Introduction to Bioinformatics

Dan Lopresti
Computer Science and Engineering
Office Building C 337
dal9@lehigh.edu
In 2017 when I gave this talk ...
Motivation

“Biology easily has 500 years of exciting problems to work on.” Donald Knuth (famous computer scientist)

By developing techniques for analyzing sequence data and structures, we can attempt to understand basis of life.

http://cmgm.stanford.edu/biochem218/
What is bioinformatics? Application of methods from computer science to biology.

Why is it interesting?

- Important problems.
- Massive quantities of data.
- Great need for efficient solutions.
- Success is rewarded.
Our genetic identity 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 a language over a four character alphabet, \{A, C, G, T\}.

NLM / NIH seems to have made a mistake: this should be billions, not millions!

https://www.nlm.nih.gov/about/2017CJ.html
Set of chromosomes that determines an organism is known as its genome.

**Us**

**Mus musculus**

**Zea mays**

**Conclusion:** size does not matter! (But you already knew this. 😊)

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

How did we decipher these relationships?

An algorithm is a precisely-specified series of steps to solve a particular problem of interest.

- Develop model(s) for task at hand.
- Study inherent computational complexity:
  - Can task be phrased as an optimization problem?
  - Can it be solved efficiently? Speed, memory, etc.
  - If we can’t find good algorithm, can we prove task hard?
  - If known to be hard, is there approximation algorithm (works some of the time or comes close to optimal)?
- Conduct experimental evaluations (iterate above steps).
Macromolecules are chains of simpler molecules.

In case of proteins, these basic building blocks are amino acids.

In DNA and RNA, they are nucleotides.
National Center for Biotechnology Information (NCBI), a branch of National Institutes of Health (NIH), maintains GenBank, a worldwide repository of genetic sequence data (all publicly available DNA sequences).

Massive quantities of sequence data ⇒ need for good computational techniques.

Reading DNA

Gel electrophoresis separates mixture of molecules in a gel media by application of an electric field.

In general, molecules with similar lengths will migrate same distance.

Make DNA fragments that end at each base. Then run gel and read off sequence: ATCGTG ...
Reading DNA

Original sequence: ATCGTGTCGATAGCGCT
Sequencing a Genome

Most genomes are enormous (e.g., $10^{10}$ base pairs for human). But current sequencing technology only allows biologists to determine ~$10^3$ base pairs at a time.

Leads to some very interesting problems in bioinformatics!
Genomes can also be determined using a technique known as **shotgun sequencing**.

Computer scientists have played an important role in developing algorithms for assembling such data.

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

http://occawlonline.pearsoned.com/bookbind/pubbooks/bc_mcampbell_genomics_1/medialib/method/shotgun.html
Sequence Assembly

fragments

fragment assembly

target

original

contig

gap

contig
Sequence Assembly

Simple model of DNA assembly is *Shortest Supersequence Problem*: given set of sequences, find shortest sequence $S$ such that each of original sequences is a subsequence of $S$.

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

$$\text{ACCGT} \quad \text{CGTGC} \quad \text{TTAC}$$

$$\text{ACCGT} \quad \text{CGTGC} \quad \text{TTAC} \quad \text{TTACCGGTGC}$$
**Sequence Assembly**

**Sketch of algorithm:**

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

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

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

\[
\begin{align*}
W &\rightarrow Y \rightarrow S \\
\text{AGTATTGGCAATC} &\text{TTGGCAATCA AATCAGG} \\
\text{AGTATTGGCAATCAGG} &\\
Z &\rightarrow U \rightarrow X \\
\text{AATCGATG ATGCAAACCT} &\text{CCTTTTGG} \\
\text{AATCGATGCAAACCTTTTGG}
\end{align*}
\]
Sequence Comparison

What’s the problem? Kind of like google for biologists ...

- Given new DNA or protein sequence, biologist will want to search databases of known sequences for similarities.
- Sequence similarity can provide clues about function and evolutionary relationships.
- Databases such as GenBank are too big for manual search. To search them efficiently, we need an algorithm.

Can’t expect exact matches (i.e., not really like google):

- Genomes aren’t static: mutations, insertions, deletions.
- Human (and machine) error in reading sequencing gels.
Sequence Comparison

Why not just line up sequences and count matches?

\[
\begin{align*}
\text{AGTC} & \quad \text{ATAT} \\
\uparrow & \quad \uparrow & \quad \uparrow & \quad \uparrow & \quad \uparrow & \quad \uparrow & \quad \uparrow & \quad \uparrow & \quad \uparrow \\
\text{ATTCTGTA} & \\
\end{align*}
\]

\[\text{Difference} = 2\]

Doesn’t work well in case of deletions or insertions:

\[
\begin{align*}
\text{AGTC} & \quad \text{ATATA} \\
\uparrow & \quad \uparrow & \quad \uparrow & \quad \uparrow & \quad \uparrow & \quad \uparrow & \quad \uparrow & \quad \uparrow & \quad \uparrow \\
\text{GTC} & \quad \text{TATA} \\
\end{align*}
\]

\[\text{Difference} = 8\]

One missing symbol at start leads to large difference!
Sequence Comparison

Instead, we’ll use technique known as *dynamic programming.*

- Three basic operations: delete a single symbol, insert a single symbol, substitute one symbol for another.
- Goal: given two sequences, find shortest series of operations needed to transform one into other.

```
AGTC
AGCT
AGCTG
```

- Delete T
- Substitute G for A
Sequence Comparison

Elegant optimization algorithm builds table of values, working from shorter prefixes to longer prefixes:

\[
\begin{array}{c|cccccc}
\text{sequence } s & \epsilon & \text{sequence } t \\
\hline
0 & 0 & \text{cost of inserting } t \\
\hline
\end{array}
\]

\[
\text{cost of deleting } s
\]

\[
d [i, j] = \min \begin{cases} 
  d [i-1, j] + 1 \\
  d [i, j-1] + 1 \\
  d [i-1, j-1] + \begin{cases} 
    0 & \text{if } s[i] = t[j] \\
    1 & \text{if } s[i] \neq t[j]
  \end{cases}
\end{cases}
\]
Genome Rearrangements

- 99% of mouse genes have homologues in human genome.
- 96% of mouse genes are in same relative location.
- Mouse genome can be broken up into 300 synteny blocks which, when rearranged, yield human genome.
- Provides a way to think about evolutionary relationships.

Recall what we saw earlier:
Reversal Distance

**Human Chromosome X**

1 2 3 4 5

Cut and reverse

-3 -2 -1 4 5

Cut and reverse

-3 -2 -5 -4 1

Cut and reverse

-3 5 2 -4 1

**Mouse Chromosome X**

Reversal distance is minimum number of steps needed.
Interesting Sidenote

Early work on related problem, sorting by prefix reversals, was done in 1970’s by Christos Papadimitriou, a professor now at UC Berkeley, and one “William H. Gates” ...

Yes, that Bill Gates ...
History of Chromosome X

Hypothesized reversals
Building the “Tree of Life”

Scientists build phylogenetic trees to help understand evolutionary relationships. Reversal distance often used.

Note: trees are “best guesses” and certainly contain errors!
DNA Microarrays

- Allows simultaneous measurement of transcription level for every gene in a genome (gene expression).
- Differential expression, want to find genes that behave similarly over time.
- One microarray can test ~10k genes.
- Data obtained much faster than we can process it!
- Must find ways to uncover patterns.

\[ \text{green} = \text{repressed} \]
\[ \text{red} = \text{induced} \]
Using DNA Microarrays

- Track sample over time to see change in gene expression.
- Track two different samples under same conditions to see difference in gene expressions.

Each cell represents one gene’s expression over time

http://www.bioalgorithms.info/presentations/Ch10_Clustering.ppt
DNA Microarrays

*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 \( k \) points that minimizes mean-squared distance from each data point to nearest center.

**Sketch of algorithm:**

- Choose \( k \) initial center points randomly and cluster data.
- Calculate new centers for clusters using points in cluster.
- Re-cluster all data using new center points.
- Repeat second two steps until no data points change clusters, 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 data points change clusters.

From “Data Analysis Tools for DNA Microarrays” by Sorin Draghici.
Example of Hierarchical Clustering

Different genes that express similarly

Why Study Bioinformatics?

- Many unanswered questions ⇒ opportunities to make fundamental contributions (+ become rich and famous).
- Stretch your creativity and problem-solving skills.
- Cross-disciplinary teams: work with interesting people.
- Participate in unlocking the mysteries of life itself.
- Make the world a better place.
Intro to Bioinformatics

CSE 308 / BioE 308 covers:
- Intro to molecular biology & algorithms,
- Genetic sequence comparison & alignment,
- Sequencing & assembly of DNA,
- DNA microarrays,
- Gene regulatory networks,
- Genome annotation,
- Transcription factor binding site prediction,
- Standard formats and sources for genomic data, etc.

Questions: chen@cse.lehigh.edu

CSE 308 is not a programming course! It’s for BioS, BioE, CSE, and Math students.
CSE 307 / BioE 307 covers:

- Geometric modeling for proteins,
- Structure alignment & protein folding,
- Protein surfaces, cavities, electrostatics,
- Protein-protein and protein-DNA
- Interfaces and interactions,
- Protein structure prediction, simulation, docking,
- Structural bioinformatics in pharmaceutical discovery,
- Function annotation, active site prediction, etc.

Questions: chen@cse.lehigh.edu
Thank you!

Beat Lafayette!