**Protein Engineering** vol. 7 no.3 pp. 331-333, 1994 ©Oxford University Press

# A heuristic procedure for the detection of locally similar substructures of two equivalent structures.

**Mihaly Mezei**

Department of Physiology and Biophysics, Mount Sinai School of Medicine, CUNY, New York, NY 10029, USA.

**A heuristic procedure is presented for the detection of locally similar substructures in two equivalent molecules, using elementary graph theoretical concepts. The capabilities of the procedure are demonstrated on the comparison of the calmodulin backbone in the X-ray structure and in a structure obtained from molecular dynamics.**

*Key words:* calmodulin/connected graph/molecular similarity

## Introduction.

Comparison of molecules of similar composition is a frequently occurring problem in the study of macromolecules. In most instances the comparison involves different molecules and the aim is to detect similar motifs between them and/or find correspondence between various parts of the two molecules. In this case the principal difficulty is the matching of the atoms in the two molecules (alignment). This is done based on a combined knowledge of chemical similarities and conformational similarities observed between substructures. Efforts of this type have been reviewed recently in Overington (1992). However, for comparisons where the assignment is not a problem (e.g. the structures compared are two conformations of the same molecule) the following question (in some sense the opposite of the one discussed above) can be raised: given the equivalence between the atoms the two structure, how can one (quickly) detect substructures that are significantly more similar to each other than molecules themselves.

Quantitative comparison requires a measure of similarity. The widely used measure is the RMS (Rao and Rossmann, 1992) between the two structures consisting of $N$ atoms:

$$RMS = \left[ \sum_{i=1}^{N} (x_i - y_i)^2 / \sum_{i=1}^{N} w_i \right]^{1/2}$$

where $x_i$, $y_i$ are the coordinates of the $i$-th atoms in two molecules considered (or known) to be equivalent and $w_i$ are their (positive) weight factors. Zero RMS clearly guarantees the identity of the two structures and a low RMS ensures strong similarity with reasonable certainty, but the converse is not true: consider two structures that contain two identical substructures connected by a bond but the torsion angle around the bond is different in the two structures. For such situations the question of identifying similar substructures (raised above) arises naturally. This note describes a heuristic procedure based on elementary graph theory that is able to detect substructures of significantly stronger similarity than the

overall RMS would indicate. It does so by identifying bonds that act as hinges between the domains of conserved structures. It is clear that the problem can be solved by a brute force approach by generating all the possible partitions of the molecule into substructures and comparing them, but this would escalate the computational requirements: considering just the selection of one substructure consisting of a contiguous stretch of a protein backbone requires $N*(N-1)/2$ RMS calculations, and to sort out the possible partition combinations requires further significant computational effort.

**Theory**.

In order to determine which bonds are the likely candidates to be hinges between conserved substructures, the proposed procedure first determines a local bond RMS for each bond in the molecule. This bond RMS is defined as the minimized (with respect to relative position and orientation) RMS for the subset of atoms consisting of the two atoms forming the bond and all their neighbors (thus, in most cases a subset of eight atoms or less). Clearly, if the bond is part of a perfectly conserved substructure, this bond RMS is zero while for a 'hinge bond' between two conserved structures it will be significantly larger than zero. In practice, there will be a distribution of bond RMS values and the bonds with the largest values are the best hinge candidates. The algorithm requires at this point a bond RMS threshold value $\mathrm{RMS}_{\mathrm{lim}}$ — any bond with bond RMS higher than $\mathrm{RMS}_{\mathrm{lim}}$ will be treated as a hinge.

Next, a graph of the molecules (having already made the connection between the atoms in the two structures, the graph is the same for both molecules) is considered whose vertices are the atoms of the molecule and edges are drawn between any two atoms that are chemically bonded. For a molecule, this is a connected graph (i.e. one can find a succession of bonds that connects any two atoms). The edges corresponding to bonds defined as hinge bonds by the $\mathrm{RMS}_{\mathrm{lim}}$ value selected are deleted from the graph and the substructures corresponding to the connected subraphs of the graph thus obtained are considered candidates for conserved substructures.

Once the substructures are determined, their minimized RMS can be computed and compared with their contribution to the original overall RMS. The procedure can be repeated with various cut-off values — the smaller the threshold value, the smaller the RMS of the resulting substructures are expected to be. Depending on the particular application, one can decide what an 'acceptable' RMS for local similarity is. This decision can be helped by the consideration of other measures of similarity, like the maximum distance between the equivalent atoms and by the visual examination of the 'fit' of the substructures on a graphics terminal. The threshold value $\mathrm{RMS}_{\mathrm{lim}}$ can then be adjusted accordingly and new substructures can be generated.

Notice also that the procedure is more likely to give 'false positives' than to miss finding substructures with low RMS since it is rather unlikely to find compensating large local RMS in a (sub)structure. False positives, however, can be easily screened out based on their local RMS or maximum distance value.

**Table I.** Calmodulin conserved substructures found with different $RMS_{lim}$ values.

| | $RMS_{lim}$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.15 | | | 0.2 | | | 0.3 | | | 0.4 | | |
| Helix | Resno | RMS | MXD | Resno | RMS | MXD | Resno | RMS | MXD | Resno | RMS | MXD |
| 6-19 | 7-18 | 0.4 | 1.4 | 5-18 | 0.5 | 1.5 | 1-23 | 3.4 | 7.0 | 1-44 | 3.3 | 8.3 |
| 29-36 | | | | 26-41 | 0.6 | 1.4 | 26-44 | 0.9 | 2.7 | | | |
| 45-53 | 45-59 | 1.2 | 2.2 | 45-62 | 1.4 | 2.5 | 45-62 | 1.4 | 4.5 | 45-93 | 6.9 | 15.5 |
| 65-92 | 62-73 | 0.6 | 1.3 | 62-77 | 0.9 | 1.7 | 62-82 | 2.0 | 4.2 | 96-117 | 1.5 | 3.0 |
| 102-112 | | | | 82-93 | 0.3 | 0.8 | 82-93 | 0.3 | 0.8 | | | |
| 118-126 | | | | | | | 96-117 | 1.5 | 3.0 | | | |
| 138-144 | | | | | | | 118-148 | 1.6 | 3.6 | 117-148 | 1.6 | 3.4 |

The substructures are characterized with the initial and final residue number (columns headed by Resno), by the RMS between the substructures in the two molecules compared, RMS (in Å) and by the maximum distance found between equivalent atoms, MXD. Substructures shorter than 10 residues are not shown. The column headed by Helix gives the residue ranges for the $\alpha$-helices found in the crystal structure.

**Computational details**.

The determination of the optimal RMS was performed in two steps. First, both molecules were translated so that the center of masses are at the origin of the coordinate system. As we used constant weights $w_i$ in the RMS calculations, the center of mass calculation was also done with assuming identical atomic masses. Next, the orientation of the second molecule was determined that minimized the RMS. The optimal orientation can be obtained by the analytical procedure of Kabsch (1976). For large molecules the bond list can be obtained efficiently by the cell index method of Quentrec and Brot (1973).

The connected subgraphs of the graph were determined using a depth-first-search approach customarily used to 'visit' all vertices of a graph (Brassard and Bratley, 1988), assuming that for each atom a list of its neighbors is prepared. The algorithm is described in Appendix 1. A C program called SIMLOC, available from the author on request, has been written to perform these calculations. In programming the algorithm generating the connected subgraphs, the rather simplistic check of comparing the total number of atoms on the lists with the number of atoms in the molecule was found very useful in eliminating coding errors.

**Results and discussion**.

The first test was the comparison of the $C_7$ and $\alpha_R$ conformations of the alanine dipeptide, $CH_3CONH\text{-}CH(CH_3)\text{-}CONHCH_3$. The two conformations differ only in the torsion angles around the two bonds indicated above, i.e. these bonds are the hinges discussed above. Except for the two bonds around which the torsions occur, the bond RMS values were indeed all found to be zero. Hence choosing the bond RMS threshold value to be less than the smaller of the two, resulted in the proper partitioning of the molecule into the three substructures $CH_3CONH$, $CH(CH_3)$, and $CONHCH_3$.

The procedure was also tested on a protein. The backbone structure of calmodulin obtained from the Brookhaven protein data bank (with residues 1-4 and 148 obtained from

model building) was compared with the corresponding backbone from structure generated with molecular dynamics by Mehler *et al.* (1991). The RMS between the two backbone structures was determined as 14.1 Å and the maximum distance between equivalent atoms was 26.6 Å. The superimposition of the two backbones are also shown on Figure 1. Table I shows the conserved substructures that are at least ten residues long found by our procedure through systematic variation of the bond RMS threshold value $\mathrm{RMS}_{\mathrm{lim}}$. Interestingly, several of the substructures found closely match the first four $\alpha$-helices found in the crystal structure.

The maximum distance between corresponding atoms is consistently about two to three times the corresponding RMS value. This is the expected behavior of random error, since assuming a Gaussian distribution of the distances between equivalent atoms, the probability of finding a value that is over twice or thrice the RMS value is 5% and 1%, respectively.

The superimposition of the backbone segments 45-62 and 82-93 are also shown on Figure 2. The RMS between the substructures was 1.35 and 0.33, respectively, and quality of overlaps reflect this difference.

The results presented thus demonstrate the capability of the proposed procedure to find substructures with low RMS. This is particularly gratifying since even the calmodulin calculations took less than a minute on a DECstation 5000/25 workstation. Furthermore, apart from the initial generation of the bond list, the computer time required by each step is linear either in the number of bonds or in the number of atoms and therefore application to significantly larger systems should not present any problems.

**Acknowledgments**.

This work was supported by NIH grant #R55-GM43500. I thank E. Mehler for providing the calmodulin data. The figures were generated with the InsightII program of Biosym Technologies.

**References**.

Brassard, G. and Bratley, P. (1988) *Algorithimics*. Prentice Hall, Englewood Cliffs, NJ.

Kabsch, W., (1976) *Acta Cryst.,* **A32**, 922–923.

Mehler, E.L., Pascal-Ahuir, J.L. and Weinstein, H., *Protein Engineering*, (1991) **4**, 625-637.

J.P. Overington, *Current Opin. Struct. Biol.,* (1992) **2**, 394–401.

Quentrec, B., and Brot, C., *J. Comput. Phys.,* **13,** 430–432.

S.T. Rao, and M.G. Rossmann, *J. Mol. Bio.,* (1973) **76**, 241–256.

*Received August 8, 1993; revised October 15, 1993; accepted October 31, 1993*

## Appendix 1

*The determination of connected subgraphs.*

The procedure used to determine the connected subgraphs of a graph consisted of the following steps:

(i) For each atom, set a flag to UNUSED.

(ii) A selection of a connected subgraph begins with the first atom still flagged UNUSED. This atom starts the list of atoms in the new cluster. Inclusion into the list results also in flagging the atom USED. A pointer ROOT is set to this place on the list.

(iii) The last included atom is examined, and if it has an UNUSED neighbor, that neighbor is included in the list. This is repeated until the last atom included has no UNUSED neighbor.

(iv) Next, the algorithm goes back to the atom pointed to by ROOT. If that atom has no more UNUSED neighbors, ROOT is incremented until either an atom with an UNUSED neighbor is found on the list or ROOT points to the last atom on the list. In the former case, that neighbor is included and the algorithm continues with step 3. In the latter case the list of vertices of this connected subraph is complete.

(v) If there are no more UNUSED atoms, the partitioning is completed and the algorithm stops. Otherwise, the algorithm is to be repeated from step 2.

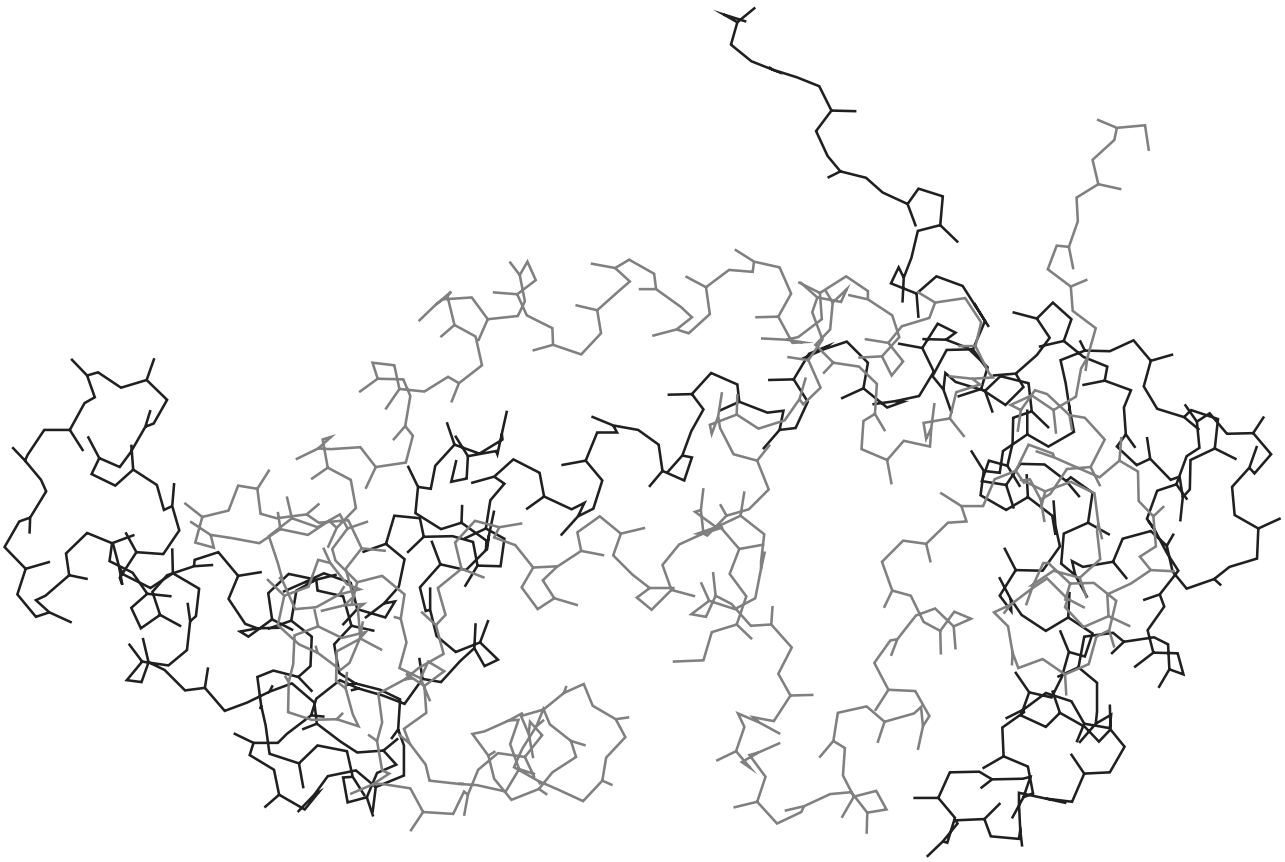

Figure 1. Superimposition of the X-ray (full lines) and model calmodulin backbones.

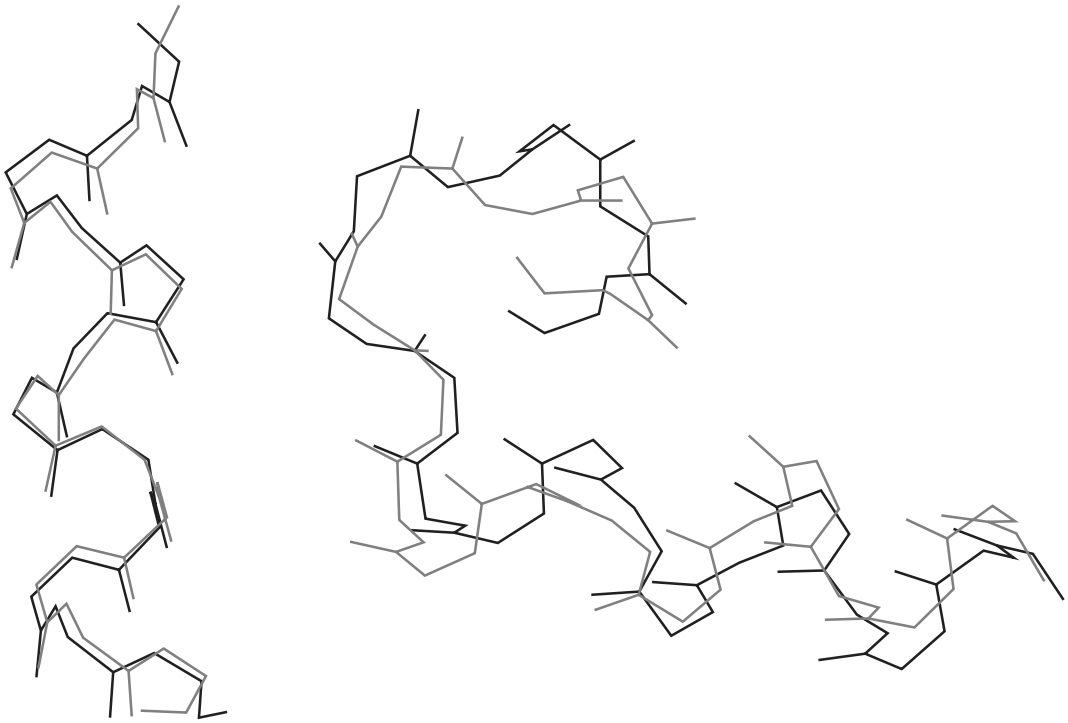

Figure 2. Superimposition of segments 45-62 (right) and 82-93 (left) from the X-ray (full line) and model (shaded line) backbones.