1. Doing Biology by Comparing Whole Genomes

- the opportunity: by aligning (large parts of) genomes, we can potentially identify important genomic features such as
  - phenotypic differences
  - genes
  - other “signals”
1. Doing Biology by Comparing Whole Genomes

- example of uncovering important phenotypic differences
  - identifying islands of pathogenicity in E. coli
  - Perna et al., *Nature* 2001

- the Twinscan gene finder uses conservation between two closely related genomes to do gene prediction
  - Korf et al., *Bioinformatics* 2001
1. Doing Biology by Comparing Whole Genomes

- technical challenges
  - scaling up alignment algorithms
  - detecting genome rearrangements
  - identifying what’s conserved for functional reasons

"Of course there's more to science than just hurting animals, but frankly it’s the part I like best."  – scientist in Dilbert

2. Learning and Modeling Biological Networks

- the opportunity: high-throughput technologies open up the possibility for characterizing cells/organisms at the systems level
- understanding the networks of interactions and pathways that are involved in
  - metabolism
  - gene regulation
  - signaling

“My view of biology is we don't know shit.”  
– Craig Venter, in New Yorker, 6/12/00
2. Modeling Networks

- example: Segal et al, *Nature Genetics* 2003
2. Modeling Networks

• example: Prof. John Yin’s work (UW ChemE)

<table>
<thead>
<tr>
<th>Concept</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein rate</td>
<td>$v_1 = \frac{dP_1}{dt}$</td>
</tr>
<tr>
<td>Translation rate</td>
<td>$v_T = k_T R$</td>
</tr>
<tr>
<td>Dynamic deviation factor (DDF)</td>
<td>$D = (v_1 - v_2) \nu$</td>
</tr>
<tr>
<td>Protein correlation coefficient (PCC)</td>
<td>$C_p = \frac{\sum x_i y_i dt}{\left( \sum x_i^2 dt \sum y_i^2 dt \right)^{1/2}}$</td>
</tr>
</tbody>
</table>

2. Modeling Networks

• technical challenges
  – data are noisy
  – data are incomplete – characterize a limited range of conditions
  – important aspects of the system typically not measured
  – model spaces are large and complex
3. Developing and Using Ontologies

- challenge: annotating genomes with a rich, controlled vocabularies of terms

“Scientists would rather share each other’s underwear than use each other’s nomenclature.”
– Keith Yamamoto, Biochemist

3. Developing and Using Ontologies

- example: The Gene Ontology (GO)
  - GO has thousands of terms organized in three areas
    - *biological process*: biological objective to which the gene product contributes
    - *molecular function*: biochemical activity of a gene product
    - *cellular compartment*: place in the cell where a gene product is active
  - being used to annotate genes by most of the genome databases (SGD, Flybase, WormBase, RGD, etc)
3. Developing and Using Ontologies

- a piece of the Gene Ontology

4. Biomedical Text Analysis

- challenge: much of the relevant data for doing biology/medicine is represented only in natural language text
- some tasks in biomedical text analysis
  - extracting keywords/keyphrases for annotating gene/protein families
    - identifying relationships that are implicitly, but not explicitly described in the literature
      [Swanson & Smalheiser, 1997]
    - recognizing and extracting instances of entities and relations of interest
      [Chang et al. 2001]
4. Biomedical Text Analysis: The Information Extraction Task

Analysis of Yeast PRP20 Mutations and Functional Complementation by the Human Homologue RCC1, a Protein Involved in the Control of Chromosome Condensation

Fleischmann M, Clark M, Forrester W, Wickens M, Nishimoto T, Aebi M

Mutations in the PRP20 gene of yeast show a pleitropic phenotype, in which both mRNA metabolism and nuclear structure are affected... By immunofluorescence microscopy the PRP20 protein was localized in the nucleus. Expression of the RCC1 protein can complement the temperature-sensitive phenotype of PRP20 mutants...

protein(PRP20)
subcellular-localization(PRP20, nucleus)

Representing Sentences as Nested Sequences of Tokens

<table>
<thead>
<tr>
<th>NP_segment</th>
<th>adjective</th>
<th>noun</th>
<th>Our results</th>
</tr>
</thead>
<tbody>
<tr>
<td>VP_segment</td>
<td>verb</td>
<td>c_m</td>
<td>that</td>
</tr>
<tr>
<td>NP_segment: PROTEIN</td>
<td>noun</td>
<td>unknown: PROTEIN</td>
<td>protein Bed1</td>
</tr>
<tr>
<td>VP_segment</td>
<td>cop</td>
<td>verb</td>
<td>is</td>
</tr>
<tr>
<td>prep</td>
<td>prep</td>
<td>in</td>
<td></td>
</tr>
<tr>
<td>NP_segment: LOCATION</td>
<td>art</td>
<td>unknown: LOCATION</td>
<td>ER</td>
</tr>
</tbody>
</table>
5. Molecular Medicine

- opportunity: using molecular-level data to improve
  - diagnosis
  - prognosis
  - treatment (including drugs)

- relevant data include
  - genotypes (genomic sequences)
  - mRNA quantities (e.g. from microarrays)
  - protein quantities
5. Molecular Medicine

**HIV Mutations & Drug Resistance**

- HIV mutations appear spontaneously and continuously in a given patient
- high rate of mutation: it is estimated that in a full-blown AIDS case, every single point mutation occurs every day! [Condra et al., 1995]
- current FDA-approved drugs target two HIV proteins
  - reverse transcriptase
  - protease
- what happens when a mutant is resistant to a drug?
  - it will outcompete other strains, become the dominant strain
  - *selective drug resistance*

**HIV Drug Therapies**

- typical treatment: drug combination
  - as many as 4 simultaneous drugs
  - typically not more than 3 because of side-effects and toxicity
- circa 1998, there were 407 different combination treatments, each consisting of 4 or fewer drugs
- the CTSHIV System [Lathrop et al., 1998]
  - given: HIV strains infecting a patient
  - do: recommend a customized drug combination \( D \)
Evaluating Resistance to D

• Q: given an HIV mutant and a drug combination $D$, how should we assess the mutant’s resistance to $D$?

<table>
<thead>
<tr>
<th>mutant</th>
<th>resistance</th>
<th>drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGGTACGATGGACA</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td></td>
</tr>
</tbody>
</table>

• A: the resistance of the least resisted drug

Predicting Nearby Resistant Mutants

• consider hypothesized mutants in addition to current sequences

- AGGTACGATAGACA
- AGGCTCGATAGACA
- AGGTACGATAGACA
- AAGTACGAAGAACA
- ACGTACGATAGTCA

- consider current hypothesized mutants

- AGGTACGATGGACA
- AGGTACGATCGACA
- AGGTACGATGGACA

• use rules to decide which mutants to consider
Evaluating Resistance to D

• Q: given a population of mutants and a drug combination $D$, how should we assess the population’s resistance to $D$?

<table>
<thead>
<tr>
<th>mutant</th>
<th>resistance</th>
<th>drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGGTACGATGGACA</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>AGGTACGAAGGACA</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td></td>
</tr>
</tbody>
</table>

• A: the resistance of the most resistant mutant

Ranking Alternatives

• rank alternative drug combinations as follows:

$$f(D) = \sqrt{\text{CurrWt}^2(D) + \text{MutScore}^2(D)}$$
Results

• phase 1 clinical trial
  – 14 patients that had detectable viral load and failure of at least one previous regimen
  – at end of first year
    • 9 had no detectable viral load
    • 1 had 25X reduction in viral load
    • 2 treatment failures
    • 2 withdrew