



DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD

Computational biology: plus c'est la même chose, plus ça change

The Harvard community has made this article openly available.
[Please share](#) how this access benefits you. Your story matters.

Citation	Huttenhower, Curtis. 2011. Computational biology: plus c'est la même chose, plus ça change. <i>Genome Biology</i> 12(8): 307.
Published Version	doi:10.1186/gb-2011-12-8-307
Accessed	February 19, 2015 10:48:32 AM EST
Citable Link	http://nrs.harvard.edu/urn-3:HUL.InstRepos:10576037
Terms of Use	This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

(Article begins on next page)

MEETING REPORT

Computational biology: plus c'est la même chose, plus ça change

Curtis Huttenhower*

Abstract

A report on the joint 19th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)/10th Annual European Conference on Computational Biology (ECCB) meetings and the 7th International Society for Computational Biology Student Council Symposium, Vienna, Austria, 15-19 July 2011.

As of this writing, the average desktop computer processor contains on the order of one billion transistors. Although the growth of the semiconductor industry has frequently been compared to that of high-throughput sequencing, the field perhaps shares a number of deeper traits with this year's joint International Conference on Intelligent Systems for Molecular Biology (ISMB) and European Conference on Computational Biology (ECCB). Bioinformatics, like data storage, is becoming ever more adept at organizing tremendous quantities of information - not simply producing data, but intelligently cataloging and interpreting it. Chip design, like biology, requires not simply the presence of billions of independent parts, but their careful orchestration into an information-processing unit through controlled interactions. And like quantum computing, goals such as translational genomics remain tantalizingly on the horizon, close enough to see their potential but not yet realized in day-to-day practice. These three topics appeared consistently throughout the ISMB/ECCB conference, which hosted nearly 2,000 attendees and was immediately preceded by the accompanying Student Council Symposium (SCS). Here, I summarize the themes, highlights, and events of these two of the many joint conferences organized by the International Society for Computational Biology (ISCB).

The data deluge: still keeping our heads above water

Bioinformatics has been dealing with an exponential growth in data since its coalescence as a field in the 1980s, making the Senior Scientist Award keynote with which Michael Ashburner closed the conference particularly appropriate. This retrospective by the 'father of ontologies in biology', to quote the introduction by ISCB president Burkhard Rost, detailed the remarkable expansion of computational biology since Ashburner's start as a Cambridge undergraduate 50 years ago. His tongue-in-cheek career summary included the point at which international data transfer rates on the emerging ARPANET became more practical than hard-copy print-outs (apparently 19.2 kbaud); the initial £13 million funding of the European Bioinformatics Institute in response to a half-page proposal; and of course the founding of the Gene Ontology Consortium by a small, enthusiastic group of researchers at the 6th ISMB in 1998. Ashburner concluded by urging scientists to communicate their research openly to the community and to the public at large. The degree to which bioinformatics already embraces this philosophy was, fortunately, evident in ISMB's Technology Track, which highlighted popular open tools throughout the conference, particularly mainstays of sequence analysis including Ensembl, the University of California Santa Cruz's Genome Browser, and Galaxy.

While the conference was ended by the Senior Scientist Award, it was opened by a group of over 80 scientists at the beginnings of their careers, specifically the ISCB Student Council Symposium. This student-run symposium has accompanied ISMB/ECCB since 2005 and annually features presentations, posters, and faculty keynotes focused on graduate student research. Both the SCS Best Presentation award to Amit Deshwar and its Best Poster award to Benjamin Kwan highlighted the need for novel algorithms as data continue to grow, both in gene expression repositories and as whole-genome sequences. The Symposium was anchored by Ivet Bahar's senior faculty keynote, which demonstrated several ways in which she has used massive data to discover new protein folds and to enrich our overall understanding of structural genomics. The areas of functional and structural genomics both include some of the

*Correspondence: chuttenh@hsph.harvard.edu
Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, USA

longest-standing challenges in bioinformatics and, along with the historical perspective of Ashburner's keynote, these presentations emphasized the continuing power of new algorithms in combination with ever-increasing data availability.

Nowhere at ISMB was this more evident than in high-throughput sequencing, an area highlighted by over 25 talks and by the HiTSeq special interest group meeting. Transcriptomics and RNA sequencing were of particular interest this year, with a comprehensive description of developments in sequencing and other high-throughput technologies provided by Janet Thornton's ECCB 10th anniversary keynote. Thornton, a self-professed 'data junkie', characterized the EBI's goal as understanding life from individual biomolecules through to interaction networks, information processing circuits, predictive simulations, and their impacts on human health. The talk mentioned the EBI's 12 petabytes of diverse information - and, amazingly, that these data continue to double every 5 months. Thornton described the breadth of these data and detailed methods for predictive biochemical modeling of specific ligase enzyme families. She closed by emphasizing that current methods only scratch the surface of what is possible. Fundamental questions regarding the evolution of the human reactome (complete reaction catalog) and the composition of a core reactome necessary for all life remain unanswered but, for the first time, are perhaps within reach.

Biological networks: unraveling a tangled web

Two years ago during the 2009 ISMB Overton Prize keynote, Trey Ideker jokingly suggested that his predictions in a 2006 review regarding the future of biological network analysis had, like most futurology, been perhaps a bit optimistic. This year's conference proved him more prescient than was apparent, as many of the most important questions in computational biology continue to revolve around network models. This was apparent from the very beginning of ISMB/ECCB 2011, which was opened by Bonnie Berger's keynote on the critical roles of good algorithms in massive data mining. Berger outlined three challenges addressed by recent work: compressive genomics for rapid data retrieval, medical genomics for establishing clinically reliable signal-to-noise levels, and network alignment for understanding interactome evolution. Each of these represents a cutting-edge application of computer science to biology not possible without sufficient data, appropriate network models, and efficient algorithms, all areas of active research.

A significant feature of bioinformatic network analysis, however, is that it enables not only novel network mining algorithms, but also new ways of formulating biological systems and modeling functional, molecular, and evolutionary mechanisms. The first was demonstrated by

Chad Myers in one of two SCS junior faculty keynotes, in which he explained experimental and computational techniques for the construction of a genome-wide genetic interaction network in *Saccharomyces cerevisiae*. In a related collaboration presented during the ISMB Highlights track, Philip Kim focused on the specific biological roles of intrinsically disordered proteins, demonstrating that these fall into classes of flexible, constrained, and non-conserved disorder. In yet another example of creative network modeling, Maureen Stolzer discussed recent research on the evolution of multi-domain proteins, using co-occurrence analysis to show that even well-studied families such as the kinases acquire functionality through domain shuffling. Although these are only a few of the network construction and investigation topics introduced at ISMB, they emphasize that although the area has been a perennial favorite of bioinformatics, it continues to be a rich source both of computational methods and of biological understanding.

Correspondingly, in what has been described as a 'systems biology tour de force', Louis Serrano unified these themes in his keynote on the comprehensive investigation of metabolism in *Mycoplasma pneumoniae*. He described the motivation for this study as an effort to reproduce a microbe *in silico*; it is not yet possible to simulate an entire human being, but a tiny obligate parasite might be feasible. An initial automated metabolic reconstruction recovered over 80% of the final reaction catalog, but the remainder was unraveled using literature curation. These reactions were validated (and improved) using flux balance analysis, which was in turn validated (and improved again) using nuclear magnetic resonance and mass spectrometry metabolomics. Placing computational and experimental techniques back to back at every step has provided detailed genomic, transcriptional, interaction, and metabolic maps of this organism - but other studies continue today on post-translational regulation, stochastic regulation, and proteomics. Serrano pointed out that over 25% of *M. pneumoniae*'s proteome still remains poorly characterized, and presentations throughout the conference and at the preceding Automated Function Prediction meeting provided additional methods for molecular function prediction. The importance of coupling computational work with detailed experimental validation was also clear, however. Adding to the emphasis on this area's significance, Sara Berthoumieux's work on improving metabolic network reconstruction by accounting for the incompleteness of high-throughput datasets garnered her the Ian Lawson Van Toch Memorial Award for the best student paper.

Translational bioinformatics on the horizon

Only a year after the *New York Times* ran a skeptical article proclaiming, 'Consumers slow to embrace the age

of genomics,' it remained clear at ISMB that the field is in fact squarely in the midst of a transition to translational genomic applications. It is of necessity a slow and careful transition, however, as emphasized by Alfonso Valencia in his ISCB Fellow keynote detailing bioinformatic challenges in personalized cancer treatment. He distinguished between the high-throughput methods at which computational biology excels, including protein docking, chromatin mapping, and evolutionary phylostratification, and the low-noise, high-reliability inferences vital to treatment in the clinic. As he stated, these are difficult to produce quickly or automatically - but contrary to the *New York Times*' expectation, successful studies have resulted in improved, targeted treatments for individual leukemia patients. Echoing Ashburner's call for open research, Valencia pointed out that the absolute certainty required for validation of clinical recommendations is best achieved by collaboration, data sharing, and consolidation throughout the community.

Methodological talks throughout the conference covered areas in which researchers continue to bridge these remaining gaps. The Biological Literature, Information and Knowledge (BioLINK) special session in particular featured presentations on data integration and interoperability across the computational, biological, and medical fields, and it was strikingly the first year in which these have converged to the point of including the session within ISMB itself. Ankur Parikh and Wei Wu, joint recipients of the JBI best translational bioinformatics paper award, reached a further level of specificity with TREEGL, a novel algorithm for reconstruction of human breast cancer cellular lineages from sparse gene expression data. Looking at the earliest stages of translation, my junior faculty keynote at the SCS discussed techniques for characterizing metabolic function in the microbial communities of the human microbiome, dysfunctions of which are increasingly implicated in disease. Translational research topics were pervasive and spanned from molecular mechanisms at the level of individual amino acids to the categorization of patient phenotypes in the clinic, reflecting the field's aspiration and expectation of extending lifespans and saving lives.

The ISCB's Overton Prize is awarded annually to an early- to mid-career investigator who has contributed

significantly to the field of computational biology and who exemplifies the research, education, and service goals of the Society as a whole. Olga Troyanskaya, this year's recipient, presented a keynote that brought together the data integration methodology, biological network modeling, and translational applications highlighted by ISMB. Troyanskaya described a research path that began almost 10 years ago with the first work on integrative bioinformatics in the simple model *S. cerevisiae*, proceeded into its systems biology and that of more complex metazoan models, and continues today with the goal of automatically reconstructing complete pathway activity maps specific to individual tissues and cellular lineages in human disease. Summarizing her own work and the state of the field, Troyanskaya concluded that most of computational biology's achievements may still not reach the clinic within 5 years - but we should expect to see them there within the next decade.

We are frequently shown that bioinformatics and processor design share the mixed blessing of exponential growth, but we are less often reminded of a corollary shared feature: no matter how much we know today, it will be outdated tomorrow. Learning from tremendous data collections, understanding biological network models, and bridging molecular biology with human health have been themes both of the field and of the ISMB/ECCB conferences since their inception. The knowledge and results accumulated within these areas, however, have also grown exponentially; as suggested by the title of this report, even the decades-old field of sequence analysis continues to grow through both new data and novel methodology. We look forward to new surprises at next year's ECCB in Basel, Switzerland, and ISMB in Long Beach, California.

Acknowledgements

I would like to thank Karen Dowell, Daniela Börnigen, and Nicola Segata for their invaluable help in summarizing ISMB/ECCB and the SCS.

Published: 23 August 2011

doi:10.1186/gb-2011-12-8-307

Cite this article as: Huttenhower C: Computational biology: plus c'est la même chose, plus ça change. *Genome Biology* 2011, 12:307.