Precision Clinical Medicine, 4(2), 2021, 80–84

10.1093/pcmedi/pbab007
Advance Access Publication Date: 16 April 2021
Perspective

PERSPECTIVE

The fourth scientific discovery paradigm for precision medicine and healthcare: Challenges ahead

Li Shen[1], Jinwei Bai[2], Jiao Wang[1] and Bairong Shen[1,*]

[1]Institutes for Systems Genetics, Frontiers Science Center for Disease-related Molecular Network, West China Hospital, Sichuan University, Chengdu 610041, China
[2]Library of West-China Hospital, Sichuan University, Chengdu 610041, China

*Correspondence: Bairong Shen, bairong.shen@scu.edu.cn

Abstract

With the progression of modern information techniques, such as next generation sequencing (NGS), Internet of Everything (IoE) based smart sensors, and artificial intelligence algorithms, data-intensive research and applications are emerging as the fourth paradigm for scientific discovery. However, we face many challenges to practical application of this paradigm. In this article, 10 challenges to data-intensive discovery and applications in precision medicine and healthcare are summarized and the future perspectives on next generation medicine are discussed.

Key words: data-intensive scientific discovery; the fourth paradigm; biomedical data diversity; precision medicine and healthcare

Introduction

The scientific discovery paradigm (SDP) provides a mature and routine framework for asking scientific questions, developing methods or strategies to answer such questions, and also includes ways to explain the experimental results or the observed data. In the last two decades, the SDP in life sciences has shifted fast, especially with progression of the human genome project. A paradigm shift, sometimes also called a 'scientific revolution', occurs when the existing paradigm cannot efficiently solve the challenges faced and a new paradigm is needed to deal with the challenges. For example, in bioinformatics, a small paradigm shift we refer to here as the bioinformatics scientific research model (SRM), emerged with accumulation of DNA sequencing data. Since then, a batch of new genes has been discovered by pattern identification with models trained using known gene structure patterns. Well-known bioinformatics tools and databases including CLUSTAL W,[1] MEGA,[2] PDB[3] *etc.*, were developed within the bioinformatics SRM. Traditional experimental paradigms can only discover new genes one by one through time-consuming and labor-intensive methods. Complex biological systems, however, often function by interactions between many genes, proteins, or other components via pathways, modules, or networks. Bioinformatics has contributed to acceleration in life sciences by fast, efficient, high throughput, and computational methods, enabling investigation of biological and medical problems at systemic levels. The microarray, yeast two-hybrid assay, and evolutionary modeling promoted the paradigm shifting to systems biology, which

Received: 29 March 2021; **Revised:** 13 April 2021; **Accepted:** 13 April 2021

aimed to reconstruct the interaction or synergistic network to explain emergence properties in a system. Systems biological tools such as gene ontology,[4] KEGG[5] and Cytoscape,[6] *etc.*, were then developed and widely used. But for clinical translation, genome functional discoveries cannot be applied directly to treatment of patients because of heterogeneities among diseases and patients. Cell-line or animal-model based biological findings need to be validated with patient samples before clinical applications. Therefore, translational and precision medicine SRMs have been proposed to integrate genotypic and phenotypic information for personalized prediction and treatment of diseases.[7,8]

Although paradigms in life sciences have shifted frequently in the past 20 years, data accumulation is always the driving force for scientific revolution. In the future, data will remain one of the most essential parts for successful scientific paradigm shifts; however, the quality, quantity, and diversity of biomedical data will pose key challenges for our future precision medicine and healthcare.

The fourth paradigm: data-intensive scientific discovery

As shown in Fig. 1A, paradigm shifts in the life sciences over the last two decades present a very salient characteristic, i.e. more and more data are needed for scientific discovery in life sciences. The bioinformatics SRM emerged with progression of the human genome project. As more DNA sequencing data were accumulated, gene structures in the genome could be compared and the DNA string patterns specific to protein coding genes, non-coding RNAs, and the regulatory elements therefore could be identified for prediction of new genes. Since then, many databases have been established for investigations of biological questions. Compared with traditional biostatistics discipline, which can do nothing with a single DNA or protein sequence, bioinformatics tools provide researchers with enormous DNA information resources for ortholog or paralog screening, phylogenetic tree construction, 3D structure modeling, functional specificity estimation, and so on.[9–12] For the systems biology SRM, the first step was reconstruction of the biological network by top-down or bottom-up strategies, where multiple points or correlated data are demanded to infer the interactions between nodes and the structures of networks or systems.[13–15] The translational medicine and precision medicine SRMs further need clinical and personalized data for deep phenotyping and personalized diagnosis and treatment of patients.[16–18]

Figure 1B displays the four SDPs, with traditional SDPs including experimental, theoretical, and computational ones. Compared with the three traditional paradigms, the fourth paradigm, i.e. data-intensive SDP, has emerged in modern technologies, including high throughput sequencing, cloud computing, smart sensors, digital medicine, Internet of Everything, next generation artificial intelligence, and so on.[19,20] Especially for complex and heterogeneous systems such as ecological systems, cancer, and many chronic diseases, it is difficult to describe and understand such complexity using simple rules or theories. The fourth paradigm will be an important complementary solution to the other three paradigms. Complex and chronic diseases are often caused by interactions of many factors such as genetic events, lifestyles, and environmental factors. The fourth paradigm provides a way to deal with personalized diagnosis and treatment with huge amounts of patient information by calculating similarities between the query patients and profiles in the targeted databases. However, the prerequisite for the success of this paradigm is that the data accumulated are 'big' enough to cover all possibilities. Furthermore, the science based on this big data needs new algorithms for discoveries of new rules, principles, key players, and mechanisms for the understanding and controlling of the life systems.

Biomedical data diversity and standardization

As the fourth paradigm for scientific research is characterized by the intensity of data, high quality data accumulation will be an essential step for data-driven personalized and precision medicine. Biomedical data such as data at molecular, cellular, tissue, individual, and population levels, are usually diverse and heterogeneous. These could be basic scientific data from laboratories, or real world clinical or health status data, and the data could be dynamic, evolutionary, and spatiotemporal. The following three challenges will be faced in application of the fourth paradigm to precision medicine practice.

Challenge 1: Data standardization for communication

To collect big biomedical data, the data formats, terminologies, and relationships should be standardized.[21] Clinical scientists need to share their data and information, especially for rare disease description, to improve the diversity and representativeness of the collected data. The biomedical data are huge considering the disease types, personalized genetics, dynamic lifestyles, environmental factors, as well as the synonyms and complex relationships.

Challenge 2: Data sharing and privacy preservation

It has been reported that several genes could be combined with known personal information to re-identify personal features, such as 3D facial reconstruction, inference of voices, and family names. To protect personal privacy, the clinical information need to be desensitized, or

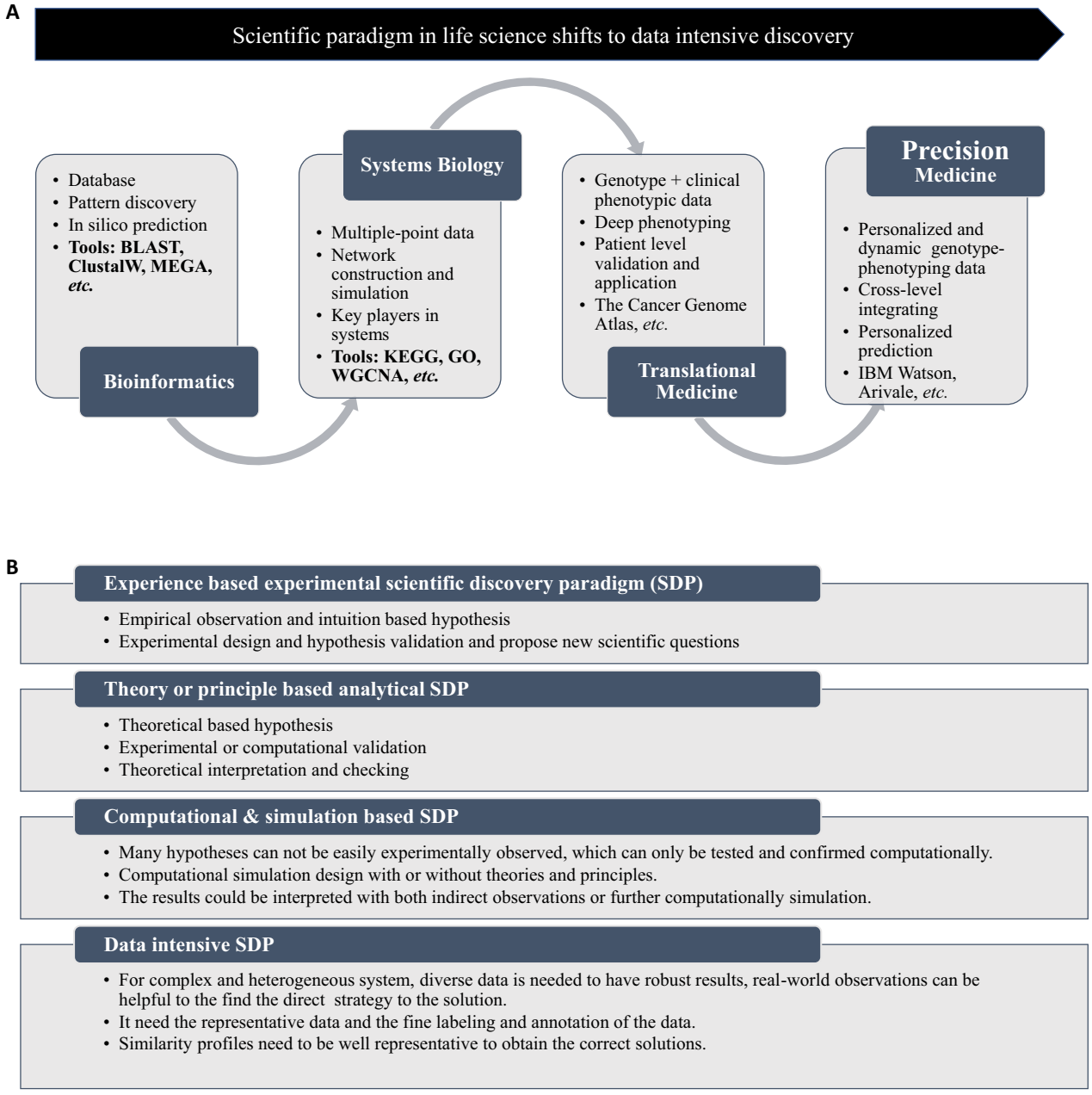

Figure 1. (A) Scientific paradigm shifts in last two decades. (B) Four scientific discovery paradigms.

perhaps even 'noise' introduced to the data via differential privacy to secure multi-party computation and sharing.

Challenge 3: Data measurability and spatiotemporal signals

Spatiotemporal molecular medicine is becoming the new discipline for investigation of dynamic and evolutionary human health status. Although modern smart sensor technologies can be applied to collection of physiological information and molecular level information, it is not easy to collect dynamic signals and data. Investigations such as gene expression and microbiota ecological dynamics are still very challenging.

The explainable model and actionable key data

To understand the complexity and identify specific patterns hidden in the big data, personalized models are necessary. Data-intensive algorithms, such as deep learning models, need to be explainable and actionable for precision medicine and healthcare.[22]

Challenge 4: Phenotype plasticity and model robustness

The health or disease status is determined by complex interactions of many genetic, lifestyle, and environmental factors.[23] One disease phenotype could be associated with many genotypic factors, and the methods or solutions to transfer the disease to health status are not unique. The model constructions are not the same as traditional one-to-one mode but should be one-to-N mode, considering the plasticity of phenotype.

Challenge 5: Explainable artificial intelligence (XAI) and precision medicine practice

Most traditional machine learning (ML) algorithms are models for classification, somewhat 'black boxes' in that their mechanisms and explanations remain unknown. It is difficult to apply these ML models to design of personalized treatment. XAI will be helpful to 'open the black box', facilitating trust of patients or clinical doctors in use of AI predictions in precision medicine practice.

Challenge 6: Clinical observation and real world data-driven scientific discovery

For precision medicine and healthcare, clinical observation/questions and real world data are the two main resources in hospitals, which cannot be obtained from laboratories. There remains a challenge to propose good clinical questions for scientific investigation as these require insights and experiences from both clinical and basic science.

Challenge 7: Experimental or computational verifiability

Discoveries of biomarkers, drug targets, and other key players based on the fourth paradigm need to be verified and validated with experiments (including clinical practice) or computational-aided simulations.[24] These could require further improvement and re-validation before they can be safely and widely applied to medicine and healthcare practice.

Challenge 8: From data and knowledge to general principles

Although an XAI-based model can explain an observation, further exploration in data-intensive research will be required to discover the general principles underlying the observed patterns. The principles can then be used to guide design of the strategies for better treatment of diseases.[25] The fourth SDP is a complement to the other three paradigms and these can be integrated with each other to accelerate discovery in medicine.

Translational application and cross-disciplinary education

Even if all scientific discoveries are aimed at applications, we are still short of qualified persons for the fourth SDP practice. The last two challenges concern application and education.

Challenge 9: Smart application of data-intensive SDP to healthcare

As we have limited medical resources to combat widespread chronic diseases, data-intensive scientific discovery could be transferred to smart patient self-administration, especially chronic disease monitoring and controlling. Knowledge-guided chatbots could offer a way to improve the quality of diagnosis, outpatient consultation, and referral as well as treatment.[26]

Challenge 10: Education and training for data-intensive SDP

To overcome the nine challenges in the life sciences as stated above, we need well-educated and trained clinicians and scientists. The next generation of medical doctors, researchers, and even patients, should be equipped with knowledge on data standardization, data security, knowledgebases, algorithms and models, *etc.*, for cross-disciplinary studies using data-intensive SDP.

Conclusions and future perspectives

The first three SDPs have been applied in most scientific fields, including physics, chemistry, engineering, *etc.* The fourth SDP is emerging and will evolve with big data science and technology. Data diversity and heterogeneity remain two main challenges in the life sciences. Disease profiles and data spaces for biomedical data are very big and still expanding with evolution of interactions between genetics, lifestyles, and environments.

Two well-known efforts, IBM Watson and Arivale's wellness project,[27,28] have reported failures in healthcare, the main reason being that the data collected for their artificial intelligent modeling or analytics are not representative when faced with complex and personalized application. The healthcare industries need more well-labeled data, knowledge-guided models,[29] and experienced human resources. With integration of the four paradigms, the challenges for the new paradigm applications, on the other hand, are also the opportunities for efforts to develop ontologies for standardization of data, to build knowledge databases for explainable artificial intelligence modeling, and to dig into the genotyping-phenotyping relationship for precision applications for precision medicine and healthcare practice.

Acknowledgements

This work was supported by the regional innovation cooperation between Sichuan and Guangxi Provinces (Grant No. 2020YFQ0019) and the National Natural Science Foundation of China (Grant No. 32070671).

Conflict of interest

None declared. In addition, as an Editorial Board Member of ***Precision Clinical Medicine***, the corresponding author Bairong Shen was blinded from reviewing and making decision on this manuscript.

References

1. Larkin MA, Blackshields G, Brown NP, *et al*. Clustal W and clustal X version 2.0. *Bioinformatics* 2007;**23**:2947–8. doi:10.1093/bioinformatics/btm404.
2. Kumar S, Stecher G, Tamura K. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol* 2016;**33**:1870–4. doi:10.1093/molbev/msw054.
3. Burley SK, Bhikadiya C, Bi C, *et al*. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res* 2021;**49**:D437–d51. doi:10.1093/nar/gkaa1038.
4. Ashburner M, Ball CA, Blake JA, *et al*. Gene Ontology: tool for the unification of biology. *Nat Genet* 2000;**25**:25–9. doi:10.1038/75556.
5. Kanehisa M, Furumichi M, Sato Y, *et al*. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res* 2021;**49**:D545–d51. doi:10.1093/nar/gkaa970.
6. Shannon P, Markiel A, Ozier O, *et al*. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504. doi:10.1101/gr.1239303.
7. Day M, Rutkowski JL, Feuerstein GZ. Translational medicine—A paradigm shift in modern drug discovery and development: The role of biomarkers. *Adv Exp Med Biol* 2009;**655** :1–12. doi:10.1007/978-1-4419-1132-2_1.
8. Loscalzo J. Precision medicine a new paradigm for diagnosis and management of hypertension? *Circ Res* 2019;**124**:987–9. doi:10.1161/circresaha.119.314403.
9. Shen B, Vihinen M. RankViaContact: ranking and visualization of amino acid contacts. *Bioinformatics* 2003;**19**:2161–2. doi:10.1093/bioinformatics/btg293.
10. Shen B, Vihinen M. Conservation and covariance in PH domain sequences: physicochemical profile and information theoretical analysis of XLA-causing mutations in the Btk PH domain. *Protein Eng Des Sel* 2004;**17**:267–76. doi:10.1093/protein/gzh030.
11. Shen B, Bai J, Vihinen M. Physicochemical feature-based classification of amino acid mutations. *Protein Eng Des Sel* 2008;**21**:37–44. doi:10.1093/protein/gzm084.
12. Yan W, Hu G, Liang Z, *et al*. Node-Weighted amino acid network strategy for characterization and identification of protein functional residues. *J Chem Inf Model* 2018;**58**:2024–32. doi:10.1021/acs.jcim.8b00146.
13. Tang Y, Yan W, Chen J, *et al*. Identification of novel microRNA regulatory pathways associated with heterogeneous prostate cancer. *BMC Syst Biol* 2013;**7**:S6. doi:10.1186/1752-0509-7-s3-s6.
14. Wang Y, Chen J, Li Q, *et al*. Identifying novel prostate cancer associated pathways based on integrative microarray data analysis. *Comput Biol Chem* 2011;**35**:151–8. doi:10.1016/j.compbiolchem.2011.04.003.
15. Chen J, Sun M, Shen B. Deciphering oncogenic drivers: from single genes to integrated pathways. *Brief Bioinform* 2015;**16**:413–28. doi:10.1093/bib/bbu039.
16. Qian F, Guo J, Jiang Z, *et al*. Translational Bioinformatics for Cholangiocarcinoma: Opportunities and Challenges. *Int J Biol Sci* 2018;**14**:920–9. doi:10.7150/ijbs.24622.
17. Qi X, Lin Y, Chen J, *et al*. The landscape of emerging ceRNA crosstalks in colorectal cancer: Systems biological perspectives and translational applications. *Clin Transl Med* 2020;**10**:e153. doi:10.1002/ctm2.153.
18. Qian F, Wang J, Wang Y, *et al*. MiR-378a-3p as a putative biomarker for hepatocellular carcinoma diagnosis and prognosis: Computational screening with experimental validation. *Clin Transl Med* 2021;**11**:e307. doi:10.1002/ctm2.307.
19. Collins JP. The fourth paradigm Data-Intensive scientific discovery. *Science* 2010;**327**:1455–6. doi:10.1126/science.1186123.
20. Nielsen M. The fourth paradigm: Data-Intensive scientific discovery. *Nature* 2009;**462**:722–3. doi:10.1038/462722a.
21. Chen Y, Yu C, Liu X, *et al*. PCLiON: An ontology for data standardization and sharing of prostate cancer associated lifestyles. *Int J Med Inform* 2021;**145**:104332. doi:10.1016/j.ijmedinf.2020.104332.
22. Shen B, Lin Y, Bi C, *et al*. Translational informatics for Parkinson's disease: from big biomedical data to small actionable alterations. *Genomics Proteomics Bioinformatics* 2019;**17**:415–29. doi:10.1016/j.gpb.2018.10.007.
23. Shen L, Ye B, Sun H, *et al*. Systems health: A transition from disease management toward health promotion. *Adv Exp Med Biol* 2017;**1028**:149–64. doi:10.1007/978-981-10-6041-0_9.
24. Lin Y, Qian F, Shen L, *et al*. Computer-aided biomarker discovery for precision medicine: data resources, models and applications. *Brief Bioinform* 2019;**20**:952–75. doi:10.1093/bib/bbx158.
25. Zhang W, Landback P, Gschwend AR, *et al*. New genes drive the evolution of gene interaction networks in the human and mouse genomes. *Genome Biol* 2015;**16**:202. doi:10.1186/s13059-015-0772-4.
26. Bai J, Shen L, Sun H, *et al*. Physiological informatics: Collection and analyses of data from wearable sensors and smartphone for healthcare. *Adv Exp Med Biol* 2017;**1028**:17–37. doi:10.1007/978-981-10-6041-0_2.
27. Fiala C, Diamandis EP. The outcomes of scientific debates should be published: The Arivale story. *J Appl Lab Med* 2020;**5**:1070–5. doi:10.1093/jalm/jfaa110.
28. Jiang F, Jiang Y, Zhi H, *et al*. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2017;**2**:230–43. doi:10.1136/svn-2017-000101.
29. Shen L, Lin Y, Sun Z, *et al*. Knowledge-Guided bioinformatics model for identifying autism spectrum disorder diagnostic MicroRNA biomarkers. *Sci Rep* 2016;**6**:39663. doi:10.1038/srep39663.