10 Evolutionary Genomics of Gene Expression

I. King Jordan and Leonardo Mariño-Ramírez

National Center for Biotechnology Information, National Institutes of Health, 8600 Rockville Pike, Bethesda MD 20894, USA jordan@ncbi.nlm.nih.gov

The study of evolution at the molecular level has focused primarily on changes in gene (protein) sequences over time [1]. Of course, phenotype is influenced not only by the sequence of genes but also by their expression patterns, i.e. the amplitude, timing, and spatial distribution of transcription. Thus, changes in gene expression are likely to be equally as important as sequence changes in evolution; indeed, the significance of gene expression divergence to the evolutionary process has been recognized for some time [2–4]. However, gene expression data have only recently accumulated to the levels needed for systematic evolutionary studies. This has been due to the application of new high throughput techniques that measure gene expression levels for thousands of genes simultaneously [5–7], as well as the development of database resources needed to handle such data [8–10]. The availability of these expression data, together with the long standing interest in the evolutionary significance of gene expression, has stimulated numerous recent studies on gene expression divergence.

This chapter will provide a guide for the study gene expression divergence. The emphasis will be placed on an integrated approach to the study of evolutionary genomics that considers both gene sequence and gene expression divergence and explores the relationship between those two aspects of the evolutionary process. The body of the chapter will be broken down into three sections. The first two body sections, on sequence divergence and on gene expression divergence, will be tutorial in nature and cover specific methodological techniques involved in the study of gene sequence and gene expression divergence. In most cases, descriptions of methods will focus on the most straightforward and widely available techniques. The third body section, on integrated analysis, will be more conceptual in nature and deal with select examples of how gene sequence and gene expression divergence analyses have been used to address fundamental evolutionary questions.

10.1 Sequence Divergence

Methods of gene (protein) sequence analysis have been covered in great detail elsewhere. Here, some of the basic techniques and issues related to sequence

analysis will be covered. A key concept for sequence analysis is that of functional, and/or selective, constraint. A functionally constrained gene is one that encodes a protein that performs a critical function for the organism, and a functionally constrained residue (site) is one that plays an important role in the function of the molecule. Changes to the sequence of such genes, or sites, are likely to be deleterious, i.e. they will reduce the fitness of the organism, and therefore will be removed by natural selection. Thus, genes, or specific sites within a gene, that are under greater selective constraint evolve more slowly, while genes (sites) that are subject to relatively less constraint evolve more rapidly. As such, comparisons of gene divergence levels can be used to make inferences about the strength of functional constraint and the relative action of natural selection.

The availability of complete genome sequences provides great utility for the comparative analysis of gene divergence levels because it allows for the controlled comparison of divergence levels for thousands genes. To compare gene divergence levels between complete genome sequences one needs to: (i) identify orthologous genes, (ii) align gene (protein) sequences, and (iii) calculate substitution rates. Each of these tasks will be briefly covered below.

10.1.1 Ortholog Identification

Genes that share a common ancestor are said to be homologous, and homologous genes can be defined as orthologous or paralogous [11]. Orthologs are genes that diverged due to a speciation event, while paralogs are genes that diverged due to a gene duplication event. Considering two species, such as human and mouse, orthologs can be thought of colloquially as the pair corresponding genes, i.e. those that perform the same function, in each genome. If both genomes are completely sequenced, then pairs of orthologous genes can be identified using sequence similarity. This is the so-called 'reciprocal best hits' approach [12, 13]. To implement this approach, each protein sequence encoded by genome A is compared individually to the entire set of protein sequences encoded by genome B using a sequence similarity comparison tool, such as BLAST [14, 15] or FASTA [16]. Protein sequences are typically used because they are more sensitive for sequence similarity comparisons. Then, for each protein from genome A, the protein with the highest similarity from genome B is recorded as its best hit. The same process is repeated in the opposite direction, with each protein from genome B compared individually to all proteins from genome A and the best hits recorded. Orthologous pairs are then identified as those pairs of proteins that are each others best hits in the reciprocal sequence similarity searches.

This simple approach works quite well for closely related genomes. However, there some caveats to ortholog identification that one should be aware of. Ortholog identification between distantly related genomes is less accurate due to problems with sequence similarity comparisons as well as the phenomenon of gene loss where the corresponding member of an orthologous pair is lost

in one lineage. Furthermore, the reciprocal best hits approach defined above identifies one-to-one orthologs. However, orthologous relationships may also be one-to-many or many-to-many due to lineage-specific gene duplications that occur subsequent to the speciation event under consideration. These complex relationships can confound attempts to identify orthologs using only reciprocal best hits. Ortholog databases, such as the Clusters of Orthologous Groups database [17], use complex algorithms that post process reciprocal best hits between multiple complete genomes and allow for the representation of many-to-many orthologous relationships. A recently developed method for pairwise genome comparison, the reciprocal smallest distance algorithm, has been shown to identify many orthologs missed by reciprocal best hits [18]. In some rare cases, lineage-specific gene duplication followed by differential loss of alternate paralogous copies in the different genomes can result in erroneous identification of orthologs [19]. One way to avoid this problem is to use an *ad hoc* approach whereby the distribution of divergence levels between orthologs is considered and the most divergent outliers are removed.

10.1.2 Sequence Alignment

Once orthologous gene (protein) pairs are identified they need to be aligned before divergence levels can be calculated. The Clustal series is the most widely used group of programs for sequence alignment [20]. Clustal uses a heuristic approach for building multiple sequence alignments, but the initial pairwise alignment step is based on dynamic programming algorithm that guarantees an optimal solution. Given its ready availability and the reliability of its pairwise alignment step, Clustal is a good choice for aligning pairs of orthologs.

10.1.3 Sequence Distance Calculation

Gene (protein) divergence levels are calculated as sequence distances, which measure the numbers of differences between sequences normalized by their lengths. The simplest sequence distance measure is the p-distance. The p-distance is simply the proportion of differences between any two gene sequences and it is defined as

$$p = \frac{n_{\mathrm{d}}}{n}\,, \tag{10.1}$$

where n_{d} is the number of differences between the sequence and n is the number of sites being compared. Alignment sites that contain gaps are usually ignored when calculating p-distances. The problem with the p-distance is that it tends to undercount the number of changes that have occurred between sequences. For example, multiple changes at a single site will only be counted as one difference. Parallel substitutions that lead to the same residue will not be counted at all. A number of different sequence distance measures have been developed that attempt to account for multiple substitutions and give

a more accurate measure of sequence divergence. Examples of a few of these will be covered below for nucleotide and protein sequences.

Nucleotide Sequences

Estimates of nucleotide divergence levels that account for multiple substitutions are based on mathematical models of the substitution process. The simplest such model is the Jukes-Cantor model. In this case, it is assumed that nucleotide frequencies are equal and that nucleotide substitutions are all equally probable. The Jukes-Cantor nucleotide distance (d) [21] can be calculated simply from the p-distance,

$$d = -\frac{3}{4}\ln[1 - (4/3)p]\,. \tag{10.2}$$

Nucleotide substitution models become progressively more complicated by separately parameterizing different aspects of the substitution process. For instance, the Kimura two-parameter (K2P) method [22] accounts for the fact that transitions, changes from purine-to-purine or from pyrimidine-to-pyrimidine, occur at different rates than transversions, changes between purines and pyrimidines. The K2P distance can be calculated as

$$d = -\frac{1}{2}\ln(1 - 2P - Q) - \frac{1}{4}\ln(1 - 2Q)\,, \tag{10.3}$$

where P and Q are the proportional differences between the sequences due to transitions (P) and transversions (Q). The Felsenstein (F81) model extends Jukes-Cantor by allowing for unequal nucleotide frequencies [23]. Both of these methods are merged in the Hasegawa, Kishino, and Yano (HKY85) model that allows for unequal base frequencies as well as different transition and transversion rates [24]. The most nuanced nucleotide substitution model is the General reversible model where nucleotide frequencies are unequal and all six pairs of substitution rates are free to vary [25, 26]. More detailed expositions of nucleotide substitutions models can be found in Refs. [27, 28].

Synonymous versus Non-Synonymous Substitutions

One of the great advantages of using nucleotide sequence distances is the information that they can provide regarding the action of natural selection. The effects of selection on nucleotide coding sequences (CDSs) can be gleaned by comparing levels of synonymous (S) versus non-synonymous (N) sequence divergence [29]. S changes are substitutions in the CDS that do not change the encoded amino acid sequence, while N changes are CDS substitutions that result in amino acid differences. Thus, N changes may change the structure and/or function of the encoded protein, while S changes are largely silent. Since natural selection exerts its influence based on detectable phenotypic

differences, N changes are subject to the effects of selection while S differences are, for the most part, invisible to natural selection. As such, S changes tend to be freer to accumulate than N changes. The proportion of synonymous (p_S) and non-synonymous (p_N) differences can be calculated as the number of S (S_d), or the number of N (N_d), differences normalized by the number of S, or N, sites

$$p_S = \frac{S_\mathrm{d}}{S} \tag{10.4a}$$

$$p_N = \frac{N_\mathrm{d}}{N}\,. \tag{10.4b}$$

These measures also needed to account for multiple substitutions to be more accurate. The p_S and p_N values can be used with the Jukes-Cantor model (Eq. (10.1)) to calculate d_S and d_N respectively. This approach is employed in the Nei-Gojobori method [30] for calculating d_S and d_N. Other methods take into account factors such as transition versus transversion differences as well as the nuances of the genetic code to try and achieve the most accurate S and N distance measures possible. Several of these methods are implemented in the program MEGA [31]. The program PAML [32] also calculates d_S and d_N using a maximum likelihood based approach. Users should note that depending on the method employed, d_S may be referred to as K_s and d_N may be referred to as K_a.

In general, when sequence pairs are compared, $d_N/d_S \ll 1$ is indicative of purifying selection, or removal of deleterious changes, $d_N/d_S \approx 1$ suggest the absence of natural selection, and $d_N/d_S \gg 1$ is indicate of adaptive, or diversifying selection, based on the fixation of beneficial N sequence changes.

Protein Sequences

As with nucleotide sequences, protein sequence divergence levels can be calculated using the proportion of differences (p-distance) between sequences, but this measure will underestimate the true amount of divergence for all but the most closely related sequences. One correction for multiple amino acid substitutions is the Poisson corrected (PC) distance [27] where the probability of k amino acid substitutions at a given site is considered to follow a Poisson distribution. PC can be calculated from the p-distance as

$$d = -\ln(1-p)\,. \tag{10.5}$$

However, most models of the amino acid substitution process are empirical rather than analytical. Empirical models take advantage of the availability of many related protein sequences that can be reliably aligned. Given a set of protein sequence alignments, the relative probabilities of exchange between any two amino acid residues can be calculated. These probabilities can be placed into a substitution matrix which can be employed in the calculation of distances between protein sequences. Commonly employed empirical models

are based on the PAM [33] and JTT [34] amino acid substitution matrices. Distances based on empirical models are more accurate than those obtained with simple analytical models like the PC distance, because they more closely reflect biological reality.

Another important consideration when calculating divergence levels between protein sequences is the fact that changes may accumulate at vastly different rates in different regions of the sequence. As described previously, sites along a sequence that are more functionally constrained will change more slowly than sites that are less constrained. The gamma distance correction is one way to account for this rate variation across sites. The gamma correction is based on the observation that the substitution rate across sites can be considered to vary according to a gamma distribution [35]. The shape of this distribution is governed by a single parameter α. The gamma distance (d_{G}) can be calculated from the p-distance as

$$d_{\mathrm{G}} = \alpha \left[(1-p)^{-1/\alpha} - 1 \right] . \tag{10.6}$$

The lower the value of α, the more severe the correction for multiple substitutions is. A gamma distance correction can also be used together with an empirical based model of amino acid substitution. For instance, the JTT model can be used together with a gamma correction and this is represented as JTT+Γ. Several protein divergence calculation methods are implemented in the program MEGA.

10.2 Gene Expression Divergence

High-throughput techniques for measuring gene expression levels, such as microarray and serial analysis of gene expression (SAGE) approaches, have resulted in an explosion of gene expression data. This section will focus on how microarray data collected from different species can be used to calculate gene expression divergence. In order to illustrate the methods that can be used in the evolutionary analysis of gene expression data, examples will be taken from our own work based on the analysis of the Novartis Foundations' mammalian gene expression atlas [36, 37]. The gene expression atlas reports expression levels for thousands of human and mouse genes based on the use of Affymetrix microarray experiments. There are two versions of the atlas and this chapter focuses primarily on the second and more recent version, sometimes referred to as GNF2. In GNF2, the results of replicate experiments across a wide variety of tissue and cell-line samples are reported. The human data set has expression level measurements for 44,744 probes over 79 distinct tissue samples, while the mouse data set includes expression level measurements for 36,181 probes over 61 distinct tissue samples; in both data sets, each tissue sample is represented by two replicate experiments. Given this abundance of comparative expression data, GNF2 is a phenomenal resource for

the analysis of gene expression divergence. Used together with human-mouse comparative sequence analysis, investigation of the GNF2 data can reveal much about the evolution of gene expression and the relationship between gene sequence and gene expression divergence.

10.2.1 Database Sources

There are a number of database sources available for extracting gene expression data. The Novartis Foundation hosts its own database [38] that allows for targeted querying of the gene expression atlas along with downloading of entire expression datasets. The Novartis site also provides valuable information concerning the expression atlas including useful tips for the analysis and interpretation of their data as well as information on the source of tissue samples used.

The UCSC genome browser [39] has integrated the gene expression atlas data into its browser tracks for the human and mouse genomes. As with everything on their site, the primary expression data is available for download either as a SQL tables or as tabulator delimited text files. In addition, one particularly nice feature provided by the UCSC genome browser is the mapping of Affymetrix probes identifiers to Genbank accessions for known human and mouse genes. The UCSC genome browser also provides a number of other canonical gene expression datasets for other model organisms such as yeast and *Drosophila*.

The Gene Expression Omnibus (GEO) database [40] at the National Center for Biotechnology Information (NCBI) also provides the GNF2 data. This data can be queried with search terms, and users can also download the primary expression data as tabulator delimited text files. GEO is a vast repository of gene expression data and one of its strengths is the many ways that the data can be queried. For instance, users can identify all datasets generated from a specific microarray platform. This can be quite useful for extracting datasets that can be readily compared. GEO also provides nice graphical views of expression data results as well as targeted comparisons between user selected samples from a given dataset.

There are a couple of other database resources of note that store and disseminate microarray expression data including the Stanford Microarray Database [41] and the Human Gene Expression Index [42].

10.2.2 Probe-to-Gene Mapping

One of the first challenges in using GNF2, or any results based on the Affymetrix platform for that matter, is probe-to-gene mapping. Affymetrix microarray technology is based on oligonucleotide probes that are designed to correspond to specific genes, and each probe is labeled with a unique identifier (id). The primary expression data are generally distributed along with

these Affymetrix ids. Since comparative, between species, analysis of gene expression data involves a gene-centric approach, the user is faced with the task of mapping these probe ids to gene ids such as Genbank or Refseq accessions. Fortunately, Affymetrix probe-to-gene mapping keys can be downloaded from the Affymetrix web site [43]. Alternatively, the UCSC Genome Browser provides probe-to-gene mapping for GNF2. Finally, since the sequences of the probes are often provided along with expression data, users could always do their own probe-to-gene mapping but this would be somewhat labor intensive.

Users should be aware that sometimes multiple probes will map to a single gene. In such cases, the user may decide to average the expression values for the multiple probes to come up with one set of gene-specific values. Another, somewhat arbitrary, approach that is sometimes used is to take the probe that yields the highest expression value as the best one for a given gene. Far more problematic is the fact that, in some cases, a single probe will map to multiple unique genes; because of their inherent ambiguity, these data should not be used for the analysis of gene expression divergence.

10.2.3 Structure of the Data

The GNF2 data, as well as many other microarray datasets, are structured as a table with probe-specific expression data in rows and experiment (sample)-specific data in the columns. Thus, for any particular probe, or gene, the expression data consists of an array of values, one for each experimental condition. These are gene expression profiles, and they can be thought of as vectors in n-dimensional space, where n is the number of distinct samples in the data set. For instance, for any gene i, with expression levels recorded across n samples, its expression profile can be represented as

$$\text{gene}_i = [X_{i1}, X_{i2}, X_{i3}, \ldots, X_{in}]\,, \tag{10.7}$$

where X_{ij} is the expression value for gene i in the experiment j. It is these gene-specific expression profiles that can be compared within or between species to measure gene expression divergence. When comparing profiles between species, it is essential that the identity j and number n of the samples is identical in the vectors. For example, in GNF2 the human and mouse datasets share 28 common tissue samples. These 28 samples can be arranged in the same order across the gene expression profiles so that vectors can be meaningfully compared between species.

10.2.4 Transformation and Normalization

Two critical issues with respect to microarray data are transformation and normalization. The details on transformation and normalization are outside the scope of this chapter but they have been treated in depth elsewhere [44–46]; here, these matters will be covered briefly. For microarray

data, transformation generally involves taking the logarithm to the base 2 ($\log_2$) of the expression value. This procedure is typically used to transform ratios as in the case where the expression data are represented as a ratio of one experimental condition over a reference experimental condition. In this case, $\log_2$ transformation treats up-regulation and down-regulation equally and represents them in a very regular and intuitive way. For example, a ratio of 1 would mean no change in expression across conditions and the $\log_2$ value would be 0. A ratio of 4 would yield a $\log_2$ value of 2, while a ratio of $1/4$ gives a $\log_2$ value of -2. Thus, using log transformation, the magnitude of the deviations in up and down-regulation are symmetrical around 0. The GNF2 data, however, consist of absolute expression value measurements as opposed to ratios. Nevertheless, the relative levels of gene expression for each sample j across a gene i expression profile is often represented as a ratio of the absolute expression value j over the median value of all n expression values in the profile. The use of $\log_2$ transformation for these kinds of profile median ratios has the same useful effect of mapping changes in gene-specific relative expression in a symmetrical and continuous way around 0. The basic idea behind normalization is to control for systematic variation between experimental conditions that affect the recorded levels of expression. This is particularly important when comparing expression levels, or profiles, of orthologous genes between species. Since the conditions under which the experiments for different species were conducted are sure to differ, it is essential that relative, as opposed to absolute, expression values are compared between experiments. One straightforward way to do this is to mean, or median, center the results for each microarray. This consists of dividing each individual expression by the mean, or median, expression value for the entire array.

10.2.5 Measuring Divergence

Expression Level and Breadth

Gene expression divergence can be measured in several different ways. Perhaps the most intuitive of these methods is to compare differences in gene expression level or amplitude. When comparing between species, it is important to ensure that the same tissues or samples are being compared and that the data between species has been appropriately normalized. Given a gene-specific profile of expression levels across a set of different tissues or samples, the expression level can be taken as the maximum or the average of the tissue-specific expression values. While both of these measures fairly represent the amplitude of gene expression, using the average (or the sum) over all tissue-specific values has the disadvantage of conflating gene expression level with gene expression breadth. Gene expression breadth is a measure of how widely a gene is expressed and can be counted simply as the number of tissues in which a gene is expressed at or above some threshold. Obviously if expression breadth is to be accurately compared between species, then it

helps to ensure that the same set of tissues is being compared for each species. For example, the GNF2 has a total of 28 tissues that are shared between the human and mouse experiments. Expression breadth would simply be measured as the number of j tissue samples out of 28 where gene i is expressed at or above some threshold. GNF2 provides absolute expression levels in arbitrary units that are referred to as signal intensity values. For the GNF2 data, a signal intensity value ≥ 350 can be taken to (approximately) indicate that a gene i is expressed in a tissue j. In addition to signal intensity values (i.e. absolute expression levels), the Novartis site also provides presence absence calls for each gene i and condition j (X_{ij}). These calls simply indicate whether or not a gene i can be considered to be expressed in tissue j with a certain level of statistical confidence. Thus, another approach to determine expression breadth is simply to use the presence/absence calls for all X_{ij}.

Gene Expression Profiles

As described earlier, a gene expression profile represents the levels of expression for gene i over all experiments (tissues) j. Most comparisons of gene expression patterns consist of quantitative measures of the similarity or difference between gene expression profile vectors. Two of the most commonly used metrics for comparing expression profile vectors are the Euclidean distance and the Pearson correlation coefficient. The Euclidean distance is geometric measure of the straight line distance of two points. The higher the Euclidean distance, the more different the gene expression profiles are. For instance, if comparing two genes A and B that have two-dimensional gene expression profile vectors $A = [a_1, a_2]$ and $B = [b_1, b_2]$, the Euclidean distance d_{E} would be calculated as

$$d_{\mathrm{E}} = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}\,. \tag{10.8}$$

For an expression profile vector of n-dimensions, the Euclidean distance d_{E} would be calculated as

$$d_{\mathrm{E}} = \sqrt{\sum_{j=1}^{n} (a_j - b_j)^2}\,. \tag{10.9}$$

The Euclidean distance is particularly sensitive to changes in the magnitude of gene expression. Comparison of genes with identical relative expression levels across n tissues may actually yield quite large Euclidean distances if their absolute expression levels differ substantially. One way to get around this is to use relative expression levels by mean, or median, centering the expression levels for each gene-specific vector.

The Pearson correlation coefficient is also widely used in comparing gene expression profile vectors, and it measures the strength of the linear relationship between the vectors being compared. Pearson correlation coefficient

values scale from −1 to +1, where −1 would correspond to the exact opposite expression pattern and +1 would indicate an identical expression pattern. Use of the Pearson correlation coefficient assumes that the data are normally distributed so log transformation of expression data is advised when comparing profiles with this method. It is also important to note that the Pearson correlation coefficient works best when comparing genes that are differentially expressed, i.e. when there are substantial differences in expression levels across the n samples being considered. Genes that are ubiquitously expressed, such as housekeeping genes, can obviously be considered to have very similar expression patterns. However, because there may be no discernable linear relationship between up and down expression across tissues for such evenly expressed genes, comparison of these genes using the Pearson correlation coefficient will often result in values around 0 indicating no correlation. On the other hand, the Pearson correlation coefficient is very good at identifying genes with similar tissue-specific expression patterns. There are many forms for the Pearson correlation coefficient r. Given two n-dimensional gene expression profiles vectors for genes A and B, where $A = [a_1, a_2, \ldots, a_n]$ and $B = [b_1, b_2, \ldots, b_n]$, r can be calculated as

$$r = \frac{\sum_{j=1}^{n} a_j\, b_j - \frac{1}{n}\sum_{j=1}^{n} a_j \sum_{j=1}^{n} b_j}{\sqrt{\sum_{j=1}^{n} a_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} a_j\right)^2}\sqrt{\sum_{j=1}^{n} b_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} b_j\right)^2}}. \tag{10.10}$$

Two other useful distance measures that are often employed include the Hamming distance and mutual information both of which are useful for considering expression data that has been rendered discrete such as presence/absence calls that can be represented as binary expression profiles.

10.2.6 Clustering and Visualization

Clustering and visualization are important components of gene expression divergence analysis, which will nevertheless be treated in only the most cursory manner here. For more detailed treatment of these issues, users can consult [44, 47–49]. The idea behind clustering is simply to group genes with similar expression profiles together. Clustering approaches can be classified as hierarchical or non-hierarchical. Hierarchical methods group profiles into clusters and also specify the relationships among the profiles within clusters, while non-hierarchical methods simply define clusters of related expression profiles with no specification of the within group relationships. Hierarchical clustering methods can be agglomerative, where profiles are successively joined until they are all connected, or divisive, where the entire set of profiles is considered as a single cluster that is progressively broken down. Non-hierarchical

clustering, on the other hand, starts with a pre-defined number of groups and then proceeds to partition the profiles into these discrete groups. Examples of non-hierarchical clustering are K-means clustering and self organizing maps.

Visualization provides a very intuitive way for the user to identify similarly expressed genes. In visualization, each sample j in a gene profile i is assigned a color that indicates its relative level of expression. A typical color scheme that is employed in visualization is to label relatively high expression levels (or up-regulated) as red and relatively low expression levels (or down-regulated) as green. This allows for ready identification of genes that have similar patters of up and down expression across their respective profiles. Visualization is often combined with clustering techniques to define related sets of genes. There are many software packages that combine clustering and visualization techniques. One freely available program that we have found to be quite useful is the TIGR Multiexperiment Viewer (MEV) [50].

10.3 Integrated Analysis

This section will treat a few select examples from the literature that illustrate how integrated gene sequence and gene expression divergence analyses can be used to address fundamental evolutionary questions. This survey highlights new findings regarding the evolution of gene expression as well as some of the open questions that have been raised in this relatively new area of inquiry.

10.3.1 Sequence versus Expression Divergence

We have explored the intersection of gene expression and gene sequence divergence in two recent publications of our own [51, 52]. Both of these papers dealt with mammalian evolution and combined genomic sequence analysis with analysis of gene expression data from the Novartis gene expression atlas. The first of these studies took a network-based approach to the study of gene co-expression [52]. Human gene expression profiles were compared and genes that were found to be co-expressed were linked in a network. The topology of the resulting human gene co-expression network was shown to have scale-free properties that imply evolutionary self-organization via preferential node attachment. When rates of sequence evolution between human and mouse orthologs were overlayed on the co-expression network, genes with numerous co-expressed partners, so-called 'hubs' of the network, were found to evolve more slowly, on average, than genes with fewer co-expressed partners. Furthermore, co-expressed genes were demonstrated to have co-evolved in the sense that they have similar rates of evolution. These observations indicate that the strength of selective constraints on gene sequences is strongly influenced by the topology of the gene co-expression network. This connection is strong for the coding regions and $3'$ untranslated regions (UTRs), but the $5'$ UTRs appear to evolve under a different regime. An interesting exception to

this trend was found for the relationship between gene sequence divergence and gene expression profile divergence. We found no correlation between the rate of gene sequence divergence and the extent of gene expression profile divergence between human and mouse. This suggests that distinct modes of natural selection might govern sequence versus expression divergence. Our current work is focused on the possibility that the evolution of gene expression may be driven by adaptation-driven divergence characterized convergent evolution of gene expression patterns.

In a related study, two different aspects of gene expression divergence were related to gene sequence divergence [51]. Changes in the expression level, or the amplitude of expression, between human-mouse orthologs were shown to be correlated with levels of gene sequence divergence that are determined largely by purifying selection. However, consistent with the previously described work, evolutionary changes of tissue-specific gene expression profiles did not show such a correlation with sequence divergence. This is despite the fact that divergence of both gene expression levels and profiles were significantly lower for orthologous human-mouse gene pairs than for pairs of randomly chosen human and mouse genes. Together, these findings indicate that while purifying selection is acting to constrain gene expression divergence, there is also likely to be a neutral component in evolution of gene expression. This may be particularly true for tissues where the expression of a given gene is low and functionally irrelevant. Neutral evolution of gene expression is explored in more detail in the following section. One prediction of the neutral model of gene expression divergence is a regular, clock-like accumulation of gene expression changes. Relative rate tests of the gene expression divergence among human-mouse-rat orthologous gene sets did reveal clock-like evolution for gene sequence divergence, and to a lesser extent for gene expression level divergence, but not for the divergence of tissue-specific gene expression profiles. These results suggest that the evolution of tissue-specific expression profiles may be influenced by adaptively driven changes that tend to accumulate at an uneven tempo over time.

10.3.2 Neutral Changes in Gene Expression

Neutral evolutionary changes are those that do not confer any selective advantage or disadvantage. The neutral theory of molecular evolution holds that most changes between gene sequences are neutral, with respect to organismic fitness, and accumulate due to the random fixation of variants. The relative influence of adaptively driven changes versus neutral evolution of gene sequences was an historically contentious issue that led to many fruitful areas of inquiry. It appears that a similar debate in the literature is emerging over the relative contributions of these two evolutionary modes – selection driven versus neutral – to the evolution of gene expression.

Only very recently, due to the systematic analysis of high-throughput gene expression data sets, has the neutral frame of reference started to be

applied in earnest to the evolution of gene expression patterns. In one particularly provocative study, Khaitovich et al. evaluated the divergence of gene expression patterns within and between several mammalian species and concluded that the evolution of gene expression patterns is largely neutral [53]. They based their conclusion on several observations. First of all, expression levels between species were found to accumulate approximately linearly as a function of time; this pattern held for comparisons of both primate species and of mouse species. Secondly, divergence of expression levels between human and chimp were found to be the same for pseudogenes, which evolve under no selective constraint, as for (intact) non-pseudogenes. Finally, expression level differences within species were shown to be strongly correlated with expression level differences between species, suggesting that the same neutral evolutionary process are involved in the evolution of gene expression both within and between species. The conclusion that gene expression divergence is primarily neutral has substantial implications for the study of biological evolution and function. For instance, the clock-like accumulation of gene expression changes may allow for detailed inferences on the evolution of different tissues and organ systems based on changes in gene expression patterns among them. On the other hand, if gene expression changes are neutral then the application of expression data to functional inferences may be limited.

Another recent study, by Yanai et al., also concluded that gene expression divergence between species may be dominated by neutral evolution [54]. This work took advantage of the first mammalian gene expression atlas provided by Novartis to assess the rate of gene expression pattern divergence between mammalian species. Orthologous pairs of human-mouse genes were identified and their expression patterns across multiple tissues were quantified. Expression patterns were represented as profiles that reflect the relative expression levels in different tissues, and both distance measures and correlations between profiles were measured. Several surprising results came out of this analysis. None of the gene expression profiles that were most similar between human and mouse corresponded to orthologous genes. In fact, expression profiles between orthologous genes were found to diverge so rapidly that their differences are comparable to those seen between duplicated genes (paralogs) and between random gene pairs. Even the corresponding tissues between the two species did not show similar patterns of gene-specific expression levels; all human tissues were more similar to one another as were all mouse tissues. Such rapid divergence in gene expression can be attributed to the effects of natural selection based on adaptively beneficial functional differences, so-called positive selection, or to random drift based on functionally indistinguishable (i.e. neutral) differences. The authors favor the neutral model for several reasons including the presence of orthologous gene pairs that are not presumed to have changed function and the even distribution of expression differences across tissues.

Contrary to the results suggesting neutral evolution of gene expression, a few other recent studies have pointed to an important role for natural selection in constraining, and perhaps driving, gene expression divergence. Fraser et al. focused on random fluctuations in gene expression that produce noise in protein levels [55]. They investigated the biological significance of this noise, specifically asking whether fluctuations in gene expression are biologically relevant and thus subject to natural selection. To investigate this issue, they tested two specific hypotheses, namely that two classes of genes, (i) essential genes and (ii) genes that encode members of multi-subunit protein complexes, should both be particularly sensitive to random fluctuations (noise) in gene expression levels. Combined computational analyses of yeast gene and protein expression levels, gene knock-out effects, and protein-protein interaction data were used towards this end. The rationale behind the test was that essential genes and genes that encode members of protein complexes should be particularly subject to the effects of natural selection – indeed these classes of genes tend to show reduced rates of evolution consistent with strong purifying selection – and if fluctuations in expression levels are significant, these too should be under strong selection for the gene classes in question. They found that, for both tests and over a large range of protein production levels, these two classes of genes show significantly and substantially lower levels of noise in protein expression than other genes. From this, it was concluded that noise in gene expression is biologically relevant and is subject to the effects of natural selection. This conclusion seemingly stands in stark contrast to those of the two studies summarized above, both of which conclude that gene expression levels are neutral with respect to organismic fitness. However, the results summarized here may not be inconsistent with a neutral model of gene expression. Indeed, natural selection does have an important role under the neutral model of evolution, but its effect is to reduce, rather than enhance, levels of diversity. Thus, genes (or positions in gene sequences) that are more functionally constrained are expected to evolve more slowly than those that are less functionally constrained and this prediction has been born out time and time again. The noise in protein expression, while biologically relevant in the sense that it is deleterious, conforms to the neutral pattern with genes that are presumably more functionally constrained showing less variation in protein production.

Yet another recent study examined patterns of gene expression polymorphism (within species changes) and gene expression divergence (between species changes) in a number of datasets from different eukaryotic species including *Drosophila*, mice, and primates [56]. Here again, the evolution of gene expression was considered with respect to the neutral model of change. Using a number of different measures, the authors found that gene expression levels tend be evolutionarily stable in that they change very little over time. This stability strongly implies that gene expression levels are subject to selective constraint. However, there are substantial differences in the rates

at which gene expression changes, and different functional classes of genes were shown to have distinct characteristic levels of change. The authors also put forward a model that explains how changes in gene expression could be driven by directional selection.

10.3.3 Evolutionary Conservation of Gene Expression

The accumulation of genome scale expression data sets is beginning to provide opportunities for cross-species comparisons of gene expression patterns. One recent report provides an example of how such studies can reveal patterns of evolutionary conservation of gene expression [57]. The authors of this work compiled large-scale gene expression data sets from six diverse species: *Escherichia coli*, *Saccharomyces cerevisiae*, *Arabidposis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Homo sapiens*. Their analysis included expression data for more than 40,000 genes under 2,000 experimental conditions, and the study is also notable for its combination of gene expression and sequence data analysis. Correlations between condition-specific gene expression patterns were used to identify co-expressed genes within species. Co-expression networks, where genes are the nodes and they are connected if significantly co-expressed, were found to have similar connectivity across species. The distributions of network connectivity were found to follow a power-law similar to other biological networks such as protein-protein interaction and metabolic networks. The scale-free nature of these distributions, along with their conservation between species, suggests that a fundamental mechanism is involved in the evolution of gene expression in different domains of life. Pairwise correlations between genes were also compared to the correlations between homologous gene pairs in different species and a significant fraction was found to be similar. The utility of this homologous co-expression similarity with respect to functional annotation of genes was demonstrated. In addition, sets of highly connected genes were enriched for genes that are essential and posses a high number of homologous sequences in other organisms. Expression data were broken down into modules that consist of co-expressed genes and the particular expression conditions that give rise to their co-regulation. While some groups of functionally related genes show up as conserved modules in multiple species, many of the expression modules vary widely across species and so probably contribute to evolutionary diversification.

Stuart et al. conducted a similar study of the cross-species conservation of gene expression and identified pairs of genes that are co-expressed across more than 3,000 microarray experiments conducted for humans, flies, worms, and yeast [58]. A total of 22,163 evolutionarily conserved co-expression relationships were identified in this way. Links between co-expressed genes were used to build a co-expression network and this approach revealed network components that were specific for different levels of diversification, such as

ancient versus more recently evolved connections. The conservation of expression patterns suggests that co-expression gene pairs are functionally related and the functional similarity provides the mechanistic basis of the selection for maintained co-expression. In light of this finding, the authors demonstrate how the co-expression relationships can be used to provide evidence for the involvement of new genes in specific cellular functions including cell cycle, secretion, and protein expression. Notably, a few specific predictions generated based on conserved co-expression were tested and confirmed.

References

1. W.H. Li: Molecular Evolution (Sinauer Associates, Sunderland 1997)
2. R.J. Britten, E.H. Davidson: Science **165**, 349 (1969)
3. R.J. Britten, E.H. Davidson: Q. Rev. Biol. **46**, 111 (1971)
4. M.C. King, A.C. Wilson: Science **188**, 107 (1975)
5. V.E. Velculescu, L. Zhang, B. Vogelstein et al: Science **270**, 484 (1995)
6. M. Schena, D. Shalon, R.W. Davis et al: Science **270**, 467 (1995)
7. M.D. Adams, J.M. Kelley, J.D. Gocayne et al: Science **252**, 1651 (1991)
8. D. Karolchik, R. Baertsch, M. Diekhans et al: Nucleic Acids Res. **31**, 51 (2003)
9. R. Edgar, M. Domrachev, A.E. Lash: Nucleic Acids Res. **30**, 207 (2002)
10. J. Gollub, C.A. Ball, G. Binkley et al: Nucleic Acids Res. **31**, 94 (2003)
11. W.M. Fitch: Syst. Zool. **19**, 99 (1970)
12. M.C. Rivera, R. Jain, J.E. Moore et al: Proc. Natl. Acad. Sci. U S A **95**, 6239 (1998)
13. R.L. Tatusov, E.V. Koonin, D.J. Lipman: Science **278**, 631 (1997)
14. S.F. Altschul, W. Gish, W. Miller et al: J. Mol. Biol. **215**, 403 (1990)
15. S.F. Altschul, T.L. Madden, A.A. Schaffer et al: Nucleic Acids Res. **25**, 3389 (1997)
16. W.R. Pearson, D.J. Lipman: Proc. Natl. Acad. Sci. U S A **85**, 2444 (1988)
17. E.V. Koonin, N.D. Fedorova, J.D. Jackson et al: Genome Biol. **5**, R7 (2004)
18. D.P. Wall, H.B. Fraser, A.E. Hirsh: Bioinformatics **19**, 1710 (2003)
19. I.K. Jordan, Y.I. Wolf, E.V. Koonin: BMC Evol. Biol. **4**, 22 (2004)
20. R. Chenna, H. Sugawara, T. Koike et al: Nucleic Acids Res. **31**, 3497 (2003)
21. T.H. Jukes, C.R. Cantor: Evolution of protein molecules. In Mammalian Protein Metabolism, ed by H.D. Munro (Academic Press New York 1969)
22. M. Kimura: J. Mol. Evol. **16**, 111 (1980)
23. J. Felsenstein: J. Mol. Evol. **17**, 368 (1981)
24. M. Hasegawa, H. Kishino, T. Yano: J. Mol. Evol. **22**, 160 (1985)
25. F. Rodriguez, J.L. Oliver, A. Marin et al: J. Theor. Biol. **142**, 485 (1990)
26. Z. Yang, N. Goldman, A. Friday: Mol. Biol. Evol. **11**, 316 (1994)
27. M. Nei, S. Kumar: Molecular Evolution and Phylogenetics (Oxford University Press, New York 2000)
28. R.D.M. Page, E.C. Holmes: Molecular Evolution: a Phylogenetic Approach (Blackwell Science, Malden 1998)
29. L.D. Hurst: Trends Genet. **18**, 486 (2002)
30. M. Nei, T. Gojobori: Mol. Biol. Evol. **3**, 418 (1986)

31. S. Kumar, K. Tamura, M. Nei: Brief. Bioinform. **5**, 150 (2004)
32. Z. Yang: Comput. Appl. Biosci. **13**, 555 (1997)
33. M.O. Dayhoff, R.M. Schwartz, B.C. Orcutt: A model of evolutionary change in proteins. In Atlas of Protein Sequence and Structure, ed by M.O. Dayhoff (National Biomedical Research Foundation Washington, D.C. 1978)
34. D.T. Jones, W.R. Taylor, J.M. Thornton: Comput. Appl. Biosci. **8**, 275 (1992)
35. T. Uzzell, K.W. Corbin: Science **172**, 1089 (1971)
36. A.I. Su, T. Wiltshire, S. Batalov et al: Proc. Natl. Acad. Sci. U S A **101**, 6062 (2004)
37. A.I. Su, M.P. Cooke, K.A. Ching et al: Proc. Natl. Acad. Sci. U S A **99**, 4465 (2002)
38. `http://symatlas.gnf.org/SymAtlas/`
39. `http://genome.ucsc.edu/`
40. `http://www.ncbi.nlm.nih.gov/geo/`
41. `http://genome-www5.stanford.edu/`
42. `http://df216w01.mgh.harvard.edu/hio/`
43. `http://www.affymetrix.com/support/index.affx`
44. M.M. Babu: An Introduction to Microarray Data Analysis. In Computational Genomics: Theory and Application, ed by R.P. Grant (Horizon Bioscience Norwich 2004)
45. H.C. Causton, J. Quackenbush, A. Brazma: Microarray Gene Expression Data Analysis: a Beginner's Guide (Blackwell Science, Malden 2003)
46. J. Quackenbush: Nat Genet. **32 Suppl**, 496 (2002)
47. R. Xu, D. Wunsch: IEEE Trans. Neural. Netw. **16**, 645 (2005)
48. D.R. Gilbert, M. Schroeder, J. van Helden: Trends Biotechnol. **18**, 487 (2000)
49. R.B. Altman, S. Raychaudhuri: Curr. Opin. Struct. Biol. **11**, 340 (2001)
50. `http://www.tm4.org/mev.html`
51. I.K. Jordan, L. Marino-Ramirez, E.V. Koonin: Gene **345**, 119 (2005)
52. I.K. Jordan, L. Marino-Ramirez, Y.I. Wolf et al: Mol. Biol. Evol. **21**, 2058 (2004)
53. P. Khaitovich, G. Weiss, M. Lachmann et al: PLoS Biol. **2**, E132 (2004)
54. I. Yanai, D. Graur, R. Ophir: Omics **8**, 15 (2004)
55. H.B. Fraser, A.E. Hirsh, G. Giaever et al: PLoS Biol. **2**, e137 (2004)
56. B. Lemos, C.D. Meiklejohn, M. Caceres et al: Evolution Int. J. Org. Evolution **59**, 126 (2005)
57. S. Bergmann, J. Ihmels, N. Barkai: PLoS Biol. **2**, E9 (2004)
58. J.M. Stuart, E. Segal, D. Koller et al: Science **302**, 249 (2003)