Protein Structure Similarity

Lecturer: Prof. Latombe

Scribe: Michael Lesnick [mlesnick@stanford.edu], borrowing liberally in some places from notes by Adreas Sundquist, Chen Ding, Huang Yang, and Meghna Agrawal, and from a couple of web resources.

Introduction

A protein's 3D structure largely determines its functional properties. As a result, knowledge of the 3D structure of a protein can yield useful information about the functional properties of the protein. In particular, structural similarity between proteins is a very good predictor of functional similarity.

Since a protein's amino acid sequence determines 3D structure, which in turn determines protein function, one might think that sequence similarity is also very good predictor of functional similarity, but this turns out to be less the case than with structural similarity. Vastly different amino acid sequences can yield very different structures, and similar sequences can sometimes yield dissimilar structures. Thus, sequence similarity is a far less reliable predictor of functional similarity than structural similarity is.

In this lecture, we discuss methods for protein structure acquisition, some key concepts in protein structure similarity comparison, and some applications of protein structure similarity comparison.

Tools for Structure Prediction and Determination

In order to classify proteins according to structure, we must first know the structures of the proteins in question. Protein structure is acquired using both experimental methods and computational methods which predict 3D structure from sequence information, but at this time the computational methods lag far behind the experimental methods in terms of power. However, experimental techniques can be costly, slow, and unusable for the acquisition of the structure of some proteins, so better computational techniques for predicting protein structure from sequence information are quite desirable. Right now, only about 10% of known protein sequences have had their structures determined.

The Protein Data Bank (www.pdb.org) is a freely accessible database of 3-D protein structures. Begun in 1971 with 7 structures, it now has nearly 40000 structures, with the yearly number of structures added to the database increasing each year. In 2004 and 2005 over 10000 new structures were added to the database. As of relatively recently, of the structures in the PDB, only 3% were obtained using computational models.

Computational Techniques for Structure Determination

The main computational structure prediction techniques are ab initio- techniques, homology modeling, and threading.

Ab Initio Methods- These techniques attempt to determine protein structure from scratch by finding the global minimum of an energy function defined on the space of possible structural conformations of the protein. With present methods these techniques are extremely computationally costly and thus have been used only for very small proteins.

Homology Modeling- is based on the reasonable assumption that two homologous proteins will share very similar structures. Given the amino acid sequence of an unknown structure and the solved structure of a homologous protein, each amino acid in the solved structure is mutated, computationally, into the corresponding amino acid from the unknown structure. [Source of this description of Homology Modeling: Wikipedia]

Threading- Given the amino acid sequence of a protein of interest, one attempts to align the sequence to each amino acid sequence in a library of template proteins of known structure in such a way that a quasi-energy score or other score is minimized. The score of an alignment is defined in such a way that the value of the score reflects the extent to which the given alignment predicts a structural similarity of the protein of interest to the template protein. Best structural alignment scores are computed for template protein and the template with the best score amongst all templates is returned. Threading relies on the fact that there are far more proteins than folds, so that a given protein of unknown structure is likely to have structure quite similar to that of a protein of known structure. See http://www.stanford.edu/class/cs273/refs/torda_chapter_proteomics.pdf for more info.

Experimental Techniques for Structure Determination

X-ray Diffraction Crystallography- This is the most widely used method for protein structure determination. In this method the protein is crystallized and an X-ray beam is projected on the crystals. It interacts with the electronic cloud of the crystal to produce diffracted X-ray beams. The diffraction pattern is obtained on a phosphor screen and an electron density map is generated from it which is used to create the 3D structure of the protein from the map. The map tends to be fuzzy in some parts (due to the problem of phasing loops) but the software used can usually predict up to 90% of the structure correctly and the rest is computed manually. This method is expensive and takes time, sometimes longer than an year. It is useful for determining the structure of relatively large proteins but the proteins have to be folded. Also, it requires the protein in form of a crystal and not every protein can be crystallized. 82% of all structures in the Protein Data Bank were obtained with X-ray Diffraction Crystallography**.**

Nuclear Magnetic Resonance Spectroscopy- NMR spectroscopy allows structure determination in solution under conditions that approximate the physiological environment of a protein. It is based on the observation of physical phenomena exhibited when nuclei absorb energy from a radio frequency source at certain characteristic frequencies in the presence of strong external magnetic fields. The position of the nuclei

in the molecule effects the electronic environment of the nucleus and thus affects the absorption frequency. The frequency differences observed in the resultant spectrum can be used to determine the molecular structure of the sample. NMR has low sensitivity and the data obtained is noisy. It is used for smaller proteins. [some of this section is adapted from http://www.process-nmr.com/process_nmr_faq.htm]

Key Definitions and Issues in Structural Similarity Comparison

Definition of 3D Molecular Structure: We represent the 3D molecular structure of a protein as a collection of (possibly typed) atoms or groups of atoms in some given 3D relative placement. The placement of a group of atoms is defined by the position of a reference point (e.g. the center of a particular atom in the group) and the orientation of a reference direction. When we say that the atoms or groups of atoms may be typed, we simply mean that we may choose to label each point representing an atom or group of atoms in the structure with a tag indicating what atom or group of atoms the point is representing.

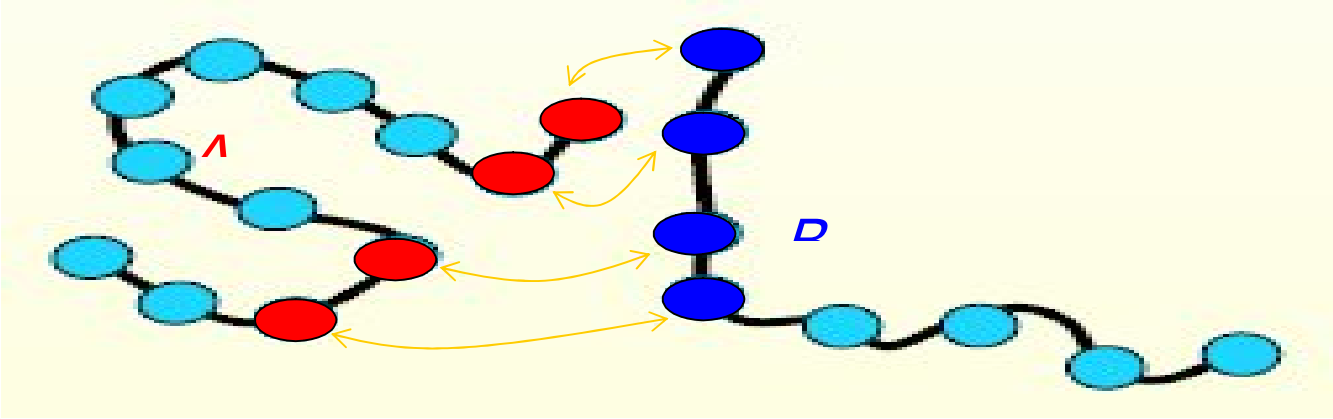
Definition of a Matching Between Two Structures
Two structures match if and only if we have:
1.Correspondence—There is a one to one map between elements of the structure
2.Alignment—There exists a rigid body transform T such that the RMSD between elements in A and those in T(B) is less than some threshold ε,

In practice a complete match of this sort between two proteins is rarely possible; in many cases of interest, two proteins may be only locally similar, may be of different sizes, or may differ structurally in other ways despite significant structurally similarity in other respects. In these cases a complete match of the two proteins is clearly too much to ask for. We can, however, hope for a partial match of two proteins.

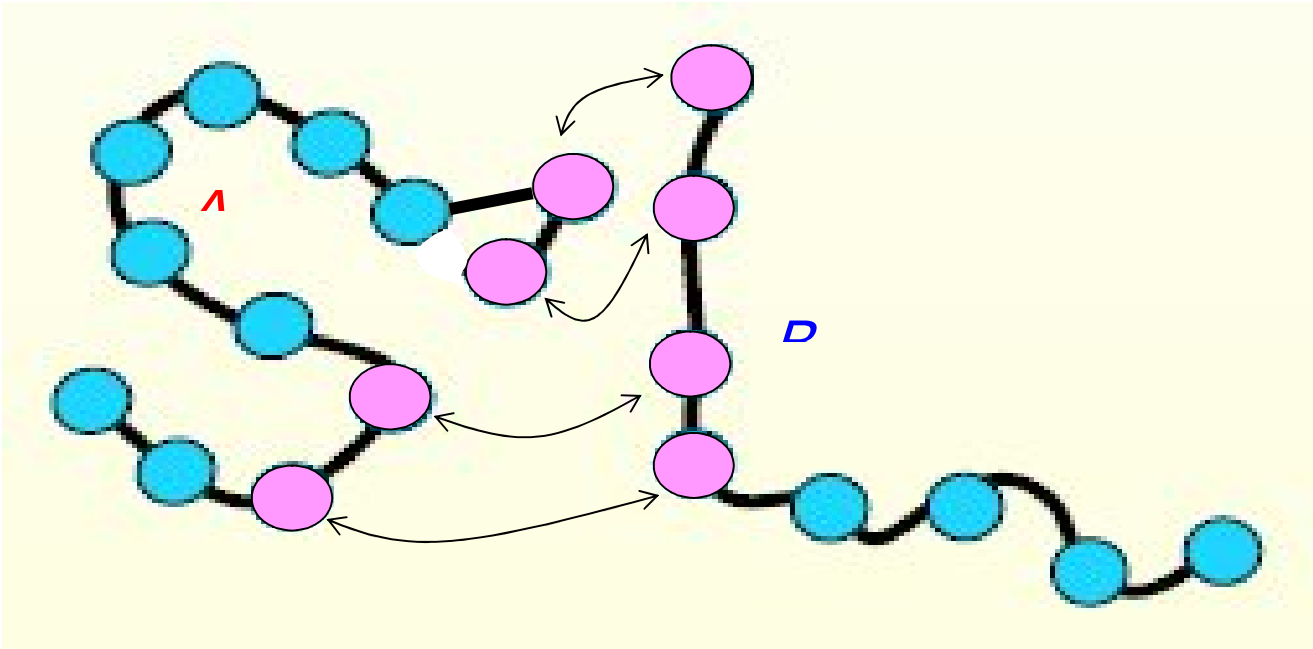
We say we have a partial matching between two proteins A and B when we have a substructure σ (A) of A and a substructure σ (B) of B such that there is a correspondence between σ (A) and σ (B) and an alignment T of A with B such that the RMSD between elements in σ (A) and those in T(σ (B)) is less than some threshold ε. We call the substructures σ (A) and σ (B) **the supports** of A and B, respectively. When a support is small, we refer to it as a motif. Generally, we do not require the support of a protein to be connected; the support may have two or more components which may not lie contiguously on the protein, and this can add to the challenge of the problem of finding a partial matching, as discussed below.

In formulating the problem of partial matching as above, a problem dual to that of finding the transform which minimizes RMSD arises—namely that of choosing the supports of A and B. Clearly there is a tradeoff between the allowed size of the supports of the two structures being aligned and the size of the RMSD. A common solution is to declare at the outset some maximum value ε of the RMSD and to then find the largest supports of A and B such that the RMSD between A and B with respect to those supports is less than ε.

Beyond the size of the support and the RMSD calculated from a match, there are a number of other issues that should be considered in the development of a measure of partial similarity between two proteins. For one, there may be multiple partial matches between substructures of 2 proteins. Secondly, if non-contiguous supports are permitted, one must consider the matter of whether and how to penalize to penalize for gaps in the supports of A and B, such as that depicted in the partial match below.



Third, one must consider how and whether to penalize matches where a subregion of the support of B has its orientation with respect to the backbone of B flipped relative to the orientation of the corresponding substructure of A with respect to the backbone of A, as in the picture below.



Fourth, we must decide to what extent our scoring method will adjust the score according to preference for type or backbone sequence matching. Fifth, we may wish to weight correspondences along accessible parts of the protein surface more heavily, since on average the geometry of these parts is more responsible for functional properties of the

protein than the geometry of the occluded parts of the surface. Sixth, we must decide whether our similarity measure should calculate a RMSD, or arrive at its score using another similarity measure.

RMSD is by no means the only way to score similarity, and there is no consensus on what the best method is, but RMSD does have the advantage of being computationally very convenient. To offer an example of an alternative measure of similarity, below the formula for RMSD is compared with a different similarity measure used by the structure comparison software STRUCTAL. Note that RMSD is actually a dissimilarity measure (the more dissimilar the two structures being compared, the higher value it gives, so that in practice we'd want to take our measure of similarity to be 1/x, where x is the value output by RMSD). STRUCTUAL's measure, on the other hand, gives higher values when the two proteins being compared are more similar.

$$\min_T \sqrt{\frac{1}{|\sigma(T)|}\sum_{i\in\sigma(T)}\left(\|a_i - T(b_i)\|^2\right)}$$

RMSD dissimilarity measure

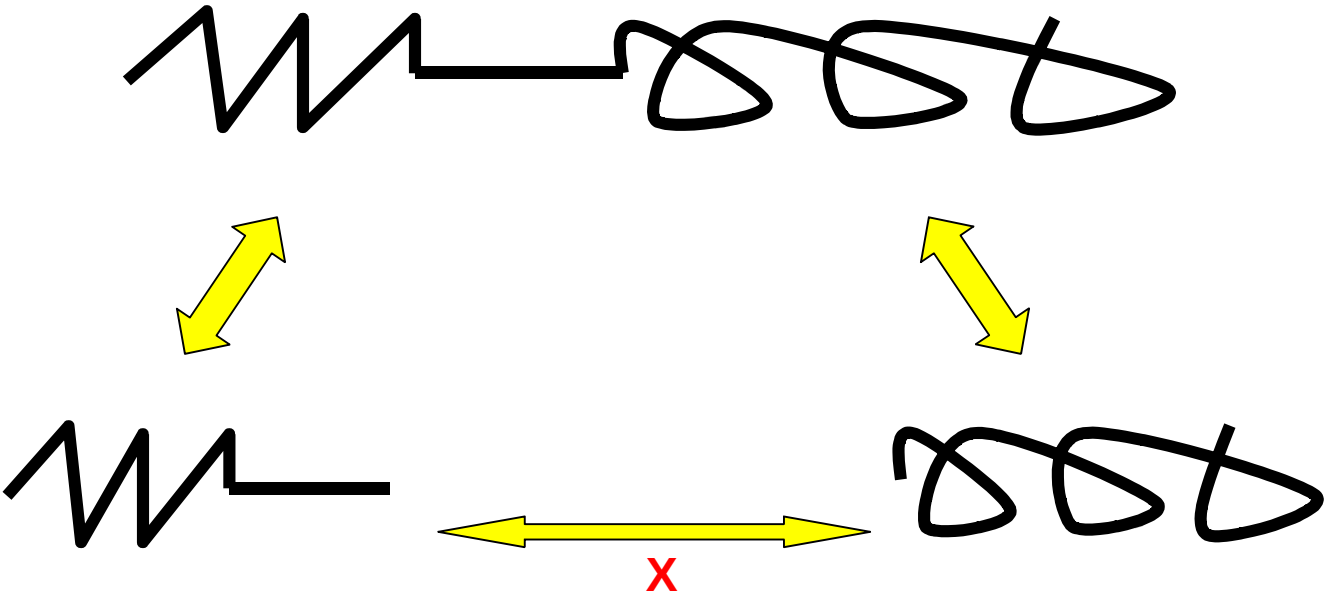
$$\max_T \left[\sum_{i\in\sigma(T)} \frac{A}{1+\left(\frac{\|a_i - T(b_i)\|}{B}\right)^2} - NGAP/2\right]$$

STRUCTAL's similarity measure

As the above discussion of the myriad issues the designer of a partial similarity measure must consider suggests, there are many ways to design such a measure. See **A.C.M. May. Toward more meaningful hierarchical classification of amino acids scoring functions. Protein Engineering, 12:707-712, 1999** for a review 37 different protein structure similarity measures.

Computationally assessing protein structure similarity is a difficult problem. The difficulty can be seen as reflection of the fact that measuring partial similarity is an ill-posed problem; there are many ways in which two 3D structures can be similar, and depending on the application of interest, similarities between certain aspects of geometric structure may be of more interest than others. The fact that there is no single way of deciding which aspects of structure to give importance to in choosing a quantitative measure of structural similarity accounts for much of the difficulty of comparing proteins according to structure.

Whatever our choice of similarity measures, though, it is not likely to define a metric on protein structures; we cannot expect the triangle equality to be satisfied, as the picture below illustrates more clearly than words ever could.



It turns out that with respect to all partial similarity measures of interest, finding an optimal partial match between two proteins (i.e. a choice of supports, a correspondence between them, and a transformation aligning corresponding parts of the supports) is NP-Hard. Thus we must be satisfied with approximate/heuristic solutions to the problem. There is probably not a single best solution to computing partial matchings; rather, specific algorithms are best suited to specific applications. But even so, there are general algorithmic principles that hold across different application areas.

To close this section, we'll mention that though we often are interested in using a similarity measure more sophisticated than RMSD in computing a partial matching of two proteins, one useful method is to compute a preliminary approximate matching using RMSD and then adjust the computed transform to maximize the score of the more sophisticated similarity measure. Methods for computing similarity will be discussed in more detail next lecture.

Applications of Structure Similarity Analysis

Though all structural similarity algorithms have a similar goal at their core, there are several different particular applications in biology today that call for somewhat different approaches.

Problem #1: Matching of Protein Structures
Given two molecules A and B, we seek substructures between A and B as large as possible while at the same time being "similar" (typically measured in RMSD).

Though the problem is stated as comparing one molecule to another, often this algorithm is used in one-to-many searches for similarity. For example, given a particular molecule, we might want to search all known proteins in the Protein Data Bank (PDB) for similarities. Or, we may want to group proteins in the PDB by doing many-to-many comparisons and clustering based on similarities.

Problem #2: Protein Classification

Besides finding similar substructures, proteins can be compared by their overall structure, i.e. classifying proteins into a hierarchy to determine similarities. Traditionally, these classifications are done manually with the aid of some automated tools, and take into account information that biologists have on the function and origin of the proteins. An example of this is the Structural Classification of Proteins (SCOP) database. It is felt that the SCOP database does a better job of classifying proteins according to structure than automated methods have been able to thus far. However, as the number of known structures is growing rapidly, we are approaching the point where the number of new structures will be to many to be classified by hand, so good automatic methods for structure comparison and identification are becoming increasingly important.

Several automated classifiers have been designed, among them are CATH (Class, Architecture, Topology, and Homologous superfamily) and FSSP (Families of Structurally Similar Proteins). As an example of how these work, the CATH protein hierarchy separates proteins at level 1 by "class" (i.e. whether the protein contains only alpha helices or beta strands or both), at level 2 by "architecture" (the gross orientation of secondary structures, currently done manually), at level 3 by "topology" (the connections between and numbers of secondary structures), and at the lowest level by "homologous superfamilies" (which takes into account structural and functional similarities between proteins).

One thing to note is that hierarchies obtained by automatic methods may be quite different from classifications designed manually because of the additional depth of knowledge that biologists have in relating proteins. Also, the splitting in the hierarchies is determined by our set of known proteins and so might by biased because there are only some proteins that we have currently been able to crystallize (i.e. determine the positions of the atoms in the molecule).

Problem #3: Finding Motif in Protein Structure

This problem aims to determine whether a motif, consisting of a small collection of atoms, matches anywhere in a very large protein. Note that the pieces of the motif are not necessarily connected, so we may not be able to constrain the search to consecutive atoms in the protein.

Often times it is difficult to isolate such a small motif, so we augment our search by using feature "types" (i.e. requiring that we have matches between types of atoms or between types groups of atoms as well as between locations of atoms). This dramatically simplifies our problem since there are much fewer candidate sites in the protein that matches that combination of atom types.

Problem #4: Finding Pharmacaphore in Ligands
A ligand is a molecule that binds to another molecule to form a larger compound. For proteins, this can have the effect of inhibiting the proteins function or catalyzing its activities. Therefore, ligands are important in drug design.
Given a set of ligands that are known to have the same activity (i.e. they all have the same effect on a protein or bind to the same site), we would like to find a substructure common to all the ligands (a pharmacaphore). Ligands are typically flexible molecules, meaning they might be in one of several conformations when they bind to the protein. Thus, for each ligand we give a set of low-energy conformations (which are more likely to react with the protein) and require that the pharmacaphore exist in at least one conformation for each ligand.

This problem is one of the key problems in drug design: if we observe that a set of ligands produce the desired activity, solving this problem will hopefully elucidate the essence of the interacting substructure and allow us to design better drugs.

Problem #5: Search for Ligands Containing a Pharmacaphore
This problem is related to the previous, but now we are already given the pharmacaphore and would like to find all the ligands in a database that contain it. Pharmaceutical companies typically have databases of 100,000s of flexible ligands and some of their low-energy conformations. By searching for pharmacaphores with known interaction properties with a protein, we can potentially find ligands that are better. This process gives chemists a better starting point for trying to improve drugs.

For your reference, here is a list of existing software for computing partial matches between protein structures.
C_α atoms:
DALI [Holm and Sander, 1993]
STRUCTAL [Gerstein and Levitt, 1996]
MINAREA [Falicov and Cohen, 1996]
CE [Shindyalov and Bourne, 1998]
ProtDex [Aung,Fu and Tan, 2003]

Secondary structure elements and C_α atoms:
VAST [Gibrat et al., 1996]
LOCK [Singh and Brutlag, 1996]
3dSEARCH [Singh and Brutlag, 1999]