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.

Key words: Gene expression, Human evolution, Natural selection, Network, Self-organization, Substitution rate.

E-mail: koonin@ncbi.nlm.nih.gov.

Mol. Biol. Evol. 21(11):2058–2070. 2004
doi:10.1093/molbev/msh222
Advance Access publication July 28, 2004

2000, 2001; Luscombe et al. 2002; Ravasz et al. 2002). Studies of the statistical properties of the topologies of such networks suggest specific mechanisms that govern their evolution. In particular, many biological networks show scale-free topological properties that can be explained by a model of network growth via preferential attachment of new nodes to existing nodes that are already highly connected (Barabasi and Albert 1999). At the genomic level, gene duplication is thought to underlie the phenomenon of preferential attachment (Rzhetsky and Gomez 2001; Bhan, Galas, and Dewey 2002; Barabasi and Oltvai 2004). Existing highly connected nodes (i.e., genes or proteins) are more likely, simply by virtue of their large number of connections, to be linked to nodes that are duplicated. Because the duplicated nodes are expected to maintain the same links as the ancestral singleton, the connectivity of a highly connected node will increase with duplication (Barabasi and Oltvai 2004). This process alone can lead to network growth by preferential attachment. The ubiquity of scale-free network topological patterns suggests that network growth by preferential attachment, via mechanisms such as gene duplication, is a fundamental and conserved evolutionary process.

In this work, we attempted to integrate the two perspectives of natural selection and physical self-organization to analyze the evolution of gene sequence and expression patterns. Comparison of human and mouse genome sequence data was combined with the analysis of gene expression profiles that were derived from microarray experiments on a number of tissues in both species. Human gene expression profiles were used to reconstruct a network of coexpressed genes, and we demonstrate the effect of the network topology on the strength of natural selection as well as an unexpected relationship between human-mouse gene sequence and gene expression divergence. These results underscore the influence of expression network self-organization on gene evolution and suggest that distinct mechanisms are responsible for the evolution of expression patterns and gene sequences.

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 corresponded to the probe identifiers. The results of microarray experiments on cancerous tissues were removed to yield profiles of normal mammalian transcriptomes. This left data from a total of 63 human and 89 mouse microarray experiments that were subsequently analyzed. Levels of expression, recorded as average difference (AD) values, for redundant experiments (i.e., studies of the same tissue samples) were averaged before analysis. Because negative AD values represent more noise than signal, AD values were clipped at a value of 20. For the purpose of determining the breadth of expression, an AD value of 200 was taken as a threshold to consider a gene to be expressed in a given tissue (Su et al. 2002), and the number of tissues where a gene was expressed was counted. For the purpose of determining the level of expression, the sum of the $\log_2$ normalized AD values over all tissues was taken. The similarity between gene expression patterns was determined using the Pearson correlation coefficient (r) (Eisen et al. 1998). Before this correlation analysis, unexpressed genes (AD $<$ 200 in all tissues) and nondifferentially expressed genes (maximum AD/minimum AD $<$ 10) were removed from the data matrices. In addition, to control for the effects of the overall level of gene expression, AD values were normalized for each gene by taking the $\log_2$ value of the ratio of the tissue-specific AD value/median AD value for all tissues.

Pairwise correlations between gene expression patterns were used to derive the gene coexpression network. The fit of the network node degree distribution—that is, the frequency distribution of the number of genes, $f(n)$, that have n coexpressed genes—to a theoretical distribution (the generalized Pareto distribution here) was done as previously described (Karev et al. 2002; Koonin, Wolf, and Karev 2002). It should be noted that, because the available node degree data span only approximately 2.5 orders of magnitude, a formal possibility remains that the coexpression data are also compatible with models other than the scale-free model adopted here. The clustering coefficient (C) was calculated for each node as 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).

Human and mouse gene sequences were extracted from 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 ($r \geq 0.7$) 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 its statistical significance (P-value at $n = 10$ for all comparisons), was computed for each comparison (table 1). Regular correlation coefficients (rXY) and partial correlation coefficients (rXY·Z) were Fisher Z–transformed, and the significance of the differences between them was calculated using a z-test.

Phyletic 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 phyletic pattern corresponds to the presence/absence state of the KOG member proteins among the seven eukaryotic species included in the database. For each phyletic 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 phyletic 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_i^{G}) and gene loss (P_i^{L}) were calculated using the total number of gene gains and losses mapped to each branch (Koonin et al. 2004) according to the following formula:

$$P_i^{\mathrm{G}} = G_i \sum_j N_j \Big/ N_i \sum_j G_j, \qquad P_i^{\mathrm{L}} = L_i \sum_j N_j \Big/ N_i \sum_j L_j$$

where N_i, G_i, and L_i are the numbers of present genes, gains, and losses mapped to the ith 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 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). The negative correlation between expression levels and substitution rates was more pronounced for dN than for dS (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 2*a*). Furthermore, both 3′ UTRs and 5′ UTRs were found to evolve somewhat slower on average than the synonymous positions (see supplementary figures 2*b* 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

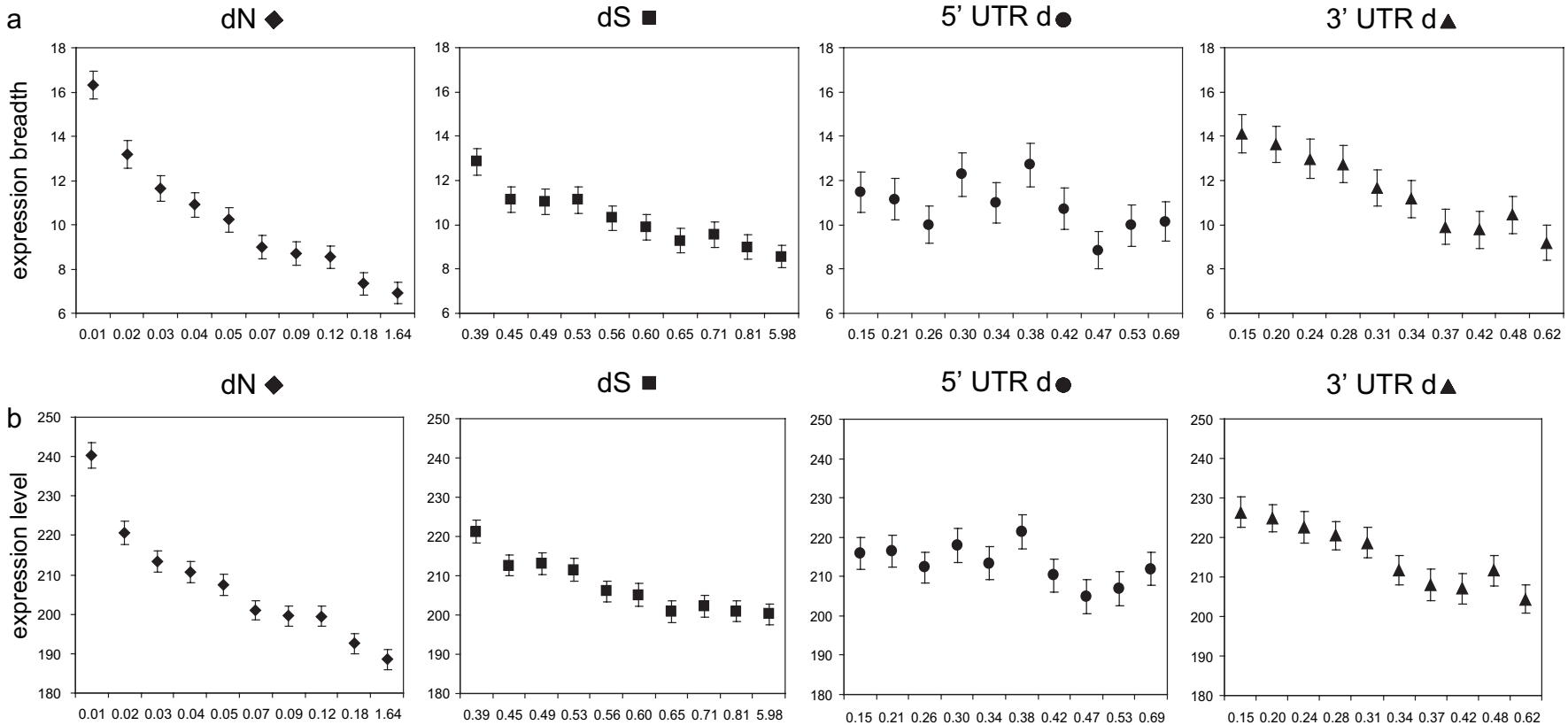

FIG. 1.—The dependence between expression breadth, expression level, and substitution rates of human genes. (*a–d*) Average (± standard error) human gene expression breadth values for 10 ascending bins of human-mouse orthologous gene substitution rates. (*e–h*) Average (± standard error) human gene expression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

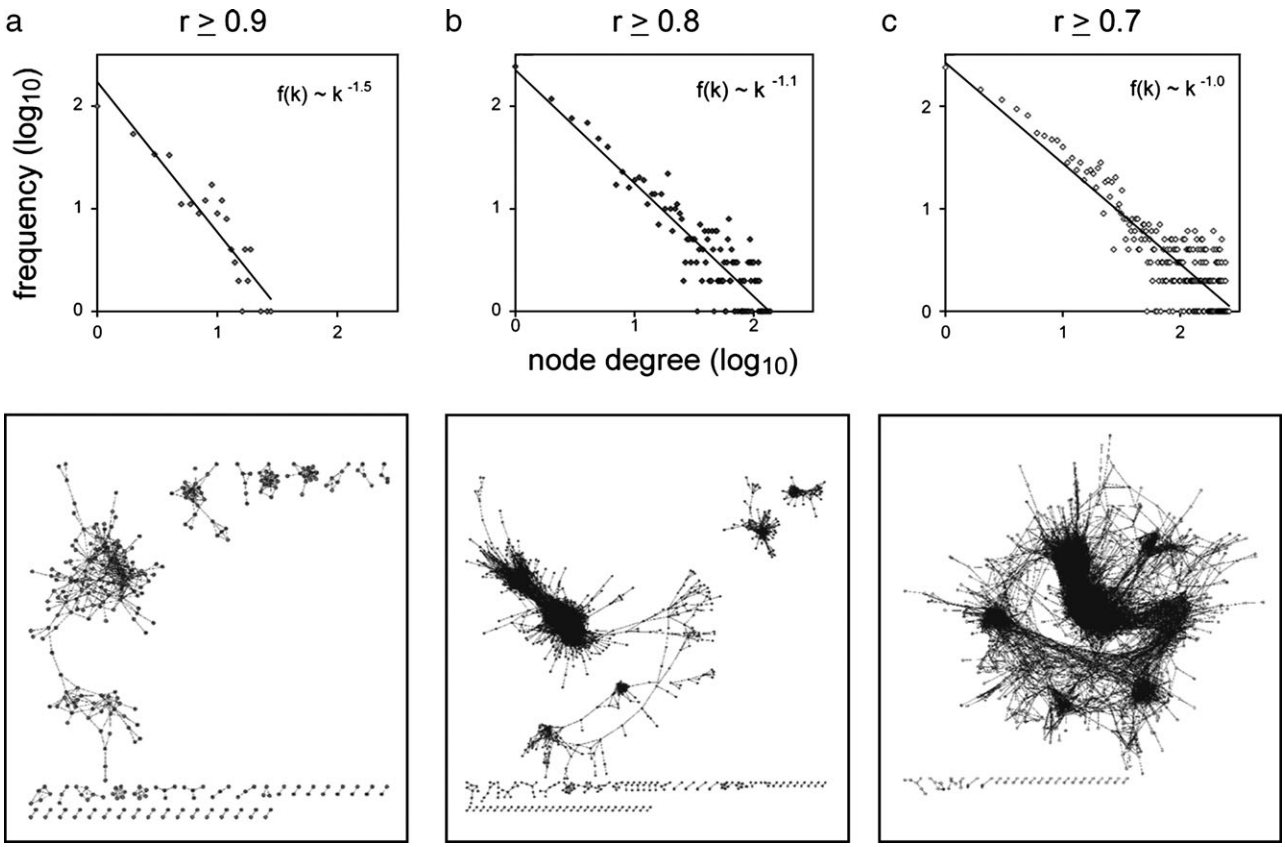

FIG. 2.—The human gene coexpression network. Node degree distributions and network topologies are shown for networks where coexpressed genes (nodes) are linked if (*a*) $r \geq 0.9$, (*b*) $r \geq 0.8$, and (*c*) $r \geq 0.7$.

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 corresponding network topologies, for $r \geq 0.9$, 0.8, and 0.7, are shown in figure 2. Because the distribution for $r \geq 0.7$ 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 $r \geq 0.7$ is shown at ftp://ftp.ncbi.nih.gov/pub/koonin/Jordan/MBE-04–0138.R1/SupplementaryTable1.tab.

The node degree distribution for $r \geq 0.7$ displays a good fit to a generalized Pareto distribution where $f(n) \approx (n + 1.34)^{-1.17}$ (fig. 3*a*). 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 coexpression and connect nodes in the coexpression network. If the tissue-specific expression pattern of gene A is correlated 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.

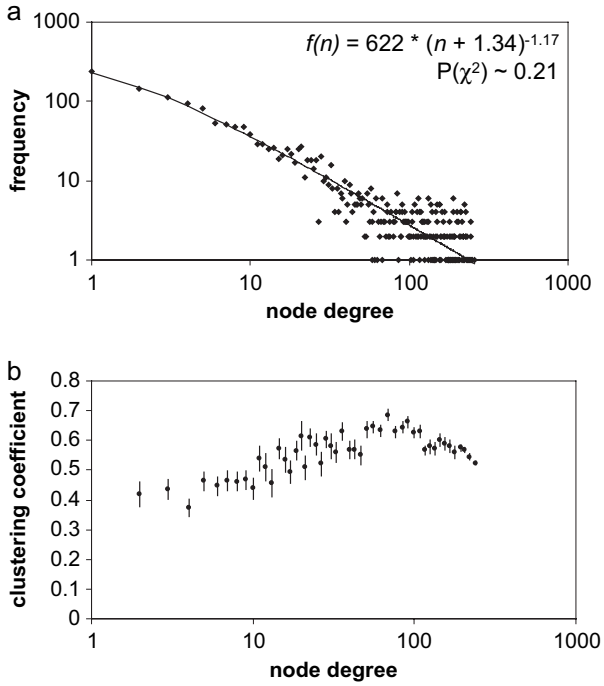

FIG. 3.—Statistical properties of the human gene coexpression network. (*a*) Node degree distribution. Number of genes, $f(n)$, that have exactly n coexpressed genes. Equation and line for the best fitting distribution are shown. Probability associated with the χ^2 test for the goodness of fit of the 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.

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 $C \ll 1$ 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. 3*b*). $C(n) \approx$ constant, as seen here, is consistent with network growth by simple preferential attachment of nodes rather than growth by the hierarchical addition of modules that is thought to be characterized by $C(n) \approx n^{-1}$ (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 5*a*). 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).

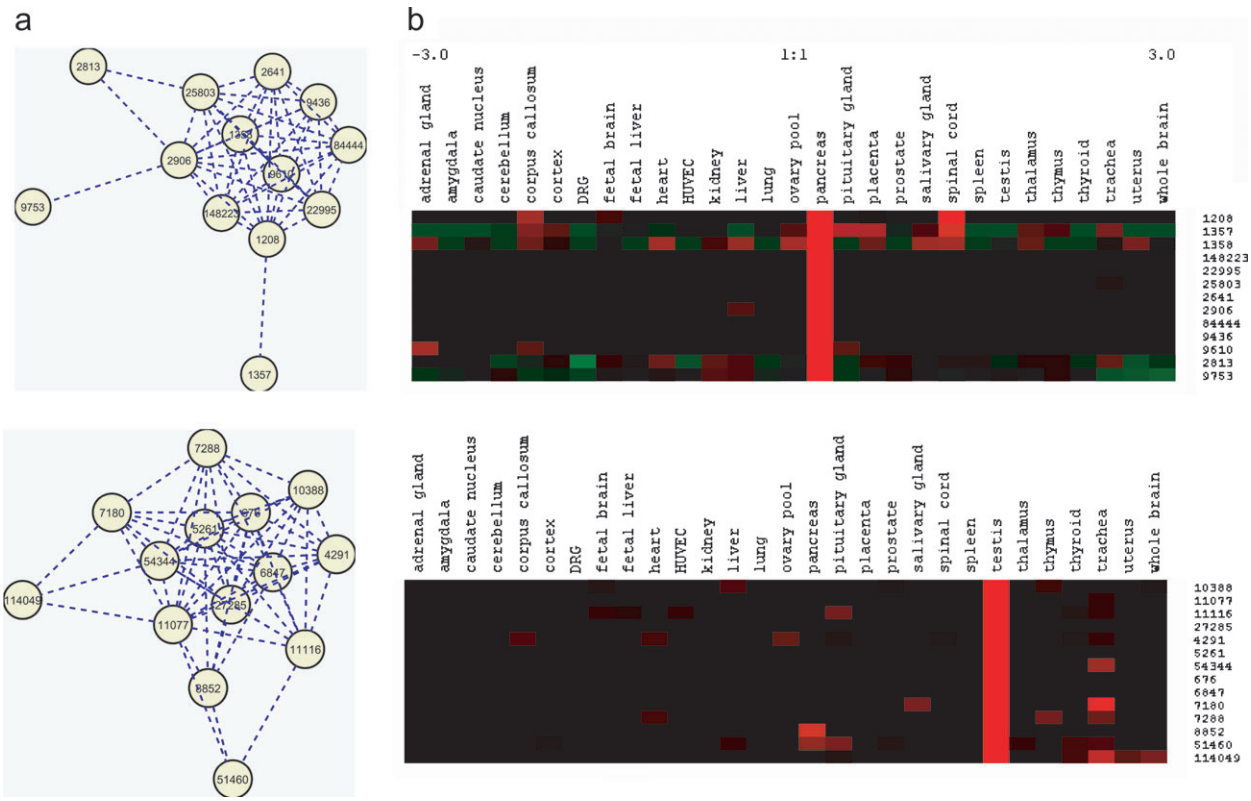

FIG. 4.—Two highly connected gene expression clusters form the human gene coexpression network. (*a*) Nodes correspond to genes and are labeled with LocusLink ID numbers. All links between genes correspond to coexpression at $r \geq 0.9$. (*b*) Tissue-specific gene expression patterns of the cluster members. Genes (rows) are labeled with LocusLink ID numbers. Relative levels ($\log_2$ ratios) of expression are shown for each gene in each tissue; green cells show underexpression relative to the median, black cells show median expression levels, and red cells show overexpression.

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, rXY·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, rXY·Z is not significantly smaller than rXY ($0.38 < P < 0.56$). 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 5*b*).

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

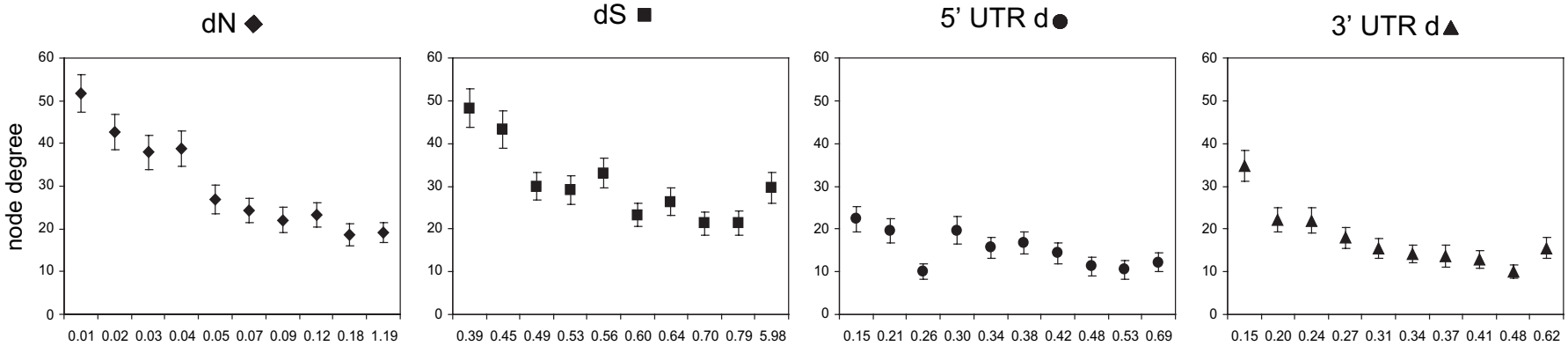

FIG. 5.—The dependence between the node degree in the human coexpression network and substitution rate. Average (± standard error) numbers of coexpressed human genes ($r \geq 0.7$) per gene for 10 ascending bins of human-mouse orthologous gene substitution rates are shown.

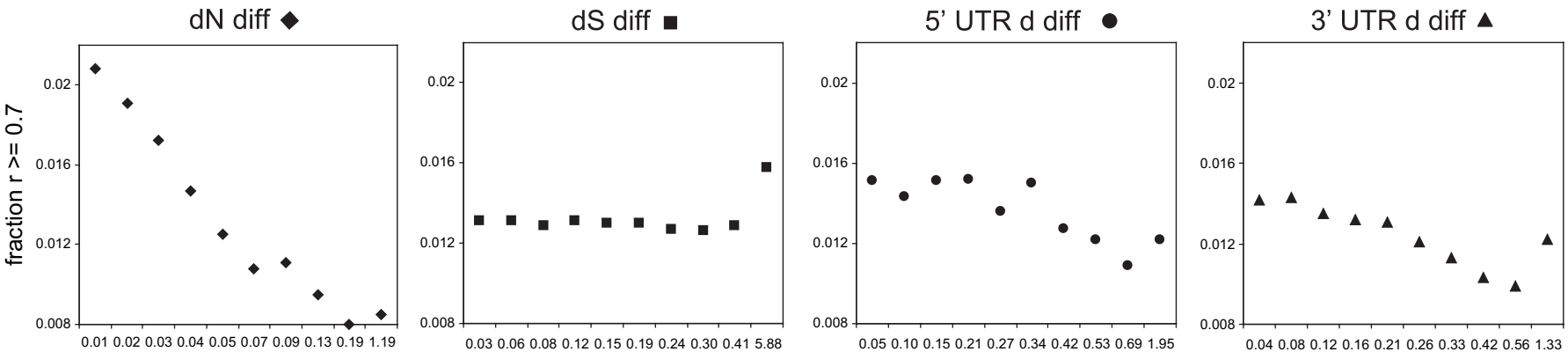

FIG. 6.—Coexpressed human genes have similar substitution rates. Fractions of pairwise correlation coefficient values, where $r \geq 0.7$ between human genes, are shown for 10 ascending bins of pairwise human-mouse orthologous gene substitution rate differences.

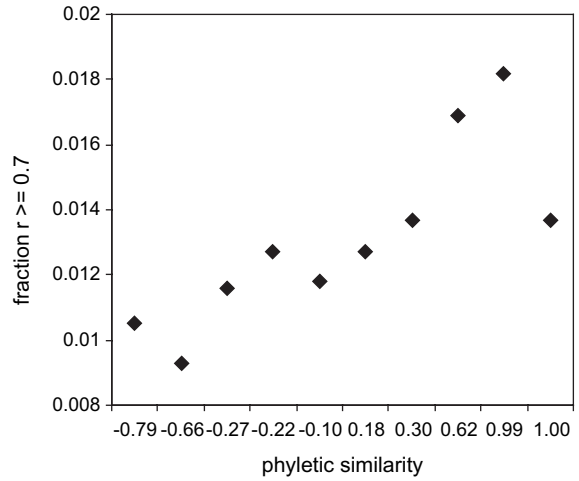

FIG. 7.—Coexpressed human genes have similar phyletic patterns. The fractions of coexpressed gene pairs ($r \geq 0.7$) are shown for 10 ascending bins of phyletic similarity values (see *Materials and Methods*).

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

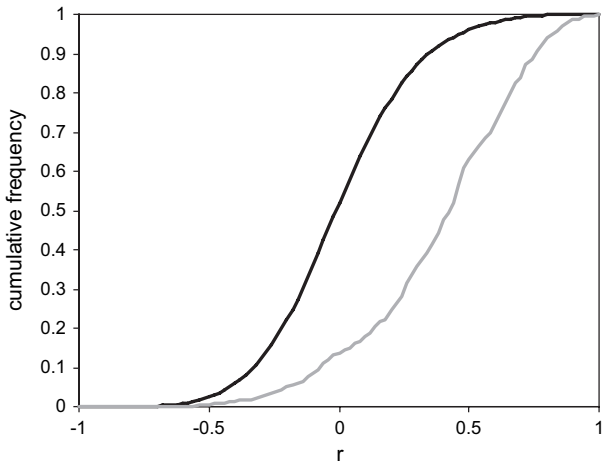

FIG. 8.—Cumulative distributions of pairwise correlation coefficient values (r). The black line is the distribution for the comparisons of nonorthologous genes within the human and mouse gene sets. The gray line is the distribution for orthologous human-mouse gene comparisons.

Table 1
Evolutionary Rate Versus Gene Expression Comparisons

Comparison	R[a]	P[b]	n[c]
dN vs. expression breadth	−1.000	5.5e–7	3,906
dS vs. expression breadth	−0.976	2.1e–5	3,906
5′ UTR d vs. expression breadth	−0.539	ns	1,121
3′ UTR d vs. expression breadth	−0.964	4.9e–5	1,438
dN vs. expression level	−1.000	5.5e–7	3,906
dS vs. expression level	−0.952	1.1e–4	3,906
5′ UTR d vs. expression level	−0.624	ns	1,121
3′ UTR d vs. expression level	−0.964	4.9e–5	1,438
dN vs. number coexpressed genes	−0.964	4.9e–5	2,307
dS vs. number coexpressed genes	−0.746	1.7e–2	2,307
5′ UTR d vs. number coexpressed genes	−0.588	ns	1,119
3′ UTR d vs. number coexpressed genes	−0.818	5.8e–3	1,436
dN difference vs. fraction $r \geq 0.7$	−0.976	2.1e–5	2,659,971
dS difference vs. fraction $r \geq 0.7$	−0.249	ns	2,659,971
5′ UTR d difference vs. fraction $r \geq 0.7$	−0.812	5.8e–3	625,521
3′ UTR d difference vs. fraction $r \geq 0.7$	−0.867	2.2e–3	1,030,330
dN vs. expression divergence	0.139	ns	551
dS vs. expression divergence	−0.079	ns	551
5′ UTR d vs. expression divergence	0.164	ns	273
3′ UTR d vs. expression divergence	−0.552	ns	382

[a] Spearman rank correlation coefficient for the average expression parameter value against the 10 bins of substitution rate values (see figures 1 and 3–5).

[b] Level of significance (n = 10) associated with R values; ns = nonsignificant (i.e., $P \leq 0.05$)

[c] The total number of comparisons that were binned.

(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′ UTR), and d(3′ 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

Table 2
Examples of Two Human Gene Coexpression Network Clusters

Locus Link ID/GenBank accession[a]	Gene Symbol[b]	Annotation[c]	Physiological Function
A Pancreas-Specific Gene Cluster			
1208/P16233	CLPS	colipase preproprotein	Pancreatic digestive enzyme
1357/NP_001859	CPA1	pancreatic carboxypeptidase A1 precursor	Pancreatic digestive enzyme
1358/NP_001860	CPA2	pancreatic carboxypeptidase A2 precursor	Pancreatic digestive enzyme
2641/P01275	GCG	glucagon preproprotein	Pancreatic islet hormone
2813/AAB19240	GP2	glycoprotein 2 (zymogen granule membrane)	The major membrane protein in the secretory granule of the exocrine pancreas
2906/AAB49992	GRIN2D	N-methyl-D-aspartate receptor subunit 2D precursor	Implicated in pancreatic hormone secretionPMID: 7768362
148223/ NP_689695		hypothetical protein FLJ36666	Uncharacterized protein
25803/NP_036523	SPDEF	SAM pointed domain–containing ets transcription factor	Prostate epithelium-specific transcription factor
84444/NP_115871		histone methyltransferase DOT1L	Regulator of gene dynamics and chromatin activity
9436/NP_004819	NCR2	natural cytotoxicity triggering receptor–2	Mediator of killer cells'cytotoxicity
9610/AAB67270	RIN1	*ras* inhibitor RIN1	A dominant negative inhibitor of the RAS signaling pathway
9753/AAH41661	ZNF305	zinc finger protein–305	Predicted transcription regulator
A Testis-Specific Gene Cluster			
6847/Q15431	SYCP1	synaptonemal complex protein–1	Major component of the transverse filaments of synaptonemal complexes during meiotic prophase
10388/ Q9BX26	SYCP2	synaptonemal complex protein–2	Major component of the axial/lateral elements of synaptonemal complexes during meiotic prophase
27285/NP_055281	TEKT2	tektin 2	Major sperm microtubule protein
8852/NP_647450	AKAP4	A-kinase anchor protein–4 isoform 2	Major sperm sheath protein
676/ AAB87862	BRDT	testis-specific bromodomain protein	Testis-specific chromatin-associated protein involved in chromatin remodeling
11077/ O75031	HSF2BP	heat shock transcription factor–2 binding protein	Implicated in modulating HSF2 activation in testis
7180/P16562	CRISP2	testis-specific protein–1	Probable sperm-coating glycoprotein
11116/NP_919410	FGFR1OP	FGFR1 oncogene partner isoform b	Implicated in the proliferation and differentiation of the erythroid lineage
4291/P58340	MLF1	myeloid leukemia factor–1	Uncharacterized protein implicated in cell proliferation
5261/AAH02541	PHKG2	phosphorylase kinase, gamma 2 (testis)	Testis/liver-specific regulator of glycogen metabolism
54344/ O94777	DPM3	dolichyl-phosphate mannosyltransferase polypeptide–3 isoform 2	Regulator of dolichyl-phosphate mannose biosynthesis. Might be important for sperm maturation PMID: 6231179
7288/O00295	TULP2	tubby-like protein–2	Retina-specific and testis-specific heterotrimeric G-protein–responsive intracellular signaling factor
51460/NP_057413	SFMBT1	Scm-like with four mbt domains–1	Polycomb group transcription regulator
114049/ NP_684281	WBSCR22	Williams Beuren syndrome chromosome region 22 protein	tRNA 5-cytosine methylase

[a] Locus Link identification numbers and GenBank accessions from NCBI's LocusLink database, http://www.ncbi.nlm.nih.gov/LocusLink/index.html.
[b] Official gene symbol from the Human Genome Organization (HUGO), http://www.gene.ucl.ac.uk/nomenclature/
[c] Official HUGO gene name.

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 ($P \ll 10^{-10}$). 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

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 coexpression 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

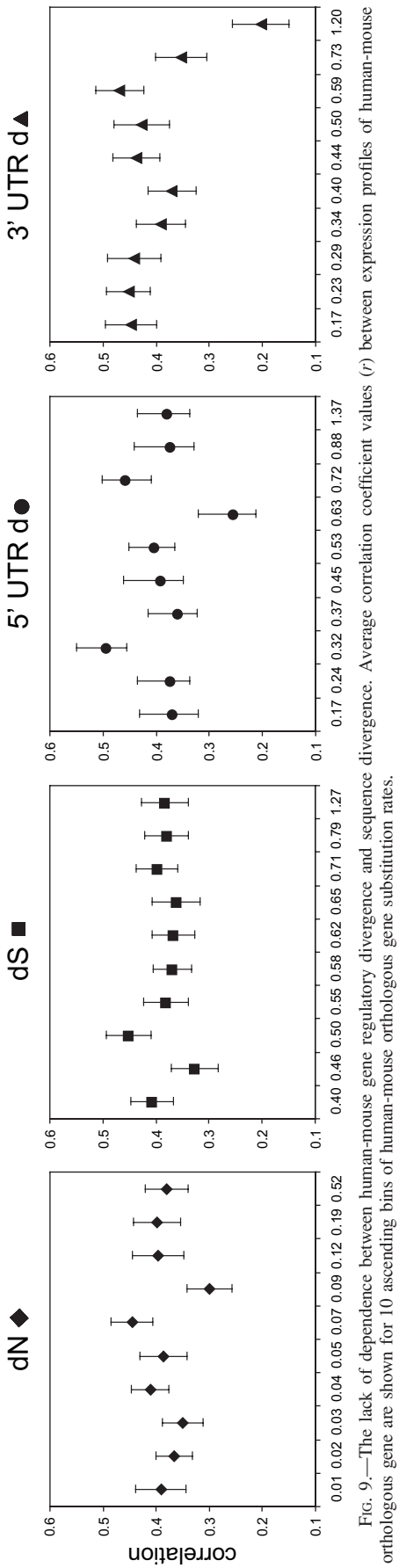

FIG. 9.—The lack of dependence between human-mouse gene regulatory divergence and sequence divergence. Average correlation coefficient values (r) between expression profiles of human-mouse orthologous gene are shown for 10 ascending bins of human-mouse orthologous gene substitution rates.

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 distributions 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 r^2 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 network ($r \geq 0.9$) 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 $r \geq 0.7$) 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 $r \geq 0.7$ between mouse genes, are shown for 10 ascending bins of pairwise human-mouse orthologous gene substitution rate differences.

Literature Cited

Agrawal, H. 2002. Extreme self-organization in networks constructed from gene expression data. Phys. Rev. Lett. **89**:268702.

Ashburner, M., C. A. Ball, J. A. Blake et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. **25**:25–29.

Barabasi, A. L., and R. Albert. 1999. Emergence of scaling in random networks. Science **286**:509–512.

Barabasi, A. L., and Z. N. Oltvai. 2004. Network biology: understanding the cell's functional organization. Nat. Rev. Genet. **5**:101–113.

Bergmann, S., J. Ihmels, and N. Barkai. 2004. Similarities and differences in genome-wide expression data of six organisms. PLoS Biol. **2**:0085–0093.

Bhan, A., D. J. Galas, and T. G. Dewey. 2002. A duplication growth model of gene expression networks. Bioinformatics **18**:1486–1493.

Bloom, J. D., and C. Adami. 2003. Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets. BMC Evol. Biol. **3**:21.

Britten, R. J., and E. H. Davidson. 1969. Gene regulation for higher cells: a theory. Science **165**:349–357.

Carbone, A., A. Zinovyev, and F. Kepes. 2003. Codon adaptation index as a measure of dominating codon bias. Bioinformatics **19**:2005–2015.

Darwin, C. 1859. On the origin of species. John Murray, London.

Duret, L., and D. Mouchiroud. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. **17**:68–74.

Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA **95**:14863–14868.

Farris, J. S. 1977. Phylogenetic analysis under Dollo's law. Syst. Zool. **26**:77–88.

Fraser, H. B., A. E. Hirsh, L. M. Steinmetz, C. Scharfe, and M. W. Feldman. 2002. Evolutionary rate in the protein interaction network. Science **296**:750–752.

Fraser, H. B., D. P. Wall, and A. E. Hirsh. 2003. A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evol. Biol. **3**:11.

Giot, L., J. S. Bader, C. Brouwer et al. 2003. A protein interaction map of *Drosophila melanogaster*. Science **302**:1727 1736.

Gu, Z., D. Nicolae, H. H. Lu, and W. H. Li. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet. **18**:609–613.

Hedges, S. B., H. Chen, S. Kumar, D. Y. Wang, A. S. Thompson, and H. Watanabe. 2001. A genomic timescale for the origin of eukaryotes. BMC Evol. Biol. **1**:4.

Higgins, D. G., J. D. Thompson, and T. J. Gibson. 1996. Using CLUSTAL for multiple sequence alignments. Methods Enzymol. **266**:383–402.

Hirsh, A. E., and H. B. Fraser. 2001. Protein dispensability and rate of evolution. Nature **411**:1046–1049.

Ho, Y., A. Gruhler, A. Heilbut et al. 2002. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature **415**:180–183.

Hoyle, D. C., M. Rattray, R. Jupp, and A. Brass. 2002. Making sense of microarray data distributions. Bioinformatics **18**:576–584.

Jeong, H., S. P. Mason, A. L. Barabasi, and Z. N. Oltvai. 2001. Lethality and centrality in protein networks. Nature **411**:41–42.

Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi. 2000. The large-scale organization of metabolic networks. Nature **407**:651–654.

Jordan, I. K., I. B. Rogozin, Y. I. Wolf, and E. V. Koonin. 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. **12**:962–968.

Jordan, I. K., Y. I. Wolf, and E. V. Koonin. 2003. No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol. Biol. **3**:1.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. N. Munro, ed. Mammalian protein metabolism. Academic, New York.

Kamath, R. S., A. G. Fraser, Y. Dong et al. 2003. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. Nature **421**:231–237.

Karev, G. P., Y. I. Wolf, A. Y. Rzhetsky, F. S. Berezovskaya, and E. V. Koonin. 2002. Birth and death of protein domains: a simple model of evolution explains power law behavior. BMC Evol. Biol. **2**:18.

Khaitovich, P., G. Weiss, M. Lachmann, I. Hellman, W. Enard, B. Muetzel, U. Wiekner, W. Ansorge, and S. Paabo. 2004. A neutral model of transcriptome evolution. PLOS Biol. **2**:0682–0689.

King, M. C., and A. C. Wilson. 1975. Evolution at two levels in humans and chimpanzees. Science **188**:107–116.

Koonin, E. V., N. D. Fedorova, J. D. Jackson et al. 2004. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol. **5**:R7.

Koonin, E. V., Y. I. Wolf, and G. P. Karev. 2002. The structure of the protein universe and genome evolution. Nature **420**:218–223.

Krylov, D. M., Y. I. Wolf, I. B. Rogozin, and E. V. Koonin. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res. **13**:2229–2235.

Kuznetsov, V. A., G. D. Knott, and R. F. Bonner. 2002. General statistics of stochastic process of gene expression in eukaryotic cells. Genetics **161**:1321–1332.

Lander, E., S. L. M. Linton, B. Birren et al. 2001. Initial sequencing and analysis of the human genome. Nature **409**: 860–921.

Li, S., C. M. Armstrong, N. Bertin et al. 2004. A map of the interactome network of the metazoan *C. elegans*. Science **303**: 540–543.

Li, W. H. 1997. Molecular evolution. Sinauer Associates, Sunderland, Mass.

Luscombe, N. M., J. Qian, Z. Zhang, T. Johnson, and M. Gerstein. 2002. The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. Genome Biol.3:RESEARCH0040.

Makova, K. D., and W. H. Li. 2003. Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Res. **13**:1638–1645.

Nei, M., and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**:418–426.

Pal, C., B. Papp, and L. D. Hurst. 2001. Highly expressed genes in yeast evolve slowly. Genetics **158**:927–931.

———. 2003. Genomic function: rate of evolution and gene dispensability. Nature **421**:496–497.

Pennisi, E. 2003. Systems biology: tracing life's circuitry. Science **302**:1646–1649.

Ravasz, E., A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi. 2002. Hierarchical organization of modularity in metabolic networks. Science **297**:1551–1555.

Rzhetsky, A., and S. M. Gomez. 2001. Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. Bioinformatics **17**:988–996.

Su, A. I., M. P. Cooke, K. A. Ching et al. 2002. Large-scale analysis of the human and mouse transcriptomes. Proc. Natl. Acad. Sci. USA **99**:4465–4470.

Tatusov, R. L., N. D. Fedorova, J. D. Jackson et al. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4**:41.

Ueda, H. R., S. Hayshi, S. Matsuyama, T. Yomo, S. Hashimoto, S. A. Kay, J. B. Hogenesch, and M. Iino. 2004. Universality and flexibility in gene expression from bacteria to human. Proc. Natl. Acad. Sci. USA **101**:3765–3769.

Wagner, A. 2000. Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. Proc. Natl. Acad. Sci. USA **97**:6579–6584.

———. 2003. How the global structure of protein interaction networks evolves. Proc. R. Soc. Lond. B Biol. Sci. **270**:457–466.

Waterston, R. H. K. Lindblad-Toh, E. Birney et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420**:520–562.

Yanai, I., D. Graur, and R. Ophir. 2004. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. OMICS **8**:15–24.

Yang, J., Z. Gu, and W. H. Li. 2003. Rate of protein evolution versus fitness effect of gene deletion. Mol. Biol. Evol. **20**:772–774.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl. Biosci. **13**:555–556.

Zhang, L., and W. H. Li. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol. Biol. Evol. **21**:236–239.

Peer Bork, Associate Editor

Accepted July 14, 2004