Bioinformatics
William S. Bush, Ph.D.
Assistant Professor of Biomedical Informatics
Center for Human Genetics Research
Vanderbilt University, Nashville, Tennessee USA
http://www.sigevo.org/gecco-2012/
william.s.bush@vanderbilt.edu
gettinggeneticsdone.com
chgr.mc.vanderbilt.edu/bushlab

Bioinformatics
Mark S Boguski
National Center for Biotechnology Information, Bethesda, USA

Computer databases, networks and software tools are essential materials and methods for biomedical research and are involved in almost every aspect of disease gene mapping and positional cloning. Public databases of DNA and protein sequences and genetic and physical map information are increasing rapidly in size and complexity and are also improving in quality, comprehensiveness, interoperability and access. A new generation of software tools for navigating through the biomedical literature has become available. Programs for sequence homology searching and genetic map construction have become more sophisticated, yet easier to use. Global computer networks are bringing state-of-the-art capabilities to all.

Current Opinion in Genetics and Development 1994, 4:383–388

About the Instructor

- William S. Bush is an Assistant Professor of Biomedical Informatics in the Center for Human Genetics Research at Vanderbilt University. Using a combination of bioinformatics, basic statistical approaches, and more advanced data mining and machine learning techniques, he studies how patterns of genomic variation influence the function of both individual genes and entire biological systems.
GENOTYPE TO PHENOTYPE

DNA

Chromatin

Table 1. Selected HapMap and HapMap-CapGamer servers.

<table>
<thead>
<tr>
<th>Resource</th>
<th>WWW Database Resource Location of HapMap-CapGamer</th>
<th>Features and comments</th>
</tr>
</thead>
</table>
| Genotype to Phenotype | NIH Common Fund http://commonfund.nih.gov/epigenomics/figure.aspx | 5

Alberts et al. 2002
Prediction of Chromatin States
Ernst et al. 2011

Mapping and analysis of chromatin state dynamics in nine human cell types

T ranscription
Alberts et al. 2002

Regulation of T ranscription
Alberts et al. 2002
Transcriptional Complexity

26 JUNE 2009  VOL 324  SCIENCE

Diversity and Complexity in DNA Recognition by Transcription Factors

Gwenda Badis,1,8 Michael F. Berger,1,2,8 Anthony A. Phillipsakis,1,3,4+, Shaheneye Talukder,1,5,8 Andrew R. Gehke,2,8 Savina A. Jaeger,2,8 Esther T. Chien,3,8 Elena Mietler,2,8 Anastasia Vedenko,2,8 Xiaoyu Chen,5 Huma Karamtev,7 Chi-Fong Wang,7 David Chou,4 Daniel E. Newburger,4 Quaid Morris,4,5,8+ Timothy R. Hughes,4,5,8+ Martha L. Bulik-Sullivan8,9,10

Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project

The ENCODE Project Consortium*

NATURE Vol 447/14 June 2007

Translation

Alberts et al. 2002
Translation

Control of Gene Expression

Protein Folding
Amino Acids

<table>
<thead>
<tr>
<th>AMINO ACID</th>
<th>SIDE CHAIN</th>
<th>AMINO ACID</th>
<th>SIDE CHAIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asparagine</td>
<td>Asn N</td>
<td>Alanine</td>
<td>Ala A</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>Glu E</td>
<td>Glycine</td>
<td>Gly G</td>
</tr>
<tr>
<td>Arginine</td>
<td>Arg R</td>
<td>Valine</td>
<td>Val V</td>
</tr>
<tr>
<td>Lysine</td>
<td>Lys K</td>
<td>Leucine</td>
<td>Leu L</td>
</tr>
<tr>
<td>Histidine</td>
<td>His H</td>
<td>Isoleucine</td>
<td>Ile I</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Gin Q</td>
<td>Proline</td>
<td>Pro P</td>
</tr>
<tr>
<td>Serine</td>
<td>Ser S</td>
<td>Phenylalanine</td>
<td>Phe F</td>
</tr>
<tr>
<td>Threonine</td>
<td>Thr T</td>
<td>Methionine</td>
<td>Met M</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Tyr V</td>
<td>Tryptophan</td>
<td>Trp W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cysteine</td>
<td>Cys C</td>
</tr>
</tbody>
</table>

Polar Amino Acids

Nonpolar Amino Acids

Protein Structural Prediction

DisProt News
- Current release: 5.3
- Released on: 08/29/2010
- Number of proteins: 52,364
- Number of disordered regions: 12,233
- Latest additions:
  - GyrA intermediate chain, octameric [PDB 5gyn]
  - Perm5 family homolog
  - Translocase of chloroplast 159, chloroplast HSC
  - Inner membrane protein AUB0001, chloroplast protein [PDB 5jta]
  - Dihydropyrimidinase [PDB 5gja]
  - More...

Developed at IST

www.disprot.org
Impact of Protein Changes

A method and server for predicting damaging missense mutations
Irene A. Antipul1,7, Stefwit Schmitt1,7, Locnart Meurice1,7, Haney E. Romano1,7, Ana Granados1,7, Paul Buh8, Alyx S. Buddroo8 & Sondi R. Senger2

Biomolecular Interactions Drive Biology

Interactome
Interactome

Some Big Questions

Genetics and Life

THE PHENOGNETIC LOGIC OF LIFE
Genetic Architecture


http://tolweb.org/tree/

The Tree of Life

Molecular Phylogenetics

Molecular Phylogenetics

Multiple alignment

1. AGGCCAAGCCATAGCTGTC
2. AGGCAAGAGCATACCTGAC
3. AGGCCAGACATTACCTGAC
4. AGGCAAGAGCATACCTGAC

Distance matrix

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>0.2</td>
<td>0.05</td>
<td>0.15</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>0.15</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>-</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
The Human Genome Project

Single Nucleotide Polymorphisms

<table>
<thead>
<tr>
<th>Subject #1</th>
<th>Subject #2</th>
<th>Subject #3</th>
</tr>
</thead>
<tbody>
<tr>
<td>-- AGGCTCA --</td>
<td>-- AGGCTCA --</td>
<td>-- AGGCTCA --</td>
</tr>
<tr>
<td>-- AGGTCA --</td>
<td>-- AGGTCA --</td>
<td></td>
</tr>
</tbody>
</table>

Two alleles (G and C)

Three genotypes (GG, GC, CC)

Single Nucleotide Polymorphisms

- ~ 1 SNP every 100 bp
- ~ 30 million SNPs
- ~500,000 SNPs in coding DNA
  - Synonymous (silent)
  - Nonsynonymous
    - Deleterious effect
    - Beneficial effect
    - No effect
Capturing SNPs

Whole genome genotyping technologies on the BeadArray™ platform

Frank J. Steemers and Kevin L. Cudenberg
Illumina, Inc. San Diego, CA, USA

M EASURING RNA

BREAKTHROUGH OF THE YEAR
Human Genetic Variation

Equipped with faster, cheaper technologies for sequencing DNA and assessing variation in genomes on scales ranging from one to millions of bases, researchers are finding out how truly different we are from one another.

The unveiling of the human genome almost 2 years ago cast the first light on our complete genetic makeup. Since then, each new genome sequenced—and each individual studied—has illuminated our genetic landscape in ever more detail. In 2007, researchers came to appreciate the extent to which our genomes differ from person to person and the implications of this variation for understanding the genetics of complex diseases and personal traits.

Less than a year ago, the big news was triangulating variation between us and our primate cousins to get a better handle on genetic changes along the evolutionary tree that led to humans. Now, we have moved from asking what in our DNA makes us human to striving to know what in my DNA makes me me.

Mol Interv. 2002

cDNA Microarrays

1151
Heatmaps

Mol Interv. 2002

MEASURING PROTEINS

2D Gels and Mass Spectrometry


Protein Profiling in Tissues

Am. J. Pathol. 2004
Nanopore Technology

http://www.nanoporetech.com

1000 Genomes Project

http://nihroadmap.nih.gov/hmp/

Our other genomes

http://nihroadmap.nih.gov/hmp/

The Human Microbiome Project

Peter J. Turnbaugh, Ruth E. Ley, Micah Hamady, Claire M. Fraser-Liggett, Rob Knight & Jeffrey I. Gordon

A strategy to understand the microbial components of the human genetic and metabolic landscape and how they contribute to normal physiology and predisposition to disease.
Bioinformatics: Databases


Mining Biomolecular Patterns

- Can we classify and/or predict biological and clinical endpoints using genetic, genomic, and/or proteomic data?
- Which biomolecules are important?
- What is their pattern or statistical relationship?

Objectives

Data → Variable Selection → Statistical Modeling → Prediction Classification

\[ l = a_1x_1 + a_2x_2 \]

Tumor A → Iterate → Tumor B
Objectives

- Data
- Variable Selection
- Statistical Modeling
- Prediction Classification

- $l = a_1x_1 + a_2x_2$

Hypothesis Testing

“The truth is out there”

Truth

$H_0$ False

$H_0$ True

Type I error

Type II error

Type III error

Decision

Accept $H_0$

Yes!

Reject $H_0$

Type I Error

Type II Error

Permutation Testing

- Many data-driven methods are nonparametric and model-free.

- Permutation testing can be used to assess statistical significance to allow formal hypothesis testing.

- Basic Idea: Randomize data so it is consistent with null hypothesis.
**Permutation Testing**

Distribution of Statistic under the Null Hypothesis from Many Permutations

![Graph showing the distribution of a statistic under the null hypothesis, with a critical region indicated by the area under the curve where the value of Z falls.]

- Distribution Known
- Population
- Random Sample
- Inference

- Distribution Unknown
- Population
- Random Sample
- Inference
- Random Samples

---

**Health is a Complex System**


- 50 Research Groups
- 14,000 cases and 3,000 shared controls
- 500,000 SNPs
- Seven complex human diseases:
  - bipolar disorder (BD)
  - coronary artery disease (CAD)
  - Crohn’s disease (CD)
  - hypertension (HT)
  - rheumatoid arthritis (RA)
  - type 1 diabetes (T1D)
  - type 2 diabetes (T2D)
Personal Genome Project

We foresee a day when many individuals will want to get their own genome sequenced so that they may use this information to understand such things as their individual risk profiles for disease, their physical and biological characteristics, and their personal ancestries. To get to this point will require a critical mass of interested users, tools for obtaining and interpreting genome information, and supportive policy, research, and service communities. To catalyze these developments, we launched the Personal Genome Project (PGP).

Type 2 Diabetes
**Type 2 Diabetes**


**A Genetics Company Fails, Its Research Too Complex**

*The New York Times*

DeCode Genetics, a pioneering company that used the Icelandic population as its guinea pigs in detecting disease-causing mutations, filed for bankruptcy on Tuesday.

The company's demise suggests that the medical promise of the human genome may take much longer to be fulfilled than its sponsors had hoped. Based in Reykjavik, Iceland, it was founded in 1996 by Dr. Kari Stefansson, a research neurologist who worked at the University of Chicago and at Harvard. After the human genome sequence was achieved in 2003, Dr. Stefansson quickly realized that Iceland's excellent medical records, combined with the genealogical information available on its close-knit population, provided a fine test bed for seeking the roots of genetically complex diseases like cancer, diabetes and schizophrenia.

**Missing heritability and strategies for finding the underlying causes of complex disease**

*Nature Reviews Genetics*

How should we solve the problem of 'missing heritability' in complex diseases?

Jason H. Moore. The case of the missing heritability for common human diseases should not be a mystery to anyone given the inherent complexity of the relationship between genotype and phenotype.

The time is now to philosophically and analytically retool for a complex genetic architecture or we will continue to undermine our research promises of human genetics.

Indeed, life, and thus genetics, is complicated and some will soon ask, as theonomists have, whether we are trying to predict the unpredictable.
Challenges

- **Complexity** – nonlinearity, heterogeneity
- **Dimensionality** – multiple genetic risk factors
- **Scale** – millions of attributes

Expert Knowledge is Critical

Using Expert Knowledge in GP

- Multi-Objective Fitness (GPTP’06)
  - Fitness = F(Accuracy + Knowledge)
- Recombination (EvoBIO’07)
  - Recombine trees with good building blocks
- Mutation (PRIB’07)
  - Mutate trees with poor building blocks
- Sensible Initialization (CEC’09)
  - Initialize trees with good building blocks
GPTP'07

Does Complexity Matter?

GUIDELINES

From artificial evolution to computational evolution: a research agenda

Wolfgang Banzhaf, Gaiaume Beslon, Steffen Christensen, James A. Foster, François Képès, Virginie Lefort, Julian F. Miller, Miroslav Radman and Jeremy J. Ramsden

A Computational Evolution System – GPTP'08

Mutation Probability
P = 0.1

Mutation Operator
DeleteOp
ChangeOp
ChangeOpArg

Solution Operator
ADD
ADD
DELETE
COPY

Solution

A Computational Evolution System – GPTP'09

A Computational Evolution System – GPTP'10

Mutation Probability
P = 0.1

Mutation Operator
DeleteOp
ChangeOp
ChangeOpArg

Solution Operator
ADD
ADD
DELETE
COPY

Solution

Environmental Noise
Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases

Kristine A. Putlin - Jason H. Moore

STRING
- 2.5 million proteins
- 630 organisms

Experimental Design and Analysis

Datasets = 100 per model
Strong, medium and weak PxP interaction
Runs = 100 per dataset
Generations = 1000
Grid Size = 18x18
Control 1: no expert knowledge
Control 2: attributes with weak PxP interaction
Measured success in finding correct attributes
Results

Table 1-1. Percentage of datasets in which CES successfully identified the correct SNP pairings as the most frequent, for the four confidence score scenarios considered.

<table>
<thead>
<tr>
<th>Confidence Score of PPI</th>
<th>With Expert Knowledge</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.998</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>0.963</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>0.933</td>
<td>96%</td>
<td>0%</td>
</tr>
<tr>
<td>0.916</td>
<td>80%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Summary

- Our CES is able to learn to exploit protein-protein interactions as a source of domain-specific knowledge.
- Can this scale to $10^6$ or more attributes?
Graph Partitioning

Motif Discovery

Sequence Alignment

Latent Variable Discovery

- PCA for ethnic ancestry
- Microarray analysis
The Future

Collaborative Approach Needed

Scientific Visualization

Information Visualization
Introductory Science and Mathematics Education for 21st-Century Biologists

William Bialek and David Botstein

Gallileo wrote that “the book of nature is written in the language of mathematics”.

His quantitative approach to understanding the natural world arguably marks the beginning of modern science. Nearly 400 years later, the fragmented teaching of science in our universities still leaves biology outside the quantitative and mathematical culture that has come to define the physical sciences and engineering. This strikes us as particularly inappropriate at a time when opportunities for quantitative thinking about biological systems are exploding. We propose that a way out of this dilemma is a unified introductory science curriculum that fully incorporates mathematics and quantitative thinking.

These traditions have resulted in a deep bifurcation in culture and quantitative competence among the scientific disciplines. On one branch are mathematics, the physical sciences, and engineering. Scientists educated along this branch achieve a high level of quantitative expertise. They generally have some mastery over and comfort with not only multivariate calculus and differential equa-

- chgr.mc.vanderbilt.edu/bushlab
- www.gettinggeneticsdone.com
- william.s.bush@vanderbilt.edu