"Good" Markers and "True" Maps: Combinatorial Algorithms for Genetic and Radiation Hybrid Mapping

Humberto Ortiz Zuazaga
University of Texas
Health Science Center
at San Antonio
`humberto@momo.uthscsa.edu`

Joint work with Rosemarie Plaetke, UTHSCSA

Human Genome Mapping

- Goal: sequence the entire human genome.

- Human genome consists of 3 billion base pairs (too big).

- Subgoal: place 30,000 markers evenly on the entire genome.

These genetic markers serve as landmarks in the search for disease causing genes, and help direct the sequencing of the genome.

Computational Biology

- The systematic study of computational problems arising in biology.

- Includes elements of biology, computer science, mathematics, and operations research.

Pedigree Report for Some Chromosome 1 Markers

```
Pedigree: 1331

Grandpa        Grandma               Grandpa         Grandma           Marker
------------------------------------------------------------------------------
  1  1           1  1                  1  2             2  3            UT851
  2  4           3  4                  4  5             1  4           UT1398
  1  3           3  3                  2  3             3  4           UT1243
  2  4           2  3                  1  2             4  4           UT1234
------------------------------------------------------------------------------

           Dad                                  Mom
------------------------------------------------------------------------------
        P 1  1 M                             P 1  2 M                    UT851
        P 2  3 M                             P 5  4 M                   UT1398
        P 1  3 M                             P 3  3 M                   UT1243
        P 2  3 M                             P 2  4 M                   UT1234
------------------------------------------------------------------------------

Pedigree: 1331
Children: 3 -
------------------------------------------------------------------------------
  1  2 M     1  2 M     1  2 M     1  2 M     1  2 M     1  2 M        UT851
P 2  5 P   P 2  4 M   M 3  4 M   P 2  5 P   P 2  4 M   M 3  4 M       UT1398
M 3  3     M 3  3     M 3  3     P 1  3     P 1  3     P 1  3         UT1243
P 2  2 P   P 2  4 M   M 3  4 M   P 2  2 P   P 2  4 M   M 3  4 M       UT1234
------------------------------------------------------------------------------
```

Genetic Mapping

- Goal: The determination of orders and distances among markers on a chromosome based on the observed patterns of inheritance of the alleles of the markers in three generation pedigrees.

- Problem: For a variety of reasons the genotypic information is not complete, and not all crosses in human pedigrees are informative. In addition, the time required to order markers grows exponentially with the number of markers.

- Solution: Only use "good" markers to make maps. Biologists already have a notion of a "framework" map, a map of a subset of the markers which has very high odds against inversion of adjacent markers.

Counting Obligate Breaks as an Estimate of Genetic Distance

A simple estimate of the genetic distance between two markers is the number of observed recombinations between the markers in the data set. For the first two markers in our sample pedigree we would have:

UT851	U	M	U	M	U	M	U	M	U	M	U	M
UT1398	P	P	P	M	M	M	P	P	P	M	M	M
Breaks		1						1				

for a total of 2 breaks.

This technique based on counting the number of recombinations is known as meiotic breakpoint analysis (BPA).

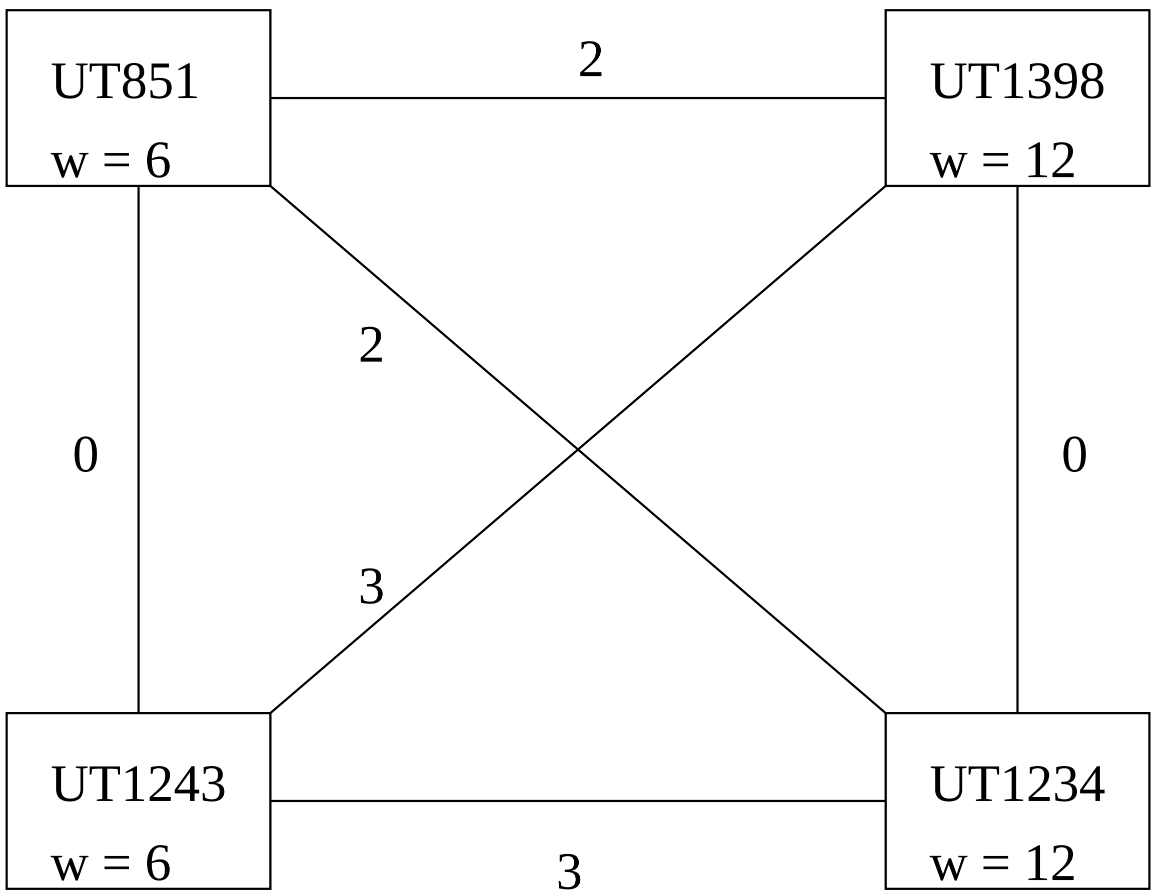
Selecting Genetic Markers With `wclique`

We have implemented an algorithm for screening markers for genetic mapping by transforming the marker selection problem into a maximum weighted clique (MWC) problem.

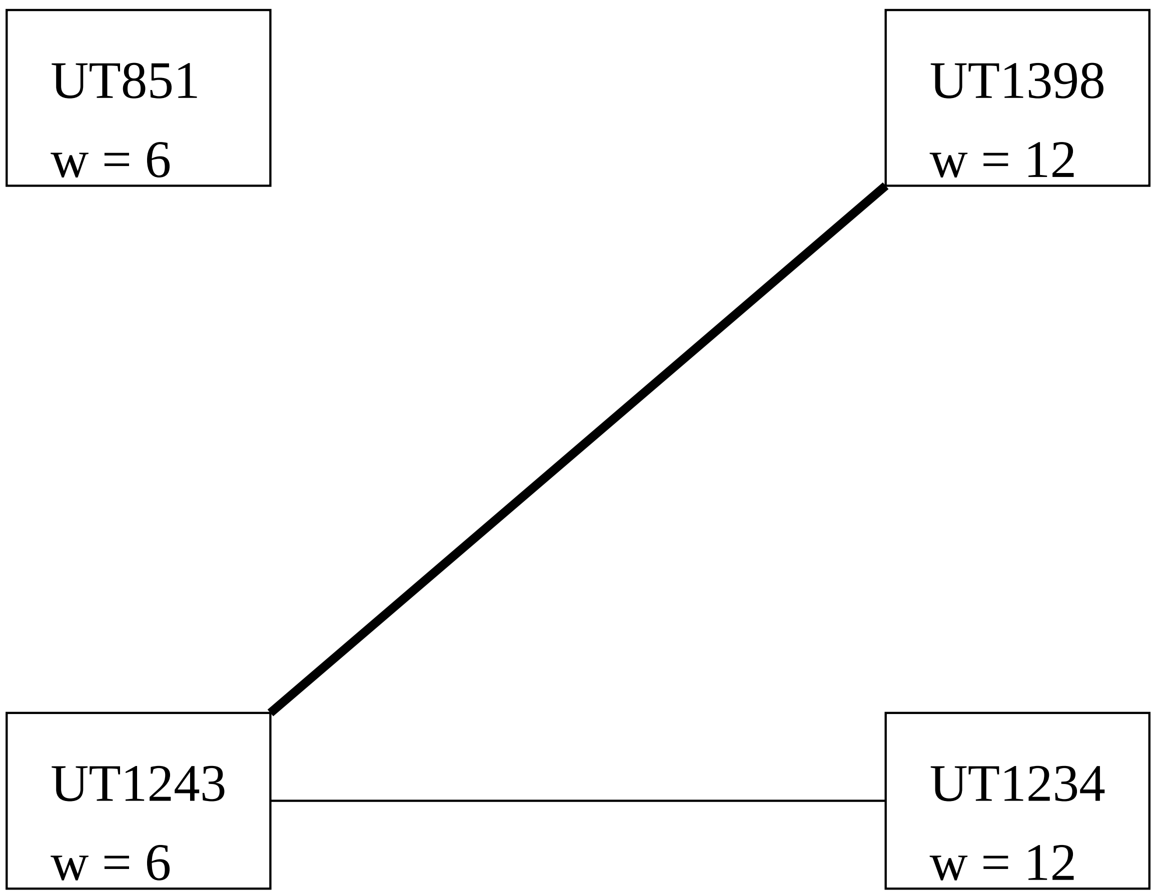
Each marker becomes a node of a graph. The weight of the node corresponds to the frequency of known phases (*i.e.*, the total count of P and M phases) for this marker. This measure directly reflects how informative each marker is for linkage analysis.

Two nodes in the graph are connected by an edge whose weight is the number of breaks between the corresponding markers. This is a heuristic estimate for genetic distance, but has been shown to result in correct marker orders as the number of gametes tends to infinity.

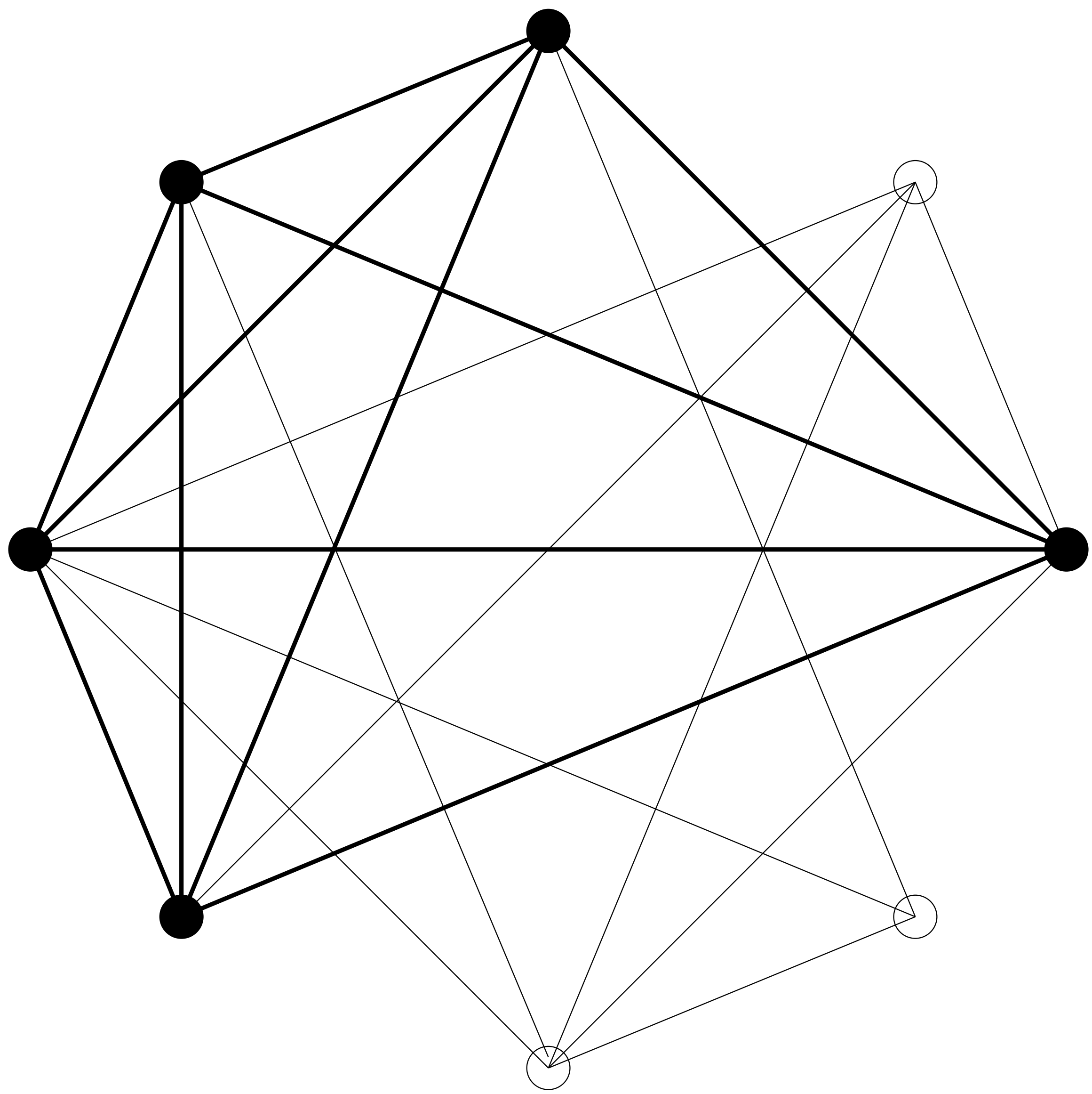
A Small Distance Graph



A Small Distance Graph and a Clique



A Larger Distance Graph and Clique



The Maximal Weighted Clique problem is NP-Complete

The MWC is a well known graph problem, extensively studied in computer science. Unfortunately, it belongs to the class of NP-complete problems, for which there is unlikely to be an efficient algorithm.

Building a linear map by ordering genetic markers so as to minimize the number of recombination events in a set of gametes can also be cast as a graph problem, the traveling salesman problem (TSP), which is also NP-complete.

But I Still Need a Map

- Exact algorithms can work on small sets of markers.

- Local search techniques can find near optimal solutions for some of these problems, at the cost of not knowing if an optimal solution was ever found. The best heuristics for TSP can find a solution with 1.05 times the optimal cost.

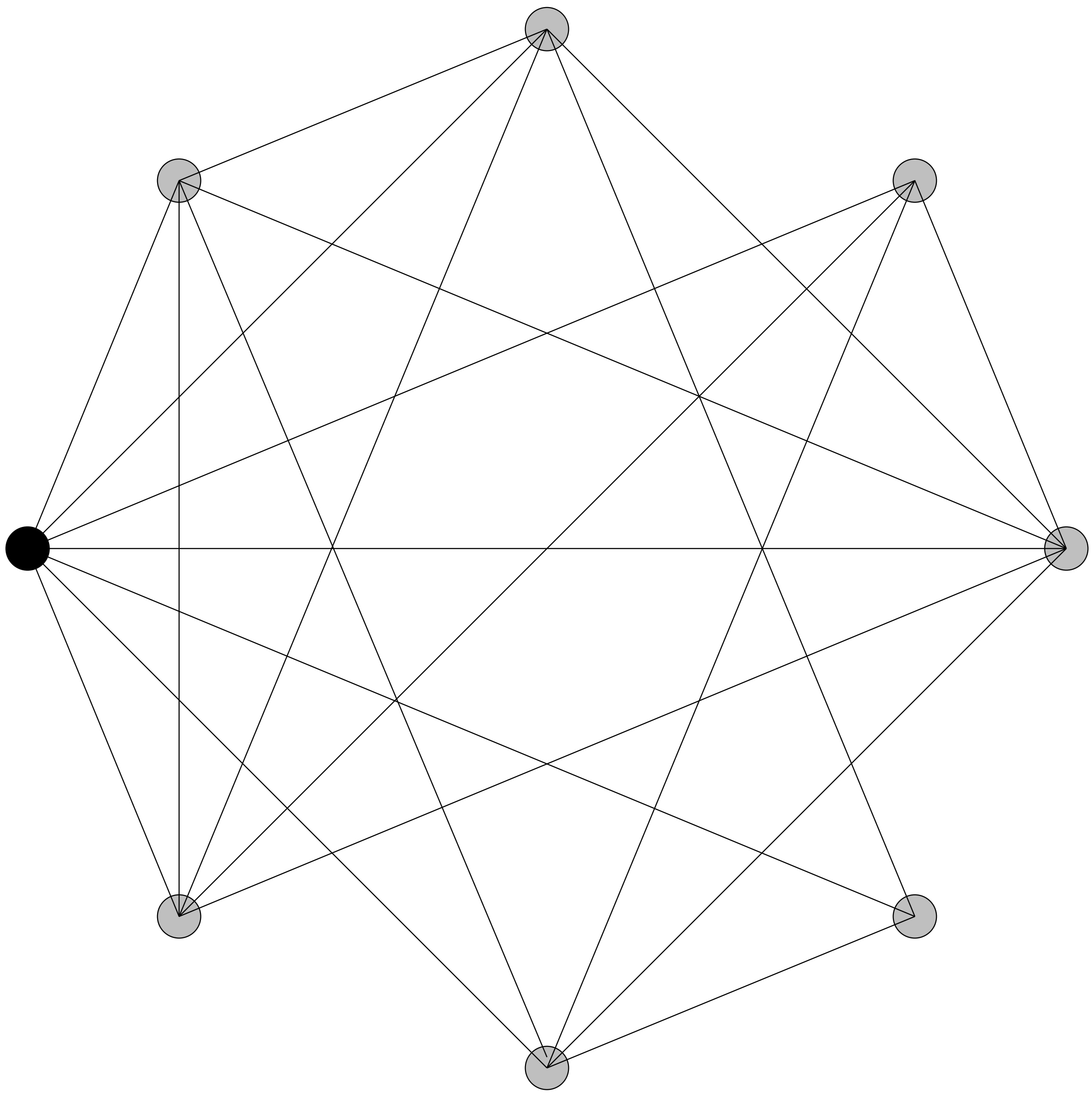
- A change in the formulation of the problems can enable other algorithms to be used. For example, if the data had no errors, was complete, and no double recombination events occurred, ordering genetic markers would be equivalent to the consecutive ones problem (C1P) for which there are linear time algorithms.

Exact Algorithm for `wclique`

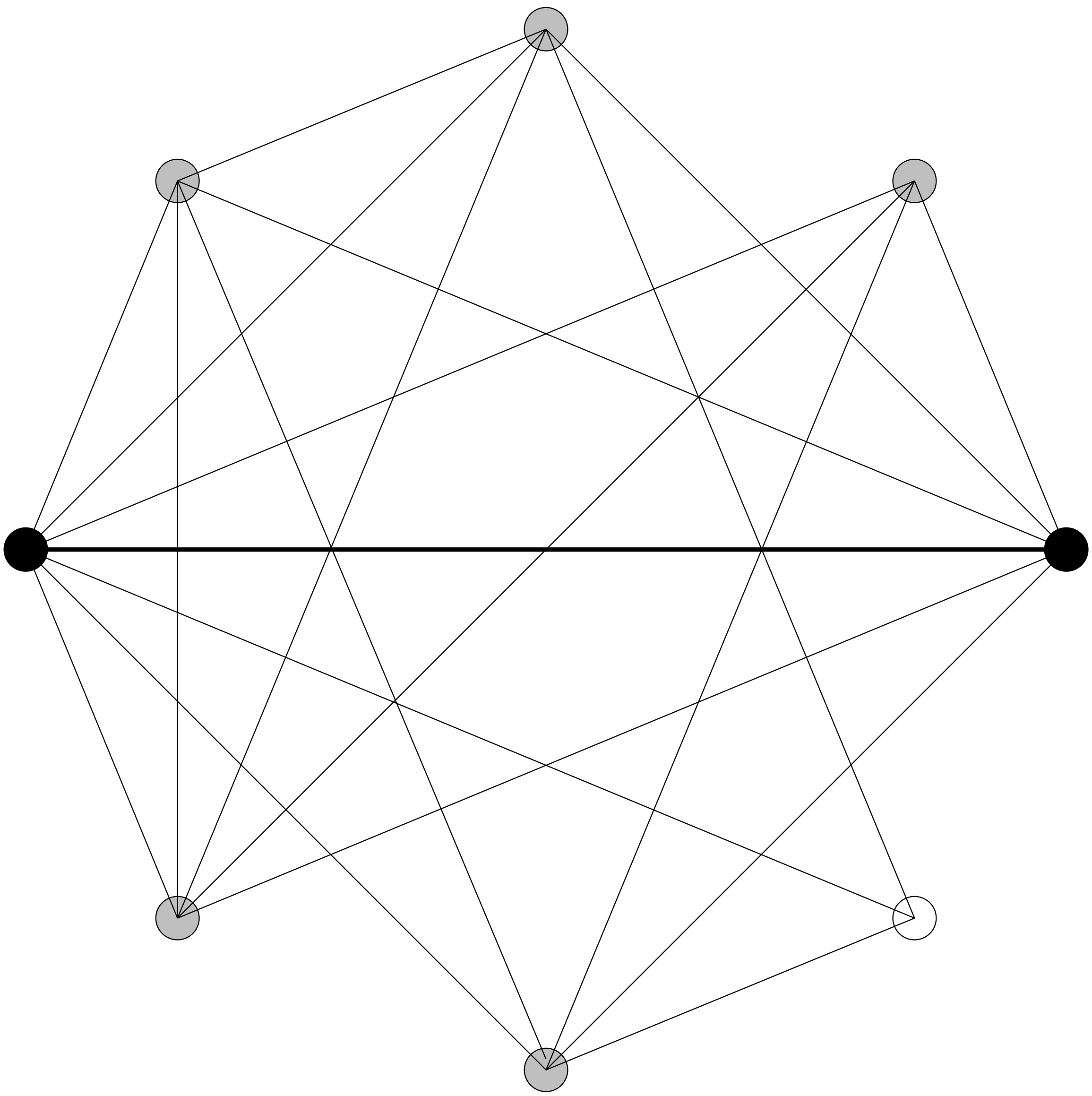
We find the maximum weighted clique using a partial enumeration algorithm based on the `dfmax` algorithm of (Applegate & Johnson 1993) for the maximum cardinality clique.

procedure `wclique`(graph G,
  vertex set C, vertex set T, bound B)
  if $\Sigma_{c \in C} w(c) + \Sigma_{t \in T} w(t) \geq B$ **then**
    for each node v in T **do**
      $T' := T - v$
      delete nodes from T' that are not adjacent to v in G
      if $\Sigma_{c \in C} w(c) + \Sigma_{t \in T'} w(t) + w(v) \geq B$
        then $B :=$ `wclique`$(G, C \cup v, T', B)$
      endif
      $T := T - v$
    endfor
    if $\Sigma_{c \in C} w(c) \geq B$ **then**
      B := $\Sigma_{c \in C} w(c)$
      output C
    endif
  endif
  return B
end

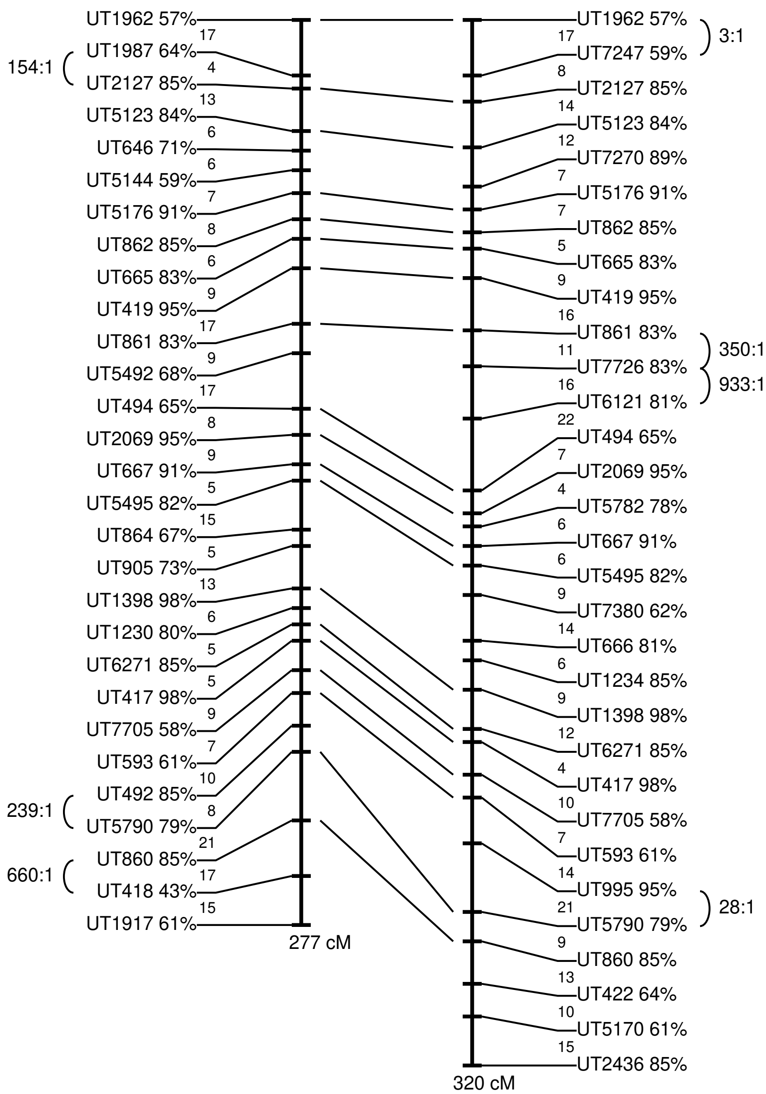
How to Grow a Clique: Step 1



How to Grow a Clique: Step 2



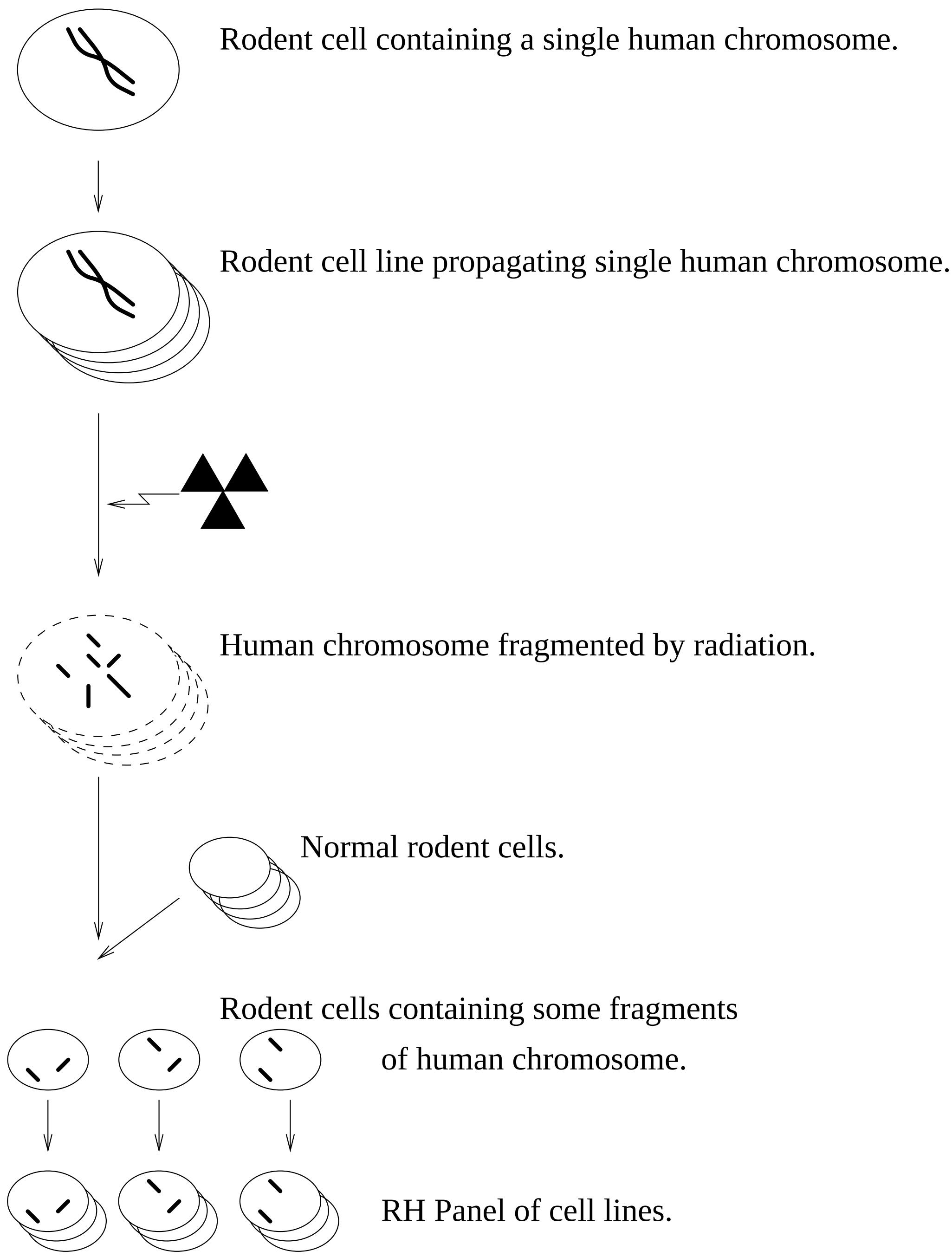
Comparison of MLA Maps of Hand-Selected and `wclique`-Selected Markers



Radiation Hybrid Mapping

In recent years a new technique for mapping has become very widespread. Radiation hybrid (RH) mapping allows researchers to construct very fine scale maps by analyzing radiation induced breaks in human chromosomes. This technique does not require the analysis of large pedigrees needed in genetic mapping.

Constructing an RH panel



Sample RH panel results

Each cell line in the RH panel is tested for the presence “P” or absence “M” of each of the markers. The experiments are done in duplicate, and conflicting or missing results are marked “U”.

RAF	U	P	M	M	P	P	M	M	P
D3S12	M	P	P	M	M	P	P	M	U
Breaks			1		1		1		

Distances between markers can be estimated by counting the number of obligate breaks, as shown above.

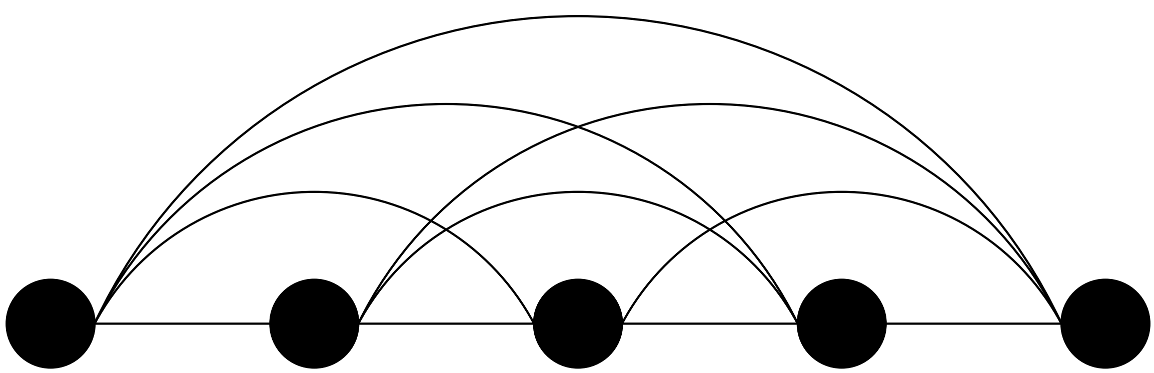
RH Mapping

- Goal: Construct a linear map of markers that explains the retention patterns.

- Problem: Many markers cannot be placed unambiguously. Several "good" maps explain the data equally well. Different algorithms find different "good" maps.

- Solution: Find a "true" map. One for which many algorithms agree on the order of the markers.

A Distance Graph with the Nodes Arranged in Map Order



Definition of a Consistent Map

Given a graph with the nodes arranged in map order:

- A **short edge** connects two adjacent markers.

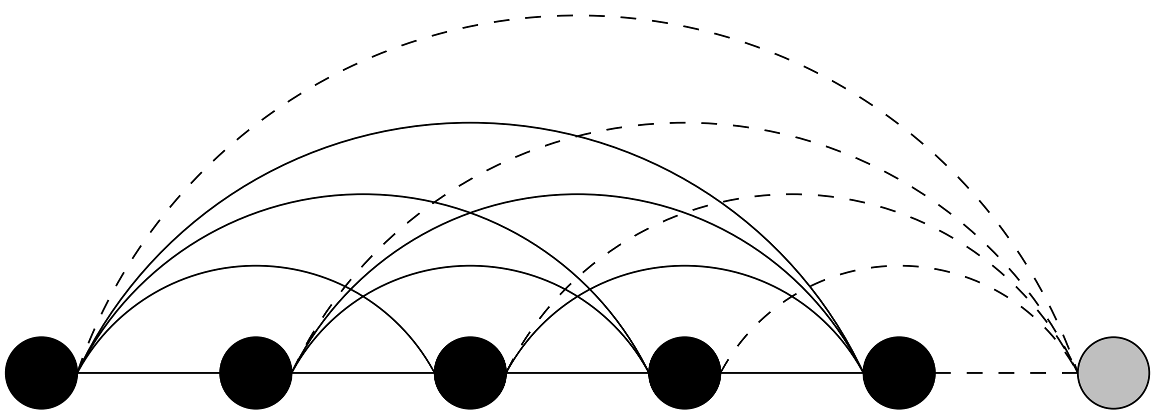
- A **long edge** is not short.

- A long edge **spans** a short edge if it starts at the same marker or sooner and ends at the same marker or later.

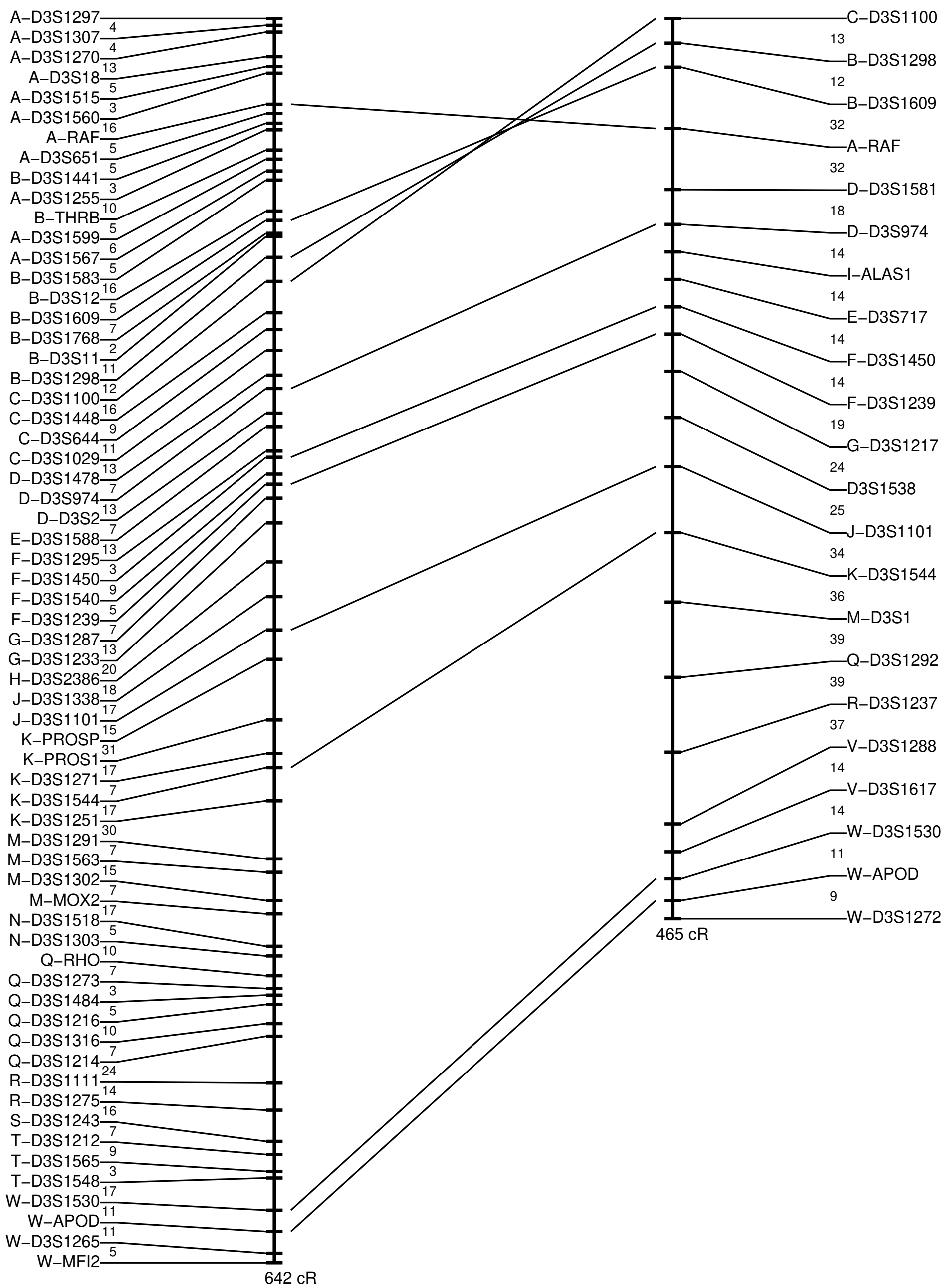
- A long edge is **consistent** if it's weight is greater than those weights of all the short edges it spans.

- A graph is consistent if all it's long edges are consistent.

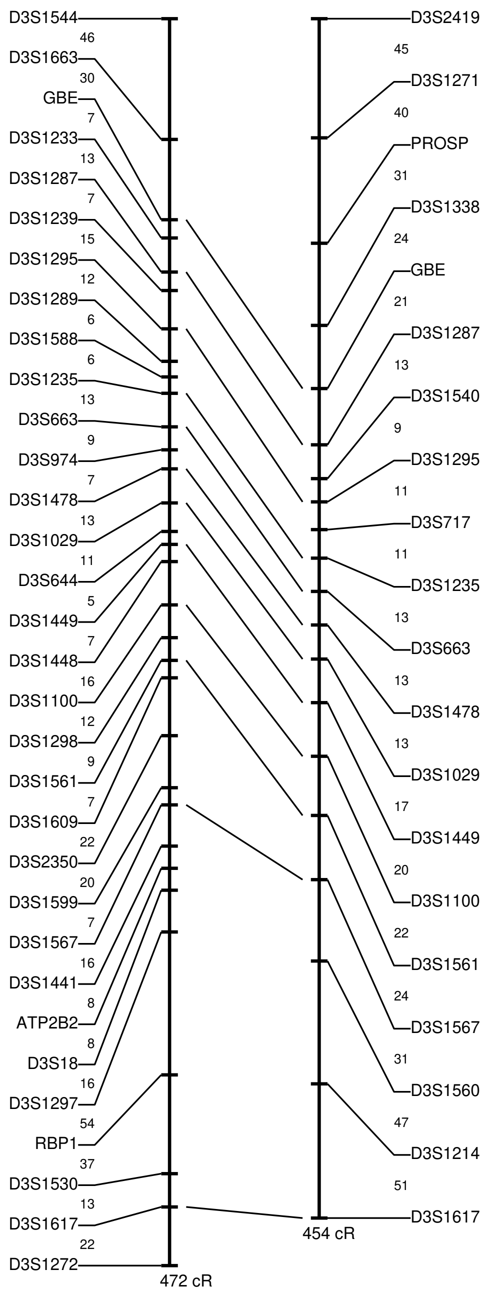
Growing a Consistent Map



Map Comparison of Hand-Selected Markers and Consistent Map



Map Comparison of RHMAPPER Map and Consistent Map



Summary

- Formalizing biological problems may provide new approaches to solving them.

- Most biologically relevant problems remain intractable.

- Wclique is an effective way of screening markers for genetic framework mapping. It speeds the initial phases of map construction.

- Consistent maps deserve more study, the techniques may be extended to construct high quality maps.

References

1. Applegate D and Johnson DS (1993). `dfmax`, A computer program for the maximum clique problem. Available via anonymous ftp from: `<URL:ftp://dimacs.rutgers.edu/pub/challenge/graph/solvers/>`.

2. Ben-Dor A and Chor B (1997). On constructing radiation hybrid maps. *Proceedings of the First Annual International Conference on Computational Molecular Biology (RECOMB)*, pages 17–26.

3. Boehnke M, Lange K, Cox DR (1991). Statistical methods for multipoint radiation hybrid mapping. *American Journal of Human Genetics*, **49**:1174–1188.

4. Stein L, Kruglyak L, Slonim D and Lander E. (1995) "RHMAPPER", unpublished software, Whitehead Institute/MIT Center for Genome Research. Available at: `<URL:http://www.genome.wi.mit.edu/ftp/pub/software/rhmapper/>`.

5. Thompson EA (1987). Crossover counts and likelihood in multipoint linkage analysis. *IMA Journal of Mathematics Applied in Medicine and Biology*, **4**:93–108.