Generating Peptide Candidates from Protein Sequence Databases for Protein Identification via Mass Spectrometry

Nathan Edwards
Informatics Research

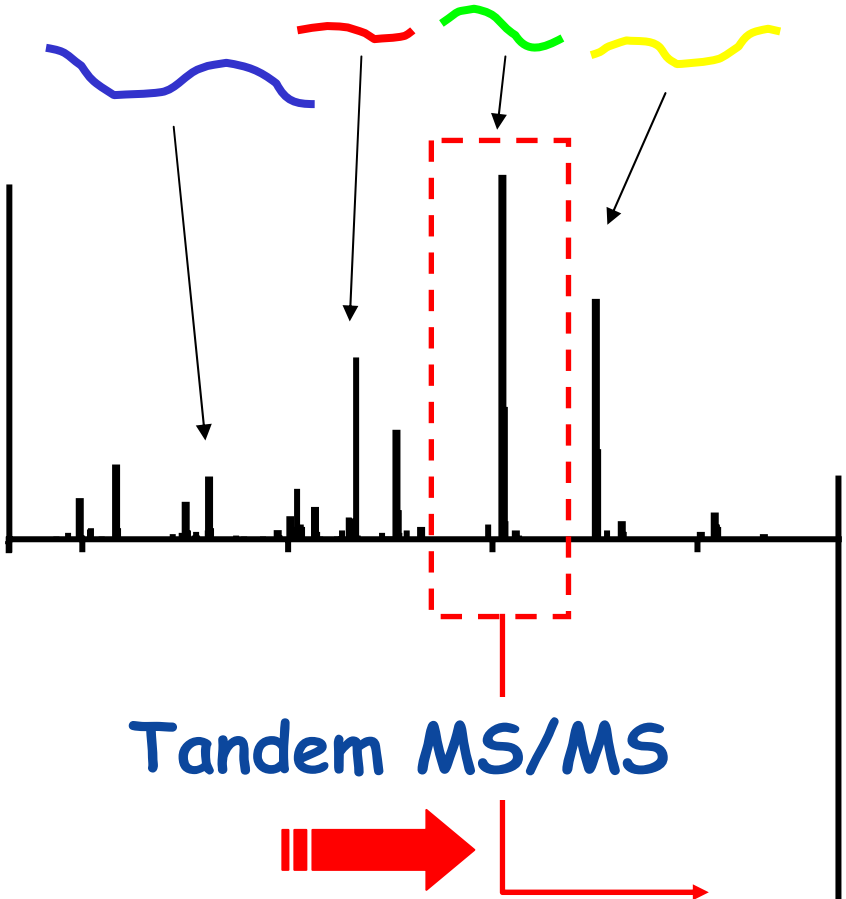
Applied Biosystems

Protein Identification

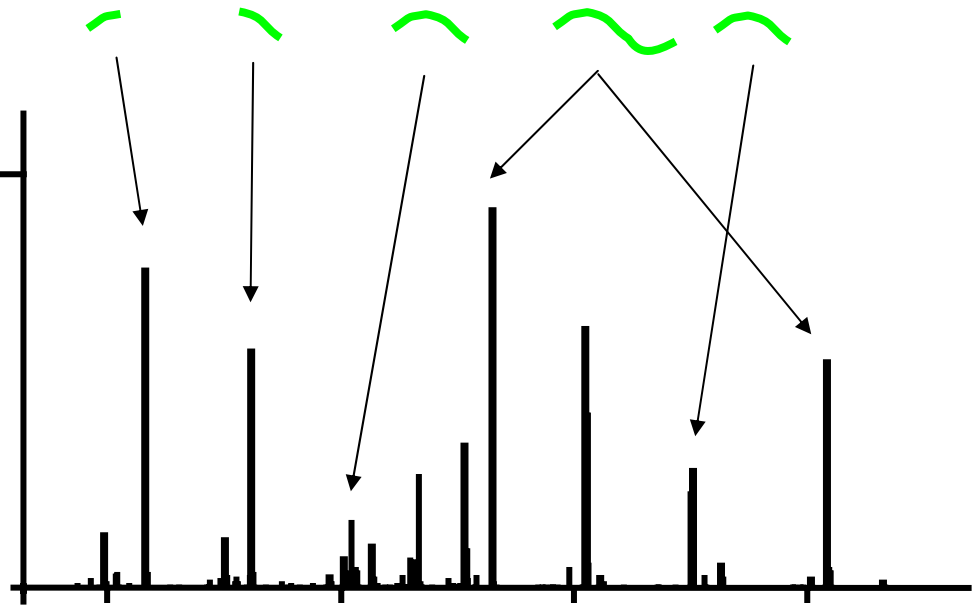
- Turns mass spectrometry into proteomics
- Sequence is link to identity, annotation, literature, genomics
  - Proteomics workflows interrogate more than mass
  - Quality of AA sequence databases sequence & annotation varies wildly
- Protein identification is not BLAST!

Applied Biosystems

LC-MS/MS for Protein Id



LC-MS/MS: 1 MS spectrum followed by 2-5 Tandem MS/MS spectra every 5-10 sec.



Applied Biosystems

LC-MS/MS for Protein Id

- 1 experiment produces 1000's of MS/MS spectra
- Suitable for complex mixtures
- 100's-1000's of proteins identified from a single experiment
- **High-throughput protein identification!**

Applied Biosystems

Sequence Database Search Engines

Input: Set of MS/MS spectra and associated parent ion masses

Output: Peptide sequence for each spectrum

1. Generate peptide candidates from a protein or genomic sequence database
2. Score and rank the peptide candidates

Sequence Database Search Engines

Input: Set of MS/MS spectra and associated parent ion masses

Output: Peptide sequence for each spectrum

1. Generate peptide candidates from a protein or genomic sequence database
2. Score and rank the peptide candidates

Peptide Candidate Generation

Input: Sequence σ (length n), from alphabet **A**
(Additive) mass μ(a) for a **2 A**
Query masses $M_1,\ldots,M_k$

Output: All (distinct) pairs of query masses i and subsequences ω with

$$\sum_{j=1}^{|\omega|} \mu(\omega_j) = M_i$$

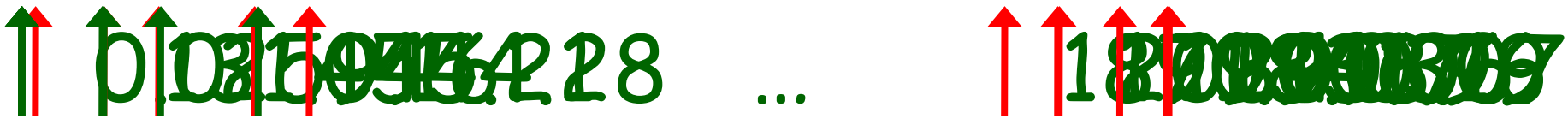
Peptide Candidate Generation and Peptide Id

- Sequence databases contain many individual proteins
- Must avoid redundant scoring
- Protein context is important

Simple Linear Scan

Query Mass = 2018.07

MKWVTFISLLFLFSSAYSRGV…



Output: WVTFISLLFLFSSAYSR

Sequential Linear Scan

- O(nk) time
- Simple to implement
- Easy to track protein context
- Poor data locality
- Redundant candidates
- String scanning problem

Simultaneous Linear Scan

Max Query Mass = 2018.07

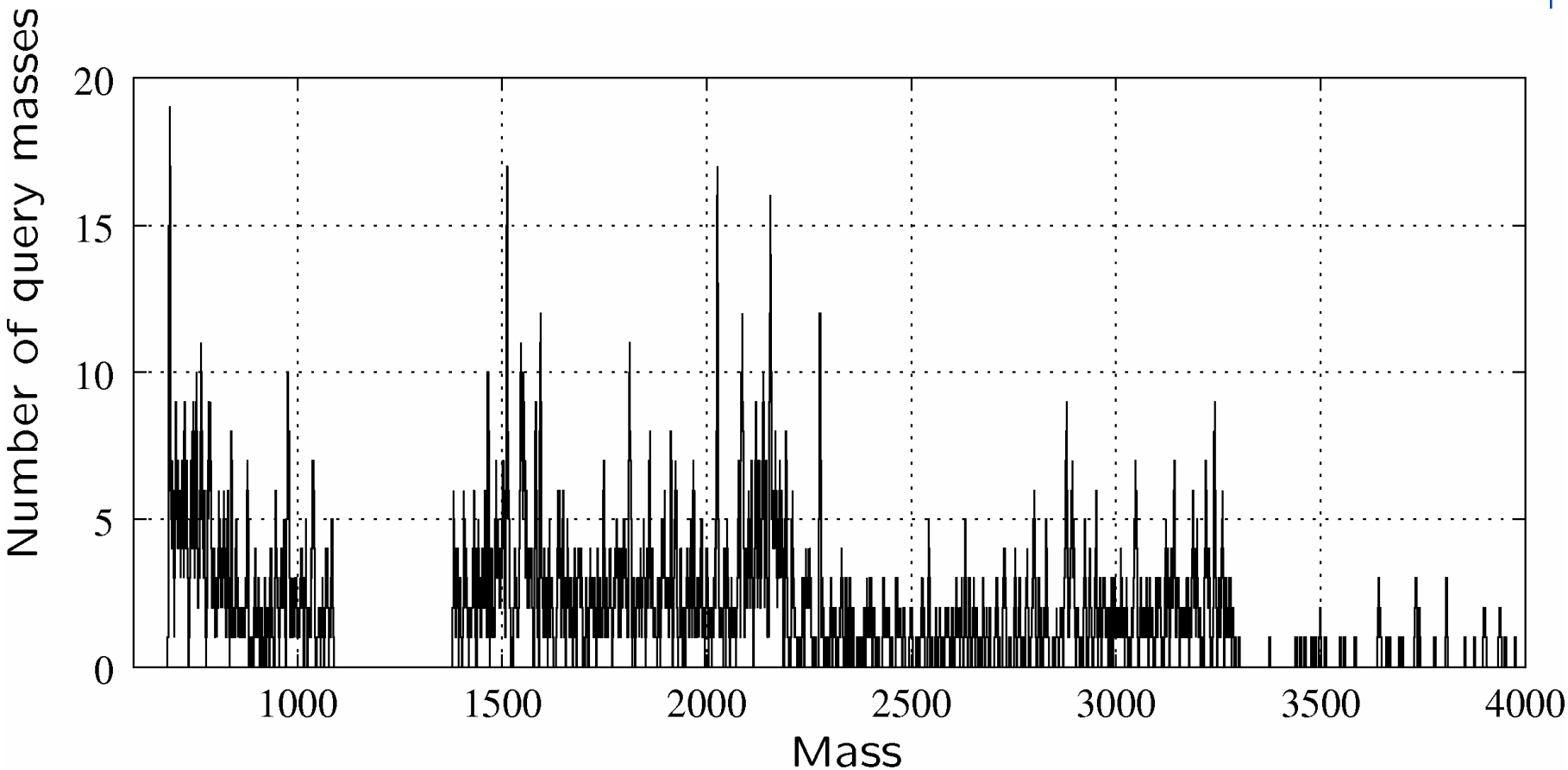
Lookup each candidate mass in turn.

Applied Biosystems

Simultaneous Linear Scan

- O(k log k + n L log k) time
- Simple to implement
- Easy to track protein context
- Better data locality
- Redundant candidates
- Now a query mass lookup problem!

Overlap Plot from a LC/MS/MS Experiment



Applied Biosystems

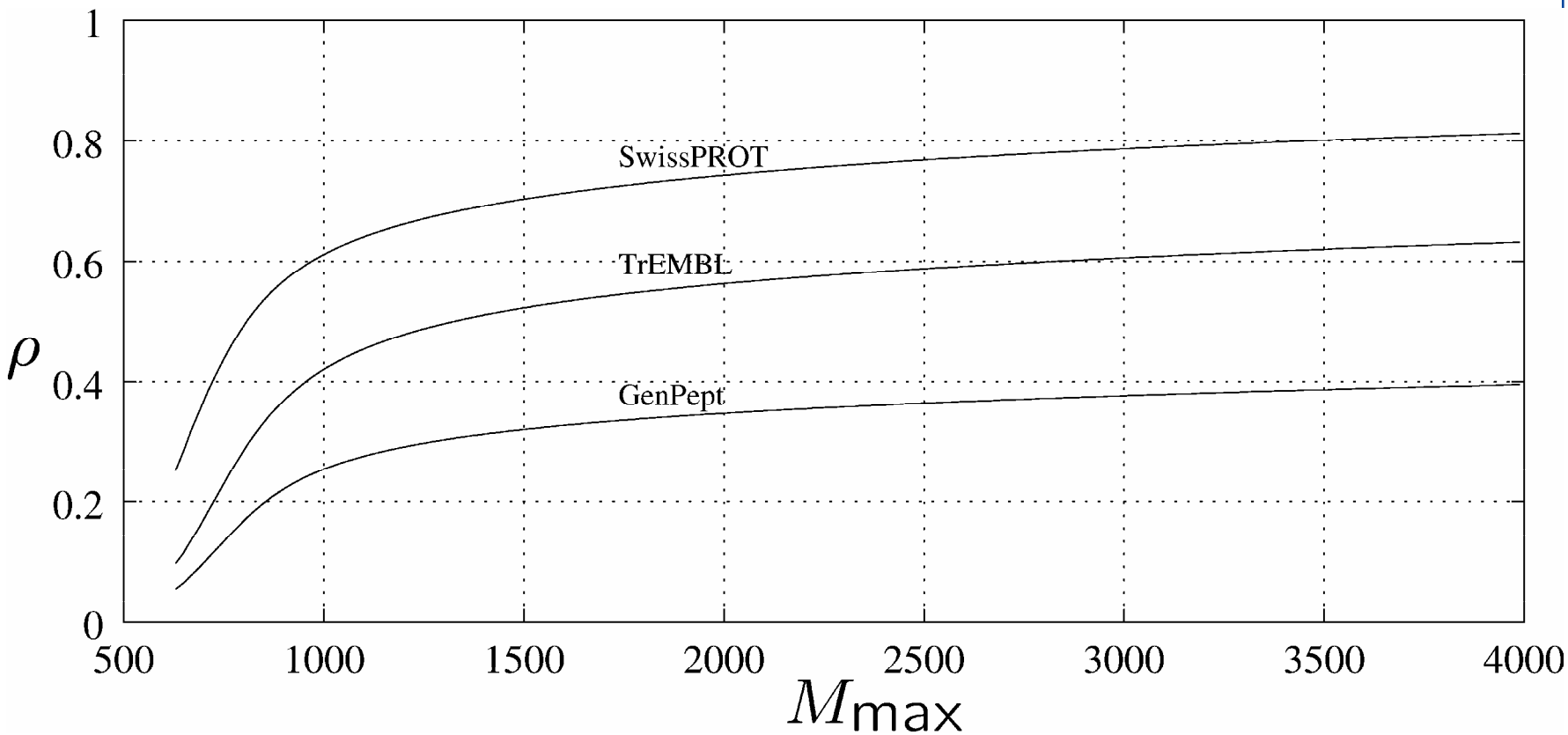
Redundant Candidate Elimination

- Must avoid repeat scoring of the same peptide candidate

- Want to avoid generating redundant candidates

- Non-redundant sequence databases contain lots of substring redundancy!

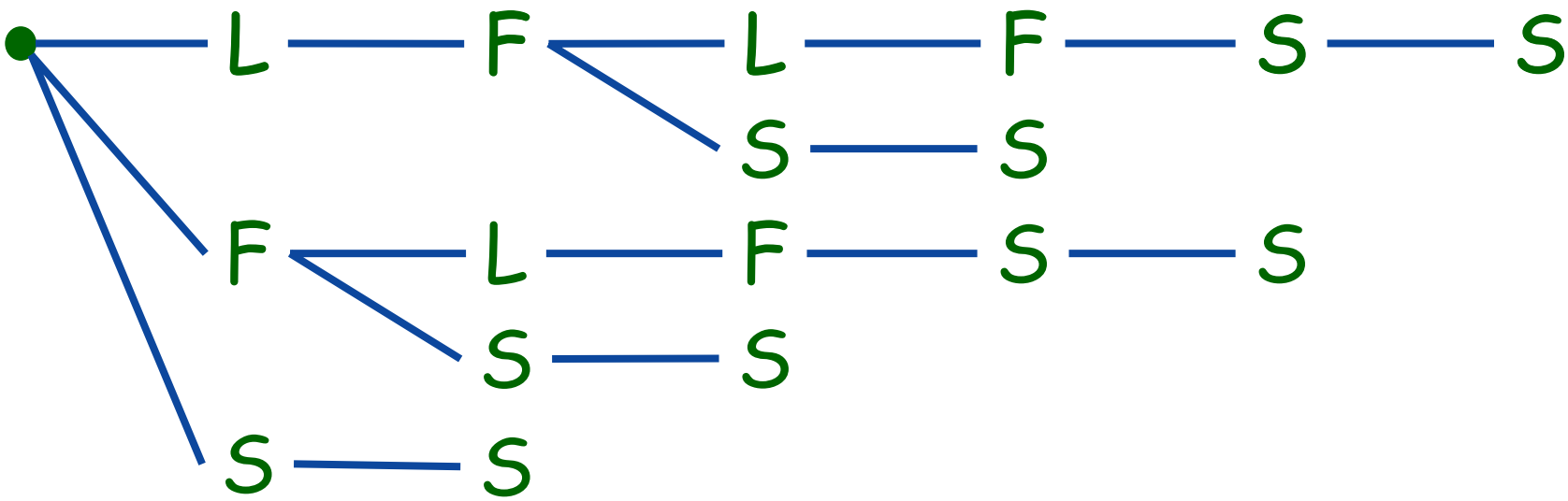
Substring Density (ρ)



Applied Biosystems

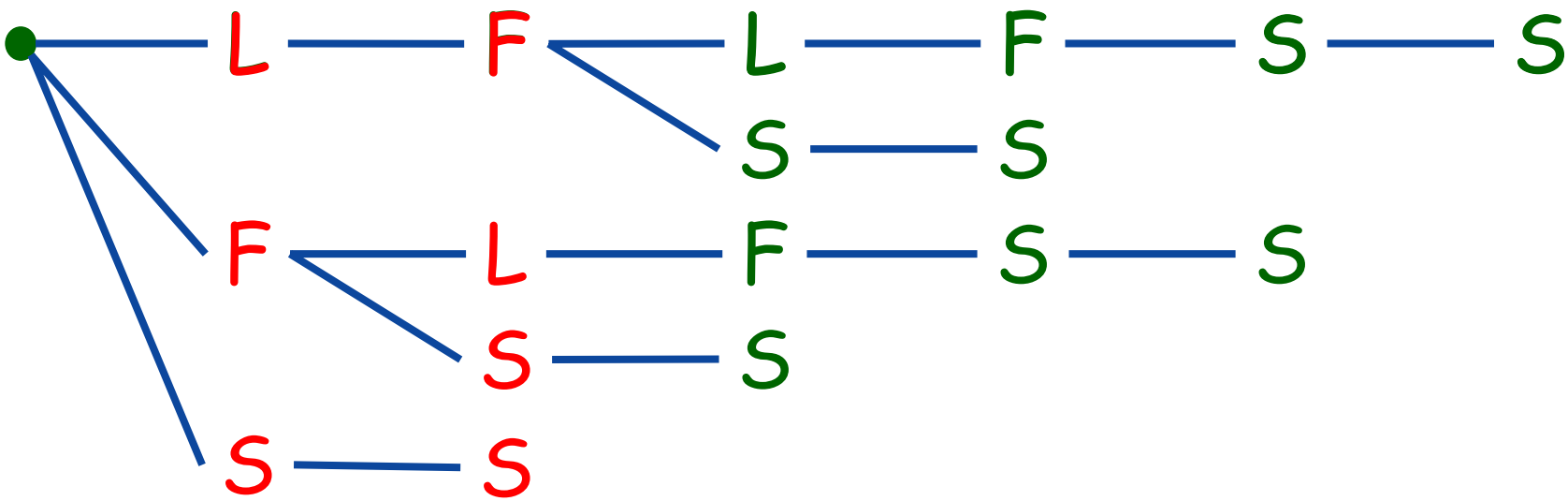
Redundant Candidate Elimination

- Suffix trees represent all distinct substrings of a string.



Redundant Candidate Elimination

- Suffix trees represent all distinct substrings of a string.



Suffix-Tree Traversal

- O(k log k + n L ρ log k) time
- Redundancy eliminated
- Tricky to implement well
- Memory overhead ¼ 5n
- Protein context more involved
- Data locality hard to quantify
- Must preprocess sequence db
- Still a query mass lookup problem!

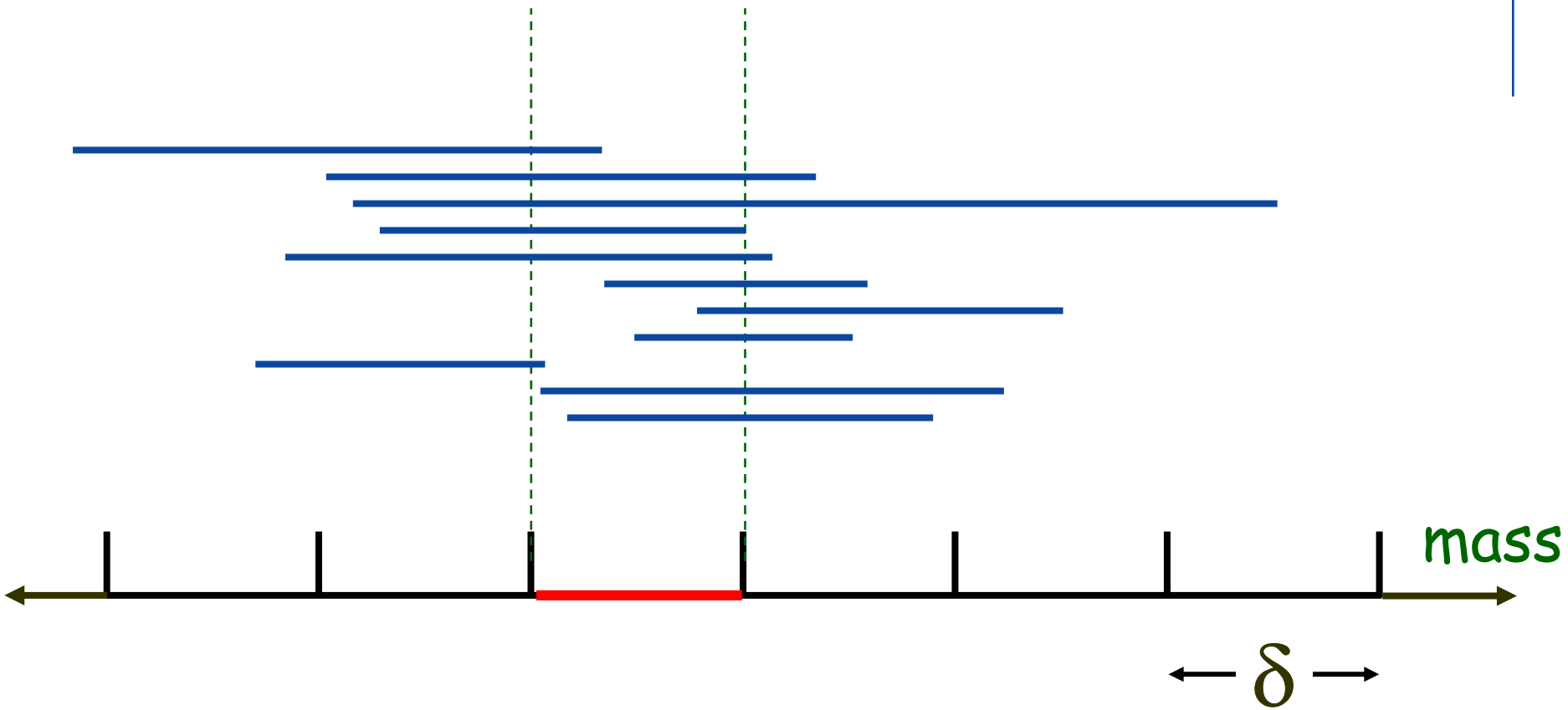
Fast Query Mass Lookup

- With (small) integer weights, $O(M_{max} + k + n\,L\,\rho\,O)$ time is possible
- Use a query mass lookup table!

- Can we achieve this for real weights and non-uniform tolerances?

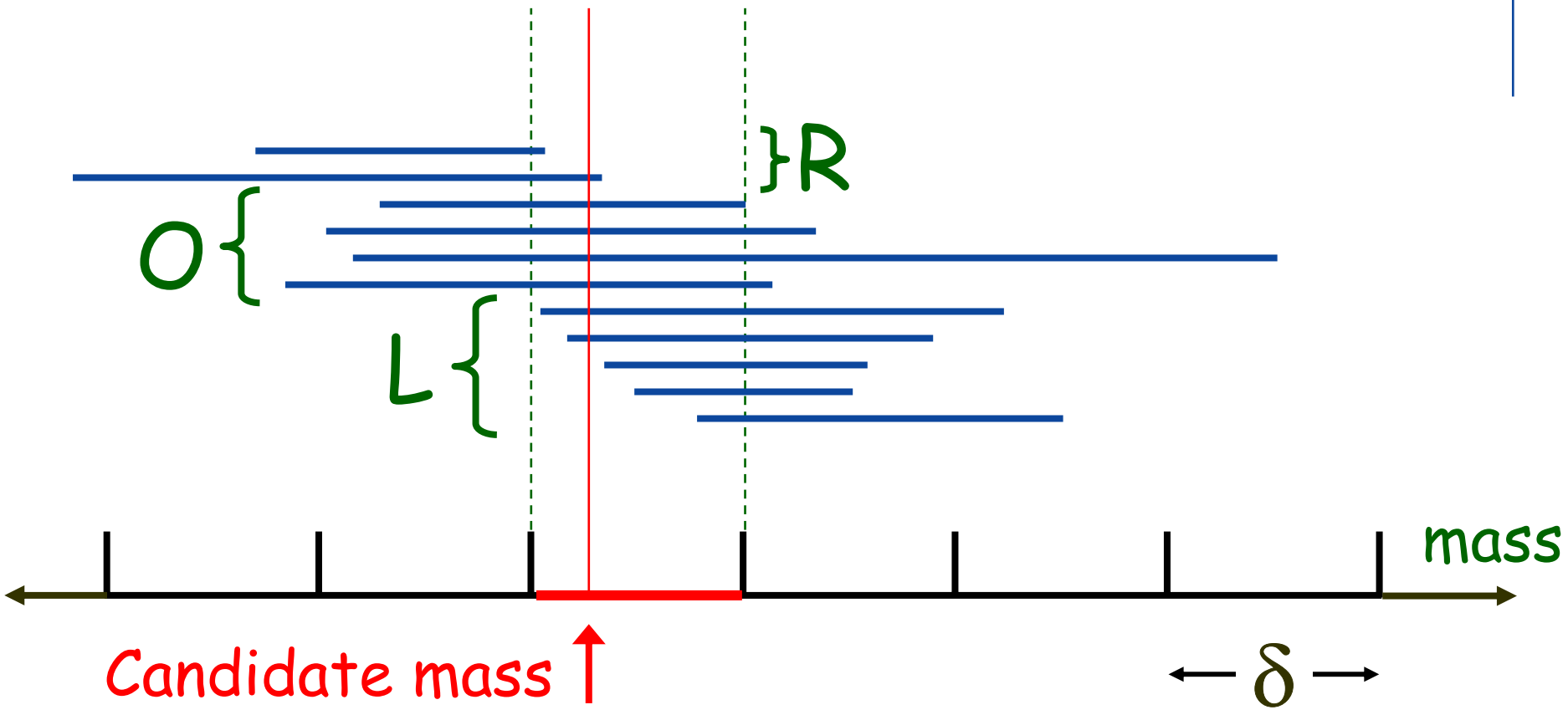
YES!

Fast Query Mass Lookup



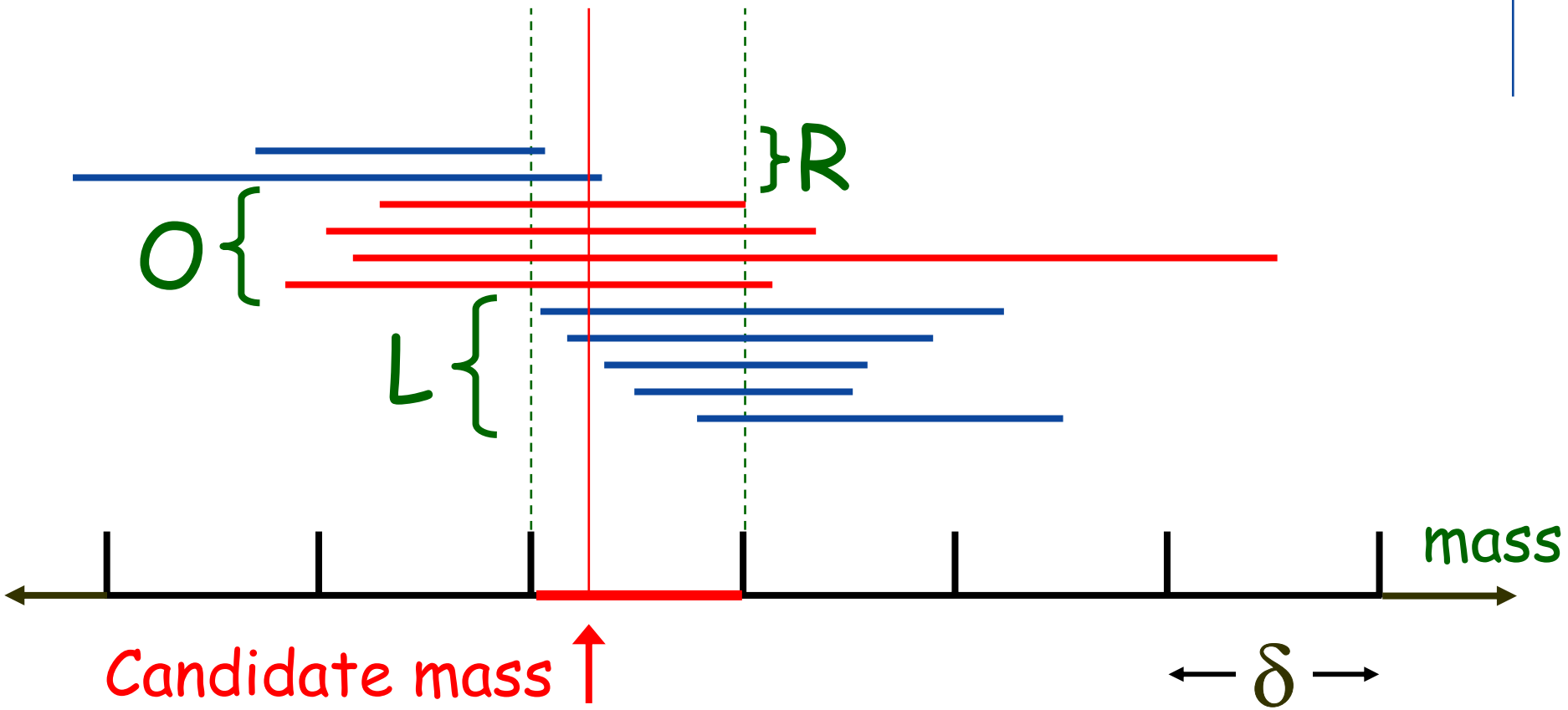
Applied Biosystems

Fast Query Mass Lookup



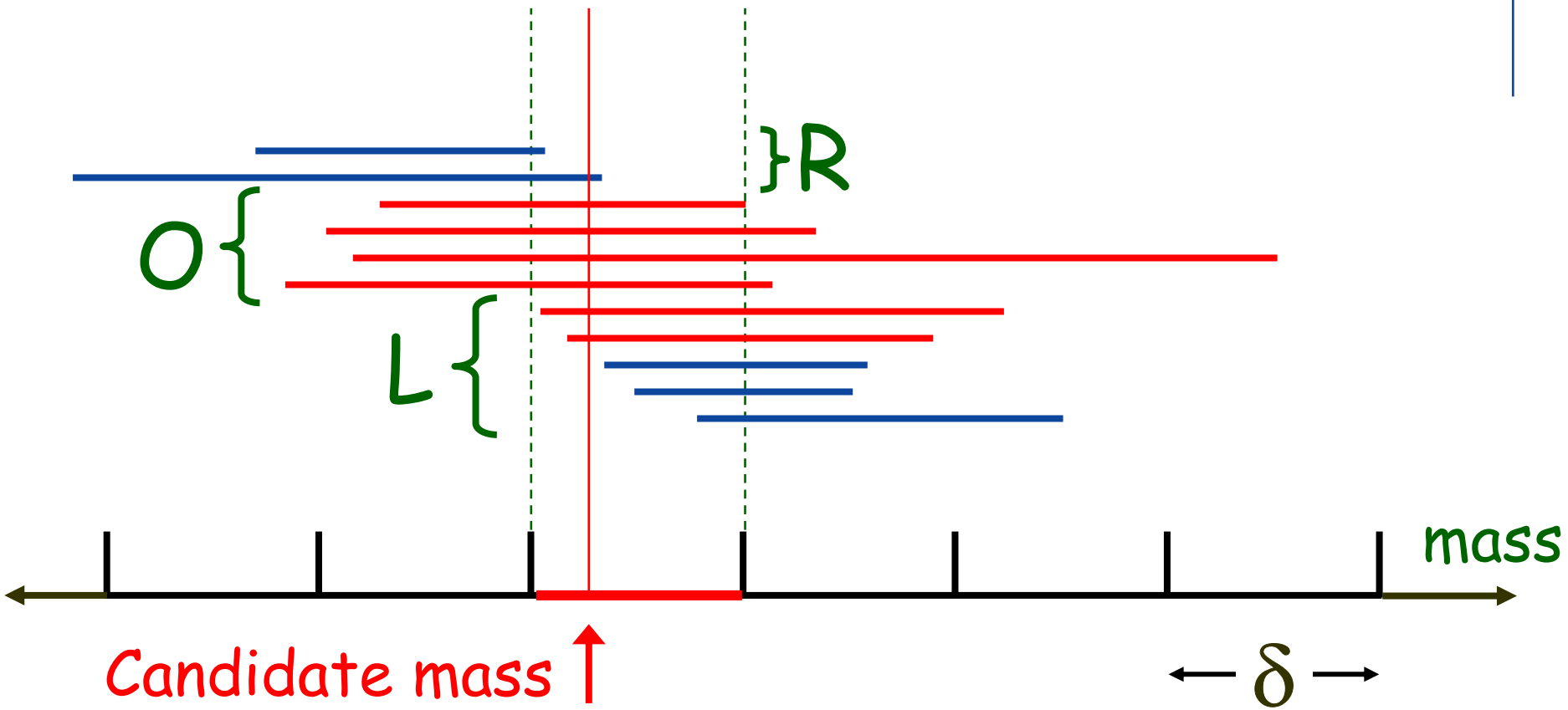
Applied Biosystems

Fast Query Mass Lookup



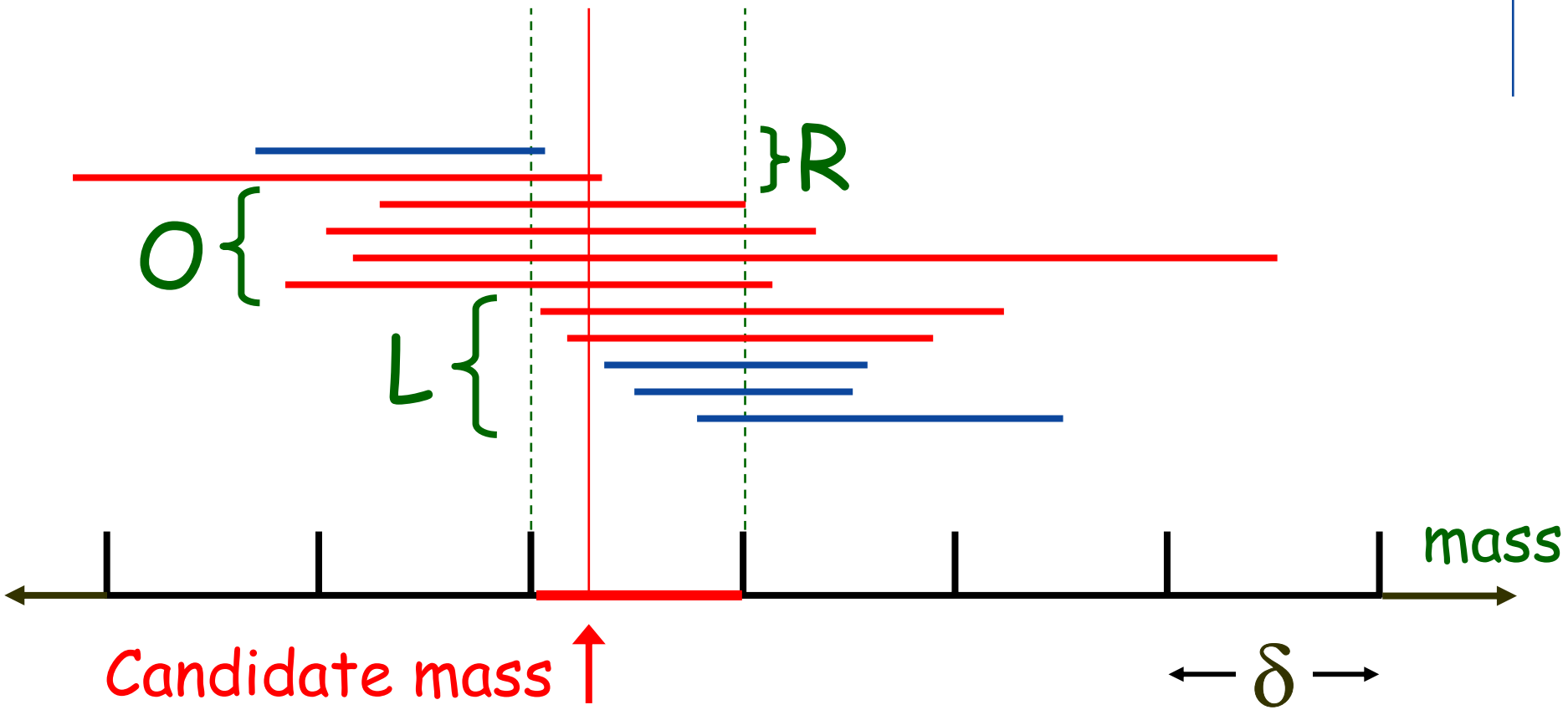
Applied Biosystems

Fast Query Mass Lookup



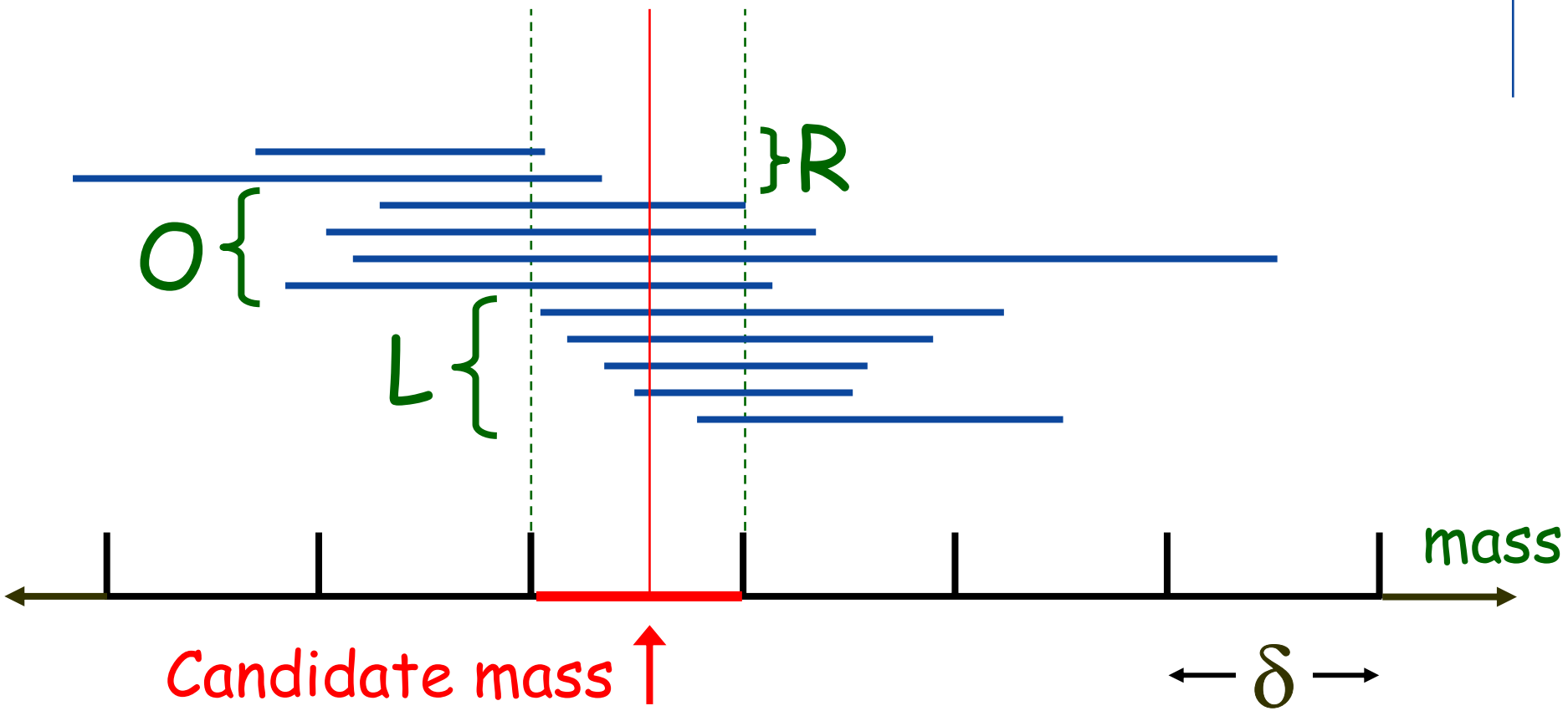
Applied Biosystems

Fast Query Mass Lookup



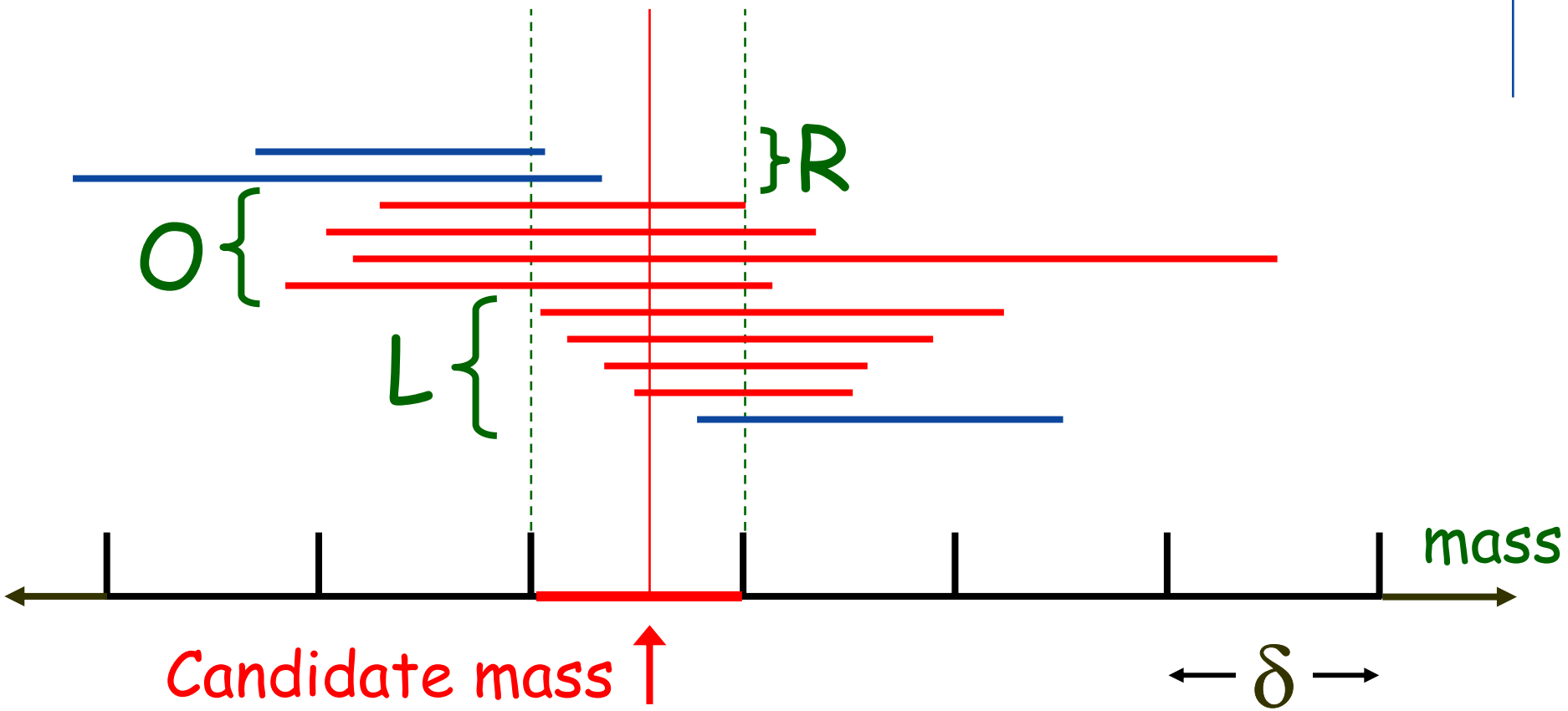
Applied Biosystems

Fast Query Mass Lookup



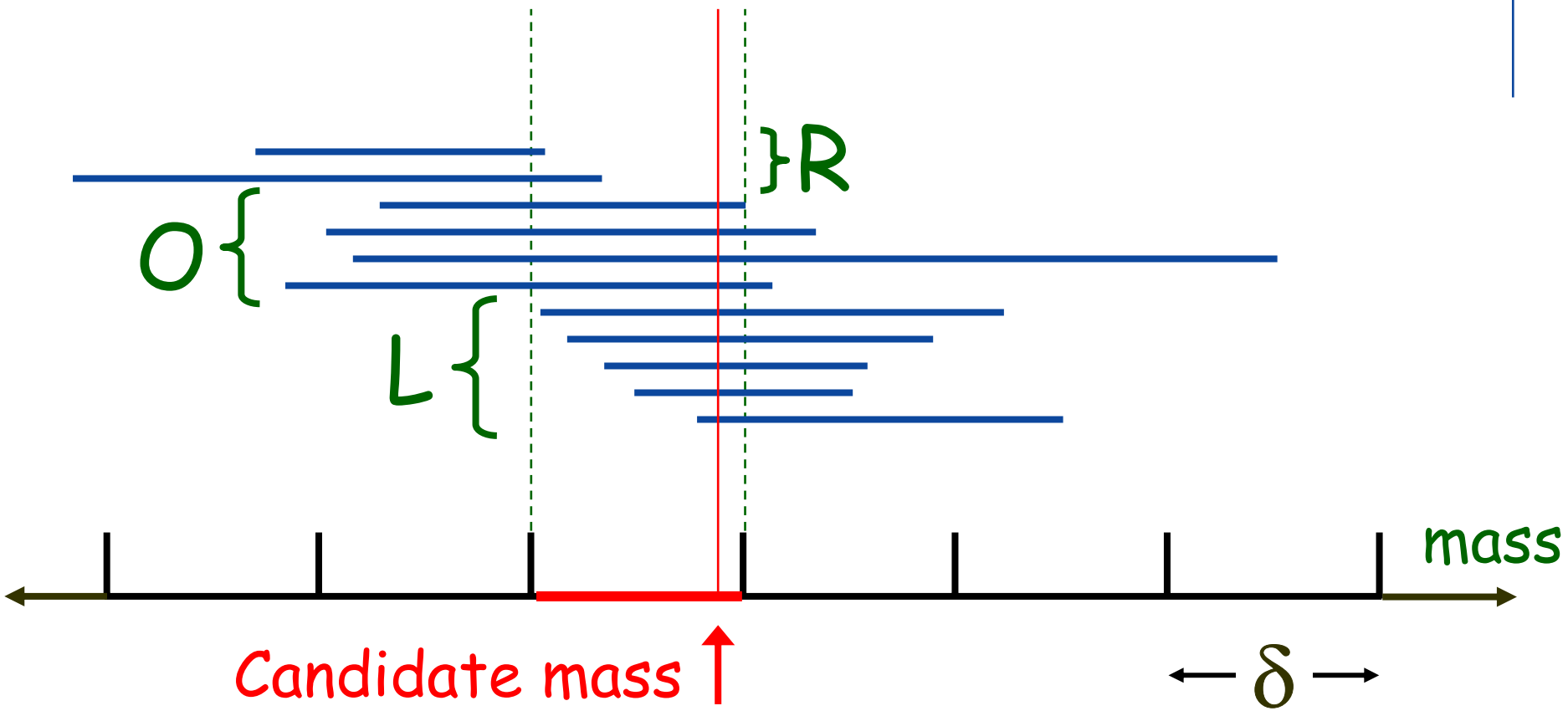
Applied Biosystems

Fast Query Mass Lookup



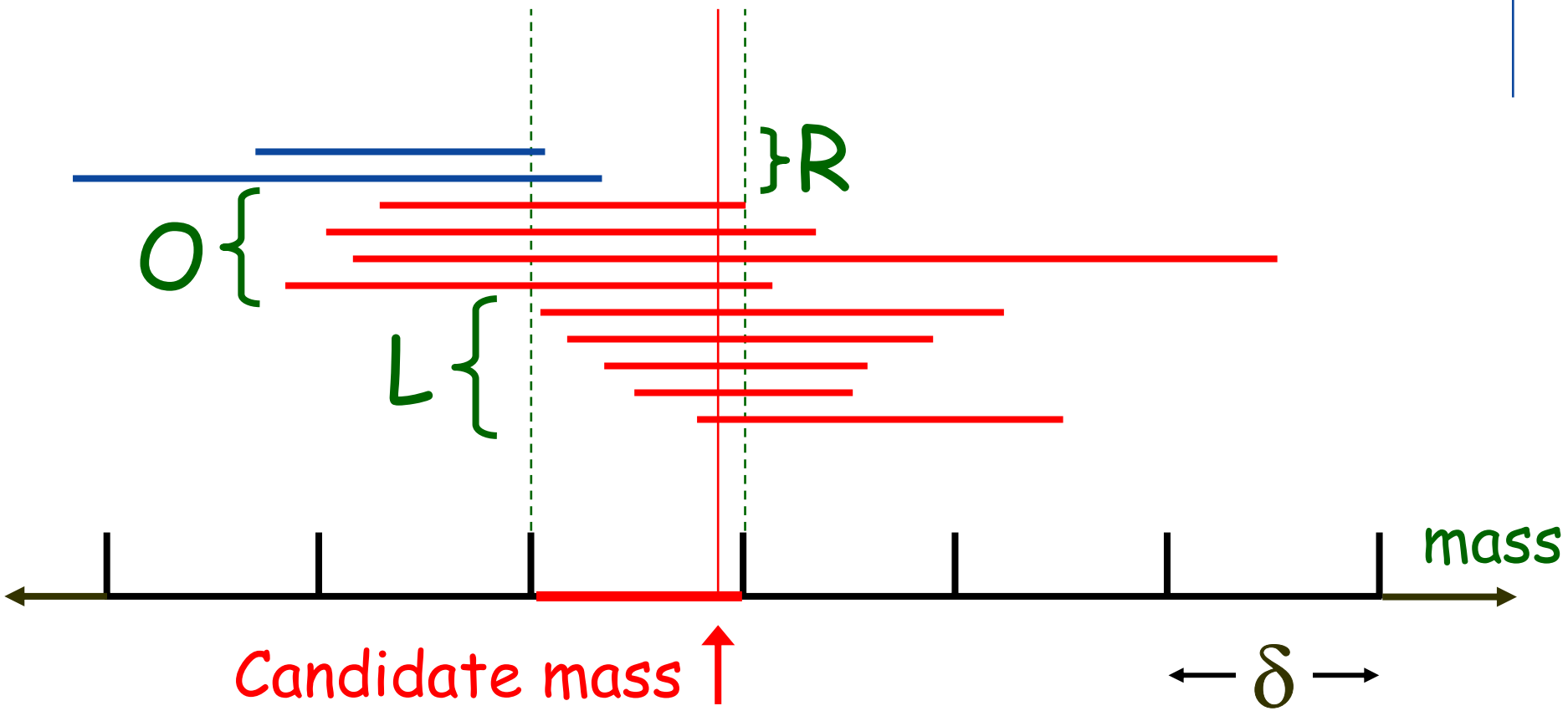
Applied Biosystems

Fast Query Mass Lookup



Applied Biosystems

Fast Query Mass Lookup



Applied Biosystems

Fast Query Mass Lookup

- Must have $\delta \cdot I_{min}$
- Table size is $O(M_{max}/\delta + k\, I_{max}/\delta)$
- Practical for typical parameters

- Running time:

  Table construction + $O(n\, L\, \rho\, O)$

  is dominated by size of output

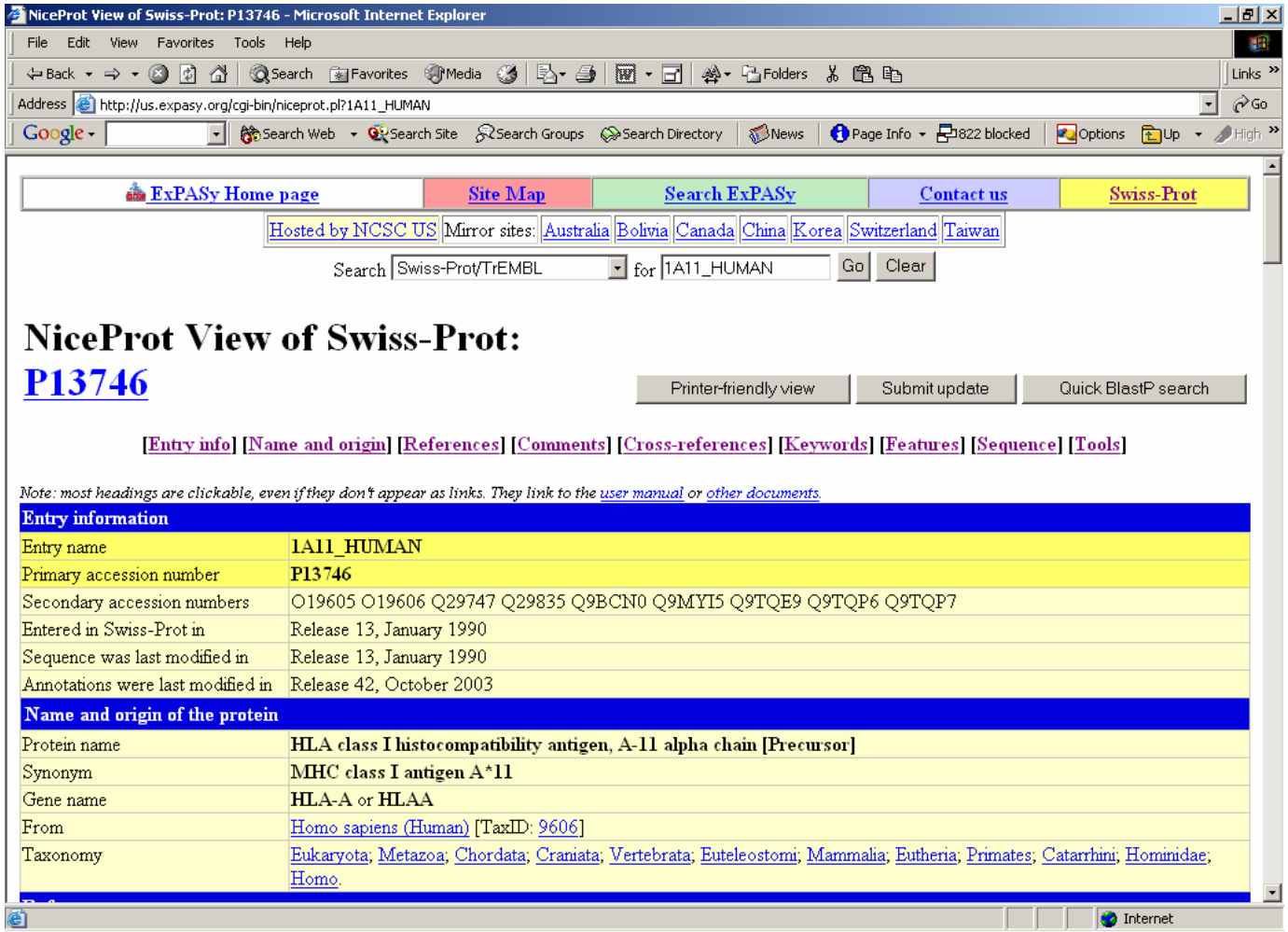
Observations

- Peptide candidate generation is a key subproblem.
- Must eliminate substring redundancy.
- As k increases, peptide candidate generation becomes an interval lookup problem.
- Run time dominated by output size.

Sequence Database Search Engines

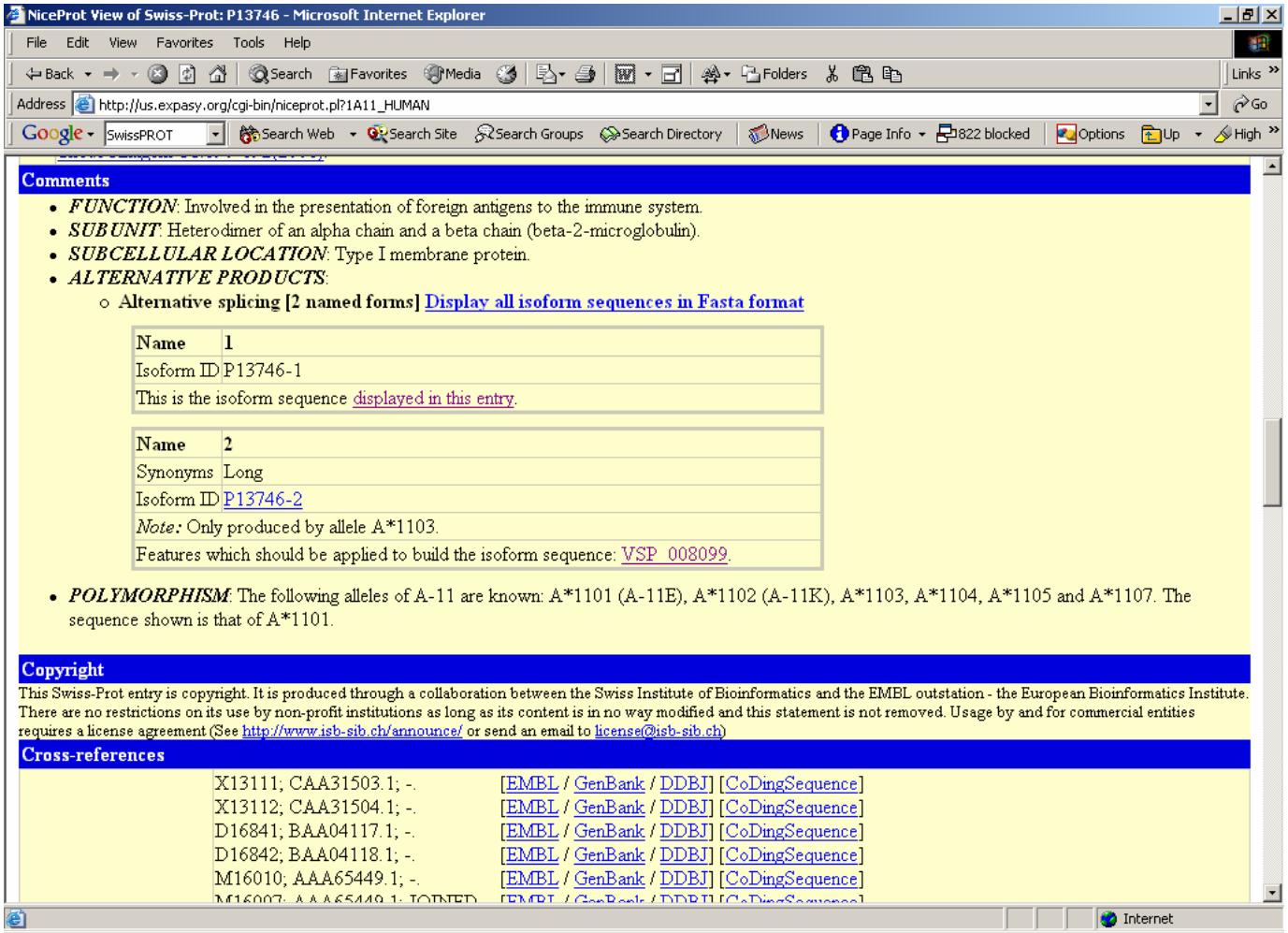
- What if peptide isn't in database?
- Need richer set of peptide candidates
  - Protein isoforms, sequence variants, SNPs, alternate splice forms
  - Some have phenotypic or clinical annotations

Swiss-Prot



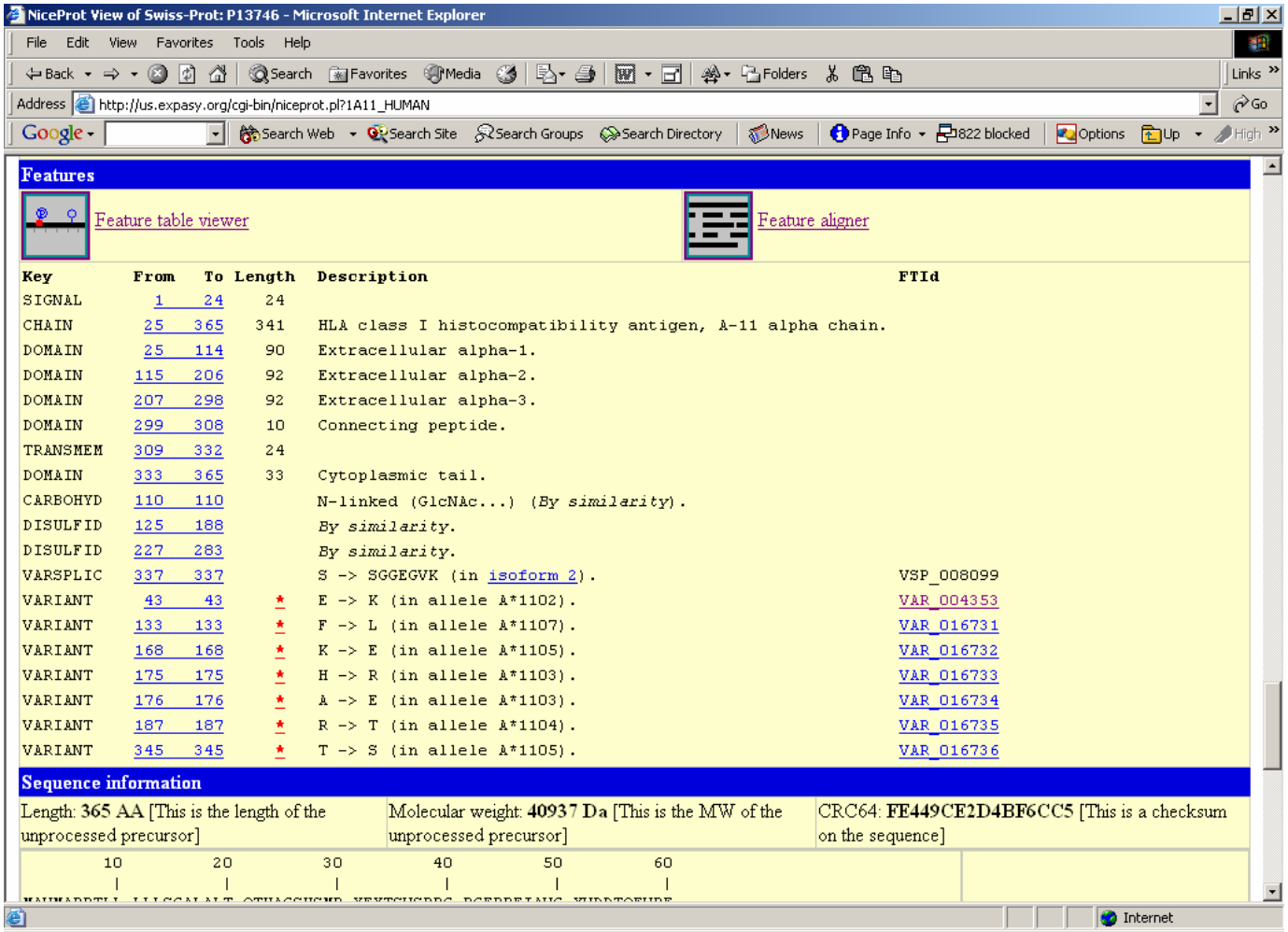
Applied Biosystems

Swiss-Prot Variant Annotations



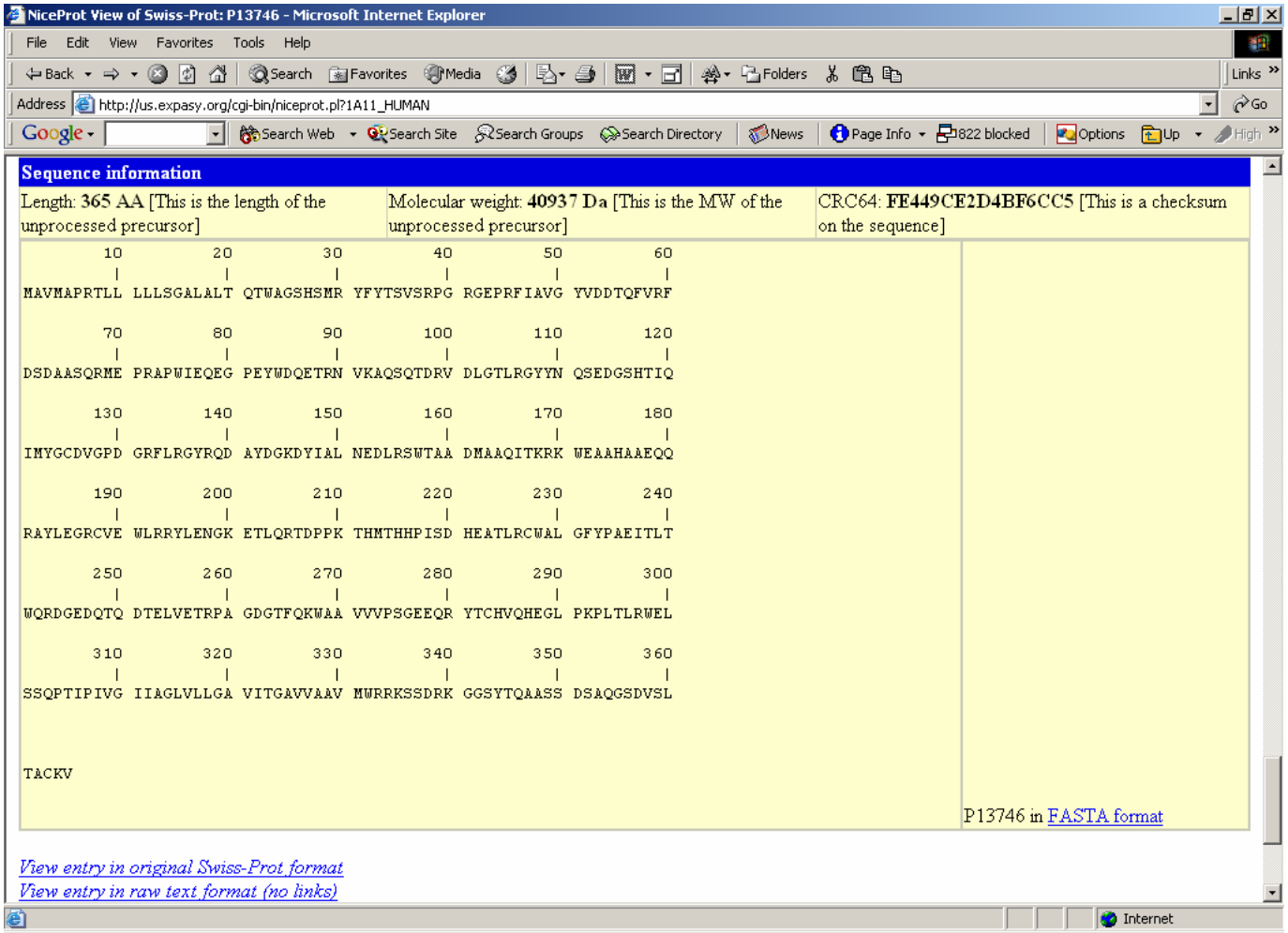
Applied Biosystems

Swiss-Prot Variant Annotations



Applied Biosystems

Swiss-Prot Sequence



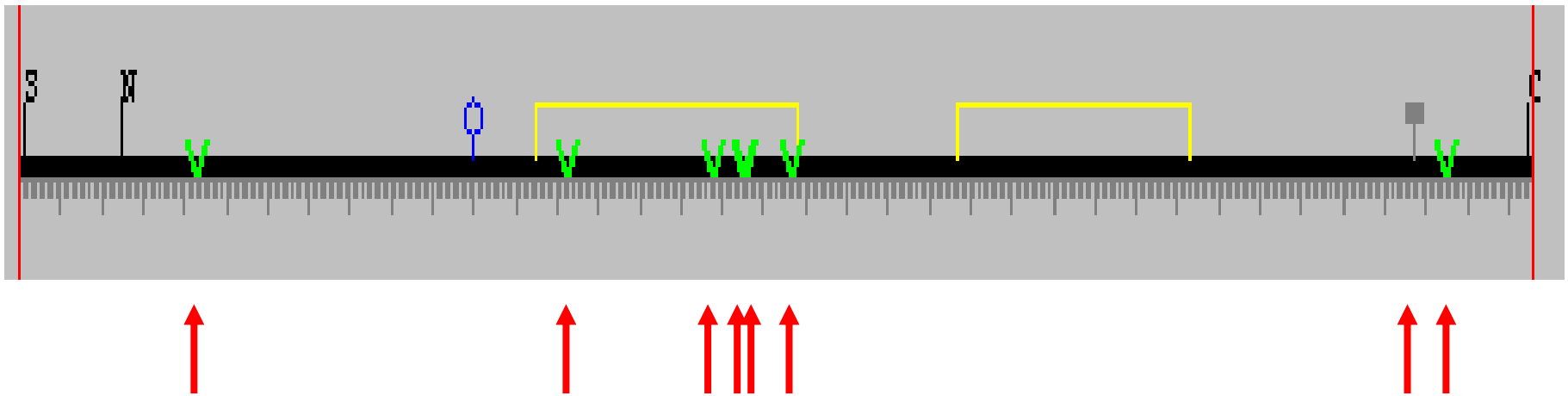
Applied Biosystems

Swiss-Prot

- VarSplic enumerates all variants, conflicts, isoforms
- Swiss-Prot sequence size:
  - 56 Mb
- VarSplic sequence size:
  - 90 Mb
- How many more peptide candidates?

Swiss-Prot
Variant Annotations

Feature viewer



Variants

Applied Biosystems

Swiss-Prot VarSplic Output

```
P13746-00-01-00      MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFYTSVSRPGRGEPRFIAVGYVDDTQFVRF
P13746-01-01-00      MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFYTSVSRPGRGEPRFIAVGYVDDTQFVRF
P13746-00-00-00      MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFYTSVSRPGRGEPRFIAVGYVDDTQFVRF
P13746-00-03-00      MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFYTSVSRPGRGEPRFIAVGYVDDTQFVRF
P13746-01-03-00      MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFYTSVSRPGRGEPRFIAVGYVDDTQFVRF
P13746-00-04-00      MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFYTSVSRPGRGKPRFIAVGYVDDTQFVRF
P13746-01-04-00      MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFYTSVSRPGRGKPRFIAVGYVDDTQFVRF
P13746-00-05-00      MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFYTSVSRPGRGEPRFIAVGYVDDTQFVRF
P13746-01-05-00      MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFYTSVSRPGRGEPRFIAVGYVDDTQFVRF
P13746-01-00-00      MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFYTSVSRPGRGEPRFIAVGYVDDTQFVRF
P13746-00-02-00      MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFYTSVSRPGRGEPRFIAVGYVDDTQFVRF
P13746-01-02-00      MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFYTSVSRPGRGEPRFIAVGYVDDTQFVRF
                     ****************************************:*****************
```

Applied Biosystems

Swiss-Prot VarSplic Output

```
P13746-00-01-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSS------DRKGGSYTQAASSDSAQ
P13746-01-01-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSSGGEGVKDRKGGSYTQAASSDSAQ
P13746-00-00-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSS------DRKGGSYTQAASSDSAQ
P13746-00-03-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSS------DRKGGSYTQAASSDSAQ
P13746-01-03-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSSGGEGVKDRKGGSYTQAASSDSAQ
P13746-00-04-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSS------DRKGGSYTQAASSDSAQ
P13746-01-04-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSSGGEGVKDRKGGSYTQAASSDSAQ
P13746-00-05-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSS------DRKGGSYTQAASSDSAQ
P13746-01-05-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSSGGEGVKDRKGGSYTQAASSDSAQ
P13746-01-00-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSSGGEGVKDRKGGSYTQAASSDSAQ
P13746-00-02-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSS------DRKGGSYSQAASSDSAQ
P13746-01-02-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSSGGEGVKDRKGGSYSQAASSDSAQ
                     ************************************       *******:*********
```

Peptide Candidates

- Parent ion
  - Typically < 3000 Da
- Tryptic Peptides
  - Cut at K or R
- Search engines
  - Don't handle > 4+ well
  - Long peptides don't fragment well
- # of distinct 30-mers upper bounds total peptide content

Peptide Candidates

- At most 2% additional peptides in ~ 1.6 times as much sequence

Sequence Database	Swiss-Prot	VarSplic
Size	56 Mb	90 Mb
30-mers (N_{30})	44 Mb	45 Mb
Overhead	27%	97%

Applied Biosystems

Sequence Database Compression

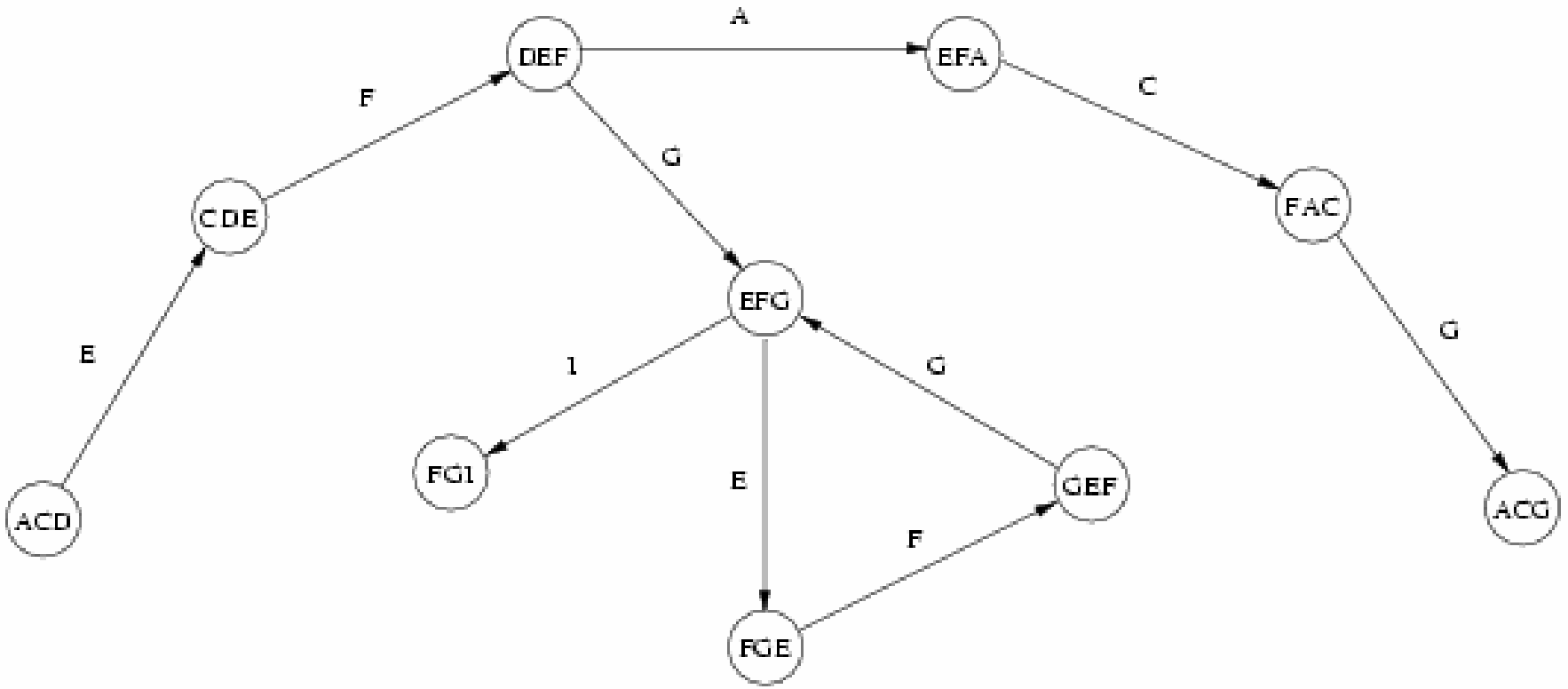
Construct sequence database that is

- Complete
  - All 30-mers are present
- Correct
  - No other 30-mers are present
- Compact
  - No 30-mer is present more than once

Sequence Database Compression

Sequence Database	Swiss-Prot	VarSplic
Original Size	55 Mb	90 Mb
Distinct 30-mers	44 Mb	45 Mb
Overhead	27%	97%
C^3 Size	53 Mb	54 Mb
C^3 Overhead	19%	20%
C^3 Compression	93%	61%
Compression LB	79%	51%

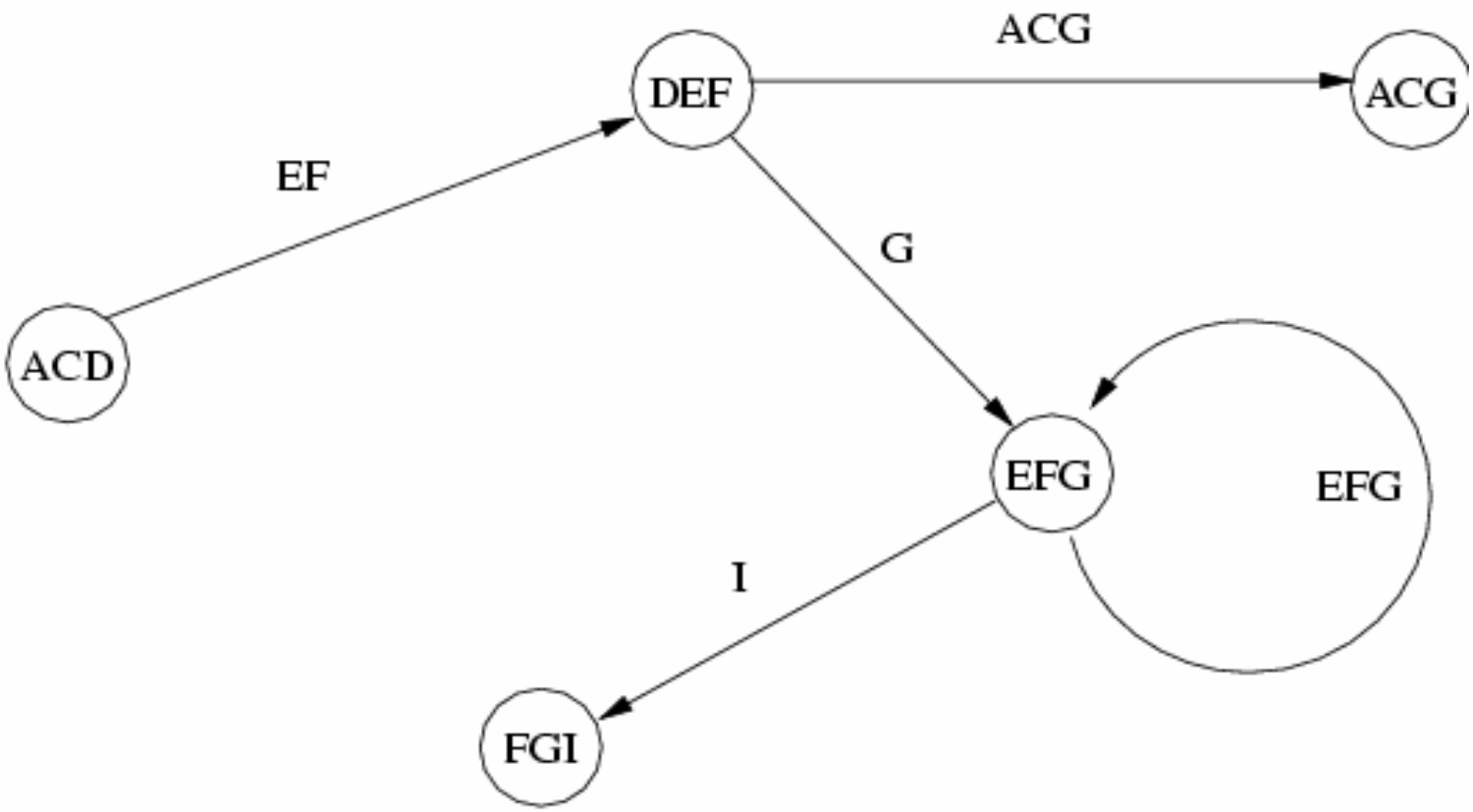
SBH-graph



ACDEFGI, ACDEFACG, DEFGEFGI

Applied Biosystems

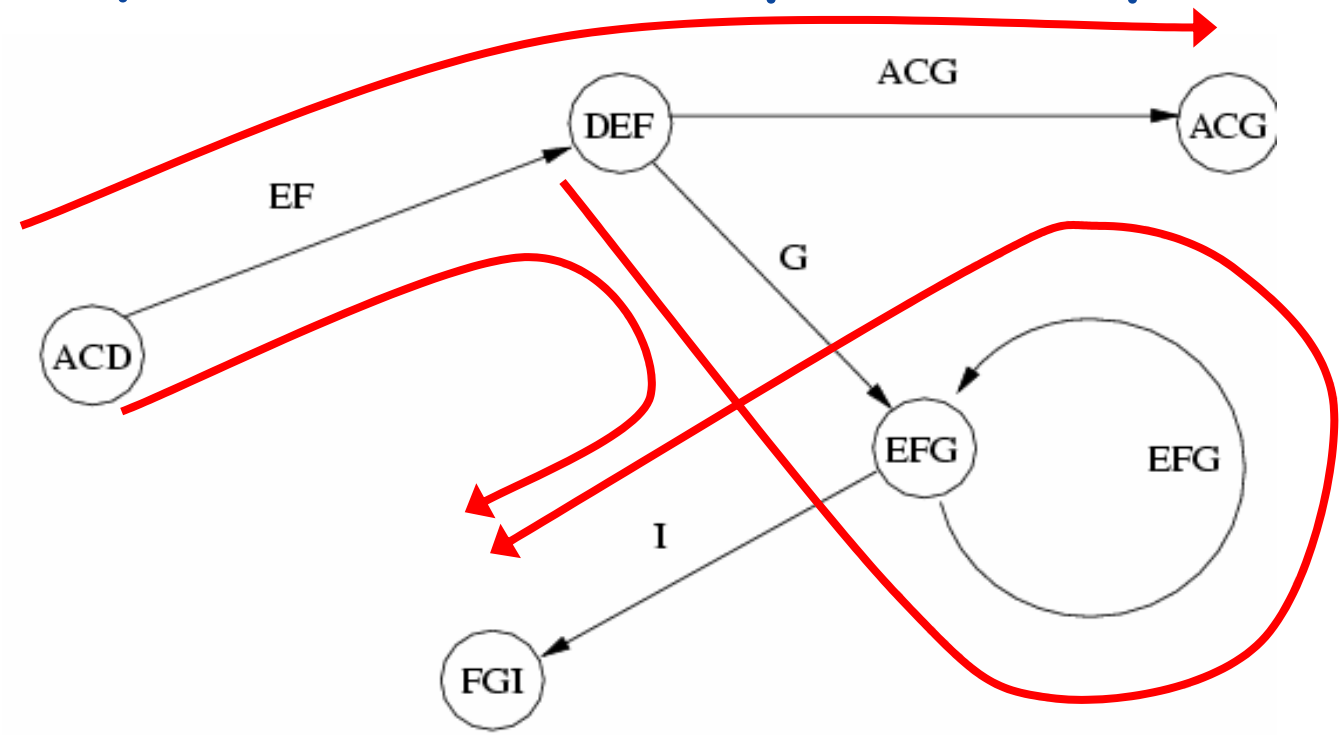
Compressed SBH-graph



ACDEFGI, ACDEFACG, DEFGEFGI

Sequence Databases & CSBH-graphs

- Sequences correspond to paths



ACDEFGI, ACDEFACG, DEFGEFGI

Sequence Databases & CSBH-graphs

- Complete
  - All edges are on some path
- Correct
  - Output path sequence only
- Compact
  - No edge is used more than once

- C^3 Path Set uses all edges exactly once.

Size of C^3 Path Set for k-mers

- Each path costs
  - (k-1)-mer + path sequence + EOS
- Sequence database with p paths
  - $N_k + p\ k$

- Minimize sequence database size by minimizing number of paths
  - subject to C^3 constraints

Best case senario…

…if CSBH-graph admits an Eulerian path.

Sequence database size

$$(k-1) + N_k + 1$$

How many paths are required if the CSBH-graph is not Eulerian?

Non-Eulerian Components

- Net degree
  - b(v) = # in edges - # out edges
- Total degree surplus
  - $B_+ = \sum_{b(v)>0} b(v)$
- For each path
  - Start node's net degree +1
  - End node's net degree -1
  - Otherwise, net degree: no change
- To reduce all nodes to net degree 0, must have at least B_+ paths.

Components w/ $B_+(C) == 0$

- Balanced component must have Eulerian tour, so require exactly one path.
- m balanced components

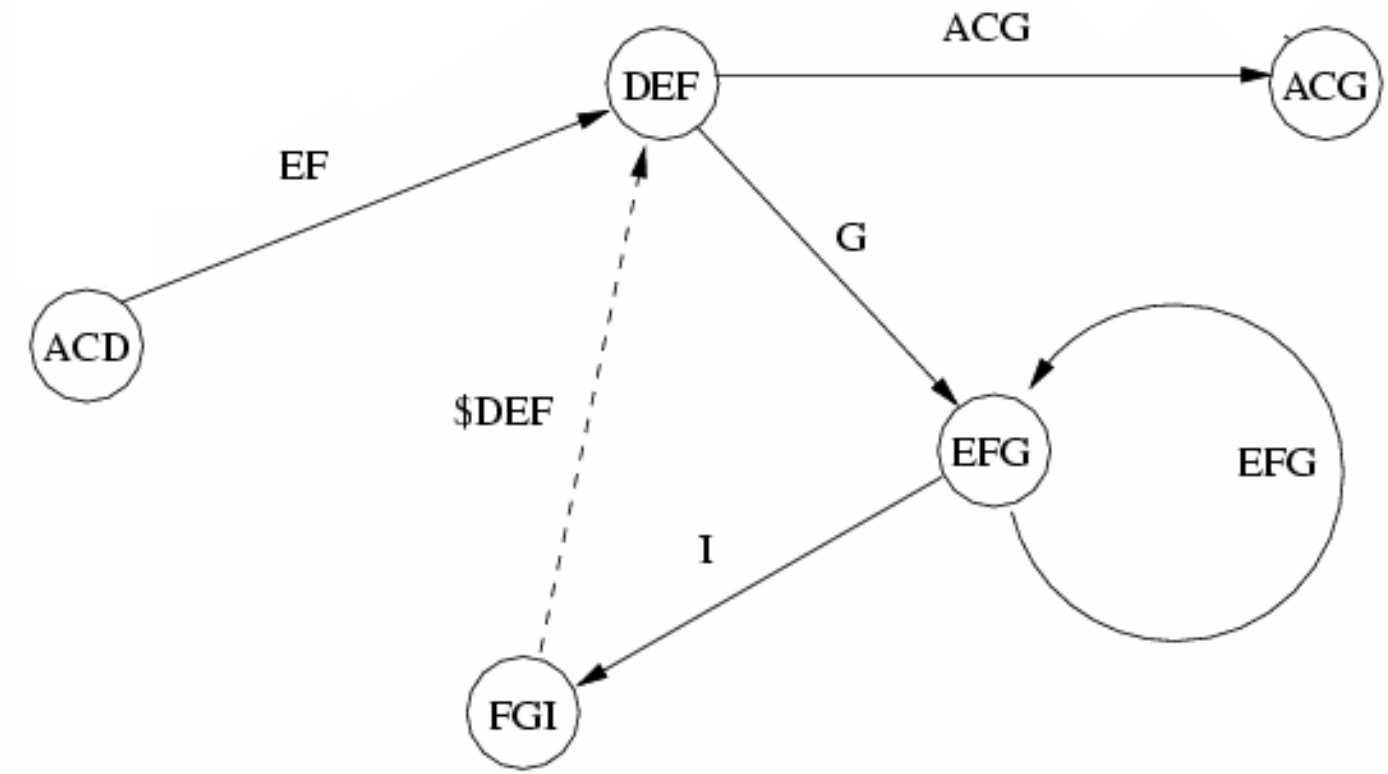
Paths Lower Bound

The C^3 path set must contain at least $B_+ + m$ paths.

This lower bound is achievable!

Just add $(B_+ - 1)$ "restart" edges to non-Eulerian components

Achieving Path Lower Bound



AA Sequence Databases

Sequence Database	Sequence Length	Distinct 30-mers	Overhead
IPI-HUMAN	20358846	12115520	68%
IPI	54145883	29769766	81%
Swiss-Prot	56454588	44374286	27%
Swiss-Prot-VS	89541275	45307827	97%
UniProt	472581860	274510105	72%
UniProt-VS	506796094	275391669	84%
MSDB	481919777	276523755	74%
NRP	495502241	283160529	75%
NCBI-nr	619132252	378721915	63%
UnionNR	674700840	385369671	75%
Union	2157353500	385369671	460%

Minimum Size C^3 Sequence Database

Sequence Database	C^3 30-mer Enumeration	Overhead	Compression	Compression Bound
IPI-HUMAN	13854679	14.35%	68.05%	59.51%
IPI	37961385	27.52%	70.11%	54.98%
Swiss-Prot	52662145	18.68%	93.28%	78.60%
Swiss-Prot-VS	54534356	20.36%	60.90%	50.60%
UniProt	337119564	22.81%	71.34%	58.09%
UniProt-VS	338890778	23.06%	66.87%	54.34%
MSDB	342924164	24.01%	71.16%	57.38%
NRP	351600578	24.17%	70.96%	57.15%
NCBI-nr	463517034	22.39%	74.87%	61.17%
UnionNR	473665310	22.91%	70.20%	57.12%
Union	473665310	22.91%	21.96%	17.86%

Implementation

- Suitable for use by Mascot, SEQUEST, …
  - FASTA format
- All connection to protein context is lost
  - Must do exact string search to find peptides in original database

Extensions

- Drop compactness constraint!
  - Reuse edges rather than starting a new path
  - Similar to the "Chinese Postman Problem"
  - Solvable to optimality using a network flow formulation.

Other Ideas

- We can drop correctness too!
  - Equivalent to shortest substring on the set of 30-mers
- 30-mer subsets
  - …containing two tryptic sites?
  - …containing Cysteine?
- Smaller suffix-tree oracles for short queries