

MSPathFinder Scoring

Manuscript:

“Informed-Proteomics: open-source software package for top-down proteomics”

Nature Methods (2017) doi:10.1038/nmeth.4388

<https://www.ncbi.nlm.nih.gov/pubmed/28783154>

MSPathFinder: Database Search



CPTAC_Intact_CR33A_24Aug15_Ba
ne_15-02-06-RZ.raw
Xcalibur Raw File

Raw File



H_sapiens_Uniprot_SPROT_2015-10
-14.fasta
FASTA File

Fasta file



Mods.txt

Modifications

MSPathFinder

ProMex
(Feature Finder)



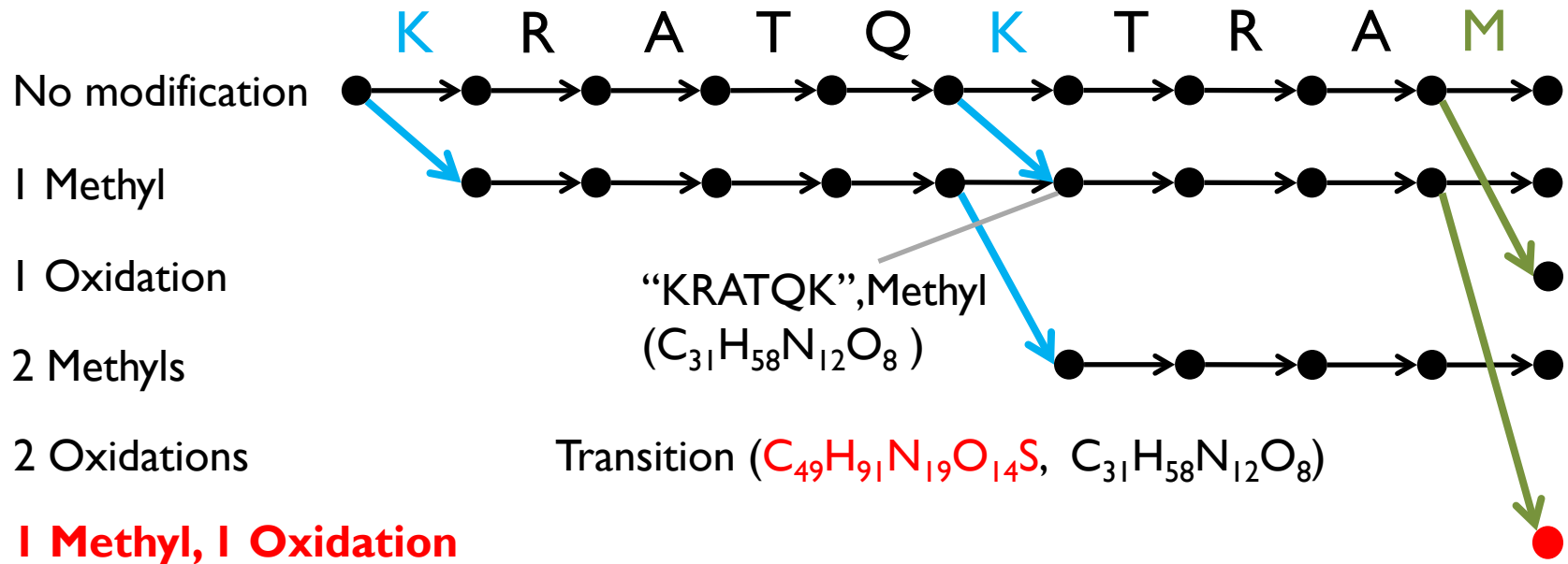
Sequence Graphs
(Proteoform explorer)



LcMsSpectator
(Visualization)

Proteoform-Spectrum Matches (PrSMs)

Sequence Graph



All proteoforms are represented as paths

Exploring > 50 trillion proteoforms (from the same protein)
in less than a minute using a graph algorithm

Internal Cleavages

Protein in a database

No cleavage or N-term single residue cleavage

% in
total ID

#Sequences
derived from
a database

25%

112K

Single internal cleavage (+ N-term single residue cleavage)

60%

3M

Multiple internal cleavages

15%

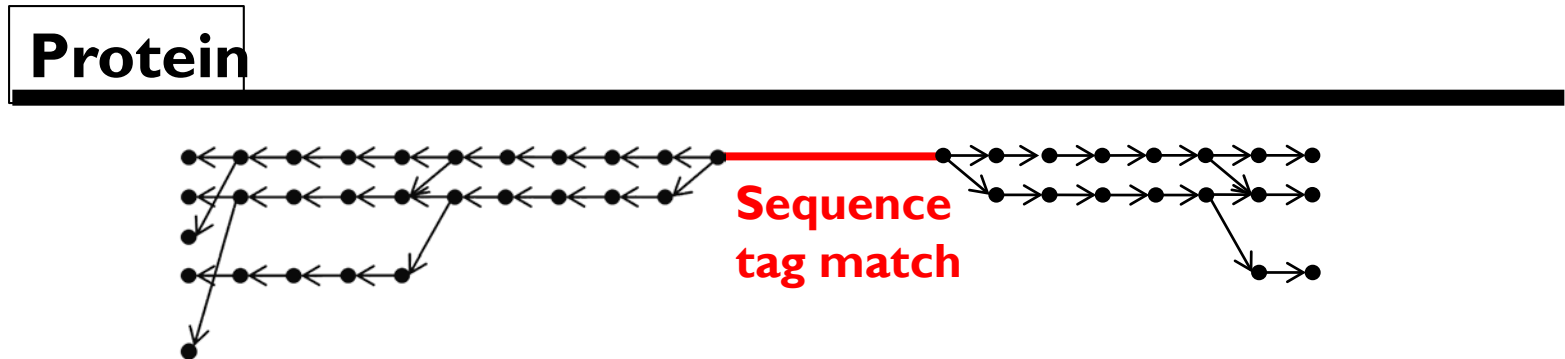
223M

99% search time

Salmonella database containing 5,634 proteins

Sequence Tag-based Search

- Cover multiple cleavages



- Generate short de novo sequence tags
- Find proteins matching the sequence tags
- Extend sequence tag matches using sequence graphs

MSPathFinder Scoring Model

MSPathScore (S, P)

$$\begin{aligned}
 = & \sum_{i \in \alpha} [W_{match}^p h + W_{intensity}^p I_i + W_{dist}^p D_i + W_{error}^p E_i] \\
 & + \sum_{i \in \beta} [W_{match}^s h + W_{intensity}^s I_i + W_{dist}^s D_i + W_{error}^s E_i] \\
 & + \sum_{\substack{(i,j) \in (\alpha \cup \beta) \\ i \neq j}} W_{compl} IsComplement(i,j) + W_{consecutive} IsConsecutive(i,j)
 \end{aligned}$$

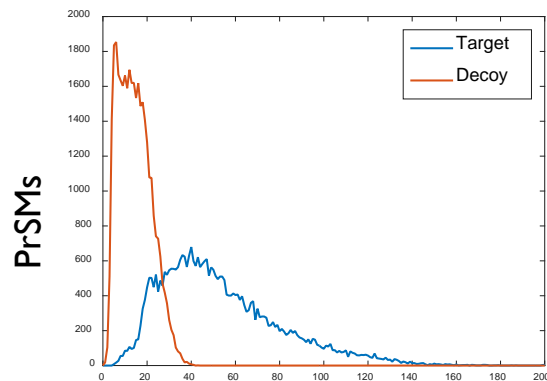
α and β : Sets of prefix and suffix fragment ion matches, respectively

I : Normalized intensity

D : Isotope envelope similarity

E : Mass error

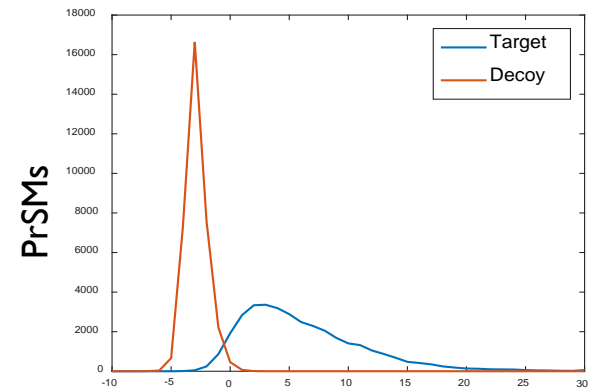
Weight parameters **{W}** are trained by Logistic Regression



#Matched Fragment Ions



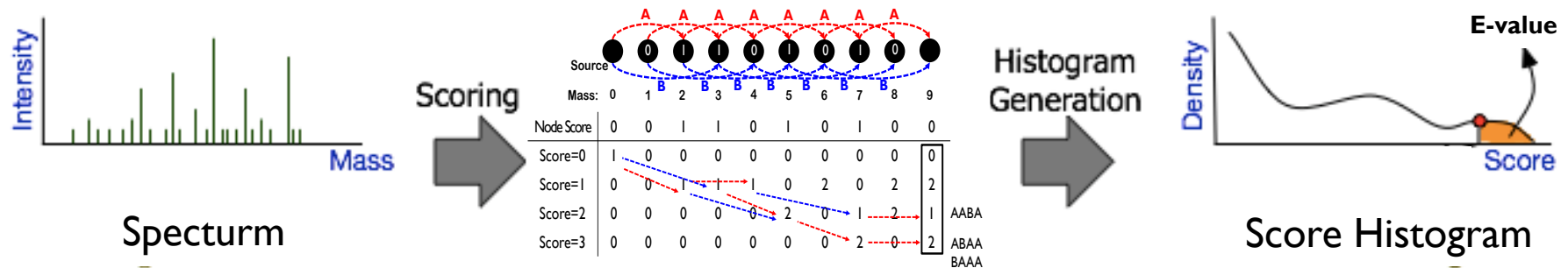
Training data
(>30,000 PrSMs)
 Human, Bacteria;
 ETD, CID



MSPathScore

Statistical Significance of Protein-Spectrum Match (PrSM)

- Generating Function Approach



Computing the (spectrum-specific) score histogram of **all Proteins**

Raw score (MSPath score) → **E-value**