Structure Alignment of Two Proteins

[16]Structural alignment is a form of sequence alignment based on comparison of shape. These alignments attempt to establish equivalences between two or more polymer structures based on their shape and three-dimensional conformation. Because structural similarity is conserved more than sequence similarity, it can be used as a more powerful "telescope" to look back to earlier evolutionary history.

[12]Protein comparison generally involves three steps: (1) deciding what feature to compare, (2) deciding how to compare the chosen feature, and (3) determining whether the feature shows a significant similarity compared to chance.

[15]In contrast to sequence alignment, the first step outlined above is more complicated with three-dimensional protein structures. One can compare the coordinates of C_{alpha} atoms, secondary structure elements (SSEs), or internal mappings such as a contact map or distance matrix.

[3]The most direct approach to accomplish step two is to move the set of points representing one structure as a rigid body over the other, and look for equivalent residues. However, this can only be achieved for relatively similar structures and will fail to detect local similarities of structures sharing common substructures. To avoid this problem, the structures can be broken into secondary structure elements, but this can lead to situations in which the global alignment can be missed. Recent work has focused on combining the local and global criteria in a hierarchical approach. These

methods proceed by first defining a list of equivalent positions in the two structures, from which a structural alignment can be derived. This initial equivalence set is defined by methods such as [13]dynamic programming, [9]comparison of distance matrices, and [18]fragment matching. Optimization of this equivalence set is performed using dynamic programming, Monte Carlo algorithms or simulated annealing, a genetic algorithm and incremental combinatorial extension of the optimal path. The majority of the current methods for protein structure alignment quantify the quality of the alignment on the basis of geometric properties of the set of points representing the structures. Some of these methods compare the respective distance matrices of each structure, trying to match the corresponding intramolecular distances for selected aligned substructures; whereas others compare the structures directly after superposition of aligned substructures, trying to match the positions of corresponding atoms.

In order to execute step (3), several scoring schemes have been proposed. [8]Taylor and Orengo defined a distance or similarity score in the form a/(D+b), where D is the difference between the two intramolecular distances, and a and b are arbitrarily defined constant values. [9]Holm and Sander defined a similarity score as (a–[D/<D>])exp(–[<D>/b]2), where <D> is the average of the two intramolecular distances. [6]Rossmann and Argos, and Russell and Barton used a score exp(–[D/a]2)exp(–[S/a]2), where S takes into account local neighbors for each pair of atoms. As another example of a scoring scheme for minimizing intermolecular distances, Levitt and co-workers defined a [12]score a/(1+[R/b]2), where R is the distance between a pair of corresponding atoms in the two structures. At this stage, there is no clear evidence as to which score performs best.

There are several problems associated with the structure alignment procedures described above.

First of all, in literature, thoughtful discussions of the issues[1] associated with the structural alignment of two proteins have been presented emphasizing the fact that there are many ambiguities associated with the problem. This difficulty is reflected in the numerous measures that have been designed to quantify similarity. Root-mean-square deviation (RMSD) of C_{alpha} coordinates is the most common measure of similarity but many others, such as distance or contact maps, have been introduced. Algorithms such as Dali[9], SSAP[5] are widely used, each optimizing a different measure of similarity in its alignment procedure. To some extent, the similarity measure and algorithm of choice are dictated by the application. For example, is one interested in aligning only protein cores, which might be appropriate for fold recognition applications, or entire proteins; are differences in loop geometry of interest? Different answers to these questions might well dictate the use of different algorithms and of different measures of similarity.

Secondly, as discussed in the lecture, unlike the optimization used for sequence alignment, which is globally convergent, structural alignment optimization is not. This is the case for sequence alignment because the optimum match for one part of a sequence is not affected by the match for any other part. Structural alignment fails to converge globally because the possible matches for different segments are tightly linked as they are part of the same rigid 3D structure. For this reason, the alignment found by a structural alignment algorithm can depend on the initial equivalences, whereas in sequence alignment there is no such dependence.

[16]Finally, there are a few drawbacks to using root mean square distance between sets of corresponding atoms after two structures have been superimposed as the measure of the similarity of two macromolecules. (a) the actual number obtained depends rather critically on the set of atoms that is chosen for the calculation. (b) there may be a problem in defining the most suitable superpositioning of two molecules. (c) there may be isolated regions (e.g., flexible loops) which display differences and cause high r.m.s.d, values, whereas, on the whole, the structures are very similar. (d) the use of only C_{alpha} atoms, which is common practice, involves loss of detail with respect to the actual similarity of the geometry of the main chain. (e) when all non-H atoms are used in a comparison, there are trivial naming conventions to cope with which are easily overlooked. For instance, if a tyrosine side chain was 'flipped' during simulated annealing refinement in one molecule, but not the other, the two residues may have an r.m.s.d, exceeding 1 Angstrom, even though their conformations are chemically and structurally identical.

When given the structures of two proteins to align, I would first determine whether computational time will be an issue to consider. If the running time is not a concern, a polynomial-time algorithm [19] proposed by Kolodny and Linial that optimizes both the correspondences (a sequence of residue pairs that a sequential alignment of the two substructures yields) and rigid transformations exists. The algorithm focuses on

STRUCTAL scores that evaluate the similarity of two structures by explicitly applying a rigid transformation to one and then comparing the transformed structure with the other. For such scores, the optimization problem is to find transformations and correspondences of (near) optimal score. After calculating the optimal score for a substantial number of rotations and translations, it then sifts through these scores to find the best ones with near globally maximum scores. Because their algorithm is not heuristic, it guarantees finding ε (percent error) -approximations to all solutions of the protein structural-alignment problem. However, because of its highly computationally demanding nature (O((length of the protein)^10)/(ε^6)), it is too slow to be a useful everyday tool for comparing a large number of structures.

As of today, this approximate solution is the most accurate algorithm for structure alignment, as the "best" solution does not exist. [19]This is because proteins are flexible and constantly fluctuating about a mean position. Consequently, protein coordinates obtained from physical experiments are merely approximations to a "true" position and are necessarily noisy. Therefore, seeking exact solutions based on experimental data is both impossible and pointless.

If computational time does become an issue, several publicly available programs utilizing heuristics can be used to facilitate the structural alignment. These tools include SSAP, STRUCTAL, DALI, LSQMAN, CE, and SSM. It is important to keep in mind, however, that none of the above-mentioned heuristics guarantees finding an optimal alignment with respect to any scoring function.

The largest and most comprehensive comparison of protein structure alignment methods as of today was performed by Kolodny and co-workers in 2005[21]. In this study, the performance of structure alignment tools was evaluated by aligning all 8,581,970 protein structure pairs in a test set of 2930 protein domains specially selected from CATH v.2.4 to ensure sequence diversity. It was found that when comparing more similar structures, STRUCAL, LSQMAN, and SSM performed the best; while STRUCTAL, CE and SSM missed fewest alignments when aligning less similar structures. In addition, the running time for each program on all pairs of structures was also calculated, showing that SSM and LSQMAN were the fastest, followed by STRUCAL and CE. Therefore, STRUCAL, LSQMAN, SSM, and CE are the four programs that consistently demonstrated the highest performance in both speed and accuracy. A brief description for each of four algorithms is presented below:

[12] The STRUCTAL algorithm starts with an initial alignment of the backbone C_{alpha} atoms of the two structures according to one of a number of possible heuristics (aligning the beginnings, the ends, random segments, by sequence similarity, etc). Then a two step process is repeated until convergence. First, a dynamic programming algorithm analogous to the Needleman and Wunsch sequence alignment algorithm finds the correspondence between the two structures that yields the highest score. Second, an optimal relative orientation is computed for the two structures, based on the previously computed correspondence.

[21]LSQMAN iteratively searches for a rigid body transformation that superimposes the structures. The initial transformation is calculated by optimally superimposing the first residues of the secondary structure elementsin the two structures. Once the

structures are superimposed, LSQMAN starts by searching for a long alignment, where matching residues are within 6A ° of each other, and the minimum fragment length is four residues. Given the alignment, an optimal transformation is calculated, starting a new iteration.

[18] Secondary Structure Matching (SSM) iteratively searches for a rigid body transformation that superimposes the structures; it then finds an optimal alignment for this transformation. The initial transformations are found by matching substructures in the three-dimensional graphs that describe the structures in terms of secondary structure elements and their relative position and orientation. SSM then iteratively finds a correspondence of nearby C_{alpha} atom pairs, one from each structure, and optimally superimposes the corresponding sets. The procedure for finding the correspondence is greedy in nature. First it matches nearby residues of matched SSEs, then nearby residues of non-matched SSEs, and then more nearby residues that are not part of SSEs.

[2]The combinatorial extension (CE) method breaks each structure in the query set into a series of fragments that it then attempts to reassemble into a complete alignment. A series of pairwise combinations of fragments called aligned fragment pairs, or AFPs, are used to define a similarity matrix through which an optimal path is generated to identify the final alignment. Only AFPs that meet given criteria for local similarity are included in the matrix. An alignment path is calculated as the optimal path through the similarity matrix by linearly progressing through the sequences and extending the alignment with the next possible high-scoring AFP pair.

In Kolodny's study, the performance of each program was judged by the following criteria: number of residues matched, number of gaps, and the length of alignment of gaps. Since these alignment properties are not independent, researchers devised alignment scores that balance these values, which are the SAS, SI, MI, GSAS geometric scores, small values of which indicate a good alignment. Because all of the programs being studied output sufficient information that enable geometric scores to be calculated, geometric measures allow direct comparison and evaluation of the different alignments of a pair of structures found by different methods.

Based on the results obtained from Kolodny's study, I will use the following procedures to align two proteins using heuristics based tools:

(1) Use STRUCAL, LSQMAN, SSM, and CE to align the two proteins, record the best alignment output from each program.

(2) From each output, calculate the SAS, SI, MI, GSAS scores for each alignment generated from (1).

(3) Out of the four alignments produced in (1), the one that has the lowest mean geometric score (ie. (SAS+SI+MI+GSAS)/4) is labeled the best alignment.

Note that in step (1), instead of merely relying on the top performing program, STRUCTAL, to generate the best alignment, all four programs will be used. This is necessary because each method optimizes a different measure of similarity in its alignment procedure, and their performance will highly depends on the structures of the specific compounds that are compared. Therefore, combining the results from all of the four programs, which are responsible for generating over 90 percent of the best alignments in Kolodny's study, will statistically enhance the probability of obtaining an optimal alignment. In step (2), an alternative way of assessing the performance of each program is to use ROC curve, which is what Sierk & Pearson[11] used in their previous study in evaluating the performance levels of various structure alignment tools. However, geometric comparison is more sensitive than ROC curve, as Kolodny pointed out in his paper, because with ROC approach, the quality of the alignments is not taken into account: sometimes a method that finds less good alignments scores better than a method that finds better alignments.

1. Godzik A. (1996) The structural alignment of proteins: is there a unique answer? *Protein Sci* 5:1325-8.
2. Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* 11: 739-747.
3. Zhang Y, Skolnick J. (2005) TM-align: A protein structure alignment algorithm based on TM-score. *Nucleic Acids Research* 33: 2302-2309.
4. Zhang Y, Skolnick J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins* 57: 702-710.

5. Taylor WR (1996) SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymology* 266:617-35
6. Levitt M., Gerstein M. (1998). A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl Acad. Sci*. 95:5913–5920
7. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH—a hierarchic classification of protein domain structures. *Structure*, 5: 1093–1108
8. Taylor,W. R. & Orengo, C. A. (1989). Protein structure alignment. *J. Mol. Biol*. 208: 1–22.
9. Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol*. 233: 123–138.
10. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*. 247: 536–540.
11. Sierk, M. L. & Pearson, W. R. (2004). Sensitivity and selectivity in protein structure comparison. *Protein Sci*. 13: 773–785.
12. M. Gerstein and M. Levitt. Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. *Protein Science*, 7:445–456: 1998.
13. Subbiah, S., Laurents, D. V. & Levitt, M. (1993).Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. *Curr. Biol.* 3: 141–148.
14. Holm, L. & Park, J. (2000). DaliLite workbench for protein structure comparison. *Bionformatics*, 16: 566–567.
15. Yang, A. S. & Honig, B. (2000). An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structure alignment and quantitative measure for protein structural distance. *J. Mol. Biol.* 301: 665–678.
16. Koehl, P. (2001). Protein structure similarities. *Curr. Opin. Struct. Biol.* 11: 348–353.
17. Satow, Y., Cohen, G. H., Padlan, E. A. & Davies, D. R. (1986). Phosphocholine binding immunoglobulin Fab McPC603. *J. Mol. Biol*. 190: 593–604.
18. Krissinel E, Henrick K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr*. 60: 2256-68
19. Kolodny, R. & Linial, N. (2004) Approximate protein structural alignment in polynomial time. *Proc. Natl Acad. Sci.* 101: 12201–12206.
20. Subbiah, S., Laurents, D. V. & Levitt, M. (1993). Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. *Curr. Biol.* 3: 141–148.
21. Kolodny R., Koehl P.,Levitt M. (2005). Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J Mol Biol.* 346:1173-88