Published online: 2019-08-16

Contributions from the 2018 Literature on Bioinformatics and Translational Informatics

Malika Smaïl-Tabbone[1], Bastien Rance[2], Section Editors for the IMIA Yearbook Section on Bioinformatics and Translational Informatics

[1] Loria UMR 7503, Université de Lorraine, CNRS, Inria Nancy Grand-Est, Nancy, France
[2] HEGP, AP-HP; Université Paris Descartes, Université de Paris; UMRS 1138 Centre de Recherche des Cordeliers INSERM, Paris, France

Summary

Objectives: To summarize recent research and select the best papers published in 2018 in the field of Bioinformatics and Translational Informatics (BTI) for the corresponding section of the International Medical Informatics Association (IMIA) Yearbook.

Methods: A literature review was performed for retrieving from PubMed papers indexed with keywords and free terms related to BTI. Independent review allowed the two section editors to select a list of 14 candidate best papers which were subsequently peer-reviewed. A final consensus meeting gathering the whole IMIA Yearbook editorial committee was organized to finally decide on the selection of the best papers.

Results: Among the 636 retrieved papers published in 2018 in the various subareas of BTI, the review process selected four best papers. The first paper presents a computational method to identify molecular markers for targeted treatment of acute myeloid leukemia using multi-omics data (genome-wide gene expression profiles) and in vitro sensitivity to 160 chemotherapy drugs. The second paper describes a deep neural network approach to predict the survival of patients suffering from glioma on the basis of digitalised pathology images and genomics biomarkers. The authors of the third paper adopt a pan-cancer approach to take benefit of multi-omics data for drug repurposing. The fourth paper presents a graph-based semi-supervised method to accurate phenotype classification applied to ovarian cancer.

Conclusions: Thanks to the normalization of open data and open science practices, research in BTI continues to develop and mature. Noteworthy achievements are sophisticated applications of leading edge machine-learning methods dedicated to personalized medicine.

Keywords

International Medical Informatics Association Yearbook, bioinformatics and translational informatics, artificial intelligence

Yearb Med Inform 2019:190-4
http://dx.doi.org/10.1055/s-0039-1677945

Introduction

Within the 2019 International Medical Informatics Association (IMIA) Yearbook, the goal of the Bioinformatics and Translational Informatics (BTI) section is to provide an overview of research trends from 2018 publications that demonstrated excellent research about various aspects of bioinformatics methods and techniques to advance clinical care. In 2008, the American Medical Informatics Association (AMIA) has defined translational bioinformatics as "*… the development of storage, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data into proactive, predictive, preventative, and participatory health"* [1]. First priorities addressed storage and retrieval, and focused analytics of high-throughput data motivating numerous research and development studies in the last decade. Today, the topic is clearly coming of age with more ambitious objectives (such as pan-cancer approaches, multi-omics analyses, drug repurposing) which make use, among others, of the most advanced computational methods such as Artificial Intelligence and Deep Learning- this year's special theme for the IMIA Yearbook.

Paper Selection Method

Following the method described in [2], a comprehensive review of articles published in 2018 and addressing various subtopics for BTI was conducted. The selection was performed by querying MEDLINE via PubMed (from NCBI, National Center for Biotechnology Information) with a set of predefined Medical Subject Headings (MeSH) descriptors along with free terms: Translational informatics; Translational bioinformatics; Bioinformatics; Computational molecular biology; Computing Methodologies; Information storage and retrieval; Pattern recognition, Automated; Medical informatics, Algorithms; Translational medical Research; Genetics, Medical; Precision Medicine; Personalized medicine; Molecular Medicine; Genomic medicine; Medical genetics; Medical genomics; Clinical genomics; Genetics; Genomics; Next-generation sequencing; High throughput sequencing; Transcriptome; Transcriptomics; Proteome; Proteomics; Proteogenomics; Epigenomics; Metabolomics; Metagenomics; Large-scale datasets; Big data; Omics; and Multi-omics. Bibliographic databases were searched on February 24th, 2019 for papers published in 2018, considering the electronic publication date. The original set of 636 references was reviewed jointly by the two section editors to select a consensual list. Hence, 42 (respectively 45) references were selected by the first (second) section editor based on the title and abstract of papers. Among the 16 papers in common, three were excluded due to moderate interest and critical length. Following the IMIA Yearbook process, the 13 candidate best papers were peer-reviewed by the IMIA Yearbook editors and external reviewers (at least three reviewers per paper). Four papers were finally selected as best papers (Table 1). A content summary of these best papers can be found in the appendix of this synopsis.

Description of Candidate Best Papers and Best Papers

Rapid content analysis of the 636 retrieved references revealed a large proportion of papers dealing with identification and routine use in clinical settings of genetic variants in connection with various diseases. Through their choices, section editors wanted to shed lights on three research trends and two emerging topics in BTI field which are presented in the following [3-15].

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2019 in the section 'Bioinformatics and Translational Informatics'. The articles are listed in alphabetical order of the first author's surname.

Section Bioinformatics and Translational Informatics
▪ Lee SI, Celik S, Logsdon BA, Lundberg SM, Martins TJ, Oehler VG, Estey EH, Miller CP, Chien S, Dai J, Saxena A, Blau CA, Becker PS. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. Nat Commun 2018 Jan;9(1):42.
▪ Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Velázquez Vega JE, Brat DJ, Cooper LAD. Predicting cancer outcomes from histology and genomics using convolutional networks. Proc Natl Acad Sci U S A 2018;115(13):E2970-E2979.
▪ Sengupta S, Sun SQ, Huang KL, Oh C, Bailey MH, Varghese R, Wyczalkowski MA, Ning J, Tripathi P, Mc Michael JF, Johnson KJ, Kandoth C, Welch J, Ma C, Wendl MC, Payne SH, Fenyö D, Townsend RR, Dipersio JF, Chen F, Ding L. Integrative omics analyses broaden treatment targets in human cancer. Genome Med 2018 Jul 27;10(1):60.
▪ Torshizi AD, Petzold LR. Graph-based semi-supervised learning with genomic data integration using condition-responsive genes applied to phenotype classification. J Am Med Inform Assoc 2018;25(1):99-108.

Trend 1: Artificial Intelligence and Deep Learning at the service of translational informatics

While machine learning has been used in medical informatics for a few decades, the year 2018 has seen a renewed interest for the field. New methods, including Neural Networks, have been largely adopted by the medical community in virtually all the realms of medicine. In the recent year, the most spectacular results have been obtained for images analysis (and especially for photographic image analysis, e.g., digital pathology images, retina pictures...), but numerous other fields are investigated. Three of the 13 candidate best papers are directly using artificial intelligence methods. The contribution by Mobadersany et al., presents a method to predict the survival of patients based on digital pathology images as well as genomics biomarkers [8]. The work by Lee K. et al., [5] used deep neural network and a word embedding built with word2vec in order to identify pharmacogenomics relations in the literature. Finally Lee SI. et al., described the MERGE algorithm (see Trend 2 for further details) to identify treatments against acute myeloid leukemias [6]. In their paper, Mobadersany et al., developed a neural network based approach to predict the survival of patient on the basis of digitalized pathology images and genomics biomarkers. The authors describe a Survival Convolutional Neural Network (SCNN) designed to predict the survival of patients suffering from glioma. Networks were trained using public data coming from the Cancer Genome Atlas (TCGA) datasets. To help the interpretation and the understanding of the prediction, the authors use a heat map visualization illustrating the structures identified as important by the neural networks.

Trend 2: Pan-cancer approach and integration of multi-omics data for more insightful analyses

The large availability of high throughput data in the context of biological and clinical research but also during normal patient care has enabled the development of new approaches to classify diseases and to identify potentially better or new treatments for known diseases. The best paper candidates highlight this trend in four contributions using multi-omics datasets (leveraging gene expression, methylation, proteomics, next generation sequencing (NGS),…etc.). These new approaches are generating a great amount of interest and a large body of work as represented by the selection of three of the four best papers of the BTI section.

In their work Sengupta *et al*., adopted a pan-cancer approach to take benefit of multi-omics data for drug repurposing [11]. Their goal is to identify drugs approved by the Food and Drug Administration (FDA) for cancer location not yet mentioned in the approval. The authors rely on the Database of Evidence for Precision Oncology (DEPO), a tool they built and presented in [18], to link druggability to genomic, transcriptomic, and proteomic biomarkers. They used a pan-cancer cohort of more than 6,500 tumors to identify those with potential druggable markers. The authors rely on the DEPO database (integrating genomic, transcriptomic, and proteomic data, as well as clinical data over several types of cancer samples) and structural alignment tools for identifying tumors with potentially druggable biomarkers consisting of drug-associated mutations, micro-RNA expression outliers, and protein/phosphoprotein expression outliers.

As opposed to the two pan-cancer contributions, the next two candidates focus on specific tumor locations : in [3] (a candidate best paper), the authors rely on a multi-omics approach to define stage and predict treatment outcome for patients suffering from Crohn's disease. Selected as best paper, Lee SI *et al*., [6] present a statistical method to identify molecular markers for targeted treatment of acute myeloid leukemia using omics data (genome-wide gene expression profiles) and in vitro sensitivity to 160 chemotherapy drugs. They describe the MERGE algorithm (standing for mutation, expression hubs, known regulators, genomic copy number variation, and methylation) a computational method to identify gene expression markers using multi-omics data. In a nutshell, MERGE learns from data the contribution of five key features (mutation associated to acute myeloid leukemia, hubness in a gene expression network…) to the drive of gene potentially implicated in cancer progression. The work by Torshizi and Petzold [13], detailed in the next trend section, also makes use of multi-omics.

Trend 3: Pathway-level versus gene-level analyses

A trend which is becoming established concerns the early integration of background biomolecular knowledge in the BTI context as initiated in previous publications [17]. As an alternative to single-gene analyses of patient genomic data, interesting papers published in 2018 adopt the pathway level to identify relevant biomarkers or to improve phenotype classification. These studies go from the assumption stating that cohort patients exhibit homogeneity at the transcript level (deregulated expression of a set of genes) to the same assumption at the pathway level (deregulated pathway expression). This is especially relevant in complex diseases or heterogeneous patient populations where conventional data analytics provided rather poor results in terms of precision medicine, so far [16].

The first paper selected as the best paper authored by Torshizi and Petzold presents a graph-based semi-supervised method to phenotype classification [13]. Several graphs of labeled and unlabeled samples are built on features sets corresponding to distinct genomic levels namely, gene expression, DNA methylation, and micro-RNA. Additional graphs add pathway knowledge - for each considered genomic level - through the use of condition-responsive genes (CORGs) defined in [17]. The method was applied on ovarian cancer data (from the Human genome Atlas) and the comparative evaluation results show that the classification accuracy in terms of survival is effectively improved when integrating transcriptomic, epigenetic, and pathway knowledge. A noteworthy advantage of the proposed approach is its capacity to address the positive example sparsity problem.

Zaim *et al*., propose in their paper (a candidate best paper) a statistical framework to revisit the biomarker discovery process [15]. They show through a POC simulation how to discover common pathways in a single-subject approach and the gained advantages over usual cohort-based approaches in specific scenarios.

Emerging Topic 1: Towards clinician-friendly infrastructures for data integration and analyses

The growing complexity of biological and clinical datasets requires new tools to help researchers in data management and exploration process. This year saw a number of contributions providing experts with innovative ways to interact with multi-omics and clinical datasets. Among the 13 candidate best papers, two were addressing the issue of data exploration [9,4], and one described a large national infrastructure hosting data and samples at a national level [7]. Moscatelli *et al*., present an infrastructure design to simplify the exploration of clinical data, both structured (e.g., laboratory data) and unstructured (free text clinical reports) [9]. Their infrastructure relies on a NoSQL structure, and manages anonymization and machine learning layers to assist the research in the mining of data. Krempel *et al*., introduce the CancerSysDB, a web-based application designed to host multi-omics pan-cancer data, and to simplify the querying process [4]. CancerSysDB is open-source and can host both public datasets, and especially data coming from the TCGA, but also private datasets through the use of a self-hosted instance of the system. Workflows can be connected to the application and dynamically uploaded to the application.

Emerging Topic 2: Ethical and methodological issues raised in BTI practices

Even though they are not new, ethical questions are still raising by the merger of research studies and clinical care practices. A certain exacerbation can be noticed may be due to the development of private companies offering genetic services. Among the 13 candidate best papers, we have a collection of three papers addressing ethical questions in several countries settings [10, 12, 14]. An additional paper covers lawful procedures of access to biobanks and electronic health records in Taiwan as the authors present the Taiwan Biobank involving biopsy samples from 200,000 participants (patients and citizens) and discuss possible solutions for ensuring both broad access and privacy preservation [7]. The first paper of the collection provides an interesting analysis of the results of a survey over nine sites implementing translational genomics (Clinical Sequencing Evidence-Generating Research (CSER) consortium) [14]. The results -although limited to the CSER consortium participants- are insightful regarding several issues raised at the interface between sequencing-based research and clinical care such as informed consent procedures, clinician and researcher roles, disclosure of primary results and secondary findings, storage of results in the medical records, payment for services, and overall characterization of the research-clinical interface. The second paper informs on the current debate around disclosure of genomic secondary findings [10]. The reported opinions and positions are those of UK participants to genome sequencing program namely the rare disease genomic medicine multidisciplinary team involved in the 100,000 Genomes project. The paper answers some important questions facing multidisciplinary care boards. Concerning the disclosure procedure of secondary findings, the conclusion of UK experts is in agreement with the US CSER survey results [14]. As for the third paper, Stoeklé *et al*., discuss interesting ethical issues from the perspective of the evolution of *tumor boards* to *molecular tumor boards* in French medical systems [12]. This evolution is a consequence of two important changes: NGS techniques which henceforth generate whole-genome sequencing data at low costs, and machine learning approaches which open huge perspectives for analysing patient data. The authors discuss how to improve patient confidence and trust in academic medical centers to prevent commercial private companies to exploit exclusively for-profit sensitive genetic data. The suggested means include the use of information technology (IT) for more efficient acquisition of informed consent from patients and better communication modalities with researchers and clinicians.

Conclusion and Outlook

A few interesting papers published in 2018 in the BTI scope matched the IMIA Yearbook special topic “Artificial Intelligence”. These

papers illustrate well the complexity and the constraints induced by the deployment of deep-learning techniques, especially in the context of multidisciplinary and personalized care (including molecular characterization of tumors...). Further intelligent approaches are expected in coming years, combining semantic web languages with clinical omics data and biomolecular knowledge for extracting self-explanatory and actionable knowledge nuggets in clinical settings. It is worth noting that many contributions keep on relying on public datasets (such TCGA…), as well as open tools and systems. In this context, the emergence of clinician-friendly exploration and analysis platforms should bring closer the clinical, translational, and bioinformatics communities.

Acknowledgments
We would like to thank Adrien Ugon for his support and the reviewers for their participation to the selection process of the BTI section of the IMIA Yearbook.

References

1. Butte AJ. Translational bioinformatics: Coming of age. J Am Med Inform Assoc 2008;15(6):709–14.
2. Lamy JB, Séroussi B, Griffon N, Kerdelhué G, Jaulent MC, Bouaud J. Toward a formalization of the process to select IMIA Yearbook best papers. Methods Inf Med 2015;54(2)135-44.
3. Douglas GM, Hansen R, Jones CMA., Dunn KA, Comeau AM, Bielawski JP, et al. Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn's disease. Microbiome 2018 Jan;6(1):13.
4. Krempel R, Kulkarni P, Yim A, Lang U, Habermann B, Frommolt P. Integrative analysis and machine learning on cancer genomics data using the Cancer Systems Biology Database (CancerSysDB). BMC Bioinformatics 2018 Apr;19(1):156.
5. Lee K, Kim B, Choi Y, Kim S, Shin W, Lee S, et al. Deep learning of mutation-gene-drug relations from the literature. BMC Bioinformatics 2018 Jan;19(1):21.
6. Lee SI, Celik S, Logsdon BA, Lundberg SM, Martins TJ, Oehler VG, et al. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. Nat Commun 2018 Jan;9(1):42.
7. Lin JC, Fan CT, Liao CC, Chen YS. Taiwan Biobank: making cross-database convergence possible in the Big Data era. GigaScience 2018 Jan;7(1):1-4.
8. Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Velázquez Vega JE, et al. Predicting cancer outcomes from histology and genomics using convolutional networks. Proc Natl Acad Sci U S A 2018;115(13):E2970-E2979.
9. Moscatelli M, Manconi A, Pessina M, Fellegara G, Rampoldi S, Milanesi L, et al. An infrastructure for precision medicine through analysis of big data. BMC Bioinformatics 2018 Oct;19(Suppl 10):351.
10. Ormondroyd E, Mackley MP, Blair E, Craft J, Knight JC, Taylor JC, et al. Not pathogenic until proven otherwise": perspectives of UK clinical genomics professionals toward secondary findings in context of a Genomic Medicine Multidisciplinary Team and the 100,000 Genomes Project. Genet Med 2018 Mar;20(3):320-8.
11. Sengupta S, Sun SQ, Huang KL, Oh C, Bailey MH, Varghese R, et al. Integrative omics analyses broaden treatment targets in human cancer. Genome Med 2018 Jul 27;10(1):60.
12. Stoeklé HC, Mamzer-Bruneel MF, Frouart CH, Le Tourneau C, Laurent-Puig P, Vogt G, et al. Molecular Tumor Boards: Ethical Issues in the New Era of Data Medicine. Sci Eng Ethics 2018 Feb;24(1):307-22.
13. Torshizi AD, Petzold LR. Graph-based semi-supervised learning with genomic data integration using condition-responsive genes applied to phenotype classification. J Am Med Inform Assoc 2018 Jan;25(1):99–108.
14. Wolf SM, Amendola LM, Berg JS, Chung WK, Clayton EW, Green RC, et al Navigating the research-clinical interface in genomic medicine: analysis from the CSER Consortium. Genet Med 2018 Apr;20(5):545-53.
15. Zaim SR, Li Q, Schissler AG, Lussier YA. Emergence of pathway-level composite biomarkers from converging gene set signals of heterogeneous transcriptomic responses. Pac Symp Biocomput 2018;23:484-95.
16. Hristova VA, Chan DW. Cancer biomarker discovery and translation: proteomics and beyond. Expert Rev Proteomics 2019 Feb;16(2):93-103.
17. Lee E, Chuang HY, Kim JW, Ideker T, Lee D. Inferring pathway activity toward precise disease classification. PLoS Comput Biol 2008;4(11):e1000217.
18. Sun SQ, Mashl RJ, Sengupta S, Scott AD, Wang W, Batra P, et al. Database of evidence for precision oncology portal. Bioinformatics 2018;34(24):4315–7.

Correspondence to:
Malika Smaïl-Tabonne
Loria UMR 7503
Université de Lorraine
Nancy, France
E-mail: malika.smail@loria.fr

Appendix: Content Summaries of Selected Best Papers for the 2019 IMIA Yearbook, Section Bioinformatics and Translational Informatics

Lee SI, Celik S, Logsdon BA, Lundberg SM, Martins TJ, Oehler VG, Estey EH, Miller CP, Chien S, Dai J, Saxena A, Blau CA, Becker PS

A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia

Nat Commun 2018 Jan;9(1):42

The authors present in this paper a statistical method to identify molecular markers for targeted treatment of acute myeloid leukemia using omics data (genome-wide gene expression profiles) and in vitro sensitivity to 160 chemotherapy drugs. They describe the MERGE algorithm (standing for mutation, expression hubs, known regulators, genomic copy number variation, and methylation) a computational method to identify gene expression markers using multi-omics data. In a nutshell, MERGE learns from data the contribution of five key features (*e.g.*, mutation associated to acute myeloid leukemia, hubness in a gene expression network) to the drive of gene potentially implicated in cancer progression. A complete approach is designed ranging from data collection and method development to both *in silico* and *in vivo* validation.

Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Velázquez Vega JE, Brat DJ, Cooper LAD

Predicting cancer outcomes from histology and genomics using convolutional networks

Proc Natl Acad Sci U S A 2018;115(13):E2970-E2979

This paper presents a method to predict the survival of patients based on digital pathology images as well as genomics biomarkers. The authors developed a neural network based approach to predict the survival of patient on the basis of digitalized pathology images and genomics biomarkers. The authors describe a Survival Convolutional Neural Network (SCNN) designed to predict the survival of patients suffering from glioma. The networks are trained using public data coming from the TCGA datasets. To help in the interpretation and understanding of the prediction, the authors use a heat map visualization highlighting the structures identified as important by the neural networks.

Sengupta S, Sun SQ, Huang KL, Oh C, Bailey MH, Varghese R, Wyczalkowski MA, Ning J, Tripathi P, Mc Michael JF, Johnson KJ, Kandoth C, Welch J, Ma C, Wendl MC, Payne SH, Fenyö D, Townsend RR, Dipersio JF, Chen F, Ding L

Integrative omics analyses broaden treatment targets in human cancer

Genome Med 2018 Jul 27;10(1):60

In this work the authors adopt a pan-cancer approach to take benefit of multi-omics data for drug repurposing. Their goal is to identify drugs approved by the Food and Drug Administration for cancer location not yet mentioned in the approval. The authors rely on the Database of Evidence for Precision Oncology (DEPO), a tool they built, to link druggability to genomic, transcriptomic, and proteomic biomarkers. They used a pan-cancer cohort of more than 6,500 tumors to identify tumor with potential druggable markers. The authors rely on the DEPO database (integrating genomic, transcriptomic, proteomic data, and clinical data over several types of cancer samples) and structural alignment tools for identifying tumors with potentially druggable biomarkers consisting of drug-associated mutations, micro-RNA expression outliers, and protein/phosphoprotein expression outliers. Orthogonal validation of putative biomarkers was performed thanks to the large-scale drug screening dataset GDSC (Genomics of Drug Sensitivity in Cancer).

Torshizi AD, Petzold LR

Graph-based semi-supervised learning with genomic data integration using condition-responsive genes applied to phenotype classification

J Am Med Inform Assoc 2018;25(1):99-108

This paper presents a graph-based semi-supervised method to phenotype classification of samples. Several graphs of labeled and unlabeled samples are built on features sets corresponding to distinct genomic levels (gene expression, DNA methylation, micro-RNA). Additional graphs add pathway knowledge -for each considered genomic level- through the use of condition-responsive genes (CORGs). CORGs are, for each pathway, the most discriminative set of genes containing the highest statistical signal level. The authors define three feature sets corresponding to different subsets of CORGs (the whole sets, the top P-value ranked genes, the top ranked genes according to their frequency in all CORGs). A weighted integration of the various graphs is performed before the semi-supervised learning based on the K nearest neighbors principles. The method was applied on ovarian cancer data from the Human genome Atlas. The conducted experiments assessed the added value of the method compared to the existing ones. The results also show that the classification accuracy is effectively improved when integrating transcriptomic, epigenetic, and pathway knowledge.