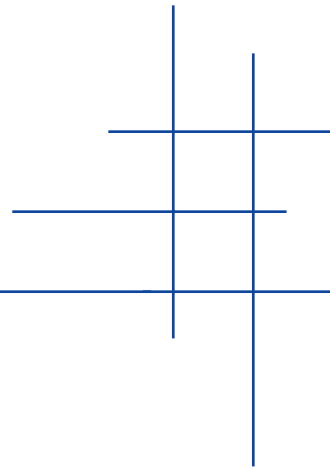


Generating Peptide Candidates from Protein Sequence Databases for Protein Identification via Mass Spectrometry

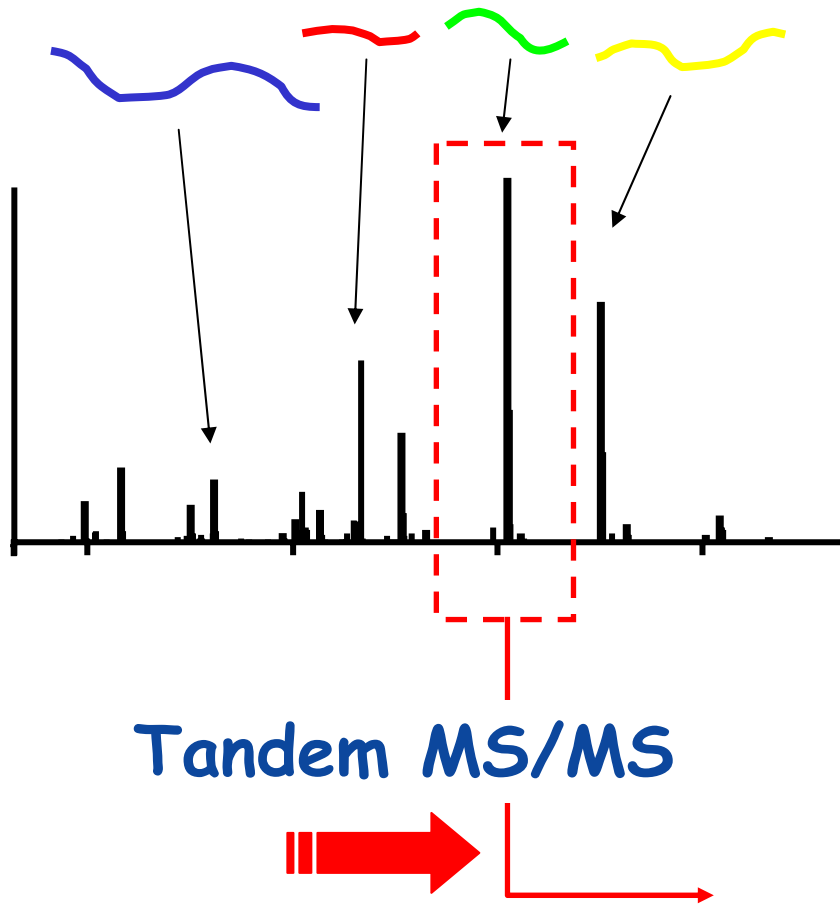
Nathan Edwards
Informatics Research

Protein Identification

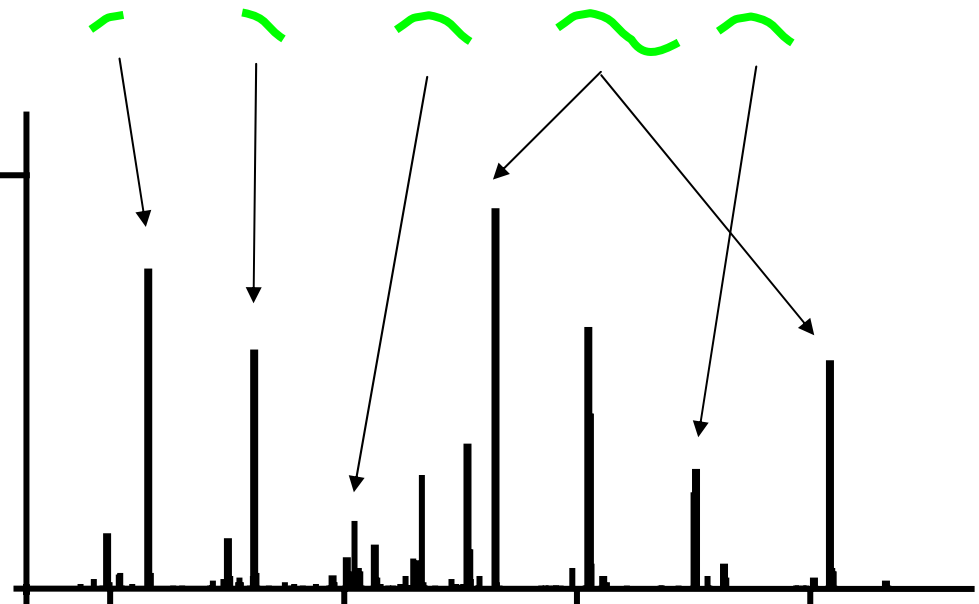


- Turns mass spectrometry into proteomics
- Sequence is link to identity, annotation, literature, genomics
 - Proteomics workflows interrogate more than mass
 - Quality of AA sequence databases sequence & annotation varies wildly
- Protein identification is not BLAST!

LC-MS/MS for Protein Id



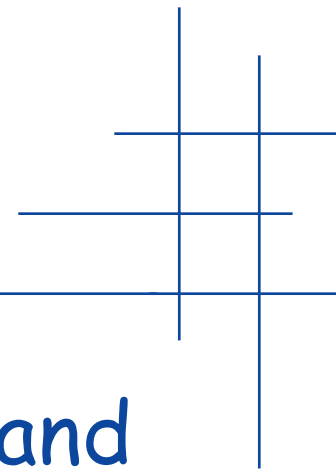
LC-MS/MS: 1 MS spectrum followed by 2-5 Tandem MS/MS spectra every 5-10 sec.



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!**

Sequence Database Search Engines

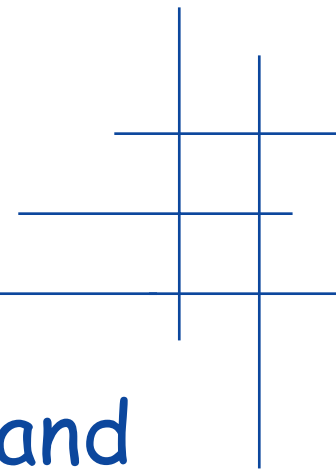


Input: Set of MS/MS spectra and associated parent ion masses

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

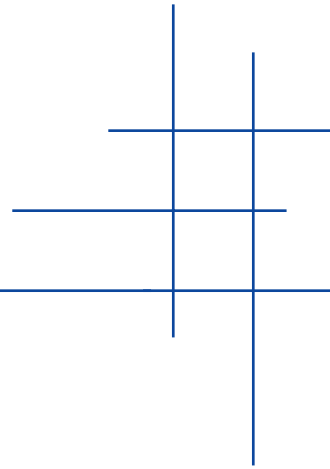
1. Generate peptide candidates from a protein or genomic sequence database
2. Score and rank the peptide candidates

Peptide Candidate Generation

Input: Sequence σ (length n),
from alphabet \mathbf{A}
(Additive) mass $\mu(a)$ for $a \in \mathbf{A}$
Query masses M_1, \dots, M_k

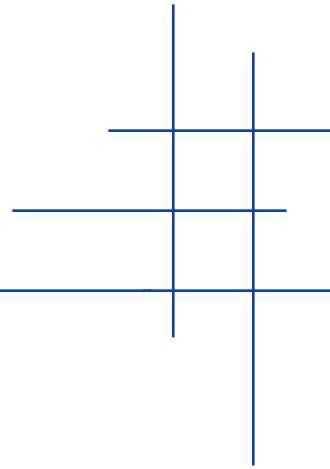
Output: All (distinct) pairs of query
masses i and subsequences ω
with
$$\sum_{j=1}^{|\omega|} \mu(\omega_j) = M_i$$

Peptide Candidate Generation and Peptide Id



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

Simple Linear Scan



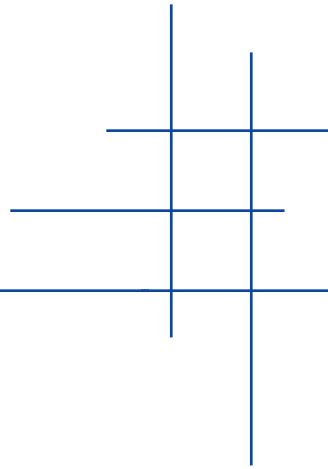
Query Mass = 2018.07

MKWVTFISLLFLFSSAYSRGV...

↑↑ ~~P~~ ~~D~~ ~~S~~ ~~L~~ ~~L~~ ~~F~~ ~~L~~ ~~F~~ ~~S~~ ~~S~~ ~~A~~ ~~Y~~ ~~S~~ ~~R~~ ~~G~~ ~~V~~ ... ↑↑ ~~1~~ ~~8~~ ~~2~~ ~~0~~ ~~2~~ ~~2~~ ~~1~~ ~~7~~ ~~0~~ ~~9~~

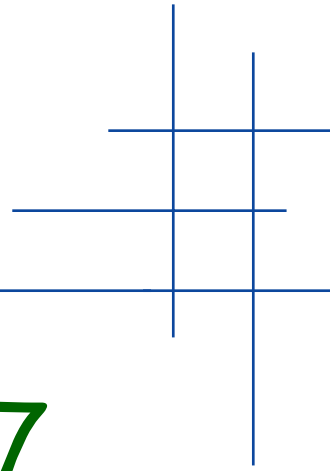
Output: WVTFISLLFLFSSAYSR

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



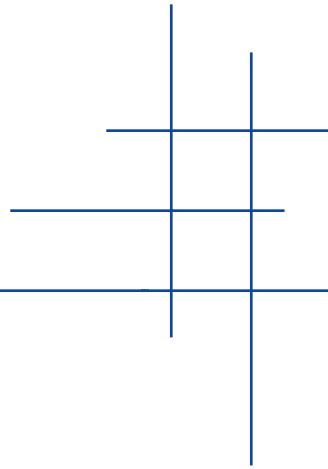
Max Query Mass = 2018.07

MKWVTFISLLFLFSSAYSRGV...

↑↑ ~~1012.128~~ ... ↑↑ 1801.07

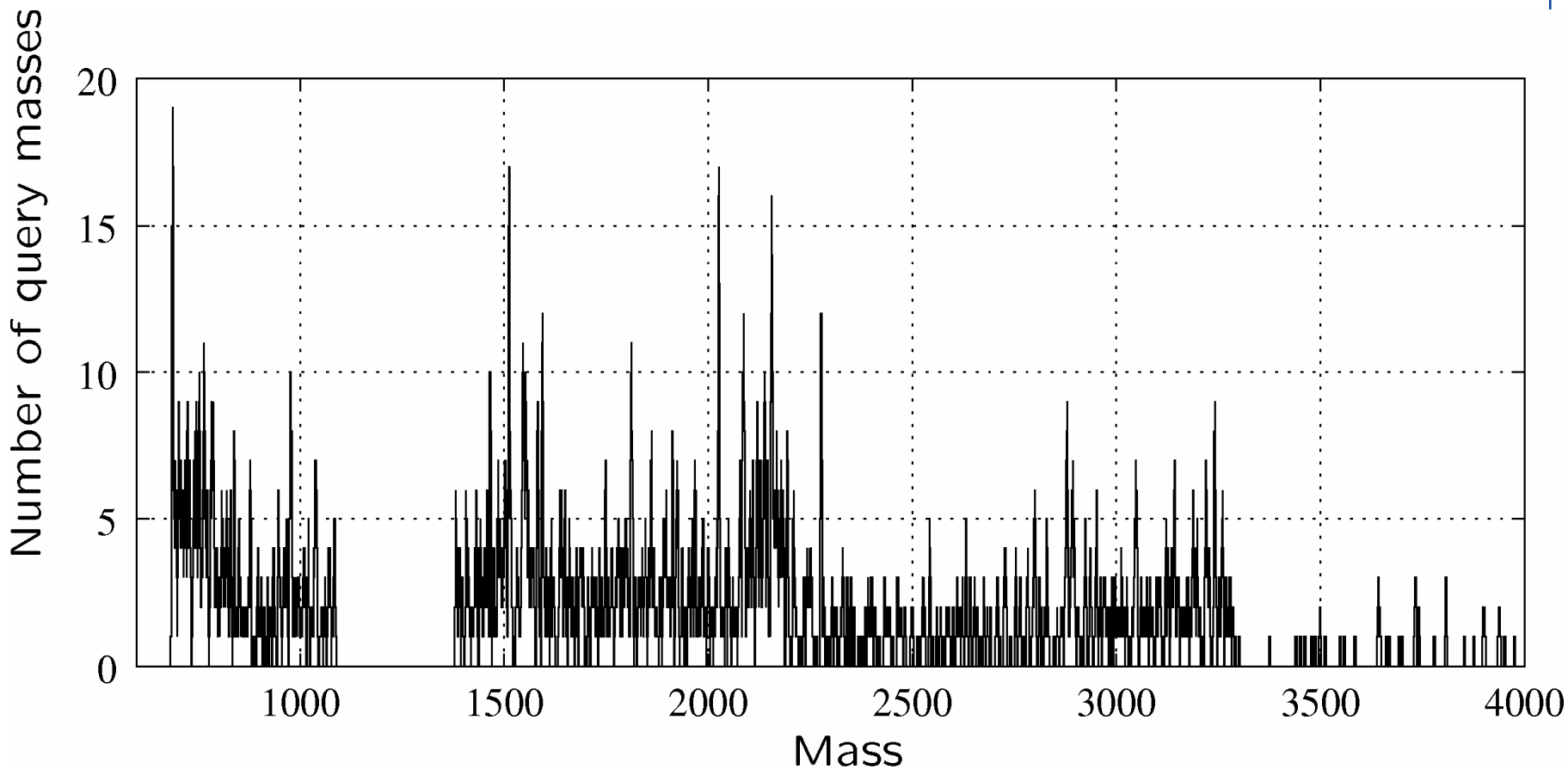
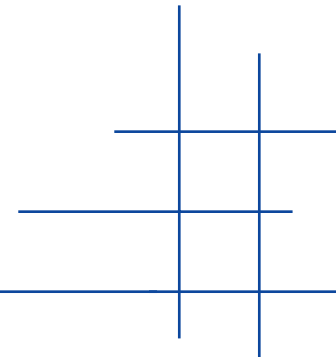
Lookup each candidate mass in turn.

Simultaneous Linear Scan

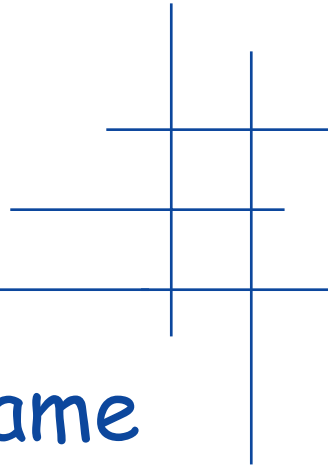


- $O(k \log k + n L \log k)$ time
- Simple to implement
- Easy to track protein context
- Better data locality
- Redundant candidates
- Now a query mass lookup problem!

Overlap Plot from a LC/MS/MS Experiment

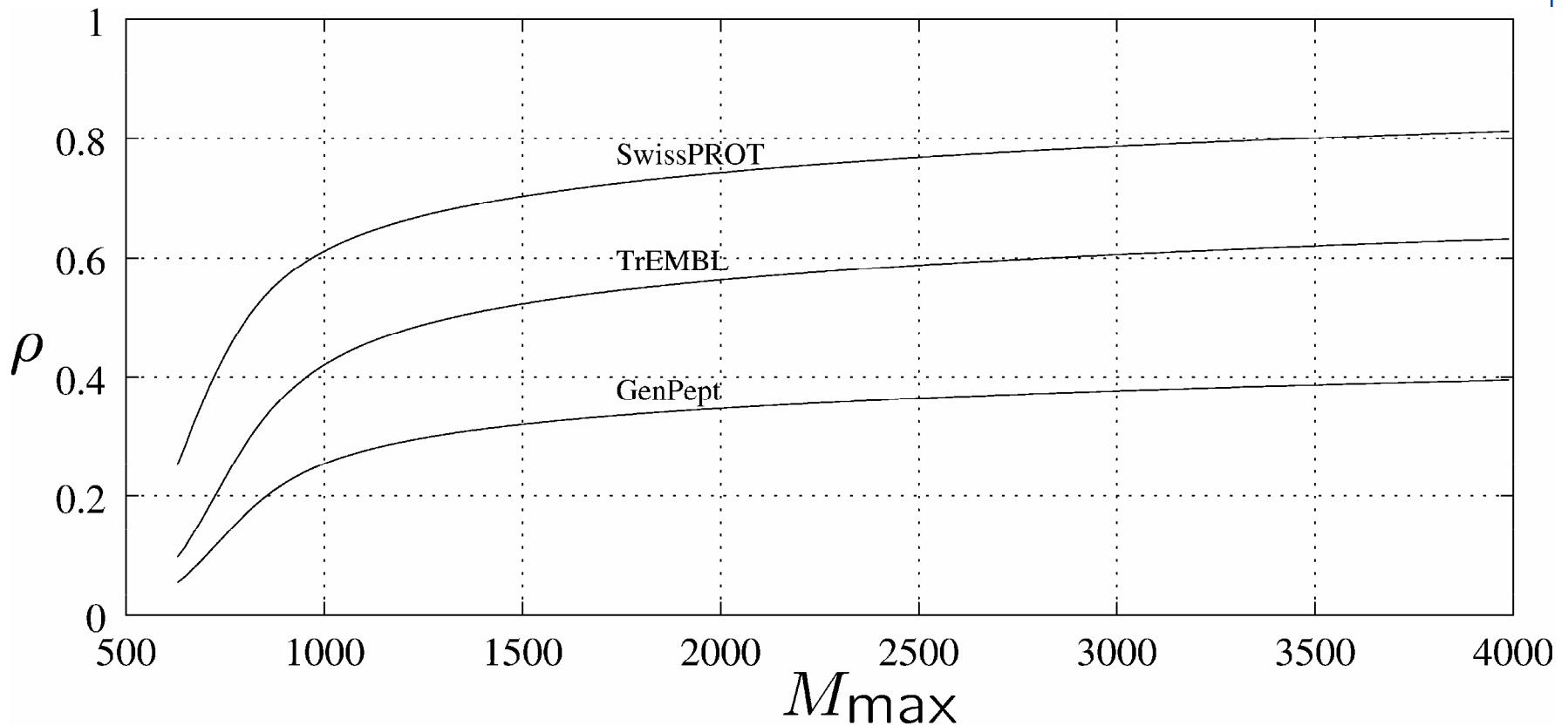


Redundant Candidate Elimination

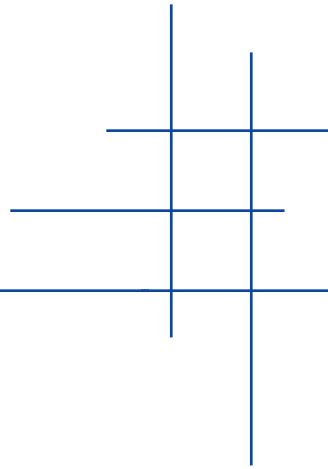


- Must **avoid repeat scoring** of the same peptide candidate
- Want to **avoid generating redundant candidates**
- Non-redundant sequence databases contain lots of **substring redundancy!**

Substring Density (ρ)



Suffix-Tree Traversal



- $O(k \log k + n L \rho \log k)$ time
- Redundancy eliminated
- Tricky to implement well
- Memory overhead $\frac{1}{4} 5n$
- Protein context more involved
- Data locality hard to quantify
- Must preprocess sequence db
- Still a query mass lookup problem!

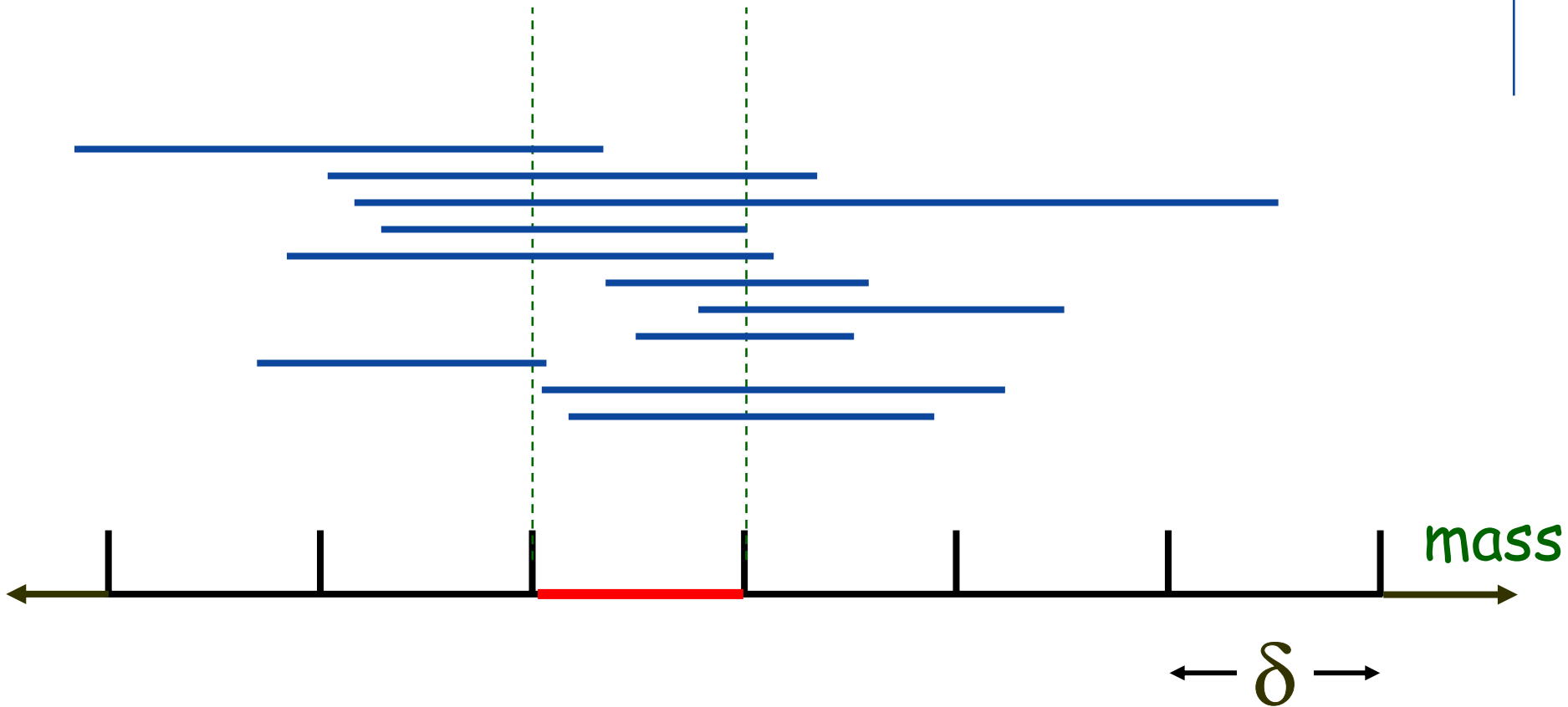
Fast Query Mass Lookup



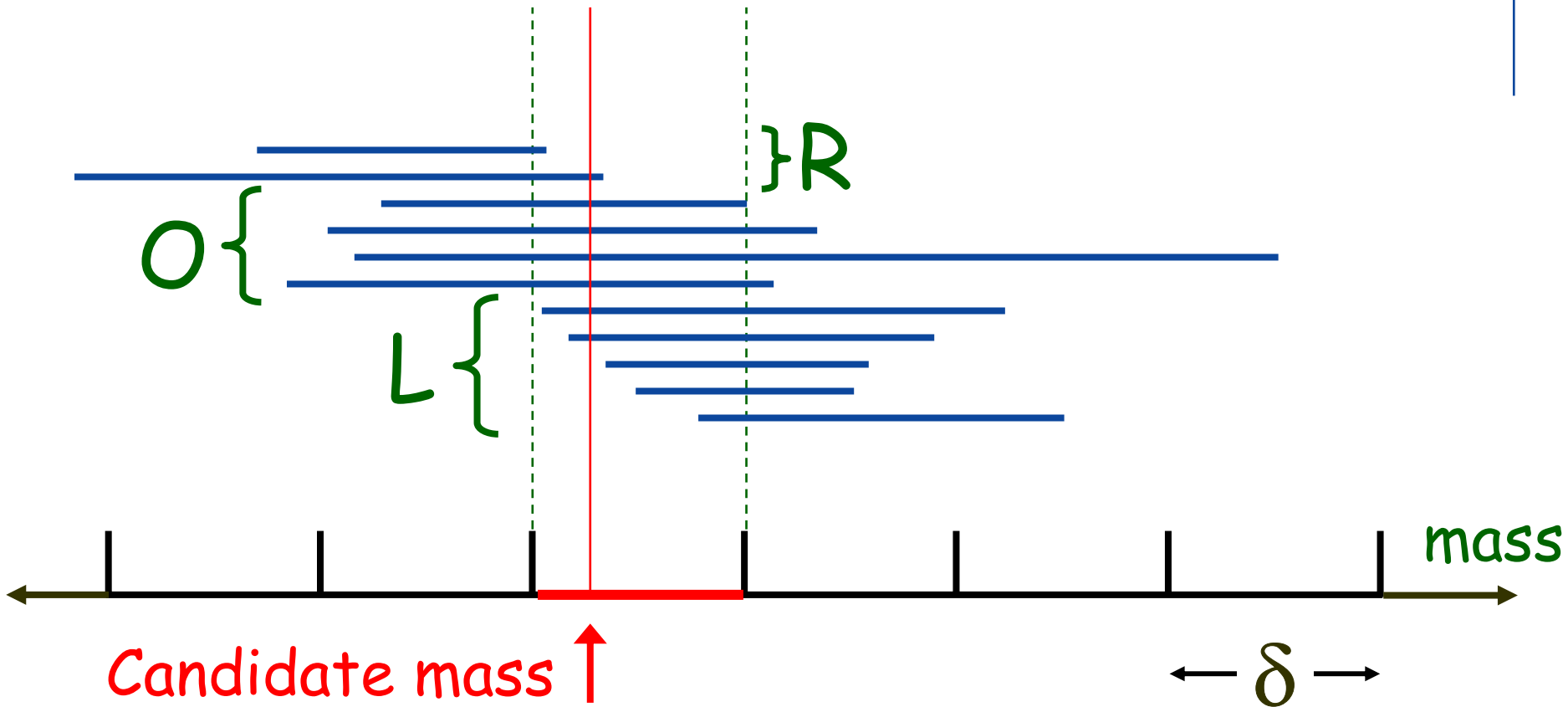
- With (small) integer weights, $O(M_{\max} + k + n L \rho O)$ time is possible
- Use a query mass lookup table!
- Can we achieve this for real weights and non-uniform tolerances?

YES!

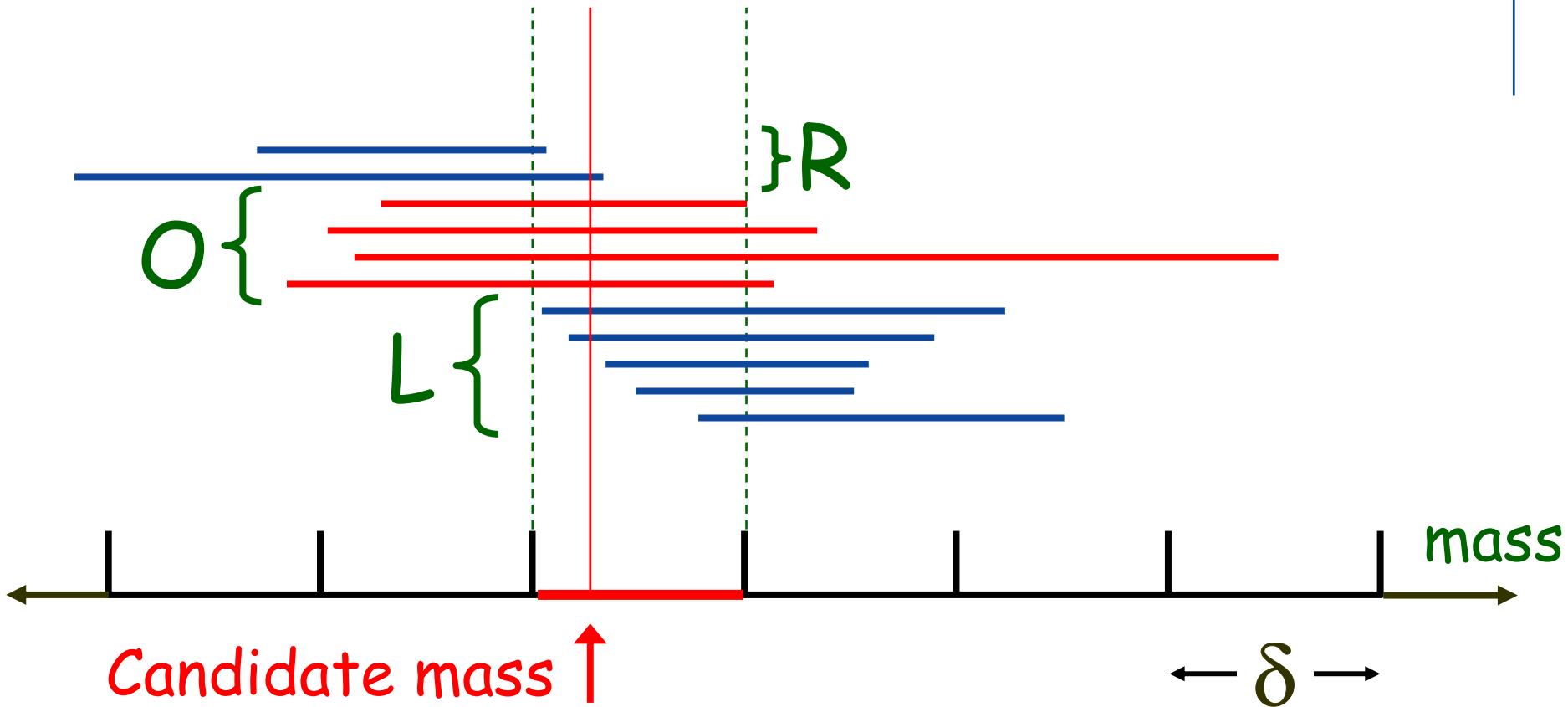
Fast Query Mass Lookup



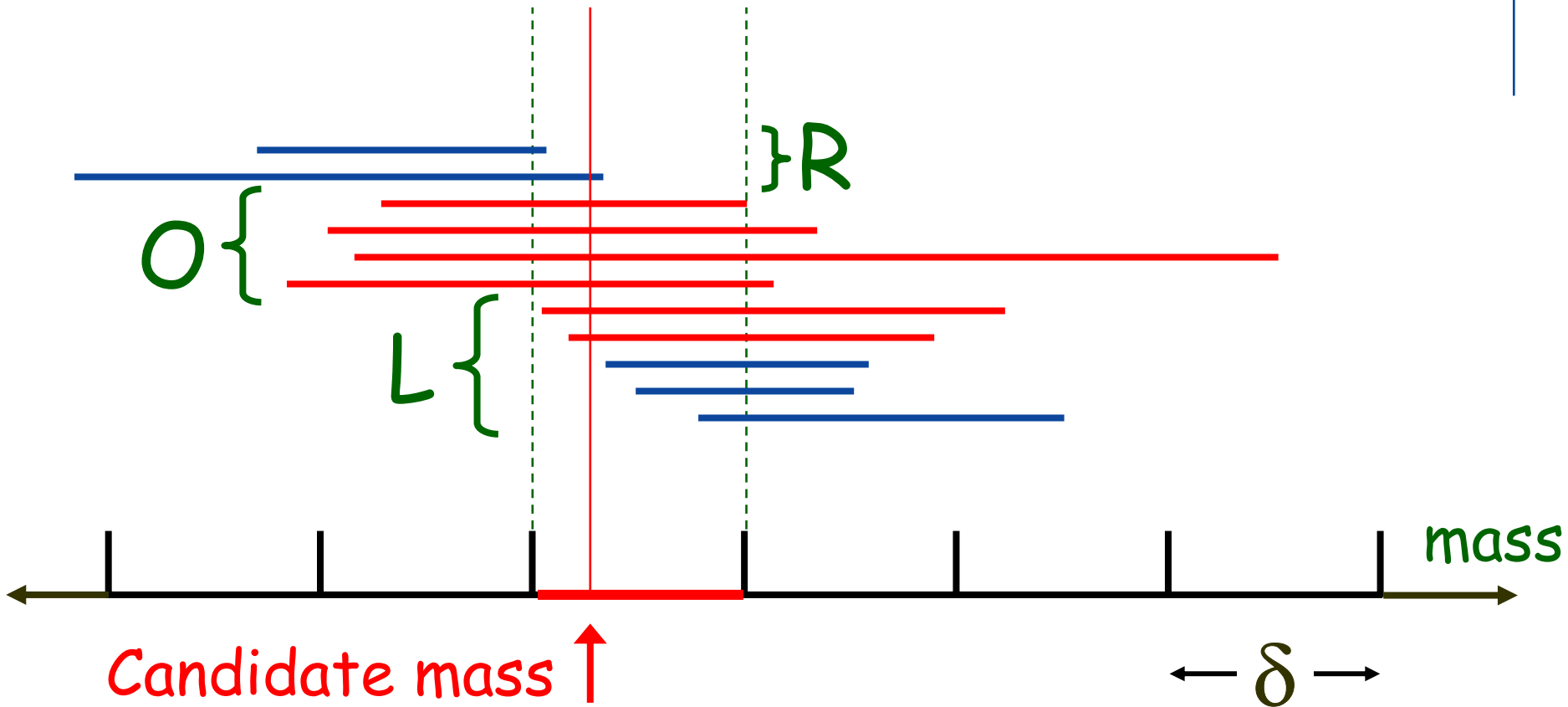
Fast Query Mass Lookup



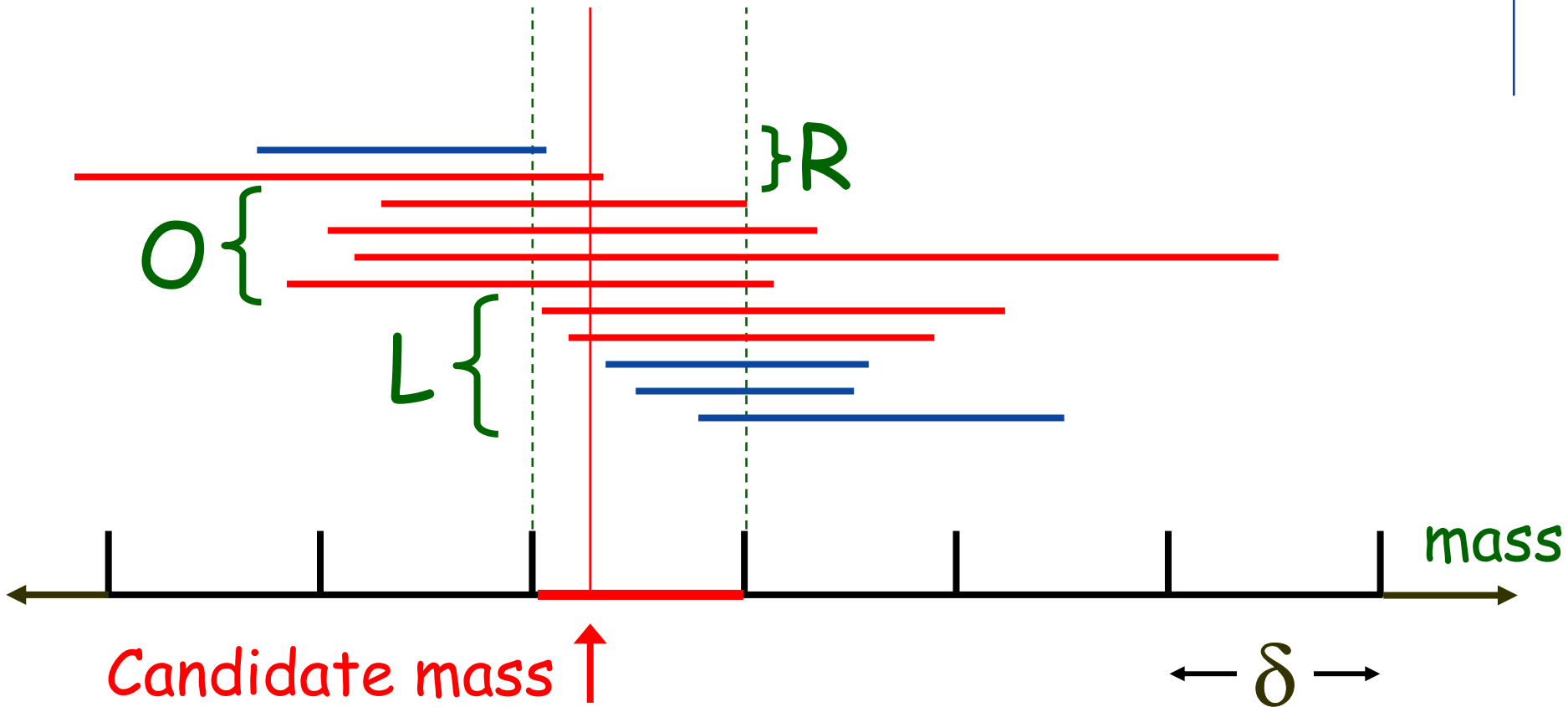
Fast Query Mass Lookup



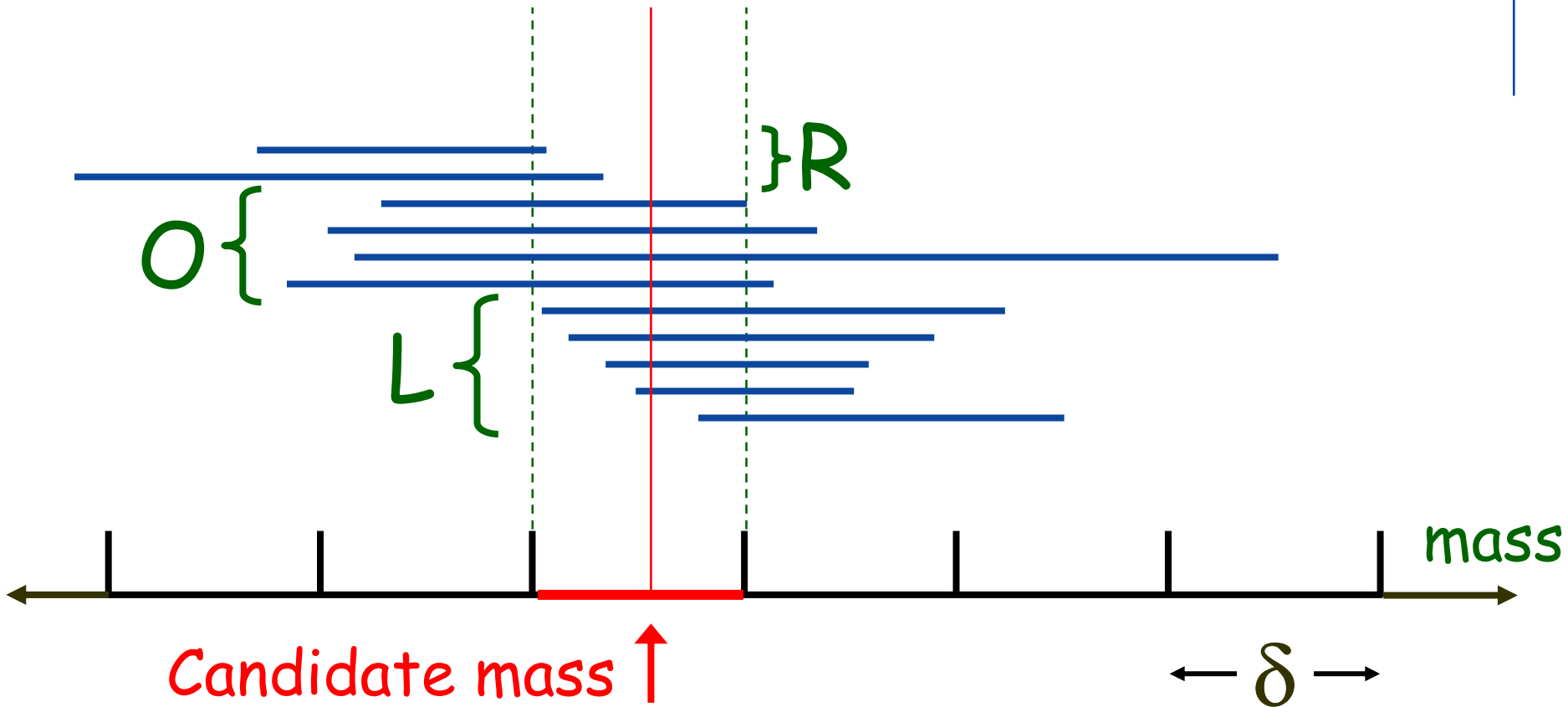
Fast Query Mass Lookup



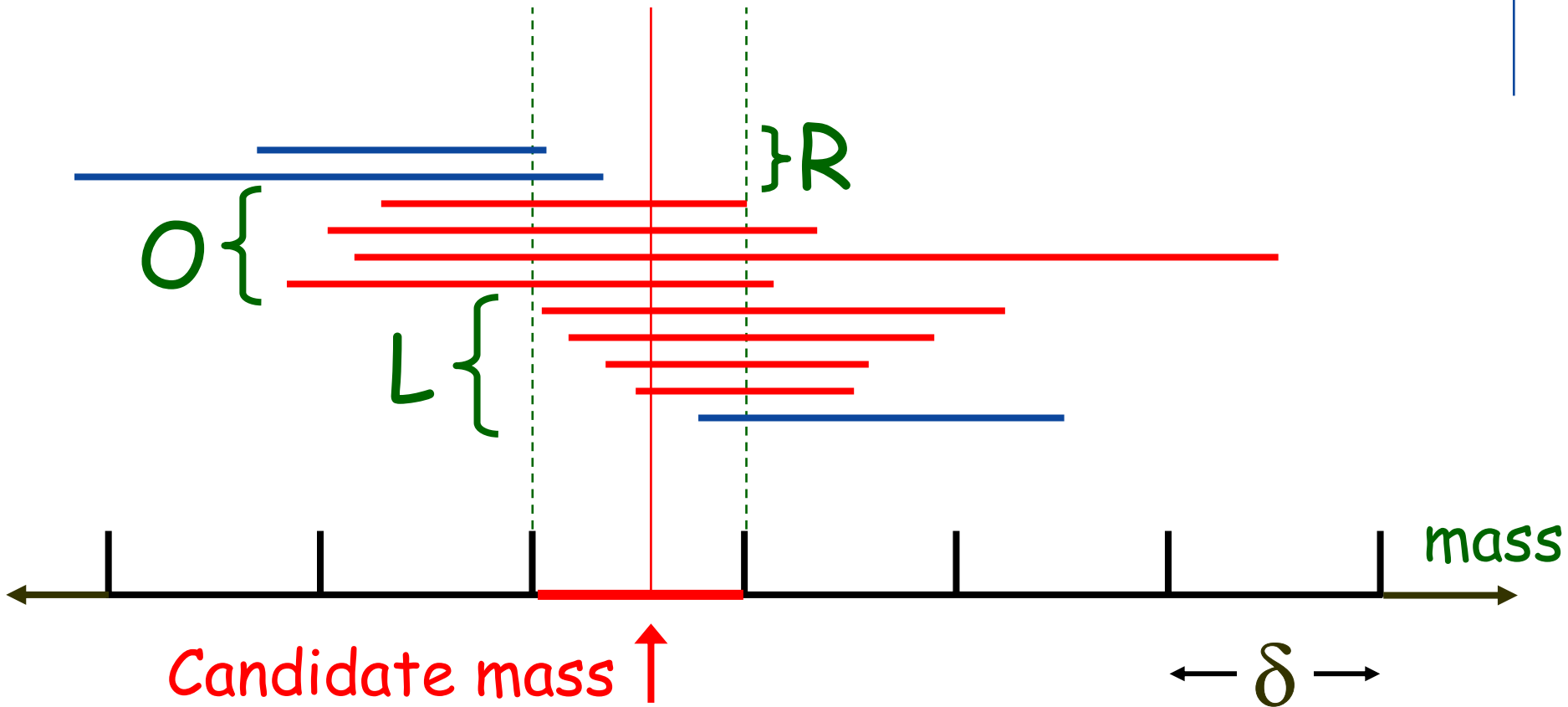
Fast Query Mass Lookup



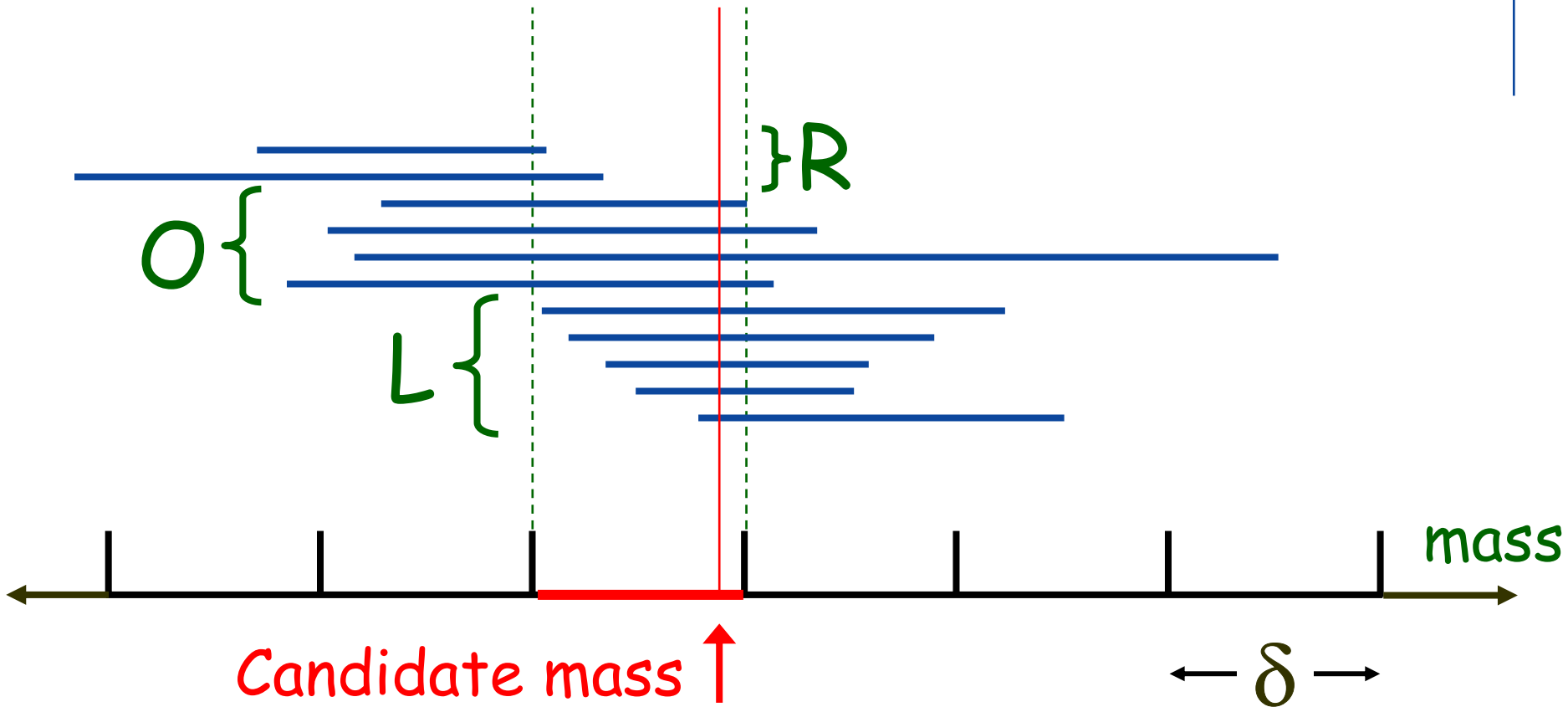
Fast Query Mass Lookup



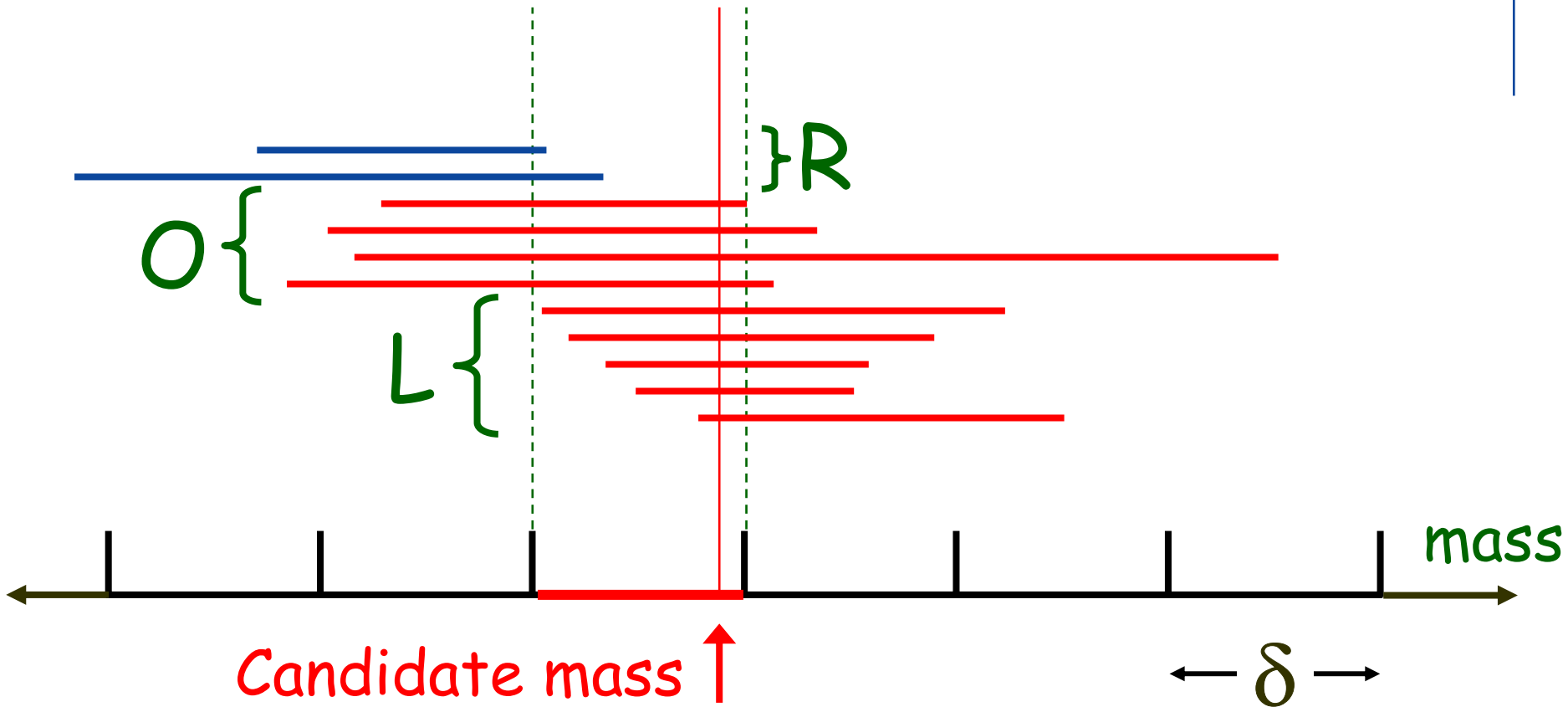
Fast Query Mass Lookup



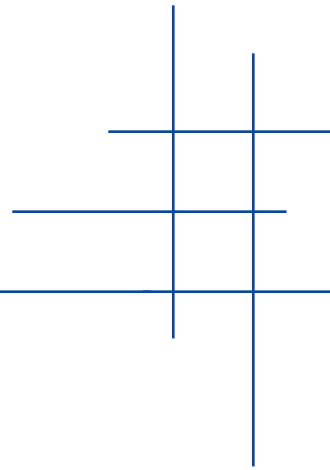
Fast Query Mass Lookup



Fast Query Mass Lookup



Fast Query Mass Lookup

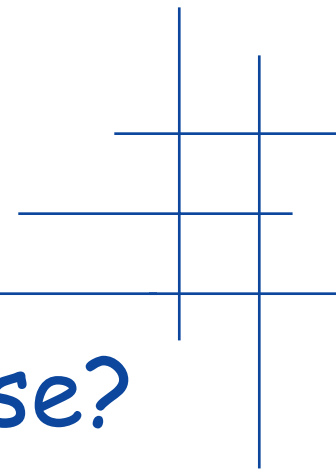


- Must have $\delta \cdot I_{\min}$
- Table size is $O(M_{\max}/\delta + k I_{\max}/\delta)$
- Practical for typical parameters
- Running time:
 - Table construction + $O(n L \rho O)$
 - is dominated by size of output

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

NiceProt View of Swiss-Prot: P13746 - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Search Favorites Media Print Folders Links

Address http://us.expasy.org/cgi-bin/niceprot.pl?1A11_HUMAN Go

Google Search Web Search Site Search Groups Search Directory News Page Info 822 blocked Options Up High

[ExPASy Home page](#) [Site Map](#) [Search ExPASy](#) [Contact us](#) [Swiss-Prot](#)

Hosted by [NCSC US](#) Mirror sites: [Australia](#) [Bolivia](#) [Canada](#) [China](#) [Korea](#) [Switzerland](#) [Taiwan](#)

Search for

NiceProt View of Swiss-Prot: P13746

[\[Entry info\]](#) [\[Name and origin\]](#) [\[References\]](#) [\[Comments\]](#) [\[Cross-references\]](#) [\[Keywords\]](#) [\[Features\]](#) [\[Sequence\]](#) [\[Tools\]](#)

Note: most headings are clickable, even if they don't appear as links. They link to the user manual or other documents.

Entry information

Entry name	1A11_HUMAN
Primary accession number	P13746
Secondary accession numbers	O19605 O19606 Q29747 Q29835 Q9BCN0 Q9MYI5 Q9TQE9 Q9TQP6 Q9TQP7
Entered in Swiss-Prot in	Release 13, January 1990
Sequence was last modified in	Release 13, January 1990
Annotations were last modified in	Release 42, October 2003

Name and origin of the protein

Protein name	HLA class I histocompatibility antigen, A-11 alpha chain [Precursor]
Synonym	MHC class I antigen A*11
Gene name	HLA-A or HLAA
From	Homo sapiens (Human) [TaxID: 9606]
Taxonomy	Eukaryota ; Metazoa ; Chordata ; Craniata ; Vertebrata ; Euteleostomi ; Mammalia ; Eutheria ; Primates ; Catarrhini ; Hominidae ; Homo .

Swiss-Prot Variant Annotations

Comments

- **FUNCTION:** Involved in the presentation of foreign antigens to the immune system.
- **SUBUNIT:** Heterodimer of an alpha chain and a beta chain (beta-2-microglobulin).
- **SUBCELLULAR LOCATION:** Type I membrane protein.
- **ALTERNATIVE PRODUCTS:**
 - Alternative splicing [2 named forms] [Display all isoform sequences in Fasta format](#)

Name	1
Isoform ID	P13746-1
This is the isoform sequence displayed in this entry .	

Name	2
Synonyms	Long
Isoform ID	P13746-2
<i>Note:</i> Only produced by allele A*1103.	
Features which should be applied to build the isoform sequence: VSP_008099 .	

- **POLYMORPHISM:** The following alleles of A-11 are known: A*1101 (A-11E), A*1102 (A-11K), A*1103, A*1104, A*1105 and A*1107. The sequence shown is that of A*1101.

Copyright

This Swiss-Prot entry is copyright. It is produced through a collaboration between the Swiss Institute of Bioinformatics and the EMBL outstation - the European Bioinformatics Institute. There are no restrictions on its use by non-profit institutions as long as its content is in no way modified and this statement is not removed. Usage by and for commercial entities requires a license agreement (See <http://www.isb-sib.ch/announce/> or send an email to license@isb-sib.ch)

Cross-references

X13111; CAA31503.1; -	[EMBL / GenBank / DDBJ] [CoDingSequence]
X13112; CAA31504.1; -	[EMBL / GenBank / DDBJ] [CoDingSequence]
D16841; BAA04117.1; -	[EMBL / GenBank / DDBJ] [CoDingSequence]
D16842; BAA04118.1; -	[EMBL / GenBank / DDBJ] [CoDingSequence]
M16010; AAA65449.1; -	[EMBL / GenBank / DDBJ] [CoDingSequence]
M16007; AAA65449.1; JOINED	[EMBL / GenBank / DDBJ] [CoDingSequence]

Swiss-Prot Variant Annotations

NiceProt View of Swiss-Prot: P13746 - Microsoft Internet Explorer

Address: http://us.expasy.org/cgi-bin/niceprot.pl?P1A11_HUMAN

Features

[Feature table viewer](#) [Feature aligner](#)

Key	From	To	Length	Description	FTId
SIGNAL	1	24	24		
CHAIN	25	365	341	HLA class I histocompatibility antigen, A-11 alpha chain.	
DOMAIN	25	114	90	Extracellular alpha-1.	
DOMAIN	115	206	92	Extracellular alpha-2.	
DOMAIN	207	298	92	Extracellular alpha-3.	
DOMAIN	299	308	10	Connecting peptide.	
TRANSMEM	309	332	24		
DOMAIN	333	365	33	Cytoplasmic tail.	
CARBOHYD	110	110		N-linked (GlcNAc...) (By similarity).	
DISULFID	125	188		By similarity.	
DISULFID	227	283		By similarity.	
VARSPPLIC	337	337		S -> SGGEGVK (in isoform 2).	VSP_008099
VARIANT	43	43	*	E -> K (in allele A*1102).	VAR_004353
VARIANT	133	133	*	F -> L (in allele A*1107).	VAR_016731
VARIANT	168	168	*	K -> E (in allele A*1105).	VAR_016732
VARIANT	175	175	*	H -> R (in allele A*1103).	VAR_016733
VARIANT	176	176	*	A -> E (in allele A*1103).	VAR_016734
VARIANT	187	187	*	R -> T (in allele A*1104).	VAR_016735
VARIANT	345	345	*	T -> S (in allele A*1105).	VAR_016736

Sequence information

Length: 365 AA [This is the length of the unprocessed precursor] Molecular weight: 40937 Da [This is the MW of the unprocessed precursor] CRC64: FE449CE2D4BF6CC5 [This is a checksum on the sequence]

10 20 30 40 50 60

MANKAPRLLIILGALALTEGTHAGGSHRVEYFCHSDRDCRERRELAHQWIDTQSFRE

Swiss-Prot Sequence

Sequence information

Length: 365 AA [This is the length of the unprocessed precursor] Molecular weight: 40937 Da [This is the MW of the unprocessed precursor] CRC64: FE449CE2D4BF6CC5 [This is a checksum on the sequence]

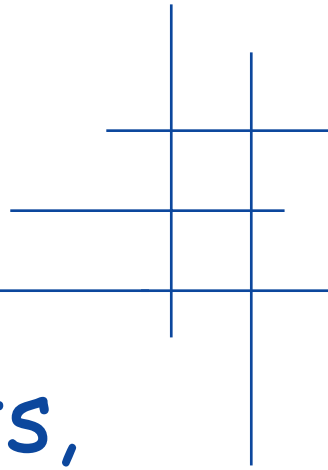
10	20	30	40	50	60
MAVMAPRTLL	LLLSGALALT	QTWAGSHSMR	YFYTSVSRPG	RGEPRFIAVG	YVDDTQFVRF
70	80	90	100	110	120
DSDAASQRME	PRAPWIEQEG	PEYWDQETRN	VKAQSQTRV	DLGTLRGYYN	QSEDGSHTIQ
130	140	150	160	170	180
IMYGCDVGPD	GRFLRGYRQD	AYDGKDYIAL	NEDLRSWTA A	DMAAQITKRK	WEAHAAEQQ
190	200	210	220	230	240
RAYLEGRCVE	WLRRYLENGK	ETLQRTDPPK	THMTHHPISD	HEATLRCWAL	GFYPAEITLT
250	260	270	280	290	300
WQRDGEDQTQ	DTELVETRPA	GDGTFQKWA A	VVVP SGEEQR	YTCHVQHEGL	PKPLTLRWEL
310	320	330	340	350	360
SSQPTPIVIG	IAGLVLLGA	VITGAVVA AV	MWRKSSDRK	GGSYTAASS	DSAQGSVSL

TACKV

P13746 in [FASTA format](#)

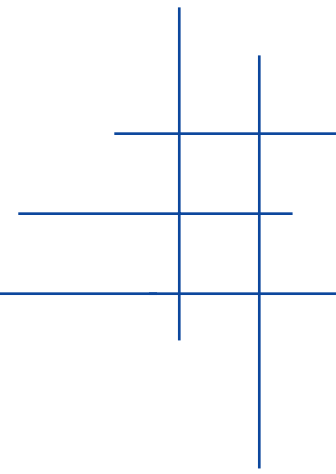
[View entry in original Swiss-Prot format](#)
[View entry in raw text format \(no links\)](#)

Swiss-Prot

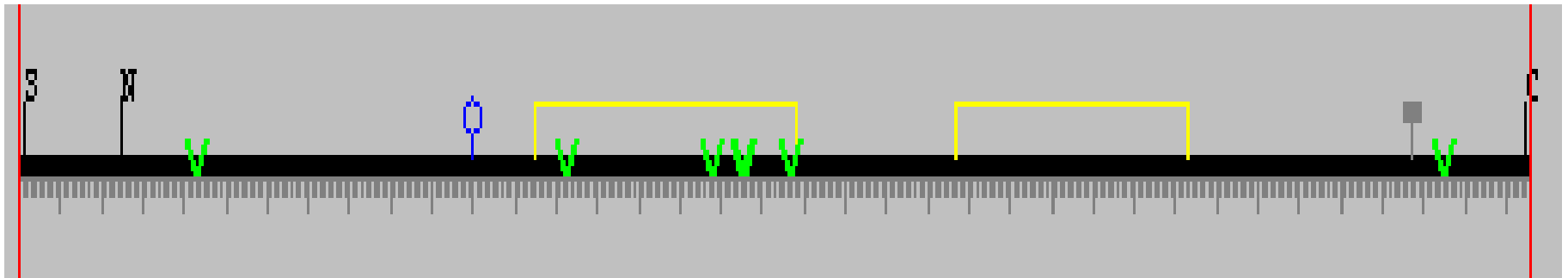


- VarSplic enumerates all variants, conflicts, isoforms
- Swiss-Prot sequence size:
 - 56 Mb
- VarSplic sequence size:
 - 90 Mb
- How many more peptide candidates?

Swiss-Prot Variant Annotations

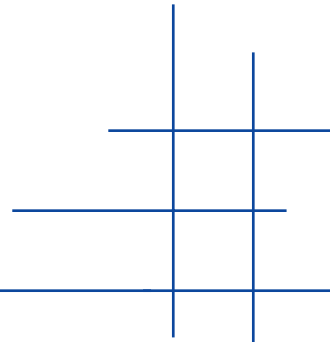


Feature viewer



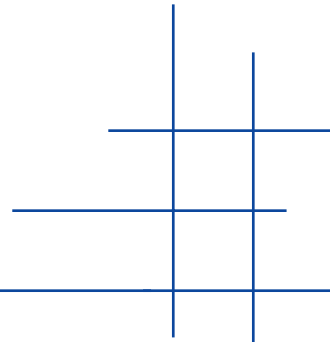
Variants

Swiss-Prot VarSplic Output



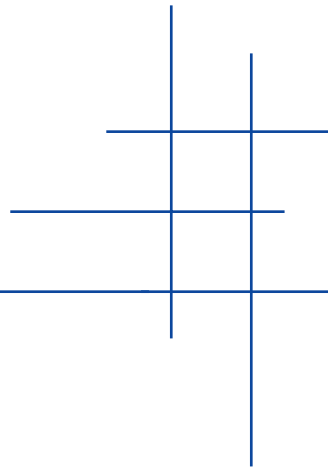
```
P13746-00-01-00      MAVMAPRTL L L L L L S G A L A L T Q T W A G S H S M R Y F Y T S V S R P G R G E P R F I A V G Y V D D T Q F V R F
P13746-01-01-00      MAVMAPRTL L L L L L S G A L A L T Q T W A G S H S M R Y F Y T S V S R P G R G E P R F I A V G Y V D D T Q F V R F
P13746-00-00-00      MAVMAPRTL L L L L L S G A L A L T Q T W A G S H S M R Y F Y T S V S R P G R G E P R F I A V G Y V D D T Q F V R F
P13746-00-03-00      MAVMAPRTL L L L L L S G A L A L T Q T W A G S H S M R Y F Y T S V S R P G R G E P R F I A V G Y V D D T Q F V R F
P13746-01-03-00      MAVMAPRTL L L L L L S G A L A L T Q T W A G S H S M R Y F Y T S V S R P G R G E P R F I A V G Y V D D T Q F V R F
P13746-00-04-00      MAVMAPRTL L L L L L S G A L A L T Q T W A G S H S M R Y F Y T S V S R P G R G K P R F I A V G Y V D D T Q F V R F
P13746-01-04-00      MAVMAPRTL L L L L L S G A L A L T Q T W A G S H S M R Y F Y T S V S R P G R G K P R F I A V G Y V D D T Q F V R F
P13746-00-05-00      MAVMAPRTL L L L L L S G A L A L T Q T W A G S H S M R Y F Y T S V S R P G R G E P R F I A V G Y V D D T Q F V R F
P13746-01-05-00      MAVMAPRTL L L L L L S G A L A L T Q T W A G S H S M R Y F Y T S V S R P G R G E P R F I A V G Y V D D T Q F V R F
P13746-01-00-00      MAVMAPRTL L L L L L S G A L A L T Q T W A G S H S M R Y F Y T S V S R P G R G E P R F I A V G Y V D D T Q F V R F
P13746-00-02-00      MAVMAPRTL L L L L L S G A L A L T Q T W A G S H S M R Y F Y T S V S R P G R G E P R F I A V G Y V D D T Q F V R F
P13746-01-02-00      MAVMAPRTL L L L L L S G A L A L T Q T W A G S H S M R Y F Y T S V S R P G R G E P R F I A V G Y V D D T Q F V R F
***** : *****
```

Swiss-Prot VarSplic Output



```
P13746-00-01-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRKSS-----DRKGGSYTQAASSDSAQ
P13746-01-01-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRKSSGGEGVKDRKGGSYTQAASSDSAQ
P13746-00-00-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRKSS-----DRKGGSYTQAASSDSAQ
P13746-00-03-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRKSS-----DRKGGSYTQAASSDSAQ
P13746-01-03-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRKSSGGEGVKDRKGGSYTQAASSDSAQ
P13746-00-04-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRKSS-----DRKGGSYTQAASSDSAQ
P13746-01-04-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRKSSGGEGVKDRKGGSYTQAASSDSAQ
P13746-00-05-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRKSS-----DRKGGSYTQAASSDSAQ
P13746-01-05-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRKSSGGEGVKDRKGGSYTQAASSDSAQ
P13746-01-00-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRKSSGGEGVKDRKGGSYTQAASSDSAQ
P13746-00-02-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRKSS-----DRKGGSYSQAASSDSAQ
P13746-01-02-00      SSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRKSSGGEGVKDRKGGSYSQAASSDSAQ
*****
***** :
```

Peptide Candidates



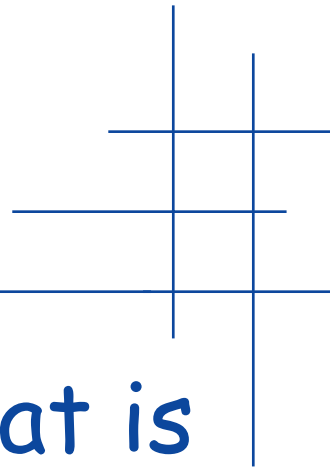
- Parent ion
 - Typically < 3000 Da
- Tryptic Peptides
 - Cut at K or R
- Search engines
 - Don't handle > 4+ well
 - Long peptides don't fragment well
- # of distinct 30-mers upper bounds total peptide content

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%

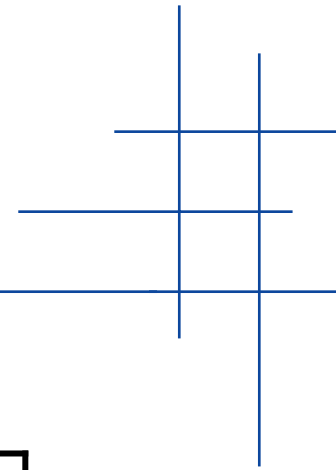
Sequence Database Compression



Construct sequence database that is

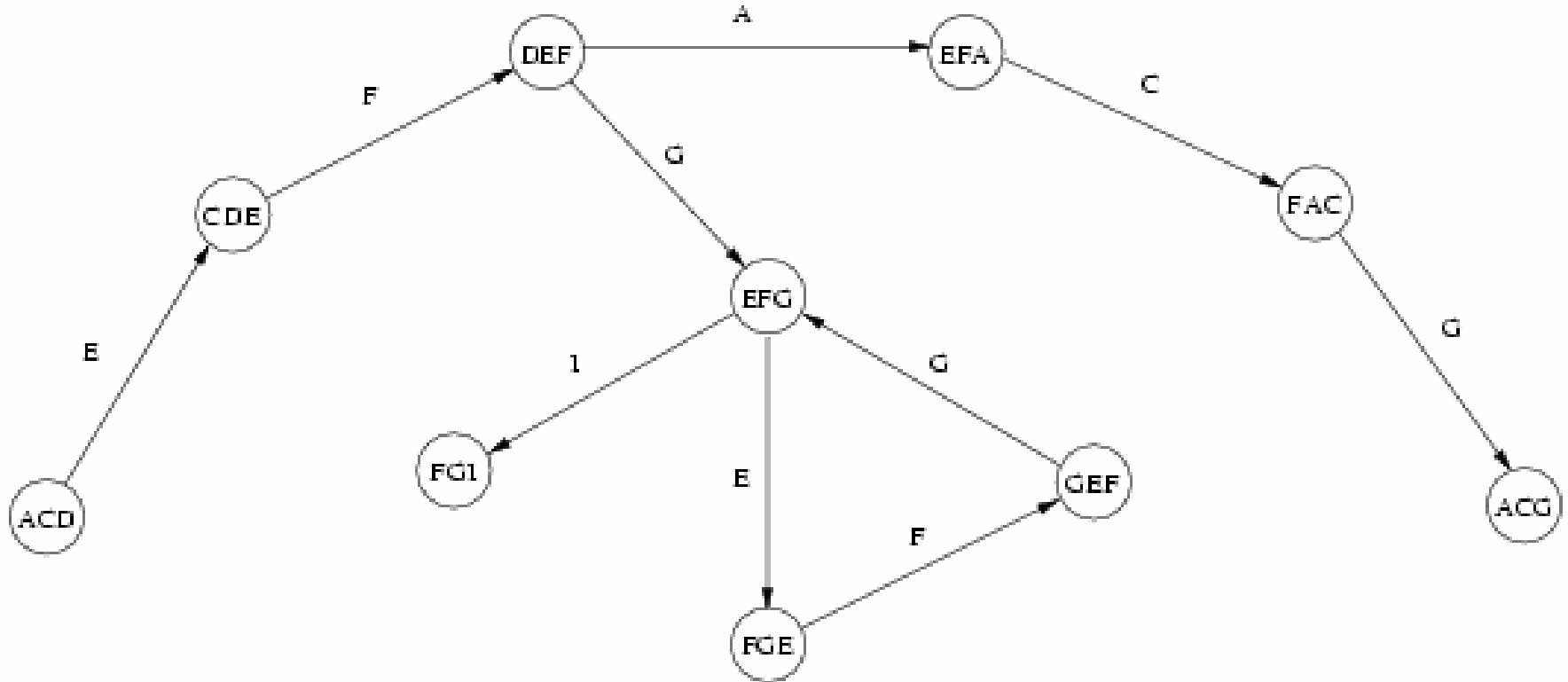
- **Complete**
 - All 30-mers are present
- **Correct**
 - No other 30-mers are present
- **Compact**
 - No 30-mer is present more than once

Sequence Database Compression



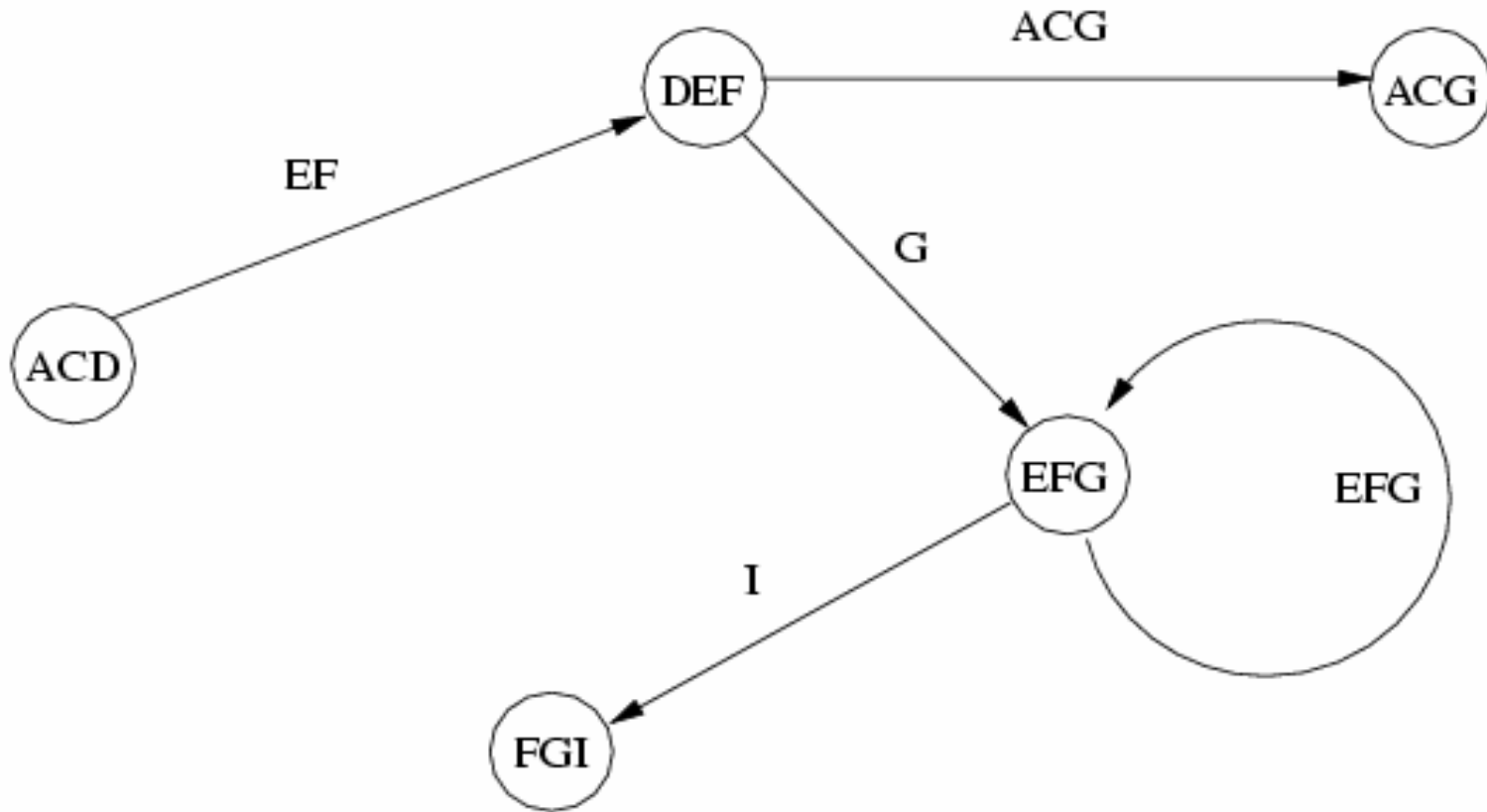
Sequence Database	Swiss-Prot	VarSplic
Original Size	55 Mb	90 Mb
Distinct 30-mers	44 Mb	45 Mb
Overhead	27%	97%
C^3 Size	53 Mb	54 Mb
C^3 Overhead	19%	20%
C^3 Compression	93%	61%
Compression LB	79%	51%

SBH-graph



ACDEFGI, ACDEFACG, DEFGEFGI

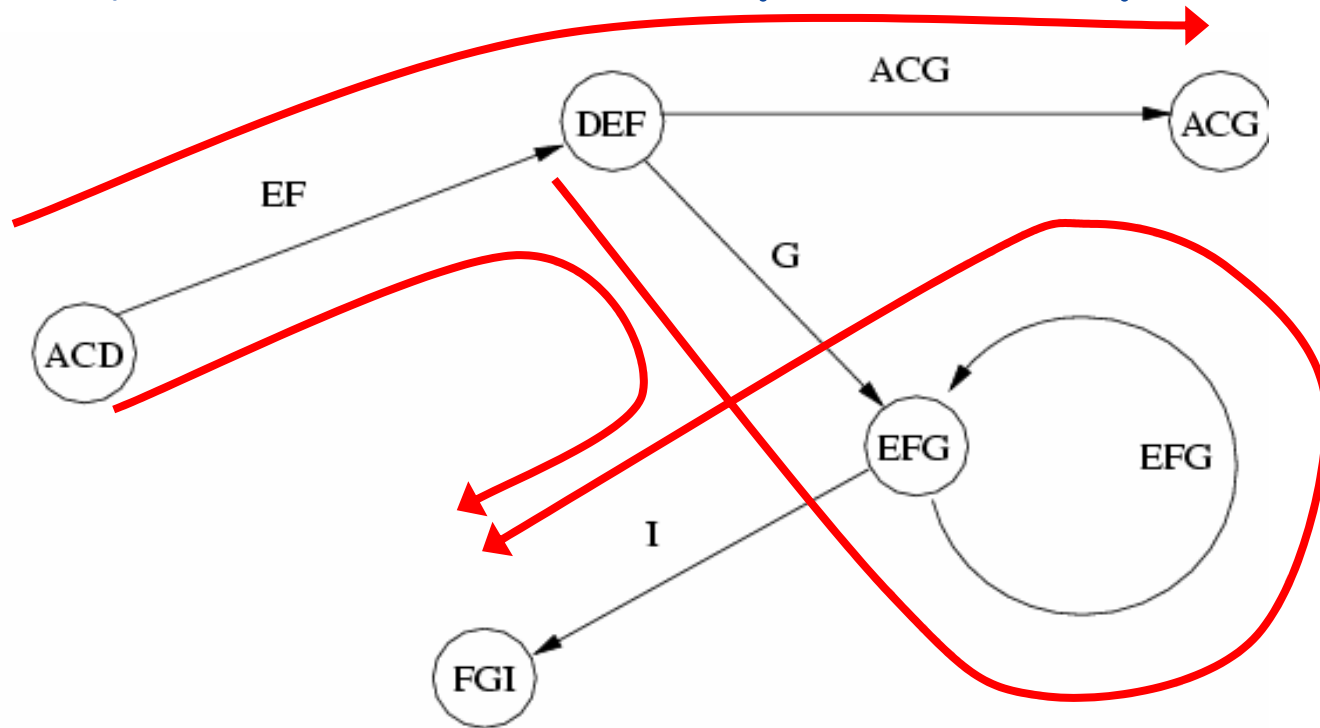
Compressed SBH-graph



ACDEFGI, ACDEFACG, DEFGEFGI

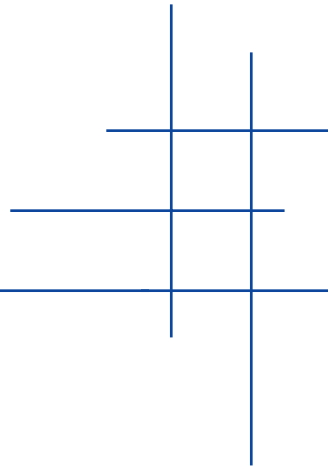
Sequence Databases & CSBH-graphs

- 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

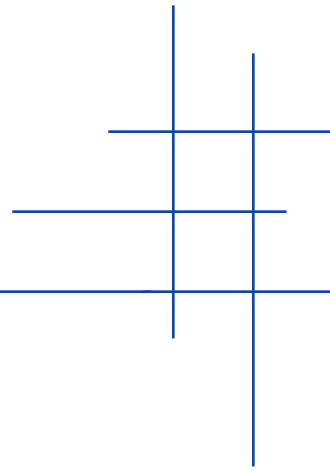
- Sequence database with p paths

$$N_k + p k$$

- Minimize sequence database size
by minimizing number of paths

- subject to C^3 constraints

Best case senario...



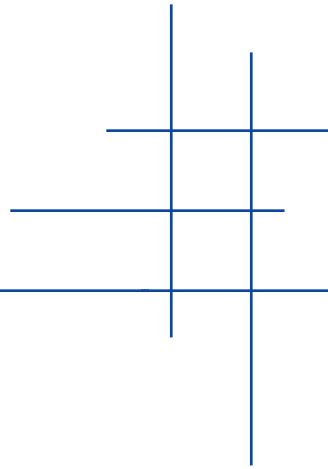
...if CSBH-graph admits an Eulerian path.

Sequence database size

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

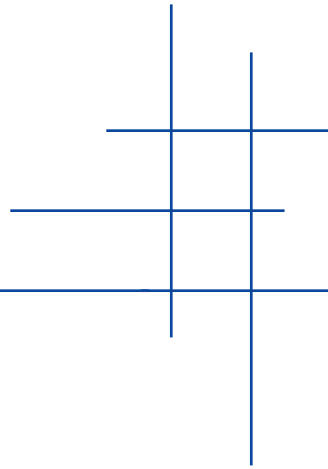
How many paths are required if the CSBH-graph is not Eulerian?

Non-Eulerian Components



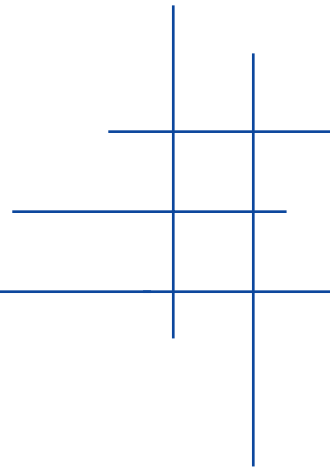
- Net degree
 - $b(v) = \# \text{ in edges} - \# \text{ out edges}$
- Total degree surplus
 - $B_+ = \sum_{b(v) > 0} b(v)$
- For each path
 - Start node's net degree +1
 - End node's net degree -1
 - Otherwise, net degree: no change
- To reduce all nodes to net degree 0, must have at least B_+ paths.

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



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

Paths Lower Bound

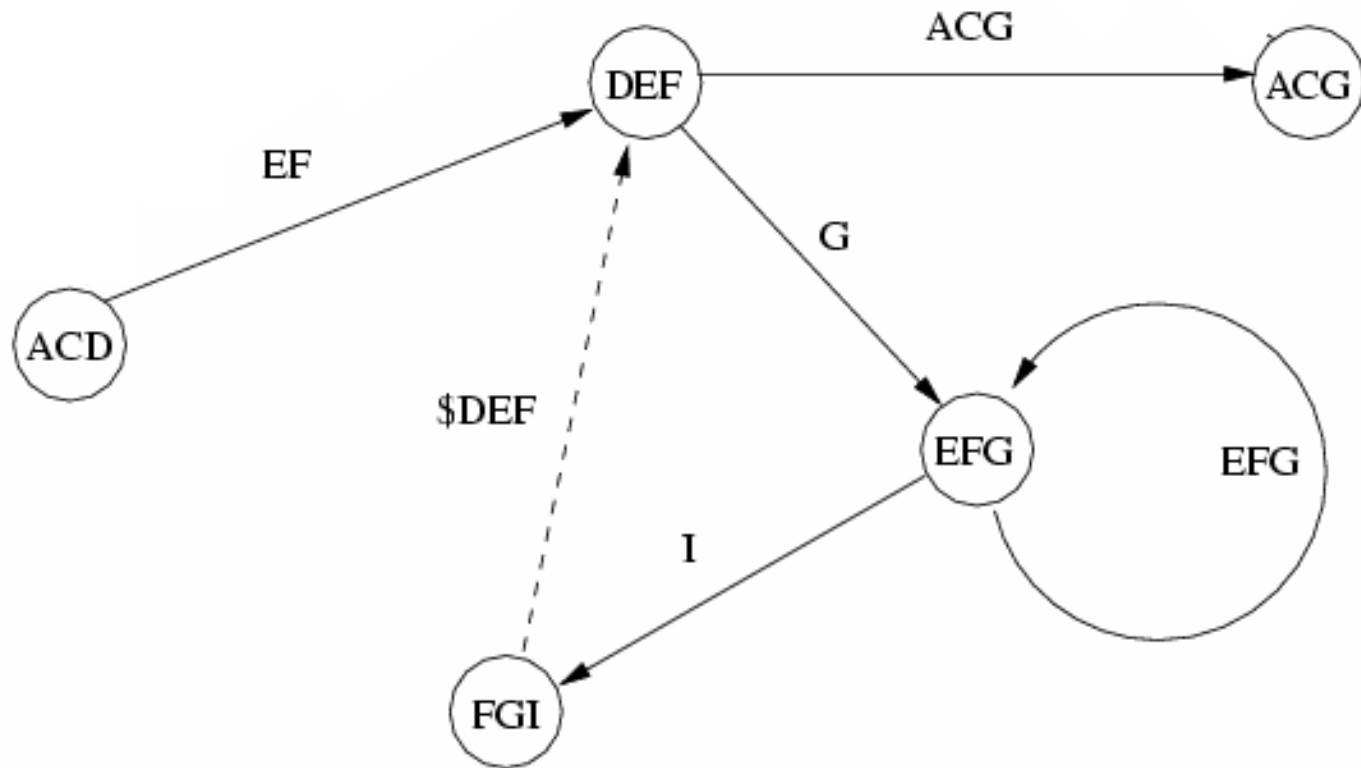


The C^3 path set must contain at least $B_+ + m$ paths.

This lower bound is achievable!

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

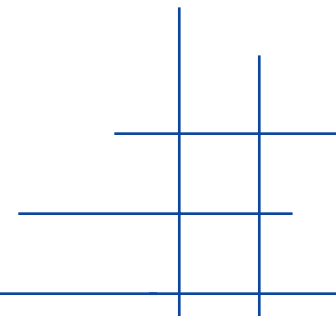
Achieving Path Lower Bound



AA Sequence Databases

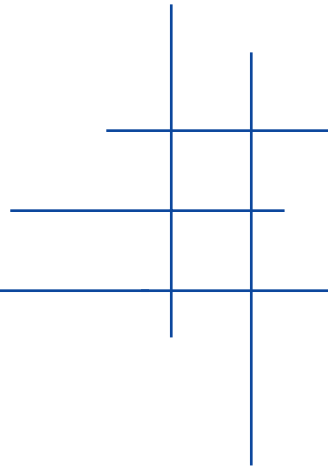
Sequence Database	Sequence Length	Distinct 30-mers	Overhead
IPI-HUMAN	20358846	12115520	68%
IPI	54145883	29769766	81%
Swiss-Prot	56454588	44374286	27%
Swiss-Prot-VS	89541275	45307827	97%
UniProt	472581860	274510105	72%
UniProt-VS	506796094	275391669	84%
MSDB	481919777	276523755	74%
NRP	495502241	283160529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

Minimum Size C³ Sequence Database



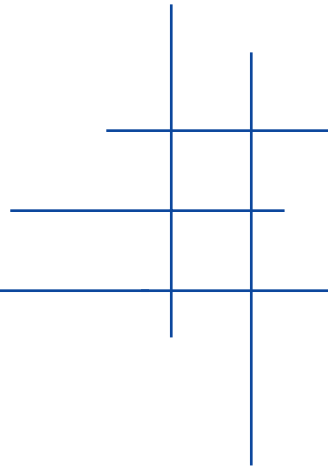
Sequence Database	C ³ 30-mer Enumeration	Overhead	Compression	Compression Bound
IPI-HUMAN	13854679	14.35%	68.05%	59.51%
IPI	37961385	27.52%	70.11%	54.98%
Swiss-Prot	52662145	18.68%	93.28%	78.60%
Swiss-Prot-VS	54534356	20.36%	60.90%	50.60%
UniProt	337119564	22.81%	71.34%	58.09%
UniProt-VS	338890778	23.06%	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



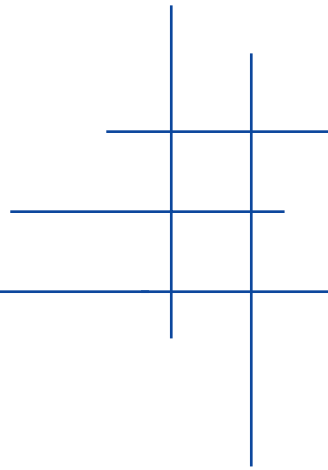
- Suitable for use by Mascot, SEQUEST, ...
 - FASTA format
- All connection to protein context is lost
 - Must do exact string search to find peptides in original database

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