Fu and Horton *Genome Medicine* 2010, **2**:7
http://genomemedicine.com/content/2/1/7

Genome **Medicine**

MEETING REPORT

Genome informatics: advances in theory and practice

Szu-chin Fu* and Paul Horton*

Abstract

A report on the 20th International Conference on Genome Informatics, Yokohama, Japan, 14-16 December 2009.

Introduction

In December 2009, 323 people from 11 countries attended the 20th annual International Conference on Genome Informatics, also known as 'GIW' from its former moniker, the 'Genome Informatics Workshop'.

GIW is both a venerable and a timely conference. Venerable because it boasts an almost 20 year history as a venue to report advances in bioinformatics - yet also timely as we head into the era of personal genomics. The authors of this report are a similar mixture; one of us is fortunate to have the distinction of presenting at the very first GIW, while the other can see things with fresher eyes. Together we attempt to summarize some of the results presented at GIW 2009.

Informatics for next-generation sequencing

Primary data analysis

Kouichi Kimura and Asako Koike (Central Research Laboratory, Hitachi Ltd, Kokubunji, Japan) received one of two 'best paper' awards, with a novel localized data structure for suffix arrays, which are a way of enabling efficient searching of large amounts of text such as genome sequences. This new structure efficiently combines lexical information, which makes suffix arrays so powerful and flexible for string matching, with the positional information needed for tasks such as genomic mapping of pair-end reads or sequences produced by splicing events, or chaining nearby short exact matches for gapped alignment. In addition to being theoretically elegant, they demonstrated that the method can increase the speed of mapping pair-end reads by a factor of two to three.

Edward Wijaya (AIST Computational Biology Research Center, Tokyo, Japan) presented RECOUNT, a program to correct transcriptome sequence read counts by subtracting counts that are likely to be the product of sequencing errors. RECOUNT is an efficient implementation of the Expectation Maximization algorithm of Beißbarth *et al.* that can process large next-generation sequencer datasets. Wijaya showed that the method can increase the proportion of mappable tags and, more importantly, avoid some false inferences of expression that would be made with uncorrected data.

Sequence analysis

In his keynote address, Sean Eddy (Howard Hughes Medical Institute Janelia Farm Research Campus, Ashburn, USA) introduced HMMER3, a major update to his popular HMMER software package for hidden Markov model-based search and analysis of similar protein sequences. His two main points were that HMMER3 is now nearly as fast as BLAST, and that it can use 'forward' scores for sequence similarity and give accurate E-values for them. The speed increase is accomplished through the use of BLAST-like heuristics to quickly identify promising matches through ungapped alignment and by using vector parallel instructions, such as SSE2, on Intel microprocessors. The use of 'forward' scores combines the theoretical work of Terry Hwa (University of California at San Diego), Ralf Bundschuh (Ohio State University) and others with empirical testing.

A discussion of 'forward' scores, which sum over all possible alignments, versus methods such as Smith-Waterman (and also BLAST), which consider only the highest scoring alignment, would not have been out of place during the early years of GIW. Yet so-called 'probabilistic' alignment scoring schemes, including 'forward' scores and alignments based on some kind of posterior decoding of such scores, have experienced a renaissance in recent years. We think this demonstrates that some ideas simply take many years to get ironed out by the community.

Medically relevant databases

Next-generation sequencing and other high-throughput measurement technologies provide an ever increasing mass of data at the biomolecular, cellular, and tissue

*Correspondence: Szu-chin Fu. Email: szuchin.fu@gmail.com;
Paul Horton. Email: horton-p@aist.go.jp
Computational Biology Research Center, AIST, and The University of Tokyo, Graduate School of Frontier Sciences, 2-42 Aomi, Tokyo, 135-0064, Japan

BioMed Central © 2010 BioMed Central Ltd

levels. Much of these data have implications for medical research, but they require extensive organization and cross-referencing to be useful in practice.

The winner of the other 'best paper' award was a presentation on recent extensions of VarDB, a database of antigenic sequence variation, by Nelson Hayes (Institute for Chemical Research, Kyoto University, Uji, Japan). VarDB contains more than 62,000 sequences organized by organism and gene family. A unified Ajax-based interface links these data to a variety of analysis and visualization tools, including BLAST, PSI-BLAST, MEME, and Jmol. Codon usage analysis tools are provided to find rapidly evolving regions or search for constraints on sequence variation acting at the DNA or mRNA level. Plugins allow one to view various aspects of the data, such as the chromosomal distribution of potentially antigenic genes or the three-dimensional position of substitutions superimposed on solved protein structures.

In his keynote address, Minoru Kanehisa (Institute for Chemical Research, Kyoto University), one of the founders of GIW, presented the latest developments of the KEGG family of databases. The KEGG DRUG database [http://www.genome.jp/kegg/drug/] provides molecular networks of target and other drug-interacting molecules. It includes the 'Chemical Structure Transformation Network', which holds information on the biosynthetic pathways of natural products and the historical development of many drugs - that is, what lead compounds or existing drugs they are based on. KEGG DRUG also contains chemical structures of all Japanese drugs, including traditional Chinese medicine and 'crude drugs' (unrefined medications in their natural form), as well as most prescription drugs in the US. The KEGG DISEASE database [http://www.genome.jp/kegg/disease/] lists disease genes and other relevant molecules, such as environmental factors, diagnostic markers and therapeutic drugs. It provides some useful information for diseases that are not characterized well enough to draw pathway maps. KEGG MEDICUS integrates the KEGG DRUG and KEGG DISEASE databases and aims to facilitate analyses of network-disease associations.

Conclusions

The algorithmic and software advances presented at this conference will facilitate the transformation of raw sequencer data into reliable sequences and statistically sound inferences about how those sequences relate to previous knowledge. Furthermore, the databases presented will provide access to such knowledge cross-linked to multiple views and contexts.

These advances will certainly have an impact on basic molecular biology, but also on genome medicine. In the near future a medical center may be able to map patient or pathogen sample genome or transcript sequences to their reference genomes with a localized suffix array, correct their abundance counts with RECOUNT, model pathogen protein sequences with HMM3, analyze pathogen antigenic sites with VarDB and give special attention to changes in disease-related genes found in KEGG DISEASE.

GIW covers a broad range of bioinformatic theory and practice, solving old problems and introducing new ones. In December 2010, GIW will celebrate its 20th birthday at the 21st annual conference, appropriately in the ancient but also modern city of Hangzhou, China.

Competing interests

One author (PH) is a coauthor on the work presented by Edward Wijaya mentioned in this report. Neither author was a member of the organizing committee of this meeting.

Published: 26 January 2010

doi:10.1186/gm128
Cite this article as:Fu S-c, Horton P: Genome informatics: advances in theory and practice. *Genome Medicine* 2010, **2**:7.