Published online: 2020-08-21

Contributions from the 2019 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 de Paris, UMRS 1138 Centre de Recherche des Cordeliers, INSERM, Paris, France

Summary

Objectives: Summarize recent research and select the best papers published in 2019 in the field of Bioinformatics and Translational Informatics (BTI) for the corresponding section of the International Medical Informatics Association 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 section editors to select a list of 15 candidate best papers which were subsequently peer-reviewed. A final consensus meeting gathering the whole Yearbook editorial committee was organized to finally decide on the selection of the best papers.

Results: Among the 931 retrieved papers covering the various subareas of BTI, the review process selected four best papers. The first paper presents a logical modeling of cancer pathways. Using their tools, the authors are able to identify two known behaviours of tumors. The second paper describes a deep-learning approach to predicting resistance to antibiotics in Mycobacterium tuberculosis. The authors of the third paper introduce a Genomic Global Positioning System (GPS) enabling comparison of genomic data with other individuals or genomics databases while preserving privacy. The fourth paper presents a multi-omics and temporal sequence-based approach to provide a better understanding of the sequence of events leading to Alzheimer's Disease.

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, ethical issues

Yearb Med Inform 2020:188-92
http://dx.doi.org/10.1055/s-0040-1702002

1 Introduction

Within the 2020 International Medical Informatics Association (IMIA) Yearbook of Medical Informatics, the goal of the Bioinformatics and Translational Informatics (BTI) section is to provide an overview of research trends from 2019 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]. The first priorities addressed storage, retrieval, and focused analytics of high-throughput data that motivated 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.

2 Paper Selection Method

Following the method described in [2], a comprehensive review of articles published in 2019 and addressing various subtopics for BTI was conducted. The selection was performed by querying MEDLINE via PubMed (from NCBI, National Center for Biotechnology Information, NLM, NIH) with a set of predefined 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 January 20th, 2020 for papers published in 2019, considering the electronic publication date. The original set of 931 references was reviewed jointly by the two section editors to select a consensual list. Hence, 41 and 34 references were selected by MS, respectively BR, based on the title and the abstract of the papers. Six articles were in common to both selections, and 11 were accepted in one selection and pending in the other. Section editors agreed on 15 papers out of the 17 pre-selected references which were subsequently peer-reviewed by the IMIA Yearbook editors and external reviewers (at least four reviews 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.

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2020 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
▪ Béal J, Montagud A, Traynard P, Barillot E, Calzone L. Personalization of logical models with multi-omics data allows clinical stratification of patients. Front Physiol 2019 Jan 24;9:1965.
▪ Chen ML, Doddi A, Royer J, Freschi L, Schito M, Ezewudo M, Kohane IS, Beam A, Farhat M. Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in Mycobacterium tuberculosis resistance prediction. EBioMedicine 2019 May;43:356–69.
▪ Kim K, Baik H, Jang CS, Roh JK, Eskin E, Han B. Genomic GPS: using genetic distance from individuals to public data for genomic analysis without disclosing personal genomes. Genome Biol 2019 Dec;20(1):175.
▪ Marttinen M, Paananen J, Neme A, Vikram M, Takalo M, Natune T, Paldanius KMA, Mäkinen P, Bremang M, Kurki MI, Rauramaa T, Leinonen V, Soininen H, Haapasalo A, Pike I, Hiltunen M. A multiomic approach to characterize the temporal sequence in Alzheimer's disease-related pathology. Neurobiol Dis 2019;124:45468.

3 Description of Candidate Best Papers and Best Papers

A rapid content analysis of the 931 retrieved references revealed a large proportion of studies 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 light on three research trends and one emerging topic in the BTI field which are presented in the following [3–17]. The research trends and emerging topics cover three different dimensions: methods, application domain, and purpose. Many of the selected articles are present in more than one trend or emerging topic.

3.1 Trend 1: Approaches Based on Machine-learning Methods

The 2019 selection confirms the massive impact of machine learning and neural networks in the biomedical informatics field. Machine learning methods in general, and more specifically deep learning, have been largely adopted by the biomedical community. Last year, successes were mostly obtained in the field of image analysis; new fields of applications, including drug sensitivity or resistance prediction or genomics, have been explored successfully in the current edition. While machine learning methods were mostly referring to tools coming from the computer science community, it is worth noting that many articles used the overall term artificial intelligence for any predictive system, including classical statistical approaches such as regression. The use of machine learning libraries such as keras, tensorflow, or scikit-learn is also clear and could further generalize the adoption of machine learning methods.

Chen et al., [6] use a neural architecture combining a wide and deep neural network (DNN) to tackle the problem of multidrug resistance in tuberculosis. To that end, the authors combined the wide and deep neural network with features derived from genomic variant data (both rare and known to cause resistance). Liu et al., [14] also approache the problem of sensitivity to drugs using a neural network approach. They designed a twin-convolutional network taking as input SMILE sequences and data derived from gene expression to predict the response of a drug. Leveraging already existing components, the authors came up with a novel DNN architecture for a complex problem. The paper also illustrates well the difficulty of tuning the hyperparameters of DNNs. In [5], Béal et al,. demonstrate the potential of using logical modeling for systems biology to support precision medicine thanks to a sophisticated tailoring protocol. Their so-called PROFILE approach helps for patient stratification which is shown to be correlated with patient grouping on NPI (Nottingham Prognostic Index) and survival time. Various strategies are proposed to use patient omics data (mutations, CNA, expression, etc.) to personalize a generic logical model of cancer signaling pathways through stochastic simulations. The paper uses the METABRIC breast cancer as a proof of concept and allows to see promising directions to build patient models with mechanistic insights. Esteban-Medina et al., [8] used public gene expression data and a list of genes that are the target of approved drugs to identify potential causal relationships between proteins and cell activities. They rely on a Multi-Output Random Forest regressor available in scikit-learn and an optimization strategy built on top to predict a circuit activity across the disease pathway. Wan et al., [17] used logistic regression and Support Vector Model (SVM), associated with dimension reduction methods (principal component analysis and truncated singular-value decomposition) to build a signature of early stage colorectal cancer using whole genome sequencing data. They used an original confounder-controlled cross validation procedure for robust generalization estimation. Ibrahim et al., [11] used a statistical approach leveraging LASSO attribute selection and Monte-Carlo cross validation simulation to identify variables predictive of acute kidney injuries using proteomics data.

The work by Adam *et al.,* [3] indirectly tackles the question of machine learning. The authors emphasized the importance of early standardization and data quality control during data collection to enable later data exploitation using machine learning methods.

3.2 Trend 2: Genomics, Proteomics, and Multi-omics for the Exploration of a Wide Range of Diseases

This year's selection and overall review process clearly highlighted a trend already observed last year: the increased use of multi-omics approaches. Nearly all the arti-

cles of the top-15 selection used some form of omics data, and most of the works relied on multi-omics approaches. The use of publicly available data already mentioned in last year's BTI section is confirmed. While the main domain of application of the field remains cancer research, numerous other domains of applications have been explored. A large variety of methods are used, along with many sources of data ranging from sequencing technologies that have been largely adopted over the last decade, to new methods including circulating DNA sequencing and multiplex proteomics.

Among the many applications of BTI in biomedical research, the selection highlighted common tasks:

- Early stage detection of cancer (e.g., using circulating biomarkers) and recurrence detection [16, 17] ;
- Patient stratification and drug sensitivity prediction, and simulation models [7, 9, 10, 14].

The paper authored by Marttinen et al., [15], which belongs to our top-4 best papers, proposes a multi-omics and temporal sequence-based approach to provide a better understanding of the sequence of events leading to Alzheimer's Disease (AD). The authors coupled transcriptomic and phosphoproteomic data to determine the temporal sequence of changes in microRNA, protein, and phosphopeptide expression levels from human temporal cortical samples, with varying stages of AD. This approach highlighted a significant sequence of key functions occurring at the considered stages of the disease.

A quite odd paper is the one authored by Chung et al., [7] where a tool, called OmicsSIMLA, is proposed to simulate genomics (SNPs and CNV), epigenomics (i.e., bisulfite sequencing), transcriptomics, and proteomics data at the whole-genome level. Both the relationships between different types of omics data and between multi-omics data and disease status have been modeled. If the tool is adopted, OmicsSIMLA would be useful to generate benchmarks to evaluate the performance of methods analysing multi-omics data methods. Sample sizes can also be estimated when planning a new multi-omics disease study.

3.3 Trend 3: Drug Repositioning and Large-scale Prediction of Drug Sensitivity

The availability of high throughput technologies (in particular in genomics and proteomics) increased the volume of open data, along with the use of artificial intelligence methods enabling new tools to identify drugs and potential targets of drugs in different diseases.

Fernández-Navarro *et al.,* [9] used PanDrug, a previously developed pharmacological resource, to identify targeted treatments against cancers. PanDrug proposes a large set of drug-target associations and offers a score on gene cancer relevance and drug target to guide the selection of the best suited treatment. In their work, the authors combined PanDrug with information coming from RNA and DNA sequencing to suggest new therapies in a case of acute T-Cell lymphoblastic leukemia. Graim *et al.,* [10] proposed PLATYPUS a machine learning approach to identify drug sensitivity signatures from cancer cell-lines databases. To overcome the problem of missing data and data sparsity both at the learning and prediction times, the authors developed an approach relying on a concept called *multiple view learning,* a semi-supervised method, and leveraged multi-omics data (expression, CNV, mutation, and also Sample- and Patient- specific information). Liu *et al.,* [14], as mentioned before, also relied on cell line data but in combination with features extracted from SMILES representation of drugs. Esteban-Medina *et al,.* [8] used a machine learning approach, combined with data from KEGG, Orphanet, GEO, GTEx and DrugBank to identify drugs that could have an effect on signaling the circuits that cause the treatment of the Fanconi anemia. Selected in the top-4 papers, Chen *et al*., [6] used a neural network to predict drug resistance to antibiotics in *Mycobacterium tuberculosis*. They leveraged whole-genome sequencing of the pathogene, as well as rare variants and known drug resistant variants to predict the resistance to 10 anti-tuberculosis drugs. They achieved AUCs over 93% for first-line drugs, and 89% for second-line drugs.

3.4 Emerging Topic: Ethical Issues Raised in BTI Practices

This year's special topic, ethics, was already an emerging topic in the previous edition. The coverage of ethical issues remains relatively low, with two articles among the 15 pre-selected in the section. While the large use of artificial intelligence methods requires and will continue to require ethical reflection and research, it may seem surprising to observe such a low representation in the final selection. We had indeed hypothesized that the culture of data use (including management of private personal information), the secondary use of public data, and the associated regulations (GDPR in Europe, HIPAA in the USA) would have enabled an early emergence of ethical concern in the field.

In this year's top-4 selection, Kim *et al.,* [13] introduced the Genomics GPS, a method analogical to transport GPS, to enable comparison of genomics information between individuals or an individual and a group without disclosing sensitive genomics information. The second paper illustrating this emerging topic comes from Kim et al. [12] and tackled the issue of racial representation disparity in population genomic sequencing programs. This topic is particularly relevant in the age of artificial intelligence and data-driven models for which biased training datasets can lead to erroneous models. The authors proposed a method to quantify the difference in the composition of ethnicity in four genomic databases relative to epidemiological data on the US population.

4 Conclusion and Outlook

A few interesting papers published in 2019 in the BTI scope matched the special topic "Ethics in Health Informatics" of the 2020 IMIA Yearbook of Medical Informatics. 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. We noticed a good diversity of methodologies, ranging from traditional regression-based approaches to

logical modeling of biological systems, to support several tasks related to precision medicine such as early stage detection of cancer, disease recurrence detection, and drug sensitivity prediction.

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 as GEO, KEGG, Orphanet…), as well as open machine learning libraries, tools, and systems.

Acknowledgments

We would like to thank Adrien Ugon for his support and the reviewers for their participation in 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 Dec;15(6):709–14.
2. Lamy J-B, Séroussi B, Griffon N, Kerdelhué G, Jaulent M-C, Bouaud J. Toward a formalization of the process to select IMIA Yearbook best papers. Methods Inf Med 2015;54(2):135–44.
3. Adam TJ, Chi C-L. Big Data Cohort Extraction for Personalized Statin Treatment and Machine Learning. In: Larson RS, Oprea TI, editors. Bioinformatics and Drug Discovery [Internet]. New York, NY: Springer New York; 2019 [cited 2020 May 25]. p. 255–72. (Methods Mol Biol 2019;1939.255-72). Available from: http://link.springer.com/10.1007/978-1-4939-9089-4_14
4. Barroso-Sousa R, Guo H, Srivastava P, James T, Birch W, Siu LL, et al. Utilization of tumor genomics in clinical practice: an international survey among ASCO members. Future Oncol 2019 Jul;15(21):2463–70.
5. Béal J, Montagud A, Traynard P, Barillot E, Calzone L. Personalization of Logical Models With Multi-Omics Data Allows Clinical Stratification of Patients. Front Physiol 2019 Jan 24;9:1965.
6. Chen ML, Doddi A, Royer J, Freschi L, Schito M, Ezewudo M, et al. Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in Mycobacterium tuberculosis resistance prediction. EBioMedicine 2019 May;43:356–69.
7. Chung R-H, Kang C-Y. A multi-omics data simulator for complex disease studies and its application to evaluate multi-omics data analysis methods for disease classification. GigaScience [Internet]. 2019 May 1 [cited 2020 May 25];8(5):giz045. Available from: https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giz045/5480572
8. Esteban-Medina M, Peña-Chilet M, Loucera C, Dopazo J. Exploring the druggable space around the Fanconi anemia pathway using machine learning and mechanistic models. BMC Bioinformatics. 2019 Dec;20(1):370.
9. Fernández-Navarro P, López-Nieva P, Piñeiro-Yañez E, Carreño-Tarragona G, Martinez-López J, Sánchez Pérez R, et al. The use of PanDrugs to prioritize anticancer drug treatments in a case of T-ALL based on individual genomic data. BMC Cancer 2019 Dec;19(1):1005.
10. Graim K, Friedl V, Houlahan KE, Stuart JM. PLATYPUS: A Multiple-View Learning Predictive Framework for Cancer Drug Sensitivity Prediction. Pac Symp Biocomput 2019;24:136–47.
11. Ibrahim NE, McCarthy CP, Shrestha S, Gaggin HK, Mukai R, Magaret CA, et al. A clinical, proteomics, and artificial intelligence-driven model to predict acute kidney injury in patients undergoing coronary angiography. Clin Cardiol 2019 Feb;42(2):292–8.
12. Kim IE, Sarkar IN. Racial Representation Disparity of Population-Level Genomic Sequencing Efforts. Stud Health Technol Inform 2019 Aug 21;264:974–8.
13. Kim K, Baik H, Jang CS, Roh JK, Eskin E, Han B. Genomic GPS: using genetic distance from individuals to public data for genomic analysis without disclosing personal genomes. Genome Biol 2019 Dec;20(1):175.
14. Liu P, Li H, Li S, Leung K-S. Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network. BMC Bioinformatics. 2019 Dec;20(1):408.
15. Marttinen M, Paananen J, Neme A, Mitra V, Takalo M, Natunen T, et al. A multiomic approach to characterize the temporal sequence in Alzheimer's disease-related pathology. Neurobiol Dis 2019 Apr;124:454–68.
16. Ruan J, Jahid Md-J, Gu F, Lei C, Huang Y-W, Hsu Y-T, et al. A novel algorithm for network-based prediction of cancer recurrence. Genomics 2019 Jan;111(1):17–23.
17. Wan N, Weinberg D, Liu T-Y, Niehaus K, Ariazi EA, Delubac D, et al. Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA. BMC Cancer 2019 Dec;19(1):832.

Correspondence to:
Malika Smaïl
Campus scientifique BP 239
LORIA Bâtiment B (équipe Capsid)
54506 Vandoeuvre-lès-Nancy Cedex
France
E-mail: malika.smail@loria.fr

Content Summaries of Best Papers Selected for the IMIA Yearbook Section on Bioinformatics and Translational Informatics

Béal J, Montagud A, Traynard P, Barillot E, Calzone L

Personalization of logical models with multi-omics data allows clinical stratification of patients

Front Physiol 24 Jan 2019;9:1965

This paper presents a new methodology - called PROFILE - to adapt generic logical models of cancer pathways to a particular biological sample (*i.e.,* patient's tumor). The authors use as a proof of concept a published model of cancer signaling pathways of breast cancer. The approach integrates mutation data, CNA (copy number alterations), and expression data to personalize the logical model to a patient's profile. The simulation of the resulting models (using the MaBoSS program) shows a good correlation with patient's subgrouping on NPI (Nottingham Prognostic Index) and survival time. This paper illustrates the potential of using logical modeling (and concepts of systems biology) for precision medicine as it can eventually facilitate the choice of patient-specific drug treatment thanks to a self-explanatory model.

Chen ML, Doddi A, Royer J, Freschi L, Schito M, Ezewudo M, Kohane IS, Beam A, Farhat M

Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in Mycobacterium tuberculosis resistance prediction

EBioMedicine 2019 May;43:356–69

Multidrug-resistant tuberculosis (MDR-TB) is still a public health challenge because of the lengthy current culture-based antimicrobial susceptibility testing due to in vitro growth of *mycobacterium tuberculosis* (MTB). As an alternative to recent molecular tests for MDR-TB criticized for low sensitivity and the small number of tested drugs, the authors propose a whole-genome sequencing approach to capture both common and rare mutations responsible for drug resistance. They use a neural architecture combining a wide and deep neural network (WDNN) compared to simpler classifiers such as logistic regression and random forests. They leverage whole-genome sequencing of the pathogen, as well as rare variants and known drug resistant variants to predict the resistance to 10 anti-tuberculosis drugs. They achieved AUCs over 93% for first-line drugs, and 89% for second-line drugs.

Kim K, Baik H, Jang CS, Roh JK, Eskin E, Han B

Genomic GPS: using genetic distance from individuals to public data for genomic analysis without disclosing personal genomes

Genome Biol 2019 Dec;20(1):175

Genomic global positioning system (GPS) applies the multilateration technique commonly used in the GPS to genomic data. This framework allows to calculate genetic distances from considered samples to reference samples (public data), and share this information with others. This sharing enables certain types of genomic analysis, such as identifying sample overlaps, close relatives, mapping to geographical origin without disclosing personal genomes. This innovative approach allows achieving a good balance between open genomic data sharing and privacy protection.

Marttinen M, Paananen J, Neme A, Vikram M, Takalo M, Natune T, Paldanius KMA, Mäkinen P, Bremang M, Kurki MI, Rauramaa T, Leinonen V, Soininen H, Haapasalo A, Pike I, Hiltunen M

A multiomic approach to characterize the temporal sequence in Alzheimer's disease-related pathology

Neurobiol Dis 2019;124:45468

This paper proposes a multi-omics and temporal sequence-based approach to provide a better understanding of the sequence of events leading to Alzheimer's Disease (AD). The authors coupled transcriptomic and phosphoproteomic data to determine the temporal sequence of changes in microRNA, protein, and phosphopeptide expression levels from human temporal cortical samples, with varying stages of the AD. This approach highlighted a significant sequence of key functions occurring at the considered stages of the disease, namely: *(i)* fluctuation in synaptic and mitochondrial functions as the earliest pathological events in brain samples with AD-related pathology, and *(ii)* the increased expression of inflammation and extracellular matrix-associated gene products. The authors made use of decision trees and random forests for identifying potential biomarkers predicting the disease degree.