Cell, Vol. 111, 13–16, October 4, 2002, Copyright ©2002 by Cell Press

Sequence First. Ask Questions Later.

Genome Minireview

Arend Sidow[1]
Department of Pathology
Department of Genetics
SUMC R248B
300 Pasteur Drive
Stanford University
Stanford, California 94305

Comparative sequence analyses of eukaryotic genes and genomic regions are beginning to provide a wealth of information that is directly relevant to human biology. Functional changes that set us apart from apes are identifiable, as are functional constraints in proteins and genomic elements that arose in our relatively distant phylogenetic past.

With respect to understanding human gene function, the use of experimentally amenable model organisms has perhaps been the most important paradigm in basic biomedical research. Comparative sequence analysis of model organism genes and genomes has recently been emerging as an approach that is complementary to experimentation. Its biomedical importance will grow in the foreseeable future, but fulfilling its ultimate potential will require whole-genome sequence from species that are at best marginal with respect to experimentation.

Consider our human lineage (red line, Figure 1A) and our past advances in biological organization (green, Figure 1A). The traits we share with our model organism relatives define the level of biological organization for which they are models with respect to human biology. Each of these traits is the result of a molecular collaboration of a vast number of genomic elements, some of which are expressed as gene products. Each element arose at some point during our evolutionary history as the result of mutations in an ancestral population. Once an element confers a selective advantage through an advantageous phenotype, it may increase in frequency and eventually become fixed, thus contributing to evolutionary change in that lineage. It is then under selective constraint and generally exhibits a slower rate of evolution than nonfunctional DNA (Li, 1997).

Conservation of Functional Elements

This depressed evolutionary rate of functional elements facilitates identification of previously uncharacterized ones by comparative sequence analysis. For example, a novel apolipoprotein gene, APOAV, was found by aligning the fully sequenced mouse apolipoprotein cluster with its orthologous region on human 11q23 (Pennacchio et al., 2001). It was shown that human single-nucleotide polymorphisms (SNPs) in APOAV are associated with elevated triglyceride levels, which was consistent with results from transgenic and knockout experiments on mice carried out as part of the same study. Using similar comparative methods, another group identified putative regulatory regions in the Stem Cell Leukemia locus (Göttgens et al., 2001). Because pairwise comparisons have limited power, chick, *Fugu*, and zebrafish were included as well in a second study (Göttgens et al., 2002) that identified several conserved promoter elements essential for proper spatiotemporal expression of SCL.

Such inclusion of additional sequences is currently a favored strategy, but it has to be borne in mind that the most distantly related species in a comparison determine what kind of element can be identified as conserved. This is because orthologs of an element can only be found in the descendants of the ancestor in which it first evolved (Figure 1B). This is obvious for proteins that are markers for a particular level of biological organization, such as eukaryotic cell biology or metazoan signaling (left two trees, Figure 1B), but the same applies to regulatory elements (right two trees, Figure 1B). The set of sequenced genomes is still too sparse to adequately cover many levels of biological organization (Figure 1A). In particular, identification of the majority of functional elements relevant to human biology requires placental genomes beyond those of human, mouse, and rat.

Quantification of Constraints

Building a parts list is important, but multiple sequence alignments by themselves do not quantify conservation and allow only limited inference as to which conserved functional element is more constrained than another. By contrast, estimates of past rates of evolution provide statistical power to quantify the strength of constraints with high resolution (Sumiyama et al., 2001; Simon et al., 2002; Pupko et al., 2002). Generating such estimates requires robust multiple sequence alignments in which the proportion of unambiguously aligned positions is maximized. These analyses also require sufficient sequence diversity, measured in substitutions per site over the tree relating the sequences, to distinguish more and less constrained regions within the functional element (Figure 1C). Orthologs that poison alignments with too many insertions or deletions, because of an accelerated rate of evolution, or because the last common ancestor is too distant are not used (struck out in Figure 1C). This pertains regardless of whether, for example, a slowly evolving eukaryotic protein (Figure 1C, left tree) or a quickly evolving mammalian regulatory element (middle tree) is analyzed. On the other hand, a functional element may be so constrained that any ortholog can be reliably aligned (right tree), but each contributes only a small amount of sequence variation. In that case, more sequences can be obtained to capture the same number of substitutions per site as in a smaller sample of a more quickly evolving element (middle tree).

Proteins

Having obtained an alignment, local evolutionary rates are estimated to quantify regional constraint. For proteins, constraints can be structural (folding or packing) or functional (catalysis or interaction). Generally, the strongest constraints reside in the regions of most important function (Simon et al., 2002). If structural data are available, functional constraints can be dissociated from structural constraints. This is illustrated by the DNA

[1] Correspondence: arend@stanford.edu

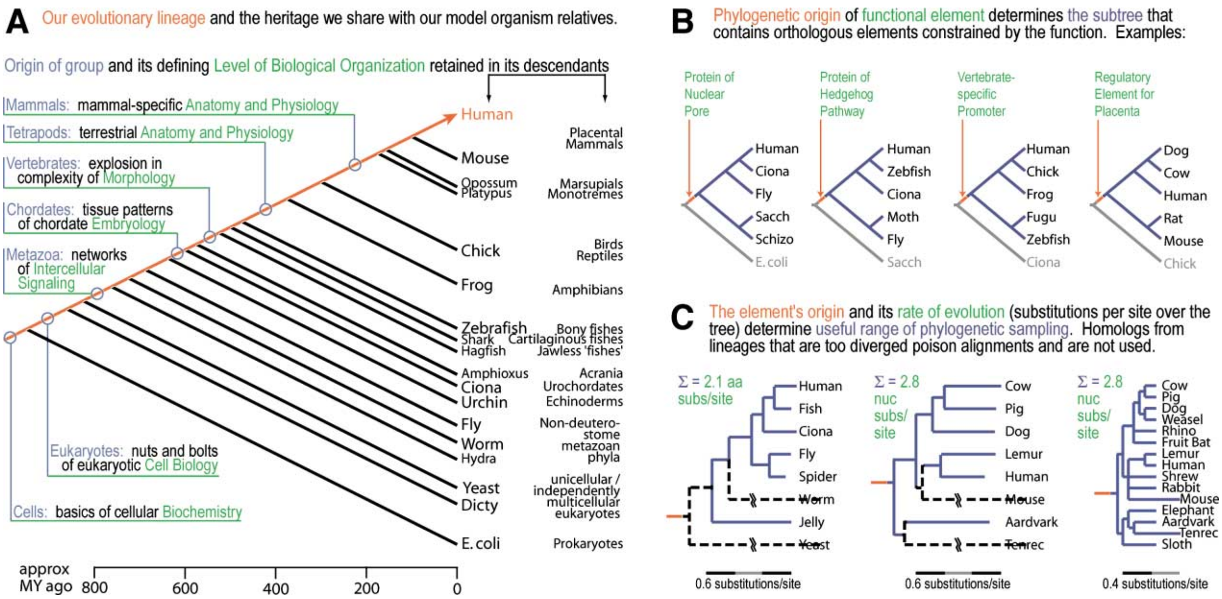
A
Our evolutionary lineage and the heritage we share with our model organism relatives.
Origin of group and its defining Level of Biological Organization retained in its descendants
Mammals: mammal-specific Anatomy and Physiology
Tetrapods: terrestrial Anatomy and Physiology
Vertebrates: explosion in complexity of Morphology
Chordates: tissue patterns of chordate Embryology
Metazoa: networks of Intercellular Signaling
Eukaryotes: nuts and bolts of eukaryotic Cell Biology
Cells: basics of cellular Biochemistry
Human
Mouse
Opossum
Platypus
Chick
Frog
Zebrafish
Shark
Hagfish
Amphioxus
Ciona
Urchin
Fly
Worm
Hydra
Yeast
Dicty
E.coli
Placental Mammals
Marsupials
Monotremes
Birds
Reptiles
Amphibians
Bony fishes
Cartilaginous fishes
Jawless 'fishes'
Acrania
Urochordates
Echinoderms
Non-deutero-stome metazoan phyla
unicellular / independently multicellular eukaryotes
Prokaryotes
approx MY ago
800
600
400
200
0
B
Phylogenetic origin of functional element determines the subtree that contains orthologous elements constrained by the function. Examples:
Protein of Nuclear Pore
Human
Ciona
Fly
Sacch
Schizo
E.coli
Protein of Hedgehog Pathway
Human
Zebfish
Ciona
Moth
Fly
Sacch
Vertebrate-specific Promoter
Human
Chick
Frog
Fugu
Zebfish
Ciona
Regulatory Element for Placenta
Dog
Cow
Human
Rat
Mouse
Chick
C
The element's origin and its rate of evolution (substitutions per site over the tree) determine useful range of phylogenetic sampling. Homologs from lineages that are too diverged poison alignments and are not used.
Σ = 2.1 aa subs/site
Human
Fish
Ciona
Fly
Spider
~~Worm~~
Jelly
~~Yeast~~
0.6 substitutions/site
Σ = 2.8 nuc subs/site
Cow
Pig
Dog
Lemur
Human
~~Mouse~~
Aardvark
~~Tenrec~~
0.6 substitutions/site
Σ = 2.8 nuc subs/site
Cow
Pig
Dog
Weasel
Rhino
Fruit Bat
Lemur
Human
Shrew
Rabbit
Mouse
Elephant
Aardvark
Tenrec
Sloth
0.4 substitutions/site

Figure 1. Concepts in Phylogeny as It Relates to Comparative Genomics

(A) Tree of select organisms (large font: whole-genome sequence obtained or slated for sequencing) and the higher taxonomic groups they represent, drawn to emphasize major innovations in our evolutionary history. Notice that there is something of an evolutionary ladder, but its rungs are not extant organisms but rather our common ancestors with them. The advances could either be described in terms of the organismal biology or the functional elements encoding it.

(B) Illustration of the relationship between origin of functional elements and the maximum set of organisms in which they may be found.

(C) Illustration of the relationship of rate of evolution and phylogenetic sampling for capturing variation in a functional element. Struck-out branches signify highly diverged sequences that are not used in the underlying multiple sequence alignments. Variation is measured in substitutions per site (amino acids, aa; nucleotides, nuc) over the entire tree relating the sequences.

binding domain of vertebrate Mybs, which consists of three Myb repeats with identical folds (Tahirov et al., 2002). The functional differences between these structurally equivalent domains are reflected in their evolutionary rates. The first repeat binds DNA nonspecifically at phosphates; it evolves most quickly. Repeats 2 and 3 bind DNA with specific base contacts and evolve more slowly. Constraints can also be quantified at higher resolution than at the level of domains. For example, within Myb repeats 2 and 3, the most constrained regions contain the residues that specifically contact DNA. On the surface of repeat 3, angled away from the DNA, is another strongly constrained region (Figure 2A). The basis for this constraint is currently unknown, but its strength provokes the hypothesis that the region is important for Myb function.

Differences in evolutionary rates between structurally equivalent EGF-like domains (EGF) in the Notch ligands Delta and Serrate also illustrate the distinction between structural and functional constraint. The function of both paralogs is to bind Notch and trigger signaling, but they elicit different responses even when they signal to equivalent cells (Panin et al., 1997). The basis for this difference is currently unknown, but the evolutionary rates of their extracellular domains (Figure 2B) suggest a specific hypothesis. In Delta, EGF2/3 evolve most slowly, followed by EGF5, whereas in Serrate, EGF15 evolves most slowly, followed by EGF2/3. A testable model based on these data is that the differences in biological activity between the two paralogs are primarily encoded in EGF5 (Delta) and EGF15 (Serrate). Thus, while homology information identifies all these structurally equivalent repeats as conserved, rate analyses uncover distinctions among them that are likely due to differences in function.

Regions with no similarity to known domains can be identified as important solely on the basis of their slow rate of evolution. For example, among the most strongly constrained regions within both Delta and Serrate are four that are N-terminal to the DSL domain (Figure 2B). They are not detectably homologous to anything else in the protein databases, but given the independent maintenance of the constraints in both paralogs, they probably fulfill an important function that is likely related to binding Notch. In summary, by making predictions relevant to function on the basis of sequence alone, quantification of constraints goes well beyond recording conservation and structural homology.

Genomic DNA

Constraints can also be quantified in multiple alignments of genomic DNA (Sumiyama et al., 2001). For example, the promoter of the cMet gene, which is part of the CFTR region that has been sequenced from several mammals by Eric Green's group, contains a previously unidentified element under strong constraint about 150 base pairs upstream of a known transcriptional start site (Figure 2C). High-resolution quantification of constraints in mammalian genomes, at the base pair level, will have a particularly important application: since the rate of evolution of a position is inversely proportional to the deleteriousness of past polymorphisms affecting it, the deleteriousness of a SNP in that position will be inferable from its past rate of evolution. Base pair-specific esti-

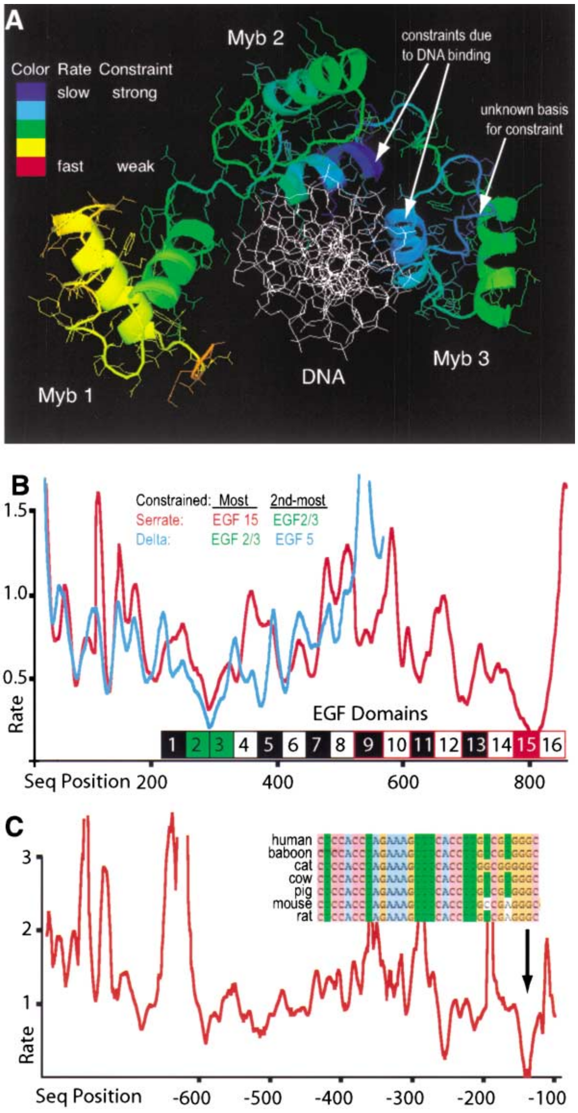

Figure 2. Examples of Quantification of Constraints

(A) Crystal structure of cMyb DNA binding domain bound to DNA (Tahirov et al., 2002) with local rates of evolution encoded in color. Greatest constraint is in dark blue.

(B) Local rates of evolution in the extracellular domain of the Notch ligands Delta and Serrate (Simon et al., 2002). Greatest constraints correspond to the troughs in the plot. Rates are normalized by division by the average rate and plotted as a function of their position in the alignments. Plots are truncated at the transmembrane domain. Boxes and numbers give approximate positions of EGF-like domains. Delta has 8 EGF repeats; Serrate has 16.

(C) Rates of evolution estimated in 12-base windows over an alignment of the cMet promoter.

mates of evolutionary rates could be obtained from alignments of several mammalian genomes.

Sampling Adequate Diversity

Mammalian phylogeny (Madsen et al., 2001; Murphy et al., 2001) suggests that ten or more additional placental genomes (that are not closely related to us or to each other) would be needed to obtain first estimates of evolutionary rates at the base pair level. For analyses of proteins, the useful phylogenetic range is from deeply divergent eukaryotes to invertebrates, depending on the origin of the protein and on its average evolutionary rate (Figure 1). In addition, a thorough sampling of vertebrate genomes will allow analyses of the proteins and promoters that arose at the origin of vertebrates, when our genomic and organismal complexity exploded and led to a doubling of the number of protein-coding genes. Wide phylogenetic diversity would also facilitate analyses of noncoding RNAs, which use different computational methodology (Eddy, 2002). None of the species need to be experimental models. Ironically, some experimental models may not be as generally useful for experimental analyses as a handful of well-placed, experimentally useless organisms that provide a good sampling of sequence diversity from which testable hypotheses about function can be derived.

Primates

Where does this leave primates, our closest relatives, whose genomes are so similar to ours that their amount of sequence divergence is minute compared to what a single bat, sloth, or aardvark would contribute (Figure 3A)? Rather than facilitating identification of constraints in functional elements that arose prior to the diversification of mammalian orders, primate sequences primarily allow identification of very recent changes in the human lineage. Of particular interest is the small number of changes that occurred since our last common ancestor with chimp, a small fraction of which must have effected the biological changes that set us apart from apes (Figure 3A).

Two recent landmark studies give a glimpse of these important differences. In the first (Enard et al., 2002a), gene expression in brains and livers from chimp, human, and macaque was quantified. The amount of difference between the species that was attributable to the human lineage was much greater in brain than in liver. This is consistent with a disproportionate amount of functionally important evolution affecting the brain in the human lineage. Will it be possible to identify the causative nucleotide changes and to determine which were positively selected because of an advantageous phenotype? The second study, which focused on the forkhead domain gene FOXP2 (Enard et al., 2002b), suggests that it may be, at least for very recent changes. Individuals with an amino acid change in a highly conserved residue in the forkhead domain suffer from a severe speech and language disorder, which had allowed identification of this gene by positional cloning (Lai et al., 2001).

Why Aren't We Apes?

On a hunch, Enard et al. (2002b) set out to see if there were any differences in FOXP2 between human and chimps. Surprisingly, two amino acid substitutions were found that were shown to have occurred on the human lineage. Given the tiny amount of neutral change since our last common ancestor with chimp (Figure 3A) and the paucity of amino acid substitutions during FOXP2 evolution in mammals, this was shown to be in excess over expectation. Thus, either human FOXP2 is not very constrained any more and accumulates missense substitutions at a higher rate, or these changes conferred an advantage and were positively selected for. If the mutations were fixed as a result of strong selection (a so-called selective sweep) that occurred recently, current variation in the human population in loci closely linked

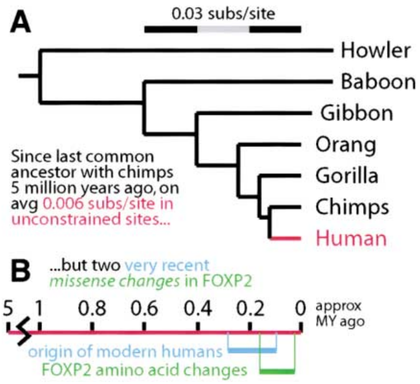
A
0.03 subs/site
Howler
Baboon
Gibbon
Orang
Gorilla
Chimps
Human
Since last common ancestor with chimps 5 million years ago, on avg 0.006 subs/site in unconstrained sites...
B
...but two very recent missense changes in FOXP2
5
1
0.8
0.6
0.4
0.2
0
approx MY ago
origin of modern humans
FOXP2 amino acid changes

Figure 3. The Most Recent Part of the Human Lineage in Context with Primate Phylogeny

(A) Representative primates and the approximate number of substitutions per site in the tree relating them.

(B) Approximate origin of modern humans in relationship to the estimated time when the two amino acid changes in FOXP2 were fixed in the human population.

to the selected locus should be diminished because a selective sweep also drives linked alleles to fixation (by so-called hitchhiking). The variation around the selected locus is then characterized by an excess of both rare and frequent alleles (Fay and Wu, 2000), exactly what was found around the FOXP2 locus. The diminished variation around FOXP2 is inconsistent with the amino acid changes having no consequence; instead, it is consistent with a selective sweep at the time modern humans evolved (Figure 3B).

The comparison with the chimp, therefore, was critical in two different ways. First, in conjunction with at least one outgroup sequence, it allowed “allocation” of the missense changes to the human lineage since our last common ancestor. Second, it allowed the determination of the ancestral state of human polymorphisms, and therefore the distinction between high- and low-frequency alleles. The latter is rather important for any kind of study of human variants in which knowing whether selection was involved may shed light on the epidemiology of disease. Whether the selected changes in FOXP2 contributed to the evolution of human language is an open question, but given its essential role in language development, it seems likely that the advantageous phenotype underlying the positive selection involved an improved ability to communicate by sound.

Conclusions

In-depth comparative analyses that are based on a large but realistic amount of sequence data will inform biomedicine in at least three important ways. First, functional research will be aided by the identification of functional elements and by the quantification of the strength of constraints within them. Second, at a finer level, the deleteriousness of any SNP will be predictable if a sufficient number of mammalian genomes are available. Finally, comparisons with our closest relatives will identify loci whose advantageous mutations were selected for and turned into fixed differences that set us apart from apes.

Selected Reading

Eddy, S. (2002). Cell *109*, 137–140.

Enard, W., Khaitovich, P., Klose, J., Zöllner, S., Heissig, F., Giavalisco, P., Nieselt-Struwe, K., Muchmore, E., Varki, A., Ravid, R., et al. (2002a). Science *296*, 340–343.

Enard, W., Przeworski, M., Fisher, S., Lai, C.S.L., Wiebe, V., Kitano, T., Monaco, A.P., and Pääbo, S. (2002b). Nature *418*, 869–872.

Fay, J.C., and Wu, C.-I. (2000). Genetics *155*, 1405–1413.

Göttgens, B., Gilbert, J.G.R., Barton, L.M., Grafham, D., Rogers, J., Bentley, D.R., and Green, A.R. (2001). Genome Res. *11*, 87–97.

Göttgens, B., Barton, L.M., Chapman, M.A., Sinclair, A.M., Knudsen, B., Grafham, D., Gilbert, J.G.R., Rogers, J., Bentley, D.R., and Green, A.R. (2002). Genome Res. *11*, 749–759.

Lai, C.S.L., Fisher, S.E., Hurst, J.A., Vargha-Khadem, F., and Monaco, A.P. (2001). Nature *413*, 519–523.

Li, W.-H. (1997). Molecular Evolution. (Sunderland, MA: Sinauer Associates).

Madsen, O., Scally, M., Douady, C.J., Kao, D.J., DeBry, R.W., Adkins, R., Amrine, H.N., Stanhope, M.J., de Jong, W.W., and Springer, M.S. (2001). Nature *409*, 614–618.

Murphy, W.J., Eizirik, E., Johnson, W.E., Zhang, Y.P., Ryder, O.A., and O’Brien, S.J. (2001). Nature *409*, 610–614.

Panin, V.M., Papayannopoulos, V., Wilson, R., and Irvine, K.D. (1997). Nature *387*, 908–912.

Pennacchio, L.A., Olivier, M., Hubacek, J.A., Cohen, J.C., Cox, D.R., Fruchart, J.-C., Krauss, R.M., and Rubin, E.M. (2001). Science *294*, 169–173.

Pupko, T., Bell, R.E., Mayrose, I., Glaser, F., and Ben-Tal, N. (2002). Bioinformatics *18*, S71–S77.

Simon, A.L., Stone, E.A., and Sidow, A. (2002). Proc. Natl. Acad. Sci. USA *99*, 2912–2917.

Sumiyama, K., Kim, C.-B., and Ruddle, F.H. (2001). Genomics *71*, 260–262.

Tahirov, T.H., Sato, K., Ichikawa-Iwata, E., Sasaki, M., Inoue-Bungo, T., Shiina, M., Kimura, K., Takata, S., Fujikawa, A., Morii, H., et al. (2002). Cell *108*, 57–70.