Managing Large-Scale Genomic Datasets and Translation into Clinical Practice

T. Lecroq, L. F. Soualmia, 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, Mont-Saint-Aignan, France

Summary

Objective: To summarize excellent current research in the field of Bioinformatics and Translational Informatics with application in the health domain.

Method: We provide a synopsis of the articles selected for the IMIA Yearbook 2014, from which we attempt to derive a synthetic overview of current and future activities in the field. 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 evaluated independently the set of 1,851 articles and 15 articles were retained for peer-review.

Results: The selection and evaluation process of this Yearbook's section on Bioinformatics and Translational Informatics yielded three excellent articles regarding data management and genome medicine. In the first article, the authors present VEST (Variant Effect Scoring Tool) which is a supervised machine learning tool for prioritizing variants found in exome sequencing projects that are more likely involved in human Mendelian diseases. In the second article, the authors show how to infer surnames of male individuals by crossing anonymous publicly available genomic data from the Y chromosome and public genealogy data banks. The third article presents a statistical framework called iCluster+ that can perform pattern discovery in integrated cancer genomic data. This framework was able to determine different tumor subtypes in colon cancer.

Conclusions: The current research activities still attest the continuous convergence of Bioinformatics and Medical Informatics, with a focus this year on large-scale biological, genomic, and Electronic Health Records data. Indeed, there is a need for powerful tools for managing and interpreting complex data, but also a need for user-friendly tools developed for the clinicians in their daily practice. All 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.

Keywords

Translational medical research, computational biology, gene expression, genome, medical informatics, big data management

Yearb Med Inform 2014: 212-4
http://dx.doi.org/10.15265/IY-2014-0039
Published online August 15, 2014

Introduction

As mentioned in the last year Yearbook [1], main ongoing work in Bioinformatics and Translational Informatics are related to Genome Medicine *i.e.* the identification from data of genes and mutations underlying human diseases by pursuing the research "from bedside to bench" [2]. In fact, the availability of large-scale 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 research activities is also a special interest to the management of "Big Data": ongoing genomic tests keep on generating massive amounts of data and our connected world propose novel requirements on transforming healthcare from reactive and hospital-centered, to preventive, proactive, evidence-based, and personalized medicine [3] with a focus on well-being rather than disease. Typical high-throughput sequencing identifies 3 to 10 million variants per individual [4], which requires in some Electronic Health Records (EHRs) 5 to 10GB storage. For instance, the 1000 Genome Project [5] has identified tens of millions of different genomic variants. As highlighted by Starren et al. [6], the clinical impact of the variants is mostly unknown and thus necessitates systems that dynamically reanalyze and interpret stored static genomic results in the context of evolving knowledge and clinical findings. In parallel, several tools have been developed. For instance, the web server Huvariome [7] can help to distinguish true new genomic variations from known variations. The web portal canEvolve [8] is developed for facilitating meta-analysis of oncogenomics datasets with a central access that provides information extracted from 90 cancer genomic studies and several platforms. Other applications are dedicated to large-scale data. The GEnomes Management Application (GEM.app) [9] allows to annotate, manage, visualize, and analyze large genomic datasets (1,600 whole exomes from 50 phenotypes). PreDPI-Ki [10] is a web server developed to predict drug-target interactions for drug discovery. The CRAVAT web server [11] assists the search for gene diseases in large-scale exome sequencing studies (whole exome data for Mendelian disorders) with the help of the Variant Effect Scoring Tool (VEST) a supervised machine learning classifier. The clinical translation of the bioinformatics results to an individual patient, which is poised to personalize medicine, is one of the challenges of the field. To manage and represent complex data and relationships efficiently, Farley et al. [12] proposed the BioIntelligence Framework that relies on a NoSQL database [13]. The framework creates a hypergraph-like store of public knowledge and when combined with personal genome and patient information, it derives a personalized genome-based knowledge store for clinical translation and discovery research.

All these studies and sequencing projects produce huge amounts of data, as for instance the Sequence Read Archive [14] of the NCBI. Most of the datasets are publicly accessible, and this poses the problems of the confidentiality of the data and the reuse of the results of personal genomics out of the scope of the clinical translation or research. In their study, Gymrek et al. [15] demonstrate the feasibility to track

back the identities of multiple participants in public sequencing projects. The authors show how to infer surnames of male individuals by crossing anonymous publicly available genomic data from the Y chromosome and public genealogy data banks. Therefore; the results of genomic tests may place the patient at several risks (discrimination, or stigmatization) and, as stated by Korf and Rehm, raising new ethical, legal, and social issues [16].

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2014 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
▪ Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the Variant Effect Scoring Tool. BMC Genomics 2013,14(Suppl 3):S3.
▪ Gymrec M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science 2013 Jan 18;339:321-4.
▪ Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, Powers RS, Ladanyi M, Shen R. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc Natl Acad Sci USA 2013 Mar 12;110(11):4245-50.

Best Paper Selection

The best paper selection for the section Bioinformatics and Translational Informatics follows a generic method, commonly used in all sections of the IMIA Yearbook 2014. As in past years, the search was performed on MEDLINE. The Boolean query included 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 2013 were considered; we excluded the publications types reviews, editorials, comments, letters to the editors, etc. We limited the search to 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 2012 Impact Factors). This year, the PubMed query yielded a set of 1,851 articles that were evaluated separately by each section editor (TL & LFS) using the BibReview tool [17]. BibReview takes into entry a PubMed file (in XML format) and shows all metadata for each article. A user can tag with the help of the interface, the articles as "Accepted", "Conflict" or "Reject" (on text, abstract or title). The results of several reviewers can be merged and the results can be filtered. A set of 41 articles was tagged "Accepted" by the section editors, from which 15 articles are selected for peer-review. As mentioned in the introduction, Genome Medicine and complex disease analysis characterize the field and the researches. Finally after evaluation three papers [11, 15, 18] wereretained by the reviewers.

In the first article [11], Carter et al. present VEST (Variant Effect Scoring Tool), which is a supervised machine learning tool for prioritizing variants found in exome sequencing project that are more likely involved in human Mendelian diseases. In the second article, [15] Gymer et al. show how to infer surnames of male individuals by crossing anonymous publicly available genomic data from the Y chromosome and public genealogy data banks. The third article [18] presents the work of Mo et al., a statistical framework called iCluster+ that can perform pattern discovery in integrated cancer genomic data. This framework was able to determine different tumor subtypes in colon cancer.

Table 1 lists the three 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.

Conclusions and Outlook

The current research activities still attest the continuous convergence of Bioinformatics and Medical Informatics, with a focus this year on large-scale biologic, genomic, and Electronic Health Records datasets. 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 recent research and development efforts are contributing to the challenge of impacting the results clinically and even going towards personalized medicine in the near future. Some frameworks already allow this translation but other issues (legal or ethical) must be taken into consideration when treating personal genomic 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 of the IMIA Yearbook.

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. 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.
3. Tian Q, Prics ND, Hood L. Systems cancer medicine: towards realization of predictive, preventive, personalized and participatory (P4) medicine. J Intern Med 2012 Feb;271(2):111-21.
4. Pelak K, Shianna KV, Ge D, Maia JM, Zhu M, Smith JP, et al. The characterization of twenty sequenced human genomes. PLoS Genet 2010 Sep 9;6(9) e1001111.
5. 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 Nov 1;491(7422):56-65.
6. Starren J, Williams MS, Bottinger EP. Crossing the omic chasm: a time for omic ancillary systems.

JAMA 2013 Mar 27;309(12):1237-8.
7. Stubbs A, McClellan EA, Horsman S, Hiltemann SD, Palli I, Nouwens Set al. Huvariome: a web server resource of whole genome next-generation sequencing allelic frequencies to aid in pathological candidate gene selection. J Clin Bioinform 2012 Nov 19;2(1):19.
8. Samur MK, Yan Z, Wang X, Cao Q, Munshi NC, Li C, et al. canEvolve: a web portal for integrative oncogenomics. PLoS One 2013;8(2):e56228.
9. 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 Jun;34(6):842-6.
10. Cao DS, Liang YZ, Deng Z, Hu QN, He M, Xu QS, Zhou GH, et al. Genome-scale screening of drug-target associations relevant to Ki using a chemogenomics approach. PLoS One 2013;8(4):e57680.
11. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the Variant Effect Scoring Tool. BMC Genomics 2013;14(Suppl 3):S3.
12. Farley T, Kiefer J, Lee P, Von Hoff D, Trent JM, Colbourn C, et al. The BioIntelligence Framework: a new computational platform for biomedical knowledge computing. J Am Med Inform Assoc 2013 Jan 1;20(1):128-33.
13. Pokorny J. NoSQL databases: a step to database scalability in web environment. International Journal of Web Information Systems 2013;9(1):69-82.
14. Kodama Y, Shumway M, Leinonen R. International Nucleotide Sequence Database Collaboration. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res 2012 Jan;40(-Database issue):D54-6.
15. Gymrec M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science 2013 Jan 18;339:321-4.
16. Korf BR, Rehm HL. New approaches to molecular diagnosis. JAMA 2013 Apr 10; 309(14): 1511-21.
17. Lamy JB. BibReview. http://extranet-limbio.smbh.univ-paris13.fr/html/limbio/claroline/backends/download.

Correspondence to:
Thierry Lecroq
Normandie Univ., University of Rouen
NormaSTIC FR CNRS 3638, IRIB and LITIS EA 4108
Information Processing in Biology & Health
76821 Mont-Saint-Aignan Cedex, France
E-mail: thierry.lecroq@univ-rouen.fr

Appendix: Content Summaries of Selected Best Papers for the IMIA Yearbook 2014, Section Bioinformatics and Translational Informatics

Carter H, Douville C, Stenson PD, Cooper DN, Karchin R

Identifying Mendelian disease genes with the Variant Effect Scoring Tool

BMC Genomics 2013,14(Suppl 3):S3

High throughput Next Generation Sequencing of exomes identifies a large number of variants, typically these kind of studies find hundreds to thousands variants per individual. Thus, it is fundamental to have tools for finding the variants that have a functional impact on the corresponding proteins and are likely to be involved in human Mendelian diseases. The authors present VEST (Variant Effect Scoring Tool) that can prioritize rare missense variants identified in whole exome sequencing studies. It uses a supervised machine learning based algorithm called Random Forest. The authors exhibit experimental results that show that VEST outperforms existing tools such as PolyPhen2 and SIFT4.0. VEST computes variant score p-values that can be aggregated at the gene level using Stouffer's Z-score. VEST is available as a stand-alone software package and is hosted by a web server.

Gymrec M, McGuire AL, Golan D, Halperin E, Erlich Y

Identifying personal genomes by surname inference

Science 2013 Jan 18;339:321-4

More and more genomic data are available on Internet. In this paper, the authors show how they recovered, with very high probability, surnames of individuals from anonymous genomic data. For that they crossed data from highly polymorphic short tandem repeats on the Y chromosome with genetic genealogy databases in the United States of America that contain a large number of surname-haplotype records and they also performed Internet searches. This study only relies on publicly accessible Internet resources. The authors show that this can be done with Next Generation Sequencing data even at low coverage. The re-identification of a single person takes only a few hours. This surname inference from personal genomes questions the privacy of de-identified public data sets.

Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, Powers RS, Ladanyi M, Shen R

Pattern discovery and cancer gene identification in integrated cancer genomic data

Proc Natl Acad Sci USA 2013 Mar 12;110(11):4245-50

In the last few years, there were considerable efforts for collecting genomic data for many types of cancer. The next challenge is mining these enormous amounts of data to unravel the biological mechanisms at the origin of these cancers: for instance to identify the genes involved in the processes. One goal is to discover therapeutic targets. Unfortunately cancer genome alterations are highly heterogeneous. In this paper, the authors present a statistical framework named iCluster+ that integrates diverse discrete and continuous genomic, epigenomic, and transcriptomic data related to cancers. This constitutes a significant enhancement of the iCluster method. This new method can perform pattern discovery using somatic mutations, copy number variations, and gene expression. It uses a linear regression with a set of latent variables that represent distinct biological factors. Using data from the Cancer Cell Line Encyclopedia and The Cancer Genome Atlas, this method was able to reveal distinct tumor subtypes in colon cancer.