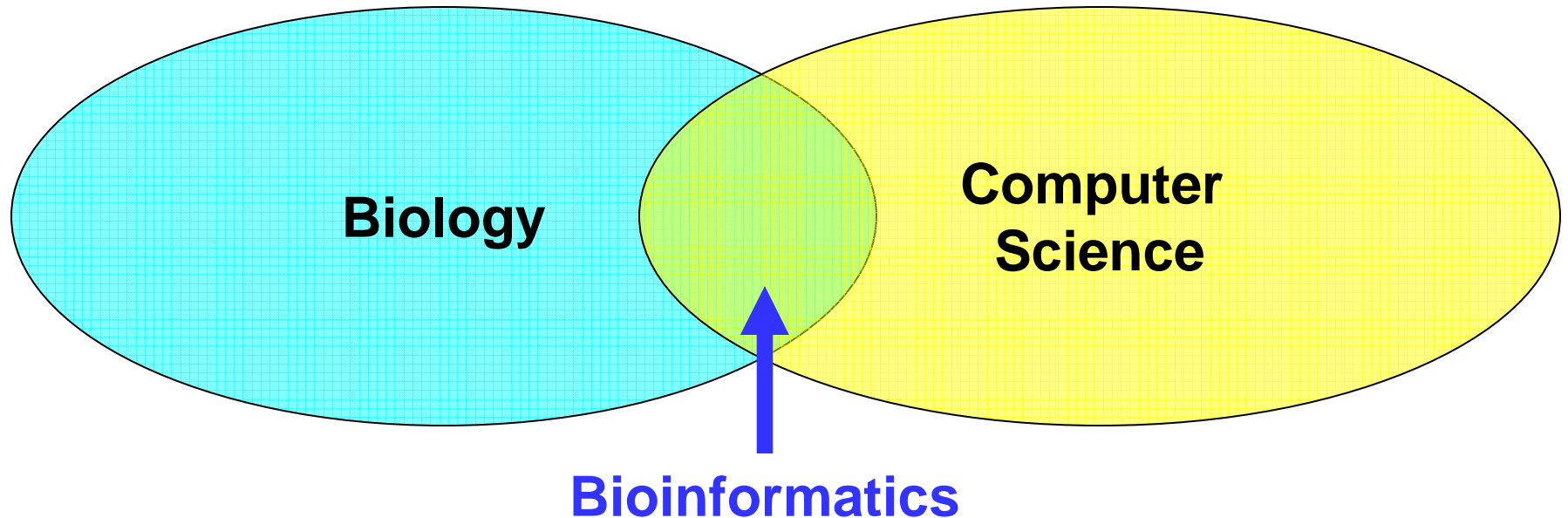


Some Experiences with Formulating Algorithmic Problems for Bioinformatics

Ming-Yang Kao

**Electrical Engineering and Computer Science
Northwestern University**

Collaboration between Biology and Computer Science



Two different cultures



Various questions about the value of a research project.

Two Frequently Asked or Overheard Questions

1. *What is the “science” in this bioinformatics project?*
2. *Are you solving the right problem?*

An Algorithmic Research Perspective

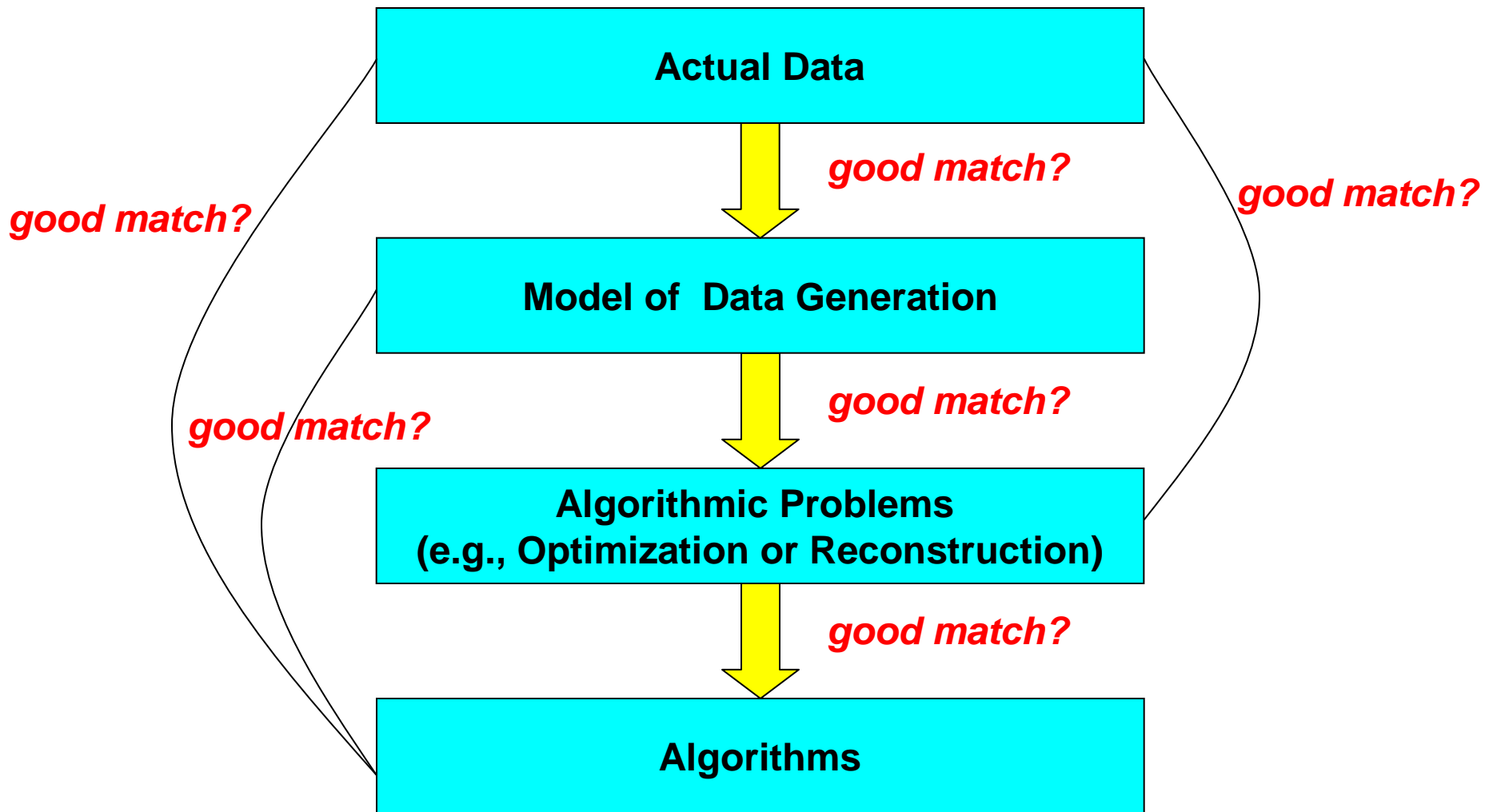
Q1: What is the “science” in this bioinformatics project?

- *Understand the computational complexity of a bioinformatics (or biology) problem.*

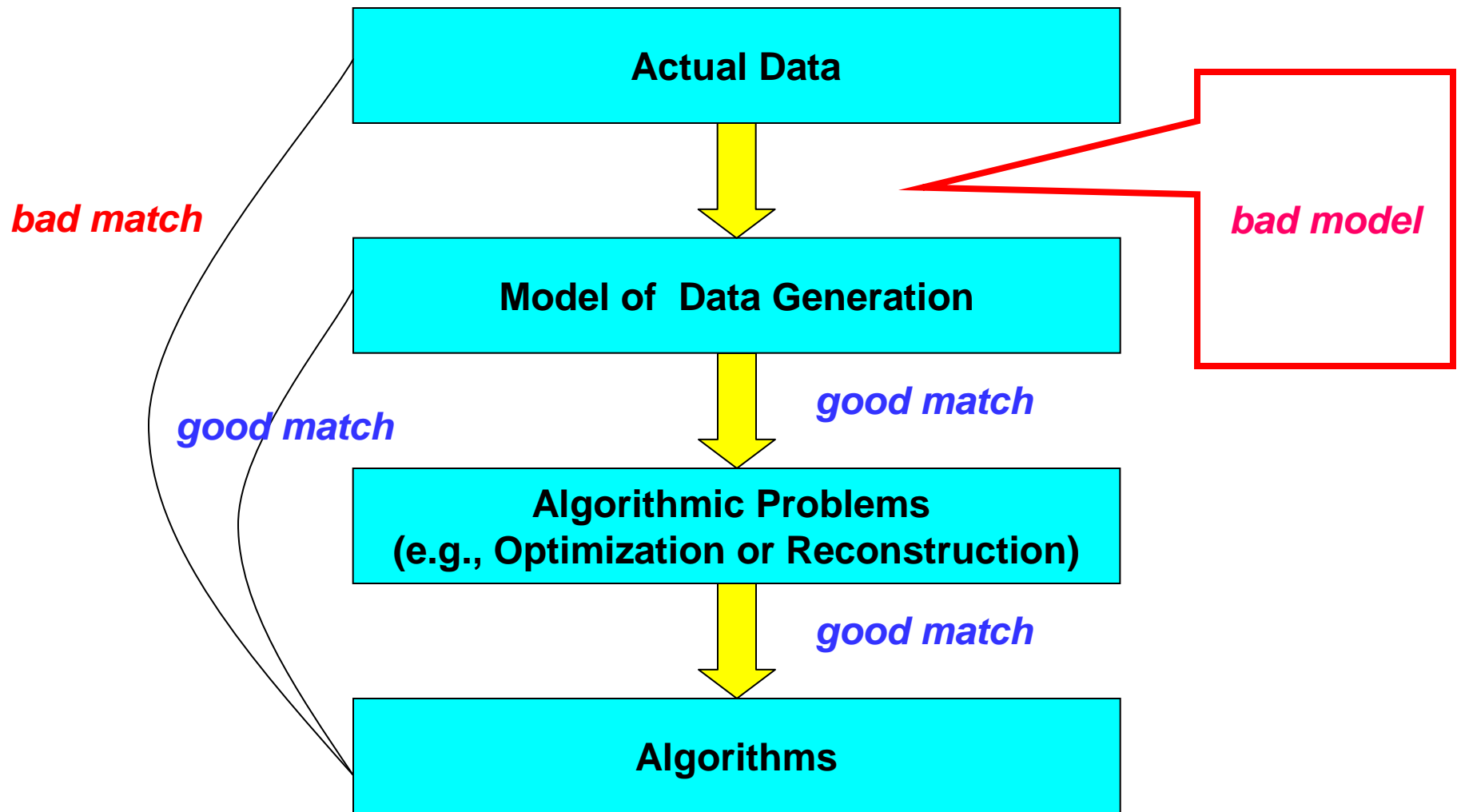
Q2: Are you solving the right problem?

- *What can go wrong in the process of formulating and solving a problem?*
- *Which problem is the right problem, e.g., optimization or reconstruction?*
- *What can difficult bioinformatics problems do for computer science (or biology)?*

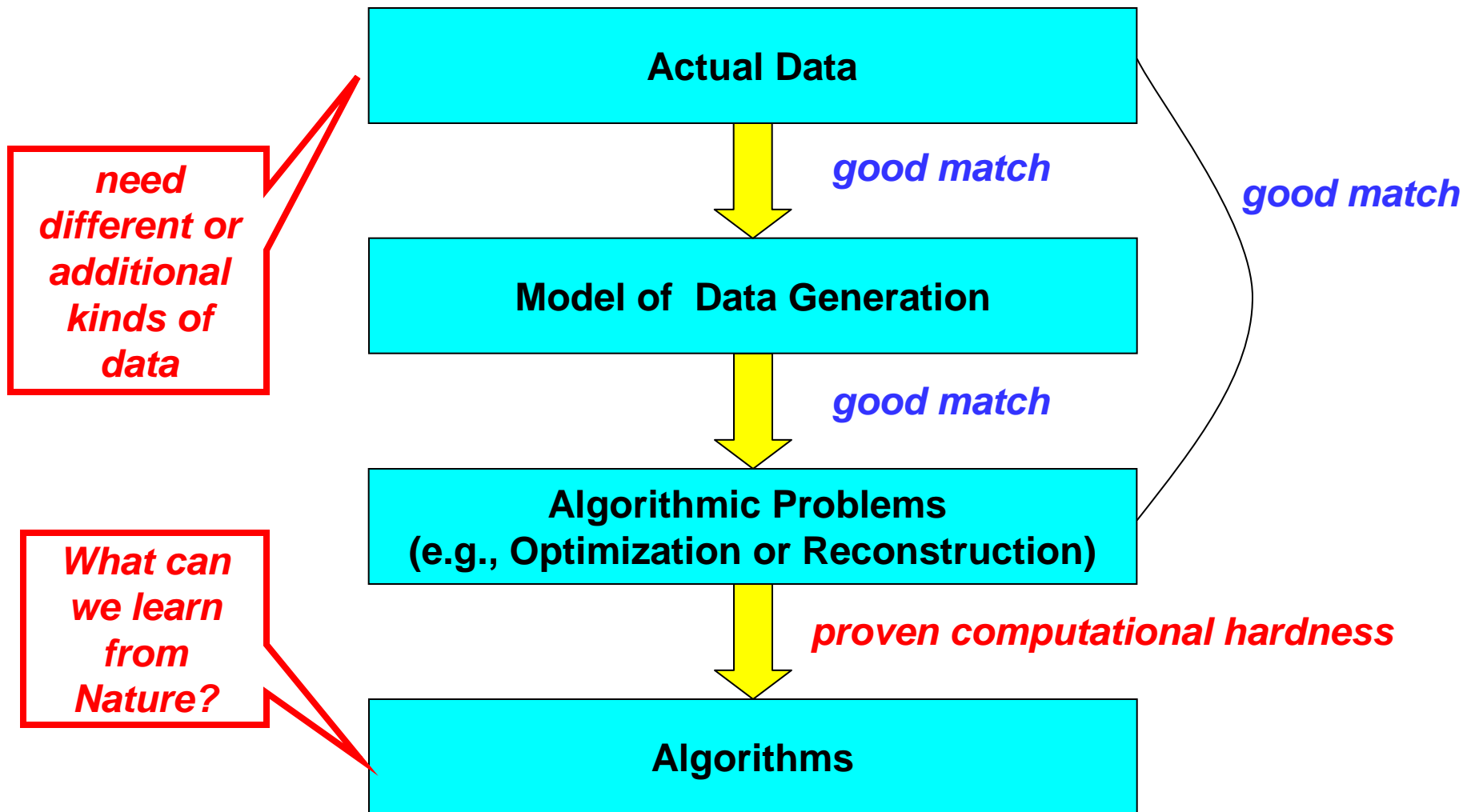
A General Framework for Algorithmic Research for Bioinformatics



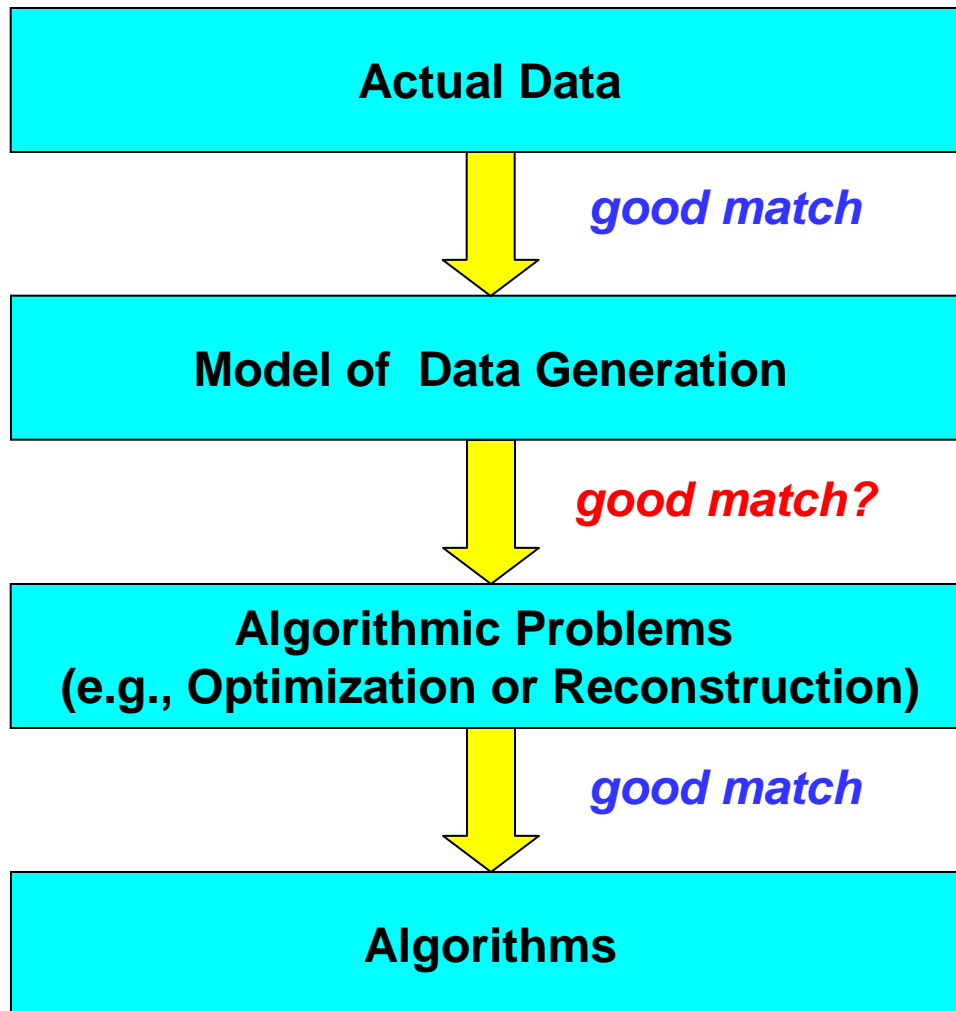
Case #1



Case #2



Case #3

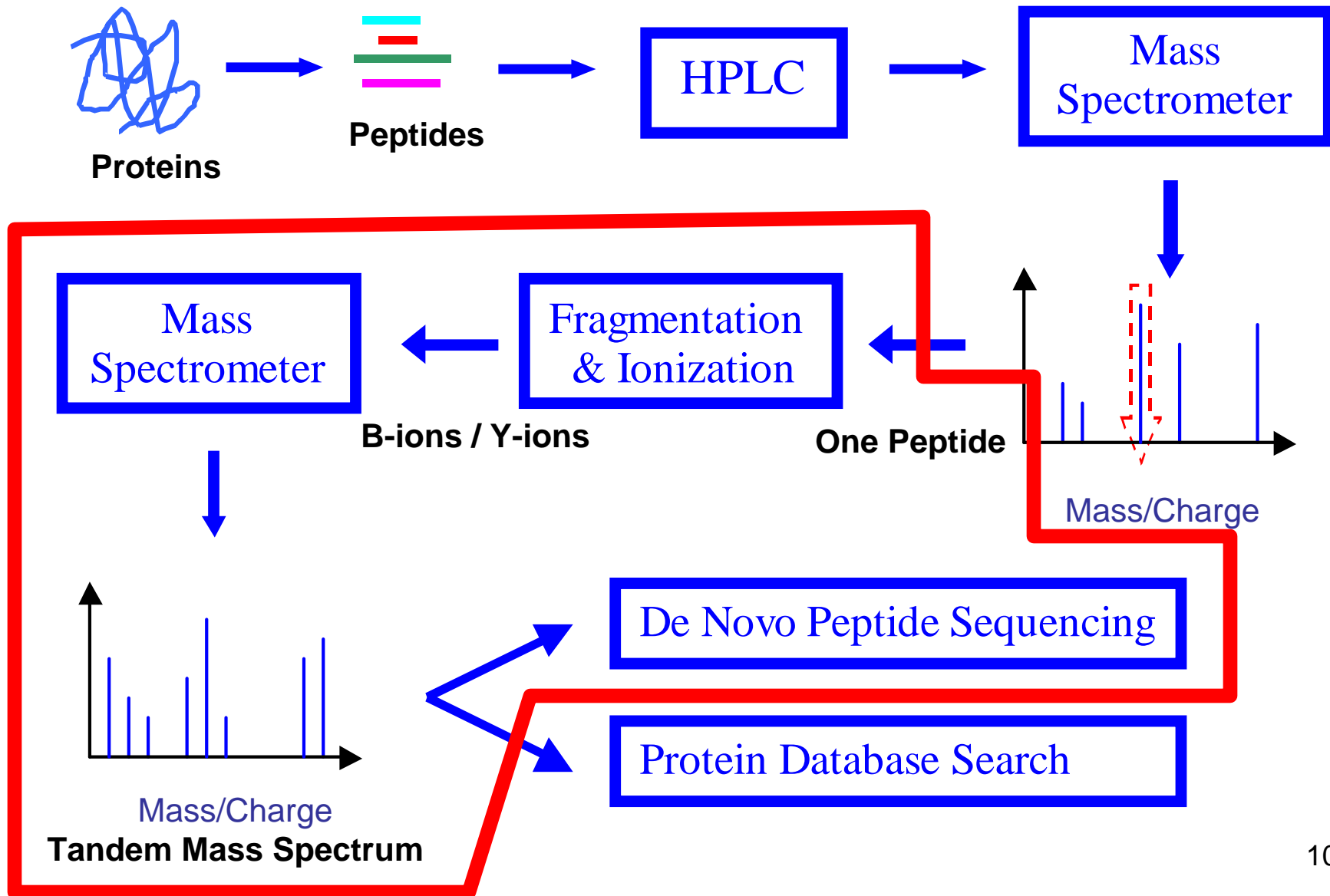


An Example

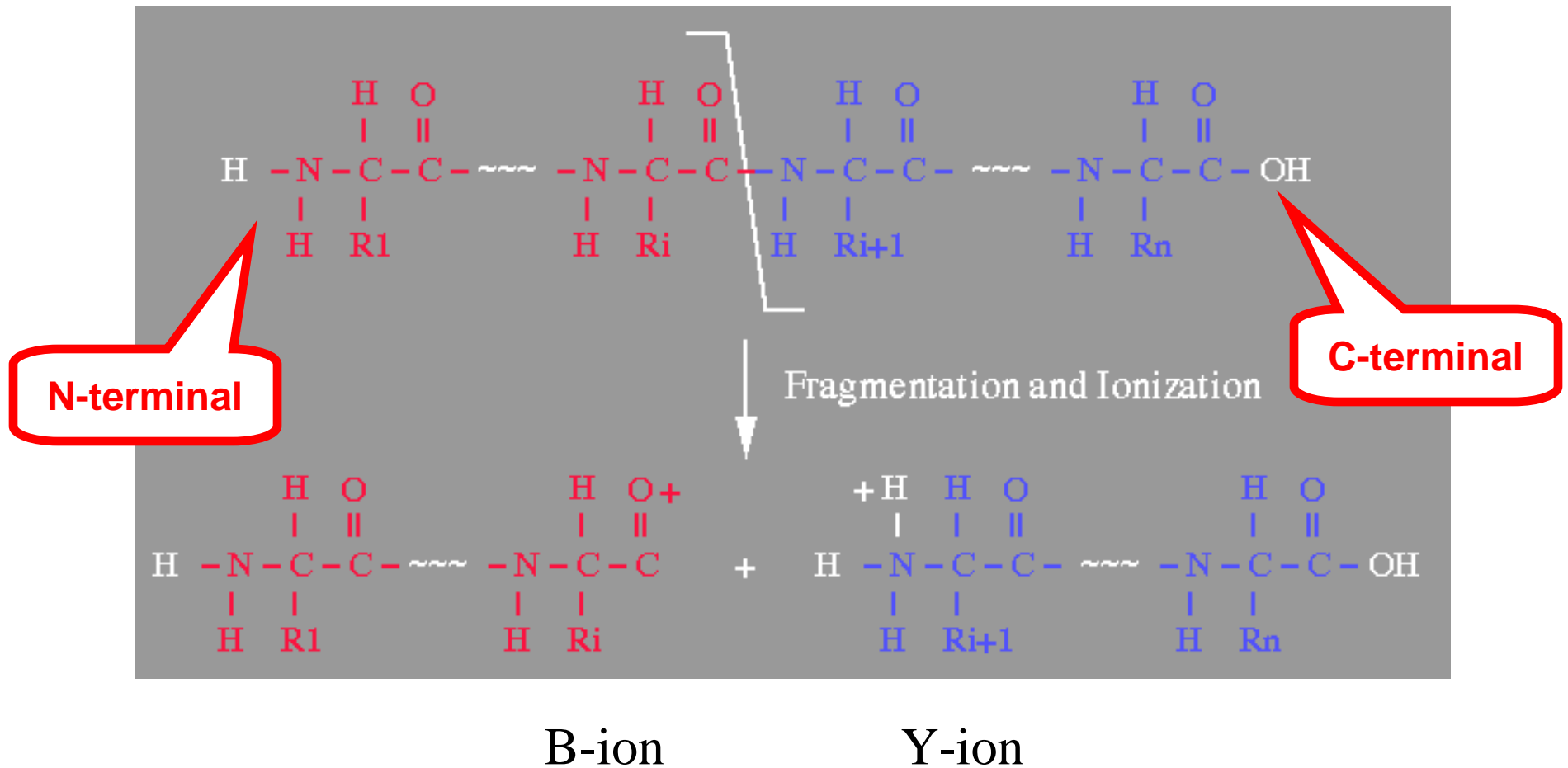
De Novo Protein Peptide Sequencing

Optimization or Reconstruction?

Protein Identification: HPLC-MS-MS



Peptide Fragmentation and Ionization



Complementary: $\text{Mass}(\text{B-ion}) + \text{Mass}(\text{Y-ion}) = \text{Mass}(\text{peptide}) + 4\text{H} + \text{O}$

B-ions and Y-ions

Fragmentation

Peptide ($\alpha \cdot R_1 - R_2 - R_3 \cdot \beta$)

All **b** – ions

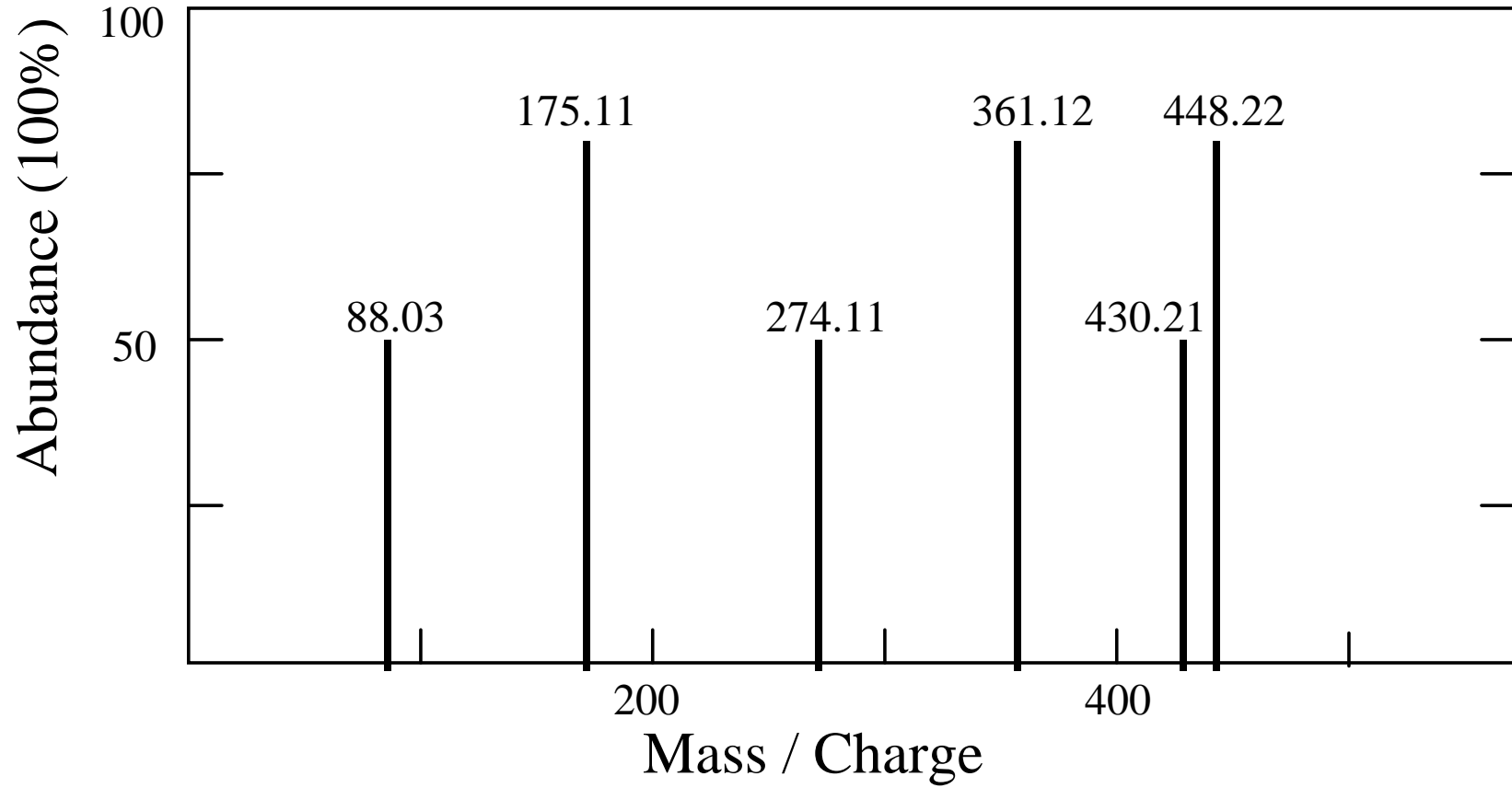
b_1	$(\alpha \cdot R_1)^+$
b_2	$(\alpha \cdot R_1 - R_2)^+$
b_3	$(\alpha \cdot R_1 - R_2 - R_3)^+$

All **y** – ions

y_1	$(R_3 \cdot \beta)^+$
y_2	$(R_2 - R_3 \cdot \beta)^+$
y_3	$(R_1 - R_2 - R_3 \cdot \beta)^+$

$$\alpha = 1, \beta = 19$$

Tandem Mass Spectrum

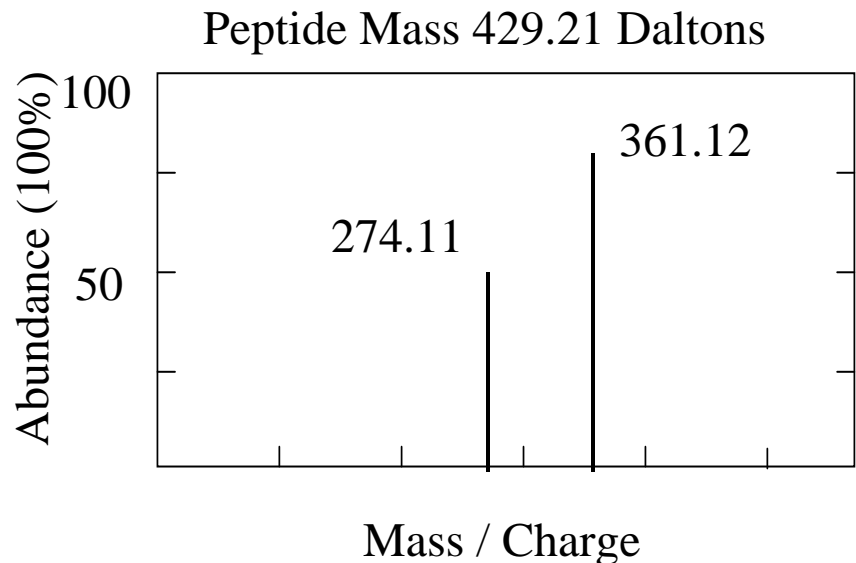


Amino Acid Mass Table

A	71.08		M	131.19
C	103.14		N	114.10
D	115.09		P	97.12
E	129.12		Q	128.13
F	147.18		R	156.19
G	57.05		S	87.08
H	137.14		T	101.11
I	113.16		V	99.13
K	128.17		W	186.21
L	113.16		Y	163.18

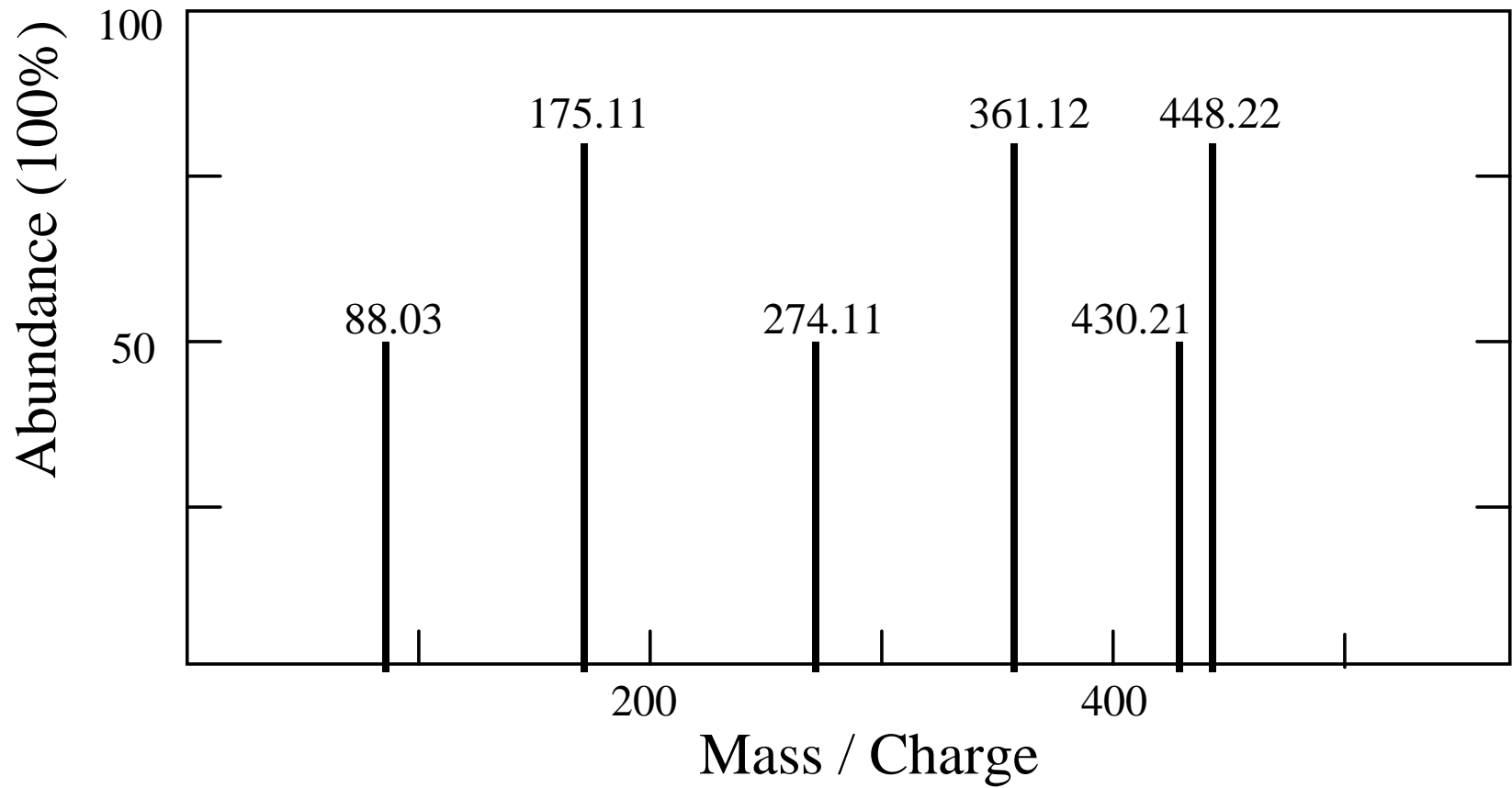
De Novo Peptide Sequencing Problem

- **Input:**
 - (1) the mass W of an unknown target peptide, and
 - (2) a set S of the masses of some or all b-ions and y-ions of the peptide.
- **Output:** a peptide P such that
 - (1) $\text{mass}(P) = \underline{W}$ and
 - (2) S is a subset of all the ion masses of P .



$P = \text{SWR}$,
 $\text{Mass}(P) = 429.21$,
 $\text{Ions}(P) =$
 $\{88.03, 175.11, \mathbf{274.11},$
 $\mathbf{361.12}, 430.21, 448.22\}$

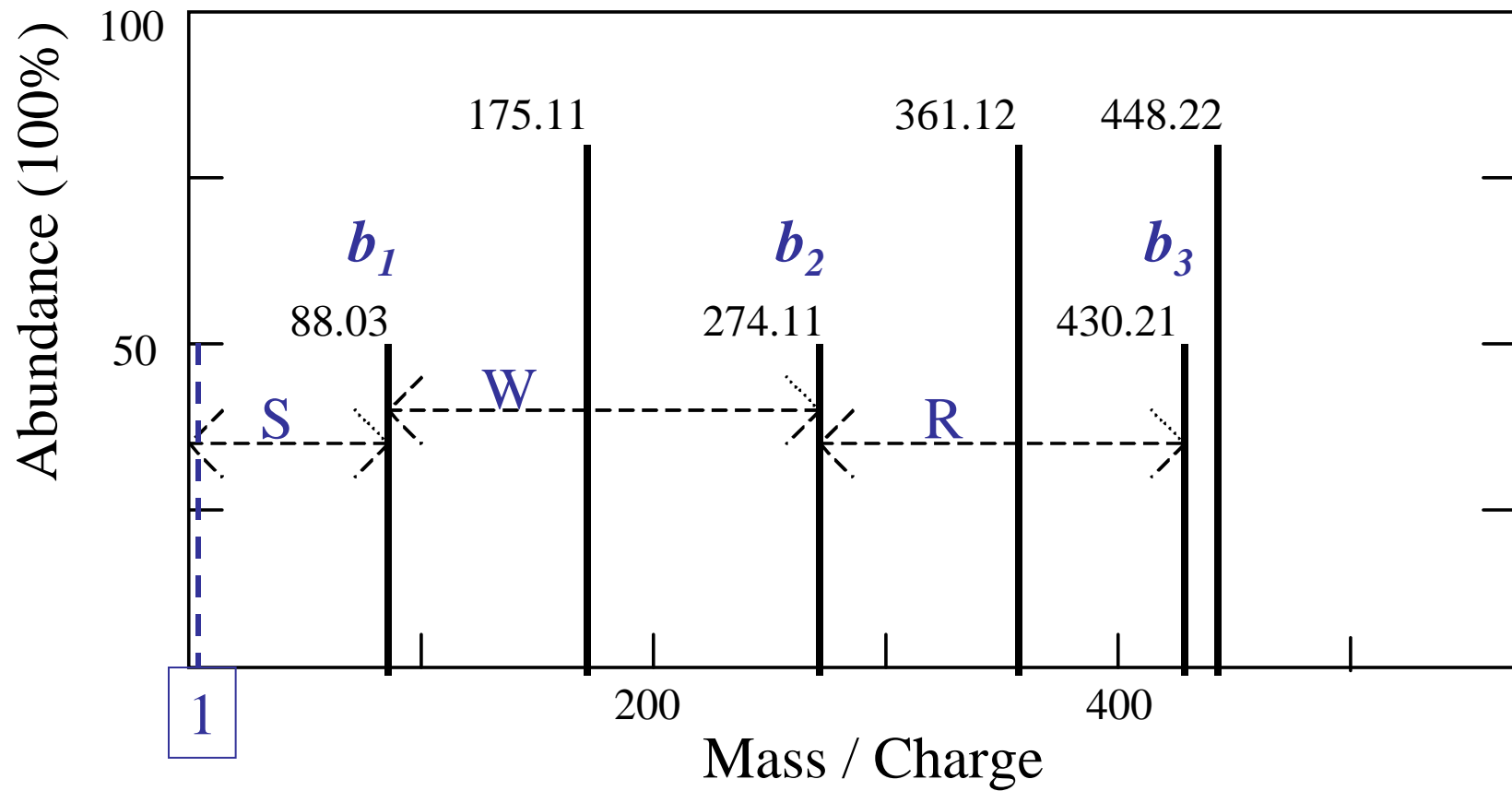
A Complete Tandem Mass Spectrum



Peptide Mass 429.21 Daltons

Observation #1

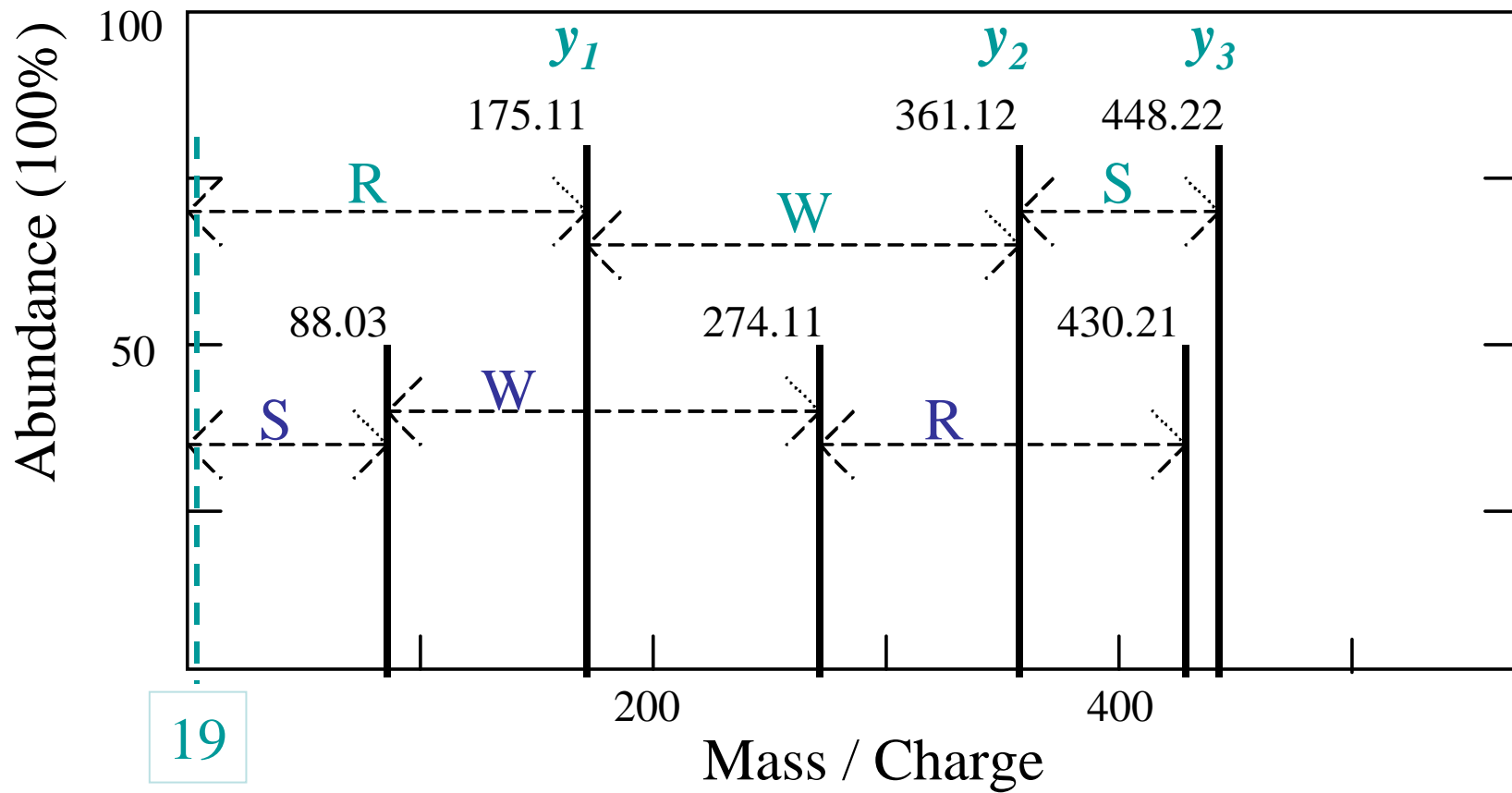
All B-ions form a forward mass ladder.



Peptide Mass 429.21 Daltons

Observation #2

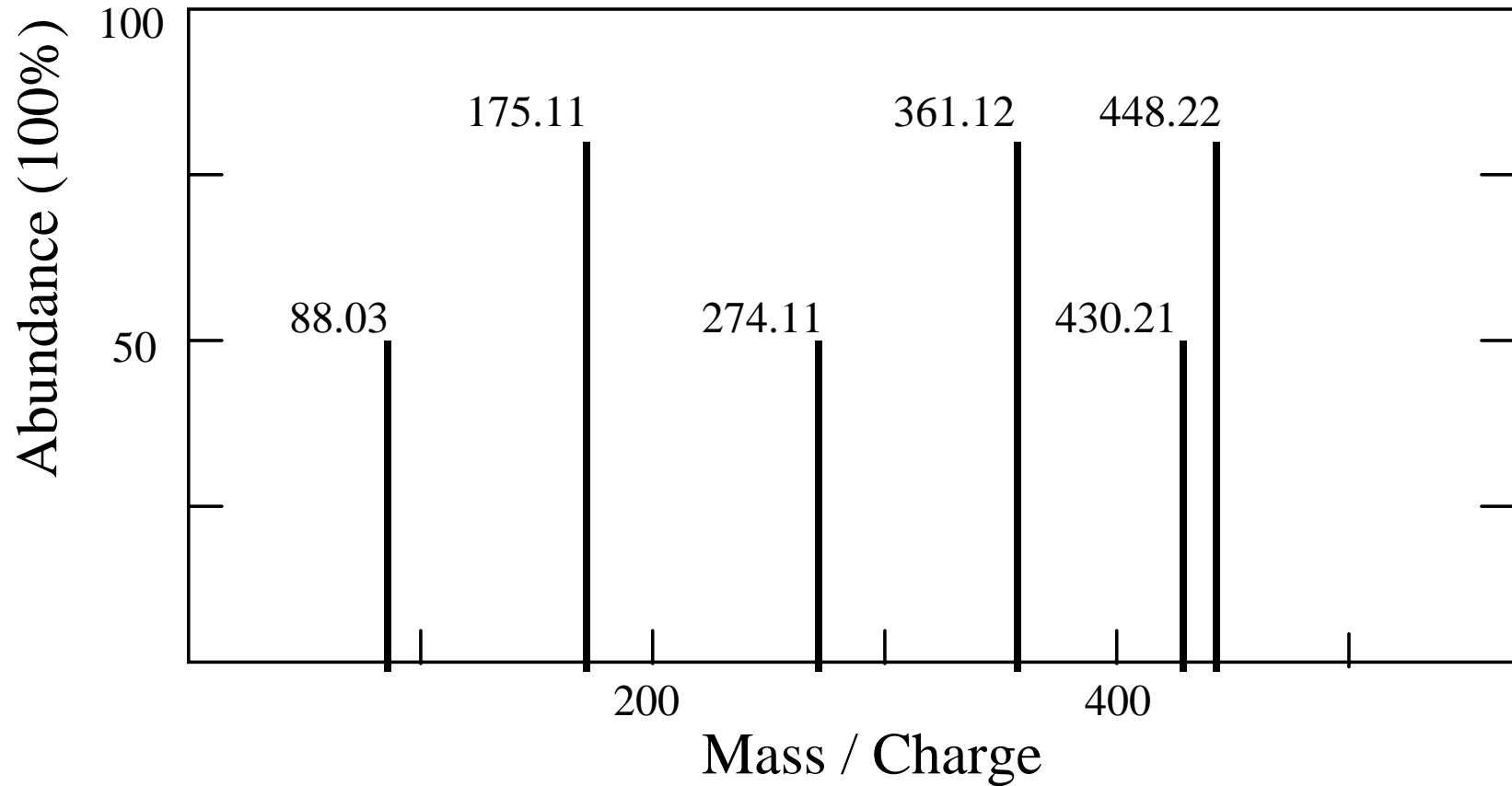
All Y-ions form a reverse mass ladder.



Peptide Mass 429.21 Daltons

Difficulty #1

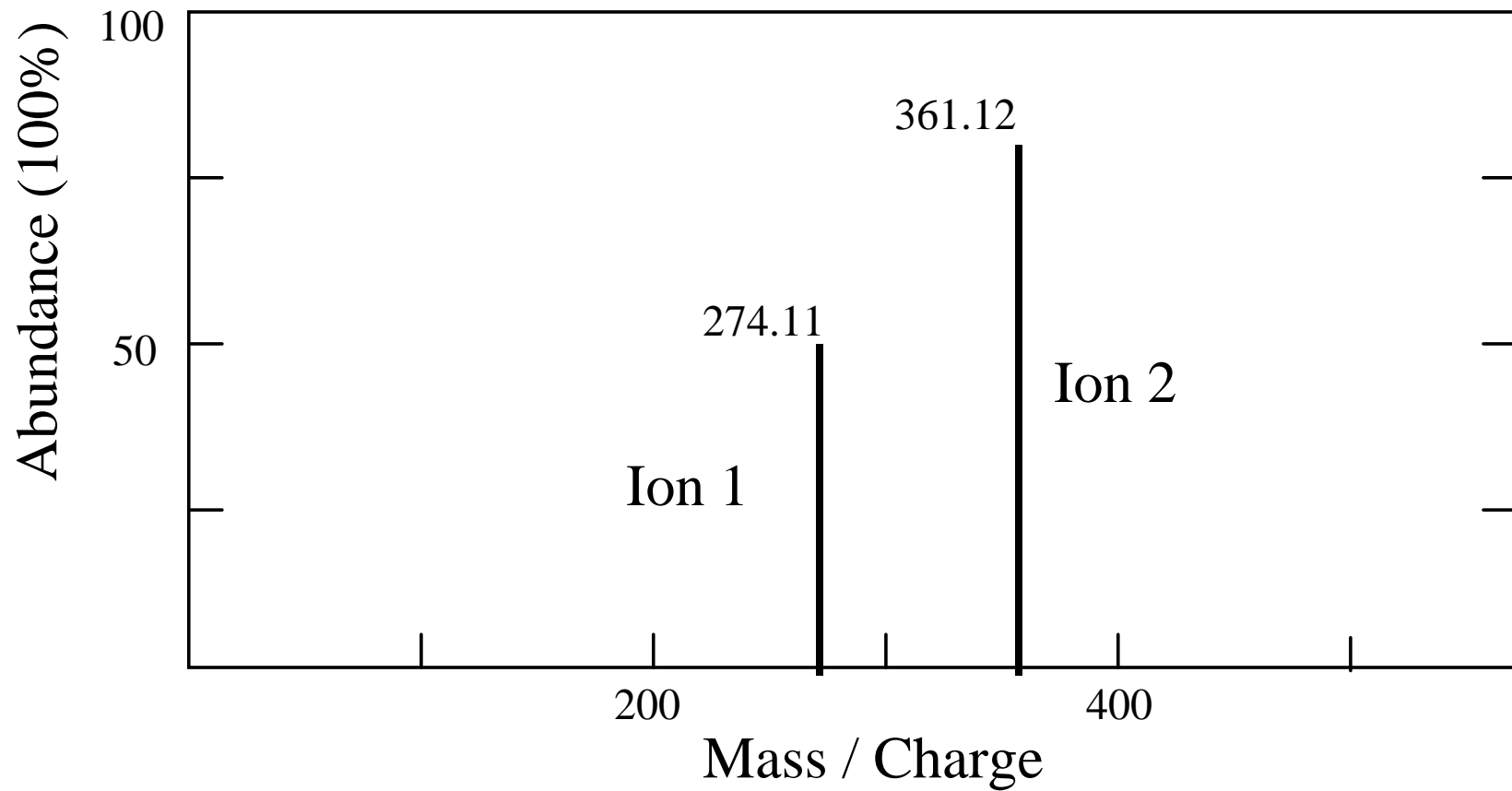
It is unknown whether an ion is a B-ion or a Y-ion.



Peptide Mass 429.21 Daltons

Difficulty #2

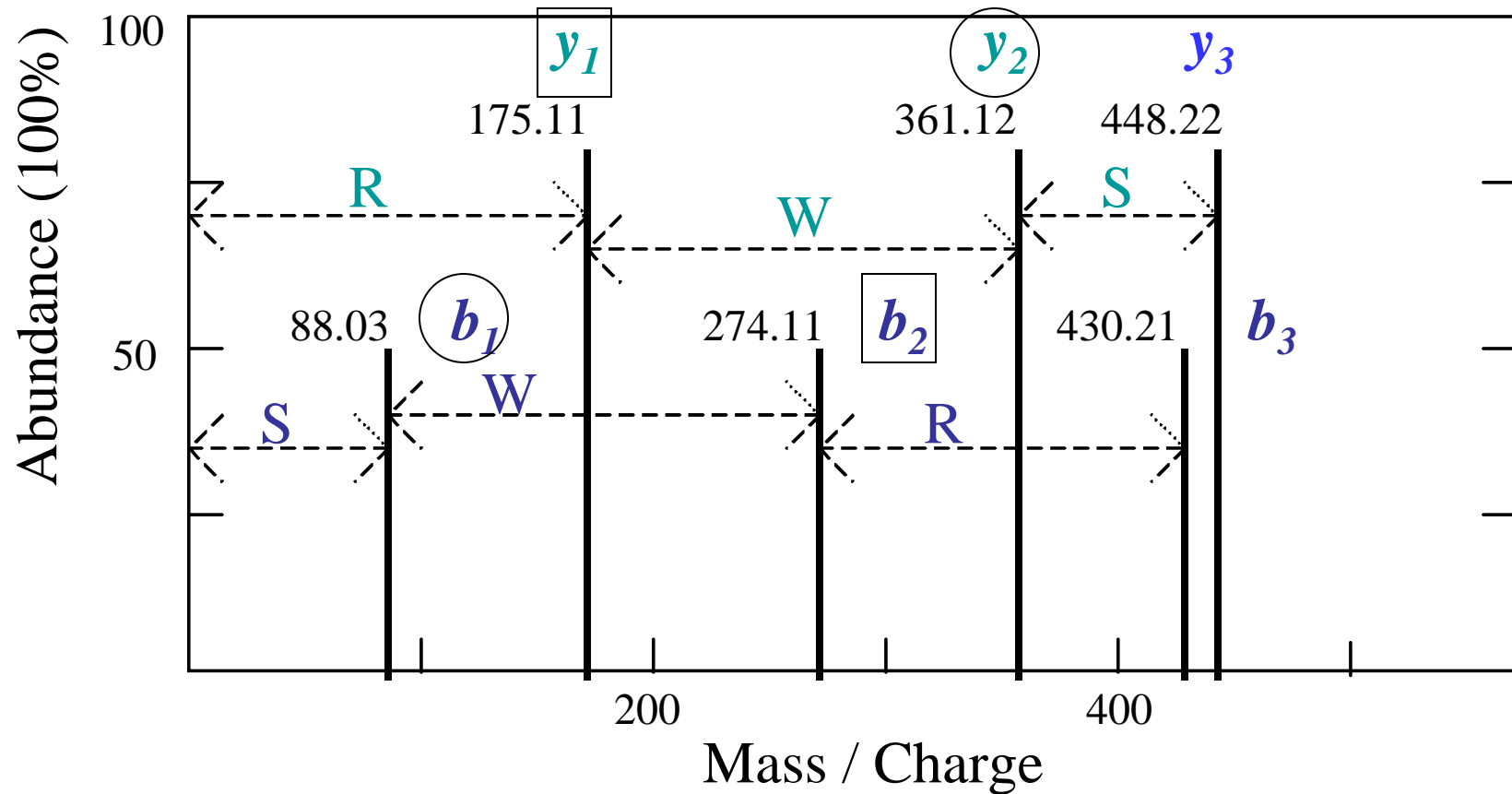
There are missing ions.



Peptide Mass 429.21 Daltons

Feature #3

Complementary Ion Pairs: b_1/y_2 and b_2/y_1



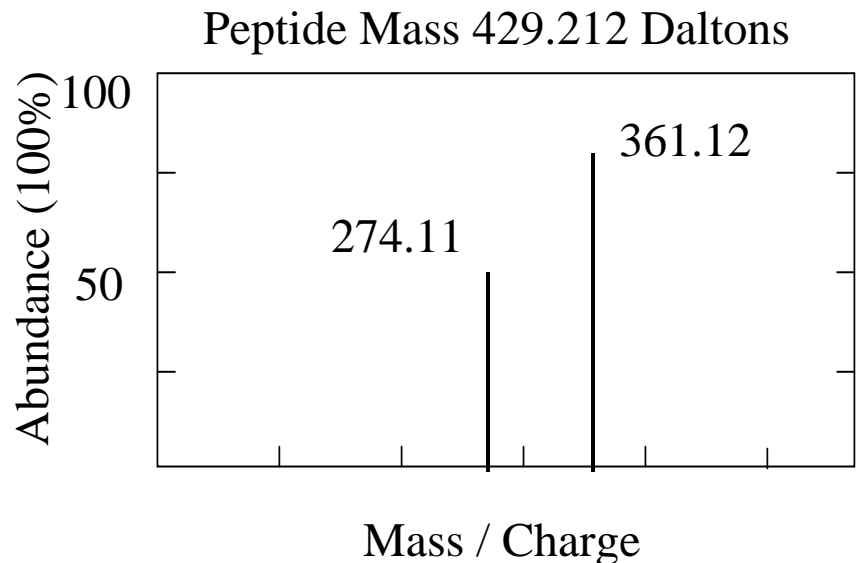
Peptide Mass 429.21 Daltons

Formulating the Algorithmic Problem

1. T = an alphabet of 20 characters a_1, a_2, \dots, a_{20} .
2. two special characters: alpha and beta.
3. the mass of alpha = 1,
the mass of beta = 19,
the mass of a_i is m_i .
4. A **peptide sequence** is $x_1, x_2, x_3, \dots, x_{n-1}, x_n$, where each x_i is from T .
5. A **b-ion** is $x_0, x_1, x_2, \dots, x_i$ for some $1 \leq i \leq n$, where $x_0 = \text{alpha}$.
6. A **y-ion** is $x_i, \dots, x_{n-2}, x_{n-1}, x_n, x_{n+1}$ for some $1 \leq i \leq n$, where $x_{n+1} = \text{beta}$.

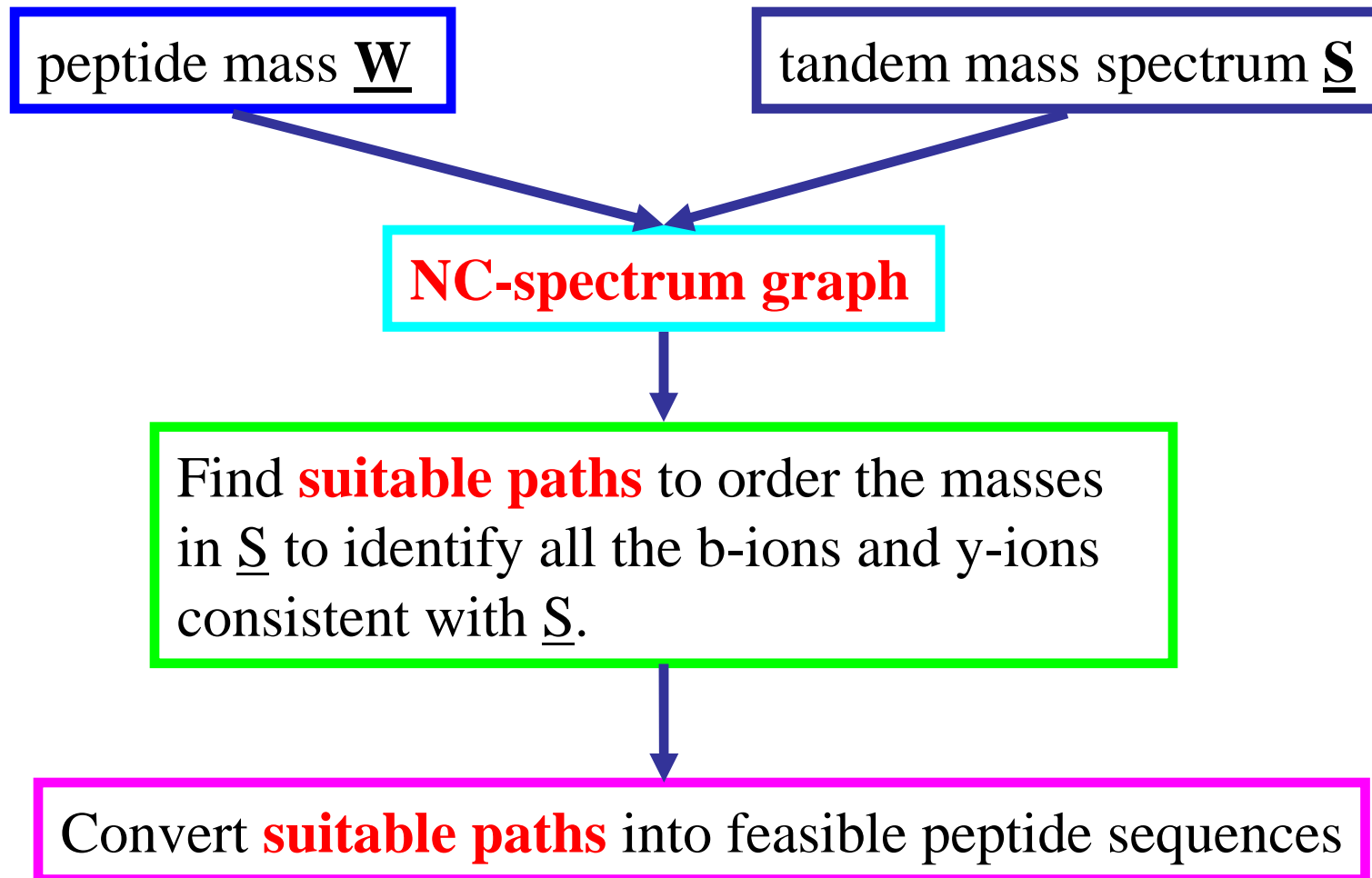
De Novo Peptide Sequencing Problem

- **Input:**
 - (1) the mass W of an unknown target peptide, and
 - (2) a set S of the masses of some or all b-ions and y-ions of the peptide.
- **Output:** a peptide P such that
 - (1) $\text{mass}(P) = \underline{W}$ and
 - (2) S is a subset of all the ion masses of P .

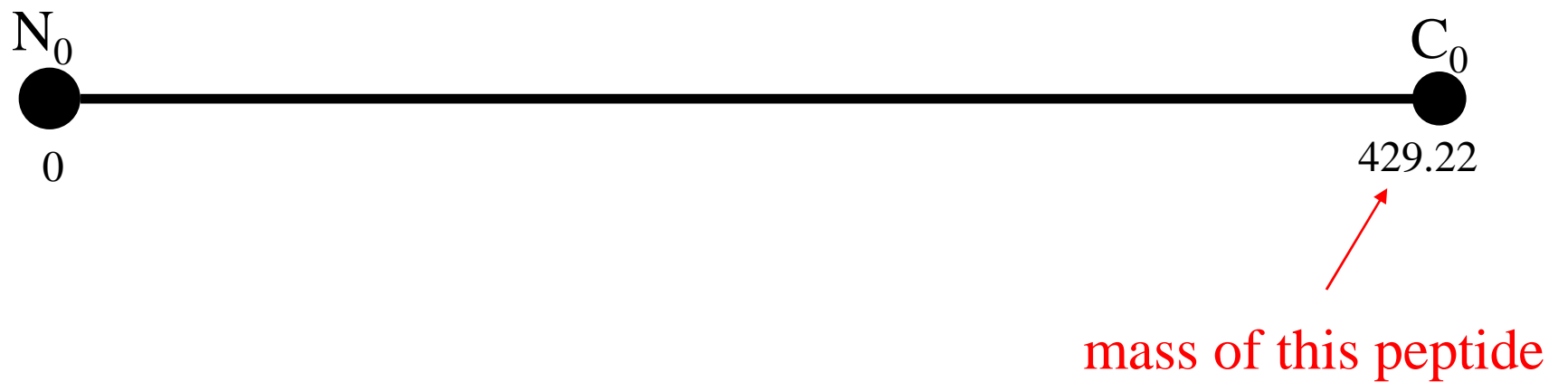


$P = \text{SWR}$,
 $\text{Mass}(P) = 429.21$,
 $\text{Ions}(P) =$
 $\{88.03, 175.11, \mathbf{274.11},$
 $\mathbf{361.12}, 430.21, 448.22\}$

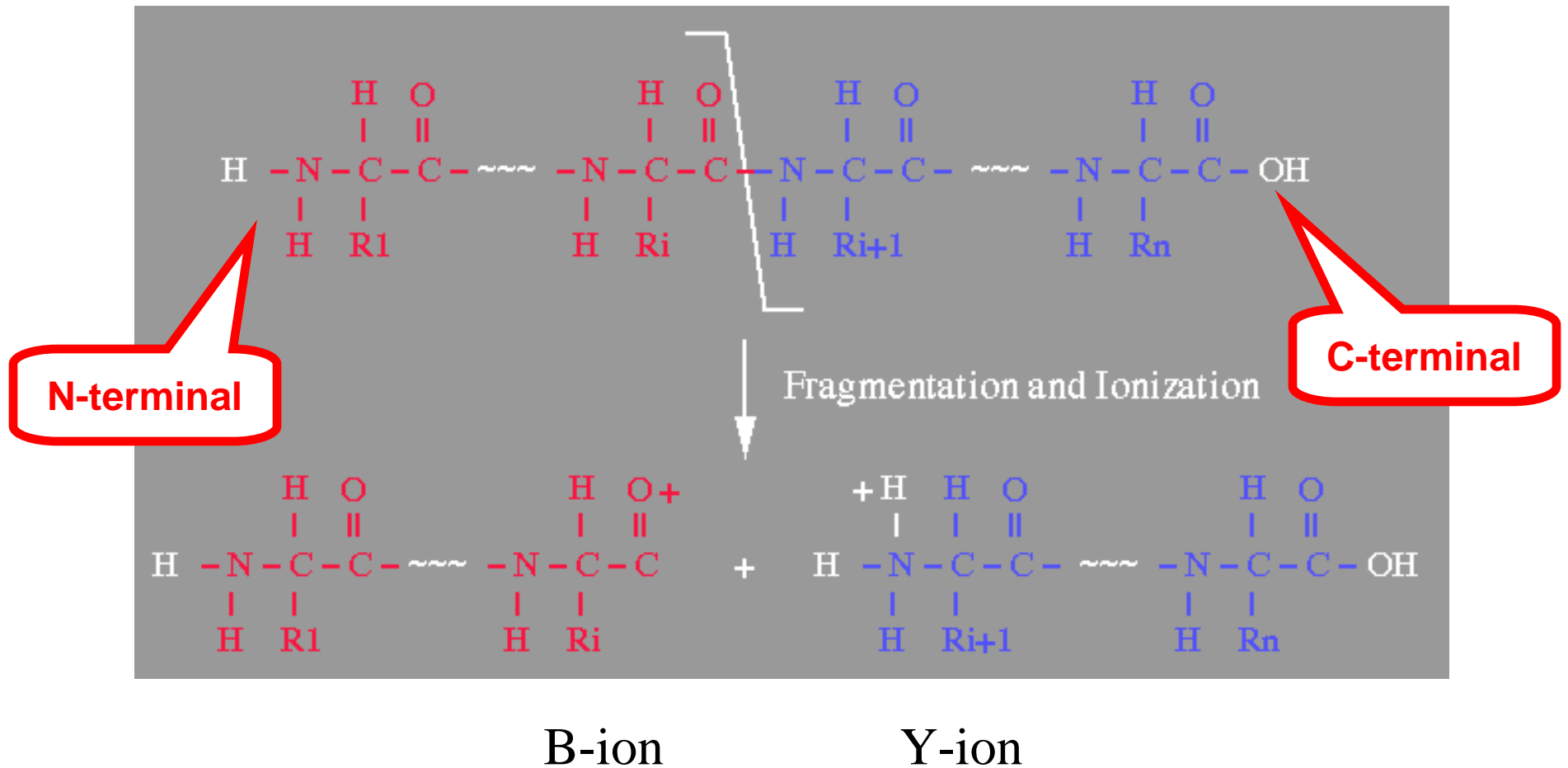
Basic Computing Scheme



NC-Spectrum Graph: Nodes (1)



Peptide Fragmentation and Ionization

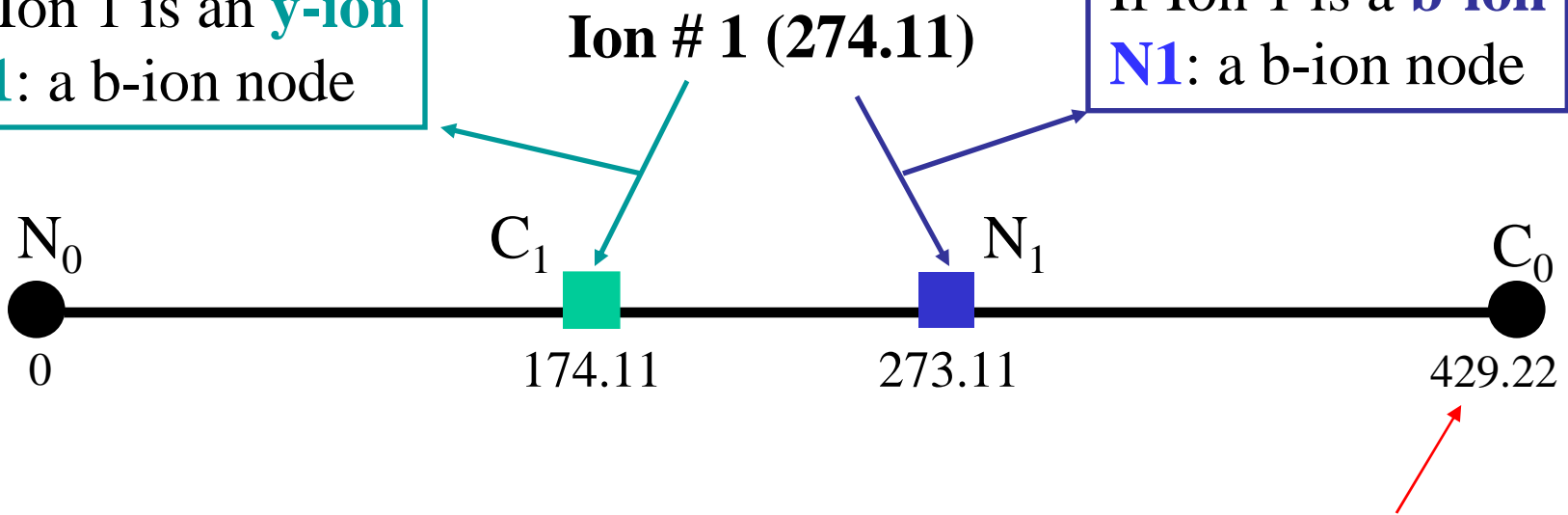


Complementary: $\text{Mass}(\text{B-ion}) + \text{Mass}(\text{Y-ion}) = \text{Mass}(\text{peptide}) + 4\text{H} + \text{O}$

NC-Spectrum Graph: Nodes (2)

Assumption 1:
If Ion 1 is an **y-ion**
C1: a b-ion node

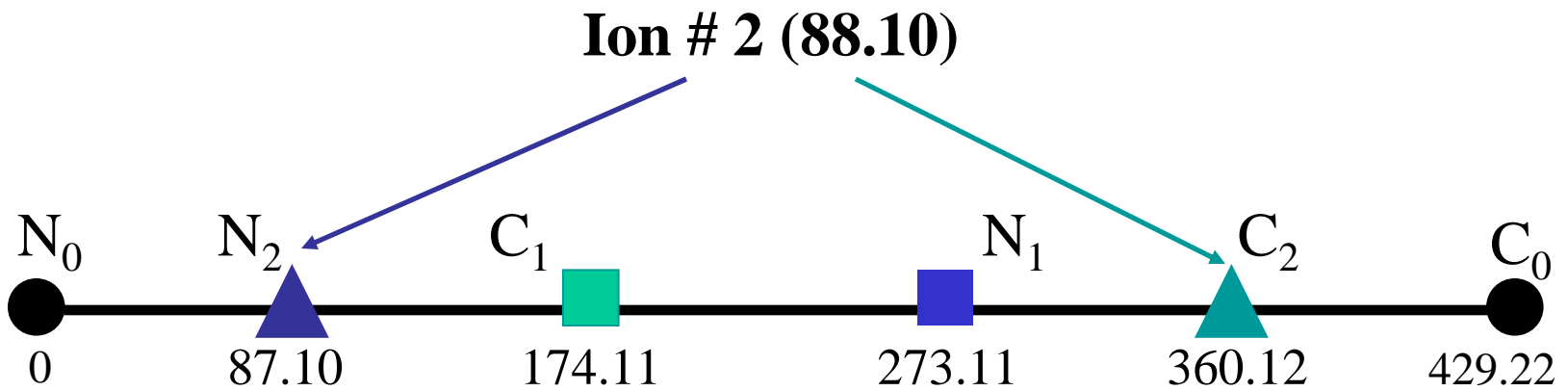
Assumption 2:
If Ion 1 is a **b-ion**
N1: a b-ion node



mass of this peptide

$$\text{mass}(\text{■}) + \text{mass}(\text{■}) = \text{mass}(P) + 18$$

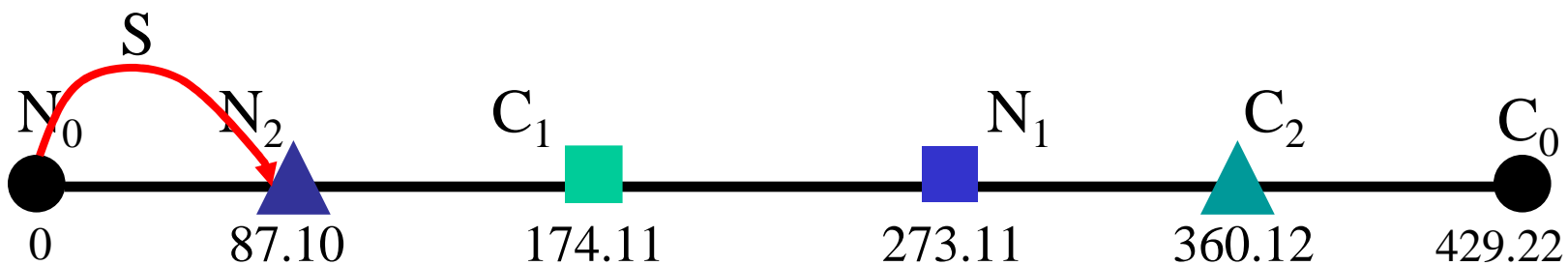
NC-Spectrum Graph: Nodes (3)



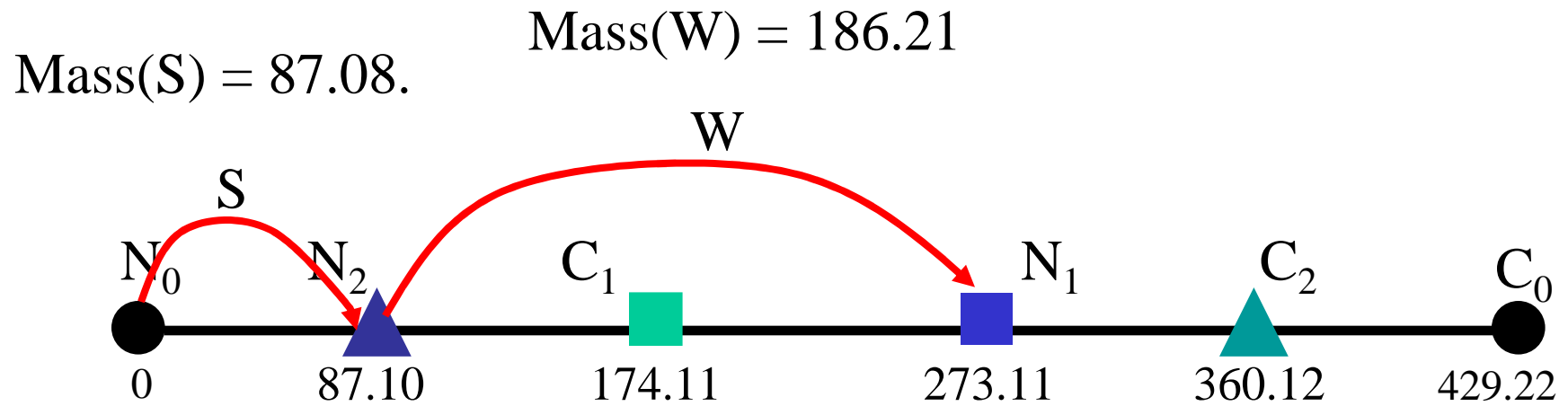
$$\text{mass}(\triangle) + \text{mass}(\triangle) = \text{mass}(\mathbf{P}) + 18$$

NC-Spectrum Graph: Edges (1)

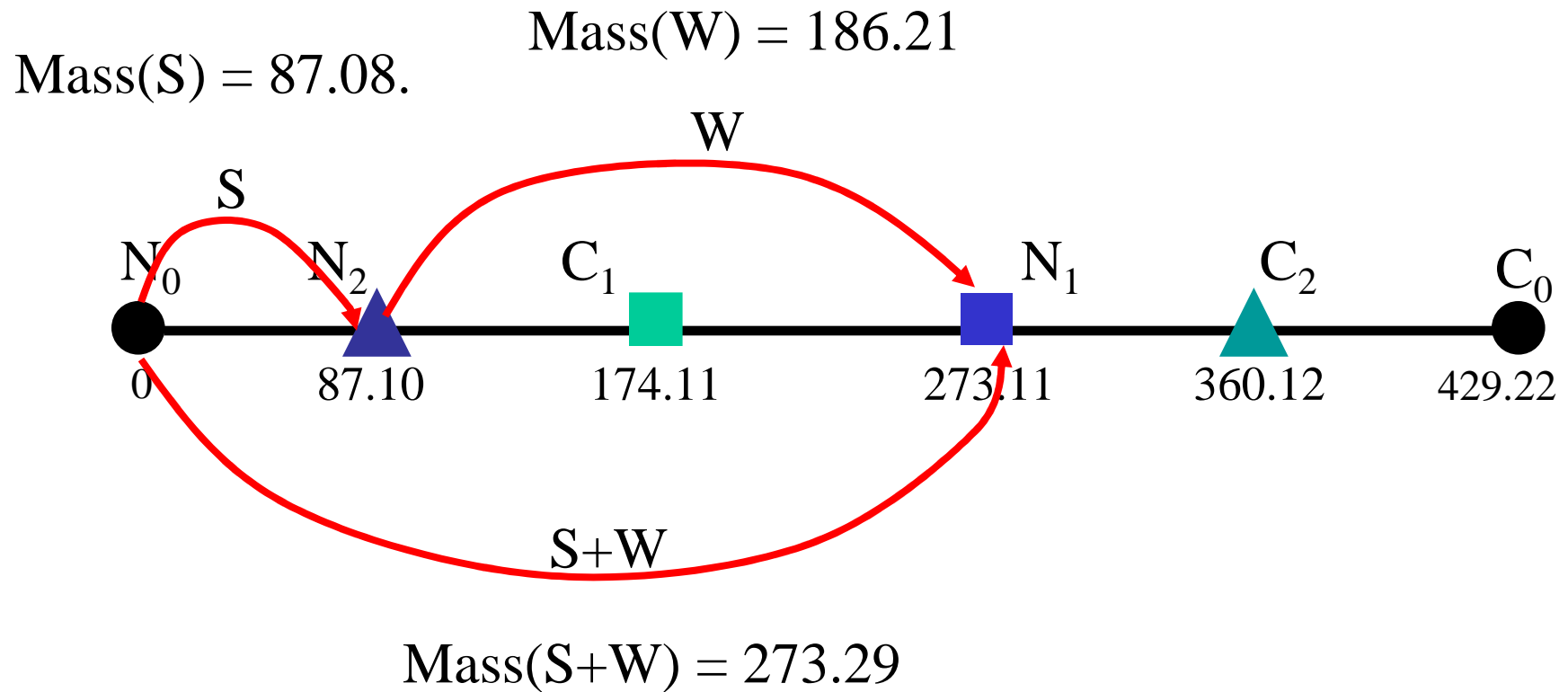
Mass(S) = 87.08.



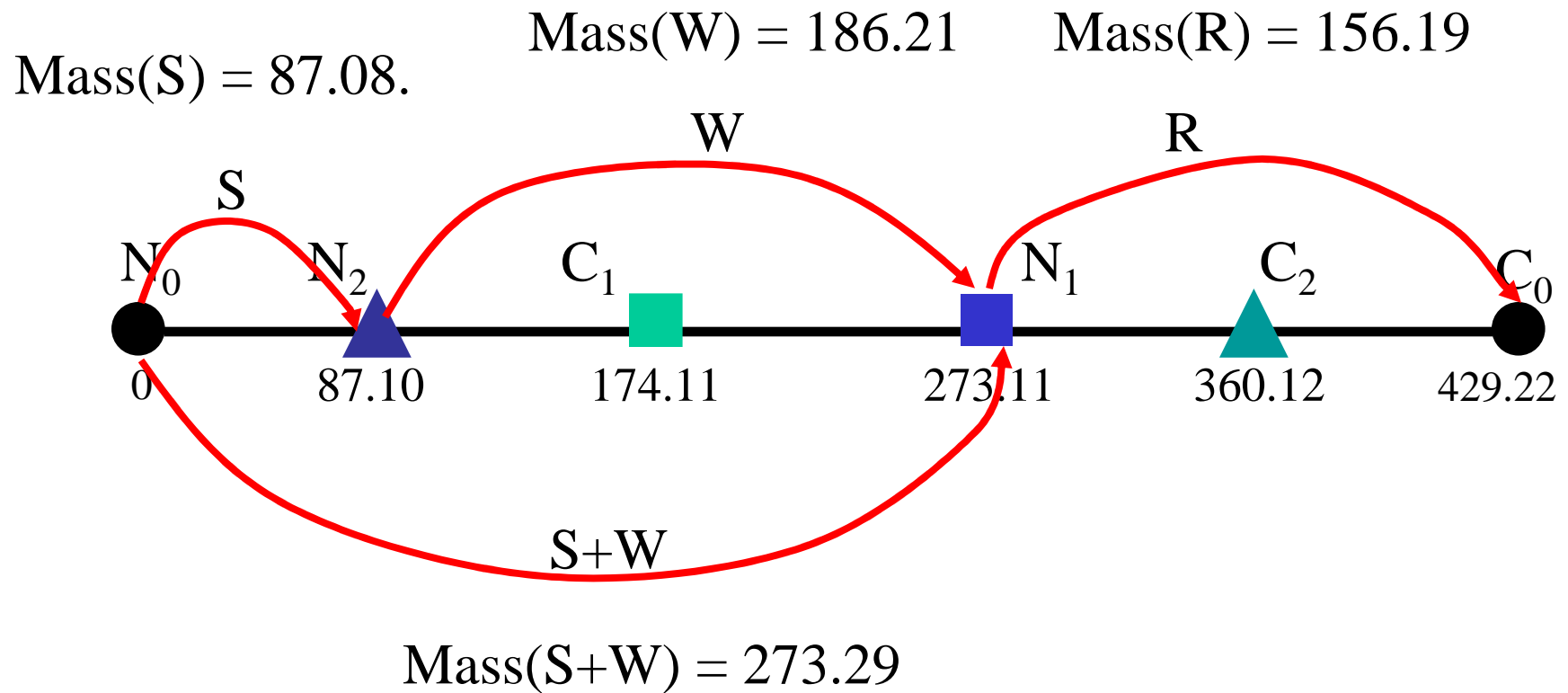
NC-Spectrum Graph: Edges (2)



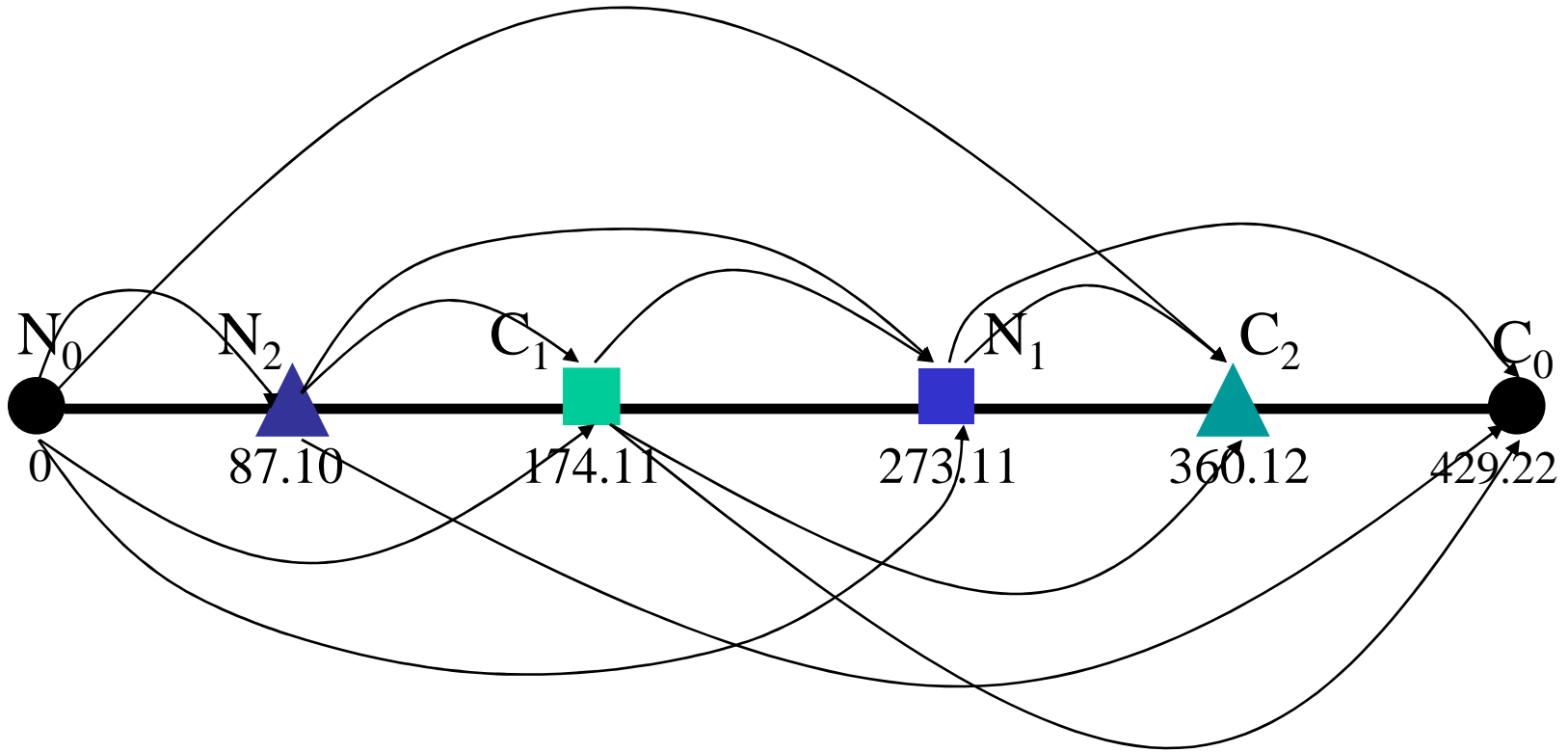
NC-Spectrum Graph: Edges (3)



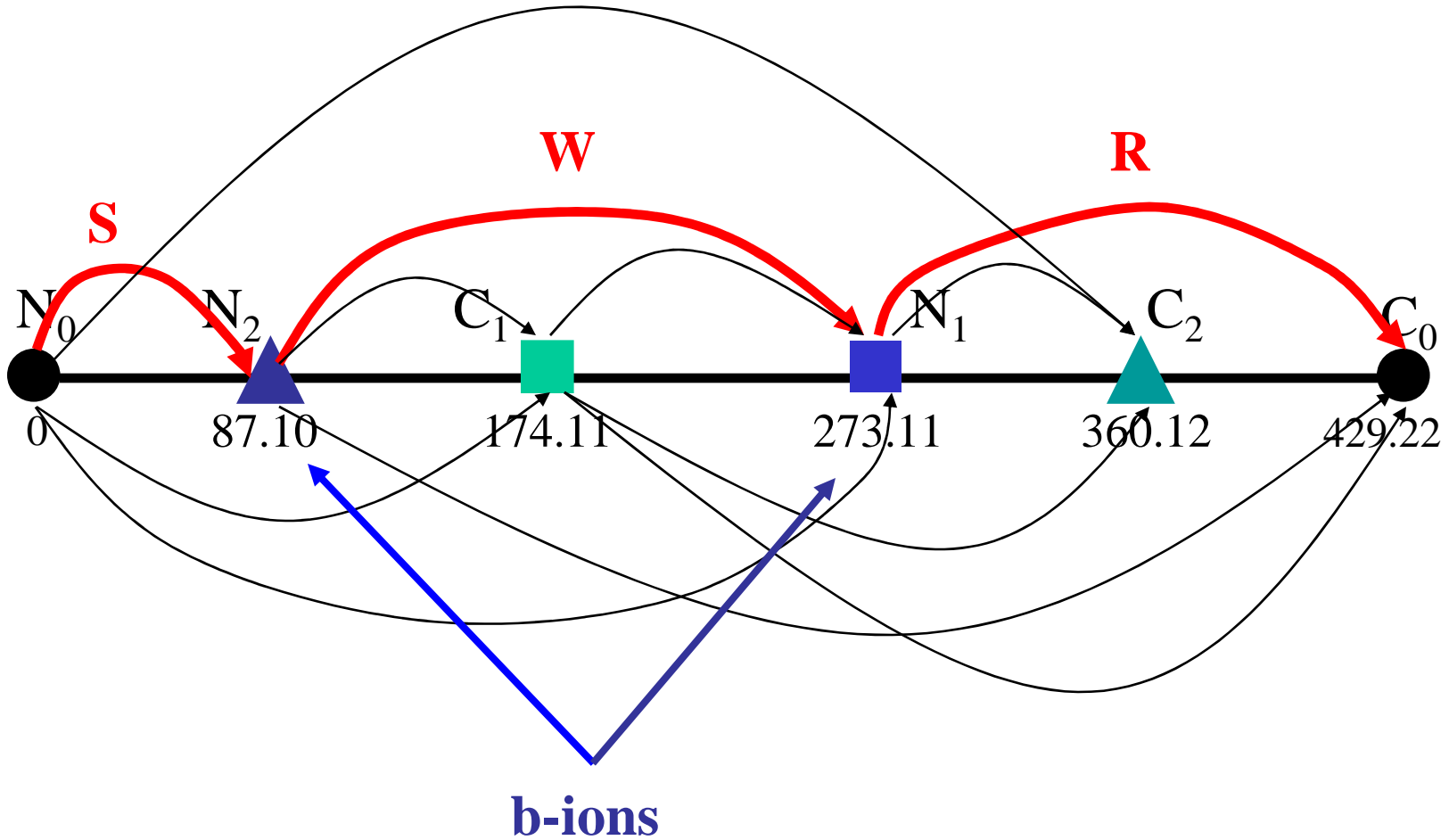
NC-Spectrum Graph: Edges (4)



NC-Spectrum Graph



NC-Spectrum Graph: Paths = Sequences



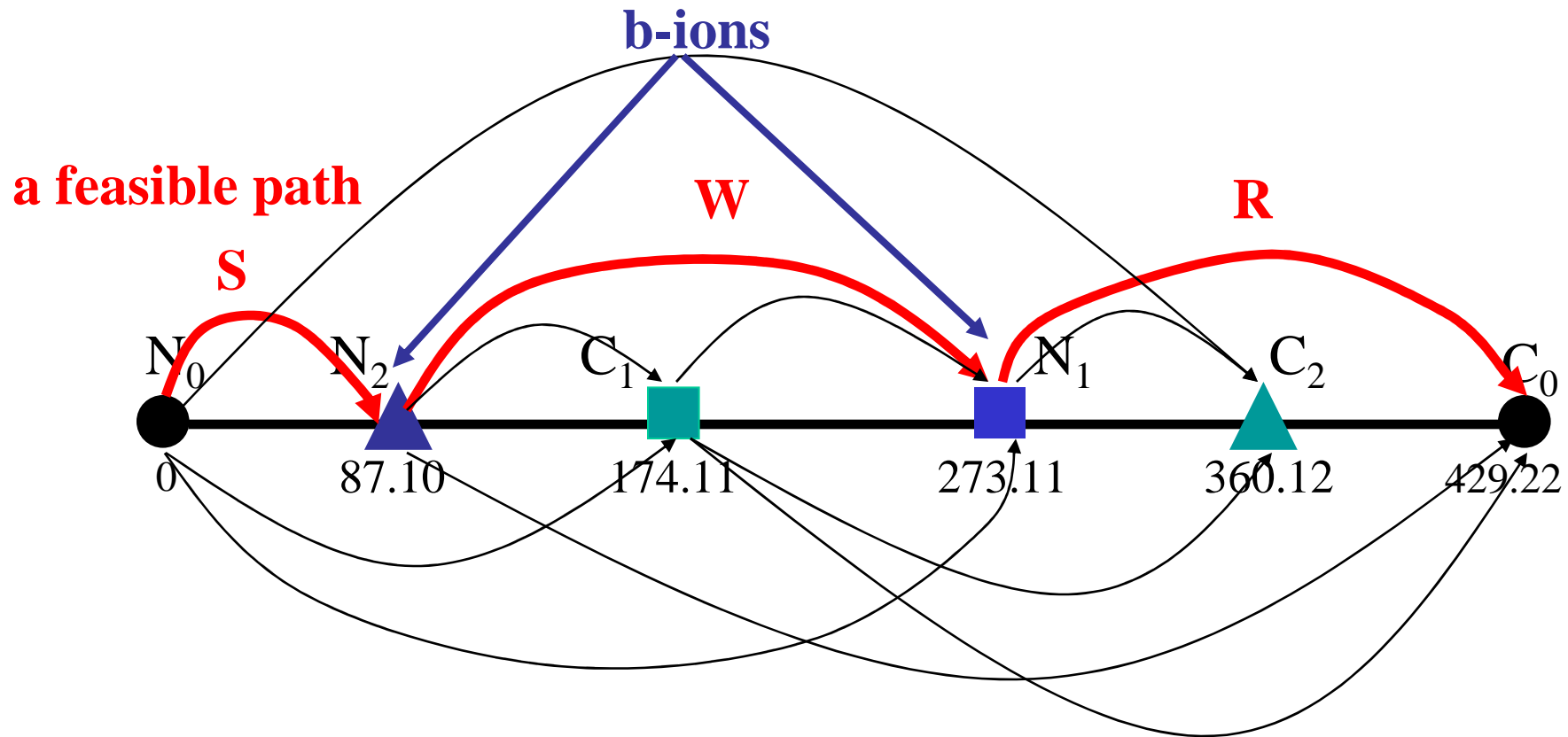
Peptide Sequencing: an Optimization Problem

Input: an NC-spectrum graph G .

Output: a longest path from N_0 to C_0 .

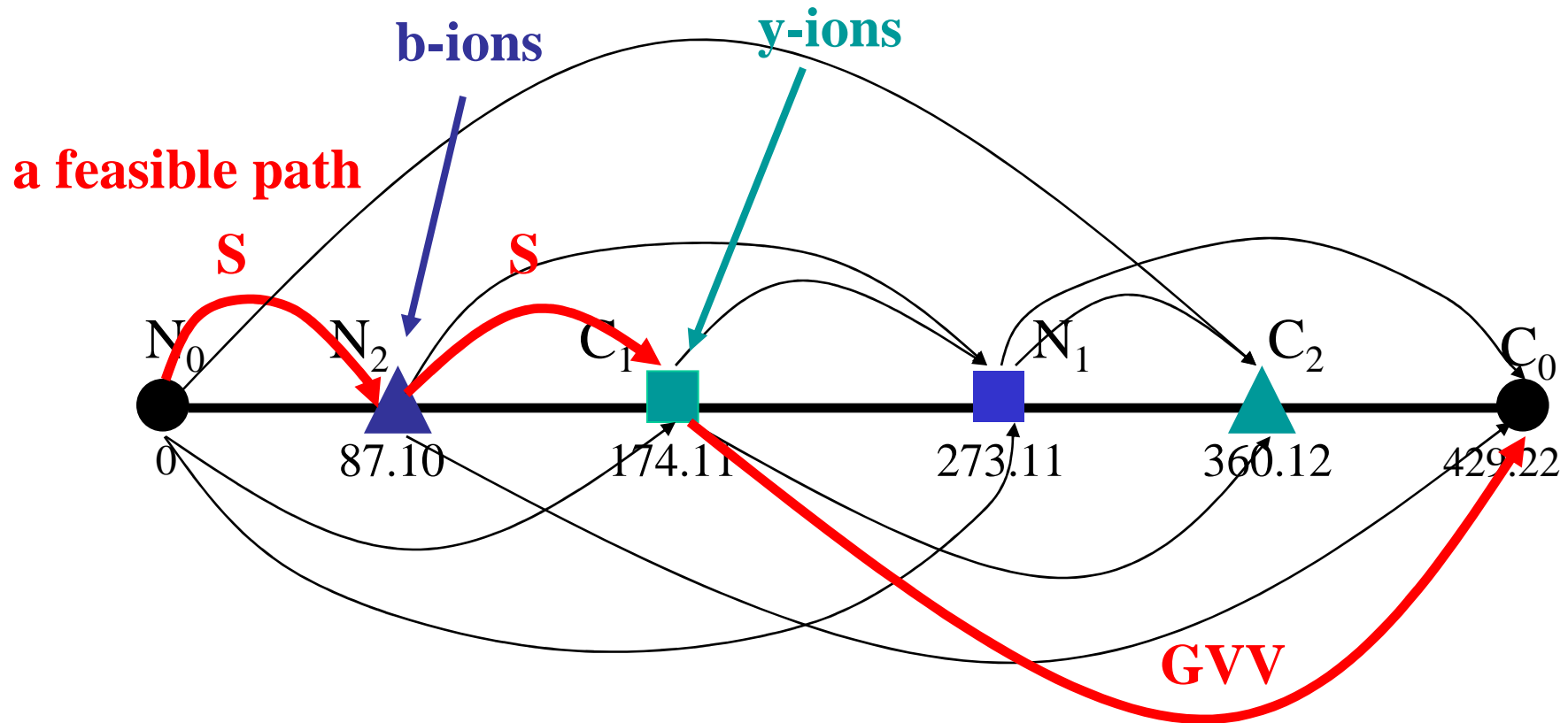
Wrong Formulation!

NC-Spectrum Graph: A Feasible Path (1)



Definition: A **feasible path** is a path from N_0 to C_0 that goes through exactly one node for each pair (either N_j or C_j).

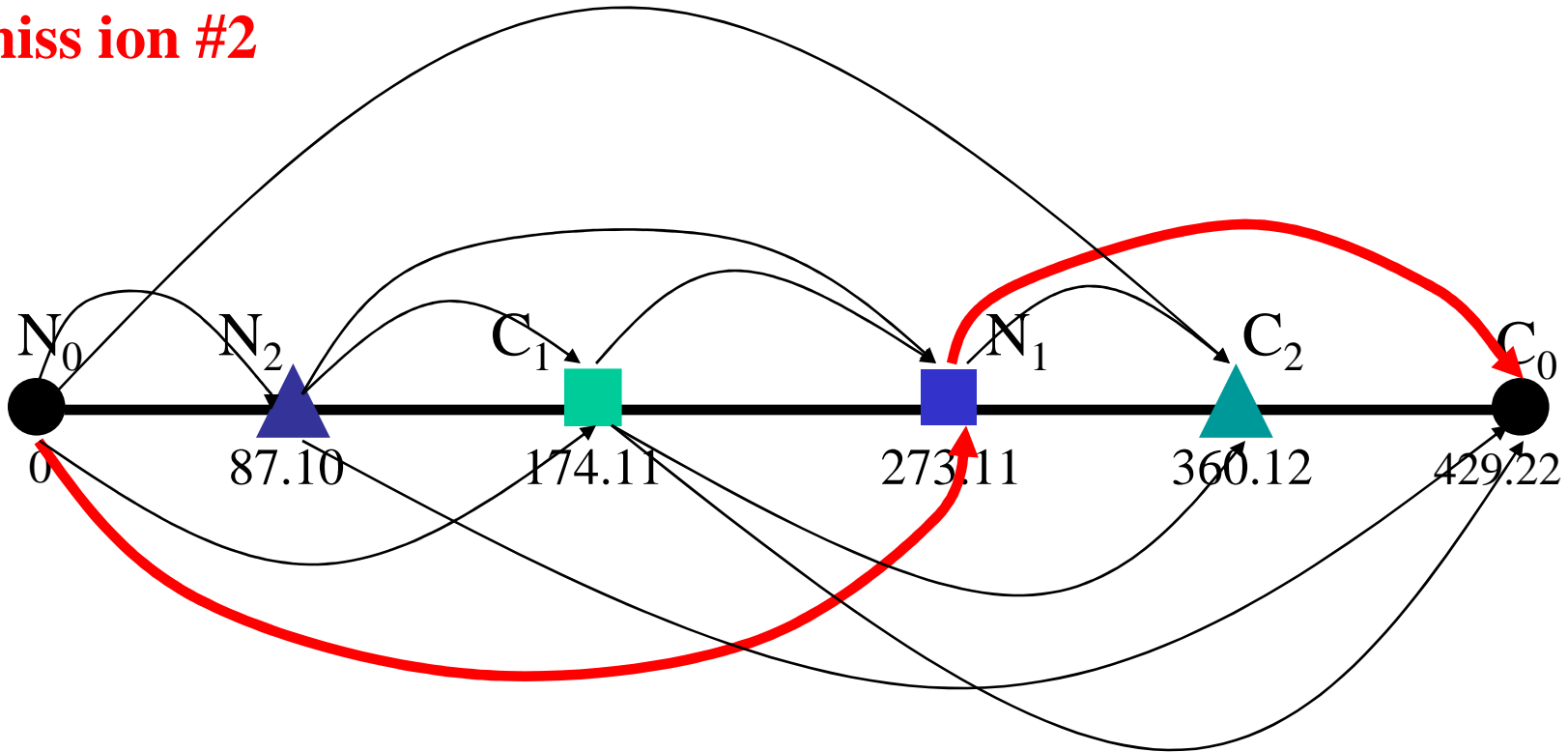
NC-Spectrum Graph: A Feasible Path (2)



Definition: A **feasible path** is a path from N_0 to C_0 that goes through exactly one node for each pair (either N_j or C_j).

NC-Spectrum Graph: Not A Feasible Path (1)

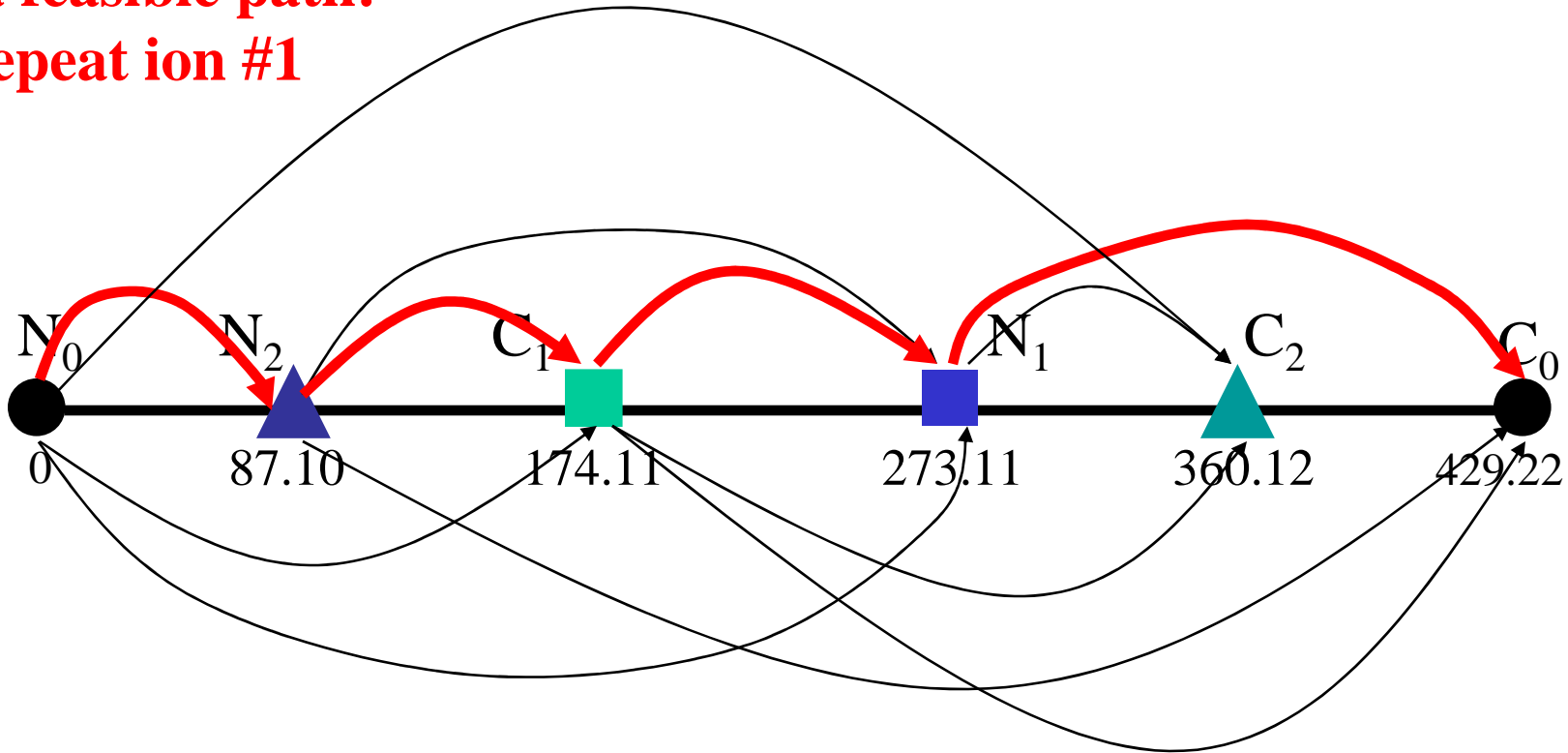
not a feasible path:
(1) miss ion #2



Definition: A **feasible path** is a path from N_0 to C_0 that goes through exactly one node for each pair (either N_j or C_j).

NC-Spectrum Graph: Not A Feasible Path (2)

not a feasible path:
(2) repeat ion #1



Definition: A **feasible path** is a path from N_0 to C_0 that goes through exactly one node for each pair (either N_j or C_j).

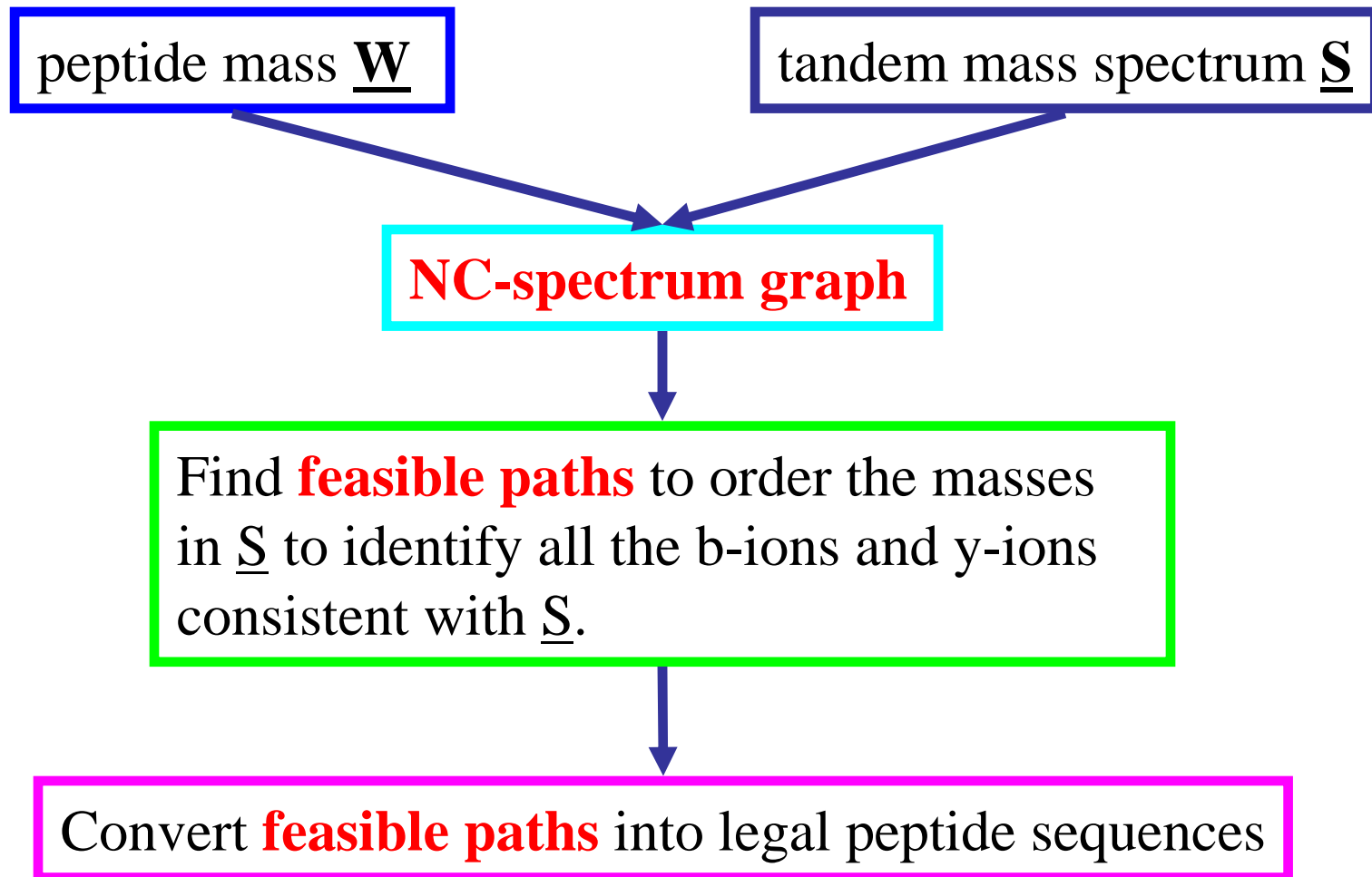
Peptide Sequencing: a Reconstruction Problem

Input: an NC-spectrum graph G .

Output: a feasible path from N_0 to C_0 .

Correct Formulation!

Basic Computing Scheme



Algorithmic Result: De Novo Peptide Sequencing

Input: an NC-spectrum graph $G=(V,E)$

Output: a feasible path (or all feasible paths)

Time Complexity: $O(|V|+|E|)$ time

Ideas: Dynamic Programming + Pre-processing

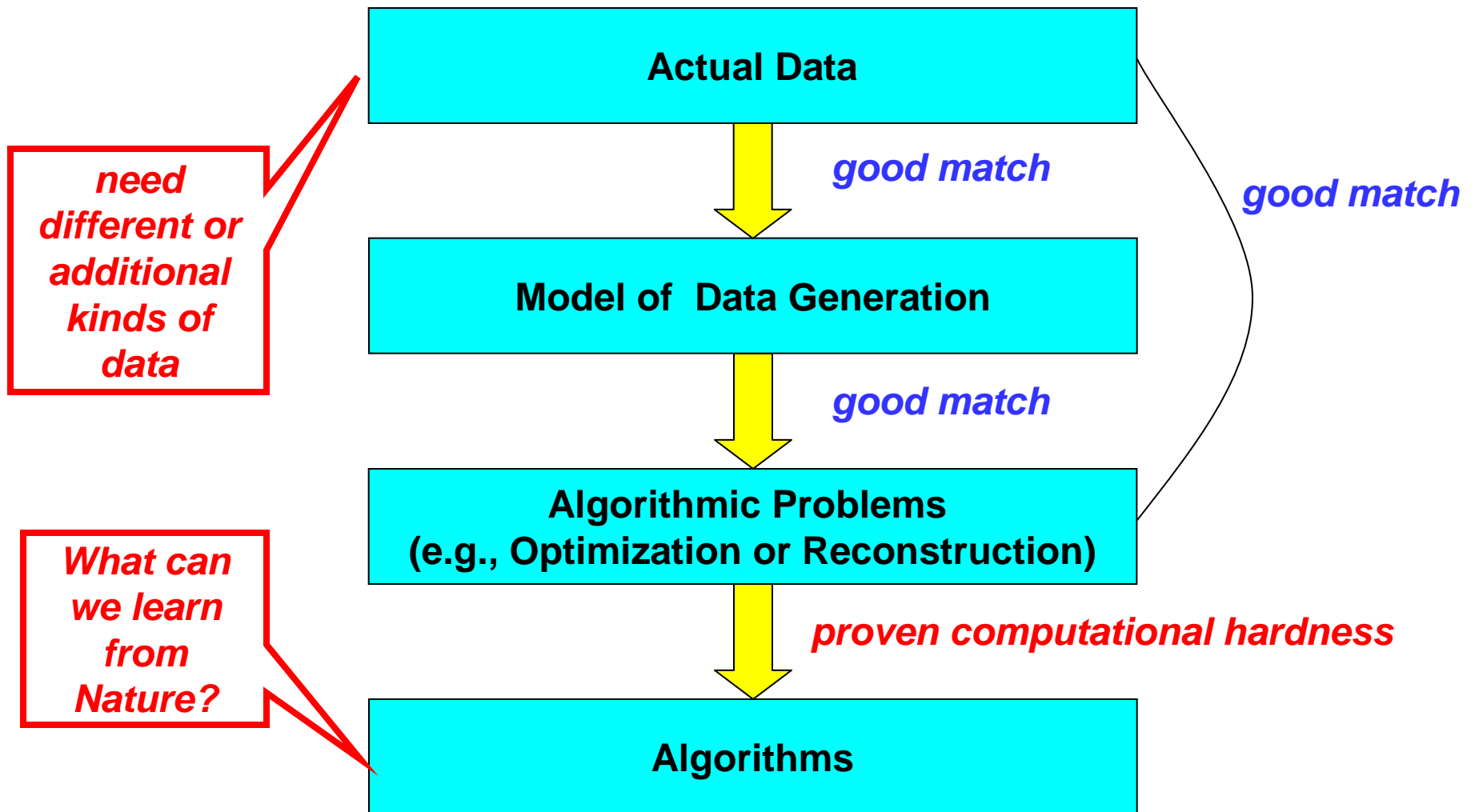
Generalization: The problem formulation and the algorithm can be generalized to handle various errors and modifications.

What Can Bioinformatics Do for Computer Science?

Bioinformatics has been generating many computationally interesting algorithmic problems.

Going beyond this ...

Case #2



What Can Bioinformatics Do for Computer Science?

- **Example: Protein Folding**
- **Biological Fact: Nature can fold a protein very fast.**
- **Current Computer Science Belief: Protein Folding is NP-hard**
- **Possible Consequences:**
 - 1. Our problem formulation is incorrect (e.g., incomplete modeling),**

or

 - 2. By imitating nature, we will find a much more powerful computational paradigm than silicon-based computing.**

THE END

THANK YOU!