SeqBIM

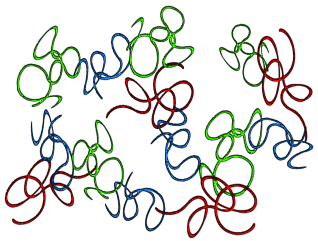
An integrative model for the protein inference problem

Emile Benoist[1], Guillaume Fertin[1], Géraldine Jean[1], Dominique Tessier[2]

1.Nantes Université, LS2N (UMR 6004), Nantes, France
2.Inrae, BIA (UR 1268), Nantes, France
{emile.benoist,geraldine.jean,guillaume.fertin}@univ-nantes.fr,
dominique.tessier@inrae.fr

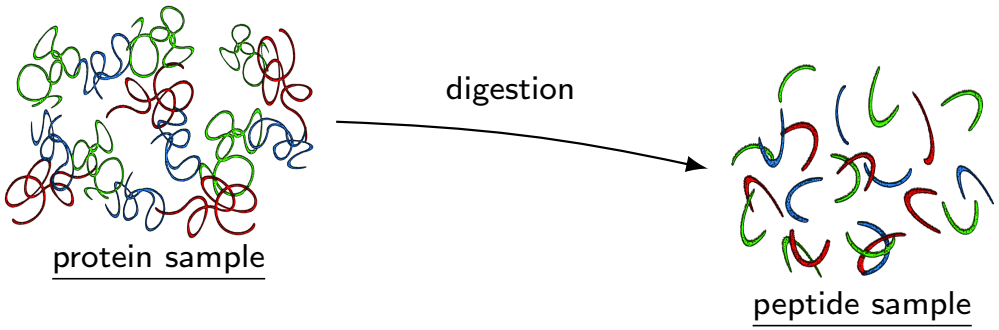
Thursday, November 17

The protein inference problem



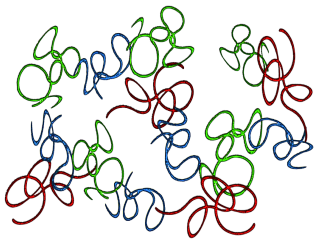
protein sample

The protein inference problem



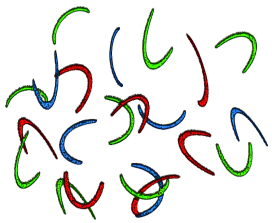
The protein inference problem

*https://fr.wikipedia.org/wiki/Spectrometrie_de_masse



protein sample

digestion



peptide sample

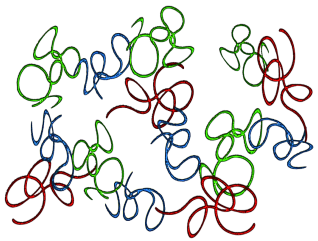
input

mass spectrometer*

output

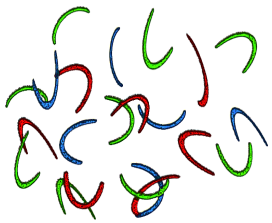
set of MS2 spectra (S)

The protein inference problem



protein sample

digestion



peptide sample

input

mass spectrometer

output

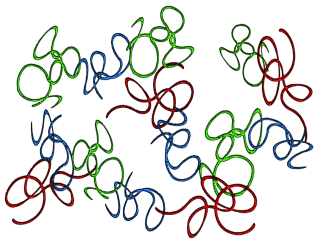
set of MS2 spectra (S)

protein database

protein loading

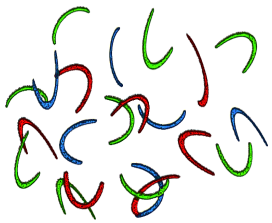
set of proteins (P)

The protein inference problem



protein sample

digestion



peptide sample

input

mass spectrometer

output

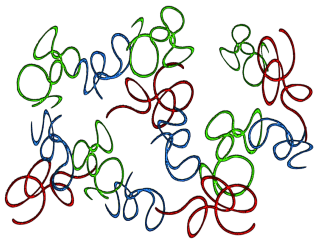
set of MS2 spectra (S)

in-silico digestion

set of peptides (**P**)

set of proteins (P)

The protein inference problem



protein sample

digestion

peptide sample

input

mass spectrometer

output

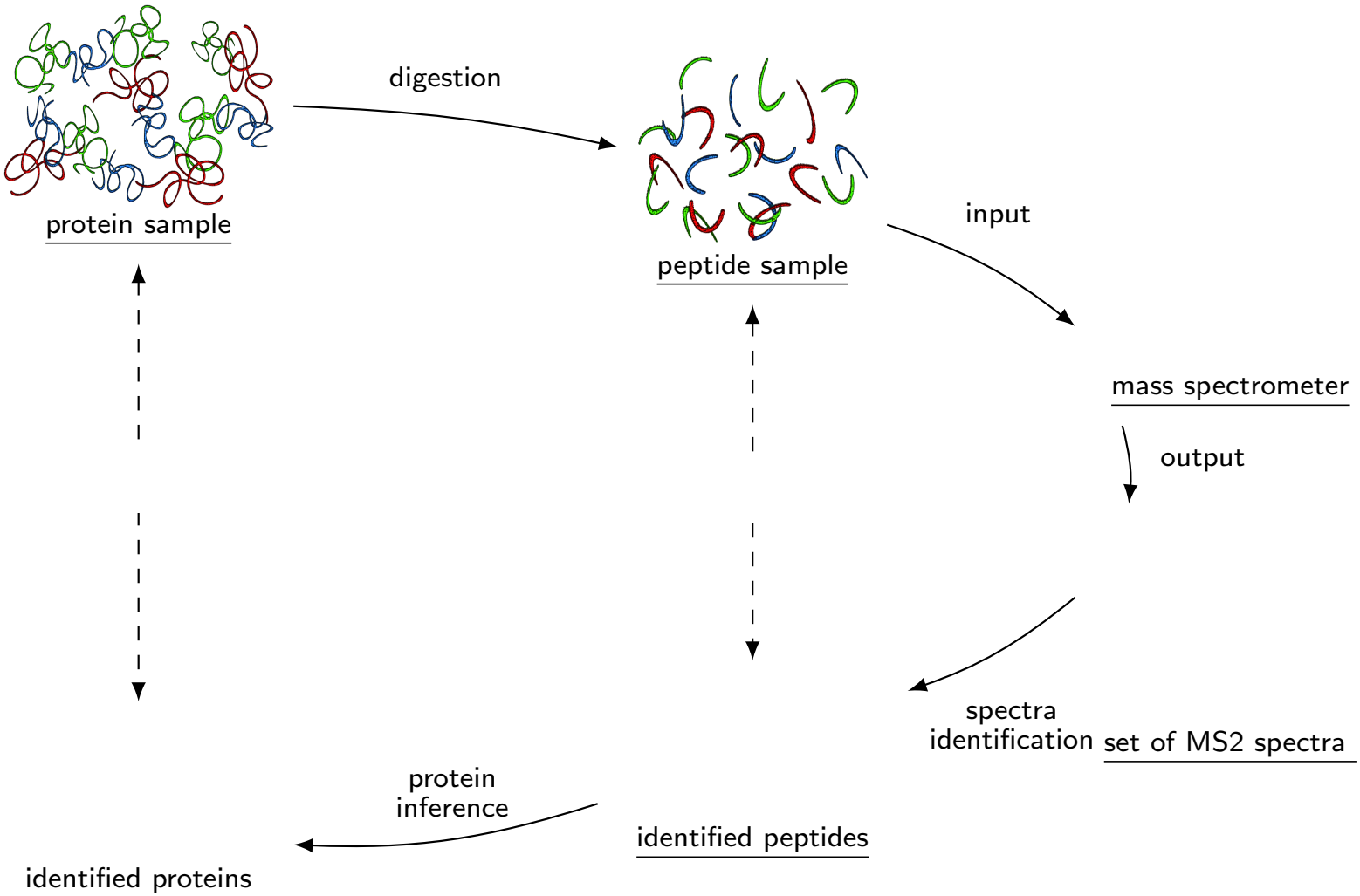
set of MS2 spectra

spectra identification

identified peptides

set of proteins (P)

The protein inference problem



Existing models

Parsimonious models (Occam's razor principle):

Infers the minimum number of proteins that can "explain" all identified peptides [IDPicker, DBParser, MassSieve, LDFA]

Optimistic models:

Infers all proteins meeting a simple criterion [DTASelect]

Statistical models:

Compute for each candidate protein, the probability that the protein is present in the sample [ProteinProphet, MSBayesPro, ProteinLP]

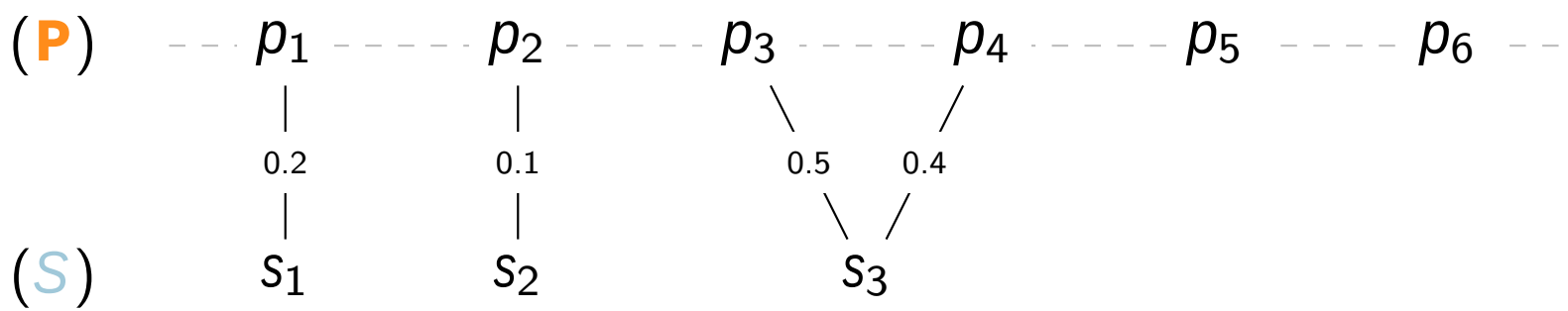
Improvement idea

Each protein inference model relies on a set of identified peptides, themselves identified from all the MS2 spectra independently of the proteins of the database

Improvement idea

Each protein inference model relies on a set of identified peptides, themselves identified from all the MS2 spectra independently of the proteins of the database

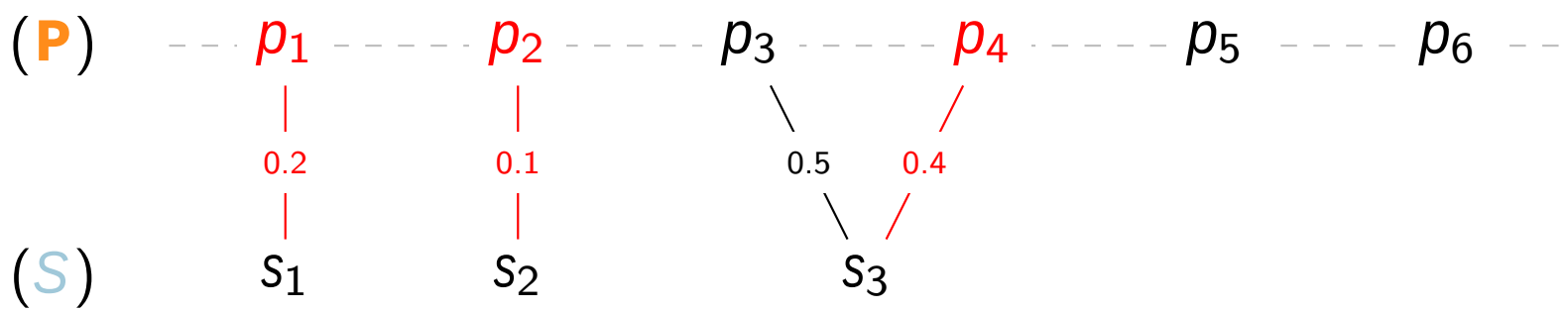
The origin of the peptides (proteins from the database) is an interesting source of information :



Improvement idea

Each protein inference model relies on a set of identified peptides, themselves identified from all the MS2 spectra independently of the proteins of the database

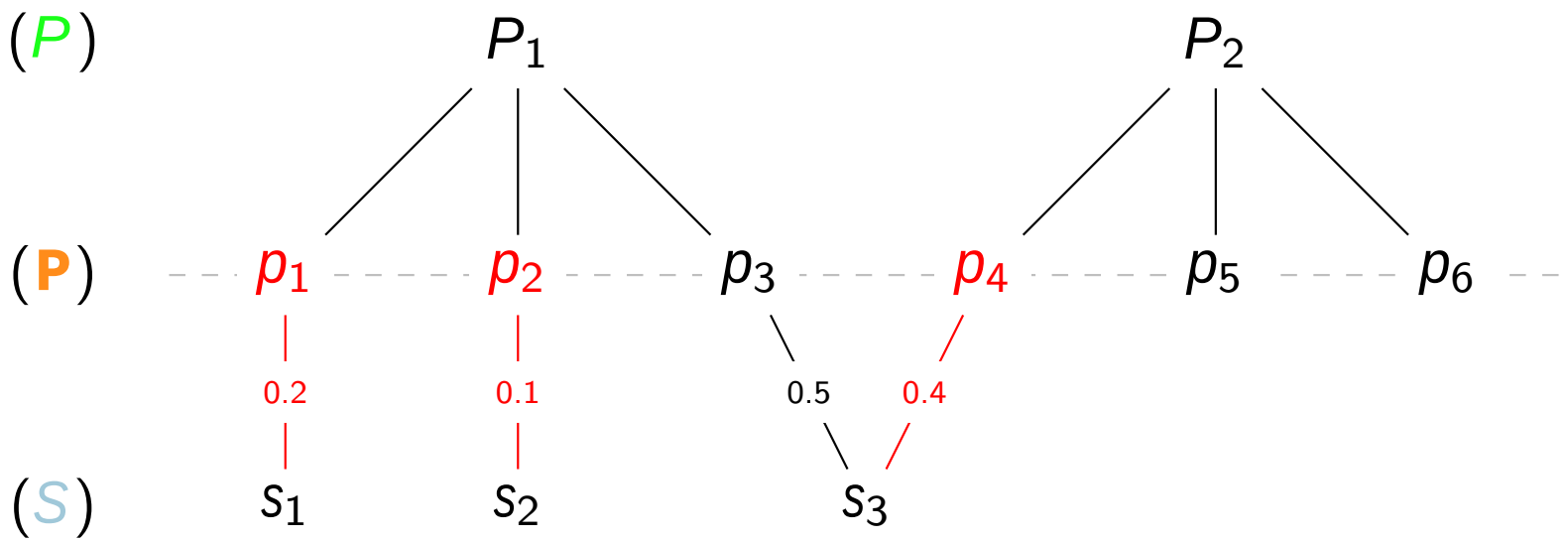
The origin of the peptides (proteins from the database) is an interesting source of information :



Improvement idea

Each protein inference model relies on a set of identified peptides, themselves identified from all the MS2 spectra independently of the proteins of the database

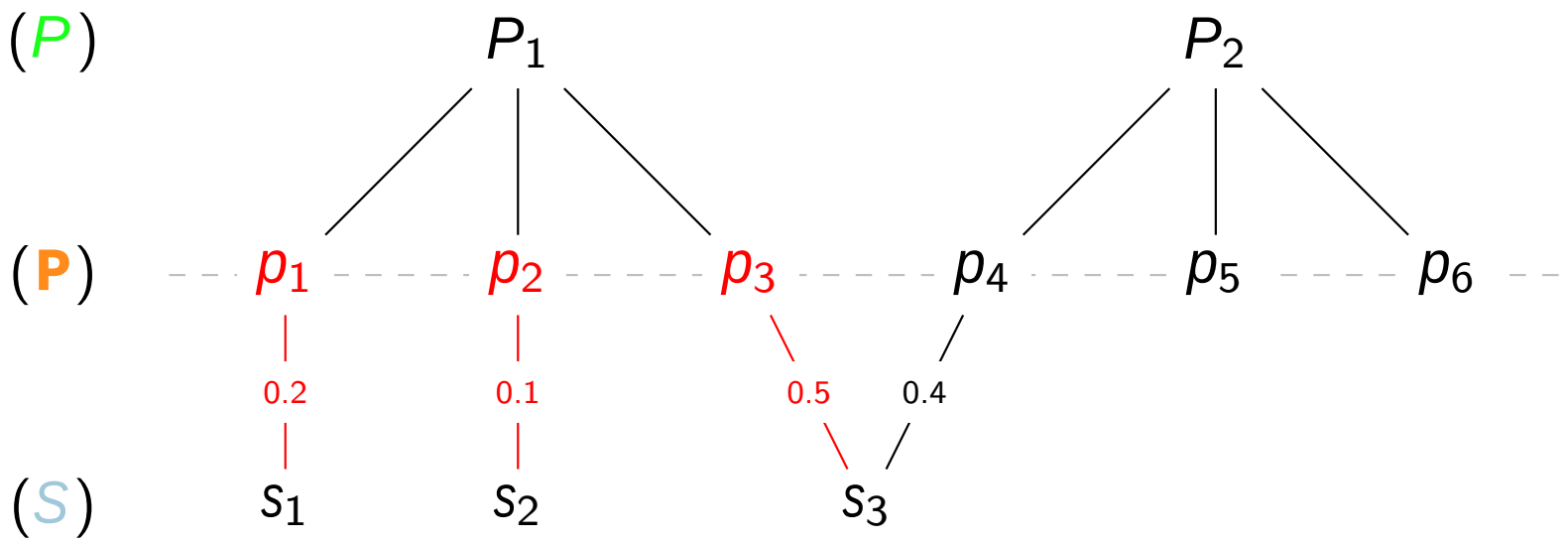
The origin of the peptides (proteins from the database) is an interesting source of information :



Improvement idea

Each protein inference model relies on a set of identified peptides, themselves identified from all the MS2 spectra independently of the proteins of the database

The origin of the peptides (proteins from the database) is an interesting source of information :

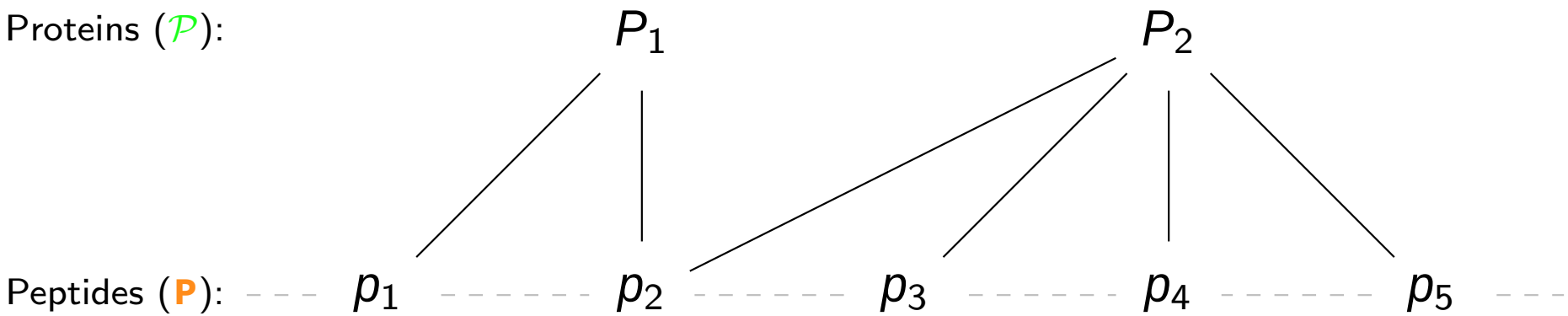


The Global Spectra Interpretation (GSI)

Proteins ($\mathcal{P}$): P_1 P_2

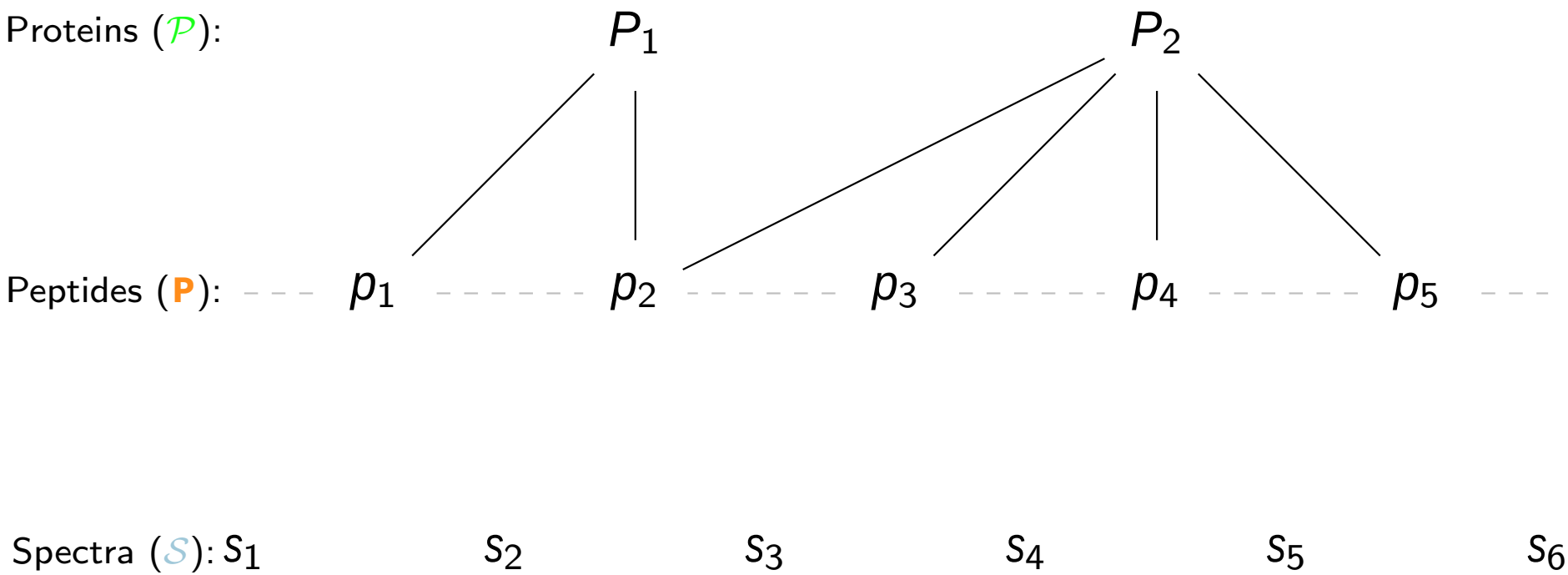
Proteins loaded from a database

The Global Spectra Interpretation (GSI)



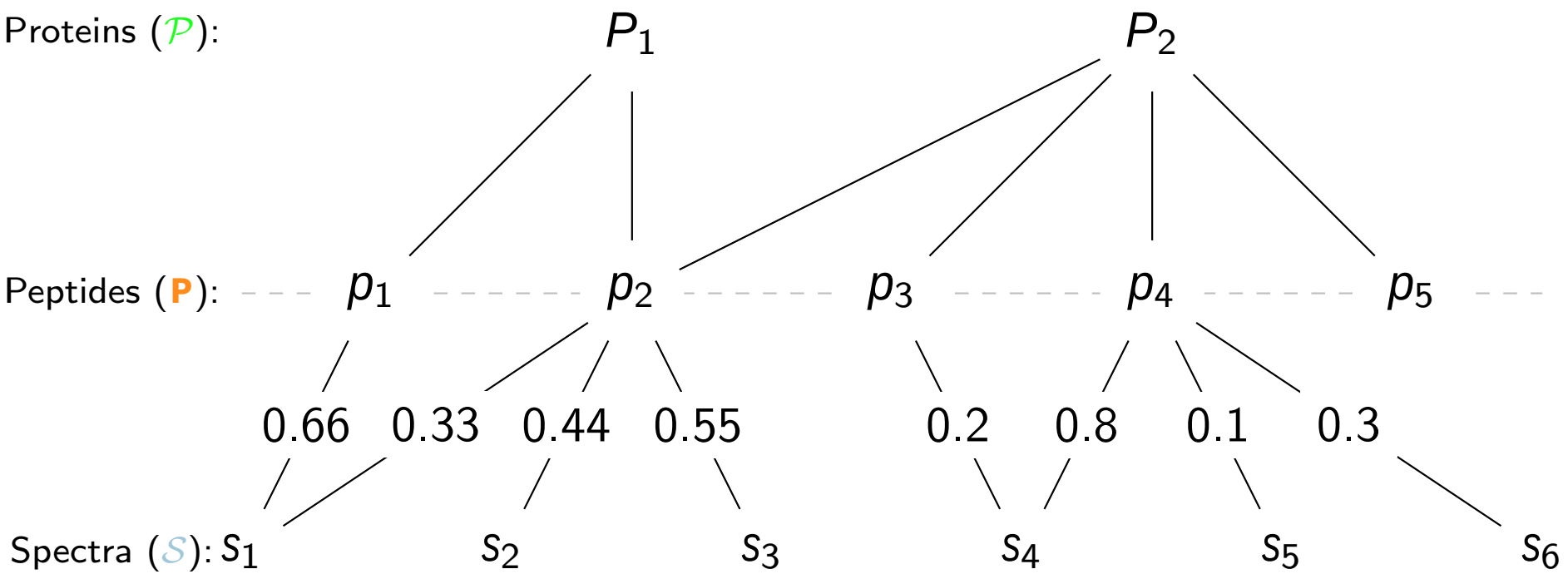
In-silico digestion (trypsine) of each protein

The Global Spectra Interpretation (GSI)



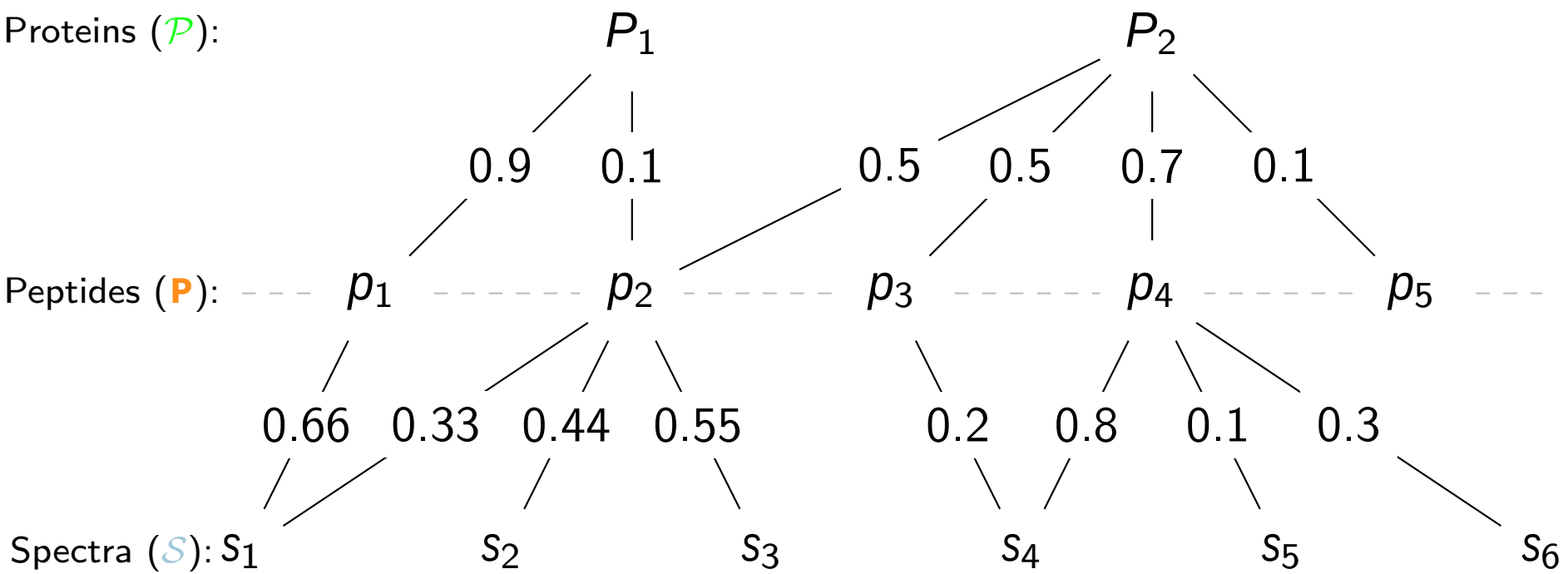
The set of MS2 spectra obtained by mass spectrometry

The Global Spectra Interpretation (GSI)



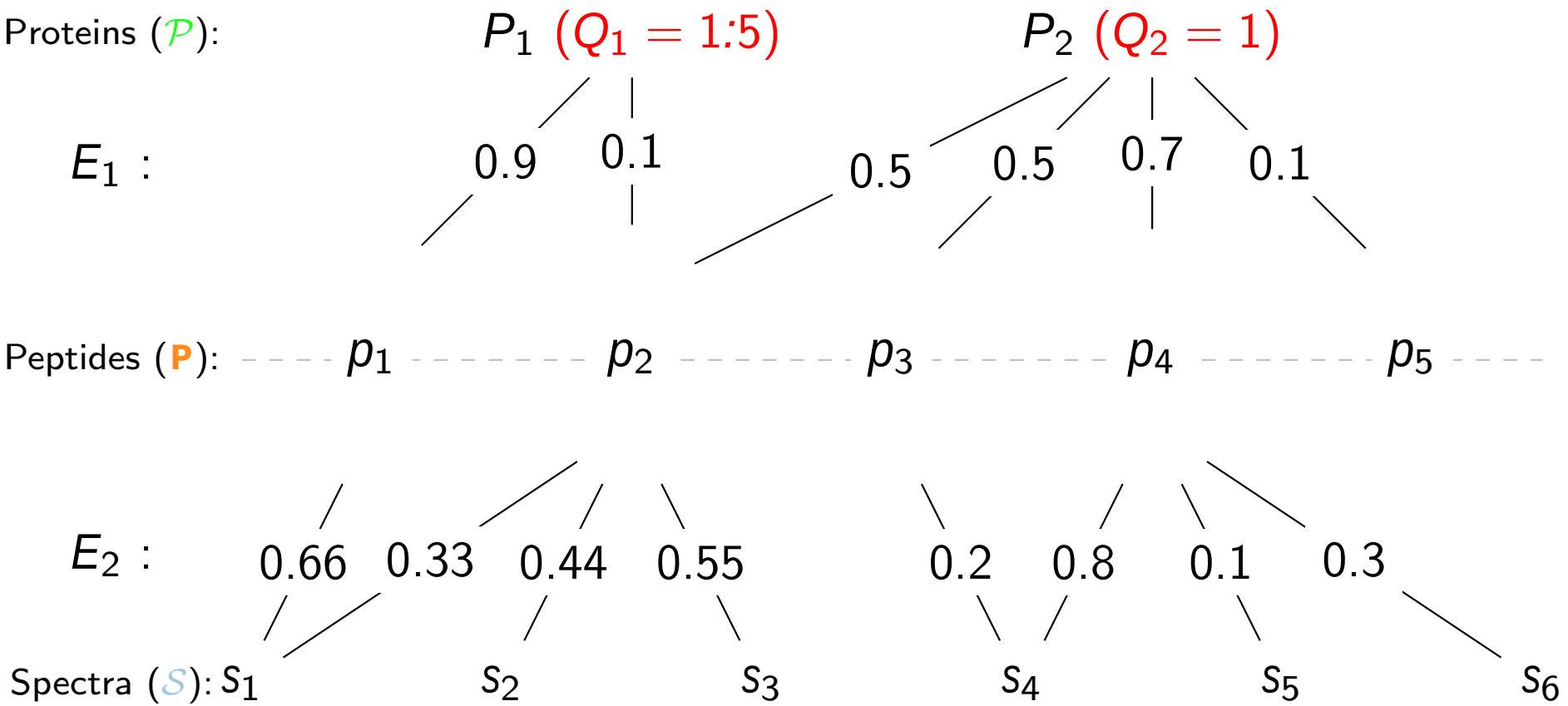
A set of scores (PSM) provide by software comparing MS2 spectra and peptides

The Global Spectra Interpretation (GSI)



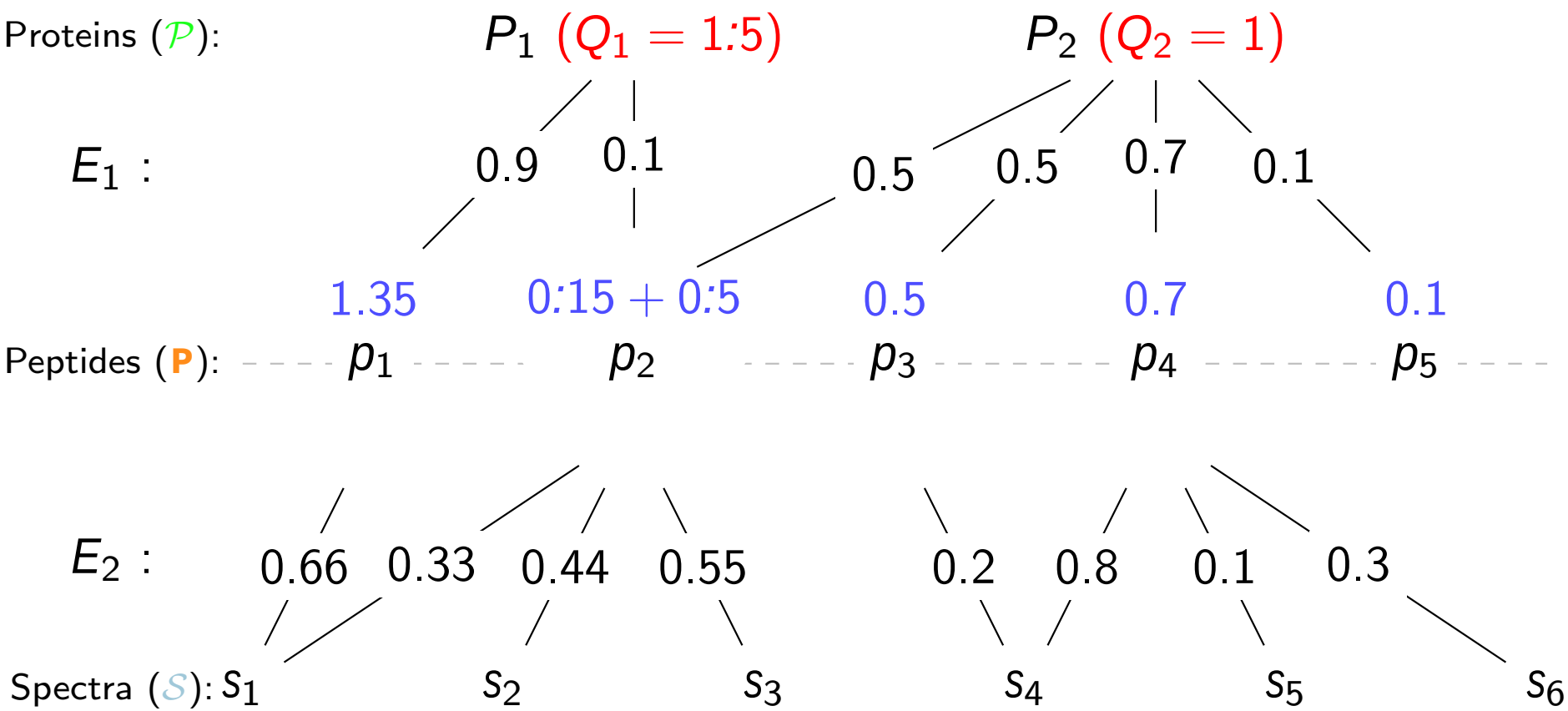
$\mathrm{P}(P_i; p_j)$: the probability that a spectrum corresponding to p_j exists in S given that P_i has en abundance of 1 in the sample

Solution example



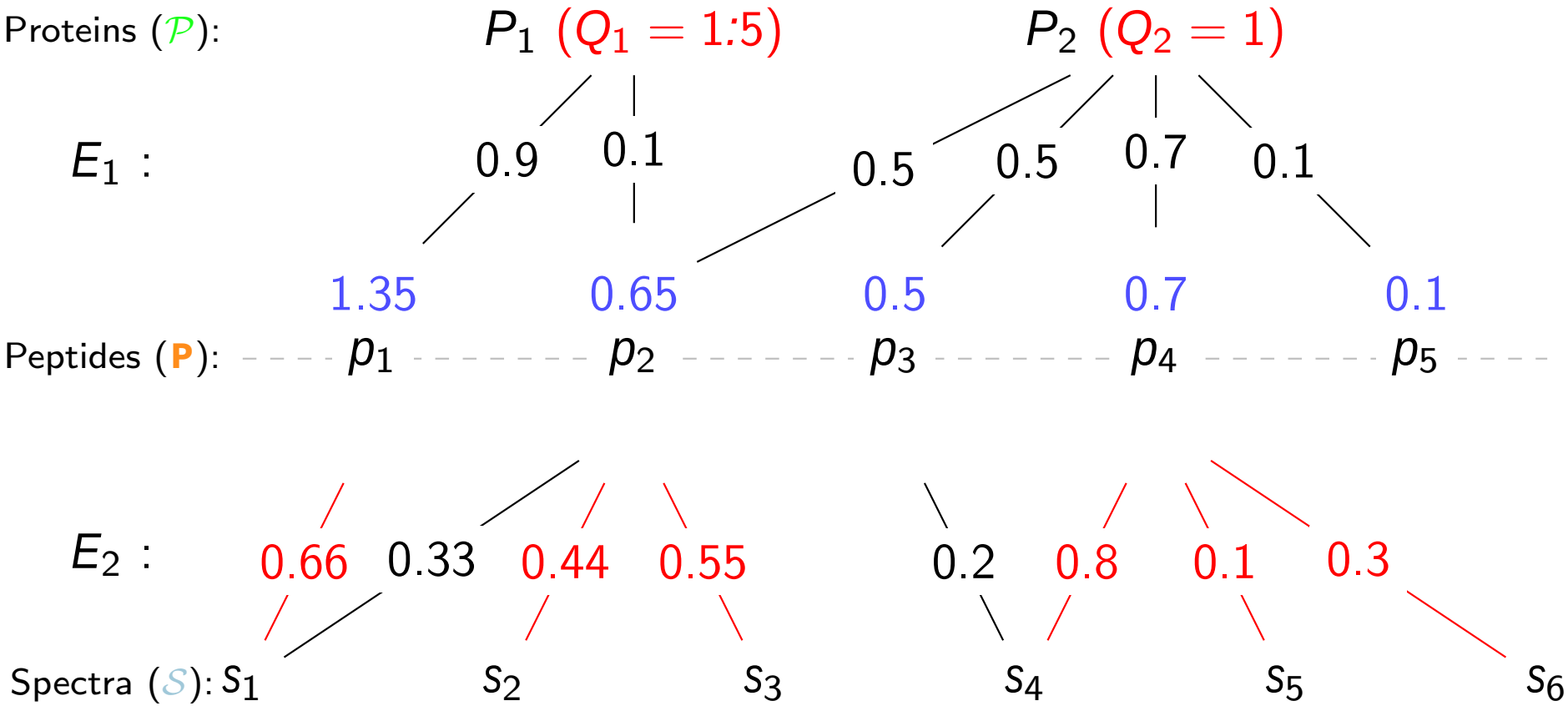
A quantity (or abundance) for each protein in P. It represents the assumed abundance of the protein in the sample

Solution example



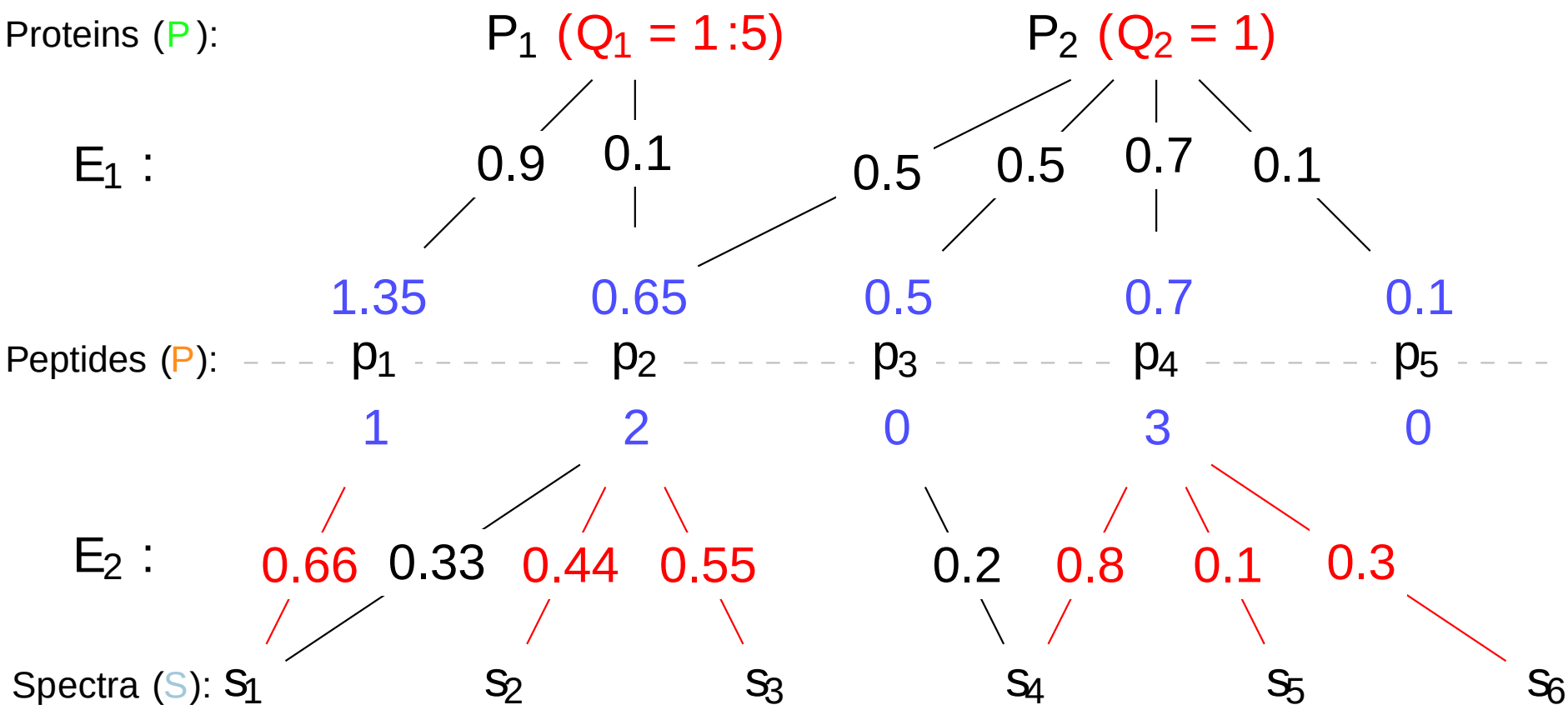
For each peptide p_j in **P** : $\overline{q_j} = \sum_{(P_i;p_j)} (\mathrm{P}(P_i;p_j) \quad Q_i)$

Solution example



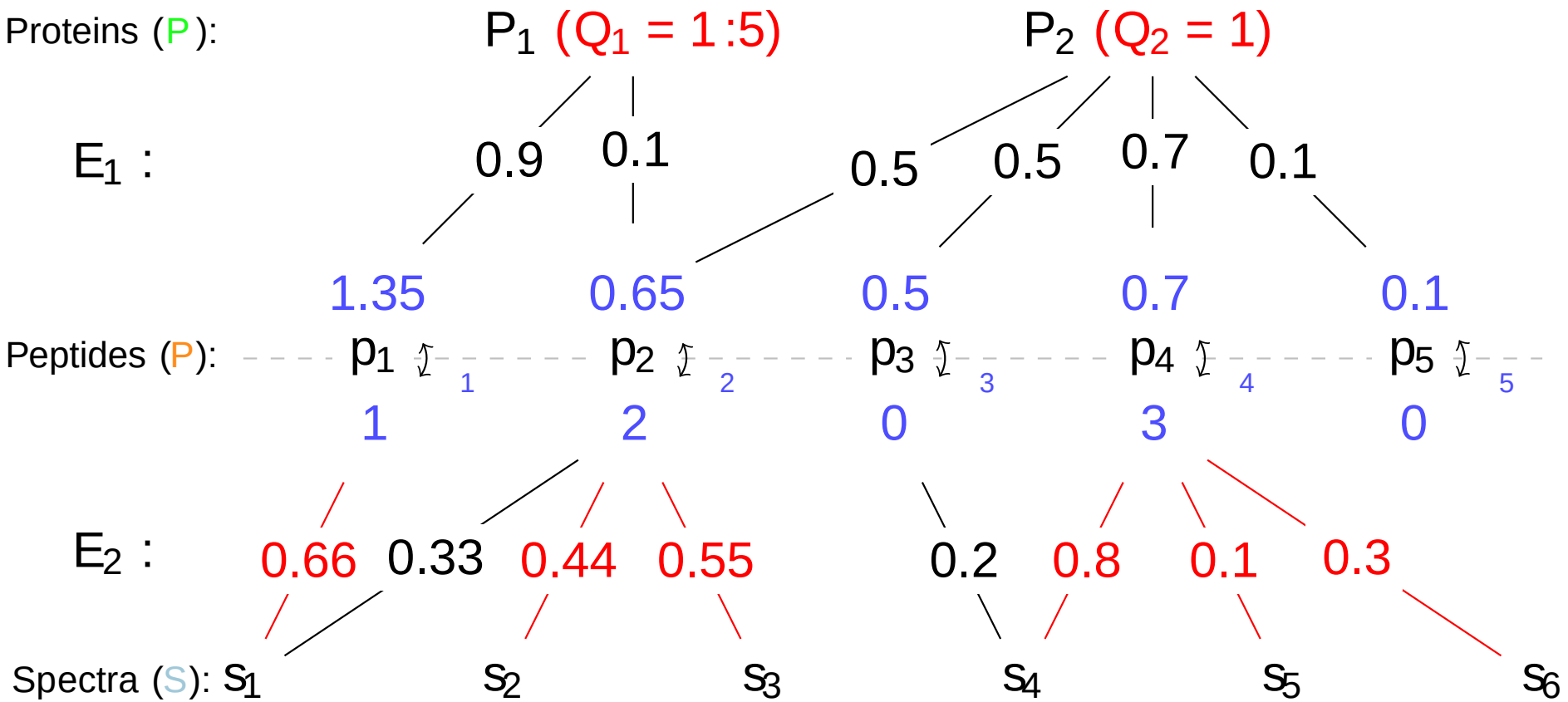
One PSM for each spectrum in S

Solution example



For each peptide p_j in P : q_j = number of selected (red) edges adjacent to p_j

Solution example



$_j$ = j$\overline{q_j}$ $\underline{q_j}$j ($_1$ = j1:35 1j = 0:35)

Objective function

Minimize the di erence ( ) between the number of times a peptide is identi ed and the number of times it is expected to be identi ed given the quantities of the proteins

Objective function

Minimize the di erence ( ) between the number of times a peptide is identi ed and the number of times it is expected to be identi ed given the quantities of the proteins

Minimize the score on the chosen edges (red edges)

Objective function

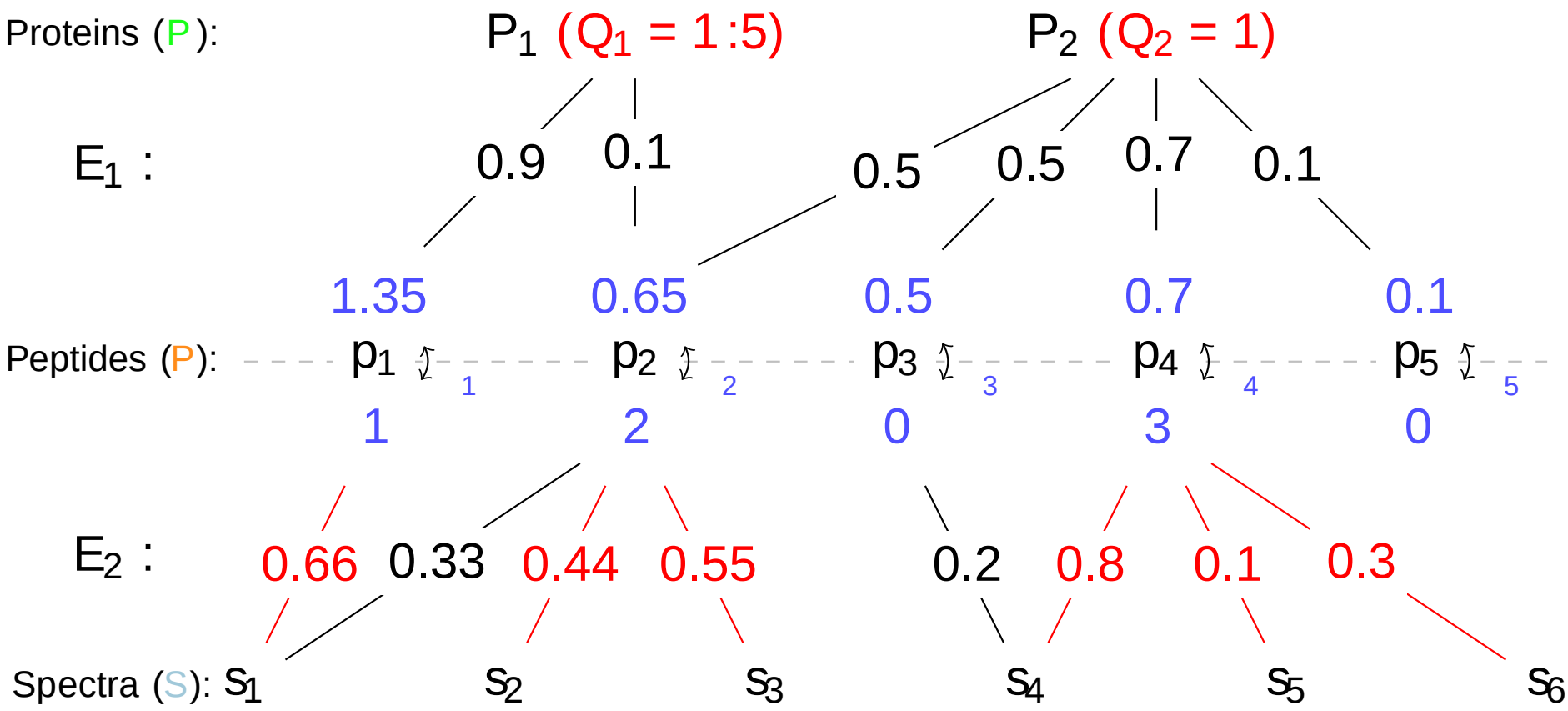
Minimize the di erence ( ) between the number of times a peptide is identi ed and the number of times it is expected to be identi ed given the quantities of the proteins

Minimize the score on the chosen edges (red edges)

$$\min z = \quad {}_1 \sum_{i=1}^{m} \quad {}_i + \quad {}_2 \sum_{k=1}^{\grave{}} \quad S(s_k)$$

m: number of peptide nodes, ` : number of spectrum nodes, : the weighting factors

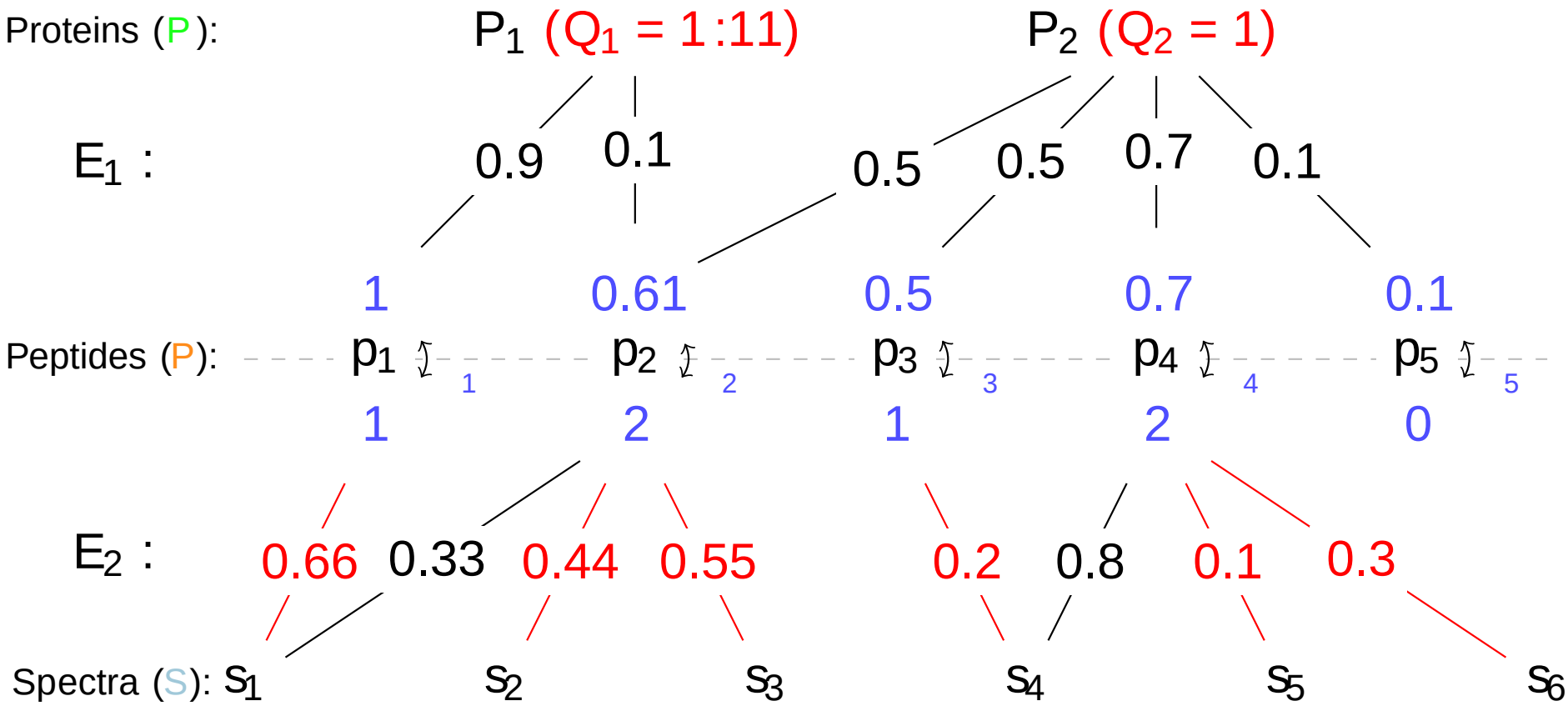
Objective function



$$z = {}_1\,0{:}35 + 1{:}35 + 0{:}5 + 2{:}3 + 0{:}1 + {}_2\,0{:}66 + 0{:}44 + 0{:}55 + 0{:}8 + 0{:}1 + 0{:}3$$

$$= 4{:}6\;{}_1 + 2{:}85\;{}_2$$

Objective function



$$z = {}_1\,0 + 1:39 + 0:5 + 1:3 + 0:1 + {}_2\,0{:}66 + 0:44 + 0:55 + 0:2 + 0:1 + 0:3$$
$$= 3:29\ {}_1 + 2:25\ {}_2$$

In what way the GSI is an integrative model?

One of our goals is to evolve the model by taking into account other sources of information in order to improve its performance

In what way the GSI is an integrative model?

One of our goals is to evolve the model by taking into account other sources of information in order to improve its performance

The probabilities on the edges between proteins and peptides can carry a lot of information :

- the probability of the peptide to be ionized
- the probability of a missed cleavage around the peptide
- the probability that the peptide will be a ected by modi cations
- : : :

In what way the GSI is an integrative model?

One of our goals is to evolve the model by taking into account other sources of information in order to improve its performance

The probabilities on the edges between proteins and peptides can carry a lot of information :

- the probability of the peptide to be ionized
- the probability of a missed cleavage around the peptide
- the probability that the peptide will be a ected by modi cations
- : : :

For now, this probability is xed to 1 (each peptide has the same probability to be identi ed).

GSI complexity

The Global Spectra Interpretation is an NP-hard problem (reduction from the RXC3 : a restricted version of the Exact-Set-Cover problem) :
under the assumption that P ≠ NP, there is no polynomial algorithm that can solve the GSI problem

GSI complexity

The Global Spectra Interpretation is an NP-hard problem (reduction from the RXC3 : a restricted version of the Exact-Set-Cover problem) :
under the assumption that P ≠ NP, there is no polynomial algorithm that can solve the GSI problem

The GSI is FPT with respect to the number of ambiguous edges (an edge adjacent to an ambiguous spectrum) :
there exist an algorithm that solve the GSI in a time that is exponential only on the number of ambiguous edges

GSI complexity

The Global Spectra Interpretation is an NP-hard problem (reduction from the RXC3 : a restricted version of the Exact-Set-Cover problem) :
under the assumption that P ≠ NP, there is no polynomial algorithm that can solve the GSI problem

The GSI is FPT with respect to the number of ambiguous edges (an edge adjacent to an ambiguous spectrum) :
there exist an algorithm that solve the GSI in a time that is exponential only on the number of ambiguous edges

Moving from theory to practice !

Using the mixed-integer linear programming to solve the GSI problem

Mixed-integer linear programming problem

$$\min z = \alpha_1 \sum_{p_j \in P} \beta_j + \alpha_2 \sum_{s_k \in S} \sum_{(p_j;s_k) \in E_2} S(p_j;s_k)\, x_{j;k}$$

subject to :

C.1 $\quad \beta_j \geq \overline{q_j} - \underline{q_j} \qquad \forall p_j \in P$

C.2 $\quad \beta_j \geq \underline{q_j} - \overline{q_j} \qquad \forall p_j \in P$

C.3 $\quad \sum_{(p_j;s_k) \in E_2} x_{j;k} = 1 \qquad \forall s_k \in S$

$$\overline{q_j} = \sum_{(P_i;p_j) \in E_1} Q_i \, P(P_i;p_j) \qquad \forall p_j \in P$$

$$\underline{q_j} = \sum_{(p_j;s_k) \in E_2} x_{j;k} \qquad \forall p_j \in P$$

Q_i : the quantity assigned to the protein P_i

β_j : the difference between $\overline{q_j}$ and $\underline{q_j}$

$x_{j;k}$: 1 if the edge $(p_j;s_k)$ is selected, 0 otherwise

Explanation of constraints

For each $p_j \in P$, α_j must be equal to $|q_j - \overline{q_j}|$

Absolute value is not a linear function but can be expressed as follows :

C.1 $\alpha_j \geq \overline{q_j} - \underline{q_j} \quad \forall p_j \in P$

C.2 $\alpha_j \geq \underline{q_j} - \overline{q_j} \quad \forall p_j \in P$

$\Leftrightarrow \alpha_j \geq |\underline{q_j} - \overline{q_j}| \quad \forall p_j \in P$

Explanation of constraints

For each $p_j \in P$, δ_j must be equal to $|q_j - \overline{q_j}|$

Absolute value is not a linear function but can be expressed as follows :

C.1 $\quad \delta_j \geq \overline{q_j} - q_j \quad \forall p_j \in P$

C.2 $\quad \delta_j \geq q_j - \overline{q_j} \quad \forall p_j \in P$

$$\min z = \sum_{p_j \in P} \delta_j + \ldots \qquad \Leftrightarrow \qquad \delta_j \geq |q_j - \overline{q_j}| \quad \forall p_j \in P$$

Explanation of constraints

For each $p_j \in P$, ε_j must be equal to $|q_j - \overline{q_j}|$

Absolute value is not a linear function but can be expressed as follows :

$$\text{C.1} \quad \varepsilon_j \geq \overline{q_j} - q_j \quad \forall p_j \in P$$

$$\text{C.2} \quad \varepsilon_j \geq q_j - \overline{q_j} \quad \forall p_j \in P$$

$$\Leftrightarrow \quad \varepsilon_j \geq |q_j - \overline{q_j}| \quad \forall p_j \in P$$

$$\min z = \sum_{p_j \in P} \varepsilon_j + \ldots \quad \Rightarrow \quad \varepsilon_j = |q_j - \overline{q_j}| \quad \forall p_j \in P$$

Each spectrum must be identified to exactly one peptide :

$$\text{C.3} \quad \sum_{(p_j, s_k) \in E_2} x_{p_j, s_k} = 1 \quad \forall s_k \in S$$

Some running time

Fixed parameters :

- the number of proteins : 10375
- the number of peptides : 370759

Variable parameters :

- the number of spectra
- the number of edges per spectrum
- the scores on the edges
- the values of $_1$ and $_2$

Some running time

Fixed parameters :

- the number of proteins : 10375
- the number of peptides : 370759

Variable parameters :

- the number of spectra
- the number of edges per spectrum
- the scores on the edges
- the values of $_1$ and $_2$

A rst series of tests was carried out on an instance containing only one edge per spectrum. Even with a high number of spectra, resolution times do not exceed a few seconds (10 seconds).

Some running time

2 edges per spectrum :

3 edges per spectrum :

The higher the ambiguity (scores close to each other), the higher the execution times

number of proteins : 10735 / number of peptides : 370759 / $_1$ = 0 :5 / $_2$ = 0 :5

Some running time

Scores = 0.5 / 0.5 :

Scores = 0.2 / 0.8 :

The higher $_1$ compared to $_2$, the higher the execution times

number of proteins : 10735 / number of peptides : 370759 / number of edges per spectrum : 2

What about performance

Two questions remain open :

- How well the model is able to correct errors of a spectrum identi cation tool ?

What about performance

Two questions remain open :

- How well the model is able to correct errors of a spectrum identi cation tool ?
- Is the model capable of correctly retrieving proteins from a sample?

What about performance

Two questions remain open :

- How well the model is able to correct errors of a spectrum identi cation tool ?
- Is the model capable of correctly retrieving proteins from a sample?

The next few months will consist of answering these 2 questions by performing tests on real data sets.

Conclusion

Contribution :

- We propose a new model for the protein inference problem that can keep several candidate peptides per spectrum
- We demonstrate the scaling of our model with an exact resolution

Work to do :

- Perform the tests on the performance of the model
- Integration of information through probabilities on edges between proteins and peptides
- Evolve the model by the integration of new source of information

Thank you for your attention