Lecture 18
Quantitative Analysis of Regulatory Variation

Bruce Walsh. jbwalsh@u.arizona.edu. University of Arizona.

ECOL 519A, April 2007. University of Arizona

GENE REGULATION IS A COMPLEX TRAIT

Quantitative-genetic approaches are often assumed to be restricted to phenotypic characters such as body weight, height, or some measure of shape. However, they apply equally well to molecular characters, such as mRNA and protein abundance. With the recent advent of functional genomics tools (such as microarrays) the power of quantitative genetics is now being widely applied to genomics.

Genomics loosely refers to analysis of whole-genome sequence data (essentially a static analysis), while **functional genomics** is more concerned with the dynamics of how the genome and cell interact. This includes the analysis of the **transcriptome** (the entire collection of transcripts and their dynamics in the cell), the **proteome** (the dynamics of the entire collection of proteins), and the **metabolome** (the dynamics of metabolism in the cell). When studying how a trait evolves, the holy grail is to be able to predict phenotypic change solely from DNA sequence change. Given that we still have only a minor understanding of the full flavor of development, we are a long way from this goal. Our first intermediate step is a better understanding of gene regulation.

QTLs Involved in Protein Regulation

As we saw in Lecture 7, molecular evolution is largely decoupled from phenotypic evolution, with frogs showing twice the sequence divergence, but only a fraction of the phenotypic divergence, of mammals. Given that most evolutionary biologists believe that changes in **gene regulation**, as opposed to **structural changes** (i.e., protein sequence changes), underlie many phenotypic changes, this decoupling is perhaps not unexpected. The era of genomics has now given us tools to look at certain features of gene regulation across the entire genome, and hence the quantitative analysis of regulatory variation is currently a very active research area.

Even before genomics, there were QTL mapping studies examining the genetic control of variation in gene regulation. In particular, Damerval et al. (1994) analyzed the spot volumes of 72 anonymous proteins (from a specific seed tissue in maize) separated by high-resolution 2-D polyacrylamide gel electrophoresis. Genes controlling protein volume are, by definition, **regulatory genes** influencing the amount of that protein. Sixty F_2 individuals were scored with 76 RFLP markers, and both ML-interval mapping and single-marker ANOVA detected a total of 70 QTLs affecting 46 of the 72 proteins. Of these 46 proteins, 25 were influenced by two or more QTLs (up to a maximum of five). Of the 70 detected QTLs, 33 showed strict additivity, while the remaining 37 showed at least some dominance. The amount of variation in protein volume accounted for by a single QTL ranged from 16% (the lower detection limit for this sample size) to 67%, and the cumulative variation accounted for by all detected QTLs for each protein ranged from 37% to 90%. Perhaps the most striking observation was the presence of significant epistasis. Four proteins had QTLs that were only detected through epistasis (their single-locus effects were not significant). In all, 14% of the 72 proteins showed detectable epistasis.

With the explosive growth of microarray analysis, a number of studies are starting to look at the entire transcriptome of an organism, generating a high-dimensional **gene expression phenotype**. This allows us to do several interesting things. The first is a refined approach to search for genes influencing the trait. Potentially more interesting, the trait of gene expression itself can be widely examined. Is it heritable? How many **eQTLs** (for **expression QTLs**; the term **ECE**, **Expression**

Control Element, is also used) underlying typical changes? How is within- and between-population variation in expression connected (if at all)? Are there any connections between the rate of divergence of gene expression versus the rate of protein sequence divergence? All of these are questions that we are starting to answer. Before examining these, we review some basic features of microarrays.

MICROARRAYS

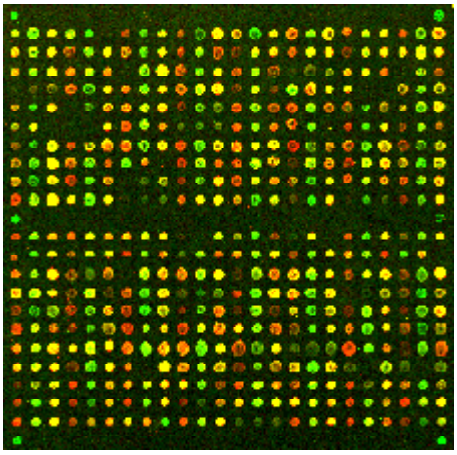

Figure 18.1 A **spotted microarray**. Here, each spot represents a fairly short cDNA sequence corresponding to a gene. mRNA from one treatment is treated to fluoresce in green, while mRNA from the other treatment will fluoresce in red. Yellow spots indicate roughly equal expression in the two treatments.

A Brief Overview Of The Technology

The basic idea behind microarrays is *hybridization*: one extracts mRNA from a collection of cells of interest and then hybridizes it to a series of probes for target genes of interest. Each individual hybridization reaction occurs on an array (Figure 18.1), and is referred to as a **spot** or a **feature**. A typical array may contain thousands of spots. In theory, at one extreme all the roughly 30,000 genes within a mammalian genome can be spotted onto a single microarray. However, given that the spot limit for a typical array is around this number, this would result in only a single replicate for each gene, extremely poor experimental design. Good experimental design technique should be practiced by including multiple technical replicates of each gene on the same slide. In this case scoring an entire genome requires several microarrays. The (sample-adjusted) hybridization levels for any particular gene are then compared across two (or more) different samples (**treatments**), which could be different cell types, time points, or individuals. In this fashion, any set of genes of interest, up to the entire genome, can be simultaneously compared for differences in relative expression across the treatments of interest.

Two different approaches have been used for generating the probes in an array. With **synthetic oligonucleotide arrays**, one uses sequence information to chemically synthesize the probe sequence (oligonucleotide), often directly onto the slide/chip/membrane, for example, by using photolithography, as is the approach of **Affymetrix** and **Agilent chips**. The other approach is a **spotted cDNA array**, which uses PCR to generate cDNA probes which are then spotted / printed on the array using a high-precision robot. We will focus on spotted microarrays, although our comments equally apply to synthetic oligonucleotide arrays.

Following extraction, mRNA from two treatments under comparison are reverse-transcribed into cDNA for hybridization on the arrays, using two different fluorescent dyes (Cy3 and Cy5) to mark the cDNA from the two treatments. The cDNA generated from one treatment fluoresces green

and the cDNA from the other fluoresces red under different wavelengths of light. The total cDNA from both samples is then run over the array, resulting in a grid of small green, yellow, and red spots (Figure 18.1). Genes with a yellow spot (feature) indicate roughly equal levels of mRNA in both samples for that feature, while those that appear green or red have the majority of mRNA from just one sample. Formally, the levels of fluorescent intensity are measured on both fluorescent wavelengths (or **channels**) for the two dyes. A digital image is produced for each channel and an estimate of the relative intensities of both colors is generated for each spot.

Analysis of Microarray Data

As one might imagine, there are a number of image-processing and standardization issues in correcting both samples to allow for a proper comparison. Parmigiani et al. (2003) provides an excellent review of the various issues, which we will regard as being (largely) solved for our discussion, although this is still an area of active research.

After the preparation and image processing stages, the critical issue of assessing statistical significance of observed differences in expression arises. Shockingly, many of the initial microarrays were run with absolutely no replication. While the molecular technology was at the very cutting edge, the initial experimental designs were certainly stone-age. It is now fairly widely appreciated that at least some replication must occur. An ideal design involves multiple spots for the same gene within an array (i.e., **technical replication**), and several replicates of each complete array representing different biological samples (i.e., **biological replication**). Further, it has been shown that there are **dye-gene interactions**, so that one can obtain different levels of expression for a gene simply from which dye is used. As a result, a **dye-swapping design**, or a **loop design** (Kerr and Churchill, 2001), is good practice, where one replicates the comparisons swapping the allocation of the Cy3 and Cy5 dyes over the two treatments.

The resulting data can easily be handled as simple linear models not unlike those used by plant breeders in multi-regional field testing of new cultivators (e.g., Gauch 1992) . For example, consider gene i (the index i may run into the thousands) from a sample of type j (the treatment, for simplicity we consider two types). In a well-designed experiment, there are multiple arrays and replicate spots of each gene within an array. The resulting level of expression y_{ijkl} for replicate l of gene i in treatment j on array k is just

$$y_{ijkl} = u + A_k + R_{l(k)} + T_j + G_i + TG_{ij} + e_{ijkl} \tag{18.1}$$

where y is the log of the spot intensity. The A and R terms control for array and location within array effects, T is the average treatment effect (the average level of mRNA over all genes in that treatment), and G the average gene effect (the average level of that gene over all treatments). What is of interest are the within gene TG contrast terms, corresponding to the gene by treatment interaction, indicating changes in the level of expression of the gene over treatments.

Significance of the TG interaction term for any particular gene is using examined using standard t-tests, which are sometimes modified by using an adjusted pooled variance. The issue of multiple comparisons can be tricky. A strict Bonferroni bound for an experiment-wide significance of α would require each TG tested using a p value of α/n (this bound incorrectly assumes each test = gene is independent). Since the number of genes (comparisons) n is typically in the thousands, this gives a very strict bound, and hence very poor power. Using the **false discovery rate** (or **FDR**) provide a nice solution to this problem. The reason for doing a microarray experiment in the first place is that one expects a number (perhaps a very sizable fraction) of the tested genes to differ. Hence, the more appropriate error control we wish is not the probability of a false positive given a null (p), but rather the probability q of a null *given a significant test*. The motivation for this approach was introduced in Lecture 3 in our discussion of the posterior error rate. Thus, if our set of (say) 600 genes has a false discovery rate of $q = 0.05$, then only $600 \cdot 0.05 = 30$ of these are expected to be false discoveries.

Example 8.1. As example of the advantage of using false discovery rates, consider Storey and Tibshirani 's (2003) analysis of a microarray data set from Hedenfalk et al. comparing BRCA1- and BRCA2 mutation positive tumors. A total of 3,226 genes were examined. Setting a critical p value (for any particular test) of 0.001 detects 51 significant genes (i.e., those with differential expression between the two types of tumors). Assuming the hypotheses being tested are independent (which is unlikely as expression can be highly correlated across sets of genes), the probability of at least one false positive is $1 - (1 - .0001)^{3226} = 0.96$, while the expected number of false-positives is 0.001· 3226 = 3.2, or 6% of the declared significant differences.

Setting a FDR rate of $\delta = 0.05$, Storey and Tibshirani detected 160 genes showing significant differences in expression. Of these 160, 8 (5%) are expected to be false-positives. Notice that, compared to the Bonferroni correction (51 genes, 6% false positives), over three times as many genes are detected, with a lower FDR rate. Further, Storey and Tibshirani estimate the fraction of nulls (genes with no difference in expression) at 67%, so that 33% (or roughly 1000 of the 3226 genes) are likely differentially expressed.

To contrast the distinction between p and q values, consider the MSH2 gene, which has q value of 0.013 and p value of 5.50×10^{-5}. This p value implies that the probability of seeing at least this level of difference in expression given the null hypothesis (no difference in expression) is 5.50×10^{-5}. Conversely, $q = 0.013$ implies that 1.3% of genes that show differences in expression that are as or more extreme (i.e., whose p values are at least as small) as that for MSH2 are false positives.

The final, and perhaps most vexing analysis issue, is that of clustering and classification. **Clustering**, detecting groups of co-expressed genes (in an attempt to draw some inference about the underlying regulatory networks or response mechanisms) is largely done in an ad-hoc fashion. For example, one standard approach is **k-means clustering** (Tavazoie et al., 1999) wherein one specifies the number of clusters k in advance, after which least-squares, or other approaches, are used to find the optimal classifiers for the clusters. Given that a typical microarray experiment involves measurement of hundreds to thousands of genes with very few (typically far less than a dozen) replicates, the very nature of these data will result in clusters. This additional noise complicates attempts to deduce any underlying biological clustering. Other common approaches to clustering are self organizing maps (SOMs, Tamayo et al., 1999), model based clustering (Yeung et al., 2001), and hierarchical clustering (Eisen et al., 1998).

The related problem of **classification**, finding those genes at which changes in mRNA expression level predict phenotype, rests on a far firmer statistical foundation. For example, one could use logistic regression to estimate those genes contributing the most information in classifying the identity of a binary phenotype (Lee et al., 2003). Classification is the issue that is of more immediate concern to a breeder, as we would like to use microarrays to find genes of interest for selection and their subsequent incorporation into elite lines.

Microarray Analysis Is Best Regarded As An EDA Approach

Given that the number of tests of significance (for non-zero gene by treatment interactions) greatly exceeds the number of data vectors, there are very serious issues with multiple tests. A further complication is that we expect many of the tests to be correlated as gene expression, by its very nature, is often highly correlated. Standard approaches for controlling the experiment-wide p value are often extremely overly-conservative, mostly ignoring the problem of highly correlated tests, although Benjamini and Yekutieli (2001) address the treatment of dependence in multiple hypothesis testing. Further, the reason for performing a microarray experiment is that we indeed expect a large number, but not large in relation to the total number of contrasts, of the T X G interactions (the TG terms in Equation 18.1) to be significant.

A more reasonable approach is to consider a microarray experiment as an exploratory data analysis tool, with the goal of using the results from one experiment to produce a reduced set of

genes for future consideration. For a chip testing 10,000 genes, extracting a set of 200 potential candidates is a very significant reduction. In such a setting, instead of controlling the experimental-wide p value, we should instead aim to control either the false discovery rate or proportion of false positives, (Storey and Tibshirani, 2003a,b; Fernando et al., 2004).

Problems (and Pitfalls) of Gene Discovery via Microarray Analysis

With a complete genome sequence in hand, arrays can be constructed to monitor levels of mRNA expression at conceivably every protein-coding gene. Given that the majority of genes discovered in genomic sequence scans are of unknown function, microarrays obviously offer a powerful approach for detecting potential candidates genes. A suitably designed microarray experiment can generate a list of genes whose expression is significantly different in high vs. low lines (or individuals), over different environments, or in different tissues. This is an extremely exciting opportunity and hence it is not surprising that biologists have been very interested in the potential of microarrays.

But just how informative are microarrays for suggesting candidate loci? There are several critical limitations. The first concern is that the wrong treatments may be considered. The difference in some trait may be due to differential expression of genes in very different tissues from the target tissue in which the trait is apparently expressed. A good (conceptual) example is to think of grain yield in plants. An exhaustive microarray study focusing on flower and seed expression would completely miss changes in expression in the roots that might result in more resources for the plant, and hence greater yield. One can similarly imagine comparable situations in animal systems.

A second caveat is that microarrays simply score mRNA levels, and other levels for the control of gene expression may result in the differences between trait values, for example post-mRNA processing or post-translation processing of proteins. Microarray systems that use a series of small non-overlapping probes within each gene (e.g., Affymetrix chips) may catch differences in mRNA processing. Spotted cDNA or long oligo microarrays, on the other hand, use much larger probes, and hence a variety of differentially-processed mRNAs from the same gene may yield the same expression signal. Probe size is critical as with a longer probe (70-100 nucleotides for long oligo arrays and >1,000 nucleotides for cDNA arrays), a single-base pair change typically will not result in a change in hybridization pattern. In such cases, two alleles showing the same level of mRNA expression, but producing different protein products can still show the same amount of hybridization, and hence be missed by a microarray screen of candidate loci based on differential expression.

A far greater complication is that an observed change in the expression level at gene X may actually be caused by a **trans-acting** factor from gene Y. Indeed, it appears that the majority of observed changes in experiments using recombinant inbred lines of mice for QTL mapping of microarray changes, wherein the expression level at each gene is treated as the trait of interest for mapping, are due to trans-acting factors (David Threadgill, personal communication). Analysis of such experiments often do show QTLs for expression level in a particular gene mapping to that gene itself (a **cis-acting** control factor). However, there are very often QTLs for expression that map to very different locations from the target gene (trans-acting factors). An especially interesting class of trans-acting factors that appear in joint QTL-microarray analyses are so-called **master controller** or **global regulators**, wherein a single genomic location influences expression levels for a very large number of genes.

An interesting graphical way to display information from a joint QTL-microarray experiment, a **waterfall plot** (Figure 18.2), is given by Pomp et al. (2004). Here, one plots the genomic location for the gene whose expression is being considered on one axis and genomic locations for any QTLs for expression of this gene on the other axis. Points on the diagonal indicate cis-acting factors, presumably either within the gene itself or very tightly linked to it. Off-diagonal elements indicate trans-acting factors, wherein the gene influencing levels of expression at the target is some distance from the target. Master-control genes are indicated by a either a horizontal or vertical run of spots, depending on whether QTL position is plotted on the horizontal or vertical axes (respectively).

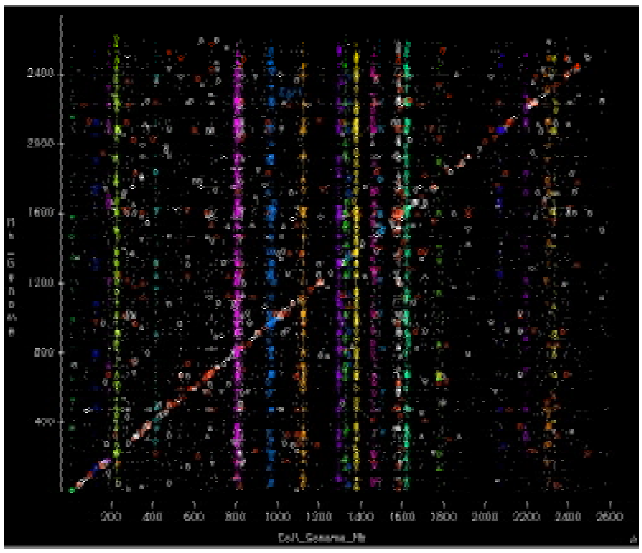

Figure 18.2: A **waterfall plot**. Here the horizontal axis represents gene location for an eQTL, while the vertical location represents the target for that eQTL. There are two notable features. First, the *diagonal line* represents *cis-acting* eQTLs, those that act on their own gene. However, note the second major feature, the vertical lines. These present single *trans-acting* eQTLs that influence a large number of genes.

GENERAL PATTERNS OF TRANSCRIPTIONAL VARIATION

While the description of genome-wide expression patterns is still in its infancy, some generalizations appear to be emerging. It, of course, remains to be seen over the next several years if these turn out to be general features. When considering the conservation (or lack thereof) of expression, there are three different features that can change: **level** (amount) of the transcript, **breadth** (the tissue distribution), and **temporal** (the timing of the transcription). Most evolutionary studies to date have focused largely on the amount, followed by the breadth. While developmental biologists have focused on termporal changes, evolutionary studies are running behind on this dimension.

Gene Expression Levels are Typically Highly Heritable

Before considering expression levels as potential quantitative trait, we must first ask if they are heritable, especially when considering a snapshot of the amount of mRNA for several thousand genes at once. The good news is that between-individual variation in expression appears not only to be heritable, but often highly heritable. Brem et al. (2002) looked at roughly 6200 genes in a cross between a wild and a lab population of yeast. Over 1500 genes were differentially expressed, and for many of these genes the expression difference segregated in yeast haploid spores (classic evidence of genetic differences). A follow-up study (Brem and Kruglyak 2005) of 5700 of the original genes detected at least one eQTL at almost half (2,984) of these. Further, variation was highly heritable, with over half (3,546) of the genes having broad-sense heritability exceeding 70% for variance in expression level. Looking over maize, mice, and humans, Schadt et. al. (2003) found that 40% of genes had at least one eQTL, and 3% had more than three. Cheunh et al (2003) looked at five highly variable genes in three human groups: 49 unrelated individuals, full sibs, and monozygotic twins. Genes showed less variability in expression in more closely related individuals, again evidence of genetic variation underlying variation in expression. Finally, Morely et al. (2004) followed 3,554 human genes, and for 1000 of these, significant evidence of linkage to specific chromosomal regions was seen.

Correlations Between Rates of Regulatory (Transcriptional) and Sequence Divergence

As we have seen (Lecture 7), genes vary in their amount of sequence conservation. An obvious question is whether there are any correlations between expression and sequence conservation. Several such correlations have been observed:

• *Codon usage bias tracks level of expression in many, but not all, organisms*. In unicellular organisms (such as bacteria and yeast), strong **codon usage bias** is seen, wherein the frequency of the different synonymous codons for a particular amino acid are not what is expected from the frequency of nucleotides in the gene. For example, for 4-way (all third bases give the same amino acid) synonymous codons, we might expect codons ending with A, T, C, G will simply track the frequency of these nucleotides. Such is not the case. Rather, it is generally the case that one or two codons are much more frequent than expected by chance. The model proposed to account for this (which has considerable support) is that the most frequent codons for a particular amino acid correspond to the most frequent tRNAs for that amino acid, speeding up translation. One common explanation for strong codon bias in uni-cellular organisms, and weaker (often much, much weaker) biases in multi-cellular organisms is that a site only experiences selection if $4N_e|\,s\,| >> 1$. Thus, humans and yeast might have the same intrinsic s value for codon selection, but our effective population sizes are too small for most of these changes to be effectively under selection, resulting in a much smaller codon bias. Strong codon bias is seen in bacteria, yeast, and *Drosophila*. In mammals, codon usage bias is very small, often not significant. For example, Duret and Mouchiround (2000) found that mammalian silent substitution rates do not vary with expression pattern, even in widely expressed genes. Conversely, for humans Cheunh et al. (2003) found a (small) positive correlation between gene expression level and codon usage bias (at synonymous sites). In *Arabidopsis*, Wright et al. (2004) found no correlation between expression level and K_s, but a significant negative correlation between expression level and K_A.

• *Genes that are expressed in more tissues are more conserved (show less divergence) compared to genes that are expressed in one or a few tissues.* In a wide survey, Subramanian and Kumar (2004) found (repeating the findings of numerous others) that highly expressed genes evolve slowly, but genes with low expression levels show a large variation in protein sequence divergence.

• *More highly expressed genes tend to evolve slower*. Pal et al. (2001) looked in yeast (which being unicellular has no tissues) and observed that the more highly expressed genes evolve slower.

• *Genes expressed later in development evolve more rapidly than those expressed earlier*. This feature was seen by Cutter and Ward (2005) in *C. elegans*, who also observed that genes expression transiently during embryogenesis evolve faster than other embryonic transcripts.

• *Mixed signals on correlations between expression profile divergence and sequence divergence*. While some studies did not find a correlation between expression profile divergence (expression in divergent tissues) and sequence divergence (Jordon et al 2005, Wagner 2000, Concant and Wagner 2004), others did (Gu et al 2002, Makove and Li 2003). Hence, while the *amount* of gene expression seems to be correlated with sequence conservation, the *distribution* of expression (across tissues) is weakly correlated, if at all.

Thus, there are several consistent correlations between expression and sequence conservation. It is unclear if expression is the causal factor imposing constraints on sequence divergence, or if other factors are involved the affect the conservation rates of both, such as common structural constraints.

Correlations Between Regulatory Divergence and Expression Level/Pattern

We now turn from correlations between expression patterns and sequence conservation to correlations between expression patterns and expression conservation. Here, the data is just starting to come in and some of the potentially associations are presently rather murky.

- *Within- and between-species variation in expression level is generally correlated*. Genes with the largest within-populations variation tend to show the greatest between-species differences. However, there are genes that show an unusually high within-species variance that does not translate into a high between-species variance. These findings were obtained by Lemos et al. (2005), who looked two closely related laboratory strains of mice, as well as human versus chimp, two species of mice, and two species of *Drosophila*. Khaitovich et al. (2004) also observed this pattern in several primates (human, chimp, oangutan, rhesus macaque).

- *The conservation of expression level of a gene in directly proportional to its range of tissue expression.* This has been seen in mammalian genes (human vs. mouse comparison, Yang et al. 2005; human-chimp by Khaitovich et al. 2005) and in *Arabidopsis* (*thaliana* vs. *lyrata*, Wright et al. 2004.) This observation runs counter to the logic that a tissue-specific gene is usually highly specific, and hence its expression is likely be conserved across species. Khaitovich et al. (2005) further note that widely expressed genes also tend to differ less among individuals than genes expressed in single tissues.

- *Different tissues can show different rates of expression level conservation*. Khaitovich et al. (2005) looked at gene expression in five tissues in humans and chimps, finding significant differences between tissues in conservation of expression level. The brain showed the least divergence, the liver the most. Genes expressed in only a single tissue show the highest expression and sequence divergence, genes expressed in all five tissues show the lowest divergence.

Does Divergence in Expression Follow a Neutral Model?

The neutral theory framework, which is the standard null model for sequence data (Lectures 7, 8), has also been proposed as a model for divergence in expression level (Khaitovich et al. 2004, 2005). Here the null model for strict neutrality (i.e., no removal of deleterious mutations) would be Brownian motion (Lecture 7) of expression level.

One general feature consistent with a neutral view is the observation that the within-population variance in gene expression is positively correlated with expression divergence between species. If one holds that, because of differing constraints, the neutral mutation rate (i.e., the neutral part of the mutational variance σ_m^2 for expression levels) varies over loci, we expect such a correlation as the within and between variance both scale with the mutation rate. The known relative rates of expression changes might be insightful here. Fay et al. (2004) found 5448 genes with expression changes in yeast, giving a ratio of 0.887 expression changes per synonymous substitution. This rate is much higher than that for non-synonymous substitutions (0.175) as well as the rate of inter-genic substitutions (0.291 per synonymous substitution). If one thinks in terms of constraints, expression changes are only slightly more constrained than synonymous sites, and far less constrained that coding or intergene regions.

Potentially more striking support was offered by Khaitovich et al. (2004), who looked at the expression of roughly 12,000 genes in six humans, three chimpanzees, one orangutan, and a rhesus macaque. First, they observed that the variance in gene expression increased linearly with time, as predicted under a Brownian motion model (Lecture 7). Second, they observed that rates of expression divergence between species do not differ significantly between intact genes and expressed pseudogenes.

Countering this, three studies have suggested evidence of stabilizing selection on the expression

levels of most genes. Two studies involve mutation accumulation experiments. Rifkin et al. (2005) estimated the mutational variance in gene expression for 12 lines of *D. melanogaster* that have been accumulating mutations for 200 generations. Significant expression variance as found at 39% of the genes. Estimates of mutational heritability had medium values of 2.5 x 10^{-5}. While they did find that genes with higher σ_m^2 values tend to have larger between species difference, the observed between-species divergence was far less than predicted under drift (given the estimated σ_m^2 values). Denver et al. (2005) compared levels of gene expression in mutation-accumulation (MA) lines versus natural isolates (NI) of *C. elegans*. They found that 9% (660/7014) of genes in MA lines had significant differential expression (across MA lines), while only 2% (118 of 5,588) of genes in NI lines did. Although the natural isolates have been diverging for significantly greater time periods than the MA lines, their variance in expression was considerably less. Denver et al. interpreted this observation as support for stabilizing selection on expression in the NI lines, as (by design) mutation accumulation lines have very low effective population sizes (typically less than 5), and hence weak selection (Lecture 7). Assuming the natural isolates are mutation-drift equilibrium, the ratio of the standing genetic variance in expression (estimated from the NI lines) to the mutation variance (estimated from the MA lines) should be (Lecture 7)

$$\frac{\sigma_G^2}{\sigma_m^2} = \frac{4N_e\sigma_m^2}{\sigma_m^2} = 4N_e \tag{18.2}$$

(The factor of four, instead of two, arises because *elegans* is largely selfing.) For all genes, this ratio was far below its neutral expectation (taking $N_e = 17,500$, which is likely a significant underestimate). Denver et al. also found evidence that most of the MA transcriptional differences were due primarily to a few trains-acting mutations with multiple down-stream effects.

The third study suggesting stabilizing selection was by Lemos et al. (2005), who looked two closely related laboratory strains of mice, as well as human versus chimp, two species of mice, and two species of *Drosophila*. Using standard tests for drift in a trait (Lecture 7), they took estimates (based on observed divergence) of $\sigma_m^2/\sigma_e^2 < 10^{-4}$ as evidence for stabilizing selection, while estimates of $\sigma_m^2/\sigma_e^2 > 10^{-2}$ were taken as evidence for directional selection. Under this criteria, the vast bulk (60% - 90%) of genes showed evidence for stabilizing selection, with 1% to 25% (depending on the species comparison) being compatible with drift, and 1% to 10% showing evidence of directional selection. Lemos et al. also suggest that the striking observation of Khaitovich et al. that pseudogenes show the same divergence in expression as functional genes could be misleading, as these so-called pseudogenes have been conserved over several million years and hence likely are functional.

It is perhaps not surprising that the pure drift model does not fit the data, as there are certainly constraints on expression. That σ_m^2 values estimated from lines with very small N_e do not fit the divergence data is also not surprising, as we have previously discussed (Lecture 7) that the key is the *neutral fraction* of σ_m^2, which is likely far less than the value estimated from MA lines (which are largely governed by drift). The data are certainly consistent with the view of deleterious mutation being removed by selection with drift occurring on the remaining fraction. Expression levels with more constraints have lower neutral mutations rates and hence diverge at slower rates.

ANALYSIS OF PATHWAYS

Microarrays provide a snapshot of the global pattern of mRNA expression, and while this is certainly quite helpful, it is by no means provides a definitive picture of a cell's regulatory network. A variety of other tools from molecular biology are available for looking at how gene products interact within a cell. One is two-hybrid screens (to be discussed shortly) for looking at protein-protein interactions. Another is the **2-D protein gel**, the protein equivalent of a microarray, giving a snapshot of all proteins within a cell and providing a rough quantification of their concentrations. Other approaches, such as **fluorescent energy resistance transfer** (**FRET**) and **fluorescence recovery after**

photobeaching (**FRAP**) have been developed for estimating how close two proteins come to each other in a cell (FRET) and for following individual protein products around in the living cell (FRAP). Coupled with microarrays, these and other genomics tools offer the very real promise of a far greater elucidation of genetic pathways than we current possess.

Two-Hybrid Screens: Construction Protein-Protein Interaction Maps

Two-hybrid screens, originally developed in yeast (Fields and Song, 1989), allow one to examine which proteins physically interact within the cell. In a two-hybrid screen, a reporter gene is expressed if and only if the two proteins of interest come into direct physical contact (Figure 18.3). In both yeast and the nematode, tests of all pair-wise interactions among all known genes have generated two-hybrid interaction maps, detailing which proteins appear to interact with one another within the cell (e.g. Wagner, 2000), Figure 18.5)

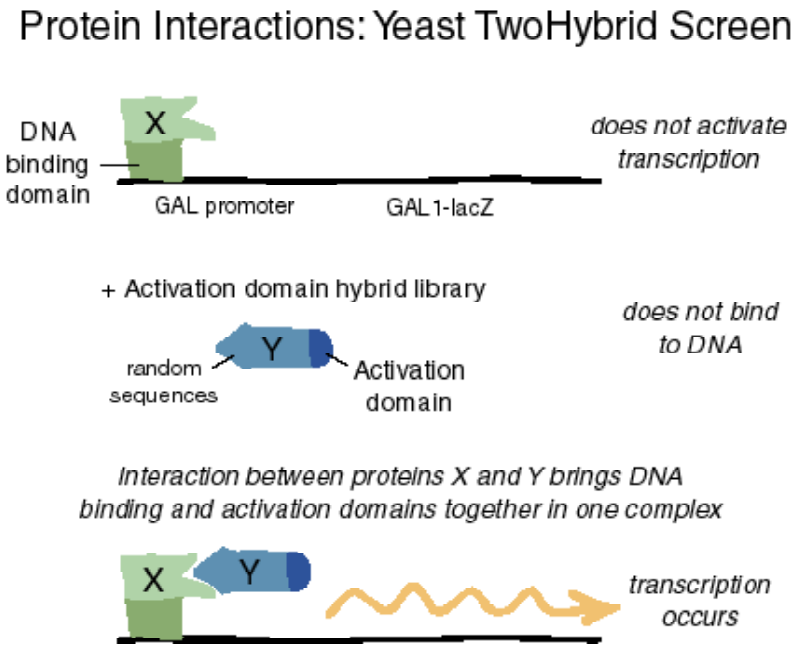

Figure 18.3 The yeast two-hybrid screen takes advantage of the nature of the GAL promoter, which consists of a DNA binding domain and an activation domain that occurs on a separate protein that must interact with the DNA binding protein for transcription to occur. Through genetic engineering, we can replace the interaction region on the binding protein and the activation domain with sequences from two target proteins of interest. If these two proteins physically interact, the activation domain binds to the DNA binding protein, and transcription occurs, in this case through a lac-z reporter gene. One can (very laboriously) use this approach to screen all potential combinations of proteins, allow construction of a protein-protein interaction map (Figure 18.5).

Flux and Pathways

A major quest among functional genomicists is to use the static snapshot of the entire genome offered by microarray and two-dimensional protein gels to at least partially reconstruct the topology of cellular networks. One standard approach is to follow expression over time, another to exploit perturbations in the mRNA or protein levels of particular genes to see how the network responds. While progress has been slow to date using such approaches, the inventiveness of molecular biolo-

gists will undoubtedly provide additional tools for estimation of network topologies. Key statistical issues of such a reconstruction are still a largely open question, although tools from phylogenetics, such as bootstrapping (Felsenstein, 1983), may prove useful.

Figure 18.4 Flux through a typical linear pathway, A through F represent products in the pathway, while e_1 to e_4 are the enzymes that move the products through the pathway.

Just what information can we exploit from a regulatory topology? Consider the simple linear network given in Figure 18.4. Suppose the ultimate goal is to increase the production of the product F. The **flux** of a pathway is the (steady state) rate at which a product is produced. The gene products e_1 through e_4 (typically proteins, although these could also be structural RNAs) move inputs through the pathway to produce F. If Figure 18.4 represents the topology of the gene network for the trait of interest in a model system, it immediately suggests strong candidate loci (the loci coding for e_1 through e_4) to test for association between variation at these loci and the trait of interest. Of course, as suggested from microarray studies, any number of other genes may influence the expression levels of the genes responsible for e_1 through e_4. We can go a step further and ask how best to increase the flux through the system, although this moves us from simple issues of the topology to the much more complex issue of the dynamics of the network. In particular, if we could engineer an up- (or down-) regulated gene in this pathway, which gene should be the target? One choice would be to up-regulate the gene(s) responsible for the production of e_1, front-loading the pathway. However, one might also argue that up-regulation of e_4 may have a greater impact on the flux. The point here is that while the topology is certainly useful, by itself it does not immediately suggest the rate-limiting step(s) in the pathway. Deduction of the key limiting steps in a pathway requires not only the topology, but additional information as well. Fortunately, there is a fairly large body of theoretical literature on the control of metabolism (e.g., Fell 1997, Hofmeyr and Westerhoff 2001) and the results from such metabolic control theory provide some very critical insights, and tools, for exploiting pathway information.

Kascer-Burns Sensitivity Analysis

In a landmark paper, Kacser and Burns (1973) developed the basic framework for the study of flux through biochemical pathways. Prior to Kacser and Burns, there was the widespread notion that most pathways were limited by one or more rate-limiting steps. As Figure 18.4 shows, the topology of a pathway tells us which substrates are transformed to the next in the chain to ultimately end up in the flux. The interesting question is which step is most rate-limiting in the pathway? Is it the initial step, with enzyme e_1 transforming A into B. Or is it the final step (enzyme e_4?) Or perhaps some other step? If we could genetically over-express any of these enzymes, which would have the greatest inapt on flux?

Kacser and Burns quantified the actual limitation (or equivalently, the regulation), imposed at each step in a pathway by introducing the concept of a **flux control coefficient** C_i^j for the flux at step i (F_i) in a pathway due to enzyme j (E_j),

$$C_i^j = \frac{\partial F_i / F_i}{\partial E_j / E_j} = \frac{\partial \ln F_i}{\partial \ln E_j} \tag{18.3}$$

Roughly speaking, C_i^j is the percent change in flux (through i) divided by percent change from a small change in the activity (or amount) of j.

Control coefficients provide a quantitative description of the critical control points within a pathway (and indeed, for any part of the pathway). The largest change in flux is generated by increasing the amount (or activity) of the element (gene, gene product, or protein) with the largest control coefficient. A rather remarkable feature of control coefficients is the **Kacser-Burns Flux Summation Theorem**, which states that the sum of the control coefficients over all elements contributing to the flux at step j sums to one, viz.,

$$\sum_i C_i^j = 1 \tag{18.4}$$

Since control coefficients are generally positive (see below), the total control of flux through a pathway is distributed over the components, and large C values (i.e., close to one) are expected to be rare. Rather, the maximal value of a control coefficient for any element within a pathway is likely modest at best, resulting in rate-limiting steps being rather rare. In general, the control of the flux through a pathway or system is shared by all members of that pathway or system. In a quantitative-genetics framework, genes of modest effect on flux are expected to be much more common than major genes (those with large control coefficients). If pathways are highly branched or negative control is exerted by one (or more) members within a network, negative flux coefficients can occur (wherein an increase in a gene product reduces the flux).

The second feature following from the summation theorem is that if a particular control coefficient is greatly increased in value, this decreases the values of other control coefficients in the system as flux control is shared by all members of the pathway. Hence, *control coefficients are not intrinsic properties of an enzyme (or gene), but rather properties of a (local) system*. This is much akin to the average effects of an allele, which are also population-dependent. The interested reader is pointed to Kacser and Burns (1981) for a description of the relationship between metabolic control theory and allelic forms. Likewise, control coefficients evolve, again akin to the average effect of an allele changing as the population evolves.

The control coefficients can be related to the factor by which flux can be increased by the **Small-Kacser Theorem** (1993): an r-fold increase in activity (or amount) of gene E results in an f-fold increase in the flux through element j, where

$$f = \frac{1}{1 - \dfrac{r-1}{r} C_E^j} \tag{18.5a}$$

In the limit of an overwhelming amount of E, the maximum that the flux is increases is just

$$f = \frac{1}{1 - C_E^j} \tag{18.5b}$$

Thus, even an infinite increase in the activity/amount of E only results in a 2-fold increase in flux if the control coefficient is 0.5 (which is a rather large value). A four-fold increase in flux requires a control coefficient of 0.75, and a 10-fold increase a control coefficient of 0.9.

Just how extendable metabolic control theory will prove to be for gene networks remains unclear, as it makes a key assumption that the system is in steady state. Non-steady states may turn out to be the norm in many gene networks, yet metabolic control theory has been shown valid for near steady states (Liao et al., 1997). Despite this (and other) limitations, something akin to metabolic control theory will be developed for dissecting gene networks, and analogs to control coefficients will likely be key components in such a theory.

REGULATORY NETWORKS AND GRAPH THEORY

One mathematical representation of a regulatory (or metabolic) network is given by a **graph** (Figure 18.5), a collection of **nodes** and **edges**. Nodes are the elements in the pathway (such as proteins,

genes, or metabolic products), while edges denote interactions between these. For example, if proteins A and B interact, we connect the A and B nodes with a line (an edge). A node with multiple edges means that the regulatory element interacts with a number of other elements. If we are looking at a metabolic network, the nodes may be substrates, and edges the enzymes that act upon then. Hence, the edges may themselves be biological entities.

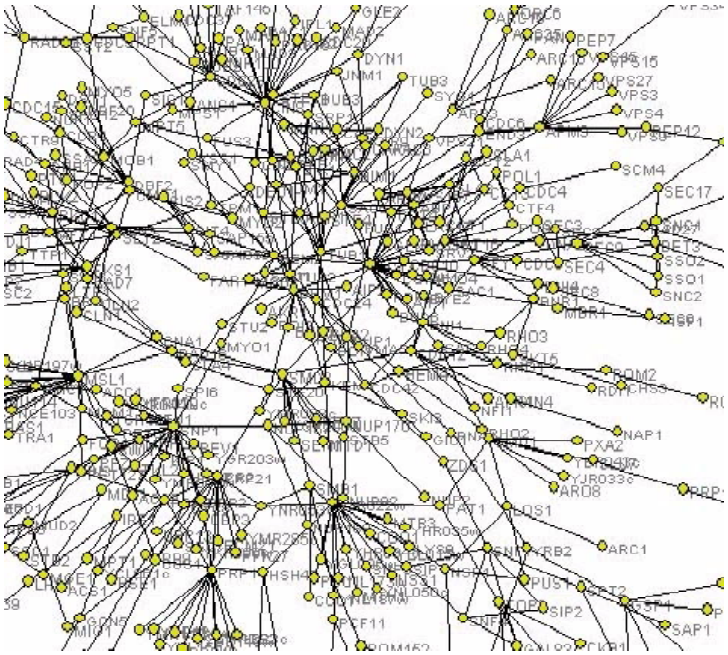

Figure 18.5. A section of the protein-protein interaction map in yeast, as deduced from two-hybrid screens.

A graph represents the basic **topology** of the network, while the actual **dynamics** (rate and types of interactions) is obviously overlaid upon this topology to give the full dynamical representation of a regulatory network. An intermediate step between the simple topology and the full-blown hyper-dimensional dynamical system is a **directed graph**, where an edge is replaced by a single- (or double-) headed arrow. A single-headed arrow pointing from node A to node B means that A influences B, while a double-headed arrow between the two indicates that the two elements influence each other.

The graph that forms the topology of a network can be expressed as a matrix. The topology matrix $\mathbf{M}$ contains only zeros and ones. If the ij-th element of $\mathbf{M}$ is one, this implies gene i influences gene j. Note that influences can be asymmetric, with $M_{ij} = 1$ while $M_{ji} = 0$. Non-zero elements in $\mathbf{M}^k$ denote elements that influence a gene after k-steps through a pathway. When the dynamic rules of a pathway are known, the simple matrix of zeros and ones is replaced by a vector-valued function, taking in the current state space at each gene and returning the new state of the system. Obviously, estimating the topology is the first (and easiest) step in completely describing a network.

Erdos-Renyi Random Graphs and Random Boolean Networks

The first step at understanding the topology of regulatory (or metabolic) networks is to consider the properties of **random** (or **Erdos-Renyi**) graphs. For such ER graphs, one starts with n of nodes, and then randomly assign edges to them (i.e., randomly connect the nodes), with p the probability

that two randomly-chosen nodes are connected. One central property of a graph is its **degree distribution** $P(k)$, the probability distribution that a given node is linked to k other nodes. Under ER graphs, $P(k)$ is Poisson,

$$P(k) = \frac{(z)^k}{k!} \exp(-z) \tag{18.6}$$

where z is the average number of nodes that a randomly-chosen node is connect to. For ER graphs, $z = np$.

A second property of a graph is how well connected the nodes are to each other, its **path length distribution**, the distribution of the minimum number of steps between any two nodes in the graph. For an ER graph, the average path length d is roughly given by

$$d_{ER} \sim \frac{\ln(n)}{\ln(z)} \tag{18.7}$$

A final important property of a graph is its **cluster coefficient** C, the probability that two neighbors of a give node are also neighbors of each other. With an ER graph,

$$C_{ER} \sim \frac{z}{n} \tag{18.8}$$

ER graphs show a very striking *phase transition* at $z = 1$ (on average, one connection between two randomly-chosen nodes). As Figure 18.6 shows, when $z = 1$, a **giant component** forms in the graph where a large fraction of the nodes in the graph are interconnected. When $z < 1$, most nodes are isolated from each other. When $z > 1$, the fraction of nodes in the giant component rapidly increases, quickly approaching one.

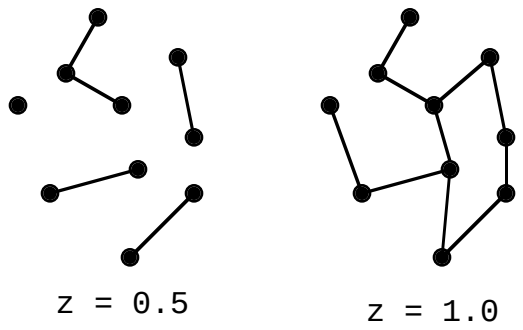

Figure 18.6: A phase transition occurs in random (i.e., ER) graphs when the average number of connections z reaches one. When $z < 1$, most nodes are isolated from each other. When $z = 1$, the graph shows a giant component, wherein a large fraction of the nodes are interconnected, and the fraction of the nodes in the giant component rapidly increases as z increases above one.

The biological implications of a giant component in a graph is that most elements influence most other elements, and the regulatory network moves from a series of discrete, non-overlapping compartments to a single integrated network. In a classic paper, Kaufmann (1967) modeled gene circuits as **random Boolean networks**. Recall that Boolean simply means a zero/one (or on/off) variable. In Kaufmann's model, genes were randomly connected with circuits (edges) and simple on-off rules applied (for example, if your neighbor is one, you are off). He found that such random Boolean networks showed giant regulatory components, features that appear to show complex *order* and give all the appearance of being highly evolved. The key is that random graphs *show considerable structure*, and *one must be extremely careful in inferring evolution simply because a graph shows a complex, and highly integrated, structure*.

Small Worlds, Scale-Free Graphs, and Power Laws

So what is the structure of real regulatory/metabolic graphs? Do they significantly depart from the properties of random (ER) graphs? Regulatory and metabolic networks examined to date show two critical features. First, they are **small-world graphs**, which means that the path distance between any two nodes is small. Think about the "six-degrees of separation" hypothesis – any two individuals in the world can be connect through just (on average) six individuals. We live in a small world. Thus social relationship networks are very similar to metabolic and regulatory networks.. Formally, a graph is said to show the **small world property** when the average path length d is close to that for an ER graph. Note that a fully regular graph (a highly ordered graph where each node is connected to exactly z of its nearest neighbors) can be transformed into a small-world graph by rewiring a small fraction of the nodes at random. Just a small amount of disorder in an otherwise highly regular system can generate the small-world property. The biological key is that *small-world graphs propagate information very efficiently*.

The second critical feature of metabolic and regulatory graphs is that the degree distribution follows a **power law**, with

$$P(k) \sim k^{-\gamma} \tag{18.9}$$

this is in contrast this to the Poisson distribution for ER graphs (Figure 18.7). The key feature of a degree distribution following a power law is that *a few nodes will have very many connections*. Graphs whose degree distribution follow a power law are said to be **scale free graphs**, whose topologies dominated by a few highly connected nodes (called **hubs**). You can see biological examples of hubs in the protein-protein graphs (Figure 18.5). You can also see the presence of hubs in the waterfall diagram (Figure 18.2), where a few eQTLs influence a large number of genes (the vertical lines in the plot).

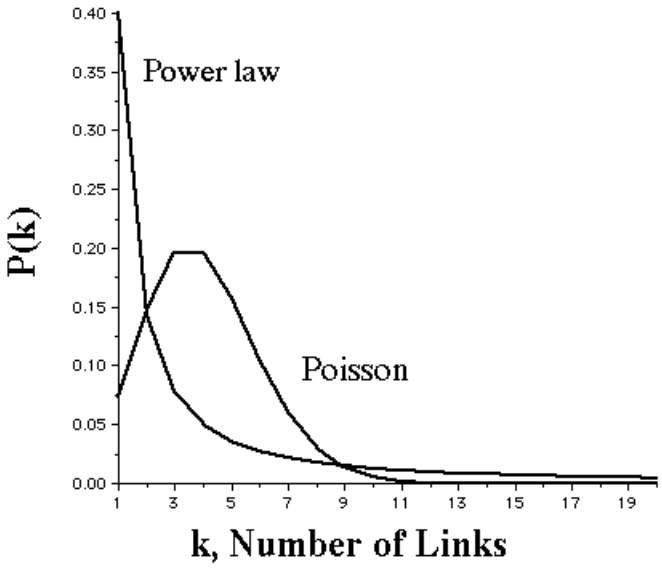

Figure 18.7. The degree distribution $P(k)$, the probability that a random node is connected to exactly k other nodes, under a random (ER) graph and under a power law. Under a random graph, $P(k)$ has a peak at z (the average number of connections), falling off exponentially above this peak. Under a power law distribution, most nodes are linked to just a few others, but the curve falls off more slowly, so that a few nodes are linked to a very large number of other nodes. Such highly connected nodes are called hubs.

A log-log plot of $P(k)$ versus k will be linear if a power law holds. The yeast protein-protein network follows a power law with $\gamma = 2.2$. Likewise, Jeong et al (2000) examined the metabolic networks of 43 organisms (spanning the three kingdoms of life), and found that all followed a power law, again with $\gamma = 2.2$. Thus, a wide range of metabolic pathways in highly divergent organism showed a very similar pattern.

Scale-free graphs show the very important property in that they are fairly robust to perturbations – if we knock out a random component, does the network still work? While scale-free networks have a few very venerable nodes – namely the hubs – most random nodes can be removed with little effect on this system. This was seen in yeast, where extensive experiments using **knock-outs** found that most genes can be deleted with no apparent effect on the cell phenotype. Thus, the scale-free structure of many regulatory graphs gives then a robustness. It is quite possible that much of homeostasis seem in biological systems may be a simple consequence of the network structure, rather than a highly evolved feature.

How might scale-free graphs evolve? The answer turns out to be very simple. When we add new nodes, we simply have a slight preference to attach them to already established nodes. This is exactly what is thought to happen as gene duplication adds new nodes (elements) to a network. Jeong et al. (2000) offer some support for this model, by looking at the hubs in their survey of the metabolic networks of 43 species. Here the nodes were substrates for metabolic reactions. They found that the ranking of the most connected substrates was essentially the same for all 43 species studied. This is especially striking in that only roughly four percent of the substrates were present in all of the studied species. Conversely, species-specific substrates turn out to be much less connected. This is consistent with a model of evolution where ancient features of the pathway serve as hubs, with new features being preferentially connected to these ancient hubs.