ASCQ_ME : un nouveau moteur d'identification de protéines par PMF à partir de spectres MS bruts

Jean-Charles BOISSON[1], Laetitia JOURDAN[1], El-Ghazali TALBI[1] et Christian ROLANDO[2]

Université des Sciences et Technologies de Lille (France)

[1] LIFL: Laboratoire d'Informatique Fondamentale de Lille UMR CNRS 8022

[2] LCOM: Laboratoire de Chimie Organique et Macromoléculaire UMR CNRS 8009

Parcours

- Master recherche informatique :
  - Spécialité algorithmique et bioinformatique.
  - Stage au sein de l'équipe OPAC (Optimisation Parallèle Coopérative) :
    - Méthodes d'optimisation pour l'identification automatique de protéines.
- Thèse au sein de l'équipe OPAC :
  - Méthodes d'optimisation pour la bioinformatique :
    - Protéomique.

Plan

- L'identification à partir de spectres MS :
  - \ Par Peptide Mass Fingerprinting (PMF).
- Schéma global :
  - Le processus de digestion.
  - La simulation de spectres.
  - Le scoring développé.
- Objectifs des nouvelles approches.
- Conclusions et perspectives.

Exploitation des spectres MS

- Spectre (masse/charge) / intensité.
- Un spectre MS ∫ digestion d'une protéine.
- Un pic ∫ un peptide.
- Extraction pics mono isotopiques
  \ liste de masses.

- Identification avec les bases de données (BD) :
  - Basée sur différent type de scores :
    - Paramétrique ∫ scoring « brut ».
    - Probabiliste ∫ tests d'hypothèses.
    - « Méta » scoring.
  - Indice de confiance (z-score, e-value).

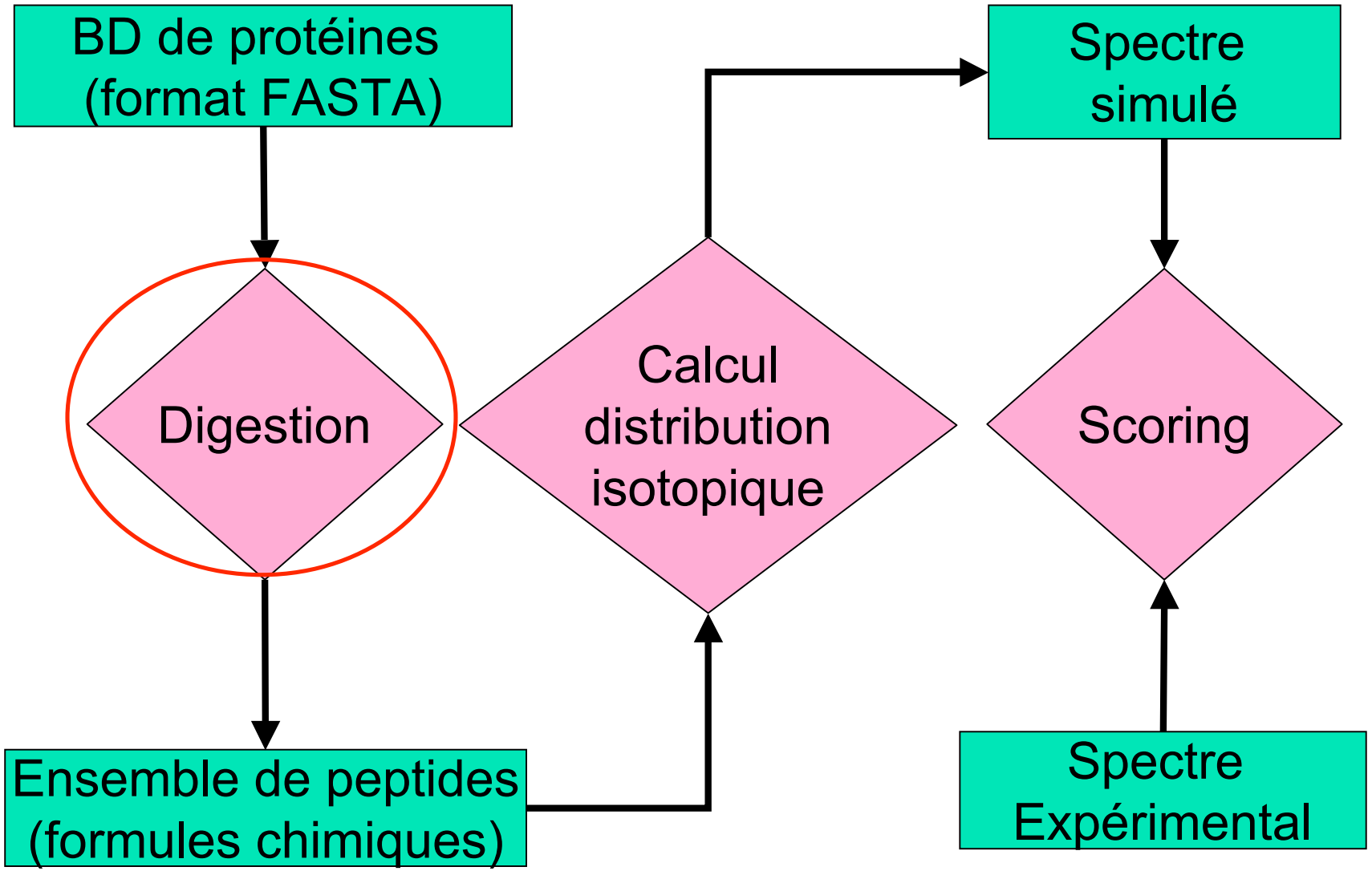
Difficultés : données

- Présence de bruit :
  - Due aux étapes en amont :
    - Extraction des protéines.
    - Séparation des protéines.
- Mauvaise calibration du spectromètre.
- Biais au moment de l'extraction des pics mono isotopiques.

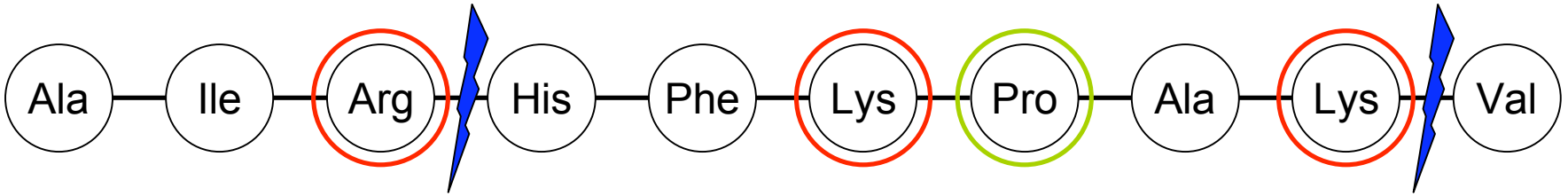
Difficultés : exploitation des données

- Erreurs ou lacunes des BD.
- Coût de la simulation des processus biologiques.
- Suivi de l'évolution des technologies :
  - Quantité de données à traiter.
  - Compatibilité avec les technologies existantes.
  - Ou spécifique à une technologie.
- Trouver de nouvelles approches ?

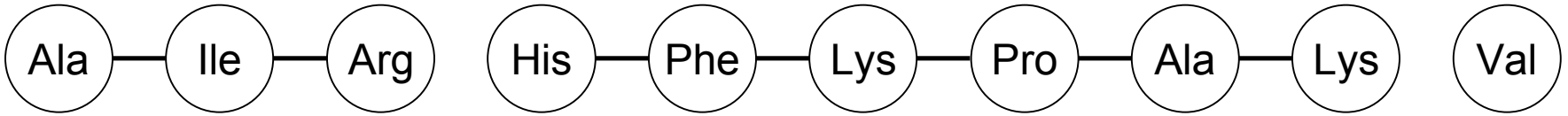
Schéma global



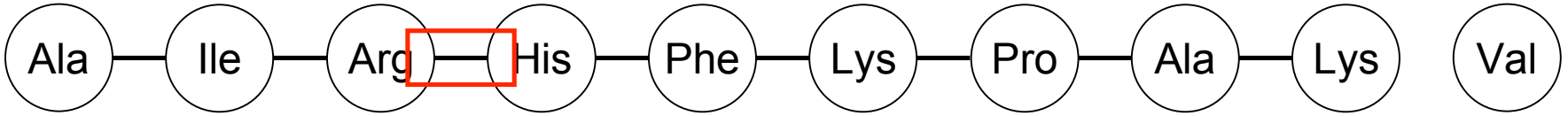
La digestion : principe

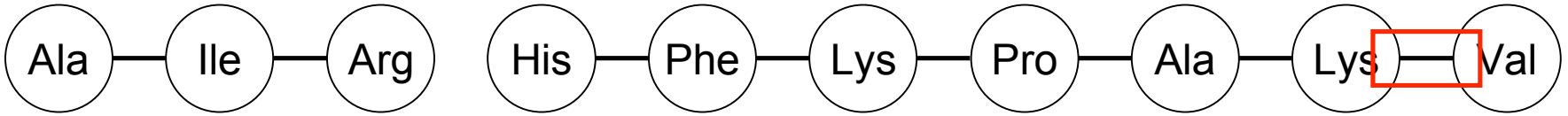


0 miss cleavage :



1 miss cleavage :

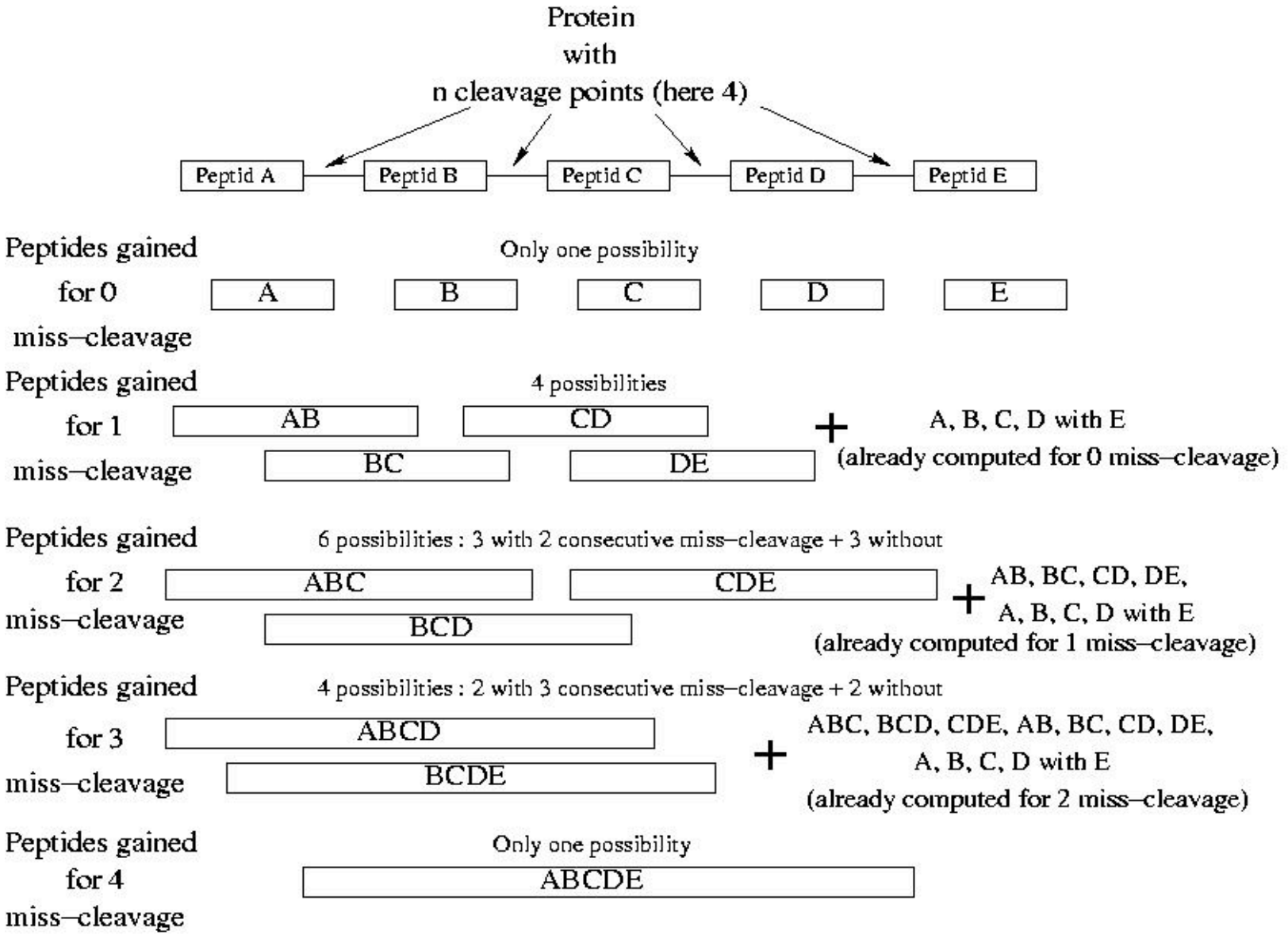




La digestion : simulation

- Mise en place d'un algorithme itératif linéaire :
  - Prise en compte des miss-cleavages.
  - Indépendant de la grammaire de coupure de l'enzyme.
  - Complexité avec *n* le nombre d'acides aminés, *miss* le nombre de miss-cleavages et *m* le nombre de points de coupes :
    - Au pire : $O(n^2)$ (\ cas globalement impossible).
    - En moyenne : O(n + miss * (m+1)) avec miss << m << n.
- Formalisation d'une preuve du parcours complet de l'arbre de digestion :
  - Réponse aux attentes de biologiste.

Parcours complet de l'arbre de digestion



La digestion : difficulté

- Cas des modifications post-traductionnelles :
  - Traitement supplémentaire.
  - Modifications fixes :
    - Linéaire en temps et en espace.
  - Modifications variables :
    - N sites sur un peptide \ 2^N nouveaux peptides.
    - Explosion de manière exponentielle du nombre de peptides.

ADP

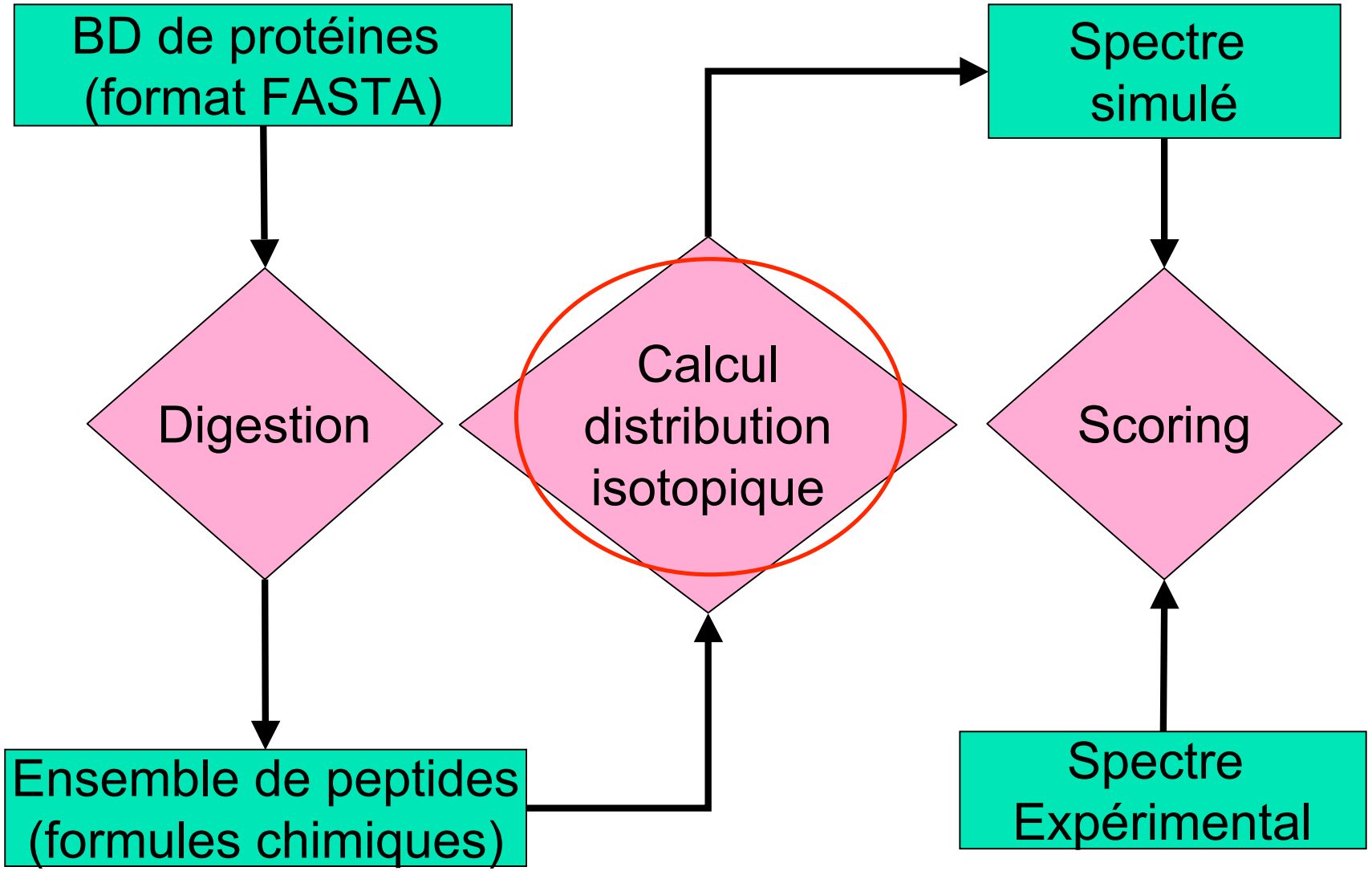
- ADP ∫ Automatic Digestion Process :
  - Digestion sans limites pour les paramètres.
  - Indépendantes des grammaires des enzymes.
  - Modifications post-traductionnelles fixes ou variables.

- Code source libre (licence CeCILL).

Phase de digestion dans ASCQ_ME

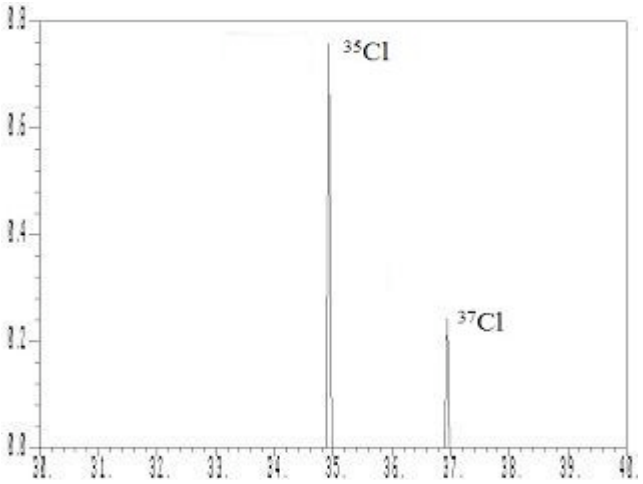
- Héritée d'ADP.
- Version online :
  - Limitation du nombre de miss cleavage (10).
  - Limitation dans la configuration des modifications post-traductionnelles.
- Version offline ⇒ aucune limitation.
- Preuve de la complétude de l'arbre de digestion disponible.

Schéma global

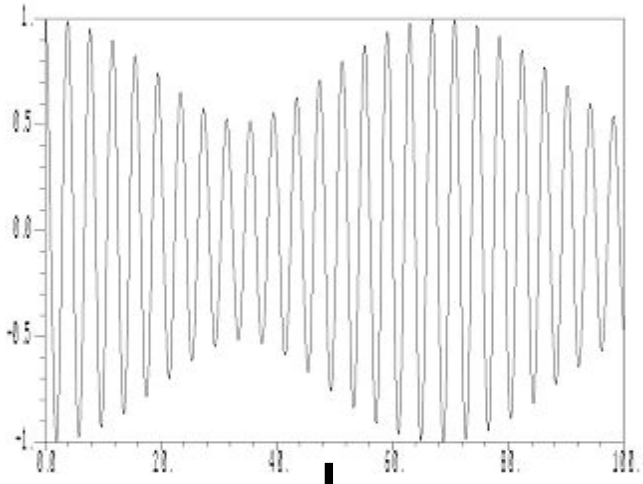


Génération des spectres théoriques (1/2)

- Ex. : la distribution de Cl_2.

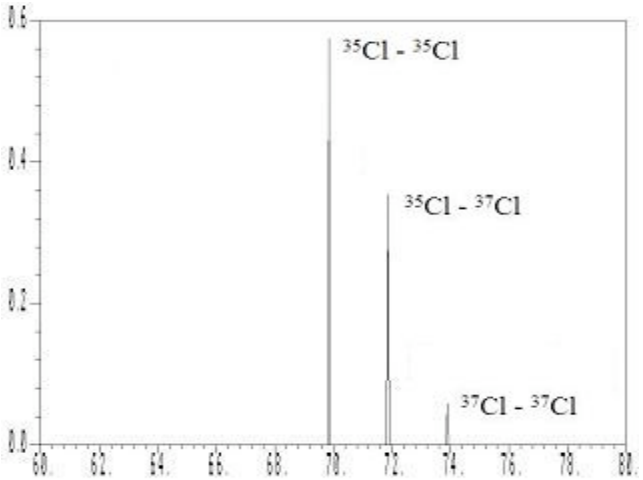


Transformée de Fourier

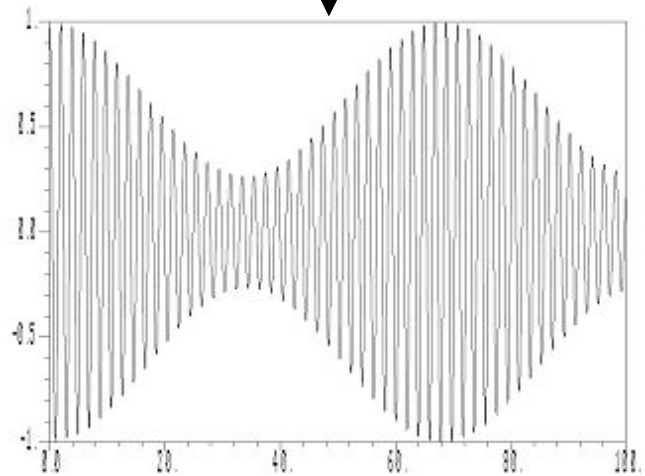


Mise à la puissance

Nb d'atomes (ici 2)



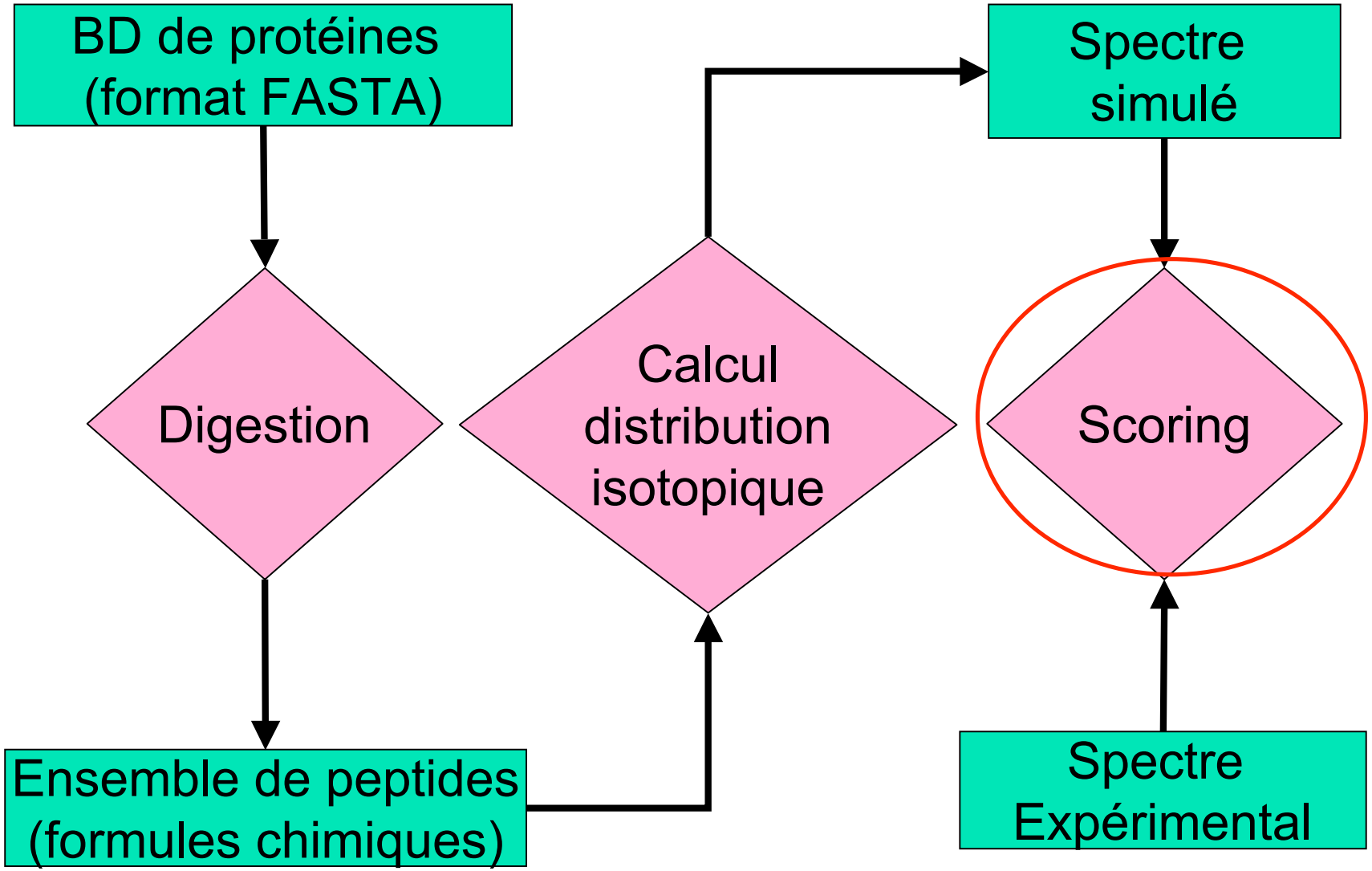
Transformée De Fourier inverse



Génération des spectres théoriques (2/2)

- Pour chaque peptide \ formule chimique.
- Pour chaque formule chimique (CxHyOz…) :
  - Pour chaque élément:
    - Distribution originel (ex: C_1).
    - Transformée de Fourier \ espace de Fourier.
    - Mise à la puissance dans l'espace de Fourier :
      - Ex: C_5 ∫ $C_1 * C_{1*} C_{1*} C_{1*} C_1$
    - Transformée de Fourier inverse \ espace Euclidien.
    - Addition de la distribution de chaque élément :
      - Ex: Cx + Hy + Oz + …
- Addition de chaque spectre peptidique.
- Fonction gaussienne pour ajuster la largeur des pics.

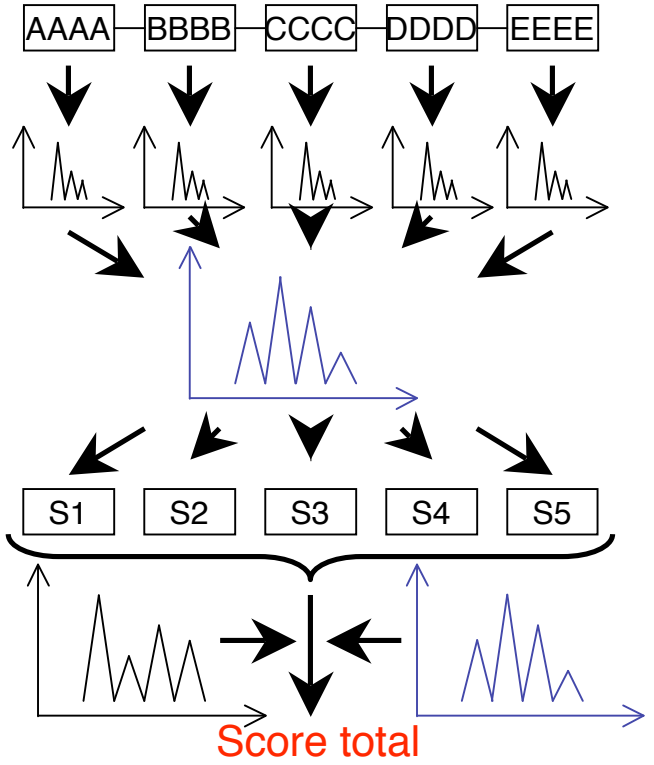
Schéma global



Scoring (maximisation)

- Comparer un spectre expérimental et un spectre simulé.

- Spectre expérimental.
- Génération des spectres théoriques:
  - Liste de peptides.
  - Calcul des distributions isotopiques.
  - Comparaison avec le spectre exp.
  - Score partiel pour chaque peptide.
- Score total:
  - Selon la répartition des scores partiels.
  - Selon les peptides absent du côté simulé.



Version offline

- Application en ligne de commande.
- Utilisation d'un fichier de configuration.
- Toutes les données peuvent être facilement étendue :
  - Grammaire des enzymes utilisables.
  - Modification post-traductionnelles activables.
  - …

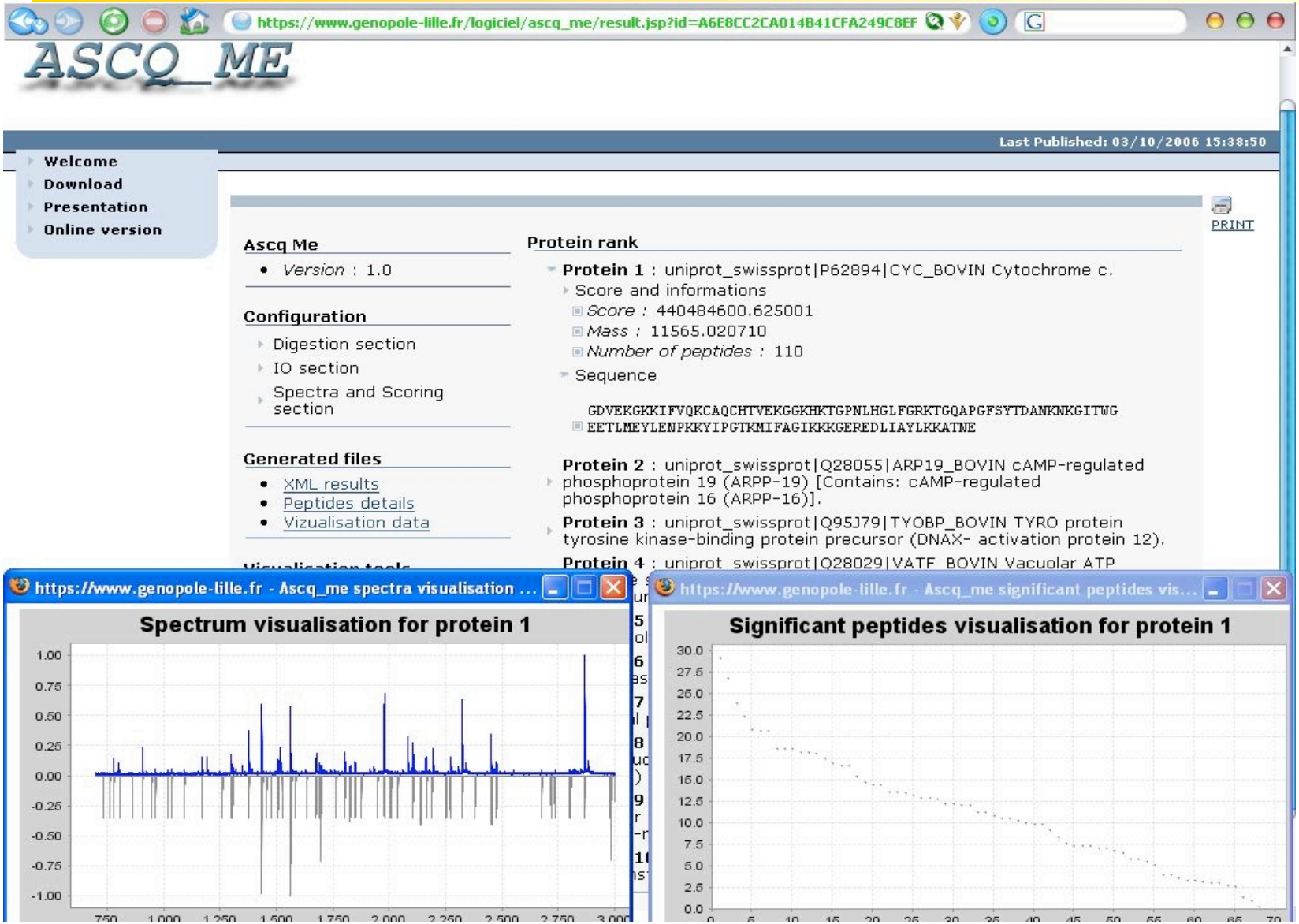
Exemple d'un fichier de configuration

```
# Original author : Jean-Charles BOISSON (2005)
#
# Default configuration file for ASCQ_ME
# all the line with a '#' for first character as commentaries
#
# this file is splitted in different sections, all of them respect a
# specific order (so don't change it) and a semantic
# for all the parameters, the style is:
#          parameter name (space) = (space) value or the key word DEFAULT
#
# to change a value, delete the key word DEFAULT and put a value
# ex : You want to accord 1 miss-cleavage so the line
# MISS_CLEAVAGE = DEFAULT 0 become MISS_CLEAVAGE = 1
# to restore the old value, write MISS_CLEAVAGE = DEFAULT
#
#####################################
# Digestion section
#####################################
#
# Give the number of miss-clevage allowed
# DEFAULT = 0
MISS_CLEAVAGE = DEFAULT
#
# Indicate if all the possible miss-cleavage have to be used
# in this case the parameters MISS_CLEAVAGE isn't used
# DEFAULT = FALSE
FULL_DIGEST = DEFAULT
#
# Indicate the charge to add on peptid result of digestion
# i.e. give the chemical formula
# DEFAULT = H
PEPTID_CHARGE = DEFAULT H
```

Version online

- Enregistrement en tant qu'utilisateur.
- Configuration complète via une interface graphique.
- Signalisation de la fin de la requête par mail.
- Visualisation des résultats :
  - Meilleures protéines obtenues.
  - Corrélation spectrale.
  - Participation de chaque peptide dans le score.

ASCQ-ME: capture d'écran



Nouvelles approches : enjeux (1/2)

- Fournir des outils :
  - Performants.
  - Rapide.
  - Pouvant évoluer.
- Créer des « méta » outils :
  - Combinant d'autres outils.
  - Automatisant l'utilisation de différentes approches.
  - Permettre à l'utilisateur une plus grande interactivité.
- Offrir des outils complètement disponible à toute la communauté (et pas forcément des « boites noires »).

Nouvelles approches : enjeux (2/2)

- Face à la quantité de données :
  - Méthodes dédiées ∫ réduction de l'espace de recherche :
    - Recherche de modifications post-traductionnelles.
    - Recherche de motifs.
    - …
  - Utilisation d'une plus grande puissance de calcul ∫ garder l'espace de recherche \ « force brute » :
    - Cluster de machines.
    - Notion de Grille.
      - Adapter les méthodes existantes.
      - Créer de nouvelles méthodes dédiées au calcul sur grille.

FT-MS instrument 9.4 Tesla, 2006

Coût 1.000.000 euros.

Avec NANO-LC \ 1 Go de données / heure.

Conclusion et perspectives (1/3)

- Version actuelle :
  - Basée sur des spectres MALDI.
  - Score en phase de test.
  - Licence en cours de relecture avant de disposer des sources complètes en téléchargement.
  - Processus peu rapide.
  - Pas de calcul réel de l'intensité.

Conclusion et perspectives (2/3)

- Perspectives :
  - Tests et adaptation aux autres types de spectres.
  - Prise en compte des décalages dans la comparaison des spectres.
  - Parallélisation de l'application.
  - Ajout d'un z-score (en cours de test).
  - Calcul des vrais intensités dans la génération des spectres.

Conclusion et perspectives (3/3)

- Futures extensions possibles :
  - Simulateur de spectres MS/MS :

    \ Identification à partir de spectrométrie de masse en tandem.

  - Intégration d'une approche par « *de novo protein sequencing* »

    \ Génération de séquence protéique à partir de spectre MS et MS/MS.

Remerciements

- Plateforme protéomique :
  - Participation à toutes les manipulations dans le processus d'identification de protéines.
- Pierre Laurence :
  - Inspiration et mise en ligne.
- Nicolas Dolet & Nicolas Gruszczynska :
  - Génération de spectres.
- Grégory Vanuynsberghe :
  - Étude statistique et développement d'un z-score.

Questions ?

Merci de votre attention.

Vers du « de novo protein sequencing »

- Mise au point d'une méthode en 3 temps pour séquencer une protéine :
  - À partir de spectres MS et MS/MS.
- 3 étapes :
  - Obtention des formules chimiques des peptides expérimentaux.
  - Obtention des séquences peptidiques.
  - Ordonnancement des peptides.

Comparaison globale (1/2)

	PE dans BDs	VPE	PEN
Spectre MS	PE Séquence trouvées	Protéine originale Séquence non	Pas de résultats
Spectres MS/MS (DNS+blast)	PE Séquence trouvées	Protéine originale Séquence partiellement	Séquence partielle pour chaque peptide

PE : protéine expérimentale. BD : base de données.

VPE: protéine expérimentale ∫ variant d'une protéine des BDs.

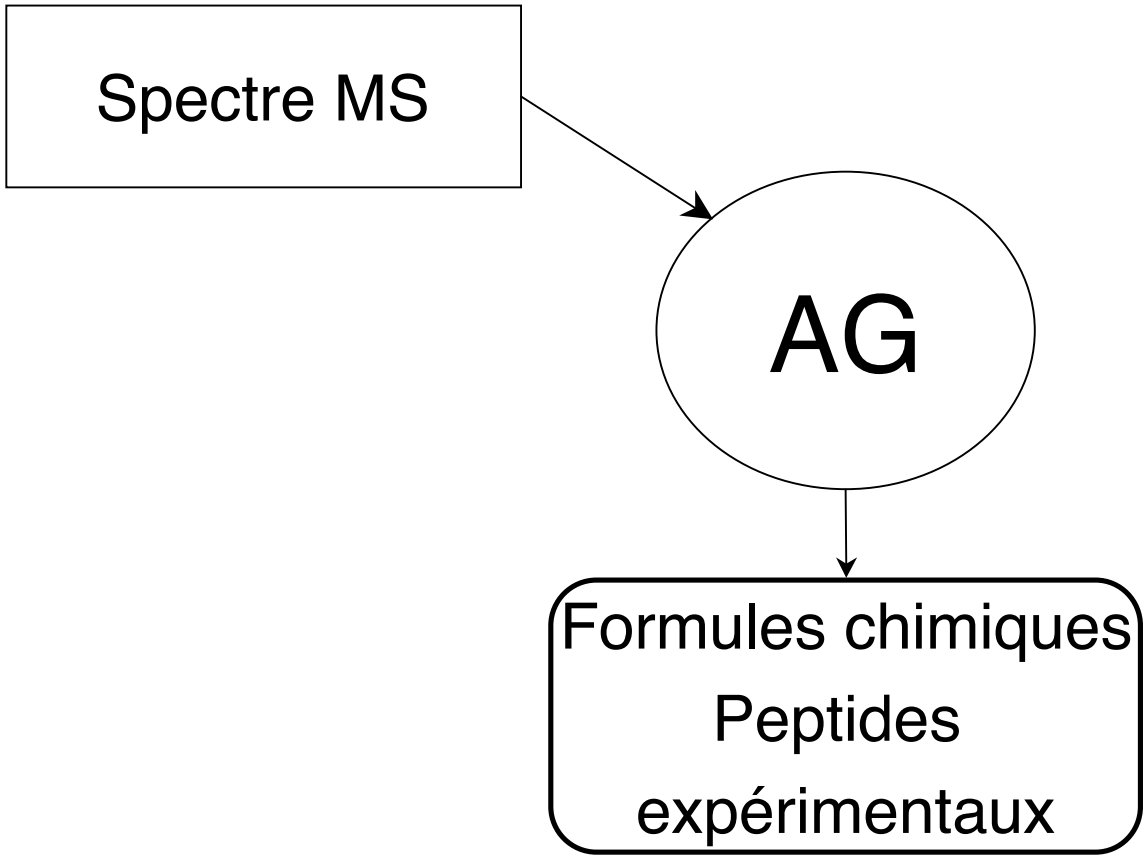
PEN: protéine expérimentale ∫ absentes des BDs.

Comparaison globale (2/2)

	PE dans BDs	VPE	PEN
Spectre MS	PE Séquence trouvées	Protéine originale Séquence non	Pas de résultats
Spectres MS/MS (DNS+blast)	PE Séquence trouvées	Protéine originale Séquence partiellement	Séquence partielle pour chaque peptide
Spectres MS + MS/MS	Séquence trouvée PE	Séquence trouvée Protéine originale	Séquence trouvée

Etape 1 : algorithme génétique

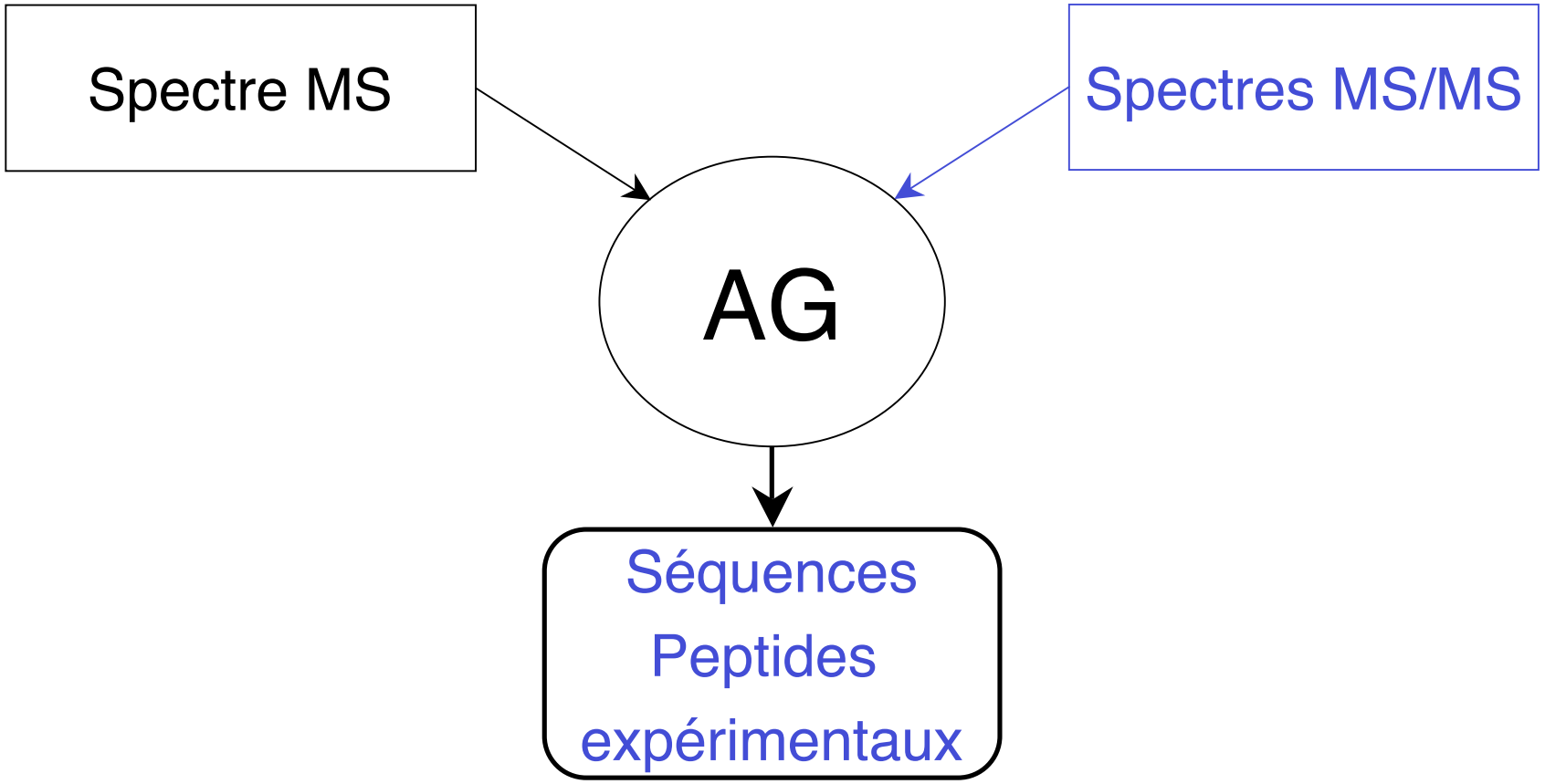
Déjà implémenté



Etape 2 : intensification

Déjà implémenté

En test

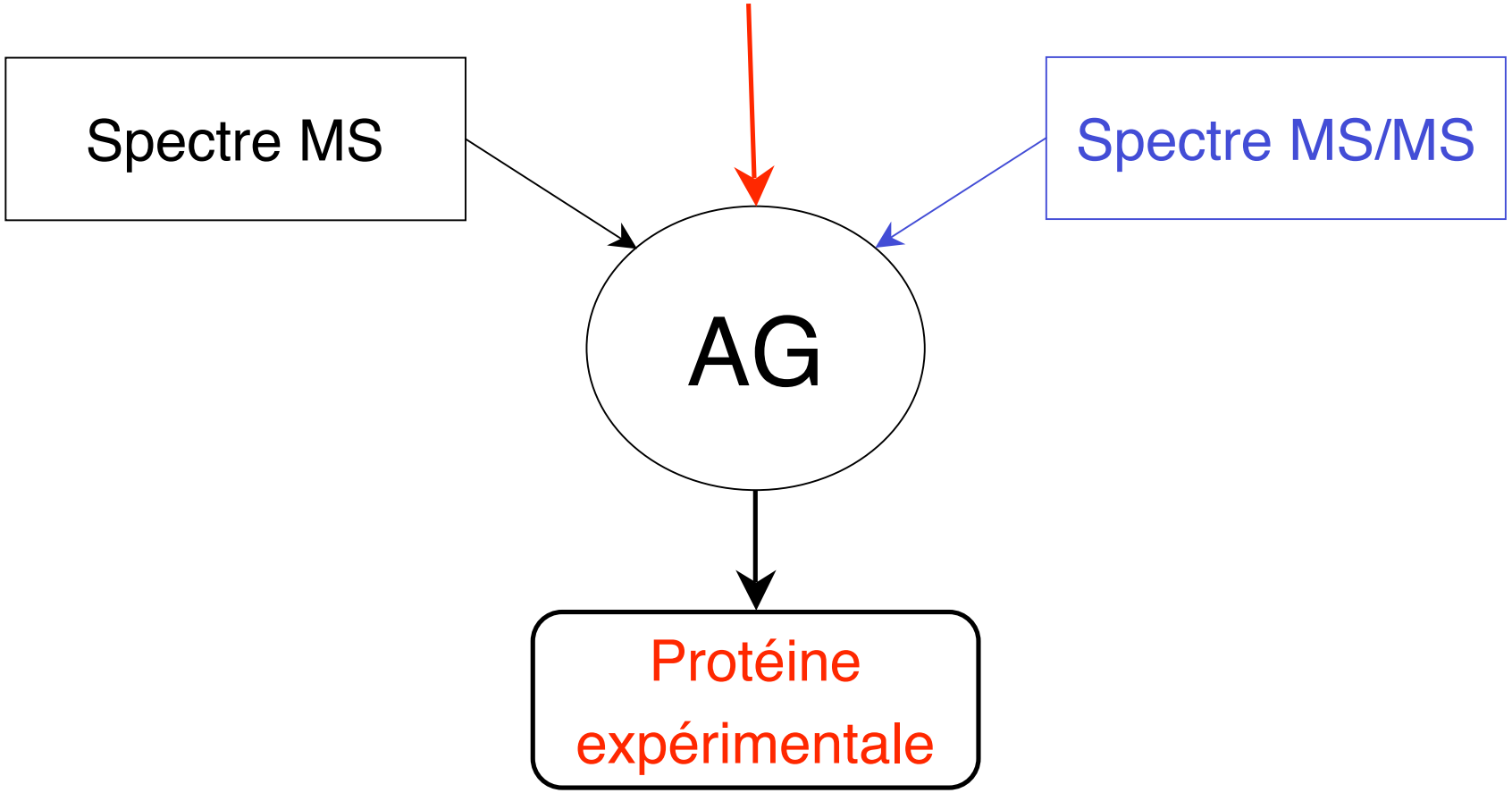


Etape 3 : ordonnancement

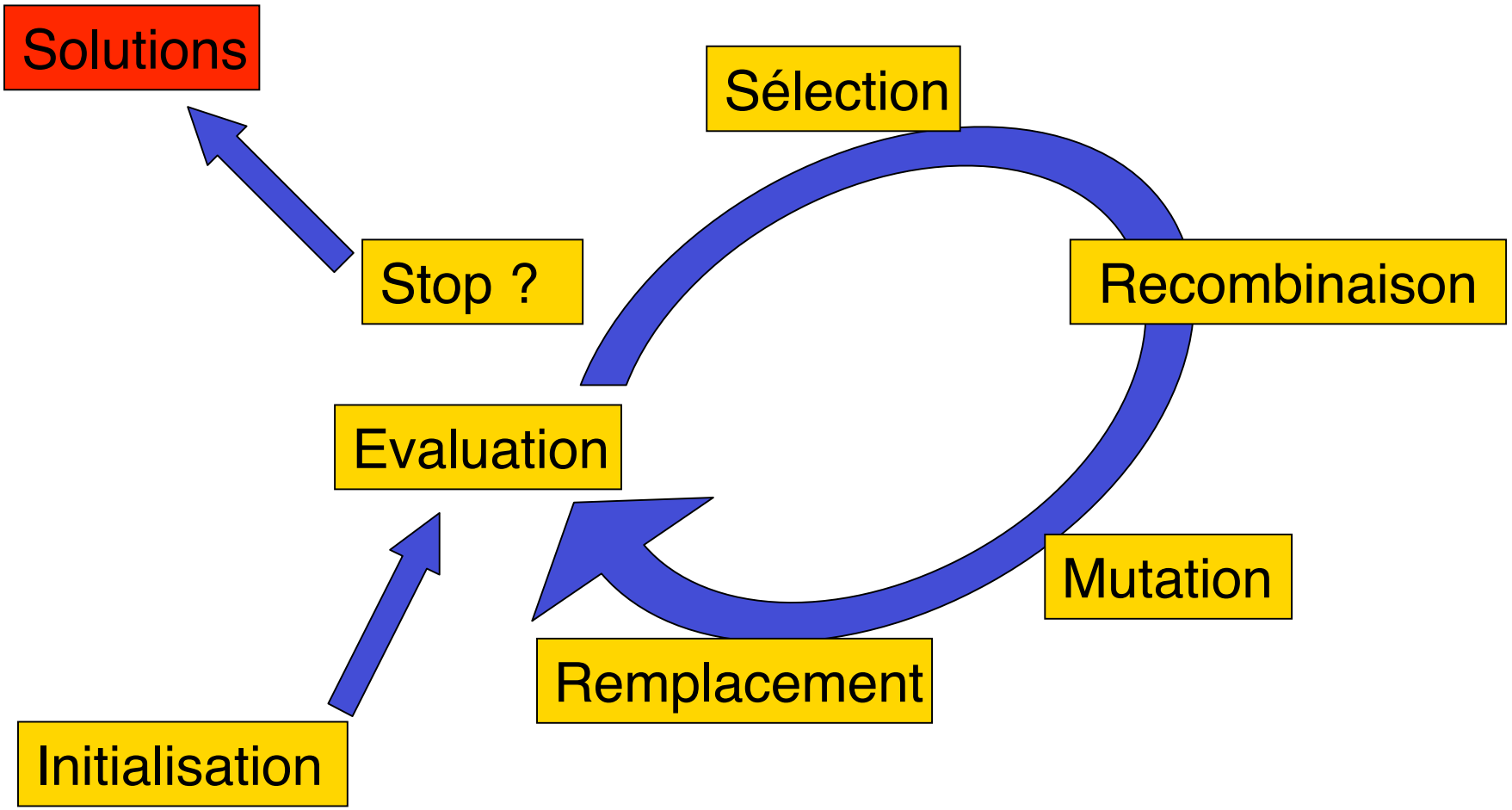
Déjà implémenté

En test

En test



Algorithme génétique



Caractéristiques de l'approche

- Utilisation de méta heuristiques :
  - 1ère étape ∫ AG avec le scoring d'ASCQ-ME.
  - 2ème et 3ème étape ∫ 2 autres méta heuristiques.
- Processus non déterministes.
- Guider la recherche.
- Nécessité d'un grosse puissance de calcul :
  - 1ère étape \ run > 2 jours en séquentiel.
  - Utilité des grilles de calculs \ test sur Grid5000.