Computational Personalized Medicine in Cancer Research in the -Omics Data Era

Duc-Hau Le, Quynh Diep Nguyen

School of Computer Science and Engineering, Thuyloi University, Hanoi, Vietnam

Correspondence: Duc-Hau Le, hauldhut@gmail.com

Communication: received 29 October 2019, revised 16 April 2020, accepted 6 May 2020

Digital Object Identifier: 10.32913/mic-ict-research.v2020.n1.899

Abstract: Omics data (e.g., genomics, transcriptomics, proteomics, epigenomics, etc . . .) generated from high-throughput next-generation sequencers in the big human genome, and cancer genome projects have changed the way to study personalized medicine. In the future, personalized medicine will not be limited to diagnosis and treatment based on a few known disease-associated mutations on some genes, but will rely on whole molecular characteristics of patients by integrating their –omics data. In this study, we draw a big picture of personalized medicine research in cancer research of the –omics data era, including –omics databases, challenges of data fusion to solve two major problems in personalized medicine, i.e., personalized diagnosis and treatment. These problems are approached as patient stratification and drug response prediction based on the –omics data by computational methods.

Keywords: *Personalized medicine, omics data integration, patient stratification, drug response prediction, computational method.*

I. Introduction

In biomedicine, with high-throughput technologies such as next-generation sequencing (NGS), we are witnessing the exponential growth of biomedical data (e.g., –omics data). To date, it is widely accepted for most diseases that their genetic etiology cannot be explained with one or a few genes/variants, but thousands of whole-genome/variome (i.e., genomics/variomics data). Besides, it is still not enough if the only genome is considered, but also transcriptome (e.g., transcriptomics data), because this data reflects how much change in a genome can affect the expression of a gene in different tissues. More importantly, how the gene expression translated into proteins is a critical issue since protein functions decide how a cell works. Therefore, besides genomics and transcriptomics, proteomics data should also be considered when studying the disease etiologies. It should be noted that the central dogma in biology states protein synthesis as DNA→RNA→Protein. However, only $1 \rightarrow 2\%$ of DNA directly encodes for proteins. The major part ($98 \rightarrow 99\%$ of DNA) are non-coding regions, which do not directly synthesize the proteins but regulate the synthesis process as well as other biological processes and cellular functions. Therefore, epigenome (i.e., epigenomics data) should be taken into account for the disease etiologies. Moreover, cells function under chemical reactions between metabolites; thus, metabolome (i.e., metabolomics data) is also critical. Finally, cellular components such as genes, proteins, and metabolites,. . . do not function alone, but they interoperate in the form of biological networks to make the cells robust against external stimulus or mutations; thus, interactomics data is a final "piece of cake" for a big picture of the –omics data.

In this study, we draw a big picture of personalized medicine in cancer research of the –omics data era. Personalized medicine is a potential approach in biomedicine with the goal is to find the right treatment (i.e., right drug, right dose, and right time) for each patient based on their biological characteristics. Different patients have different biological characteristics in terms of –omics data; thus, the development of the right methods for diagnosis and treatment is a challenging task due to the diversity and complexity of the –omics data. However, to fully understand the biological characteristics of each patient, their –omics data should be combined together in the right ways. Therefore, besides introducing –omics databases, we also discuss how to integrate the –omics data for solving the two main problems (i.e., diagnosis and treatment) in personalized medicine.

II. The –omics data

With the rapid development of NGS technologies, the cost for sequencing is significantly going down, and the –omics data are growing exponentially. Biomedical research nowadays is not only based on the genomics data, but also on other –omics data such as transcriptomics,

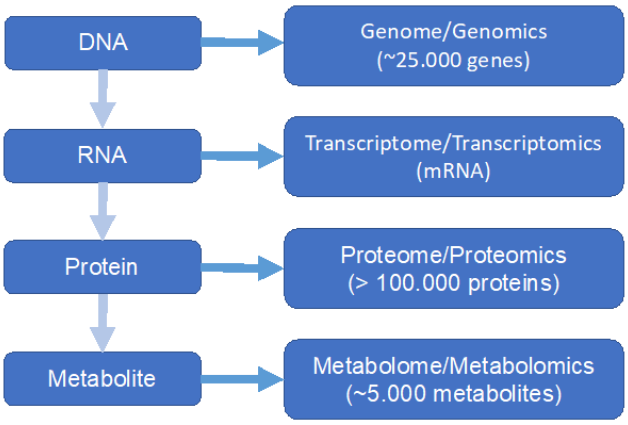

Figure 1. The –omics layers.

proteomics, and metabolomics [1]. Figure 1 shows the central dogma in biology with the corresponding –omics data layers.

A genome is understood as whole genetic information on DNA of a species. The human genome contains about three billion base pairs (bp), which encode for about 25,000 genes. Coding regions occupy approximately $1 \rightarrow 2\%$ of the genome, and the remaining is non-coding regions [2]. Change in a genome (variants) can happen at a single point, such as SNP (Single Nucleotide Polymorphism) or short insertion and deletion (short indel). Those changes can be detected by traditional techniques such as Sanger or modern ones such as microarray and NGS. If the changes happen in protein-coding regions of DNA, they can be synonymous, missense, nonsense, and frameshift. When they occur in the remaining regions, those changes may affect the level of gene expression (e.g., up and down) or RNA splicing process. Besides such small changes, larger changes are also observed, such as structure variants (SV), where the insertions and deletions are found in large regions (100 – 1000 bp). In addition, copy number variant (CNV), inversion, and translocation are also found. For those changes, the microarray or NGS technologies are required for the detection.

A transcriptome includes all RNAs (coding RNAs such as mRNA and non-coding RNAs) in a cell. RNA reflects how much genetic information (i.e., embedded in genes) from DNA is expressed. To measure the level of gene expression, microarray and RNA-seq are widely used. The microarray technology is used to detect known transcripts and isoforms with established and standard pipelines/tools. Meanwhile, the RNA-Seq, as a modern technology, can be used to detect de novo transcripts and isoforms as well as structural variants. However, it also yields big and complex data, thus requires more sophisticated pipelines/tools.

A proteome is all proteins in a cell at a specific time of development. Two main problems that are widely studied for proteins are protein structure and protein interaction network. A total of 20 types of amino acids can be translated by the protein synthesis process. In the structural view, a protein is a sequence of amino acids (aka polypeptide), which can be in different shapes such as linear (level 1) and fold (level 4). The structural features decide the protein's functions. For the second problem, physical interactions between proteins by chemical bonds form the protein interaction networks or protein complexes. These also decide the functions of cells.

A metabolome is mentioned as small chemical compounds such as endogenous and exogenous metabolites.

Figure 1 shows that, for humans, there are about 25,000 genes in the genome, more than 100,000 proteins, and 5,000 metabolites [3]. Those data generated from large-scale biomedical projects such as human genomes, cancer genomes, and cell lines have yielded big–omics databases. In the next section, we are going to introduce some of such projects briefly.

III. The –omics database

The –omics databases are generated from large-scale projects on the human genome, such as UK's 100,000 Genomes [4] and GenomeAsia 100K [5]. In addition, disease-specific projects such as those for cancer genome have largely contributed their data to the research community. Here are some critical projects which are widely used in cancer research.

The Center for Cancer Genomics of National Cancer Institute in the US had launched a number of big projects on cancers such as TCGA (The Cancer Genome Atlas) [6], CGCI (Cancer Genome Characterization Initiative) and TARGET (Therapeutically Applicable Research to Generate Effective Treatments). These projects are now managed by Genomic Data Commons data portal (GDC, https://portal.gdc.cancer.gov) to foster precise medicine in cancer. The portal has deposited data for more than 40 cancer projects on 61 organs, with 32,555 samples [7]. A number of –omics data has been generated from the projects such as single point somatic/germline mutation, structural somatic mutations, gene/microRNA expression by microarray, gene expression by RNA-Seq, and DNA methylation data. Besides, the Catalogue Of Somatic Mutations In Cancer project (COSMIC) is the largest and most comprehensive database of cancer somatic mutations on human [8]. Major data of COSMIC is collected and curated by experts. It also provides 3D structures of the mutations as well as related genes. In addition, more than 1,000 mutation profiles of cell lines are also accumulated. Finally, the

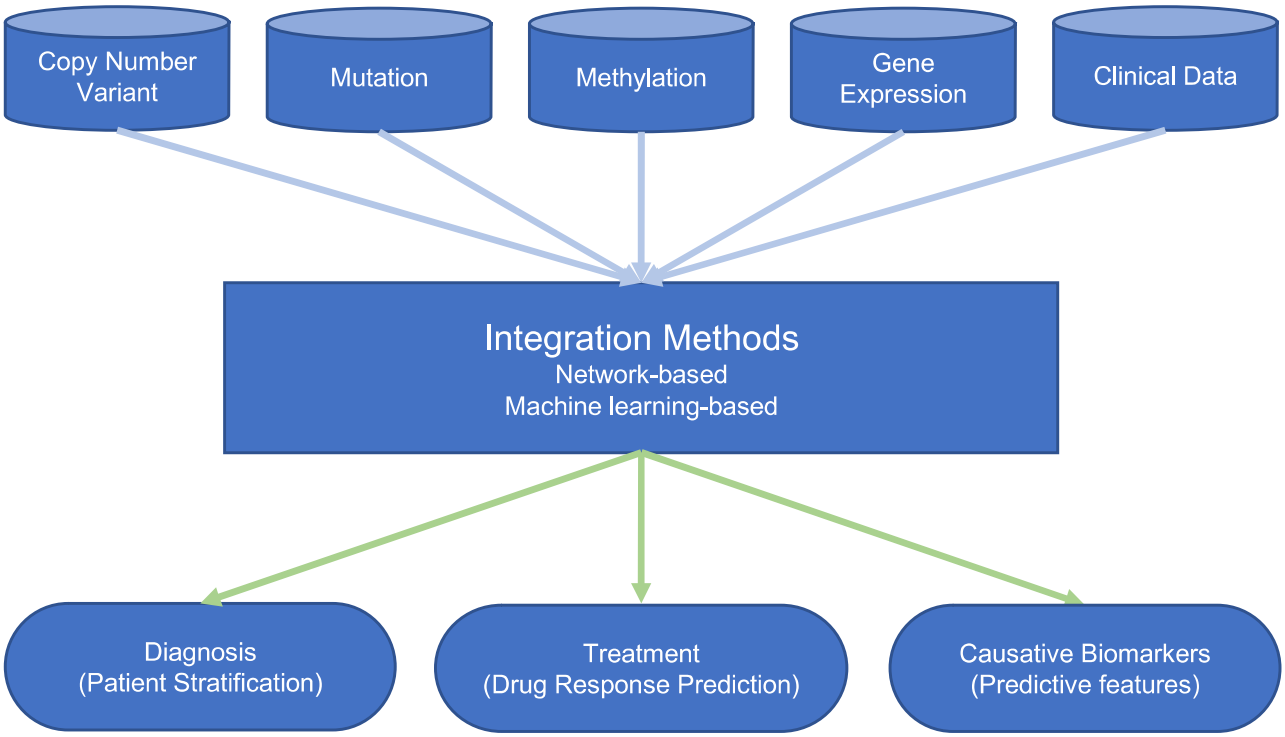

Figure 2. Integration of the –omics data in personalized medicine problems.

International Cancer Genome Consortium (ICGC) aimed to define the genomes of 25,000 primary untreated cancers, as its data portal [9] makes global genomic data sharing for cancer possible, providing the international community with comprehensive genomic data for many cancer types.

In parallel with the abovementioned projects on cancer tumors, projects on testing drugs/compounds on cell lines were also launched. The outcome of those projects are databases of the cell line's biological characteristics and responses of the tested drugs/compounds. These data help foster research on pharmacogenomics, which is the most promising direction in personalized medicine. Two typical projects are Cancer Cell Line Encyclopedia (CCLE) [10] and Genomics of Drug Sensitivity in Cancer (GDSC) [11]. CCLE aims to identify –omics and pharmacological properties of cancer cell lines, then developing analysis tools for predicting drug responses based on the –omics data. A potential application of this is to stratify patients based on their –omics and pharmacological characteristics. CCLE provides the –omics data including single point, indel, and structural mutations and the gene expression of more than 1,000 cancer cell lines as well as the pharmacological properties of 24 drugs. Similarly, GDSC freely published those data for about 1,000 cancer cell lines and 265 drugs/compounds.

IV. Integration of the –omics data

It is obvious that the future of biomedical research will mainly rely on the –omics data because the biological characteristics of patients are fully considered. Therefore, with the assumption that "More is better", recent biomedical research on the –omics data is trying to combine as many –omics data and other clinical data as possible to solve the two major problems in personalized medicine, i.e., diagnosis (aka patient stratification) and treatment (aka drug response prediction) (see next section) [1, 3, 12]. In addition, the identification of causative biomarkers (aka predictive features) is also addressed (Figure 2).

Data integration is approached by machine learning-based [13] network-based [14] methods. The machine learning-based methods rely on unsupervised and supervised techniques. The typical unsupervised methods, such as matrix factorization, aim at subtyping cell lines using the –omics data with the final goal is stratifying patients. Meanwhile, with supervised methods, labelled data (e.g., patients/cell lines with known drug responses or subgroups) is needed for training prediction models. The trained models are then used to predict for patients/cell lines with unknown drug responses and subgroups. In parallel, the networkbased methods often rely on the functional relationships among biomedical entities in the same or different

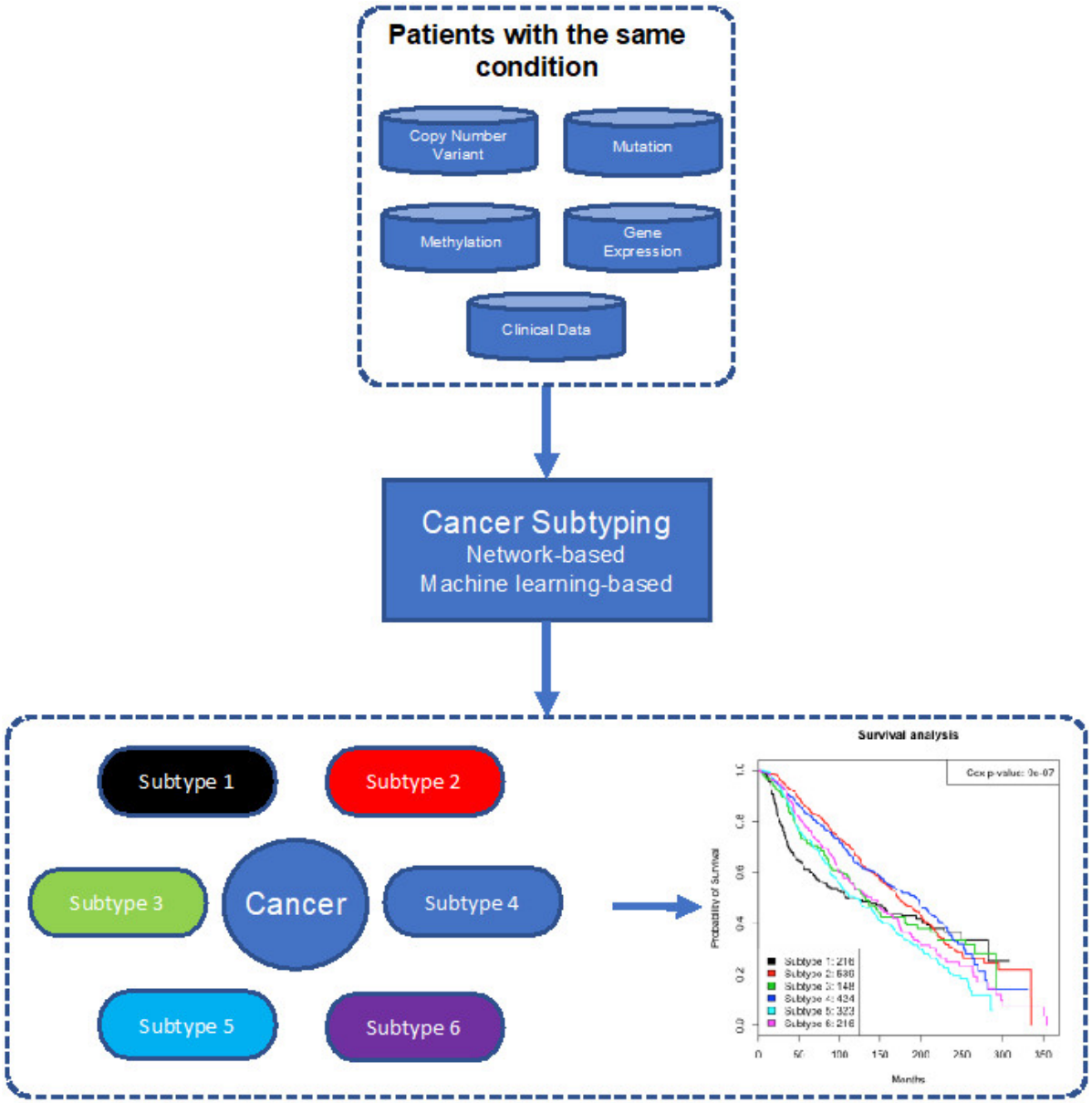

Figure 3. Patient stratification using –omics data.

–omics layers [14].

Along with the advantages of integration, there are also some challenges in integrating the whole genome, transcriptome, and proteome data are also faced with. An obvious problem is "highdimensional data" since thousands to millions of genes/proteins and variants must be investigated. Therefore, techniques in dimension reduction [15] or feature selection [16] have been proposed to select predictive features. Those features are also important biomarkers for diseases. In contrast with the high dimensionality, the number of samples is relatively small (i.e., often thousands). Thus the most challenging problem when dealing with the integration of largescale –omics data is "small n, large p" (a small number of sample, but high dimension). This raises a lot of problems requiring computational solutions [17].

In summary, the data integration aims to systematically assess the biological characteristics of patients; thus, the application of identifying cancer genes [18] and cancer diagnosis [19] in personalized medicine is approached more precisely. In the next sections, we introduce how the two major problems in computational personalized medicine, i.e., diagnosis (patient stratification) and treatment (drug response prediction), are approached using the –omics data.

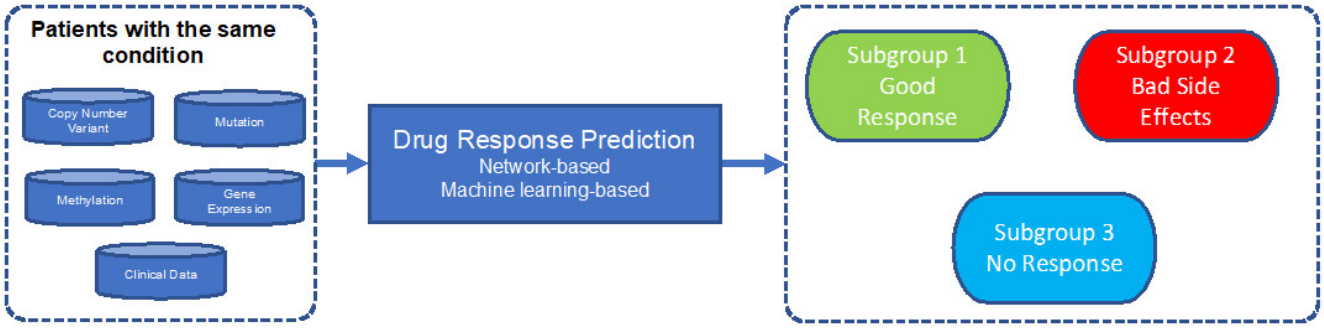

Figure 4. Drug response prediction using –omics data.

V. Patient Stratification Using –omics Data

In this section, we introduce how to stratify patients using the –omics data. Cancer is selected for study as it is a complex disease. The genetic etiologies are also diversified among different cancers/subtypes of the same cancer. Even cells of the same cancer tumor may also act differently (Figure 3). The patient stratification has been studied based on molecular subtyping for some types of cancer, such as bladder urothelial carcinoma [20] and breast cancer [21].

As the abovementioned of the rapid growth of the –omics data, this fosters the development of computational methods for patient stratification based on molecular profiles and towards clinical application [22]. The problem is generally solved by the following four steps: i) data preprocessing, ii) clustering, iii) supervised classification, and iv) subtype characterization. For the first step, raw –omics data from high-throughput sequencers or genotyper are qualifiedly controlled, standardized, and transformed into the right formats for the next steps. Then, the data clustering is performed based on either genes/proteins or samples. If these two types of data are used simultaneously, then biclustering techniques are used. In addition, triclustering techniques are used if other factors, such as time, are considered. In the third step, supervised classification techniques are used to determine which subtype/subgroup a patient belongs to. Finally, the identified subtypes are characterized by causative biomarkers, involved pathways, and survival time.

Two main approaches have been proposed for this problem, including i) machine learningbased and ii) network-based methods. A common point of the two approaches is that they both rely on the similarity between samples (i.e., the similarity of patient's molecular profiles). The similarity can be represented as similarity matrices in machine learning-based methods or similarity networks in network-based methods. By representing as similarity matrices, clustering techniques on a matrix such as matrix factorization are widely used [23, 24]. For network-based methods, network clustering techniques such as module/community detection are often used. Those techniques can be applied to different types of biological networks such as patient similarity network [25], protein interaction network [26], and gene network [27]. Recent methods have relied on both matrix and network clustering techniques [28] or matrix clustering techniques but guided by a network [23, 29].

Although many methods have been proposed, a number of tools have been developed for the patient stratification problem. However, different methods could lead to different subtypes. This is caused by the complexity of the –omics data (i.e., small n, large p). In addition, the cancer tumor is heterogeneous; thus, the molecular profiles (i.e., the –omics data) are different for even cells derived from the same cancer types.

VI. Drug Response Prediction Using –omics Data

Besides the requirement of precisely stratifying patients based on their –omics data, personalized medicine also requires selecting the right drugs for the right treatment for each patient. Patients with the same conditions/cancer types may respond differently to the same drugs due to their different molecular profiles. Thus, the prediction of drug response using the –omics data is an important step for selecting the right drugs (Figure 4).

Similar to the patient stratification problem, many methods have also been proposed for predicting drug response for patients/cell lines [30]. The drug response is measured by dose level to inhibit 50% of the disease's bioactivity (IC50), or they are under the dose-response curve (AUC). They are both continuous values. Thus, the drug response prediction is often approached by regression techniques. However, response values can be categorized into some levels, such as good response, no response, and bad side effects (Figure 4); thus, it can be formulated as a classification problem. The main difference between the two main

problems in personalized medicine is that the patient stratification is usually based on tumor data from tumor/patient-based projects such as TCGA. Meanwhile, the drug response prediction uses cell line and drug response data from drug trial projects such as CCLE and GDSC.

Generally, machine learning- and network-based methods are often proposed for the drug response prediction. Network-based methods are usually based on similarity networks of drugs and cell lines and local [31] or global [32] graph traversal algorithms. In contrast to a few network-based methods, many machine learning-based methods have been proposed for the drug response prediction problem. Indeed, a challenge was organized for research groups over the world [33]. Interestingly, the winner over 44 submissions is a method integrating the –omics data (including single point mutation, structural mutation, gene expression by microarray and RNA-Seq technologies, methylation, and protein data) using multiple kernel learning technique [34]. Other methods also show that response prediction for multiple drugs simultaneously achieve better performance than that for a single drug, because functional and structural similarity among drugs is taken into account [34, 35]. In addition, gene expression data is more dominant than the others [33]. Finally, until now, computational methods for the drug response prediction have been proposed mostly for cancer cell lines. Thus to translate them to clinical application, a recent method has built the prediction model using the data from cell lines in GDSC, then use the built model for predicting drug response for patients in TCGA [36].

VII. Conclusions

Nowadays, the rapid development of high-throughput technologies and large-scale genome projects have generated a large amount of the –omics data (i.e., the –omics era). This has changed the ways to computationally approach the problems in personalized medicine. To fully understand the biological characteristics of patients, their molecular profiles at the –ome scale has been studied. Thus, the –omics data has been integrated into computational methods to solve the problems in personalized medicine. The two major problems in medicine (i.e., diagnosis and treatment) are formulated as two problems in computational space (i.e., patient stratification and drug response prediction, respectively). Although current studies of the two problems target different objects, i.e., the patient stratification mainly focuses on patient data from the patient/tumor projects; meanwhile, the drug response prediction mostly works with artificial patients/tumors (i.e., cell lines). However, they are both personalized based on molecular profiles of each patient/tumor/cell line. Integration of the –omics data algorithmically faces with the "small n, large p" problem. The object (i.e., cancer) itself is a complex disease, which is heterogeneous between cancer types and even cells in the same tumor. In addition, unexpected changes in characteristics of cell lines during culture may limit the translation of research results on cell lines to patients. Fortunately, many big human genome and disease genome projects have been launched and freely published the data for the research community. In parallel, state-of-the-art techniques in computational sciences (e.g., artificial intelligence, statistics) have fostered the application of computational methods to study problems in medicine. This could open a brighter future for personalized medicine in cancer research of the –omics data era. Personalized medicine is a broad research area and application. Indeed, besides biological characteristics of the patients, their clinical data, environment, and lifestyles are also important factors in tailoring the individual treatments. In addition, personalized medicine approaches are not only limited to cancers, but also be used to diagnose and treat other disorders such as rare diseases, which are strongly linked to molecular alterations. Furthermore, besides the abovementioned –omics data, metagenomics and metatranscriptomics should also be worthy of studying personalized medicine since there exist interactions between humans and the microbiome.

Acknowledgment

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.01-2017.14.

References

[1] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton *et al.*, "The sequence of the human genome," *science*, vol. 291, no. 5507, pp. 1304–1351, 2001.

[2] C. Manzoni, D. A. Kia, J. Vandrovcova, J. Hardy, N. W. Wood, P. A. Lewis *et al.*, "Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences," *Briefings in Bioinformatics*, vol. 19, no. 2, pp. 286–302, 2018.

[3] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski *et al.*, "GENCODE: The reference human genome annotation for the ENCODE project," *Genome Research*, vol. 22, no. 9, pp. 1760–1774, 2012.

[4] R. P. Horgan and L. C. Kenny, "Omic technologies: Genomics, transcriptomics, proteomics and metabolomics," *The Obstetrician & Gynaecologist*, vol. 13, no. 3, pp. 189–195, 2011.

[5] G. N. Samuel and B. Farsides, "The UK's 100,000 Genomes Project: Manifesting policymakers' expectations," *New Genetics and Society*, vol. 36, no. 4, pp. 336–353, 2017.

[6] GenomeAsia100K Consortium *et al.*, "The GenomeAsia 100K Project enables genetic discoveries across Asia," *Nature*, vol. 576, no. 7785, pp. 106–111, 2019.

[7] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott *et al.*, "The cancer genome

atlas pan-cancer analysis project," *Nature Genetics*, vol. 45, no. 10, p. 1113, 2013.

[8] S. Deorowicz, A. Danek, and M. Niemiec, "GDC 2: Compression of large collections of genomes," *Scientific Reports*, vol. 5, p. 11565, 2015.

[9] S. A. Forbes, D. Beare, P. Gunasekaran, K. Leung, N. Bindal, H. Boutselakis *et al.*, "COSMIC: Exploring the world's knowledge of somatic mutations in human cancer," *Nucleic Acids Research*, vol. 43, no. D1, pp. D805–D811, 2015.

[10] J. Zhang, J. Baran, A. Cros, J. M. Guberman, S. Haider, J. Hsu *et al.*, "International Cancer Genome Consortium Data Portal – a one-stop shop for cancer genomics data," *Database*, vol. 2011, 2011.

[11] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim *et al.*, "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, pp. 603–607, 2012.

[12] W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes *et al.*, "Genomics of Drug Sensitivity in Cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells," *Nucleic Acids Research*, vol. 41, no. D1, pp. D955–D961, 2012.

[13] S. Huang, K. Chaudhary, and L. X. Garmire, "More is better: Recent progress in multi-omics data integration methods," *Frontiers in Genetics*, vol. 8, p. 84, 2017.

[14] Y. Li, F.-X. Wu, and A. Ngom, "A review on machine learning principles for multi-view biological data integration," *Briefings in Bioinformatics*, vol. 19, no. 2, pp. 325–340, 2018.

[15] J. Yan, S. L. Risacher, L. Shen, and A. J. Saykin, "Network approaches to systems biology analysis of complex disease: Integrative methods for multi-omics data," *Briefings in Bioinformatics*, vol. 19, no. 6, pp. 1370–1381, 2018.

[16] C. Meng, O. A. Zeleznik, G. G. Thallinger, B. Kuster, A. M. Gholami, and A. C. Culhane, "Dimension reduction techniques for the integrative analysis of multi-omics data," *Briefings in Bioinformatics*, vol. 17, no. 4, pp. 628–641, 2016.

[17] F. Rohart, B. Gautier, A. Singh, and K.-A. Lê Cao, "mixOmics: An R package for 'omics feature selection and multiple data integration," *PLoS Computational Biology*, vol. 13, no. 11, p. e1005752, 2017.

[18] M. Bersanelli, E. Mosca, D. Remondini, E. Giampieri, C. Sala, G. Castellani *et al.*, "Methods for the integration of multi-omics data: Mathematical aspects," *BMC Bioinformatics*, vol. 17, no. S2, p. S15, 2016.

[19] C. Dimitrakopoulos, S. K. Hindupur, L. Häfliger, J. Behr, H. Montazeri, M. N. Hall *et al.*, "Network-based integration of multi-omics data for prioritizing cancer genes," *Bioinformatics*, vol. 34, no. 14, pp. 2441–2448, 2018.

[20] Q. Zhao, X. Shi, Y. Xie, J. Huang, B. Shia, and S. Ma, "Combining multidimensional genomic measurements for predicting cancer prognosis: Observations from TCGA," *Briefings in Bioinformatics*, vol. 16, no. 2, pp. 291–303, 2015.

[21] Q. Mo, F. Nikolos, F. Chen, Z. Tramel, Y.-C. Lee, K. Hayashi *et al.*, "Prognostic power of a tumor differentiation gene signature for bladder urothelial carcinomas," *Journal of the National Cancer Institute*, vol. 110, no. 5, pp. 448–459, 2018.

[22] M. Cortet, A. Bertaut, F. Molinié, S. Bara, F. Beltjens, C. Coutant *et al.*, "Trends in molecular subtypes of breast cancer: Description of incidence rates between 2007 and 2012 from three French registries," *BMC Cancer*, vol. 18, no. 1, p. 161, 2018.

[23] L. Zhao, V. H. Lee, M. K. Ng, H. Yan, and M. F. Bijlsma, "Molecular subtyping of cancer: Current status and moving toward clinical applications," *Briefings in Bioinformatics*, vol. 20, no. 2, pp. 572–584, 2019.

[24] M. Hofree, J. P. Shen, H. Carter, A. Gross, and T. Ideker, "Network-based stratification of tumor mutations," *Nature methods*, vol. 10, no. 11, pp. 1108–1115, 2013.

[25] Z. He, J. Zhang, X. Yuan, Z. Liu, B. Liu, S. Tuo *et al.*, "Network based stratification of major cancers by integrating somatic mutation and gene expression data," *PloS One*, vol. 12, no. 5, 2017.

[26] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno *et al.*, "Similarity network fusion for aggregating data types on a genomic scale," *Nature Methods*, vol. 11, no. 3, pp. 333–337, 2014.

[27] F. Zhang, C. Ren, K. K. Lau, Z. Zheng, G. Lu, Z. Yi *et al.*, "A network medicine approach to build a comprehensive atlas for the prognosis of human cancer," *Briefings in Bioinformatics*, vol. 17, no. 6, pp. 1044–1059, 2016.

[28] M. Le Morvan, A. Zinovyev, and J.-P. Vert, "NetNorM: Capturing cancer-relevant information in somatic exome mutation data with gene networks for cancer stratification and prognosis," *PLoS Computational Biology*, vol. 13, no. 6, p. e1005573, 2017.

[29] C. R. Planey and O. Gevaert, "CoINcIDE: A framework for discovery of patient subtypes across multiple datasets," *Genome Medicine*, vol. 8, no. 1, pp. 1–17, 2016.

[30] G. Yu, X. Yu, and J. Wang, "Network-aided Bi-Clustering for discovering cancer subtypes," *Scientific Reports*, vol. 7, no. 1, pp. 1–15, 2017.

[31] F. Azuaje, "Computational models for predicting drug responses in cancer research," *Briefings in Bioinformatics*, vol. 18, no. 5, pp. 820–829, 2017.

[32] N. Zhang, H. Wang, Y. Fang, J. Wang, X. Zheng, and X. S. Liu, "Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model," *PLoS Computational Biology*, vol. 11, no. 9, 2015.

[33] D.-H. Le and V.-H. Pham, "Drug response prediction by globally capturing drug and cell line information in a heterogeneous network," *Journal of Molecular Biology*, vol. 430, no. 18, pp. 2993–3004, 2018.

[34] J. C. Costello, L. M. Heiser, E. Georgii, M. Gönen, M. P. Menden, N. J. Wang *et al.*, "A community effort to assess and improve drug sensitivity prediction algorithms," *Nature Biotechnology*, vol. 32, no. 12, pp. 1202–1212, 2014.

[35] M. Ammad-ud din, S. A. Khan, D. Malani, A. Murumägi, O. Kallioniemi, T. Aittokallio *et al.*, "Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization," *Bioinformatics*, vol. 32, no. 17, pp. i455–i463, 2016.

[36] D. Le and D. Nguyen-Ngoc, "Multi-task regression learning for prediction of response against a panel of anti-cancer drugs in personalized medicine," in *Proceedings of the International Conference on Multimedia Analysis and Pattern Recognition*, Ho Chi Minh City, Vietnam, Apr. 2018.

Le Duc Hau obtained his PhD degree in Bioinformatics from University of Ulsan, Republic of Korea in 2012. He is now leading the Department of Computational Biomedicine, Vingroup Big Data Institute, VietNam. He has been focusing on proposing computational methods for disease- and drug-related problems in personalized medicine, especially on identification of disease-associated biomarkers, prediction of drug targets and response. In parallel, he has been developing bioinformatics tools. So far, he hasmore than fifty papers published in well-recognized journals and conferences, nearly a half of those are in ISI-indexed journals. In addition, he has been a member of program committees and reviewer of several international conferences/journals. Moreover, he is a principal investigator and a key member of some national/ministry-level projects. Specially, he is the principal investigator of the biggest genome project in Vietnam (i.e., building databases of genomic variants for Vietnamese population). Finally, he has been collaborating with some well-recognized international research institutes.

Quynh Diep Nguyen obtained her PhD degree in Information Technology from the Institute of Information Technology - The Vietnam Academy of Science and Technology in 2015. She is a lecturer at the School of Computer Science and Engineering, Thuyloi University. She has been focusing on computational methods for reconstructing the metabolic networks. So far, she has more than fifteen papers in journals and conferences published . Moreover, she is a member of some national/ministry-level projects which research on computational methods for uncovering latent knowledge from high-throughput biological data.