PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2010) **277**, 1403–1408
doi:10.1098/rspb.2009.1865
Published online 6 January 2010

Systemic factors dominate mammal protein evolution

Alexander E. Vinogradov*

Institute of Cytology, Russian Academy of Sciences, St Petersburg 194064, Russia

Proteins encoded by highly expressed genes evolve more slowly. This correlation is thought to arise owing to purifying selection against toxicity of misfolded proteins (that should be more crucial for highly expressed genes). It is now widely accepted that this individual (by-gene) effect is a dominant cause in protein evolution. Here, I show that in mammals, the evolutionary rate of a protein is much more strongly related to the evolutionary rate of coexpressed proteins (and proteins of the same biological pathway) than to the expression level of its encoding gene. The complexity of gene regulation (estimated by the numbers of transcription factor targets and regulatory microRNA targets in the encoding gene) is another important cause, which is much stronger than gene expression level. Proteins encoded by complexly regulated genes evolve more slowly. The intronic length and the ratio of intronic to coding sequence lengths also correlate negatively with protein evolutionary rate (which contradicts the expectation from the negative link between expression level and evolutionary rate). One more important factor, which is much stronger than gene expression level, is evolutionary age. More recent proteins evolve faster, and expression level of an encoding gene becomes quite a minor cause in the evolution of mammal proteins of metazoan origin. These data suggest that, in contrast to a widespread opinion, systemic factors dominate mammal protein evolution.

Keywords: protein evolution; gene expression; cell integrity; coevolution

1. INTRODUCTION

The causes of among-protein variation in the evolutionary rate are thought of as one of the most important problems in the molecular evolution (Wolf *et al*. 2008). Furthermore, the evolutionary rate can serve as an indicator of even more significant aspects. For instance, whether a protein evolves mostly by itself (as was assumed in 'bean-bag' genetics) or whether cell integrity is a more influential factor of protein evolution. Paradoxically, even after the advent of systems biology, the prevailing opinion is still in favour of the former suggestion. Despite the fact that protein evolutionary rate was found to correlate with certain systemic parameters such as the number of protein interactions of a given protein, its position in protein interaction network, evolutionary rate of coexpressed proteins, gene dispensability, subcellular location (Fraser *et al*. 2002; Jordan *et al*. 2004; Batada *et al*. 2006; Julenius & Pedersen 2006; Koonin & Wolf 2006; Makino & Gojobori 2006; Wolf 2006), it was concluded that it is the expression level of an encoding gene that is a major determinant of protein evolutionary rate (Drummond *et al*. 2006; Kawahara & Imanishi 2007; Drummond & Wilke 2008; Powers & Balch 2008). This relationship is negative, the higher the expression level, the lower the evolutionary rate. It is believed that this regularity arises owing to negative (purifying) selection for robustness to protein misfolding because of the toxicity of misfolded proteins, which should be more important for highly expressed proteins (Drummond & Wilke 2008; Powers & Balch 2008; Wolf *et al*. 2008). In other words, this regularity is supposed to be an individual (by-gene) effect. The other factors are assumed to play a minor role or even be just the results of their correlation with gene expression level (Drummond *et al*. 2006; Koonin & Wolf 2006; Hakes *et al*. 2007). For instance, it has been suggested that the similarity of evolutionary rates of interacting proteins is a result of their similarity in gene expression levels (Fraser *et al*. 2004; Hakes *et al*. 2007). However, recent work reported that in the fused proteins, the expression level and the structural–functional constraints participate comparably in the determination of protein evolutionary rate (Wolf *et al*. 2008). Here, I show in a straightforward genome-wide analysis that systemic factors dominate mammal protein evolution. Compared with them, expression level of an encoding gene is a minor cause, especially for proteins of metazoan origin. The expression level of coexpressed genes produces a similarly weak effect.

2. MATERIAL AND METHODS

(a) *Expression and coexpression data*

The data on human and mouse gene expression were taken from the Gene Expression Atlas (Su *et al*. 2004). They present the results of high-density oligonucleotide microarray experiments performed uniformly for all tissues of the same species (79 human and 61 mouse tissues). Only probes that presented the well-characterized protein-coding genes, i.e. with links to the entrez gene (Sayers *et al*. 2009), were used (total 16 554 human and 17 108 mouse genes). The among-tissues correlation of gene expression levels was used for determination of coexpressed genes. A low cutoff of the Pearson correlation coefficient ($r > 0.3$) was chosen for determination of coexpressed genes but the parameters of coexpressed genes were tested in a wide range of weights. The weights were equal to correlation coefficient raised to different powers (from 0 to 12). In a special analysis, a

*aevin@mail.cytspb.rssi.ru

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2009.1865 or via http://rspb.royalsocietypublishing.org.

Received 14 October 2009
Accepted 7 December 2009

negative correlation ($r < -0.3$) was tested (with powers in the range from 0 to 10). The exploration of the weights range is preferable to the study of the range of correlation coefficient cutoffs, with the advantages that more genes are included in the analysis and the differences in the strength of coexpression links are accounted for. The range of actual weights varied from zero-fold for power = 0 to about million-fold for power = 12, i.e. the usage of high powers practically means the exclusion of weakly correlated links. In other words, the variation of weights was similar to variation of correlation cutoffs but in a more finely adjusted way.

The following parameters were calculated using these expression data: the average (over all tissues) expression level of a given gene ('expression I'), the weighted degree (number of coexpressed genes) of a given gene ('degree I'), the weighted average expression level of coexpressed genes ('expression coexpr-avg'), and the weighted average degree of coexpressed genes ('degree coexpr-avg'). The main results are presented with expression data standardized with the MAS5 algorithm. (The data standardized with the gcRMA algorithm were also tested.)

As a completely independent type of data, in a separate analysis, the libraries of expressed sequence tag (EST) sequences were used for determination of gene expression levels and coexpression memberships. The clusters of EST sequences corresponding to genes were extracted from the Unigene database (Sayers *et al.* 2009). Only libraries, which were obtained from normal tissues and contain more than 10 000 sequences, were used. The count of sequences in each library was normalized to 50 000 (roughly a mean library size). The normalized counts were used as gene expression levels (zero levels were assigned to genes absent in a library). All the above-mentioned parameters (expression I, degree I, expression coexpr-avg, degree coexpr-avg) were determined using these expression levels.

As another independent type of data, the known protein interactions were taken from the STRING database (Jensen *et al.* 2009). They were used for determination of the parameters of coexpression membership (degree I, expression coexpr-avg, degree coexpr-avg). There were no weights here. The expression data for these calculations were taken from the Gene Expression Atlas.

In a separate analysis, human biological pathways were used for definition of supposedly coevolved proteins. The data on pathways (i.e. lists of genes belonging to different pathways) were taken from the Molecular Signatures Database (Subramanian *et al.* 2007). They present the compilation of pathways from different databases (Kegg, Reactome, Biocarta, HumanCyc, GenMapp; totally 639 pathway gene sets). Genes belonging to the same pathway were treated as 'coexpressed' for calculation of the above-mentioned parameters (degree I, expression coexpr-avg, degree coexpr-avg). There were no weights here. The expression data for these calculations were taken from the Gene Expression Atlas.

The data on the transcription factor targets (615 gene sets) and the regulatory microRNA targets in 3′-UTR (222 gene sets) were also taken from the Molecular Signatures Database.

(b) *Evolutionary rate*

The human and mouse protein sequences (and corresponding coding and intronic sequences) were taken from the RefSeq database (Sayers *et al.* 2009). The orthology of human–mouse genes was determined by all human against all mouse (and vice versa) protein matching using the Smith–Waterman algorithm implemented in the 'ssearch' program of the FASTA package with the 'shuffled' calculation of statistical significance (Pearson 1999). The reciprocal best hits of the longest proteins of human–mouse gene pairs were treated as orthologous pairs. (In total, there were 16 211 such pairs.) The human–mouse evolutionary distances were determined by protein sequence alignment using the CLUSTALW program with default parameters (Larkin *et al.* 2007). These distances were used as the main evolutionary rate parameter (because protein distances require less assumptions than nucleotide substitutions models). In additional analyses, the rates of synonymous and non-synonymous nucleotide substitutions were determined using the PAML program with default parameters (Yang 2007). The human–mouse alignments of coding sequences used as input for this program were obtained with the REVTRANS program (Wernersson & Pedersen 2003), using the CLUSTALW protein alignments and the coding sequences. The ratios of non-synonymous to synonymous nucleotide substitution rates (dN/dS) were taken as protein evolutionary rates. These ratios were obtained using two substitution models: Nei–Gojobori and Yang–Nielsen. The results obtained with each model were analysed separately. As with gene expression data, there were two parameters of evolutionary rate: evolutionary rate of a given protein ('evolrate I', which was the dependent variable in the analysis), and weighted average evolutionary rate of proteins encoded by coexpressed genes or genes of the same biological pathway ('evolrate coexpr-avg', which was one of predictor variables).

(c) *Evolutionary origin*

Gene evolutionary origin was determined as described (Vinogradov 2009). Briefly, this determination was based on the clusters of orthologous groups (COG) (eukaryotic orthologous groups, KOG) orthologous gene groups (Koonin *et al.* 2004) as presented in the STRING database with addition of non-supervised orthologous groups (NOG) orthologous groups (Jensen *et al.* 2009), and the NCBI phylogenetic tree (Sayers *et al.* 2009). Twelve evolutionary stages were taken, determined by the following phylogenetic branching: cellular organisms, Eukaryota, Fungi/Metazoa group, Bilateria, Coelomata, Chordata, Vertebrata, Tetrapoda, Amniota, Mammalia, Eutheria, Primates (or Rodentia). A gene was regarded as appearing at a corresponding evolutionary stage if it had relatives in the same COG (KOG, NOG) group in the phylogenetic lineages branched off after this stage and there were no relatives in the lineages branched off earlier.

(d) *Analysis*

The statistical analyses were done using a general linear model (GLM, which is a generalization of ANOVA) with type III sums of squares, where the effect of each tested variable does not depend on the order in which it is introduced into the model (in contrast, for instance, to multiple regression). Type III sums of squares measure the marginal contribution of each predictor variable, assuming it was added last (i.e. the effect that remains independent of the effects of other tested variables). (If predictor variables are tested separately, their *F*-values can be higher.) The Fisher ratios of variance explained by a given parameter to error variance (*F*-values) were used for estimation of relative effects of

Table 1. The *F*-values for different predictor variables in the general linear model (GLM) with the evolutionary rate of a protein (protein distance) as dependent variable (human). (Significance levels for different *F*-values for $n > 1000$: 6.7, $p < 10^{-2}$; 10.9, $p < 10^{-3}$; 15.2, $p < 10^{-4}$; 24, $p < 10^{-6}$; 42, $p < 10^{-10}$; 68, $p < 10^{-16}$.)

parameter	all genes ($n = 13\,577$)	pre-metazoan ($n = 8698$)	metazoan ($n = 4879$)	sign
expression I	24.8	51.4	4.7	−
expression coexpr-avg	24.1	10.6	10.0	+
degree I	8.3	16.0	0	−
degree coexpr-avg	25.3	10.8	9.4	+
evolutionary rate coexpr-avg	1945.7	1016.6	620.7	+
evolutionary origin I	1166.8	0.4	1341.0	+

Table 2. The *F*-values for different predictor variables in the GLM with the evolutionary rate of a protein as dependent variable (dN/dS for two different nucleotide substitution models).

	human		mouse		
parameter	Nei–Gojobori	Yang–Nielsen	Nei–Gojobori	Yang–Nielsen	sign
expression I	10.5	7.9	45.0	70.8	−
expression coexpr-avg	21.1	14.5	71.2	74.8	+
degree I	8.3	7.5	1.2	0.6	−
degree coexpr-avg	22.3	23.2	0.4	0	+
evolutionary rate coexpr-avg	1772.8	1516.1	1525.0	1058.3	+
evolutionary origin I	1316.5	1098.9	1259.1	1292.7	+

parameters. The expression level, degree, number of transcription factor targets, number of regulatory microRNA targets, evolutionary origin, intronic sequence length and the ratio of intronic to coding sequence lengths were used log-transformed. However, these and other parameters were also tested in different forms (non-transformed, log-transformed, standardized, ranked), all producing a qualitatively similar picture. The analyses were done using the StatGraphics Centurion package (Statpoint Technologies, Inc.). The analysis of overrepresented gene ontology categories was done as described (Vinogradov 2009).

3. RESULTS

(a) *Evolutionary rate of coexpressed genes*

The average parameters of coexpressed genes were determined in the wide range of weights (§2). With increasing of weight, the influence of evolutionary rate of coexpressed genes was slightly increasing, reaching a plateau when the weight power was equal to about 8 (electronic supplementary material, figures S1 and S2). Therefore, this power was chosen for presentation of results. However, in the whole range of weights, the effect of evolutionary rate of coexpressed genes and evolutionary origin was above an order of magnitude higher than the effect of expression level (of a given gene or its coexpressed genes) (electronic supplementary material, figures S1 and S2). This difference becomes even higher in genes of metazoan origin (table 1; electronic supplementary material, tables S1–S3). If instead of the average among-tissues expression, the maximum expression level (in any tissue) was used, its effect was even weaker (not shown).

If the ratio of non-synonymous to synonymous nucleotide substitution rates (dN/dS) was taken as protein evolutionary rate (which was assumedly corrected for mutation rate because of division by dS), the picture was similar (table 2; electronic supplementary material, table S4). Interestingly, when genes whose expression negatively correlates with expression of a given gene were taken as 'anti-coexpression' membership, the effect of evolutionary rate of anti-coexpressed proteins was also stronger than the effect of gene expression level. Yet, the direction of this effect changed sign, which supports the results obtained with positive coexpression membership (electronic supplementary material, table S5).

It should be noted that both the expression level and the coexpression membership were determined using the same dataset of microarray gene expression measurements across the same number of tissues (§2), i.e. they were determined with the same accuracy. Similarly, when the expression level and the coexpression membership were determined using a completely independent type of data (EST database), the effect of evolutionary rate of coexpressed genes and evolutionary origin was above an order of magnitude higher than the effect of gene expression level (electronic supplementary material, tables S6 and S7).

If protein interactions were taken for determination of coexpression membership, the picture was similar (electronic supplementary material, tables S8 and S9). Also, the picture was similar when biological pathways were used for definition of supposedly coevolved proteins (electronic supplementary material, table S10).

It should be noted that expression level of coexpressed genes produces a very weak effect (similar to the effect of expression of a given gene). Moreover, if all studied parameters are introduced to the model simultaneously, the sign of this effect is opposite to the sign of the effect of expression of a given gene (tables 1 and 2; electronic supplementary material, tables S1–S4 and S6–S10).

Table 3. The *F*-values for different predictor variables in the GLM with the evolutionary rate of a protein (protein distance) as dependent variable (human). (The last six variables were tested separately.)

parameter	$n = 7754$	$n = 5526$	$n = 12\,630$	$n = 12\,630$	$n = 4213$	$n = 2830$	sign
expression I	5.7	9.9	48.5	41.3	4.1	2.7	−
expression coexpr-avg	14.6	7.3	14.4	17.8	18.5	12.7	+
degree I	1.1	0.5	5.9	6.8	4.1	1.7	−
degree coexpr-avg	0.2	2.5	28.2	27.1	5.5	4.8	+
evolutionary rate coexpr-avg	863.5	286.6	1670.0	1713.5	741.8	523.6	+
evolutionary origin I	823.1	335.9	945.9	1049.4	478.7	401.3	+
number of transcription factor targets	725.2	—	—	—	—	—	−
number of regulatory microRNA targets	—	331.9	—	—	—	—	−
intronic length	—	—	180.0	—	—	—	−
ratio of intronic to coding sequence lengths	—	—	—	235.6	—	—	−
number of biological pathways (all)	—	—	—	—	5.1	—	−
number of Kegg pathways	—	—	—	—	—	12.8	−

(**b**) ***Gene regulation complexity***

Human genes, for which transcription factor targets or regulatory microRNA targets are known, as well as genes with introns, represent subsets of the total dataset. Therefore, they were analysed separately. The number of transcription factor targets, the number of regulatory microRNA targets, the length of intronic sequence and the ratio of intronic to coding sequence lengths all negatively correlate with protein evolutionary rate (table 3; electronic supplementary material, tables S11 and S12). The effect of the number of transcription factor targets is comparable with the effect of evolutionary rate of coexpressed proteins and evolutionary age. The effect of other parameters is weaker but also highly significant. Interestingly, genes with more complex regulation are themselves involved mostly in regulation and development (electronic supplementary material, tables S13 and S14).

In contrast to the regulatory parameters (numbers of transcription factor targets and microRNA targets) and the intronic length, the number of biological pathways in which a given gene is involved shows only a very weak, marginally significant effect (table 3; electronic supplementary material, tables S11 and S12). This is true both for the total pathways dataset and for the Kegg pathways taken separately (to avoid possible duplications in the total dataset).

It is noteworthy that genes with no listed transcription factor targets (which can be considered as genes with a lower number of really existing targets) evolve faster than genes with known targets (16.55 ± 0.32 versus 12.37 ± 0.24; Mann–Whitney $p < 10^{-12}$). Similarly, genes with no listed regulatory microRNA targets evolve faster than genes with known targets (17.29 ± 0.14 versus 9.59 ± 0.23; $p < 10^{-12}$). In a similar vein, genes without introns evolve faster than genes with introns (17.14 ± 0.87 versus 13.96 ± 0.20; $p < 10^{-12}$). These facts support the conclusion that all these parameters (number of transcription factor targets, number of regulatory microRNA targets and intronic length) are associated with retardation of protein evolution.

4. DISCUSSION

The obtained data show that in mammals, the evolutionary rate of a protein is much more strongly related to the evolutionary rate of coexpressed proteins (and proteins of the same biological pathway) than to the expression level of its encoding gene. This regularity holds in all variants of analysis: for four different types of data (microarray measurements of gene expression, ESTs estimations of gene expression, protein interactions and biological pathways) and three estimations of evolutionary rate (protein distances and the ratios of non-synonymous to synonymous nucleotide substitution rates obtained with two substitution models). These data suggest that mutual coordination with other proteins is more important in protein evolution than an individual (by gene) effect (the toxicity of protein misfolding). This does not mean the prevalence of positive selection: the mutual effect of coexpressed proteins might be due to either positive or negative (purifying) selection. It just means that some protein coalitions (coexpression memberships) evolve more slowly while others evolve more quickly. In this regards, it is interesting that genes with negatively correlated expression show the negative relation of their evolutionary rates. One could hardly say that about a direct negative effect. A more simple explanation is that these proteins just belong to disparate coalitions, which differ in their evolutionary rates.

The expression level of coexpressed genes produces a similarly weak effect. Furthermore, if all studied parameters are introduced to the model simultaneously, the sign of this effect is opposite to the sign of the effect of expression of a given gene. These data contradict the assumption that the similarity of evolutionary rates of interacting proteins is a result of their similarity in expression levels (Fraser *et al.* 2004; Hakes *et al.* 2007).

One more important factor, which is stronger than expression level, is evolutionary age. More recent proteins evolve more quickly. This effect of evolutionary origin cannot be explained by the link between evolutionary origin and expression level because both parameters were introduced into the model simultaneously. A more likely explanation is that the more ancient proteins are members of more conservative coalitions (e.g. they are involved in the 'deeper' layers of organismal systems).

The complexity of gene regulation (estimated by the numbers of transcription factor targets and regulatory microRNA targets in the encoding gene) is another

important cause, which is stronger than gene expression level. Proteins encoded by complexly regulated genes evolve more slowly. This can be explained by a putative phenomenon, which could be called 'regulatory inertia': the change in the more complexly regulated gene requires a greater number of coordinated changes in genes involved in its regulation. Also, more complexly regulated genes are themselves involved mostly in regulation and development. Therefore, changes in them may cause more profound changes in the organismal systems. This suggests stronger purifying selection on more complexly regulated genes. Interestingly, the number of biological pathways in which a protein is involved, produces a much weaker effect. Thus, it is the regulatory inertia that is important for evolutionary rate, not just the involvement of a given protein in many processes.

The intronic length and the ratio of intronic to coding sequence lengths also negatively correlate with protein evolutionary rate. It is known that highly expressed genes have a shorter intronic sequence and a lower ratio of intronic to coding sequence lengths (Castillo-Davis *et al.* 2002; Eisenberg & Levanon 2003; Urrutia & Hurst 2003; Vinogradov 2004). Were the 'misfolding toxicity' effect prevailing, one could expect a positive relationship between intronic length and protein evolutionary rate. However, the relationship is negative. The longer introns suggest more complex regulation of gene expression (Vinogradov 2004, 2006; Sironi *et al.* 2005*a*,*b*). Therefore, the negative relationship between intronic length and protein evolutionary rate can be interpreted in a similar way with transcription factor targets and regulatory microRNA targets (i.e. as the effect of regulatory inertia).

Summing up, the presented data suggest that systemic factors (coordination with coexpressed proteins, evolutionary origin and complexity of gene regulation) prevail in mammal protein evolution over individual (by gene) effects.

This work was supported by the Russian Foundation for Basic Research (RFBR).

REFERENCES

Batada, N. N., Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B. J., Hurst, L. D. & Tyers, M. 2006 Stratus not altocumulus: a new view of the yeast protein interaction network. *PLoS Biol.* **4**, e317. (doi:10.1371/journal.pbio.0040317)

Castillo-Davis, C. I., Mekhedov, S. L., Hartl, D. L., Koonin, E. V. & Kondrashov, F. A. 2002 Selection for short introns in highly expressed genes. *Nat. Genet.* **31**, 415–418.

Drummond, D. A. & Wilke, C. O. 2008 Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**, 341–352. (doi:10.1016/j.cell.2008.05.042)

Drummond, D. A., Raval, A. & Wilke, C. O. 2006 A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.* **23**, 327–337. (doi:10.1093/molbev/msj038)

Eisenberg, E. & Levanon, E. Y. 2003 Human housekeeping genes are compact. *Trends Genet.* **19**, 362–365. (doi:10.1016/S0168-9525(03)00140-9)

Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C. & Feldman, M. W. 2002 Evolutionary rate in the protein interaction network. *Science* **296**, 750–752. (doi:10.1126/science.1068696)

Fraser, H. B., Hirsh, A. E., Wall, D. P. & Eisen, M. B. 2004 Coevolution of gene expression among interacting proteins. *Proc. Natl Acad. Sci. USA* **101**, 9033–9038. (doi:10.1073/pnas.0402591101)

Hakes, L., Lovell, S. C., Oliver, S. G. & Robertson, D. L. 2007 Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proc. Natl Acad. Sci. USA* **104**, 7999–8004. (doi:10.1073/pnas.0609962104)

Jensen, L. J. *et al.* 2009 STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* **37**, D412–D416. (doi:10.1093/nar/gkn760)

Jordan, I. K., Marino-Ramirez, L., Wolf, Y. I. & Koonin, E. V. 2004 Conservation and coevolution in the scale-free human gene coexpression network. *Mol. Biol. Evol.* **21**, 2058–2070. (doi:10.1093/molbev/msh222)

Julenius, K. & Pedersen, A. G. 2006 Protein evolution is faster outside the cell. *Mol. Biol. Evol.* **23**, 2039–2048. (doi:10.1093/molbev/msl081)

Kawahara, Y. & Imanishi, T. 2007 A genome-wide survey of changes in protein evolutionary rates across four closely related species of *Saccharomyces sensu stricto* group. *BMC Evol. Biol.* **7**, 9.

Koonin, E. V. & Wolf, Y. I. 2006 Evolutionary systems biology: links between gene evolution and function. *Curr. Opin. Biotechnol.* **17**, 481–487. (doi:10.1016/j.copbio.2006.08.003)

Koonin, E. V. *et al.* 2004 A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* **5**, R7. (doi:10.1186/gb-2004-5-2-r7)

Larkin, M. A. *et al.* 2007 CLUSTAL W and CLUSTAL X version 2.0. *Bioinformatics* **23**, 2947–2948. (doi:10.1093/bioinformatics/btm404)

Makino, T. & Gojobori, T. 2006 The evolutionary rate of a protein is influenced by features of the interacting partners. *Mol. Biol. Evol.* **23**, 784–789. (doi:10.1093/molbev/msj090)

Pearson, W. R. 1999 Flexible similarity searching with the FASTA3 program package. In *Bioinformatics methods and protocols* (eds S. Misener & S. A. Krawetz), pp. 185–219. Totowa, NJ: Humana Press.

Powers, E. T. & Balch, W. E. 2008 Costly mistakes: translational infidelity and protein homeostasis. *Cell* **134**, 204–206. (doi:10.1016/j.cell.2008.07.005)

Sayers, E. W. *et al.* 2009 Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **37**, D5–D15. (doi:10.1093/nar/gkn741)

Sironi, M., Menozzi, G., Comi, G. P., Bresolin, N., Cagliani, R. & Pozzoli, U. 2005*a* Fixation of conserved sequences shapes human intron size and influences transposon–insertion dynamics. *Trends Genet.* **21**, 484–488. (doi:10.1016/j.tig.2005.06.009)

Sironi, M., Menozzi, G., Comi, G. P., Cagliani, R., Bresolin, N. & Pozzoli, U. 2005*b* Analysis of intronic conserved elements indicates that functional complexity might represent a major source of negative selection on non-coding sequences. *Hum. Mol. Genet.* **14**, 2533–2546. (doi:10.1093/hmg/ddi257)

Su, A. I. *et al.* 2004 A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA* **101**, 6062–6067. (doi:10.1073/pnas.0400782101)

Subramanian, A., Kuehn, H., Gould, J., Tamayo, P. & Mesirov, J. P. 2007 GSEA-P: a desktop application for gene set enrichment analysis. *Bioinformatics* **23**, 3251–3253. (doi:10.1093/bioinformatics/btm369)

Urrutia, A. O. & Hurst, L. D. 2003 The signature of selection mediated by expression on human genes. *Genome Res.* **13**, 2260–2264. (doi:10.1101/gr.641103)

Vinogradov, A. E. 2004 Compactness of human housekeeping genes: selection for economy or genomic design? *Trends Genet.* **20**, 248–253. (doi:10.1016/j.tig.2004.03.006)

Vinogradov, A. E. 2006 'Genome design' model: evidence from conserved intronic sequence in human–mouse comparison. *Genome Res.* **16**, 347–354. (doi:10.1101/gr.4318206)

Vinogradov, A. E. 2009 Global versus local centrality in evolution of yeast protein network. *J. Mol. Evol.* **68**, 192–196. (doi:10.1007/s00239-008-9185-2)

Wernersson, R. & Pedersen, A. G. 2003 RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res.* **31**, 3537–3539. (doi:10.1093/nar/gkg609)

Wolf, Y. I. 2006 Coping with the quantitative genomics 'elephant': the correlation between the gene dispensability and evolution rate. *Trends Genet.* **22**, 354–357. (doi:10.1016/j.tig.2006.04.009)

Wolf, M. Y., Wolf, Y. I. & Koonin, E. V. 2008 Comparable contributions of structural–functional constraints and expression level to the rate of protein sequence evolution. *Biol. Direct.* **3**, 40. (doi:10.1186/1745-6150-3-40)

Yang, Z. 2007 PAML 4: a program package for phylogenetic analysis likelihood. *Mol. Biol. Evol.* **24**, 1586–1591. (doi:10.1093/molbev/msm088)