PROPOSED WORKSHOP "COMPUTATIONAL AND STATISTICAL GENOMICS"

Jenny's edits appear in red. Arrows appear in the margin when I've inserted a comma, which, I, think, we, need, more, of! ⇐

Objectives

The main objective of this workshop is to formulate and address important statistical problems in the analysis of biomedical and high-throughput genomic data, including DNA chip, ChIP-chip, ⇐
SNP, whole-genome sequence and proteomic data. A distinct goal is to facilitate the interaction between biologists performing genome scale experiments ("wet-lab" researchers) and statisticians with expertise and interest in genomics ("dry-lab" researchers). Substantive collaborations between wet and dry lab scientists are vital for transforming the massive amount of data produced by new technologies into important biological discoveries.

The workshop is intended to foster deeper connections between the two research communities and to be a forum for (1) the dissemination of cutting-edge developments, including new high-throughput biological assays and novel statistical methodologies and (2) the identification of open problems in the analysis of these data. These challenges include not only analyzing genotypic data, but also relating these to phenotypic data, such as biological and clinical outcomes, and ⇐
further relating both to meta-data from WWW databases, such as PubMed and Gene Ontology ⇐
(GO) Consortium.

We anticipate that this workshop will enable statisticians to articulate theoretically grounded statistical formulations of existing and emerging computational biology problems; create an exceptional opportunity for exchanging ideas between the communities; and help to shape the future of this dynamic field. Input from wet-lab computational biologists is absolutely crucial for development of appropriate statistical methodologies. For this reason, we have targeted areas that are relatively new to statisticians, as well as areas that have already been greatly influenced by statistical approaches. ⇐
These include phylogenomics, computational population genetics, comparative genomics, microarray technologies and protein structure prediction. We anticipate that the interaction between statisticians and wet-lab computational biologists will lead to major advances in computational and statistical genomics.

Relevance, importance and timeliness

It is now well accepted that the capacity to generate genome data has far outpaced our ability to analyze and interpret it. The rapid development of new and existing high-throughput technologies is allowing biologists to investigate biological processes on an ever- growing scale. Statistical genomics adapts well to these changes, due to the incredible interest of statisticians in these ⇐
methodological challenges. Since this is a relatively new and rapidly developing field, it enjoys an above-average representation of young talent and women. Rapid communication between wet-lab computational biologists and statisticians is absolutely vital. While there are several well-established

computational biology conferences (e.g., ISMB and RECOMB), the primary quantitative discipline has historically been computer science, not statistical science. Likewise, even though major statistical conferences often have sessions on computational biology, the number of those are too few, and the audience is almost exclusively statisticians. Hence, a workshop that specifically brings wet-lab biologists and statisticians together is sorely needed and BIRS provides an unbeatable environment for this task.

Subject area overview

Modern high-throughput technologies are changing the face of biomedical and life science research. Today, researchers investigating a molecule or process in any given organism (including the human) often have the complete DNA sequence of that organism. However, determining the DNA sequence is just the first step in understanding the structure of a genome and the functions of its genes. Where labs used to focus on single genes and proteins, they now aim to integrate vast amounts of ever-changing types of data to study complicated entities, such as protein complexes and regulatory networks. The dawning of the "post-genomic" era of biology requires an interdisciplinary approach. Cooperations among statistics, mathematics, computer science and biology are vital to the future of genome biology. A recent news feature article in Nature (August 7, 2003) emphasizes the importance of sound statistical analysis of genomic data. In reference to the "growing number of statistical experts [familiar with] ... the complexities of microarrays", the article concludes with some advice: "In the meantime,the message to biologists is clear: if you want to work with microarrays, you need to find yourself one of these precious experts and don't wait until after you've collected your data." This observation is even more true for other areas of computational and statistical genomics, such as phylogenomics, protein structure prediction, computational population genetics, and comparative genomics which we highlight later in this section. Better qu ⇐

The pace at which new technologies and data acquisition methods emerge makes computational and statistical genomics an extremely dynamic field. It is our goal to bring wet-lab biologists with interest in computational biology and statisticians working in several aspects of statistical genomics together in this workshop. This would serve as a great opportunity (1) to summarize new advances in biological technologies and state of the art statistical methodologies addressing relevant challenges, (2) to criticize and discuss limitations of the existing methodologies and formulations, (3) to explore ways to solve these issues, and (4) to discuss areas where more interaction among the two communities is needed.

We list below several areas of biological investigation that are fueled by technological advances and require rigorous statistical and computational analysis. There are no strict borders between topics, since most share high dimensional multivariate data that are similar in nature and biological discoveries are often achieved through merging of various sources of data and perspectives. The workshop will focus around these five topics. For each topic, related statistical problems such as parameter specification, estimation, inference and testing, model selection, and statistical computing issues will be addressed. We are aiming to have at least one well-known plenary wet-lab computational biologist and one statistician to speak on each of these topics. Aside from regular talks, poster and software demonstration sessions will provide researchers to present current applications and results on these topics. We have pre-invited some of the possible speakers from both communities including Terry Speed (University of California, Berkeley), John Quackenbush (TIGR), David M. Rocke (University of California, Davis), Wyeth Wasserman (University of British Columbia, Vancouver), Tim Hughes (University of Toronto), Rafael Irizarry (Johns Hopkins), Robert Gentleman (Harvard University), Jason Lieb (University of North Carolina, Chapel Hill), Todd Lowe (University of California, Santa Cruz), and David Hinds (Perlegen Sciences). All of these researchers shared our enthusiasm in such a workshop and showed great interest in participating. ⇐

Can we really do anything about this now? Our pre-invitation list and our longer list of participants does not have nearly the female representation we enjoyed in 2004.

- *Phylogenomics.* Phylogenomics is a new emerging field that combines *genomics* and *molecular phylogenetics*, which are two major fields in the life sciences. Completion of whole genome sequencing projects provides scientists with a unique opportunity to study the origin and evolution of genomes and facilitates improvement of functional predictions for uncharacterized genes by evolutionary analysis. Relevant statistical research includes statistical models for evolution, construction and estimation of evolutionary trees, confidence sets of trees, and statistical models for sequence alignment.
- *Computational population genetics.* Single nucleotide polymorphism (SNPs) are the most simple form and most prevalent source of genetic polymorphism in the human genome. The advent of SNP genotyping and haplotyping technologies are leading to accumulation of massive amounts of SNP data spanning a variety of species. One of the challenges faced by researchers in this field is how to relate such multimillion dimensional genotypic profiles to both biological and clinical phenotypes, such as disease and drug reaction. As more information accumulates, analysis of the emerging complex data requires comprehensive statistical methodologies capable of dealing with challenging issues such as censoring and causality.
- *Comparative genomics.* Comparison of genomes between species aids in every step of the genomic analysis. Some of the areas that gained attention are identification of the differences between related genomes including presence and absence of genes and pathways, and regulatory sequence signals. By incorporating multiple species sequence data with other sources of high-throughput genomic data, scientists are trying to understand how sequence features control the activities of genes and how these features are organized into modules. Evolutionary conservation of regulatory modules and gene expression are also among aspects that can assist us in understanding gene function and regulatory pathways.
- *Microarray technologies.* The types of biological investigations that microarrays enable are increasing rapidly. Today, microarrays are used in areas from gene expression profiling to protein expression profiling and whole-genome profiling of interactions between DNA-binding proteins and DNA. Statisticians have already contributed immensely in improving design and analysis of gene expression microarrays and similar challenges are inevitably arising for other platforms, such as the SNP chips used for high-throughput genome sequencing, that utilize microarray technology.
- *Protein structure prediction.* A crucial step in understanding of human biology is the characterization of all human proteins. This requires the knowledge of their three dimensional structures since structural information leads to protein function and aids in drug design. Protein structure prediction is one of the more elusive goals of computational biology where rigorous modern statistical techniques such as prediction tools for high-dimensional data can provide improvement and new perspectives. Up to date, there are relatively few statisticians involved in this area, thus exposure of more statisticians to this challenging and exciting field is essential.

A list of possible participants and their affiliation

(1) Joe Felsenstein, University of Washington.
(2) Adam Seipel∗, University of California, Santa Cruz.

(3) Bret Larget, University of Wisconsin, Madison.
(4) Lior Pachter, University of California, Berkeley.
(5) Leonid Kruglyak, University of Washington, Fred Hutchinson.
(6) Charles Kooperberg, University of Washington, Fred Hutchinson.
(7) David Hinds, Perlegen Sciences.
(8) Jurg Ott, Rockefeller University.
(9) Hao Li, University of California, San Francisco
(10) Michael Eisen, University of California, Berkeley, LBL.
(11) Wing Wong, Harvard University.
(12) Todd Lowe∗, University of California, Santa Cruz.
(13) Jun Liu, Harvard University.
(14) Tim Hughes∗, University of Toronto.
(15) Jason Lieb∗, University of North Carolina, Chapel Hill.
(16) Joe Derisi, University of California, San Francisco.
(17) Terry Speed∗, University of California, Berkeley.
(18) Rafael Irizarry∗, Johns Hopkins.
(19) Simon Cawley, Affymetrix.
(20) David M. Rocke∗, University of California, Davis.
(21) John Quackenbush∗, TIGR.
(22) Aad van der Vaart, Vrije Universiteit.
(23) Michael A. Newton, University of Wisconsin, Madison.
(24) Heping Zhang, Yale University.
(25) Hongyu Zhao, Yale University.
(26) Neil Clarke, Johns Hopkins University.
(27) Pat Brown, Stanford University.
(28) Wyeth Wasserman∗, University of British Columbia, Vancouver.
(29) Steven Brenner, University of California, Berkeley.
(30) Andrej Sali, University of California, San Francisco.
(31) David Baker, University of Washington.
(32) Ingo Ruczinski, Johns Hopkins University.
(33) Robert Gentleman∗, Harvard University.
(34) Tim Hubbard, The Wellcome Trust Sanger Institute.

Additional comments

A 5-day workshop, organized by Jennifer Bryan, Sandrine Dudoit and Mark J. van der Laan in 2004 at BIRS, brought together statisticians working in different genomics related aspects of statistics. The workshop was a huge success and covered several broad areas of biological investigation that relied on statistical and computational methods. An outstanding aspect of the workshop was participation of many young researchers and women. 23 of the 39 participants were in the category of graduate student/postdoctoral researcher/assistant professor. Moreover, 13 among these were women. We would like to build on this extremely successful workshop but adapt its scope to new emerging fields of statistical genomics and extend the invitee list to wet-lab computational biologists to enable close communications between two communities.

Our preferred week (August 13 - 17) would be particularly advantageous, since the Joint Statistical Meetings (the largest statistics conference in North America) will be concluding in Seattle, WA on August 11.

Dates

Preferred dates.

- August 13 - 17, 2006.
- June 11 - 15, 2006.
- June 3 - 7, 2006.
- July 9 - 13, 2006.
- September 13 - 17 or October 8 - 12 (off-season dates).

Impossible dates.

- May 28 - 31, 2006; Statistical Society of Canada.
- August 6 - 11, 2006; ISMB.
- July 30 - August 4, 2006; IMS.
- August 6 - 10, 2006; JSM.
- Late June, 2006; WNAR.
- March 19 - 22, 2006; ENAR.