

Spectrum Graph approaches for de novo peptide identification



CHAO JI

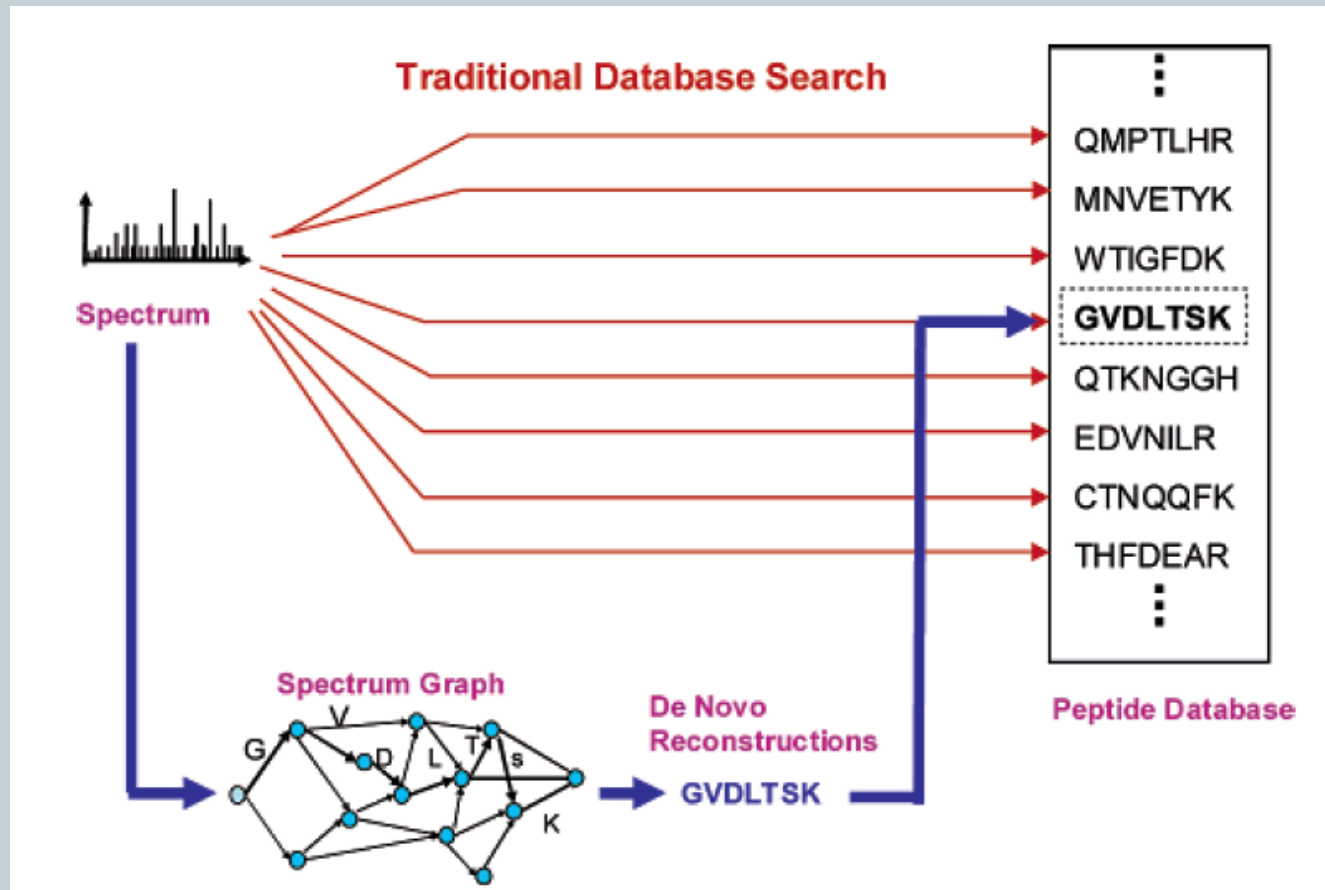
RADIVOJAC' LAB MEETING

Outline



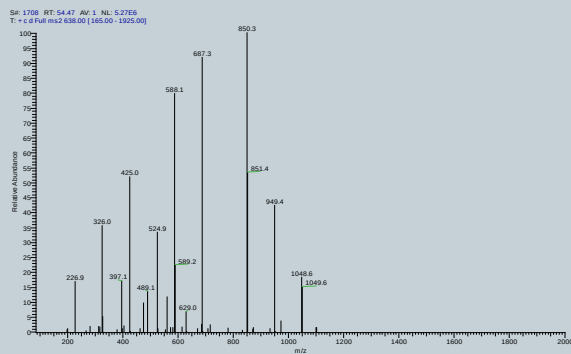
- Spectrum Graph
- Issues with current de novo approaches and suggested ways for improving de novo searching results

De novo vs database search

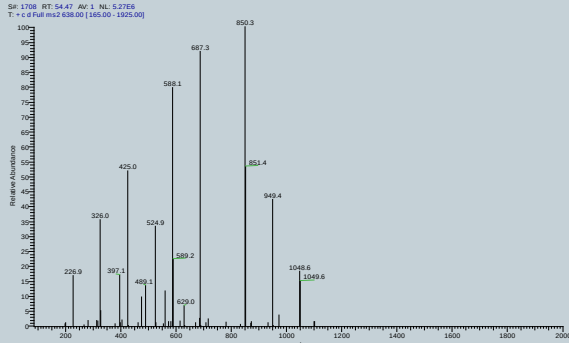


Ari M. Frank et al. 2007

De novo vs database search



De novo vs database search



Database of known peptides

MDERHILNM, KLQWVCSDL,
PTYWASDL, ENQIKRSACVM,
TLACHGGEM, NGALPQWRT,
HLLERTKMNVV, GGPASSDA,
GGLITGMQSD, MQPLMNWE,
ALKIIMNVRT, **AVGELTK**,
HEWAILF, GHNLWAMNAC,
GVFGSVLRA, EKLNKAATYIN..

Database of known peptides is much smaller

Database of all peptides = 20^n

AAAAAAAA,AAAAAAAC,AAAAAAD,AAAAAAAE,
AAAAAAAG,AAAAAAF,AAAAAAAH,AAAAAAI,

: : :

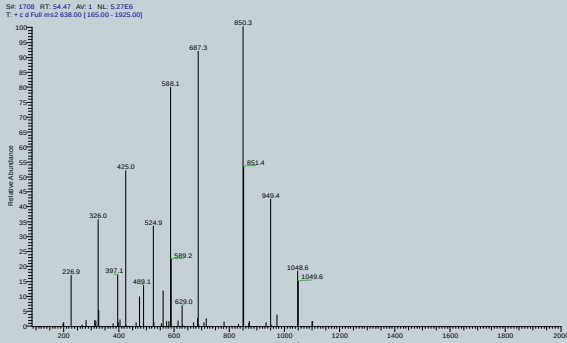
AVGELTI, **AVGELTK**, AVGELTL, AVGELTM,

: : :

YYYYYYYYS,YYYYYYYT,YYYYYYYV,YYYYYYY

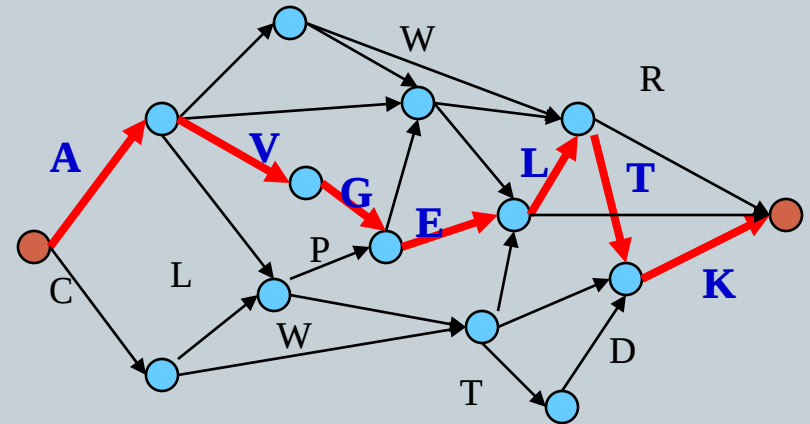
Theoretical database is huge

De novo vs database search



Database of known peptides

MDERHILNM, KLQWVCSDL,
PTYWASDL, ENQIKRSACVM,
TLACHGGEM, NGALPQWRT,
HLLERTKMNVV, GGPASSDA,
GGLITGMQSD, MQPLMNWE,
ALKIIMNVRT, **AVGELTK**,
HEWAILF, GHNLWAMNAC,
GVFGSVLRA, EKLNKAATYIN..

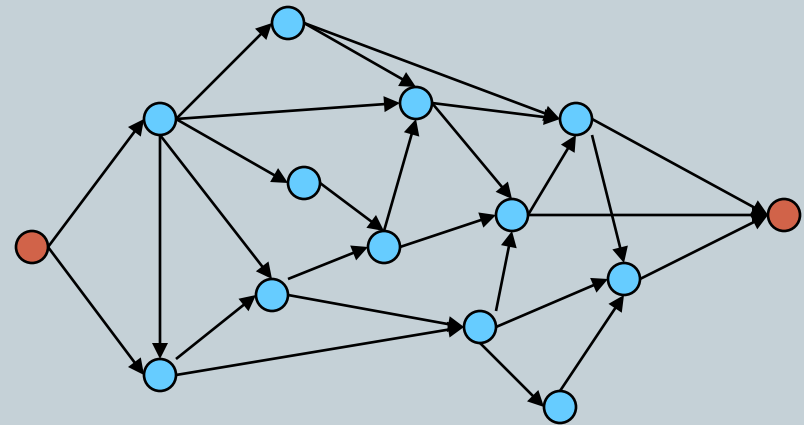
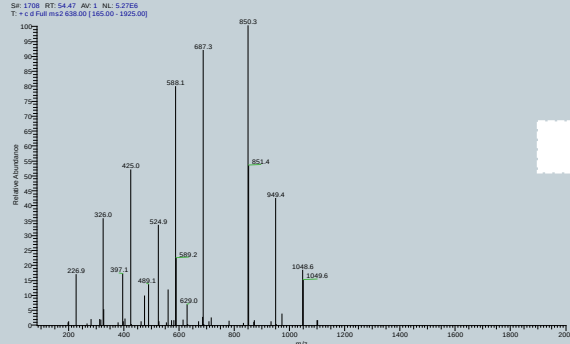


Searching a SG
is much more
efficient!

Spectrum Graph



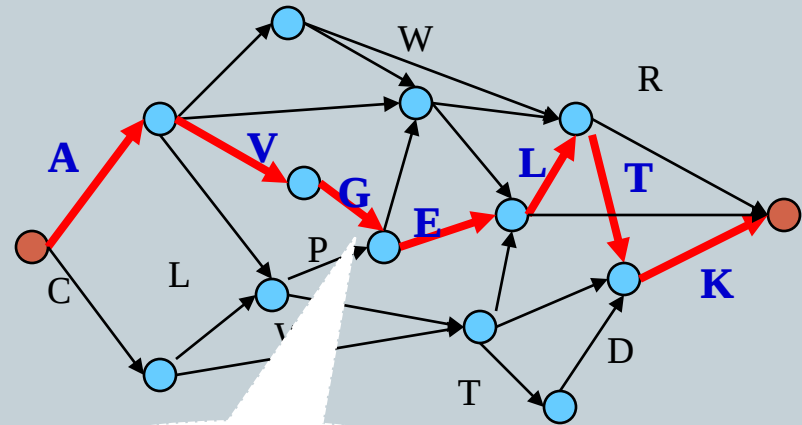
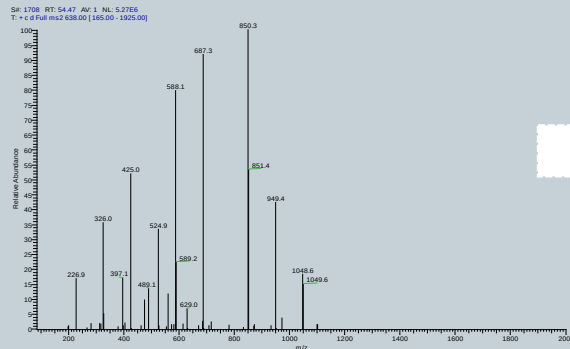
- An experimental spectrum is represented by spectrum graph (SG).
- Vertices = m/z values, edges = mass of amino acid(s)



Spectrum Graph



- An experimental spectrum is represented by spectrum graph (SG).
- Vertices = m/z values, edges = mass of amino acid(s)

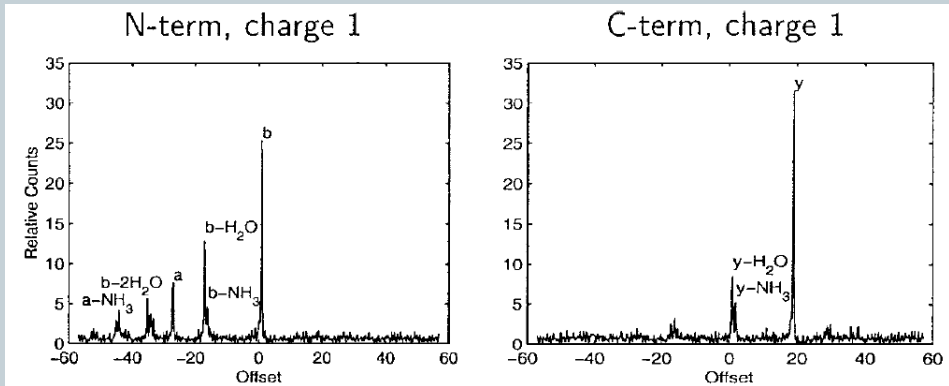


Goal is to find a path corresponding to the correct sequence

Construction of SG



- Learning ion types from spectrum-peptide matches
- For all $S = \{s_1, s_2, \dots, s_n\}$ and peptide, P plot the histogram of $x_{ij} = m(P_i) - s_j$



Offset Frequency Function(OFF)

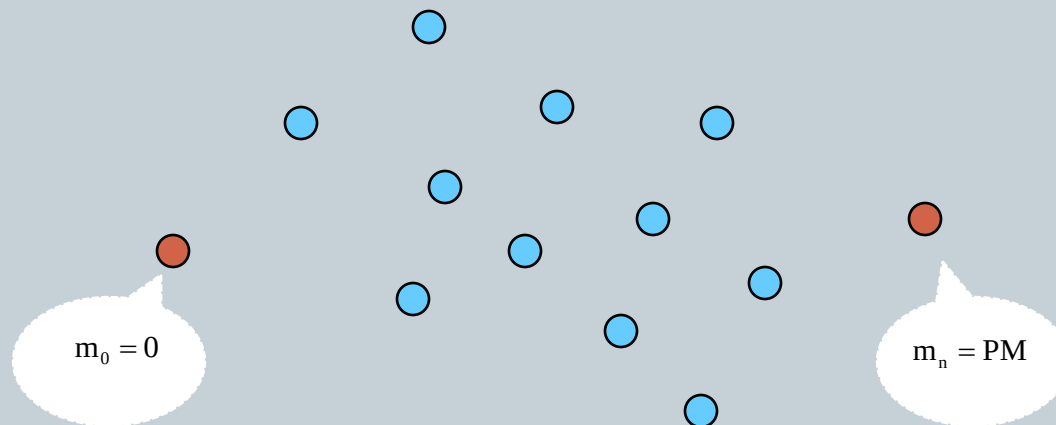
Offset	Integer offset	Count	Filtered count	Probability	Average intensity	Term	Ion
18.85	19	604	604	0.6895	4.5457	C	y
0.85	1	568	565	0.6484	2.2788	N	b
-17.05	-17	338	338	0.3858	1.1966	N	b-H ₂ O
0.90	1	248	231	0.2831	0.5716	C	y-H ₂ O
-27.15	-27	204	200	0.2329	0.7537	N	a
20.05	20	183	180	0.2089	3.4699	C	y ²
-16.15	-16	159	152	0.1815	1.2766	N	b-NH ₃
1.90	2	131	114	0.1495	0.6680	C	y-NH ₃
-35.20	-35	151	141	0.1724	0.5253	N	b-H ₂ O-H ₂ O
-34.20	-34	134	131	0.1530	0.4736	N	b-H ₂ O-NH ₃
-44.25	-44	129	126	0.1473	0.5516	N	a-NH ₃
-45.15	-45	107	98	0.1221	0.4820	N	a-H ₂ O
2.30	2	102	95	0.1164	1.7460	C	y ² -H ₂ O
-16.10	-16	97	84	0.1107	0.4913	C	y-H ₂ O-NH ₃
-17.15	-17	91	71	0.1039	0.4935	C	y-H ₂ O-H ₂ O

FIG. 4. Plots of the offset frequency functions. Horizontal axes represent offsets between peaks in spectra and masses of partial peptide molecules. Vertical axes represent normalized offset counts with 1 being the average count. The only significant offsets for doubly charged ions correspond to y and y-H₂O ions.

Construction of SG



- Select K -most abundant Ions from OFF $\Delta = \{\delta_1, \delta_2, \dots, \delta_k\}$
- For each s_i create $s_i + \delta_1, s_i + \delta_2, \dots, s_i + \delta_k$
- Merge close pairs $s_i + \delta_j, s_{i'} + \delta_{j'}$
- Create vertices for empty peptide and full peptide
- Each vertex also has a weight proportional to its intensity
- Each edge has weight proportional to the weights of incident vertices
- $\text{Score}(\text{path}) = \text{sum of weights of edges}$



Construction of SG



- How to create vertices
- How to create edges
- How to score paths
- How to find best paths

Ideal tandem mass spectrum is noise-free and contains only b- y- ions

NC-spectrum graph



- $G = (V, E)$ where $V = \{N_0, N_1, \dots, N_k, C_0, C_1, \dots, C_k\}$
- Each peak in S generates a pair of vertices
- Coordinates of vertices: the sum of mass of prefix residues, or the sum of mass suffix residues (+18)

$$\text{cord}(x) = \begin{cases} 0 & x = N_0; \\ W - 18 & x = C_0; \\ w_j - 1 & x = N_j \quad \text{for } j = 1, \dots, k; \\ W - w_j + 1 & x = C_j \quad \text{for } j = 1, \dots, k. \end{cases}$$

Constructing NC-Spectrum Graph



- Vertices: $V = \{x_0, x_1, \dots, x_k, y_k, \dots, y_1, y_0\}$
- Edges: For $x, y \in V$, $E(x, y) = 1$, if
$$\begin{cases} x + y \neq W - 18 \\ x < y \\ y - x = \text{mass of aa}(s) \end{cases}$$
- Mass array $A[m] = 1$, if m equals mass of $\text{aa}(s)$ can be computed in $O(m)$
- Constructed NC-Spectrum Graph in $O(k^2)$ if mass array is pre-computed

Search High-scoring paths in SG



- An exhaustive search of all paths still takes $O(2^k)$
- Use dynamic programming

Algorithm Compute- $Q(G)$

1. Initialize $Q(i, j) = 0$ for all $0 \leq i, j \leq k$;
2. For $j = 1$ to k
3. If $E(y_j, y_0) = 1$, then $Q(0, j) = \max\{Q(0, j), s(y_j, y_0)\}$;
4. If $E(x_0, x_j) = 1$, then $Q(j, 0) = \max\{Q(j, 0), s(x_0, x_j)\}$;
5. For $i = 1$ to $j - 1$
 - (a) For every $E(y_j, y_p) = 1$ and $Q(i, p) > 0$, $Q(i, j) = \max\{Q(i, j), Q(i, p) + s(y_j, y_p)\}$;
 - (b) For every $E(x_p, x_j) = 1$ and $Q(p, i) > 0$, $Q(j, i) = \max\{Q(j, i), Q(p, i) + s(x_p, x_j)\}$.

- Highest-scoring path can be found in $O(|V| |E|)$ time and $O(|V|^2)$

Two steps in de novo sequencing



- Narrow down the search space to a pool of candidate sequences
- A scoring function designed to discriminate between correct and incorrect solutions

Two steps in de novo sequencing



- Narrow down the search space to a pool of candidate sequences

Spectrum Graph

- A scoring function designed to discriminate between correct and incorrect solutions

Two steps in de novo sequencing



- Narrow down the search space to a pool of candidate sequences

Spectrum Graph

- A scoring function designed to discriminate between correct and incorrect solutions

Likelihood ratio e.g.
PepNovo,

Issues with de novo approaches



- Low accuracy as opposed to database search (Only 70-75% of amino acids correctly called, 30% of identified peptides being correct)
- Accuracy deteriorates for long sequences

algorithm	average accuracy	average length	predictions with correct subsequences of length at least x							
			$x = 3$	$x = 4$	$x = 5$	$x = 6$	$x = 7$	$x = 8$	$x = 9$	$x = 10$
PepNovo	0.727	10.30	0.946	0.871	0.800	0.654	0.525	0.411	0.271	0.193
Sherenga	0.690	8.65	0.821	0.711	0.564	0.364	0.279	0.207	0.121	0.071
Peaks	0.673	10.32	0.889	0.814	0.689	0.575	0.482	0.371	0.275	0.179
Lutefisk	0.566	8.79	0.661	0.521	0.425	0.339	0.268	0.200	0.104	0.057

^a Cumulative results for 280 test spectra: the average accuracy of predicted amino acids, average prediction length, and proportions of predictions that had a correct subsequence of length at least x , for $3 \leq x \leq 10$.

Issues with de novo approaches



- Incomplete fragmentation/low mass accuracy

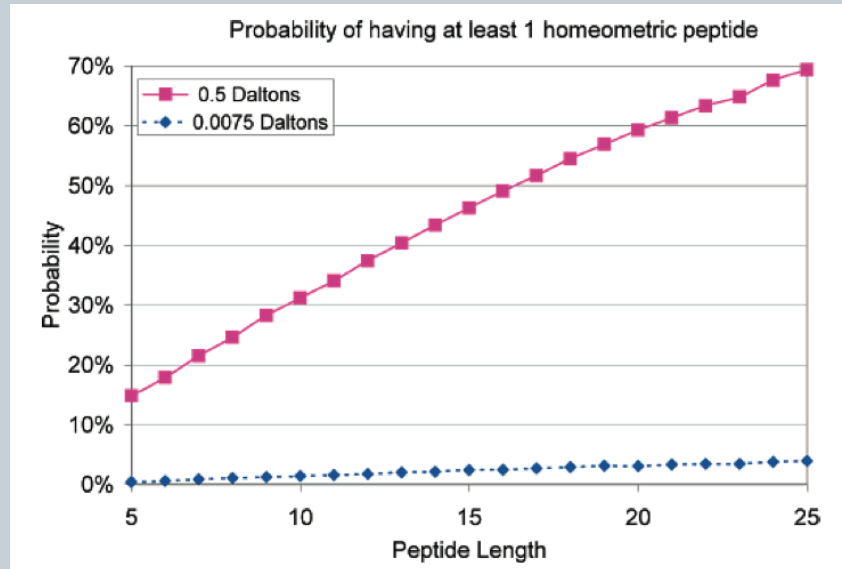
Use high-resolution instruments (i.e. HCD, ECD, ETD)

79.8% of the correctly identified HCD spectra contain full or almost full cleavage information, compared with CID (62.8%)

Issues with de novo approaches

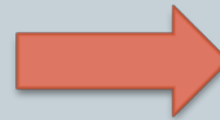
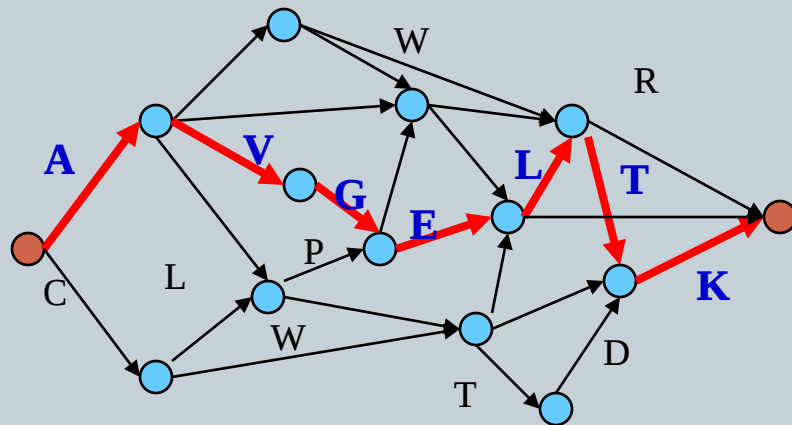


- Two peptides are *homeometric* if their theoretical spectra contains the same set of m/z values





- Applying more rigorous PSM scoring function



MDERHILNM, KLQWVCSDL,
PTYWASDL, ENQIKRSACVM,
TLACHGGEM, NGALPQWRT,
HLLERTKMNVV, GGPASSDA,
GGLITGMQSD, MQPLMNWE,
ALKIIMNVRT, **AVGELTK**,
HEWAILF, GHNLWAMNAC,
GVFGSVLRA, EKLNKAATYIN..

More rigorous
PSM scoring
function



ALKIIMNVRT, **AVGELTK**



- Make use of the correlations of outputs of de novo sequencing

```
>> 0 2712  Cmpd 4338, +MSn(535.98), 87.08 min (SQS 0.831)
#Index  RnkScr  PnvScr  N-Gap  C-Gap  [M+H]  Charge  Sequence
0       9.519   134.637  214.928  0.000   1070.425  2       ALLGLQLK
1       9.118   120.359  285.799  0.000   1070.011  2       LLGLQLK
2       8.859   133.525  214.928  0.000   1070.425  2       ALLGLQNK
3       8.529   119.246  285.799  0.000   1070.011  2       LLGLQNK
4       8.331   138.635  214.928  0.000   1070.425  2       ALLGLELK
5       8.176   130.896  214.928  0.000   1070.425  2       ALLGLKNK
6       8.061   131.088  113.372  0.000   1070.425  2       TALLGLQLK
7       8.001   123.310  57.010   0.000   1070.011  2       TGALLGLELK
8       7.816   137.074  86.612   0.000   1070.425  2       QALLGLQLK
9       7.796   135.086  113.372  0.000   1070.425  2       TALLGLELK
10      7.728   133.548  214.928  0.000   1070.425  2       ALLGLKLK
11      7.720   127.128  285.799  0.000   1070.011  2       LLGLELK
12      7.710   131.000  214.928  0.000   1070.425  2       ALLGLTLR
13      7.596   135.961  86.612   0.000   1070.425  2       QALLGLQNK
14      7.536   119.001  57.010   0.000   1070.011  2       TQLLGLELK
15      7.272   126.133  86.612   0.000   1070.425  2       QALLGLGALK
16      7.162   129.975  113.372  0.000   1070.425  2       TALLGLQNK
17      7.147   126.144  214.928  0.000   1070.425  2       ALLGTLLR
18      7.136   131.625  0.000     0.000   1070.011  2       GTGALLGLQLK
19      7.122   134.406  0.000     0.000   1070.425  2       NTALLGLELK
20      7.121   146.630  0.000     0.000   1070.425  2       NTALLGLQLK
21      7.101   141.072  86.612   0.000   1070.425  2       QALLGLELK
22      7.016   119.302  285.799  0.000   1070.011  2       LLGLKLK
23      6.960   135.689  0.000     0.000   1070.425  2       NTALLGLGALK
```

The confidence of the correct sequence can be boosted by subsequences with higher scores