Genomics Proteomics Bioinformatics 13 (2015) 1–3

HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

PERSPECTIVE

On Bioinformatic Resources

Runsheng Chen [*,a]

CAS Key Laboratory of RNA Biology, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China

Received 23 January 2015; accepted 15 February 2015
Available online 2 March 2015

A starting point of curating bioinformatic resources for the public is marked by the establishment of the US National Center for Biotechnology Information (NCBI) in 1988 [1]. One of its many purposes is certainly to echo the initiative of the Human Genome Project (HGP) — when two landmark reports were published at the same time: "*Mapping and Sequencing the Human Genome*" by the National Research Council [2] and "*Mapping Our Genes — The Genome Project*: *How Big*, *How Fast*?" by the US Congress [3].

As HGP is prepared to scale up its sequencing operations about 5 years into the Project, a new discipline – bioinformatics – became inevitable. It emerged originally for processing and sharing genome sequences and was thus known as genome informatics. In the first five-year plan of HGP, it was pointed out that genome informatics is a scientific discipline that encompasses all aspects of genome information acquisition, processing, storage, distribution, analysis, and interpretation [4]. This insightful statement was footnoted by the launch of GenBank at NCBI in 1992, a database of nucleotide sequences, in collaboration with its international partners at the European Molecular Biology Laboratory (EMBL) and the DNA Data Bank of Japan (DDBJ). Now, some 20 years have passed, not only GenBank still prospers but other databases on human genomics and biology have emerged to fill up all data landscapes [5].

Genomics together with many other research fields of life sciences had entered *the Era of Large-scale Data Acquisition* in the early 90's. The Era was led by the fast accumulation of human genomic sequences and followed by similar data from other large model organisms [6]. As soon as HGP declared the human genome sequenced in 2001 [7], the HapMap Project [8] was announced to sequence hundreds of human genomes in a diverse population background. Microbial genomics has also been pursued into both metagenomics and pangenomics [9,10]. The big data collection efforts have been extending all the way into its vertical path — from DNA to RNA to Protein [11,12]. Another expansion of the efforts is horizontally toward biological functions and disease relevance [13–15]. In order to annotate biological functions at different levels of anatomy and physiology — cell, organ, tissue, and systems — for biomedical applications, data integration and interpretation become more relevant, and thus the birth of systems biology, which takes account of data from all omics fields and deciphers them in a context of biological networks. In recent years, with strong funding and rapid development, the sequence technology has been advancing significantly, resulting in dramatic cost reduction and explosion of data accumulation. Therefore, a large number of bioinformatic resources become obvious and there is demand for some comprehensive digestions.

There are several data types being curated in bioinformatic resources world-wide. The first of them is classified as scale-up omics data, which includes genome-wide data from the vertical path as mentioned above: genome, transcriptome, proteome, and metabolome, just to name a few. Each of these data may have different disease relevance, such as DNA variations for cancer predisposition and protein structures for drug development. The second type of data resources includes data generated by molecular biologists that tend to look at molecular details, such as transcription factors and signal transduction pathways, where interactions among macromolecules and

[*] Corresponding author.
E-mail: chenrs@sun5.ibp.ac.cn (Chen R).
[a] ORCID: 0000-0002-1640-6481.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

http://dx.doi.org/10.1016/j.gpb.2015.02.002

structure–function relationships are highly valued. The third type of data resources concerns bioinformatic tools (algorithms and software packages), which include those for sequence assembly and stitching, gene prediction and annotation, protein structure prediction, multiple sequence alignment, phylogeny, and network analysis. The fourth type is the primary and secondary literature that contains research papers, monographs, reviews, conference reports, and even personal websites and blogs. It also includes publically-available educational materials, such as training courses, online tutorials, and e-books. The last, but not the least, refers to web-based gateways of various institutions, such as research institutes, academic journals, and topic-focusing websites. All the aforementioned types of information are mainly stored in a variety of bioinformatic databases, accessible through the internet.

Another characteristic trend of bioinformatic resources is the increasing bodies of contributors and users from all fields of life sciences. First, it calls for more and better user-friendly tools. In this case, bioinformaticians are held responsible for building better hubs — usually containing databases and toolkits — for users from the diverse research fields. Second, it forces the contributors and users to start with their own urgent demands and to be innovative for mining data, providing tools, and understanding information imbedded in the data. In the meantime, the crowdfunding activities often lead to resource allocations and thus generate new fields or disciplines. Bioinformatics itself is already a complicated field, and its integration and reaching-out to chemistry efforts have given birth to systems biology and synthetic biology, respectively. Obviously, the demands coming from medicine and pharmaceutics are composed of molecular markers for disease diagnostics and prognostics, as well as drug targets and treatment strategies. The birth of translational medicine represents a field of research seeking medical applications for genomic data. The ultimate goal of genomic data and bioinformatics is to pave a way for medicine to be personalized and more precision and thus the new path — precision medicine. Other extensions include applications in a much broader settings, such as environment, energy, and agriculture. Third, the complexity of the data and information, as well as the conversion of both into knowledge, are all seeking for more intensive collaborations involving experts and talents from relevant disciplines, other than biologists, such as mathematicians, physicists, chemists, and computer scientists.

I would like to end this perspective by emphasizing a couple of the major issues of bioinformatics, encountered over my career path. On one hand, we still face the same challenge as the past three decades — best annotating the large number of experimental data; one of the examples is non-coding sequences in the genomes, such as the human genome. There are at most 3% of the human genomes are "*gene coding*", and the rest is classified as non-coding, sometimes the "dark matter" without functional definitions. The full interpretation should include all sequence elements of a genome. On the other hand, we still do not have enough high-quality data for biological applications, such as in the case of cancer studies, where only thousands of samples have been sequenced so far for the most complicated disease threatening all members of our mankind on a daily basis. Most important for the field of bioinformatics is that as more high-quality data are generated, we apparently are stepping into the "Big Data Era". Therefore, we need to raise the bar higher for better data and more sophisticated tools. *Genomics*, *Proteomics & Bioinformatics* (GPB) is apparently doing its job by evaluating the existing databases, algorithms, and toolkits (see related articles in this and the upcoming issues). Through summarizing the quotation and unique features of the databases and tools, the authors are trying to provide users clear ideas as to what may be the best to use and tool developers as the right directions for new initiatives. For instance, biological networks, which are time-dependent and non-linear, have to be built with full participation of the relevant parties — not only proteins and enzymes but also RNA molecules that are truly involved in cellular complexes for appropriate functions. As a final note, we would like to see some consolidations where the current bioinformatic resources are organized by applications, ranked based on user feedbacks, and evaluations by all users.

Completing interests

The author declared that there is no competing interest.

Acknowledgements

This work was supported by the National High-tech R&D Program of China (863 Program; Grant Nos: 2012AA020402 and 2012AA02A202).

References

[1] Smith K. A brief history of NCBI's formation and growth. In: The NCBI handbook [Internet]. Bethesda: National Center for Biotechnology Information; 2013, http://www.ncbi.nlm.nih.gov/books/NBK148949/.

[2] National Research Council (US) Committee on Mapping and Sequencing the Human Genome. Mapping and sequencing the human genome. Washington (DC): National Academies Press; 1988, http://www.ncbi.nlm.nih.gov/books/NBK218252/.

[3] US Congress, Office of Technology Assessment. Mapping our genes — the Genome Project: how big, how fast. Washington (DC): US Government Printing Office; 1988, http://www.ornl.gov/sci/techresources/Human_Genome/publicat/OTAreport.pdf.

[4] US Department of Health and Human Services, US Department of Energy. Understanding our genetic inheritance — the US Human Genome Project: the first five years FY 1991–1995; 1990. http://www.ornl.gov/sci/techresources/Human_Genome/project/5yrplan/firstfiveyears.pdf.

[5] Zou D, Ma L, Yu J, Zhang Z. Biological databases for human research. Genomics Proteomics Bioinformatics 2015;13:55–63.

[6] Tang B, Wang Y, Zhu J, Zhao W. Web resources for model organism studies. Genomics Proteomics Bioinformatics 2015;13:64–8.

[7] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.

[8] International HapMap Consortium. The International HapMap Project. Nature 2003;426:789–96.

[9] Sun Q, Liu L, Wu L, Li W, Liu Q, Zhang J, et al. Web resources for microbial data. Genomics Proteomics Bioinformatics 2015;13:69–72.

[10] Xiao J, Zhang Z, Wu J, Yu J. A brief review of software tools for pangenomics. Genomics Proteomics Bioinformatics 2015;13:73–6.

[11] Chen T, Zhao J, Ma J, Zhu Y. Web resources for mass spectrometry-based proteomics. Genomics Proteomics Bioinformatics 2015;13:36–9.

[12] Zhao D, Wu J, Zhou Y, Gong W, Xiao J, et al. WikiCell: a unified resource platform for human transcriptomics research. Omics 2012;16:357–62.

[13] Zhang G, Zhang Y, Ling Y, Jia J. Web resources for pharmacogenomics. Genomics Proteomics Bioinformatics 2015;13:51–4.

[14] Yang Y, Dong X, Xie B, Ding N, Chen J, Li Y, et al. Databases and web tools for cancer genomics study. Genomics Proteomics Bioinformatics 2015;13:46–50.

[15] Wei T, Peng X, Ye L, Wang J, Song F, Bai Z, et al. Web resources for stem cell research. Genomics Proteomics Bioinformatics 2015;13:40–5.