Getting closer to efficient gene discovery, in silico

Lorenzo Segovia

Driven in part by the enormous amount of raw data being generated by the growing number of genome sequencing projects—and the commensurate interest in the potential of such data to be eventually translated into marketable pharmaceuticals—applying computer methods for gene identification and function assignment, or in silico biology as it is now known, has quickly become one of the most active areas in biotechnology. It is also one of the most exciting as demonstrated by the lively discussions and wealth of new information presented at a recent meeting*, organized by Mark Borodovsky (Georgia Institute of Technology, Atlanta) and Eugene Koonin of the National Center for Biotechnology Information/National Institutes of Health (NCBI/NIH; Bethesda, MD), to review the state-of-the-art and examine the most recent advances in this fast-moving field.

Since prokaryotes genes are encoded as single open reading frames (ORFs), their identification should, in principle, be straightforward. The problem lies in distinguishing coding from noncoding ORFs. Programs such as MagPie or algorithms that use rules based on codon usage or dinucleotide frequencies in ORFs, work well for "typical" genes. But, as pointed out by Jean Michel Claverie (SGI-CNRS Marseille, France), they perform considerably less satisfactorily when it comes to "atypical" genes, probably products of lateral transfer, which are quite common and have a tendency to be lost when the general rules are applied.

In eukaryotes, the problem of gene identification is compounded by the fact that the coding sequences may be distributed in as many as 40 or more exons over as much as 100 kb of DNA. Victor Solovyev (European Bioinformatics Institute, Cambridge, UK), Steven Salzberg (Johns Hopkins University, Baltimore, MD), and Mark Borodovsky—the developers of three of the more widely used programs (FGENES, MORGAN, and GeneMark, respectively)—reviewed the performance of their software in the context of a group of related computational approaches, such as GENSCAN (currently recognized as the most accurate), GeneID, GenParser, Genie, GRAIL, and PROCRUSTES.

All, in their latest incarnations, have improved intron/exon prediction that can reach 60% accuracy. Greater accuracy, they suggested, will come with better knowledge of the splicing mechanisms and their guiding signals. This observation—the necessity of joining theoretical and experimental approaches—was one of the important and recurring themes of the meeting.

Figure 1. The Georgia Institute of Technology, where the conference took place last November.

James Fickett (SmithKline Beecham,King of Prussia, PA) and Gary Stormo (University of Colorado, Boulder) analyzed another actively explored approach to gene discernment. Known as pattern recognition, the technique makes use of functional DNA motifs such as the promoters, operator regions, and other binding sites that characterize transcriptional regions. Such methods are, of course, dependent on having bona fide binding or promoter sequences. This remains a serious limitation because these sites tend to be poorly characterized experimentally. Currently, only 15–53% of promoters can be accurately predicted from the available data. The situation with regard to other transcription-related binding sites is also is need of significant improvement. Often, binding sites defined in silico, from data obtained in vitro, fail to function in vivo. One possible way to avoid this kind of caveat is to consider physical properties of DNA associated with transcription. For example, Soren Brunak (Technical University of Denmark, Lyngby) has shown that the transcription origins in a collection of well defined eukaryotic genes have a particular "bendability" profile, thought to be involved in nucleosome phasing, that could be used to identify the beginning of genes.

Even the most accurate gene identification program needs to be coupled to accurate information about the function of its predicted proteins, and reciprocally, knowledge of protein functional determinants can be used to improve gene discovery programs. Experimental approaches, using knockout collections or high throughput mRNA characterization in different tissues under different conditions, are now being joined by increasingly powerful computational methods for protein analysis. Steven Altschul (NCBI/NIH; Bethesda, MD), presented the latest versions of his BLAST family of search programs, which are widely used to find homologs in databases. The extent of homology can often be a good indicator of function, and where many close homologs are found function can be assigned quite confidently.

When the levels of homology fall, several relatively new approaches can be followed, as described by Philip Bucher (Swiss Institute for Experimental Cancer Research, Epalinges sur Lausanne, Switzerland). He showed that generalized profiles from complete sequences (based, e.g., on hidden Markov models; HMMs) are often effective tools in identifying and characterizing highly divergent proteins. HMMs have been used to generate comprehensive profile collections, such as PFAM, with which new sequences can be functionally annotated. An alternative, offered by Michael Gribskov (San Diego Supercomputer Center, CA) and Steven Henikoff (Fred Hutchinson Cancer Research Center, Seattle, WA), is to generate particular motifs based on short homology units that characterize an entire protein family. These motifs can be used to build a kind of gapped profile, which in turn can be used to identify distantly related sequence families or sequences, and thus potential functions.

However, even function assignment is not the ultimate goal. Incorporating these functions into an integrated whole requires the ability to perform accurate metabolic reconstructions. Missing functions derived from these reconstructions can then be pinpointed and investigated more intensively, for example, by looking for nonorthologous displacement (nonhomologous proteins with identical functions). An interesting derivative of these analyses, discussed by Eugene Koonin, has been the generation of databases of clusters of orthologous or homologous groups. These can be scanned with any sequences of interest and their relation to both close and remote homologs easily dissected.

Propelled at the moment by the excitement of learning how cells work and interact at levels of detail unimaginable even a few years ago, and the more worldly impetus provided by the huge economic investment in "functional genomics," in silico gene discovery is also an important new meeting ground for mathematicians, computer scientists, physicists, as well as theoretical and experimental biologists. Such ground cannot fail to be extremely fertile.

Lorenzo Segovia is an investigator at the Institute of Biotechnology, UNAM, Cuernavaca, Mexico (lorenzo@ibt.unam.mx).

*Gene Discovery in Silico, Georgia Institute of Technology, Atlanta, GA, November 6–9, 1997.