Invited Talk

Comparative Genomics: Is It Changing the Paradigm of Evolutionary Biology?

Eugene V. Koonin
koonin@ncbi.nlm.nih.gov

National Center for Biotechnology Information, National Library of Medicine
National Institutes of Health
Bethesda MD 20894, USA

About 20 complete genome sequences of cellular life forms — bacteria, archaea an eukaryotes — are currently available, and many more are in the pipeline. Considerable comparative analysis of these genomes has already been performed, and while even more challenging work lies ahead, it is fair to ask at this juncture, what is the impact of this research on biology in general. In my opinion, comparative analysis of complete genome has already affected our ideas of what biological evolution is to such an extent that it is appropriate to claim a paradigm shift in evolutionary biology.

Computer analysis of complete genomes of unicellular organisms shows that protein sequences are in general highly conserved in evolution, with at least 70% of them containing ancient conserved regions. This allows us to delineate families of orthologs across a wide phylogenetic range and in many cases, predict protein functions with reasonable confidence. Once a robust set of such orthologous families is established, it is possible to examine the pattern of phylogenetic representation (or for brevity, simply 'phylogenetic pattern') for each of them. Such an examination readily shows that for the great majority of orthologous families, the phylogenetic distribution is quite patchy and in many cases unexpected. Only $\sim$100 families, the majority of them including components of the translation machinery, are universally conserved in all sequenced genomes. These observations indicate that horizontal gene transfer and lineage-specific gene loss are not inconsequential evolutionary quirks but rather prevailing forces of evolution, at least in the prokaryotic world. On many occasions, in detailed studies of protein superfamilies and even entire functional systems, such as those for DNA repair and programmed cell death, it is now possible to reconstruct detailed evolutionary scenarios that account for a number of distinct events of horizontal gene transfer and gene loss. These detailed studies show that at the level of individual genes, horizontal gene transfer and gene loss are complemented by numerous recombination events that manifest in domain rearrangement at the protein level. Previously, such rearrangements were associate primarily with exon shuffling but the analysis of complete genomes shows that they are critically important also in the prokaryotic world where this mechanism is not operative.

Examination of phylogenetic patterns for families of orthologous proteins also results in more specific conclusions some of which may have far-reaching consequences. In particular, it is now clear that the basic DNA replication machineries (that is, the replicative DNA polymerases, primases, helicases, and several other proteins) in bacteria and in archaea/eukaryotes are ***not*** orthologous and may have evolved independently. This is in a sharp contrast with the remarkable conservation of the components of the translation apparatus and the core transcription machinery. The simplest, even if unconventional interpretation of these observations is that the common ancestor of all extant cellular life forms actually had an RNA genome that, however, encoded an advanced translation machinery, the main proteins currently used for transcription (which at that time might have been involved in genome replication) and probably a considerable repertoire of metabolic enzymes. Taking into account the common chemical knowledge that long RNA molecules are quite unstable and the conclusions on the major role of gene sampling discussed above, we arrive to an unexpected, speculative but not unplausible picture of the common ancestor of all known life (the so-called cenancestor). The cenancestor might not even have had a defined genome in the sense all modern organisms have it but rather could be a loose collection of a large number of relatively small, in a sense virus-like, RNA-based genetic elements.

Further genome sequencing, particularly of genomes of deep-branching organisms will put these concepts to test and in any case, will add more substance for critical analysis. This must be complemented by further developments of methods for theoretical and ultimately experimental analysis of evolution on the basis of multiple genome comparison.

Acknowledgements

I thank L. Aravind for numerous stimulating discussions, without which these ideas could not have been developed in a clear form. I am aware and appreciative of the major contribution of Carl Woese to our understanding of the possible nature of the Universal Ancestor [5].

References

[1] Mushegian, A.R. and Koonin, E.V., A minimal gene set for cellular life derived by comparison of complete bacterial genomes, *Proc. Natl. Acad. Sci. U.S.A.*, 93:10268–10273, 1996.

[2] Koonin, E.V., Mushegian, A.R., Galperin, M.Y., and Walker, D.R., Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea, *Mol. Microbiol.*, 25:619–637, 1997.

[3] Tatusov, R.L., Koonin, E.V., and Lipman, D.J., A genomic perspective on protein families, *Science*, 278:631–637, 1997.

[4] Huynen, M.A. and Bork, P., Measuring genome evolution, *Proc. Natl. Acad. Sci. U.S.A.*, 95:5849–5856, 1998.

[5] Woese, C. The universal ancestor, *Proc. Natl. Acad. Sci. U.S.A.*, 95:6854–6859, 1998.