



Beyond the Tsunami: Developing the Infrastructure to Deal with Life Sciences Data

CHRISTOPHER
SOUTHAN

GRAHAM
CAMERON

EMBL-European
Bioinformatics Institute

SCIENTIFIC REVOLUTIONS ARE DIFFICULT TO QUANTIFY, but the rate of data generation in science has increased so profoundly that we can simply examine a single area of the life sciences to appreciate the magnitude of this effect across all of them. Figure 1 on the next page tracks the dramatic increase in the number of individual bases submitted to the European Molecular Biology Laboratory Nucleotide Sequence Database¹ (EMBL-Bank) by the global experimental community. This submission rate is currently growing at 200% per annum.

Custodianship of the data is held by the International Nucleotide Sequence Database Collaboration (INSDC), which consists of the DNA Data Bank of Japan (DDBJ), GenBank in the U.S., and EMBL-Bank in the UK. These three repositories exchange new data on a daily basis. As of May 2009, the totals stood at approximately 250 billion bases in 160 million entries.

A recent submission to EMBL-Bank, accession number FJ982430, illustrates the speed of data generation and the effectiveness of the global bioinformatics infrastructure in responding to a health crisis. It includes the complete H1 subunit sequence of 1,699 bases from the first case of novel H1N1 influenza virus in Denmark. This was submitted on May 4, 2009, within days of

¹ www.ebi.ac.uk/embl

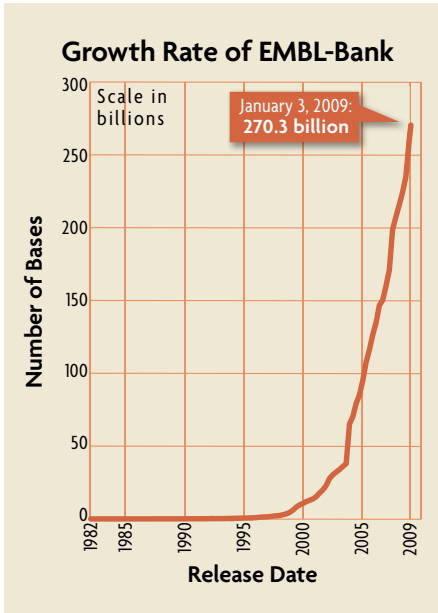


FIGURE 1.
Growth in the number of bases deposited in EMBL-Bank from 1982 to the beginning of 2009.

Genome campus include scientists who generate data and administer the databases into which it flows, biocurators who provide annotations, bioinformaticians who develop analytical tools, and research groups that seek biological insights and consolidate them through further experimentation. Consequently, it is a community in which issues surrounding computing infrastructure, data storage, and mining are confronted on a daily basis, and in which both local and global collaborative solutions are continually explored.

The collective name for the nucleotide sequencing information service is the European Nucleotide Archive [1]. It includes EMBL-Bank and three other repositories that were set up for new types of data generation: the Trace Archive for trace data from first-generation capillary instruments, the Short Read Archive for data from next-generation sequencing instruments, and a pilot Trace Assembly Archive that stores alignments of sequencing reads with links to finished genomic sequences in EMBL-Bank. Data from all archives are exchanged regularly with the National Center for Biotechnology Information in the U.S. Figure 2 compares the sizes of

the infected person being diagnosed. Many more virus subunit sequences have been submitted from the U.S., Italy, Mexico, Canada, Denmark, and Israel since the beginning of the 2009 global H1N1 pandemic.

EMBL-Bank is hosted at the European Bioinformatics Institute (EBI), an academic organization based in Cambridge, UK, that forms part of the European Molecular Biology Laboratory (EMBL). The EBI is a center for both research and services in bioinformatics. It hosts biological data, including nucleic acid, protein sequences, and macromolecular structures. The neighboring Wellcome Trust Sanger Institute generates about 8 percent of the world's sequencing data output. Both of these institutions on the Wellcome Trust

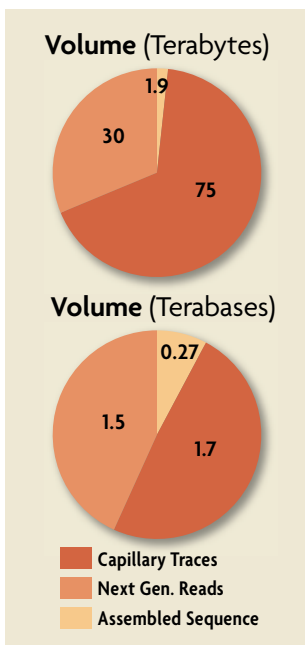


FIGURE 2.
The size in data volume and nucleotide numbers of EMBL-Bank, the Trace Archive, and the Short Read Archive as of May 2009.

EMBL-Bank, the Trace Archive, and the Short Read Archive.

THE CHALLENGE OF NEXT-GENERATION SEQUENCING

The introduction in 2005 of so-called next-generation sequencing instruments that are capable of producing millions of DNA sequence reads in a single run has not only led to a huge increase in genetic information but has also placed bioinformatics, and life sciences research in general, at the leading edge of infrastructure development for the storage, movement, analysis, interpretation, and visualization of petabyte-scale datasets [2]. The Short Read Archive, the European repository for accepting data from these machines, received 30 terabytes (TB) of data in the first six months of operation—equivalent to almost 30% of the entire EMBL-Bank content accumulated over the 28 years since data collection began. The uptake of new instruments and technical developments will not only increase submissions to this archive manifold within a few years, but it will also preclude the arrival of “next-next-generation” DNA sequencing systems [3].

To meet this demand, the EBI has increased storage from 2,500 TB (2.5 PB) in 2008 to 5,000 TB (5 PB) in 2009—an approximate annual doubling. Even if the capacity keeps pace, bottlenecks might emerge as I/O limitations move to other points in the infrastructure. For example, at this scale, traditional backup becomes impractically slow. Indeed, a hypothetical total data loss at the EBI is estimated to require months of restore time. This means that streamed replication of the original data is becoming a more efficient option, with copies being stored at multiple locations. Another bottleneck example is that technical advances in data transfer speeds now exceed the capacity to write out to disks—about 70 megabits/sec, with no imminent expectation of major performance increases. The problem can be ameliorated by writing to multiple disks, but at a considerable increase in cost.

This inexorable load increase necessitates continual assessment of the balance



between submitting derived data to the repositories and storing raw instrument output locally. Scientists at all stages of the process, experimentalists, instrument operators, datacenter administrators, bioinformaticians, and biologists who analyze the results will need to be involved in decisions about storage strategies. For example, in laboratories running high-throughput sequencing instruments, the cost of storing raw data for a particular experiment is already approaching that of repeating the experiment. Researchers may balk at the idea of deleting raw data after processing, but this is a pragmatic option that has to be considered. Less controversial solutions involve a triage of data reduction options between raw output, base calls, sequence reads, assemblies, and genome consensus sequences. An example of such a solution is FASTQ, a text-based format for storing both a nucleotide sequence and its corresponding quality scores, both encoded with a single ASCII character. Developed by the Sanger Institute, it has recently become a standard for storing the output of next-generation sequencing instruments. It can produce a 200-fold reduction in data volume—that is, 99.5% of the raw data can be discarded. Even more compressed sequence data representations are in development.

GENOMES: ROLLING OFF THE PRODUCTION LINE

The production of complete genomes is rapidly advancing our understanding of biology and evolution. The impressive progress is illustrated in Figure 3, which depicts the increase of genome sequencing projects in the Genomes OnLine Database (GOLD).

While the figure was generated based on all global sequencing projects, many of these genomes are available for analysis on the Ensembl Web site hosted jointly by the EBI and the Sanger Institute. The graph shows that, by 2010, well over 5,000 genome projects will have been initiated and more than 1,000 will have produced complete assemblies. A recent significant example is the bovine genome [4], which followed the chicken and will soon be joined by all other major agricultural species. These will not only help advance our understanding of mammalian evolution and domestication, but they will also accelerate genetic improvements for farming and food production.

RESEQUENCING THE HUMAN GENOME: ANOTHER DATA SCALE-UP

Recent genome-wide studies of human genetic variation have advanced our understanding of common human diseases. This has motivated the formation of an international consortium to develop a comprehensive catalogue of sequence variants in

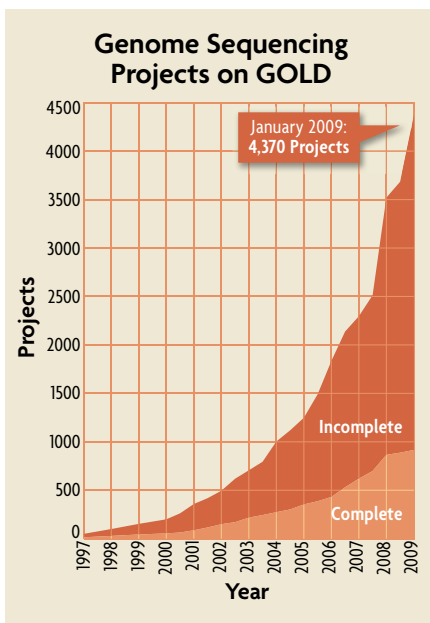


FIGURE 3. The increase in both initiated and completed genome projects since 1997 in the Genomes OnLine Database (GOLD). Courtesy of GOLD.

to more than two human genomes (at 2.85 billion per human) every 24 hours. The completed dataset of 6 trillion DNA bases will be 60 times more sequence data than that shown earlier in Figure 1.

THE RAISON D'ÊTRE OF MANAGING DATA: CONVERSION TO NEW KNOWLEDGE

Even before the arrival of the draft human genome in 2001, biological databases were moving from the periphery to the center of modern life sciences research, leading to the problem that the capacity to mine data has fallen behind our ability to generate it. As a result, there is a pressing need for new methods to fully exploit not only genomic data but also other high-throughput result sets deposited in databases. These result sets are also becoming more hypothesis-neutral compared with traditional small-scale, focused experiments. Usage statistics for EBI services, shown in Figure 4 on the next page, show that the biological community, sup-

multiple human populations. Over the next three years, the Sanger Institute, BGI Shenzhen in China, and the National Human Genome Research Institute's Large-Scale Genome Sequencing Program in the U.S. are planning to sequence a minimum of 1,000 human genomes.

In 2008, the pilot phase of the project generated approximately 1 terabase (trillion bases) of sequence data per month; the number is expected to double in 2009. The total generated will be about 20 terabases. The requirement of about 30 bytes of disk storage per base of sequence can be extrapolated to about 500 TB of data for the entire project. By comparison, the original human genome project took about 10 years to generate about 40 gigabases (billion bases) of DNA sequence. Over the next two years, up to 10 billion bases will be sequenced per day, equating

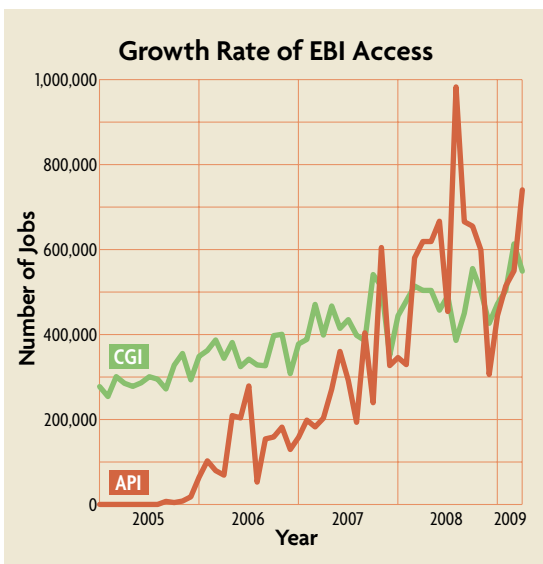


FIGURE 4. Web accesses (Common Gateway Interface [CGI]) and Web services usage (application programming interface [API]) recorded on EBI servers from 2005 to 2009.

ported by the bioinformatics specialists they collaborate with, are accessing these resources in increasing numbers.

The Web pages associated with the 63 databases hosted at the EBI now receive over 3.5 million hits per day, representing more than half a million independent users per month. While this does not match the increase in rates of data accumulation, evidence for a strong increase in data mining is provided by the Web services' programmatic access figures, which are approaching 1 million jobs per month.

To further facilitate data use,

the EBI is developing, using open standards, the EB-eye search system to provide a single entry point. By indexing in various formats (e.g., flat files, XML dumps, and OBO format), the system provides fast access and allows the user to search globally across all EBI databases or individually in selected resources.

EUROPEAN PLANS FOR CONSOLIDATING INFRASTRUCTURE

EBI resources are effectively responding to increasing demand from both the generators and users of data, but increases in scale for the life sciences across the whole of Europe require long-term planning. This is the mission of the ELIXIR project, which aims to ensure a reliable distributed infrastructure to maximize access to biological information that is currently distributed in more than 500 databases throughout Europe. The project addresses not only data management problems but also sustainable funding to maintain the data collections and global collaborations. It is also expected to put in place processes for developing collections for new data

types, supporting interoperability of bioinformatics tools, and developing bioinformatics standards and ontologies.

The development of ELIXIR parallels the transition to a new phase in which high-performance, data-intensive computing is becoming essential to progress in the life sciences [5]. By definition, the consequences for research cannot be predicted with certainty. However, some pointers can be given. By mining not only the increasingly comprehensive datasets generated by genome sequencing mentioned above but also transcript data, proteomics information, and structural genomics output, biologists will obtain new insights into the processes of life and their evolution. This will in turn facilitate new predictive power for synthetic biology and systems biology. Beyond its profound impact on the future of academic research, this data-driven progress will also translate to the more applied areas of science—such as pharmaceutical research, biotechnology, medicine, public health, agriculture, and environmental science—to improve the quality of life for everyone.

REFERENCES

- [1] G. Cochrane et al., “Petabyte-scale innovations at the European Nucleotide Archive,” *Nucleic Acids Res.*, vol. 37, pp. D19–25, Jan. 2009, doi: 10.1093/nar/gkn765.
- [2] E. R. Mardis, “The impact of next-generation sequencing technology on genetics,” *Trends Genet.*, vol. 24, no. 3, pp. 133–141, Mar. 2008, doi: 10.1016/j.physletb.2003.10.071.
- [3] N. Blow, “DNA sequencing: generation next-next,” *Nat. Methods*, vol. 5, pp. 267–274, 2008, doi: 10.1038/nmeth0308-267.
- [4] Bovine Genome Sequencing and Analysis Consortium, “The genome sequence of taurine cattle: a window to ruminant biology and evolution,” *Science*, vol. 324, no. 5926, pp. 522–528, Apr. 24, 2009, doi: 10.1145/1327452.1327492.
- [5] G. Bell, T. Hey, and A. Szalay, “Beyond the Data Deluge,” *Science*, vol. 323, no. 5919, pp. 1297–1298, Mar. 6, 2009, doi: 10.1126/science.1170411.