Life @ Lehigh

Lehigh Life Days • April 12, 2006

Dan Lopresti
Associate Professor
Office PL 404B
dal9@lehigh.edu
Bioinformatics

What is bioinformatics?
Application of techniques from computer science to problems from biology.

Why is it interesting?
- Important problems.
- Massive quantities of data.
- Desperate need for efficient solutions.
- Success is rewarded.
“Biology easily has 500 years of exciting problems to work on.”
Donald Knuth

By developing techniques for analyzing this data, we can attempt to understand genetic nature of diseases, evolution of life itself, etc.

http://cmgm.stanford.edu/biochem218/
Opportunity

Genetic identity of most organisms is encoded in long molecules made up of four basic units, the nucleic acids:

1. Adenine,
2. Cytosine,
3. Guanine,
4. Thymine.

To first approximation, DNA is language over 4 character alphabet, \{A, C, G, T\}.

Complete set of chromosomes that determines an organism is known as its *genome*.

<table>
<thead>
<tr>
<th>GenBank Release 121.0 — December 15, 2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species</td>
</tr>
<tr>
<td>---------------------------------------</td>
</tr>
<tr>
<td>Homo sapiens</td>
</tr>
<tr>
<td>Mus musculus</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
</tr>
<tr>
<td>Tetraodon nigroviridis</td>
</tr>
<tr>
<td>Oryza sativa</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
</tr>
<tr>
<td>Bos taurus</td>
</tr>
<tr>
<td>Glycine max</td>
</tr>
<tr>
<td>Medicago truncatula</td>
</tr>
<tr>
<td>Trypanosoma brucei</td>
</tr>
<tr>
<td>Lycopersicon esculentum</td>
</tr>
<tr>
<td>Giardia intestinalis</td>
</tr>
<tr>
<td>Strongylocentrotus purpur</td>
</tr>
<tr>
<td>Entamoeba histolytica</td>
</tr>
<tr>
<td>Hordeum vulgare</td>
</tr>
<tr>
<td>Danio rerio</td>
</tr>
<tr>
<td>Zea mays</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
</tr>
</tbody>
</table>

http://www.cbs.dtu.dk/databases/DOGS/
http://www.nsrl.ttu.edu/tmot1/mus_musc.htm
http://www.oardc.ohio-state.edu/seedid/single.asp?strID=324
Comparative Genomics

Mouse and Human Genetic Similarities

Mouse chromosomes

Human chromosomes

Courtesy Lisa Stubbs
Oak Ridge National Laboratory

Gel electrophoresis is the process of separating a mixture of molecules in a gel media by application of an electric field. In general, DNA molecules with similar lengths will migrate the same distance. This is known as Sanger sequencing.

First cut DNA at each base: A, C, G, T. Then run gel and read off sequence: ATCGTG ...
**Reading DNA**

Original sequence: ATCGTGTGATAGCGCT

<table>
<thead>
<tr>
<th>G</th>
<th>ATCG</th>
<th>ATCGTG</th>
<th>ATCGTGTCG</th>
<th>ATCGTGTCGATAG</th>
<th>ATCGTGTCGATAGCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>ATCGTG</td>
<td>ATCGTGTCG</td>
<td>ATCGTGTCGATAG</td>
<td>ATCGTGTCGATAGCG</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>ATCGTGTCG</td>
<td>ATCGTGTCGATAG</td>
<td>ATCGTGTCGATAGCG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>ATCGTGTCGATAG</td>
<td>ATCGTGTCGATAGCG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>ATCGTGTCGATAGCG</td>
<td>ATCGTGTCGATAGCG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>ATCGTGTCGATAGCGCT</td>
<td>ATCGTGTCGATAGCGCT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>ATCGTGTCGATAGCGC</td>
<td>ATCGTGTCGATAGCGC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>ATCGTGTCGATAGC</td>
<td>ATCGTGTCGATAGC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>ATCGTGTCGATAG</td>
<td>ATCGTGTCGATAG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>ATCGTGTCG</td>
<td>ATCGTGTCG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>ATCGTG</td>
<td>ATCGTG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>ATCG</td>
<td>ATCG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>ATCGT</td>
<td>ATCGT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>ATCGTG</td>
<td>ATCGTG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>ATCGTGT</td>
<td>ATCGTGT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>ATCGTGTG</td>
<td>ATCGTGTG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>ATCGTGTGT</td>
<td>ATCGTGTGT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>ATCGTGTGTC</td>
<td>ATCGTGTGTC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>ATCGTGTGTCG</td>
<td>ATCGTGTGTCG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>ATCGTGTGTCGATAG</td>
<td>ATCGTGTGTCGATAG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>ATCGTGTGTCGATAGCG</td>
<td>ATCGTGTGTCGATAGCG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>ATCGTGTGTCGATAGCGC</td>
<td>ATCGTGTGTCGATAGCGC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>ATCGTGTGTCGATAGCGCT</td>
<td>ATCGTGTGTCGATAGCGCT</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Unfortunately, current sequencing technologies can only read 700 nucleotides at a time.

For genomes, we use *shotgun sequencing*, which breaks a chromosome into overlapping short sequences which must then be reassembled.

It's kind of like putting together a jigsaw puzzle with millions of pieces (a lot of which are “blue sky”).
Sequence Assembly

fragments

fragment assembly

target

original

contig

gap
A simple model of DNA assembly is the *Shortest Supersequence Problem*: given a set of sequences, find the shortest sequence $S$ such that each of original sequences appears as subsequence of $S$.

Look for overlap between *prefix* of one sequence and *suffix* of another:
Sequence Assembly

Sketch of algorithm (procedure for assembling fragments):

- Create an *overlap graph* in which every node represents a fragment and edges indicate overlap.
- Determine which overlaps will be used in the final assembly: find an optimal collection of paths in overlap graph.

\[
\begin{align*}
W & = \text{AGTATGGCAATC} \\
Z & = \text{AATCGATG} \\
U & = \text{ATGCAAACCT} \\
X & = \text{CCTTTTGG} \\
Y & = \text{TTGGCAATCA} \\
S & = \text{AATCAGG}
\end{align*}
\]
Sequence Assembly

- Look for paths of maximum weight: use *greedy* algorithm to select edge with highest weight at every step.
- Selected edge must connect nodes with in- and out-degrees $\leq 1$.
- May end up with set of paths: each corresponds to a contig.

\[
\begin{align*}
W &\rightarrow Y \rightarrow S \\
Z &\rightarrow U \rightarrow X
\end{align*}
\]

AGTATTGGCAATC
TTGGCAATCA
AATCAGG

AGTATTGGCAATCAGG

AATCGATG
ATGCAAACCT
CCTTTTGG

AATCGATGCAAAACCTTTTGG
DNA Microarrays

- Allows simultaneous measurement of the level of transcription for every gene in a genome (gene expression).
- Differential expression, changes over time.
- Single microarray can test ~10k genes.
- Data obtained faster than can be processed.
- Want to find genes that behave similarly.
- A pattern discovery problem.

green = repressed
red = induced
Visualizing Microarray Data

Clustering Microarray Data

*K-means clustering* is one way to organize this data:

- Given set of $n$ data points and an integer $k$.
- We want to find set of $k$ center points that minimizes mean-squared distance from each data point to its nearest cluster center.

Sketch of algorithm:

- Choose $k$ initial center points randomly and cluster data.
- Calculate new centers for each cluster using points in cluster.
- Re-cluster all data using new center points.
- Repeat second two steps until no data points are moved from one cluster to another or some other convergence criterion is met.
Clustering Microarray Data

- Pick $k = 2$ centers at random.
- Cluster data around these center points.
- Re-calculate centers based on current clusters.

From “Data Analysis Tools for DNA Microarrays” by Sorin Draghici.
Clustering Microarray Data

- Re-cluster data around new center points.

- Repeat last two steps until no more data points are moved into a different cluster.

From “Data Analysis Tools for DNA Microarrays” by Sorin Draghici.
Building the “Tree of Life”

Scientists build phylogenetic trees in an attempt to understand evolutionary relationships.

Note: these trees are “best guesses” and certainly contain some errors!
**Why Study Bioinformatics?**

- Still many urgent open problems ⇒ lots of opportunities to make fundamental contributions (and become rich and famous).
- Stretch your creativity and problem-solving skills to the limit.
- Join a cross-disciplinary team – work with interesting people.
- Participate in unlocking the mysteries of life itself.
- Make the world a better place.
My Advice on Colleges

• Gather all the data. Then trust your instincts.
• Whatever your decision, college will probably be the best years of your life – take full advantage of it.
• Beyond the obvious criteria, look for opportunities to become engaged in cutting-edge research, whatever your major.
• Seek out projects that will attract attention and have an impact.
Breaking a Civil War Secret Code

The Civil War letter ...

... encrypted – not yet broken.

The players ...

Major General Fitz-John Porter
author of letter, blamed for Union loss at Second Bull Run, court-martialed in 1863

Anson Stager
inventor of cypher system and later an early leader in U.S. telecommunications

Major General George McClellan
recipient of letter
Breaking a Civil War Secret Code

The news story ...

"Lehigh team works to crack Civil War code"

… finale yet to be written.

The current software tool ...

… promising, but needs work.

Hint: I'm looking for help ...
Better Electronic Voting

E-voting has generated some controversy recently ...

Maryland votes “yes” for paper trail ... while Pennsylvania votes “no.”
E-voting: what's the right answer?

- Take a critical look at all aspects of the problem.
- Examine both security and usability issues.
- Build a prototype of an e-voting system that includes a reliable Voter Verified Paper Audit Trail (VVPAT).
- Some critics claim it can't be done: we disagree.
- Yet another undergraduate project.
- Of fundamental importance because our democracy depends on fair and transparent elections.
Evaluating Biometric Security

- Cryptographic key broken into shares and mixed with random data.
- Features extracted from user's speech or handwriting.
- Only input from true user will select correct shares to yield proper key.
Evaluating Biometric Security

Biometrics may be vulnerable to attacks using generative models.

- Some current systems at risk.
- Results for handwriting show machine can equal performance of skilled human forger:

Use our experience to improve biometrics, increase security.
Last but not least ...

Why attend Lehigh?
Because this is a great place to be a student!

Thank you!