Bioinformatics Methods and Tools to Advance Clinical Care

Findings from the Yearbook 2015 Section on Bioinformatics and Translational Informatics

L. F. Soualmia, T. Lecroq, Section Editors for the IMIA Yearbook Section on Bioinformatics and Translational Informatics

Normandie Univ., University of Rouen, NormaSTIC FR CNRS 3638, IRIB and LITIS EA 4108, Information Processing in Biology & Health, Saint Étienne du Rouvray, France

Summary

Objectives: To summarize excellent current research 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 2015, 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 1,594 articles and the evaluation results were merged for retaining 15 articles for peer-review.

Results: The selection and evaluation process of this Yearbook's section on Bioinformatics and Translational Informatics yielded four excellent articles regarding data management and genome medicine that are mainly tool-based papers. In the first article, the authors present PPISURV a tool for uncovering the role of specific genes in cancer survival outcome. The second article describes the classifier PredictSNP which combines six performing tools for predicting disease-related mutations. In the third article, by presenting a high-coverage map of the human proteome using high resolution mass spectrometry, the authors highlight the need for using mass spectrometry to complement genome annotation. The fourth article is also related to patient survival and decision support. The authors present datamining methods of large-scale datasets of past transplants. The objective is to identify chances of survival.

Conclusions: The current research activities still attest the continuous convergence of Bioinformatics and Medical Informatics, with a focus this year on dedicated tools and methods to advance clinical care. Indeed, there is a need for powerful tools for managing and interpreting complex, large-scale genomic and biological datasets, but also a need for user-friendly tools developed for the clinicians in their daily practice. All the recent research and development efforts contribute to the challenge of impacting clinically the obtained results towards a personalized medicine.

Keywords

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

Yearb Med Inform 2015;10:170-3
http://dx.doi.org/10.15265/IY-2015-026
Published online August 13, 2015

Introduction

As already mentioned [1-2], 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" [3], the management of "Big Data" [4] and personalized medicine [5]. 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. Akan et al. [6] presented a study, in which whole genome and transcriptome data for three human cancer cell lines were analyzed in conjunction with protein data. The authors demonstrate the advantage for integrative analysis for identifying tumor-related genes. Among several results, another direct use of these high throughput technologies in patient cares could help to diagnose cancer without biopsy [7]. Analogously to NGS, mass-spectrometry allows proteomic studies and characterization [8] which carry biological information that is not accessible by genomics [9]. Until recently, few efforts characterize the human proteome because of the non-publicly-availability of the proteomic data. ProteomicsDB [10] is a in-memory database designed for the real-time analysis of big data (https://www.proteomicsdb.org). Wilhelm et al. [10] present also a draft of the human proteome assembled using disparate but huge high quality proteomic data. Similarly to the human genome projects [11-12-13], an issue is to address proteome coverage and resolution. Kim et al. [14] developed also a draft map of the human proteome using high-resolution mass spectrometry. They identified proteins encoded by 17,294 human genes, accounting for 84% of the annotated protein-coding genes in the human genome. On the other hand, several bioinformatics tools and methods are developed to advance clinical care by studying disease-related genes. For example, PPISURV [15] is a free online datamining tool that correlates expression of an input gene interaction with cancer survival by employing several public databases (http://www.bioprofiling.de/PPISURV). The BioMet Toolbox [16] provides a web-user interface for metabolic pathways and omics analysis. PredictSNP [17] is a classifier for predicting disease related-mutations which user-friendly web interface enables the freely access to several prediction tools but also to datasets.

Method

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 2015. As for the last two 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 2014 (from 01/01/2014 to 12/31/2014) 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).

Results

This year, the PubMed query yielded a set of 1,594 articles (vs. 1,851 last year) that were evaluated separately by each section editor (LFS & TL) using the BibReview tool and the generic method described by Lamy et al. in [18]. 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 top-5 accepted papers to compose a set of 15 articles for peer-review. Ten reviewers, specialized in the field, consider the 15 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 [19]. Finally, after evaluation four papers [14-15-17-20] were retained by the reviewers.

In the first article [15], Antonov et al. present PPISURV, a free online datamining tool which correlates the expression, not only of a single gene, but of the interactome (the expression of the gene interaction partners) of an input gene with the survival rates of patients suffering from cancer. All the processed patient datasets are publicly available. In the second article, [17] Bendl et al. describe the classifier PredictSNP which combines six performing tools for predicting disease-related mutations. The third article [14] concerns the work of Kim et al. in which the authors present a high-coverage map of the human proteome using high-resolution mass spectrometry. The authors highlight the need for using mass spectrometry to complement genome annotation. The large human proteome catalogue is an interactive web-based resource (http://www.humanproteomemap.org). In the fourth article [20], Taati el al. perform datamining in large datasets of records related to bone marrow cell transplantation. The objective is to identify the patients with the highest chance of survival. The method can be involved in a decision support system.

Table 1 lists the four papers selected as best papers for the section Bioinformatics and Translational Informatics. 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 2015 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
▪ Antonov AV, Krestyaninova M, Knight RA, Rodchenkov I, Melino G, Barlev NA. PPIsurv: a novel bioinformatics tool for uncovering the hidden role of specific genes in cancer survival outcome. Oncogene 2014 Mar 27; 33(13):1621-8.
▪ Bendl J, Stourac J, Salanda O, Pavelka A, Wieben ED, Zendulka J, Brezovsky J, Damborsky J. PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. PLoS Comput Biol. 2014 Jan;10(1):e1003440.
▪ Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, Thomas JK, Muthusamy B, Leal-Rojas P, Kumar P, Sahasrabuddhe NA, Balakrishnan L, Advani J, George B, Renuse S, Selvan LD, Patil AH, Nanjappa V, Radhakrishnan A, Prasad S, Subbannayya T, Raju R, Kumar M, Sreenivasamurthy SK, Marimuthu A, Sathe GJ, Chavan S, Datta KK, Subbannayya Y, Sahu A, Yelamanchi SD, Jayaram S, Rajagopalan P, Sharma J, Murthy KR, Syed N, Goel R, Khan AA, Ahmad S, Dey G, Mudgal K, Chatterjee A, Huang TC, Zhong J, Wu X, Shaw PG, Freed D, Zahari MS, Mukherjee KK, Shankar S, Mahadevan A, Lam H, Mitchell CJ, Shankar SK, Satishchandra P, Schroeder JT, Sirdeshmukh R, Maitra A, Leach SD, Drake CG, Halushka MK, Prasad TS, Hruban RH, Kerr CL, Bader GD, Iacobuzio-Donahue CA, Gowda H, Pandey A. A draft map of the human proteome. Nature 2014 May 29; 509(7502):575-81.
▪ Taati B, Snoek J, Aleman D, Ghavamzadeh A. Data mining in bone marrow transplant records to identify patients with high odds of survival. IEEE J Biomed Health Inform 2014 Jan; 18(1):21-7.

Conclusions and Outlook

The current research activities still attest the continuous convergence of Bioinformatics and Medical Informatics. The availability of genomic data from 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. The clinical translation of the bioinformatics results to an individual patient, which is poised to personalized medicine, is one of the challenges of the field. There is a need for powerful tools for managing efficiently and interpreting complex data, but also a need for user-friendly tools developed for the clinicians in their daily practice (identifying gene cancer cells for therapy, inherited cancer risk, risk factors for rare and common diseases, personalizing the choice of drug and dosage …etc). The tools and datasets presented in the four selected papers of the section are publicly available, which allow reproducible processing on similar datasets. Electronic Health Records are a valuable source of information for knowledge discovery. The integration of genomic data into Electronic Health Records as well as the development of genomic tests and their increasing clinical utility can change the medical decision process [21]. The Electronic Medical Records and Genomics consortium [22] share biobanks and genotype data to process high-throughput phenotyping of patient cohorts. The recent research and development efforts are contributing to the challenge of impacting clinically the results and even going towards a personalized medicine in the near future. Some frameworks allow this translation and problems, related to the availability of publicly available datasets that we have mentioned last year [2], seem now to be partially resolved. A valuable research breakthrough research is the human proteome draft-map, using mass spectrometry, which will complement the already

available human genome and transcriptome data in advancing genomic medicine and clinical care. However, the full map of the human proteome is still years away.

Nowadays, thanks to different high throughput technologies, it is possible to obtain huge quantities of biological data. A first difficulty consists in identifying, among all these data, which are the most relevant regarding a given pathology. Many studies were conducted this last year in that direction. For instance, in [23] Buggert et al. studied a cohort of 47 untreated HIV-infected individuals and 21 controls. Using mainly statistical tools they demonstrate that, among the usual laboratory parameters, the CD4/CD8 ratio is the most suitable laboratory predictor of combined T cell pathogenesis in HIV infection. McClay et al. [24] conducted a methylome-wide association study (MWAS) of 718 human individuals (aged 25-92 y; mean=55) using NGS technologies. With about 67.3 millions reads per individual they obtained measurements for the about 27 million autosomal CpG sites in the human genome. From these data, they were able to identify several new age-associated differentially methylated regions. In [25] Kamaraj and Purohit performed a computational screening of disease-mutations in the Oculocutaneous albinism type 2 (OCA2) gene. They considered 95 non-synonymous SNVs, identified one specific variation and using molecular dynamics simulation they showed prominent loss of stability and rise in mutant flexibility in the corresponding P protein. A second difficulty consists in being able to integrate all the relevant data. In [26], Kohl et al. presented a practical data processing workflow for integrating transcriptomics and proteomics data sets obtained from patients suffering from hepatocellular carcimona. They also developed a software, called CrossPlatformCommander, that implements some of the tasks of the workflow. This workflow could serve as a template when considering data coming from different high throughput technologies. There is no doubt that in the near future there will be more efforts for identifying relevant data among those generated by different techniques and for integrating these relevant data in user-friendly tools. This is essential to help clinicians in their daily practice.

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 2015 editors and managing editors Marie-Christine Jaulent, Brigitte Séroussi and Christoph U Lehmann.

References

1. Lecroq T, Soualmia LF. From genome sequencing to bedside. Findings from the section on bioinformatics and translational informatics. Yearb Med Inform 2013;8(1):175-7.
2. Lecroq T, Soualmia LF. Managing large-scale genomic datasets and translation into clinical practice. Yearb Med Inform 2014;9(1):212-4.
3. Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. J Am Med Inform Assoc 2012 Mar-Apr; 19(2):181-5.
4. Chute CG, Ullman-Cullere M, Wood GM, Lin SM, He M, Pathak J. Some experiences and opportunities for big data in translational research. Genetic Med 2013;15(10):802-9.
5. Capriotti E, Nehrt NL, Kann MG, Bromberg Y. Bioinformatics for personal genome interpretation. Brief Bioinform 2012;13(4):495-512.
6. Akan P, Alexeyenko A, Costea PI, Hedberg L, Solnestam BW, Lundin S, et al. Comprehensive analysis of the genome transcriptome and proteome landscapes of three tumor cell lines. Genome Med 2012;4(11):86.
7. Leary RJ, Sausen M, Kinde I, Papadopoulos N, Carpten JD, Craig D, et al. Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. Sci Transl Med 2012;4(162):162ra154.
8. Bensimon A, Heck AJ, Aebersold R. Mass spectrometry-based proteomics and network biology. Ann Rev Biochem 2012;81:379-405.
9. Cravatt BF, Simon GM, Yates JR. The biological impact of mass-spectrometry-based proteomics. Nature 2007;450:991-1000.
10. Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, et al. Mass-spectrometry-based draft of the human proteome. Nature 2014;509(7502):582-7.
11. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature 2012;491(7422):56-65.
12. Gonzalez MA, Lebrigio RF, Van Booven D, Ulloa RH, Powell E, Speziani F, et al. Genomes Management Application (GEM.app): a new software tool for large-scale collaborative genome analysis. Hum Mutat 2013;34(6):842-6.
13. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 2012;22(9):1760-74.
14. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, et al. A draft map of the human proteome. Nature 2014;509(7502):575-81.
15. Antonov AV, Krestyaninova M, Knight RA, Rodchenkov I, Melino G, Barlev NA. PPISURV: a novel bioinformatics tool for uncovering the hidden role of specific genes in cancer survival outcome. Oncogene 2014;33(13):1621-8.
16. Garcia-Albornoz M, Thankaswamy-Kosalai S, Nilsson A, Väremo L, Nookaew I, Nielsen J. BioMet Toolbox 2.0: genome-wide analysis of metabolism and omics data. Nucleic Acids Res 2014;42 (Web Server issue):W175-81.
17. Bendl J, Stourac J, Salanda O, Pavelka A, Wieben ED, Zendulka J, et al. PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. PLoS Comput Biol 2014;10(1):e1003440.
18. 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.
19. Ammenwerth E, Wolff AC, Knaup P, Ulmer H, Skonetzki S, van Bemmel JH, Mc Cray AT, Haux R, Kulikowski C, et al. Developing and evaluating criteria to help reviewers of biomedical informatics manuscripts. JAMIA 2003;10:512-4.
20. Taati B, Snoek J, Aleman D, Ghavamzadeh A. Data mining in bone marrow transplant records to identify patients with high odds of survival. IEEE J Biomed Health Inform 2014;18(1):21-7.
21. Tarczy-Hornoch P, Amendola L, Aronson SJ, Garraway L, Gray S, Grundmeier RW, et al. A survey of informatics approaches to whole-exome and whole-genome clinical reporting in the electronic health record. Genet Med 2013;15(10):824-32.
22. Sleiman P, Bradfield J, Mentch F, Almoguera B, Connolly J, Hakonarson H. Assessing the functional consequence of loss of function variants using electronic medical record and large-scale genomics consortium efforts. Front Genet 2014;5:105.
23. Buggert M, Frederiksen J, Noyan K, Svärd J, Barqasho B, Sönnerborg A, et al. Multiparametric bioinformatics distinguish the CD4/CD8 ratio as a suitable laboratory predictor of combined T cell pathogenesis in HIV infection. J Immunol 2014 Mar 1;192(5):2099-108.
24. McClay JL, Aberg KA, Clark SL, Nerella S, Kumar G, Xie LY, et al. A methylome-wide study of aging using massively parallel sequencing of the methyl-CpG-enriched genomic fraction from blood in over 700 subjects. Hum Mol Genet 2014 Mar 1;23(5):1175-85.
25. Kamaraj B, Purohit R. Computational screening of disease-associated mutations in OCA2 gene. Cell Biochem Biophys 2014 Jan;68(1):97-109.
26. Kohl M, Megger DA, Trippler M, Meckel H, Ahrens M, Bracht T, et al. A practical data processing workflow for multi-OMICS projects. Biochim Biophys Acta 2014 Jan;1844(1 Pt A):52-62.

Correspondence to:
Dr Lina F. Soualmia
Normandie Univ., Rouen University and Hospital
SIBM & LITIS EA 4108, Information Processing in Biology & Health
1, rue de Germont, Cour Leschevin porte 21
76031 Rouen Cedex, France
Tel : +33 232 885 869
E-mail: Lina.Soualmia@chu-rouen.fr

Appendix: Content Summaries of Selected Best Papers for the IMIA Yearbook 2015, Section Bioinformatics and Translational Informatics

Antonov AV, Krestyaninova M, Knight RA, Rodchenkov I, Melino G, Barlev NA

PPIsurv: a novel bioinformatics tool for uncovering the hidden role of specific genes in cancer survival outcome

Oncogene 2014 Mar 27; 33(13):1621-8

In this paper the authors present a free online data mining tool which correlates the expression, not only of a single gene but of the interactome of an input gene with the survival rates of patients suffering from cancer. The patients data come from more than 40 publicly available clinical expression data sets of about 8,000 patients. The data for gene interactomes also come from public databases. The gene data are statistically linked to cancer type data then the tool can, for instance, help to decide if an input gene upregulation is associated with positive or negative outcomes. The tool, named PPISURV, is available at http://www.bioprofiling.de/PPISURV. Users can specify their own source of interactions.

Bendl J, Stourac J, Salanda O, Pavelka A, Wieben ED, Zendulka J, Brezovsky J, Damborsky J

PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations

PLoS Comput Biol. 2014 Jan;10(1):e1003440

The main problem in medicine nowadays is not to identify all the variations in the genome of a human individual anymore but to identify among all those genetic variations which are those that can explain a given phenotype. In this paper the authors build a benchmark dataset of over 43,000 Single Nucleotides Variations (SNVs) for evaluating 8 tools that predict the effect of the variations: the 6 best were combined in a new consensus classifier, named PredictSNP, that significantly improves the prediction performance. Basically the individual tools classify the effect of the variations into two classes: neutral or deleterious together with a confidence score. PredictSNP computes a prediction consensus by combining all the individual predictions and confidence scores. The authors compared their tool with other consensus classifiers to confirm the gain in prediction performance. This new tool is available at http://loschmidt.chemi.muni.cz/predictsnp.

Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, Thomas JK, Muthusamy B, Leal-Rojas P, Kumar P, Sahasrabuddhe NA, Balakrishnan L, Advani J, George B, Renuse S, Selvan LD, Patil AH, Nanjappa V, Radhakrishnan A, Prasad S, Subbannayya T, Raju R, Kumar M, Sreenivasamurthy SK, Marimuthu A, Sathe GJ, Chavan S, Datta KK, Subbannayya Y, Sahu A, Yelamanchi SD, Jayaram S, Rajagopalan P, Sharma J, Murthy KR, Syed N, Goel R, Khan AA, Ahmad S, Dey G, Mudgal K, Chatterjee A, Huang TC, Zhong J, Wu X, Shaw PG, Freed D, Zahari MS, Mukherjee KK, Shankar S, Mahadevan A, Lam H, Mitchell CJ, Shankar SK, Satishchandra P, Schroeder JT, Sirdeshmukh R, Maitra A, Leach SD, Drake CG, Halushka MK, Prasad TS, Hruban RH, Kerr CL, Bader GD, Iacobuzio-Donahue CA, Gowda H, Pandey A

A draft map of the human proteome

Nature 2014 May 29; 509(7502):575-81

The goal of the study related in this paper is to build a comprehensive catalogue of the human proteome. Using high-resolution mass-spectrometry the authors were able to identify proteins encoded by 17,294 human genes, accounting for 84% of the annotated protein-coding genes, by profiling 30 histologically normal human cell and tissue types: 17 adult tissues, 7 fetal tissues and 6 haematopoietic cells. This study enabled to discover novel protein-coding regions including translated pseudogenes, non-coding RNAs and new Open Reading Frames. The findings of the study have been made available in a interactive web portal (http://www.humanproteomemap.org). These data complement the knowledge previously collected on the genomic level and are likely to help new discoveries on the genes and proteins interactions.

Taati B, Snoek J, Aleman D, Ghavamzadeh A

Data mining in bone marrow transplant records to identify patients with high odds of survival

IEEE J Biomed Health Inform 2014 Jan; 18(1):21-7

The goal of the study presented in this paper aims at being able to estimate the survival status of patients undergoing a bone marrow stem cell transplant in order to identify patients with very high chances of survival with high accuracy. This can have an impact on the donor matching strategy. The authors used different supervised learning techniques (regression linear classifiers, random forests and support vector machines), bayesian optimization together with parameter tuning for analyzing records up to 120 measurements from 1,751 patients that already had a transplant. There was still 22.3% of missing values after filtering. The main result of this study is that patients with the highest chances of success can be predicted with high accuracy. The authors also claim that the methodology used in the paper is transferable to the processing of similar data.