Published in final edited form as:

Curr Opin Chem Biol. 2001 February ; 5(1): 86–89.

Genome comparisons highlight similarity and diversity within the eukaryotic kingdoms

Catherine A Ball and J Michael Cherry*

Department of Genetics, Stanford University School of Medicine, Stanford, California 94305-5120, USA

Abstract

In 2000, the number of completely sequenced eukaryotic genomes increased to four. The addition of *Drosophila* and *Arabidopsis* into this cohort permits additional insights into the processes that have shaped evolution. Analysis and comparisons of both completed genomes and partially sequenced genomes have already shed light on mechanisms such as gene duplication and gene loss that have long been hypothesized to be major forces in speciation. Indeed, duplicate gene pairs in *Saccharomyces*, *Arabidopsis*, *Caenorhabditis* and *Drosophila* are high: 30%, 60%, 48% and 40%, respectively. Evidence of horizontal gene-transfer, thought to be a major evolutionary force in bacteria, has been found in *Arabidopsis*. The release of the ‘first draft’ of the human genome sequence in 2000 heralds a new stage of biological study. Understanding the as-yet-unannotated human genome will be largely based on conclusions, techniques and tools developed during the analysis and comparison of the genome of these four model organisms.

Introduction

Biologists long-ago embraced sequence comparison as a valuable method for studying genes, proteins and the evolution of species. The recent explosion of sequence data caused by worldwide genomic sequencing efforts promises to provide new insights into biology. Comparisons of completed genome sequences enable biologists to ask questions that were impossible with small sets of sequences. Large-scale sequence comparisons can identify shared biological processes and biochemical functions and uncover the genetic basis of variation and speciation. Although there are many advantages to having a completely sequenced genome for study, the sequence of a single genome is not as powerful alone as it is when compared with others.

Results discussed in this review generally take advantage of commonly used software such as Blast [1] and Clustalw [2]. In other cases, specialized software has been developed that might have practical applications for other researchers. One area in which novel resources have been developed is in the organization of structured definitions and controlled vocabularies for the annotation of biological processes and molecular functions [3•].

This review concentrates on genomic comparisons of eukaryotes, with emphasis on studies that identify potential orthologs (proteins that share a common ancestor), use information from one genome to understand another, and look into sources of variation. Although not as numerous as the genomes of completed prokaryotic organisms, sequenced eukaryotic genomes now number four (*Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila*

melanogaster and *Arabidopsis thaliana*), with a fifth (*Schizosaccharomyces pombe*) to be released imminently. The genomic sequence of *Homo sapiens*, although in its first draft, is not fully assembled and is yet to be annotated to the same extent as that of the model organism sequences mentioned above.

Orthology

Confident identification of potential orthologs relies heavily on a high-quality survey of the genome. Pairwise comparisons of completed genomes have helped identify the sets of genes that appear to be common within eukaryotes. Orthologous gene products are those that have evolved in vertical descent from a common ancestral gene product. Without a comprehensive survey of a genome, it is essentially impossible to differentiate orthologs from paralogs, gene products whose similarity to one another arose from duplication of an ancestral gene. When analyzed with respect to the biological processes in which they participate, orthologous proteins are most often involved in 'core biological processes' — for example, metabolism, protein folding, synthesis and transport [4,5•,6•]. Because these core processes are essential for every eukaryotic cell, their conservation is not surprising. The identification of orthologs provides insight into a protein's molecular function in different species. Putative orthology can be determined as proposed by Rubin *et al.* [5•]. They analyzed all protein sequences from *D. melanogaster*, *C. elegans* and *S. cerevisiae*. A core set of proteins, and their family members, were determined by their presence in all three organisms using Blast E-values, the expected probability of randomly finding a match with a similar quantity, at least as significant as 1×10^{-10} and requiring 80% of the query protein to be contained within the reported similarity. Rubin *et al.* found 1833 proteins from the unicellular *S. cerevisiae* in this core set. Interestingly, the number of gene products present in this group was remarkably similar for the multicellular organisms: 3033 *D. melanogaster* proteins and 3229 *C. elegans* proteins.

Comparisons of multiple genome sequences can reveal possible evolutionary relationships. For example, Tekai *et al.* [7•] constructed a phylogenetic tree using 20 fully sequenced genomes as well as the partial sets of the mouse, human and *S. pombe* genomes. By comparing predicted open reading frames, they were able to form a tree that places the Archaea in closer proximity to the Bacteria than to the Eukaryotes. Further, the rate of duplication or deletion events seemed to be similar in related organisms. Wang *et al.* [8] estimated that the last common ancestor of plants, animals and fungi existed 1.6 billion years ago. Phylogenetic analysis suggests that plants diverged before the separation of animals and fungi [9•].

Conservation of syntenous regions

Evidence for conservation of synteny, genome collinearity, gives researchers good evidence about possible evolutionary mechanisms and phylogenetic relationships. Observing a large number of gene structures and orders conserved between species suggests an evolutionary relationship. Identification of chromosomal regions of conserved synteny will probably aid the annotation of genes in different species and help in gene/exon prediction.

By examining the genetic or physical locations of murine orthologs of genes on human chromosomes 21 [10•] and 22 [11•], researchers have detected many regions that share genomic organization within sections of mouse chromosomes. Not only can examining the genomic structure give insight into phylogenetic relationships between species, but identifying regions of conserved synteny can also provide good targets for locating the orthologs of genes that currently lack counterparts in the other species. Batzoglou *et al.* [12•] report increased success in exon prediction when looking at syntenic regions of mouse and

human chromosomes. This type of technique will probably become invaluable for gene prediction in the human genome sequence.

Variation

Although the identification of similarities between different genomes is very interesting, benefit can also be gained from examining the variation between genomes that might result in speciation. Recent publications investigate the roles that gene families, gene loss and gene amplification have played in the speciation of eukaryotes.

Amplification of gene families

As mentioned above, the number of genes comprising core biological functions is similar between the sequenced multicellular eukaryotes. However, some genomes show vastly different numbers of predicted genes. Although the *A. thaliana* genome is similar in size to *C. elegans* and *D. melanogaster*, their gene numbers are quite different: ~26,000, ~19,000, and ~14,000, respectively.

There are large differences between *D. melanogaster*, *C. elegans* and *A. thaliana* in the number of gene products associated with multicellular development and cellular processes, with many examples of gene families that are amplified and specialized in one organism. *D. melanogaster* and *A. thaliana* both show phenomenal amplification of gene families within their genomes. More than 25,000 genes have been annotated in the *A. thaliana* genome sequence [6•]. Of these genes, there appear to be only 11,000 gene families. *A. thaliana* displays remarkable redundancy: only 35% of the genes are unique within the genome and 37% of the gene families have more than five members. In comparison, the *C. elegans* and *D. melanogaster* genomes encode fewer genes, but similar numbers of gene families. The authors suggest that the large number of tandemly repeated genes in *A. thaliana* gives evidence that unequal crossing over is more common in plants, or that a large genome size is simply more easily tolerated by plants. Duplication of genes and gene families give an organism the ability to evolve independently highly specialized versions of a gene product.

Riechmann *et al.* [9••] worked with complete genome sequences from the three eukaryotic kingdoms (animal, *D. melanogaster* and *C. elegans*; plant, *A. thaliana*; and fungi, *S. cerevisiae*) in search of transcription factors. Interestingly, a large fraction of the identified transcription factors were found to be representatives of lineage-specific gene families. That is, 45% of identified *A. thaliana* transcription factors, 47% from *C. elegans* and 32% from *S. cerevisiae* were in a gene family unique to that kingdom. As with so much in this field, an interesting puzzle is created: why are only 14% of *D. melanogaster* transcription factors unique to animals?

Gene loss

Multiple lines of evidence suggest that some species have lost genes and/or gene families that were retained by their close relatives. Comparisons of the *S. cerevisiae* genome to those of other fungi (*S. pombe* and *Neurospora crassa*) show an interesting role for gene loss in speciation. When compared with the other genomes, *S. cerevisiae* appears to have lost many gene families.

Aravind *et al.* [13•] used Blast comparisons against *S. pombe* sequences to show that the *S. cerevisiae* genome has co-eliminated functionally related groups of proteins, such as the signalosome and spliceosome components, as well as genes involved in nuclear maintenance and protein folding/processing. They further predict that *S. cerevisiae* lacks the post-transcriptional silencing system that includes an RNA-dependent RNA polymerase and

suggest that the loss of this system resulted in increased retrotransposition, which in turn caused loss of introns and permitted the loss of some spliceosome components.

A similar comparison between *N. crassa* and *S. cerevisiae* [14•] shows that *S. cerevisiae* has lost genes involved in ion homeostasis, the cytoskeleton, protein synthesis and some metabolic enzymes. At the time the Aravind *et al.* and Braun *et al.* studies were performed, the genomic sequences of *S. pombe* and *N. crassa* were incomplete, preventing detection of potential gene loss in those organisms.

Substantial proportions of *N. crassa* genes are ‘orphans’ — without clear homologs in any species. Braun *et al.* [14•] suggest that the ‘streamlining’ of the *S. cerevisiae* genome in relation to the more complex *N. crassa* is consistent with the proposal that unicellular fungi arose from multicellular ancestors. However, simple gene loss in *S. cerevisiae* cannot explain the increased number of total or even orphan genes that are present in the *N. crassa* genome.

Gene expression

Control of gene expression is a vehicle that can be used by evolution to provide variation. Although it is evident that coding sequences are often well conserved between species, the cross-species relationships of non-coding or regulatory regions is less well understood. Likewise, reliable methods of evaluating similarities between non-coding regions are still being developed. Jareborg *et al.* [15•] examined the non-coding regions of 77 mouse/human orthologs and found that they could detect similar sequences in 50% of 5' UTRs (untranslated regions) and 56% of 3' UTRs. The first 100 base pairs upstream of transcription start sites showed similarity in 70% of the orthologous pairs. Wasserman *et al.* [16•] compared the genomic sequences for 28 sets of co-regulated skeletal muscle genes, facilitating the identification of transcription-factor-binding sites in their non-coding regions. Dubchak *et al.* [17•] were able to identify conserved non-coding sequences when comparing syntenous regions of the human, mouse and dog genomes (interleukin-4 and interleukin-13).

We predict that a new branch of comparative genomics will be the analysis of gene expression patterns. Genes that have similar responses in different organisms could possibly be involved in similar biological processes, whereas genes with similar sequences, but dissimilar expression patterns are likely to have divergent roles. Establishing protocols and algorithms to identify relationships between regulatory regions will be one of the future challenges to genome analysis.

Human disease genes

Homologs to human disease genes in yeast, nematode, fly and *Arabidopsis* suggest that many of these disease genes are really involved in what we would classify as core biological processes — these processes would be expected to have significant effects on an organism if perturbed. That mutations in human disease genes cause disease in humans while often causing lethality in other organisms suggests that humans have evolved a considerable duplication of function.

When compared with a list of 289 human disease genes, the *A. thaliana* genome had 139 similar sequences. Rubin *et al.* [5••] and Fortini *et al.* [18•] found that *D. melanogaster* has 178 potential homologs to a list of 287 human disease genes. The classes of human disease genes most likely to have *D. melanogaster* homologs were those involved in cancers, malformation syndromes and metabolic, renal, and neurological diseases. Less represented

in the *D. melanogaster* genome were homologs of human genes involved in endocrine, hematological or immune system diseases.

Conclusions

The diversity observed in biology is recorded in the genomes of all organisms. Through analysis and study of these genomes, we are able to make hypotheses about the relationships between organisms, the shared mechanisms encoded by orthologous genes and the striking divergence of some pathways between kingdoms. Both computational and experimental studies are needed to help fill in the blanks. For example, only 9% of *A. thaliana* proteins have had an experimentally determined function assigned. Likewise, predicting human genes and associating them with biological information is an ongoing challenge. Lessons learned from previous genome studies and information from direct comparisons can greatly assist in gene prediction. We can use orthology to make putative identification of the molecular functions of uncharacterized gene products; however, we have much to do before biological processes can be assigned.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
 - of outstanding interest
1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410. [PubMed: 2231712]
 2. Higgins DG, Thompson JD, Gibson TJ. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol* 1996;266:383–402. [PubMed: 8743695]
 3. Gene Ontology Consortium. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;25:25–29. [PubMed: 10802651] A descriptive tool for annotation of gene products is explained in this paper. This tool is being used by the *Saccharomyces* Genome Database, Mouse Genome Informatics, Flybase, The *Arabidopsis* Information Resource, and Wormbase projects. The Gene Ontology (www.geneontology.org) was used by Rubin *et al.* [5••] for classification of fly proteins.
 4. Chervitz SA, Aravind L, Sherlock G, Ball CA, Koonin EV, Dwight SS, Harris MA, Dolinski K, Mohr S, Smith T, et al. Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science* 1998;282:2022–2028. [PubMed: 9851918]
 - 5••. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W, et al. Comparative genomics of the eukaryotes. *Science* 2000;287:2204–2215. [PubMed: 10731134] This comparison of fly and worm proteomes also provides a discussion of the similarity of the fly proteome to mammalian proteins.
 - 6•. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 2000;408:796–815. [PubMed: 11130711] In this publication, the first completed plant genomic sequence is described. *A. thaliana* has almost twice as many genes as *D. melanogaster*, contained within a similar genome size.
 - 7•. Tekaia F, Lazcano A, Dujon B. The genomic tree as revealed from whole proteome comparisons. *Genome Res* 1999;9:550–557. [PubMed: 10400922] In this analysis, 20 completed genomic sequences were compared to produce a phylogenetic tree based on similarity and duplications observed in their proteome.
 8. Wang DY, Kumar S, Hedges SB. Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proc R Soc Lond B Biol Sci* 1999;266:163–171.
 - 9••. Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR. *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science* 2000;290:2105–2110. [PubMed: 11118137] This first comparison of the *A.*

thaliana complete proteome with those of fly, worm and yeast provides an illustration of divergence of transcription regulators within the three eukaryotic kingdoms.

- 10•. Hattori M, Fujiyama A, Taylor TD, Watanabe H, Yada T, Park HS, Toyoda A, Ishii K, Totoki Y, Choi DK, et al. The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. *Nature* 2000;405:311–319. [PubMed: 10830953] In this report on the completion of human chromosome 21 genomic sequence, the authors discuss human/mouse regions of conserved synteny with include shared telomeric structures.
- 11•. Dunham I, Shimizu N, Roe BA, Chisoe S, Hunt AR, Collins JE, Bruskewich R, Beare DM, Clamp M, Smink LJ, et al. The DNA sequence of human chromosome 22. *Nature* 1999;402:489–495. [PubMed: 10591208] Regions of human/mouse conserved synteny are described in this report on the completion of human chromosome 22 genomic sequence.
- 12•. Batzoglou S, Pachter L, Mesirov JP, Berger B, Lander ES. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res* 2000;10:950–958. [PubMed: 10899144] In a study of gene prediction using conserved regions between human and mouse as an aid, the authors report that using cross-species similarity provides increased specificity of their exon prediction algorithm when compared with Genscan.
- 13•. Aravind L, Watanabe H, Lipman DJ, Koonin EV. Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc Natl Acad Sci USA* 2000;97:11319–11324. [PubMed: 11016957] In this comparison of the *S. cerevisiae* proteome with a subset of *S. pombe* sequences available in March 2000, the authors find evidence for coelimination and codivergence in significant cellular pathways.
- 14•. Braun EL, Halpern AL, Nelson MA, Natvig DO. Large-scale comparison of fungal sequence information: mechanisms of innovation in *Neurospora crassa* and gene loss in *Saccharomyces cerevisiae*. *Genome Res* 2000;10:416–430. [PubMed: 10779483] In this comparison of the *S. cerevisiae* complete proteome with the subset of *N. crassa* proteins (as of September 1999), the authors describe potential gene loss and divergence between *N. crassa* and *S. cerevisiae*.
- 15•. Jareborg N, Birney E, Durbin R. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res* 1999;9:815–824. [PubMed: 10508839] This study explores the conservation between non-coding regions of 77 human/mouse genomic pairs assembled from EMBL. It reports the 3' UTR to be most similar between human/mouse pairs (56% alignable regions). There were 50% and 26% alignable regions found in the 5' UTR and intron sets, respectively.
- 16•. Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE. Human-mouse genome comparisons to locate regulatory sites. *Nat Genet* 2000;26:225–228. [PubMed: 11017083] This paper describes the use of a Bayesian sampling algorithm with Gibbs sampling to detect potential regulatory binding-sites.
- 17•. Dubchak I, Brudno M, Loots GG, Pachter L, Mayor C, Rubin EM, Frazer KA. Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res* 2000;10:1304–1306. [PubMed: 10984448] In this comparison of regulatory sequences of orthologous 200 kb regions of human, mouse and dog genomes, 16 non-coding sequences were found to be similar to one another.
- 18•. Fortini ME, Skupski MP, Boguski MS, Hariharan IK. A survey of human disease gene counterparts in the *Drosophila* genome. *J Cell Biol* 2000;150:F23–F30. [PubMed: 10908582] *Drosophila* genes with significant homology to human disease genes are surveyed in this paper.

Abbreviation

UTR untranslated region