of calculating the r.m.s. deviation of best N-atom substructures of two known proteins as a function of N. It is not clear either how to choose a single point on this curve to characterize the pair of structures, or how to derive a single parameter for the entire curve. In general, there can even be multiple equal optima: more than one N-atom substructure with the same r.m.s. deviation.

3 Assessment methods for *ab initio* predictions at CASP–2

The numerous methods for comparing protein structures when the overall folding pattern is largely correct are appropriate for assessing entries in the homology modelling and fold recognition categories. In the *ab initio* category, only for target 42 (NK-lysin) were there predictions of such a quality. Other *ab initio* submissions were not successful in a global sense, and were more difficult to assess. In some cases the submissions were fragmentary. In others they were not compact – containing regions of secondary structure with flexible linkers; in effect a prediction of secondary structure expressed in the format of a three-dimensional coordinate set. As assessor, I sought a procedure that would identify well-fitting fragments, limiting consideration to methods that are perspicuous rather than subtle [1].

I calculated local r.m.s. deviations in a running window for several values of the window length, and plotted these together on a single graph. Figure 3 shows such a graph for target 42, prediction AB979-1 by D. Jones. In interpreting this graph, it is useful to keep in mind the experimental secondary structure of the target: helices extend from residues 3–17, 23–36, 40–51, 57–62, and 66–73. These are shown as horizontal lines at the top of the graph; the disulfide bridges are indicated also.

There are four curves in the chart: the lowest, in a solid line, is the r.m.s. deviation in a running window of width 6; above it are results for windows of widths 15, 25 and 40. Consider first the top graph, shown in a dotted line, which corresponds to a window of 40 residues. The lowest point, at the left end of the curve, shows that the r.m.s. deviation of residues 1–40 is 3.5 Å. The highest point, near the right end, shows that the r.m.s. deviation of the C-terminal 40 residues is above 6 Å. The overall structure of the N-terminal part of the molecule is better predicted than the C-terminal part.

The curve below the top, corresponding to a window of 25 residues, shows that for several 25-residue regions around residue 20, the r.m.s. deviations are below 2.0 Å. These correspond to the helix hairpin 3–17 / 23–36 which is correctly predicted in this submission (see Figure 4). The r.m.s. deviation of the Cα atoms of residues 5–16 and 23–35 is 2.0 Å. The minimum in this curve at residue 42 probably reflects the disulfide bridge between residues 35 and 45. The curves remain high for the C-terminus of the molecule. The third curve down from the top, for a window of 15 residues, is roughly parallel to the preceding one.

The lowest curve, with a window of 6 residues, is sensitive to very local structure. It is in effect a measure of correct prediction of secondary structure and turn geometry. We can see that the local structure of the C-terminus from residues 7–30 is quite good, and that the region from residue 43–50 is also good. However it is clear that the region around residue 40 is quite bad, and reflects the fact that the turn between helices 23–36 and 40–51 is poorly predicted. The C-terminus of the molecule is not predicted correctly *even* at a local level.

To summarize the information contained in these curves:

- The curve with window 6 reveals the quality of the prediction of the local structure – regions of helix and strand, and turns.
- The curve with window 15 reveals the quality of the prediction of long elements of secondary struc-

ture – usually helices – or of short elements of supersecondary structure: β–hairpins or pairs of successive helices.

- The curve with window 25 reveals the quality of the prediction of elements of supersecondary structure.
- The curve with window 40 reveals the quality of the prediction of global aspects of the fold, albeit below the domain level.

It is obvious that the method is not limited to this choice of numbers, which are arbitrary but not unreasonable. A "three-dimensional" graph of r.m.s. deviation as a function of initial residue and window length would give a complete picture of the situation (Figure 5.)

To see how these graphs appear for a prediction with isolated local success, consider prediction AB405 of *E. coli* L-fucose isomerase, target 22 (Fig. 6). It is clear that the success of the prediction is limited to a region around residue 250, and although other regions of the chain may have reasonable local structure (see lowest curve in Fig. 6, and Fig, 7), the overall structure of any 40–residue region remains above 9 Å except for the region around residue 250 (where there is a region of about 100 residues for which the r.m.s. deviation in the 40–residue window remains well below 10 Å), and minor dips at residues 34, 160, 379 and 428.

These graphs give a reasonably clear picture of the local accuracy of the predicted structures. But is it possible to derive from these data a single parameter to present as a quantitative assessment of the predictions?

Consider first the minimum value for each window size. The value of the minimum indicates the *best* local regions at each level of the structural hierarchy. It answers the question – are there *any* regions that fit well? – for each window size. Because the goal is a prediction of the global structure it is suggested that the minimum in the 40-residue window be taken as an index of quality. (Other uses of these numbers would reflect other qualities of the predicted structures. Looking at the *distribution* of numbers in the 16- or 25-residue column would rank entries according to correct prediction of *local* structure; ultimately (for window lengths < 10) this overlaps with checking secondary structure prediction.)

The minimum value for the 6-residue window is not very informative, because given the shortness of the segment, any correct element of secondary structure will drive this number very low. For the 6-residue window it is better to report the *mean* value which in some sense represents an average prediction of secondary structure.

4 Deficiencies of the methods used at CASP–2

The methods for assessment of *ab initio* prediction used at CASP–2 were designed under two special conditions: (1) Knowing, by careful inspection of the structures on computer displays, the general nature of relationship between submissions and correct structures, and (2) Under the constraint that the method must be simple and obvious to ensure that people recognized that they were being treated fairly. Both these features were pragmatic aspects of the situation.

What is really wanted is the determination of *all* correct structural features, by a procedure that is not *ad hoc* and which can be proven to work in all circumstances. The development of such a method is the subject of the remainder of this report.

5 Extraction of common substructures

Consider the problem of extraction of common substructures in the case of known alignment. The problems addressed are the relationship between different measures of structural similarity, and the development of algorithms that can ensure that *all* substructures with a specified degree of similarity or better can be extracted efficiently. It must be emphasized that the task of finding *one* common similar substructure is much simpler than that of guaranteeing the identification of *all* common substructures with similarity smaller than a specified threshold, for some measure of similarity.

One question that arises in structure comparison is what measure of similarity to use. Many authors have used the root-mean-square deviation upon optimal superposition as an index of similarity, defined above. Algorithms and software exist for the efficient calculation of the required least-squares superposition in the case of known alignment (see [2]). and considerable experience in interpreting the values has been developed. An alternative measure of similarity is the maximum element of the difference distance matrix. Again let p_i and q_i be sets of n points in three-dimensional space, with correspondences $p_i \longleftrightarrow q_i, i = 1, \ldots N$. The difference matrices D_p and D_q are defined as $D_p(i,j) = |p_i - p_j|$ and $D_q(i,j) = |q_i - q_j|$, and the difference distance matrix is $\Delta D_{pq}(i,j) = |D_p(i,j) - D_q(i,j)|$. The measure of similarity used in Nichols *et al.* [12] and Lesk [13] (is $\max_{ij} \Delta D_{pq}(i,j)$. Algorithms based on distance matrices are somewhat more powerful for solving the combinatorial problems that arise in addressing the latter problem [14].

This measure has provided the basis for algorithms and software for extracting common subsets of similar structure according to this parameter. Nichols *et al.* and Lesk have developed software with the following facilities:

(1) Find the largest subsets of corresponding points from the two sets such that for all points in the subsets $|D_{ij}(p) - D_{ij}(q)| \leq$ a prespecified value δ.
(2) Find *all* subsets of corresponding points from the two sets such that for all points in the subsets $|D_{ij}(p) - D_{ij}(q)| \leq$ a prespecified value δ.

Note that there are in general many small subsets with this property. Therefore it is useful to be able to specify:
(3) Find *all* subsets *containing at least N corresponding points* from the two sets such that for all points in the subsets $|D_{ij}(p) - D_{ij}(q)| \leq$ a prespecified value δ.

6 Relationships among different measures of similarity

Consider optimal superposition in other norms than L_2 (which corresponds to the optimal r.m.s. deviation) including, in particular, optimal fitting in the L_∞, or Chebyshev, norm. The results of Chebyshev fitting can be used to relate the r.m.s. deviation and the maximum element of the difference distance matrix [15, 16].

The optimization problem for Chebyshev superposition is:

$$\text{Find } \Delta_\infty = \min_{\mathbf{R},\mathbf{t}}\{\max|\mathbf{R}p_i + \mathbf{t} - q_i|\}$$

where again $\mathbf{R}$ and $\mathbf{t}$ specify a proper rotation and translation. In words, superposition according to the Chebyshev norm is finding a relative position and orientation in which the *maximum* difference between corresponding points is a *minimum*.

It should be noted that for fitting in different norms L_n, the higher the value of n the more sensitive the result will be to outliers. Indeed, in common practice using programs that calculate the root-mean-square deviation (equivalent to fitting in the L_2 norm), superpositions are "tuned" by removing outliers from the list of aligned atoms, either by hand or automatically. Of all norms, the Chebyshev norm is the *most* sensitive to outliers, and in most cases it would not be the method of choice for superposing molecules. However, there is one situation in which Chebyshev superposition may provide useful information. In a discussion of the structural consequences of mutations in T4 lysozyme [17], Chothia and Gerstein 18] pointed out that the claim that the structural changes were small, based on calculations of the root-mean-square deviation, was unjustified because of relatively large local deformations to which the r.m.s. deviation was inadequately sensitive. It was a case of the noise getting lost in the signal! In principle, optimal superposition in the Chebyshev sense could have avoided this.

The properties of Chebyshev superposition imply two inequalities relating the root-mean-square deviation and the maximal element of the difference distance matrix: First, given a threshold $\epsilon > 0$, we can compute a number $\delta > 0$ that depends on ϵ and on the coordinates of one of the point sets, such that if the root-mean-square deviation is $\leq \epsilon$, all elements of the difference distance matrix have absolute values $\leq \delta$. Conversely, we can compute a (different) number δ that depends on ϵ and on the elements of the distance matrices of the two point sets, such that if all elements of the difference distance matrix have absolute values $\leq \delta$, then the root-mean-square deviation is $\leq \epsilon$ [15,16].
This relation between the r.m.s. deviation and the maximal element of the difference distance matrix means that we can use the efficient algorithms based on difference distance matrices to extract *all* common substructures with r.m.s. deviation < some prespecified value.

7 Conclusions

The results presented here clarify the relationships between the different measures of structural similarity currently in use, and can be applied to the development of algorithms and software for determining geometrically similar subsets of two point sets. Such software will have application to the assessment of predictions of three-dimensional structures of proteins, as it is guaranteed to find all possible substructures (optionally: of at least a specified size) that fit to within a specified r.m.s. deviation. It will also be useful in analyzing conformational changes in known protein structures.

ACKNOWLEDGEMENT

I thank The Wellcome Trust for generous support,

REFERENCES

1. Lesk, A.M. "CASP–2: Report on *ab* initio predictions." *Proteins: Structure, Function and Genetics,* in press.

2. Lesk, A.M. "Computational Molecular Biology." In: *Encyclopedia of Computer Science and Technology,* A. Kent and J.G. Williams, (eds.) New York, Marcel Dekker, Inc. Volume 31, pp. 101–165, 1994.

3. Levine, M., Stuart, D., Williams, J. "A method for systematic comparison of the three-dimensional structures of proteins and some results." *Acta crystallographica* Vol. A40, pp. 600–610, 1984.

4. Karpen, M.E., de Haseth, P.L. Neet, K.E. "Comparing short protein substructures by a method based on backbone torsion angles." *Proteins: Structure, Function and Genetics,* Vol. 6, pp. 155–167, 1989.

5. Sippl, M.J. "On the problem of comparing protein structures. Development and applications of a new method for the assessment of structural similarities of polypeptide conformations." *J. Mol. Biol.,* Vol. 156, pp. 359-388, 1982.

6. Liebman, M. N., Venanzi, C.A., Weinstein, H. "Structural analysis of carboxypeptidase A and its complexes with inhibitors as a basis for modelling enzyme recognition and specificity." *Biopolymers,* Vol. 24, pp. 1721–1758, 1985.

7. Taylor, W., Orengo, C. "Protein structure alignment." *J. Mol. Biol.,* Vol. 208, 1–22, 1989.

8. Alexandrov, N.N., Takahashi, K., Gō, N. Common spatial arrangement of backbone fragments in homologous and non-homologous proteins. *J. Mol. Biol.,* Vol. 225, pp. 5–9, 1992.

9. Vriend, G., Sander, C. Detection of common three-dimensional substructures in proteins. *Proteins: Structure, Function and Genetics,* Vol. 11, pp. 52–58, 1991.

10. Fischer, D., Bachar, O., Nussinov, R., Wolfson H.J. "An efficient automated computer vision based technique for detection of three-dimensional structural motifs in proteins." *J. Biomol. Str. Dyn.,* Vol. 9, pp. 769–789, 1992.

11. Bachar, O., Fischer, D., Nussinov, R., Wolfson, H.J. "A computer vision based technique for 3-D sequence independent structural comparison of proteins." *Prot. Eng.,* Vol. 6, pp. 279–288, 1993.

12. Nichols, W.L, Rose, G.D., Ten Eyck, L.F., Zimm, B.H. "Rigid domains in proteins: an algorithmic approach to their identification." *Proteins: Structure, Function and Genetics,* Vol. 23, pp. 38–48, 1995.

13. Lesk, A.M. "Three-dimensional pattern matching in protein structure analysis In: *Combinatorial Pattern Matching,* Z. Galil & E. Ukkonen, eds. Lecture Notes in Computer Science 937. Springer-Verlag, Berlin. pp. 248–260, 1995.

14. Holm, L., Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.,* Vol.233, pp 123–138, 1993.

15. Lesk, A.M. "Extraction of well-fitting substructures: r.m.s. deviation and the difference distance matrix." *Folding and Design,* Vol. 2, pp. S12–S14., 1997.

16. Lesk, A.M. "Extraction of geometrically similar substructures: Least-squares and Chebyshev fitting, and the difference distance matrix." *Proteins: Structure, Function and Genetics,* in press.

17. Gassner, N.C., Baase, W.A., Matthews, B.W. "A test of the "jigsaw puzzle" model for protein folding by multiple methionine substitutions within the core of T4 lysozyme." *Proc. Nat. Acad. Sci.,* Vol. 93, pp. 12155–12158, 1996.

18. Chothia, C., Gerstein, M. "How far can sequences diverge?" *Nature,* Vol. 385, 579–581, 1997.

19. Lesk, A.M., Chothia,C. "The response of protein structures to amino acid sequence changes." *Phil. Trans. Roy. Soc. London,* Vol. 317, pp. 345–356, 1986.

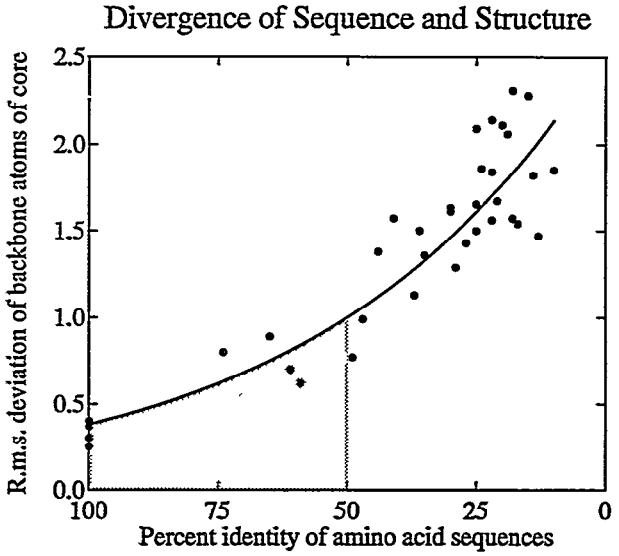

Figure 1. Relation between the divergence of sequence and divergence of structure in homologous proteins. The shaded area defines roughly the region for which homology modelling can produce results of useful quality (see [19]).

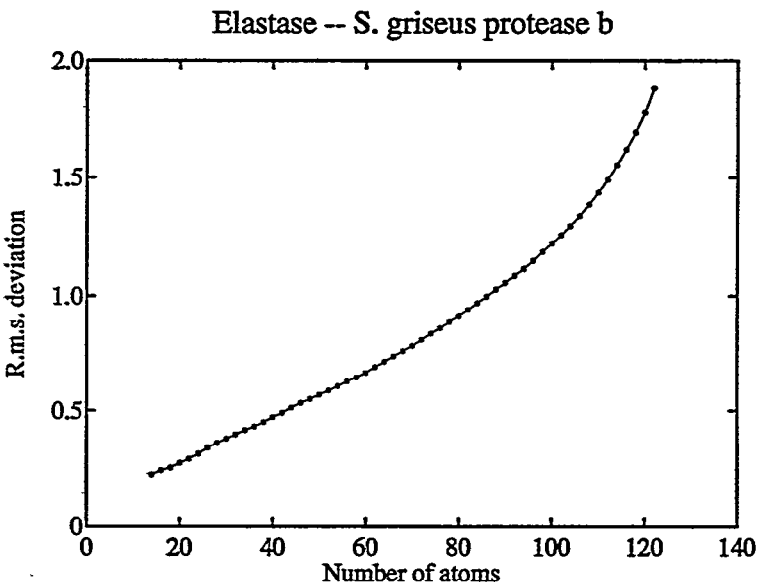

Figure 2. The r.m.s. deviation of the best-fitting N-atom substructure of two proteins depends on N. As a result the problem of finding the maximal common substructure is not well posed.

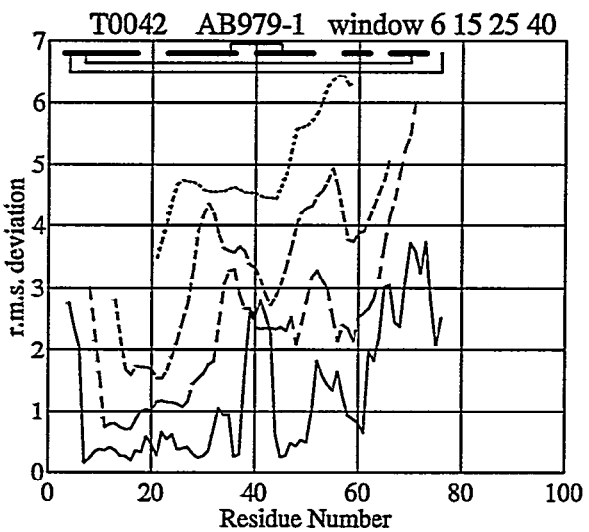

Figure 3. R.m.s deviations of Cα atoms in running windows of lengths 6, 15, 25 and 40 for prediction AB979-1 of NK-lysin by D. Jones. Horizontal lines indicate helices in the observed structure; links indicate disulfide bridges.

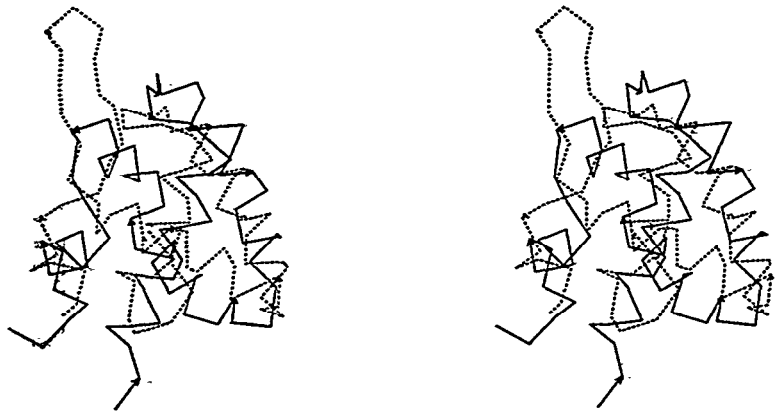

Figure 4. Superposition of Cα traces of experimental structure of NK-lysin (solid lines) and prediction AB979-1 by D. Jones (broken lines).

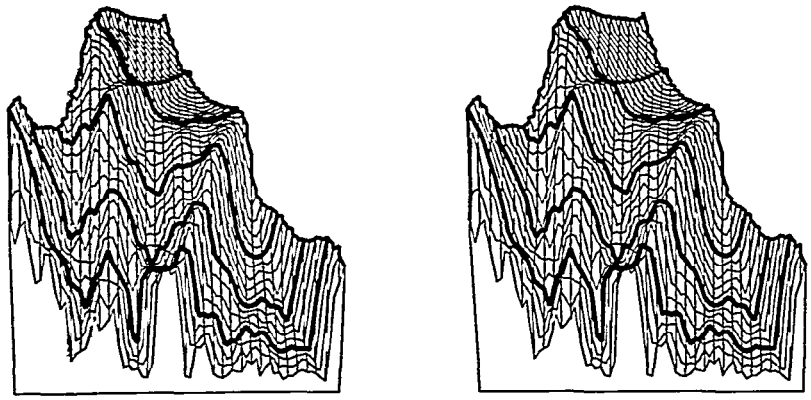

Figure 5. Three-dimensional graph of r.m.s. deviation as a function of window length and position for prediction AB979-1 of D. Jones (cf. Fig. 3).

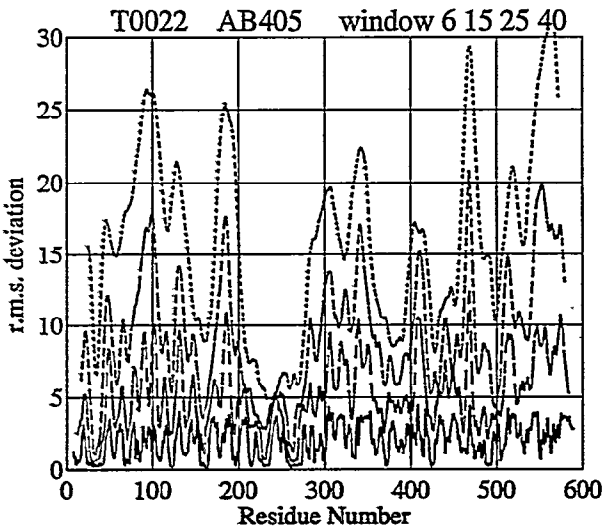

Figure 6. R.m.s deviations of Cα atoms in running windows of lengths 6, 15, 25 and 40 for prediction AB405 by C. Bystroff and D. Baker of *E. coli* L-Fucose Isomerase.

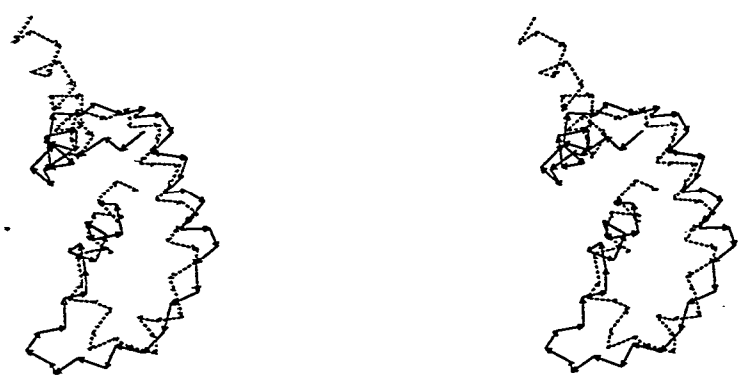

Figure 7. Superposition of Cα traces of experimental structure of *E. coli* (solid lines) and prediction AB405 by C. Bystroff and D. Baker (broken lines).