Genomics Proteomics Bioinformatics 12 (2014) 187–189

HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

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

PERSPECTIVE

Big Biological Data: Challenges and Opportunities

Yixue Li *, Luonan Chen *

Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China

Received 1 September 2014; revised 1 October 2014; accepted 1 October 2014
Available online 14 October 2014

In "Omics" era of the life sciences, data is presented in many forms, which represent the information at various levels of biological systems, including data about genome, transcriptome, epigenome, proteome, metabolome, molecular imaging, molecular pathways, different population of people and clinical/medical records. The biological data is big, and its scale has already been well beyond petabyte (PB) even exabyte (EB). Nobody doubts that the biological data will create huge amount of values, if scientists can overcome many challenges, *e.g.*, how to handle the complexity of information, how to integrate the data from very heterogeneous resources, what kind of principles or standards to be adopted when facing with the big data. Tools and techniques for analyzing big biological data enable us to translate massive amount of information into a better understanding of the basic biomedical mechanisms, which can be further applied to translational or personalized medicine.

Today, big data is one of the hottest topics in information science, but its concept can be misleading or confusing. The name itself suggests huge amount of data, which, however, represents only one aspect. In general, big data has four important features, so called four V's: volume of data, velocity of processing the data, variability of data sources, and veracity of the data quality. These four hallmarks of big data require to be characterized by special theory and technology; however, currently there is no satisfactory solution. Now, more biologists are involved with the big data due to the rapid advance of high-throughput biotechnologies. As an example, the Human Genome Project utilized the expertise, infrastructure, and people from 20 institutions and took 13 years of work with over $3 billion to determine the whole genome structure of approximately three billion nucleotides. But now we can sequence a whole human genome for $1000 and within three days. We have spent decades struggling to collect enough biological and biomedical data, but when big data overwhelms us, are we ready to face the challenge? The new bottleneck to this problem in biology is how to reveal the essential mechanisms of biological systems by understanding the big noisy data. Life sciences today need more robust, expressive, computable, quantitative, accurate and precise ways to handle the big data. As a matter of fact, recent works in this area have already brought remarkable advantage and opportunities, which implies the central roles of bioinformatics and bioinformaticians in the future research of the biological and biomedical fields. In the following text, we describe several aspects of big biological data based on our recent studies.

Expanding volume of the big biological data and its bonanza

With the increasingly accumulated large volumes of information about human, animals or microbe, researchers are starting to grapple with massive datasets and further elucidate the fundamental implications of those datasets in biology. For instance, a recent genetics study about six dog breeds to decipher adaption mechanisms under hypoxia in highland area has produced 3.2 terabytes (TB) genome sequencing data of 60 dogs from different altitudes along the "Ancient Tea Horse Road" [1]. After the data analysis, an important nonsynonymous mutation, G305S, was found on gene *EPAS1*enocding endothelial Per-Arnt-Sim (PAS) domain protein 1 and this mutation is most possibly involved in the adaption mechanisms of hypoxia. In a milestone study of using next-generation sequencing technology, Jay Shendure and his co-workers captured 12 human exomes [2]. They obtained over

* Corresponding authors.
E-mail: yxli@sibs.ac.cn (Li Y), lnchen@sibs.ac.cn (Chen L).

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.2014.10.001

30 gigabyte (GB) DNA sequencing data and found a SNP in gene *MYH3* straightforward, which is causative for a monogenic human disease called Freeman–Sheldon syndrome (FSS). Another impressive big data based biomedicine study was performed by Lupski and his colleagues [3]. In this work, 12 pedigree samples were sequenced to acquire their whole genome DNA data for finding causative variations related to Charcot-Marie-Tooth (CMT) disease. Over one TB of genome sequencing data were analyzed, and then a novel missense mutation Y169H on gene *SH3TC2* was found associated with CMT disease. All such studies show that big biological data has become an essential part of biological discovery and biomedical research [4], and clearly, the aforementioned discoveries are inconceivable without big biological data. Exponential growth in the amount of biological data undoubtedly means potential bonanza opportunities; nonetheless, we need to develop revolutionary measures for data management, analysis and accessibility.

Hypothesis-driven study is still a key for big biological data mining

Biological big data, in general, has the similar properties to the 4Vs of big data, in particular, at molecular levels. But, unlike the data gathered by Google, WeChat and Ali Baba, the big biological data is highly heterogeneous; even within the data, there exist intrinsic structures determined by various biological principles and experiment designs. Because of the 4-V features of big data, association or correlation rather than causal relationships are expected to be built among certain elements, such as genes, proteins, and pathways, across the whole big data. However, biological studies always need to know driving force or causal relationship among biological elements, which form complex biological systems. Many studies show that the existing intrinsic structures determined by various biological principles and experiment designs have already provided biological data miners or curators possible ways to identify causal relationship among biological molecules in big biological data. In this case, ''hypothesis-driven study'' is a key for big biological data mining, which can reduce the CPU time of data mining and the occupation of computing resources effectively. In other words, it is a key how we can rely on some reasonable hypotheses to guide our big biological data mining by solving 4-V problems efficiently. A remarkable example can be seen from a research study by Yuan and her colleagues [5]. They found that very diverse outputs are often generated when the same gene expression data is analyzed using different algorithms, *i.e.*, low overlap and substantial false positives. The problem results from the extreme heterogeneousness of gene expression data and there is no guarantee that a pure statistical model will solve it. A recent effort was made to present a methodology, aimed to circumvent the limitations of pure statistical models and general gene expression data analysis strategy. The method was based on a simple biological assumption: ''If a number of genes that are conservatively co-expressed emerge as a dynamically-cooperative group across certain biological processes, these genes are most likely functionally closely related with physiological and pathological processes'' [5]. Then, according to this ''hypothesis'', the data mining is just to be converted to finding those gene clusters with strongly cooperative and conservative properties across cancer progression stages.

Computational systems biology plays a central role in big biological data era

There are four key features related with systems biology: integration – the whole biological system is more than the sum of its parts; network – biological function is a phenotype of inherent biological network structure; emergency – new function is emerged through interactions among the elements of the biological system; interference – it implies correlation or coherence between biological molecules in a single biological pathway or between several biological pathways. The nature of biological big data can be summarized as: hierarchy – data is generated at different levels ranging from molecules, cells, tissues to systems; heterogeneous – data is generated using different methods ranging from genetics, physiology, pathology to imaging; complexity – data can be simultaneously recorded in the forms of multi-level information from over thousands of cells or even more; dynamics – biological processes or states change with conditions and over time. It is undeniable that the association study only is too superficial to meet the needs of scientists, and our aspirations are to reveal the driving force or causal relationship among biological elements, which can be used for deciphering the mechanisms of biological processes and diseases, such as cancer, diabetes, and Alzheimer's disease. The main challenge for big data mining then would be how we can achieve a transition from association study to causality study. From this point of view, computational systems biology [6,7] provides a new way for system-wide study and could play a key role in such a transition in big-data era.

The impacts of engineering, cooperation, standardization and pipeline to big data analysis

Because of the nature of big biological data, conducting research in life science to some extent has to change its style in the era of big data, *e.g.*, from academic exploration individually to more cooperative study in systematic, standardized and pipelining ways. The main challenges here could be to establish interoperable databases, make sustainable tools available to the research community, create tool development centers, construct resources and infrastructure, such as cloud computing to serve the huge amount of researches, generate standards, vocabularies and ontologies of big biological data, develop new systems of infrastructure and tools, and obtain buy-in from the scientific community, such as cloud service. Clearly, aforementioned challenges can be solved in a more engineering manner, and a well-designed experiment system matching some systematic, standardizing data processing pipeline will be an important factor for a successful study.

Big-data medicine by dynamical network biomarkers

It is commonly recognized that a complicated living organism cannot be completely appreciated by merely analyzing individual components. Phenotypes and functions of an organism are ultimately determined by interactions between these components or networks in terms of structures and dynamics [2]. Network and dynamics are two key aspects in computational systems biology [6,7–13]. However, majority of traditional

research focuses on the static and statistic properties (*e.g.*, GWAS) of big data, rather than the essential dynamics and networks of life in living organisms. Generally, a disease is a problem resulting not from malfunction of individual molecules but from failure of the relevant system or network, which can be considered as a set of interactions among molecules. Thus, rather than single molecules, the networks are stable forms as biomarkers to reliably characterize complex diseases. The era of big data [14,15] provides great opportunities for predictive, preventive, personalized and participatory (P4) medicine, which is expected to lead to big-data medicine. The study of network and interactions of biological elements rather than biological elements themselves, can capture the previously-unobserved features at the levels of both network (or edges) and dynamics. Therefore, with the demand from both theoretical and clinic aspects, biomarkers are evolving from single molecules (*e.g.*, individual genes) to multiple molecules (*e.g.*, gene set), associated molecules (*e.g.*, molecule network) and dynamical interactive molecules (*e.g.*, dynamical molecule network) due to the availability of big data, in particular, high-dimensional data, which can be categorized as node biomarkers [14,15], network-based biomarkers [16–18], network biomarkers [19,20] and dynamical network biomarkers (DNBs) [21,22], respectively. By exploiting the network information from big data, recent studies on EdgeMarker [14,20] demonstrate that non-differentially expressed genes, which are usually ignored by traditional methods, can be as informative as differentially expressed genes in terms of classifying different biological conditions or phenotypes of samples. By exploiting the dynamical information from big data, a novel biomarker, DNB, was recently developed [22]. In contrast to the disease state detected by traditional biomarkers, DNB is able to identify the pre-disease state before the occurrence or serious deterioration of diseases, which can actually be used to prevent from further disease progression before deteriorating into their irreversible states [21–24]. In other words, by high-dimensional data (such as gene expression, RNA-seq, protein expression, and imaging data), this new type of biomarkers can achieve the early diagnosis of ''pre-disease'' state or ''un-occurring disease'' state, which is a concept raised in ''Yellow Emperor's Canon of Internal Medicine'' (one of the earliest books for Traditional Chinese Medicine) [14].

Acknowledgements

This project was partially supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB13040700), the National Program on Key Basic Research Project (973 Program, Grant No. 2014CB910504) and the National Natural Science Foundation of China (NSFC) (Grant Nos. 61134013, 91130032, 61103075 and 91029301).

References

[1] Gou X, Wang Z, Li N, Qiu F, Xu Z, Yan D, et al. Whole-genome sequencing of six dog breeds from continuous altitudes reveals adaption to high-altitude hypoxia. Genome Res 2014;24:1308–15.
[2] Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 2009;461:272–6.
[3] Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, et al. Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. N Engl J Med 2010;362:1181–91.
[4] Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, et al. Big data: the future of biocuration. Nature 2008;455:47–50.
[5] Yuan L, Ding G, Chen YE, Chen Z, Li Y. A novel strategy for deciphering dynamic conservation of gene expression relationship. J Mol Cell Biol 2012;4:177–9.
[6] Chen L, Wang R, Zhang X. Biomolecular networks: methods and applications in systems biology. Hoboken: John Wiley & Sons; 2009.
[7] Chen L, Wang R, Li C, Aihara K. Modelling biomolecular networks in cells: structures and dynamics. London: Springer-Verlag; 2010.
[8] Song W, Wang J, Yang Y, Jing N, Zhang X, Chen L. Rewiring drug-activated p53-regulatory network from suppressing to promoting tumorigenesis. J Mol Cell Biol 2012;4:197–206.
[9] Zeng T, Wang DC, Wang X, Xu F, Chen L. Prediction of dynamical drug sensitivity and resistance by module network rewiring-analysis based on transcriptional profiling. Drug Resist Update 2014;17:64–76.
[10] Liu B, Yuan Z, Aihara K, Chen L. Reinitiation enhances reliable transcriptional responses in eukaryotes. J R Soc Interface 2014;11:20140326.
[11] Wu FX, Wu L, Wang J, Liu J, Chen L. Transittability of complex networks and its applications to regulatory biomolecular networks. Sci Rep 2014;4:4819.
[12] Zhu H, Rao RS, Zeng T, Chen L. Reconstructing dynamic gene regulatory networks from sample-based transcriptional data. Nucleic Acids Res 2012;40:10657–67.
[13] Ma H, Zhou T, Aihara K, Chen L. Predicting time-series from short-term high-dimensional data. Int J Bifurcat Chaos 2014;24:1430033.
[14] Zeng T, Zhang W, Yu X, Liu X, Li M, Liu R, et al. Edge biomarkers for classification and prediction of phenotypes. Sci China Life Sci 2014;57:1103–14.
[15] Liu R, Wang X, Aihara K, Chen L. Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers. Med Res Rev 2014;34:4555–78.
[16] Ren X, Wang Y, Chen L, Zhang XS, Jin Q. EllipsoidFN: a tool for identifying a heterogeneous set of cancer biomarkers based on gene expressions. Nucleic Acids Res 2013;41:e53.
[17] Zeng T, Zhang CC, Zhang W, Liu R, Liu J, Chen L. Deciphering early development of complex diseases by progressive module network. Methods 2014;67:334–43.
[18] Wen Z, Liu ZP, Liu Z, Zhang Y, Chen L. An integrated approach to identify causal network modules of complex diseases with application to colorectal cancer. J Am Med Inform Assoc 2013;20:659–67.
[19] Yu X, Li G, Chen L. Prediction and early diagnosis of complex diseases by edge-network. Bioinformatics 2014;30:852–9.
[20] Zhang W, Zeng T, Chen L. EdgeMarker: identifying differentially correlated molecule pairs as edge-biomarkers. J Theor Biol 2014. http://dx.doi.org/10.1016/j.jtbi.2014.05.041.
[21] Liu R, Yu X, Liu X, Xu D, Aihara K, Chen L. Identifying critical transitions of complex diseases based on a single sample. Bioinformatics 2014;30:1579–86.
[22] Chen L, Liu R, Liu ZP, Li M, Aihara K. Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. Sci Rep 2012;2:342.
[23] Teschendorff AE, Liu X, Caren H, Pollard SM, Beck S, Widschwendter M, et al. The dynamics of DNA methylation covariation patterns in carcinogenesis. PLoS Comput Biol 2014;10:e1003709.
[24] Li M, Zeng T, Liu R, Chen L. Detecting tissue-specific early-warning signals for complex diseases based on dynamical network biomarkers: study of type-2 diabetes by cross-tissue analysis. Brief Bioinform 2013;15:229–43.