Thoughts on Computational Biology at Yale Related to Research, Education & Infrastructure

Mark Gerstein
Computational Biology at Yale

World Class Research

Support: Unit w. MS & PhD staff

Education: CBB PhD Program

Computational Infrastructure (HPC + BioCompute Proposal)
(Molecular) BIOINFORMATICS

**BIOLOGY**

- Data Mining
  - Sequence & Genome Analysis
  - Other 'omic & Network Analyses
  - Medical & Translational Informatics
  - 3D Structure Analysis
  - Systems Analysis

**INFORMATICS**

- Modeling & Simulation

What is Bioinformatics?

• (Molecular) **Bioinformatics**

• One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **“informatics” techniques** (derived from disciplines such as CS, stats & physics) to **organize, analyze, model & understand the information associated** with these molecules, **on a large-scale**.

• Bioinformatics is a practical discipline with many **applications**.

What **Information** to Organize?

- **Sequences** (DNA & Protein)
- 3D Structures
- Network & Pathway Connectivity
- Phylogenetic tree relationships
- Large-scale gene expression & functional genomics data
- Phenotypic data & medical records....
Internet Hosts
(adapted from D Brutlag, Stanford & http://navigators.com/stats.html)

Proteins
Suzek, B. E. et al.

'68 '95 '02 '06
Sequencing Data Explosion: Going to $0/base

From ‘00 to ~’ 20, cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

[Sboner et al. (‘11) GenomeBiology]
Chip Technology

Features per Slide

Features per chip

Chip Technology

transistors

oligo features
General Types of “Informatics” techniques in Computational Biology

• Databases
  - Building, Querying
  - Representing Complex data

• Data mining
  - Machine Learning techniques
  - Clustering & Tree construction
  - Rapid Text String Comparison & textmining
  - Detailed statistics of significance & association

• Network Analysis
  - Analysis of Topology (eg Hubs)
  - Predicting Connectivity

• Structure Analysis & Geometry
  - Graphics (Surfaces, Volumes)
  - Comparison & 3D Matching (Vision, recognition, docking)

• Physical Modeling
  - Newtonian Mechanics
  - Electrostatics
  - Numerical Algorithms
  - Simulation
  - Modeling Chemical Reactions & Cellular Processes
Defining the Boundaries of the Field

(Determining the "Support Vectors")
Are They or Aren’t They Comp. Bio.? (#1)

- (YES?) Digital Libraries & Medical Record Analysis
  ◦ Automated Bibliographic Search and Textual Comparison
  ◦ Knowledge bases for biological literature
- (YES?) Motif Discovery Using Gibb's Sampling
- (YES?) Methods for Structure Determination
  ◦ Computational Crystallography
    • Refinement
  ◦ NMR Structure Determination
    • (YES) Distance Geometry
- (YES?) Metabolic Pathway Simulation
- (NO?) The DNA Computer
Are They or Aren’t They Comp. Bio.? (#1, Answers)

- **(YES?)** Digital Libraries & Medical Record Analysis
  - Automated Bibliographic Search and Textual Comparison
  - Knowledge bases for biological literature
- **(YES)** Motif Discovery Using Gibb's Sampling
- **(NO?)** Methods for Structure Determination
  - Computational Crystallography
    - Refinement
  - NMR Structure Determination
    - **(YES)** Distance Geometry
- **(YES)** Metabolic Pathway Simulation
- **(NO)** The DNA Computer
Are They or Aren’t They Comp. Bio.? (#2)

- ( ) Gene identification by sequence characteristics
  ◊ Prediction of splice sites
- ( ) DNA methods in forensics
- ( ) Modeling of Populations of Organisms
  ◊ Ecological Modeling (predator & prey)
- ( ) Modeling the nervous system
  ◊ Computational neuroscience
  ◊ Understanding how brains think & using this to make a better computer
- ( ) Molecular phenotype discovery – looking for gene expression signatures of cancer
  ◊ What if it included non-molecular data such as age?
Are They or Aren’t They Comp. Bio.? (#2, Answers)

• **(YES)** Gene identification by sequence characteristics
  ◊ Prediction of splice sites
• **(YES)** DNA methods in forensics
• **(NO)** Modeling of Populations of Organisms
  ◊ Ecological Modeling (predator & prey)
• **(NO?)** Modeling the nervous system
  ◊ Computational neuroscience
  ◊ Understanding how brains think & using this to make a better computer
• **(YES)** Molecular phenotype discovery – looking for gene expression signatures of cancer
  ◊ What if it included non-molecular data such as age?
Are They or Aren’t They Comp. Bio.? (#3)

- ( ) RNA structure prediction
- ( ) Radiological Image Processing
  ◊ Computational Representations for Human Anatomy (visible human)
- ( ) Artificial Life Simulations
  ◊ Artificial Immunology / Computer Security
  ◊ ( ) Genetic Algorithms in molecular biology
- ( ) Homology Modeling & Drug Docking
- ( ) Char. drugs & other small molecules (QSAR)
- ( ) Computerized Diagnosis based on Pedigrees
- ( ) Processing of NextGen sequencing image files
- ( ) Module finding in protein networks
Are They or Aren’t They Comp. Bio.? (#3, Answers)

- **(YES)** RNA structure prediction
- **(NO)** Radiological Image Processing
  ◊ Computational Representations for Human Anatomy (visible human)
- **(NO)** Artificial Life Simulations
  ◊ Artificial Immunology / Computer Security
  ◊ **(NO?)** Genetic Algorithms in molecular biology
- **(YES)** Homology Modeling & Drug Docking
- **(YES)** Char. drugs & other small molecules (QSAR)
- **(NO)** Computerized Diagnosis based on Pedigrees
- **(NO)** Processing of NextGen sequencing image files
- **(YES)** Module finding in protein networks
Computational Biology at Yale

World Class Research

Support: Unit w. MS & PhD staff

Education: CBB PhD Program

Computational Infrastructure (HPC + BioCompute Proposal)
• History
  - Started in '02 1st as BBS track & in '03 then as a PhD granting program
  - by M Gerstein & P Miller

• Curr. Structure
  - co-DGses
    M Gerstein [MB&B & CS] & H Zhao [Public Health, Genetics & Stats]
  - DGAs (M Krauthammer & C O'Hern)

History & Current Structure of PhD Program

• Key Numbers
  - 77 matriculated, 34 graduated so far
  - 3 in PEB
  - ~7 students/yr (~40% non-US)
Inputs

- CBB Graduates – Undergrad Majors

<table>
<thead>
<tr>
<th>Biology</th>
<th>Bioinformatics</th>
<th>Informatics</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>3</td>
<td>15</td>
<td>5</td>
</tr>
</tbody>
</table>

- CBB Current Students – Undergrad Majors

<table>
<thead>
<tr>
<th>Biology</th>
<th>Bioinformatics</th>
<th>Informatics</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>8</td>
<td>8</td>
<td>1</td>
</tr>
</tbody>
</table>

- Admissions
  - '14 numbers
    - XXX131162 % US accepted,
    - XXX131162 % foreign accepted,
    - XXX131162 % of the accepts come

- XXXXXXXX – See Shadow
Curriculum: Courses & Competency in Core CBB, Biological Sciences & Informatics

• 10 Courses in Three Core Areas of Competency
  – Computational Biology & Bioinformatics (3 grad courses)
    • CBB 752b Bioinformatics: Practical Applications of Simulation & Data Mining
    • CBB 740a Clinical and Translational Informatics
    • CBB 562a Dynamical Systems in Biology
  – Biological sciences (2 grad courses)
  – Informatics - e.g., CS, stats, app. math (2 grad courses)
  – Electives (2 undergrad or grad courses, in any of the above)

• Competency of incoming students (need to take courses to get to this level)
  – Biology & Natural Science: introductory biology, biochemistry, chemistry
  – CS: introduction to CS, data structures & programming techniques
  – Math & Stat: introduction to probability and statistical inference, multivariate calculus and linear algebra

Students studying over whole campus

Labs of CBB students (incl. rotations) (*=PhD advisor, incl. jt.)

<table>
<thead>
<tr>
<th>Location</th>
<th>Faculty</th>
</tr>
</thead>
</table>
Program is doing well from Grad. Sch. Surveys & Rankings

XXXXXXXX – See Shadow
Program is doing well from Grad. Sch. Surveys & Rankings
### Outputs

- Over last 7 yrs
- Some faculty; many in industry, split betw. traditional bioinfo. route in biotech/pharma & more general "data-science" business positions

<table>
<thead>
<tr>
<th>Year</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002-2007</td>
<td>Assoc Professor, ASU</td>
</tr>
<tr>
<td>2002-2007</td>
<td>Asst Professor, UT</td>
</tr>
<tr>
<td>2005-2010</td>
<td>UCLA Lecturer</td>
</tr>
<tr>
<td>2009-2014</td>
<td>Asst Professor, UNC</td>
</tr>
<tr>
<td>2006-2012</td>
<td>Assoc Bioinformatics Scientist, Children's Hospital of Philadelphia</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002-2008</td>
<td>Postdoc, Stanford University</td>
</tr>
<tr>
<td>2002-2009</td>
<td>Postdoc, Dana Farber Institute</td>
</tr>
<tr>
<td>2004-2010</td>
<td>Resident in General Surgery, Yale</td>
</tr>
<tr>
<td>2007-2012</td>
<td>Computational Biologist, Broad Institute, MA</td>
</tr>
<tr>
<td>2008-2013</td>
<td>Postdoc, Stanford University</td>
</tr>
<tr>
<td>2006-2013</td>
<td>Programmer Analyst II, Yale University</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002-2007</td>
<td>Sr. Bioinformatics Scientist, Illumina</td>
</tr>
<tr>
<td>2004-2009</td>
<td>Data Integration Officer, St. Jude, Memphis</td>
</tr>
<tr>
<td>2003-2010</td>
<td>Scientist, Celgene</td>
</tr>
<tr>
<td>2004-2010</td>
<td>Quantitative Trader, Laurion Capital Mgt</td>
</tr>
<tr>
<td>2005-2010</td>
<td>Director of Informatics, Bina Technologies Inc.</td>
</tr>
<tr>
<td>2005-2010</td>
<td>Investigator, Novartis Institutes for BioMedical Research</td>
</tr>
<tr>
<td>2004-2010</td>
<td>Sr. Developer, Schrodinger, Inc.</td>
</tr>
<tr>
<td>2006-2011</td>
<td>Assoc Principal Scientist, Merck Company</td>
</tr>
<tr>
<td>2005-2011</td>
<td>Product Manager &amp; Bioinformatics Analyst, 5AM Solutions</td>
</tr>
<tr>
<td>2005-2011</td>
<td>Financial firm in Beijing</td>
</tr>
<tr>
<td>2006-2011</td>
<td>Quantitative Analyst, Google</td>
</tr>
<tr>
<td>2005-2011</td>
<td>Data Analyst/NLP Specialist, Elsevier</td>
</tr>
<tr>
<td>2007-2012</td>
<td>Lead Bioinformatics R&amp;D Developer, Regeneron Pharmaceuticals Inc.</td>
</tr>
<tr>
<td>2006-2012</td>
<td>Software Developer, Berkeley Nat Lab</td>
</tr>
<tr>
<td>2009-2012</td>
<td>Information Technology and Services, Germany</td>
</tr>
<tr>
<td>2008-2013</td>
<td>Economic Modeling Senior, Freddie Mac</td>
</tr>
<tr>
<td>2007-2013</td>
<td>Analytics Consultant, SeqWise Next Generation Sequencing Consulting</td>
</tr>
<tr>
<td>2008-2014</td>
<td>Research Scientist, GE Global Research</td>
</tr>
<tr>
<td>2008-2014</td>
<td>Bioinformatics Scientist, Illumina</td>
</tr>
<tr>
<td>2009-2014</td>
<td>Senior Consulting Engineer, Attivio, Inc.</td>
</tr>
<tr>
<td>Year Range</td>
<td>Institution/Company</td>
</tr>
<tr>
<td>--------------</td>
<td>--------------------------------------</td>
</tr>
<tr>
<td>1996–2001</td>
<td>Bank of America</td>
</tr>
<tr>
<td>1997–2002</td>
<td>Goldman Sachs</td>
</tr>
<tr>
<td>1998–2003</td>
<td>Psychogenics</td>
</tr>
<tr>
<td>1999–2004</td>
<td>Pearl Cohen</td>
</tr>
<tr>
<td>2002–2007</td>
<td>Illumina</td>
</tr>
<tr>
<td>2003–2005</td>
<td>Bristol-Myers Squibb</td>
</tr>
<tr>
<td>2002–2010</td>
<td>J.P. Morgan</td>
</tr>
<tr>
<td>2005–2011</td>
<td>MF Global</td>
</tr>
<tr>
<td>2005–2010</td>
<td>23andme</td>
</tr>
<tr>
<td>2003–2007</td>
<td>NEA</td>
</tr>
<tr>
<td>2006–2011</td>
<td>E.L.K. Consulting</td>
</tr>
<tr>
<td>2007–2012</td>
<td>LEK Consulting</td>
</tr>
<tr>
<td>2008–2014</td>
<td>23andme</td>
</tr>
<tr>
<td>2009–2014</td>
<td>Illumina</td>
</tr>
</tbody>
</table>

**Faculty**

- Johns Hopkins
- McGill U
- Yale
- Univ. College London
- U of Toronto
- Miami U.
- McGill U
- Cincinnati Children's Hospital
- Royal Inst. of Technology, Sweden
- Albert Einstein College of Medicine
- U of London
- U of Toronto
- Albert Einstein College of Medicine
- EMBL
- Cornell Medical School
- Tsinghua University
- Dartmouth University
- Mayo Clinic/U of Minnesota
- Weill Cornell Medical College
- NYU (Shanghai)

**Industry**

- Goldman Sachs
- Incyte
- Sigma-Aldrich
- ExxonMobil
- Genellogic
- McKinsey Consulting
- UCB Pharma
- McKinsey Consulting
- Glaxosmithkline
- British Telecom
- Quantitative consulting & writing
- BASF
- NEC
- BioMarin Pharmaceutical

**Output Dataset**

- EBI (Cambridge)
- Cornell U
- Uppsala U
- CUHK

**Majority of industry positions in generalized data-science rather than traditional bioinfo. in biotech/pharma**

**Of 25 faculty positions split betw. bio, cs & bioinfo & later incr.**
US programs in Bioinformatics

For more information see: http://blog.gerstein.info/2014/05/updated-listing-of-us-programs-in.html
Computational Biology at Yale

World Class Research

Support: Unit w. MS & PhD staff

Education: CBB PhD Program

Computational Infrastructure (HPC + BioCompute Proposal)
Yale Life Sciences HPC

• Current workhorses
  – BulldogN [W Campus Seq. Ctr.]: 2Pb, 2.6K cores
    • used by ~20 groups (at 1% level) w/ 5 big users on each (~5% level)
  – Louise [300 George]: 1Pb, 3.5K cores
    • Similar usage profile to BulldogN ("20 & 5")
  – Omega: 1.4Pb, 8.5K cores
    • Phys. Sci. cluster, small use by ~10 bio. groups

• Future
  – Grace: 1 Pb, 1.6K cores
  – Louise & BulldogN to fold into Grace,
    most compute hardware moving to WC
  – Expanding Grace storage
    & mounting it on all clusters as a shared resource
XXXXXXX – See Shadow

- XXXXXXX – See Shadow
XXXXXXX – See Shadow
Technical Architecture

• XXXXXXX – See Shadow
Cancer Genomics & PDX Use Case

• Importance of topic obvious
• JAX is rapidly accruing genomics data for many PDX (Patient-derived xenograft models) samples
  – Expect the scale of data in next year to be 100-200 TB.
• Desire to analyze data, collaborate, merge data & compare with public cancer genomics information
At Yale: Researchers developing systems for analyzing cancer genomes

- Variant Calling
- Recurrence Analysis
- Mutation Prioritization
- All req. access to many sequenced genomes for context

[Khurana et al., *Science* ('13)]
TCGA endpoint: ~2.5 Petabytes
~1.5 PB exome
~1 PB whole genome

SRA >1 petabyte

1000 Genomes
A Deep Catalog of Human Genetic Variation

National Human Genome Research Institute

ADSP

Sofia, 2-28-14
TCGA: What’s in a petabyte?

>30 TCGA Cancer Types
>73K Experiments
>11K Patients

https://cghub.ucsc.edu/
Biocompute Comparables

- **Princeton** (only FAS)
  - Della Cluster - 2816 cores, 2PB storage

- **Columbia** (FAS+med+seq. ctr.)
  - C2B2 - 6336 CPU cores, 73,728 GPU cores, 1.4PB storage
  - NY Genome Center - 2,000 CPU cores, 2PB storage

- **Harvard**
  - Odyssey Cluster - 60,000 cores, 79,872 CPU cores, 14PB storage
  - Massachusetts Green High Performance Computing Center
    - Incl. part of Odyssey
    - MIT, Harvard, NEU, BU, UMASS
    - $95M

- **Texas**
  - Texas Advanced Computing Center (TACC): 203K CPU cores, 319K GPU cores, 14PB storage, 200Tb of RAM!

(Extracted from public websites)
Computational Biology at Yale

World Class Research

Support: Unit w. MS & PhD staff

Education: CBB PhD Program

Computational Infrastructure (HPC + BioCompute Proposal)
Computational Biology at Yale

World Class Research

Support: Unit w. MS & PhD staff

Education: CBB PhD Program

Computational Infrastructure
(HPC + BioCompute Proposal)
• Current PhD program with many students & grads (>75,>35)
  – Balanced combination of Bio., Informatics & focused Bioinformatics
  – "Happy" students & diverse outcomes
  – Rise of Data Science as a driver for education
  – Students studying over whole campus

• Importance of robust computational infrastructure
  – Expertise for cloud computing
  – Necessary to tackle future problems in cancer genomics
  – More so than physical buildings!

Key points & challenges

• Challenge: Quality People!
  – Importance of getting highest quality faculty, students & computational staff
  – Often it's hard for people outside the field to judge & recruit

• Challenge: Unifying 3 locations for CBB at Yale
  – "Embedding" computational faculty, students & fellows but still giving them a coherent identify
    • Addressed by program, but what for faculty & postdocs ?
    • XXXXXXX – See Shadow
Info about content in this slide pack

• General PERMISSIONS
  – This Presentation is copyright Mark Gerstein, Yale University, 2014.
    – Feel free to use slides & images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).
    – Paper references in the talk were mostly from Papers.GersteinLab.org.

• For SeqUniverse slide, please contact Heidi Sofia, NHGRI

• PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info .
  – In particular, many of the images have particular EXIF tags, such as kwpotppt , that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwopotppt