Pacific Symposium on Biocomputing 12:112-114(2007)

COMPUTATIONAL APPROACHES TO METABOLOMICS: AN INTRODUCTION

DAVID S. WISHART

Depts. of Computing Science and Biological Sciences, University of Alberta, 2-21 Athabasca Hall, Edmonton, AB, T6G 2E8, Canada

RUSSELL GREINER

Dept. of Computing Science, University of Alberta, 2-21 Athabasca Hall, Edmonton, AB, T6G 2E8, Canada

1. Session Background and Motivation

This marks the first time that the Pacific Symposium in Biocomputing has hosted a session specifically devoted to the emerging computational needs of metabolomics. Metabolomics, or metabonomics as it is sometimes called, is a relatively new field of "omics" research concerned with the high-throughput identification and quantification of the small molecule metabolites in the metabolome (i.e. the complete complement of all small molecule metabolites found in a specific cell, organ or organism). It is a close counterpart to the genome, the transcriptome and the proteome. Together these four "omes" constitute the building blocks of systems biology. Even though metabolomics is primarily concerned with tracking and identifying chemicals as opposed to genes or proteins, it still shares many of the same computational needs with genomics, proteomics and transcriptomics. For instance, just like the other "omics" fields, metabolomics needs electronically accessible and searchable databases, it needs software to handle or process data from various high-throughput instruments such as NMR spectrometers or mass spectrometers, it needs laboratory information management systems (LIMS) to manage the data, and it needs software tools to predict or find information about metabolite properties, pathways, relationships or functions.

These computational needs are just beginning to be addressed by members of the metabolomics community. As a result we believed that a PSB session devoted to this topic could address a number of important issues concerning both the emerging computational needs and the nascent computational trends in metabolomics. This year we solicited papers that focused specifically on describing novel methods for the acquisition, management and analysis of metabolomic data. We were particularly interested in papers that covered one of the five following topics: 1) metabolomics databases; 2) metabolomics LIMS; 3) spectral analysis tools for metabolomics; 4) medical or applied metabolomics

and 5) metabolic data mining. In total there were 15 papers submitted to this session, with 8 papers accepted (5 for oral presentation). The papers we received covered an enormous range of metabolomics and computational topics and most were of very high quality. All papers underwent rigorous peer review with up to 4 reviewers for each paper. We are particularly grateful to Lori Querengesser who helped coordinate the review process and the 31 expert reviewers who provided thoughtful and informative comments for each paper.

2. Session Summary

Among the papers accepted for publication in this year's proceedings are two manuscripts concerned with metabolomic databases and metabolomic laboratory information management systems (LIMS). The paper by Markley *et al* describes a suite of databases and LIM systems developed at the University of Wisconsin (Madison) including: 1) the BioMagResBank metabolomics database – which contains experimental NMR data on ~250 metabolites; 2) the Madison Metabolomics Consortium Database (MMCD) database – which contains literature derived data on ~10,000 Arabadopsis-related metabolites and 3) SESAME – a LIM system designed to handle the unique and diverse data needs of metabolomics researchers. Nicely complementing this work by the Madison group is the paper by Scholz and Fiehn, which describes a generalizable metabolomics LIM system called SETUPX. This XML-based system supports nearly every aspect of metabolomic-based lab workflow management, data entry and data processing. It also makes use of publicly available taxonomic and ontology repositories to ensure data integrity and logical consistency. These and other features will likely make SETUPX a gold standard for the design and implementation of other metabolomics LIMS.

Another area of active research in computational metabolomics is data mining and automated data retrieval. This year's proceedings includes two such papers. The manuscript by Knox *et al* describes a metabolome annotation tool, called BioSpider, that specifically seeks out chemical and biological data through text mining and data extraction. It access snippets of data from dozens of websites and electronic databases and assembles them into comprehensive (80+ data fields) "metabo-cards". Given the dearth of electronically accessible data about most metabolites, this tool will likely be very popular within the metabolomics community. A related manuscript by Ganesan *et al*. provides a nice overview and a critical assessment of various data harvesting and data profiling tools in genomics and proteomics and their potential applications to metabolic profiling.

Two papers were also accepted into this year's proceedings which describe and assess the application of existing metabolomic software to real-world medical problems. The paper by Yoon *et al* describes a novel and very sophisticated approach to using metabolic flux profiling to characterize the metabolic fate of the anti-diabetic drug, troglitazone, in the liver. The authors

use experimental data combined with flux balance analysis and metabolic reaction network analysis to unearth some of the underlying reasons for this drug's hepatotoxicty. This provides a superb example of the potential impact that computational metabolomics could have in the area of pharmacology and pharmacotoxicity. A related paper by Yang *et al* describes the application of targeted metabolite identification and ^{13}C isotopomer analysis towards distinguishing between cells derived from normal and cancerous breast tissue. Using existing computational tools, the authors identified a number of important changes in metabolic pathways and the redistrubution of metabolic fluxes that may point to new diagnostic and therapeutic targets.

The paper by Chang *et al* describes the development and comparison of a targeted profiling technique that allows 1D ^{1}H NMR spectra of metabolite mixtures (i.e. blood, urine, tissue extracts) to be analyzed and the compounds within the mixtures to be identified and quantified. The method uses a library of pure metabolite spectra that can be used to fit experimentally collected NMR spectra. The authors compare this targeted profiling method to other spectral analysis tools and demonstrate that their method is generally more robust and often more useful for metabolite analysis.

Finally, the paper by Karakoc *et al* addresses an interesting question about what chemical characteristics distinguish plant or animal metabolites from bacterial or fungal metabolites. This is an important question as many drugs and drug precursors are derived from plants, bacteria and fungi. Likewise, with the advent of high throughput metabolomics, many metabolites are being identified but their biological origins are unclear or unknown. The authors of this paper use classical cheminformatic techniques (QSAR, clustering, neural network analysis) to show that plant, fungal, bacterial and mammalian metabolites do have characteristic features that can distinguish one from another.

Overall, the manuscripts appearing in this year's proceedings provide a nice cross-section of the activities and advances being made in computational metabolomics. No doubt, as the field matures, the focus will likely shift from developing tools and platforms for metabolite data extraction and metabolite data storage to the development of software to aid in the interpretation of that data. Over time, we are hopeful that metabolomic studies will become more integrated with genomic or proteomic studies and that software tools will eventually be developed to aid scientists in the prediction of the physiological or metabolic consequences of drugs, foods or genetic lesions.

Acknowledgments

We would like to thank Genome Alberta, a division of Genome Canada, for their financial support and technical assistance in making this session possible.