

# *Evolutionary Docking of Proteins -* *Computational prediction of functional interaction networks*

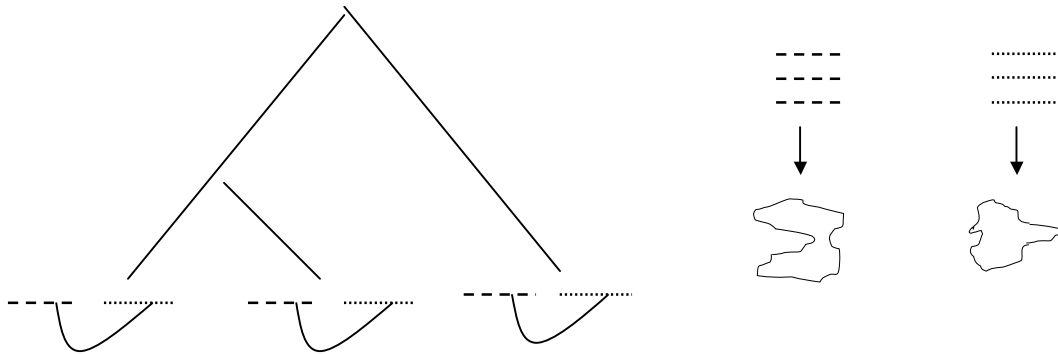
22.5.2007

The coupled evolution of pairs of positions is the cornerstone in the prediction of RNA secondary structure (Knudsen and Hein, 1999) and thus the position of RNA genes in DNA sequences. There have been similar attempts to use such dependencies to predict which protein residues were physically close (Abe and Mamitsuka, 1993; Chiang et al. 2006, Pollock et al., 1999; Bickel et al., 2002). However, these attempts have clearly been much less successful in protein than in RNA. Nevertheless, success of this could be very useful and we here propose a new version of this problem that differs from old versions in two aspects: Firstly, the application is the interaction between known structures. In this case focussed towards structures predicted from structural genomics projects (<http://www.sgc.ox.ac.uk/>). Secondly, it is coupled to another predictive method namely docking (Brooijmans and Kuntz, 2003) that by geometric or physical characterization tries to find complementary regions in structures that are likely to interact. This last facet allows a ranking of the search space (all possible pairs of residues) by using the structural knowledge to immediately discard sets of interactions that are physically impossible.

The project will need following components:

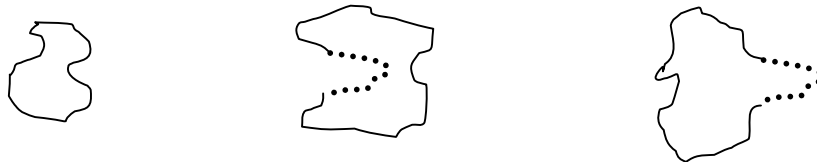
- A set of structures of interest
- For each pair of structures a large set of protein sequences
- A method for find physically complementary regions in structures

If successful the method will predict positions that interact among pairs of structures. Clearly the success depends on positions actually co-evolving and having sequences enough to actually detect this. To avoid serious loss of statistical power it is important to avoid a too-many test problem. This is solved by coupling the evolutionary predictions to a docking method. This both prevents the investigations of unlikely interactions and allows simultaneous testing of positions that are likely to be part of the same complementary surfaces.



Left: We have a pair of proteins where sequences are known for a series of species (3 in this illustration) and we want to find positions (connected by curve) correlated evolution. For real data sets it is advantageous if 10-100 variants are known for each pair. The total number of proteins could be hundreds to thousands and all pairs would have to be considered.

Right: For each structure the corresponding sequences can be aligned to each other and to the structure.



Docking will find complementary surfaces here illustrated with dotted lines. This can be purely geometric or also consider physical-chemical properties. The former is clearly simpler, but both are computationally hard.

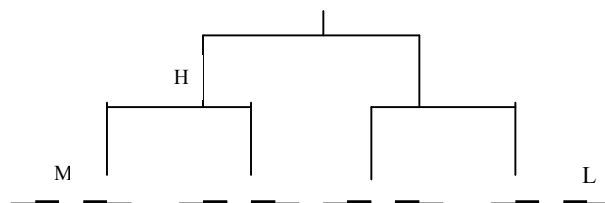
The over all structure of this problem is:

- a. for ALL pairs of proteins
- b. for ALL pairs of substructures
- c. Answer the question: Do the substructures evolve correlated?

- a. implies pairwise structure and sequence alignments. There are a series of programs doing combined Structure and Sequence Alignment, probably starting with Taylor et al. from 1994. These are publically available.
- b. implies protein-protein docking which again can be done by a series of publically available programs. Either a ranking or a cutt-off will have to be chosen.
- c. Models of evolution are almost all continous time markov chains and in most cases based on individual elements – nucleotides, amino acids or codons. The model is then characterized by a rate matrix,  $Q$ , that for amino acids has dimension  $20 \times 20$ . The transition probabilities for can then be calculated by  $e^{tQ}$ . When evolution is correlated then a model is needed on pairs of amino acids, which in principle is a  $400 \times 400$  matrix. Both the single and double amino acids models avoid a full parametrization, since this will lead to an impossibly large number of parameters. The single amino acid model uses an empirical rate matrix that has estimated individual entries from a large analysis of existing databases (Li, 1997). For RNA di-nucleotide models (Kirby et al., 1995) have been made from single nucleotide models by simple combination of single nucleotide models and adding a parameter that diminishes rates away from base pairing and increases events that restores base pairing. A similar approach can be adopted for proteins by a parameter,  $\lambda$ , that multiplied on entries creating pairs that can interact and is divided to entries that destroys such a pair.

### Simulation sub-project:

Motivated by the fact that finding correlations in protein structure evolution has been so much harder than in proteins, we suggest to start the total project with some simulations that should give some feel for how strong correlations should be to be detected and would result in software that will be needed later. Let  $N$  be the number of proteins (say 100),  $L$  their lengths (say 500),  $M$  the number of matching residue pairs between two structures that interact (say 8),  $K$  the number of sequences available for each structure (say 16),  $H$  branch lengths in phylogeny relating the sequences (say 0.2 expected event per position) and finally let  $R$  be the parameter describing the strength of the interaction (say 0.2). The sub-project has all the components needed in the real data analysis, except docking.



Here we have 4 sequences for each of the 2 protein structures and the  $K$  correlated positions are shown as a thick interval of the sequences each  $L$  long. The tree is bifurcating and each edge represents  $H$  expected events duration.

**Application to real data:** Structures can be found and we suggest that we apply analysis to 3 sets of data of increasing difficulty. First a pair of protein structures that are known to interact. Second to a set where some pairs are known to interact. Third to a set of newly determined structures about which nothing is known, which could for instance be the ones determined within the last year by local and international structural genomics projects.

### 2 month pilot project:

Week 1: Read references below. The books should only be used for reference.

Week 2: Plan the remaining period in detail. During this week also expand the present project description to 3000 words.

Week 3-4: Simulation sub-project

Week 5-6: Structure-Sequence Alignments, Docking and Search for correlated positions.

Week 7-8: Application of simulation to specific hypothesis and final report write-up. Report writing is encouraged to be done continuously during the period.

### References

- Abe and Mamitsuka (1997) "Predicting protein secondary structure based on stochastic tree grammars" *Machine Learn.* 29:275-301.
- Bickel P et al. (2002) "Finding important sites in protein sequences" *PNAS* 99:23:14764-71
- Brooijmans N and I Kuntz (2003) "Molecular Recognition and Docking Algorithms" *Annu.Rev.Biophys.Biomol.Struct.* 32:335-73.
- Chiang, Joshi and Searls (2006) "Grammatical Representation of Macromolecular Structure" *J. Compu. Biol.* 13.5. 1077-100.
- Kirby DA, S Muse and S Wolfgang (1995) "Maintainance of pre-mRNA secondary structure by epistatic selection" *PNAS* 92:9047-51.
- Knudsen, B. and J.J.Hein (1999) "Using stochastic context free grammars and molecular evolution to predict RNA secondary structure" *Bioinformatics* vol. 15.5 15.6:446-454
- Li, W-H. (1997) "Molecular Evolution" Sinuaer
- Pollock, Goldman and Taylor, WR (1999) "Coevolving Protein Residues: Maximum Likelihood Identification and Relationship to Structure" *J.Mol.Biol.*287:187-198
- Taylor WR, Flores TP, Orengo CA. (1994). Multiple protein structure alignment. *Protein Sci* 3(10):1858-70
- Amnon Horovitz's, Nir Ben-Tal's and Richard Aldrich's, Alfonso Valencia labs
- Lockless & co: *Science* 286:295-299, *Nat.Str.Biol.*, 10:59-69, *Nature*, 437:512-518