Protein Engineering vol.1 no.1 pp.77–78, 1986

PROTOCOL

Alignment of the amino acid sequences of distantly related proteins using variable gap penalties

Arthur M.Lesk[1,2], Michael Levitt[3] and Cyrus Chothia[1,4]

[1]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, [3]Department of Chemical Physics, The Weizmann Institute of Science, 76100 Rehovot, Israel, and [4]Christopher Ingold Laboratory, University College London, 20 Gordon Street, London WC1H 0AJ, UK

[2]Permanent address: Fairleigh Dickinson University, Teaneck-Hackensack Campus, Teaneck, NJ 07666, USA

An essential step in assessing and analysing the relationship between two or more proteins is the establishment of an alignment of their amino acid sequences. The prediction of an unknown protein structure by model building from the known structure of a homologue also requires the correct correspondence between the residues. Using sequence information alone, the optimal alignment determined by the procedure of Needleman and Wunsch (1970) or equivalent modifications, with scoring parameters expressed in the mutation data matrix MD (Orcutt *et al.*, 1984), gives correct results for closely related sequences. For distantly related proteins, however, this procedure gives results inconsistent with the alignments derived from superpositions of the 3-D structures; this has been observed in the globins (Lesk and Chothia, 1980), the cytochromes *c* (Dickerson, 1980), the serine proteases (Delbaere *et al.*, 1979; Greer, 1981; Read *et al.*, 1984) and plastocyanin and azurin (Chothia and Lesk, 1982).

Consider the alignment of the α-strand of human haemoglobin A with lupin leghaemoglobin. Figure 1 compares the optimal alignment produced by the implementation of the Needleman–Wunsch–Sellars algorithm in the PIR program ALIGN, using a matrix bias of 2, as recommended by Orcutt *et al.* (1984), and a gap penalty of 10, with the alignment based on the 3-D structures of these molecules produced by superpositions to determine which residues had similar relative spatial dispositions in the two structures (Lesk and Chothia, 1980). Figure 1 also indicates the positions of the helices in the α-strand of human haemoglobin A. It is clear that the algorithm has not hesitated to put insertions within helical regions.

Because of the importance of the stability of the structures of the packing of helix–helix interfaces, insertions and deletions are not observed to occur in the interiors of helical regions of proteins. They would entirely change the pattern of residue–residue packing, and destroy the complementarity of the occluding surfaces. (Insertions and deletions can occur at the ends of helical regions.) It is possible to apply this insight into the mechanism of protein evolution to the alignment of distantly related sequences by a modification of the Needleman–Wunsch procedure. This method depends on knowing the 3-D structure of one member of a protein family. Instead of applying a gap penalty uniform throughout the sequence, we increase the gap penalty within helical regions. Because regions of secondary structure in related proteins do not always retain the same boundaries, we adopt the following relative gap penalties:

Structural position of residue	Relative gap penalty
Not in helix or strand of sheet	1
Last two residues of helix or strand of sheet	2
Interior of helix or strand of sheet	8

The alignment of human haemoglobin (α-chain) with lupin leghaemoglobin produced using variable gap penalties is also shown in Figure 1. The number of errors, relative to the alignment based on the 3-D structure, is reduced from 78 to 10.

This method is applicable to the alignment of a family of protein sequences whenever at least one structure is known. The serine protease inhibitors (serpins) are such a case: the structure of α_1-antitrypsin has been determined by Loebermann *et al.* (1984). Ten other sequences are known. In collaboration with Carrell and Boswell (1986), we have aligned the sequences of this family, using a secondary structure of α_1-antitrypsin to choose secondary structure gap weights. The method produced an alignment that needed only minor editing to be satisfactory. Unlike control runs with uniform gap penalties, the non-uniform gap weights aligned certain residues known to be homologous on functional grounds. It would have been possible to force the alignment of these residues without using non-uniform gap penalties, either by separate alignment of the regions between them, or by defining new amino acid types as 'calibration marks' and choosing scoring parameters that ensure that these special amino acids were aligned. However, the procedure utilized achieved this result more conveniently.

The globins have very different sequences but are similar in the extent and distribution of their regions of secondary structures; it is likely that this is true of the serine protease inhibitors also. These structural characteristics contribute to the success of the alignment method. Other families, such as plastocyanins and azurins, the variable and constant immunoglobulin domains, and the bacterial and mitochondrial cytochromes *c*, each contain members that differ appreciably in the number of α-helices or strands of β-sheet. In these three cases the alignments produced with gap penalties assigned from secondary structures are not inferior to that produced with constant gap penalties, but the improvement is minimal. The method proposed, in its current form, produces no significant improvement in alignment in cases where there are large, extensive deletions involving significant elements of secondary structure.

In summary, the analysis of relationships among distantly related proteins depends on the availability of structural information. The use of variable gap weights can be a useful way to bring structural information to bear on the problem of alignment of distantly related sequences.

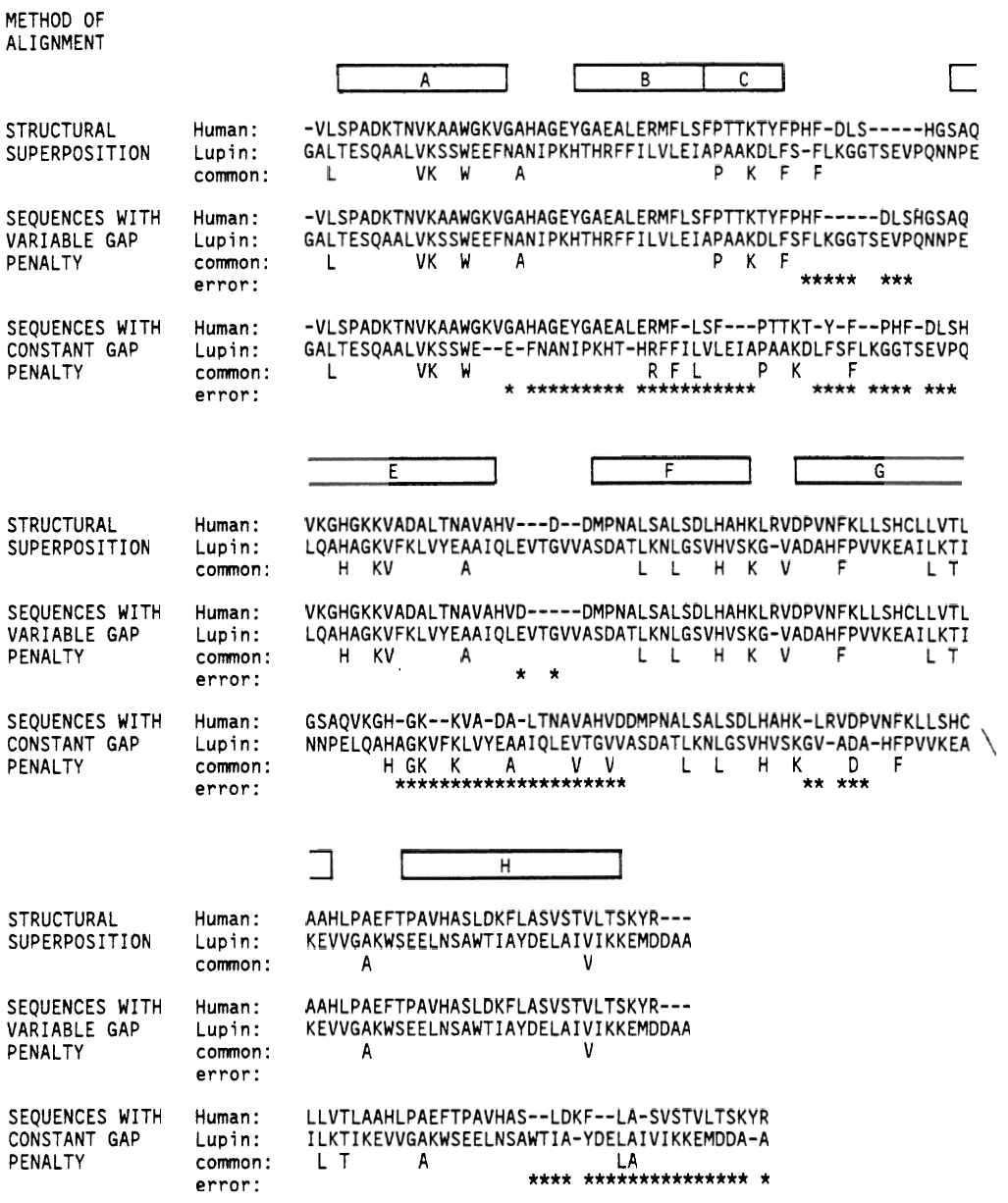

Fig. 1. Alignments of the sequences of human haemoglobin (α-chain), and lupin leghaemoglobin, based on **(a)** superpositions of the crystal structures; **(b)** sequence information alone, using constant gap penalties; and **(c)** variable gap penalties chosen tc inhibit insertions or deletions within a region of secondary structure.

Acknowledgements

We thank the Royal Society, the US National Science Foundation (PCM83-20171), the National Institute of General Medical Sciences (GM25435) and the European Molecular Biology Organization for support.

References

Carrell,R.W. and Boswell,D.R. (1986) In Barrett,A. and Salvesen,G. (eds), *Proteinase Inhibitors*. Elsevier, Amsterdam, in press.

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

Delbaere,L.T.J., Brayer,G.D. and James,M.N.G. (1979) *Nature*, **279**, 165–168.

Dickerson,R.E. (1980) In Sigman,D.S. and Brazier,M.A.B. (eds), *Evolution of Protein Structure and Function. UCLA Forum Med. Sci.*, Academic Press, New York, Vol. 21, pp. 173–202.

Greer,J. (1981) *J. Mol. Biol.*, **153**, 1027–1042.

Lesk,A.M. and Chothia,C. (1980) *J. Mol. Biol.*, **136**, 225–270.

Loebermann,H., Tokuoka,R., Deisenhofer,J. and Huber,R. (1984) *J. Mol. Biol.*, **177**, 531–556.

Needleman,S.B. and Wunsch,C.B. (1970) *J. Mol. Biol.*, **48**, 443–453.

Orcutt,B.C., Dayhoff,M.O., George,D.A. and Barker,W.C. (1984) *User's Guide for the Alignment Score Program of the Protein Identification Resource (PIR)*. National Biomedical Research Foundation, Washington, DC, PIR Report ALI-1284.

Read,R.P., Brayer,G.D., Jurašek,L. and James,M.N.G. (1984) *Biochemistry*, **23**, 6570–6575.

Received on 5 June 1986