



Public sphere and the sustainability of the bioinformatics promise

Marcelo Leite

Ciências Sociais, Instituto de Filosofia e Ciências Humanas, Unicamp,
Campinas, SP, Brasil
Corresponding author: M. Leite
E-mail: mleite@post.harvard.edu

Genet. Mol. Res. 3 (4): 575-581 (2004)
Received October 4, 2004
Accepted December 14, 2004
Published December 30, 2004

ABSTRACT. The literature about genomics and bioinformatics achievements in high-impact journals such as *Nature* and *Science* has raised disproportionate expectations amongst the general public about fast and revolutionary drugs and breakthroughs in biomedicine. However, the yield obtained by database mining activities has been modest, as reported in the February 2001 issues of these journals featuring the completion of human genome draft sequences by the Human Genome Project Consortium and the company Celera. I have compared changes in rethoric employed by molecular biologists in 2001 and in April 2003, when the final sequence was announced. The comparison suggests that researchers are concerned about the sustainability of society's investment in this field, though not explicitly.

Key words: Genetic determinism, Human genome, Bioinformatics, Public understand of science, Science and technology studies, Genomics

At face value, the human genome sequence is not much more than an apparently endless string of chemical letters, or a book written in a foreign and undeciphered language. However, based on accumulated knowledge about certain particularities in sequences containing genes, bioinformaticians have identified many candidate genes and even probable functions for quite a few of them. The present study of computer-aided, *in silico* genome analysis in search genes of interest is sometimes referred to as *mining*. One of the main hopes among bioinformaticians and molecular biologists has always been to turn knowledge of the complete DNA sequence of the species into great discoveries, which would be bound to revolutionize both biology and medicine. This is the way that this quest has been presented to the lay public.

A great number of articles about the genome appeared in the groundbreaking issues of the journals *Nature* and *Science* during the second week of February 2001, and many among them focused on preliminary mining results obtained from the draft sequences then available, results which were to a certain extent disappointing. Nine articles published in *Nature* that week were summarized in a 10th one by Birney et al. (2001), who rated them paradoxically as both frustrating and fruitful:

- The group dedicated to mining genes associated with the traffic of substances through the cell membrane had been able to find only a few;
- Those in charge of kinase-dependent cyclins, an important class of cell signals, came out empty-handed, finding no genes specifying a cyclin that had not already been described in the literature;
- Last, but not least, no new genes related to cancer had been uncovered.

Nevertheless, at the end of the article the authors reaffirm their optimism: "... there are many undiscovered treasures in the current data set waiting to be found by intuition, hard work and experimental verification. Good luck, and happy hunting!" (Birney et al., 2001)

One of the reasons behind latent public frustration with genomics and bioinformatics lies undoubtedly in the widespread identification between *biological function* and *specification of protein(s)*, which is at the root of the very notion of a *genetic code*. Even today, it is not unusual to find abbreviated definitions of genes as the piece of DNA carrying the *code* for a protein, which does not make much sense anymore, since it has been known for decades that genomic DNA can also specify RNA sequences never to be translated into proteins. Moreover, one of the things made manifest by genome sequencing is that vertebrate complexity seems to stem from an elaborate system of genome *regulation*, rather than genes themselves. A comparison of human draft sequences with those of other species makes it clear that they differ from the more "primitive" ones, for instance by the presence of much longer introns, which leads to the conclusion that introns indeed have a function, though a poorly understood one.

Even so, many a geneticist remains attached, when addressing the public, to the traditional and simplistic definition *gene equals protein equals function*, to the point of risking recycling the doomed metaphor of *junk-DNA* with the superimposition of a slightly less derogative metaphor taken from good old cybernetics: "Nearly all of the increase in gene size in human compared with the fly or worm is due to the introns becoming much longer (about 50 kb versus 5 kb). The protein-encoding exons, on the other hand, are roughly the same size. This *decrease in signal (exon) to noise (intron) ratio* in the human genome leads to misprediction by computational gene-finding strategies" (Birney et al., 2001, p. 827; emphasis added).

What has become more and more evident after the publication of the draft sequences of 2001 and the announcement of the final one in 2003 is how far we still are from the long-awaited coming of theoretical, virtual biology, one in which investigation of functional variation - both in health and disease - would be pursued solely *in silico*. Annotation remains dependent on painstaking lab work for the validation of genes, “wet” work that may have been greatly abbreviated, but has not been made obsolete. On the other hand, there are now tens of thousands of genes awaiting this laborious scrutiny. Back to the lab bench, then:

“Although these searches highlight the power of the new genomic information, they also reveal important limitations. In particular, the existence of a related gene sequence does not mean that there is a corresponding protein: the sequence could be a non-expressed pseudogene. (...) Expression studies will be required to complement genomic information. A final caveat is that many of the factors are components of multi-subunit complexes. Sometimes the same factor is present in multiple complexes whose activities differ substantially. Thus, the full value of the genomic information can be realized only when it is coupled with appropriate biochemical studies” (Tupler et al., 2001, p. 833).

Those shortcomings are partially acknowledged in the article by the Human Genome Project (HGP) Consortium in *Nature* (Lander et al., 2001, p. 907 and p. 913). Other authors also felt compelled to point out the impossibility of analyzing the human genome solely by computational means (Bork and Copley, 2001, p. 819; Galas, 2001, p. 1257 and p. 1259). Some went as far as confessing some exasperation with the dual grip of sequencing and bioinformatics on biology, as did Tom Pollard in a quote for a news story published by *Nature*, confessing his fear that it might delay the unavoidable “wet” work in the lab without which biology would be unable to complete the understanding of physiology in the big picture (Butler, 2001, p. 760). A similar view is presented in his own article for the same issue of *Nature*, when he tries to overcome the Big Science version of biological research: “This [annotation] is one case where small science will yield a better product than the industrial approach required for sequencing” (Pollard, 2001, p. 843).

Most authors contributing to the historical issues of *Nature* and *Science* seem to foresee problems lurking from within the genomics/bioinformatics research strategy, although they surely do not go as far as to devalue it, which would indeed be nonsense. Baltimore (2001, p. 815), with the authority of an early critic of the HGP, who by its completion had become a well-balanced supporter, adeptly summarizes this ambiguity stating that post-sequencing analysis allows many global questions to be addressed, whereas details - in a nutshell, exactly what matters in any type of research - remain largely unaddressed:

“...it is clear that we do not gain our undoubted complexity over worms and plants by using many more genes. Understanding what does give us our complexity - our enormous behavioral repertoire, ability to produce conscious action, remarkable physical coordination (shared with other vertebrates), precisely tuned alterations in response to external variations of the environment, learning, memory. . . need I go on? - remains a challenge for the future” (Baltimore, 2001, p. 816).

The fact is that genomics is not only a biological research strategy, but also a technoscientific system in the making, which only starts to face some difficulties and opposition

beyond the institutional setting where it has been seeded. In spite of all the enthusiasm from the side of venture capitalists with the virtuoso duo genomics/bioinformatics during the high-tech bubble in the late 1990s, one already notices some signs of concern over its sustainability as a reliable source of revenue.

The alert about biotechnology performance as an investment came from Malakoff and Service (2001), writing in the news section of *Science*. They open their report mentioning the announcement by Millennium and Bayer in January 2001, with much fanfare, of a new wonder antitumoral drug, which was about to go through phase I clinical tests, only eight months after the target gene identification, thereby jumping two years in the usual drug licensing process. Both companies had boasted the new medicine as an industry landmark. The authors acknowledge that the Millennium/Bayer announcement could well be a sign that genomics was indeed beginning to fulfill its promises, but recommended caution: "Such expansive claims are not unusual in the biotechnology industry, which for more than a decade has hyped the profit making potential of sequencing human genes, only to see many of those claims founder in a sea of red ink" (Malakoff and Service, 2001, p. 1193). They identify three main fields in the sector: *tool-making* companies (DNA chips, automatic sequencers, etc.); gene finders (*genomics*) and *information* providers, and *drug* developers. According to their analysis, business opportunities at that implementation stage tended to be better for the first company type. When evaluating the outlook for the two other types, many of them initiated by academic researchers with the itch of pioneers, the news story in *Science* launches considerations that would later prove premonitory: "...the companies still have to show that they can move that speedily on a routine, sustained basis. Even then, some observers were skeptical that early agility would translate into substantially shorter drug development cycles, as major delays often occur during clinical trials and in the regulatory process" (Malakoff and Service, 2001, p. 1203).

Early in 2002, Craig Venter stepped down from Celera, when the company decided to try a reorientation towards a more traditional approach to drug development, after reaping too meager revenues with the genomic information provider model. The following year another high-profile researcher turned businessman, William Haseltine from Human Genome Sciences, felt the need to step down as well.

Then came in 2003 the actual completion of the HGP, when the sequence ceased to be a mere draft. Many articles by world genomic celebrities then published offer good examples of a renewed effort geared to the public sphere with the intent of reinvigorating biology's reputation as *Big Science*. This is indeed the case of an article penned by leaders of the three leading institutions engaged in the HGP (NIH, the National Institutes of Health, and DOE, the Department of Energy, from the U.S., and the Wellcome Trust, from the U.K.), respectively, Francis Collins, Aristides Patrinos and Michael Morgan. Their vocabulary remains as hyperbolic as in 2001: plenty of references to revolutions, new eras, visionary adventures, monumental scales, eternal benefits to mankind, exciting scientific challenges, etc. Even with formal concessions to the role of gene-environment interactions and to the complexity inherent to the genome, the ill-disguised purpose of the text was to justify and secure a continuous flow of research money, so that old promises could finally be met:

"The millions of people around the world who supported our quest to sequence the human genome did so with the expectation that it would benefit humankind. Now, at the dawning of the genome era, it is critical that we encourage the same intensity toward deriving medical benefits

from the genome that has characterized the historic effort to obtain the sequence. If research support continues at vigorous levels, we imagine that genome science will soon begin revealing the mysteries of hereditary factors in heart disease, cancer, diabetes, schizophrenia, and a host of other conditions” (Collins et al., 2003b, p. 290).

However, in 2003 the international panorama was completely different from that of 2001, after U.S. president George W. Bush’s election, September 11, and the Afghanistan and Iraq wars. Mentions in articles of the HGP as a concerted international effort, for instance, almost completely vanished. The U.K. was visibly intent to profit the most, during the celebrations around the 50th anniversary of the double helix, from the fact that Watson and Crick’s landmark work had been carried out at the Cavendish Laboratory in Cambridge. The international and cooperative moral high ground from which HGP leaders looked down to Celera’s alleged commercialism started to decline.

Nature and *Science* commemorated the 50th anniversary of the Watson and Crick article with a set of articles in April 2003. The most relevant to understand the future of genomics and bioinformatics was the paper in *Nature* by Collins et al. (2003a), from the U.S. National Human Genome Research Institute (NHGRI), and the work by five authors from the U.S. DOE in *Science* (Frazier et al., 2003). The outlook emerging from these articles is a sort of genomic Tordesilhas Treaty, with a reattribution of sequencing and research territories between NHGRI (human health and comparative genomics of animal species) and DOE (genomics applied to environment and energy, dealing with microorganisms and plants), somewhat as Spain and Portugal divided in 1494 domains still unknown in America.

With an installed capacity to sequence 15 to 20 human-scale genomes in five years (Collins et al., 2003a, p. 844), NHGRI had acquired enough technological momentum to start competing directly with former partners abroad. Its leaders devised a self-sustaining plan for the future in which the Book of Life metaphor is replaced by that of a Life Building, in which HGP is sort of downgraded to the role of mere foundations to support three main floors: Genomics to Biology, Genomics to Health, and Genomics to Society (Collins et al., 2003a, p. 836). First floor: the goal is to understand the human genome’s peculiar architecture, compiling a catalogue of all its functional elements (not only genes in the strict “coding” definition of the term). Second floor: to apply genome structural information to characterize diseases, in order to generate a new and molecular taxonomy for them, as well as to develop new therapeutic approaches. Third floor: to project genomic knowledge beyond the clinical context, extracting conclusions relevant to the racial, ethnic and behavioral fields, and debating foreseeable ethical consequences and limits surrounding these uses.

It must be more than a coincidence that DOE’s program presented in *Science* was named Genomes to Life (Frazier et al., 2003, p. 290), as if it were destined to add a fourth floor to NHGRI’s building, or maybe a neighboring edifice. The project in this case seems to be extending genomic nets over two fields that are crucial to the economy in their interdependence with nature, clean source energy and environmental quality, by sequencing plants and entire microbial communities in the hope that they will teach us ancestral biochemical lessons on how to deal with extreme environmental degradation: “A central goal of this program is to understand microbes and microbe communities, and their molecular machines and controls at the molecular level well enough that we can use them to address DOE and national needs” (Frazier et al., 2003, p. 291). Biology in the *Big Science* version enters the field of covert self-justification based on a set of

more fashionable values, e.g., concerns about U.S. homeland security: “Knowledge is power, and we must develop a comprehensive understanding of biological systems if we are to use their abilities effectively to meet daunting societal challenges” (Frazier et al., 2003, p. 293).

Symptomatically, new partners added to DOE’s team were summoned among former foes of HGP, Craig Venter and his collaborators, now housed by the Institute for Biological Energy Alternatives (IBEA), launched with a set of objectives all too close to the Genomes to Life program, as can be gleaned from its website (www.bioenergyalts.org/about.html):

“The Institute for Biological Energy Alternatives (IBEA) is a research-based institution dedicated to exploring solutions for carbon sequestration using microbes, microbial pathways, and plants. For example, genomics could be applied to enhance the ability of terrestrial and oceanic microbial communities to remove carbon from the atmosphere. IBEA will develop and use microbial pathways and microbial metabolism to produce fuels with higher energy content in an environmentally sound fashion. IBEA will undertake genome engineering to better understand the evolution of cellular life and how these cell components function together in a living system”.

One of the first results from the IBEA/DOE partnership was announced in April 2004: the bulk sequencing of all genomes belonging to microorganisms found in a sample of Sargasso Sea water, containing an estimated number of at least 1,800 species (148 of them unknown to scientists). From these tangled genomes IBEA was able to mine more than 1.2 million previously undescribed genes, identified solely on the basis of computational methods (Venter et al., 2004, p. 66), of which not less than 782 are probably involved in specifying photo-receptor proteins and, therefore, in the tapping of solar light.

Be it for the sake of human health or the global environment, a new genomic era has indeed begun - the age of wholesale genomics, dedicated to the appropriation of genes in the old Gold Rush fashion. It remains to be established if, reenacting HGP’s difficulties in keeping up with induced expectations in the public sphere, the new age will deliver all which is being promised and be able to overcome unavoidable ethical and political concerns normally raised in diverse societal sectors against this high-technology version of common encroachment.

However, the brand new skyscrapers which genomicists and bioinformaticians want to erect on public ground would rest on shaky foundations, if renovated expectations come to be seen as or effectively turn into a forward escape of sorts, sheer amplification of the maximalist rhetoric that has so far fueled the unstoppable march of molecular biology on the path to hegemony. The time has come for molecular biologists to face a sort of concern more typical of another, not long ago also very fashionable branch of biology, ecology: in a nutshell, *sustainability*.

REFERENCES

- Baltimore, D.** (2001). Our genome unveiled. *Nature* 409: 814-816.
- Birney, E., Bateman, A., Clamp, M.E. and Hubbard, T.J.** (2001). Mining the draft human genome. *Nature* 409: 827-828.
- Bork, P. and Copley, R.** (2001). Filling in the gaps. *Nature* 409: 818-820.
- Butler, D.** (2001). Are you ready for the revolution? *Nature* 409: 758-760.
- Collins, F., Green, E.D., Guttmacher, A.E. and Guyer, M.S.** (2003a). A vision for the future of genomics research. A blueprint for the genomic era. *Nature* 422: 835-847.
- Collins, F., Morgan, M. and Patrinos, A.** (2003b). The Human Genome Project: Lessons from large-scale biology. *Science* 300: 286-296.

- Frazier, M.E., Johnson, G.M., Thomassen, D.G., Oliver, C.E. and Patrinos, A.** (2003). Realizing the potential of the genome revolution. The Genomes to Life Program. *Science 300*: 290-293.
- Galas, D.J.** (2001). Making sense of the sequence. *Science 291*: 1257-1260.
- Lander, E.S., Linton, L.M., Birren, B. et al. and the International Human Genome Sequencing Consortium** (2001). Initial sequencing and analysis of the human genome. *Nature 409*: 860-921.
- Malakoff, D. and Service, R.F.** (2001). Genomania meets the bottom line. *Science 291*: 1193-1203.
- Pollard, T.D.** (2001). Genomics, the cytoskeleton and motility. *Nature 409*: 842-843.
- Tupler, R., Perini, G. and Green, M.** (2001). Expressing the human genome. *Nature 409*: 832-833.
- Venter, J., Remington, K., Heidelberg, J.F. et al.** (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science 304*: 66-74.