

CS 262 – Computational Genomics
 Prof. Serafim Batzoglou
Functional Genomics & Network Integration
 March 14, 2006 –Guest Lecture: Balaji S. Srinivasan
 Scribe: Mike Polcari

What we've learned this semester - sequencing & aligning genomic data – has a unifying goal. That is: to use the genomic data as a platform for biological analysis. This lecture covers the following topics:

Functional Genomics: How can we use methods of functional genomics to assay individual proteins or make inferences about pairs of proteins?

Network Integration: How can we stack multiple data-sets together in systematic ways, to discover?

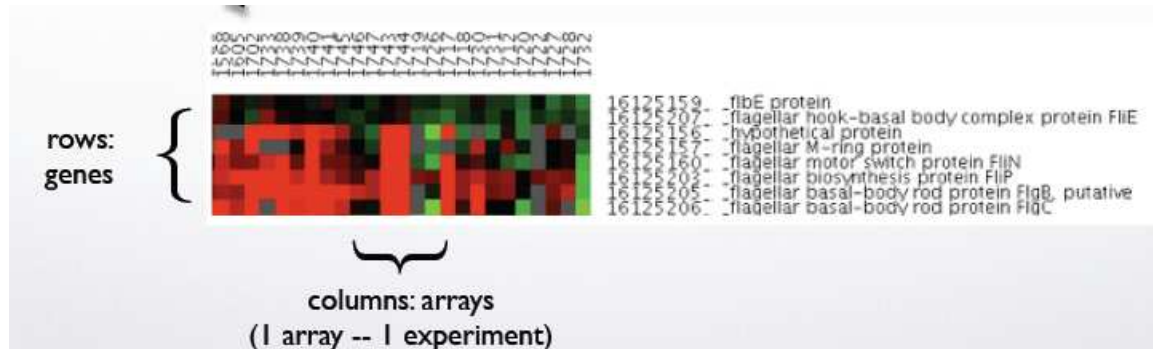
Further Directions

Functional genomics covers:

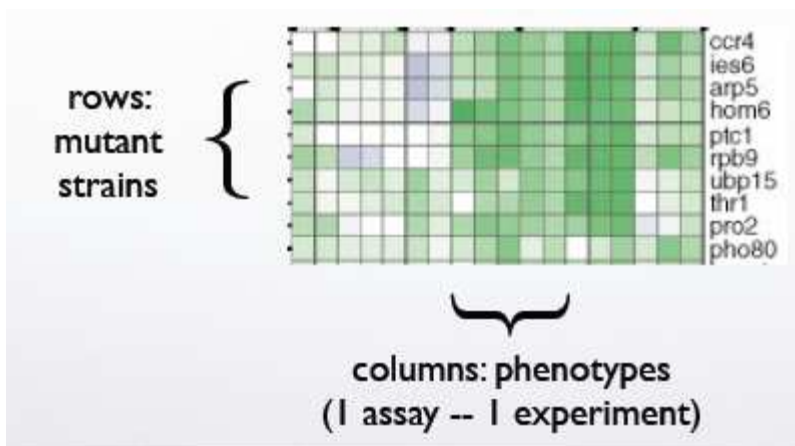
- determining function on a genomic scale. This has only been possible in last 10 years with whole genomes
- large scale assays for DNA, RNA, protein properties
- The “*omics*” (proteomics,metabolomics, glycomics) are very fashionable

Examples:

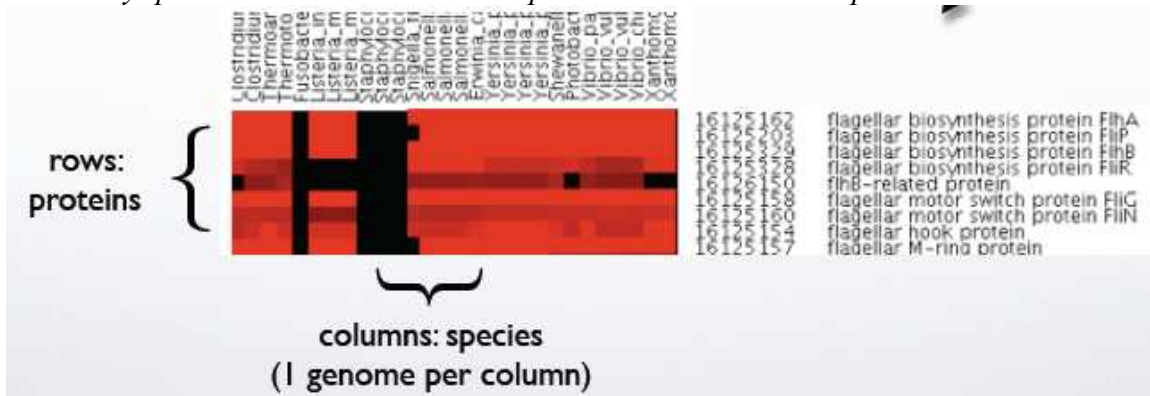
Micro-arrays – study different times, conditions etc in each experiment and cluster genes based on their reaction.



Pheno-arrays – take mutant strains and see how they survive –this can tell us: what is the function of the underlying deleted gene?



Taxonomy- presence or absence across species can tell us what a protein does.



Each of these techniques can yield answers to particular questions:

When is a particular gene expressed?

What phenotype does a particular mutant exhibit?

Which species have a particular protein?

However, what if we want to discover *system level* info? To discover system level info, we need to discover pair wise similarities. *ie* How does the above data co-vary with other genes/proteins/etc? Which genes co-express with a particular gene?

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

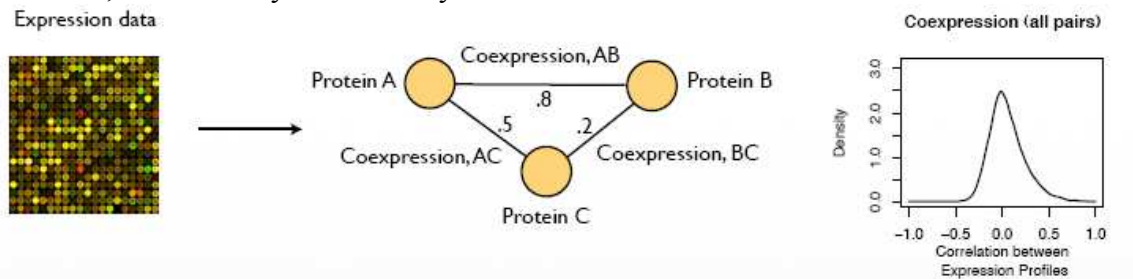
System level info is clearly very interesting but is (as usual) computationally demanding. For a typical microbe with 3700 genes, there are 6.9×10^6 pairs of genes.

Even with just pairwise correlations from just one dataset, we can discover revealing information. For example, by correlating gene-expression time during the cell-cycle, we

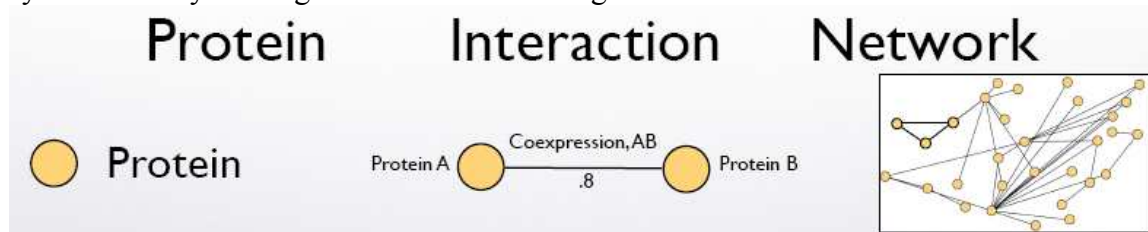
can identify different modules & group them into hierarchy of function. Imagine what we could discover if we used multiple data sources!
 Functional genomics can provide both gene level and system level info, how can we systematically combine them for even more-revealing discoveries?

Network Integration

When co-expression & co-inheritance tell us different things about which proteins should be linked, we need a systematic way to resolve & combine results from different datasets.

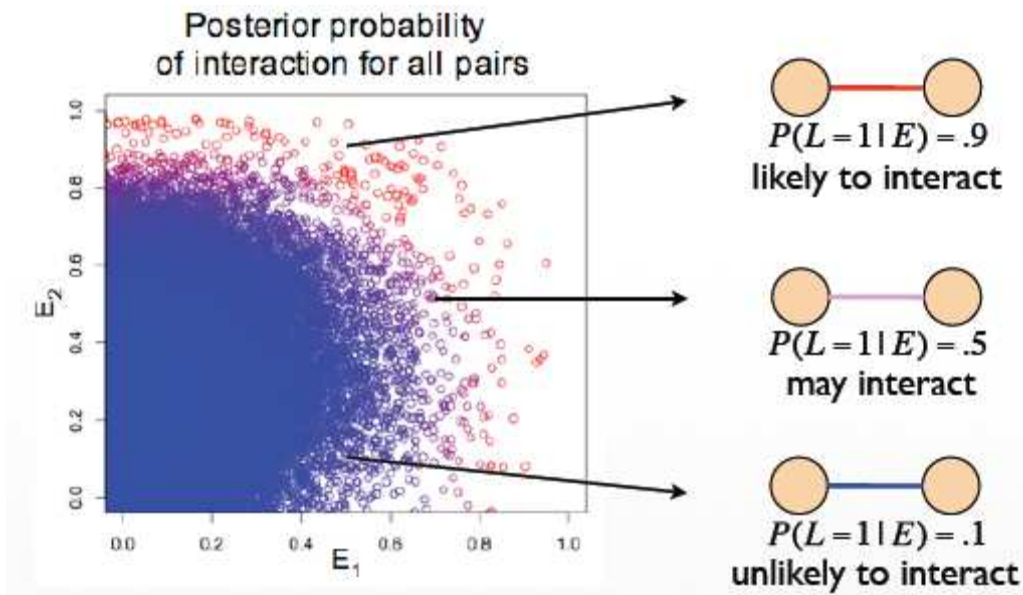


Functional genomics provides dozens of ways to discover covariance – we need a systematic way to integrate all of these findings:

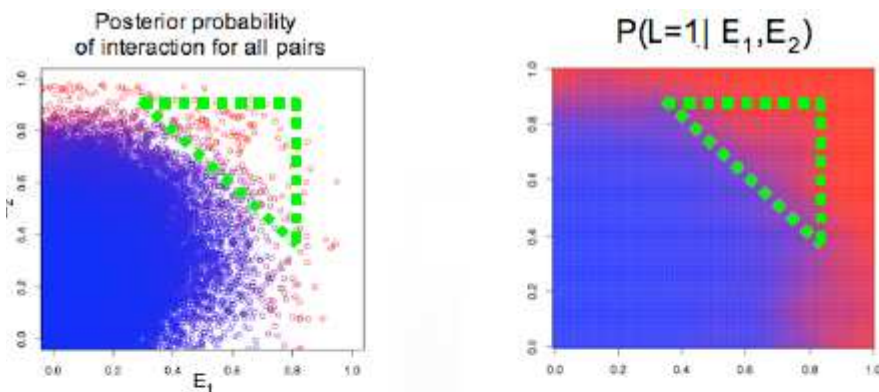


We can create many protein interaction networks – each corresponding to the results from a functional genomics experiment. Each of these networks are on completely different scales – how do we combine them? *Training sets*

We assay different methods on known true/false interactions. By measuring these known covariances & how they relate to functional linkage, we can determine higher-dimensional decision rules for categorizing proteins as functionally linked. We can take another step back and view this as a classical machine-learning problem where the dozens of functional genomics readings represent dozens of dimensions of the input data & a binary classification Machine Learning algorithm (an SVM, for example) is used to classify proteins as either functionally linked or not.



Network integration can provide insight into ‘hidden biology’ - linkages that would not be evident if either of the data-dimensions is used in isolation:



$$P(L = 1 | E_1, E_2) > .5$$

$$P(L = 1 | E_1) < .5 \quad P(L = 1 | E_2) < .5$$

Joint density reveals **hidden biology**

Subtle interactions missed if data analyzed in isolation

Moderate evidence from multiple sources is often better than strong evidence from one source

30-60% of interactions fall in this region!

Scope & Scale – it scales to very large datasets:

Applied to 230 sequenced microbes

3-8 dimensions

~4.89 billion possible protein pairs, but only 2.81 million predicted interactions.

Example of an application:

We can make interesting predictions about *Caulobacter crescentus* – in particular, what are the interaction partners of mreB ?

When looking at this through our network browser, we can see strong linkages to proteins we'd expect (which is encouraging). More interestingly, we can see moderate linkages that would have been missed without network integration. One of these linkages was later confirmed!

The moral of the story: We can discover new & interesting relationships!

Now that we have hundreds of interactive networks – what can we do with them?

1) Network alignment: Can we systematically compare interaction networks to find conserved components?

2) Can we suggest proteins within a species to assay in the lab?

Network alignment:

Sequence alignment is about finding conserved proteins – network alignment is about finding conserved modules – both conserved sequence & interaction. If we can discover something about the interaction or behavior of a module in a particular species, we can use network alignment to translate those findings to a forward hypothesis in novel organisms.

Protein recommender:

Use a partial set of proteins to recommend other proteins which are likely involved. These recommendations are then tested experimentally.