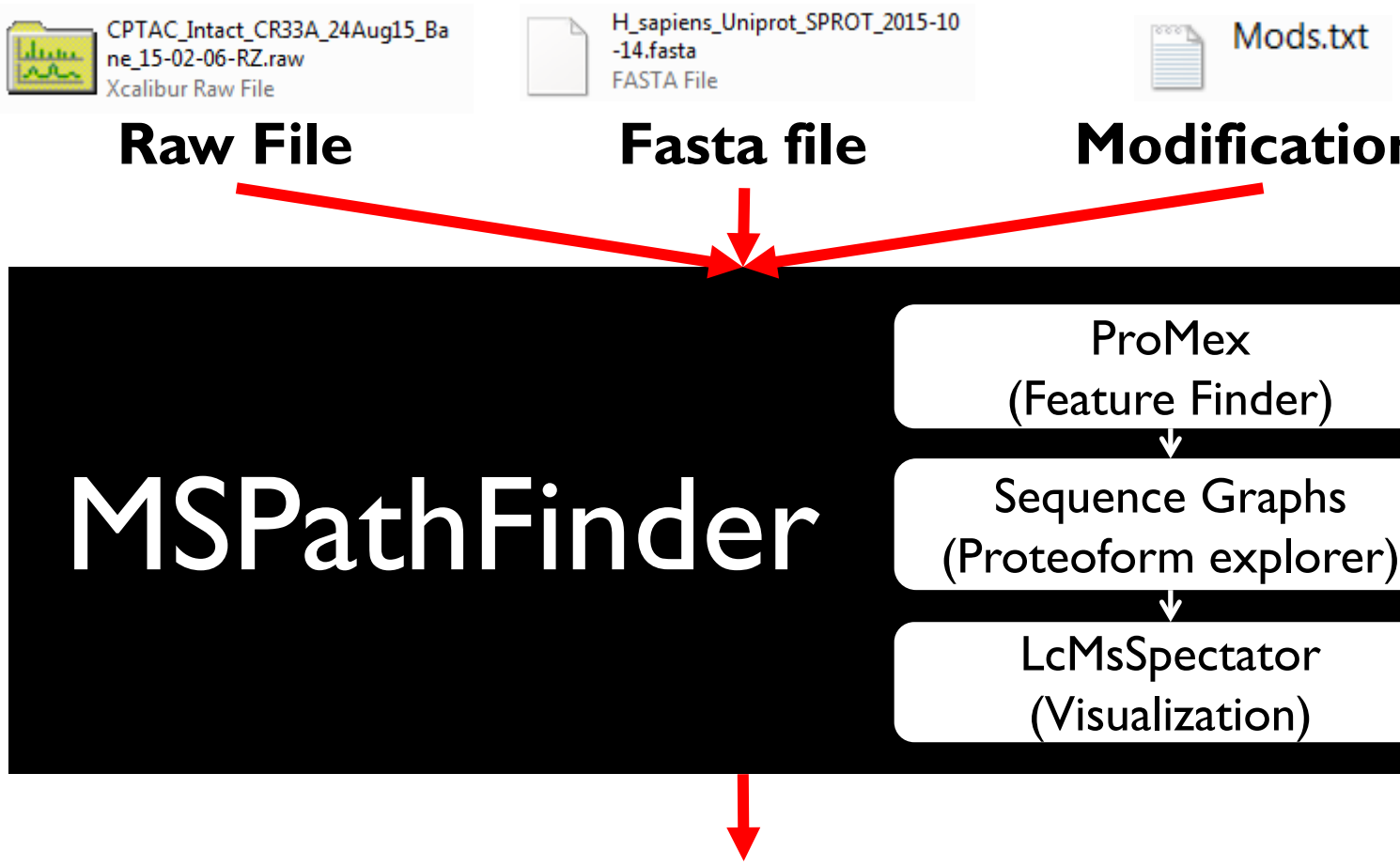
MSPathFinder Scoring

Manuscript:

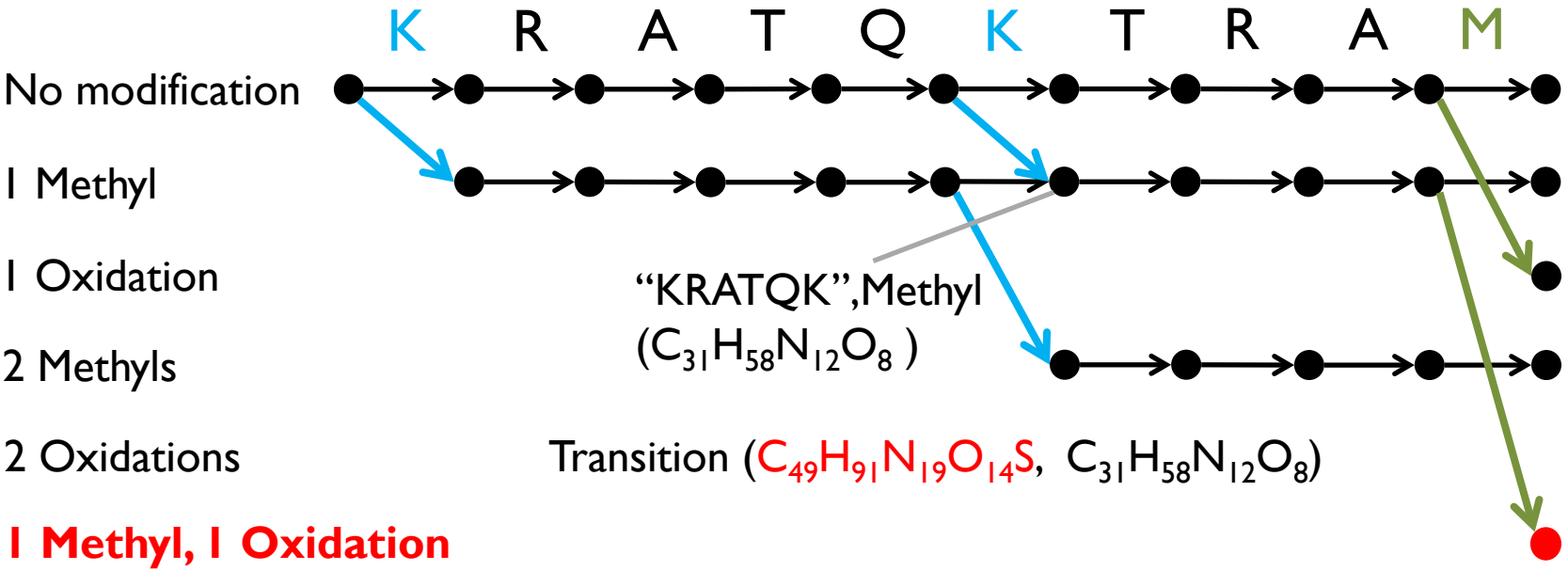
"Informed-Proteomics: open-source software package for top-down proteomics" Nature Methods (2017) doi:10.1038/nmeth.4388

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

MSPathFinder: Database Search



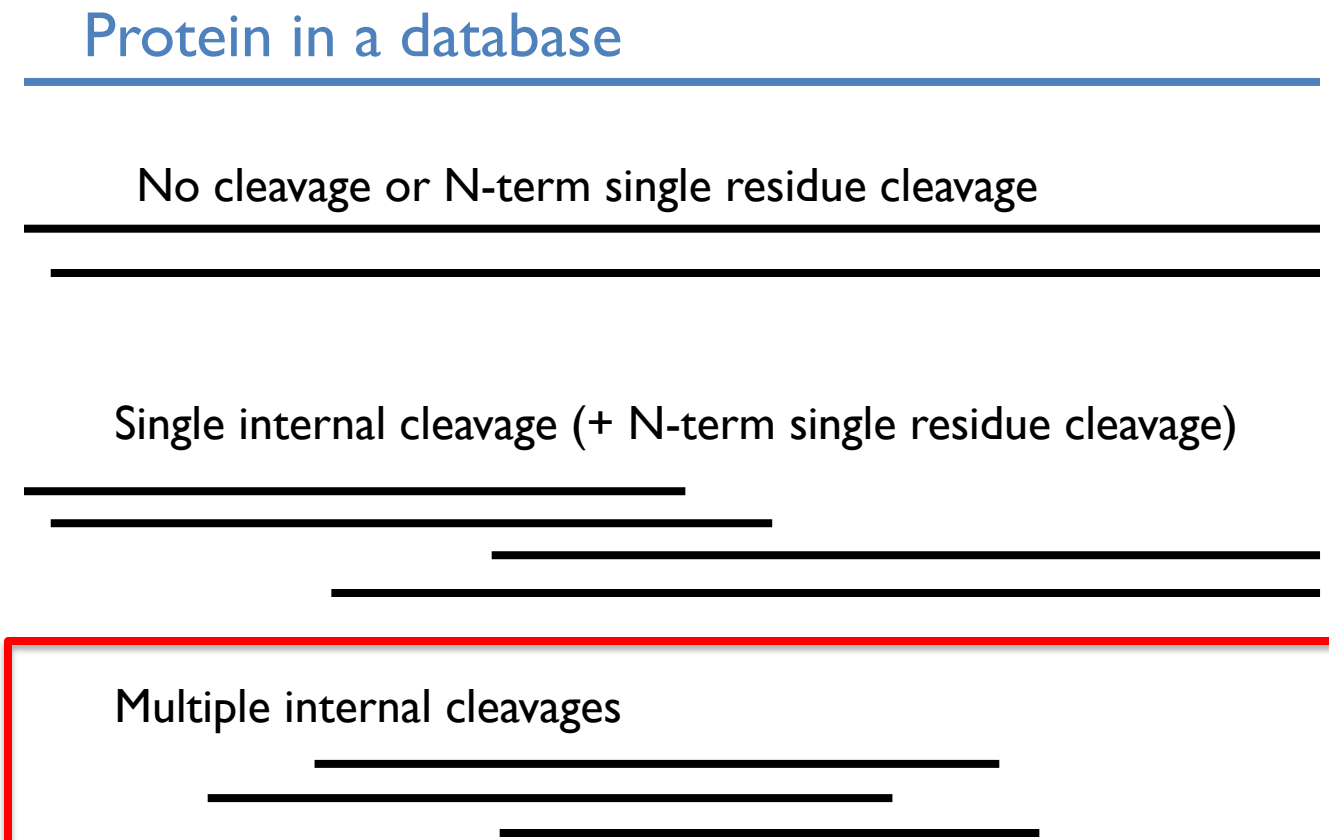
Sequence Graph



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

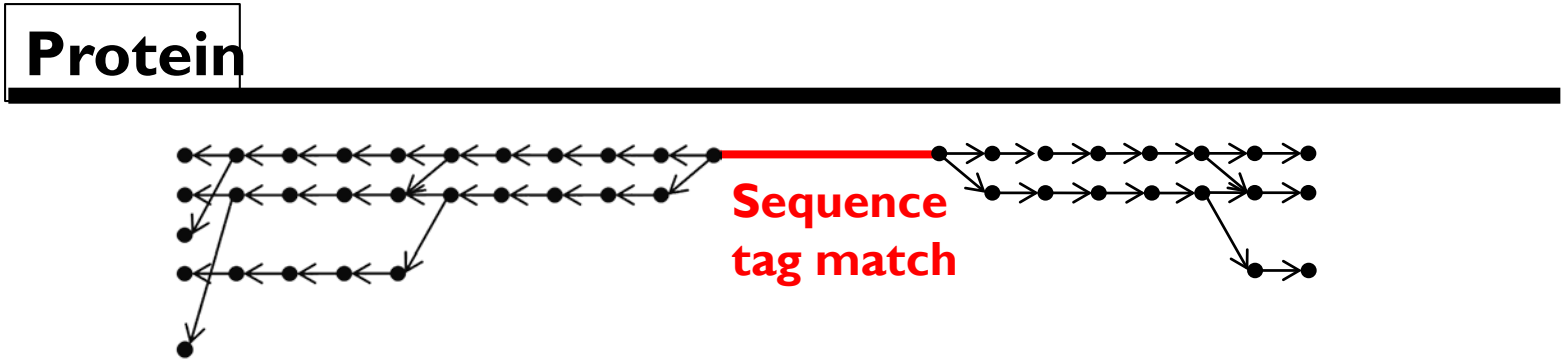


99% search time

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)

$$= \sum_{i \in \alpha} \left[W^{p}_{matc\,h} + W^{p}_{intensity}\, I_i + W^{p}_{dist}\, D_i + W^{p}_{error}\, E_i \right]$$
$$+ \sum_{i \in \beta} \left[W^{s}_{matc\,h} + W^{s}_{intensity}\, I_i + W^{s}_{dist}\, D_i + W^{s}_{error}\, E_i \right]$$
$$+ \sum_{\substack{(i,j) \in (\alpha \cup \beta) \\ i \neq j}} W_{compl}\, IsComplement(i,j) + W_{consecutive}\, IsConsecutive(i,j)$$

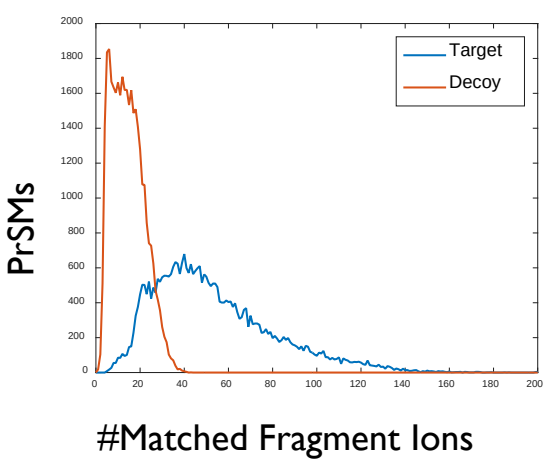
α and β : Sets of prefix and suffix fragment ion matches, respectively

I : Normalized intensity

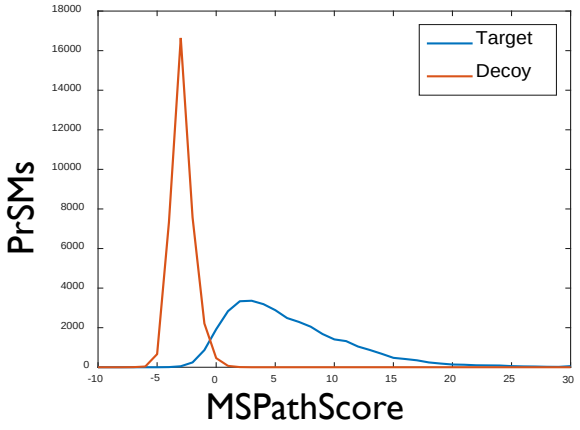
D : Isotope envelope similarity

E : Mass error

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

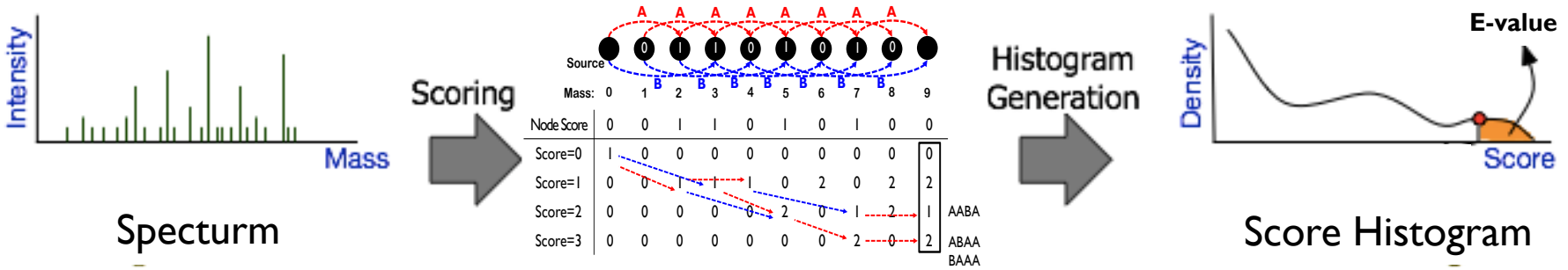


Training data (>30,000 PrSMs)
Human,Bateria;
ETD,CID



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**