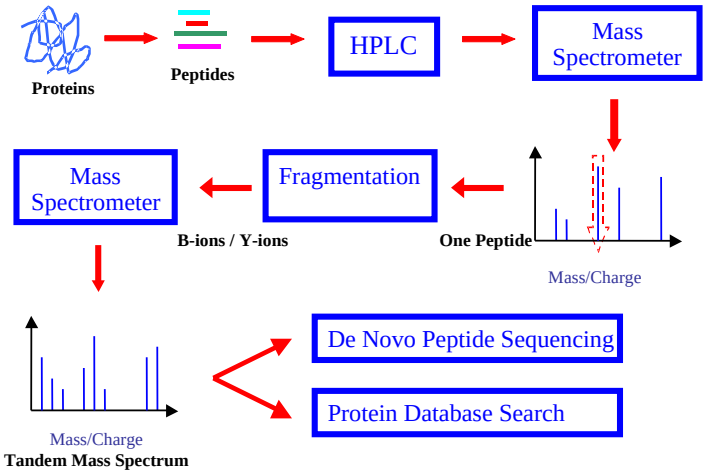
Algorithms for Peptide Sequencing

Tim Ting Chen

University of Southern California

Peptide Identification: HPLC-MS-MS



Goal: Spectrum ➔Peptide

Why?

- The sequences in protein databases are not accurate. Most of them come from gene-finding programs.
- Modifications to amino acids: RNA editing, post-translational modifications.
- SNPs in coding regions that change amino acids.

Algorithms

- Rich Johnson: Tree-based search
- Dancik et al: Spectrum graph
- Chen et al: dynamic programming
- Bafna et al and Ma et al: dynamic programming for multiple ions.
- Lu and Chen: suboptimal algorithm

***De novo* Peptide Sequencing Problem**:

Given a peptide mass W, an error range e, and a spectrum S, ask for a peptide P, such that
(1) $|\text{mass}(P)-W|<e$, and
(2) Let T be a set of all ion masses of P. Then S and T are optimally correlated.

Given
e = 0.5
W = 429.100
S = {199.022,274.31,361.01}

→

$\boldsymbol{P}$ = SWR,
Mass($\boldsymbol{P}$) = 429.212,
B-ions($\boldsymbol{P}$) ={88.033,274.112}
Y-ions($\boldsymbol{P}$)={175.113,361.121}
T ={88.033,175.113,274.112,361.121}

Chen et al., *Journal of computational Biology*, 2001

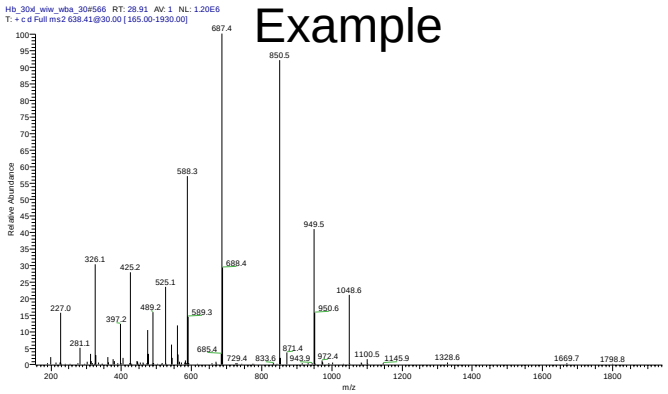
General Methodology

- Identify candidate peptides using a simple scoring function such as the number of matches, or a simple set of ions such as b and y –ions only.
- Score candidate peptides using a better scoring function

Key Ideas

- Complementary ions: if the peptide length is n, then the k-th b-ion and the (n-k)-th y-ion are called the *complementary ions*.
- Ladder ions: the k-th b-ion and the (k+1)-th b-ion forms one ladder because the difference between them equals exactly the mass of the (k+1)-th amino acid. (similar idea for the y-ions)

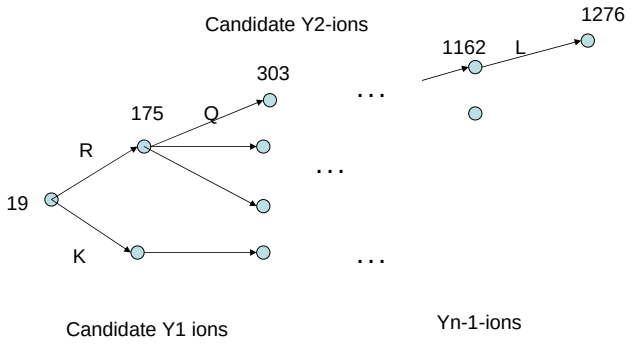
Example



Possible Fragment Ions from Cleavages at Peptide Bonds

"b-ions"	114.17	227.33	326.46	425.59	588.77	685.89	872.10	973.21	1101.34	
	L	L	V	V	Y	P	W	T	Q	R
"y-ions"	1162.38	1049.22	950.09	850.96	687.78	590.66	404.45	303.34	175.21	

Method 1: Tree Based Search



Tree-Based Search

- *Search complementary ions*: When searching the Y*k* ion, we need to search W-mass(b*n-k*) in case that the Y*k* ion is missing.
- *Missing ion is allowed*: the mass of an edge can be the mass of two or more amino acids.
- *Backtrack:* when we can not extend the tree further.

Limitations for Tree-based Search

- Time and space: the search space can be huge and it will take a very long time to sequence one spectrum if the spectrum is very dense.

  Solution: pruning the spectrum first
- Complementary ions: it is possible to interpret one mass peak as both b-ion and y-ion.

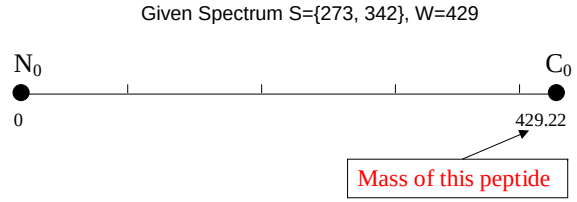
  Solution: keep track of the peaks in the tree but it will make the program even slower.

Method 2: Spectrum graph-based search

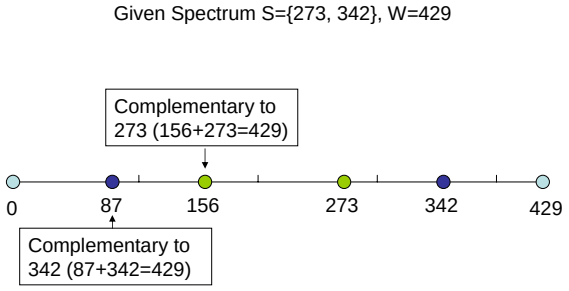
- Convert all peaks into b-ion masses.

  each mass peak *m* can be a b-ion or an y-ion, if it is a b-ion, the mass is *m-1*, if it is an y-ion, the mass of the complementary b-ion is *W-(m-19)+1*.
- Convert every peak into two nodes labeled with a mass on a line.
- Draw an edge between two nodes if their mass difference equals the mass of some amino acid.
- Goal: find the longest paths.

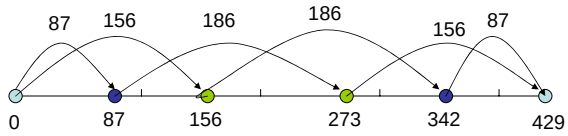
Spectrum Graph: Nodes



Construction of a spectrum graph: Nodes

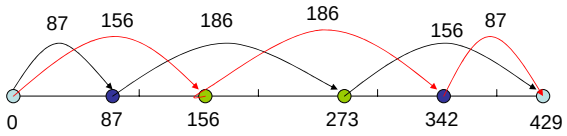


Construction of a spectrum graph: Edges



S:87, W=186, R=156

Construction of a spectrum graph: the longest Path



Formulation of the problem

Given a spectrum graph, find the longest (highest-scoring) path.

Algorithm

We consider the longest path.

Rationale. Let c(a) be the longest path from the start to the node a. Then

c(b)=max {c(a)+1: a→b is an edge}

Computation.

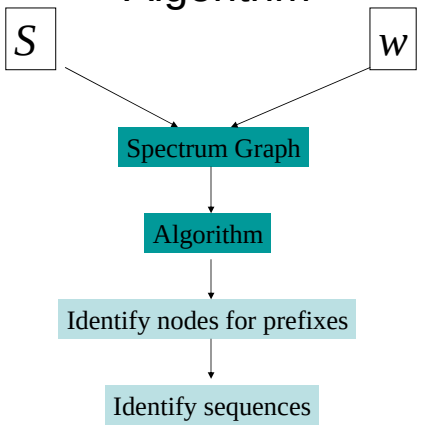
1, list nodes from left to right
2, compute c()
3, report the c() of the right end node
4, back-track to find the path.

Complexity: O(|E|) time and space.

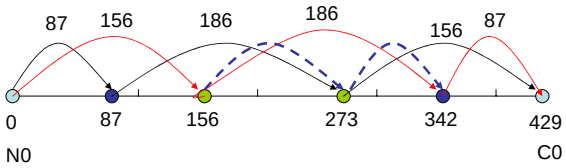
Pros and cons

- Allow missing ions by allowing edges longer than one amino acid.
- Can use a simple scoring function by putting weights on edges or nodes
- It is very fast and space efficient.
- However, it can interpret a peak as both a b-ion and a y-ion

Method 3: Dynamic Programming Algorithm



Construction of a spectrum graph: Feasible Path



Define: A *feasible path* is a path from N_0 to C_0 that goes through exactly one node for each pair.

Formulation of the problem

Given a spectrum graph, find the longest path that goes through every pair of nodes at most once.

Solution: Dynamic Programming

Given an NC-Spectrum Graph *G=(V,E)*

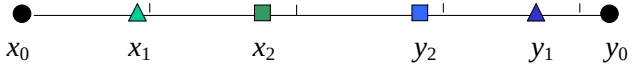
Algorithms: $O(|V|^2)$ time and $O(|V|^2)$ space.

Key Idea

- If we start with a blank spectrum, every peptide with mass *W* will be our candidate.
- If we add one peak into the spectrum at a time, we can narrow down the number of candidate peptides.
- In fact, it is very easy to update our candidates if we just add one peak (using a recursion)

DP Algorithm: order of mass peaks

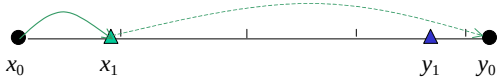
Let the nodes from left to right be $x_0,\ldots, x_k,x_k,\ldots,y_0$



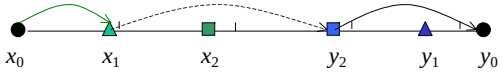
Initialize the graph with x_0 and y_0 only
For $i=1$ to k
Loop: add x_i and y_i and compute all feasible paths.

DP: update path

Loop: Use *the feasible paths of* $x_0,\ldots, x_i,y_i,\ldots,y_0$ to compute *the feasible paths of* $x_0,\ldots, x_{i+1},y_{i+1},\ldots,y_0$



Add nodes x_2 and y_2 and compute



Generalization of the model

- Edges can be a sum of multiple numbers.
- Put weights into nodes.
- Allow noise in the data set.
- Allow tolerance for a match.

Current Status

- We can sequence part of the peptides if the quality of a spectrum is good.
- It is still a hard problem.

Reference

1. Taylor JA, Johnson RS. (1997) Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom;11(9)*:1067-75.
2. Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA. (1999) De novo peptide sequencing via tandem mass spectrometry.*J Comput Biol*; 6(3-4):327-42.
3. Chen, T., Kao, M.Y., Tepel, M., Rush, J., and Church, G.M. (2001) A Dynamic Programming Approach to De Novo Peptide Sequencing via Tandem Mass Spectrometry. *Journal of Computational Biology*, 8(3):325-337.
4. Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, Lajoie G (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom.* 2003;17(20):2337-42.
5. Vineet Bafna, Nathan Edwards: On de novo interpretation of tandem mass spectra for peptide identification. *RECOMB 2003*: 9-18
6. Lu, B. and Chen, T. (2003) A Suboptimal Algorithm for De novo Peptide Sequencing via Tandem Mass Spectrometry. *Journal of Computational Biology*, 10(1):1-12.