Determining protein function from comparative genome analysis

Dissertation submitted by: **Parag Ghosh**

Date: 16.11.01

Introduction:

Biologists of this decade are confronted with the problem of synthesizing the enormous amount of data accumulated from genome sequences, into some useful knowledge. It is estimated that around 40% of the open reading frames in a fully sequenced organism have no known function at the biochemical level and are unrelated to any known gene. Consequently, a shift of emphasis is now occurring from genome mapping and sequencing to determination of genome function. This is the area known as functional genomics.

Functional genomics offers the key to integrating DNA sequence information with multiple disciplines such as drug discovery, environmental monitoring, cancer, aging, evolution and many more. Since the impact of mutations depends on the context in which the genes exist, understanding genomics is pivotal to making rational judgements about risks and consequences of mutations. As a result, much research is now targeting the identification of genes and mutations and the dynamic processes that lead to their expression as proteins. Functional genomics rely on both biochemical experimentation and computational methods to determine the function of proteins.

Aim:

This essay aims at identifying proteins that participate in a functional pathway. The underlying assumption is that proteins that function together in a pathway or structural complex are likely to preserved together or eliminated together in organisms during the process of evolution. This property of correlated evolution is studied here by characterizing each protein by its phylogenetic profile, a string that encodes the presence or absence of a protein in every genome. This method of phylogenetic profile not only brings out the functional correlations between proteins but also helps us to predict the function of uncharacterized proteins.

Method:

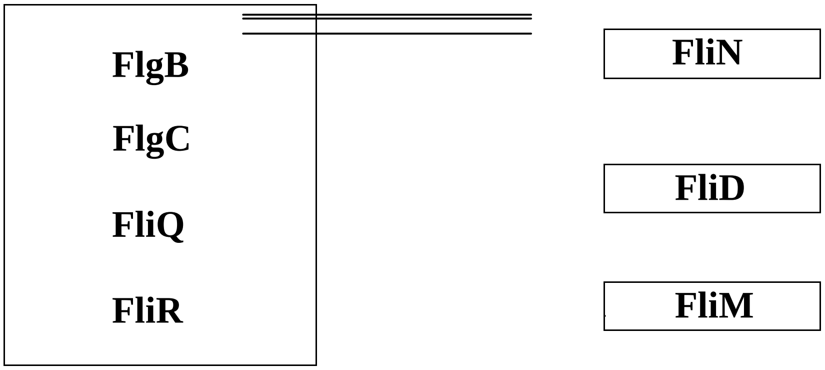
One of the computational methods used for establishing functional linkages between proteins is the method of phylogenetic profile. A phylogenetic profile describes the presence or absence of a particular protein across a set of organisms whose genomes have been sequenced. If two proteins have the same phylogenetic profile in all surveyed genomes, it is inferred that the two proteins have a functional link.

A heuristic argument can be given to support the claim: Each test on the presence or absence of a protein can yield two possible outcomes. Now if there are *n* fully sequenced genomes then there ought to be 2^n phylogenetic profiles. Currently there are about 30 fully sequenced genomes, meaning there ought to be 2^{30} possible phylogenetic profiles for a protein's phylogenetic profile to be a unique characterization of its distribution among genomes. Now this number far exceeds the number of protein families indicating that not all the outcomes we have considered are allowed. Here lies the idea of correlated evolution.

To outline the method of constructing phylogenetic profiles, I have selected 6 proteins from the flagellar apparatus of E.Coli. For each protein I used the **Swiss-Prot** database to search for homologues in different organisms. The organisms I have studied are: Aquifex aeolicus, Bacillus subtilis, Borrelia burgdorferi (Lyme disease spirochete), Buchnera aphidicola (subsp. Acyrthosiphon pisum) (Acyrthosiphon pisum symbiotic bacterium) and Salmonella typhimurium. The presence or absence of each protein is indicated by a 1 or 0, respectively. Since we have considered 5 genomes here, so each profile is a string 5 bits long. These profiles are then compared bit by bit and all proteins having the same profile are clustered together in one box. Profiles differing by 1 bit have been connected by lines in the figure.

Phylogenetic profiles of 7 proteins of E.Coli

	FlgB	FlgC	FliD	FliN	FliQ	FliM	FliR
Aquifex	1	1	1	1	1	0	1
Bacillus	1	1	1	0	1	1	1
Borrelia	1	1	1	1	1	1	1
Buchneria	1	1	0	1	1	1	1
Salmonella	1	1	1	1	1	1	1



Functional correlations obtained between the 7 proteins

(Profiles differing by 1 bit are connected by lines)

I have also listed the functions of some of the proteins as are listed in the **Swiss-Prot** Database:

FliR: Role in flagellar biosynthesis.

FliQ: Required for the assembly of the rivet at the earliest stage of flagellar biosynthesis.

FlgB: Role in flagellar biosynthesis.

FlgC: Role in flagellar biosynthesis.

FliD: Required for the morphogenesis and for the elongation of the flagellar filament by facilitating polymerization of the flagellin monomers at the tip of growing filament. Forms a capping structure, which prevents flagellin subunits (transported through the central channel of the flagellum) from leaking out without polymerization at the distal end.

FliM: FliM is one of three proteins (FliG, FliN, FliM) that form a switch complex that is proposed to be located at the base of the basal body. This complex interacts with the chey and chez chemotaxis proteins, in addition to contacting components of the motor that determine the direction of flagellar rotation.

FliN: FliN is one of three proteins (FliG, FliN, FliM) that form a switch complex that is proposed to be located at the base of the basal body. This complex interacts with the chey and chez chemotaxis proteins, in addition to contacting components of the motor that determine the direction of flagellar rotation.

Results:

As can be seen by comparing the functions of these proteins, the first 4 take part in flagellar biosynthesis and are functionally linked. This also what we obtain by comparing their phylogenetic profiles. The other proteins considered here are closely related to the first 4 proteins (all of them being flagellar proteins) and hence differ only by 1 bit from the profiles of the first 4 proteins.

The results obtained also closely match the results obtained from a much rigorous and detailed analysis by Pellegrini *et al.* (2)

The fact that the proposed scheme works out pretty well has been verified by Eisenberg *et al.* (1) by a statistical test. The method is called “keyword recovery” where one compares the keyword annotations for both members of each pair of proteins linked by one of the methods. This is possible only where both members of the pair have known functions. When the keywords for both members agree, it is called ‘keyword recovery’.

Eisenberg *et al.* (1) compared the signal-to-noise ratio of keyword recovery for different sets of yeast proteins and compared them with the experimental values. It is seen that the method of phylogenetic profiles have fair reliability in general and excellent reliability when two or more of these methods agree on a link.

Conclusion:

The phylogenetic profile of a protein describes the presence or absence of homologues in organisms. Proteins that make up structural complexes have identical profiles. Also proteins that take part in a metabolic pathway are likely to have similar profiles. Hence by studying the phylogenetic profiles one can not only obtain the functional linkages between proteins but also predict the functions of uncharacterized proteins.

References:

1. Eisenberg D., Marcotte E.M., Xenarios I. and Yeates T.O. (2000) *Nature* **405**, 823-826
2. Pellegrini M., Marcotte E.M., Thompson M.J., Eisenberg D. and Yeates T.O. (1999) *Proc. Natl. Acad. Sci. USA* **96**, 4285-4288
3. Gaasterland T. and Regan M.A. (1998) *Microb. Comp. Genomics* **3**, 177-192
4. Altshul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. and Lipman D.J. (1997) *Nucleic Acid Res.* **25,** 3389-3402
5. **Swiss-Prot:** http://www.expasy.ch/sprot/sprot-top.html
6. http://www.tigr.org