Large-Scale Physical Genome Mapping

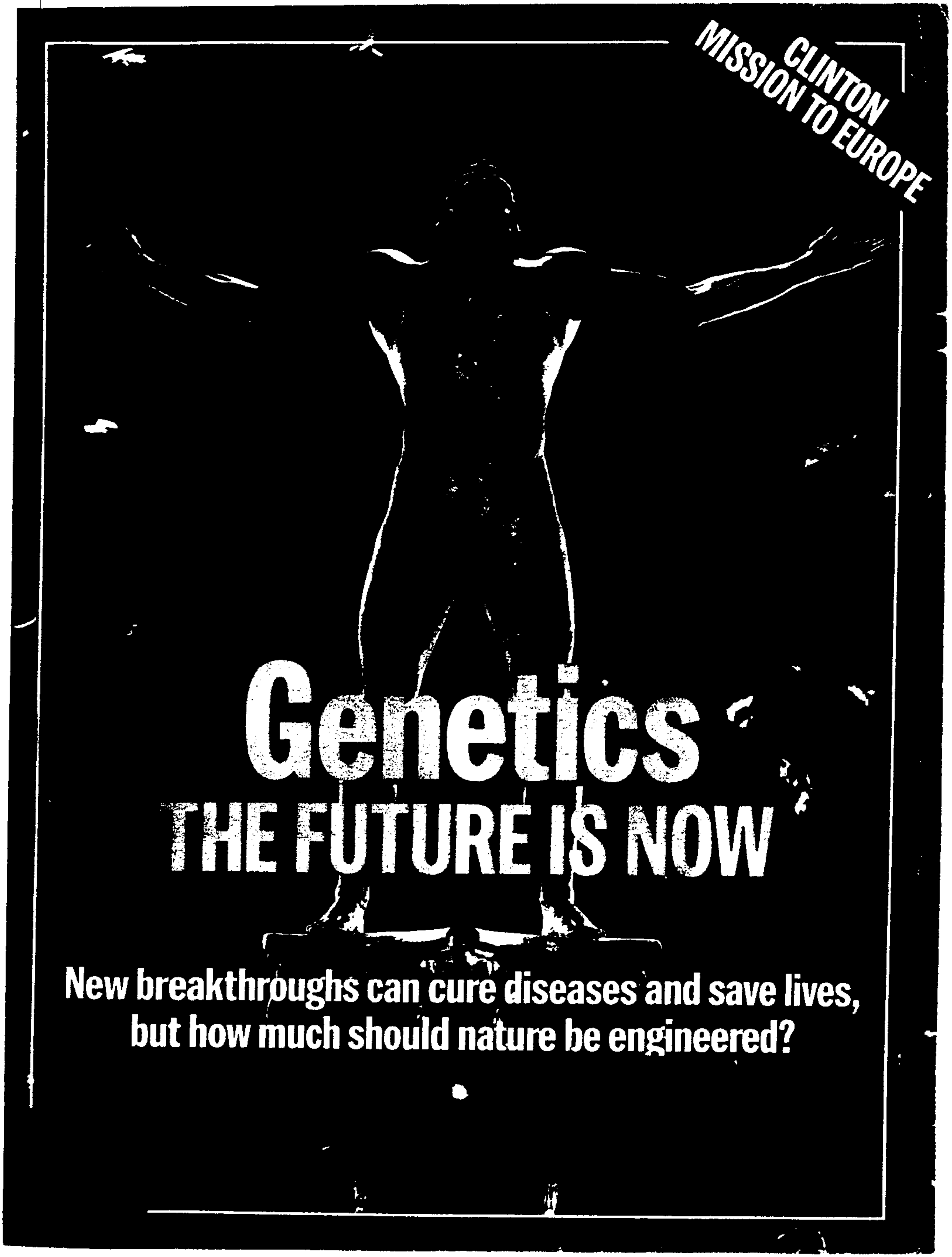
Anthony Bonner

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 groph.
  - Graph algorithms
  - Graph visualization
- Summary



CLINTON MISSION TO EUROPE

Genetics

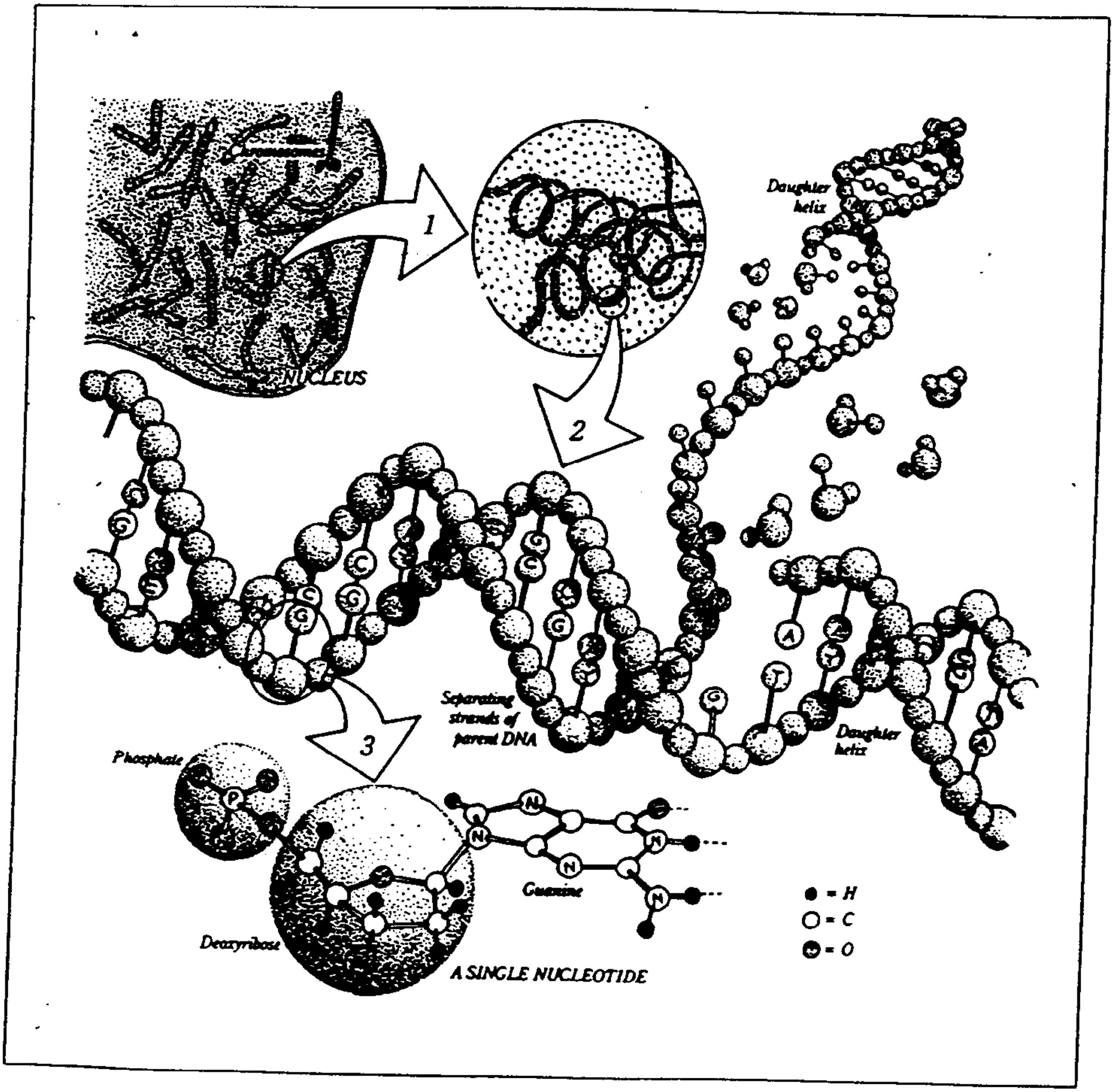
THE FUTURE IS NOW

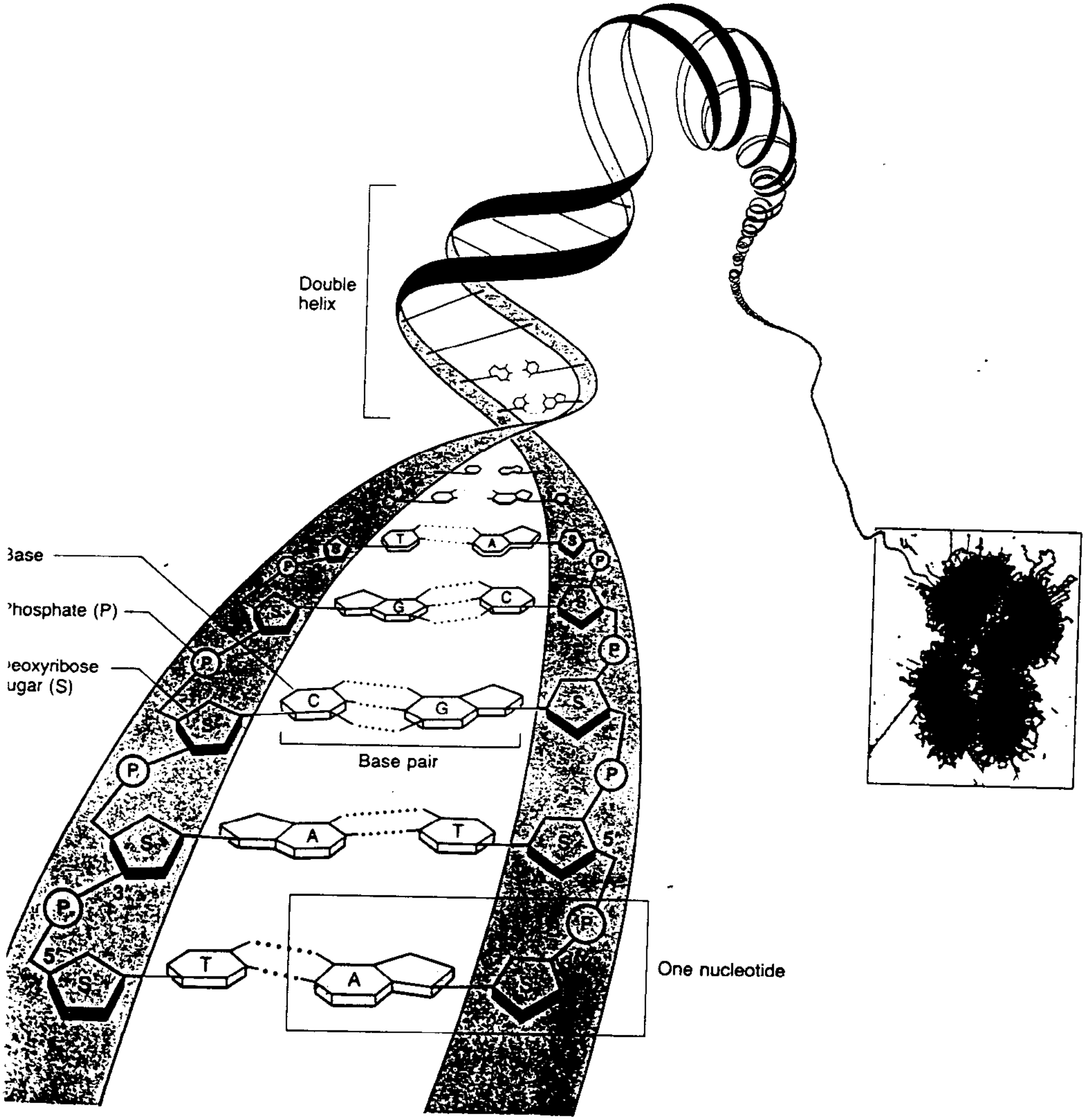
New breakthroughs can cure diseases and save lives, but how much should nature be engineered?

The Human Genome at Four Levels of Detail.

Primer on Molecular Genetics

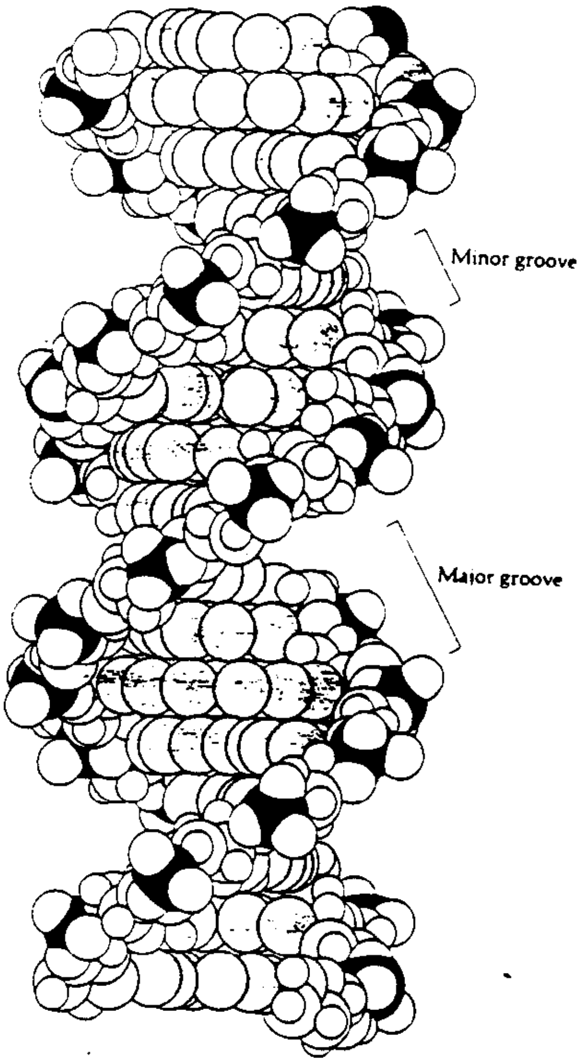
Appendix A





'e 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:

```
   1  GTAAGCAGGT TGTGGTTGAG AAAGGAAAGT GTGAAACAGG GACCCAGAGG
  51  GAGAGGTGGG GGGATGGCGC TGCTCAGTTT GGTTTGAGGG ACTTGCTTCT
 101  CTGACCAAGG TAGGAGGATA CTAACTTCTT CCCAAACTGC CATCACTGGA
 151  GACATAGTAA GGGGTAAGAA AGTGTGTCCG GGCAACTGAT AAGGATTCCC
 201  TGCACCTAGG GGAAGCACAA CCCAGCCCCA GAATCTCAGG GGCCCTAACA
 251  AGTTTTACTG GGTAGAGCAA GCACAAACCA GCCAATGAGT AACTGCTCCA
 301  AGGGCGTGTC CACCCTGCCT GGAGGACAGC CCTTGGAGGG CATATAAGTG
 351  CTACTTGCTG CAGGTCCAAG ACACTTCTGA TTCTGACAGA CTCAGGAAGA
 401  AACCATGGTG CTCTCTGGGG AAGACAAAAG CAACATCAAG GCTGCCTGGG
 451  GGAAGATTGG TGGCCATGGT GCTGAATATG GAGCTGAAGC CCTGGAAAGG
 501  TGAGAACAGG ACCTTGATCT GTAAGGATCA CAGGATCCAA TATGGACCTG
 551  GCACTCGCTC AGTGGGCAGC TTCTAACTAT GCTTTTCTGT GACCTCAACT
 601  TCTCTTCTCT CCTTCTCCCA GGATGTTTGC TAGCTTCCCC ACCACCAAGA
 651  CCTACTTTCC TCACTTTGAT GTAAGCCACG GCTCTGCCCA GGTCAAGGGT
 701  CACGGCAAGA AGGTCGCCGA TGCGCTGGCC AGTGCTGCAG GCCACCTCGA
 751  TGACCTGCCC GGTGCCTTGT CTGCTCTGAG CGACCTGCAT GCCCACAAGC
 801  TGCGTGTGGA TCCCGTCAAC TTCAAGGTAT GCGCTGGGAC CTGGCAGGCG
 851  GCATCTGGGA CCCCTAGGAA GGGCTTGGGG GTCCTCGTGC CCAAGGCAGG
 901  GAACATAGTG GTCCCAGGAA GGGGAGCAGA GGCATCAGGG TGTCCACTTT
 951  GTCTCCGCAG CTCCTGAGCC ACTGCCTGCT GGTGACCTTG GCTAGCCACC
1001  ACCCTGCCGA TTTCACCCCC GCGGTACATG CCTCTCTGGA CAAATTCCTT
1051  GCCTCTGTGA GCACCGTGCT GACCTCCAAG TACCGTTAAG CTGCCTTCTG
1101  CGGGGCTTGC CTTCTGGCCA TGCCCTTCTT CTCTCCCTTG CACCTGTACC
1151  TCTTGGTCTT TGAATAAAGC CTGAGTAGGA AGAAGCCTGC ATGCCTGGTT
1201  CTCTGCGTCT GCAAAGGTGT CATGTTTAGT GTGGGGATGC CGCAGCTCAT
1251  TTGCCATGGG GCAGTAAAGA CAAGGTTCAG AGCAAAAAGC ATAATTGGAT
1301  GCCTACACAC ACACACATAT GTCTTCTGAG TCTGGGCAGC AGTCCCTCCC
1351  AAGCCCTCCA CTGACAGCCA TGTGTCTTCT CCTCGAGCCA AAGAAGCCAA
```

Humans cells are estimated to contain 3×10^9 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

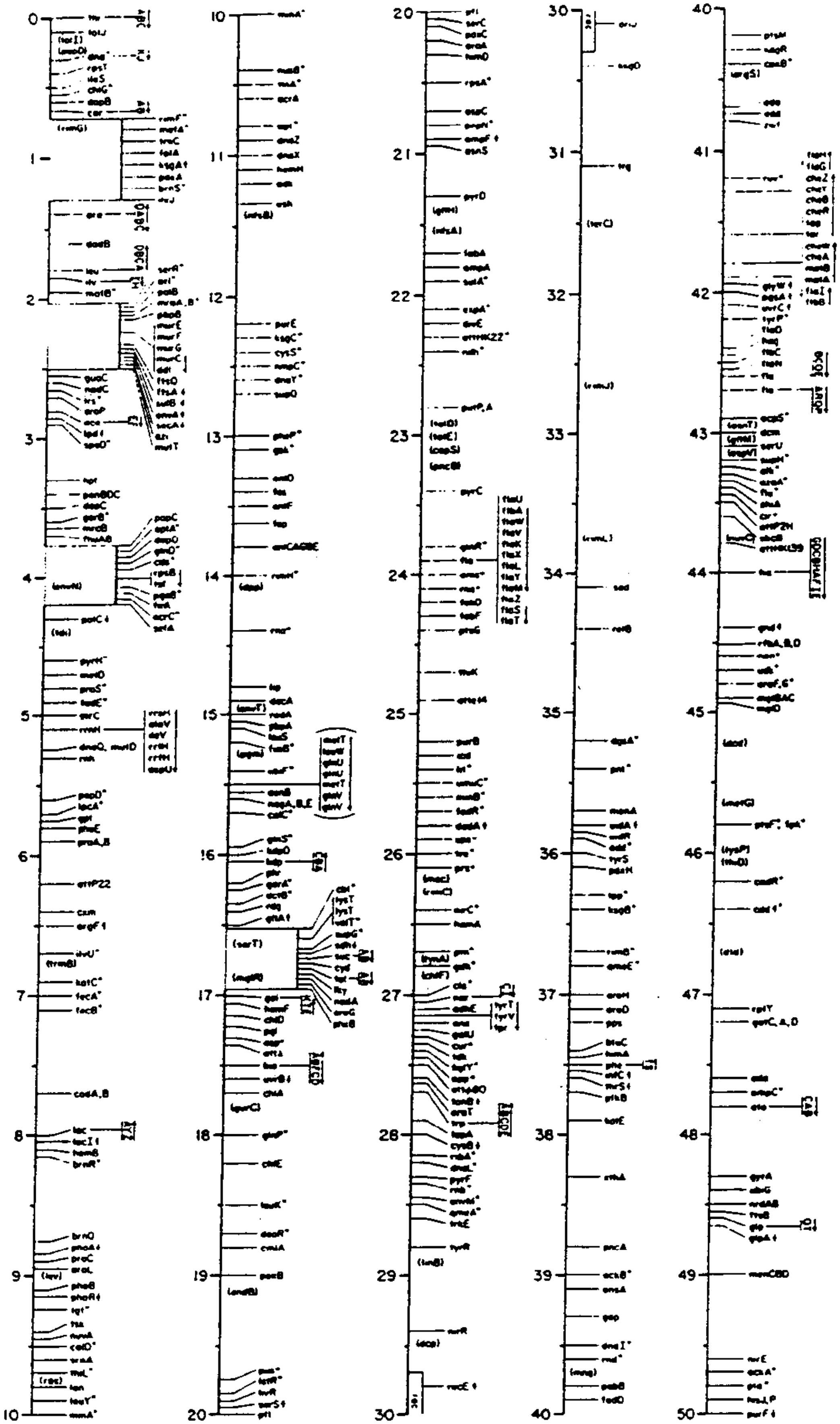
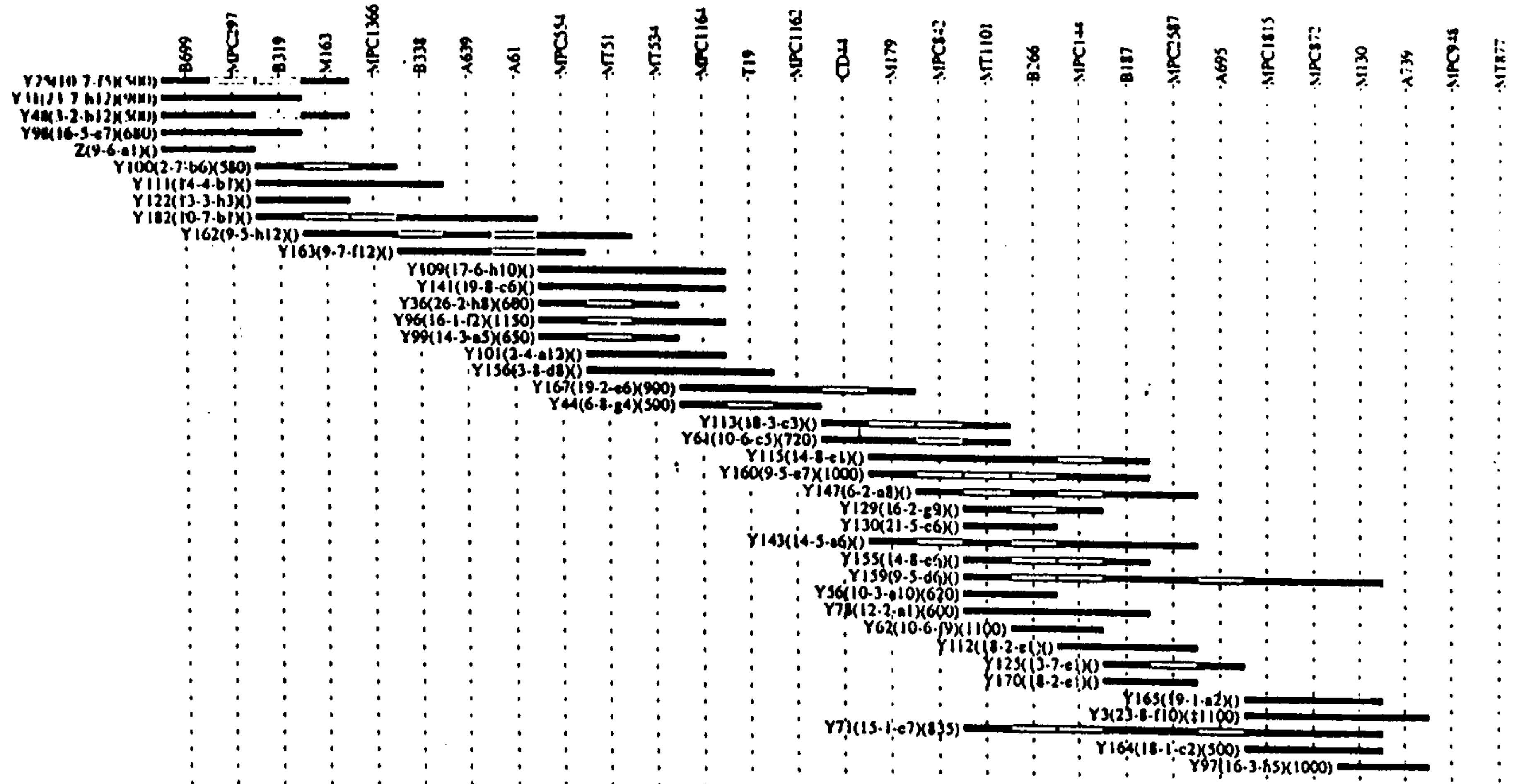


Figure 7-22
The complete genetic map of *E. coli.* [Courtesy of Barbara J. Bachman, *Microbiol. Rev.* 47 (1983):180]

Genetic Linkage Map



MOUSE CHR. 2
PETER GROOT

page 99 - Thu Nov 18 11:13:09 1993

Tiny Portion of a Physical Genome Map

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:

$$\left.\begin{array}{l}\bullet\ \text{Noisy data}\\ \bullet\ \text{Anomalous data}\end{array}\right\}\ \text{data cleansing}$$

$$\left.\begin{array}{l}\bullet\ \text{Many Forms of data}\\ \bullet\ \text{Subtle interelationships}\end{array}\right\}\ \begin{array}{c}\text{data}\\ \text{integration}\end{array}$$

Overlap

There are many ways to determine whether two clones overlap.

Some common methods:

- Finger prints
- STS content
- ALU PCR

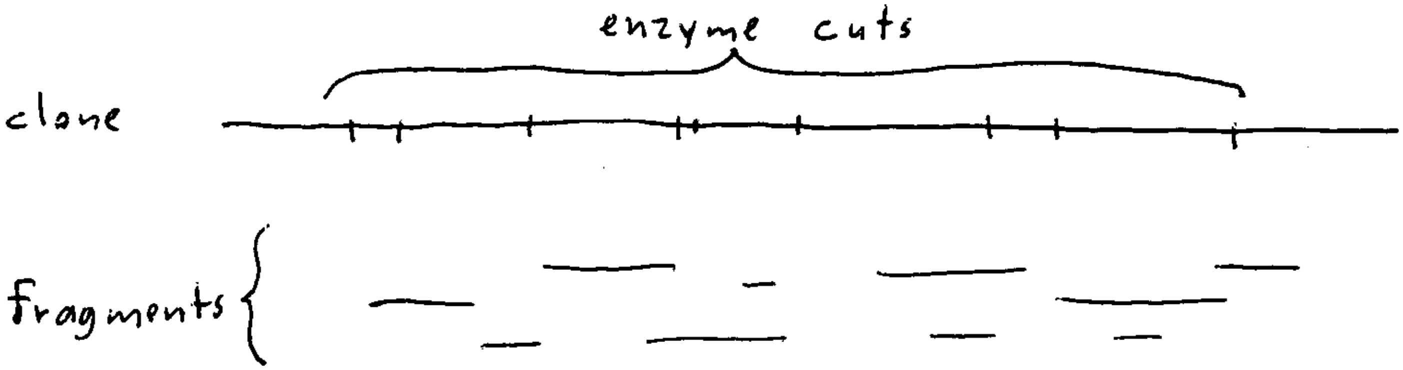
Finger Prints

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

- eg. cut a clone at every occurrence of atcgat and gatc (complete digestion

- Measure the lengths of the resulting fragm

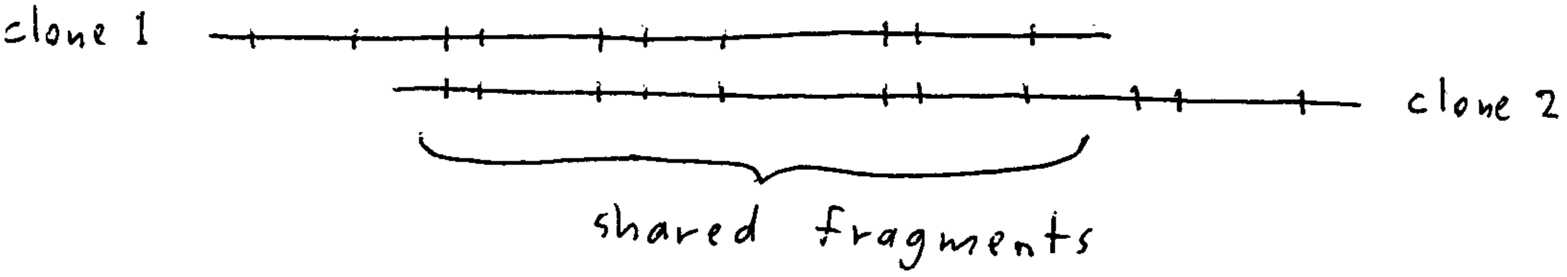
- The set of lengths is called a finger print.



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

Finger Print Overlap

IF the Fingerprints of two clones have many lengths in common, Then the clones probably overlap.



Fingerprint 1 = {4, 5, 6, 7, 8, 10, 11, 12, 15, 19, 21}

Fingerprint 2 = {4, 5, 6, 7, 8, 9, 12, 15, 19, 21, 23, 25}

Fingerprint 1 $\cap$ Fingerprint 2

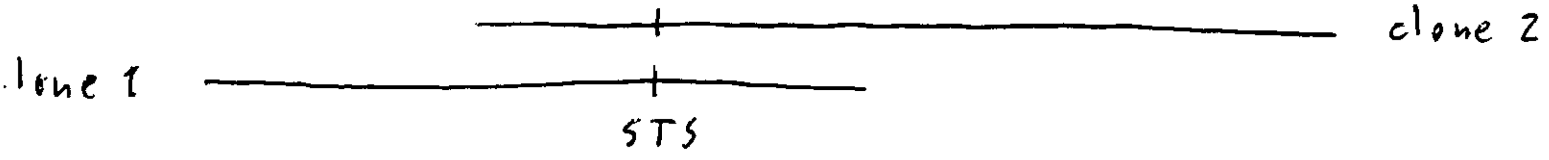
= {4, 5, 6, 7, 8, 12, 15, 19, 21}

STS Content

- 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.

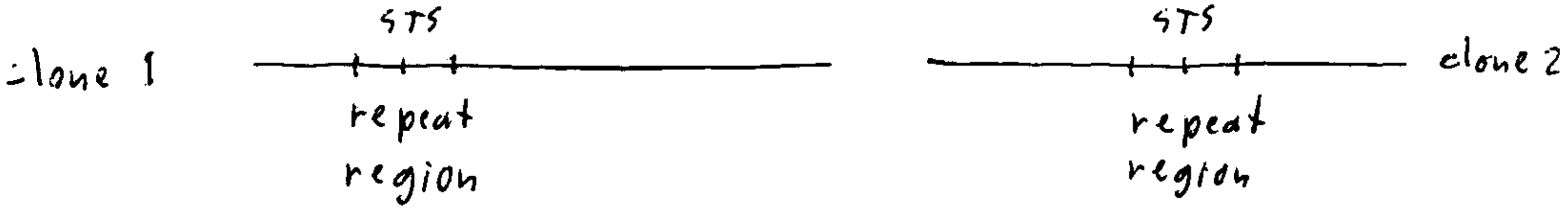


Problems: Biological Anomalies

- A genome is <u>not</u> a <u>random</u> sequence.

- Many regions (subsequences) <u>repeat</u>.

- IF an STS hybridizes with a repeat region, then clones that are far apart may appear to overlap.



Problms: Experimental Error

False Negatives:

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

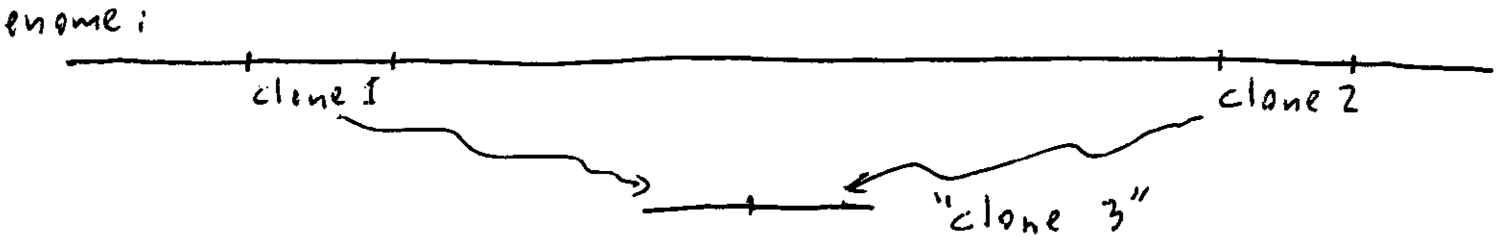
False Positives:

- An STS should <u>not</u> hybridize with a clone, but does.
- So, an overlap may erroneously be inferred.

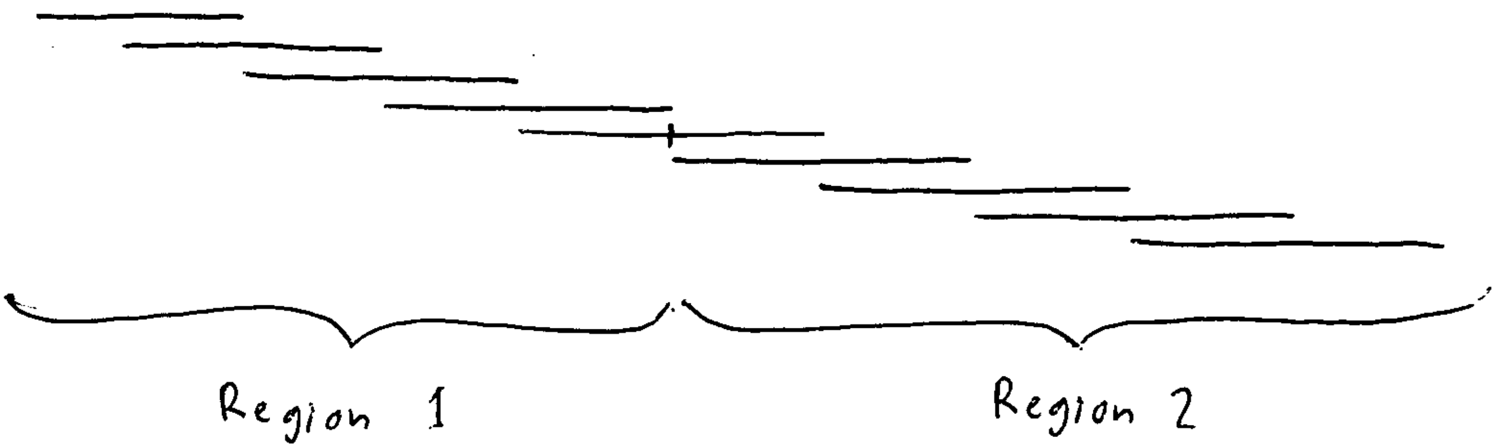
Problems: Chimerism

- 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.



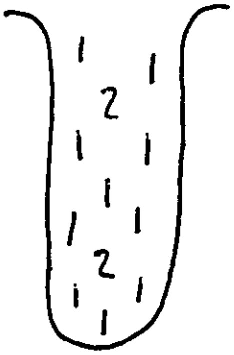
- Chimeric clones make discontiguous maps look contiguous:



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.

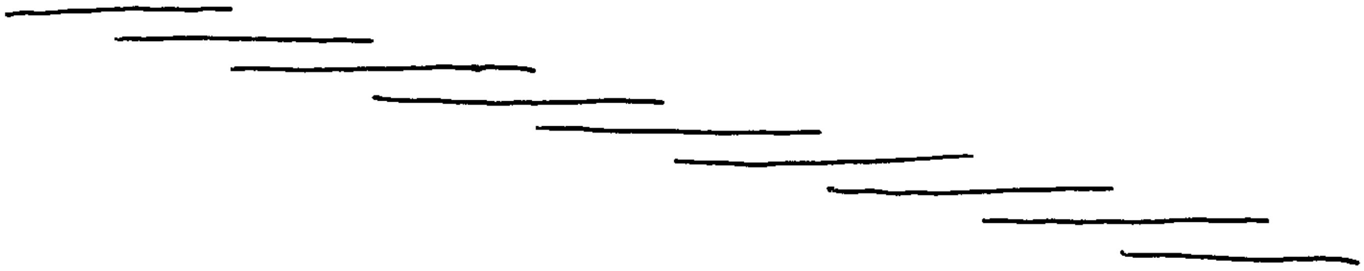
Identity Errors

- 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

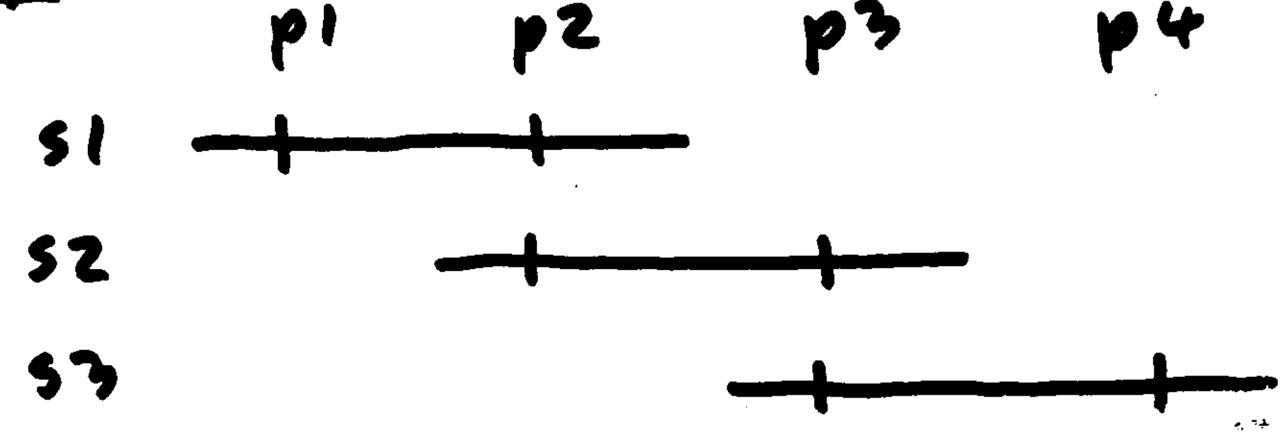
(noise-Free case)

Probe P "hits" DNA segment S.

data:

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

map:



False Negatives

eg.

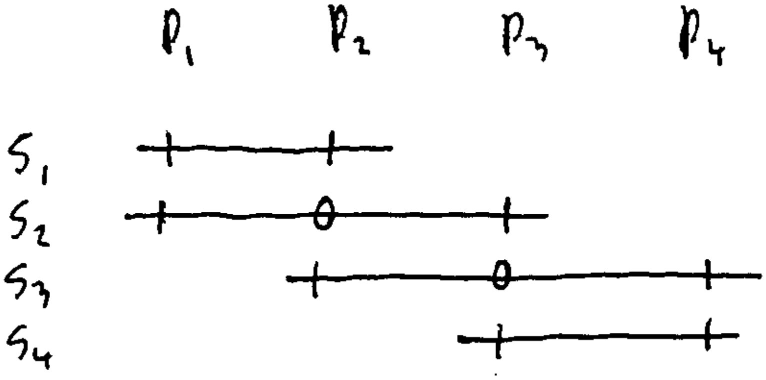
Data:

	P_1	P_2	P_3	P_4
S_1	X	X		
S_2	X		X	
S_3		X		X
S_4			X	X

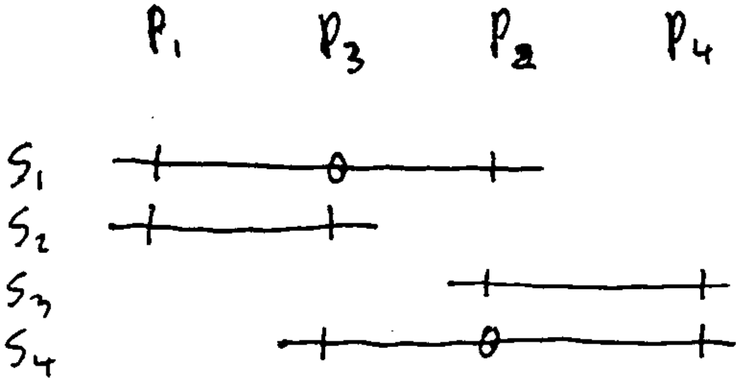
No map is consistent with this data.

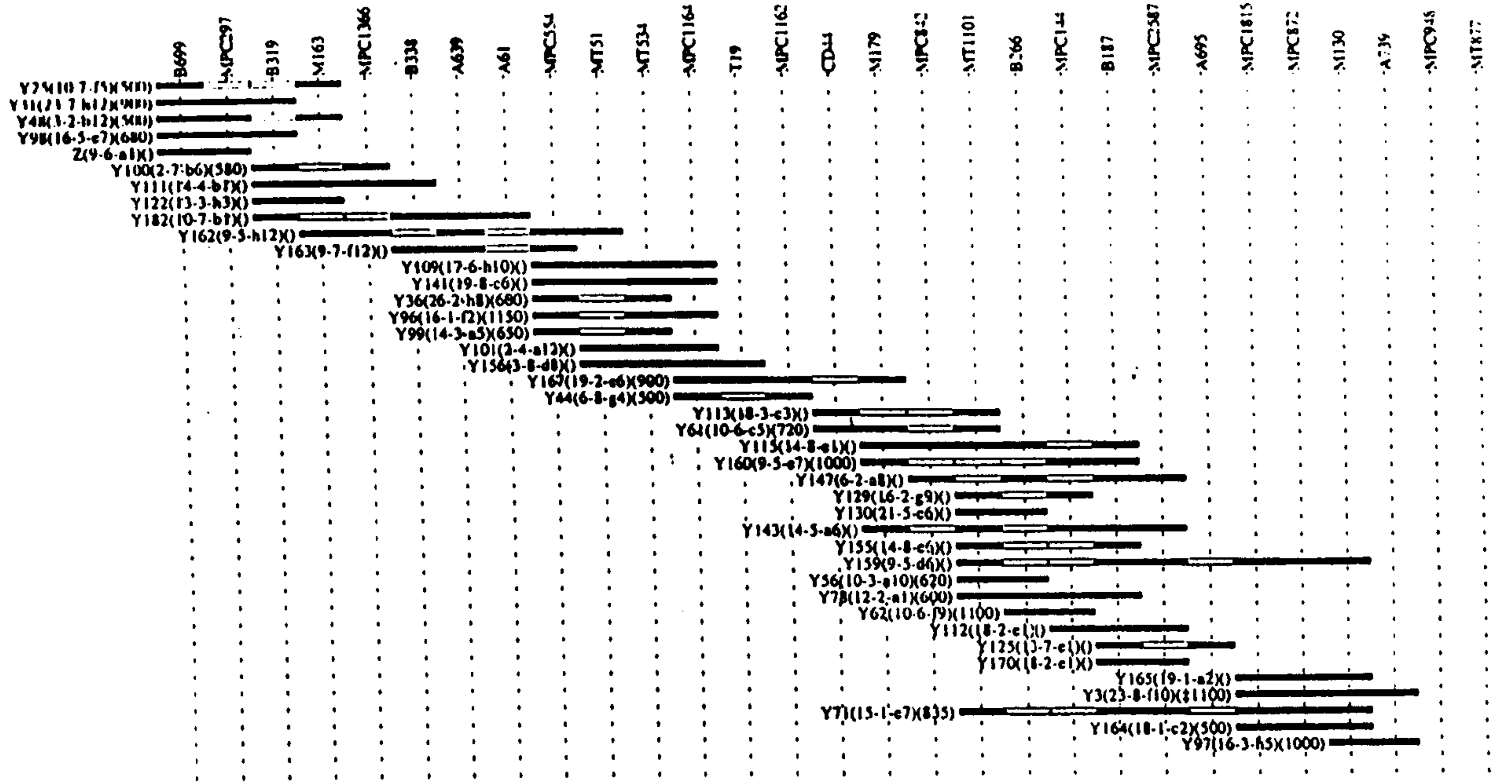
eg.

Map 1:



Map 2:





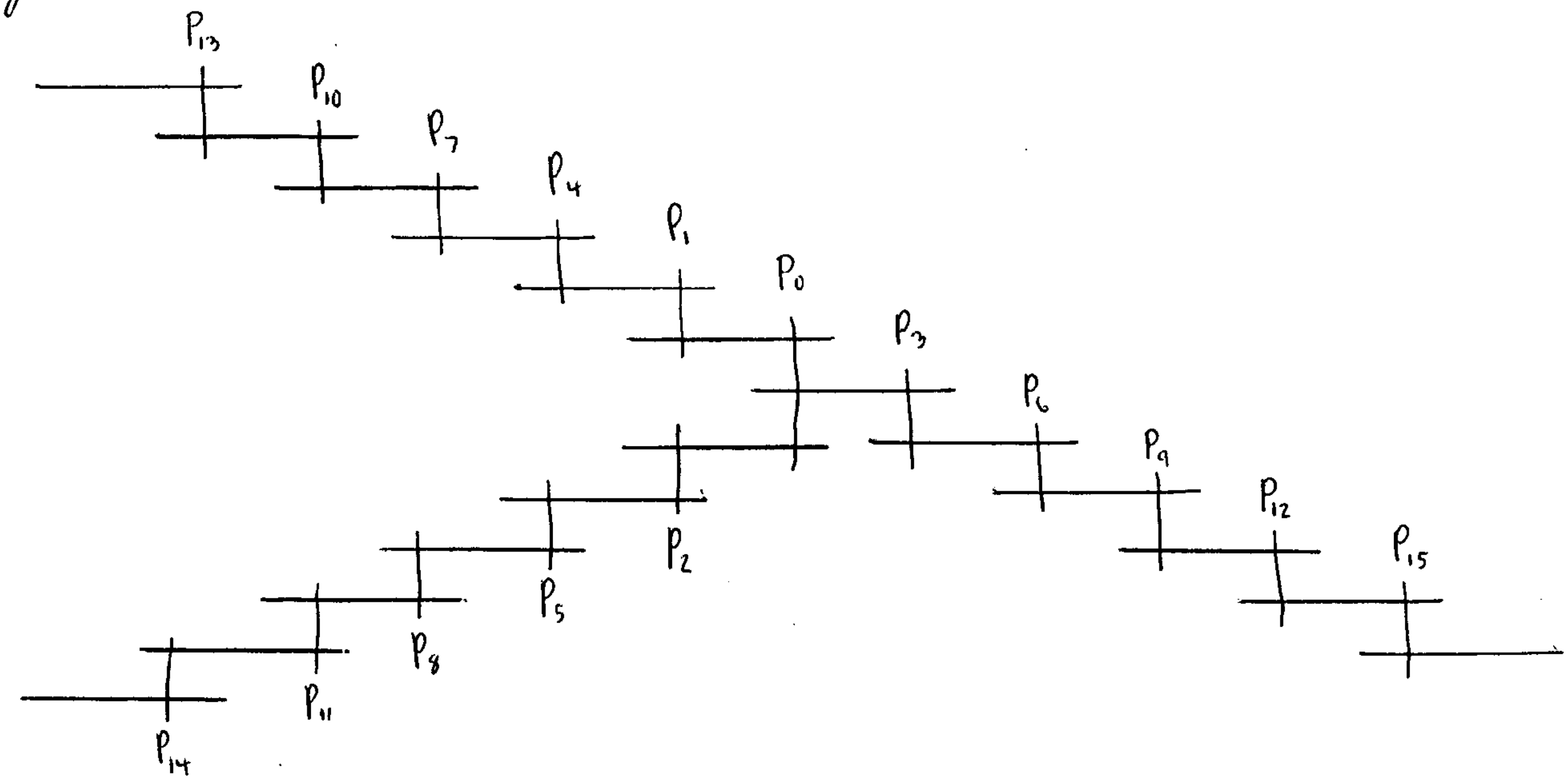
MOUSE CHR. 2
PETER GROOT

page 99 · Thu Nov 18 11:13:09 1993

Tiny Portion of a Physical Genome Map

False Positives

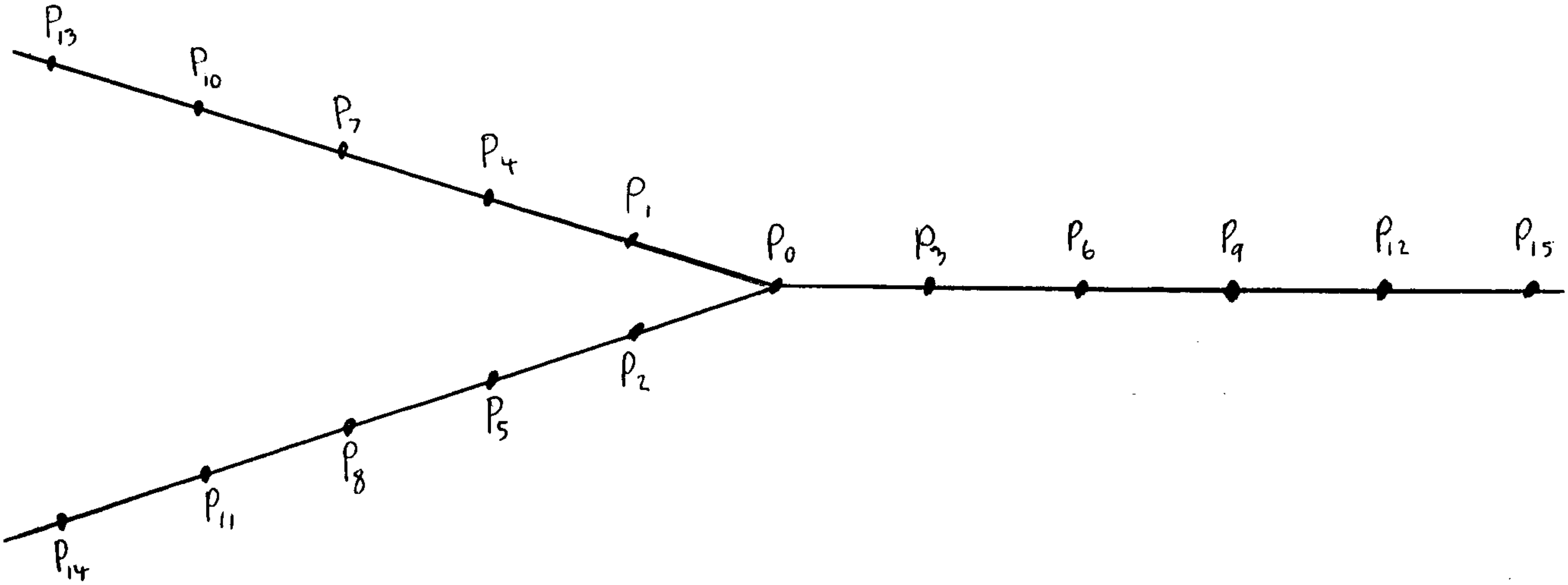
eg.



Unfortunately, no linear order of the probes is even approximately correct.

Observation

The data can be summarized as a graph in which each node is a probe.



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, anomolies.
  - 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.

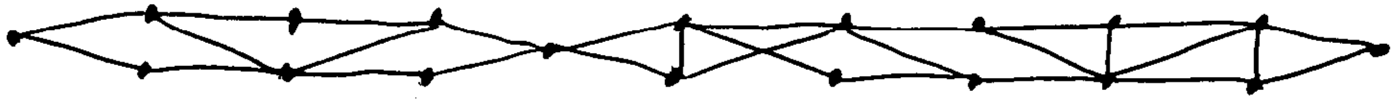
- Maps are full of errors.

Our Approach

- Abstract the genome data as a graph.

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

eg.



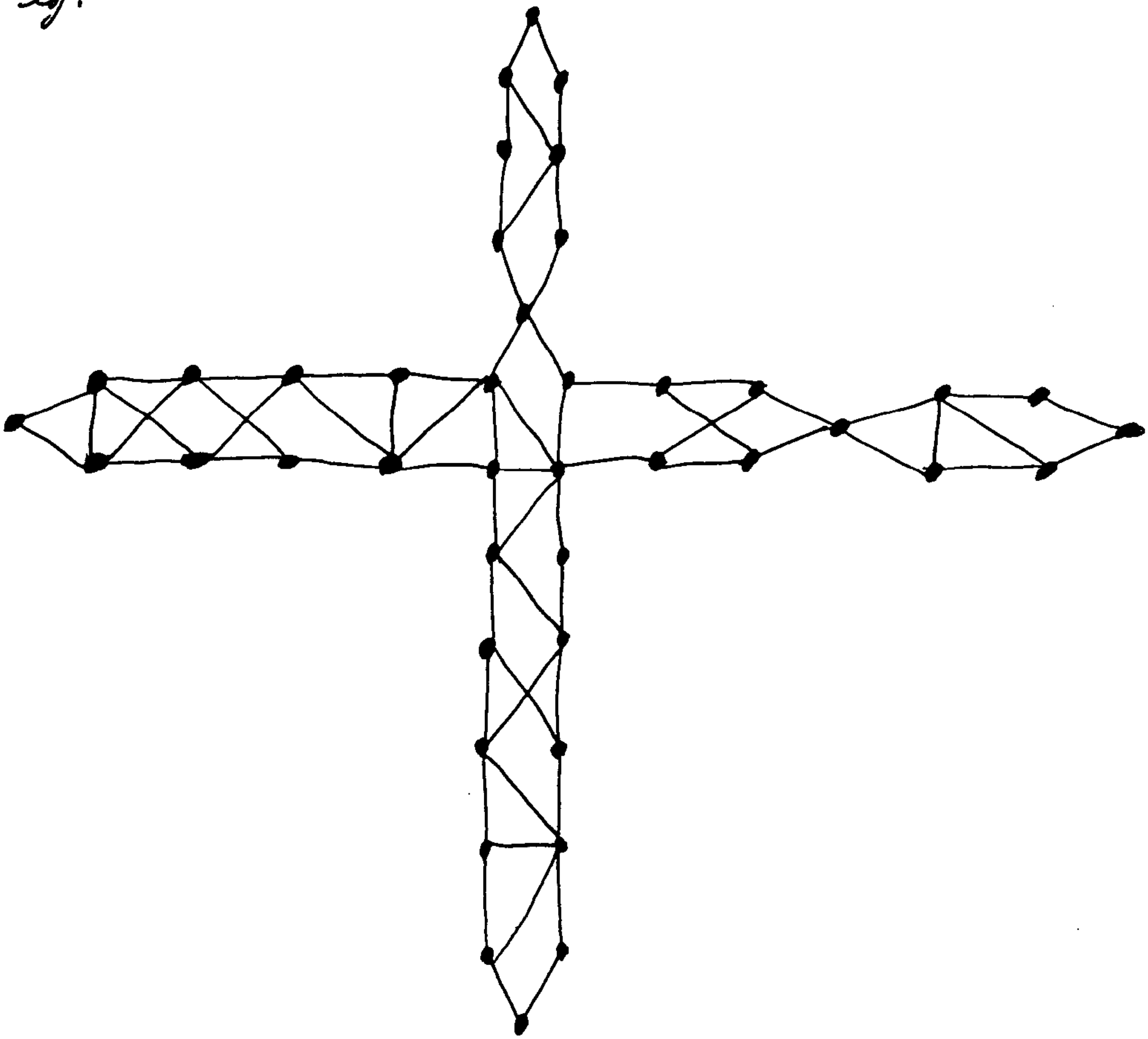
- Each node represents a point on the genome.

- An edge means that two nodes are "close" together on the genome.

Problem

Noise in the data distorts the linear structure of the graph:

eg.



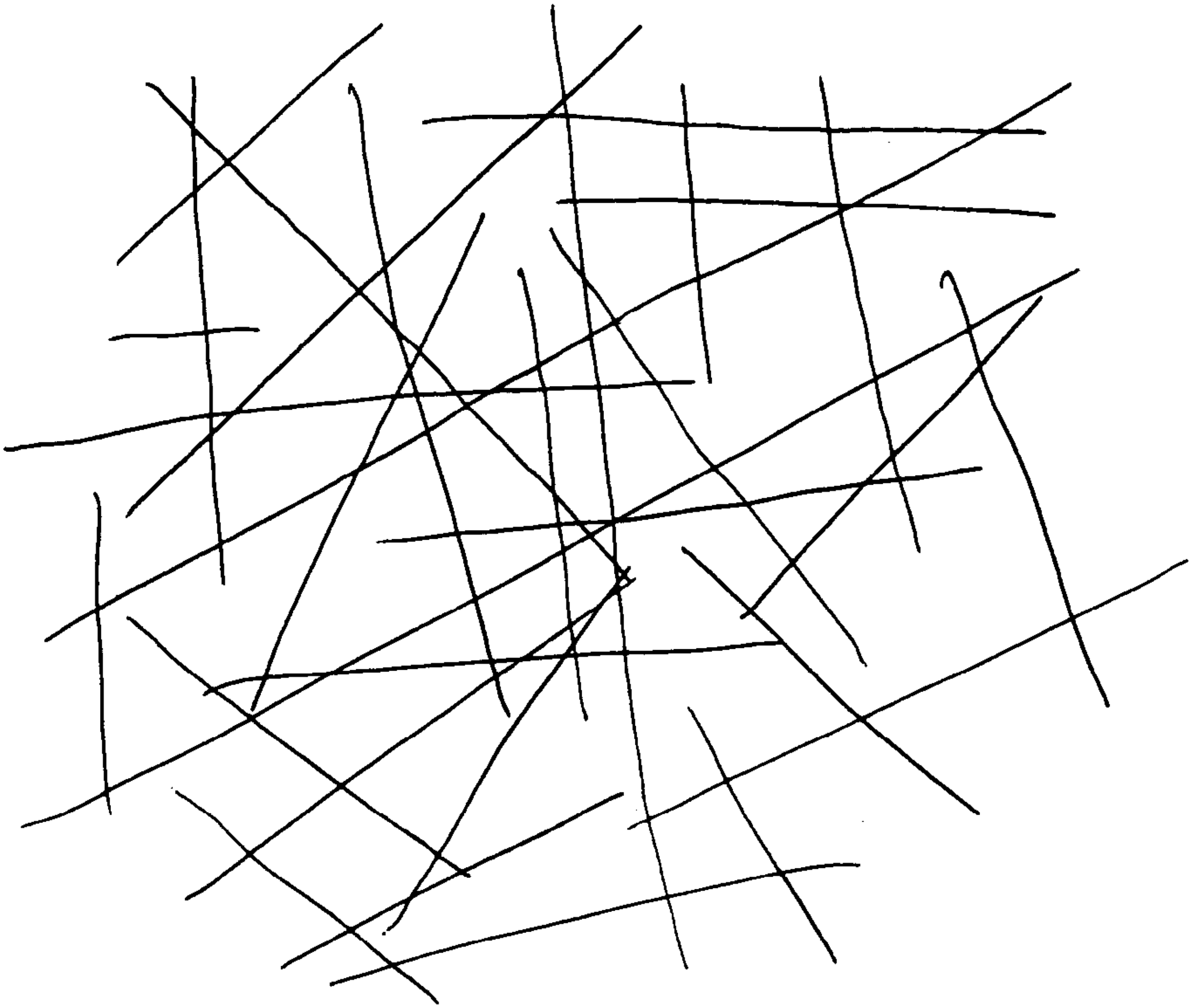
Real Genome Graphs

eg. 25,000 nodes

1,000,000 edges

Structure: Like a plate of spaghetti.

eg.



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.

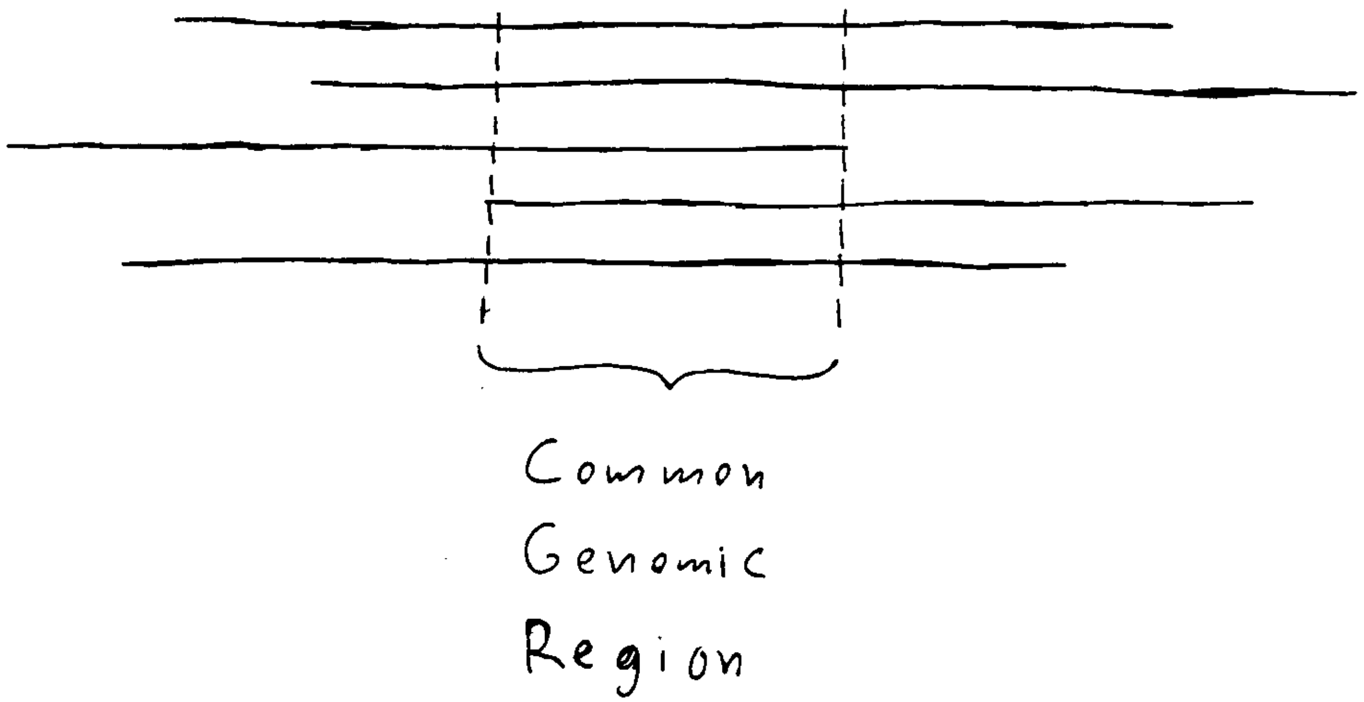
Graph Generation

(Data Integration)

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

<u>Definition</u>: A <u>cluster</u> is a maximal set of mutually overlapping clones.

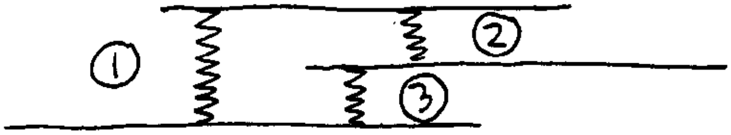
eg.



Observation 1

Clusters Filter out False overlaps.

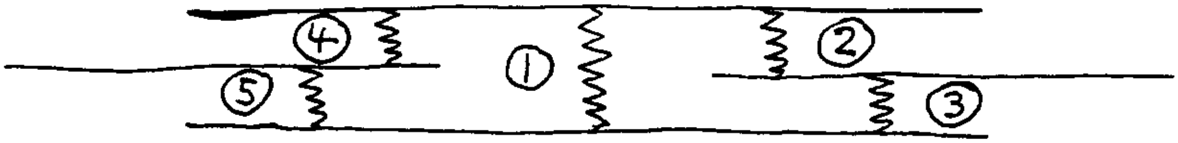
eg.



Overlap ① is corroborated by overlaps ② and ③.

Larger clusters have more corroboratio

eg.

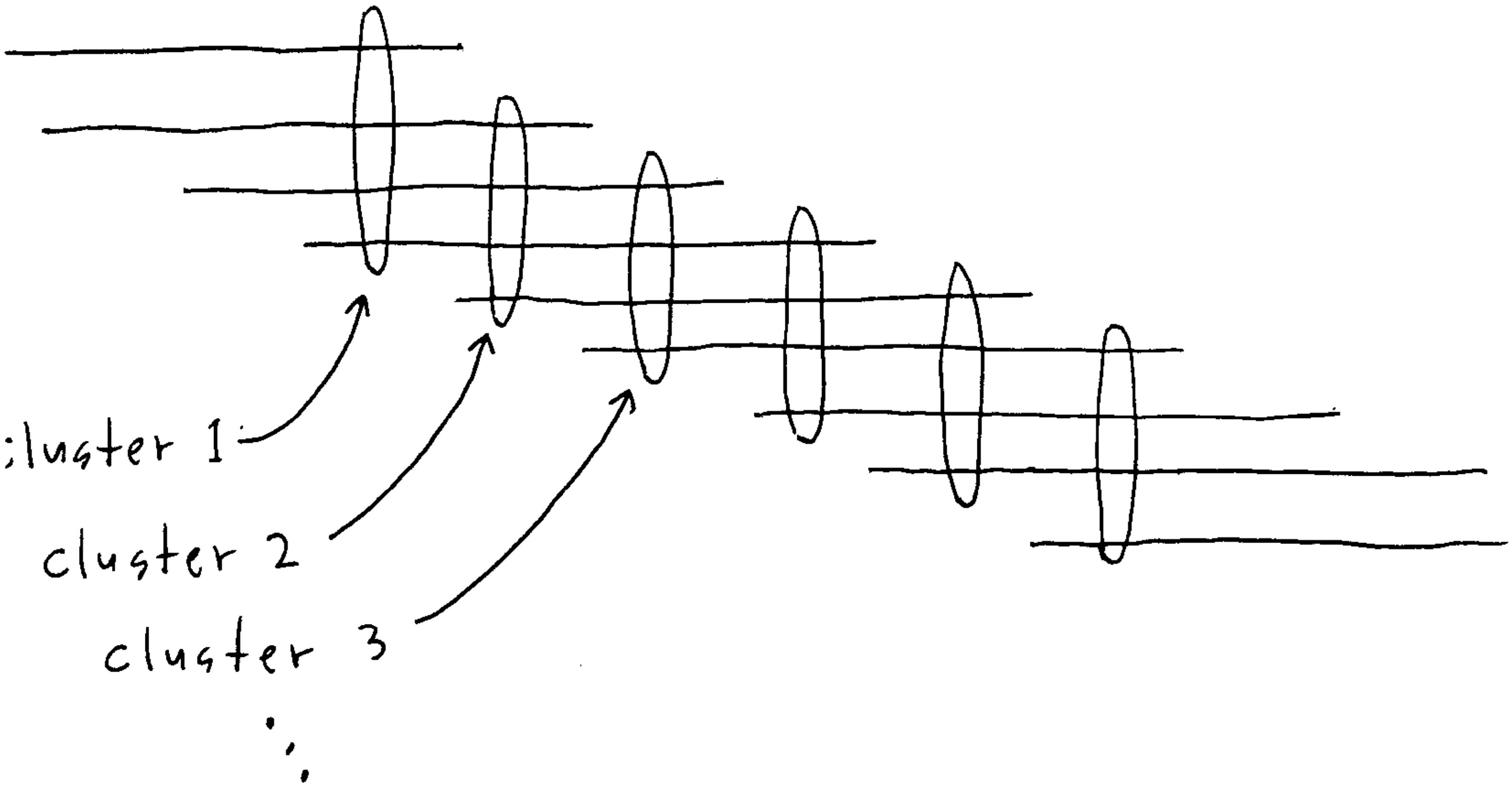


Overlap ① is corroborated by overlaps ② and ③, and by overlaps ④ and ⑤

Observation 2:

Average cluster size
= Average depth of clone coverage.

eg.

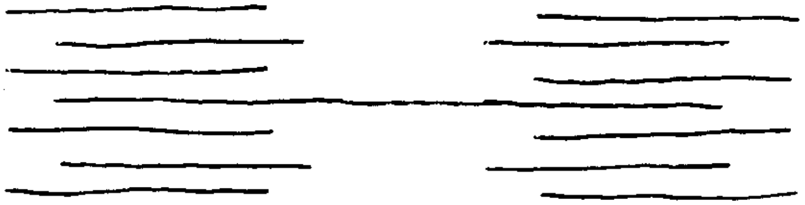


Cluster Proximity

Single Linkage:

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

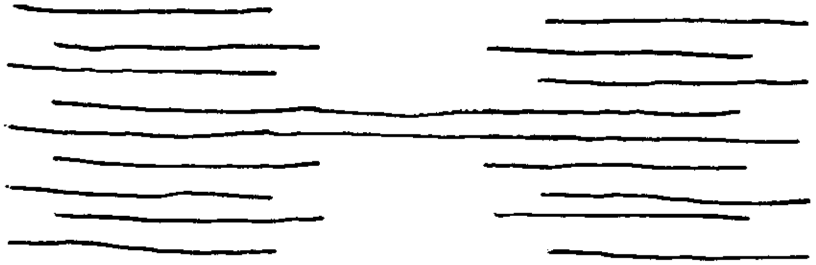
eg.



error prone

Double Linkage:

Two clusters are close if they have <u>two</u> clones in common.



much less error prone

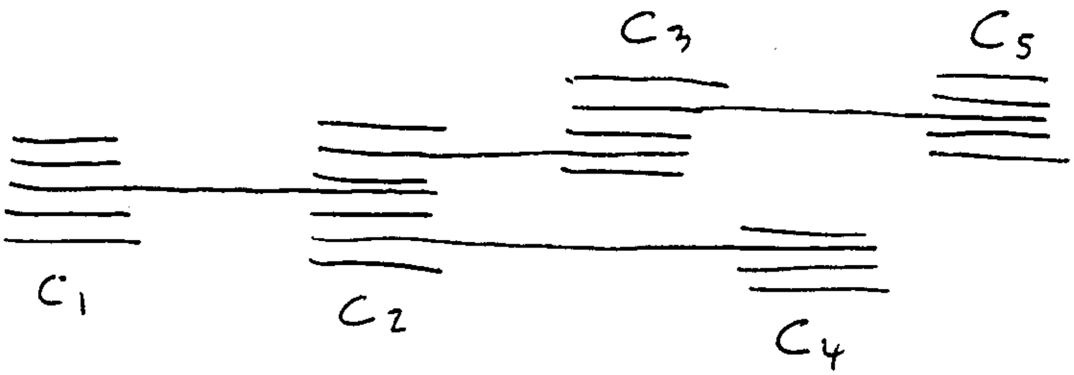
Many kinds of linkage are possible.

Cluster Graphs

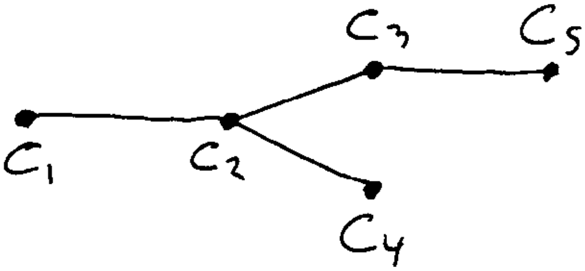
- Each node is a cluster.

- An edge between two nodes means the two clusters are close.

eg.



clusters



single-linkage cluster graph.

Map-Assembly Phases

(1) Overlap (cluster formation)

(2) Linkage (graph manipulation)

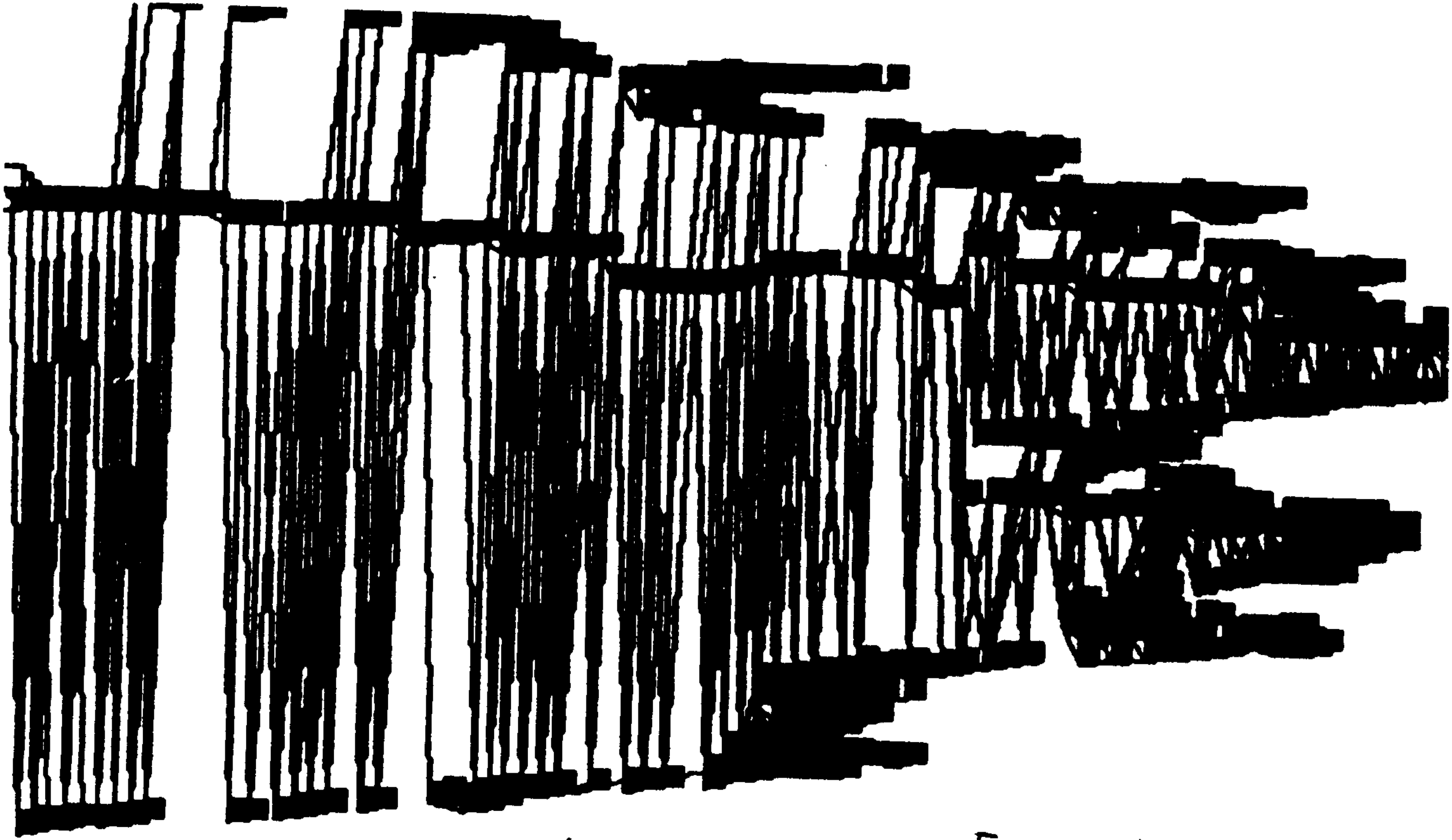
(3) Ordering (cluster ordering)

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.
    - 1 large graph instead of 23 smaller ones.
    - ie, a big mess.



Example 1

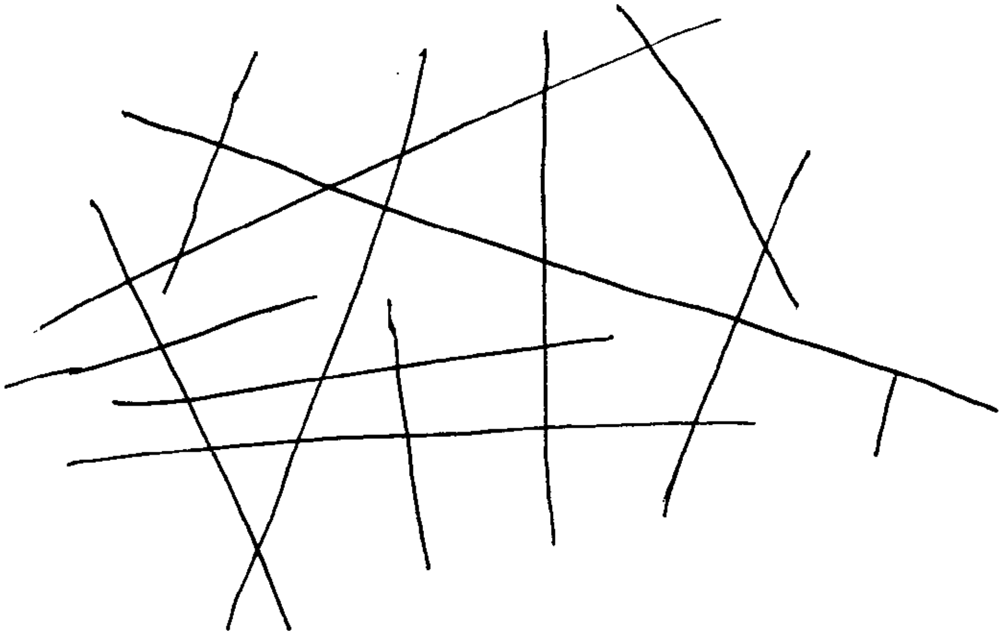
1,463 nodes

11,017 edges

Graph Abstraction (Simplification)

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

eg,

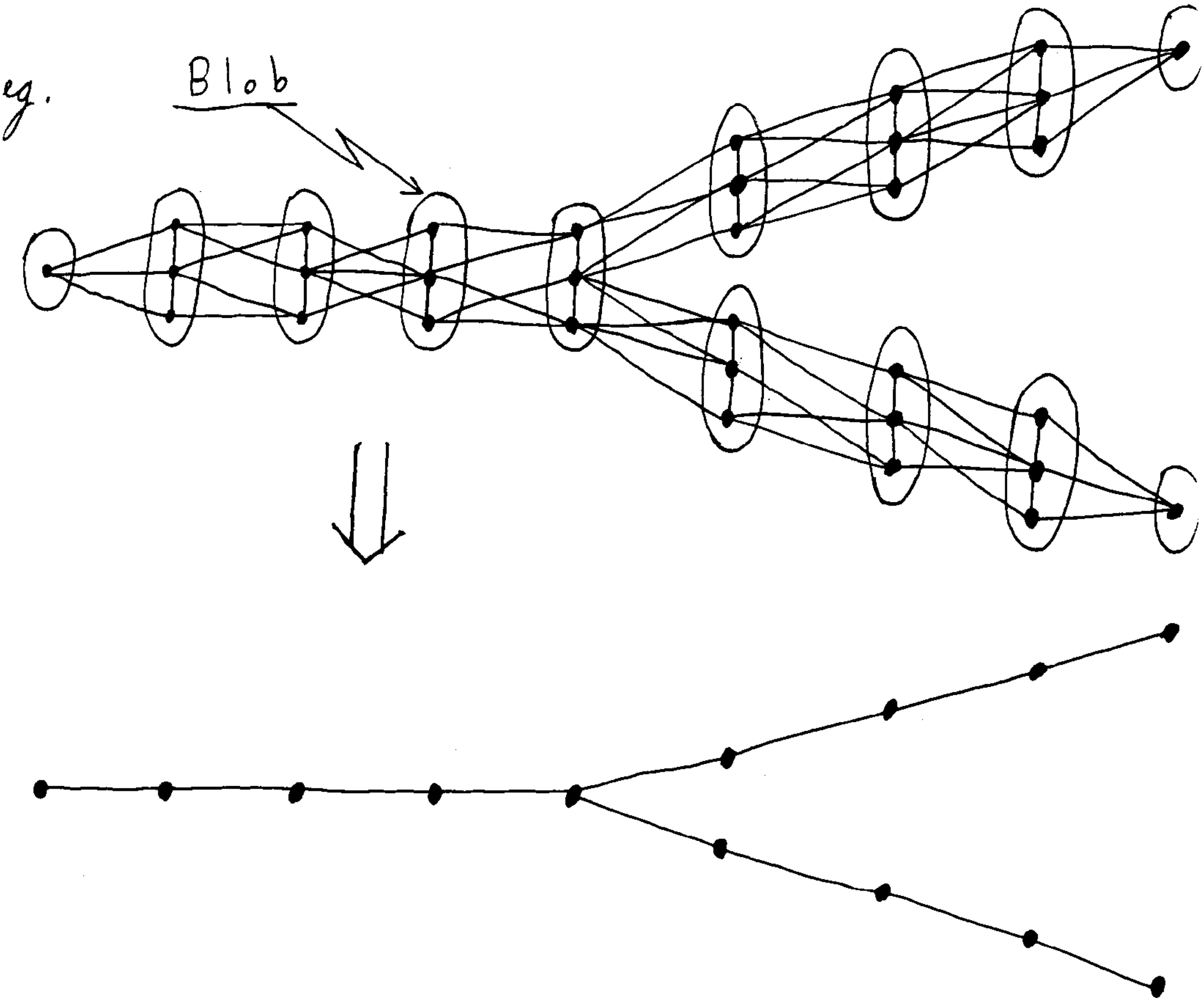


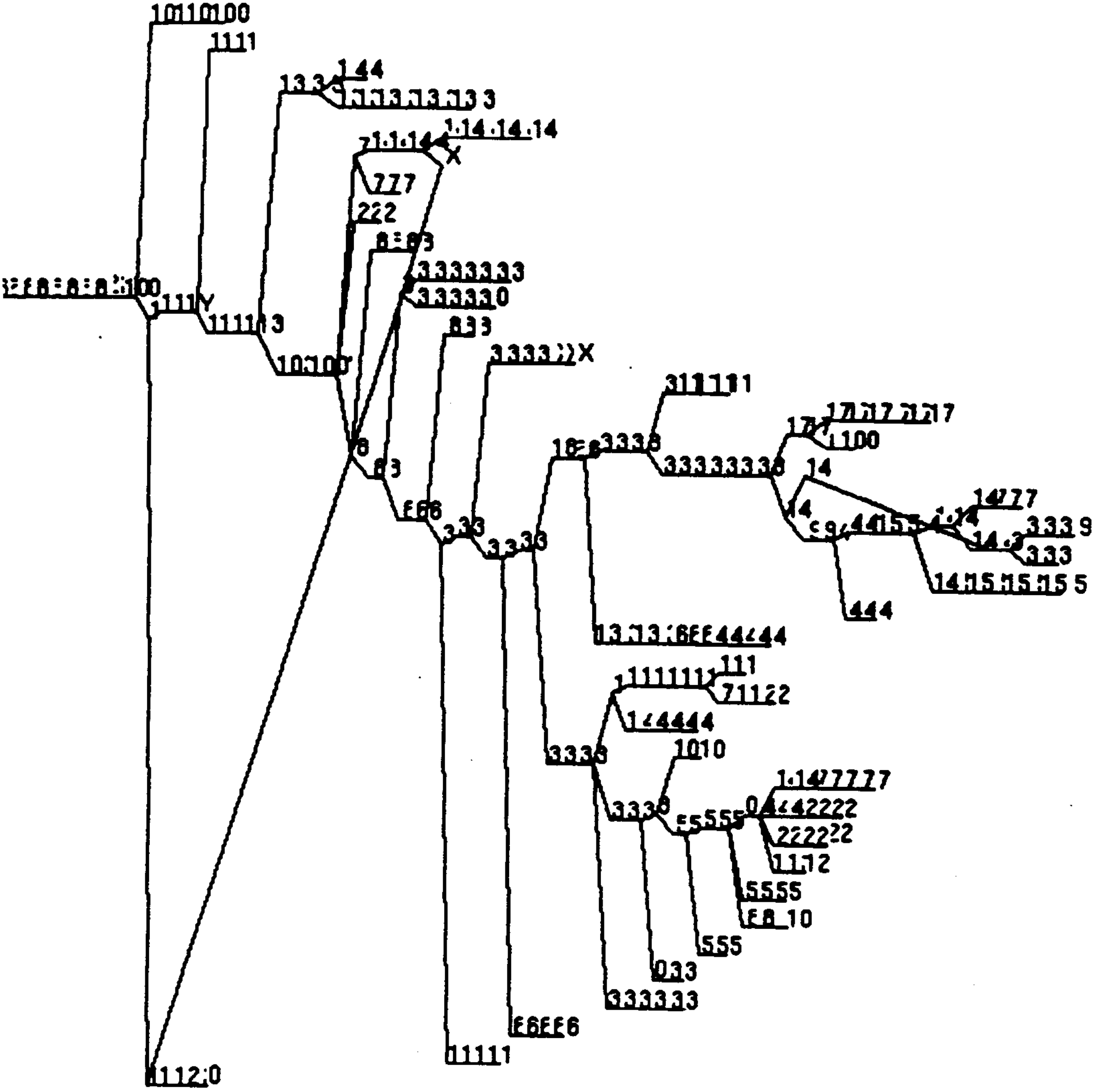
Next Step: Extract and display this structure.

Graph Abstraction

Idea: Coalesce groups of neighboring nodes into single nodes.

eg.





Abstracted sts graph of the human genome.

(301 nodes, 302 edges)

Example 2

Using all human mapping data from

- Whitehead / MIT
- Ceph/Genethon

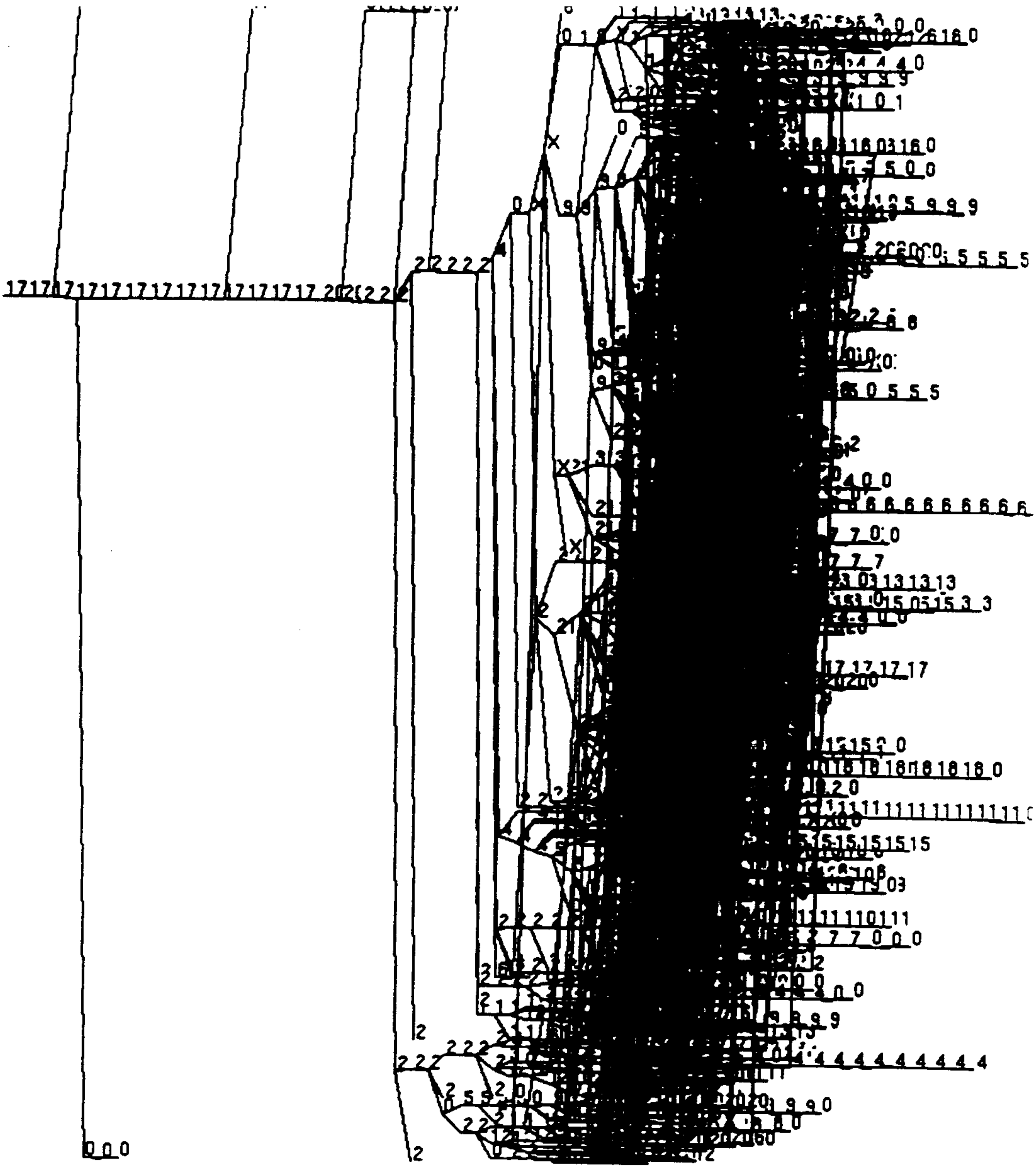
including

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

the cluster graph has

- 450,342 edges,
- 27,379 nodes.

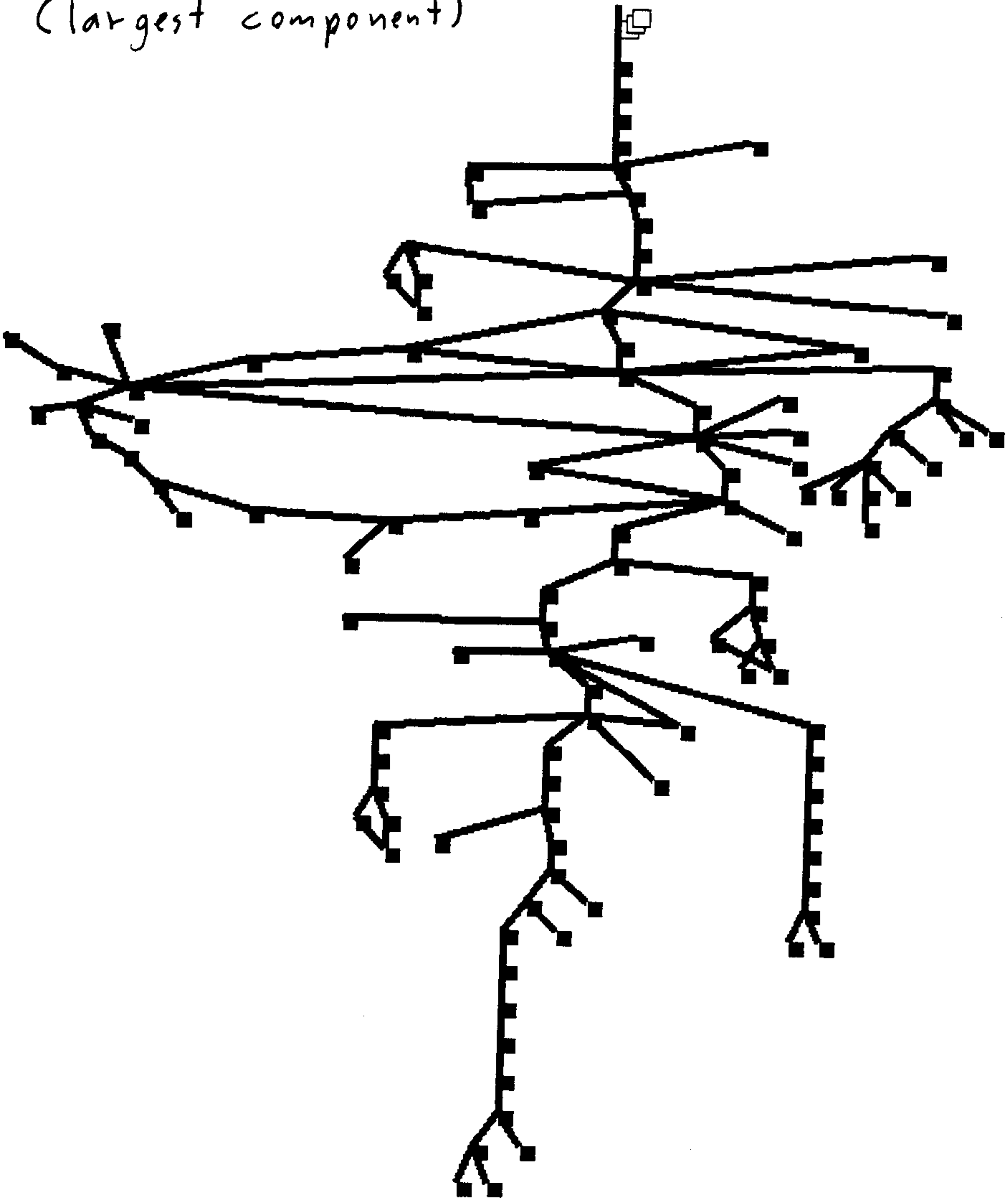
Most edges (432,888) are concentrated in a single connected component.



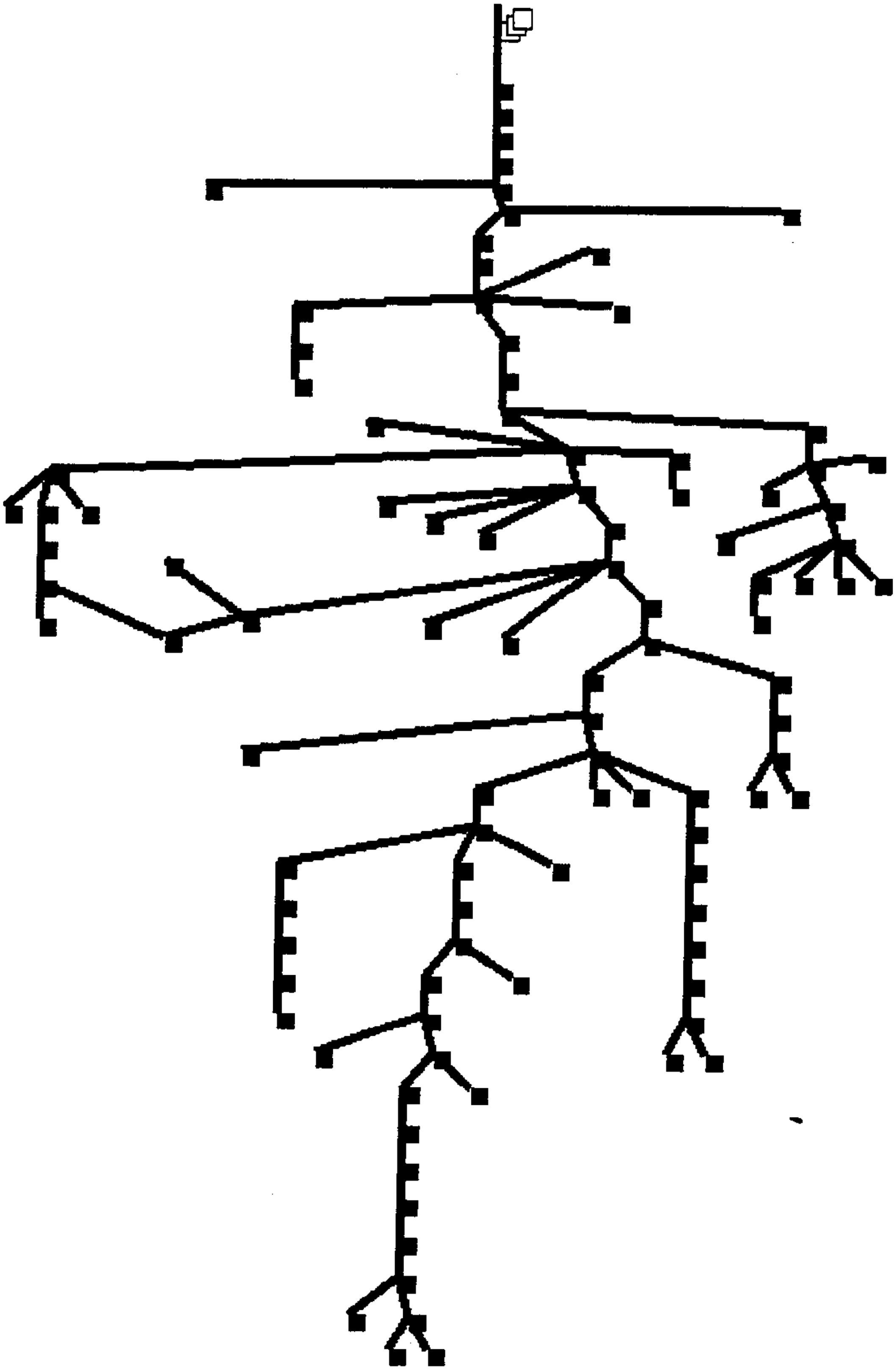
Abstracted cluster graph of human genome.
(4621 nodes, 5002 edges)

Abstracted cluster graph of Human Chromosome 7

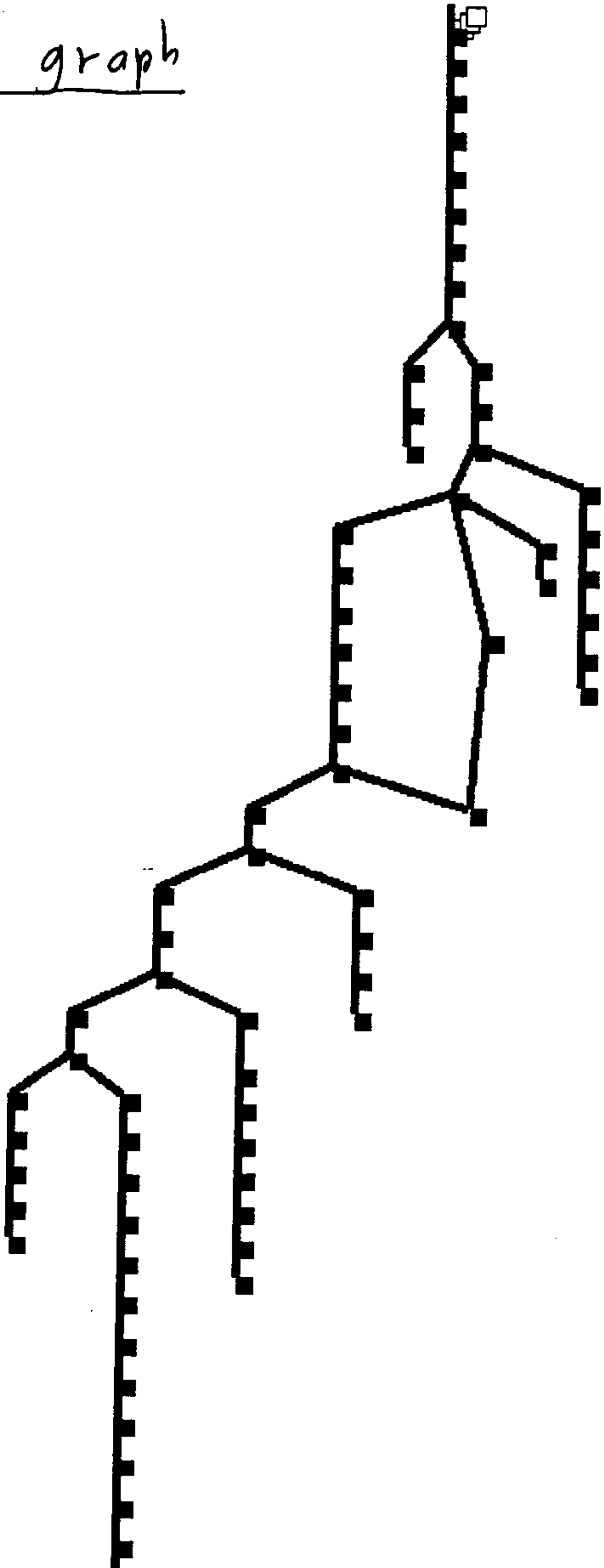
(largest component)

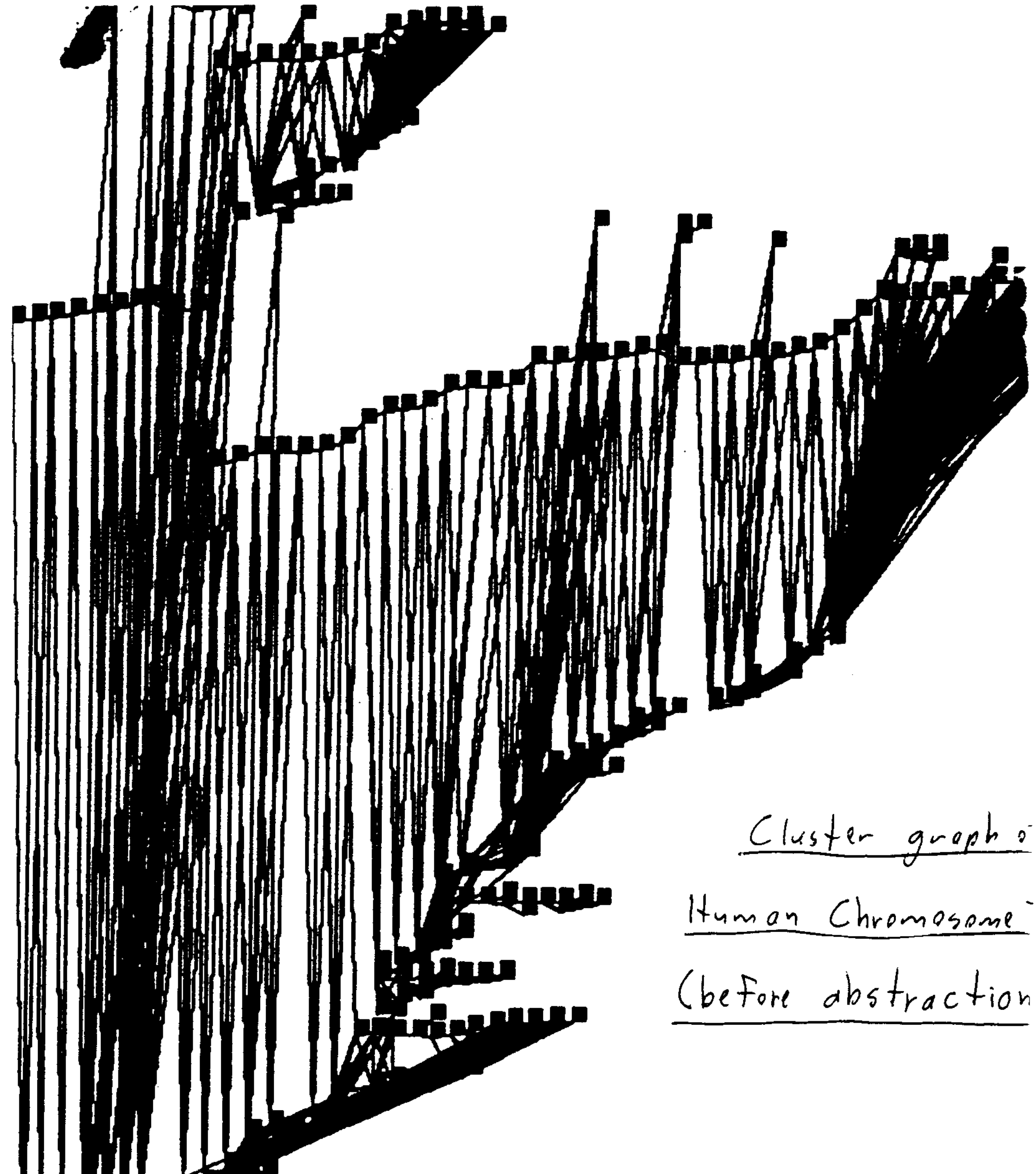


Partially trimmed graph



Fully trimmed graph



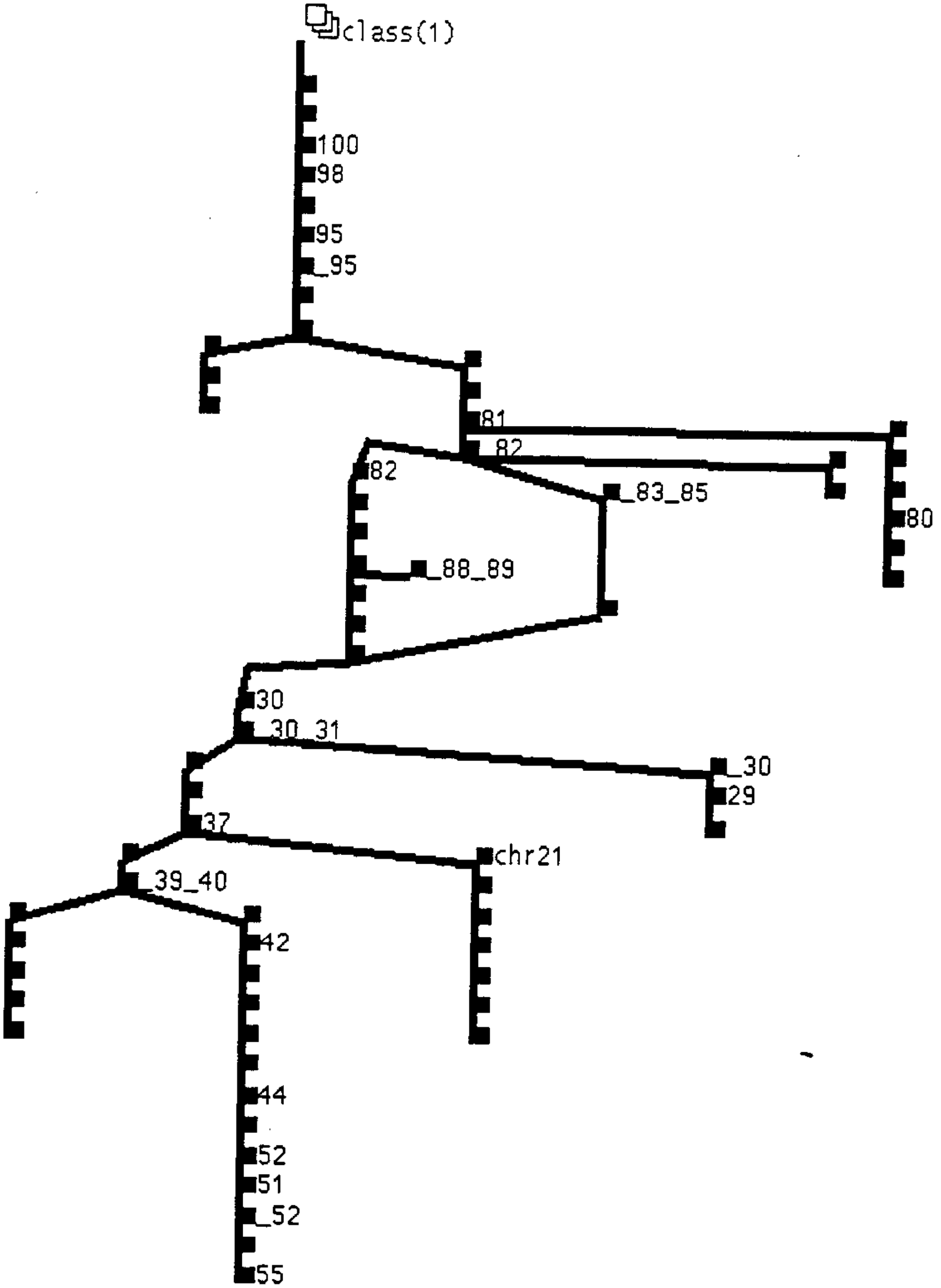


Cluster graph o

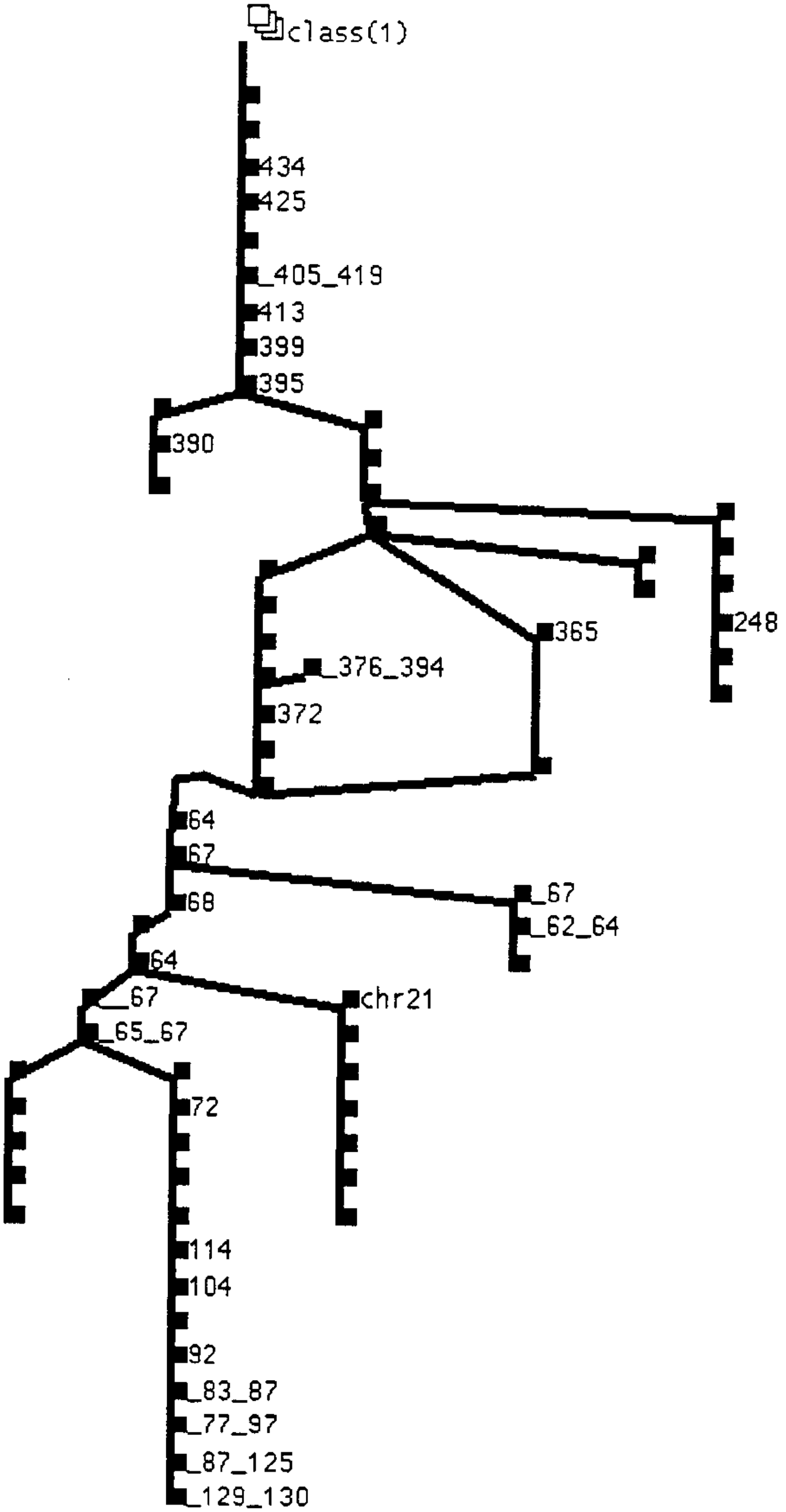
Human Chromosome

(before abstraction

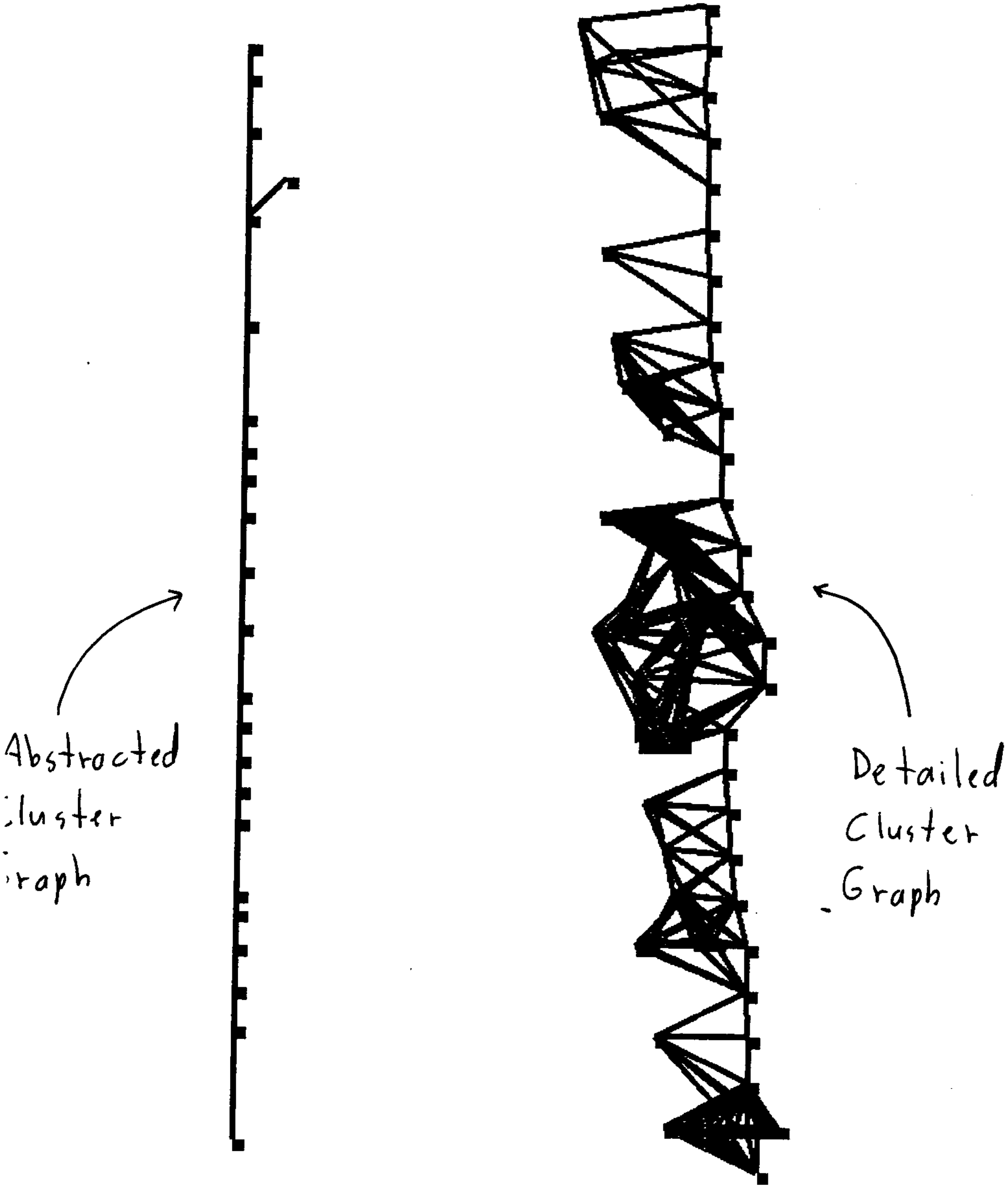
Graph of Chromosome 7
with genetically-mapped positions



Graph of Chromosome 7 with radiation-hybrid positions



Contig of Human Chromosome 7



Summary

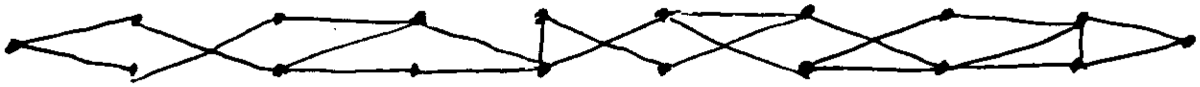
Cluster Graphs:

- 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:

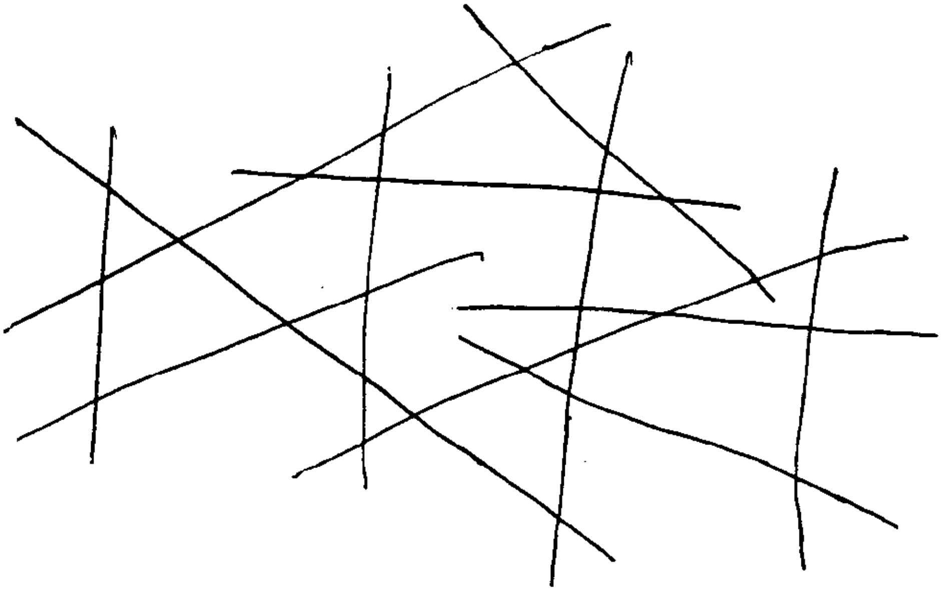
<u>overlap</u>   <u>linkage</u>   <u>ordering</u>

Errors and Anomolies

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

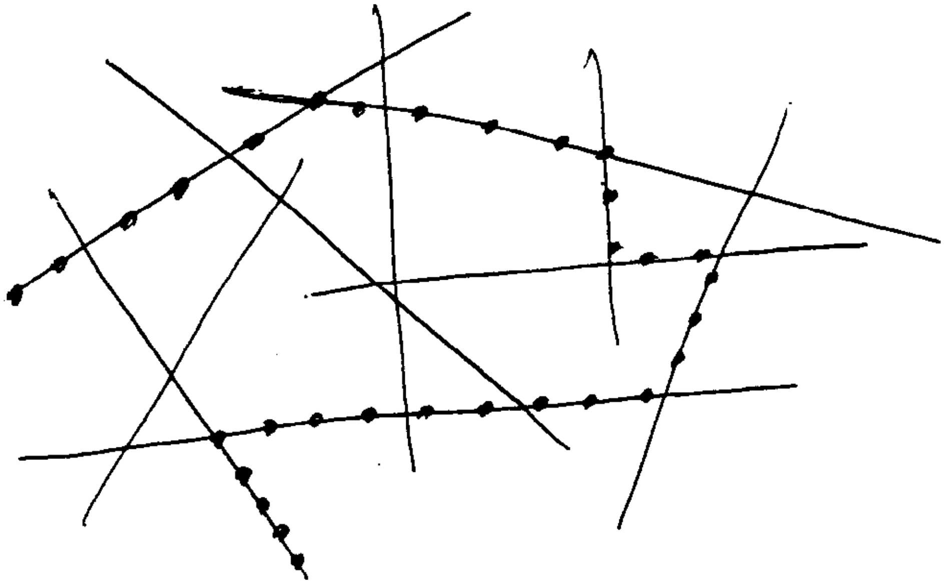
eg. 

- 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 (eg, chr. assignments, genetic maps, rh maps, long linear subgraphs can be extracte

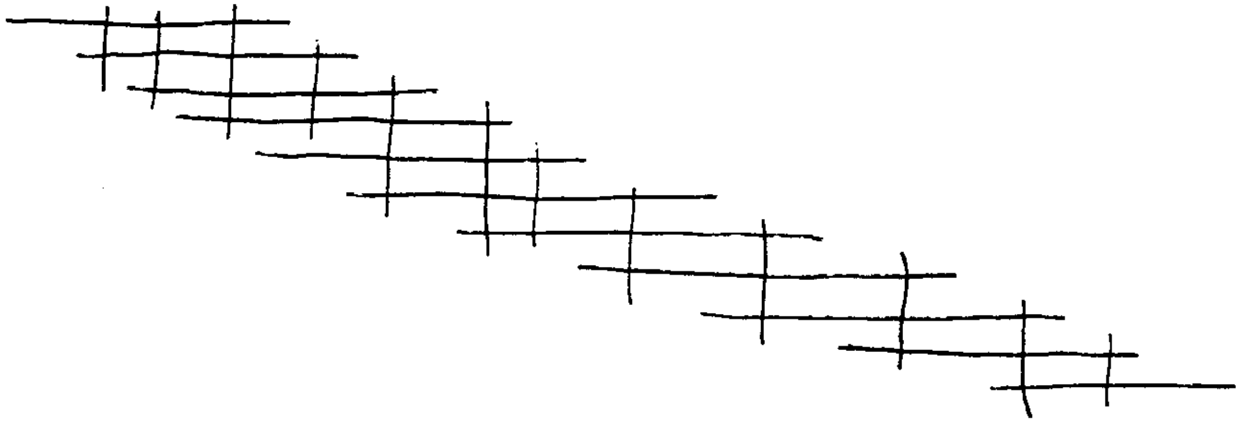


- These subgraphs represent genomic contigs.

Map Assembly

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

eg.



ie, Cluster graphs help to clean up data and remove anomolies.

Our Publications

Available on the Web, at

www.cs.toronto.edu/~bonner

under three categories:

- Genome Mapping
- Sequence Databases
- Laboratory Workflow