molecular biology is information science .......

-- Leroy Hood, RECOMB'00
Bioinformatics
(or computational biology)

This interdisciplinary science ... is about providing computational support to studies on linking the behavior of cells, organisms and populations to the information encoded in the genomes.

Bioinformatics

- It is about using computational techniques to
  - interpret biological data
  - predict structures and functions of biological entities
  - model the dynamic behavior of biological processes and systems
  - …..

- People have used mathematical or computational techniques to solve biological problems since early 1900’s
  - e.g., evolution and genetic analyses by R.A. Fisher, J.B.S. Haldane, S. Wright

- So what is new?
A Historical Perspective

- Realization of the existence of “genes” in our cells by Hermann Müller, a student of Morgan’s (1921)

- Understanding of the physical natures of genes by
  - F. Sanger (e.g., 1949),
  - E. Chargraff (e.g., 1950),
  - J. Kendrew (e.g., 1958)
in 40’ and 50’s
A Historical Perspective

- Understanding of the double helical structure of DNA by James Watson and Frances Crick in 1953

- Development of sequencing technology, first of proteins and then of genomic, based on the work of
  - F. Sanger on sequencing of insulin (1956),
  - W Gilbert and A. Maxam on sequencing of Lactose operator (1977)

which demonstrated that the genetic sequence of a genome, including human’s, is sequence-able!
A Historical Perspective

- Development of the *science* of analyzing protein and DNA sequences, particularly in
  - protein sequence analyses and evolution by Margaret Dayhoff (60's)
  - phylogenetic analyses and comparative sequence analyses by W. Fitch and E. Margoliash (1967) and by R. Doolittle (1983)
A Historical Perspective

- Development of sequence comparison algorithms
  - Needleman and Wunsch (1970)
  - Smith and Waterman (1981)

- Organization of biological data into databases
  - GENBANK (1982) of DNA sequences
  - Protein Data Bank (PDB, 1973) of protein structures

- Computational methods for gene finding in genomic sequences
  - Work by Borodovsky, Claverie, Uberbacher from mid-80’s to early 90’s
A Historical Perspective

- Sequencing of Human and other genomes
- Development of “high-throughput” measurement technologies
  - microarray chips for functional states of genes
  - two-hybrid systems for protein-protein interactions
  - structural genomics for structure determination
  - ........
A Historical Perspective

- These "high-throughput" probing technologies and others are being used to generate enormous amounts of data about the existence, the structures, the functional states, the relationships of biological molecules and machineries but …..

- ….. what are these data telling us? …..
So what is new?

It is the amount & the type of biological data about the cellular states, and molecular structures and functions, generated by high-throughput technologies, that have driven the rapid advancement of bioinformatics!
An Example of Computation for Biology

- **Goal:** to understand the phosphorus assimilation process in *Synechococcus* (a microbe in ocean) at molecular level
  - Observation: When the living environment is low in phosphate, *Synechococcus* cells will turn on different machineries with higher affinity for in-taking particular forms of phosphate
  - Question: how is this regulatory mechanism implemented at the molecular level?

- **Hypothesis driven research**
  - Need to come up with a hypothesized model
  - Search for genes that may be involved in the proposed model
    - perturbation (gene knock-out/mutation) and phenotypic observations (data analysis)
  - .................
  - Model revision based on collected data, and iterate!
An Example of Computation for Biology

- X years ago, .... to search for which genes are possibly involved in this process, researchers had to
  - remove various parts of DNA sequence,
  - then observe if they may have any relevance

```
acggtcgtacgtacgttacgtgctagccgataatccagtgtgagatacatcatacatcgaacacatgaggccgtcagatgatcc....
```

This could be a very lengthy process ......
An Example of Computation for Biology

- Since the inception of Human genome project (1986), computational scientists have developed computer programs to locate genes in a long stretch of DNA sequence
  - GRAIL, Gene-Scan, Glimmer, .......

acggtcgatcgtacgtgttagccgataatccagtgtgatagatacatacatcgaaacacatgagggcgtcgatagatgc......

- With gene-prediction programs, researchers only needed to knock-out regions predicted to be genes in their search for understanding of phosphorus assimilation process

Great saving in time .........
An Example of Computation for Biology

- Over the years, many genes have been thoroughly studied in different organisms, e.g., human, mouse, fly, ..., rice, ...  
  - their biological functions have been identified and documented

- Computational scientists have developed computer programs to associate newly identified genes to genes with known functions!
  - Existing methods can associate > 60% of newly identified genes to genes with known functions

- Now, researchers only need to knock-out genes with possibly relevant functions in their search for understanding of a particular biological process .......

Saving even more time ........
An Example of Computation for Biology

- **Deciphering a biological process**, e.g., phosphorus assimilation, involves more than just gene finding/functional assignments
  - interactions between gene products (proteins)
  - dynamic behavior of genes throughout a biological process
  - mechanistic issues re. how a protein implements its function
  - ........

- **Computational programs are springing out that can**
  - predict if two proteins are interacting with each other
  - model the dynamic behavior of metabolic pathways ....
  - predict protein structures to help answer mechanistic issues of its function
  - ........

... and many more yet to come .....
Computation for Biology

computation can help quickly narrow down the search space

Searching a needle in a haystack …
Examples of “Computation for Biology”

• Suggesting functions of newly identified genes
  - It was known that mutations of NF1 are associated with inherited disease neurofibromatosis 1; but little is known about the molecular basis of the disease.
  - Sequence search found that NF1 is homologous to a yeast protein called Ira, which is a GAP-type protein and known to regulate the function of a second type of protein called Ras.
  - Hypothesis: NF1 regulates Ras in human cell; follow-up experiments verified this.
Examples of “Computation for Biology”

• Computer-assisted drug design
  • 3D structure models of G protein-coupled receptors were used to computationally screen 100,000+ compounds as possible drug targets and 100 were selected
  • Follow-up experiments confirmed a high hit rate of 12%-21%

Examples of “Computation for Biology”

- Computational analysis of *Plasmodium falciparum* metabolism
  - *Plasmodium* causes human malaria
  - computational prediction of metabolic pathways of plasmodium
  - computational simulations have helped to identify 216 “chokepoints” in this pathway model
  - among all 24 previously suggested drug targets, 21 target at the “chokepoints”
  - among the three popular drugs for malaria, they all targeted at the “chokepoints”

Yeh I. et. al. Genome Research 2004, 14:917-24
Computation for Biology

- Computation can help quickly reduce the search space of “searching a needle in a haystack”
- Help rationally design experiments
- Help to formulate new hypotheses
- …

It is becoming an essential part of the tool kit of modern biology!
What Can Computation Do for Biology

- **Evolution-based computation -- deriving structural, functional and evolutionary information**
  - Sequence alignment algorithms for detection of homologues, functional inference, etc.
  - Multiple sequence alignment algorithm: detection of functional sites, understanding of genome rearrangement, etc.
  - Phylogeny reconstruction algorithm: estimation of evolutionary distance, evolutionary tree reconstruction, etc.
  - Protein threading algorithms: identification of remote homologues, prediction of protein 3D structures, etc.
  - ....
**What Can Computation Do for Biology**

Multiple protein sequence alignment

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Protein Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSJ849/133-201</td>
<td>RKOSFPK -- FKTSRTH- -- YrDNRETJAMK - FYHTGKFEK -- RCSSCKFV</td>
</tr>
<tr>
<td>O3964S/89-157</td>
<td>RKDFSRO -- yKFKAKSKTNS -- IDLADMGC - FYYGTYKTEV -- RCAYGILV</td>
</tr>
<tr>
<td>OSYMP8/85-153</td>
<td>RASKRA -- FAKOGK -- YgSDANALEACG - FTYWRCFAA -- OCSRQGNY</td>
</tr>
<tr>
<td>OS6307/85-152</td>
<td>RKFSSS -- FKVARRFQkSHnKLIDLEMSEK - FCYFGKKAOL -- RCYGGKVV</td>
</tr>
<tr>
<td>IAP2_NPYOP/85-150</td>
<td>RKRFAQ -- FKVARRFQkSRaFEWDMLERG - FCYFGK -- RL -- RCAGCKVV</td>
</tr>
<tr>
<td>Q65368/85-150</td>
<td>RKRFAQ -- FKVARRFQkSRaFEWDMLERG - FCYFGK -- RL -- RCAGCKVV</td>
</tr>
<tr>
<td>082438/85-152</td>
<td>RKRFAQ -- FKVARRFQkSRaFEWDMLERG - FCYFGK -- RL -- RCAGCKVV</td>
</tr>
<tr>
<td>IAP2_NPVAI/85-152</td>
<td>RKRFAQ -- FKVARRFQkSRaFEWDMLERG - FCYFGK -- RL -- RCAGCKVV</td>
</tr>
<tr>
<td>089789/95-159</td>
<td>RKRFAQ -- FKVARRFQkSRaFEWDMLERG - FCYFGK -- RL -- RCAGCKVV</td>
</tr>
<tr>
<td>MRC90/90-185</td>
<td>RKRFAQ -- FKVARRFQkSRaFEWDMLERG - FCYFGK -- RL -- RCAGCKVV</td>
</tr>
<tr>
<td>OS6380/130-188</td>
<td>RKRFAQ -- FKVARRFQkSRaFEWDMLERG - FCYFGK -- RL -- RCAGCKVV</td>
</tr>
<tr>
<td>MRC90/90-185</td>
<td>RKRFAQ -- FKVARRFQkSRaFEWDMLERG - FCYFGK -- RL -- RCAGCKVV</td>
</tr>
</tbody>
</table>

Conserved sites and hence possibly functional sites

**Phylogenetic tree**

[Link to Georgia Cancer Coalition]
What Can Computation Do for Biology

- **Physics-based computation** – helping to derive information about functional mechanisms
  - *Ab initio* folding of protein structures – calculation of structural conformation that has the lowest potential energy
  - Protein docking calculation – docking of two protein structures that gives the best physical and geometric complementarity
  - Simulation of metabolic pathways – network simulation based on biomass conservation, kinetics equations, …
  - ……..
What Can Computation Do for Biology

Molecular dynamics simulation

Metabolic pathway simulation

Figure 1 - Simplified metabolic pathway activated by a signal transduction cascade.
What Can Computation Do for Biology

- **Statistics-based computation** – discovering patterns in unstructured data
  - Computational methods for gene finding – recognition of distinguishing statistical features of coding regions from non-coding regions
  - Computational methods for promoter identification – recognition of distinguishing features of promoter regions from non-promoter regions
  - Computational methods for microarray gene expression data interpretation – identification of significant patterns in a data set, which are possibly associated with a particular experimental condition
  - .......
What Can Computation Do for Biology

gene finding

gene chip data analysis

proteomic data analysis
What Can Computation Do for Biology

- **Bioinformatic support for modern biology**
  - Data storage, management, representation, and dissimilation: modern biology generates great amount of data
  - Data interpretation for analytical instruments – x-ray crystallographic/NMR data interpretation, mass spec. data interpretation, DNA sequencer data interpretation, ….
  - Computational infrastructure for modern biology – sample tracking systems, laboratory information management systems, …
  - ……
What Can Computation Do for Biology?
What Can Computation Do for Biology

- data management/process; data mining; modeling; prediction; hypothesis formulation

<table>
<thead>
<tr>
<th>bioinformatics</th>
</tr>
</thead>
<tbody>
<tr>
<td>engineering aspect</td>
</tr>
</tbody>
</table>

an indispensable part of biological science

genes, proteins, protein complexes, pathways, cells, organisms, eco-systems .............
What Can Biology Do for Computational Science

- *What computational science is to molecular biology is like what mathematics has been to physics* ......
Change of Paradigm

• The human genome sequencing project has led to fundamental changes in how biological science is done!
  - It represents biology’s first foray into ‘big science’ – *Science*, editorial, 2003

• The coordinated efforts, funded by federal and private agencies, in “high-throughput” production of biological data beyond sequences have fueled the rapid transition of biology from “cottage industry” science to “big science”
  - functional genomic data
  - structural genomic data
  - proteomic data
  - haplotype mapping
  - metabolomic data
  - .....
Change of Paradigm

high-throughput data production

data mining and interpretation

models or hypotheses

domain expertise

rational experimental design

specific data generation
Integrative Biology

• “Howdy, want to do biology together?”
Examples of Integrated Computational and Experimental Biology

- Structure prediction of *vitronectin* (a three-domain protein) through combined computational and experimental work.

- Xu lab did computational prediction of each of the domains of *vitronectin*.

- Peterson lab of UTK did NMR work to determine a few cross-links between domains.

- Together we did the structure prediction.

We are carrying out similar work with Prestegard lab in a systematic manner.
Examples of Integrated Computational and Experimental Biology

- Operon prediction in prokaryotic genomes
- Xu lab did operon prediction in *cyanobacteria*
- Palenik lab of UCSD did experimental validation of some of the predicted operons
- Together we predicted operon prediction of cyanobacteria at genome scale

Xu lab and Adams lab are carrying out similar studies on *P. furiosus*
Examples of Integrated Computational and Experimental Biology

- Prediction of nitrogen assimilation pathway and its cross-talk network with photosynthesis pathway in *cyanobacteria*

- Xu lab did the initial prediction of component genes and possible interactions among the gene products in these pathway/network

- Palenik lab of UCSD did a number of microarray chips, knockout experiments and proteomic experiments, based on our initial predictions

- Through a few iterations, we together have predicted the nitrogen assimilation pathway and its cross-talk network with photosynthesis
Cross-talk Network between Nitrogen Assimilation and Photosynthesis

- **Light**
- **Nutrients**
- **CO₂**

**Photosystem**
- PsbA3, ApcF, PetF2, ...
- PetH, Cytb6/f, CP43, ...

**Calvin cycle**
- RbcL, RbcS, Icd

**ATP**
- NADPH⁺

**Other pathways**
- Hypothetical proteins

**Transcription factor**
- NtcA regulon
- Non-ntcA regulon

**Transporter genes**
- NrtP, NrtP
- Urt, Urt
- Cyn, Cyn

**Protein**
- Synt273, Synt2289
- PetH, 2468, 2469, 2474

**Other pathways**
- Cyanate
- Urea
- Urea

**Translation/translocation**
- GS, PII

**Regulation**
- Cyanase
- Urease

**DNA**
Examples of Integrated Computational and Experimental Biology

- Understanding the mechanisms of alternative splicing in eukaryotes
  - >80% of proteins in human cells are splicing variants

- Incorrect balance of different isoforms in a cell could cause diseases
  - cystic fibrosis (CF)
  - spinal muscular atrophy (SMA)
  - Frontotemporal dementia
Examples of Integrated Computational and Experimental Biology

- Our lab has been doing prediction of regulatory elements involved in regulation of alternative splicing
“Systems Biology”

- Integration of computational, experimental biology and high-throughput probing technologies allows scientists to study “systems” rather than individual elements.

- One argument is that a “system” is more than just the sum of all individual parts; it could have emerging properties of its own, based on the complex (2-, 3-, …, n-body) interactions among its parts.

- While systems emerging properties have been studied in other disciplines, such as stock market, not much has been discussed in the setting of biology.
“Systems Biology”

• Used previously published data to construct initial model of the pathway model

• Refined the model using various experimental data

• Validated the final pathway model through comparing dynamics simulation results and time-course data (metabolites)

• Revealed a number of interesting features of the pathway, useful for guiding further experimental design
“Systems Biology”

• ..... a “system” is probably more than just the sum of all individual parts; it could have emerging properties of its own
Like physics, where general rules and laws are taught at the start, biology will surely be presented to future generations of students as a set of basic systems... duplicated and adapted to a very wide range of cellular and organismic functions, following basic evolutionary principles constrained by Earth’s geological history.

– Temple Smith, Current Topics in Computational Molecular Biology
Take-Home Message

• MANY biological databases about functions and structures of biomolecules, processes and systems have been developed and provided with powerful search engines
  – so much information is at the fingertips of biologists
  – why NOT to use them?

• By fully combining experimental techniques with computation, biologists can probably do much more than what they can do using experiments (or computation) alone

• More than ever, biological studies need infrastructure such as bioinformatics and high-throughput measurement technology, to carry out complex system-level studies
That is all, Folks.

THANK YOU!