Opinion

Genomics: what is realistically achievable?

Ross Overbeek

Address: Integrated Genomics Inc., 2201 W Campbell Park Drive, Chicago, IL 60612, USA. E-mail: Ross@IntegratedGenomics.com

Published: 28 July 2000

Genome ***Biology*** 2000, **1(2)**:comment2002.1–2002.3

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2000/1/2/comment/2002

We now have a large and growing number of sequenced genomes. It is widely understood that this presents research opportunities and promises to change the way biology advances, but the magnitude and nature of the opportunities is, for the most part, poorly understood. In this short piece, I wish to examine the following two questions: First, how quickly will sequence data be produced? Second, what impact will this have on our understanding of the sequenced organisms?

Since I am a computer scientist by training, I tend to think of the current situation in which the field of genomics is being driven forward by rapid technological advances as quite analogous to the sequence of events in computing that were triggered by advances in microcomputer and network technologies. I distinctly remember the early period in which it seemed clear to most computer scientists (including myself) that technical advances were very desirable and interesting, but could have little impact on either the fundamental research issues or the overall advance of the field. Most of us completely underestimated the impact of exponential price improvements in key-enabling technologies. Certainly no one that I know of foresaw in any detail the current world of computing (although a few had rare insights into the potential). As we face the world generated by the web, we should remember that as late as the early 1990s common wisdom indicated that 'movies on demand' would be the application that drove increased network bandwidth.

How quickly will sequence data be produced?

Most readers will be aware that the amount of actual sequence data available to the research community has doubled roughly every 1.5 years for at least the last decade. This is completely analogous to the doubling of microprocessor speed every 1.5 years in the computing community. The question in such situations is: how long can such growth be sustained? It is clear that there are very real limits in both cases, and there is certainly no physical law dictating that such growth will continue.

There are actually three closely related 'laws' that should be considered: (1) the amount of available DNA sequence data will double every 18 months; (2) the number of available genomes will double every 18 months; (3) the cost of sequence will drop by a factor of 2 every 18 months.

They are all closely related, but quite distinct. Over the last decade, the first law has been approximately accurate. The number of available genomes has actually exceeded the number predicted by law 2 since 1995, and the costs have dropped somewhat less than predicted by the third law. The trends driving this rapid increase are improvements in sequencing technology, expansion of the market, and increased public funding. These three laws simply estimate the impact of these trends. On the assumption that the amounts of funding stay fixed, laws 1 and 3 are equivalent. But this almost certainly will not be the case. During the last decade we have gone through a period in which the tangible applications have been minimal - they have all been part of a projected future. This produced a situation in which funding was dictated largely by political issues worldwide, and the levels expanded gradually. I expect the amounts of available funding to expand rapidly as applications start to expand; conversely, if application areas do not meet expectations it will contract somewhat.

What would it mean for the third law to hold for 15 years? To pursue this issue, first note that a drop in price by a factor of 2 every 18 months implies a drop of about 10-fold every 5 years. Sequence now costs somewhere between $0.10 and $0.30 per basepair (bp), depending on quality of coverage and annotation. This would imply (as illustrated in Table 1) that the cost per base in 1995 would have been between $1 and $3 (which is reasonably accurate) and that it will be about 0.001-0.003 cents per base in 2020.

That is, the cost to sequence the equivalent of a human genome will drop from about $3-9 billion in 1995 to about $30,000-$90,000 in 2020. If the trend were to even

Table 1

The predicted decrease in cost of DNA sequencing

Year	Cost (cents) per base
1995	100-300
2005	1-3
2010	0.1-0.3
2015	0.01-0.03
2020	0.001-0.003

approximate truth, by 2025 we would expect many individuals to have their genomes sequenced for medical reasons.

There are a number of complicating factors relating to this analysis - most notably how much money will be spent and how much of it will be public. It is obvious that estimates of these amounts will be extremely speculative. Let us, for purposes of analysis, suppose that $1 billion per year worldwide will be spent on average. It would be easy to argue that the real amount will be less, but given the potential application areas it is equally easy to argue that it will rapidly grow to an order of magnitude more. Let us expand our original table with a column giving the rate at which sequence accumulates under the assumption of a fixed funding rate of $1 billion per year (Table 2).

That is, now we are accumulating the equivalent of one to three human genomes per year, and by 2020 we will be accumulating on the order of 10,000 times as much data per year. This is under the assumption of a constant funding level of $1 billion per year. It seems far more likely to me that the application areas will explode by 2010, driving the investments in sequencing up by at least an order of magnitude. This would allow the sequencing of 100,000 human genome equivalents per year by about 2020.

There are certainly many issues that would suggest the above analysis is misguided. One of the most obvious is that a majority of the funding may rapidly shift from sequencing to functional analysis of genomes. Since the overall amount of funding is constrained (and at this point fairly tightly constrained), this could reduce the projected advances in sequencing capacity by a factor of two to three.

More importantly, could we do anything useful with 100 trillion basepairs of data per year? In the application areas of medicine and agriculture, it seems likely that the presence of these volumes of sequence would produce major changes in industries in which $1 billion per year is relatively small change. So, the answer is: yes, we could justify this investment in terms of application areas of importance.

What impact will this have on our understanding of the sequenced organisms?

It is likely that the initial investments in sequencing will be made in a few application areas of narrow interest to the research community. The bulk of the data will be studied by researchers focused on a few questions of huge practical, but of little scientific, importance. But the impact of driving the prices of sequence down will be rapidly felt within the entire research community. While the vast majority of sequence produced will have minimal scientific impact, the rapid drop in cost of sequencing will profoundly affect biology. It will allow the sequencing of a large, diverse collection of organisms, and this body of sequence data will support the comparative analysis that will reshape our understanding of biology.

To illustrate, let me just consider one area - biochemistry - and restrict the discussion to just the prokaryotes. It came as a surprise to me that many aspects of what everyone would call 'core biochemistry' remain unclear. This realization occurred at a workshop in which a number of us were trying to enumerate the set of genes required by 'core machinery'; when we reached the topic of coenzyme metabolism, the fact that there were numerous gaps in current understanding became obvious. We do not have to search beyond basic pathways from common textbooks to illustrate our ignorance; at this point we do not even know the initial steps of aromatic amino acid biosynthesis used by Archaea. I cite these areas just as examples. If you were to ask a good biochemist how many genes corresponding to the core machinery remain unclassified, you will probably begin a most interesting discussion. An initial period will be spent defining what is meant by core machinery, and this is an immensely important question. Once that is settled, you will probably reach the point where the question has become precise, but even an approximate answer cannot be reached. As a working hypothesis, I will suggest that there are somewhere between 50 and 200 enzymes that catalyze reactions representing central pathways that are not yet understood or cases in which the reactions are known to exist but for which no genes (in any organism) can be identified as encoding the enzyme.

Table 2

The projected rate of sequence accumulation

Year	Cost (cents) per base	Minimal rate of growth (base-pair/year)
2000	10-30	1×10^{10}
2005	1-3	1×10^{11}
2010	0.1-0.3	1×10^{12}
2015	0.01-0.03	1×10^{13}
2020	0.001-0.003	1×10^{14}

Assumes a fixed rate of funding at $1 billion/year.

Then, there is a related question. We often find an organism in which it is clear that an enzymatic activity is encoded in one or more of its genes (and examples of other genes encoding this activity in other organisms do exist), but we are unable to identify any candidate genes in the given organism. In most such cases, the organism has an alternative form of the enzyme - a gene exists that encodes the activity, but it does not resemble any of the broad and growing collection of classified genes. How many of these alternative forms exist?

So, how will the presence of many, many genomes affect our ability to clarify these (relatively few) uncharacterized areas of core biochemistry? Most biologists would expect the problem to be gradually resolved using classical techniques from biochemistry and genetics. It is clear that once one gene in any organism is properly characterized, that characterization can be projected (albeit in a somewhat error-prone way) to many other organisms; however, the rate-limiting process is assigning a function to the first representative. In my opinion, the assignment of these 'new' functions will be driven by a two-step process. The second step will be achieved by classical techniques - confirmation of hypothesized functions. The first step, the derivation of the hypotheses, will be a bioinformatics task, and it will be possible largely as a result of the availability of large numbers of genomes.

The techniques for generating the critical hypotheses from genomics data are now emerging. Here, I will summarize them briefly. In prokaryotes, functionally related genes (for example, genes that encode subunits of the same enzyme or enzymes from the same pathway) tend to cluster on the chromosome. This tendency is strong enough to allow accurate determination of the genes participating in a specific subsystem, once the number of sequenced genomes containing the biochemical subsystem exceeds about 20.

If one has a collection of sequenced genomes, X of which contain a given functional capability and Y of which do not, then the genes encoding the capability must come from the set present in all of X and none of Y. For this assertion to even be well defined requires the notion of 'a gene present in all of X' to be clear. Unfortunately, in cases in which alternative forms of an enzymatic function exist or in which paralogous genes exist within the set of organisms (with varying specificities and activities), the simplistic view inherent in the assertion breaks down. Even so, as the number of genomes grows, the thought process leading to the simple assertion becomes increasingly valuable.

As the number of sequenced genomes has grown, it has become possible to characterize regulatory sites and, hence, regulons. This technology is still emerging, but it clearly has enormous potential. It depends on having the sequence for a number of very closely related organisms, and this allows a comparative analysis of the upstream regions of genes. Currently, the enteric bacteria and, perhaps, a region in the gram-positive bacteria around *Bacillus subtilis* are becoming densely enough sampled. As the success of the technology improves and the number of densely sampled regions increases, one would expect this technology to play a major role in identifying functionally coupled genes.

As microarray data from genomic sequences become increasingly available, the expression data that will be produced will obviously be directly relevant. There will probably be other forms of data, as well, but these adequately illustrate the point: the generation of hypotheses will flow from integrating a number of such sources of data. The existence of a growing number of hypotheses will guide the rate-limiting 'wet' lab efforts.

Let me now return to the 50-200 'missing genes' corresponding to enzymatic processes that are either known to exist, but for which we have not located the relevant genes, or to pathways that have not yet been clarified. Is it reasonable to speak of identifying a majority of these genes/functions within, say, a five year period? My assertion is that it is now reasonable to speak in such terms, that to succeed will require a tight coupling of wet lab activity to bioinformatics, and that the presence of large numbers of genomes will directly begin to reshape the core research areas in biology.