© 2016 IMIA and Schattauer GmbH

Findings from the Section on Bioinformatics and Translational Informatics

H. Dauchel, T. Lecroq, Section Editors for the IMIA Yearbook Section on Bioinformatics and Translational Informatics
Normandie Univ., UNIROUEN, LITIS, Rouen, France

Summary

Objectives: To summarize excellent current research and propose a selection of best papers published in 2015 in the field of Bioinformatics and Translational Informatics with application in the health domain and clinical care.

Method: We provide a synopsis of the articles selected for the IMIA Yearbook 2016, from which we attempt to derive a synthetic overview of current and future activities in the field. As last year, a first step of selection was performed by querying MEDLINE with a list of MeSH descriptors completed by a list of terms adapted to the section. Each section editor has evaluated separately the set of 1,566 articles and the evaluation results were merged for retaining 14 articles for peer-review.

Results: The selection and evaluation process of this Yearbook's section on Bioinformatics and Translational Informatics yielded four excellent articles focusing this year on data management of large-scale datasets and genomic medicine that are mainly new method-based papers. Three articles explore the high potential of the re-analysis of previously collected data, here from The Cancer Genome Atlas project (TCGA) and one article presents an original analysis of genomic data from sub-Saharan Africa populations.

Conclusions: The current research activities in Bioinformatics and Translational Informatics with application in the health domain continues to explore new algorithms and statistical models to manage and interpret large-scale genomic datasets. From population wide genome sequencing for cataloging genomic variants to the comprehension of functional impact on pathways and molecular interactions regarding a given pathology, making sense of large genomic data requires a necessary effort to address the issue of clinical translation for precise diagnostic and personalized medicine.

Keywords

Translational medical research, computational biology, gene genome expression, genome, medical informatics

Yearb Med Inform 2016:207-10
Published online November 10, 2016

Introduction

As already mentioned [1], main ongoing works on Bioinformatics and Translational Informatics are related to Genome Medicine *i.e.* the identification from data of genes and mutations underlying human diseases and by pursuing the research on "bedside to bench", the management of "Big Data" and personalized medicine. Actually, the availability of large-scale genomic data from Next Generation Sequencing (NGS) experiments allows the analysis of the disease-related biomolecular networks, which are expected to couple genotypes and disease phenotypes to determine the biological mechanisms of complex diseases. This trend that appears a few years ago remains very pregnant. We assist now to projects consisting in the exploitation of new or previously collected data for providing new decision support with applied objectives.

In [2], the authors use a combination of high throughput sequencing (low coverage genome and deep exome sequencing) and dense hybridization technology for studying genome expression and genotyping. In [3] the authors describe the integrative analysis of 111 reference human epigenomes and elucidate how epigenetic processes contribute to human biology and disease. In [4], the authors have reconstructed the genomes of 2,504 individuals from 26 populations using a combination of low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping. They characterized a broad spectrum of genetic variation, in total over 88 million variants (84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants), all phased onto high-quality haplotypes. They describe the distribution of genetic variation across the global sample, and discuss the implications for common disease studies. The authors of [5] present a computational pipeline (ICRmax) developed to efficiently identify a minimal set of reliable interchromosomal rearrangements (ICR) from low coverage sequenced tumor genomes without the need to sequence a matched normal genome, reducing the sequencing cost and creating an opportunity to implement the use of personalized biomarkers in the routine clinical management of solid tumors. They validated their method on solid rectal tumors and simulated data. It results in an average accuracy of 68% for ICR identification. In [6] the authors present a new motif-searching algorithm, which is able to identify common motif types from the cancer networks and signal transduction networks. In [7] the authors analyzed 2158 full cancer transcriptomes from 163 diverse cancer entities to identify co-regulated genes. They found two novel oncogenes in hepatocellular carcinoma and in several tumors. TarPred [8] enables to find ligand–target interactions. It is based on a library of 533 individual targets with 179,807 active ligands. Given a compound structure, it shows the 30 best interacting targets and for each it also gives the three most similar ligands and the disease indications linked with the target. In [9] the authors realize the first comprehensive comparison of the disease genes and drug targets in the context of the human protein interactome. The authors of [10] develop a new machine learning-based method using epigenomics data for gene expression prediction in lung cancer.

Best Paper Selection

The best paper selection for the section Bioinformatics and Translational Informatics follows a generic method, commonly used in all the sections of the IMIA Yearbook 2016. As for the last three years, the search is performed on MEDLINE by querying PubMed. The Boolean query includes MeSH descriptors related to the domain of computational biology and medical genetics with a restriction to international peer-reviewed journals. Only original research articles published in 2015 (from 01/01/2015 to 12/31/2015) were considered; we excluded the publications types reviews, editorials, comments, letters to the editors …etc. We limited the search on the major MeSH descriptors to avoid a large set of articles and we completed it by non-MeSH terms searched on the titles and abstracts of the articles. However, there was no restriction on the top international peer-reviewed journals of the Bioinformatics and Translational Informatics section and Medicine (by using the 2-year Impact Factors). This year, the PubMed query yielded a set of 1,566 articles (vs. 1,594 last year) that were evaluated separately by each section editor (HD & TL) using the BibReview tool and the generic method described by Lamy et al. in [11]. BibReview takes into entry a PubMed file (in XML format) and shows all metadata for each article. A user can tag, thanks to the interface, the articles as "Accepted", "To Revise", "Conflict" or "Reject" (on text, abstract or title). The results of several reviewers can be merged and the results can be filtered. This year, only 5 articles are tagged "Accepted" in common by the two section editors. Each section editor proposed a separate list of accepted papers to compose a set of 14 common articles for peer-review. Ten reviewers, specialized in the field, consider the 14 articles as candidates for inclusion. The best papers are ranked according to criteria of: topic significance, coverage of literature, quality of research, results and presentation [12]. Finally, after evaluation four papers [13, 14, 15, 16] were retained by the reviewers.

The four papers retained focus on the management of large-scale datasets and medical genomic that are also mainly method and tool-based papers. In the first article [13], Gurdasani et al present a statistical analysis of genomic data from sub-Saharan Africa populations giving the genetic structure and identifying new loci under selection, including for malaria and hypertension. In the second article [14] Korhauer and Kendziorski describe MADGiC for Model-based Approach for identifying Driver Genes in Cancer which is a new unified empirical Bayesian model-based approach illustrated in a targeted re-analysis of ovarian and lung cancer data from The Cancer Genome Atlas project (TCGA). In the third article [15], Leiserson et al. present a pan-cancer analysis of various somatic mutation data from TCGA using HotNet2 a novel algorithm to find mutated sub-networks through protein complexes and pathways. In the fourth article [16], Zhu et al. present Zodiac a tool based on Bayesian graphical models that realizes a big-data integrative analysis of multiple genomic features from pan-cancer TCGA data and which generates new knowledge of molecular interactions in cancer.

A brief content of each one can be found in the appendix of this synopsis.

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2016 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
▪ Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, Karthikeyan S, Iles L, Pollard MO, Choudhury A, Ritchie GR, Xue Y, Asimit J, Nsubuga RN, Young EH, Pomilla C, Kivinen K, Rockett K, Kamali A, Doumatey AP, Asiki G, Seeley J, Sisay-Joof F, Jallow M, Tollman S, Mekonnen E, Ekong R, Oljira T, Bradman N, Bojang K, Ramsay M, Adeyemo A, Bekele E, Motala A, Norris SA, Pirie F, Kaleebu P, Kwiatkowski D, Tyler-Smith C, Rotimi C, Zeggini E, Sandhu MS. The African Genome Variation Project shapes medical genetics in Africa. Nature 2015 Jan 15;517(7534):327-32.
▪ Leiserson MD, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JL, Papoutsaki A, Kim Y, Niu B, McLellan M, Lawrence MS, Gonzalez-Perez A, Tamborero D, Cheng Y, Ryslik GA, Lopez-Bigas N, Getz G, Ding L, Raphael BJ. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet 2015 Feb;47(2):106-14.
▪ Korthauer KD, Kendziorski C. MADGiC: a model-based approach for identifying driver genes in cancer. Bioinformatics 2015 May 15;31(10):1526-35.
▪ Zhu Y, Xu Y, Helseth DL Jr, Gulukota K, Yang S, Pesce LL, Mitra R, Müller P, Sengupta S, Guo W, Silverstein JC, Foster I, Parsad N, White KP, Ji Y. Zodiac: A Comprehensive Depiction of Genetic Interactions in Cancer by Integrating TCGA Data. J Natl Cancer Inst 2015 May 8;107(8).

Conclusions and Outlook

During the past years there was a large number of articles dealing with proteomic but most of these papers lacked of computational methods. In the next few years there will probably a need for new computational resources for dealing with large proteomics data. A large number of important articles deals with cancer studies. We could see the emergence of pan-cancer studies such as [15], rather than studies focusing on a single type of cancer. The challenge is not only to analyze new big dataset but also to exploit various massive previously collected omics data. There exist thus two directions: one consists of studying broad spectra of information while the other is restricted to more targeted datasets. There is still a challenge for reducing the sequencing costs and create an opportunity to implement the use of personalized biomarkers in the routine clinical management. Furthermore data mining techniques may improve in the next years and it will then be possible to reconsider previously collected data. This also reinforce the need for efficiently share omics data.

Acknowledgements

We would like to acknowledge the valuable support of Martina Hutter and all the reviewers in the evaluation process of the section Bioinformatics and Translational Informatics. We also would like to greatly thank the IMIA Yearbook 2016 editors and managing editors Marie-Christine Jaulent, Brigitte Séroussi and Christoph U Lehmann.

References

1. Soualmia LF, Lecroq T. Bioinformatics Methods and Tools to advance Clinical Care. Yearb Med Inform 2015:170-3.
2. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature 2015;526(7571):68-74.
3. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M et al, Integrative analysis of 111 reference human epigenomes, Nature 518(7539)317-30 2015.
4. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature 2015;256(7571):75-81.
5. Donnard ER, Carpinetti PA, Navarro FC, Perez RO, Habr-Gama A, Parmigiani RB. ICRmax: an optimized approach to detect tumor-specific interchromosomal rearrangements for clinical application. Genomics 2015;105(5-6):265-72.
6. Hsieh WT, Tzeng KR, Ciou JS, Tsai JJ, Kuruban-jerdjit N, Huang CH, et al. Transcription factor and microRNA-regulated network motifs for cancer and signal transduction networks. BMC Syst Biol 2015; 9(Suppl. 1):S5.
7. Itzel T, Scholz P, Maass T, Krupp M, Marquardt JU, Strand S, et al. Translating bioinformatics in oncology: guilt-by-profiling analysis and identification of KIF18B and CDCA3 as novel driver genes in carcinogenesis. Bioinformatics 2015;31(2):216-24.
8. Liu X, Gao Y, Peng J, Xu Y, Wang Y, Zhou N, et al. TarPred: a web application for predicting therapeutic and side effect targets of chemical compounds. Bioinformatics 2015;31(12):2049-51.
9. Sun J, Zhu K, Zheng W, Xu H. A comparative study of disease genes and drug targets in the human protein interactome. BMC Bioinformatics 2015;16(Suppl. 5):S1.
10. Li J, Ching T, Huang S, Garmire LX. Using epigenomics data to predict gene expression in lung cancer. BMC Bioinformatics 2015;16(Suppl. 5):S10.
11. 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.
12. Ammenwerth E, Wolff AC, Knaup P, Ulmer H, Skonetzki S, van Bemmel JH, et al. Developing and evaluating criteria to help reviewers of biomedical informatics manuscripts. J Am Med Inform Assoc 2003;10(5):512-4.
13. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African Genome Variation Project shapes medical genetics in Africa. Nature 2015;517(7534)327-32.
14. Korthauer KD, Kendziorski C. MADGiC: a model-based approach for identifying driver genes in cancer. Bioinformatics 2015;31(10):1526-35.
15. Leiserson MD, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JL, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet 2015;47(2):106-14.
16. Zhu Y, Xu Y, Helseth DL Jr, Gulukota K, Yang S, Pesce LL, et al. Zodiac: A Comprehensive Depiction of Genetic Interactions in Cancer by Integrating TCGA Data. J Natl Cancer Inst 2015;107(8):djv129.

Correspondence to:
Dr Hélène Dauchel
Normandie Univ., UNIROUEN, LITIS
76000 Rouen, France
Tel : +33 235 146 389
E-mail: Helene.Dauchel@univ-rouen.fr

Appendix: Content Summaries of Selected Best Papers for the IMIA Yearbook 2016, Section Bioinformatics and Translational Informatics

Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, Karthikeyan S, Iles L, Pollard MO, Choudhury A, Ritchie GR, Xue Y, Asimit J, Nsubuga RN, Young EH, Pomilla C, Kivinen K, Rockett K, Kamali A, Doumatey AP, Asiki G, Seeley J, Sisay-Joof F, Jallow M, Tollman S, Mekonnen E, Ekong R, Oljira T, Bradman N, Bojang K, Ramsay M, Adeyemo A, Bekele E, Motala A, Norris SA, Pirie F, Kaleebu P, Kwiatkowski D, Tyler-Smith C, Rotimi C, Zeggini E, Sandhu MS

The African Genome Variation Project shapes medical genetics in Africa

Nature 2015 Jan 15;517(7534):327-32

The aim of the present article is to capture the African genetic diversity. This is done thanks to a human population study on a large-scale dataset. The African Genome Variation Project (AGVP) is an international collaboration and has the objective to study human origins and disease susceptibility. The used dataset consists of 1,481 dense genotypes and whole genome sequences (WGS) from 320 individuals across sub-Saharan Africa (SSA). A statistical analysis of population enhances the genetic structure of SSA population and identifies new loci under selection, including for malaria and hypertension. This also results in the design of a pan-African genotype array to capture the common genetic variations in Africa for further GWAS studies (large-scale medical genomic studies across the region, as well as studies of population history and evolution).

Importantly, the AGVP has evolved to help develop local resources for public health and genomic research, including strengthening research capacity, training, and collaboration across the region. This constitutes a proof-of-concept for the utility of geographically widespread genetic data within Africa to conduct medical genomic studies in Africa.

Leiserson MD, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JL, Papoutsaki A, Kim Y, Niu B, McLellan M, Lawrence MS, Gonzalez-Perez A, Tamborero D, Cheng Y, Ryslik GA, Lopez-Bigas N, Getz G, Ding L, Raphael BJ

Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes

Nat Genet 2015 Feb;47(2):106-14

The authors realize a pan cancer study of mutated networks. They use a dataset of Single Nucleotide Variations, small indels and Copy Number Variations from 3281 mutated networks from 12 different cancer types from The Cancer Genome Atlas (TCGA). They use a new algorithm called HotNet2 (HotNet diffusion oriented sub-networks) and were able to identify rare combination of somatic mutations across multiple cancers. They identify 14 mutated sub-networks. Some were already known to have an effect in cancers but the role in cancer of others had not been clearly identified before. Thus compare to other algorithms HotNet2 has a higher sensitivity and specificity on both real and simulated data. This study represents the largest network analysis of somatic aberrations across multiple cancer types.

Korthauer KD, Kendziorski C

MADGiC: a model-based approach for identifying driver genes in cancer

Bioinformatics 2015 May 15;31(10):1526-35

MADGiC for Model-based Approach for identifying Driver Genes in Cancer is a new unified empirical Bayesian model-based approach that uses both frequency and functional impact criteria (potential pathogenicity of the individual variations) but also incorporates a number of parameters to improve the background model (mutation type, gene-specific features: genomic context, replication timing region, expression of the gene...) MADGiC is implemented in R and is open source. It has been used for an analysis of ovarian and lung cancer data from The Cancer Genome Atlas (TCGA) project. Both simulation studies and case studies show that this new method can achieved better results than other known methods.

Zhu Y, Xu Y, Helseth DL Jr, Gulukota K, Yang S, Pesce LL, Mitra R, Müller P, Sengupta S, Guo W, Silverstein JC, Foster I, Parsad N, White KP, Ji Y

Zodiac: A Comprehensive Depiction of Genetic Interactions in Cancer by Integrating TCGA Data

J Natl Cancer Inst 2015 May 8;107(8)

Zodiac realizes a big-data integrative analysis on multimodal data form The Cancer Genome Atlas (TCGA) project with a goal to generate new knowledge of molecular interactions in cancer (gene-gene interactions, transcriptional regulation, and other types of molecular interplays in cancer). This is done by using statistical models. Actually Zodiac merges a prior interaction graph and TCGA data to produce a posterior graph. It results in a whole-genome and pairwise interaction map, which contains intragenic and intergenic interactions of all pairs of genes in cancer-molecular interactions for about 200 million pairs of genes in a publicly comprehensive database together with a search engine system. In addition, Zodiac allow users to customize the prior networks and update the genetic pathways of their interest.