Conservation and Coevolution in the Scale-Free Human Gene Coexpression Network

I. King Jordan, Leonardo Mariño-Ramírez, Yuri I. Wolf, and Eugene V. Koonin

National Center for Biotechnology Information, National Institutes of Health Bethesda, Maryland

The role of natural selection in biology is well appreciated. Recently, however, a critical role for physical principles of network self-organization in biological systems has been revealed. Here, we employ a systems level view of genome-scale sequence and expression data to examine the interplay between these two sources of order, natural selection and physical self-organization, in the evolution of human gene regulation. The topology of a human gene coexpression network, derived from tissue-specific expression profiles, shows scale-free properties that imply evolutionary self-organization via preferential node attachment. Genes with numerous coexpressed partners (the hubs of the coexpression network) evolve more slowly on average than genes with fewer coexpressed partners, and genes that are coexpressed show similar rates of evolution. Thus, the strength of selective constraints on gene sequences is affected by the topology of the gene coexpression 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. Surprisingly, we found no connection 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, and we propose a model, based on rapid, adaptation-driven divergence and convergent evolution of gene expression patterns, for how natural selection could influence gene expression divergence.

Introduction

The recent genomic sequencing efforts yielded detailed lists of genes and the proteins that they encode (Lander et al. 2001; Waterston et al. 2002), and functional genomics studies have gone a step further by elucidating many of the processes and interactions that individual proteins are involved in (Ho et al. 2002; Giot et al. 2003; Kamath et al. 2003; Li et al. 2004). Building on the success of these high-throughput experimental approaches, a synthetic view of how individual genes and proteins emerge and act collectively to carry out the business of the cell is needed to facilitate a deeper understanding of biological function and evolution. Systems-based approaches to biology seek to meet this challenge by emphasizing the patterns and processes that govern how collections of biological molecules are assembled and ordered (Pennisi 2003).

The agent that probably has been most often invoked to explain the ordering of biological systems over time is natural selection (Darwin 1859; Li 1997). Genome-scale studies on natural selection have detailed many of the factors that mitigate the effects of selection on the evolution of gene sequences. Such surveys rely on comparisons between evolutionary rates, which yield information about the action of natural selection, and various quantifiable functional genomic parameters. For instance, several studies have demonstrated a relationship between gene evolutionary rates and the fitness effects associated with gene knockouts. Genes with greater fitness effects (e.g., essential genes) seem to evolve more slowly, on average, than genes with smaller fitness effects (Hirsh and Fraser 2001; Jordan et al. 2002). This is taken to suggest that essential genes evolve under stronger functional constraints and, thus, a more severe purifying selection regime, than nonessential genes. Similarly, genes that encode proteins involved in numerous protein-protein interactions have been reported to be more evolutionarily conserved than genes encoding less-prolific interactors (Fraser et al. 2002; Fraser, Wall, and Hirsh 2003). A recent study that dealt with several such relationships simultaneously demonstrated correlations between different measures of evolutionary conservation and various functional genomic parameters (Krylov et al. 2003).

However, the findings of some of these evolutionary genomics studies have been challenged. The possibility that the observed effects of any one genomic parameter on evolutionary rates can be confounded by the correlations between different genomic parameters has been raised repeatedly. For example, some of the strongest correlations seen are between evolutionary rates and gene expression levels. Genes that are expressed at high levels and in numerous tissues tend to be more conserved than genes with lower and narrower expression patterns (Duret and Mouchiroud 2000; Pal, Papp, and Hurst 2001; Krylov et al. 2003; Zhang and Li 2004). When the effects of expression level are controlled for, the correlations between evolutionary rate and fitness effects as well as between evolutionary rate and the number of protein-protein interactions are mitigated (Bloom and Adami 2003; Pal, Papp, and Hurst 2003). Furthermore, when duplicate genes were removed from consideration, the relationship between fitness effects and evolutionary rate disappeared (Yang, Gu, and Li 2003). These controversies remain unsettled, and the general question of how various functional genomic parameters interact to effect evolutionary rate is open.

In addition to natural selection, an emphasis has recently been placed on the role of fundamental physical principles in imposing order on biological systems (Barabasi and Oltvai 2004). Various complex biological systems have been abstracted as networks where the nodes in the network represent the individual parts, such as proteins or metabolites, and the links in the network represent the interactions between the parts (Jeong et al.
Materials and Methods

Human and mouse gene expression levels were taken from a recently published series of Affymetrix microarray experiments (Su et al. 2002) and were retrieved from the Gene Expression Omnibus database at the National Center for Biotechnology Information (NCBI). The two data set flat files—GDS181.soft (human) and GDS182.soft (mouse)—were downloaded from ftp://ftp.ncbi.nih.gov/pub/geo/data/gds/soft/. Affymetrix probe identifiers were mapped to individual loci in the human and mouse genomes using the LocusLink database (NCBI, NIH, Bethesda). A total of 7,383 human and 6,724 mouse loci were identified using an all-against-all BlastP search of the RefSeq database (NCBI, NIH, Bethesda). Human-mouse orthologs were identified as reciprocal best Blast hits between protein sequences as previously described (Jordan, Wolf, and Koonin 2003). Human-mouse orthologous protein sequences were aligned using ClustalW (Higgins, Thompson, and Gibson 1996), and the protein alignments were used to guide alignments of the corresponding nucleotide coding sequences to ensure that they were aligned in frame. The 5′ and 3′ UTR sequences were aligned using ClustalW, and only alignments where the shortest sequence had more than 20 residues and had no more than 50% differences in the number of residues (i.e., length) between aligned sequences were used for further analysis. Synonymous (dS) and nonsynonymous (dN) substitution rates were calculated for alignments of protein-coding sequences using the Nei-Gojobori method (Nei and Gojobori 1986) implemented in the PAML package (Yang 1997). The 5′ and 3′ UTR substitution rates (d) were calculated using the Jukes-Cantor correction for multiple substitutions (Jukes and Cantor 1969). Paralogous genes were identified using an all-against-all BlastP search of the proteins in the coexpression (r ≥ 0.7) network with an e-value threshold of 10−5 and a coverage cutoff such that more than 50% of the shorter protein sequence had to be included in the high scoring segment pair.
Comparisons between substitution rates and various gene expression parameters (expression breadth, expression level, and correlations) were done by sorting the rates or rate differences in the ascending order and then binning the sorted values into 10 equal-sized bins. Average and fractional values of gene expression parameters for each substitution rate bin were compared with respect to the order of the bins and the Spearman rank correlation (R), along with the significance of the differences between them was calculated using a z-test.

Phylectic patterns for human genes were taken from the eukaryotic Clusters of Orthologous Groups of proteins (KOGs) database (Tatusov et al. 2003; Koonin et al. 2004). For each KOG, its phylectic pattern corresponds to the presence/absence state of the KOG member proteins among the seven eukaryotic species included in the database. For each phylectic pattern, a specific evolutionary scenario was determined by mapping the states (presence or absence) and events (gain or loss) to the species tree of the seven organisms in KOGs using the Dollo parsimony method (Farris 1977). Pairs of scenarios were then compared to derive values of the phylectic similarity measure. This was done by scoring the combinations of states and events on each branch of the species tree and summing across all branches. For each branch, the specific combination of states and events for a pair of scenarios was considered with respect to the branch length and the branch specific propensities for gene gain and loss. The branch lengths (MYR) on the species tree were taken as described previously (Hedges et al. 2001; Krylov et al. 2003). The branch-specific relative propensities for gene gain ($P^g_i$) and gene loss ($P^l_i$) were calculated using the total number of gene gains and losses mapped to each branch (Koonin et al. 2004) according to the following formula:

$$P^g_i = G_i \sum_j N_j / N_i \sum_j G_j, \quad P^l_i = L_i \sum_j N_j / N_i \sum_j L_j$$

where $N_j$, $G_j$, and $L_i$ are the numbers of present genes, gains, and losses mapped to the $i$th branch. The scoring schemes for all possible combinations of states and events are shown at ftp://ftp.ncbi.nih.gov/pub/koonin/Jordan/MBE-04-0138.R1/SupplementaryTable4.doc. The sum of scores across all branch lengths was divided by the sum of the respective normalization scores (see table 4 in Supplementary Material online) to provide the pattern similarity scores (range from ~1 to 1).

**Results and Discussion**

**Expression Level and Breadth Versus Sequence Divergence**

A recent large-scale microarray study that includes quantitative analysis of gene expression patterns for 31 human and 46 mouse tissues yielded detailed profiles of two mammalian transcriptomes (Su et al. 2002). We used these expression data, along with comparative human-mouse orthologous gene sequence analysis, to evaluate the relationship between gene sequence evolution and gene expression divergence on a genomic scale. For each human-mouse orthologous gene pair, the number of tissues where it is expressed (expression breadth) and total level of expression were determined (see Materials and Methods). These values were compared to several measures of human-mouse orthologous gene sequence divergence: the synonymous (dS) and nonsynonymous (dN) protein-coding sequence substitution rates as well as the substitution rates (d) for the 5′ and 3′ untranslated regions (UTRs). As reported previously (Duret and Mouchiroud 2000; Pal, Papp, and Hurst 2001; Zhang and Li 2003; Zhang and Li 2004), genes that are more widely expressed and genes that are more highly expressed are more evolutionarily conserved (i.e., evolve more slowly) than genes with narrower and lower overall levels of expression (fig. 1 and table 1, and see supplementary figure 1). This suggests that relatively slow evolution of highly expressed genes depends more on the strict functional constraints on the protein structure than on adaptation of codon usage for expression level. The connection between gene expression and evolutionary rate was far stronger for the 3′ UTRs than for the 5′ UTRs (fig. 1 and table 1, and see supplementary figure 1). This seems to indicate that, on average, 3′ UTRs contain more cis-regulatory sites that are functionally constrained in highly expressed genes than do 5′ UTRs. Overall, the 5′ and 3′ UTRs have similar rates of evolution: the median of the ratio d(5′ UTR)/d(3′ UTR) is approximately 1.12 (see supplementary figure 2a). Furthermore, both 3′ UTRs and 5′ UTRs were found to evolve somewhat slower on average than the synonymous positions (see supplementary figures 2b and c). Thus, 5′ UTRs appear to evolve under functional constraints that are nearly as strong as those that affect the 3′ UTRs and even stronger than those for the synonymous positions of the coding region. However, for the 5′ UTRs, these constraints appear to be unrelated to expression breadth and level as measured here.

**Human Gene Coexpression Network**

Tissue-specific expression patterns of human genes were further compared to identify coexpressed genes. For each differentially expressed human gene, the normalized expression levels in each of the 31 tissues were used to construct a vector, which was compared with similarly derived vectors of other human genes using the Pearson correlation coefficient ($r$) (Eisen et al. 1998). Pairs of genes with high $r$-values are considered to be coexpressed. Values of $r$ were determined for all pairs of human genes. These data were used to infer a network of coexpressed genes, where the genes are nodes that are connected by an edge if they share an $r$-value greater than or equal to a specified threshold. This was done using a series of $r$-value...
Fig. 1.—The dependence between expression breadth, expression level, and substitution rates of human genes. (a–d) Average (+ standard error) human gene expression breadth values for 10 ascending bins of human-mouse orthologous gene substitution rates. (e–h) Average (+ standard error) human gene expression level values for 10 ascending bins of human-mouse orthologous gene substitution rates.
thresholds (0.9 - 0.4), and the topological properties of
the resulting series of networks were investigated by
analyzing their node degree distributions; that is, the
frequency distributions of the number of genes, f(n), that
have n coexpressed genes. As the r-value threshold
increases, the number of edges in the network decreases,
and the node degree distribution seems to tend to a power
law distribution (see supplementary figure 3). Node degree
distributions and graphic representations of the corre-
sponding network topologies, for r/C21 0.9, 0.8, and 0.7,
are shown in figure 2. Because the distribution for r ≥ 0.7
shows a good fit to a distribution with a power law tail
while still retaining enough data for meaningful statistical
analysis (fig. 2, and see supplementary figure 3), the
threshold of 0.7 was chosen for further analysis. A list
of coexpressed gene pairs at r ≥ 0.7 is shown at ftp://
SupplementaryTable1.tab.

The node degree distribution for r ≥ 0.7 displays a
good fit to a generalized Pareto distribution where f(n) ∼
(n + 1.34)−1.17 (fig. 3a). This distribution has a power law
tail, which implies asymptotically scale-free properties
(Barabasi and Albert 1999). Such scale-free networks have
no single characteristic node degree and are dominated by
a small number of highly connected hubs (Barabasi and
Albert 1999); in this case, genes that are coexpressed with
many other genes. Similar scale-free properties have been
previously detected for node degree distributions of several
other gene expression–related parameters. For instance,
gene expression levels (Hoyle et al. 2002; Kuznetsov,
Knott, and Bonner 2002; Luscombe et al. 2002), as well as
changes in gene expression level (Ueda et al. 2004), have
been shown to follow power law distributions. Coexpression
networks derived with different techniques for several
cancers (Agrawal 2002) and for yeast, plants, and animals
(Bhan, Galas, and Dewey 2002; Bergmann, Ihmels, and
Barkai 2004) also show scale-free properties. Finally,
a number of other biological and other evolving networks,
including protein-protein interactions, metabolic networks,
social contacts networks, and the Internet (Barabasi and
Albert 1999; Jeong et al. 2000, 2001; Ravasz et al. 2002;
Barabasi and Oltvai 2004) have node degree distributions
that follow a power law and, thus, imply scale-free
properties. The principal mode of evolution of scale-free
networks is thought to be preferential attachment of new
nodes to those that are already highly connected (i.e., “the
rich get richer” or “the fit get fitter” model [Barabasi and
Albert 1999]).

Obviously, there is a transitive property to the
correlation coefficients that are used to measure coex-
pression and connect nodes in the coexpression network. If
the tissue-specific expression pattern of gene A is cor-
related with that of gene B, and the pattern of gene B is
correlated with that of gene C, then one should expect
expression of gene A to be correlated with that of gene C.
However, the level of correlation between A and C is an
open question because some groups of genes can be tightly
coregulated, whereas others are only loosely coexpressed.
To further evaluate the topological properties of the human gene coexpression network, the clustering of network nodes was analyzed. The clustering coefficient \((C)\) of a given node is the ratio of the number of the actual connections between the neighbors of the node to the number of possible connections between them (Barabasi and Oltvai 2004). The average \(C\) for the gene coexpression network is 0.452. This indicates a high degree of clustering that is typical of many scale-free networks (Barabasi and Oltvai 2004); however, the fact that \(C \approx 1\) indicates only limited transitivity in the gene coexpression network analyzed here. The shape of the dependence of \(C\) on the node degree is thought to be indicative of the mode of network growth (Barabasi and Oltvai 2004). Furthermore, a plot of the clustering coefficient, \(C(n)\), versus node degree \((n)\) shows the absence of any clear trend between the two (fig. 3b). \(C(n) \approx n^{-1}\) (Barabasi and Oltvai 2004).

The human gene coexpression network was examined to assess the functional relationships between coexpressed genes. Pairs of coexpressed genes were functionally classified using the eukaryotic KOG database (Tatusov et al. 2003) and the Gene Ontology database (Ashburner et al. 2000). Using both of these approaches, approximately 17.5% of coexpressed gene pairs were found to encode proteins with the same functional classification, a significant \((P < 0.0001)\) but relatively small excess over the random expectation (~13%). However, far more striking nonrandomness was observed when the coexpressed genes were examined for co-occurrence of different functions: certain combinations of seemingly related functions were strongly preferred (ftp://ftp.ncbi.nih.gov/pub/koonin/Jordan/MBE-04–0138.R1/SupplementaryTable2.tab). For example, signal transduction genes are coexpressed with those involved in secretion, transcription, and posttranslational protein modification much more often than expected by chance. This is likely to reflect coexpression of genes coding for proteins that function together in biochemical and signaling pathways.

A number of highly connected individual clusters from the human gene coexpression network were further examined with respect to the expression patterns and functions of their member genes. The majority of these clusters are made up of genes with narrow, if not exclusive, tissue-specific expression patterns (see supplementary figure 4). Examples of the topologies of two of these clusters, along with their tissue-specific expression patterns, are shown in figure 4. The experimentally determined and predicted functions for the member genes of these two clusters are generally consistent with their expression patterns and indicate their involvement in pancreatic and testis-related physiological functions, respectively (table 2).

Conservation and Coevolution of Coexpressed Human Genes

For each human gene with an identifiable mouse ortholog, the number of other human genes that it is coexpressed with at \(r \geq 0.7\) (ftp://ftp.ncbi.nih.gov/pub/koonin/Jordan/MBE-04–0138.R1/SupplementaryTable3.tab) was compared with its human-mouse substitution rate (fig. 5 and table 1). Genes that have a higher number of coexpressed genes (the node degree in the transcription expression network) are substantially more conserved than genes with fewer coexpressed partners. Qualitatively identical results are seen when the mouse expression data is used to compare the number of coexpressed mouse genes with human-mouse evolutionary rates (see supplementary figure 5a). As with the relationship between expression level and evolutionary rate, the negative correlation between the number of coexpressed genes and evolutionary rates was strongest for \(dN\) (fig. 5). Thus, the protein products of those genes that are hubs of the transcription coexpression network are subject to comparatively high levels of functional constraint, which could reflect their essential roles in cellular processes (Hirsh and Fraser 2001; Jordan et al. 2002; Krylov et al. 2003), multifunctionality, and/or involvement in a large number of protein-protein interactions (Fraser et al. 2002; Jordan, Wolf, and Koonin 2003). The negative correlation between \(dS\) and \(d(3’\) UTR) and the number of coexpressed genes (fig. 5) is likely to result from purifying selection maintaining tight translational regulation of genes that are highly connected in the coexpression network. In contrast, the correlation between \(d(5’\) UTR) and the number of coexpressed genes was not significant (fig. 5 and table 1).

---

**Fig. 3.**—Statistical properties of the human gene coexpression network. (a) Node degree distribution. Number of genes, \(f(n)\), that have exactly \(n\) coexpressed genes. Equation and line for the best fitting distribution are shown. Probability associated with the \(\chi^2\) test for the goodness of fit of the observed data to the theoretical generalized Pareto distribution is shown. (b) Clustering coefficient (\(y\)-axis) plotted against the node degree (\(x\)-axis). Average clustering coefficients and standard deviations of the averages for node degree bins are shown.
Although the correlation between dN and the number of coexpressed genes was nearly as strong as that between dN and expression breadth and level, for dS, the latter correlation was substantially stronger (table 1). This is not surprising, because codon usage adaptation is particularly important for highly expressed genes (Carbone, Zinovyev, and Kepes 2003). Partial correlation, \( r_{XY|Z} \) where \( X = \) number of coexpressed genes, \( Y = \) substitution rate, and \( Z = \) expression level, was used to control for the effects of expression level on the relationship between the number of coexpressed genes and substitution rates. In all cases where the substitution rate was found to be significantly correlated with the number of coexpressed genes—dN, dS, and d(3' UTR) (table 1)—the application of partial correlation did not result in any significant decrease of the correlation coefficient; that is, \( r_{XY|Z} \) is not significantly smaller than \( r_{XY} \) (0.38, \( P < 0.56 \)). Thus, the effect of the number of coexpressed genes on substitution rates is not based on any correlation between the number of coexpressed genes and the expression level.

The relationship between gene coexpression and evolutionary rate was further examined by comparing the \( r \)-values between pairs of human gene expression patterns with their pairwise gene substitution rate differences (fig. 6 and table 1). The results show that bins of human genes with similar values of dN between human and mouse (i.e., those genes that evolve at similar rates) have a greater fraction of pairwise \( r \)-values \( \geq 0.7 \). This negative correlation was substantial and statistically significant (fig. 6 and table 1). In contrast, the pairwise dS difference between human genes did not correlate with the difference in expression patterns (fig. 6). Thus, genes with similar patterns of expression tend to evolve at similar rates, and the selection that leads to such coevolution appears to operate primarily at the protein sequence level. The difference in evolution rates of 5' and 3' UTRs also negatively correlated with pairwise \( r \)-values between human gene expression patterns, although the effect, statistically significant because of the large number of analyzed gene pairs (table 1), was not nearly as pronounced as seen for dN (fig. 6). As with other comparisons between evolutionary rate and expression patterns, the magnitude of this relationship was greater for 3' UTRs, suggesting that this region is more functionally relevant than the 5' UTR with respect to the pattern of gene expression. Qualitatively identical results are seen when the mouse expression data was used to compare \( r \)-values between pairs of mouse genes with pairwise human-mouse gene substitution rate differences (see supplementary figure 5b).

The connection between gene coexpression and evolutionary conservation on a longer timescale was investigated by considering the expression data along with the patterns of phyletic distribution inferred from the recently developed collection of eukaryotic KOGs (Tatusov et al. 2003; Koonin et al. 2004). Each human gene was mapped to a specific KOG and assigned...
Fig. 5.—The dependence between the node degree in the human coexpression network and substitution rate. Average (± standard error) numbers of coexpressed human genes \( r \geq 0.7 \) per gene for 10 ascending bins of human-mouse orthologous gene substitution rates are shown.

Fig. 6.—Coexpressed human genes have similar substitution rates. Fractions of pairwise correlation coefficient values, where \( r \geq 0.7 \) between human genes, are shown for 10 ascending bins of pairwise human-mouse orthologous gene substitution rate differences.
a phyletic pattern; that is, the pattern of presence/absence of orthologous genes from the given KOG among seven eukaryotic species. The phyletic patterns of KOGs were compared to determine their similarity in terms of propensities for gene gain and loss over the approximately 1.8 billion years of eukaryotic crown group evolution (Krylov et al. 2003). A statistically significant positive correlation was detected between the similarity of evolutionary patterns of gain-loss and the level of coexpression among pairs of human genes (fig. 7 and table 1). Thus, coexpressed genes tend to evolve under similar long-term evolutionary constraints.

Human-Mouse Sequence Divergence Versus Expression Profile Divergence

Expression data collected for human and mouse were further compared to assess levels of regulatory divergence (divergence of expression profiles) between orthologous genes in the two species. From the original series of microarray experiments (Su et al. 2002), 19 tissues that were studied in both human and mouse were identified. Relative levels of gene expression across these tissues in both species were taken as vectors, and pairs of vectors for human-mouse orthologous genes were compared to measure the extent of regulatory divergence between species. In a general agreement with the original report (Su et al. 2002), the \( r \)-values for orthologs displayed a narrow distribution with the median at approximately 0.4. This was in sharp contrast to the distribution of \( r \)-values for nonorthologous human and mouse genes, which had a median of approximately 0 (fig. 8). The difference between the two distributions is highly statistically significant (\( P \ll 10^{-10} \)). Surprisingly, however, when the resulting \( r \)-values were compared with the levels of sequence divergence to determine whether sequence and regulatory divergence were correlated, no significant correlation between the human-mouse expression pattern divergence and the four different measures, \( dN \), \( dS \), \( d(5’ \text{ UTR}) \), and \( d(3’ \text{ UTR}) \), of human-mouse sequence divergence was detected (fig. 9 and table 1). This finding stands in contrast with the results of the recent analysis of the connection between the sequence and expression divergence between paralogous human genes, where a significant negative correlation was observed between the correlation coefficients of the expression profiles and sequence divergence (Makova and Li 2003). However, an earlier study of yeast duplicate genes found no correlation.
between gene sequence and gene expression pattern divergence (Wagner 2000). The apparent disparity between orthologs and paralogs in terms of expression evolution in mammals seems to suggest unexpected differences in the evolutionary modes and deserves further investigation.

Conclusion

The results described here indicate that mammalian coexpressed genes form a scale-free network, which probably evolved by self-organization based on the preferential attachment principle. One possible mechanism of this evolution is the duplication of genes together with their transcriptional control regions, as suggested for the yeast coexpression network (Bhan, Galas, and Dewey 2002). An examination of the contribution of paralogous gene pairs to the gene coexpression network analyzed here indicates that such duplicated pairs (see Materials and Methods) are found approximately two times more frequently than expected by chance ($P_{\leq 10^{-2}}$). However, the fraction of coexpressed gene pairs that are related by duplication is small (0.28%), suggesting that duplication may not be a major force affecting the structure of the gene coexpression network. This finding seems to be consistent with an analysis of yeast protein

Table 2
Examples of Two Human Gene Coexpression Network Clusters

<table>
<thead>
<tr>
<th>Locus Link ID/GenBank accession</th>
<th>Gene Symbol</th>
<th>Annotation</th>
<th>Physiological Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1208/P16233</td>
<td>CLPS</td>
<td>colipase preproprotein</td>
<td>Pancreatic digestive enzyme</td>
</tr>
<tr>
<td>1357/NP_001859</td>
<td>CPA1</td>
<td>pancreatic carboxypeptidase A1 precursor</td>
<td>Pancreatic digestive enzyme</td>
</tr>
<tr>
<td>1358/NP_001860</td>
<td>CPA2</td>
<td>pancreatic carboxypeptidase A2 precursor</td>
<td>Pancreatic digestive enzyme</td>
</tr>
<tr>
<td>2641/P01275</td>
<td>GCG</td>
<td>glucagon preproprotein</td>
<td>Pancreatic islet hormone</td>
</tr>
<tr>
<td>2813/AAB19240</td>
<td>GP2</td>
<td>glycoprotein 2 (zymogen granule membrane)</td>
<td>The major membrane protein in the exocrine pancreas</td>
</tr>
<tr>
<td>2906/AAB49992</td>
<td>GRIN2D</td>
<td>N-methyl-D-aspartate receptor subunit 2D precursor</td>
<td>Implicated in pancreatic hormone secretion</td>
</tr>
<tr>
<td>148223/NP_689695</td>
<td>SPDEF</td>
<td>SAM pointed domain-containing eukaryotic transcription factor</td>
<td>Prostate epithelium-specific transcription factor</td>
</tr>
<tr>
<td>25803/NP_036523</td>
<td>84444/NP_115871</td>
<td>histone methyltransferase DOT1L</td>
<td>Regulator of gene dynamics and chromatin activity</td>
</tr>
<tr>
<td>9436/NP_004819</td>
<td>NCR2</td>
<td>natural cytotoxicity triggering receptor–2</td>
<td>Mediator of killer cells’ cytotoxicity</td>
</tr>
<tr>
<td>9610/AAB67270</td>
<td>RIN1</td>
<td>ras inhibitor RIN1</td>
<td>A dominant negative inhibitor of the Ras signaling pathway</td>
</tr>
<tr>
<td>9753/AAH41661</td>
<td>ZNF305</td>
<td>zinc finger protein–305</td>
<td>Predicted transcription regulator</td>
</tr>
<tr>
<td>6847/Q15431</td>
<td>SYCP1</td>
<td>synaptonemal complex protein–1</td>
<td>Major component of the transverse filaments of synaptonemal complexes during meiotic prophase</td>
</tr>
<tr>
<td>10388/Q9BX26</td>
<td>SYCP2</td>
<td>synaptonemal complex protein–2</td>
<td>Major component of the axial/lateral elements of synaptonemal complexes during meiotic prophase</td>
</tr>
<tr>
<td>27285/NP_055281</td>
<td>TEK2</td>
<td>tektin 2</td>
<td>Major sperm microtubule protein</td>
</tr>
<tr>
<td>8852/NP_647450</td>
<td>AKAP4</td>
<td>A-kinase anchor protein–4 isoform 2</td>
<td>Major sperm sheath protein</td>
</tr>
<tr>
<td>676/AAB87862</td>
<td>BRDT</td>
<td>testis-specific bromodomain protein</td>
<td>Testis-specific chromatin-associated protein involved in chromatin remodeling</td>
</tr>
<tr>
<td>11077/OT5031</td>
<td>HSF2BP</td>
<td>heat shock transcription factor–2 binding protein</td>
<td>Implicated in modulating HSF2 activation in testis</td>
</tr>
<tr>
<td>7180/P16562</td>
<td>CRISP2</td>
<td>testis-specific protein–1</td>
<td>Probable sperm-coating glycoprotein</td>
</tr>
<tr>
<td>11116/NP_919410</td>
<td>FGFR1OP</td>
<td>FGFR1 oncogene partner isoform b</td>
<td>Implicated in the proliferation and differentiation of the erythroid lineage</td>
</tr>
<tr>
<td>4291/P85820</td>
<td>MLF1</td>
<td>myeloid leukemia factor–1</td>
<td>Uncharacterized protein implicated in cell proliferation</td>
</tr>
<tr>
<td>5261/AAH02541</td>
<td>PHKG2</td>
<td>phosphorylase kinase, gamma 2 (testis)</td>
<td>Testis/liver-specific regulator of glycogen metabolism</td>
</tr>
<tr>
<td>54344/O94777</td>
<td>DPM3</td>
<td>dolichyl-phosphate mannosyltransferase polypeptide–3 isoform 2</td>
<td>Regulator of dolichyl-phosphate mannosylbiosynthesis. Might be important for sperm maturation</td>
</tr>
<tr>
<td>7288/00029</td>
<td>TULP2</td>
<td>tubby-like protein–2</td>
<td>Retina-specific and testis-specific heterotrimeric G-protein–responsive intracellular signaling factor</td>
</tr>
<tr>
<td>51460/NP_057413</td>
<td>SFMBT1</td>
<td>Scm-like with four mbt domains–1</td>
<td>Polycistron group transcription regulator</td>
</tr>
<tr>
<td>114049/NP_684281</td>
<td>WBSCR22</td>
<td>Williams Beuren syndrome chromosome region 22 protein</td>
<td>iRNA 5–cytosine methylase</td>
</tr>
</tbody>
</table>


* Official gene symbol from the Human Genome Organization (HUGO), http://www.gene.ucl.ac.uk/nomenclature/

* Official HUGO gene name.
interaction networks, which showed that the network structure is not predominantly shaped by duplication (Wagner 2003).

Our results show that overall levels of expression and more specific topological properties of the gene co-expression network are clearly related to the rates of sequence evolution: highly connected network hubs tend to evolve slowly, and genes that are coexpressed show a strong tendency to evolve at similar rates. These relationships most likely reflect purifying selection, which appears to act primarily at the level of the protein sequence. Generally, these results are compatible with the notion of greater biological importance of highly connected nodes of biological networks (Jeong et al. 2001). However, two of the findings reported here appear to be distinctly surprising. Firstly, we found that the connection between gene coexpression network topology and sequence evolution held for both nonsynonymous and synonymous sites in the coding sequence and 3' UTR but not for the 5' UTR. Thus, the constraints on 5' UTR evolution appear to be unrelated to expression regulation as analyzed here. Secondly, we found that, although the expression profiles of orthologs tend to be strongly correlated, they diverged at roughly the same rate across a wide range of sequence evolution rates. Thus, unlike the properties of the gene coexpression network, evolution of an individual gene’s expression regulation after speciation seems to be uncoupled from the functional constraints on the gene’s sequence. This is generally compatible with the pregenomic notion that regulatory evolution could be the decisive factor of biological diversification between species (Britten and Davidson 1969; King and Wilson 1975) and can be taken to suggest that sequence divergence and expression pattern divergence are governed by distinct forces.

Recent studies have reported the rapid diversification of gene expression patterns and suggested that this might reflect neutral evolution of transcriptional regulation (Khaitovich et al. 2004; Yanai, Graur and Ophir 2004). The lack of correlation between sequence divergence and expression profile divergence demonstrated here could be deemed as compatible with the neutral model whereby transcription profiles of orthologs diverge rapidly upon speciation until they reach a basal level of similarity that is then maintained by purifying selection. However, it is tempting to speculate as to a distinct role for natural selection in driving expression pattern divergence. Perhaps, whereas sequence divergence levels are determined largely by purifying selection, the effects of adaptive (diversifying) selection are more prevalent at the level of gene expression. Under this hypothetical scenario, although the basal level of correlation between orthologs tends to persist, probably reflecting the conservation of the general function, the aspects of the gene expression patterns that reflect species-specific changes, governed in part by short, degenerate transcription factor–binding sites, would be expected to diverge rapidly. Indeed, our analysis of the relationship between gene coexpression in this work and gene duplication, as well as previous results (Gu et al. 2002), suggest rapid divergence of expression patterns after duplication. Once the expression patterns of homologous
genes (paralogs or orthologs) diverge, convergent evolution, which is a hallmark of adaptation and is widely evident for phenotypic characters, could cause unrelated genes to achieve similar expression patterns. This convergence could be related to the rapid de novo generation of transcription factor–binding sites in promoters that lead to slight changes in expression patterns. If the functionally relevant aspects of the ancestral expression pattern are maintained during this process, subtle expression pattern changes would be simultaneously invisible to purifying selection and serve as the raw material for repeated trials of adaptive selection. This hypothesis yields a specific prediction with regard to the evolution of promoter regions that is currently being investigated. Expression profiles are predicted to be more correlated with the distributions of specific transcription factor–binding sites along promoters than with the overall sequence divergence (i.e., relatedness) between promoters.

Acknowledgments

We thank Igor Rogozin, David Landsman, Fyodor Kondrashov, Itai Yanai, and Katerina Makova for helpful discussions.

Supplementary Material

Supplementary Figure 1.—The dependence between expression breadth, level, and substitution rates of mouse genes. (a–d) Average (± standard error) expression breadth values for 10 ascending bins of human-mouse orthologous gene substitution rates. (e–h) Average (± standard error) expression level values for 10 ascending bins of human-mouse orthologous gene substitution rates.

Supplementary Figure 2.—Frequency distributions of human-mouse substitution rate ratios. Ratios shown in log scale. (a) d(5′ UTR)/d(3′ UTR). d(5′ UTR)/d(3′ UTR) is shown in log scale. In 43% of genes, d(5′ UTR) < d(3′ UTR), whereas in 57% d(3′ UTR) < d(5′ UTR). (b) d(5′ UTR)/dS. (c) d(3′ UTR)/dS.

Supplementary Figure 3.—Frequency distribution of the number of links per node in human gene coexpression networks. For each network, the correlation coefficient value (r) threshold used to determine whether any two genes (nodes) are considered to be coexpressed (linked) are shown above the plot. The slope of the linear trend line that fits the data and the r² value indicating the goodness of fit to the power law are shown for each network plot.

Supplementary Figure 4.—Tissue-specific expression patterns for human gene coexpression networks (r > 0.9) clusters. Cluster expression patterns: 1, adult and fetal liver; 2, fetal liver; 3, testis; 4, pancreas; 5, testis and umbilical vein; 6, various brain tissues; 7, salivary gland; 8, adult liver; 9, thymus; 10, spleen; 11, lung; 12, cerebellum.

Supplementary Figure 5.—Relationship between the mouse coexpression network topology and substitution rates. (a) The dependence between the node degree and substitution rate. Average (± standard error) numbers of coexpressed genes (node degree for r ≥ 0.7) per gene for 10 ascending bins of human-mouse orthologous gene substitution rates are shown. (b) Coexpressed mouse genes have similar substitution rates. Fractions of pairwise correlation coefficient values, where r ≥ 0.7 between mouse genes, are shown for 10 ascending bins of pairwise human-mouse orthologous gene substitution rate differences.

Literature Cited


few: power-law behaviour applies to a wide variety of genomic properties. Genome Biol. 3:RESEARCH0040.


Peer Bork, Associate Editor

Accepted July 14, 2004