Large-Scale Physical Genome Mapping

Anthony Bonner

Dept. of Computer Science
University of Toronto

Joint work with the Whitehead Institute/MIT Center for Genome Research
Outline

- Introduction to Molecular Biology
- Genome Mapping
- Noisy Data
- Integrating many forms of data.
- Our approach
  - Condense mapping data into a graph.
  - Graph algorithms
  - Graph visualization
- Summary
Genetics
THE FUTURE IS NOW

New breakthroughs can cure diseases and save lives, but how much should nature be engineered?
The Human Genome at Four Levels of Detail.

Primer on Molecular Genetics

Appendix A
2.5 The arrangement and association of nucleotides in the DNA double helix.
DNA is a complex biological polymer:

which may be represented by a string of symbols from the set

\{ A G C T \}

each of which represents a chemical group (nucleotide)
A Portion of a Human Hemoglobin Gene:

<table>
<thead>
<tr>
<th></th>
<th>GTAAGCAGGT</th>
<th>TGTGGTTGAG</th>
<th>AAAGGAAAGT</th>
<th>GTGAACAGGG</th>
<th>GACCAGGAGG</th>
</tr>
</thead>
<tbody>
<tr>
<td>51</td>
<td>GAGAGGGGGG</td>
<td>GGATGCGGC</td>
<td>TGCCAGTTTT</td>
<td>GTTTTGAGGG</td>
<td>ACTTTCTCTT</td>
</tr>
<tr>
<td>101</td>
<td>CTGACCCAAGG</td>
<td>TAGGAGGATA</td>
<td>CTAACCTCTT</td>
<td>CCCAAAAGTC</td>
<td>CATCACTGGA</td>
</tr>
<tr>
<td>151</td>
<td>GACATAGTAA</td>
<td>GGG GTAAGAA</td>
<td>AGTGTGTCGG</td>
<td>GCAACTGAT</td>
<td>AAGGATTCCC</td>
</tr>
<tr>
<td>201</td>
<td>TGCCACCTAGG</td>
<td>GAGACCAACG</td>
<td>CCAAGCGCCA</td>
<td>GAATCTCAGG</td>
<td>GGGCCTAACA</td>
</tr>
<tr>
<td>251</td>
<td>AGTTTTACTG</td>
<td>GTGAGAGCAA</td>
<td>GCACAAAACA</td>
<td>GCAAATGAGT</td>
<td>AACTGCTCCA</td>
</tr>
<tr>
<td>301</td>
<td>AGGGGCGTGC</td>
<td>CACCTGCTTC</td>
<td>GAGGCAACGG</td>
<td>CTTTGAGGAG</td>
<td>CATATAAGTG</td>
</tr>
<tr>
<td>351</td>
<td>CTACTTTGCTG</td>
<td>CAGGTTCAAG</td>
<td>AACTTCTCTA</td>
<td>TTCTGACAGA</td>
<td>CTCAGGAAAG</td>
</tr>
<tr>
<td>401</td>
<td>AACCATTGTTG</td>
<td>CTCTCTGGGG</td>
<td>AAGCACAAAG</td>
<td>CAATATCAAG</td>
<td>GTGCTCTGGG</td>
</tr>
<tr>
<td>451</td>
<td>GGAAGATTGG</td>
<td>TGGCATGCTG</td>
<td>GCTGAATATG</td>
<td>GAGCTGAAAGC</td>
<td>CCTGGAAGGG</td>
</tr>
<tr>
<td>501</td>
<td>TGAGAACAGG</td>
<td>ACTTGTATCG</td>
<td>GTAAGATGCA</td>
<td>CAGATCTCCA</td>
<td>TATGGAACCT</td>
</tr>
<tr>
<td>551</td>
<td>GCACTCGCTC</td>
<td>AGTTGCGAGC</td>
<td>TTCTAATCTA</td>
<td>GCTTTTCTGT</td>
<td>GACCTCAACT</td>
</tr>
<tr>
<td>601</td>
<td>TCCTCTTCTT</td>
<td>CTTTTCTCCA</td>
<td>GAGTGTGTGC</td>
<td>TAGCTTTCCC</td>
<td>ACCAACAAGA</td>
</tr>
<tr>
<td>651</td>
<td>CCTACTTTCC</td>
<td>TCACCTTGTG</td>
<td>GTAAGCCAGG</td>
<td>GCTCTGCCCA</td>
<td>GGTCAGGGGT</td>
</tr>
<tr>
<td>701</td>
<td>CACGCGAAAG</td>
<td>AGGTGCGCGA</td>
<td>TGCCGTGCGC</td>
<td>AGGTCTGCA</td>
<td>GCCACCTCGA</td>
</tr>
<tr>
<td>751</td>
<td>TGACCTGCCG</td>
<td>GGTGCTTGTG</td>
<td>CTGCTGTAGG</td>
<td>CGACCTGCA</td>
<td>GCCCAACAGC</td>
</tr>
<tr>
<td>801</td>
<td>TGCGGTGAGA</td>
<td>TCCCCTCACG</td>
<td>TTCAAGGTAT</td>
<td>GGCTGTGGAAG</td>
<td>C'TGGACAGGG</td>
</tr>
<tr>
<td>851</td>
<td>GCATCTGAGG</td>
<td>CCCCCTAGGA</td>
<td>GGCTTTGGGG</td>
<td>GTCTCTGTCG</td>
<td>CAAAGGAGAG</td>
</tr>
<tr>
<td>901</td>
<td>GAACATAAGTG</td>
<td>GTCCAGAAAG</td>
<td>GGGGAAGCGA</td>
<td>GCCATCAGGG</td>
<td>TGTCACCTTT</td>
</tr>
<tr>
<td>951</td>
<td>GTCTCCGCGG</td>
<td>CCTGCGAGCC</td>
<td>ACTGCCTCTG</td>
<td>GTGACCTTAG</td>
<td>GCTAGCCACC</td>
</tr>
<tr>
<td>1001</td>
<td>ACCCTGCGGA</td>
<td>TTTCAACCCG</td>
<td>GCAGTGATAT</td>
<td>CCTCCTTGGG</td>
<td>CAAAT'CCTT</td>
</tr>
<tr>
<td>1051</td>
<td>GCCTTCGTGA</td>
<td>GCACAGGCTT</td>
<td>GACCTCAAGG</td>
<td>TACCGTAAGG</td>
<td>CTGCTCCTCG</td>
</tr>
<tr>
<td>1101</td>
<td>CGGGGCTTGC</td>
<td>CTTCTGCGGA</td>
<td>TGCCCTCTCT</td>
<td>CTTCTGTCG</td>
<td>CACCTGTACC</td>
</tr>
<tr>
<td>1151</td>
<td>TCTCTGGTCTT</td>
<td>GAAATAAGAC</td>
<td>CTGAGTGAAG</td>
<td>AGAAGCCTCG</td>
<td>ATGCTGTTG</td>
</tr>
<tr>
<td>1201</td>
<td>CTCTGCGTGCT</td>
<td>GCAAAGGTTG</td>
<td>CATGTTTGTG</td>
<td>GTGGGAGATG</td>
<td>CGCAGCTCAG</td>
</tr>
<tr>
<td>1251</td>
<td>TGGGATTGGG</td>
<td>GCAAGTAAGA</td>
<td>CAAGGTTGAC</td>
<td>AGCAAAAGAC</td>
<td>ATAATTTGAG</td>
</tr>
<tr>
<td>1301</td>
<td>GCCTACACAC</td>
<td>ACACACATAT</td>
<td>GTCTTTGTGG</td>
<td>TCTGGAGCAG</td>
<td>AGTCCTCCTCC</td>
</tr>
<tr>
<td>1351</td>
<td>AAGGCCCTCCA</td>
<td>CTGACAGCCA</td>
<td>TGTGCTCTCT</td>
<td>CCTCGAGCCA</td>
<td>AAGAAGCGCA</td>
</tr>
</tbody>
</table>

Humans cells are estimated to contain $3 \times 10^9$ characters of independent information.
GENOME SIZES

- E. Coli
  - 4.5 Million base pairs.
  - 4000 genes.

- Fungi, Yeasts:
  - 10-100 Million bp.

- Human:
  - 3 Billion bp.
  - 100,000 genes.
The Human Genome Project

- An International Effort to map & sequence the human genome.

- Made possible by genetic breakthroughs in 1970's & 80's.

- Biologists are now overwhelmed with genome data.

⇒ Data Management Problems
Genome Research

- Gene Hunting
- Genome Mapping
- Genome Sequencing

Related Activities:
- Evolution
- Protein Structure
- Cellular Metabolism
- Gene Regulation
Genome Maps

- Gives the location of important or easily identifiable sites on each chromosome.

Two Kinds

- Linkage Maps:
  - Course grained
  - Lots of statistics

- Physical Maps:
  - Fine grained
  - Lots of algorithms
Figure 7-22
The complete genetic map of E. coli. [Courtesy of Barbara J. Bachman, Microbiol. Rev. 47 (1983):160]

Genetic Linkage Map
Physical Gene Map

Mouse chr 2

Peter Green
Building a Physical Genome Map

- Smash (many copies of) a genome into thousands of fragments (clones).

- Experimentally determine which pairs of clones overlap.

- Computationally, use the overlap data to assemble a map ("jigsaw puzzle").

- Main problems:
  - Noisy data
  - Anomalous data
  - Many forms of data
  - Subtle interrelationships

  } data cleansing

  } data integration
Overlap

There are many ways to determine whether two clones overlap.

Some common methods:

- Finger prints
- STS content
- ALU PCR
Finger Prints

- Use restriction enzymes to cut a clone wherever a given sequence occurs.

- E.g., cut a clone at every occurrence of atcgat and gatc (complete digestion).

- Measure the lengths of the resulting fragments.

- The set of lengths is called a fingerprint.

Finger print = \{4, 6, 9, 11, 15, 19, 25, 29, 32, 37\}
Fingerprint Overlap

If the Fingerprints of two clones have many lengths in common, then the clones probably overlap.

\[ \text{Fingerprint 1} = \{4, 5, 6, 7, 8, 10, 11, 12, 15, 19, 21\} \]
\[ \text{Fingerprint 2} = \{4, 5, 6, 7, 8, 9, 12, 15, 19, 21, 23, 25\} \]
\[ \text{Fingerprint 1} \cap \text{Fingerprint 2} = \{4, 5, 6, 7, 8, 12, 15, 19, 21\} \]
An STS (Sequence Tag Site) is a fragment of genomic DNA several hundred base pairs long.

With high probability, an STS will appear only once in a random sequence of 3 Billion base pairs.

With high probability, two clones overlap if they hybridize with (hit) the same STS.
Problems: Biological Anomalies

- A genome is not a random sequence.
- Many regions (subsequences) repeat.
- IF an STS hybridizes with a repeat region, then clones that are far apart may appear to overlap.

Clone 1

\[\text{repeat region} \]

Clone 2

\[\text{repeat region} \]
Problems: Experimental Error

**False Negatives:**
- An STS should hybridize with a clone, but fails to
- So, an overlap goes undetected.

**False Positives:**
- An STS should *not* hybridize with a clone, but does.
- So, an overlap may erroneously be inferred.
Problems: Chimerism

- What appears to be a single clone, is actually two (or more) clones.

- One possible cause: When the genome is smashed into clones some clones may fuse.

Genome:

- Chimeric clones make discontiguous maps look contiguous:

Region 1

Region 2
Chimerism

- Another possible cause: Contamination

- A test tube containing (many copies of) clone 1, may also contain (some copies of) clone 2.

- Problem: Clones that overlap with either clone 1 or clone 2 will react positively with this test tube.
Summary of Problems

Biological Anomalies

- Repeat regions: The same sequence may be found in many places on a genome.

Overlap Errors

- False positives: Clones appear to overlap when actually they do not.
- False Negatives: Clones appear not to overlap when actually they do.

Identity Errors

- Chimerism: What appears to be one clone is actually two (or more) clones.
Assembling Large Physical Maps

- Given a collection of overlap data, how do we construct a map?

  eg.  
  
  
  
  
  

- Example: STS data
  
  - Ideal case (no noise)
  
  - False negatives
  
  - False positives

- Integrating many forms of overlap data
  
  - Detecting & removing errors & anomalies
STS Data
(noise-Free case)

Probe P "hits" DNA segment 5.

data: hits (p1, s1)  hits (p3, s2)
hits (p2, s1)  hits (p2, s3)
hits (p2, s2)  hits (p4, s3)

map: p1  p2  p3  p4
     s1
     s2
     s3
No map is consistent with this data.
Tiny portion of a physical genome map

Pete Groot
Mouse CH. 2
Unfortunately no linear order of the problem is even approximately correct.
The data can be summarized as a graph.

Observation
Integrated Physical Genome Maps

- Integrating many forms of physical data into a single map.

- Problems: Noise & Complexity.

- Our Approach
  - **Clusters**: Overlap, Linkage, Ordering
  - **Graphs**: Algorithms & Visualization
  - Examples

- Summary
Assembling Integrated Maps: Problems

* Complex Data:
  - Many kinds of data,
  - Data from many labs,
  - Uneven data quality,
  - Noisy data: errors, ambiguities, contradictions, anomalies,
  - Subtle relationships between various forms of data and noise.

* Large and increasing volume of data.

* Algorithms are inflexible and limited to a few forms of data and noise.

* Requires much human intervention and biological expertise.

* Maps are full of errors.
Our Approach

- Abstract the genome data as a graph.

- Ideally, the graphs are long & thin, approximately "linear."

- Each node represents a point on the genome.

- An edge means that two nodes are "close" together on the genome.
Noise in the data distorts the linear structure of the graph:

\[ \text{e.g.} \]
Real Genome Graphs

- 25,000 nodes
- 1,000,000 edges

Structure: Like a plate of spaghetti.
Research Problems

(1) Transforming genome-mapping data into a graph.

(2) Identifying contiguous paths (contigs) within the graph.

(3) Generating a genome map from these paths.

Our approach uses numerous graph algorithms and graph visualizations.
Graph Generation (Data Integration)

Many forms of physical mapping data determine whether two clones overlap.

**Definition:** A **cluster** is a maximal set of mutually overlapping clones.

**Eg.**

```
  +---+---+---+---+
  |   |   |   |   |
  +---+---+---+---+
     |   |   |   |
     +---+---+---+
          |   |
          +---+
            |
            +
```

Common Genomic Region
Observation 1

Clusters Filter out False overlaps.

e.g.,

Overlap ① is corroborated by overlaps ② and ③.

Larger clusters have more corroborati

e.g.,

Overlap ① is corroborated by overlaps ② and ③, and by overlaps ④ and ⑤.
Observation 2:

Average cluster size

= Average depth of clone coverage.

eg.

cluster 1

cluster 2

cluster 3
Cluster Proximity

Single Linkage:

Two clusters are close if they have a clone in common.

eg.

Double Linkage:

Two clusters are close if they have two clones in common.

Many kinds of linkage are possible.
Cluster Graphs

- Each node is a cluster.
- An edge between two nodes means the two clusters are close.

\[
\text{clusters}
\]

\[
C_1 \quad C_2 \quad C_3 \quad C_4 \quad C_5
\]

\[
\text{single-linkage cluster graph.}
\]
Map-Assembly Phases

(1) Overlap (cluster formation)

(2) Linkage (graph manipulation)

(3) Ordering (cluster ordering)
Example 1

- Using all STS content data from
  - Whitehead/MIT
  - Ceph/Genethon

with the double-linkage strategy, the proximity graph has
  - 1,463 nodes
  - 11,017 edges.

- Overall graph structure:
  - Many connections between STS probes on different chromosomes.
  - A large graph instead of 23 smaller ones.
  - A big mess.
Example 1

1,467 nodes    11,017 edges
Graph Abstraction (Simplification)

Although the graph is messy, it has a simple "piecewise linear" structure.

Next Step: Extract and display this structure.
Graph Abstraction

Idea: Coalesce groups of neighboring nodes into single nodes.

eg. Blob

---
Abstracted STS graph of the human genome.

(301 nodes, 302 edges)
Example 2

Using all human mapping data from
- Whitehead/MIT
- Ceph/Genethon

including
- sts content data,
- Fingerprint data,
- ALU-PCR data,

the cluster graph has
- 450,342 edges,
- 27,379 nodes.

Most edges (432,888) are concentrated in a single connected component.
Abstracted cluster graph of human genome.
(4621 nodes, 5002 edges)
Abstracted cluster graph
of Human Chromosome 7
(largest component)
Partially trimmed graph
Fully trimmed graph
Cluster graph

Human Chromosome
(between abstraction)
Graph of Chromosome 7 with genetically-mapped positions
Graph of Chromosome 7 with radiation-hybrid positions
Contig of Human Chromosome 7

Abstracted Cluster Graph

Detailed Cluster Graph
Summary

Cluster Graphs:

- Many forms of physical-mapping data determine whether two clones overlap.
- A cluster is a maximal set of mutually-overlapping clones.
- A cluster represents a point on the genome.
- Clusters filter out many false overlaps.
- Clusters reduce map-assembly to three phases:
  overlap  linkage  ordering
Errors and Anomalies

- Given ideal data, a cluster graph would be "nearly linear," ie, long & thin (in fact, an interval graph).

  eg.

- However, because of experimental error (mainly false positives), the graphs are "piecewise linear."

  eg.
- This piecewise linear structure can be automatically extracted and displayed.

- Using high-level mapping data (e.g., chr. assignments, genetic maps, rh maps), long linear subgraphs can be extracted.

- These subgraphs represent genomic contigs.
Map Assembly

- These "nearly linear" subgraphs are the input to an algorithm that generates a physical genome map, i.e., an ordering of clones & STS's.

  eg.  

  Cluster graphs help to clean up data and remove anomalies.
Our Publications

Available on the Web at

www.cs.toronto.edu/~bonner

under three categories:

- Genome Mapping
- Sequence Databases
- Laboratory Workflow