Disease genes and intracellular protein networks

S. Bortoluzzi,1 C. Romualdi,2 A. Bisognin,1 and G. A. Danieli1

1Department of Biology and 2Ricerche Interdipartimentale di Biotecnologie Innovative (CRIBI), University of Padua, Padua I-35131, Italy

Submitted 10 June 2003; accepted in final form 3 September 2003

Bortoluzzi, S., C. Romualdi, A. Bisognin, and G. A. Danieli. Disease genes and intracellular protein networks. Physiol Genomics 15: 223–227, 2003. First published September 9, 2003; 10.1152/physiolgenomics.00095.2003.—By a computational approach we reconstructed genomic transcriptional profiles of 19 different adult human tissues, based on information on activity of 27,924 genes obtained from unbiased UniGene cDNA libraries. In each considered tissue, a small number of genes resulted highly expressed or “tissue specific.” Distribution of gene expression levels in a tissue appears to follow a power law, thus suggesting a correspondence between transcriptional profile and “scale-free” topology of protein networks. The expression of 737 genes involved in Mendelian diseases was analyzed, compared with a large reference set of known human genes. Disease genes resulted significantly more expressed than expected. The possible correspondence of their products to important nodes of intracellular protein network is suggested. Auto-organization of the protein network, its stability in time in the differentiated state, and relationships with the degree of genetic variability at genome level are discussed.

IN THE LAST FEW YEARS, scale-free network models have been proposed to describe interactions underlying organization of biological systems (16, 22). Scale-free networks are characterized by robustness against perturbation, such as random depletion of network components (17, 22). Attempts to reconstruct protein-protein interaction networks by systematic analysis of protein complexes (7, 13), by two-hybrids approach (15), or by combination of computational and experimental approaches (25) produced results supporting these models.

Data concerning expression of genes in human tissues, derived from unbiased cDNA libraries, enabled in silico reconstruction of genomic expression profiles of different human tissues (3–5) and showed that a small fraction of genes account for a large proportion of the total transcriptional activity. Preliminary studies on adult skeletal muscle, heart, and retina indicated that Mendelian disease genes are relatively upregulated in affected tissues (3–5). High expression of a limited number of genes would produce a limited number of proteins in large amounts. Since it could be assumed that “importance” or connectivity of proteins is related to their abundance, genes mutated in Mendelian diseases, coding for proteins of pivotal role, should be highly expressed in disease cells and tissues.

We report here on quantitative analysis of the expression of 737 genes involved in Mendelian diseases compared with a set of 13,294 known human genes, as obtained from transcriptional profiles of 19 adult human tissues.

METHODS

Data mining and processing from UniGene, the largest available collection of human expressed genes (http://www.ncbi.nlm.nih.gov/UniGene/) (2), was performed by a fully automated procedure for expression profiles reconstruction based on Perl scripts. Only unbiased (unsubtracted and/or unnormalized) UniGene cDNA libraries pertaining to adult human tissues were considered.

An expression profile is a list of expressed genes, showing the abundance of their transcripts in a given cell or tissue. Each gene in a given profile is identified by gene name and description, UniGene cluster, LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/; 19) number, if available, and GenBank ID of the longest sequence representative of the cluster. The number of expressed sequence tags (ESTs) per UniGene cluster included in the expression profile of a given tissue was used to estimate the level of activity of the corresponding gene. Tissue expression profiles were reconstructed by using information on pool of libraries pertaining to the tissue. Only those tissues were considered for which a sufficiently large number of ESTs from unbiased libraries was available. The whole data set included 270,871 ESTs, corresponding to 27,924 UniGene “clusters”, 14,131 of which corresponded to LocusLink entries and to “known human genes.”

The distribution of number of genes according to the level of expression is exponential. For the observed distribution of number of genes with x level of expression, F(x) = ax−λ, we used the nonlinear regression method to estimate parameters characterizing the function of the number of genes per level of expression in each tissue. Estimation was done to infer parameters describing the connectivity distribution of a hypothetical network, describing the protein-protein interactions between proteins encoded by the considered genes, under a proportionality hypothesis between expression level of genes and “importance” or connectivity of the corresponding proteins.

We further considered the disease loci of the OMIM Morbid Map (ftp://ftp.ncbi.nih.gov/repository/OMIM/morbidmap) and, using a Perl script analyzing the files for the conversion of LocusLink/UniGene numbers and of LocusLink/MIM numbers, we established the exact correspondence between disease loci and UniGene clusters represented in the reconstructed expression profiles. Genes involved in tumor devel-

Article published online before print. See web site for date of publication (http://physiolgenomics.physiology.org).

Address for reprint requests and other correspondence: G. A. Danieli, Dept. of Biology, Univ. of Padua, via U. Bassi 58 B, Padua I-35131, Italy (E-mail: danieli@bio.unipd.it).
RESULTS

By a computational approach described elsewhere (3) and implemented in novel software (21), we reconstructed genomic transcriptional profiles of 19 different adult human tissues, based on information on activity of 27,924 genes, obtained from unbiased cDNA libraries (Table 1; http://telethon.bio.unipd.it/bioinfo/HGXP/). In each considered tissue, only a small proportion of genes appears highly expressed. An even smaller proportion of genes appears as tissue specific.

The number of ESTs obtained for each given transcript from a specific tissue and derived from pools of development, susceptibility or resistance to diseases, and those involved in “conditions” or in nondisease phenotypes were excluded. At the end, a data set of 737 disease genes was obtained, whose mutations cause Mendelian disorders. We compared their expression with that of LocusLink entries corresponding to 13,294 known human genes (‘reference set’), in 19 human tissues. Mendelian disease genes were excluded from the reference set.

For each given gene, average level of expression was calculated over all tissues in which its activity was detected. The average level of expression of disease genes was compared with that of genes of the reference set. Genes of the reference set and disease genes were classified according to their average level of expression in all the considered tissues (average on tissues in which the gene is expressed; more than 5 ESTs per 10,000, high; from 1 to 5 ESTs per 10,000, moderate; less than 1 EST per 10,000, weak). Frequency distribution in the two groups was compared by \( \chi^2 \) test. The same was done separately for comparing expression levels of autosomal dominant or autosomal recessive genes with those of the reference set.

The hypothesis of independence between expression level of genes and mode of inheritance of the corresponding diseases was tested by means of a \( \chi^2 \) test on a contingency table, after grouping genes corresponding to autosomal dominant or autosomal recessive diseases into the three previously described ranks of expression.

Perl software used to generate data in this manuscript is freely available upon request to the authors.

Table 1. Expression profiles of adult human tissues

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Selected cDNA Libraries</th>
<th>Genes</th>
<th>ESTs</th>
<th>Genes Contributing to 50% of Transcription</th>
<th>Genes Differentially Expressed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No.</td>
<td>%</td>
</tr>
<tr>
<td>Adipose</td>
<td>816, 10983</td>
<td>2,470</td>
<td>4,329</td>
<td>486</td>
<td>19.7</td>
</tr>
<tr>
<td>Bone</td>
<td>10410, 1124</td>
<td>4,119</td>
<td>11,381</td>
<td>489</td>
<td>11.9</td>
</tr>
<tr>
<td>Cerebellum</td>
<td>742</td>
<td>2,419</td>
<td>4,046</td>
<td>552</td>
<td>22.8</td>
</tr>
<tr>
<td>Colon</td>
<td>656, 1556</td>
<td>2,772</td>
<td>5,717</td>
<td>379</td>
<td>13.7</td>
</tr>
<tr>
<td>Hypophysis</td>
<td>6826, 6829, 6828</td>
<td>2,756</td>
<td>5,842</td>
<td>395</td>
<td>14.3</td>
</tr>
<tr>
<td>Hypothalamus</td>
<td>6814, 6811, 6812</td>
<td>3,422</td>
<td>6,484</td>
<td>722</td>
<td>21.1</td>
</tr>
<tr>
<td>Hypocampus</td>
<td>8569</td>
<td>5,336</td>
<td>15,032</td>
<td>746</td>
<td>14.0</td>
</tr>
<tr>
<td>Lens</td>
<td>7315, 10273</td>
<td>1,740</td>
<td>4,627</td>
<td>158</td>
<td>9.1</td>
</tr>
<tr>
<td>Liver</td>
<td>252, 5601</td>
<td>4,943</td>
<td>21,452</td>
<td>239</td>
<td>4.8</td>
</tr>
<tr>
<td>Lung</td>
<td>10397, 249, 6834</td>
<td>6,924</td>
<td>23,686</td>
<td>549</td>
<td>7.9</td>
</tr>
<tr>
<td>Marrow</td>
<td>6975, 6976, 931</td>
<td>3,718</td>
<td>11,627</td>
<td>230</td>
<td>6.2</td>
</tr>
<tr>
<td>Melanocyte</td>
<td>367</td>
<td>6,020</td>
<td>20,774</td>
<td>1,062</td>
<td>17.6</td>
</tr>
<tr>
<td>Muscle</td>
<td>45, 530, 6761, 9692</td>
<td>8,517</td>
<td>26,964</td>
<td>859</td>
<td>10.1</td>
</tr>
<tr>
<td>Pancreas</td>
<td>6760, 1098, 3884, 422, 83, 8840</td>
<td>6,067</td>
<td>27,430</td>
<td>336</td>
<td>5.5</td>
</tr>
<tr>
<td>Pinal gland</td>
<td>734, 342</td>
<td>3,252</td>
<td>6,275</td>
<td>673</td>
<td>20.7</td>
</tr>
<tr>
<td>Prostate</td>
<td>891, 925, 787, 573, 6763</td>
<td>5,046</td>
<td>13,535</td>
<td>528</td>
<td>10.5</td>
</tr>
<tr>
<td>Retina/choroid</td>
<td>10285, 10283, 10286, 690, 303, 433, 10966, 302, 7269</td>
<td>8,647</td>
<td>33,815</td>
<td>921</td>
<td>10.7</td>
</tr>
<tr>
<td>Testis</td>
<td>677, 1752, 6762</td>
<td>6,665</td>
<td>14,763</td>
<td>1,235</td>
<td>18.5</td>
</tr>
</tbody>
</table>

Data were mined from UniGene, by a dedicated software developed in our laboratory. Only tissues were considered, for which more than 4,000 expressed sequence tags (ESTs) from unbiased libraries were available. The whole data set included 270,871 ESTs, corresponding to 27,924 UniGene “clusters.” For each tissue, the dbEST ID of source cDNA libraries are listed, with the number of detected genes per tissue and the total number of ESTs used for the estimation of genes expression levels. In columns 5 and 6, numbers and percentages of genes contributing to the 50% of the total detected transcriptional activity are reported. In the last two columns, number of genes differentially expressed in each considered tissue is shown. Audic and Claverie test statistic was used for comparisons between the number of ESTs in the two tissues in which the gene resulted more expressed, with probability threshold set to \( \alpha = 0.001 \) or \( \alpha = 0.01 \), respectively.
unbiased cDNA libraries was used to estimate the expression level of the corresponding gene (3, 20, 21). In every considered tissue, a relatively small number of highly expressed genes appeared to account for a very large fraction of the total transcriptional activity, as shown in Fig. 1 and reported in Table 1 (columns 5 and 6). In general, slightly more than 10% of the total number of expressed genes accounts for one-half of the total transcriptional activity of any given tissue.

Genes which appeared differentially expressed in a specific tissue, below the selected significance threshold level, are presumably “tissue specific” (Table 1, columns 7 and 8, $\alpha = 0.01$ and $\alpha = 0.001$, respectively). Such genes account for less than 1% of genes expressed in a given tissue.

It has been proposed that intracellular protein-to-protein interactions form a highly inhomogeneous scale-free network in which few highly connected proteins play a central role in interactions with numerous, less connected proteins (6, 12, 16). If, under a certain degree of inaccuracy, abundance of a given protein is due to level of expression of the corresponding gene, then we should expect that distribution of protein-encoding genes according to their levels of expression fits in a scale-free network model. We observed that the distribution of number of genes with $x$ level of expression, $P(x)$, follows an exponential trend ($P(x) = ax^{-k}$) in each of the 19 considered tissues (minimum correlation coefficient of the fitting $r^2$ 0.80, average $r^2$ 0.98, median $r^2$ 0.99).

Connectivity of a network ($P(k)$) could be defined as the function describing the distribution of nodes having $k$ connections with other nodes in the network. Under the assumption that level of expression of proteins correlates with number of interactions established with other proteins, it could be supposed that $k \propto x$. We used the parameters characterizing the function of the number of genes per level of expression in each tissue, estimated with nonlinear regression method, to infer parameters describing the connectivity of the network, $P(k)$. For each of the 19 tissues considered in our study, estimated parameters of the observed distribution, $\tilde{\lambda}$ and $\tilde{a}$, resulted to be on average 1.7 and 3,334.9. Estimated $\tilde{\lambda}$ values ranged from 1.2 and 2.4 with extreme values in melanocytes and in the adipose tissue, respectively. Relative homogeneity of estimated $\tilde{\lambda}$ among the 19 different tissues suggests that genomic transcriptional profiles show common features, despite considerable biological differences among tissues.

We analyzed the expression of 737 human genes whose mutations cause Mendelian disorders, in 19 human tissues, compared with that of 13,294 known human genes. We excluded from analysis genes involved in nondisease phenotypes, psychiatric conditions, tumor development, or in susceptibility or resistance to diseases. The Supplemental Table (available at http://telethon.bio.unipd.it/bioinfo/Disease_genes/, as well as at the Physiological Genomics web site1) reports expression data of 737 selected disease genes in 19 human tissues. Average expression level of disease genes in tissues resulted significantly higher than the mean of the 13,294 genes of the reference set (1.220 vs. 0.574, $t$-test, $P = 3.298E-8$). Disease genes, classified according to their average level of expression in all the considered tissues, appeared significantly more expressed than expected (observed 16.6%, expected 3.2%; $\chi^2$ test, $P = 2.76E-56$, Table 2), when compared with the sample of genes of the reference set. When comparing expression of genes involved in autosomal dominant diseases with genes of the reference sample, the difference appeared even more impressive (observed 18.9%, expected 3.2%; $P = 2.24E-49$).

The hypothesis of independence between mode of inheritance of Mendelian diseases and level of expression of involved genes was tested by means of a $\chi^2$ test on a contingency table, after grouping genes corresponding to autosomal dominant or autosomal recessive diseases into three different ranks of gene expression. The hypothesis of independence was rejected ($P = 0.00802$), thus indicating that genes corresponding to autosomal dominant diseases tend to be highly expressed, whereas those corresponding to autosomal recessive tend to be moderately expressed, although, in general, genes involved in Mendelian diseases appear to be more expressed than expected.

**DISCUSSION**

According to data obtained by the present study, the fraction of genes accounting for observed structural

---

1The Supplementary Material for this article (expression data of 737 selected disease genes in 19 human tissues) is available online at http://physiolgenomics.physiology.org/cgi/content/full/00095.2003/DC1.
and functional differences among human tissues is probably small. A number of housekeeping genes must be ubiquitously active and some genes might be selectively overexpressed in contaminant tissues (e.g., vessels, blood, connective tissue, etc). In addition, some UniGene clusters could contain sequences derived from tissue-specific isoforms of transcripts belonging to the same gene. Therefore, the proportion of “tissue-specific” genes might be even smaller than observed. This view is in agreement with results obtained by multiple-tag experiments, indicating that differentially expressed state is characterized by high expression of a limited number of genes probably functionally related and possibly coregulated (14, 18, 24).

According to recent hypotheses, intracellular protein-to-protein interactions form a highly inhomogeneous scale-free network in which few highly connected proteins play a central role in interactions with numerous, less connected proteins (6, 12, 16).

Since abundance of a given protein is, in general, strongly correlated with level of expression of the corresponding gene (11), characteristics of protein-protein interaction network would be partially determined by the number of protein-encoding genes and by their respective levels of expression.

A very abundant protein, which likely corresponds to a highly transcribed gene, would potentially interact not only with other abundant proteins but also with rare or very rare polypeptides, thus becoming a relevant node in the network of protein-protein interactions. Statistical analysis of large-scale gene expression and protein interaction data showed that protein pairs encoded by coexpressed genes interact with each other more frequently than with random proteins and that the mean similarity of expression profiles is significantly higher for respective interacting protein pairs than for random ones (10). Moreover, Ge and colleagues (8) showed that pairs of genes encoding interacting proteins tend to be coexpressed and that gene expression data could be profitably used to refine models of protein interactions.

Highly expressed genes are expected to encode proteins representing vulnerable nodes, and therefore, their mutations are expected to be pathogenic. Recently, it was shown that removal of highly connected nodes causes a noticeable fragmentation in gene networks (22). Our observation that genes involved in Mendelian diseases are mostly highly or moderately expressed and, in particular, more expressed than expected if compared with a large set of reference genes, strongly suggests that disease genes encode for highly connected nodes of the network.

The “scale-free network” model applied to genomic expression might explain stability of many tissues in time. The structure of the protein-protein interactions network in a given cell or tissue is generated by potentiality of interactions inherent to the structure of individual proteins and to their relative abundance. Up-regulation of a limited number of genes would force a given tissue to maintain stable structural and functional characteristics during time, unless strong perturbations would occur, affecting the transcription levels of genes coding for functional nodes, or creating novel nodes by upregulation of silent genes. This would be the case of perturbations induced by activity of genes encoding transcription factors, which regulate the expression of several additional genes. Defective mutations in regulatory genes are expected to produce complex and often lethal phenotypes, infrequently reported as Mendelian diseases.

According to “scale-free network” model, errors or attacks involving the majority of nodes showing small connectivity should not significantly alter the path structure of the remaining nodes and should have scarce impact on the overall network topology (1), i.e., on function. Systematic mutagenesis in yeast provided evidence of a striking capacity of tolerating deletion of substantial number of proteins from its proteome (16).

It is known that both physiological and developmental processes of eukaryotes display considerable robustness against the effects of mutations (26). Many loss-of-function mutations of developmental genes in higher organisms show no or weak phenotypic effect (9, 23, 27). It was shown in yeast that genes whose loss of function results in a weak or no fitness effect are not more similar to their closest paralogues, both in sequence and temporal expression pattern (26). Thus gene duplications contribute little to mutational robustness on a genomic scale, whereas it seems to mainly depend on nature and organization of interactions between different proteins. Transferring these concepts to human genome, defective mutations in a relatively low number of genes encoding proteins corresponding to highly connected nodes seem to produce inherited disorders classified as Mendelian diseases. Therefore, defective mutations in genes encoding nodes with small connectivity are expected to have minor clinical consequences, selection against such variants are expected to be relatively mild, and genetic variability in these genes should be expectedly higher than in those directly involved in Mendelian diseases.

GRANTS

The financial support of Italian Ministry for Scientific and Technologic Research and of the University of Padua to G. A. Danieli and of the Italian Association for Cancer Research to S. Bortoluzzi is gratefully acknowledged.

REFERENCES


