A Survey of Bioinformatics

Humberto Ortiz Zuazaga

humberto@hpcf.upr.edu

http://www.hpcf.upr.edu/~humberto/

Bioinformatics

The application of computers to the collection, analysis, and presentation of biological information.

Taxonomy (of the talk)

- Sequence analysis
- Structure prediction
- Genetic and physical mapping
- Gene expression networks

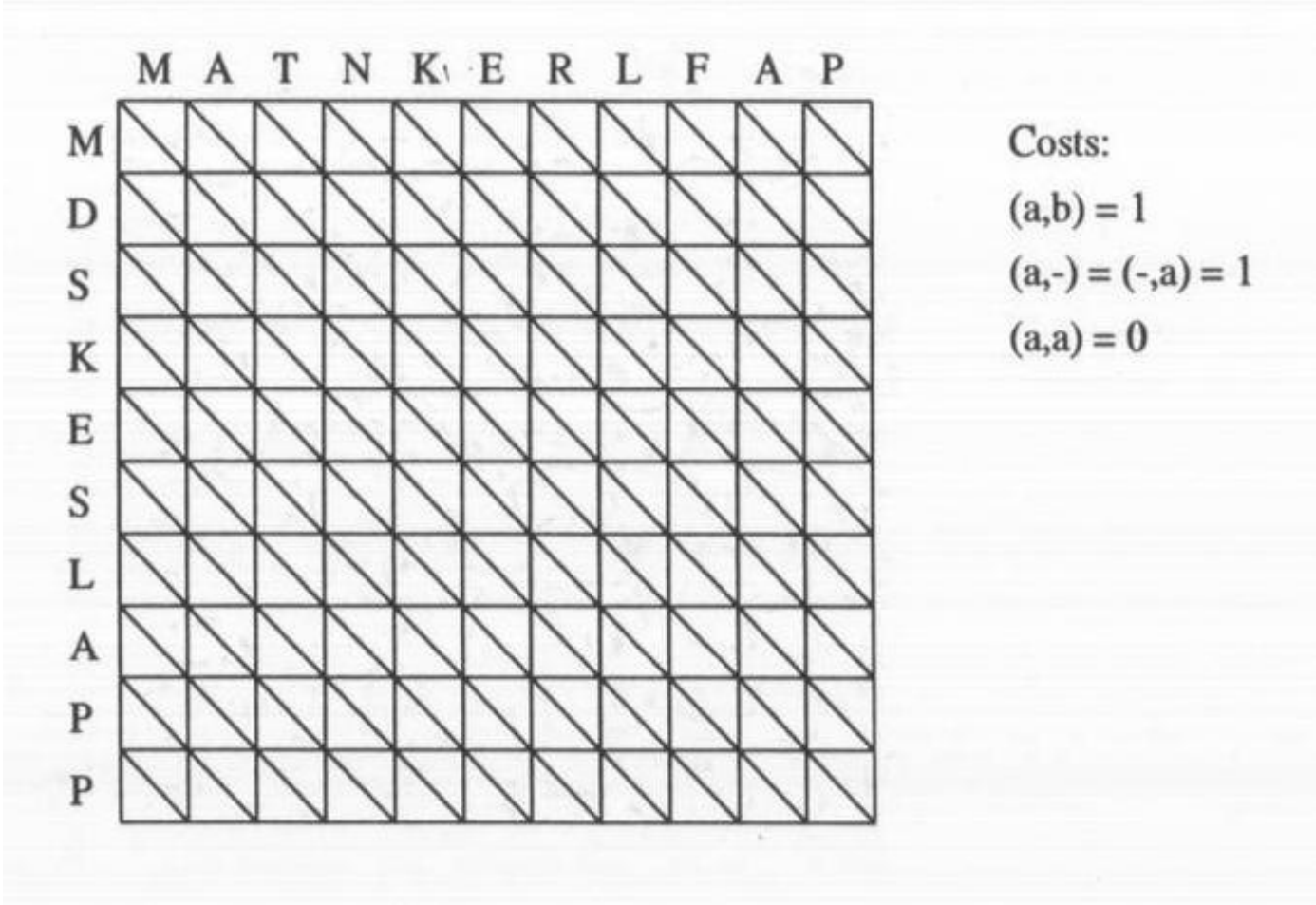
Pairwise sequence alignments

Given a pair of sequences over an alphabet Σ, and a cost function that assigns a cost to an alignment, find an alignment of the strings that minimizes the cost.

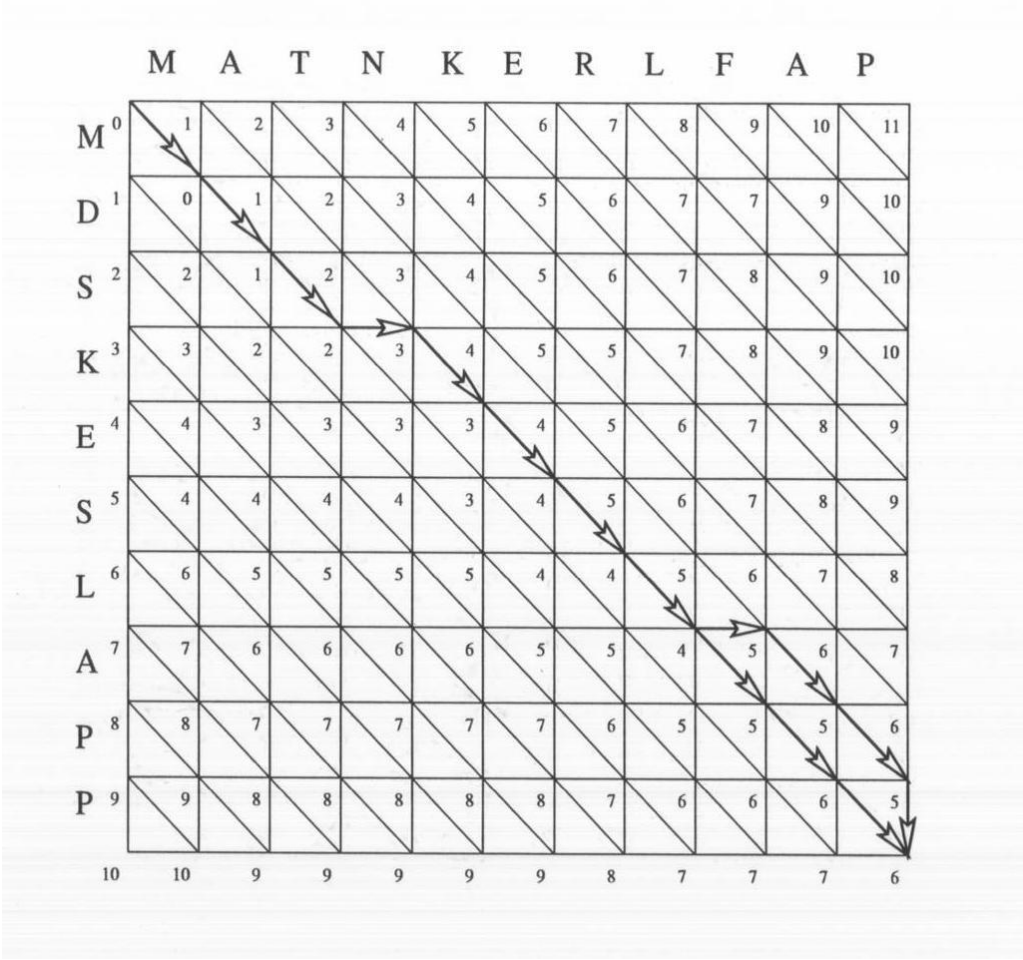
- This problem is central to many biological programs (homology searches, multiple alignments, molecular phylogeny, exon prediction, protein threading, ...)

- Usually solved by Dynamic Programming.

Alignment dynamic programming table



The solved alignment



References

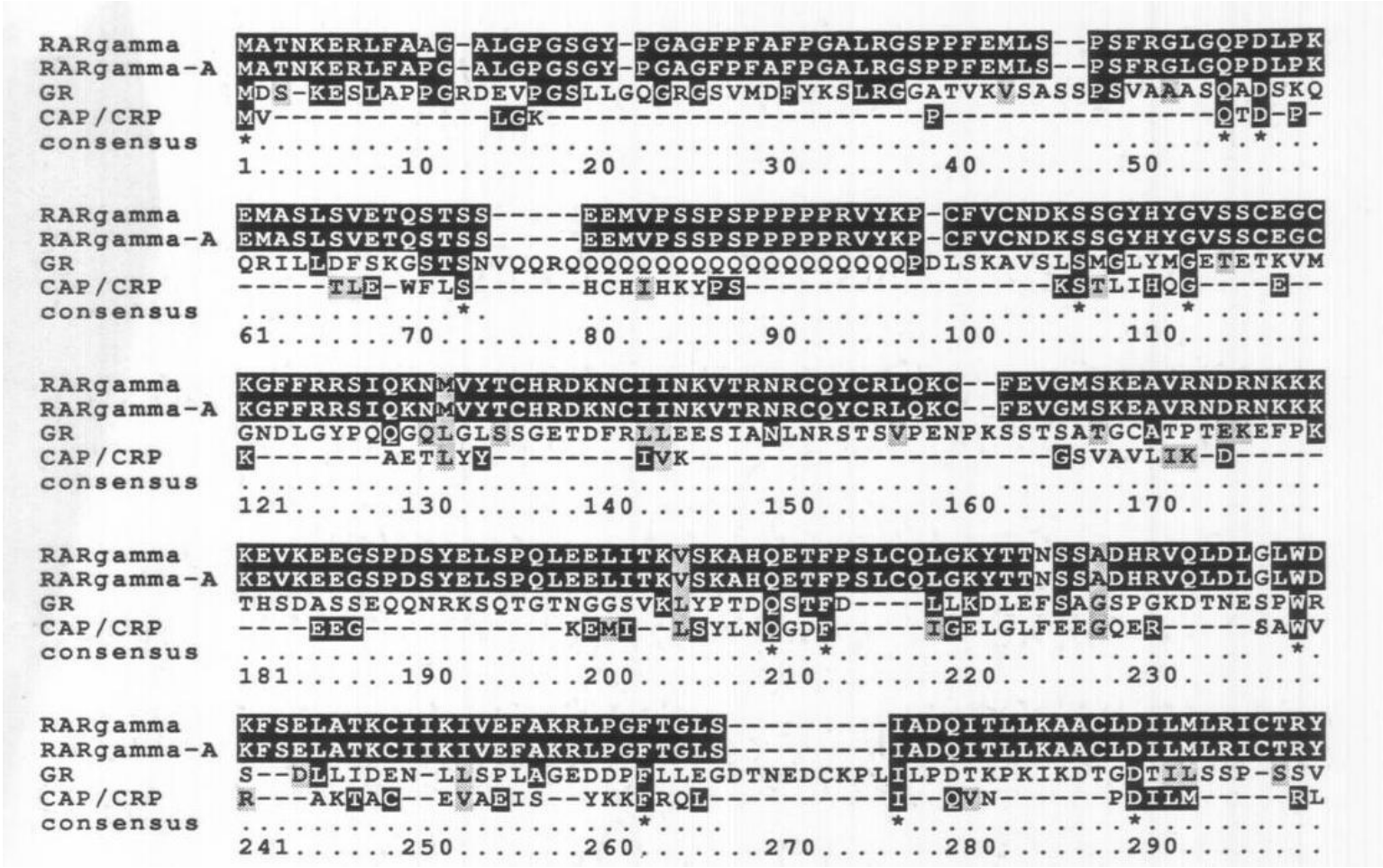
VSNS BCD Pairwise Alignments chapter.

`http://www.techfak.uni-bielefeld.de/bcd/Curric/PrwAli/prwali.html`

Multiple sequence alignments

- Assign a cost to alignments of multiple sequences.
- Sum of pairs metric, add all pairwise alignments together.
- NP-complete for the SP metric.
- Other heuristics must be used for practical programs (Divide and Conquer, HMM's).

A sample multiple alignment



References

VSNS BCD Multiple Alignment chapter

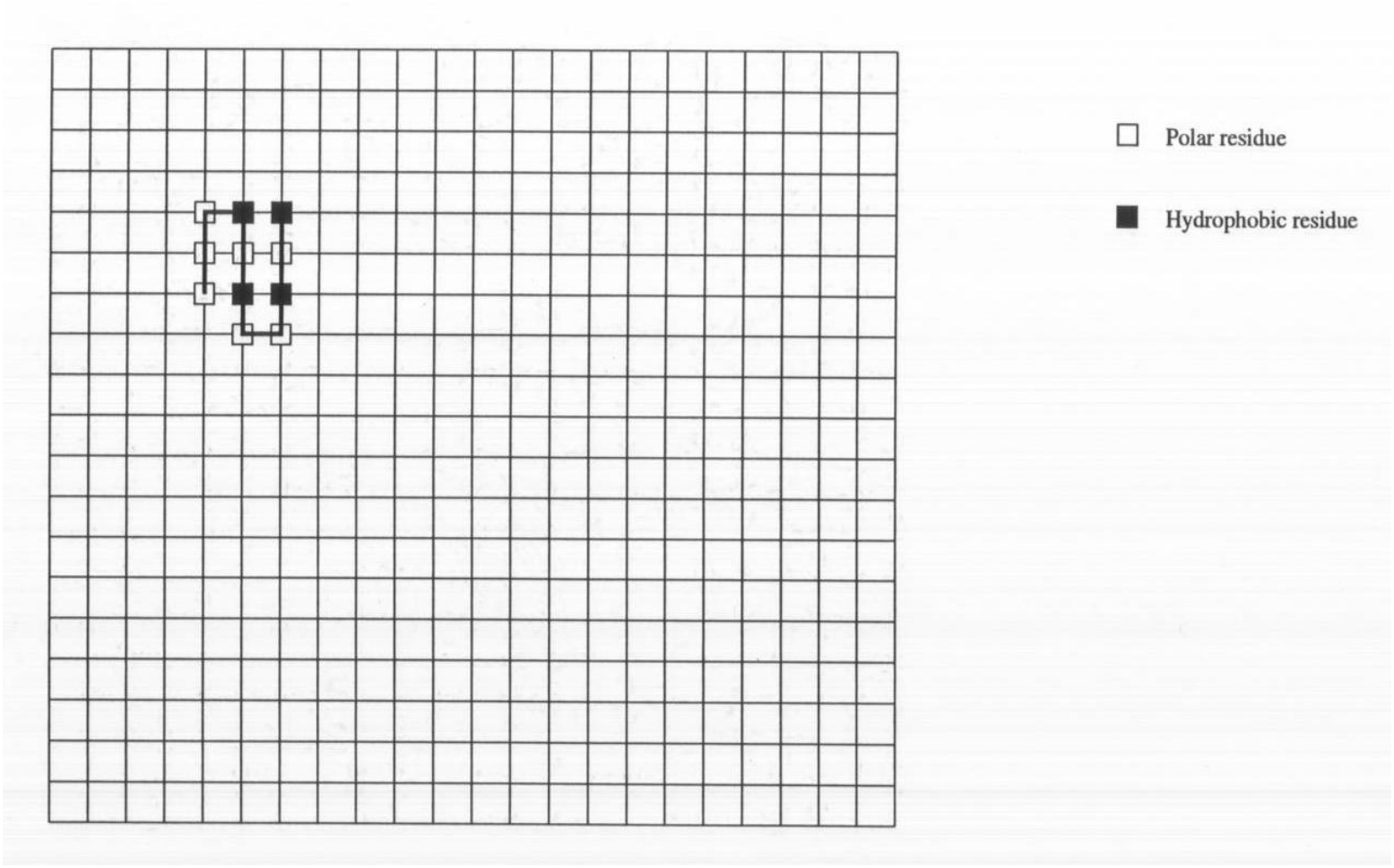
`http://www.techfak.uni-bielefeld.de/bcd/Curric/MulAli/mulali.html`

The H-P model of globular proteins

Simple model first postulated by Ken Dill [1].

- Amino acids are hydrophobic (H, nonpolar) or hydrophilic (P, polar).
- H-H contacts contribute -1 to the energy of the protein.
- all other contacts contribute 0.
- Protein structures are constrained to self-avoiding paths on a regular lattice.

H-P proteins on a regular lattice



The Protein Folding problem

Given a sequence s and an integer E, is there a fold that has $-E$ or lower energy?

- Has been shown to be NP-complete in 2D (HAMILTONIAN PATH) [2]
- MAXSNP-complete in 3D [2]
- Can be approximated to 3/8 of optimal in linear time in 3D [3]

The Inverse Protein Folding (IFP) problem

Given a target structure or conformation of a protein G, find a sequence s of length n that:

- Has G as is it's minimum energy state.

- Has the lowest degeneracy (number of other conformations with the same energy) of any possible sequence.

The Heuristic Sequence Design (HSD) problem

IFP is conjectured to be NP-complete, the best known algorithm must search over all possible conformations of all possible sequences. HSD problems try to simplify the computation by restricting the problem.

- The Canonical Method: find the sequence with at most λn hydrophobic residues.

- The Grand Canonical Method: Change the energy function so that H-H contacts have -2, an exposed H residue has 1, and all other interactions have 0.

Computational Results

- The Canonical Method is NP-complete (reduction from SUBSET-SUM), but there are algorithms that can approximate the energy of the optimal solution (OPT) to 1+OPT on 2D lattices, and 1/2 OPT on 3D lattices [4].

- The Grand Canonical Method can be solved in polynomial time.

References

1. Ken Dill. Dominant forces in protein folding. *Biochemistry* 29, pages 7133–7155, 1990.

2. P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, M. Yannakakis, On the Complexity of Protein Folding. In Proc. of the Second Annual Conference of Computational Biology (RECOMB '98).
`http://citeseer.nj.nec.com/31865.html`

3. Hart, W. and Istrail, S. Fast protein folding in the hydrophobic-hydrophilic model within three eighths of optimal. In Proceedings of the 27th Annual ACM Symposium on the Theory of Computing. 1995.
`http://citeceer.nj.nec.com/hart95fast.html`

4. William Hart. On the computational complexity of sequence design problems. In First Annual International Conference on Computational Molecular Biology (RECOMB'97), pages 128–136, 1997.
`http://citeseer.nj.nec.com/hart97computational.html`

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.

A genotyped pedigree

```
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
```

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).

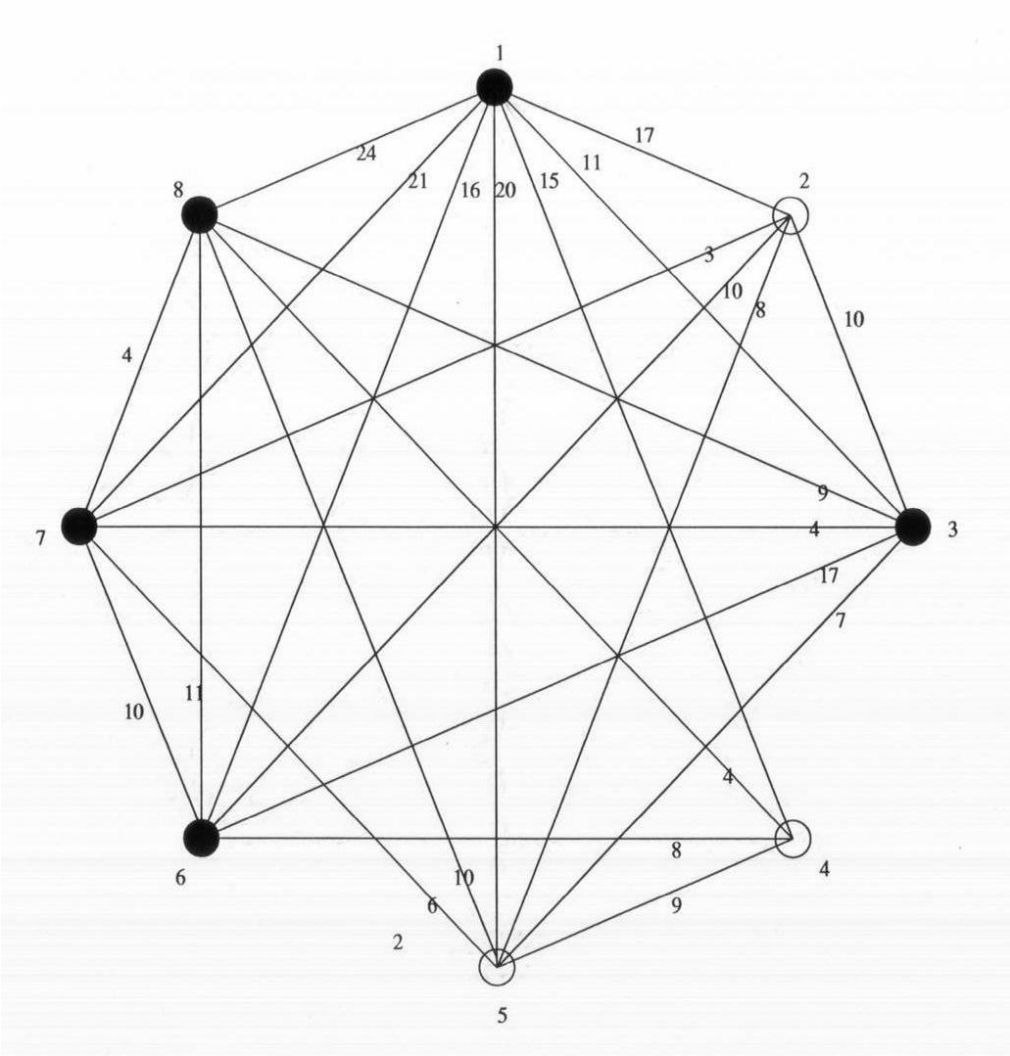
Selecting Genetic Markers With `wclique`

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



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.

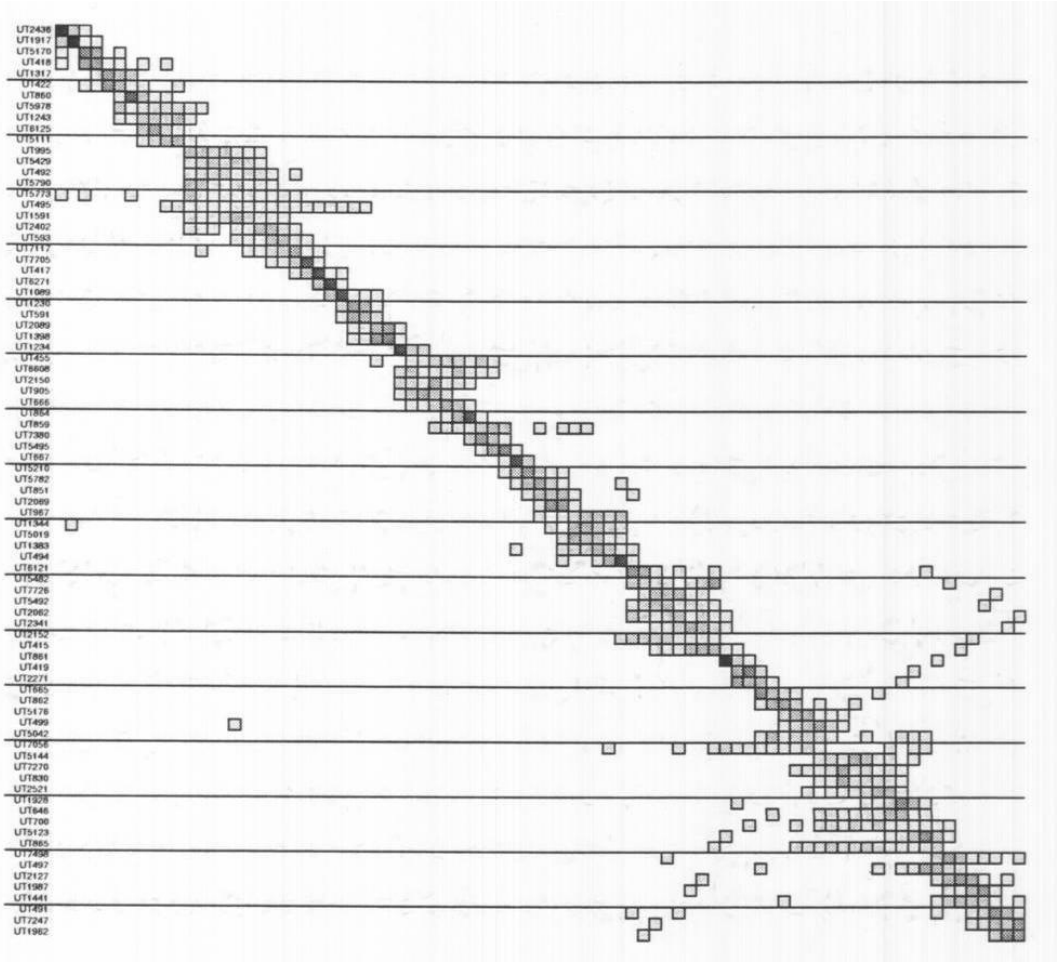
But I Still Need a Map

- Exact algorithms can work on small sets of markers.

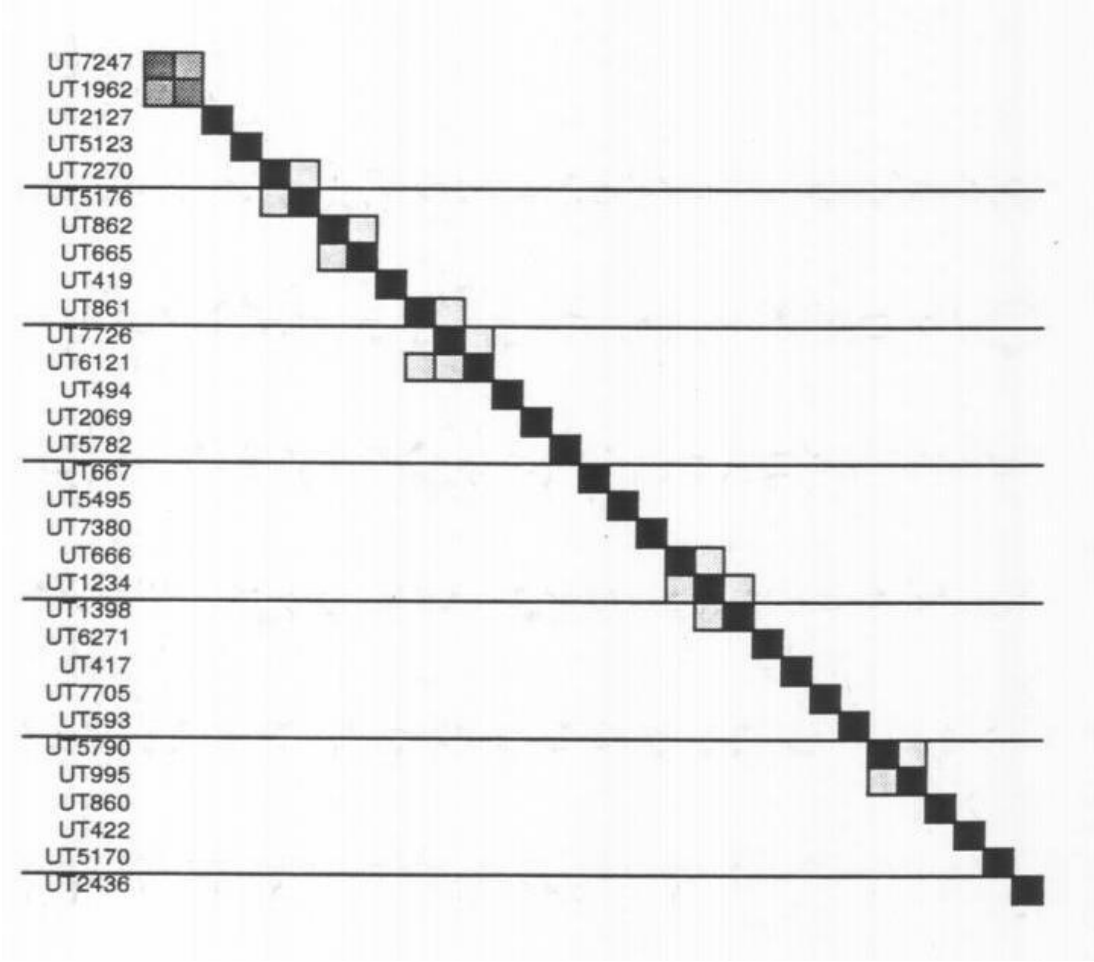
- Local search techniques can find near optimal solutions for some of these problems, at the cost of not knowing if an optimal solution was ever found. The best heuristics for TSP can find a solution with 1.05 times the optimal cost.

- 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.

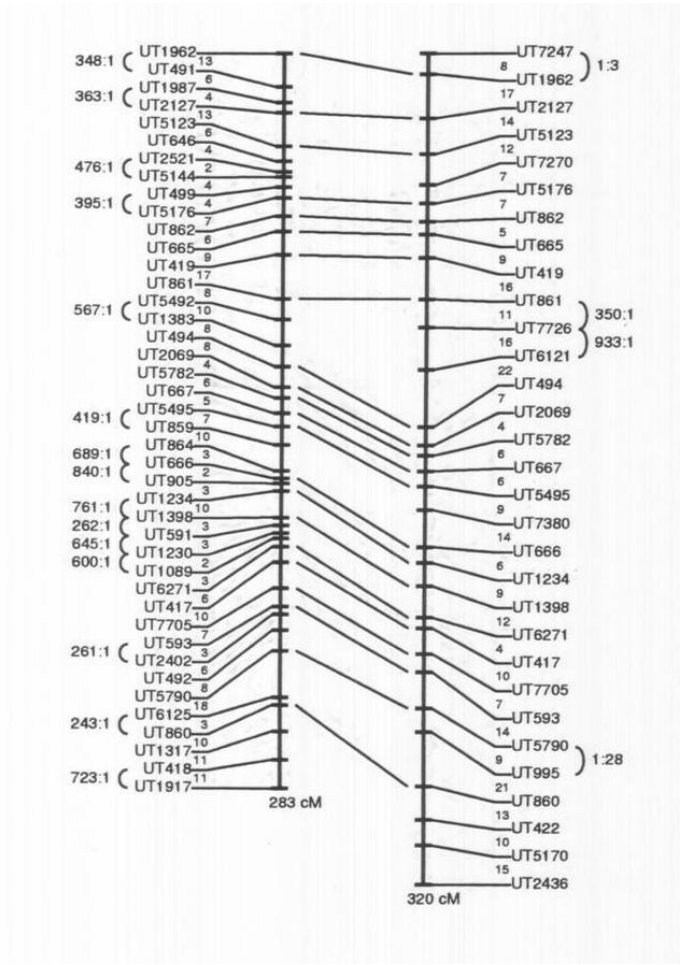
Searls plot of unselected markers



Searls plot of `wclique`-selected markers



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



References

1. H. Ortiz-Zuazaga, and R. Plaetke. Screening genetic markers with the maximum weighted clique method. Abstract presented at Genome Mapping and Sequencing. Cold Spring Harbor, May 1997.

2. S.L. Naylor, R. Plaetke, H. Ortiz-Zuazaga, P. O'Connell, B. Reus, X. He, R. Linn, S. Wood, and R.J. Leach. Construction of Framework and Radiation Hybrid Maps of Chromosomes 3 and 8. Abstract presented at Genome Mapping and Sequencing. Cold Spring Harbor, NY, May 1997.

Gene expression networks

- Complete genomes available for several species.
- 40,000 human genes, many already sequenced.
- microarrays can measure expression levels for ALL GENES in a single assay.

Boolean Genetic Network Model

In [2] we define Boolean genetic network model (BGNM):

- A *Boolean variable* takes the values 0, 1.
- A *Boolean function* is a function of Boolean variables, using the operations $\wedge$, $\vee$, $\neg$.

A *Boolean genetic network model* (BGNM) is:

- An n-tuple of Boolean variables $(x_1, \ldots, x_n)$ associated with the genes
- An n-tuple of Boolean control functions $(f_1, \ldots, f_n)$, describing how the genes are regulated

Results on Boolean Networks

- Determining if a given assignment to all the variables is consistent with a given gene network was shown to be NP-complete in [1] (by reduction from 3-SAT).

- In the worst case, $2^{(n-1)/2}$ experiments are needed

- If the indegree of each node (the genes that affect our target gene) is bound by a constant D, the cost is $O(n^{2D})$.

- For low D, [2] and [3] provide effective procedures for reverse engineering, assuming any gene may be set to any value.

Reverse Engineering Boolean Networks

1. Akutsu, S. Kuahara, T. Maruyama, O. Miyano, S. 1998. Identification of gene regulatory networks by strategic gene disruptions and gene overexpressions. Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms (SODA 98), H. Karloff, ed. ACM Press.

2. Ideker, T.E., Thorsson, V., and Karp, R.M. 2000. Discovery of regulatory interactions through perturbation: inference and experimental design. Pacific Symposium on Biocomputing 5:302-313.

3. S. Liang, S. Fuhrman and R. Somogyi. 1998. REVEAL, A General Reverse Engineering Algorithm for Inference of Genetic Network Architectures. Pacific Symposium on Biocomputing 3:18-29.

The World's Smallest Finite Field

The integers 0 and 1, with integer addition and multiplication modulo 2 form the finite field $Z_2 = \{\{0, 1\}, +, \cdot\}$.

The operators $+$ and $\cdot$ are defined as follows:

$+$	0	1
0	0	1
1	1	0

$\cdot$	0	1
0	0	0
1	0	1

Finite field equivalents to the Boolean operators

We can realize any Boolean function as an expression over Z_2:

$$\begin{aligned} X \wedge Y &= X \cdot Y \\ X \vee Y &= X + Y + X \cdot Y \\ \neg X &= 1 + X \end{aligned}$$

Finite Field Genetic Networks

Any BGNM can be converted into an equivalent model over Z_2 by realizing the Boolean functions as sums-of-products and products-of-sums, then converting the Booleans to Z_2. We now have a *finite field genetic network* (FFGN):

- An n-tuple of variables over Z_2, $(x_1, \ldots, x_n)$ associated with the genes

- An n-tuple of functions over Z_2, $(f_1, \ldots, f_n)$, describing how the genes are regulated

Publications

1. Ortiz-Zuazaga, H., Aviño-Diaz, M. A., Laubenbacher, R., Moreno O. Finite fields are better Booleans. Refereed abstract, poster to be presented at the Seventh Annual Conference on Computational Molecular Biology (RECOMB 2003), April 10–13, 2003, Germany.

2. Ortiz-Zuazaga, H., Aviño-Diaz, M. A., Corrada Bravo, C. J., Laubenbacher, R., Peña-de-Ortiz, S., Moreno O. Applications of finite fields to the study of microarray expression data. Submitted to the 11th International Conference on Intelligent Systems for Molecular Biology (ISMB 2003), June 29 to July 3, 2003, Brisbane, Australia.