Cell, Vol. 101, 573–576, June 9, 2000, Copyright ©2000 by Cell Press

# The Impact of Comparative Genomics on Our Understanding of Evolution

Minireview

**Eugene V. Koonin,* L. Aravind, and Alexey S. Kondrashov**
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Bethesda, Maryland 20894

As recently as 1995, the count of completely sequenced genomes of cellular life forms was exactly zero. The shotgun sequencing of the first genome, that of the bacterium *Haemophilus influenzae*, opened the floodgate, and 5 years later, there are about 30 bacterial, archaeal, and eukaryotic genomes to compare, with dozens more in the pipeline (http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html). The phrase "genomic revolution" has become commonplace, and is certainly appropriate in a purely quantitative sense. What may not be immediately obvious is what, if any, truly fundamental insights into the organization and evolution of biological systems come from complete genome sequences. Some of the new generalizations emerging from the "first round" of genome sequences, that hardly could have been anticipated in the pregenomic era, are discussed here.

### *Unity and Diversity of Life Revealed by Comparative Genomics*

Genome comparisons reveal the remarkable unity of all living things. In simple genomes, those of bacteria and archaea, about 70% of the genes belong to ancient conserved families that include orthologs (direct evolutionary counterparts) from organisms whose common ancestors lived billions of years ago (Tatusov et al., 2000). The same comparisons, however, also reveal striking diversity as most of these ancient families are represented only in small subsets of the sequenced genomes. Among 2100 ancient families of orthologs, identified by comparison of 21 bacterial, archaeal, and eukaryotic genomes, only 80 ($<$5%) are ubiquitous (Figure 1); most of these universal families include core components of the translation and transcription machineries.

This remarkable plasticity of the genome is perhaps the biggest news that comes directly from genome comparisons. The almost inevitable corollary is that horizontal dissemination of genes and lineage-specific gene loss are not just interesting evolutionary quirks, but major forces, at least in the evolution of prokaryotes. Gene loss is evident on many occasions, particularly in the evolution of parasites. In contrast, horizontal gene transfer, while plausible as one of the principal explanations for the mosaic phyletic distribution of most of the conserved protein families, can be difficult to prove in specific cases. Nevertheless, the correlation between an organism's lifestyle and the apparent source of horizontally acquired genes is, along with phylogenies of individual protein families, a strong argument in support of horizontal gene transfer (Aravind et al., 1998; Stephens et al., 1998). This correlation is seen in certain parasitic bacteria that have more genes which seem to be acquired from their eukaryotic hosts than their free-living relatives, and in hyperthermophilic bacteria that possess an excess of genes of apparent archaeal origin (see http://www.ncbi.nlm.nih.gov:80/cgi-bin/Entrez/taxik?gi=141 and http://www.ncbi.nlm.nih.gov:80/cgi-bin/Entrez/taxik?gi=138). It could be argued that shared genes in archaeal and bacterial hyperthermophiles are ancestral, rather than horizontally transferred. However, the fact that hyperthermophilic bacteria possess archaeal versions of many genes, along with typically bacterial ones, strongly supports the horizontal transfer interpretation.

### *The Tree of Life—Is It a Legitimate Depiction of Evolution?*

The prevalence of horizontal gene transfer and gene loss revealed by genome comparisons suggests that the "Tree of Life," the canonical representation of evolution since Darwin and Haeckel, could be misleading, at least in its deep branches. Strictly speaking, a tree cannot precisely reflect the phylogeny if there had been any horizontal gene transfer at all, but if it has been extensive, the tree can become meaningless (Doolittle, 1999). Nevertheless, hope remains that extant genomes still contain a sizable fraction of genes that have been vertically inherited throughout the entire evolutionary history from the last universal common ancestor; subunits of large macromolecular complexes, such as ribosomal proteins, are reasonable candidates for this vertically inherited component of the genomes. If so, analysis of a large number of gene families or all-against-all comparisons of protein sequences encoded in complete genomes could reveal a meaningful consensus topology. Several such analyses seem to converge at the conclusion that the large-scale phylogenetic signal has not been completely washed-out, and at least the three primary domains of life, bacteria, archaea, and eukaryotes, are robustly recovered in whole-genome trees, horizontal transfer notwithstanding (Fitz-Gibbon and House, 1999; Snel et al., 1999). On the other hand, the phylogeny of the major bacterial lineages does not seem to emerge reliably at all, suggesting pervasive horizontal transfer, rapid evolution at the onset of each domain, or most likely, both (Teichmann and Mitchison, 1999).

### *Common Ancestry and Convergence in Protein Evolution*

A somewhat less obvious conclusion drawn from the observations on genomic mosaicism is that, for most of the essential tasks in the cell, there are at least two different solutions. These solutions may involve functionally analogous proteins of different superfamilies or proteins whose structural folds are unrelated, but that possess the same biochemical activity (Galperin et al., 1998). In the course of evolution, analogous proteins could have displaced one another via combination of horizontal gene transfer and gene loss, a phenomenon called *nonorthologous gene displacement*. The case of two lysyl-tRNA synthetases, one of which, found in eukaryotes and most bacteria, belongs to class II of

*To whom correspondence should be addressed (e-mail: koonin@ncbi.nlm.nih.gov).

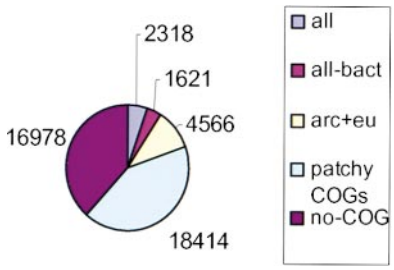


Figure 1. Phyletic Patterns in Conserved Protein Families—Indications of Pervasive Gene Loss and Lateral Transfer

The sequences of 43,897 proteins from completely sequenced genomes of 16 bacteria, 4 archaea, and yeast were classified into clusters of probable orthologs (COGs) including all species, all bacterial species (and possibly some of the archaea and/or yeast) (all-bact), all archaea and yeast (and possibly some bacteria) (arc+eu), families with less regular phylogenetic patterns (patchy COGs), and those proteins that did not belong to a family of orthologs (no-COG). The number of proteins in each category is indicated. The data are from (Tatusov et al., 2000; http://www.ncbi.nlm.nih.gov/COG/).

aminoacyl-tRNA synthetases, and the other one, recently identified in archaea and three bacterial lineages, belongs to the unrelated class I, is a perfect example of two independent evolutionary inventions used to perform the same essential function (Woese et al., 2000; Table 1). Nonorthologous displacement results in complementary phyletic patterns with some overlap, because certain genomes encode both implementations of the respective function (Table 1).

The field of molecular evolution largely operates on the premise that significant sequence similarity implies common ancestry, but the specter of sequence convergence due to common function has been raised repeatedly. The widespread occurrence of nonorthologous displacement leads to the conclusion that convergent evolution of protein domains with significant sequence similarity is highly unlikely, if possible at all.

### *Instability of Gene Order*

Rampant as horizontal gene transfer may be, protein families show considerable stability over billions of years of evolution. But this is not so for gene order in prokaryotes. Synteny is broken even between bacterial species within the same genus that possess largely the same sets of highly conserved orthologs (Himmelreich et al., 1997). At moderate evolutionary distances, long-range conservation of gene order becomes undetectable. Only a minimal number of essential operons that encode subunits of macromolecular complexes are universally conserved in the prokaryotic world (Dandekar et al., 1998). The ability of bacteria and archaea to tolerate and probably utilize recombinational redistribution of genes is amazing, and is in line with the relative unimportance of large-scale gene order. Rather unexpectedly, it seems that prokaryotic genomes, after all, can be thought of as "bags of genes," or more precisely, of operons.

In contrast, in multicellular eukaryotes, synteny is largely preserved across considerable evolutionary range, for example between actinopterygian fishes and humans (Gellner and Brenner, 1999). In some cases, such as the Hox clusters, partial synteny may endure the entire breadth of the animal kingdom. The conservation of gene order is likely to be maintained in part due to the presence of key regulatory elements in the intergenic regions, and insulation from promiscuous lateral acquisition of DNA by the separation of the germline from the soma in multicellular eukaryotes. A complementary explanation of the conservation of synteny in animals, compared to prokaryotes, is that the actual scales of evolution involved may be very different, with immeasurably more generations separating bacterial lineages than animal phyla.

Table 1. Complementary Phyletic Patterns in Gene Families—Manifestation of Nonorthologous Gene Displacement

| | Species[a] | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Archaea | | | | | E | Bacteria[b] | | | | | | | | | | | | | | |
| | Mj | Mt | Af | Ph | Ap | Sc | Aa | Tm | Dr | Ss | Ec | Bs | Mh | Hi | Hp | Uu | Mg | Bb | Tp | Ct | Rp |
| COG/protein function[c] | | | | | | | | | | | | | | | | | | | | | |
| Lysyl-tRNA synthetase class II (COG1190) | − | − | − | − | − | + | + | + | + | + | + | + | + | + | + | + | + | − | − | + | − |
| Lysyl-tRNA synthetase class I (COG1384) | + | + | + | + | + | − | − | − | − | − | − | − | − | − | − | − | − | + | + | − | + |
| Fructose-1,6-bisphosphate aldolase (COG0191)[d] | − | − | − | − | − | + | + | + | + | + | + | + | + | + | + | + | + | + | + | − | − |
| DhnA-type Fructose-1,6-bisphosphate aldolase (COG1830)[d] | + | + | + | + | + | − | + | − | − | − | + | − | − | − | − | − | − | − | − | + | − |

[a] Abbreviations: Aa, *Aquifex aeolicus*; Af, *Archaeoglobus fulgidus*; Ap, *Aeropyrum pernix*; Bb, *Borrelia burgdorferi*; Bs, *Bacillus subtilis*; Ct, *Chlamydia trachomatis*; Ec, *Escherichia coli*; Hi, *Haemophilus influenzae*; Hp, *Helicobacter pylori*; Mg, *Mycoplasma genitalium* & *pneumoniae*; Mj, *Methanococcus jannaschii*; Mh, *Methanobacterium thermoautotrophicum*; Ph, *Pyrococcus horikoshii*; Rpr, *Rickettsia prowazekii*; Sce, *Saccharomyces cerevisiae*; Ssp, *Synechocystis* sp.; Tm, *Thermotoga maritima*; Tp, *Treponema pallidum*; Uu, *Ureaplasma urealyticum*; E stands for Eukaryotes.

[b] Several completely sequenced bacterial genomes are omitted for brevity (their inclusion would not affect the phyletic patterns).

[c] The Cluster of Orthologous Groups of proteins (COG) numbers are from http://www.ncbi.nlm.nih.gov/COG/.

[d] The two aldolases are distantly related but not orthologous. Note that the patterns are not perfectly complementary because *A. aeolicus* and *E. coli* encode both types of aldolases, whereas *R. prowazekii*, which lacks glycolysis, has none.

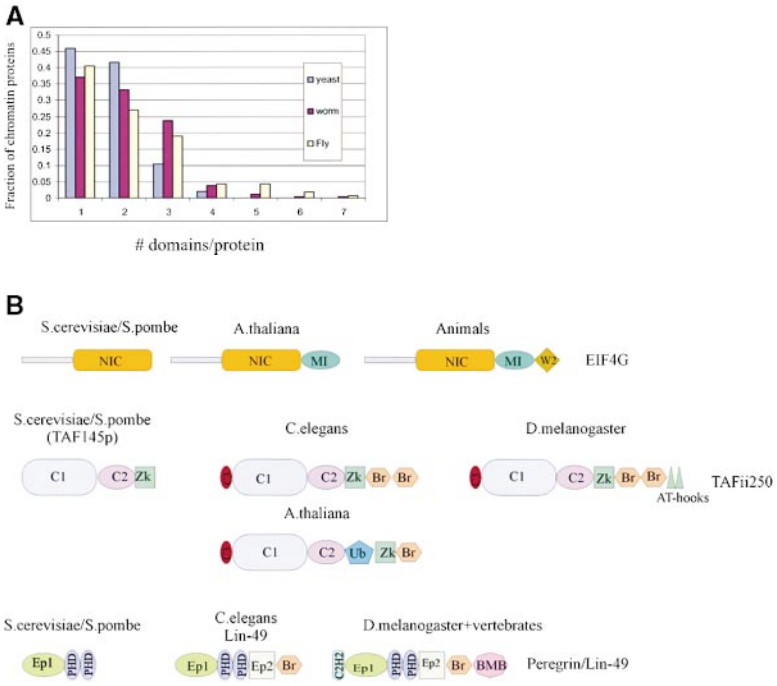


Figure 2. Domain Accretion in the Evolution of Eukaryotic Proteins

Domains were identified using the PSI-BLAST program and libraries of domain-specific sequence profiles (Chervitz et al., 1998). (A) Distribution of chromatin-associated proteins by the number of detectable, distinct domains for three eukaryotic species.

(B) Domain accretion in three orthologous sets of eukaryotic proteins—translation initiation factor eIF4G, transcription factor TAFii250, and chromatin remodeling factor Lin-49.

Distinct domains are designated by unique shape and color. Domain key: NIC (*N*MD2, e*I*F4G, *C*BP80), domain conserved in various cap binding proteins; MI, domain conserved in *M*A-3 proteins and e*I*F4G; W2, domain conserved in several translation factors (named after two invariant tryptophans); Zk, Zn knuckle; Br, Bromo domain; Ub, ubiquitin; PHD, PHD finger; C2H2, C2H2 finger; Ep1, 2, distinct conserved domains found in *E*nhancer of *P*olycomb; C1, C2, C3, uncharacterized conserved domains.

### *Domain Accretion and the Notion of Orthology*

The concept of orthology between genes is central to any comparative-genomic analysis. Intergenomic comparisons suggest, however, that this notion does not always adequately describe the relationship between genes connected by vertical descent. Indeed, orthology normally implies not only a vertical connection, but a complete structural and functional correspondence ("the same" gene in different organisms). This "strong" concept of orthology breaks down for genes coding for complex, multidomain proteins. Portions of such proteins (genes) may be related by vertical descent, but they also accrete new domains in different lineages and concomitantly acquire new functions. This phenomenon is relatively rare, albeit important, in prokaryotes, but appears to be much more prevalent in the evolution of eukaryotes. Domain accretion is manifest in a wide range of proteins, from components of the basic translation machinery to proteins involved in diverse regulatory processes, but is perhaps most prominent in proteins involved in chromatin remodeling and transcription regulation. The complexity of protein domain organization in general, and in orthologous gene lineages in particular, seems to increase with the organism's complexity (Figure 2). For example, *Drosophila* or vertebrate proteins have accreted new domains and evolved more complex domain organizations compared to nematode, and especially, fungal orthologs (Chervitz et al., 1998; Rubin et al., 2000; Figure 2). Much of the evolutionary process should be thought of and analyzed in terms of domains, rather than proteins (genes), as primary evolving units that recombine to form multiple domain architectures. Domain architectures may serve as unique evolutionary markers (shared derived characters) that help in cladistic analysis.

### *The Nature of Evolutionary Innovation—Where Do New Genes Come from?*

Even a preliminary analysis of eukaryotic genomes shows that they possess thousands of genes that have no counterparts in prokaryotes (Hutter et al., 2000; Rubin et al., 2000). What are the sources of these innovations? We certainly do not know all of them, but several can be deciphered. On many occasions, "new" domains are, in fact, old ones that have been modified to the point that their origin cannot be easily recognized. This seems to be the case with several domains involved in protein–protein interactions, such as von Willebrand A, Fibronectin type III, immunoglobulin and SH3 modules. These domains show vast proliferation in complex eukaryotes, but their distantly related homologs in prokaryotes and/or unicellular eukaryotes could be identified only using the most powerful of the available methods for sequence analysis (Ponting et al., 2000). A clue to another important source of innovation could lie in the observation that many of the new eukaryotic domains are completely $\alpha$-helical; the adaptor domains of the programmed cell death system (the death domain, the death effector domain and the Caspase recruiting domain) illustrate this point. It seems likely that such $\alpha$-helical domains have evolved from condensed coiled-coil structures that are present in considerable amounts even in prokaryotes, but are particularly abundant in eukaryotes.

On other occasions, the origin of novel eukaryotic proteins may be utterly unexpected. One of the key regulators of eukaryotic development, hedgehog, is a case in point. Hedgehog seems to have evolved by fusion of an intein, a prokaryotic mobile element involved in protein splicing, and a dramatically modified metalloprotease domain (Hall et al., 1997). Notably, in the nematodes, the intein-derived autoprocessing domain

combines with several distantly related extracellular domains, giving rise to a whole spectrum of potential signaling molecules (Aspock et al., 1999). At least a few domains with complex topology and distinct biochemical activities also appear to have been "invented" by eukaryotes, such as the SET domain, a predicted protein-lysine methylase involved in chromatin remodeling, and the cap binding translation factor eIF4E.

### *The Nature of Adaptation and Selective Forces in Evolution—Do We Understand the Connection between Genotype and Phenotype through Comparative Genomics?*

If an essential gene is displaced by a horizontally transferred one (be it related or unrelated) with the same function, the latter should for the most part be adaptive (confer a distinct evolutionary advantage to the organism) to be fixed in the population. Generally, an alien gene is likely to be at a disadvantage compared to a resident, well-integrated gene with the same function. The reverse may be true, however, after a significant change of the environment, which could apply to the origin of bacterial hyperthermophily via continuous acquisition of archaeal genes. At least one case where the nature of the advantage is clear has been described—eukaryotic isoleucyl-tRNA synthetase confers antibiotic resistance to bacteria whose genome it invades via horizontal transfer (Brown et al., 1998). Comparative genome analysis reveals other features that appear to be adaptive, such as lineage-specific expansion of gene families. A striking example of this is the expansion of the photosensory PAS domain in the photosynthetic cyanobacterial lineage, which is without any precedence in nonphotosynthetic organisms. More frequently, however, the teleology behind the presumed adaptation remains unclear.

More generally, what we realize with due humility from our first forays into complete-genome-scale comparative genomics, is that we do not truly understand the connection between the genome and the phenotype of an organism. Comparative analysis of the genomes of archaeal and bacterial thermophiles reveals numerous interesting features, including strong evidence of extensive, specific horizontal gene transfer, but fails to identify the genomic basis for thermophily. Similarly, a comparison of the genome of the extreme radioresistant bacterium *Deinococcus radiodurans* to those of other bacteria showed many peculiarities in its DNA repair and stress response systems (White et al., 1999), but failed to explain how this organism survives under irradiation of an intensity that makes pyrex glass crack. We believe that the problem of the genome–phenotype connection, which, in a sense, is the central theme of biology, can be solved only through an experimental program strategically planned on the basis of comparative-genomic results. Much of the biological research of the next few decades is likely to develop along these lines.

#### Selected Reading

Aravind, L., Tatusov, R.L., Wolf, Y.I., Walker, D.R., and Koonin, E.V. (1998). Trends Genet. *14*, 442–444.

Aspock, G., Kagoshima, H., Niklaus, G., and Burglin, T.R. (1999). Genome Res. *9*, 909–923.

Brown, J.R., Zhang, J., and Hodgson, J.E. (1998). Curr. Biol. *8*, R365–R367.

Chervitz, S.A., Aravind, L., Sherlock, G., Ball, C.A., Koonin, E.V., Dwight, S.S., Harris, M.A., Dolinski, K., Mohr, S., Smith, T., et al. (1998). Science *282*, 2022–2028.

Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Trends Biochem. Sci. *23*, 324–328.

Doolittle, W.F. (1999). Science *284*, 2124–2129.

Fitz-Gibbon, S.T., and House, C.H. (1999). Nucleic Acids Res. *27*, 4218–4222.

Galperin, M.Y., Walker, D.R., and Koonin, E.V. (1998). Genome Res. *8*, 779–790.

Gellner, K., and Brenner, S. (1999). Genome Res. *9*, 251–258.

Hall, T.M., Porter, J.A., Young, K.E., Koonin, E.V., Beachy, P.A., and Leahy, D.J. (1997). Cell *91*, 85–97.

Himmelreich, R., Plagens, H., Hilbert, H., Reiner, B., and Herrmann, R. (1997). Nucleic Acids Res. *25*, 701–712.

Hutter, H., Vogel, B.E., Plenefisch, J.D., Norris, C.R., Proenca, R.B., Spieth, J., Guo, C., Mastwal, S., Zhu, X., Scheel, J., and Hedgecock, E.M. (2000). Science *287*, 989–994.

Ponting, C.P., Schultz, J., Copley, R.R., Andrade, M.A., and Bork, P. (2000). Adv. Prot. Chem. *54*, 185–244.

Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R., Fleischmann, W., et al. (2000). Science *287*, 2204–2215.

Snel, B., Bork, P., and Huynen, M.A. (1999). Nat. Genet. *21*, 108–110.

Stephens, R.S., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L., Mitchell, W., Olinger, L., Tatusov, R.L., Zhao, Q., et al. (1998). Science *282*, 754–759.

Tatusov, R.L., Galperin, M.Y., Natale, D.A., and Koonin, E.V. (2000). Nucleic Acids Res. *28*, 33–36.

Teichmann, S.A., and Mitchison, G. (1999). J. Mol. Evol. *49*, 98–107.

White, O., Eisen, J.A., Heidelberg, J.F., Hickey, E.K., Peterson, J.D., Dodson, R.J., Haft, D.H., Gwinn, M.L., Nelson, W.C., Richardson, D.L., et al. (1999). Science *286*, 1571–1577.

Woese, C.R., Olsen, G.J., Ibba, M., and Söll, D. (2000). Microbiol. Mol. Biol. Rev. *64*, 202–236.