Intersecting Pathways in Bioinformatics and Translational Informatics: A One Health Perspective on Key Contributions and Future Directions

Mary Lauren Benton¹, Scott McGrath², Section Editors for the IMIA Yearbook Section on Bioinformatics and Translational Informatics

¹ Department of Computer Science, Baylor University, USA
² CITRIS Health, University of California Berkeley, USA

Summary

Objectives: To identify and summarize the top bioinformatics and translational informatics (BTI) papers published in 2022 for the International Medical Informatics Association (IMIA) Yearbook 2023.

Methods: We conducted a comprehensive literature search to identify the top BTI papers, resulting in a set of ten candidate papers. The candidates were reviewed by the section co-editors and external reviewers to select the top three papers from 2022.

Results: From a total of 558 papers, we identified a final candidate list of ten BTI papers for peer-review. These papers apply new statistical frameworks and experimental designs to better capture individual variability in disease and incorporate data that captures differences between single cells and across environmental exposures. In addition, they highlight the importance of model generalization across diverse cohorts and scalability to large medical centers.

Conclusions: We note several important trends in the candidate top BTI papers this year, including a continued focus on developing accurate and scalable computational models to predict disease risk across diverse cohorts and new strategies to capture the molecular heterogeneity of disease.

Keywords

Bioinformatics; genomics; transcriptomics; machine learning; precision medicine

Yeamb Med Inform 2023.99-103
http://dx.doi.org/10.1055/s-0043-1768745

1 Introduction

The year 2022 found us in a crucial moment of reckoning with our interdependence with the environment and the myriad creatures that share our planet. As such, there is an opportunity to integrate One Health into the Bioinformatics and Translational Informatics (BTI) domain. One Health is an approach that aligns with the realization that the health of humans is intrinsically connected to the health of animals and our shared environment.

Work on pursuing the One Health framework, with an emphasis on the interconnectedness of human, animal, and environmental health, has occurred for a few years now. Trinh et al. expanded the One Health approach to include entire microbial communities, showing that environmental and animal microbiomes can influence human health [1]. Timme et al. highlighted the necessity of interoperable and open data platforms to facilitate easy sharing and building upon findings in One Health [2]. Lustgarten et al. presented a review on the state and potential of veterinary informatics, shedding light on the vital role it plays in bridging veterinary medicine, human medicine, and One Health initiatives despite current limitations [3]. These significant contributions underscore the continued development in bioinformatics and translational informatics focusing on an integrative One Health perspective.

For the 2022 BTI best papers, our focus shifted to this One Health theme. In the following synopsis, we will describe our methodology for selecting the top BTI papers of the year that have made significant contributions to the literature this year, including those that improve our understanding of the One Health concept. We will also discuss three important trends highlighted by these papers. We hope that these selections will stimulate further discussion and research into this critical area of health and well-being.

2 Methods

We conducted a comprehensive literature search to identify candidates for the top bioinformatics and translational informatics (BTI) synopsis. To identify articles specifically from 2022, we restricted the query to only return articles with electronic publication dates between January 1, 2022, and December 31, 2022. In addition, we focused the query on the most relevant journals for BTI: Journal of the American Medical Informatics Association (JAMIA), Journal of Biomedical Informatics (JBI), PLoS Computational Biology, Bioinformatics, BMC Bioinformatics, BMC Systems Biology, Nature, Nature Genetics, Nature Biotechnology, Nature Methods, Science, Science Translational Medicine, Clinical Pharmacology and Therapeutics, New England Journal of Medicine, Journal of the American Medical Association (JAMA), Lancet, PLoS Genetics, and Cell.

Using the constraints above, we queried PubMed using the following Medical Subject Headings (MeSH) terms and their deriv-
3 Trends

The ten candidate best papers for 2022 illustrate recent efforts towards improving the representation of individual and molecular variability in models of disease and exemplify previous trends in BTI such as computational models for disease prevention. We provide a brief discussion of the main themes we discovered while evaluating the top BTI papers. These themes highlight important areas of research in this field where we expect to see continued growth in the coming years.

3.1 Applying Proactive Models of Risk to Prevent Disease and Enhance Patient Outcomes

The development and clinical implementation of risk prediction models continues to be a focal point of BTI research. These models apply advanced machine learning techniques to assess potential risk factors and disease subtypes to provide clinically actionable insights.

Grazioli et al. [4] propose a Multimodal Variational Information Bottleneck (MVIB) to integrate heterogeneous data modalities into a single predictive framework. The joint encoding leveraged information about microbial species abundance, microbial strain-level markers, and metabolites to generate a predictive model. This approach could potentially offer a proactive model of risk prevention by enabling early disease prediction, thereby contributing to improved patient outcomes. In particular, the incorporation of microbiome data to improve disease prediction for a range of common diseases demonstrates the importance of using One Health concepts to understand human health. Moving beyond a strictly human-centric approach has the potential to improve the impact of precision health initiatives. The survey paper in the 2022 IMIA Yearbook for the BTI section written by Cooper et al. [7] explores the interaction between the microbiome and nutrition in more detail.

The work by Wang et al. [8] enhances our understanding of the genetic factors that influence physical activity and sedentary behavior; importantly, their study provides insights that may contribute to the development of personalized interventions to promote physical activity and reduce sedentary behavior. In turn, this can help prevent disease and enhance patient outcomes. This is particularly relevant as it recognizes the interconnection between the health of individuals and their environments.

The last three papers in this section build upon a broad foundation of disease risk prediction models in the literature. Of note, each of these papers include a discussion about the performance of such approaches when applied to diverse human populations. In Wang et al. [9], the authors demonstrate the potential of genomic data in personalized disease risk assessment, specifically for cancer. By improving our understanding of genetic risk factors and their variability among different populations, healthcare providers can better predict an individual’s risk of developing cancer and take proactive measures to prevent disease progression. However, the findings also highlight the importance of inclusivity and diversity in genetic research to ensure the benefits of such risk models can be realized by all populations. Mishra et al. [10] conduct a large cross-ancestry genome-wide association study (GWAS) for stroke, and

Table 1: Selection of best papers for the 2022 IMIA Yearbook of Medical Informatics for the Bioinformatics and Translational Informatics section. The articles are listed in alphabetical order by the first author’s surname.

<table>
<thead>
<tr>
<th>Section</th>
<th>Bioinformatics and Translational Informatics</th>
</tr>
</thead>
</table>
use their results to identify new risk loci and potential drug targets. Ultimately, the authors generate high performing cross-ancestry and ancestry-specific polygenic risk scores that could improve predictions for individuals with diverse ancestries. Continuing with this theme, Wang et al. [11] conclude that implementing preemptive and sequence-based pharmacogenomics prescribing is almost universally applicable, providing a proactive model for disease prevention and enhancing patient outcomes. Additionally, it showcased the efficiency of using health care resources, a critical factor in the scalability of such models in a clinical setting. This proactive approach enables personalized medicine and a more efficient use of health care resources.

3.2 Capturing the Environmental and Molecular Heterogeneity of Disease

Building on the availability of increasingly large datasets and greater emphasis on single-cell genomics assays, many of this year’s BTI papers focus on quantifying the environmental and molecular heterogeneity of disease. By capturing and cataloging this variability, we will be able to increase the precision of future research. For example, Kuppe et al. [5] generated a comprehensive spatial and temporal map of cardiac cell types in myocardial infarction patients and controls. This approach recognizes that the disease process in myocardial infarction is not homogeneous; instead, it involves complex interactions between different cell types and their environment at different stages of the disease. The temporal and single-cell approaches allowed the authors to capture the diversity of cellular responses within the heart tissue during disease development and progression. By identifying sets of differentially expressed genes in cells that mark the border between injured and uninjured tissues, the authors characterized the unique profiles of remodeled versus functional myocardium. This provides insights into the heterogeneity of the disease process, reflecting the distinct molecular responses of cells to the injury.

Similarly, Robbe et al. [12] generated high resolution maps of copy-number alterations and other genetic variation from DNA sequences of 485 patients with chronic lymphocytic leukemia. This catalog of variation captures distinct patient subgroups within the dataset that associate with disease outcome and differential treatment responses. Understanding the heterogeneity within a single diagnostic category will improve patient outcomes by allowing for more precise treatment options.

There is also environmental heterogeneity for individuals that is not well-captured by traditional population-based estimates. Howe et al. [13] sought to address this limitation in GWAS by comparing the results from a large family-based GWAS to existing population-based estimates. They considered a range of behavioral, anthropometric, and molecular traits for their meta-analysis, finding that many phenotypes had smaller GWAS estimates in the family-based analyses. The differences in GWAS estimates were greater for anthropometric and behavioral phenotypes than for the molecular phenotypes (e.g., cholesterol levels), suggesting that leveraging different GWAS structures can better estimate associations for phenotypes influenced by demography and indirect effects. Future insights will require continued data collection for both families and unrelated individuals, as well as the application of multiple GWAS frameworks to compute accurate estimates.

3.3 Considering the Scalability of Computational Models in Clinical Settings

A final theme from the BTI candidate papers this year is an emphasis on the scalability of computational and risk prediction models for use in clinical settings. Some papers, including Wang et al. [9] and Wang et al. [11], evaluate the deployment of precision models at scale in existing academic medical centers. Others, such as Weitz et al. [6] and Tran et al. [14], develop models that can more efficiently leverage available data to produce accurate results.

For example, the convolutional neural network (CNN) trained by Weitz et al. [6] predicts gene expression directly from hematoxylin and eosin-stained whole slide images in samples from patients with prostate cancer. The final CNN predicts gene expression levels for clusters of co-expressed genes to improve performance and can be used to generate tumor scores normally derived from RNA-sequencing. This showcases a scalable solution that allows for the quantification of gene expression phenotypes directly from imaging, which is particularly beneficial in settings where full molecular phenotyping would be otherwise unattainable. This scalability is crucial for the broad application of such models in clinical settings, thereby facilitating more personalized and effective patient care.

Tran et al. [14] apply semi-supervised machine learning approaches to classify central nervous system tumors into known subclasses based on DNA methylation patterns. This allows the authors to leverage a much larger collection of unlabeled methylation data to generate pseudo-labels that can improve classification performance without requiring the expensive clinical resources required to generate fully labeled data. They find that semi-supervised learning accurately classifies tumor classes and can be used to boost performance in supervised learning models. In the future, this approach could be applied to other clinical classification problems where labeled data are limited or to leverage additional public datasets to improve performance.

4 Conclusion

Our efforts to identify the top BTI papers for 2022 have revealed a sustained focus on the development of accurate and scalable computational models for disease risk prediction across diverse cohorts, as well as novel approaches for capturing the molecular heterogeneity of disease. These papers embody the ongoing trend in BTI research towards increasing the representation of individual and molecular variability in disease models and leveraging computational models for proactive disease prevention.
We saw work where the utility of advanced machine learning techniques in assessing potential risk factors and disease subtypes in an effort generate clinically actionable insights [4, 8-10]. Moreover, they exemplify the importance of inclusive genetic research and the potential of pharmacogenomics in personalized medicine.

Other papers [5, 12, 13] accentuated the importance of capturing the environmental and molecular heterogeneity of disease. These works paved the way for more precise future research by quantifying and cataloging disease variability through single-cell genomics assays, high-resolution genetic variation maps, and different GWAS structures.

Lastly, the scalability of computational models for use in clinical settings is a key theme in the papers [6, 9, 14], which illustrate the potential for implementing precision models at scale in existing academic medical centers and how machine learning models can efficiently leverage available data to produce accurate results, facilitating more personalized and effective patient care.

These overarching themes seen across our selection of candidate papers exemplify the continued progress in this field. With such robust foundations in place, we look forward to witnessing further breakthroughs and innovative advancements in these areas in the future.

Acknowledgements
We are grateful to Martina Hutter and the other IMIA Yearbook editors for their support. We would also like to thank the peer reviewers for their participation in the selection of the top papers for the IMIA Yearbook.

References


Correspondence to:
Mary Lauren Benton
One Bear Place #97141
Waco, TX, 76798, USA
E-mail: marylauren_benton@baylor.edu

Scott McGrath
One Bear Place #97141
Waco, TX, 76798, USA
E-mail: scottmgrath@baylor.edu
Appendix: Summary of Best Papers Selected for the IMIA Yearbook 2023, Bioinformatics and Translational Informatics


Microbiome-based disease prediction with multimodal variational information bottlenecks


In this paper the authors addressed the untapped potential of multimodal machine learning in disease prediction by leveraging the diagnostic potential of gut microbial profiling. Traditionally, microbial species-relative abundances or strain-level markers extracted through shotgun metagenomic sequencing have been separately assessed in disease prediction models. Grazioli et al.’s innovative approach involved the development of a Multimodal Variational Information Bottleneck (MVIB), a deep learning model capable of integrating multiple heterogeneous data modalities into a single predictive framework. MVIB was devised to offer both efficient performance and interpretability. The model creates a joint stochastic encoding of different input data types, thereby integrating a plethora of disease-related markers. Through evaluating the model on 11 publicly available disease cohorts, the researchers achieved high classification performance, with areas under the ROC curve (AUCs) ranging from 0.80 to 0.95 for five cohorts, while maintaining medium performance for the remainder. The versatility of MVIB was demonstrated through cross-study generalization experiments, where training and testing were performed on different cohorts for the same disease. The results were comparable to a benchmark Random Forest model. Moreover, the scalability of MVIB was underscored by its ability to incorporate a third input modality, metabolomic data derived from mass spectrometry, without compromising efficiency or performance.


Spatial multi-omic map of human myocardial infarction

Nature 2022 Aug;608(7924):766-77. doi: 10.1038/s41586-022-05060-x

In this paper the authors leveraged single-cell -omics profiling to generate a temporal and spatial map of cardiac cell types in myocardial infarction patients and controls. Remodeling of cardiac tissues after myocardial infarction significantly contributes to late-stage mortality and is not well-addressed by current therapies. Limiting the negative impacts of cardiac remodeling on patients will require the development of new therapeutic approaches enabled by a more precise molecular understanding of the cell types involved in the process. The authors combined single-cell approaches, including single nucleus RNA sequencing and single nucleus assay for transposase-accessible chromatin sequencing, with spatial transcriptomics in 31 samples that spanned multiple clinical timepoints. From these data, they were able to identify major cell types in heart tissue and map these to particular histomorphological regions. Integrating these multi-omics data identified sets of differentially expressed genes in cells that marked the border between injured and uninjured tissues, and characterized the profiles of remodeled versus functional myocardium. The resulting multi-modal atlas of the human heart generates hypotheses that can facilitate new therapeutic advances and provides an important resource for the research community.


The authors developed a novel machine learning to predict gene expression directly from haematoxylin and eosin-stained whole slide images (WSIs) in samples from patients with prostate cancer. Molecular phenotyping, especially using gene expression data, is an increasingly important approach to characterize patient samples, compute clinical scores from biomarkers, and implement precision care. However, assays to generate the required data to conduct molecular phenotyping and compute clinical scores are costly to implement on a large scale. Previous work has shown that molecular phenotypes, including gene expression, can be accurately predicted from histopathology WSIs and these WSIs are routinely collected and digitized during care. In their study, Weitz et al. trained a convolutional neural network (CNN) using prostate cancer samples from TCGA PRAD to predict gene expression levels for clusters of co-expressed genes. They then applied their predictions to compute a predicted cell cycle progression (CCP) score, which correlates with cancer aggressiveness, recurrence, and mortality. The authors found a significant correlation between predicted gene expression and transcript levels measured by RNA-sequencing in more than 6,600 genes. The significantly predicted genes were enriched in pathways relevant to prostate cancer, including those involved in DNA replication, cell cycle, and metabolism. Weitz et al. computed a CCN-based CCP score using these results. The CNN-based CCP scores were prognostic in their preliminary analysis, and were correlated with tumor grade to a similar degree as the RNA-seq-based CCP. Ultimately, this work suggests that deep learning models could provide a scalable solution to quantify gene expression phenotypes directly from imaging, particularly in settings where full molecular phenotyping would be otherwise unattainable.