Molecular Systems Biology (2005) doi:10.1038/msb4100029

www.molecularsystemsbiology.com
Article number: 2005.0021

molecular systems biology

NEWS AND VIEWS

Systemic determinants of gene evolution and function

Eugene V Koonin

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Molecular Systems Biology 13 September 2005; doi:10.1038/msb4100029

What determines a gene's evolutionary rate? In particular, does it depend solely on functional constraints imposed on the structure of the encoded protein or are there higher-level factors related to the selection at the organismal level? These questions seem to be among the most fundamental ones in biology because comprehensive answers will reveal the nature of the links between genome evolution and the phenotypes of organisms. A recent study by Wall *et al* (2005) proves more convincingly than ever before that systemic determinants of gene evolution rate do exist, and an intriguing paper by Fraser (2005) sheds light on some of the underlying mechanisms. However, a recent report by Coulomb *et al* (2005) issues an important warning by showing that some of the intuitively plausible connections discovered by Systems Biology may be due to biases in the data.

Nearly 30 years ago, Wilson *et al* (1977) put forward a general proposition that may be called the rate-dispensability conjecture—the evolutionary rate should be a function of, firstly, the constraints on the function of the given gene (protein) and, secondly, the 'importance' (fitness effect of knockout or dispensability) of the gene for the organism: $R_i=f(P_i)f(Q_i)$ (R_i is the rate of evolution of the given protein, P_i is the probability that a substitution is compatible with the function of this protein, and Q_i is the probability that the organism survives and reproduces without this protein).

The prediction, thus, is that essential (indispensable) genes, on average, should evolve slower than nonessential genes. This conjecture generally follows from Kimura's neutral theory of evolution but is nontrivial given the broad variance of structural–functional constraints on proteins, regardless of their dispensability; in principle, this variance could completely explain the distribution of evolutionary rates among genes without invoking the fitness connection. Thus, empirical tests of the conjecture are of interest, and such tests have been conducted as soon as the combination of genome sequences and genome-wide knockout fitness effect data became available. The results, however, were ambiguous. The first attempt by Hurst and Smith (1999) involving only $\sim$100 orthologous human and mouse genes, for which knockout effect data in mouse were available, failed to detect the predicted connection. A subsequent study by Hirsh and Fraser (2001) dealt with $\sim$300 yeast genes, with quantitative fitness effect data taken from the results of a genome-wide measurement in yeast and the rates derived from a comparison with the nematode orthologs. These authors reported a weak but statistically significant negative correlation between the knockout fitness effect and evolution rate, in accord with the Wilson conjecture. However, when the genes were classified into two categories, essential and nonessential, no significant difference in rates was detected. In contrast, Jordan *et al* analyzed much larger sets of orthologous genes in bacteria for which knockout data were available and came to the conclusion that essential genes, indeed, on average, evolved slower than nonessential ones (Jordan *et al*, 2002). The issue has been further confounded by two studies that examined partial correlations between evolution rate, fitness effect, and expression level of a gene and concluded that the link between evolution rate and fitness effect vanished once expression level was taken into account (Pal *et al*, 2003; Rocha and Danchin, 2004). A recent study by Wall *et al* (2005) makes major strides to finally settle the issue. These authors produced robust estimates of short-term evolutionary rates for $>$3000 orthologous gene sets from four yeast species of the genus *Saccharomyces* and compared them with two independent data sets on the phenotypic effects of yeast gene knockouts and two measures of gene expression (experimentally determined mRNA abundance and codon adaptation index). Now, partial correlation analysis gave an unequivocal answer: a gene's evolutionary rate significantly depends both on its dispensability and on expression level, and the contributions of these two variables are, largely, independent. Thus,'important' genes and genes that are highly expressed tend to evolve slowly, supporting and extending Wilson's conjecture. This is not the final word on the connection between evolutionary rate, dispensability, and expression, as much work remains to be carried out to obtain reliable quantitative estimates of the strength of the dependences involved. It does seem, however, that, at least for yeast, the reality of these links is now established beyond reasonable doubt. The simple and not particularly new methodological lesson from this work is that, in many cases, careful analysis of improved data sets will do more to resolve a fundamental scientific issue than sophisticated theoretical considerations.

Gene dispensability and expression level are not the only functional variables that have been linked to the evolution rate. In the current era of Systems Biology, many researchers have been particularly intrigued by the possibility that gene evolution is affected by the topology of various interaction networks. In particular, negative correlation has been reported to exist between a gene's node degree in protein–protein interaction (Fraser *et al*, 2002) and coexpression networks (Jordan *et al*, 2004) and evolutionary rate. In other words, genes that interact with many other genes either at the level of coexpression or through physical interaction between their protein products tend to evolve slowly. However, at least the

connection between a protein's position in the interaction network and evolutionary rate has been no less contentious than the link with dispensability. Subsequent to the original report on the correlation, one re-analysis failed to confirm the overall connection although the most prolific interactors (network hubs) did seem to evolve slowly (Jordan *et al*, 2003), whereas another study denied the link altogether, suggesting that it was an artifact of protein abundance (Bloom and Adami, 2003). A recent study by Fraser (2005)seems to clarify the issue and provides an intriguing insight into the evolutionary forces that may be at play in network evolution. Fraser partitioned the interaction network hubs into two classes and showed that they dramatically differ in terms of the connection with the evolutionary rate (or, more precisely, the strength of purifying selection measured as the ratio of the rates for synonymous and nonsynonymous positions in coding sequences). It turns out that hubs that interact with numerous partners within a network module (intramodule hubs, also known under the more appealing name of 'party hubs'; Han *et al*, 2004), indeed, are strongly constrained and evolve much slower than either proteins that have no partners at all or intermodule hubs ('date hubs'; Han *et al*, 2004) that interact with partners from different modules. The intermodule hubs are only slightly more constrained than noninteractors. This observation leads to the intuitively plausible hypothesis that organization and functions of network modules tend to be conserved during evolution, whereas intermodule hubs are involved in network rewiring and could be foci of innovation.

Taken together, these recent studies make, perhaps, relatively small but concrete inroads into the domain of Evolutionary Systems Biology (Medina, 2005). This area of inquiry is just making its baby steps, and the road ahead will be long and hard. That this is so, is demonstrated by the recent analysis of Coulomb *et al* (2005), which, while not dealing directly with evolution, is an important note of caution for systems biologists. These authors take on the connection between a gene's position in biological networks, in particular, genome-wide networks of protein–protein interactions and essentiality. It seems intuitively almost obvious that genes with many connections (network hubs) are 'important' and should be essential more often than poorly connected genes; of course, this is perfectly compatible with the observations on slow evolution of both network hubs and essential genes discussed above. Indeed, such a connection between 'centrality and lethality' has been reported by several groups (Jeong *et al*, 2001); apparent links between a gene's essentiality and other topological characteristics of networks, such as clustering coefficient, also have been reported (Yu *et al*, 2004). However, Coulomb *et al* (2005) argue that these effects were caused by biases in the analyzed interaction data that contained a greater number of valid interactions for essential genes. When a supposedly unbiased data set (Ito *et al*, 2001) was analyzed, only a marginal correlation between node degree (centrality) and essentiality was detected, and no dependence at all was seen for other topological features of networks (Coulomb *et al*, 2005).

The current state of Evolutionary Systems Biology is typical of any burgeoning discipline: it is clear that there are important signals out there but our ability to discern and understand these signals is hampered both by inaccuracies and biases in the data and the inadequacy of the existing theoretical models. These difficulties notwithstanding, we should be motivated by the (I believe, reasonable) hope that, as this field matures, our one-dimensional understanding of genome evolution develops into a multidimensional picture of evolution of organisms as systems.

References

Bloom JD, Adami C (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

Coulomb S, Bauer M, Bernard D, Marsolier-Kergoat MC (2005) Gene essentiality and the topology of protein interaction networks. *Proc Biol Sci* **272:** 1721–1725

Fraser HB (2005) Modularity and evolutionary constraint on proteins. *Nat Genet* **37:** 351–352

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002) Evolutionary rate in the protein interaction network. *Science* **296:** 750–752

Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP, Vidal M (2004) Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* **430:** 88–93

Hirsh AE, Fraser HB (2001) Protein dispensability and rate of evolution. *Nature* **411:** 1046–1049

Hurst LD, Smith NG (1999) Do essential genes evolve slowly? *Curr Biol* **9:** 747–750

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA* **98:** 4569–4574

Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature* **411:** 41–42

Jordan IK, Marino-Ramirez L, Wolf YI, Koonin EV (2004) Conservation and co-evolution in the scale-free human gene co-expression network. *Mol Biol Evol* **21:** 2058–2070

Jordan IK, Rogozin IB, Wolf YI, Koonin EV (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res* **12:** 962–968

Jordan IK, Wolf YI, Koonin EV (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

Medina M (2005) Genomes, phylogeny, and evolutionary systems biology. *Proc Natl Acad Sci USA* **102** (Suppl 1)**:** 6630–6635

Pal C, Papp B, Hurst LD (2003) Genomic function: rate of evolution and gene dispensability. *Nature* **421:** 496–497; discussion 497–498

Rocha EP, Danchin A (2004) An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol Biol Evol* **21:** 108–116

Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, Feldman MW (2005) Functional genomic analysis of the rates of protein evolution. *Proc Natl Acad Sci USA* **102:** 5483–5488

Wilson AC, Carlson SS, White TJ (1977) Biochemical evolution. *Annu Rev Biochem* **46:** 573–639

Yu H, Greenbaum D, Xin Lu H, Zhu X, Gerstein M (2004) Genomic analysis of essentiality within protein networks. *Trends Genet* **20:** 227–231