Application of Sequence Alignment Methods to Multiple Structural Alignment and Superposition[1]

Arthur M. Lesk

Department of Haematology
University of Cambridge Clinical School
MRC Centre
Hills Road
Cambridge CB2 2QH
United Kingdom

e-mail: aml2@mrc-lmb.cam.ac.uk

Abstract. With the goal of developing efficient multiple structural alignment methods, we have asked which of the pairwise structure alignment methods lends itself most readily to generalization to multiple structure alignment. A simple linear encoding of the sequence and associated residue conformation can be treated by standard multiple *sequence* alignment methods.

Key words: Protein structure, multiple alignment

1 Introduction

One often wishes to analyse proteins that have similar folding patterns but too little sequence similarity to permit the alignment of their residues by sequence-based methods. Such proteins may be very distant relatives, or independently-evolved examples of the same folding pattern. For only two structures, it is possible to perform a structural alignment; that is, to identify residues that occupy similar spatial positions within the structure [GL98]. However, just as multiple sequence alignments are far more informative than pairs of aligned sequences, so the analysis of protein structures requires alignment of more than two sequences.

Most previous approaches to multiple structure alignment have been based on pairwise structural alignments. The simplest approach is to choose a master structure and align all the others to it. This has the obvious limitations of dependence on the choice of the master structure, and failure to make use of relationships between pairs of non-master sequences. Lesk & Fordham [LF96], in a study of the chymotrypsin-like serine proteases, did structural alignments of all pairs of structures, and collated the results into a common alignment table. However, the experience with those calculations suggests that it would be useful to ask whether any of the known pairwise structural

[1]This work was supported by the Wellcome Trust.

superposition methods lends itself to generalisation to a true multiple superposition approach. The problem is only to determine the residue–residue correspondences, that is, the alignment. Once the alignment is known methods are available for the multiple superposition of the molecules [SBPL92],[D92].

There have been numerous approaches to the problem of structural superposition (for a review see [GL98]). Some operate in three-dimensional space, and are based on detection of small well-fitting pieces and combining them [VS91],[ATG92]; others are based on similarity of contact matrices [HS93],[NRTZ95],[L95].

However, the methods that would seem to be most directly generalizable to multiple alignment are those that reduce the three-dimensional structural superposition problem to a one-dimensional problem. There are several ways to achieve this. One is to characterise each residue by its pattern of neighbours [LVW85],[TO89]. Another is to characterize each residue by its mainchain conformation [LSW84],[KdHN89]. (It is clear that these approaches depend on the linear nature of the polypeptide chain.) Still another is to classify each position in a polypeptide chain by its environment; this also has application to structure prediction by asking whether a particular sequence is compatible with a succession of encoded environments [BLE91].

In this report we pursue the idea that after encoding a protein by a one-dimension characterization of the successive residues, together with limited amino acid sequence information, multiple sequence alignment methods can be applied to produce a multiple structure alignment. We use a set of distantly-related globins as an example and test of feasibility of the method.

Other approaches to multiple structure alignment have been published by Russell & Barton [RB92], Taylor, Flores & Orengo [TFO94], and May & Johnson [MJ95]. Our approach is similar to that of Šali & Blundell [ŠB90].

2 Co-ordinates and Calculations

All co-ordinates are taken from the Protein Data Bank [B77]. For multiple sequence alignment we used the program map, by Huang [H94].

We assign to each residue a symbol that combines information from the amino acid sequence and from the residue conformation.

2.1 Encoding the sequence: reduced amino acid alphabet

We encode the amino acid sequence according to a reduced alphabet corresponding to physico-chemical classes of amino acids:

Table 1. Reduced alphabet based on classifying amino acids into types of similar physicochemical properties

GAST	small nonpolar
CVILP	small/medium hydrophobic
FYMW	large hydrophobic
NQH	polar
DE	charged, negative
KR	charged, positive

2.2 Encoding the conformation

We make use of Efimov's dissection of the Sasisekhan–Ramachandran diagram [E93], with modifications: The conformation of the mainchain of a protein is specified by conformational angles ϕ, ψ and ω. Values of ω are limited to narrow ranges around $+180°$ and $-180°$. Allowed ranges for ψ and ϕ are limited by steric constraints to discrete regions which can be charted in the Sasisekharan–Ramachandran plot. We use the nomenclature of Efimov [E93] but extend his regions to assign to each residue a symbol for the region to which it is closest. (Efimov's definitions cover only a subset of the possible values of ϕ and ψ.) In this way we encode the structure of a protein as a sequence of conformation states of the individual residues:

Table 2. Classification of mainchain conformations based on that of A.V. Efimov [E93]

A	α_r — right-handed $\alpha-$helix
B	β — extended strand
D	throat between α and β regions
L	α_l — left-handed $\alpha-$helix
E	bottom of $+/-$ region (in which $\phi > 0, \psi < 0$)
C	cis-peptide
X	other

From the previous two tables we have assigned to each residue one of six symbols based on its amino acid identity, and one of 7 symbols based on its conformation. By assigning a unique symbol to each possible combination of these we represent each residue by a single character in a 42-character alphabet. Each element of the substitution matrix associated with this alphabet is the sum of a contribution from change in amino acid class (see Table 1) and a contribution from change in conformation class (see Table 2) according to the following rules:

Contribution from amino acid classification:

Same class	uncharged $\leftrightarrow$ uncharged (including polar)	uncharged $\leftrightarrow$ charged
10	5	0

Contribution from conformational classification:

Same class	different class
0	-10

The initiate-gap penalty was 20 and the extend-gap penalty 5.

3 Results

We have implemented the methods described and applied them to three distantly-related globin structures: sperm whale myoglobin, bloodworm globin and leghaemoglobin from yellow lupin. The results are as follows. (The symbols, which correspond to the assignment of a unique character to each ordered pair of reduced amino acid alphabet and residue conformation, should be considered arbitrary.)

```
                 .    :    .    :    .    :    .    :    .    :    .    :
Sperm whale     HBYAYMSGGGSGMA4GYAOG--AASASYGGG4GM4AUGYAGY4NY4M4S----H4BYAY
Bloodworm       HBAAS4SGGAAAM4YGAECOVDAAGA44GGG4MGAAUGSMAAGMDNA------EACZGA
Yellow lupin   EJBYASAAGG4AAMYYMSAUG--G4SAS4MMGGGGYGCGAA4YGNAMG5EBAZHHSUTGY

                     .    :    .    :    .    :    .    :    .    :    .    :
Sperm whale    M4ABYYG44SAGAGGAAGAAGG447EUUYAYG4GGASASA--A4S7HHG4MGYMGAYAGG
Bloodworm      GAAGAA4GGASGAGAGAUGAOYA4MG---ASM4AGAG4S4CNES5TH5ASMMYGGAAAGG
Yellow lupin   GSA-SAA4GM4GGMYAAGSGYGAEHHHBZAAG4SGAAGSG--A6-DHBYASMGGG4YAGG

                     .    :    .    :    .    :    .    :    .    :    .    :
Sperm whale    SGGSA4UGAYOBAYASAAMS4AGYGM44YGAA4M4YGDNV
Bloodworm      AAMYS4GEA4OTAAA4YAMAAAMAYGAAAGGAAGS
Yellow lupin   4AG4YGGEA6NBYYGSAAMAGAMYYGAGGG44YMYYA
```

A translation of these results back to the amino acid sequence follows:

```
Sperm whale    VLSEGEWQLVLHVWAKVEADV--AGHGQDILIRLFKSHPETLEKFDRFKH----LKTEAE
Bloodworm      GLSAAQRQVIAATWKDIAGADNGAGVGKKCLIKFLSAHPQMAAVFGFS------GASDPG
Yellow lupin  GALTESQAALVKSSWEEFNANI--PKHTHRFFILVLEIAPAAKDLFSFLKGTSEVPQNNPE

                     .    :    .    :    .    :    .    :    .    :    .    :
Sperm whale    MKASEDLKKHGVTVLTALGAILKKKGHHEAELKPLAQSHA--TKHKIPIKYLEFISEAII
Bloodworm      VAALGAKVLAQIGVAVSHLGDEGKMV---AQMKAVGVRHKGYGNKHIKAQYFEPLGASLL
Yellow lupin   LQA-HAGKVFKLVYEAAIQLEVTGVVVTDATLKNLGSVHV--SK-GVADAHFPVVKEAIL

                     .    :    .    :    .    :    .    :    .    :    .    :
Sperm whale    HVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
Bloodworm      SAMEHRIGGKMNAAAKDAWAAAYADISGALISGLQS
Yellow lupin   KTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMDDAA
```

In contrast, the following results are from applying the same multiple sequence alignment program to the sequences alone:

```
Sperm whale     VLSEGEWQLVLHVWAKVE-ADV-AGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKA
Bloodworm       GLSAAQRQVIAATWKDIAGADNGAGVGKKCLIKFLSAHPQMAAVFG-FS--------GA
Yellow lupin   GALTESQAALVKSSWEEFN-ANI-PKHTHRFFILVLEIAPAAKDLFS-F--LKGTSEVPQ

                     .    :    .    :    .    :    .    :    .    :    .    :
Sperm whale    SE-DLKKHGVTVLTALG-AI--LKKKGHHEAE--LKPLAQSH---ATKHKIPIKYLEFIS
Bloodworm      SDPGVAALGAKVLAQIGVAVSHLGDEGKMVAQ--MKAVGVRHKGYGNKH-IKAQYFEPLG
Yellow lupin   NNPELQAHAGKVFKLVYEAAIQLEVTGVVVTDATLKNLGSVHVSKG----VADAHFPVVK

                     .    :    .    :    .    :    .    :    .    :    .    :
Sperm whale    EAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
Bloodworm      ASLLSAMEHRIGGKMNAAAKDAW-----------AAAYADIS--GALISGLQS
Yellow lupin   EAILKTIKEVVGAKWSEELNSAW-----------TIAYDEL----AIV--IKKEMDDAA
```

The results were checked against the published structural alignments [LC80],[BCL87], and it can be stated that the structure-based calculation performed somewhat better than the purely sequence-based one. However, extensive tests on a variety of systems are required to evaluate the effectiveness of the method properly. We suggest that the results presented here encourage further development of the approach.

Conclusions

We have designed and implemented a simple method for multiple structural alignment, using a one-dimensional representation of the conformation of a polypeptide chain, combined with the sequence, and standard multiple *sequence* alignment methods to perform the alignment.

References

[ATG92] Alexandrov, N.N., Takahashi, K. & Gō, N. (1992). Common spatial arrangement of backbone fragments in homologous and non-homologous proteins. J. Mol. Biol. 225. 5–9.

[BCL87] Bashford, D., Chothia, C. & Lesk, A.M. (1987). Determinants of a protein fold: Unique features of the globin amino acid sequences J. Mol. Biol. 196, 199–216.

[B77] Bernstein, F.C, Koetzle, T.F., Williams, G.J.B., Meyer, E.F. Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer based archival file for macromolecular structures. J. Mol. Biol. 112, 535–542.

[BLE91] Bowie, J.U., Lüthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. Science 253: 164–170.

[D92] Diamond, R. (1992). On the multiple simultaneous superposition of molecular structures by rigid body transformations. Protein Science 1, 1279–1287.

[E93] Efimov, A.V. (1993). Standard structures in proteins. Prog. Biophys. Molec. Biol. 60, 201-239.

[GL98] Gerstein, M. & Levitt, M. (1998) Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. Prot. Sci. 7, 1–12.

[HS93] Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. J. Mol. Biol. 233, 123–138.

[H94] Huang, X. (1994) On global sequence alignment. Computer Applications in the Biosciences 10, 227–235.

[KdHN89] Karpen, M.E., de Haseth, P.L. & Neet, K.E. (1989). Comparing short protein substructures by a method based on backbone torsion angles. Proteins: Structure, Function and Genetics 6, 155–167.

[LC80] Lesk, A.M. & Chothia, C. (1980). How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins J. Mol. Biol. 136, 225–270.

[L95] Lesk, A.M. (1995). 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.

[LF96] Lesk, A.M. & Fordham, W.D. (1996). Conservation and variability in the structures of serine proteases. J. Mol. Biol. 258, 501–537 (1996).

[LSW84] Levine, M., Stuart, D. & Williams, J. (1984). A method for systematic comparison of the three-dimensional structures of proteins and some results. Acta crystallographica A40, 600–610.

[LVW85] Liebman, M. N., Venanzi, C.A. & Weinstein, H. (1985). Structural analysis of carboxypeptidase A and its complexes with inhibitors as a basis for modelling enzyme recognition and specificity. Biopolymers 24, 1721–1758.

[MJ95] May, A.C.W. & Johnson, M.S. (1995). Improved genetic algorithm-based protein-structure comparisons – pairwise and multiple superpositions. Protein Engineering 8, 873–882.

[NRTZ95] Nichols, W.L, Rose, G.D, Ten Eyck, L.F. & Zimm, B.H. (1995). Rigid Domains in Proteins: An Algorithmic Approach to their Identification. Proteins: Structure, Function, Genetics 23, 38–48.

[RB92] Russell, R. B. & Barton, G. J. (1992), Multiple sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. PROTEINS: Struc. Func. Genet., 14, 309–323.

[ŠB90] Šali, A & Blundell, T.L. (1990). Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. J. Mol. Biol., 212, 203–228.

[SBPL92] Shapiro, A., Botha, J.D., Pastore, A. & Lesk, A.M. (1992). A method for multiple superposition of structures. Acta Crystallographica A48, 11–14.

[TFO94] Taylor, W.R., Flores, T.P. & Orengo, C. (1994) Multiple protein structure alignment. Prot. Sci. 3, 1858-1870.

[TO89] Taylor, W.R. & Orengo, C.A. (1989). Protein structure alignment. J. Mol. Biol. 208, 1–22.

[VS91] Vriend, G. & Sander, C. (1991). Detection of common three-dimensional substructures in proteins. Proteins: Structure, Function and Genetics 11, 52–58.