Infrastructure informatique et bio-informatique pour l'identification de protéines à haut-débit

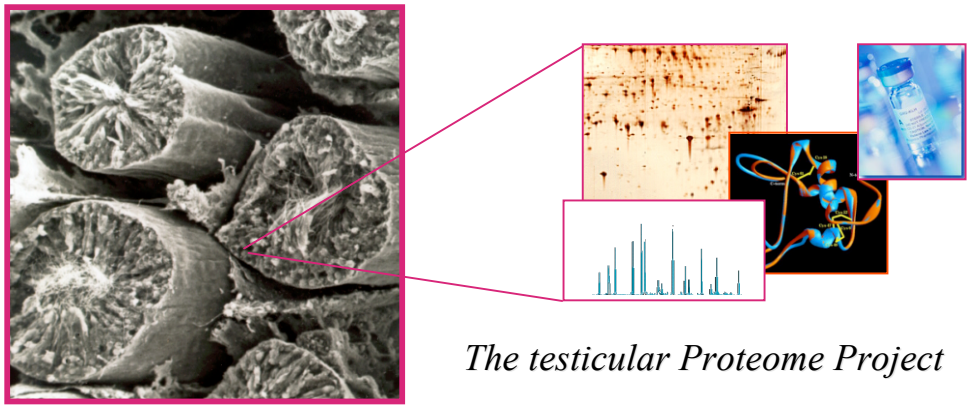
R. Lavigne – D. Ousmanou – C. Pineau

Plate-forme Protéomique OUEST-genopole®
Campus de Beaulieu, Bât. 24, 263 ave du général Leclerc, 35042 Rennes cedex

Objectifs du programme

- **Automatisation complète de la procédure d'identification de protéines à partir de gels 1D/2D**
- **Permettre aux intervenants de se concentrer sur les étapes de validation**

- **Etablir les bases technologiques pour la réalisation d'un programme de protéomique systématique**



The testicular Proteome Project

- **96% des spectromètres MALDI utilisés en mode manuel** *(source: HUPO)*

- **92% des utilisateurs régulieurs de MASCOT travaillent sur serveur distant** *(source: MatrixScience)*

- **55% des utilisateurs travaillent directement sur les bases de données originales** *(Source: HUPO)*

Workflow

Projet :
Echantillon(s)

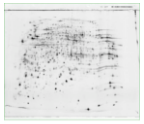
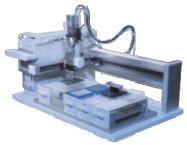
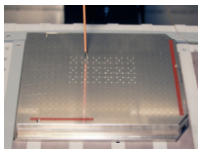


Image Master Platinum

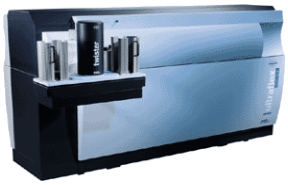
Ettan Worstation



Map II/8 robot



Anchorchip Target



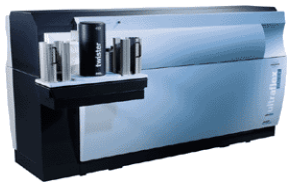
ULTRAFLEX ToF/ToF
FlexControl (measure, HPC),
FlexAnalysis
Spectre MS

ProteinScape

Moteurs de recherche : MASCOT, Phenyx
Banques de données (NCBI, ESTs…)
Cluster 8 noeuds
Configuration Linux

WARP control

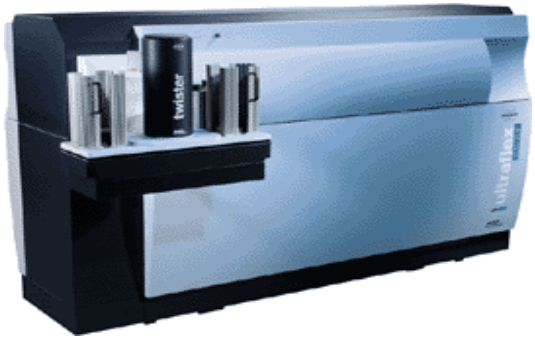
Spectre MS/MS
FlexAnalysis
FlexControl (LIFT)
ULTRAFLEX ToF/ToF



Acquisition et gestion des données de spectrométrie de masse

Génération des données - Rappels

- **Fichiers de données entrants :** ***Picklist (spots 2D, bandes 1D), image de gel***
- **Spectres de masses MS et MS/MS acquis automatiquement**
- **Liste de masses calibrées pour chaque protéine à identifier**
- **Feuille de résultats MASCOT**



Ultraflex™ ToF/ToF

Bruker Daltonics

ProteinScape™

Spectrométrie de masse
Ultraflex™ MALDI TOF/TOF Bruker Daltonics

Deux niveaux d'optimisation…

Automatisation de la méthode d'acquisition MS:

- **Calibration et précision: propriétés HPC (*High Precision Calibration*)**
- **Utilisation de l'intelligence artificielle pour l'optimisation de la puissance laser (*Fuzzy control*)**
- **Optimisation du mouvement et du nombre d'acquisition**
- **Optimisation de la gamme de masse d'acceptation des pics**

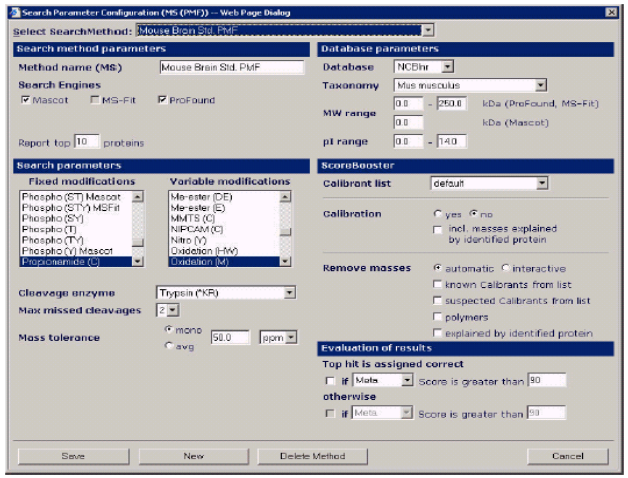
Automatisation de la méthode d'analyse:

- **Développement de macro de lancement automatique (Visual Basic)**
- **Assignements des masses utilisant la technique SNAP (Sophisticated Numerical Annotation Procedure)**

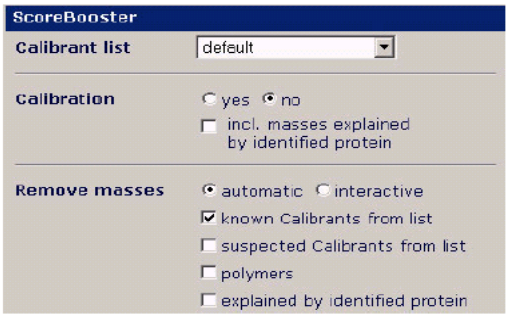
  identification et calcul du pic monoisoptopique, détermination de la ligne de base interne et du bruit de fond
- **Utilisation du fichier de calibration interne (calibration sur les pics d'autolyse de la trypsine = 2ème calibration)**

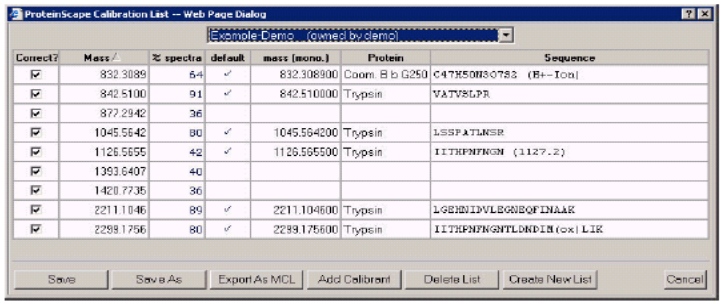
Optimisation sur ProteinScape 1.3

ScoreBooster



Configuration des paramètres de recherche pour spectres MS





Contaminants classiques, pics de l'enzyme d'autolyse (références internes)

Re-calibration "Intelligente" 3 ème calibration

Recalibration automatique des spectres

Elimination des masses connues du bruit de fond avant la recherche

Obtention de listes de masse avec une calibration très précise

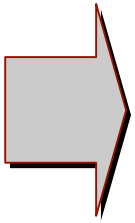
Optimisation sur ProteinScape 1.3

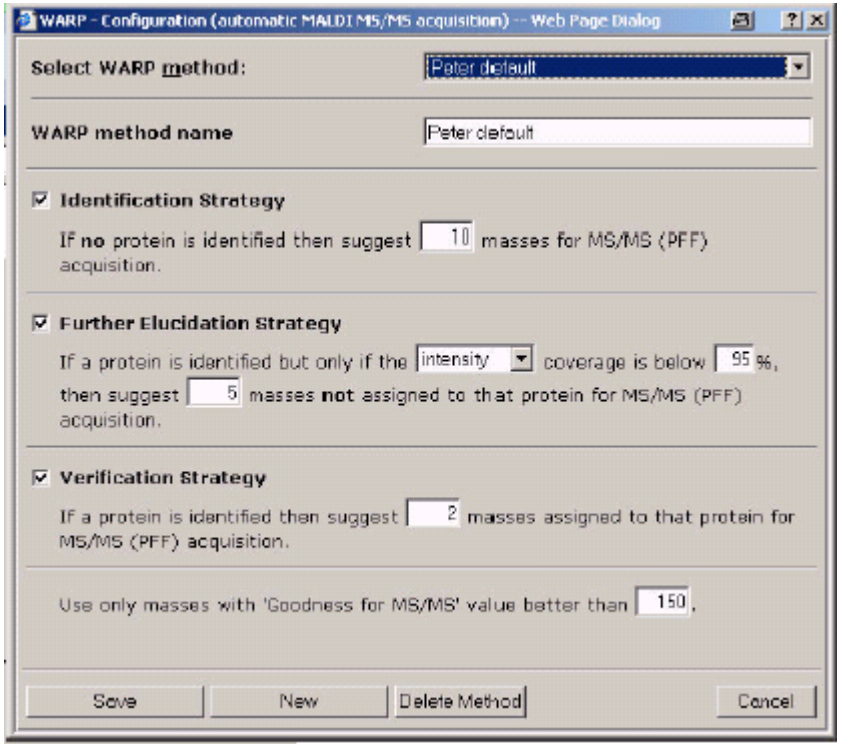
Le WARP Control (*Workflow Administration by Result-driven Processing*):

Contrôle intelligent en 2 étapes de l'instrument basé sur les résultats d'identification:

- Évaluation automatique du statut du spectre MS
- Acquisition automatique et intelligente de spectres MS/MS pour augmenter le pourcentage d'identification ou la confiance sur les identification

3 stratégies différentes d'utilisation des données MS/MS





Résultats de l'optimisation

Obtention d'une méthode d'acquisition rapide et précise

- Temps total d'acquisition et d'analyse par spectre = 26 sec
- Jusqu'à 100 pics de masses assignés
- Qualité des spectres (S/N > 35, Résolution > 10 000)

Automatisation sur 48 spots

- Temps total d'acquisition = 17,5 min
- Temps total de recherche en banques de données = 2h 57 min
- Taux d'identification de 63% sans configuration ScoreBooster

Automatisation sur 384 spots

- Temps total d'acquisition = 2h 20 min
- Temps total de recherche en banques de données = environ 23h 00
- Taux d'identification ≥ 70% avec ScoreBooster

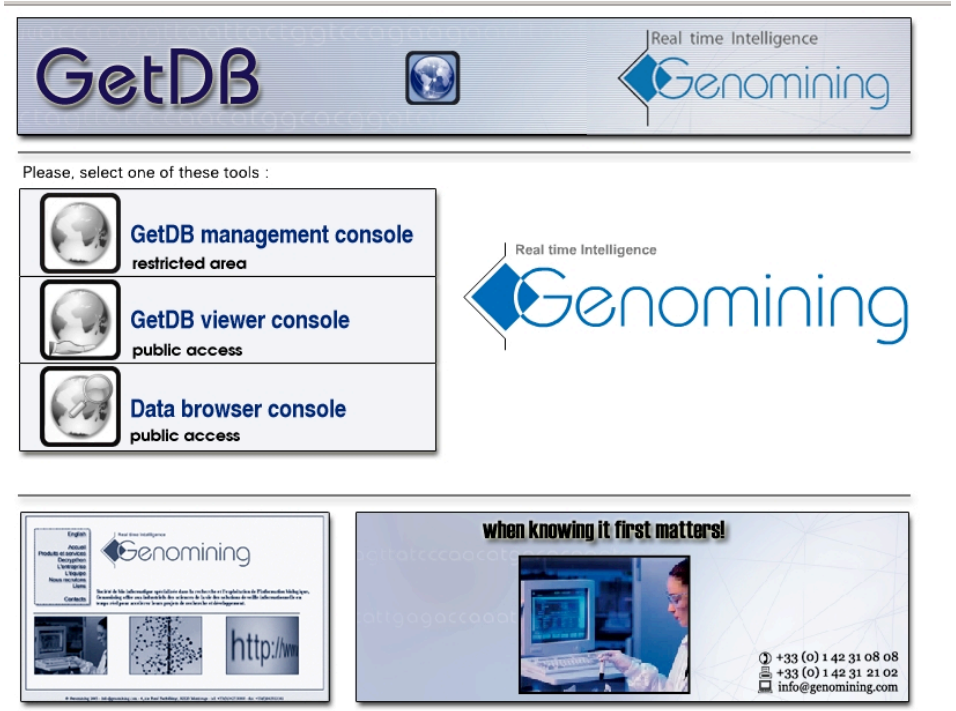
Temps incompatible avec les objectifs du haut-débit

Ne permet pas d'utiliser la configuration WARP Control

Outils de gestion et d'analyse des données

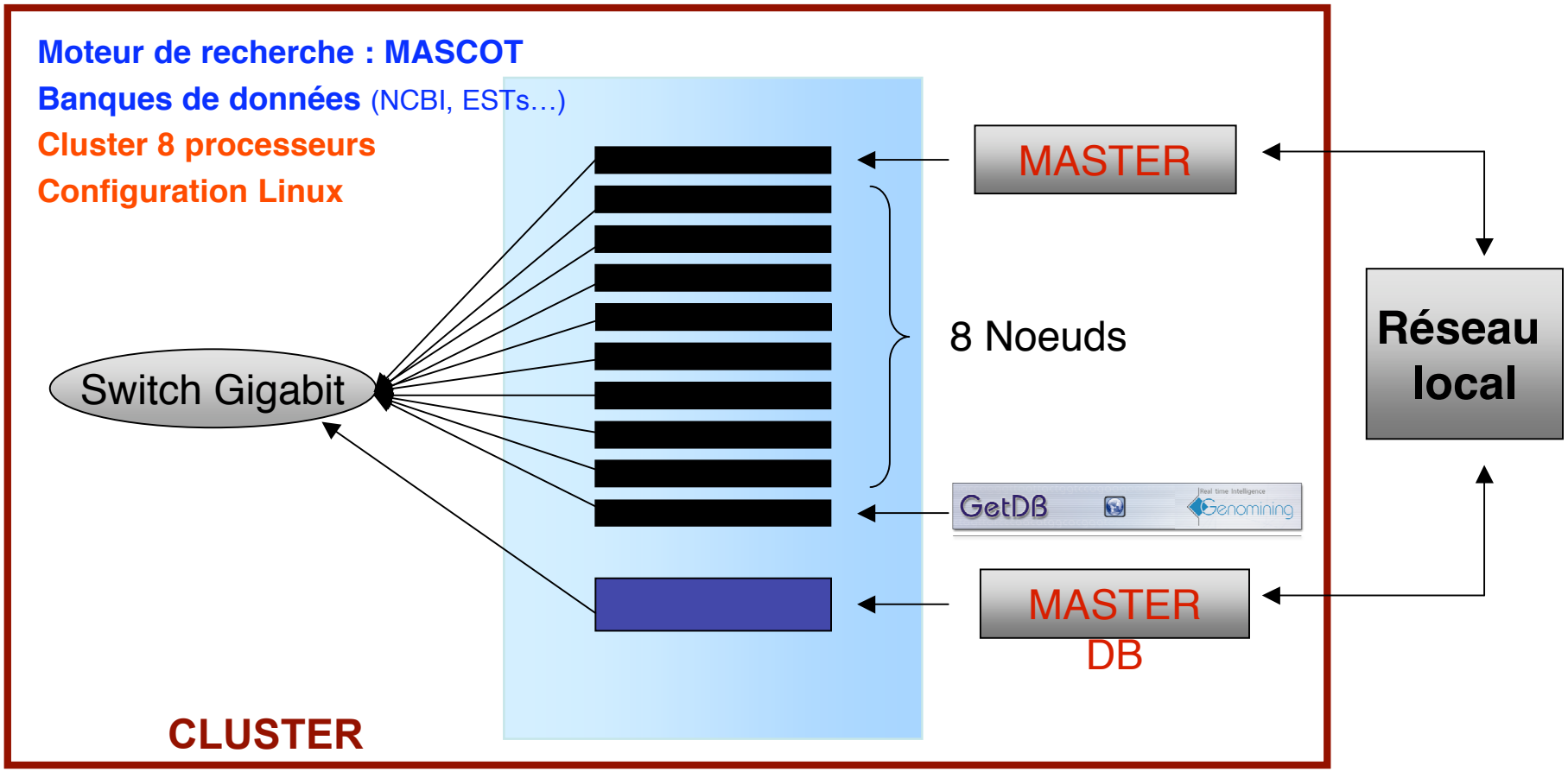
- Rapatriement et gestion des bases de données
- Interrogation des bases de données
- Gestion des résultats

Rapatriement et gestion des bases de données



- Mise à jour des banques de données
- Veille des sites sources des banques de données
- Internalisation des nouvelles versions des banques
- Mises à disposition des moteurs de recherche grâce à des Plugins adaptés

Configuration hardware *Version d'évaluation*



Master : Administration du cluster entier, gestion de Get DB…

Master DB : MASCOT, Phenyx et autres moteurs de recherche

Résultats : Infrastructure en cluster

Automatisation sur 384 spots

- Temps total d'acquisition = Temps total de recherche en banques de données
  = 2h 20 min (Sans configuration WARP Control)
- Temps d'acquisition du spectre n+1 = temps de recherche du spectre n
- Taux d'identification ≥ 70 % avec ScoreBooster
- Temps total environ 7 heures avec configuration WARP Control

Automatisation sur 1152 spots

- Temps total = environ 7 heures (sans configuration WARP control)
- Temps total avec WARP Control estimé à environ 21 heures
- Taux d'identification ≥ 70 % avec ScoreBooster

Gain de temps considérable

Possibilité d'utiliser le WARP Control

Parfaite interaction entre l'acquisition et la recherche en banques de données

Développements en cours

Software

- **Optimisation du ScoreBooster (affiner la liste des masses parasites)**
- **Optimisation de la méthode d'acquisition MS/MS**
- **Utilisation de moteurs de recherche concurrents** ***(Meta-scoring ProteinScape)***

Hardware

- ♣ **Fréquence du laser**
- ♣ **Processeur du PC contrôlant le spectromètre de masse**
- ♣ **Nombre de nœuds du cluster**

- ♣ **Technologie Flash-Disk (Collab. D. Lavenier IRISA Rennes)**

Développements en cours

Objectif:

Re-soumission automatique des spectres non identifiés de façon transparente pour les utilisateurs dès la mise à jour "significative" des banques de données

Problèmes à résoudre:

- Indiquer au système chaque mise à jour des banques
- Qu'est ce qu'une mise à jour "significative" ?
- Nombre de resoumissions versus nombre de mises à jour ?
- Risque de boucles de traitement infinies

Les partenaires impliqués dans le programme

A. Bereiziat
M. Couvet
N. Guitton

R. Lavigne
D. Ousmanou
C. Pineau

BRUKER DALTONICS

J. Glandorf
P. Rainer
P.O. Schmit
H. Thiele
P. Ufnagel

Genomining

M. Levy
W. Saurin
K. Schuller

Innova PROTEOMICS

A. Pobla
F. Bourgeon
N. Melaine

Les financements

genopole® OUEST

Inserm
Institut national de la santé et de la recherche médicale

UNIVERSITÉ DE RENNES 1

Région BRETAGNE

Développement futur (solutions software)

Extraction d'informations des données MS

- Augmentation du taux d'identification

  = Optimisation de ScoreBooster (re-calibration éprouvé des spectres MS,filtration dynamique et élimination du bruit de fond)
- Intégration de résultats issus de moteurs de recherche différent pour augmenter la confiance et l'exactitude des résultats

PMF	*Mascot*	Profound	MS-Fit

= Développement d'un Méta score

Extraction d'information des données MS/MS

- Optimisation de la méthode MS/MS d'acquisition utilisé avec WARP control :
  (L'acquisition MS/MS prend environs 5/ 6 heures pour les 30% de spectres non-identifiés)

  LIFT method evolution 1 ⟶ LIFT method evolution 2
- Intégration de résultats issus de moteurs de recherche différent pour augmenter la confiance et l'exactitude des résultats

MS/MS	*Mascot*	PHENYX	*SONAR*	SEQUEST

= Développement d'un Méta score