Data fusion for disease gene identification (and a brief primer about gene networks)

Victor Hanson-Smith
CIS607 Winter 2009
The central dogma of biology.

Some regions of DNA are coding-regions, which encode for a gene.

**DNA strand:** ATCBATTABCATCG...TTGCTAGCTATTACAGCATCACATCTAATCGATCGCTTCTAGC

A type of protein called **transcriptase** transcribes coding-regions of DNA into RNA.

**RNA:** AUGC...UUGC

The ribosome (which is a polymer of proteins) translates messenger RNA (mRNA) into proteins.
Protein molecules are diverse!

The human body is estimated to contain 25,000 different types of proteins. Here are a few of my favorite proteins, which don’t necessarily appear in the human body:

**Glycoprotein antifreeze protein in ice fish**
http://www.pnas.org/content/94/8/3811.abstract

**Steroid-hormone receptor protein in vertebrates:**
http://www.sciencemag.org/cgi/content/abstract/312/5770/97

**Opsin visual pigment protein is archaosaurs:**
http://mb.e.oxfordjournals.org/cgi/content/abstract/19/9/1483
Some regions of DNA are promoter regions, enhancer regions, or silencer regions.

When certain types of protein bind to promoter or enhancer regions, the related coding-region is able to be transcribed.

When proteins bind to silencer regions, the related coding-region will not be transcribed.

DNA strand: ATC......TABCATCG. . . TGTAGCTATT. .........GCATCACATCTAATCGATCGCTTCTAGC
The big picture: proteins interact in a network

The downstream phenotype is determined by a complex interaction of proteins promoting, enhancing, and/or silencing the expression of other proteins. For example, consider the circadian rhythm in mammals (below): heat, food, and light directly affect only a few genes, but these genes affect many other genes.

Microarray analysis measures the expression levels of RNA

By comparing the different RNA expression levels between populations, we can potentially identify which genes are responsible for the differences between the populations.

Here, researchers compare the RNA expression levels between obese and non-obese populations. They identify a set of genes which might be responsible for obesity.

These putative genes are called quantitative trait loci, or **QTL**.

One challenge with QTL/microarray analysis is identifying the root gene which causes a disease

Although QTL identifies several genes which might be the cause of some condition (i.e. a disease), some of these genes might actually be downstream in the gene-interaction network.
By using data fusion, Aerts et al. are able to better identify which genes cause a disease versus the genes which are simply related in the network.
Their approach...
See also...
fin.