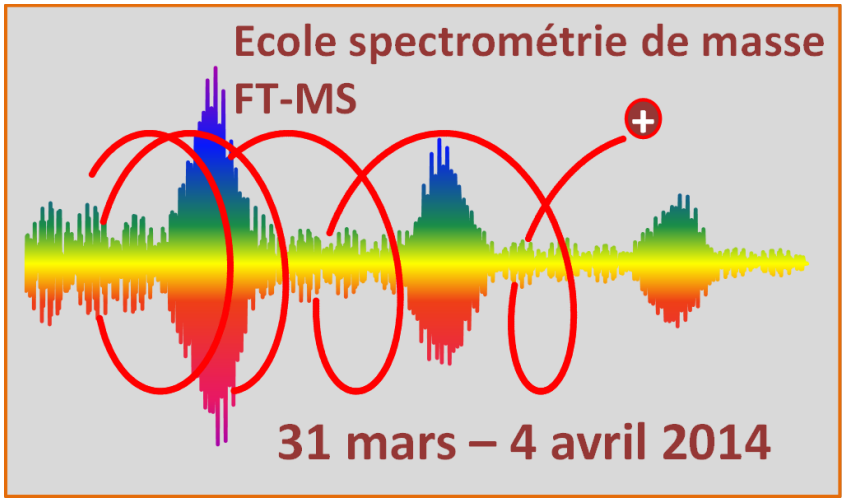
Partie I
Des spectres à l'identité des protéines

Thomas Burger

--

Etude de la Dynamique
des Protéomes



Présentation

- Ingénieur en télecom et mathématique discrète
  (Ensimag, 2004)
- Thèse CIFRE en vision par ordinateur
  (GIPSA-Lab et Orange-labs, 2007)
- Maître de conférences en informatique décisionnelle
  (2008-2011, Université de Bretagne Sud)
- Chargé de Recherche au CNRS
  (à EDyP, depuis 2011)

Plan

1. Introduction générale
2. Extraction du signal
3. Identification des peptides
4. Inférence de protéines
5. Parenthèse statistique
6. Contrôle de la qualité

- Objectifs questions et enjeux
- Pipe-line d’analyse
- Infrastructure logicielle

INTRODUCTION GENERALE

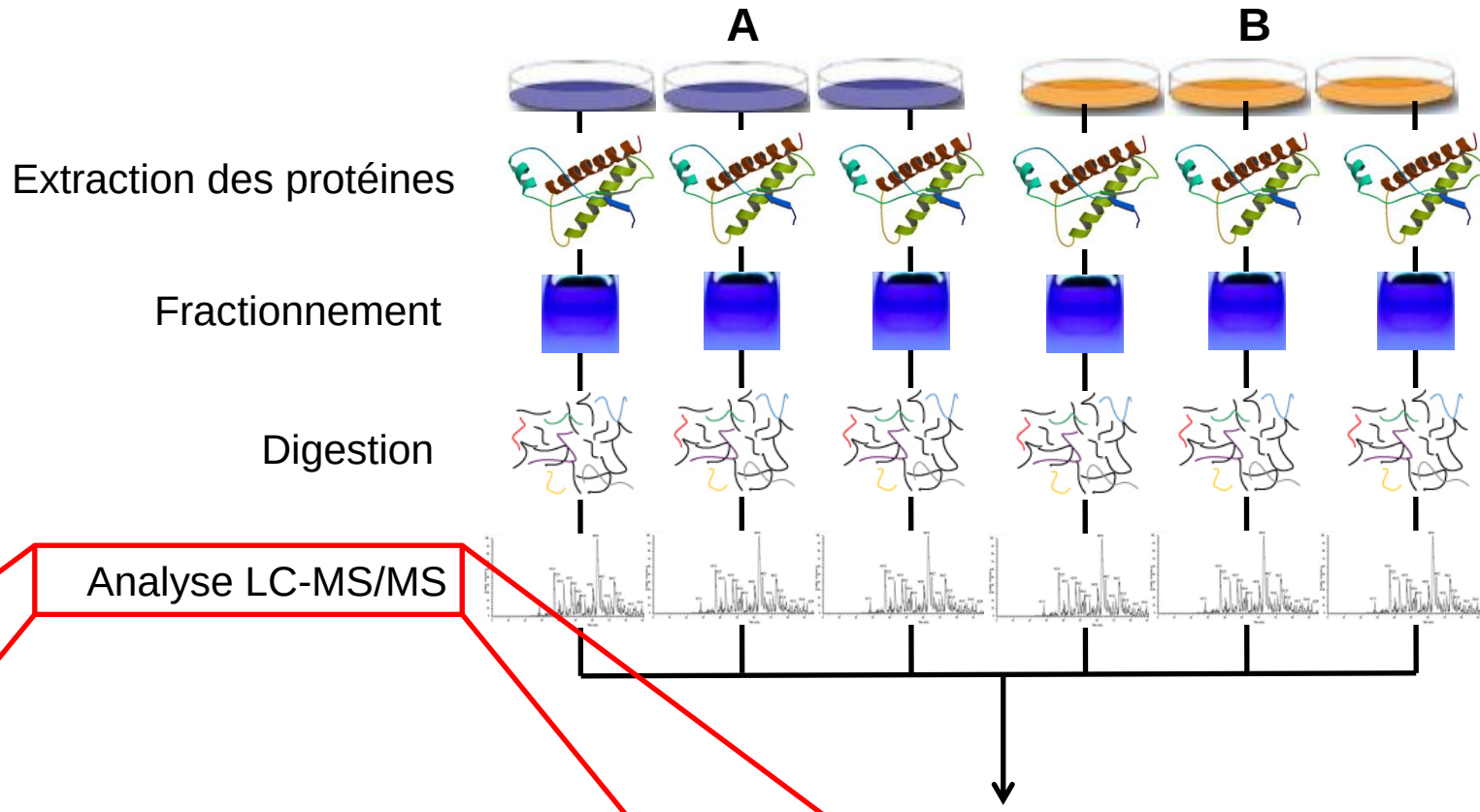
Objectifs, questions et enjeux

- Résumer les étapes “sèches” d’un pipe-line de **quantification relative label-free**

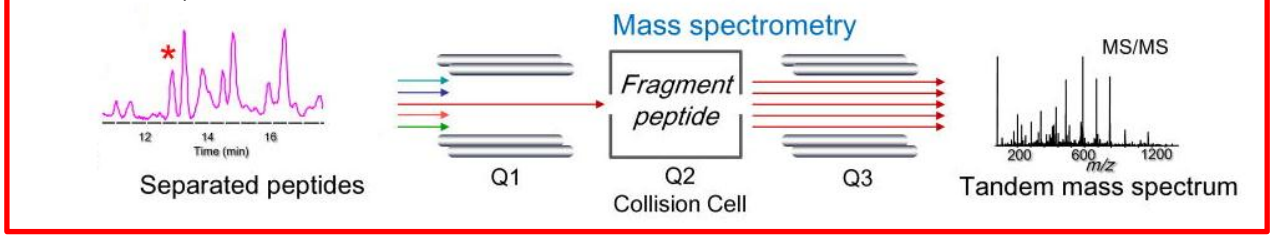
- Présenter les choix algorithmiques ou statistiques possibles

- Insister sur la maîtrise de la qualité des résultats finaux…
  Cela fait partie de la protéomique !!!

Pipe-line d'analyse (humide)



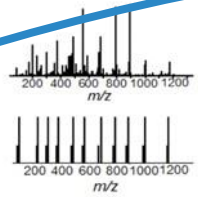
Nesvizhskii, 2011



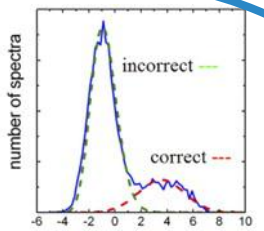
Pipe-line d'analyse (sèche)

Common Data Format

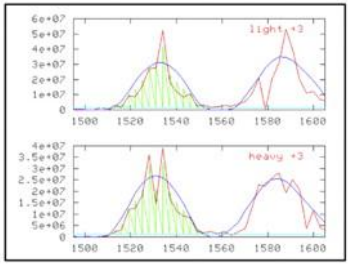
1. Conversion to and use of open data formats



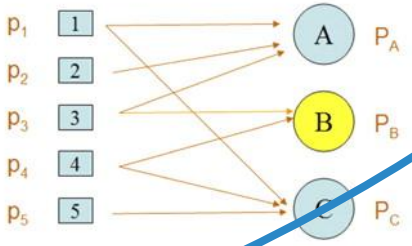
2. Spectrum identification with a search engine



3. Validation of identifications

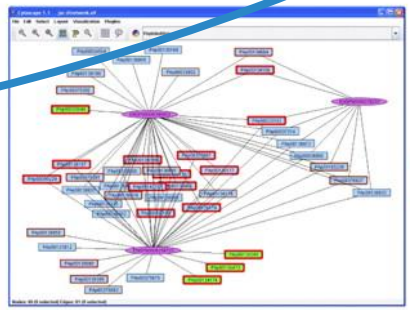


5. Quantification

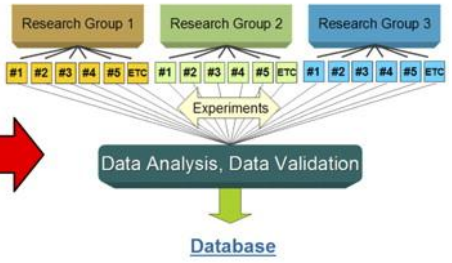


4. Protein inference

6. Organization in local data management systems



7. Interpretation of the protein lists



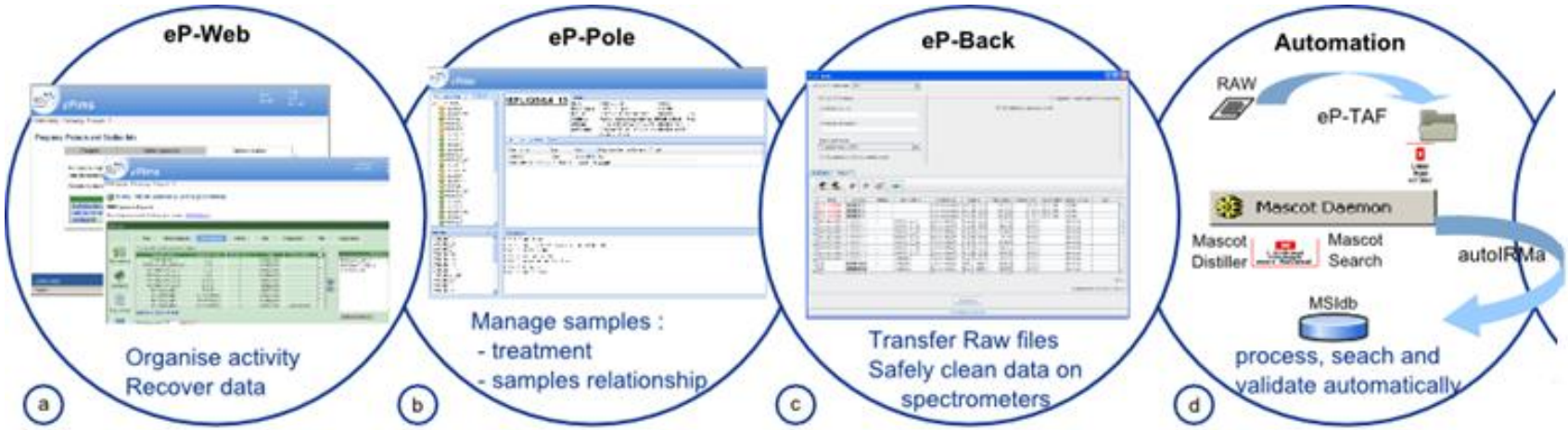
8. Transfer to public data repositories

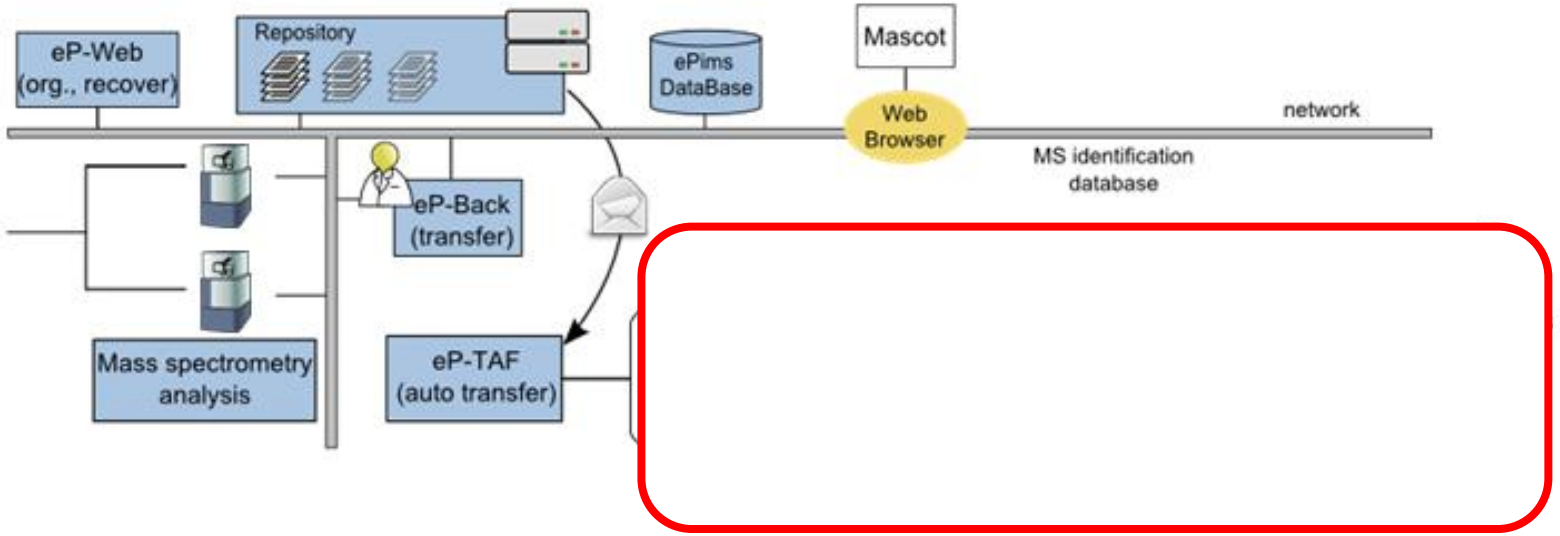
Deutsch E W et al. Physiol. Genomics 2008;33:18-25

Infrastructure logicielle: pourquoi ?

- Stockage et sécurisation des données
- Indexation pour usage ultérieure
- Contrôle qualité (ISO9001) du processus
- Automatisation des différentes étapes

Notre infrastructure logicielle





- Deisotoping
- Etats de charge
- Débruitage
- Etc.

EXTRACTION DU SIGNAL

Dans le spectro

- Beaucoup de traitements invisibles
  - Deisotoping, sélection de l'état de charge
  - Exclusion dynamique
  - Codage du signal (mode profile/centroid)
- Tout cela, dans un logiciel propriétaire
- Influence sur la qualité des résultats ?

En sortie du spectro, sur les raw

- Sur le spectre MS
  - Calcul de la masse du précurseur
  - Vérifications (état de charge, deisotoping)

- Sur le spectre MS/MS
  - Deisotoping
  - Extraction de peak-list
  - Denoizing

Quelques outils

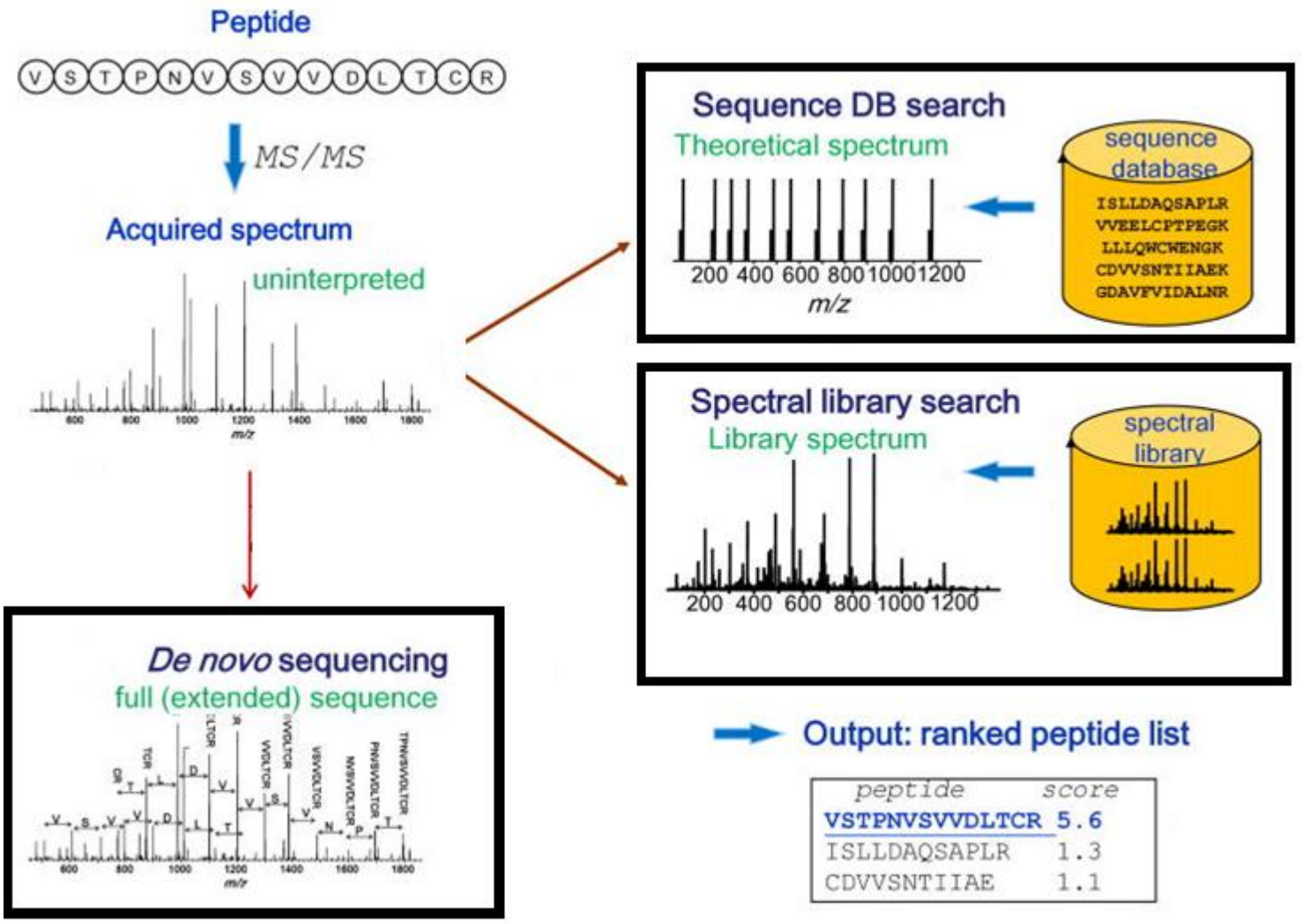
- Plusieurs suites logicielles incluant
  - Différents traitements modulaires
  - Un “daemon” permettant d’enchaîner les traitements, + synchronisation p.r. production des données

- Mascot Daemon + Mascot Distiller
- TPP, OpenMS, Scaffold, MyProMS, etc.

- Généralité sur les moteurs d'identification
- Principe de fonctionnement

IDENTIFICATION DES PEPTIDES

Stratégies



Nesvizhskii, 2011

Résultats de l'identifications

- Comment interpréter la sortie ?
  - Liste ordonnée
  - Le premier est souvent le bon… Mais pas toujours

- Outils ouverts/fermés: prix, complexité de maîtrise, durabilité/maintenance, etc.

Outils complémentaires

- Prophets:
  - recalcule de nouveaux scores
  - On sait interpréter le score
  - Les scores sont cohérents sur le dataset

- Mergers (ou outils de meta-scoring):
  - Combine le résultat de plusieurs moteurs
  - En théorie, plus robuste
  - Usine à gaz, perte de la maîtrise

- Filters:
  - Outils de prise de décision sur les PSM
  - Indispensable pour éviter un traitement manuel

Les choix d'EDyP

- A l'heure actuelle:
  - Database seach avec Mascot
  - Pas de Prophet, pas de Merger
  - Outil de filtrage: IRMa
  - Gestion des contextes expérimentaux: hEIDI

- A l'avenir:
  - Intégration de ces outils dans ProLine
  - Plusieurs outils d'identification possibles
  - Peut-être usage d'un merger…

Database search - fonctionnement

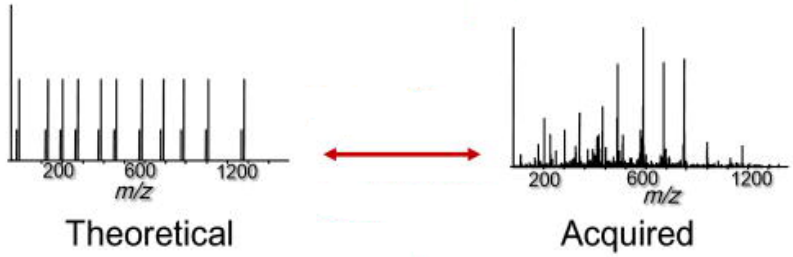
- Constitution d'une base FASTA de réference

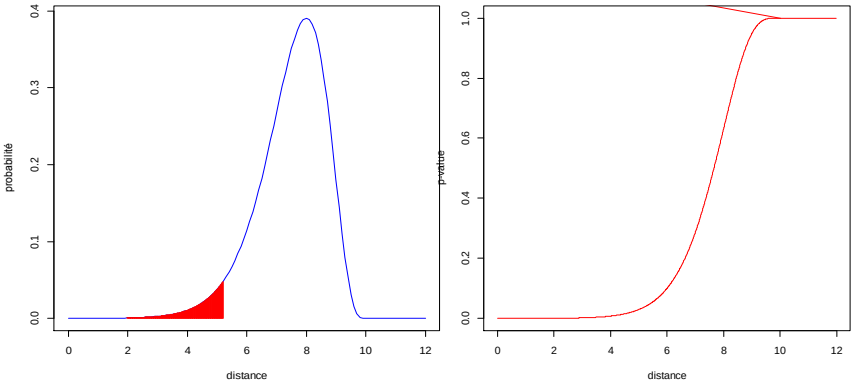
- Mesure de distance:
  - MOWSE score
  - Autocorrelation
  - Produits scalaires

- Un test d'hypothèse sur cette mesure => p-value

- La p-value ***p*** est convertie en un score ***S***:

Nesvizhskii, 2011



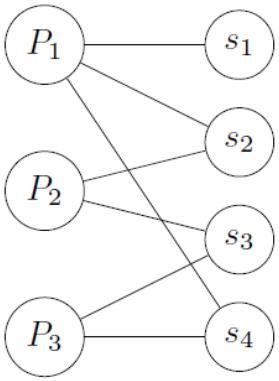


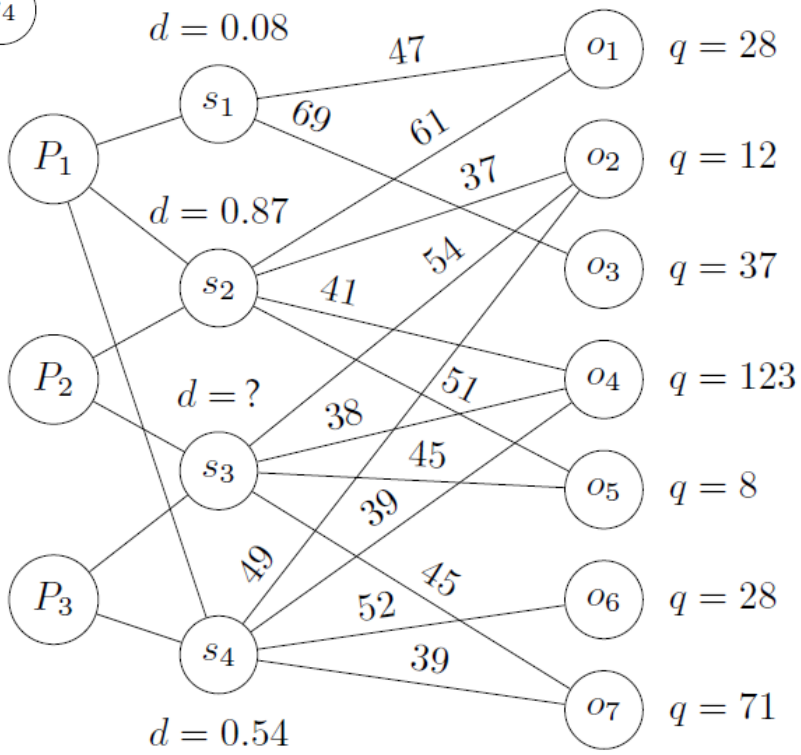
$$S = -10 \cdot \log_{10}(p)$$

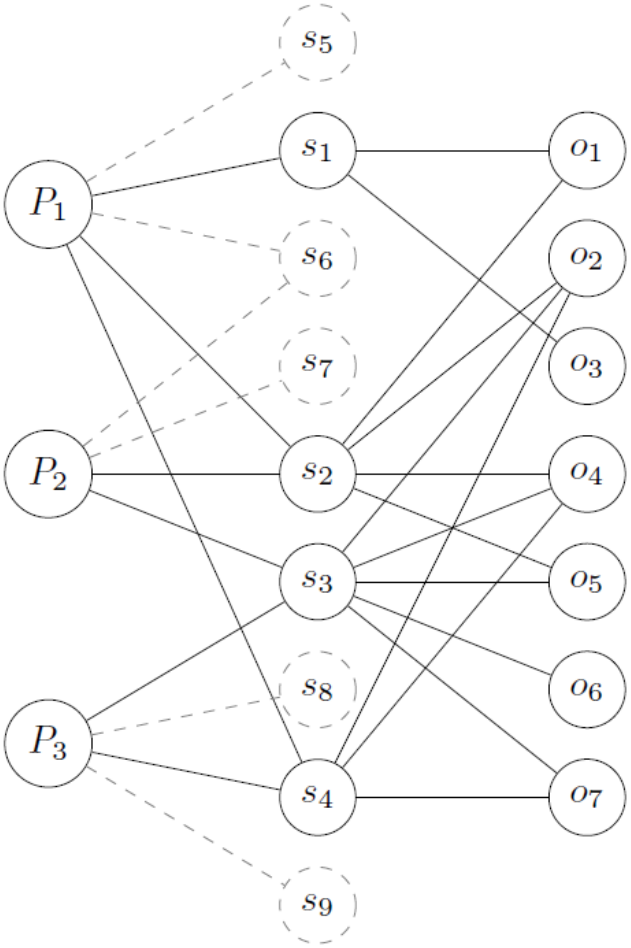
- Les différentes versions du problème
- Les outils disponibles
- Les enjeux futurs
- Les choix d'EDyP

INFÉRENCE DE PROTÉINES

Différentes versions du problème







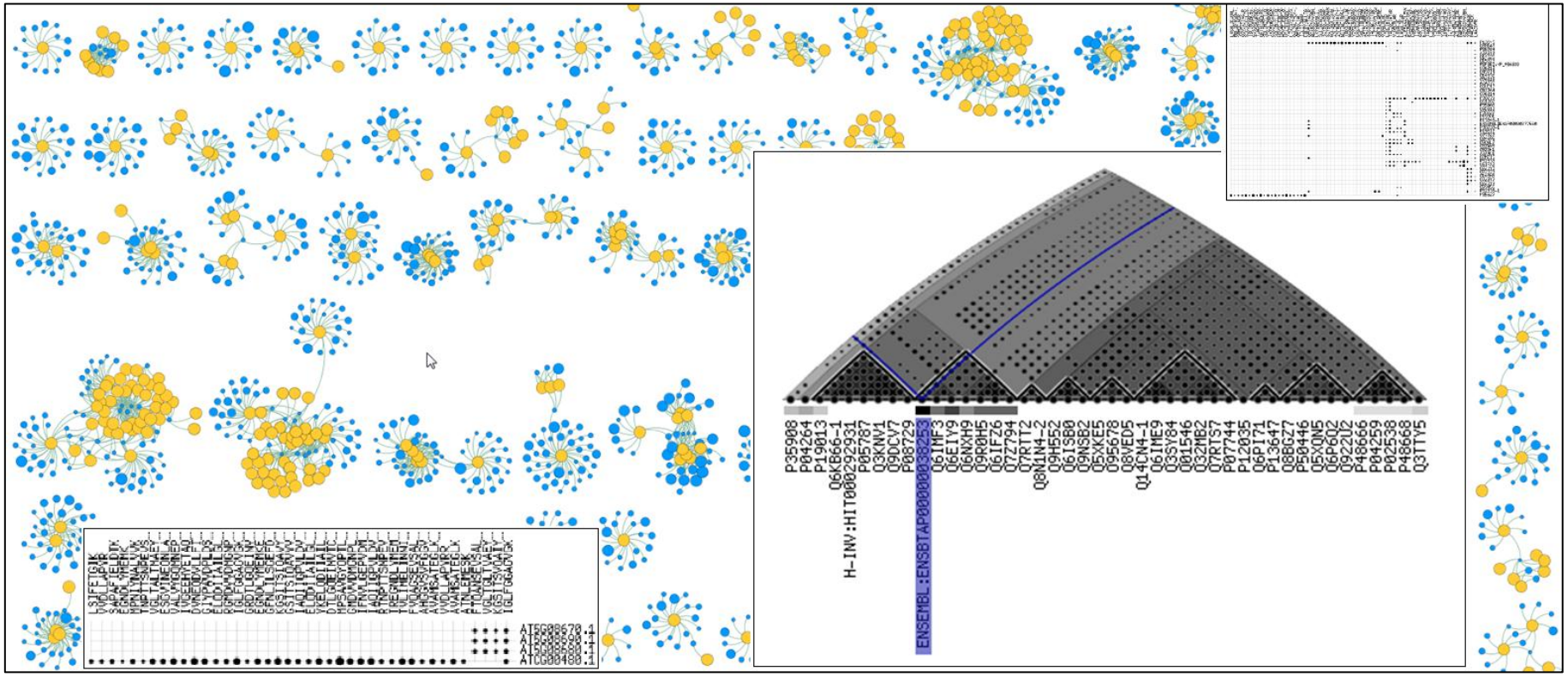
Outils disponibles

- Optimisation combinatoires
  - One and two peptides rules
  - DTASelect
  - IDPicker

- Méthodes statistiques
  - Iterative : ProteinProphet, Scafold, EBP, PANORAMICS
  - Various types of Bayesian model (nested, hierachical, etc.)

Serang & Noble, 2012

Enjeux futurs: la visual analytics



Choix d'EDyP

- Pour l'instant algorithme simple et glouton correspondant au problème le plus simple
- Tentatives d'amélioration en cours
- Incorporation d'outils de visualisation

- Beaucoup de choses atch'compliquées !!!

PARENTHÈSE STATISTIQUE:
INTRODUCTION AU TEST D'HYPOTHÈSE

Une question et une réponse

- Un spectre observé est-il suffisamment proche d'un spectre théorique pour considérer que c'est un "match" ?

- On peut répondre OUI (à tort ou à raison) ou NON (aussi à tort ou à raison)

- On peut se tromper de 2 manières différentes

Contrôler l'erreur la plus grave

- Dans notre cas:
  - On **peut** accepter un NON A TORT, i.e. oublier un PSM.
  - On **ne peut pas** accepter un OUI A TORT, i.e. considérer un "match" qui n'en est pas un.

- On définit l'**erreur de première espèce** comme l'erreur inacceptable

- *En générale on ne peut pas accepter de définir à tort un écart à la norme.*

Hypothèse nulle

- Une erreur de première espèce revient à rejeter à tort l'**hypothèse nulle** (notée H0).

- On peut donc formuler nos hypothèses:
  - **H0**: le couple observation/spectre n'est pas un PSM.
  - **H1**: (**hypothèse alternative**) il est un PSM.

- On appelle **positif** ou **découverte**, un test pour lequel H0 est fausse (contraire: **négatif**)

- On qualifie de **fausse** une décision prise à tord (contraire: **vraie**)

La question du test d'hypothèse

Est-ce que compte tenu de mes observations, je peux rejetter H0 ?

- OUI: le rejet de H0 implique l'acceptation de H1.
  le couple spectre/peptide est donc un "match"

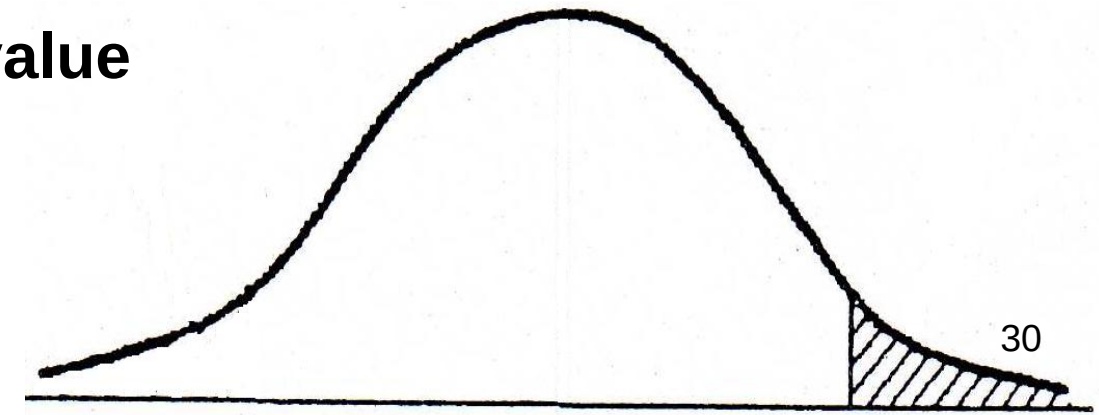
- NON: Est-ce qu'on peut accepter H0 alors ? Non plus… On ne peut donc rien dire.
  Le couple spectre/peptide est donc soit un "match", soit pas… (soit H0 est vraie, soit c'est une *erreur de seconde espèce*, donc moins grave).

Réalisation du test

- Définir une mesure de distance entre un peptide et un spectre

- Définir la distribution de cette distance quand H0 est vraie :
  Un “histogramme” des mesures pour des “non-match”

- Pour chaque peptide à tester, calculer la probabilité d’une observation au moins aussi extrême: La **p- value**



Un petit quizz…

Concrètement, une p-value de p = 0.02 pour un PSM donné s'interprète comment ?

A. Il y a moins de 2% de fausses découvertes parmi tous les PSM ayant une p-value plus faible.

B. La probabilité que ce PSM soit une fausse découverte vaut 2%.

C. N'importe quelle fausse découverte à 2% de chance d'avoir un score meilleur que ce PSM qui m'intéresse ?

D. La réponse D…

E. Mon PSM est une vraie découverte puisque $p < 0.05$.

La réponse au Quizz:

A. Il y a moins de découvertes parmi tous les PSM ayant une FDR faible.

Contrôle du nombre de fausses découvertes

B. La probabilité [illegible] est 2%.

Ce qui nous intéresse : P [H0 | Obs]

C. [illegible] un score meilleur que ce PSM qui m'intéresse ?

La p-value que le test nous donne : P [Obs | H0]

E. Mon PSM est [illegible] puisque p < 0.05.

Règle pifométrique à oublier

Comment passer de **P [H0 | Obs]** à **P [Obs | H0]** ?

Anniversaires à EDyP

- Au sein d'EDyP, **Sylvain** et **Véronique** …

- … sont tous deux nés un 6 Juin !!!
  C'est fou, non ?

Le mystère de l'aéroport St Exupéry

- Pour chaque humain, la probabilité qu'il passera devant la porte 15 du Terminal 1 de l'AST, le 6 juin 2016 est presque nulle.

- Pourtant, le 6/6/16… on aura ceci:

- **C'est fou, non ?**

Application à la protéomique

- L'anniversaire :
  - Nous voyons deux observations qui "matchent" avec tellement de précision que nous pensons cela difficilement explicable par le hasard
  - En fait, le hasard est la meilleure explication
  - **C'est une fausse découverte**

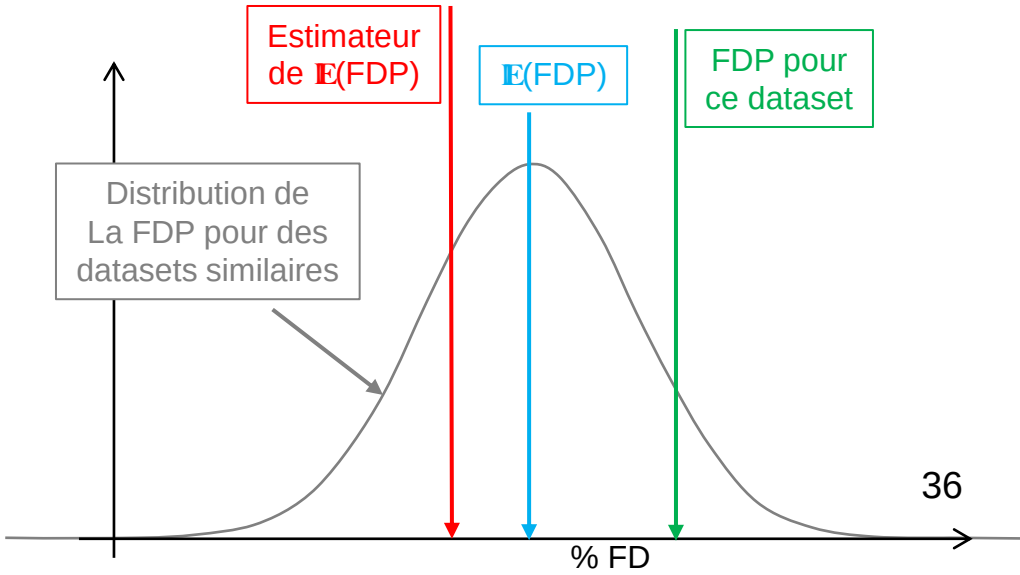
- L'AST:
  - les phénomènes rares apparaissent avec une proba forte
  - Parmi vos découvertes "fiables biologiquement", et publiées, il y a sur le nombres, quelques erreurs…
  - **Problème des tests d'hypothèse multiples**

Corrections de tests multiples (MTC)

- Besoin de MTC : Il faut pouvoir contrôler la ***False Discovery Proportion (FDP)***, alors qu'elle est inconnue.

- Les stratégies types FDR proposent:
  - De s'intéresser à l'espérance de la FDP: $\mathbb{E}$(FDP)=FDR
  - D'estimer celle-ci le mieux possible…

- Méthodes classiques:
  - Benjamini-Hochberg
  - Permutation-based FDR
  - Target-Decoy
  - Etc…



- De l’identification des peptides
- De l’identification des protéines

CONTRÔLE DE LA QUALITÉ

Pour l'identification de peptides

- Méthode classique: Target-Decoy
  - On crée des faux peptides
  - La proportion de faux PSM selectionnés donne le FDR

- Preuve que cela correspond à une stratégie de type FDR ? Qualité de l'estimateur ?

- Cela correspond à une moyenne sur…
  1 seul individu !!!

Alternatives possibles

- Avoir des bases Decoy gigantesques
- Remonter à la probabilité critique de l'identification (si possible ?) et utiliser BH ?
- Niveau de stabilités des identifications par rapport à l'élargissement des bases FASTA ?

Et sur les protéines ?

- Contrôle du FDP au niveau des peptides et publication de listes de protéines
- Que représente un FDR au niveau protéique ? Quelle est l'hypothèse nulle ?
- A FDP constant:
  - si on augmente le nombre de découverte de peptides,
  - on augmente le nombre de fausses découvertes sur les protéines !!!

Au final…

- A quoi tout cela sert-il ?
- Le nombre de fausses découvertes n'est pas fiable
- Il s'agit d'un garde fou pour éviter des publications folkloriques