

From Bioinformatics to Computational Biology

Jean-Michel Claverie

Structural and Genetic Information Laboratory, CNRS-AVENTIS UMR 1889, Marseille cedex 20, France

It is quite ironic that the uncertainty about the number of human genes (28,000–120,000) (Ewing and Green 2000; Liang et al. 2000; Roest Crollius et al. 2000) appears to increase as the determination of the human genome sequence is nearing completion. I shall contend here that this paradox reveals deep epistemological problems, and that “bioinformatics”—a term coined in 1990 to define the use of computers in sequence analysis—is no longer developing in directions relevant to biology.

After the pioneers who established the basic concepts of molecular sequence analysis (Fitch and Margoliash 1967; Needleman and Wunsch 1970; Chou and Fasman 1974), most computational biologists of my generation (the second one) embarked on their journey into the emerging discipline with the ambition to turn it into the bona fide theoretical branch of molecular biology. Having a physicist’s background, I suspect that many of us had the vision of establishing bioinformatics in a leadership role over experimental biology, similar to the supremacy that theoretical physics enjoys over experimental physics. Somewhere along the line, it seems that bioinformatics lost this ambition and became sidetracked onto what physicists would call a “phenomenological” pathway.

Let us follow the example of particle physics for a little longer. There, theoretical research has two phases (which, in fact, run in parallel). In the first phase (so-called phenomenological), a large number of physical events are recorded in huge raw databases, classified into separate groups based on statistical regularities, and then utilized to identify the most recurrent objects. Optimal database design, fast classification/clustering algorithms, and data mining software are the main area of development here. The level of knowledge gained from this phase is, for instance, that objects A and B often appear together except when C is around, or when parameter X is lower than a certain threshold; it is mostly statistical in nature. The parallel with the current state of bioinformatics is clear.

However, theoretical physics also has a subsequent, totally different phase, aiming at discovering the basic (few) rules (e.g., $E = mc^2$) underlying the relationships between the objects, their individual properties, and thus finally explaining the statistical distri-

butions of the events recorded in the databases. Once known, these rules considerably simplify the description of the database content and, more important, have a predictive power: the realm of the theory may encompass objects or events that have not been observed previously. This part of theoretical endeavor is entirely missing in current bioinformatics. As a consequence, we are still not able to agree on the number of human genes despite having the complete sequence of the human genome at hand.

Identifying precisely the 5’ and 3’ boundaries of genes (the transcription unit) in metazoan genomes, as well as the correct sequences of the resulting mRNA (“exon parsing”) has been a major challenge of bioinformatics for years. Yet, the current program performances are still totally insufficient for a reliable automated annotation (Claverie 1997; Ashburner 2000). It is interesting to recapitulate quickly the research in this area to illustrate the essential limitation plaguing modern bioinformatics. Encoding a protein imposes a variety of constraints on nucleotide sequences, which do not apply to noncoding regions of the genome. These constraints induce statistical biases of various kinds, the most discriminant of which was soon recognized to be the distribution of six nucleotide-long “words” or hexamers (Claverie and Bougueleret 1986; Fickett and Tung 1992). Initial gene parsing methods were then simply based on word frequency computation, eventually combined with the detection of splicing consensus motifs. The next generation of software implemented the same basic principles into a simulated neural network architecture (Uberbacher and Mural 1991). Finally, the last generation of software, based on hidden Markov models, added an additional refinement by computing the likelihood of the predicted gene architectures (e.g., favoring human genes with an average of seven coding exons, each 150 nucleotides long) is added (Kulp et al. 1996; Burge and Karlin, 1997)). These *ab initio* methods are used in conjunction with a search for sequence similarity with previously characterized genes or expressed sequence tags (EST). Sadly, it is often claimed that matching back cDNA to genomic sequences is the best gene identification protocol; hence, admitting that the best way to find genes is to look them up in a previously established catalog!

Thus, the two main principles behind state-of-the-art gene prediction software are (1) common statistical regularities and (2) plain sequence similarity. From an

E-MAIL Jean.Michel.Claverie@igs.cnrs-mrs.fr; **FAX** +33-4-91-16-45-49.

Article and publication are at www.genesdev.org/cgi/doi/10.1101/gad.155500.

epistemological point of view, those concepts are quite primitive. For instance, the concept of analyzing the frequency of groups of letters was actually introduced by Arab scholars around A.D. 700 to break substitution ciphering. Thus, the legendary cryptanalyst al-Kindi (Singh 1999) could still grab the essence of modern bioinformatics without much difficulty. Moreover, the above concepts are intrinsically conservative and introduce a bias in favor of the detection of genes similar to those already known. But the most fundamental limitation of the current approaches is that they bear absolutely no relationship to the actual molecular mechanisms of gene expression; when a human cell triggers the transcription of a given region of its genome, it is not because an homologous region exists in yeast, or because the transcripts (once translated) will lead to a meaningful (three-dimensional folding) amino acid sequence. Thus, current approaches are not on the pathway of a theoretical understanding of the genome, and have no predictive power beyond the realm of immediate analogy. This limitation is well illustrated by the complete failure of current programs in locating nonprotein coding genes (such as *Xist* and *H19*), which might have essential regulatory roles. The number of nonprotein coding genes is unknown, and they might constitute a significant fraction of yet anonymous EST clusters. The same fundamental problem is also attested by the near-zero performance of current methods to locate core promoter regions, as well as all other regulatory segments (Fickett and Hatzigeorgiou 1997; Stormo 2000). Yet, textbooks persist in misusing the words “signal” or “motif” to qualify genomic subsequence such as TATAAA or CCAAT, the occurrence of which are only exceptionally related to transcription. Students are left to discover the hard way that the promoter theory they were taught is not at all predictive. Finally, the presence of repeats or low complexity regions, induce intractable situations for all statistical approaches. The mere masking of the problematic regions is not a satisfactory solution as they do occur in bona fide, often important, genes.

The phenomenological approach to bioinformatics is fundamentally limited by the fallacious analogy that the human genome is a text to be deciphered. This vision, very popular with the media, is also pervading the scientific policy in the field. One often hears that bioinformatics must become “multidisciplinary,” must attract more computer scientists and mathematicians, etc., in the hope that fancier computational techniques will crack the code. However, computer cryptanalysis techniques only work at the level of symbols; the final understanding of the meaning of a message remains the privilege of its intended recipient—a human brain. For example, a simple frequency analysis will recognize a simple Cesar shift ciphering in the following message: *ZfmmpxEvdlxjmmnffuUbsabobu-*

2241qnbuNjbnjCfbdi, leading to its decoding to: *YellowDuckwillmeetTarzanat1130pmatMiamiBeach*. However, this is not truly useful if we do not know the meaning of the predefined code words: *YellowDuck*, *Tarzan*, and *MiamiBeach* (e.g., a given boat, a certain admiral, and a precise geographical location).

Similarly, for whatever DNA sequence we decipher, we have to determine its meaning from the cell’s point of view. Thus, I believe that the days of abstract DNA “numerology” are over, and theoretical biologists with a strong interest in the intricacies of the cell machinery should now reinvest the field. We now need to make educated guesses on the meaning of the code words, for example, by concentrating on the way they might interact with each other. The statistical features recognized at the level of DNA sequence have now to be related to chromatin structures, kinetic properties, and physicochemical principles of macromolecular interactions. The huge amount of information acquired recently (Szentirmay and Sawadogoo 2000) on the spatial organization of large molecular edifices such as transcriptional complexes and the spliceosome have now to be incorporated into a radically new type of bioinformatic approach. Similarly, the careful classification (requiring a detailed biological knowledge) of genes in well-defined subsets is certainly the key in identifying the multiplicity of specific regulatory motifs (Wasserman and Fickett 1998; Fickett and Wasserman 2000) resulting in biologically significant promoter regions, while the quest of a generic definition has remained elusive (Fickett and Hatzigeorgiou 1997). Clearly, taking advantage of such complex and specialized knowledge in the design of new gene prediction methods requires different skills than refining the now routine statistical “deciphering” approaches.

For those wishing to remain at a more abstract level, designing new algorithms at least logically consistent with known biochemical and cellular processes is also a worthwhile direction of research. For example, one could try to develop in silico promoter detection algorithms mimicking the formation of the pretranscriptional complex in response to the proper sequential recognition of individual sequence motifs. For this, we have to implement a multistate detector depending on the type and history of the sequence elements passing through it (e.g., having seen AAT may increase the affinity for a subsequent CGC, but only at a certain distance range). Standard approaches used until now (such as finite state machines, neural networks, or Markov models) cannot implement all the properties of such a contextual recognition process. This simple example illustrates why qualitative progress should now be given priority over the incremental improvement of current methods. Trying to achieve a reasonably accurate detection of human genes without reference to coding potential, sequence similarity, or any

property of the gene product, is certainly a good benchmark problem—both of tremendous practical and fundamental interest—on which to focus the development of new approaches.

Instead of computer scientists and mathematicians who look at the DNA as if it was the tape of a Turing machine, we now need a generation of computational biologists with a solid background in such fields as transcription, development, enzymology, microbiology, structural biology, etc. This will help bioinformatics to become a truly successful branch of biology, in pursuit of a satisfactory understanding of the function and evolution of genomic sequences in their cellular context.

ACKNOWLEDGMENTS

I thank Drs. C. Abergel, H. Ogata, and W. Fitch for their helpful comments.

REFERENCES

- Ashburner, M. 2000. A biologist's view of the Drosophila genome annotation assessment project. *Genome Res.* **10**: 391–393.
- Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**: 78–94.
- Claverie, J.-M. 1997. Computational methods for the identification of genes in vertebrate genomic sequences. *Human Mol. Genet.* **6**: 1735–1744.
- Claverie, J.M. and Bougueleret, L. 1986. Heuristic informational analysis of sequences. *Nucleic Acids Res.* **14**: 179–196.
- Chou, P.Y. and Fasman, G.D. 1974. Prediction of protein conformation. *Biochemistry* **13**: 222–245.
- Ewing, B. and Green, P. 2000. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet.* **25**: 232–234.
- Fickett, J.W. and Hatzigeorgiou, A.G. 1997. Eukaryotic promoter recognition. *Genome Res.* **7**: 861–878.
- Fickett, J.W. and Tung, C.S. 1992. Assessment of protein coding measures. *Nucleic Acids Res.* **20**: 6441–6450.
- Fickett, J.W. and Wasserman, W.W. 2000. Discovery and modeling of transcriptional regulatory regions. *Curr. Opin. Biotechnol.* **11**: 19–24.
- Fitch, W.M. and Margoliash, E. 1967. Construction of phylogenetic trees. *Science* **155**: 279–284.
- Kulp, D., Haussler, D., Reese, M.G., and Eeckman, F.H. 1996. A generalized hidden Markov model for the recognition of human genes in DNA. *Ismb* **4**: 134–142.
- Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S.L., and Quackenbush J. 2000. Gene index analysis of the human genome estimates approximately 120, 000 genes. *Nat. Genet.* **25**: 239–240.
- Needleman, S.B. and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**: 443–53.
- Roest Crollius, H., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quéher, F. et al. 2000. Estimate of human gene number provided by genome-wide analysis using tetraodon nigroviridis DNA sequence. *Nat. Genet.* **25**: 235–238.
- Singh, S. 1999. *The Code Book*. Doubleday, New York, NY.
- Stormo, G.D. 2000. Gene-finding approaches for eukaryotes. *Genome Res.* **10**: 394–397.
- Szertirmay, M.N. and Sawadogoo, M. Spatial organization of RNA polymerase II transcription in the nucleus. *Nucleic Acids Res.* **28**: 2019–2025.
- Uberbacher, E.C. and Mural, R.J. 1991. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci.* **88**: 11261–11265.
- Wasserman, W.W. and Fickett, J.W. 1998. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* **278**: 167–181.