Primer

Computational genomics

Eugene V. Koonin

We now know how to read the sequences of nucleotide letters that comprise the genome at a rather frightening speed — a several-million-base bacterial genome in several days is not a problem for one of the sequencing centers, and a billion-base eukaryotic genome can be done in less than a year. But reading a text and understanding it are two different things. So how well can we understand the genome sequences? The answer to this question is central to the entire enterprise of genomics, and this is where computational analysis of genomes takes the driver's seat. Here I will try to briefly outline some major goals, problems, challenges and approaches of computational genomics. Such a young field is already quite diverse, and in this short article I will concentrate on several issues that seem to be critical for deciphering biology from genome sequences, rather than mathematical and computer-science aspects that are well covered in several excellent books.

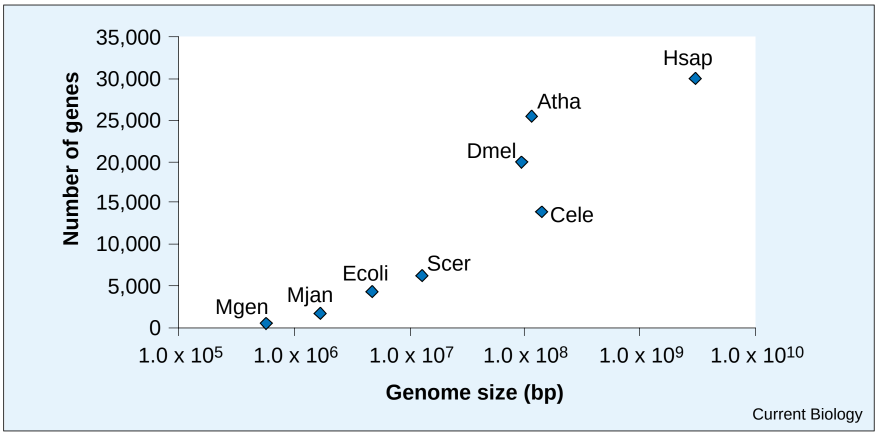
The importance of being comparative

Perhaps the most important achievement of the Human Genome Project is that it has spawned sequencing of other genomes from all walks of life, including multiple species of bacteria, archaea, animals, plants, and fungi. The connection of comparative genomics with fundamental evolutionary studies is obvious, for example, by analogy with the similar role of comparative anatomy in classical evolutionary biology. What was perhaps less clear in the early days of genomics is the crucial role of genome comparisons in interpreting the sequence of any particular genome. By now, however, we have come to realize that comparative analysis of two or more genome sequences is the surest, the easiest — once the sequences are available — and often the only path to reliable identification of genes and other important parts of the genome and prediction of their functions and interactions. This is due to a very simple, but powerful principle of the neutral theory of molecular evolution: when left to their own devices, that is, allowed to evolve neutrally, nucleotide sequences will, over millions of years, accumulate multiple changes and diverge beyond recognition. Specifically, in the eighty million years that separate humans and mice, neutrally evolving sequences, such as pseudogenes, have become completely saturated with mutations, and accordingly, have lost all information about the ancestral sequence. Whatever sequence conservation is detectable between human and mouse genome sequences, and this conservation is indeed extensive, indicates that the conserved sequences are functionally important or, in the parlance of evolutionary biology, are subject to stabilizing selection. The extent and nature of sequence conservation can provide us with vital information on the content of the genome such as the location of the genes, regulatory regions and functional sites in proteins.

First find the genes…but how?

The prevailing view of the genome is gene-centric, and although genes certainly do not tell us the whole story, there is merit to this approach. A striking aspect of comparative genomics is how relatively similar genomes of different types of organisms are in terms of the number and the repertoire of the encoded proteins and how dramatically different they are with respect to the organization of the genes themselves and the amount of extragenic sequence. Indeed, the human genome encodes only about seven times as many proteins as that of *Escherichia coli*, but the difference in the size of the genomes themselves is nearly a thousand-fold. Generally, the gene density in a genome seems to decrease consistently with the increase of the organism's complexity (Figure 1). Furthermore, eukaryotic genes are interrupted by multiple introns. However, whereas the average size of an exon is about the

Figure 1



A plot of the predicted number of protein-coding genes versus genome size. Mgen, *Mycoplasma genitalium*; Mjan, *Methnococcus jannaschii* (an archaeon); Ecoli, *Escherichia coli*; Scer, *Saccharomyces cerevisiae* (yeast); Cele, *Caenorhabditis elegans* (nematode); Atha, *Arabidopsis thaliana* (thale cress); Dmel, *Drosophila melanogaster* (fruit fly); Hsap, *Homo sapiens*.

same in all multicellular eukaryotes, around fifty codons, the characteristic length of an intron markedly increases with the organism's complexity. As if this were not enough, it appears now that alternative splicing occurs in the majority of the genes in complex eukaryotes such as humans, and for many genes, the number of distinct splice forms is huge.

All this considered, gene identification in prokaryotes and unicellular eukaryotes, on one hand, and in multicellular eukaryotes, on the other hand, are issues of completely different magnitude. The former problem should be considered effectively solved. In the early days of prokaryotic genomics, significant work has been done on statistical methods for gene recognition as exemplified by such widely used programs as GenMark and Glimmer. In brief, these methods 'learn' the distinction between coding and non-coding sequences in a given genome by comparing statistical properties, such as, for example hexanucleotide frequencies in GenMark, of experimentally identified genes and intergenic regions and applying the derived discrimination rules to new sequences.

However, with many genomes now at hand, the main role seems to belong to gene identification by homology. In most newly sequenced prokaryotic genomes, 80 to 90% of genes can be identified in this fashion. For many of these conserved genes, even the start position can be determined from alignments with homologous sequences, and this is all that is required because prokaryotic genes and many of the genes in unicellular eukaryotes are uninterrupted open reading frames. The rest of the genes that are located between the conserved ones still need to be predicted by statistical methods, but, given the limited search space, this is unlikely to produce many errors. On the whole, it appears that, for most of these genomes, a reasonably accurate gene complement has been identified.

The situation differs drastically for complex eukaryotic genomes. Powerful statistical methods such as GenScan and Genie and schemes that combine statistical analysis with homology information, such as PROCRUSTES and the latest incarnation of GenScan, have been developed. However, inspection of the gene annotations in the genomes of the nematode *Caenorhabditis elegans*, the fruit fly *Drosophila melanogaster*, the plant *Arabidopsis thaliana*, and most dramatically, the draft human genome suggests that, in general, the gene identification problem is not solved for genomes at this level of complexity. The small size of exons compared to introns and intergenic regions makes it difficult to identify exons either by homology or by the statistical properties of sequences. Many independent assessments tend to converge on the estimate of 20 to 30% of the predicted genes in the nematode containing major problems such as missing large exons or fusion of different genes, and it is likely that many others have relatively minor problems such as missing small exons. The state of gene prediction in the human genome, with its characteristic long introns and intergenic regions, is even poorer; probably a majority of genes are predicted to some extent incorrectly, except for those genes for which a complete mRNA sequence is available.

The extent of the problem seems to be such that no conceivable improvement of statistical approaches alone is likely to solve it. Some outside support is required, and it may come from two sources, namely mRNA sequences and homology information. In all likelihood, to derive a complete and accurate set of genes for a genome as complex as the human one, synergy between both types of information will be required. Both these sources are currently used in conjunction with statistical methods and have been partially incorporated in automatic procedures, but the amount of data is clearly insufficient. One might think that a complete library of mRNA sequences for the given organism would suffice to know the structure of all genes and thus would solve the problem of gene identification once and for all. However, this is unlikely to be the case because the very definition of a complete mRNA set has become murky with the realization that the majority of genes in complex eukaryotes produce alternatively spliced mRNAs, and sometimes many of these. Besides, many mRNAs are present at low levels or only during a short window in development.

It seems that multiple genome sequences from relatively close species, along with large collections of mRNA sequences, are a must for accurately delineating the gene sets of complex eukaryotes. For the human genome, the 'supporting' genome will be that of the mouse whose completion is expected in the near future. Yet it is likely that a third genome will be required — preferably that of a primate — to overcome the problems associated with the existence of unique genes, and the inevitable inaccuracies in alignments of long genomic segments.

Prediction of protein functions – domain detection and beyond

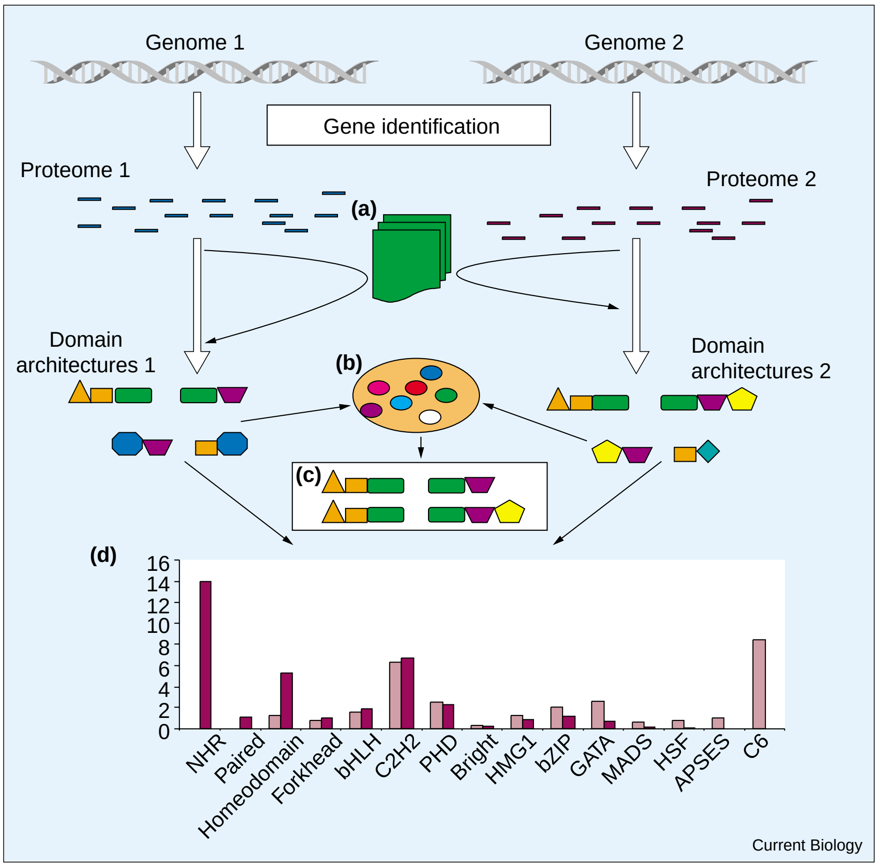
Once the protein set — the predicted proteome — for a given genome has been defined, the next major task is to analyze and interpret it. Even more than gene identification, this is all about sequence comparison. The current standard for database search, BLAST, combines high speed with robust statistics, but has limited sensitivity. However, the resolution of sequence analysis can be significantly increased by taking advantage of the information contained in sequences of multiple homologs. This is achieved in the

Position-Specific Iterating BLAST method (PSI-BLAST), an enhancement of BLAST that constructs a multiple alignment on the fly and employs it to iterate the database search, and in several methods based on the Hidden Markov Model (HMM) formalism.

These methods reveal their full potential when used in conjunction with libraries of protein sequence profiles, which may be considered to be models of protein domains. Carefully constructed profiles ensure rapid and sensitive detection of new instances of known domains and are powerful tools for genome annotation. Collections of domain profiles are employed in online systems for domain detection such as Pfam, SMART (both using HMM based methods) and CD-search which employs a modification of PSI-BLAST. These systems, particularly SMART, are designed to recognize not only sequence similarity between proteins, but also to explicitly represent the protein domain architecture, that is, the linear order of domains in the polypeptide chain. The use of these methods helps to avoid the most common pitfall of genome annotation, incorrect functional assignment on the basis of sequence similarity that involves only one domain, rather than the entire protein.

Combined with the rapid progress in protein three-dimensional structure determination, the sensitive methods for domain recognition effectively address one of the principal goals of structural genomics, assignment of structure to the maximal number of proteins encoded in each genome. This analysis allows researchers to concentrate on those structurally uncharacterized proteins that hold most promise to reveal both structural novelty, such as a new fold, and functional novelty, such as the structural basis for a critical biochemical activity. Clearly, such priority targets are those proteins that, firstly, show no detectable similarity to known structures (in addition to profile-based sequence analysis methods, this can be assessed using sequence–structure threading, which occasionally shows even greater sensitivity), and secondly, are highly conserved in evolution, which is suggestive of an important function.

Figure 2



A simplified scheme for comparative analysis of predicted proteomes. (a) Library of domain-specific profiles used for protein domain recognition; (b) database of orthologous protein families; (c) comparison of domain architectures of orthologous proteins from the analyzed genomes revealing domain accretion and rearrangements; (d) comparison of domain counts in the analyzed genomes revealing lineage-specific expansions.

Even the presently available methods for sequence and structure analysis show considerable power in detecting distant relationships between proteins. The algorithms themselves could be further improved, but perhaps the most rewarding train of development in the nearest future will involve further growth and refinement of domain databases, possibly resulting in a structural–functional classification of nearly all proteins encoded in each genome. This lofty goal may be within our reach because the estimated total number of protein folds is relatively small — about a thousand — and even the number of families with readily detectable sequence conservation is unlikely to exceed ten thousand.

Functional genomics and the evolutionary classification of genes

One of the immediate goals of comparative genomics is understanding the evolutionary trajectories of genes and integrating them into plausible evolutionary scenarios for entire genomes.

A prerequisite for this process is a phylogenetic classification of genes. A natural unit of such a classification is a cluster of orthologs — genes related by vertical descent; these clusters may also include some paralogs — genes related by intragenomic duplication. Since orthologs from different species as a rule perform the same function, a database of orthologous families, once carefully explored and annotated, becomes a powerful engine for predicting functions of genes from newly sequenced genomes. Targets for functional genomics may be defined using the system of orthologous families. The high-priority targets are those families of orthologs that are widely conserved, but whose functions cannot be predicted on the basis of sequence and structure analysis. The number of such protein families is surprisingly small, which makes their experimental characterization all the more enticing.

Recent work on the evolutionary classification of the proteins encoded in complete prokaryotic genomes shows that a significant majority of them belong to clusters of orthologs shared by genomes from distant lineages. However, most of the families are represented only in a minority of the genomes. This patchy composition of orthologous families is explained primarily by extensive, lineage-specific gene loss and horizontal gene transfer, two major evolutionary trends — at least in prokaryotes — whose scale has become apparent only through comparison of multiple genomes. On a genome scale, this unexpected complexity of the evolutionary process makes precise evolutionary reconstruction an extremely challenging task, and an algorithmic solution seems to be far beyond reach of modern computational approaches. Arguably, coping with this complexity is one of the primary goals of comparative and evolutionary genomics for the next several decades.

From genome comparison to biological adaptation

Several studies have led to the development of a basic scheme to identify quickly the significant differences between complete genomes, which might provide clues to unlock their unique adaptation strategies (Figure 2). Essentially, this strategy involves a census of protein domains and domain architectures in each of the genomes and comparison of the results. Comparisons of the sequenced eukaryotic genomes — yeast, nematode, fruit fly and human — reveal a moderate, but definite increase in the complexity of the protein repertoire parallel to the growth in the complexity of biological organization. This increased complexity is manifest both in genome-specific expansion of domain and protein families and in a trend that has been dubbed ‘domain accretion’ whereby proteins in orthologous groups tend to incorporate additional domains in more complex organisms. These observations might point to many additional interactions that could be important for unique mechanisms of signal transduction, regulation and development.

Prospects – beyond the genes

An enormous amount of work remains to be done by both computational and experimental genome biologists using the basic strategies outlined above. However, an entire uncharted field awaiting a concerted effort from comparative genomicists is the study of intergenic regions which occupy, in humans, approximately 97% of the genome. A recent comparison of the intergenic regions of two nematodes, *Caenorhabditis elegans* and *C. briggsae*, has shown that up to 20% of such sequences are conserved in evolution and hence may have an important biological function. If, as expected, this result is confirmed by comparisons of other eukaryotic genomes, this will mean that a large majority of functional DNA in the human genome is not in the genes at all. At this time, we do not have a clear idea what kind of computational methods — beyond those for alignment of long nucleotide sequences — are required for us to even start exploring possible functions of the intergenic sequences. We may be entering the post-genomic era, but it seems that there will be no shortage of genuinely important work for computational genomicists in any foreseeable future.

References

Baxevanis A, Ouelette BFF: *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins.* 2nd edn. New York: John Wiley & Sons, 2001.

Pevzner PA: *Computational Molecular Biology: An Algorithmic Approach.* Cambridge, Massachusetts: MIT Press, 2000.

Tatusov RL, Koonin EV, Lipman DJ: A genomic perspective on protein families. *Science* 1997, 278:631-637.

Frishman D, Mironov A, Mewes HW, Gelfand M: Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res* 1998, 26:2941-2947.

Bork P, Koonin EV: Predicting functions from protein sequences – where are the bottlenecks? *Nat Genet* 1998, 18:313-318.

Chervitz SA, Aravind L, Sherlock G, Ball CA, Koonin EV, Dwight SS, Harris MA, Dolinski K, Mohr S, Smith T, Weng S, *et al.*: Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science* 1998, 282:2022-2028.

Shabalina SA, Kondrashov AS: Pattern of selective constraint in ***C. elegans*** and ***C. briggsae*** genomes. *Genet Res* 1999, 74:23-30.

Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W, Cherry JM, *et al.*: Comparative genomics of the eukaryotes. *Science* 2000, 287:2204-2215

International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. *Nature* 2001, 409:860-921.

Venter JC *et al.*: The sequence of the human genome. *Science* 2001, 291:1304-1351.

http://www.expasy.ch/

http://www.ncbi.nlm.nih.gov/

http://smart.embl-heidelberg.de

Address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA.
E-mail koonin@ncbi.nlm.nih.gov