Translational Bioinformatics Embraces Big Data

N. H. Shah
Stanford Center for Biomedical Informatics Research, Stanford University School of Medicine, Stanford, California, United States of America

Summary

We review the latest trends and major developments in translational bioinformatics in the year 2011-2012. Our emphasis is on highlighting the key events in the field and pointing at promising research areas for the future. The key take-home points are:

- Translational informatics is ready to revolutionize human health and healthcare using large-scale measurements on individuals.
- Data–centric approaches that compute on massive amounts of data (often called "Big Data") to discover patterns and to make clinically relevant predictions will gain adoption.
- Research that bridges the latest multimodal measurement technologies with large amounts of electronic healthcare data is increasing; and is where new breakthroughs will occur.

Keywords

Informatics, data mining, big data

Yearb Med Inform 2012:130-4

Introduction

Summarizing an entire research field is an intrinsically hard problem and for the purpose of this survey, I rely on discussions among the Scientific Program Committee of the 2012 AMIA Summit on Translational Bioinformatics (TBI), the focus areas of the excellent submissions received at the 2012 Summit [1] and the year-in-review presentations of the past two years at the TBI Summit [2].

The key areas of activity at the 2012 Summit were focused on research that take us from base pairs to the bedside [3], with a particular emphasis on clinical implications of mining massive data-sets, and bridging the latest multimodal measurement technologies with large amounts of electronic healthcare data that are increasingly available. Among the submissions to TBI, those that stood out for their innovation were invited into a special issue of the Journal of the American Medical Informatics Association. These capture some the trends underway in translational bioinformatics. For example, Liu et al [4] demonstrated how the ability to predict Adverse Drug Reactions (ADRs) can be increased by integrating chemical, biological, and phenotypic properties of drugs. They demonstrated that data fusion approaches are promising for large-scale ADR predictions in both preclinical and post-marketing phases. Similarly, for advancing the state of the art on interpreting GWAS data, Russu et al. [5] introduced a novel Bayesian model search algorithm, Binary Outcome Stochastic Search (BOSS), for model selection when the number of predictors (e.g. SNPs) far exceeds the number of observations. Finally, advancing the science on using the genome-in-the-clinic, Morgan et al [6] constructed genomic disease risk summaries for 55 common diseases using reported gene-disease associations in the research literature. They constructed risk profiles based on the SNPs as well as based on 187 whole genome sequences and show that risk predictions derived from sequencing differ substantially from those obtained from the SNPs for several non-monogenic diseases—by as much as a factor of 20 times in some instances.

Beyond this year's conference papers, in the larger informatics community, the following significant themes emerge over the past two years: 1) the genome has arrived at the door of the clinic [7, 8]. 2) „Big Data" approaches that compute on massive amounts of data to make clinically relevant predictions are poised for breakthroughs [1, 9-11]. 3) Efforts to bridge the latest multimodal measurement technologies with large amounts of electronic healthcare data are increasing. We refer to this emerging focus area as research on *mass phenotyping*.

Genome in the Clinic

Researchers from the eMERGE project recently demonstrated that GWAS can now be performed by leveraging large amounts of EMR data [12]. For example, Kho et al showed that by using commonly available data from five different EMRs it is possible to accurately identify T2D cases and con-

trols for genetic study across multiple institutions [13]; although in some instances the algorithms need some local tweaking [13, 14].

In parallel, genomic sequencing has moved out of the research realm and established itself in the clinic. For example, at the Medical College of Wisconsin, Dr. Howard Jacob's team used exome sequencing to identify a novel casual mutation that led to successful treatment of a 6-year-old boy with an extreme form of inflammatory bowel disease [7, 8].

In this landmark study, the authors used the patient's medical history, genetic and functional data, to diagnose an X-linked inhibitor of apoptosis deficiency. Going a step ahead, they performed an allogeneic hematopoietic progenitor cell transplant based on this finding to prevent the development of life-threatening hemophagocytic lymphohistiocytosis. Since treatment, there has been no recurrence of gastrointestinal disease, suggesting this mutation drove the gastrointestinal disease. This report demonstrates the power of exome sequencing to arrive at a molecular diagnosis in an individual patient in the setting of a novel disease, and illustrates clinical use of genomic sequencing. In recognition of the importance of such systematic clinical use of genomic information, the team's activities were recently the focus of a PBS NOVA episode titled "Cracking your genetic code".

With the increasing use of genomic information in the clinic, we are bound to be faced with having to interpret sequence variations that have not been observed and cataloged before. The problem is particularly acute in the case of multigenic diseases where known variants only contribute a small amount of risk. In research that attempts to find disease-causing variants in whole genome sequences, Yandell et al, develop a Bayesian method for prioritization of coding and non-coding variants combining several sequence features. They demonstrate the ability to detect rare variants in key genes in small cohorts, and common multigenic diseases.

Given the complexity of interpreting genomic information and the lack of comprehensiveness of current genomic variation databases—which are based on a small number of individuals, cover mostly Caucasian population, and where the variant-to-disease correlations don't generalize well—patients look to the scientific community to reliably detect disease and to predict their likelihood of responding to specific drugs. As a testimonial to the importance of this task, the Institute of Medicine recently published a consensus report on the *Evolution of Translational Omics: Lessons Learned and the Path Forward*. Inspite of the challenges ahead, it is now clear that the era of using the Genome in the clinic has arrived. It remains to be seen if the actual benefit obtained from using genomic information for patient care lives up to the promise genomic data has been hyped up to.

Big Data Goes Main Stream

Another clear trend over the past year was the adoption of "Big Data" and the associated change in mindset that Big Data analysis entails [15]. Currently, the discussion of Big Data in translational informatics frequently connotes next-generation sequencing data [10, 16, 17]. However, this is beginning to change: in 2011, the use of large public datasets of various kinds increased dramatically. As an example, let us consider the research activity around data mining for predicting adverse drug events (ADEs) [18] and novel drug indications using public data [19].

Currently drug safety surveillance is based on spontaneous reporting systems (SRS), which contain reports of suspected adverse drug events seen in clinical practice. In the United States, the primary database for such reports is the Adverse Event Reporting System (AERS) database at the US Food and Drug Administration agency (FDA). This resource has been successfully mined using *disproportionality measures*, which quantify the magnitude of difference between observed and expected rates of particular drug/adverse-event pairs [20, 21].

Given the amount of data available in AERS [22], researchers are developing methods for detecting new or latent multi-drug adverse events. For example, Tatonetti et al used side effect profiles from AERS reports to infer the presence of unreported adverse events [23-25], and Came et al created a network of known drug-ADE relationships to predict yet unknown ADEs before they are found in post-market evidence [26]. Making use of molecular level data, Pouliot et al [27] generated logistic regression models to correlate and predict postmarketing ADEs based on screening data from PubChem, a public database of chemical structures of small organic molecules along with information about their biological activities. In a related effort, Vilar et al [28] devised a way to enhance existing, data mining algorithms with chemical information using molecular fingerprints—which represent molecules through a bit vector that codifies the existence of particular structural features or functional groups—to enhance ADE signals generated from adverse event reports.

There have been increasing efforts to use other data sources, such as EMRs, for the purpose of detecting ADEs [29-31] and to discover multi-drug ADEs [32]. Researchers have also used billing and claims data for active drug safety surveillance [33-35] and applied literature mining for drug safety [36]. Recently Chee et al [37] explored the use of online health forums as a source of data to identify drugs for further scrutiny. They analyzed individuals' opinions of drugs in roughly 12 million personal health messages using natural language processing and are able to identify drug withdrawals based on messages discussing them before their removal.

As an example of practical results that can result for large scale data mining, Gottlieb et al present a method for inferring novel drug indications [38] to find novel uses for existing drugs by mining drug-drug and disease-disease similarities across multiple sources; ranging from gene annotations to disease phenotype descriptions. By assessing overlap of their predicted novel uses with drugs currently in clinical trials, they show that disease-specific signatures can potentially predict new drugs. Going a step further, Sirota et al and Dudley et al actually *find* a novel use for the antiulcer drug cimetidine as a candidate therapeutic in the treatment of lung adenocarcinoma and the anticonvulsant topiramate for inflammatory bowel disease [39, 40].

Given the need for detecting latent associations in large datasets of different types, Reshef et al devised new methods for finding interesting (nonlinear) relationships between pairs of variables in very large data sets [41]. In their paper in Science, they argue that maximal information coefficient (MIC) captures wide range of associations and that their method is applicable to diverse datasets such as those about global health, gene expression, baseball, microbiota in gut with good results.

Mass Phenotyping on the Rise

Looking ahead, it is reasonable to assume that Big Data in biomedical informatics will be far more than genome sequence data or gene expression data [42-44]. 'Big Data' should be considered in a comprehensive manner, including both large amounts of *molecular measurements* on a person (e.g., sequencing) as well as small amounts of *routine measurements* on a large number of people (e.g., clinical notes, lab measurements, claims data and adverse event reports). In contrast to the buzz around genomic-data-in-the-clinic or adverse event predictions, consider the example by Frankovich et al. [45]. When the existing literature and a survey of colleagues was insufficient to guide the clinical care of a patient, Frankovich et al applied trend analysis to the EMR data from 98 patients to "learn" a data-driven guideline on how to provide care for a 13-year-old girl with systemic lupus erythematosus (SLE) [45].

In terms of size, the EMR data from 98 patients is certainly not "big" as would be the case with genome sequences from 98 individuals. However, such approaches, which analyze data that is already routinely collected, are particularly valuable when a formal guideline is not available or feasible from a practical standpoint. Outside of healthcare and medicine, a small amount of data about millions of individuals is already being collected and mined by Web companies (e.g. a typical social network profile, when exported is a couple of GB) resulting in a gold rush around analyzing this "digital exhaust"[1]

The idea of using user generated content for enhancing health and well-being is highly popular in groups such as the *Quantified Self* collaborative, which lists some 504 tools for collecting data on an individual for the purpose of self-tracking[2]. Given the rising popularity of such efforts and the increasingly sophisticated monitoring mechanisms, there is a revolution underway in terms of what phenotypic data we can collect on an individual. We define "mass phenotyping" as the collection and integration of massive amounts of diverse phenotypical information (continuous variables or categorical) in order to discover patterns which would be invisible otherwise and correlate those patterns with health and well-being.

There are already some early successes at *correlating* genotypic data with such self-collected phenotypic data by individuals. For example, Tung et al show that over 180 associations of genotypes with specific phenotypes can be replicated using self-reported phenotypic data; although with lower precision than a clinical study [46]. In a similar study done in an academic setting Roque et al mined phenotypic descriptions from the free-text of electronic medical records to cluster patients based on disease co-occurrences and to suggest genetic hooks for phenotype "syndromes" [47]. In a highly innovative study, Frost et al show the feasibility of using patient-reported outcomes to profile the safety of two drugs prescribed off-label [48]. In fact, communities of individuals that participate in such collaborative mass phenotyping can even self-organized to conduct a "trial" for the utility of lithium carbonate in the treatment of Amyotrophic lateral sclerosis (ALS) [49]. In this particular study, at 12 months after treatment, the patient community found no effect of lithium on disease progression [49].

Looking ahead

It is tantalizing to imagine how scientific inquiry would be done differently if we collect and share access to lots of data—both genomic and "routine". How will the kinds of questions we ask change when we cross a certain data-threshold? [15, 50]. For example, researchers at Carnegie Mellon University built a scene completion tool by scraping millions of other images on the Web from public sources. After the system accumulated a corpus of millions of photos, completed scenes were indistinguishable to the naked eye. The case for big data analytics has already won over the legal domain in at least one application replacing armies of lawyers with computer algorithms designed for *e-discovery*, i.e., retrieval of relevant materials for a legal case [51]. Even the liberal

[1] http://www.vlab.org/article.html?aid=304
[2] http://quantifiedself.com/about/

arts are embracing Big Data: capitalizing on Google's efforts to digitize books, researchers in the humanities are blazing new trails in *culturomics* by examining language based on the analysis of word combinations occurring in millions of digitized books through time [52].

In recognition of this emerging mass phenotyping trend, a recent US National Research Council report [53] also acknowledged the role of new modes of population based research in enabling a new understanding of human disease and health states. In the United States, the government is making a highly visible push towards promoting the use of Big Data across multiple disciplines—including translational research that bridges the latest multimodal measurement technologies with large amounts of electronic healthcare data [11]. This is an exciting time when medicine begins utilizing massive amounts of data to discover patterns, trends, and to make predictions in a manner that is a mainstay of Web-scale computing [15].

Acknowledgments

N.H.S. is funded by the US National Institute of Health Roadmap (U54 HG004028 and U54 LM008748). The ideas around mass phenotyping benefited from discussion with Lawrence Hunter and participants at the Discovery Informatics Workshop 2012, supported by the National Science Foundation.

References

1. Shah NH, Tenenbaum JD. The coming age of data-driven medicine: Translational Bioinformatics' next frontier. J Am Med Inform Assoc 2012 Jun 1;19(e1):e2-e4.
2. Altman RB, Miller KS. 2010 translational bioinformatics year in review. J Am Med Inform Assoc 2011;18(4):358-66.
3. Green ED, Guyer MS. Charting a course for genomic medicine from base pairs to bedside. Nature 2011; 470(7333):204-13.
4. Liu M, Wu Y, Chen Y, Sun J, Zhao Z, Chen XW, et al. Large-scale Prediction of Adverse Drug Reactions by Integrating Chemical, Biological, and Phenotypic Properties of Drugs. J Am Med Inform Assoc 2012 Jun 1;19(e1):e28-e35.
5. Russu A, Malovini A, Puca AA, Bellazzi R. Stochastic model search with binary outcomes for Genome-Wide Association Studies. J Am Med Inform Assoc 2012 Jun 1;19(e1):e13-e20.
6. Morgan AA, Chen R, Butte AJ. Clinical utility of sequence-based genotype compared with that derivable from genotyping arrays. J Am Med Inform Assoc 2012 Jun 1;19(e1):e21-e27
7. Worthey EA, Mayer AN, Syverson GD, Helbling D, Bonacci BB, Decker B, et al. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet Med 2011;13(3):255-62.
8. Mayer AN, Dimmock DP, Arca MJ, Bick DP, Verbsky JW, Worthey EA, et al. A timely arrival for genomic medicine. Genet Med 2011; 13(3):195-6.
9. Trelles O, Prins P, Snir M, Jansen RC. Big data, but are we ready? Nat Rev Genet 2011;12(3):224.
10. Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP. Cloud and heterogeneous computing solutions exist today for the emerging big data problems in biology. Nat Rev Genet 2011; 12(3):224.
11. Weiss R. Obama Administration Unveils „Big Data" Initiative: Announces $200 million in new R&D Investments. Washington D.C.: O.o.S.a.T. Policy, Executive Office of the President; 2012. p. 1-4.
12. Kho AN, Pacheco JA, Peissig PL, Rasmussen L, Newton KM, Weston N, et al. Electronic Medical Records for Genetic Research: Results of the eMERGE Consortium. Sci Transl Med 2011;3(79):79re1.
13. Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, Armstrong LL, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J Am Med Inform Assoc 2012;19(2):212-8.
14. Wei WQ, Leibson CL, Ransom JE, Kho AN, Caraballo PJ, Chai HS, et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. J Am Med Inform Assoc 2012;19(2):219-24.
15. Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data. IEEE Intelligent Systems 2009;24(2):8-12.
16. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, et al. Clinical assessment incorporating a personal genome. Lancet 2010;375(9725):1525-35.
17. Samani NJ, Tomaszewski M, Schunkert H. The personal genome—the future of personalised medicine? Lancet 2010;375(9725):1497-8.
18. Harpaz R, Dumouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel Data Mining Methodologies for Adverse Drug Event Discovery and Analysis. Clin Pharmacol Ther 2012. in press.
19. Lussier YA, Chen JL.The emergence of genome-based drug repositioning. Sci Transl Med 2011;3(96):96ps35.
20. Bate A, Evans SJW. Quantitative signal detection using spontaneous ADR reporting. Pharmacoepidemiol Drug Saf 2009;18(6):427-36.
21. Szarfman A, Machado SG, O'Neill RT. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. Drug Saf 2002;25(6):381-92.
22. Weiss-Smith S, Deshpande G, Chung S, Gogolak V. The FDA drug safety surveillance program: adverse event reporting trends. Arch Intern Med 2011;171(6):591-3.
23. Norén GN, Sundberg R, Bate A, Edwards IR. A statistical methodology for drug-drug interaction surveillance. Stat Med 2008;27(16):3057-70.
24. Tatonetti NP, Denny JC, Murphy SN, Fernald GH, Krishnan G, Castro V, et al. Detecting Drug Interactions From Adverse-Event Reports: Interaction Between Paroxetine and Pravastatin Increases Blood Glucose Levels. Clin Pharmacol Ther 2011;90(1):133-42.
25. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. Sci Transl Med 2012 Mar 14; 4(125):125ra31.
26. Cami A, Arnold A, Manzi S, Reis B. Predicting adverse drug events using pharmacological network models. Sci Transl Med 2011;3(114):114ra127.
27. Pouliot Y, Chiang AP, Butte AJ. Predicting adverse drug reactions using publicly available PubChem BioAssay data. Clin Pharmacol Ther 2011; 90(1):90-9.
28. Vilar S, Harpaz R, Chase HS, Costanzi S, Rabadan R, Friedman C. Facilitating adverse drug event detection in pharmacovigilance databases using molecular structure similarity: application to rhabdomyolysis. J Am Med Inform Assoc 2011;18 Suppl 1:i73-80.
29. Liu Y, LePendu P, Iyer S, Shah NH. Using Temporal Patterns in Medical Records to Discern Adverse Drug Events from Indications. In: AMIA Summit on Clinical Research Informatics, 2012. San Francisco: AMIA.
30. Lependu P, Iyer SV, Fairon C, Shah NH. Annotation Analysis for Testing Drug Safety Signals. J Biomed Semantics 2012;3 Suppl 1:S5.
31. Brownstein JS, Sordo M, Kohane IS, Mandl KD. The tell-tale heart: population-based surveillance reveals an association of rofecoxib and celecoxib with myocardial infarction. PLoS ONE 2007; 2(9):e840.
32. Harpaz R, Chase H, Friedman C. Mining multi-item drug adverse effect associations in spontaneous reporting systems. BMC Bioinformatics 2010;11 Suppl 9:S7.
33. Dore D, Seeger J, Arnold Chan K . Use of a claims-based active drug safety surveillance system to assess the risk of acute pancreatitis with exenatide or sitagliptin compared to metformin or glyburide. Curr Med Res Opin 2009;25(4):1019-27.
34. Nadkarni P. Drug safety surveillance using de-identified EMR and claims data: issues and challenges. J Am Med Inform Assoc 2010; 17(6):671-4.
35. Brown JS, Kulldorff M, Chan KA, Davis RL, Graham D, Pettus PT, et al. Early detection of adverse drug events within population-based health networks: application of sequential testing

methods. Pharmacoepidemiol Drug Saf 2007; 16(12):1275-84.
36. Shetty KD, Dalal S. Using information mining of the medical literature to improve drug safety. J Am Med Inform Assoc 2011;18(5):668-74.
37. Chee BW, Berlin R, Schatz B. Predicting adverse drug events from personal health messages. AMIA Annu Symp Proc 2011;2011:217-26
38. Gottlieb A, Stein GY, Ruppin E, Sharan R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. Mol Syst Biol 2011;7:496.
39. Sirota M, Dudley JT, Kim J, Chiang AP, Morgan AA, Sweet-Cordero A, et al. Discovery and preclinical validation of drug indications using compendia of public gene expression data. Sci Transl Med 2011;3(96):96ra77.
40. Dudley JT, Sirota M, Shenoy M, Pai RK, Roedder S, Chiang AP, et al. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. Sci Transl Med 2011;3(96):96ra76.
41. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, et al. Detecting Novel Associations in Large Data Sets. Science 2011;334(6062):1518-24.
42. Sobek M, Cleveland L, Flood S, Hall PK, King ML, Ruggles S, et al. Big Data: Large-Scale Historical Infrastructure from the Minnesota Population Center. Hist Methods 2011;44(2):61-8.
43. Fox B. Using big data for big impact. How predictive modeling can affect patient outcomes. Health Manag Technol 2012;33(1):32.
44. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Chen R, et al. Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes. Cell 2012;148(6):1293-1307.
45. Frankovich J, Longhurst CA, Sutherland SM. Evidence-based medicine in the EMR era. N Engl J Med 2011;365(19):1758-9.
46. Tung JY, Do CB, Hinds DA, Kiefer AK, Macpherson JM, Chowdry AB, et al. Efficient replication of over 180 genetic associations with self-reported medical data. PLoS ONE 2011;6(8):e23473.
47. Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T, et al. Using electronic patient records to discover disease correlations and stratify patient cohorts. Plos Comput Biol 2011;7(8):e1002141.
48. F Frost J, Okun S, Vaughan T, Heywood J, Wicks P. Patient-reported outcomes as a source of evidence in off-label prescribing: analysis of data from PatientsLikeMe. J Med Internet Res 2011;13(1):e6.
49. Wicks P, Vaughan TE, Massagli MP, Heywood J. Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. Nature Biotechnol 2011;29(5):411-4.
50. Hays J, Efros AA. Scene completion using millions of photographs. Commun ACM 2008;51(10):87-94.
51. Bringardner J. Winning the Lawsuit: Data Miners Dig for Dirt. Wired Magazine 2008(16-07).
52. Michel JB, Shen YK, Aiden AP, Veres A, Gray MK; Google Books Team, et al. Quantitative analysis of culture using millions of digitized books. Science 2011;331(6014): 176-82.
53. National Research Council, U.S.C.o.A.F. f.D.a.N.T.o.D. Toward precision medicine building a knowledge network for biomedical research and a new taxonomy of disease. 2011; Available from: http://www.worldcat.org/isbn/0309222222.

Correspondence to:
Nigam H Shah, MBBS, PhD
Stanford University School of Medicine
1265 Welch Road
Room X-229
Stanford, CA 94305, USA
Tel: +1 650 725-6236
Fax: +1 650 725-7944
E-mail: nigam@stanford.edu