

# Evolutionary Protein Structure Comparison

1.12.11

Comparing homologous objects are central to biology and the last decade have been dominated by this in the field of comparative genomics. However, many other objects than sequences in biology are homologous and can be subject to evolutionary study. Examples could be networks and structures. Protein structures have been compared for decades in a non-statistical and non-evolutionary way.

There are many principles, principles and domains of applications in this field. Are the amino acids of the structures prepared [ie aligned] or not. If they are the comparison is much easier. Can the structures be morphed or are they rigid? The latter makes the problem easier, but less realistic. Are we comparing the complete structures or search for similar sub-structures? In sequence alignment, if  $i$  is aligned to  $i'$  and  $j > i$ , then  $j$  cannot be aligned to  $j' < i'$  because we only allow insertion-deletions as position changing evolutionary events. But on billion year time scale, this limitation might not be warranted. Do we have more than 2 structures to be compared? Successful comparison has many applications, but given the very large evolutionary distances, it is also very hard and often unsuccessful.

## Major Questions/Principles

Prepared - Not Prepared  
Rigid - Flexible  
Global - Local  
Order - not order preserving  
Pairwise-Multiple  
Significance of Similarity

## Domains of Applications

Homology of Structures  
Evolutionary Relationship  
Comparative Modelling  
Consensus Structures.  
Measure of Structure Prediction

Recent years have seen a very large growth in the number of determined structure, collected in databases such as SCOP, CATH and PDB. The problem has also recently caught the attention of statisticians [Green and Mardia, 2006; and several papers by Scott Schmidler], although these models yet have to be based on an evolutionary model. But overall there is reason to believe statistical analysis of structures will get more attention in coming years.

A serious biological criticism against alignment algorithms is that they are biologically structureless - all positions are treated equivalently and no biological knowledge is incorporated into the algorithm.

McLachlan (1972) proposed a method to compare the structure of two proteins, for cases where it is clear which aa should be compared with which aa.s. Their distance measure was root mean square (RMS). The points of specific gravity of the two molecules had to be superimposed and then one molecule could be rotated so RMS was minimised. For comparison of more distant structures this method fails on two accounts: i. It supposes total rigidity of the molecule. It assumes a clear matching between the amino acids, that doesn't exist in reality. Many methods have subsequently improved upon this. We only mentioned a few:

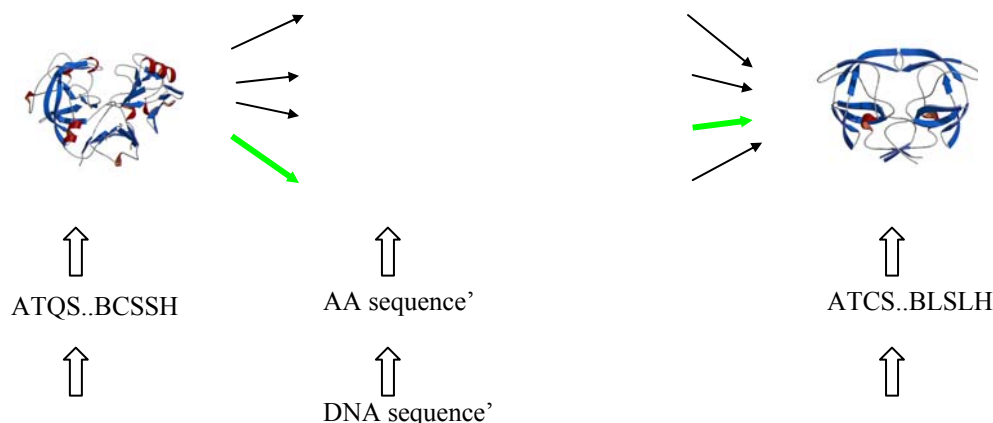
Orengo and Taylor (1989) encoded structural information about each amino acid in a vector with information such as relationship to neighboring amino acids, involvement in hydrogen bonding, area accessible to solvent and more. Each sequence was then a sequence of such vectors that could be aligned by a sequence aligning algorithm, with a net result of a structural alignment.

Holm and Sander (1994) compared not the physical structures themselves, but their distance maps. For a protein  $L$  long, the  $L*(L-1)/2$  internal distances are calculated. Obviously, identical (or mirrored!) proteins result in identical distance maps and similar structures should result in similar distance maps. How to measure the similarity of such maps is not straight forward because there is no prepairing between amino acids. But if two molecules shared a substructure, it should be possible to find a set of amino acids in both that had the same distance maps. There are a lot of subsets of amino acids to be searched for similar distance maps. Holm and Sander did not assume that the substructures matched had to be colinear (colinear: if  $S1$  is before  $S2$  in protein 1 and  $S1$  is matched to  $S1'$ , then  $S2$  must be matched to a  $S2'$  that comes after  $S1'$  in protein 2). This can be an advantage if the proteins have experienced other events than single element mutations and insertion-deletions.

Comparison would benefit tremendously by methods that modelled the evolution over time of structures explicit. The motivation for doing this has increased due to the large set of known structures presently, but has not materialized due to lack of models and computational power. Fortunately, such models have been explored extensively both in statistics and also in Molecular Dynamics (MD). Molecular Dynamics can simulate the behaviour of molecular systems with up to  $10^6$  atoms (typically  $10^3$ - $10^4$  atoms) and for time periods of up to a microsecond ( $10^{-6}$  s) dependent on details. Applications often involve dynamic paths, where both start configuration and end configuration of the system is known - for instance the catalysis of a substrate into a product. This is a well studied problem in statistics and the natural algorithms are now being used large scale in MD under the name Transition Path Sampling (TSP) (Bolhuis et al, 2002). The modelling problem described here is very similar to the TPS problem and can be explored using the same algorithm.

Evolutionary comparison clearly is ambitious, but is worth pursuing as it represents the optimal way of studying protein evolution. It is ambitious since it involves investigate *all* possible path between two protein structures. And for each path it involves predicting *all* protein structures on the steps of the path. There are methods to do this, but great care must be taken to be efficient in computations. Both paths and structures will have great redundancy allowing for reuse of calculations. Additionally, protein structures can be represented at different levels of detail, with the coarsest being representing secondary structure elements (SSEs) as labelled sticks with relationships to other SSEs, and the (almost) finest level being a full atomic representation of the structure. Making the correct

representational choice is crucial in making an interesting analysis in finite time and at the same time not having trivialized the problem.



The basic problem of evolutionary structure analysis of two homologous proteins is simple to formulate: A DNA gene sequence is translated to a protein that folds up in a structure. In evolving one structure into another, the corresponding genes are changed by substitution and insertion-deletions from the gene behind the first structure to the gene behind the second structure. Each gene on this unobservable evolutionary path has been translated into a protein that folded into a structure. There are methods that can sum over all paths from one sequence to another if we ignore structure. However, if structure is considered each step on this evolutionary path will be influenced by the quality/fitness of the structures. In this problem cannot be address without assigning some fitness to structures.

#### Possible Contents of Presentation/Report:

- **Introduction**
- **History of Structure Comparison: Data and Different Methods**
- **Evolutionary Analysis of Protein Structure**
- **Details of Models**
- **Discussion and Possible further research**

**Comment:** The presentation/report should have a strong bias towards recent statistical models, evolutionary models and evolutionary data analysis. This is mainly because this is where the field is moving and there is such an enormous number of earlier methods that a long report [like the Brown et al (1996) paper] could be written only dedicated to work up to 2000, which would not lead to exciting new ideas.

#### References

- Bolhuis et al. (2002) "TRANSITION PATH SAMPLING" An. Rev. of PhysChem Vol. 53: 291-318
- Brown, N., C.Orengo and Taylor (1996) "A Protein Structure Comparison Methodology" Computers Chem. 20:359-380.
- Chothia, and Lesk, (1986) The relationship between the divergence of sequence and structure in proteins. EMBO J., 5, 823-826.
- Green and Mardia (2006) Bayesian alignment using hierarchical models, with applications in protein bioinformatics. Biometrika, 93, pp. 235-254
- Hasegawa and Holm (2009) "Advances and pitfalls of protein structural alignment" Current Opinion in Structural Biology 19:341-348
- Holm and Sander (1996) Mapping the Protein Universe Science Vol. 273 no. 5275 pp. 595-602
- McLachlan AD: A mathematical procedure for superimposing atomic coordinates of proteins. Acta Crys A 1972, 28(6):656-657
- Meyerguz, Kleinberg, and Elber (2007) "The network of sequence flow between protein structures" PNAS 104.28 **11627-11632**
- Murzin et al. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540
- Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997). "CATH--a hierarchic classification of protein domain structures". *Structure* 5 (8): 1093-1108.
- Rodriguez A, Schmidler SC. Bayesian Protein Structure Alignment. Submitted to *Annals of Applied Statistics*
- Schmidler SC, Liu JL, Brutlag DB. Bayesian Modeling of Non-local Interactions in Protein Sequences: Prediction of  $\beta$ -Sheets. Submitted to *Journal of Computational Biology*
- Schmidler SC (2004). *Bayesian shape matching and protein structure alignment*. In *Bioinformatics, Images, and Wavelets*, edited by Akroyd, RG, Barber, S, and Mardia KV. *Proceedings of LASR 2004*, Leeds Univeristy Press, Leeds, UK
- Taylor and Orengo (1989) Protein structure alignment. *J.Mol.Biol.*208.1.1-22.