Computationally translating molecular discoveries into tools for medicine: translational bioinformatics articles now featured in *JAMIA*

Atul J Butte,[1,2] Nigam H Shah[3]

This year marks the 15th anniversary of the invention of the gene expression microarray. As mRNA transcripts serve as the blueprint within cells for making proteins, measuring mRNA levels was seen as an accurate and manageable way to investigate cell and tissue processes. Those earliest microarrays in 1995 could measure 48 transcripts in parallel in plants,[1] but within 1 year were scaled up to measure more than 1000 transcripts including those in human tissues. Today, these microarrays are essentially commodity items, commonly used to study human health and disease in hospitals and academic institutions, as well as in the biotechnology and pharmaceutical industry.

While tens of thousands of publications have already been published referencing microarrays, this is just the start. Similar arrays are already used to probe genetic differences in DNA, but even these will soon be supplanted by whole genome sequencing, where we can expect all three billion human base pairs to be sequenced for a few thousand dollars. The exponential decrease in costs for whole genome sequencing has been described as going beyond the decline we are used to from Moore's law.[2] These enormous amounts of molecular data make it clear that there is a pressing need for computational methods to analyze and interpret them.

Molecular data have never been foreign to the pages of *JAMIA* or AMIA Symposia. The second volume of *JAMIA* back in 1995 contained an article describing how the internet and newly introduced world wide web could be used to facilitate genome sequencing efforts across two academic genome centers and introduced concepts like yeast artificial chromosomes, sequence tagged sites and contigs.[3] The 2002 AMIA Fall Symposium suggested bioinformatics and medical informatics could and should be part of the same discipline of biomedical informatics.

While not every publication in bioinformatics is likely relevant to AMIA members, it is important to track those applying bioinformatics to human health and disease. Translational bioinformatics can be defined as 'the development of storage, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data into proactive, predictive, preventive, and participatory health.'[4] Indeed, translational bioinformatics has been a core strategic area of importance for AMIA since 2008. Today, as more hospital and academic medical centers embrace molecular measurements for diagnosis and planning therapies, we recognize that the community of *JAMIA* readers will need to keep abreast of new developments and applications of translational bioinformatics.[5] In the past few years, however, it has been increasingly hard to find articles on translational bioinformatics in *JAMIA*.

However, with this month's issue, we hope to start reversing the trend. Starting with a Perspective on from Neil Sarkar and colleagues (***see page 354***), we are highlighting five manuscripts in the field of translational bioinformatics, and 'opening the doors' for submissions from investigators and authors in this field.[6]

Recognizing the continued growth in bioinformatics, especially as related to human health and disease, in 2009 AMIA initiated the annual Summit on Translational Bioinformatics as a new annual meeting to address the growing need for a scientific conference to present and discuss developments and the application of methods in this field. At the 2011 Summit, several authors of top-ranked submitted poster abstracts were invited to expand their submissions into full manuscripts, which were then evaluated by *JAMIA* reviewers. One such manuscript appears in this month's issue. Hua Xu and colleagues (***see page 387***) show how dosing details locked within free-text electronic health records could be found using natural language processing, thus enabling a gene–dosing association study for warfarin dosing at Vanderbilt University.[7] This work serves as a premier model of the type of translational research possible when bioinformatics and medical informatics investigators truly collaborate.

Few institutions have the considerable resources of Vanderbilt University, which has a DNA biobank linked to a de-identified electronic health record subset, but there are a few, notably the Mayo Clinic. In a manuscript this month, Pathak and colleagues (***see page 376***) show how standardized representations of clinical characteristics extracted from institutional electronic health records can be integrated to enable large scale genetics studies between Vanderbilt University and the Mayo Clinic.[8] As the genetic risk of disease is now theorized to result from a combination of rare DNA variants,[9] larger cross-country cohorts like these will be needed to identify these variants.

Integration of information on patients, samples, or data is indeed a common theme in translational bioinformatics. David Foran and colleagues (***see page 403***) highlight a new federated software system that can handle another type of biobank for pathology samples.[10] Processed into histology cores and distributed on a tissue microarray, these samples are proving to be invaluable for the high-throughput query of specific proteins. James Chen and colleagues (***see page 392***) show how comparing the genomic differences found across publicly available data from previous prostate cancer studies can yield a single core 'signature' that can distinguish prostate cancer patients with better prognosis from those with worse prognosis.[11]

Translational bioinformatics grew out of the work of a small but cohesive group of researchers who bridged the gap between computational biology and medicine. It is with sad remembrance that we note the passing of Marco Ramoni, a pioneering spirit who was one of the

[1]Division of Systems Medicine, Department of Pediatrics, Stanford University School of Medicine, Stanford, California, USA; [2]Lucile Packard Children's Hospital, Palo Alto, California, USA; [3]Stanford Center for Biomedical Informatics Research, Stanford University School of Medicine, Stanford, California, USA

Correspondence to Dr Atul J Butte, Stanford University School of Medicine, 251 Campus Drive MS-5415, Room X-163, Stanford, CA 94305-5415, USA; abutte@stanford.edu

first Track Chairs for the AMIA Summit on Translational Bioinformatics. So that his contributions to AMIA and the field of translational bioinformatics will not be forgotten, this year the Board of Directors established the Marco Ramoni Best Paper Award, to be presented each year at the AMIA Summit on Translational Bioinformatics. The manuscript from the award winner for 2011, Wei Wei, appears in this issue of *JAMIA* (***see page 370***).[12] Selected by an external review panel chosen by the Chair of the Scientific Program Committee and again peer reviewed by *JAMIA* reviewers, it is slightly ironic that the award winning manuscript is on the application of Bayes' theorem, coincidentally similar to Dr Ramoni's own work, which is summarized by Kohane and Szolovits (***see page 367***) in an invited academic tribute to our dear colleague.[13]

Next year will see the fifth Summit on Translational Bioinformatics. With the approaching changes in the amount and diversity of datasets discussed above, we anticipate that data-centric approaches that compute on massive amounts of data (often called 'big data'[14]) to identify patterns and make clinically relevant predictions will be increasingly common in translational bioinformatics. In anticipation, the 2012 Summit on Translational Bioinformatics will have four tracks focusing on research that take us from base pairs to the bedside,[15] with a particular emphasis on the clinical implications of mining massive datasets, and bridging the latest multimodal measurement technologies using the large amounts of electronic healthcare data that are increasingly available. We invite readers to submit extended (10-page) submissions for the Summit before the deadline of August 15. The top papers will be published in *JAMIA* after peer review and the editorial office will make all efforts to have them available online first by the time the Summit takes place on March 19–21, 2012 in San Francisco.

In closing, we have continued to note arguments over how much translational bioinformatics informatics investigators and professionals really need to know. To answer this, we must consider that we are entering a decade where tens of thousands of people have already obtained samplings of their own DNA sequences from consumer genomics companies,[16] with over 700 000 RNA microarray measurements already available to the public,[17 18] and we expect that 30 000 people will have their whole genome sequenced this year alone.[19] We must not keep assuming that if we just build and provide the right generalized tools and methods, others will take them and use them the right way to improve healthcare—the field is oversaturated with tools already. If we best understand the tools and methods we build, we need to be the first to actually use those tools, and show the world what can be achieved. Instead of discussing how relevant translational bioinformatics is, we need to argue that biomedical informatics is the only field in biomedicine that is ready to revolutionize human health and healthcare using these tools and measurements.

Funding AJB is funded by the US National Library of Medicine (R01 LM009719) and the Lucile Packard Foundation for Children's Health. NHS is funded by the US National Institute of Health Roadmap (U54 HG004028).

Competing interests None.

Provenance and peer review Commissioned; internally peer reviewed.

Accepted 29 April 2011

J Am Med Inform Assoc 2011;**18**:352–353.
doi:10.1136/amiajnl-2011-000343

REFERENCES

1. **Schena M,** Shalon D, Davis RW, *et al*. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995;**270**:467–70.
2. **Wetterstrand K.** *DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program*. 2011. http://www.genome.gov/sequencingcosts (accessed 1 Apr 2011).
3. **Miller PL,** Nadkarni PM, Kidd KK, *et al*. Internet-based support for bioscience research: a collaborative genome center for human chromosome 12. *J Am Med Inform Assoc* 1995;**2**:351–64.
4. **American Medical Informatics Association.** *AMIA Strategic Plan*. 2006. http://www.amia.org/inside/stratplan (accessed 15 Mar 2007).
5. **Butte AJ.** Translational bioinformatics: coming of age. *J Am Med Inform Assoc* 2008;**15**:709–14.
6. **Sarkar IN,** Butte AJ, Lussier YA, *et al*. Translational bioinformatics: linking knowledge across biological and clinical realms. *J Am Med Inform Assoc* 2011;**18**:354–7.
7. **Xu H,** Jiang M, Oetjens M, *et al*. Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin. *J Am Med Inform Assoc* 2011;**18**:387–91.
8. **Pathak J,** Wang J, Kashyap S, *et al*. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. *J Am Med Inform Assoc* 2011;**18**:376–86.
9. **Manolio TA,** Collins FS, Cox NJ, *et al*. Finding the missing heritability of complex diseases. *Nature* 2009;**461**:747–53.
10. **Foran DJ,** Yang L, Chen W, *et al*. ImageMiner: a software system for comparative analysis of tissue microarrays using content-based image retrieval, highperformance computing, and grid technology. *J Am Med Inform Assoc* 2011;**18**:403–15.
11. **Chen JL,** Li J, Stadler WM, *et al*. Protein-network modeling of prostate cancer gene signatures reveals essential pathways in disease recurrence. *J Am Med Inform Assoc* 2011;**18**:392–402.
12. **Wei W,** Visweswaran S, Cooper GF, *et al*. The application of naive Bayes model averaging to predict Alzheimer's disease from genomewidec data. *J Am Med Inform Assoc* 2011;**18**:370–5.
13. **Kohane IS,** Szolovits P. Marco Ramoni: an appreciation of academic achievement. *J Am Med Inform Assoc* 2011;**18**:367–9.
14. **Schadt EE,** Linderman MD, Sorenson J, *et al*. Cloud and heterogeneous computing solutions exist today for the emerging big data problems in biology. *Nat Rev Genet* 2011;**12**:224.
15. **Green ED,** Guyer MS. Charting a course for genomic medicine from base pairs to bedside. *Nature* 2011;**470**:204–13.
16. **Eriksson N,** Macpherson JM, Tung JY, *et al*. Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genet* 2010;**6**:e1000993.
17. **Barrett T,** Troup DB, Wilhite SE, *et al*. NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res* 2007;**35**:D760–5.
18. **Brazma A,** Parkinson H, Sarkans U, *et al*. ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* 2003;**31**:68–71.
19. Human genome: Genomes by the thousand. *Nature* 2010;**467**:1026–7.