AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

22BIO211: Intelligence of Biological Systems - 2

FROM 20 TO MORE THAN 100 AMINO ACIDS

Dr. Manjusha Nair M
Amrita School of Computing, Amritapuri

Email : manjushanair@am.amrita.edu
Contact No: 9447745519

From 20 to more than 100 amino acids

- **Proteinogenic amino acids.**
  - *so far, we used just 20 amino acids as the building blocks of proteins*
  - *There are two additional proteinogenic amino acids, called* ***selenocysteine*** *and* ***pyrrolysine****, which are incorporated into proteins by special biosynthetic mechanisms*
- NRPs contain non-proteinogenic amino acids
  - *this expand the number of possible building blocks for antibiotic peptides from 20 to over 100.*
- This creates problems in our current approach to cyclopeptide sequencing.
- The correct peptide now must “compete” with many more incorrect ones for a place on the leaderboard, increasing the chance that the correct peptide will be cut along the way.

From 20 to more than 100 amino acids

- For example, although Tyrocidine B1 ((Val-**Lys**-Leu-Phe-Pro-Trp-Phe-Asn-Gln-Tyr), contains only proteinogenic amino acids,
  - *its close relative, Tyrocidine B*

  *(Val-**Orn**-Leu-Phe-Pro-Trp-Phe-Asn-Gln-Tyr),*
  - *contains a non-proteinogenic amino acid called ornithine (Orn).*
- Because so many non-proteinogenic amino acids exist, bioinformaticians often assume that any integer between 57 and 200 may represent the mass of an amino acid
  - *the "lightest" amino acid, glycine (Gly), has mass 57 Da*
  - *most amino acids have masses smaller than 200 Da.*

From 20 to more than 100 amino acids

- Applying LeaderboardCyclopeptideSequencing on the extended amino acid alphabet (i.e., every integer between 57 and 200 inclusively) to $Spectrum_{10}$ with N = 1000 returns 34 different linear peptides of maximum score.
- One of the highest-scoring peptides is VKLFPWFNQXZ, where X has mass 98 and Z has mass 65.
- Non-standard amino acids successfully competed with standard amino acids for the limited number of positions on the leaderboard, resulting in VKLFPWFNQXZ winning over the correct peptide VKLFPWFNQY.

From 20 to more than 100 amino acids

- LeaderboardCyclopeptideSequencing fails to identify the correct peptide even with only 10% false and missing masses.
- How can we improve?
- We must determine the amino acid composition of a peptide from its spectrum so that we may run LeaderboardCyclopeptideSequencing on this smaller alphabet of amino acids.

How can we determine which amino acids are present in an unknown peptide using only an experimental spectrum?

From 20 to more than 100 amino acids

- One way to determine the amino acid composition of a peptide from its experimental spectrum would be to take the smallest masses present in the spectrum (between 57 and 200 Da).
  - *Even if only a single amino acid mass is missing, then this approach will fail to reconstruct the peptide's amino acid composition.*
- Let's take a different approach - Using spectral convolution

Spectral Convolution - Theoretical Spectrum

- Define the **convolution** of a spectrum by taking all positive differences of masses in the spectrum.
- Spectral convolution for the theoretical spectrum of NQEL

Theoretical: 0 113 114 128 **129** 227 **242** **242** 257 355 356 370 371 484

Experimental: 0 **99** 113 114 128 227 257 **299** 355 356 370 371 484

	""	L	N	Q	E	LN	NQ	EL	QE	LNQ	ELN	QEL	NQE
	0	113	114	128	129	227	242	242	257	355	356	370	371
0													
113	**113**												
114	**114**	1											
128	**128**	15	14										
129	**129**	16	15	1									
227	227	**114**	**113**	99	98								
242	242	**129**	**128**	**114**	**113**	15							
242	242	**129**	**128**	**114**	**113**	15							
257	257	144	143	**129**	**128**	30	15	15					
355	355	242	241	227	226	**128**	**113**	**113**	98				
356	356	243	242	228	227	**129**	**114**	**114**	99	1			
370	370	257	256	242	241	143	**128**	**128**	**113**	15	14		
371	371	258	257	243	242	144	**129**	**129**	**114**	16	15	1	
484	484	371	370	356	355	257	242	242	227	**129**	**128**	**114**	**113**

Spectral Convolution

- The most frequent elements in the convolution between 57 and 200 are (multiplicities in parentheses): 113 (8), 114 (8), 128 (8), 129 (8). - L , N, Q, E
- For example, 113 (the mass of L) has multiplicity 8 in the table
- Six of the eight occurrences of 113 in the table above correspond to subpeptide pairs differing in
  - *an L: L and "";*
  - *LN and N;*
  - *EL and E;*
  - *LNQ and NQ;*
  - *QEL and QE;*
  - *NQEL and NQE.*

Theoretical: 0 113 114 128 129 227 242 242 257 355 356 370 371 484

Experimental: 0 99 113 114 128 227 257 299 355 356 370 371 484

	""	L	N	Q	E	LN	NQ	EL	QE	LNQ	ELN	QEL	NQE
	0	113	114	128	129	227	242	242	257	355	356	370	371
0													
113	113												
114	114	1											
128	128	15	14										
129	129	16	15	1									
227	227	114	113	99	98								
242	242	129	128	114	113	15							
242	242	129	128	114	113	15							
257	257	144	143	129	128	30	15	15					
355	355	242	241	227	226	128	113	113	98				
356	356	243	242	228	227	129	114	114	99	1			
370	370	257	256	242	241	143	128	128	113	15	14		
371	371	258	257	243	242	144	129	129	114	16	15	1	
484	484	371	370	356	355	257	242	242	227	129	128	114	113

L

LN − N

NQEL − NQE

Spectral Convolution - Experimental Spectrum

- Spectral convolution for the experimental spectrum of NQEL

	""	false	L	N	Q	LN	QE	false	LNQ	ELN	QEL	NQE
	0	99	113	114	128	227	257	299	355	356	370	371
0												
99	**99**											
113	**113**	14										
114	**114**	15	1									
128	**128**	29	15	14								
227	227	**128**	**114**	**113**	**99**							
257	257	158	144	143	**129**	30						
299	299	200	186	185	171	72	42					
355	355	256	242	241	227	**128**	98	56				
356	356	257	243	242	228	**129**	**99**	57	1			
370	370	271	257	256	242	143	**113**	71	15	14		
371	371	272	258	257	243	144	**114**	72	16	15	1	
484	484	385	371	370	356	257	227	185	**129**	**128**	**114**	**113**

- The most frequent elements in the convolution between 57 and 200 are (multiplicities in parentheses): 113 (4), 114 (4), 128 (4), 99 (3), 129 (3)- L,N,Q,E

Spectral Convolution

- 129 (the mass of E) pops up three times in the above convolution of the simulated spectrum, even though 129 was missing from the spectrum itself.
- Use most frequently appearing integers in the convolution as a guess for the amino acid composition of an unknown peptide.
  - *In our simulated spectrum for NQEL, the most frequent elements of the convolution in the range from 57 to 200 are (multiplicities in parentheses):*
  - *113 (4), 114 (4), 128 (4), 99 (3), 129 (3)*
- Note that these most frequent elements capture all four amino acids in NQEL.

Spectral Convolution Problem

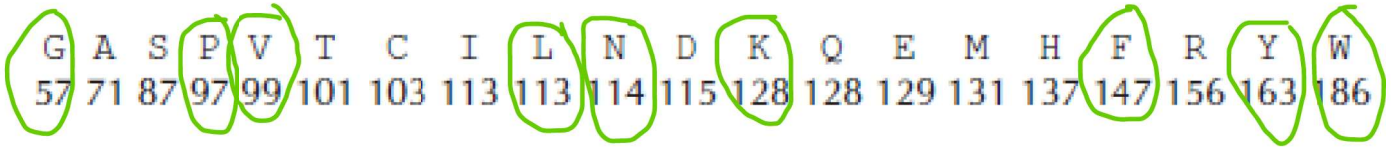
- Spectral Convolution Problem: Compute the convolution of a spectrum.
  - *Input: A collection of integers Spectrum.*
  - *Output: The list of elements in the convolution of Spectrum.*
    - If an element has multiplicity k, it should appear exactly k times; you may return the elements in any order.

Spectral Convolution Problem

- Recall that LeaderboardCyclopeptideSequencing failed to reconstruct Tyrocidine B1 from $Spectrum_{10}$ when using the extended alphabet of amino acids.
- The ten most frequent elements of its spectral convolution in the range from 57 to 200 are (with multiplicities in parentheses):

147 (35) *128 (31)* *97 (28)* *113 (28)* *114 (26)*

186 (23) *57 (21)* *163 (21)* *99 (18)* *145 (18)* ×



- Every mass in this list except for 145 captures an amino acid in Tyrocidine B1! ((VKLFPWFNQY)

Convolution Cyclopeptide Sequencing

- We now have the outline for a new cyclopeptide sequencing algorithm.
  - *Given an experimental spectrum, first compute the convolution of an experimental spectrum.*
  - *Select the M most frequent elements between 57 and 200 in the convolution to form an extended alphabet of candidate amino acid masses.*
    - In order to be fair, we should include the top M elements of the convolution "with ties".
  - *Finally, we run the algorithm LeaderboardCyclopeptideSequencing, where the amino acid masses are restricted to this alphabet.*
  - *We call this algorithm ConvolutionCyclopeptideSequencing.*

Convolution Cyclopeptide Sequencing

- ConvolutionCyclopeptideSequencing
  - *Input: An integer M, an integer N, and a collection of (possibly repeated) integers Spectrum.*
  - *Output: A cyclic peptide* ***LeaderPeptide*** *with amino acids taken only from the top M elements (and ties) of the convolution of Spectrum that fall between 57 and 200, and where the size of Leaderboard is restricted to the top N (and ties).*

Convolution Cyclopeptide Sequencing

- ConvolutionCyclopeptideSequencing (with N = 1000 and M = 20) now correctly reconstructs Tyrocidine B1 from Spectrum10.
- the true test of this algorithm is whether it will work on a noisier spectrum.
  - *Recall that our previous algorithm failed to identify the correct peptide for* $Spectrum_{25}$
- Run ConvolutionCyclopeptideSequencing on $Spectrum_{25}$ with N = 1000 and M = 20.
  - *This algorithm now correctly identifies Tyrocidine B1 from this spectrum!*

Summary

- From 20 to more than 100 amino acids
- Spectral Convolution
  - *Theoretical Spectra*
  - *Experimental Spectra*
- Spectral Convolution Problem
- Convolution Cyclopeptide Sequencing