Adventures in Bioinformatics and Computational Biology

Nathan Edwards

Informatics Research

Applied Biosystems

Outline

- Bioinformatics and Computational Biology is fun!
- Interesting problems from
  - Proteomics
  - Haplotypes
  - Multiplexed Assay Design

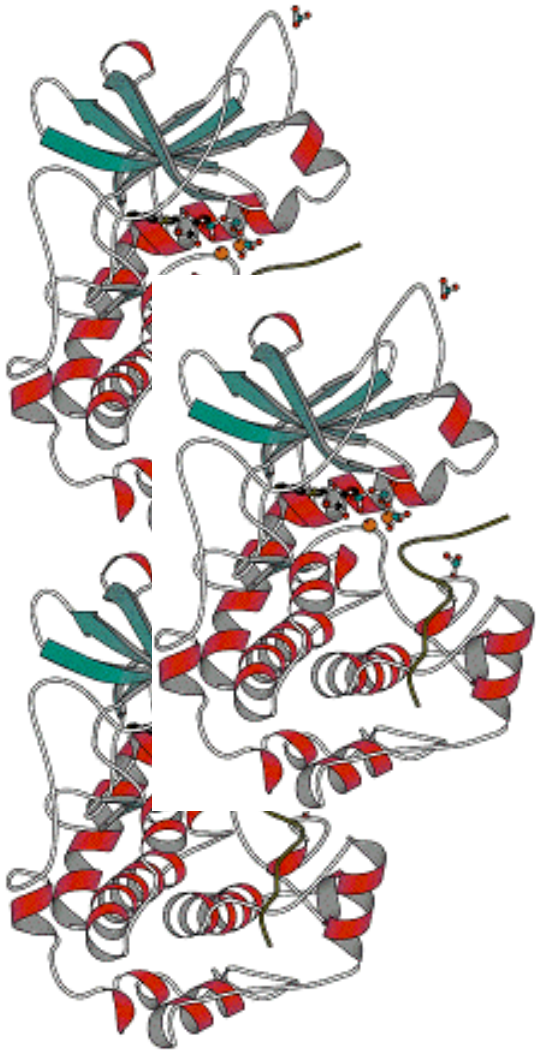
Computational Biology is Fun!

- Novel, yet-to-be-modeled problems
  - Lots of modeling tradeoffs
- Empirical performance matters
  - Sometimes “simple” works really well
  - The instances are big
  - Creative amortization pays
- Lots of (public) data to play with
- Lots of (biology) material to learn
- Lots of (grant) money to be had
- People care about the results

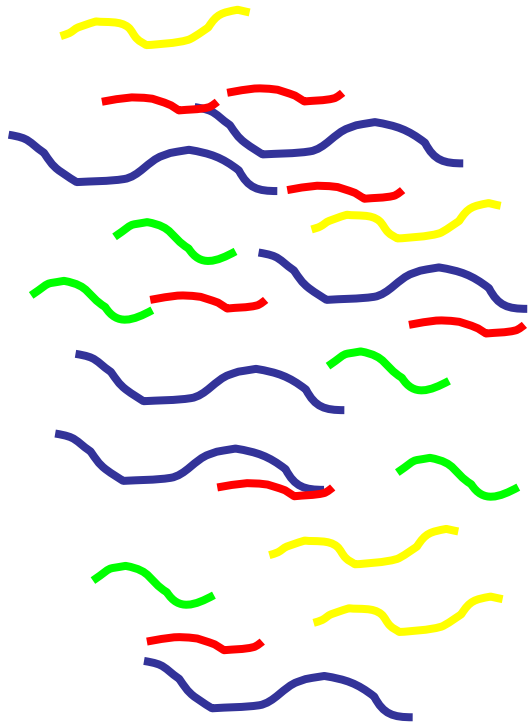
Proteomics

- Study of observed proteins
  - How much of each?
  - What protein is it?
- Usually a combination of
  - Wet-lab biological sample manipulation
  - Mass spectrometry
  - Data analysis

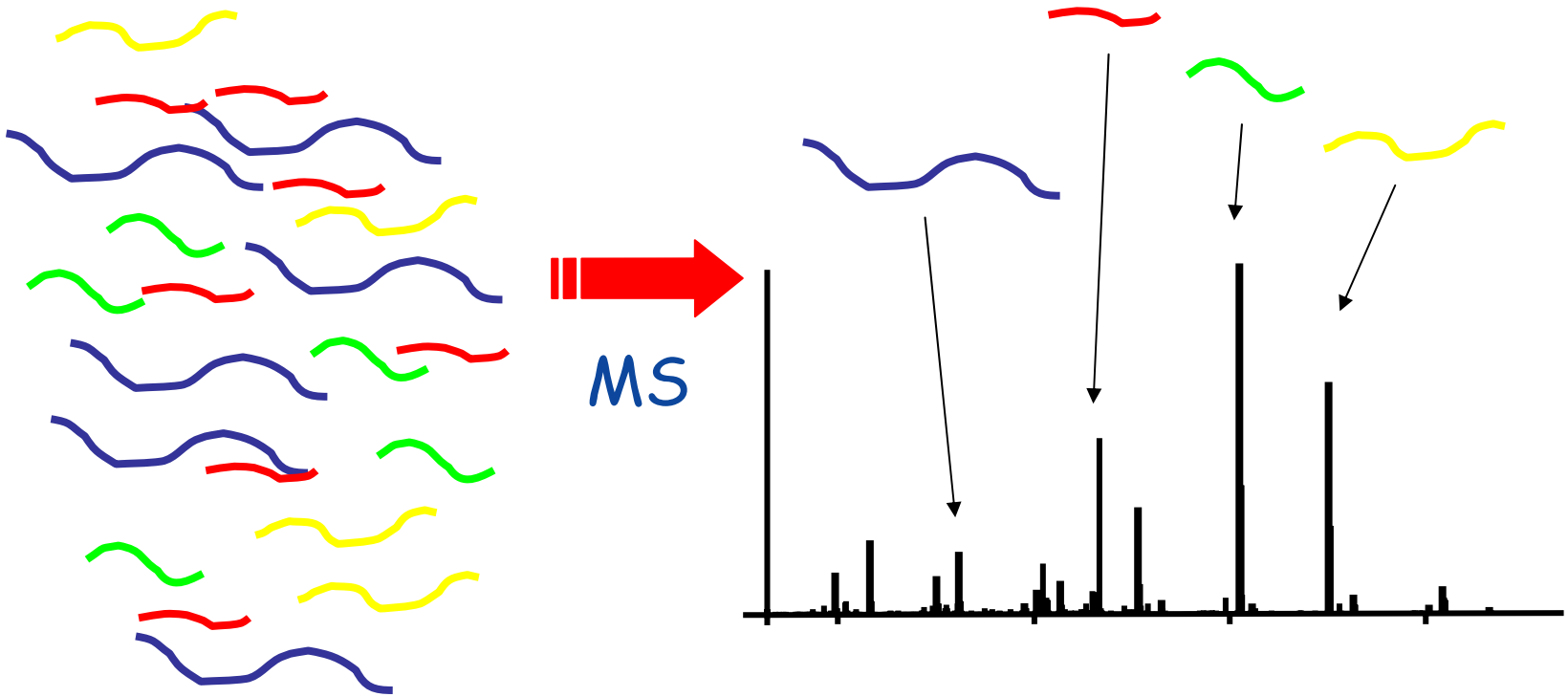
Sample Preparation



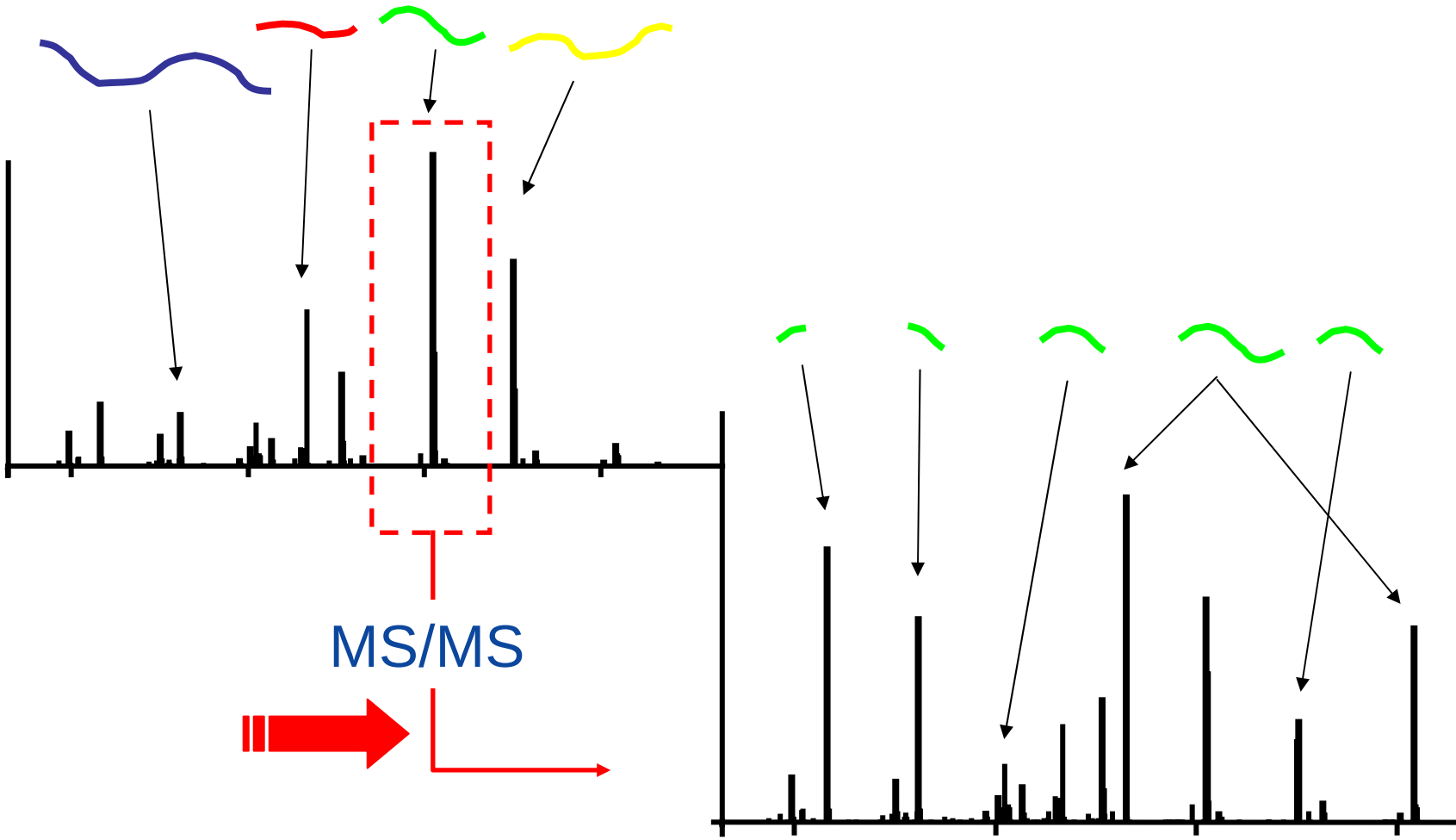
Enzymatic Digest
and
Fractionation



(Single Stage) Mass Spectrometry



Tandem Mass Spectrometry (MS/MS)



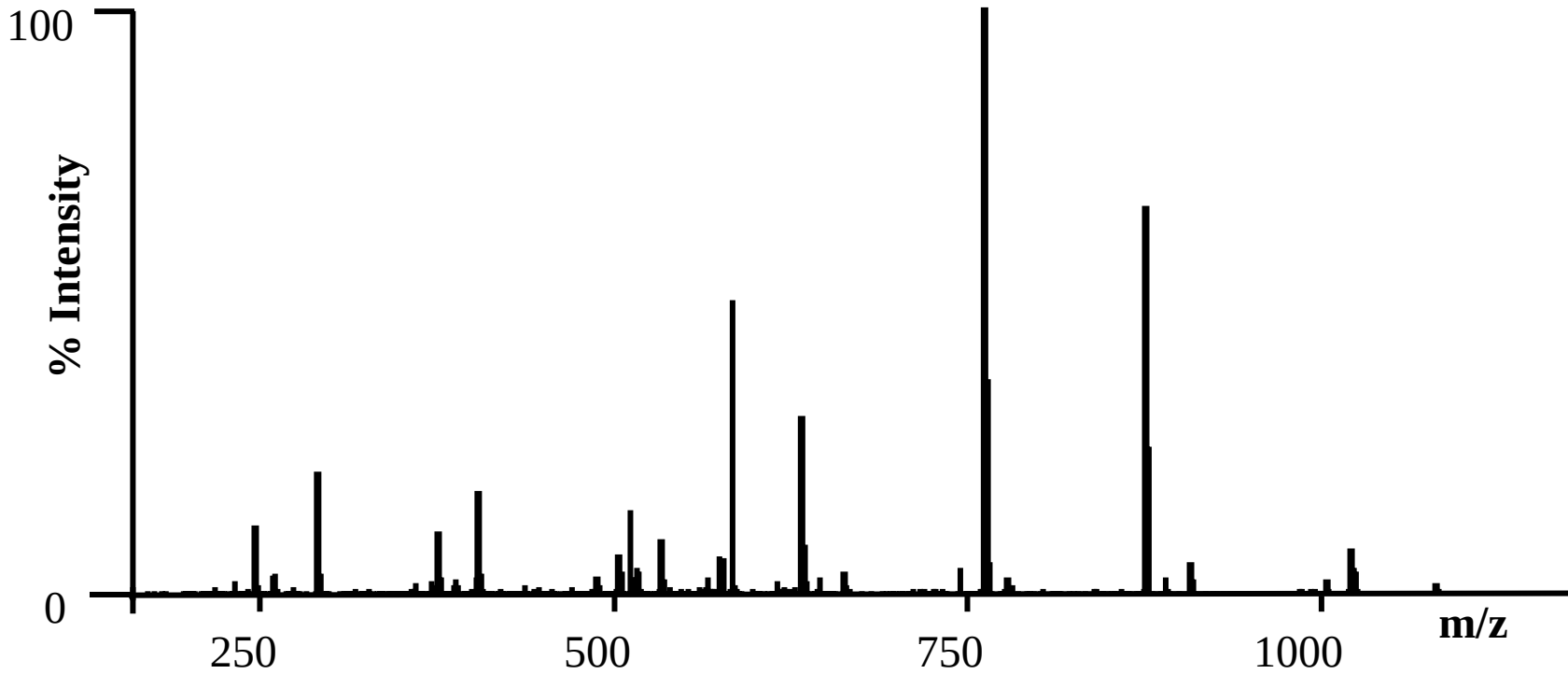
Peptide Fragmentation

Peptide: S-G-F-L-E-E-D-E-L-K

MW	ion			ion	MW
88	b_1	S	GFLEEDELK	y_9	1080
145	b_2	SG	FLEEDELK	y_8	1022
292	b_3	SGF	LEEDELK	y_7	875
405	b_4	SGFL	EEDELK	y_6	762
534	b_5	SGFLE	EDELK	y_5	633
663	b_6	SGFLEE	DELK	y_4	504
778	b_7	SGFLEED	ELK	y_3	389
907	b_8	SGFLEEDE	LK	y_2	260
1020	b_9	SGFLEEDEL	K	y_1	147

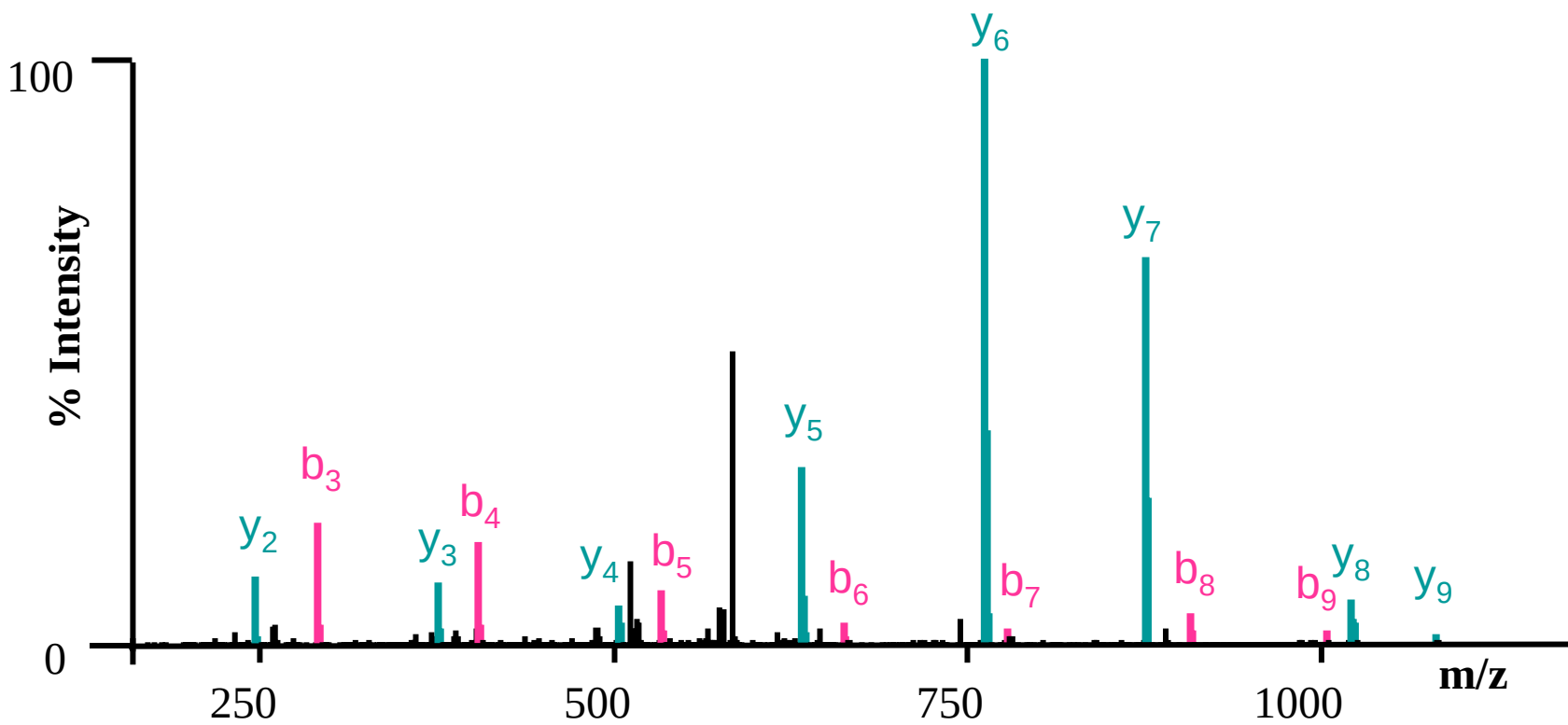
Peptide Fragmentation

88	145	292	405	534	663	778	907	1020	1166	b ions
S	G	F	L	E	E	D	E	L	K	
1166	1080	1022	875	762	633	504	389	260	147	y ions



Peptide Fragmentation

88	145	292	405	534	663	778	907	1020	1166	b ions
S	G	F	L	E	E	D	E	L	K	
1166	1080	1022	875	762	633	504	389	260	147	y ions



Peptide Identification

Given:

- The mass of the parent ion, and
- The MS/MS spectrum

Output:

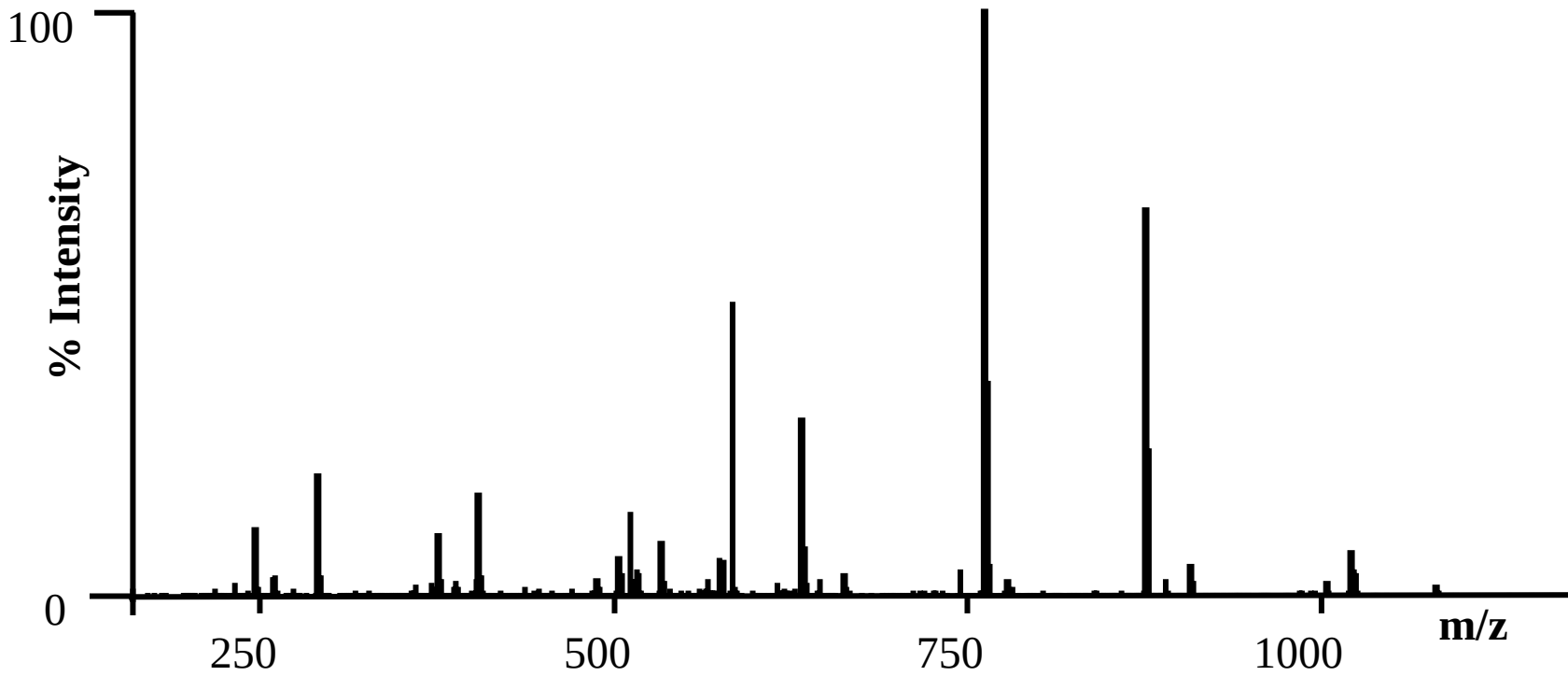
- The amino-acid sequence of the peptide

Peptide Identification

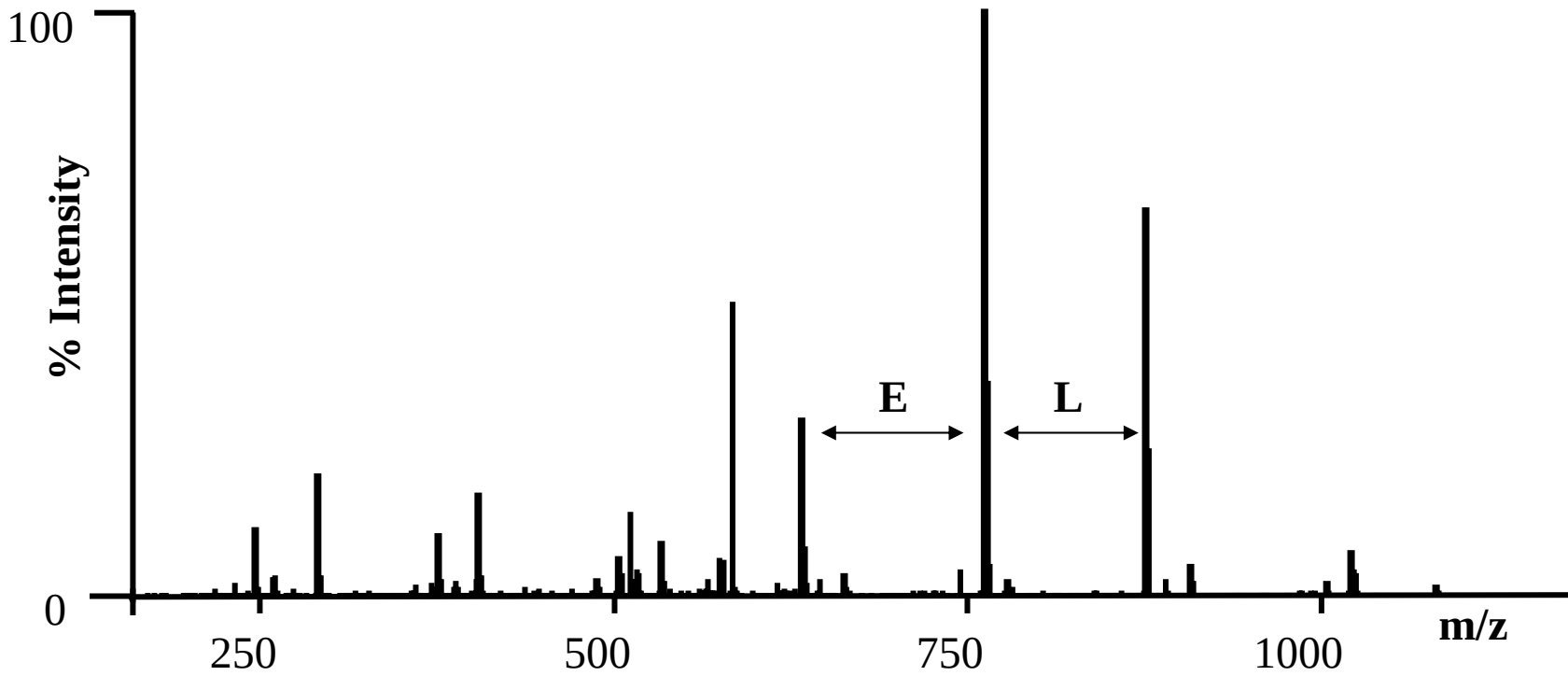
Two paradigms:

- *De novo* interpretation
- Sequence database search

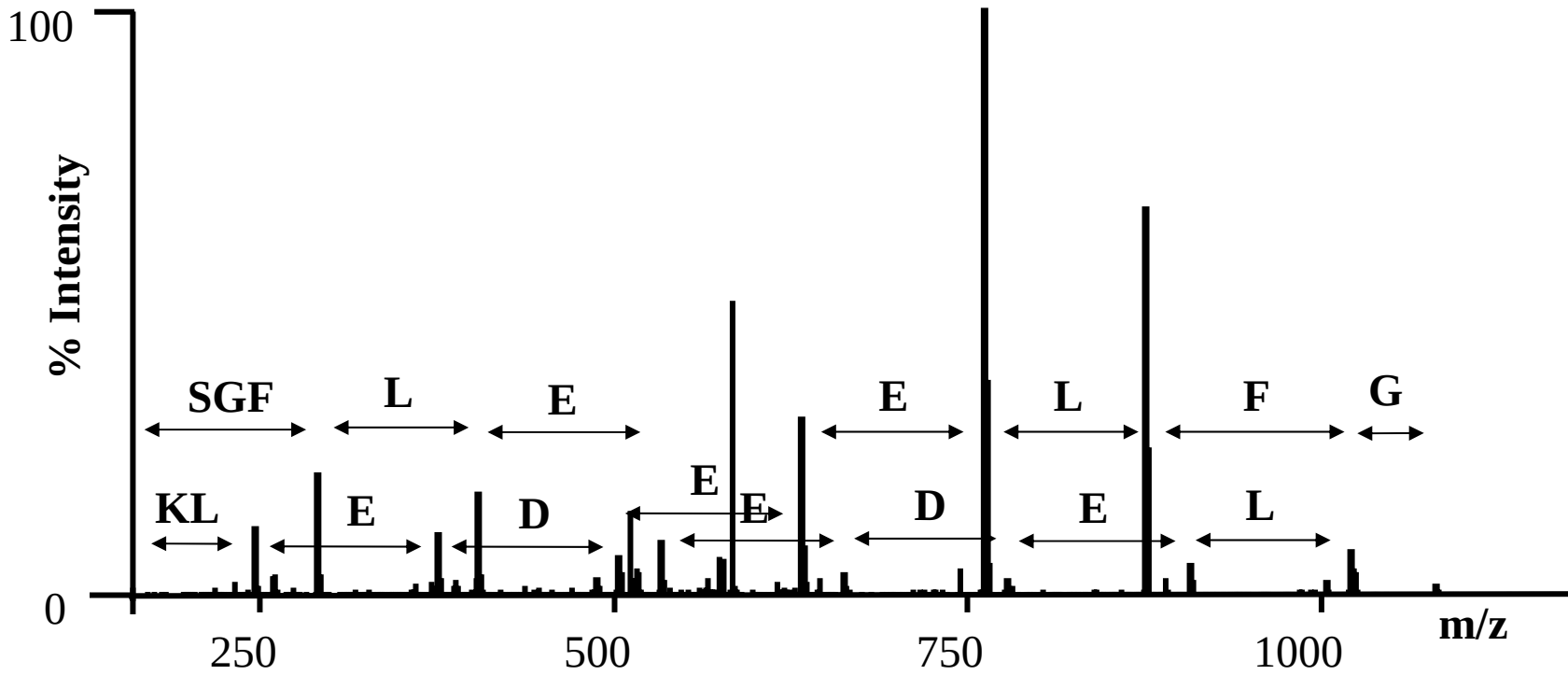
De Novo Interpretation



De Novo Interpretation



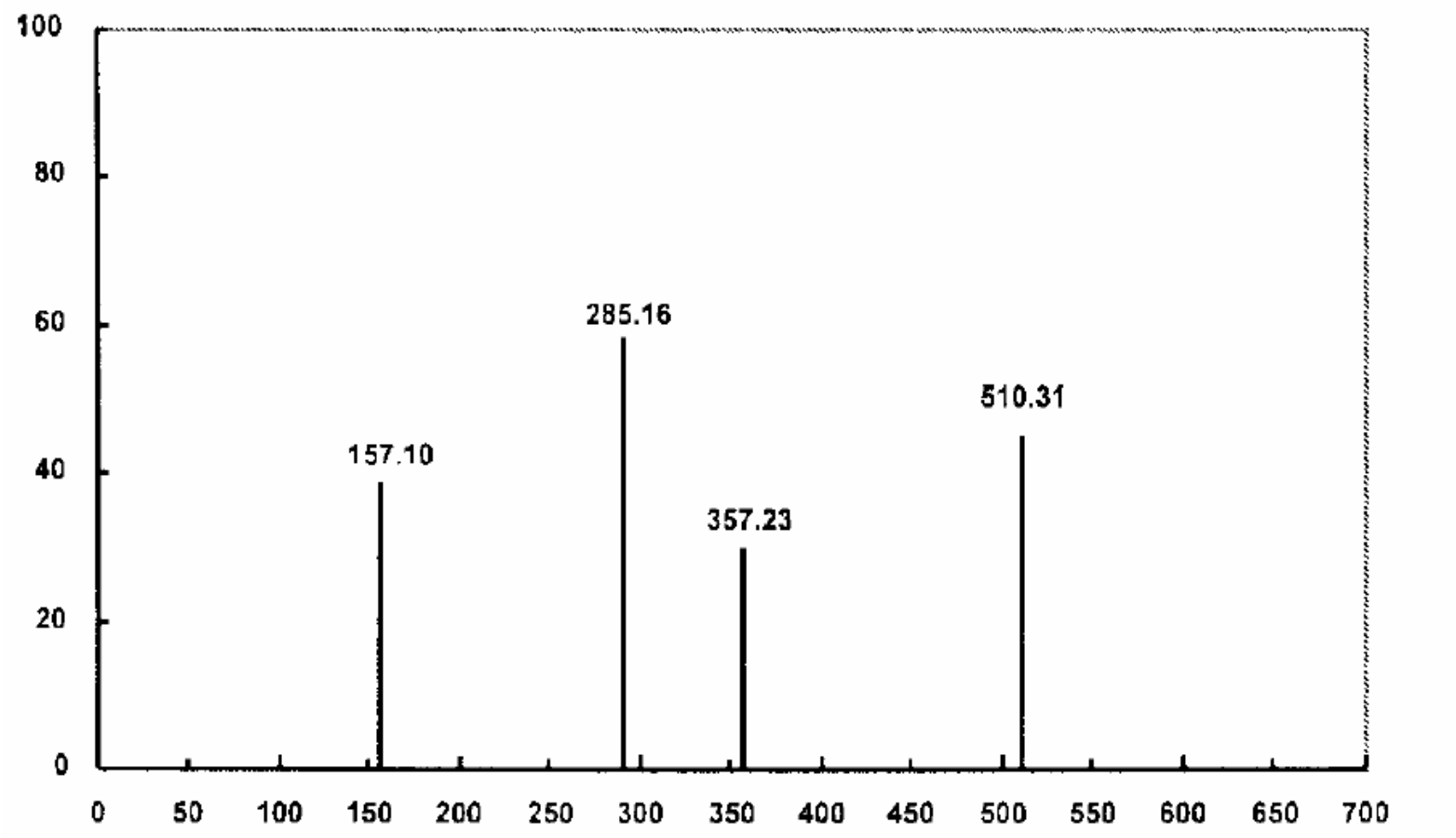
De Novo Interpretation



De Novo Interpretation

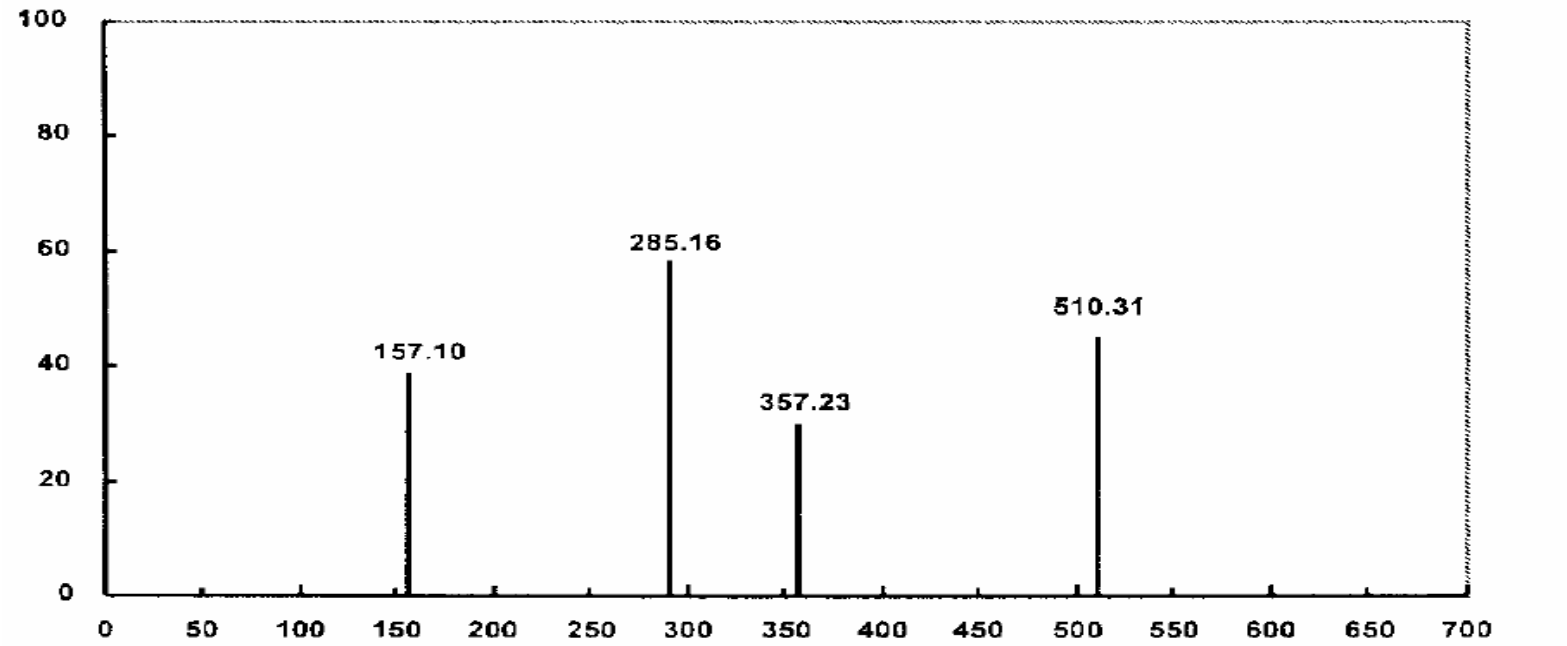
	Amino-Acid	Residual MW		Amino-Acid	Residual MW
A	Alanine	71.03712	M	Methionine	131.04049
C	Cysteine	103.00919	N	Asparagine	114.04293
D	Aspartic acid	115.02695	P	Proline	97.05277
E	Glutamic acid	129.04260	Q	Glutamine	128.05858
F	Phenylalanine	147.06842	R	Arginine	156.10112
G	Glycine	57.02147	S	Serine	87.03203
H	Histidine	137.05891	T	Threonine	101.04768
I	Isoleucine	113.08407	V	Valine	99.06842
K	Lysine	128.09497	W	Tryptophan	186.07932
L	Leucine	113.08407	Y	Tyrosine	163.06333

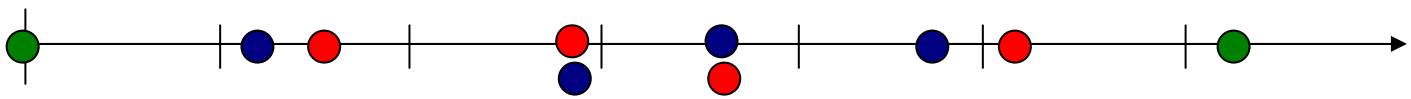
De Novo Interpretation



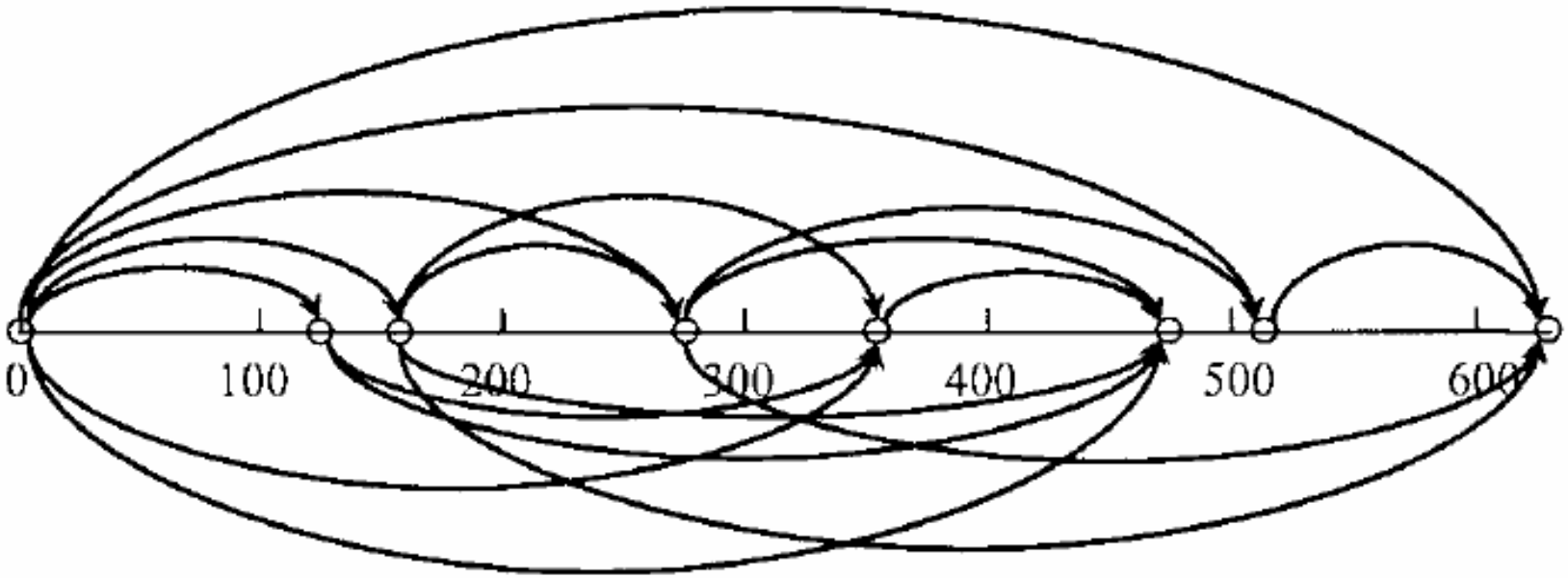
…from Lu and Chen (2003), JCB 10:1

De Novo Interpretation





De Novo Interpretation



…from Lu and Chen (2003), JCB 10:1

De Novo Interpretation

- Find *good* paths in spectrum graph
- Can't use same peak twice
  - Forbidden pairs: NP-hard
  - "Nested" forbidden pairs: Dynamic Prog.
- Simple peptide fragmentation model
- Usually many apparently good solutions
- Needs better fragmentation model
- Needs better path scoring

Sequence Database Search

- Compares peptides from a protein sequence database with spectra
- Filter peptide candidates by
  - Parent mass
  - Digest motif
- Score each peptide against spectrum
  - Generate all possible peptide fragments
  - Match putative fragments with peaks
  - Score and rank

Simple Linear Scan

Parent Mass = 2018.07

MKWVTFISLLFLFSSAYSRGV…

Output: WVTFISLLFLFSSAYSR

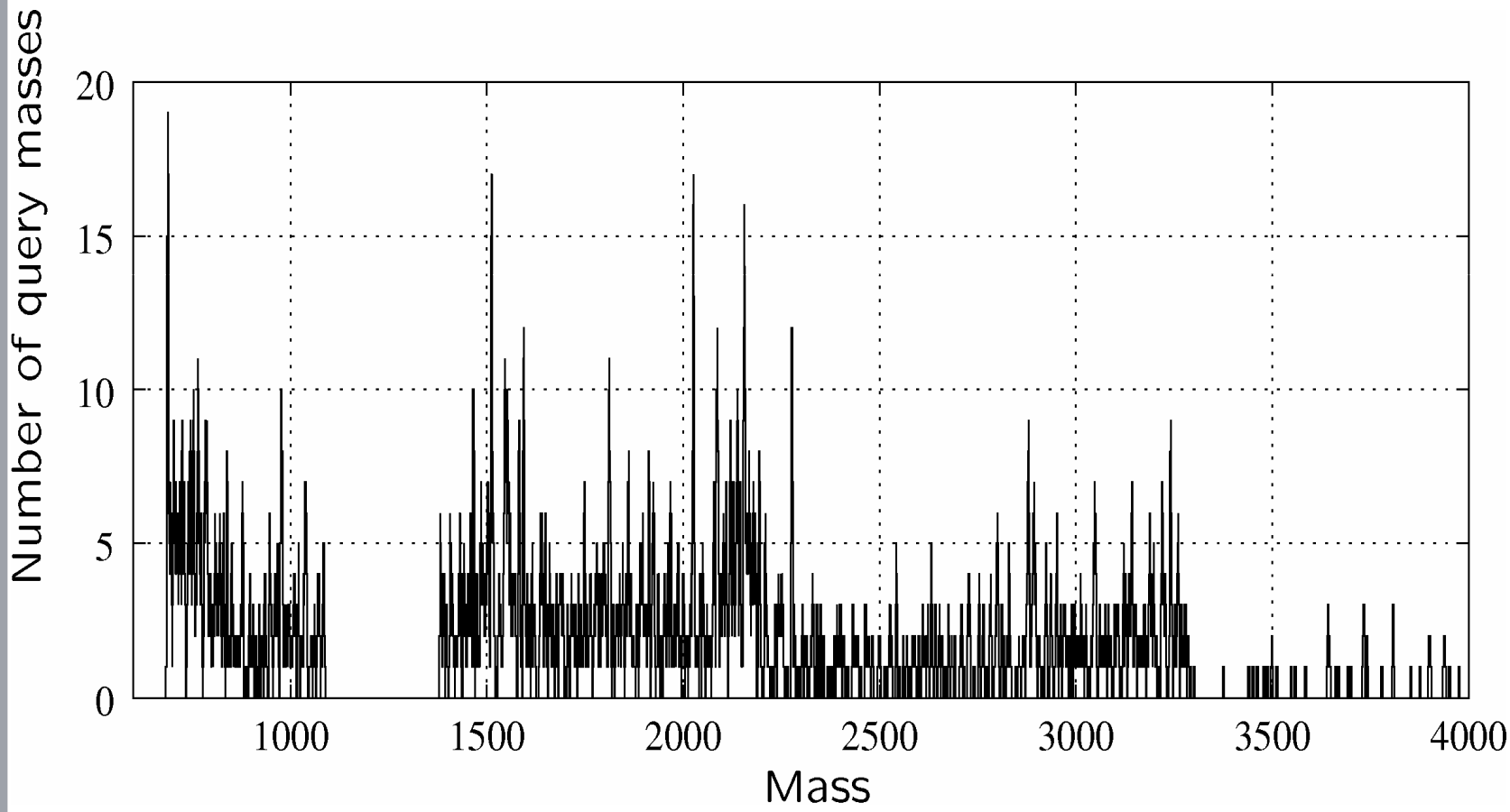
Simultaneous Linear Scan

Max Query Mass = 2018.07

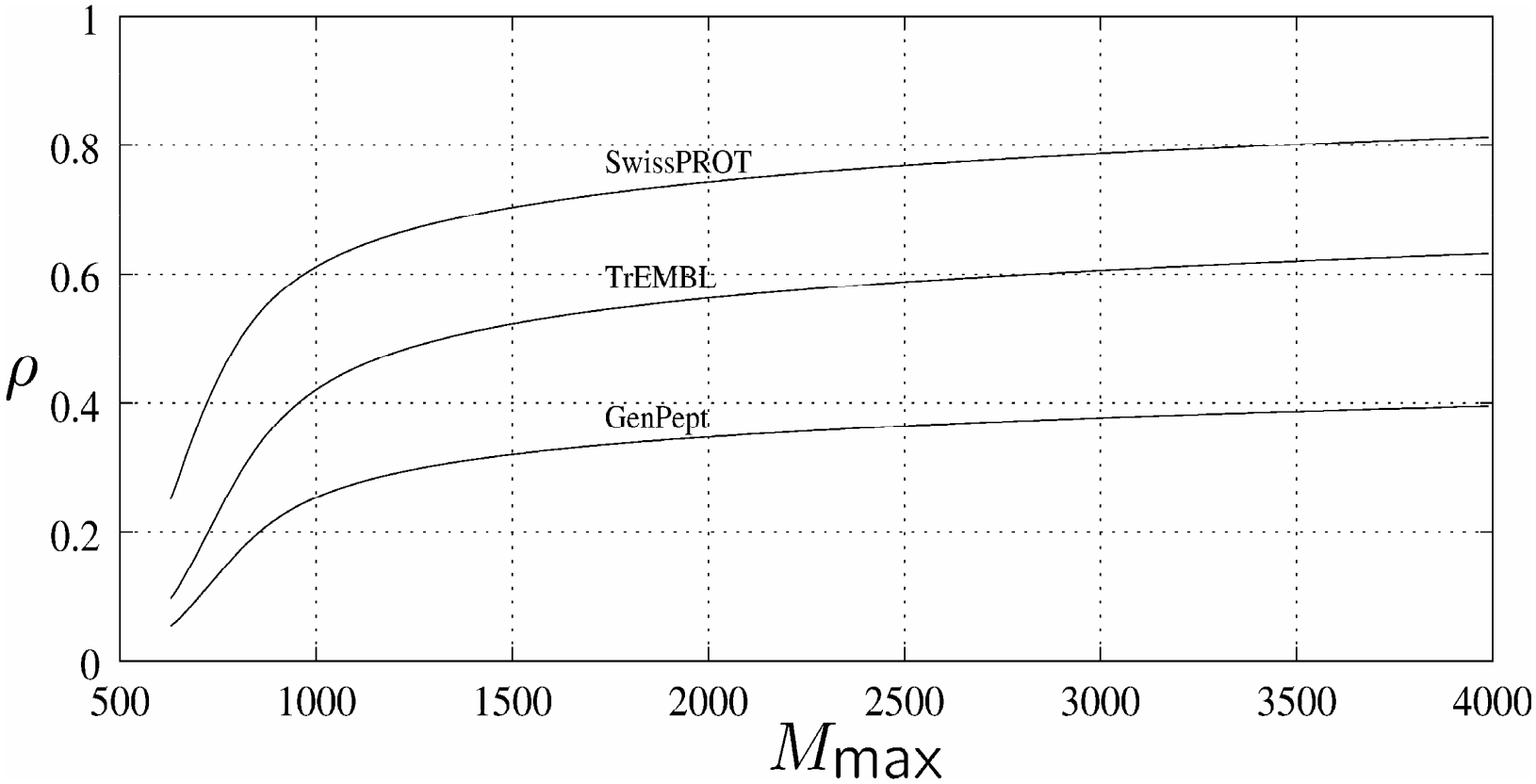
Simultaneous Linear Scan

- O(k log k + n L log k) vs O(k n)
- Now a mass lookup problem rather than a string scanning problem
- Empirically, we have to look-up *every* peptide!

LC/MS/MS Experiment Observed Parent Masses

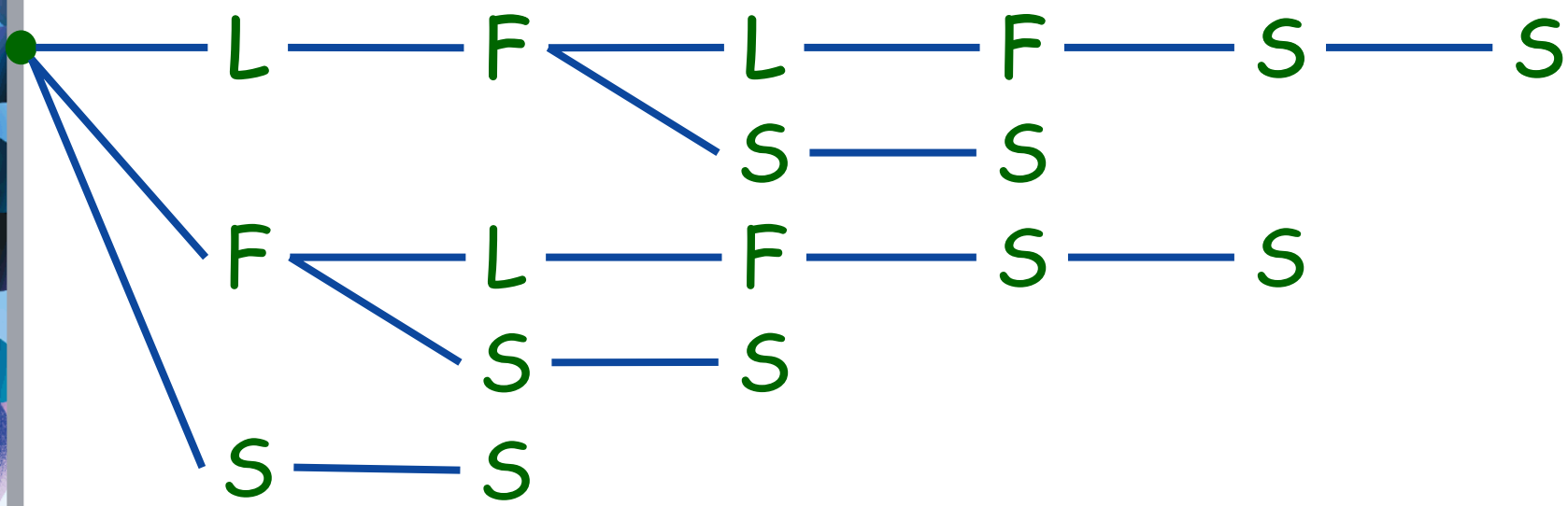


Substring Density

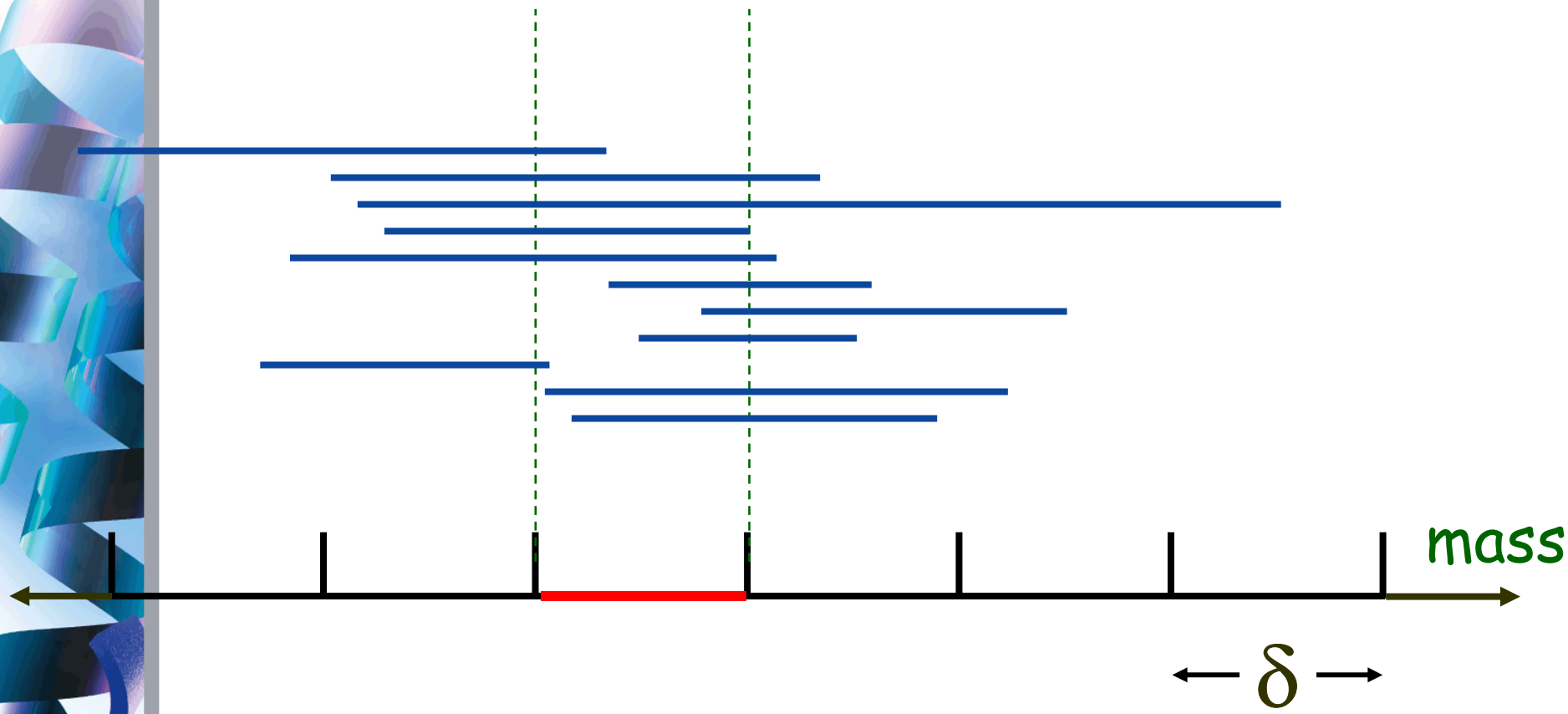


Redundant Candidate Elimination

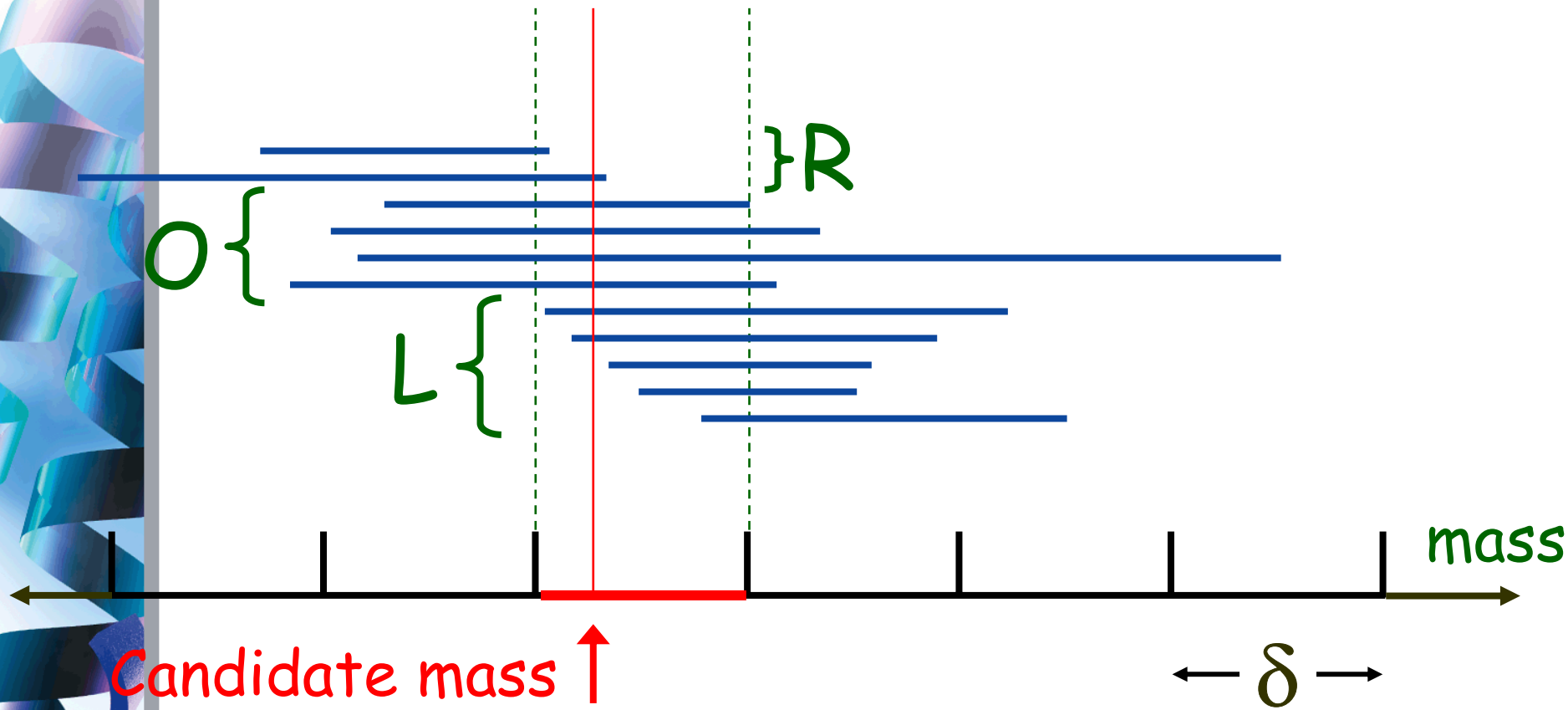
- Suffix trees represent all distinct substrings of a string.



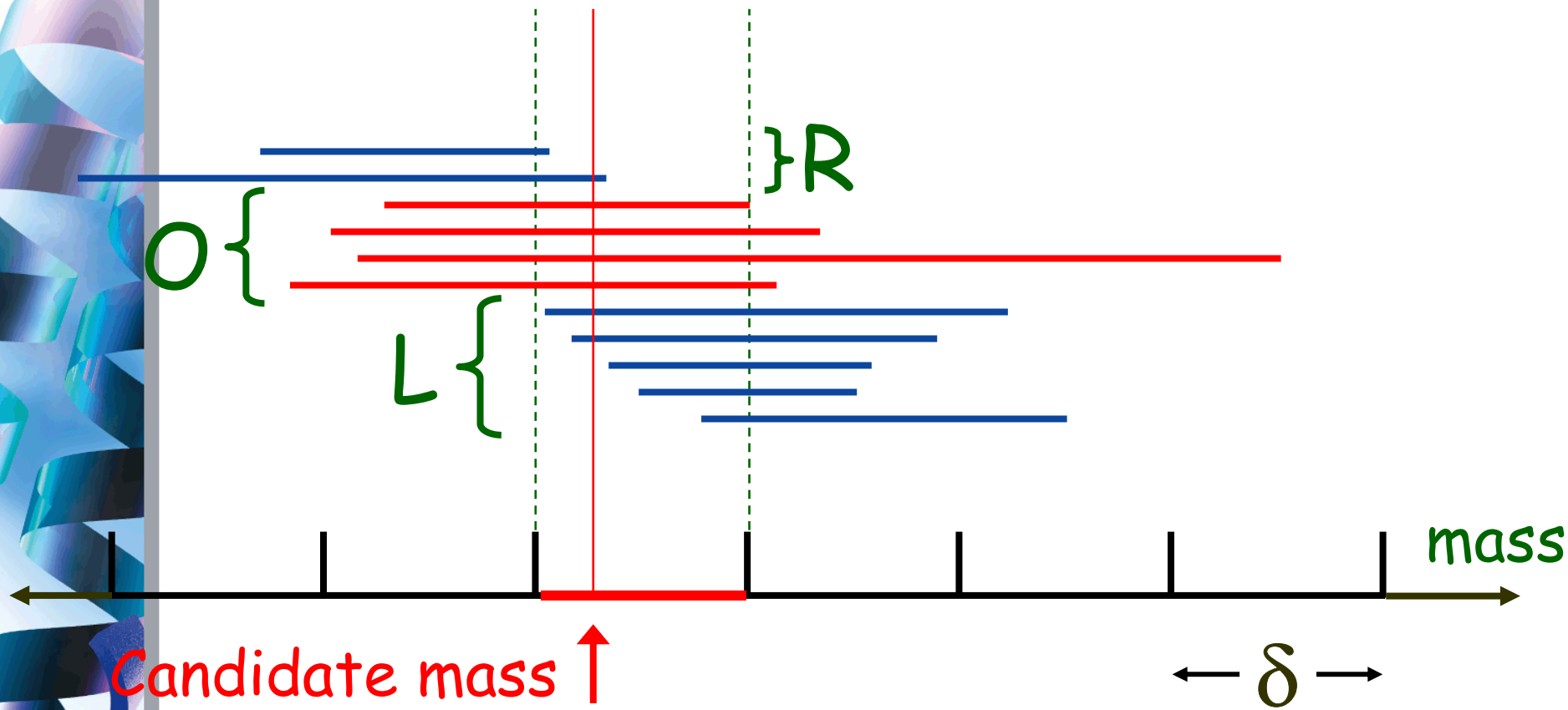
Fast Query Mass Lookup



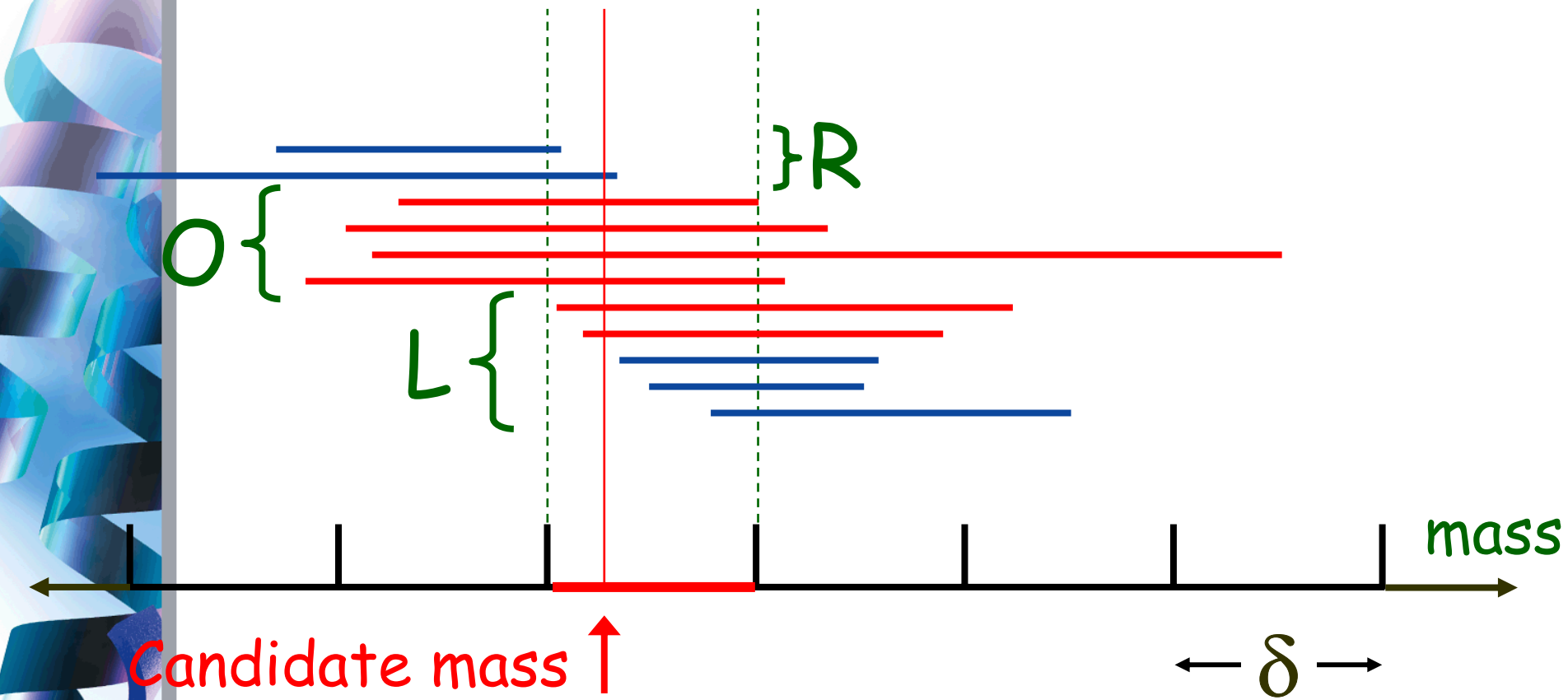
Fast Query Mass Lookup



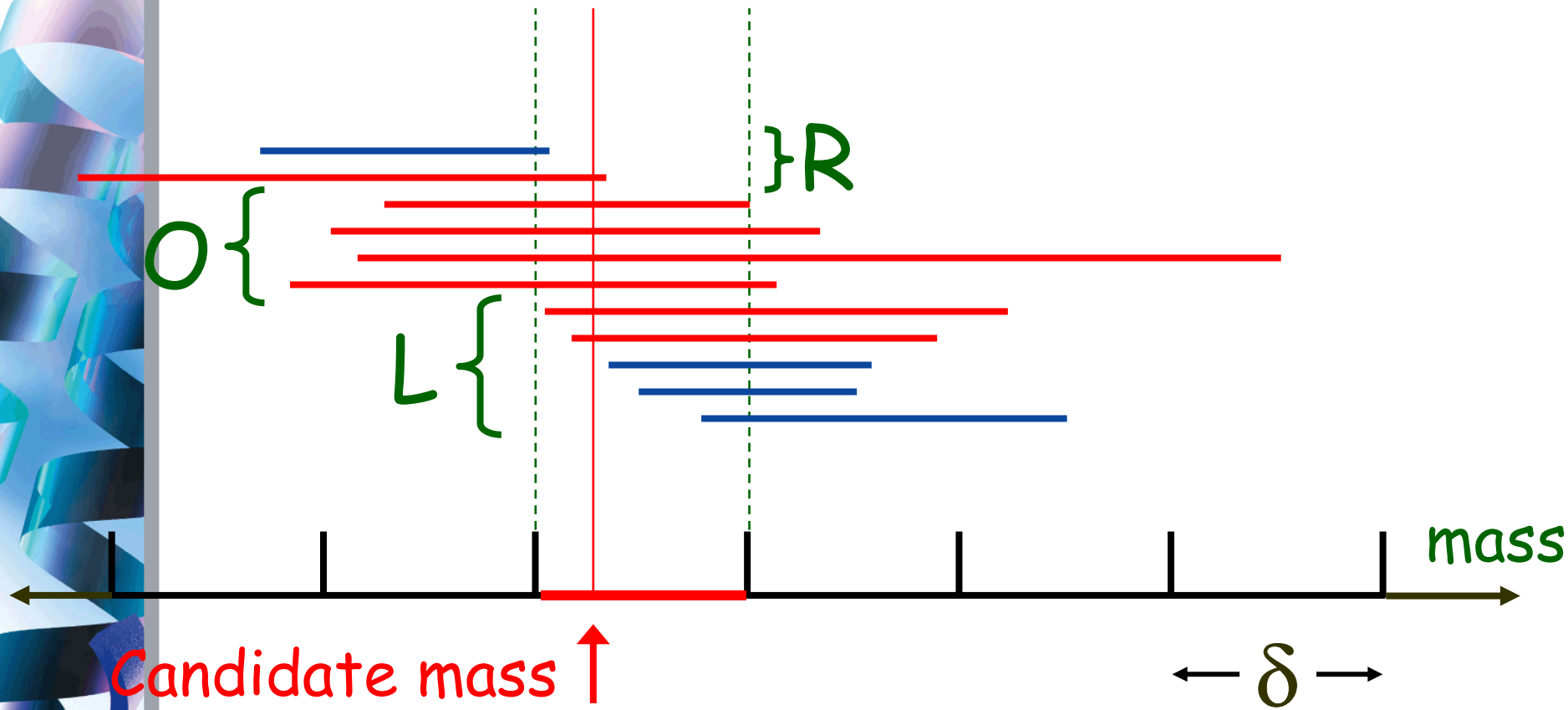
Fast Query Mass Lookup



Fast Query Mass Lookup



Fast Query Mass Lookup



Haplotypes

- Genetic variation thought responsible for many diseases
- Single nucleotide polymorphisms (SNPs) are very common
- Genetic mutation is rare
- Genetic mutation is inherited
- We have two copies of each chromosome.

Haplotypes

1:ACGACTCAGATCACTACGTACGACT

1:ACGACTCAGATAACTACGGACGACT

2:ACGACTCAGATCACTACGTACGACT

2:ACGACTCAGATCACTACGTACGACT

3:ACGAGTCAGATCACTACGTACGACT

3:ACGAGTCAGATAACTACGGACGACT

Haplotypes

1:ACGACTCAGATCACTACGTACGACT

1:ACGACTCAGATAACTACGGACGACT

2:ACGACTCAGATCACTACGTACGACT

2:ACGACTCAGATCACTACGTACGACT

3:ACGAGTCAGATCACTACGTACGACT

3:ACGAGTCAGATAACTACGGACGACT

Genotypes

1: C/C, A/C, G/T

2: C/C, C/C, T/T

3: G/G, A/C, G/T

1: 0, 2, 2

2: 0, 0, 0

3: 1, 2, 2

Haplotype Phasing Problem

Given a set of genotypes, find the haplotypes that generated them.

Parsimony:

- “Simplest” solution is “right”
- Minimum number?
- Minimum diversity?
- Respect some pedigree?

Phasing by Integer Programming

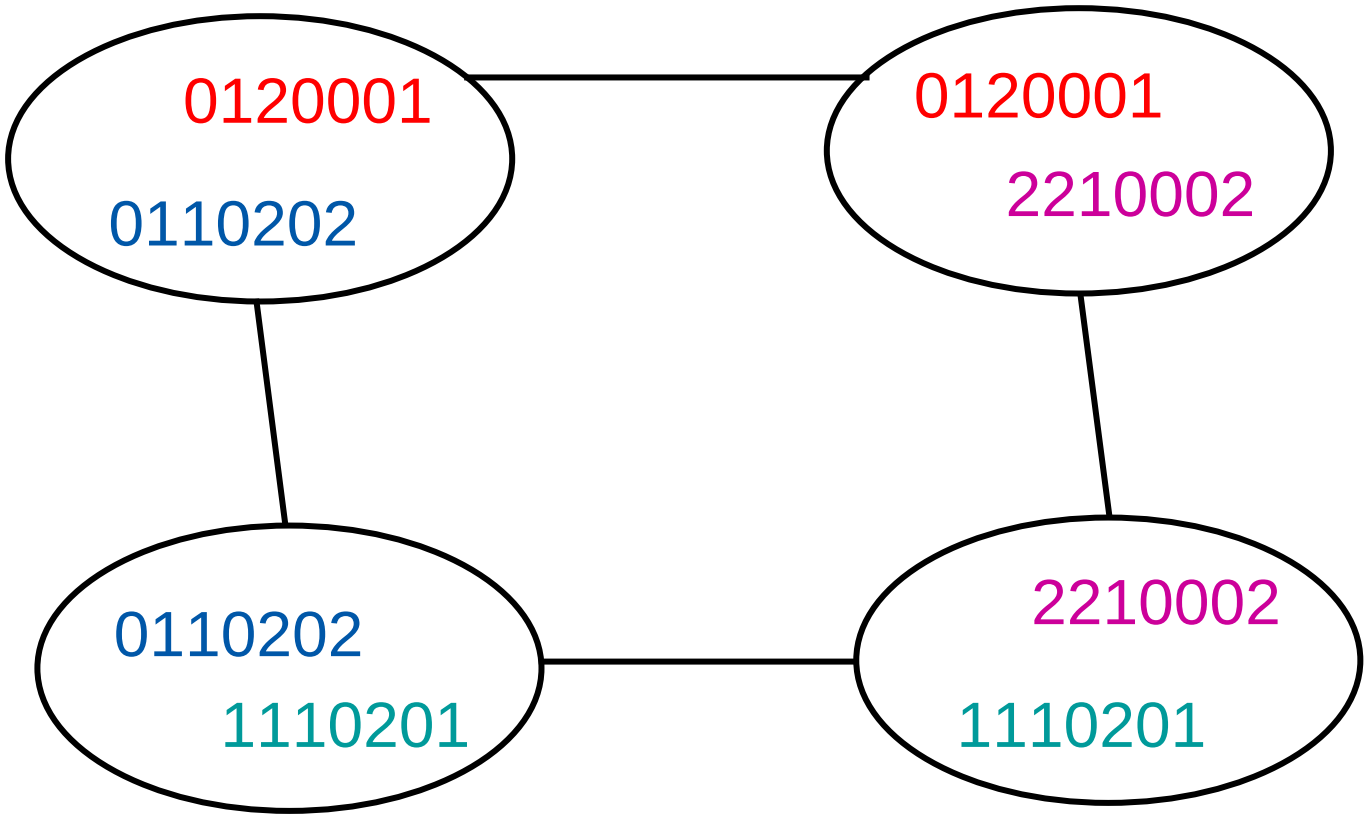
- Gusfield used an exponential size facility location like formulation
- Haplotypes/facilities (y_h)
- Phasing assignment (x_{gp})
- Must phase: $\Sigma_p\, x_{gp} = 1$ for all g
- Used haplotypes: $x_{gp} \leq y_h$ for $p = h + h'$

  $x_{gp} \leq y_{h'}$
- Minimize: $\Sigma_h\, y_h$
- Remove surplus x's & y's for tractability

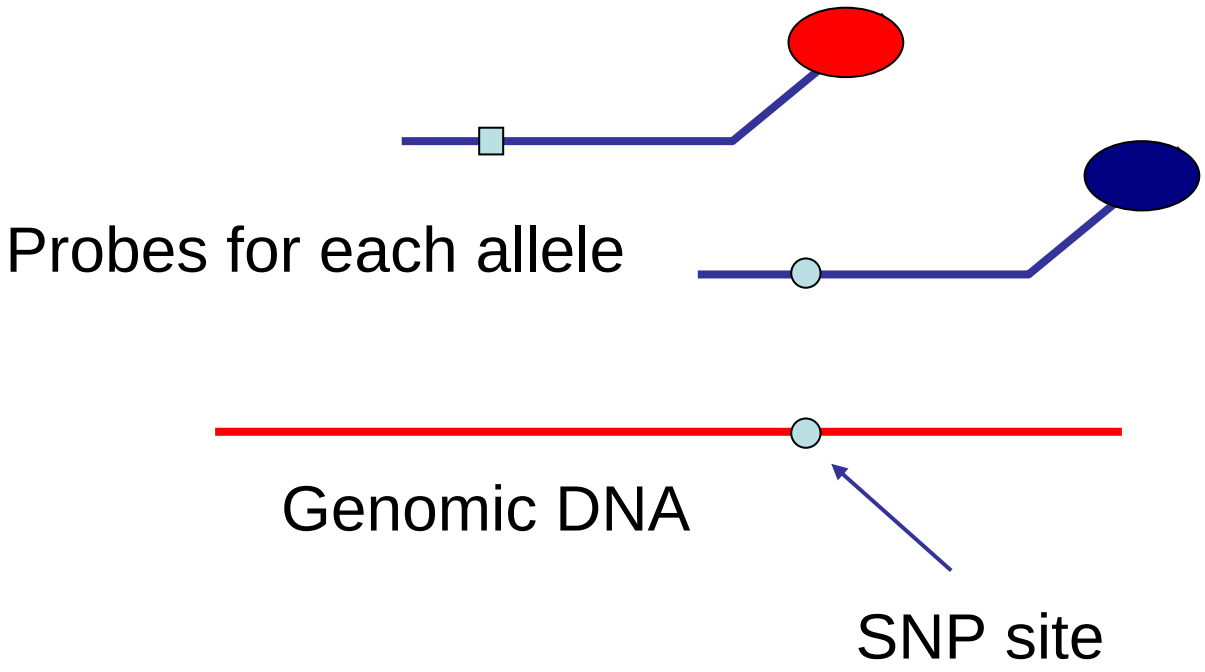
Phasing by Integer Programming

- Alternate formulation constructs equivalence classes of haplotypes
- Haplotypes h and h' are different: $d_{h,h'}$
- Many d's are set from the input data
- Minimize: $\Sigma_{h,h'} d_{h,h'}$
- Given integer solution, we can infer the values of 2's and test feasibility
- Need cutting planes for LP!

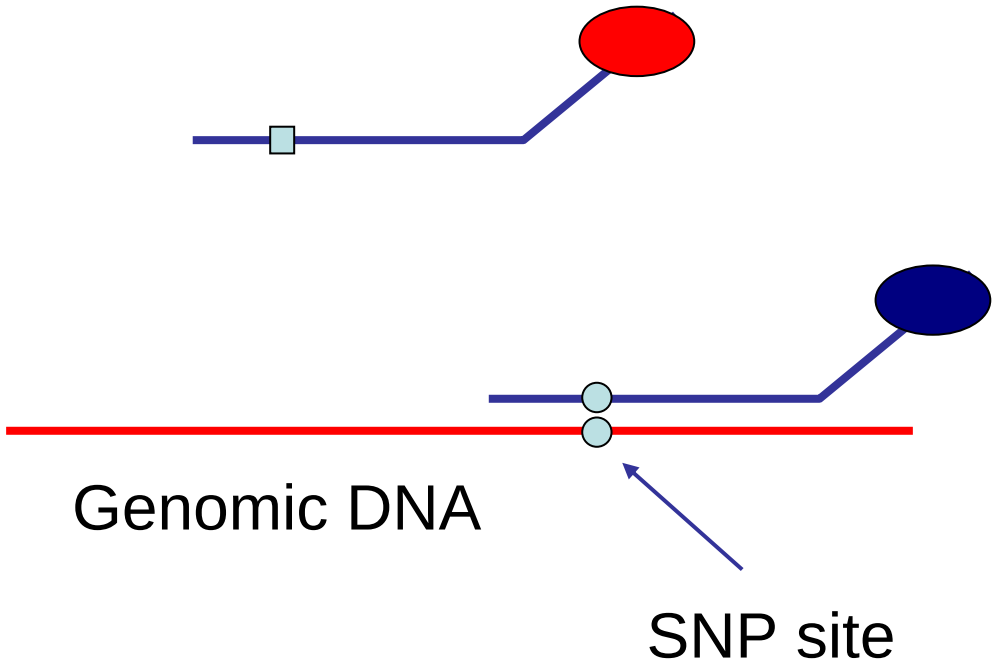
Inference from Equivalence Classes of Haplotypes



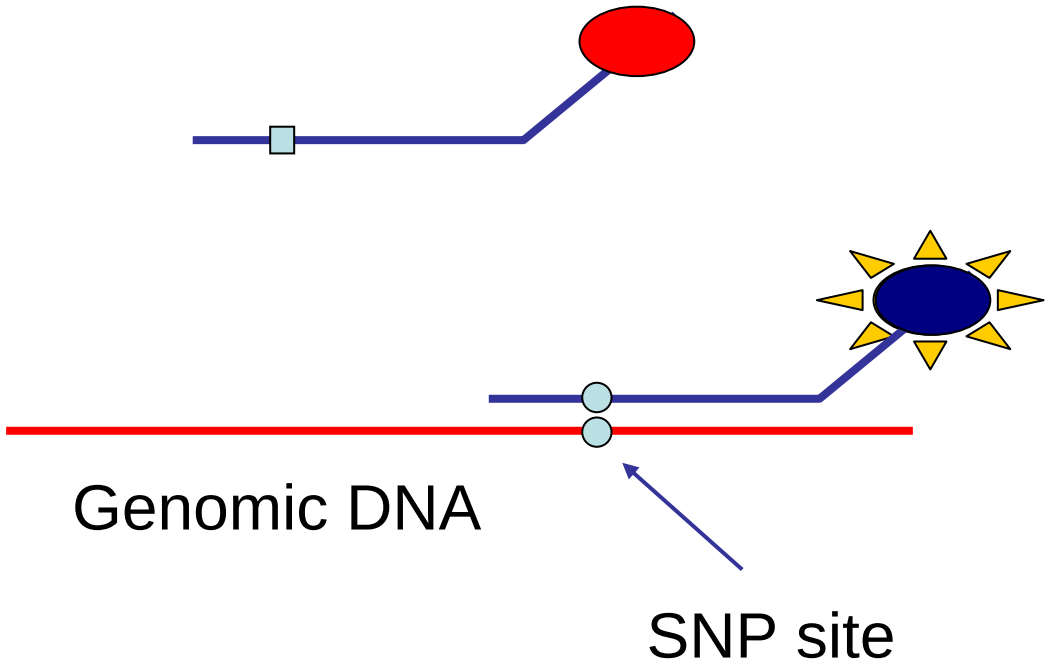
Multiplexed Assay Design



Multiplexed Assay Design



Multiplexed Assay Design



Multiplexed Assay Design

- Want to genotype n SNP sites
- Each SNP is assigned two “labels”
- Labels are read by some instrument
- Instrument can de-convolute m (« n) labels
- Sometimes SNP reagents interfere with one another

- Find feasible groups of SNPs!

Multiplexed Assay Design

- Can’t group same label SNPs together
  - Clique in conflict graph
- Solution corresponds to a coloring of the conflict graph
- Algorithm: Color cliques arbitrarily and iteratively fix solution until feasible.
- What about randomized list-coloring?
- Can we find orthogonal solutions?

Interested?

- Big names…
  - Pavel Pevzner (UCSD)
  - Dan Gusfield (UCDavis)
  - Ron Shamir (Tel Aviv)
  - Steve Skiena (Stony Brook)
- Conferences
  - RECOMB, ISMB
- Data-sources and education
  - NCBI (NIH)
  - TIGR