RENÉ PARADIS

IMPLANTATION D'UN SYSTÈME DE CUEILLETTE ET D'ANALYSE DE DONNÉES GÉNÉRÉES PAR LA PLATE-FORME PROTÉOMIQUE DU CENTRE DE RECHERCHE DU CHUL

Mémoire présenté
à la Faculté des études supérieures de l'Université Laval
dans le cadre du programme de maîtrise sur mesure en bioinformatique
pour l'obtention du grade de maître ès sciences (M.Sc.)

FACULTÉ DES ÉTUDES SUPÉRIEURES
UNIVERSITÉ LAVAL
QUÉBEC

AOÛT 2004

Résumé

Un système informatique a été développé pour centraliser la cueillette et le traitement des données provenant de la plate-forme protéomique du CHUL (Centre Hospitalier de l'Université Laval). Les informations sont saisies à l'aide d'interfaces web générées en PL/SQL, ou cueillies par des scripts Perl et des macros en VB (Visual Basic). Toutes ces données sont sauvegardées dans une base de données Oracle installée sur un serveur Unix. L'engin de recherche Mascot (Matrix Science), utilisé pour l'identification des protéines, est intégré au système. Ce dernier a grandement réduit le temps dédié à la manipulation des données et à leur analyse, augmentant ainsi la productivité de la plate-forme. La versatilité du système permet l'intégration de nouveaux algorithmes d'analyse de données provenant de divers instruments, ainsi que l'intégration des résultats d'autres plate-formes telles les micropuces, le SAGE, l'hybridation in situ et le QRT-PCR.

Avant-Propos

Au cours des années 2002 à 2004, j'ai eu l'opportunité de travailler avec une équipe très dynamique du Centre Protémique de l'Est du Québec. Ce centre est localisé au centre de recherche du Centre Hospitalier de l'Université Laval (CHUL). Pendant ces deux années et même depuis l'automne 2001, j'ai élaboré un système automatisé de cueillette et de consultation d'information provenant de la plate-forme protéomique. La bourse du CHUL attribué aux étudiants inscrits aux études supérieures m'a donné les ressources financières nécessaires pour effectuer mon projet.

Ce travail a été réalisé en bénéficiant du support de plusieurs personnes. À celles-ci, je veux les remercier grandement pour ce qu'elles ont fait. Merci à Pascal Belleau pour avoir donné de son temps dès le début du projet, à l'été 2001, et qui a contribué à l'élaboration du premier modèle conceptuel de données. Il est resté disponible pour me conseiller dans l'avancement du projet. Merci à Hugo Laliberté pour ses très belles Applets Java. Elles ont permis d'apporter une touche esthétique au système. Je remercie Benoît Hébert pour ses nombreux conseils sur le côté esthétique de la présentation de l'information et sur les corrections grammaticales du contenu. Merci à Astrid Deschênes, notre « javawoman », pour ses critiques constructives du présent mémoire, et son support au niveau de la partie Java du système. Merci à Arnaud Droit, notre toutou favori, pour ses conseils pratiques sur le contenu protéomique et informatique du travail. Je ne pourrais oublier les commentaires, critiques constructives et les rapports de problèmes de la base de données venant de l'équipe très dynamique du centre de la protéomique, merci à mes meilleurs utilisateurs beta : Tam Lehu, Isabelle Kelly, Sylvie Bourassa, Christian Lessard, Joanna Hunter, Éric Winstall et Jean-François Lemay. Je remercie mon directeur de recherche, monsieur Jean Morissette, et mon co-directeur, Docteur Guy Poirier, pour m'avoir permis d'entreprendre un tel projet.

Un merci très spécial à ma famille qui m'a supporté tout au long de mon programme de maîtrise, et m'a aidé à passer au travers de ma mauvaise expérience de mon appendicite.

En somme, cette expérience m’a fourni des ressources qui m’ont permises de grandir davantage sur le plan professionnel. Il ne faut surtout pas se décourager devant les obstacles qui entravent notre chemin. La persévérance n’existe que si l’on croit qu’elle mènera à un idéal. J’ai appris dans mes situations les moins évidentes, et je garde de cette maîtrise une expérience enrichissante et constructive. Mon désir d’élargir davantage ce que j’entreprends, ce désir d’aller au fond des choses, m’amène vers de nouveaux défis.

À mes parents, André et Suzanne

Table des matières

Liste des figures

CHAPITRE I

Introduction

1.1 Introduction à la protéomique

Plusieurs domaines scientifiques se sont développés depuis le projet de séquençage du génome humain (Science vol 291 2001), dont l'objectif était d'identifier et d'ordonner chacune des trois milliards de paires de bases de l'ADN (Acide DéoxyriboNucléique) qui composent son code génétique. La génomique est une science qui tente de comprendre la complexité du code génétique. Cette dernière s'intéresse à l'assemblage des fragments, la prédiction des gènes, l'alignement des séquences et la détection des motifs de séquences fonctionnelles. La génomique fonctionnelle est par la suite apparue pour analyser les profils d'expression des ARNm (Acide RiboNucléique messager). Elle utilise la technologie des biopuces et du SAGE (Serial Analysis of Gene Expression). Elle a permis de réaliser des projets à grande échelle d'études d'expressions géniques aidant ainsi à la compréhension des processus cellulaires et des causes des maladies à facteur génétique. La caractérisation systématique des gènes et l'annotation produisent une quantité de données estimée mille fois plus importante que le séquençage (Kearney *et al.* 2003). Mais l'étude des gènes implique aussi la compréhension et l'intégration des fonctions cellulaires à l'échelle des protéines (Aebersold *et al* .2003).

La protéomique s'intéresse aux fonctions de toutes les protéines exprimées (Tyers *et al.* 2003). C'est un défi où la biologie de l'ère post-génomique tente de comprendre comment les résultats de l'information génétique permettent d'interpréter les résultats des produits des gènes dans le temps et l'espace pour générer des fonctions biologiques (Gavin *et al.* 2002,Sanseau *et al.* 2001). Cette dernière approche entre dans un niveau de complexité plus élevé que les recherches sur le génome humain (Cook 2002, Patterson *et al.* 2003) et serait susceptible de produire jusqu'à mille fois plus de données que la génomique fonctionnelle, soit de l'ordre du petaoctet en unités de stockage informatique (Kearney *et al.* 2003). En effet, un gène peut produire plus d'un transcrit et chacun de ceux-ci peut encoder une protéine qui peut subir des modifications post-traductionnelles et se différencier davantage. De plus, les différentes interactions possibles protéine-protéine viennent multiplier les informations produites à partir d'un seul gène. Et puisque de tels phénomènes se produisent à différents moments du cycle cellulaire, une quantité stable d'ARNm pourrait ne pas corréler avec la quantité de son produit protéique dû à des effets régulatoires post-transcriptionnels et post-traductionnels (Patterson *et al.* 2003, Kearney *et al.* 2003), ce qui ajoute davantage de complexité aux fonctions protéiques.

1.2 La problématique

1.2.1 La quantité de données générées et les outils disponibles

Les domaines d'expertise en protéomique sont de plus en plus variés et les instruments de mesure utilisés, de plus en plus précis. Des instruments tels les spectromètres de masse sont utilisés pour identifier le contenu protéique des échantillons et peuvent fonctionner à haut débit produisant une quantité importante de données. Ces instruments produisent une masse d'information tellement considérable qu'il devient impossible d'analyser et valider manuellement tous ces résultats (Aebersold *et al.* 2003). La bioinformatique constitue un outil dynamique qui peut aider à modéliser les données produites par les plates-formes de recherche. En effet, celle-ci a permis de nettoyer et d'organiser rapidement l'alignement des bases lors du séquençage génétique des organismes et l'annotation de plusieurs gènes, en plus de construire des algorithmes efficaces pour la

recherche de séquences homologues (BLAST). Mais il existe peu d'outils en bioinformatique pour permettre l'analyse et l'intégration des données en protéomique. Ceci est principalement dû au fait que ce domaine est particulièrement nouveau et en rapide évolution. Cette dynamique rend difficile la définition des données clefs dans un ensemble de résultats et certains de ceux-ci n'ont de signification que dans un contexte particulier (Taylor *et al.* 2003, Aebersold *et al.* 2003).

La quantité importante de données rend non seulement leur analyse difficile, mais leur organisation, leur entreposage et la maintenance de leur cohérence représentent d'autres défis à surmonter (Goodman *et al.* 2003). Finalement, la présentation et la consultation de ces informations facilitent l'intégration des résultats en procurant des meilleurs outils d'analyse pour les scientifiques.

1.2.2 Le projet ATLAS

La plate-forme de protéomique du CHUL (Centre Hospitalier de l'Université Laval) est impliquée dans un projet financé par Génome Canada et Génome Québec dont l'objectif est d'identifier les profils d'expression des gènes et des protéines modulées par l'effet d'hormones stéroïdiennes. La plate-forme dispose d'instruments pour la préparation de gels polyacrylamides à deux dimensions à partir desquels sont prélevés des échantillons de protéines. Les échantillons sont traités par la suite par des spectromètres de masse de plusieurs types permettant d'identifier les protéines ou complexes protéiques sous différentes conditions expérimentales.

Le projet de profils d'expression au CHUL, appelé « L'Atlas des profils génomiques de l'action des stéroïdes » ou projet ATLAS, nécessitait une attention particulière en ce qui concerne la validation et l'analyse des données. Il existait déjà en protéomique des architectures d'organisation de données, comme celui de PEDRo (Proteomics Experiment Data Repository), un LIMS (Laboratory Information Management System), qui propose une idée générale de la façon de gérer les données (Taylor *et al.* 2003). D'autres systèmes, développés par des compagnies telles Proteometrix, Perkin Elmer, ABI et Micromass, sont disponibles à des prix élevés. De plus, ces programmes ne peuvent pas être modifiés et sont

dédiés aux instruments développés par les compagnies respectives (Baevis *et al.* 2002). Il n'y avait donc pas d'architecture capable de répondre spécifiquement aux besoins que nécessitait le projet Atlas compte tenu de l'instrumentation utilisée et du type de données à analyser.

Finalement, la construction d'un système en protéomique devenait nécessaire et devait remplir les fonctions suivantes :

- Cueillir automatiquement les données provenant des différents postes de travail de la plate-forme protéomique.
- Organiser et entreposer les informations dans un endroit sécuritaire.
- Permettre une validation et une analyse locale, facile et rapide des résultats.
- Permettre une consultation ergonomique des informations.
- Permettre une réanalyse rapide et facile des résultats.

1.3 Implantation de l'informatique dans la recherche biomédicale

La bioinformatique, qui est l'application de l'informatique au domaine biologique et biomédical permet de modéliser les données provenant de ces domaines. La bioinformatique a élaboré plusieurs applications et algorithmes de calcul pour la génomique et la génomique fonctionnelle(Buckingham 2003, Chait *et al.* 2001). Plusieurs entreprises, chercheurs universitaires et groupes de bioinformaticiens travaillent au développement de tels outils (Stein *et. al.,* 2002). L'accessibilité de certains d'entre eux est possible via Internet. Par exemple, les modules de Bioperl, écrits en langage Perl, sont des modules pratiques pour les calculs de points isoélectrique, la recherche de séquences homologues, l'alignement et l'arrangement graphique de séquences d'ADN (Stajich *et al.* 2002, Smith *et. al.,* 1981). Mascot (Pappin *et. al.,* 1993), MS-Fit (Clauser *et. al.,* 1999), PepTident (Wilkins *et. al.,* 1997), GPM (craig *et. al.*,2003) et Profound (Chait *et. al.,* 2000, Tang *et al.* 2000) sont des logiciels servant à l'identification de protéines d'échantillons générés par spectrométrie de masse. SeQuest (Eng *et. al.*, 1994), est un logiciel d'identification de protéines disponible qu'avec l'achat d'un spectromètre de masse. D'autre part, certaines compagnies rendent accessible, via Internet, l'utilisation de leurs logiciels pour l'identification protéique à partir d'empreintes de masses peptidiques. Plusieurs banques de données sont disponibles sur

Internet et contiennent le regroupement des annotations géniques, protéomiques ainsi que la littérature de la communauté scientifique. Les perçées en bioinformatique ont favorisé le développement de nouveaux modèles probabilistes et mathématiques tant au niveau de la génomique, de la protéomique que de la modélisation tridimentionnelle de protéines (Claverie *et al.*).

Ce mémoire décrit toutes les étapes suivies lors du développement d'un système de cueillette et d'analyse de données répondant aux attentes de la plate-forme protéomique du centre de recherche du CHUL.

CHAPITRE II

MATÉRIEL ET MÉTHODE

2.1 Description de l'instrumentation

2.1.1 Couteau à gel

Lorsque les gels à deux dimensions sont créés, un numériseur optique prend une image, en noir et blanc, de chacun des gels. L'apparition des taches du gel au numériseur est possible grâce à une coloration au Sypro Ruby, qui est révélée par un faisceau ultra violet. Ces images sont traitées par le logiciel Pdquest fourni par BioRad. Celui-ci analyse les zones foncées sur le gel, les taches de protéines et calcule leurs coordonnées, leur intensité et délimite ces zones par le calcul de courbes gaussiennes. Chaque tache possède un numéro d'identification unique pour le gel. La normalisation des taches est faite par le logiciel en superposant plusieurs réplicats de gels qui ont sensiblement les mêmes coordonnées et intensités. La sélection des taches à prélever se fait de façon automatique ou manuelle. Cette sélection est ensuite transmise au couteau à gel (spot cutter, BioRad) qui effectue les prélèvements. Les morceaux de gels sont ensuite déposés dans les puits de plaques à 96 puits. Puisque la normalisation des gels est possible, un puits peut contenir plusieurs prélèvements provenant de différents gels qui appartiennent à un même

répliquat. La plaque est finalement envoyée à l'appareil de digestion.

Figure 2.1 Couteau à gel communément appelé « spot cutter » de BioRad.

2.1.2 Appareil de digestion Mass Prep

Les échantillons provenant du couteau à gel doivent subir une série de transformations avant d'être analysés par les appareils de spectrométrie de masse. Chaque puits de la plaque contient un mélange de protéines dont les séquences en acides aminés sont digérées ou coupées en petits fragments de faibles poids moléculaires par une protéase du pancréas, la trypsine. Celle-ci reconnaît certains acides aminés. L'arginine et la lysine sont les points de coupure de la protéase sauf lorsque celles-ci suivent une proline. L'appareil qui digère les échantillons se nomme un MassPrep, fourni par Micromass. Le logiciel Winprobe II, fourni avec cet appareil, utilise un fichier contenant une suite de paramètres et d'instructions permettant la préparation des échantillons à la digestion. Cette suite de paramètres est appelée méthode de digestion. Elle contient des informations telles les solutions d'extraction, les tampons, les alkylants, les enzymes et les solutions de nettoyage. Les manipulations sont faites par un bras multi-fonctionnel à pipettes. Les échantillons digérés sont finalement transférés dans une autre plaque de 96 puits entreposée par la suite à 4° Celcius avant l’analyse par spectrométrie de masse.

Figure 2.2 Appareil de digestion MassPrep de Micromass

2.1.3 Matrix Assisted Laser Desorption-Ionisation Time-Of-Flight (MALDI-TOF)

L'empreinte de masse peptidique est la technique la plus accessible pour l'identification protéique en protéomique (Lester et al. 2002). Le MALDI-TOF est un spectromètre de masse qui génère des spectres permettant de déterminer la masse des peptides résultant de la digestion des échantillons. Le mélange contenant la digestion est transformé sous forme de cristal et déposé sur une plaque de métal. Dans la chambre du MALDI, un rayon laser percute le cristal et des fragments ionisés sont propulsés hors de la plaque et sont accélérés à haut voltage. Ils volent ensuite dans un tunnel sans champs électrique où est calculé leurs temps de passage (Time Of Flight). Les temps de passage sont convertis en masses en calculant le rapport de la masse (m) sur la charge du fragment (z) d'où (m/z). La quantité relative est enregistrée pour chaque masse. Cette quantité est normalisée à partir de la masse la plus abondante. En effet, un échantillon bombardé au laser contiendra plusieurs fragments de même masse, car la trypsine coupe toujours au même

endroit les multiples copies de la protéine. Cette signature de l'enzyme produit les spectres de masses, générés à l'aide du logiciel Voyager de ABI (Applied Biosystems). Ce logiciel exécute une suite d'instructions et de paramètres permettant à l'appareil de produire des spectres pour chacun des puits de la plaque. Cette suite d'instructions est appelée une séquence de MALDI. Les fichiers de spectre sont de format binaire .DAT et sont lisibles par le logiciel Explorer de ABI. Ces fichiers sont convertis en format texte, format compatible avec les algorithmes de recherche utilisés pour l'identification protéique. Les fichiers sont analysés par des algorithmes qui comparent les masses expérimentales à d'autres masses digérées théoriquement, venant de banques de protéines, ce qui permet d'identifier le contenu protéique des échantillons (Blackstock et al. 2000). Cette méthode d'analyse est une première étape dans l'identification protéique et révèle des limitations lorsque les fragments peptidiques sont en faibles concentrations (Blackstock et al. 2000). L'approche de spectrométrie de masse en tandem (MS/MS) est plus sensible et est entre autre utilisée pour contrer ce problème.

a)

b)

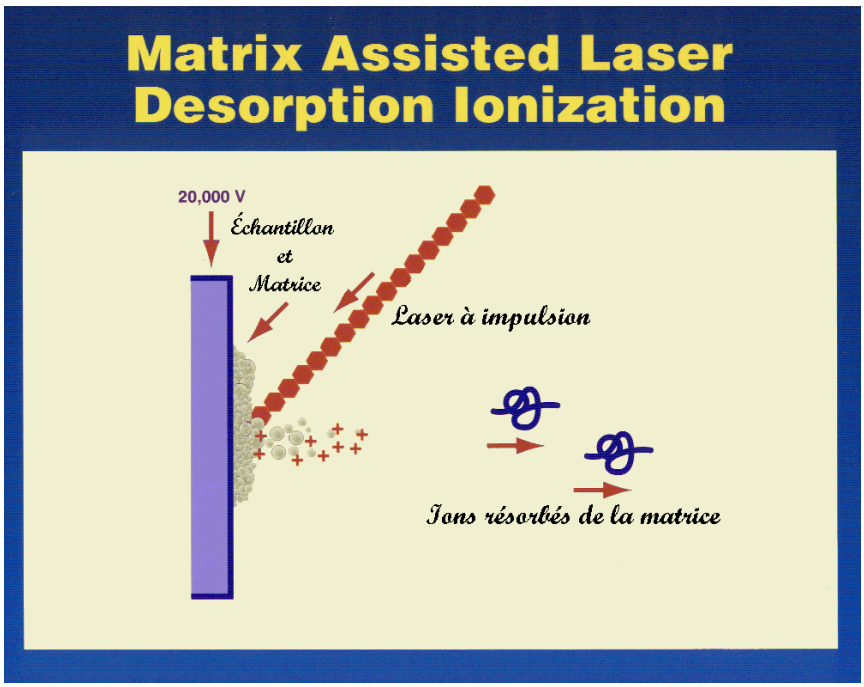


Figure 2.3 a) Spectromètre de masse de type (MALDI-TOF)de ABI. L'instrument est géré par le poste de travail. Une caméra à l'intérieur de l'instrument permet de visualiser l'échantillon ciblé par le laser. b)Principe d'ionisation d'un spectromètre de masse de type (MALDI-TOF).

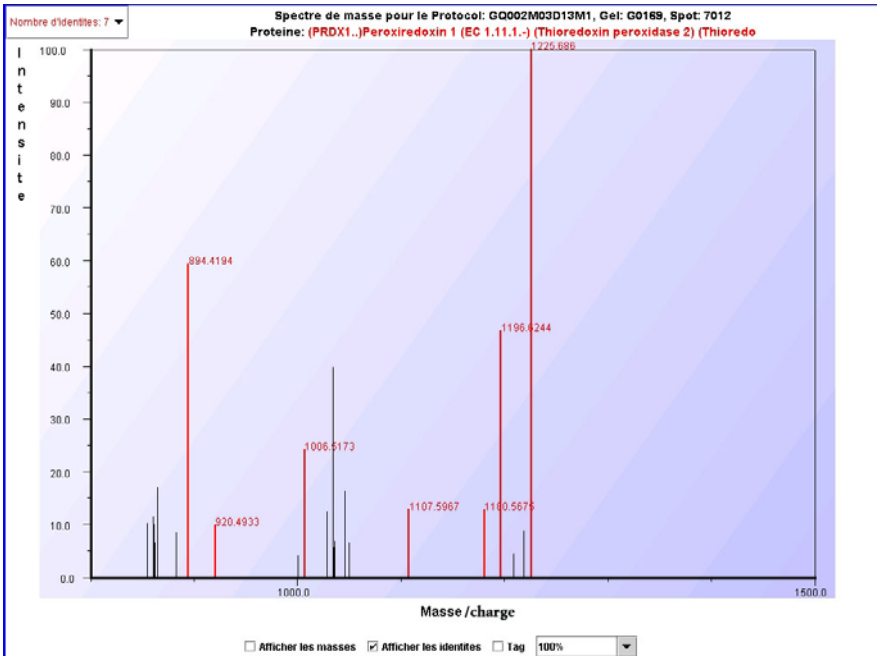
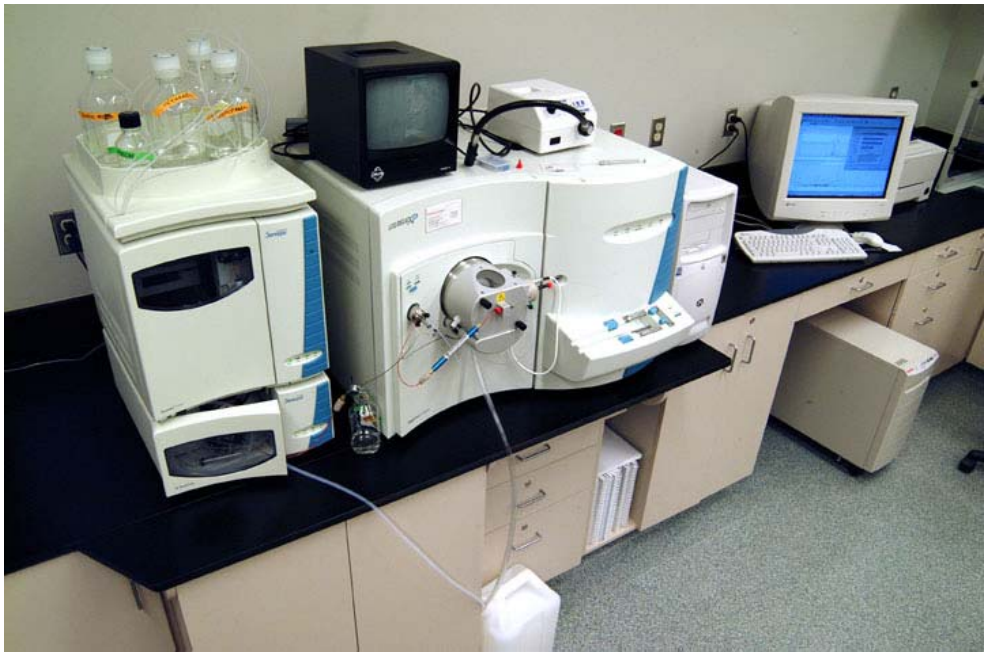


Figure 2.4 Spectre de masse MALDI-TOF. Chacune des masses est représentée en abscisses, et sa quantification relative, en ordonnées.

2.1.4 Spectromètre de masse en tandem à trappe ionique (LC - MS/MS)

Conçu par la compagnie ThermoFinnigan, cet appareil fait la séparation de ses fragments à l'aide d'un chromatographe en phase liquide (HPLC) (Blackstock et al. 2000). Les fragments se séparent à l'intérieur d'une colonne de chromatographie selon leur hydrophobicité en avançant dans un flux constant d'un gradient d'acétonitrile. Par la suite, ils entrent dans une chambre à ions remplie d'un gaz neutre, et le spectromètre génère un premier spectre de masse. Ce spectre ressemble à celui du MALDI-TOF (voir figure 2.2). Finalement, chaque masse est isolée à l'intérieur de la trappe ionique par des champs électriques variables et fragmentée par des collisions avec les atomes de gaz, ce que l'on appelle la dissociation collisionelle (collision induced dissociation). Les liens peptidiques sont rompus dans cette étape. Le produit final donne des sous-fragments de diverses longueurs ayant des masses différentes. Les sous-fragments les plus courts représentent un acide aminé. Puisque chacune des masses des acides aminés est connue et spécifique, la reconstitution des sous-fragments à leur masse originale permet de déterminer la séquence peptidique du fragment initial (voir figure 2.3). L'étape est répétée pour les autres fragments sélectionnés du spectre et pour tous les échantillons d'une séquence LCQ. Un fichier binaire d'une taille de 20 à 40 méga-octets est produit pour chaque échantillon. Ces fichiers ne sont lisibles que par le logiciel Xcalibur de la compagnie ThermoFinnigan.

a)



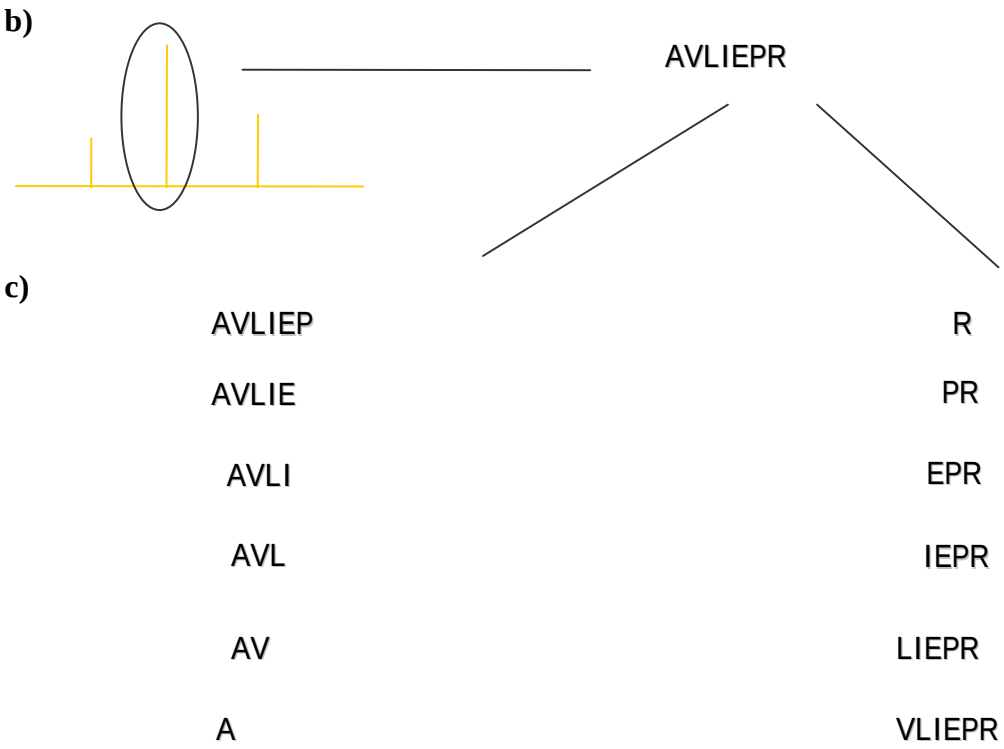


Figure 2.5 a)Spectromètre de masse en tandem (LCQ) de ThermoFinnigan.

b) Le premier spectre de masse est construit au moment ou l'échantillon entre dans la chambre à ionisation. Ces fragments sont obtenus suite à une digestion tryptique. Certaines masses du spectre sont sélectionnées une à une, et sont fragmentées de nouveau par bombardement particulaire.

c) Combinaison de sous fragments obtenus par le bombardement. L'agencement des différentes combinaisons des sous-fragments permet de reconstituer le fragment initial. L'ordre des acides aminés du fragment initial est alors déterminé.

2.2 Les Serveurs UNIX

2.2.1 Système d'exploitation UNIX

Le choix du matériel informatique utilisé pour le traitement des données est important. En effet, la sécurité des informations, la fiabilité des sauvegardes et la rapidité à leur accès sont des facteurs critiques à considérer. Le système d'exploitation UNIX est fiable et un des plus versatile de tous les systèmes d'exploitation. Il s'agit d'un logiciel de base destiné à commander l'exécution des programmes en assurant la gestion des travaux, les opérations d'entrée/sortie sur les périphériques, l'affectation des ressources aux différents processus, l'accès aux bibliothèques de programmes et aux fichiers ainsi que la comptabilité des travaux. Sa capacité à gérer systématiquement les exécutions de programmes et les connexions d'utilisateurs a fait en sorte qu'il soit beaucoup utilisé dans l'industrie. Il gère également les ressources de l'ordinateur et permet de communiquer avec d'autres machines. Plusieurs fabricants d'ordinateurs utilisent UNIX comme système d'exploitation et développent des interfaces graphiques pour en faciliter l'usage. La plate-forme de bioinformatique utilise Solaris version 2.8, le système UNIX distribué par Sun microsystems. Les programmes tels Oracle, Mascot, R, ImageMagick tournent sur Solaris et sont utilisés dans l'exploitation du traitement des données de protéomique.

2.2.2 Serveurs

Le choix d'un serveur a été fait en tenant compte des facteurs tels la robustesse du matériel, sa stabilité, sa performance, et sa facilité d'entretien. Le département de bioinformatique possède un serveur SUN 3800 à 6 processeurs divisé en 2 domaines : Tahiti et Tahaa. Tahiti, qui est un domaine composé de deux processeurs ultrasparc III tournant à 750 Mhz, est le serveur de base de données. Afin de maximiser la confidentialité et la sécurité des données, ce domaine n'est accessible que par le sous-réseau interne du CHUL. Le domaine Tahaa comprend 4 processeurs de 750 Mhz et est destiné aux traitements de données. Il est accessible au sous-réseau interne et au réseau public. Un troisième serveur,

plus petit appelé Maupiti, héberge le module IAS (9i Application Server) qui permet un accès à partir du réseau public de la base de données.

2.2.3 Sauvegarde des données

Les données sont sauvegardées sur des disques de haute performance configurés en RAID 5 (Redundant Array of Independant Disks) assurant ainsi une grande rapidité d'accès aux données et une grande sécurité. De plus, des copies de sécurité des données sont effectuées à chaque nuit pour la récupération de fichiers à une date donnée. Finalement, les données sont archivées à des périodes déterminées sur des cassettes d'archivage. Les cassettes sont dupliquées et une copie quitte l'édifice pour aller dans un endroit sécurisé; l'autre reste au centre de recherche du CHUL. Ces données peuvent toujours être remises sur disques au besoin. Les usagers se connectent via des terminaux tant aux serveurs du sous-réseau interne qu'aux serveurs du réseau public. Une quatrième machine à quatre processeurs de 400 Mhz, Borabora, sert de serveur de terminaux. Cette machine gère plus de quarante terminaux. Les serveurs sont reliés entre eux par un lien optique de 1 Gigaoctet, alors que les terminaux sont branchés par des câbles torsadés de 100 Mégaoctets.

2.3 Base de données

2.3.1 Définition

Une base de données est un ensemble d'informations structurées de façon logique pour en faciliter leur exploitation. Les données sont stockées sur un support informatique et sont accessibles par plusieurs utilisateurs à la fois.

L'architecture d'une base de données regroupe les informations aux caractéristiques semblables dans des entités nommées tables. Celles-ci sont généralement inter reliées par des références pour modéliser le type de dépendance qu'elles ont entre elles. Ce concept est appelé modèle relationnel. Un autre modèle, comme celui de l'objet-relationnel ou objet, utilise le concept d'encapsulation des données en une entité objet possédant des propriétés distinctes. Ce dernier modèle s'inspire du concept objet utilisé dans des langages de programmation orienté-objet tel le C++ et le Java, et donc inclut les propriétés de la

programmation orienté-objet. Toutes les tables du système ont été créées selon le modèle relationnel.

2.3.2 Système de gestion de base de données

Le système de gestion de base de données ou SGBD gère la saisie des données, la modification, la suppression et la consultation. Le SGBD de Oracle 9i a été choisi pour sa sécurité, sa stabilité et sa compatibilité avec plusieurs langages de programmation. Oracle utilise la norme SQL (Structured Query Language) comme langage d'interrogation avec la base de données. L'insertion et la consultation des données de la base se font au moyen d'exécution de requêtes SQL qui sont interprétées par le DML (Data Manipulation Language) d'Oracle. L'utilisation d'interfaces générées en HTML accessibles via le réseau Internet facilite l'interrogation des bases de données et offre une meilleure présentation des informations.

2.4 Langages informatiques

2.4.1 Perl

Le langage Perl est une abréviation de l'acronyme (Practical Extraction and Report Language). Ce langage a été conçu pour extraire facilement les informations contenues dans des fichiers de format texte. Cependant, Perl est beaucoup plus qu'un simple extracteur d'informations. Il offre aussi une interface polyvalente pour communiquer avec la base de données Oracle. Par exemple, les informations extraites des fichiers textes sont insérées dans la base via un module de connexion à la base Murin. Ce module utilise le protocole FTP (File Transfert Protocole) permettant une connexion à distance sur un poste de travail et la récupération des nouveaux fichiers à traiter. Ce langage peut aussi générer des fichiers de format PDF (Portable Document Format), format grandement utilisé sur différentes plates-formes informatiques. Perl est utilisé dans la majorité des scripts servant au traitement et la récupération des données. Il a été préféré aux langages C, C++ ou JAVA. Ces derniers sont efficaces dans le développement des algorithmes de calculs, mais ils ne sont pas optimisés

pour l'analyse syntaxique. Les besoins dans l'élaboration du système de la plate-forme protéomique permettaient d'exploiter la force de ce langage.

2.4.2 PL/SQL

PL/SQL (Procedural Language/Structured Query Language) est le langage compatible avec les modules d'Oracle et fut développé par cette compagnie. Toutes les interfaces HTML sont générées par le PL/SQL et sont ensuite affichées à l'écran via la passerelle d'Oracle. La base de données est directement interrogée et les données sont manipulées aisément au moyen de ce langage. PL/SQL est compilé dans Oracle et bénéficie de toute la sécurité attribuée par ce dernier. Le code de PL/SQL est soit compilé sous forme de procédures, de fonctions ou de déclencheurs. Plus de 90% des programmes PL/SQL sont compilés sous forme de procédures. Plusieurs arguments typés en entrée peuvent être passés lors de l'appel de procédures. Une procédure peut retourner plusieurs valeurs typées. Les fonctions peuvent aussi prendre plusieurs arguments en entrée mais un seul en sortie. Les déclencheurs sont des programmes s'exécutant selon certaines situations, comme à la suite d'une insertion d'une entrée dans une table. Ces différentes fonctionnalités du PL/SQL facilitent grandement la manipulation des données de la base, ce qui explique pourquoi le PL/SQL fut choisi pour le développement de programmes servant à la gestion des données protéomiques.

2.4.3 HTML

Le HTML (Hyper Texte Markup Language) est le langage de description de documents structurés à l'aide de balises et est utilisé pour afficher l'information sur les sites Internet. Puisque la consultation des données du système se fait via un portail, le HTML est l'outil utilisé pour développer les pages affichant les données à l'écran tant sous forme de textes, de fichiers sonores, que des fichiers images. Ce sont les programmes écrits en langage PL/SQL qui génèrent les lignes de HTML interprétées par le fureteur du poste de travail de l'usager.

2.4.4 Java

Java est un langage de troisième génération entièrement orienté-objet, c'est-à-dire qu'il utilise le principe de l'encapsulation pour la manipulation et la création de données. L'encapsulation est l'opération qui consiste à regrouper, dans des entités distinctes (objets), les données et les procédures (méthodes) qui les manipulent (www.grandictionnaire.com). Les objets encapsulés doivent alors être traités par ces méthodes seulement. Java est utilisé dans la construction de sites Internet pour rendre leurs contenus dynamiques et interactifs par l'implémentation de graphiques, d'images et autres effets visuels. Deux types d'approches de conception de programmes, applets et servlets, ont été utilisés pour rendre interactives certaines parties du système protéomique.

Les applets

Les applets sont des programmes en Java intégrables dans un document HTML (Deitel & Deitel 2002). Les programmes sont compilés et autonomes et sont téléchargés d'un serveur pour s'exécuter du côté client via le fureteur. Certaines pages HTML utilisées pour l'interrogation de la base de données incluent des applets qui construisent et affichent des graphiques et des images.

Les servlets

Le servlet est un module de code Java s'exécutant sur le serveur d'application et non sur le poste client (www.Java.sun.com). Ceci offre une sécurité non négligeable lorsque le servlet doit interroger une base de données, car aucun nom usager et aucun mot de passe ne voyage sur le réseau. C'est le servlet qui communique avec la base et qui retourne les informations au fureteur du client.

Notons que les graphiques auraient pu être faits en Perl, ou en Javascript, mais Java a été choisi pour sa portabilité et pour sa présence de modules conçus pour générer des graphiques. La familiarité du langage a aussi motivé ce choix.

2.4.5 Javascript

Le Javascript est un langage de programmation utilisé dans les pages de sites Internet afin de les rendre dynamiques. Un script peut être inséré dans une page de HTML pour, par exemple, valider les données d'un formulaire et retourner des messages en cas d'erreurs. L'utilisation des ressources de serveurs pour la validation de ces formulaires peut alors être évitée, optimisant ainsi les ressources globales du système. Le menu de la plate-forme protéomique a été développé en JavaScript. Il offre des options telles que le changement des couleurs lorsque la souris est sur une zone du menu, ou l'affichage d'une description déroulante lorsque la souris pointe sur le titre d'un commentaire. Ces options rendent alors les pages de HTML dynamiques.

Il est utile de rappeler que Javascript et Java sont des langages bien distincts et que la similarité de leurs noms ne repose que sur un motif commercial. Cependant, ces deux langages sont complémentaires. JavaScript contrôle les comportements du fureteur et son contenu, mais ne peut générer de graphiques ou établir une connexion réseau alors que Java ne contrôle pas le fureteur dans son ensemble, mais génère des graphiques, établit des connexions réseau et gère des processus multitâches. JavaScript, sur le poste de travail client, interagit et contrôle des applets Java intégrés dans une page web (Flanagan 2002).

2.4.6 UML

L'UML (Unified Modeling Language) est un langage standardisé de modélisation. Ce n'est pas simplement une notation pour créer des diagrammes, mais plutôt un langage complet permettant l'acquisition de connaissances (sémantique) et l'expression de savoir (syntaxe) afin de faciliter la communication. Ce langage résulte de l'unification des meilleures pratiques industrielles utilisées dans la conception de projets soit : les principes, les techniques, les méthodes et les outils (Si Albir 1998). C'est pourquoi la norme UML est largement utilisée dans les domaines publics et privés, et qu'il a été choisi pour conceptualiser le modèle relationnel du système protéomique.

2.5 Mascot et Mascot Daemon

L'automatisation d'un système d'analyse efficace de données protéomiques exigeait que les programmes d'identification des protéines soient exécutés sur des serveurs locaux et non sur des serveurs distants. Ceci permet entre autre, une plus grande liberté dans la manipulation des données entrantes et sortantes, le contrôle de la confidentialité des résultats, la possibilité de modifier l'environnement de l'algorithme de recherche et la liberté d'utiliser les options de notre choix. Ainsi, il fallait un logiciel capable de procurer tous ces éléments, en plus d'être performant, peu coûteux, compatible UNIX (Solaris) et permettant de bénéficier d'un support technique adéquat. Le logiciel Mascot de Matrixscience, est un algorithme de recherche performant d'identification de protéines qui répond le mieux à ces besoins parmi d'autres algorithmes disponibles sous licence (Pappin *et al.* 1993, Pappin *et. al.*, 1999). Cet algorithme compare des masses peptidiques obtenues expérimentalement avec celles calculées théoriquement à partir des séquences protéiques ou nucléotidiques (Choudhary *et al.* 2001,Kuster *et al.* 2001) d'une banque de protéines locales. L'algorithme Mowse (MOlecular Weight SEarch) identifie les protéines de la banque qui ont obtenu les meilleurs résultats de comparaison entre les masses théoriques et expérimentales. Chaque résultat est accompagné d'une probabilité, dont le calcul est basé sur des paramètres de recherche choisis par l'utilisateur, sur le nombre de masses identifiées ainsi que sur les longueurs et les modifications chimiques. Cette probabilité représente la possibilité d'obtenir l'identification de façon aléatoire de cette protéine. Plus sa valeur est basse, moins il y a de risques que cette identification soit fausse. Mascot tourne sur un serveur UNIX.

Le démon de Mascot (Mascot Deamon) est un programme installé sur les postes de travail Windows et conçu pour soumettre les fichiers au programme Mascot. Le démon prépare un fichier de format MIME (Multipurpose Internet Mail Extensions) et envoie ce fichier par protocole HTTP (HyperText Transfer Protocol) au serveur URL (Uniform Resource Locator) de Mascot. De plus, le démon se connecte à la base Murin via un ODBC (Open DataBase Connectivity), et envoie après chaque recherche les principaux résultats dans les tables de Mascot : Mascot_daemon_results,
Mascot_daemon_files, Mascot_daemon_parameters et Mascot_daemon_tasks.

2.6 Banques de données protéomiques

2.6.1 Les différentes banques

Il existe plusieurs banques de protéines annotées et disponibles gratuitement sur Internet (Baxevanis *et. al.,* 2001). Ces banques de données sont organisées dans une forme standard de façon à pouvoir retrouver rapidement les informations désirées et alléger le temps de la recherche. L'annotation d'une protéine comprenant son nom et ses numéros d'identification, ainsi que sa séquence en acides aminés est souvent conservée dans un fichier sous un format appelé FASTA. Le logiciel d'analyse peptidique consulte ce fichier pour identifier les protéines contenues dans l'échantillon. Le format FASTA commence avec une seule ligne d'en-tête décrivant la séquence suivie par plusieurs lignes contenant la séquence peptidique. La ligne de description commence obligatoirement par le symbole « > » et les informations descriptives sont séparées par le caractère spécial « | ». La première information identifie la banque protéique à partir de laquelle le fichier a été construit. Dans l'exemple ci-dessous représentant une entrée de format FASTA, l'information « gi » (pour GenInfo) identifie la banque de NCBI (National Center for Biotechnology Information). Cette information est obligatoirement suivie par l'identifiant unique de la séquence pour la banque désignée. Cet identifiant est représenté dans cet exemple par « 532319 ». Si l'information de la protéine a été récupérée d'une autre banque, l'identifiant de la banque source et de la protéine sont inscrits à la suite. Dans l'exemple, ces deux données sont représentés respectivement par « pir » et par « TVFV2E ». Le nom de la protéine est toujours la dernière information fournie dans la ligne de l'en-tête. Cette information est représentée ci-dessous par « TVFV2E envelope protein ». Les lignes ont une longueur limite de 80 caractères.

```
>gi|532319|pir|TVFV2E|TVFV2E envelope protein
ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVVHCTNLMNTTVTTGLLLNGSYSENRT
QIWQKHRTSNDSALILLNKHYNLTVTCKRPGNKTVLPVTIMAGLVFHSQKYNLRLRQAWC
HFPSNWKGAWKEVKEEIVNLPKERYRGTNDPKRIFFQRQWGDPETANLWFNCHGEFFYCK
MDWFLNYLNNLTVDADHNECKNTSGTKSGNKRAPGPCVQRTYVACHIRSVIIWLETISKK
TYAPPREGHLECTSTVTGMTVELNYIPKNRTNVTLSPQIESIWAAELDRYKLVEITPIGF
APTEVRRYTGGHERQKRVPFVXXXXXXXXXXXXXXXXXXXXXXVQSQHLLAGILQQQKNL
LAAVEAQQQMLKLTIWGVK
```

Exemple d'une entrée protéique en format FASTA.

Les banques de données de protéines peuvent parfois contenir des données redondantes ou des séquences incomplètes. Parmi les banques de données disponibles, les principales sont:

Swiss-Prot

SWISS-Prot est la banque contenant probablement le moins de protéines parmi les banques générales qui regroupent l'ensemble des entrées connues (Bairoch *et. al.,* 2000). Elle est la plus rigoureuse des banques disponibles et elle regroupe le maximum d'informations sur chaque protéine annotée telles leurs isoformes, leurs modifications post-traductionnelles (Jung *et. al.,* 2001) ainsi que de nombreuses références vers d'autres banques de données. (http://us.expasy.org/sprot/)

TrEMBL

TrEMBL (transcrits de EMBL[1]) est un supplément à la banque de Swiss-Prot. Elle contient les séquences annotées par ordinateur suite à la transcription de toutes les entrées nucléotidiques de la banque EMBL non encore intégrées à la banque de Swiss-Prot. Ces séquences protéiques complètes viennent de l'assemblage de EST[2] (partie codante d'un gène)ou de CDS[3] (séquence d'ADN codant obtenue à partir de l'ARN). TrEMBL donne accès à ces séquences à la communauté scientifique même si leur annotation n'est pas suffisante pour être transférée vers la banque Swiss-Prot (http://us.expasy.org/sprot/).

TrEMBL-NEW

Cette banque contient les nouvelles entrées de la banque TrEMBL. Ces dernières sont en attente d'annotation avant d'être transférées dans TrEMBL. Leur séquence peut être

[1] European Molecular Biology Laboratory. TrEMBL est la transcription en acides aminés des séquence provenant du EMBL.
[2] Coding DNA Sequence : séquence codante d'ADN obtenue par une transcription inverse de son arn (ADNc).
[3] Expressed Sequence Tag : séquence partielle d'ADN codant obtenue à partir de clones d'ARNm.

partielle. L'analyse de résultats provenant de cette banque doit être prise avec réserve. (http://us.expasy.org/sprot/)

PIR

La banque PIR (Proteine Information Ressources) contient des séquences protéiques provenant entre autre de la banque PSD (Protein Sequence Database) et couvre l'ensemble des taxons étudiés (Wu *et. al.*, 2003). Elle tient son origine du projet de « *Atlas of Protein Sequence and Structure »* (1965-1978) sous la direction de Margaret Dayhoff. (http://pir.georgetown.edu)

Genepept

Genepept est une banque dérivée du site de NCBI contenant la transcription des CDS en acides aminés. Cette transcription est accomplie par le programme Blast développé par le centre NCBI (Benson *et al.* 1998).

NCBI

La banque du NCBI regroupe des séquences protéiques provenant en grande partie de la traduction des séquences d'ADN de EST ou de contigs[4]. De plus elle intègre aussi les entrées des banques précédentes dont Genepept, Swiss-Prot, TrEMBL, PIR et PDB. Chaque entrée offre une description complète de la protéine et une liste importante de références sur celle-ci (http://www.ncbi.nlm.nih.gov/).

[4] Séquence d'ADN obtenue par l'alignement de séquences partageant des parties communes.

Leshmania major

Les banques énumérées précédemment ne contiennent pas toutes les protéines existantes. Par exemple, la banque de protéines de *Leishmania major*, un organisme séquencé par le Sanger institute (http://www.sanger.ac.uk/), possède toutes les protéines répertoriées de cet organisme.

Les banques de données décrites précédemment sont dans le format FASTA. Le nombre total des entrées protéiques présentes dans les différentes banques a été estimé en sommant le nombre d'entrées distinctes des tables publiques (voir section 2.6.2). Ce nombre dépasse le million et il est modifié à la hausse suite à la mise à jour mensuelle de notre banque de données.

2.6.2 Mise à jour des banques de données

Les banques de données mentionnées ci-haut sont mises à jour hebdomadairement sur les sites qui les hébergent. Par conséquent, la communauté scientifique doit télécharger régulièrement ces banques pour être à jour. La plate-forme protéomique possède les banques de Swiss-Prot, Trembl,Trembl- new, PIR et NCBI et les met à jour mensuellement. Ces banques sont intégrées dans l'environnement de Mascot et sont utilisées pour traiter les fichiers générés par les spectromètres. Il est tout à fait possible d'intégrer de nouvelles banques de protéines à l'engin de recherche Mascot afin de couvrir le plus d'espèces possibles.

2.7 Autres Programmes

2.7.1 ImageMagick

ImageMagick est un programme utilisé pour le traitement des images et compatible avec l'environnement Unix. Il est lancé par une ligne de commande Unix et est donc exploitable à l'intérieur de programmes ou scripts Unix. Lorsque les images des gels à deux dimensions sont générées, leur format doit être modifié pour en réduire la taille et faciliter la

visualisation. **Image2d.pl** est le script Perl qui a été développé pour appeler automatiquement ImageMagick.

2.7.2 Librairie R

La librairie R contient une suite de programmes destinés aux calculs des données et à l'affichage de graphiques de données statistiques. Plusieurs outils statistiques y sont implémentés. R est appelé par des scripts Perl pour le traitement de données protéomiques. Il est utilisé ensuite pour déterminer statistiquement si l'intensité des taches entre deux traitements est différente. Un test t de Student est effectué et la valeur de p (seuil de signification) résultant de ce test est insérée dans la base.

CHAPITRE III

MISE EN PLACE DU SYSTÈME DE CUEILLETTE ET D'ANALYSE

3.1 Les besoins et les données pertinentes de la protéomique

L'élaboration d'un système de cueillette de données cohérent nécessite la définition des besoins à combler, et la détermination des types de données qui contribueront à combler ces besoins. Les besoins principaux de départ étaient d'organiser et de regrouper les données des échantillons produits par la plate-forme des gels à deux dimensions, de faire suivre ces informations jusqu'à la plate-forme des spectromètres de masse, de récupérer les résultats de la recherche d'identification protéique et de faciliter leur analyse par des outils graphiques et des tableaux conviviaux. Tout cela devrait permettre également d'augmenter le rendement de la production en diminuant le temps d'analyse des échantillons et l'importante quantité de papier. Le système devait être fonctionnel, autant pour des données provenant du projet ATLAS que pour les données provenant d'autres projets. Les deux sous-sections énumèrent et expliquent les besoins qui ont été déterminants dans la construction du système. Les autres informations pertinentes seront expliquées dans la présentation du schéma relationnel de la base de données.

3.1.1 Les gels à deux dimensions

Les prélèvements pour le projet ATLAS sont constitués de protéines extraites des morceaux de tissu de souris. Ces protéines sont d'abord séparées par électrophorèse sur gels. Cette séparation se fait en deux dimensions : premièrement en fonction du point isoélectrique en faisant varier le pH sur un gel d'acrylamide, puis par la masse en les faisant migrer sur un gel de même composition traversé par un courant électrique. La figure 3.1 montre une image d'un gel d'acrylamide. En abscisses, les protéines sont séparées selon leur point isoélectrique, en ordonnées, selon leur masse. Le point isoélectrique est le pH où la protéine est chimiquement neutre et appelée switterion. Chaque tache contient des protéines de même type ou de types différents. En effet, certaines protéines aux fonctions distinctes peuvent avoir des propriétés physico-chimiques similaires et vont se retrouver au même endroit sur le gel. Il se produit aussi l'effet de co-migration : c'est-à-dire que des protéines s'associent par affinité et ne se séparent pas lors de la migration.

Pour connaître l'identité d'une tache, il faut la prélever en utilisant un couteau à gels « spot cutter ». La tache prélevée constitue ce qu'on appelle l'échantillon. Elle est une zone définie sur le gel indiquant la présence de matière protéique Plusieurs taches peuvent être prélevées. Il est rare qu'un gel de tissu soit fait en simplet. Les expériences sont effectuées avec un ou deux contrôles et accompagnées de plusieurs traitements avec trois à quatre exemplaires de gels par traitement. Le logiciel Pdquest permet de superposer les gels afin que leurs taches puissent être comparées entre elles. Ce processus s'appelle la normalisation des gels. Il y a évidemment des gels qui ne peuvent pas être superposés, dû à leur mauvaise qualité. La qualité d'un gel est évaluée selon la présence de taches distinctes et de leur répartition uniforme sur la surface du gel. Il arrive que toutes les taches d'un gel normalisé ne puissent pas être comparées, car leur localisation diffère entre les gels. Une association inter-gel est alors impossible. Chaque tache associée possède un numéro d'identification identique pour tous les gels normalisés. Souvent, un échantillon provient de plusieurs gels d'un même traitement. Le logiciel Pdquest calcule plusieurs informations telles les courbes gaussiennes de la densité optique (D.O.), les coordonnées et l'intensité de chaque tache. Ces données sont générées à partir d'une image de gel en format TIFF (Tag Image File Format).

Les images, ainsi que leurs données devaient être stockées afin de pouvoir les visualiser sur écran. Cette visualisation facilite l'analyse et la validation des résultats obtenus sur les taches.

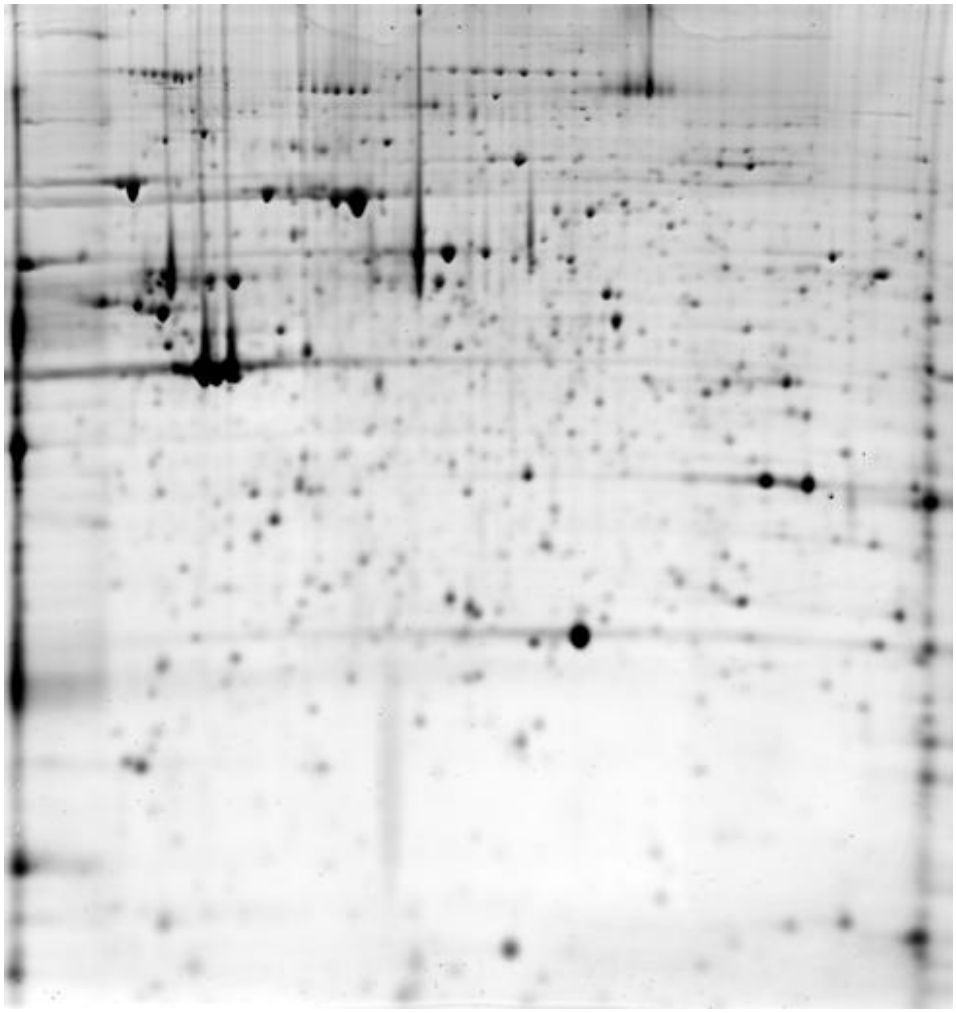


Figure 3.1 Gel d'acrylamide à deux dimensions.

Chacune des taches du gel représente théoriquement un groupe de protéines identiques. Cependant, une tache peut correspondre à plusieurs protéines différentes.

3.1.2 Les spectres de masse du MALDI-TOF

Le spectromètre de masse MALDI-TOF génère autant de spectres de masse qu'il y a d'échantillons dans une séquence de MALDI. Le logiciel Explorer (ABI) calcule la masse et l'intensité de chaque pic du spectre. Une des étapes fastidieuses de l'analyse de données consiste à vérifier les pics théoriques permettant d'identifier une protéine ou un mélange protéique avec les pics expérimentaux obtenus sur le spectre de masse. La fiabilité de l'identification est alors évaluée en tenant compte de l'intensité des pics, l'exactitude des masses et le nombre de pics retenus. Un des besoins de la plate-forme était de développer un outil capable de mettre en évidence automatiquement, pour chaque spectre de masse, les pics servant à l'identification de la protéine.

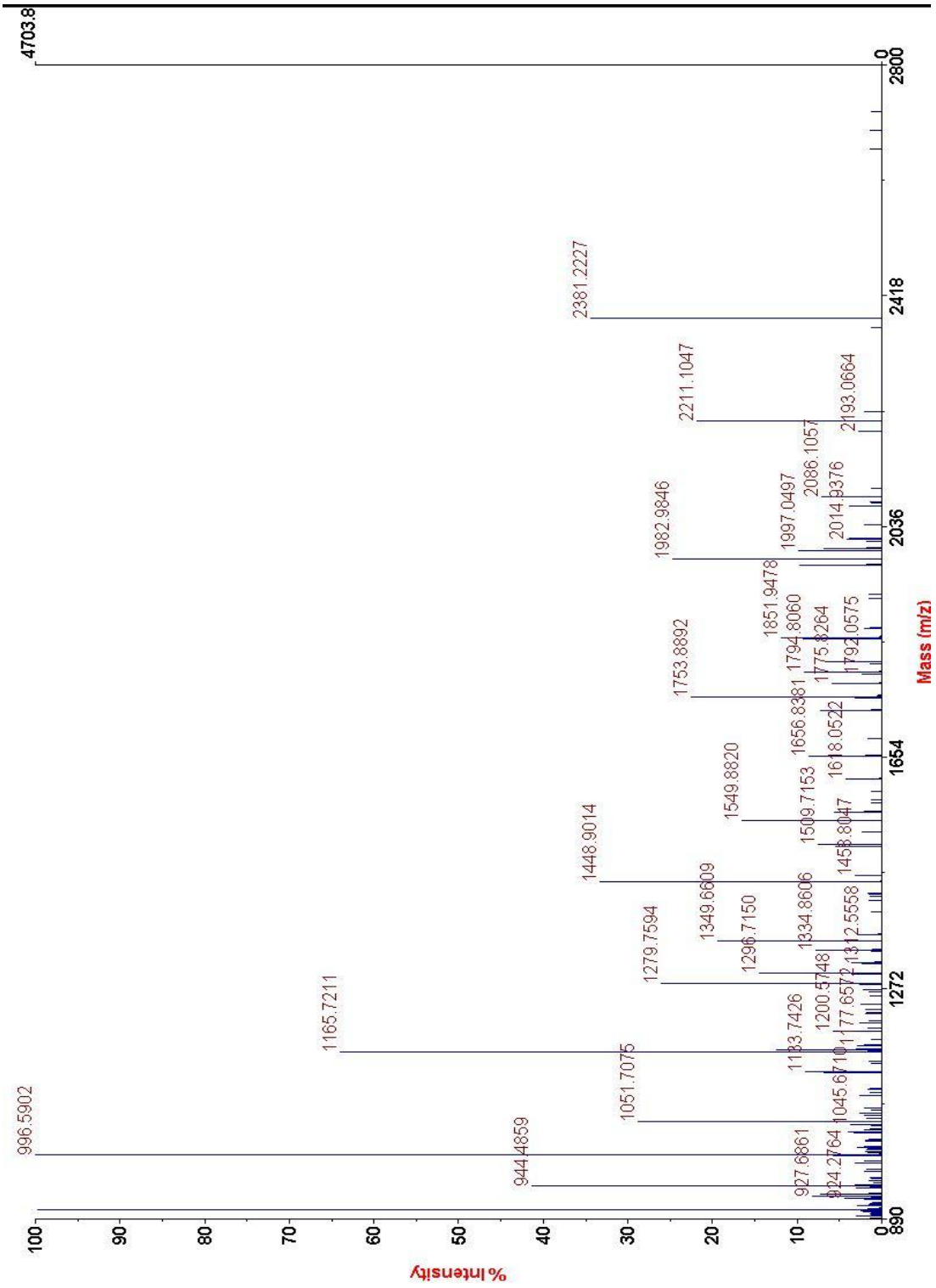


Figure 3.2 Spectre de masse MALDI-TOF généré par Explorer.

3.2 Modèle relationnel de la base de données

Le modèle relationnel est une forme de représentation à caractère sémantique permettant de spécifier une partie de la réalité au moyen d'entités et d'associations perçues entre les données (Gamache 2001). Chaque information contenue dans une base de données est appelée une entité. Les entités élémentaires ou attributs, sont regroupées dans une structure plus globale que sont les tables. Il est nécessaire d'établir un modèle conceptuel pour représenter l'organisation des entités et des relations entre toutes les données requises dans l'élaboration du système. Ce modèle conceptuel sera établi en utilisant la norme universelle UML. Cette norme permet principalement de modéliser les propriétés d'une entité et d'établir les associations entre elles dans un modèle standard. Dans le cas d'un modèle relationnel, l'entité est représentée par la table, et une ligne d'insertion des données de la table ou enregistrement, constitue un tuple. Le schéma du modèle relationnel de la plate-forme protéomique est illustré en appendice D. Les tables du schéma seront décrites en détail dans la section suivante.

La nomenclature UML que nous avons utilisée permet de retracer aisément l'utilité de chaque table ainsi que leurs entités et leurs relations. Une relation représente un lien sémantique entre les entités d'une classe X et celles d'une classe Y (Gamache 2001). Elle spécifie le nombre possible d'entités de X associées avec les entités de Y. Ce nombre n'est pas nécessairement identique dans la relation inverse. Le nom de chaque clé primaire, l'attribut unique qui identifie un tuple de la table, est composé soit par le préfixe « no_ » ou par le suffixe « _id ». La relation de la table X à Y est établie par la présence de la clé primaire de X dans la table Y. On appelle clé étrangère une clé primaire qui est contenue dans une autre table. La nomenclature relationnelle est de la forme (1.. n) et représente le nombre d'entités possibles de la classe X associées à la table Y. Dans les relations n-aires , c'est-à-dire deux tables ayant chacune une relation (1.. n) entre elles, la règle relationnelle impose la création d'une table intermédiaire pour ramener la relation à (1..n) en insérant les clés primaires de X et Y dans cette table intermédiaire. Le nom de ces tables intermédiaires faisant cette relation commence généralement par le préfixe « Inter_ ». Lorsqu'une clé primaire devient étrangère dans une autre table, elle peut avoir le même nom que dans sa propre table. L'attribut d'une clé primaire est très souvent un nombre, ce qui diminue l'espace

sur disque et augmente l'efficacité lors de l'établissement de nouvelles relations entre les tables. Certaines tables possèdent des attributs décrivant la nature d'une entrée ou enregistrement de la table. Les attributs sont identifiés par le préfixe « desc_ ». Par exemple, dans l'entité coloration, la clé primaire est coloration_id, et sa description, desc_coloration.

L'ajout de préfixes et suffixes abrégés aux attributs du schéma augmente leur cohérence et diminue le temps consacré à la saisie. En effet, l'utilisation de noms concis et descriptifs d'attributs et de variables peut considérablement réduire le temps relié au codage et à sa maintenance sans pour autant accélérer la saisie de données.

3.3 Description des tables du schéma de la base de données

Le modèle relationnel constitue une représentation schématique structurée des informations contenues dans la base de données. La base de données protéomique a été développée en utilisant ce modèle (appendice D). La section qui suit présente les tables de cette base. Pour faciliter la lecture, ces tables ont été regroupées selon le type d’information qu’elles contiennent, à savoir: les échantillons, les gels, les plaques, les redirections, les résultats et la validation des résultats.

Les Échantillons

Les tables de la section échantillon sont sans doute les plus importantes du schéma relationnel de la base de données, car elles renferment les informations sur les échantillons et sont le point de jonction de toutes les autres sections sauf celle concernant la validation des résultats.

COLORATION

Cette table contient les différents types de coloration utilisés dans les gels produits par la plate-forme. Sa clé primaire est reliée à la table Échantillon. Elle permet d'optimiser l'espace disque dans la description des échantillons.

COMMENTAIRE_ÉCHANTILLON

Cette table sert à entreposer les différents commentaires standardisés assignés à un échantillon. La table Échantillon est directement reliée à la table Commentaire_Échantillon via une table pont appelée Inter_comm_echant, puisque plusieurs commentaires peuvent être associés à un échantillon.

ÉCHANTILLON

Cette table est le cœur du schéma de la base, car les autres tables ne prennent leur sens que si elles ont une relation directe ou indirecte avec cette table. De plus, l'échantillon est l'élément premier du système de cueillette et d'analyse de données. Il est actuellement formé à partir des taches prélevées sur les gels à deux dimensions, mais sa provenance pourrait changer avec le développement de nouvelles technologies. La table Échantillon est une table pivot. La création de nouvelles tables de provenance devraient par conséquent être rattachée à celle-ci. Globalement, un échantillon provenant d'un gel 2d a un poids moléculaire apparent et un point isoélectrique apparent ou réel. Puisque l'échantillon est entreposé dans une plaque à 96 puits, le numéro de la plaque et la position qu'il occupe font partie de ses propriétés.

La plate-forme protéomique utilise un numéro séquentiel pour identifier les échantillons. Ce numéro est contenu dans l'attribut « no_interne ». Dans l'éventualité où un échantillon n'est pas caractérisé par un numéro séquentiel, comme un contrôle ou un test de

type RD (Recherche et Développement), l'attribut « autre_description » contient des valeurs en format caractères. La clé primaire de cette table est « no_échantillon ».

La séquence MALDI, la méthode de digestion, le tissu, la fraction cellulaire, la quantité utilisée, le type de contrôle et la taxonomie sont autant de caractéristiques spécifiques qui appartiennent à la table Échantillon. Si une caractéristique supplémentaire doit être ajoutée, elle le sera directement dans cette table ou dans une autre en relation avec cette dernière. La possibilité de modifier Échantillon, tout en conservant son intégrité, est nécessaire pour maintenir la versatilité de tout le système.

INTER_COMM_ECHANT

Cette table pont établit la relation n-aire entre les tables Échantillon et Commentaire_échantillon. Elle contient le numéro d'identification de l'échantillon et le numéro de son commentaire. L'attribut « origine » sert à stocker le type d'appareil concerné par le commentaire.

INTER_SEQUENCE

Cette table sert de pont entre la table Sequence et la table Échantillon. Elle fait l'association d'un échantillon avec la séquence MALDI-TOF ou MS/MS. Un échantillon fait partie d'une ou plusieurs séquences d'où l'existence de Inter_sequence. Cette table est maintenant peu utilisée, car la pertinence de sauvegarder ces informations est moindre avec l'avancement du projet ATLAS. Toutefois, elle pourrait éventuellement servir avec l'arrivée d'autres projets.

MASSE_LCQ

Cette table permettait au début de rapatrier les masses et les intensités des spectres de masse de type MS/MS. Mais comme la quantité de données par échantillon était trop importante et qu'il était difficile d'extraire les informations du logiciel Sequest, cette table est

devenue obsolète. D'autant plus que Mascot reproduit les spectres MS/MS dans ses pages d'affichage de résultats. La table est cependant conservée dans l'éventualité où un autre logiciel moins restrictif que Xcalibur permettrait de rapatrier directement les masses. Une représentation graphique d'un spectre de masse MS/MS pourrait alors être construit au moyen d'un script JAVA à partir des masses expérimentales.

MASSE_MALDI

Cette table contient les masses et les intensités expérimentales du spectre de masse MALDI-TOF appartenant à un échantillon. Ces informations proviennent d'une macro, écrite en VB (Visual Basic), située dans le logiciel Explorer du poste de travail du MALDI-TOF. Si le spectre de masse de l'échantillon est refait, l'exécution à nouveau de la macro supprime les anciennes données de masses/intensités de la table et sont remplacées par les nouvelles données.

METHODE_DIGESTION

Les échantillons sont digérés par l'appareil de digestion avant d'être envoyés vers un spectromètre de masse. Cette table contient les méthodes de digestion. L'insertion d'une séquence dans la table Séquence est corrélée par la sélection d'une méthode de digestion contenue dans Méthode_digestion. Elle est utilisée conjointement avec la table Séquence.

ORGANELLE

La table organelle a une contrainte de partition sur les tables Échantillon et Gel; c'est-à-dire qu'une fraction cellulaire n'est associée qu'à l'une des deux tables à la fois. L'association de la fraction cellulaire se fait avec la table Gel, si l'échantillon provient d'un gel, ou la table Échantillon, si aucune provenance particulière ne lui est attribuée. Ainsi, il n'est pas obligatoire qu'un échantillon provienne d'un gel pour être relié à cette table. Cette contrainte de partition est gérée par les scripts PL/SQL exécutant l'ajout des gels 2d et des échantillons (valide_test, valide_exp, edit_exp, ajout_controles).

PROVENANCE

Les échantillons sont créés de plusieurs façons. Ils proviennent soit des gels à une et deux dimensions, de chercheurs externes ou encore de tissus entiers et digérés. Cette table décrit donc ces différentes provenances et les assigne aux échantillons de la base. Les entités de cette table sont différentes de celles contenues dans la table Association, qui décrit la raison de la création des échantillons.

QUANTITE

Cette table regroupe toutes les unités de mesure utilisées dans la base Murin. Par exemple, les gels utilisent une certaine quantité de micro-grammes de tissu, et les échantillons, des femto-moles. La table Quantité est associée avec toute table contenant des unités de mesure telles la table Gel et la table Source.

SEQUENCE

Cette table sert à entreposer une séquence de spectromètre de masse. Dans les faits, cette table ainsi que la table Méthode_digestion sont peu utilisées, car les informations qu'elles contiennent sont plus ou moins nécessaires au fonctionnement du système. Toutefois, elles sont conservées en cas de modifications de protocoles de travail. Séquence et Méthode_digestion sont les tables les moins utilisées de la base.

SOURCE

La table Source définit l'échantillon contrôle en spécifiant la source et la quantité utilisée du contrôle. Elle est associée aux tables Type_source et Quantité. La raison expliquant la séparation de la source en ces deux composants est de permettre une réutilisation des données en évitant de surcharger inutilement les tables. En effet, il est plus efficace de sauvegarder une fois une quantité et de la relier par sa clé primaire, que de réécrire à chaque fois sa valeur. Lorsque l'usager construit un échantillon contrôle, il lui est

plus facile de choisir les composants du contrôle dans une liste déjà construite, au lieu de les réécrire à chaque fois en risquant de commettre des fautes d'orthographe. Les attributs de source sont le numéro d'identification de la source, le numéro associé à la source identifiant le type de contrôle et le numéro de quantité identifiant la quantité utilisée. Un échantillon ne provient que d'une source unique.

TYPE_SOURCE

Type_source regroupe les informations sur la nature des échantillons contrôles et ceux-ci sont identifiés par la clé primaire type_source_id. Le champ description_source est un code à trois lettres identifiant le contrôle. Ce code apparaît sur le nom de l'échantillon et permet de l'identifier rapidement. Détail_description est un champ de description plus approfondi du code. Par exemple, BSA serait le code à trois lettres où le terme « albumine de sérum bovin » serait la description du contrôle. Type_source est associé à la table Source, car elle lui fournit une des deux composantes requises pour l'entité source.

Les Gels

Les tables impliquées dans l’annotation des gels à deux dimensions servent dès la première étape de cueillette d’information des échantillons provenant de la plate-forme des gels.

DESCRIPTION_GEL

Les gels possèdent souvent des caractéristiques semblables. Ces caractéristiques doivent faire partie de la table Gel. Cette table sert à emmagasiner les descriptions courantes attribuées aux gels. Lorsqu'une description particulière est nécessaire, elle est directement insérée dans la table Gel. Originalement, la table Description_gel contenait des informations qui ont été regroupées par la suite en sous groupes et insérées dans de nouvelles tables : Quantités et Organelle. Ces tables sont reliées à la table Gel et Échantillon par des liens conditionnels. Description_gel joue un rôle similaire aux tables Commentaire_échantillon, Commentaire_protéine et Commentaire_plaque.

GEL

Cette table comprend les informations sur les gels à deux dimensions. Dans le cadre du projet ATLAS, le gel requiert un numéro d'identification GQ (Génome Québec) fourni par des tables externes à la plate-forme protéomique. De plus, ces tables contiennent les informations suivantes: no_protocole, sexe_contenant, no_groupe, lettre_temps, but_contenant, no_pool_contenant et no_tissu. La table gel stocke l'image du gel afin de permettre sa visualisation. La représentation de l'image à l'écran est accompagnée de l'identification des taches prélevées. Les gels regroupés dans un ensemble possèdent un numéro d'expérience et un numéro de pool permettant de les associer d'une part à une expérience et, d'autre part, à un traitement. La taille du gel, en millimètres, est importante afin de pouvoir positionner correctement les taches sur le gel lors de l'affichage. Enfin, la quantité de tissu utilisée (quantité_id) et la fraction cellulaire (organelle_id) réfèrent aux tables Quantité et Organelle.

INTER_DESCRIPTION

Inter_description établit les relations n-aire entre les tables Description_gel et Gel. Un gel peut être décrit par plusieurs descriptions et Inter_description a pour but de les associer au gel. Cette table joue un rôle similaire aux tables Inter_comm_prot, Inter_comm_echant et Inter_comm_plaque.

NUM_SPOT

Les informations sur les taches prélevées sont sauvegardées dans cette table. Cette table fait également l'association entre les gels et les échantillons. Un échantillon peut provenir de plusieurs taches appartenant à différents gels de même traitement. La clé unique de Num_spot est la combinaison unique du nom du gel d'où provient la tache et du numéro de cette tache. Le champ no_échantillon est une référence au numéro d'échantillon de la table Échantillon. La date de coupe, la qualité de l'apparence de la tache, son intensité et la

valeur du seuil de signification des différences d'intensités traitement/contrôle sont aussi des propriétés d'une tache dans un gel.

SAC

Les gels à deux dimensions sont entreposés dans des sacs de type « Zip-lock ». Un sac loge entre 12 et 15 gels. Ces gels, comme le sac, sont étiquetés de codes à barres. Le suivi des gels se fait via la table Sac, laquelle répertorie chaque gel à son sac et informe sur l'emplacement d'entreposage. Cette table est reliée à la table Gel.

SPOT

Cette table contient toutes les taches d'un gel résultant de la normalisation. Si un gel n'est pas normalisé, toutes ses taches sont répertoriées dans la table. Spot regroupe les coordonnées de la tache, son intensité, son intensité normalisée, une valeur en pourcentage sur son uniformité, son numéro de tache et le nom du gel auquel elle appartient. Les valeurs d'intensités sont utilisées pour faire un test de Student entre les taches d'un traitement et d'un contrôle. Ceci permet de vérifier si une différence des profils d'expression entre les protéines est présente. Les coordonnées servent à positionner la tache sur le gel lors de l'affichage de l'image gel à l'écran.

Les Instruments

Les informations décrivant les spectromètres de masse ainsi que celles utilisées par ces appareils sont contenues dans les tables de cette section. Ces informations servent par exemple à la validation des résultats d’identification protéique ou bien à la création de spectres de masse.

INFO_SEQUENCE

La table Info_séquence est destinée à emmagasiner deux informations requises pour bâtir une séquence de type MS/MS (voir section 2) interprétée par la suite par le logiciel

Xcalibur (ThermoFinnigan). La première information donne le chemin du poste de travail d'où proviennent les fichiers d'échantillons; la deuxième emmagasine la méthode de l'instrument utilisé. Les deux valeurs font partie du tableau de la séquence et celui-ci est généré par l'usager de la base protéomique.

INSTRUMENT

Tous les noms des instruments servant à créer des spectres de masse sont contenus dans cette table. Ainsi, lors d'une validation, une consultation, un ajout de commentaires ou une re-soumission, l'instrument concerné doit être sélectionné. La table Instrument est associée aux tables qui caractérisent leurs entités en incluant le type d'appareil. Les tables concernées sont Interprétation, Resoumission, Inter_comm_echant, Inter_redirection et Path.

PATH

Path contient les chemins de répertoires des postes de travail reliés aux spectromètres de masse. Cette table permet d'associer la provenance d'un fichier de spectre de masse à l'instrument qui l'a généré. Lors de la validation des résultats, cette table permettra d'identifier l'appareil qui les a générés. Donc la présence d’une nomenclature pour les chemins de fichiers propre à chaque instrument est une nécessité. Actuellement, l'association des instruments à leurs résultats est faite manuellement au moment de la validation.

Les Plaques

Les plaques de 96 puits sont des structures de plastiques servant de support pour les échantillons analysés par les appareils de spectrométrie de masse. Chaque puits de la plaque peut contenir un échantillon.

ASSOCIATION

Afin de différencier les plaques de 96 puits de la base, des catégories de plaques ont été créées et sauvegardées dans cette table. Ces catégories fournissent les raisons d'être de ces échantillons. Voici les catégories actuelles :

- *Protéome variable*

Les échantillons de la plaque proviennent d'expériences visant à étudier les profils d'expression protéique sous différentes conditions.

- *Protéome fixe*

Les échantillons de la plaque proviennent de gels qui ne sont pas dédiés à l'étude des profils d'expression de protéines mais servent plutôt à l'identification protéique.

- *Test*

Ces échantillons sont utilisés pour vérifier différentes modifications de protocoles expérimentaux. Ils ne sont donc pas considérés comme des échantillons servant à produire des résultats.

- *Chercheurs*

Ces échantillons proviennent des chercheurs clients du service du centre protéomique. Le type de projet et la nature de leurs échantillons sont inconnus.

La table Association est directement reliée à la table Inter-association.

COMMENTAIRE_PLAQUE

Les commentaires standardisés concernant les plaques sont emmagasinés dans cette table. Ces commentaires décrivent le contenu de la plaque, le type de protocole expérimental utilisé, son utilité et tout autre attribut descriptif d'une plaque.

INTER_ASSOCIATION

Inter_association est la table pont faisant la relation n-aire entre les tables Plaque et Association. Une plaque peut contenir des échantillons générés pour des raisons différentes. La table Inter_association distingue ces échantillons en étiquetant les puits de la plaque.

INTER_COMM_PLAQUE

Cette table établit la relation n-aire entre les tables Plaque et Commentaire_plaque. Elle contient le numéro d'identification de la plaque et le numéro de son commentaire.

PLAQUE

Cette table contient les informations relatives à une plaque : son numéro d'identification, sa description et un commentaire spécifiant, au besoin, une description particulière de la plaque. Ce dernier champ est utile lorsqu'il est nécessaire d'ajouter des commentaires autres que ceux contenus dans la table Commentaire_plaque.

Les Redirections

Ces tables sont dédiées à contenir le journal des échantillons destinés à être ré-analysés.

INTER_REDIRECTION

La table Inter_redirection fait l'association n-aire entre la table Échantillon et la table Instrument. Elle sert à mettre à jour les différents échantillons redirigés directement vers d'autres appareils, ou après avoir été dirigé au MALDI-TOF. Lors de la création d'une séquence de type MS/MS, tous les échantillons redirigés sont étiquetés dans cette table. L'étiquette de la redirection est enlevée après une nouvelle validation des échantillons via l'attribut « fait ».

RESOUMISSION

La table Resoumission permet de faire le suivi des échantillons destinés à être ré-analysés par Mascot. Il arrive que les résultats d'identification protéique d'un échantillon ne soient pas concluants et que l'échantillon doit être ré-analysé. Cette table enregistre temporairement le numéro d'identification du spectromètre de masse et de l'échantillon qui a été resoumis. L'usager peut alors connaître les échantillons qui ont été ré-analysés par Mascot. Lorsque les identifications protéiques de l'échantillon sont validées de nouveau, la table Resoumission est mise à jour en supprimant la ligne d'insertion concernant cet échantillon et ce spectromètre de masse. La section 3.6 explique en détail le principe de resoumission.

Les Résultats

Tous les résultats obtenus par le moteur de recherche Mascot sont sauvegardés dans ces tables. Ces résultats sont les identifications protéiques potentielles de chaque échantillon de la table Échantillon.

ANALYSE

Cette table est conçue pour stocker les informations globales sur les fichiers de résultats de recherche d'identification protéiques. Ces fichiers proviennent d'un moteur autre que Mascot ou Sequest. Ces informations comprennent le nom du fichier, la date de

soumission, la banque de protéines utilisées, le type d'instrument utilisé, etc. Elle sert à stocker les résultats d'identification de protéines générés par un moteur de recherche s'exécutant sur un serveur distant. Puisque ces résultats sont fournis par le site web, ils sont en format HTML.

INFO_MASSE

Cette table contient toutes les informations sur les masses théoriques, résultant d'une digestion protéique simulée, servant à appuyer une identification protéique trouvée par Mascot. Une procédure PL/SQL exécutant un script Perl ouvre le fichier de nouveaux résultats d'identification produit par l'engin Mascot et récupère les 20 premières identifications les plus probables. Chaque enregistrement de cette table est relié à la table Résultats_mascot qui contient une protéine du fichier de résultats.

MASCOT_DAEMON_FILES

Cette table fait partie de l'engin de recherche Mascot et contient les informations concernant les fichiers de spectre de masse envoyés à Mascot via son démon. Ce démon est le logiciel installé sur un poste de travail se chargeant d'envoyer les fichiers de spectre de masse au moteur Mascot situé sur le serveur Unix.

MASCOT_DAEMON_PARAMETERS

Cette table fait aussi partie du système Mascot et contient les informations sur les paramètres d'entrées de fichiers de spectre de masse envoyés à Mascot via son démon.

MASCOT_DAEMON_RESULTS

Cette table fait partie du système Mascot et permet de sauvegarder les informations concernant les résultats des fichiers des spectres de masse envoyés au démon Mascot. Cependant, elle ne contient que l'identification de la protéine ayant la plus grande probabilité

selon Mascot. Il est possible qu'une protéine ayant une probabilité moindre puisse s'avérer être la bonne identification. Des scripts se chargent alors d'amasser les informations du fichier de résultats contenus dans cette table, dont le numéro d'identification du fichier, de rapatrier les autres protéines de la liste et de les insérer dans la base. Lors de la validation, l'usager choisit s'il y a lieu, parmi la liste des protéines, celle ou celles qu'il considère comme étant la bonne identification.

MASCOT_DAEMON_TASKS

Cette table fait partie du système Mascot et permet de sauvegarder les informations sur les tâches créées dans le démon pour l'envoi des échantillons à Mascot. Une tâche est la suite d'instructions et de paramètres fournie à Mascot au lancement de la recherche.

PONT

Cette table relie un échantillon contenu dans la table Échantillon avec un résultat généré par Mascot, la table Mascot_daemon_results. Cette table est nécessaire à la validation des résultats, car les résultats générés par l'engin de recherche Mascot sont basés sur les mesures produites par les spectres de masse. Après chaque exécution du démon, des informations sur les fichiers de spectres de masse sont contenues dans la table réservée à l'engin Mascot (Mascot_daemon_results). D'autres tables utilisées par Mascot procurent des informations encore plus détaillées sur ces fichiers de spectre de masse (Mascot_daemon_parameters, Mascot_daemon_files). L'insertion des données dans Pont est effectuée par un déclencheur, **Raccord_after.pls**, qui fait l'association du fichier de spectre avec l'échantillon correspondant. Idéalement, la clé no_échantillon de la table échantillon aurait dû être contenue dans la table Mascot_daemon_results, tel que représenté par la relation (1..n). Mais puisque les insertions dans cette dernière sont faites par le démon et que toute modification de la table rend l'usage du démon Mascot impossible, il était nécessaire de créer la table Pont.

RESULTATS_MASCOT

Les résultats de Mascot concernent les informations contenues dans la liste protéique d'un fichier de résultats. Une partie de la liste est rapatriée dans cette table via des scripts en Perl. Son existence est essentielle, car elle permet à l'usager de sélectionner correctement la bonne identification protéique. Puisque la table Mascot_daemon_results ne contient qu'une partie de la première identification protéique, Resultat_mascot contient toutes les informations de ces identifications. Chaque tuple de la table s'associe à un ou plusieurs tuples de la table Info_masse qui regroupe les informations sur les masses théoriques.

La Validation des Résultats

Les résultats d'identification protéique validés pour chaque échantillon sont contenus dans les tables de cette section. Cette section est directement reliée à celle des résultats, car la validation d'une identification concerne les résultats d'identification obtenus par le moteur de recherche.

COMMENTAIRE_PROTEINE

Les commentaires de protéines sont utilisés pour qualifier une validation d'une identification. Ainsi, une protéine peut être validée avec une seule séquence peptidique, dans le cas du MS/MS, et avoir le commentaire « identifié avec un seul pic ». Le fait d'avoir séparé les commentaires protéiques des commentaires d'échantillons permet d'apporter des précisions qui ne concernent que les protéines. Lorsqu'aucune identification est associée à l'échantillon, il est évidemment impossible d'ajouter de commentaires.

Puisque les commentaires servent à décrire des entités différentes, il était plus pratique de les stocker dans des tables différentes (Commentaire_Échantillon, Commentaire_plaque, Commentaire_proteine).

INTER_COMM_PROT

Cette table pont relie les tables Commentaire_protéine et Prot_inter. Lorsqu'une protéine est validée, l'usager doit assigner au moins un commentaire afin de préciser la raison d'une identification. Ces informations sur les échantillons sont particulièrement utiles lors de consultations ultérieures. Inter_comm_prot contient le numéro d'interprétation d'un fichier, le numéro de résultat d'une protéine, le numéro du commentaire assigné et le numéro d'identification de la protéine contenu dans les tables publiques. Ce dernier champ ne contient aucune donnée, mais sa présence est nécessaire pour faire la référence à la table Prot_inter.

INTERPRÉTATION

La table Interprétation gère l'état des fichiers générés par Mascot. Elle indique si une identification d'un fichier a été validée. Elle contient également le numéro d'identification du fichier Mascot, un numéro d'interprétation, l'instrument pour lequel il y a validation ainsi que l'usager qui valide les résultats. Par conséquent, l'usager connaît le cheminement d'un échantillon de chaque appareil d'où il a été analysé.

PROT_INTER

Prot_inter est la table contenant les références protéiques suite à la validation d'une identification. Les attributs sont le numéro d'interprétation du fichier de résultats, le numéro de résultat Mascot, le type de confirmation de la validation, le numéro de référence protéique aux indexes publics et la date de validation. Prot_inter est associée aux tables Inter_comm_prot et Résultats_mascot, où la première contient les différents commentaires assignés à la protéine, et la seconde, les informations relatives à la protéine extraites du fichier de résultats Mascot.

Tables Obsolètes

CLIENT

Cette table entreposait les informations relatives aux clients qui utilisent le service de la plate-forme protéomique. Elle est devenue obsolète suite à la création d'une autre base dédiée à la gestion de la clientèle du service de protéomique.

LIEN_ECH

Cette table était utilisée pour faire la jonction entre la table Client et la table Echantillon lorsque les clients fournissent leurs échantillons au service protéomique. Elle est devenue obsolète pour la même raison que la table précédente.

LIEN_GEL

Cette table était utilisée pour faire la jonction entre la table Client et la table Gel lorsque les clients faisaient produire leurs gels d'acrylamide par la plate-forme des gels. Elle est devenue obsolète pour la même raison que les deux tables précédentes.

3.3 Programmes de cueillette automatique des données de la plate-forme 2D

3.3.1 Création et préparation des échantillons de la plate-forme *In vivo*

Le système actuel de la base de données protéomique offre deux façons de générer des échantillons. L'une se fait manuellement via un interface HTML (Hyper Text Markup Language) et concerne en grande partie les contrôles de calibration et de type RD (Recherche et Développement). L'autre façon est un processus de création automatique d'échantillons à partir des taches prélevées par la plate-forme des gels à deux dimensions. Ces taches sont des protéines extraites à partir des tissus dont l'information provient d'une base de données externe nommée *In vivo*. Cette base gère l'information sur les morceaux de

tissus prélevés immédiatement après le sacrifice des animaux et déposés dans des contenants. Ces contenants possèdent une identification dont le format est GQ001M01A13P1. Voici la signification de ce code :

GQ001 : Les 5 premiers caractères du code identifient le type de protocole. Un protocole représente une série de manipulations expérimentales impliquant plusieurs tissus d'un groupe de souris soumises à des traitements particuliers.

M : Ce caractère identifie le sexe des animaux du groupe. Le caractère M identifie les mâles, et le caractère F, les femelles.

01 : Ce nombre, combiné de 2 chiffres, identifie le traitement appliqué au groupe de souris. Dans cet exemple, 01 identifie Intact-GDX, qui correspond à une castration simulée chez les souris mâles.

A : Cette lettre identifie la durée de l'application du traitement. Par exemple, la lettre A signifie que l'animal a été sacrifié 24 heures après le traitement.

13 : Ce nombre de 2 chiffres identifie le tissu. 13 correspond au foie.

P : Cette lettre indique la destination du prélèvement. P signifie que le traitement est dédié à la plate-forme de protéomique.

1 : Ce chiffre indique le numéro de réplicat des prélèvements d'un même tissu provenant d'un groupe de souris soumis au même traitement.

3.3.2 Création des gels à deux dimensions de la plate-forme 2D

La séparation des protéines extraites des tissus est effectuée au moyen des gels d'acrylamide. Ces gels contiennent soit la totalité des protéines du tissu ou une partie

provenant des fractions sub-cellulaires. L'utilisation de ces fractions dans la préparation des gels est parfois nécessaire afin de récupérer suffisamment d'échantillons protéiques pour les analyser. Chaque expérience forme des lots de 6 à 24 gels regroupés en traitements.

Lorsque les échantillons sont générés à partir des gels à deux dimensions, les premières informations à être insérées dans la base de données concernent ces gels. La saisie des données se fait via une interface HTML. La section 4.1.1 explique en détail le mode d'emploi de cette interface. La saisie s'effectue avant même qu'une tache n'ait été prélevée sur les gels. Chaque gel est numérisé au balayeur optique pour générer des images TIFF (Tag Image File Format), d'une grosseur de 5 à 10 méga-octets. Ces images sont par la suite traitées avec le logiciel PdQuest et sauvegardées dans un répertoire du poste de travail muni du numériseur optique. Exemple, le répertoire /image contient les fichiers image identifiés d'un nom unique et significatif. Par convention, le nom des gels et des images sont de la forme A9999. Pour l'instant, la lettre permet d'identifier rapidement les gels au projet de recherche correspondant. Par exemple, la lettre G est associée à Génome Québec.

Image2d.pl est un script Perl implanté sur le serveur de base de données. Celui-ci utilise la librairie Net::FTP pour se connecter sur le poste de travail distant et se réfère à un fichier d'archives pour récupérer toutes les nouvelles images situées sur le poste de travail. Le fichier d'archives contient le nom des fichiers image récupérés par le script. Ce fichier est mis à jour lors de chaque insertion des fichiers image dans la base. Les nouvelles images de format TIFF sont transférées dans un répertoire du serveur de base de données nommé image_tif. Par la suite, le script utilise le programme ImageMagik pour transformer les images TIFF en format JPEG (Joint Photographic Expert Group), d'une grosseur de 100 à 500 kilo-octets. Ce format compresse l'image TIFF afin de diminuer sa taille. Les images résultantes occupent moins d'espace disque pour leur entreposage dans la base de données. La qualité de ces images est diminuée, mais demeure suffisante pour permettre leur consultation sur les pages HTML de la base. Les images JPEG sont sauvegardées dans la table Gel par le script Image2d.pl. Ce script utilise la libraire locale Murin.pm qui comprend des fonctions générales servant à l'interrogation de la base de données.

3.3.3 Prélèvement des taches et création des échantillons de la plate-forme protéomique

On sélectionne les images gels dont les taches sont les plus nettes. Ces images sont comparées entre elles pour normaliser les taches. Ainsi, le numéro d'une tache sur un gel est identique à la tache située au même endroit sur un autre gel. Lors du prélèvement, les informations sur chacune des taches sont sauvegardées dans un fichier de type Excel, qui est converti par la suite en format texte. Ce fichier texte possède un nom unique et significatif. Il est sauvegardé dans un répertoire du poste de travail : spot. Le script Perl, **Pro_murin.pl**, utilise les mêmes librairies que **Image2d.pl**, rapatrie les fichiers sur le serveur et stocke les informations dans la table Num_spot de la base Murin. Ces fichiers sont entreposés sur le serveur de base de données dans le répertoire dont le nom est défini dans un fichier de paramètres. **Pro_murin.pl** traite aussi les données du fichier en format texte et prépare les informations sur les échantillons qui seront insérées dans la table Échantillon.
Voici ces informations :

- Numéro de la tache
- Nom du gel
- Le poids moléculaire
- La quantification ou intensité du spot qui représente une valeur de densité optique
- La qualité de la tache ou uniformité de la tache
- Le point isoélectrique
- Le numéro de plaque de 96 puits
- La position sur la plaque
- La date de coupe
- Le numéro de pool. Ce dernier concerne les échantillons provenant d'une tache prélevée sur plus d'un gel, car un échantillon peut être créé à partir d'une ou de plusieurs taches de gels.

Le logiciel PdQuest génère un autre fichier en format Excel. Ce fichier comprend les quantifications normalisées, les coordonnées sur l'axe des abscisses et des ordonnées de toutes les taches qui ont été retenues suite à la normalisation des gels. Ce dernier fichier est

converti en format texte et inséré dans le répertoire du poste de travail dont le nom est défini dans un fichier de paramètres. Le script Perl **Coord.pl**, implémenté sur le serveur de base de données, rapatrie ce fichier dans le répertoire 2d-coord du serveur à l'aide des mêmes librairies utilisées par les scripts précédents. La mise à jour des tables Spot et Num_spot est effectuée par ce script. De plus, ce script utilise la librairie R pour appliquer des tests de t Student dans le but de détecter des différences significatives entre les intensités des taches provenant des traitements et des contrôles. La moyenne de différentes quantifications d'une même tache est calculée dans chaque traitement. Ces moyennes servent au calcul du test de t pour vérifier si les différences d'intensité entre les traitements sont statistiquement significatives. Les tests de t sont calculés uniquement sur des taches prélevées.

3.3.4 Insertion des données du spectre MALDI

La macro 5 du logiciel Explorer, écrite en Visual Basic, a été conçue pour aller extraire les masses les plus intenses d'un spectre de masse, selon un barème pré-défini, et les insérer directement dans la table Masse_maldi via un module de connexion à Oracle ODBC (Object Data Base Connection). Le barème consiste à diviser le spectre de masse en 8 parties et à sélectionner les 10 masses les plus intenses de l'intervalle. La macro est automatiquement activée lors de la création d'un spectre de masse d'une séquence MALDI-TOF. Elle est aussi exécutable manuellement. La création de macros constitue la seule façon d'extraire les données du logiciel, puisque la compagnie ne distribue pas la source de son code. Si un nouveau spectre de masse est généré pour remplacer l'ancien, la ré-exécution de la macro supprime les données insérées précédemment qui sont remplacées par les nouvelles. Chaque spectre génère un fichier texte contenant les masses les plus intenses du spectre. Ce fichier est utilisé comme paramètre d'entrée pour la recherche d'identité effectuée par le moteur Mascot.

3.3.5 Les échantillons contrôles et Recherche et Développement (RD)

La création des échantillons contrôles se fait de deux façons : l'une est effectuée manuellement via l'interface web, et l'autre automatiquement via un script Perl. La méthode manuelle permet de créer un seul échantillon contrôle à la fois, alors que la deuxième en

crée systématiquement plusieurs. Le script Perl s'appelle **Recherche_dev.pl** et il est implémenté sur le serveur de base de données. Celui-ci utilise les mêmes librairies NET::FTP pour les connexions et le transfert à distance des données, et Murin.pm pour les interrogations dans la base. Le fichier texte contenant les informations sur les échantillons RD et les contrôles résident dans un répertoire du poste de travail. Il est rapatrié sur le serveur de base de données par le script **Recherche_dev.pl**. Ce dernier insère par la suite les données dans la table Échantillon. Les échantillons contrôles diffèrent des échantillons habituels par leur nom et leur contenu. Leurs spécifications sont stockées dans la table Source. Cette dernière regroupe les références sur le type de contrôle et la quantité utilisée. Ces informations sont contenues dans les tables Type_source et Quantité.

3.4 Mascot et son démon

3.4.1 Principes de la soumission des fichiers à Mascot

Les fichiers produits par les différents spectromètres de masse ont des formats différents. Le MALDI-TOF génère des fichiers de format texte, alors que le LCQ produit des fichiers de format binaire RAW. La façon la plus conviviale de soumettre ces fichiers au moteur de recherche Mascot est de les envoyer via le démon Mascot. Le démon peut être installé sur n'importe lequel poste de travail possédant un module de communication avec la base (section 3.4.2). La principale fonction du démon est de fournir les paramètres d'entrées nécessaires et les fichiers de spectre de masse à Mascot. Le fonctionnement du logiciel de recherche Mascot requiert un fichier de paramètres de format MIME (Multipurpose Internet Mail Extensions). Ces paramètres incluent les données expérimentales du spectre de masse et d'autres informations sélectionnées dans le formulaire du démon précédant l'exécution de la recherche (voir annexe A). L'utilisation du démon est expliquée dans la partie 4.3. Cette section décrit la manipulation des fichiers de paramètres entrants, les fichiers de paramètres sortants et la façon dont ils sont récupérés par les scripts.

Les fichiers de format MIME générés par le démon sont transférés au moteur de recherche Mascot, lequel est situé sur le serveur Tahaa. L'utilisateur peut préparer des lots de fichiers de spectres de masse ayant les mêmes paramètres de recherche. Ces lots de fichiers

sont appelés des tâches. La préparation de plusieurs tâches en simultané est possible, mais Mascot n'analyse qu'un seul fichier par tâche. Ceci permet à plusieurs postes de travail d'utiliser l'engin de recherche sans surexploiter les ressources du système. Lorsque Mascot a terminé une recherche, il produit un fichier de résultats qui est sauvegardé dans un répertoire spécifique sur le serveur. C'est un répertoire qui est identifié par la date de soumission du fichier. Tous les fichiers sont identifiés par une lettre suivie d'une clé unique et séquentielle. Le contenu du fichier est de format MIME. Il contient tous les paramètres qui ont servi à générer les résultats et peuvent être réutilisés pour générer un nouveau fichier de paramètres entrant à Mascot. C'est le mode de re-soumission discuté dans la section 4. Toutes ces étapes de manipulations de fichiers se font automatiquement, excepté la création des tâches, qui elle, demande à l'usager de choisir les fichiers de spectres de masse à soumettre. Tous les fichiers de spectres doivent être inspectés afin de s'assurer de leur conformité à certains critères. Un spectre de masse de bonne qualité contient des masses (m/z qui signifie masse sur la charge du peptide) se situant entre 900 et 2500 Dalton, dont quelques masses présentes en quantités nettement supérieures à la moyenne du spectre (Figure 3.2).

3.4.2 Entreposage des résultats de Mascot dans la base

Les Démons Mascot, installés sur les postes de travail, sont directement connectés à la base Murin via un module de communication ODBC. Ce module est nécessaire au transfert des informations des postes de travail aux tables de la base utilisées par le logiciel Mascot. Les insertions dans ces tables sont transparentes aussi bien à l'usager qu'au développeur. On ne doit pas modifier ces tables, au risque de tout rendre le système inutilisable. Toutefois, un déclencheur (trigger) nommé **Raccord_after**, récupère le numéro de résultat (result_number), de la table Mascot_daemon_results, le nom du fichier qui a été soumis à Mascot et le nom du fichier des résultats. Il associe le nom du fichier à son échantillon et met à jour la table Pont. Il appelle par la suite une fonction compilée en JAVA, **Run_perl.** Cette fonction permet l'exécution de commandes externes à la base Oracle à partir de scripts PL/SQL. Le script Perl **Sommaire.pl,** situé sur Tahiti**,** est alors exécuté. Ce dernier ouvre le fichier du résultat et récupère les 20 premières identifications protéiques sur un total de 50. Les insertions des identifications se font via la librairie Murin.pm dans les tables Résultats_mascot et Info_masse. **Résultats_mascot** contient les

informations générales sur les protéines et chacune de ses masses est une entité de la table **Info_masse**. Lorsque le déclencheur a terminé son exécution, les données ont été mises à jour et sont accessibles aux usagers.

3.5 Programmes d'analyse et de validation des données

Le système de validation a été développé en grande partie en PL/SQL, puisque c'est le langage compatible avec le SGBD Oracle. Plusieurs procédures sont requises dans les différentes voies qu'empruntent les étapes de validation. La validation des identifications protéiques est spécifique aux échantillons et est effectuée de trois façons.

La première façon consiste à valider un échantillon dont aucun spectre de masse généré par le MALDI-TOF a une qualité suffisante pour être analysable. Cet échantillon est étiqueté du commentaire « Pas de spectre ». Il est alors conseillé de re-générer un autre spectre ou de le re-diriger vers un autre instrument, comme le MS/MS. Cette validation met à jour la table **Inter_redirection** et **Inter_comm_echant**. Il n'y a donc pas d'interprétation des fichiers résultats, car ceux-ci n'existent pas. Les tables utilisées dans l'interprétation des fichiers de résultats ne sont par conséquent pas touchées.

La deuxième façon consiste à valider un échantillon ayant été soumis au moteur de recherche Mascot, et pour lequel aucune identification formelle n'a été trouvée. La table **Interprétation** permet d'étiqueter le fichier résultat et de lui assigner la valeur 'non' à l'attribut « identification ». En fait, un échantillon possède un ou plusieurs fichiers de résultats pouvant provenir d'instruments différents. Chacun de ces fichiers est catalogué dans la table **Interprétation**. Ainsi, si tous les fichiers de résultats d'un échantillon sont étiquetés d'un « non » dans l'attribut identification, cela signifie qu'aucune identification n'a pu être attribuée à l'échantillon. Il est important que le système fasse la différence entre les fichiers qui proviennent des différents instruments. En effet, la validation de résultats d'un échantillon dont les données expérimentales ont été générées par un MALDI-TOF doivent être spécifiques à cet instrument. L'échantillon ne doit pas avoir ses fichiers résultats LCQ interprétés dans cette même validation, ce qui rendrait la base de données incohérente. Pour l’instant, cette distinction est faite manuellement via l’interface de validation, mais sera faite

automatiquement par reconnaissance des chemins de fichiers contenus dans la table **Path**. La procédure PL/SQL **Pas_identification** se charge de mettre à jour la table **Interprétation** et de proposer une redirection de l'échantillon vers un autre instrument.

La troisième façon est utilisée lorsqu'il y a une identification. Les protéines stockées de la base possèdent un numéro de résultat unique et ont toujours deux options de validation, soit une validation présomptive, soit une validation définitive. La procédure **Affiche_protéine** affiche la liste des identifications protéiques de tous les résultats obtenus d'un échantillon. Les informations spécifiques aux identifications protéiques retenues pour la validation sont transmises à la procédure **Ajout_protéine**. Par exemple, 107792_tentative représente une valeur concaténée contenant le numéro de résultat unique de l'identification protéique, soit « 107792 », et « tentative » représente l'attribut signifiant la validation présomptive. La procédure insert le numéro d'interprétation, le numéro de résultat, le type de validation, la date de confirmation et le numéro de la protéine correspondante dans la table **Prot_inter**. Deux types d'ajouts sont possibles dans cette table. Le premier consiste à supprimer toutes les anciennes validations de l'échantillon selon le spectromètre de masse. Ainsi, une ambiguïté telle une double validation du même résultat, est impossible. Le deuxième type d'ajout est de conserver les anciennes validations et d'ajouter une identification supplémentaire concernant l'instrument spécifié. La procédure met à jour les identifications d'échantillons par les tables **Inter_redirection**, **Re-soumission** et **Interprétation**. Pour chaque validation, la procédure exige l'insertion d'au moins un commentaire sur la protéine. Cette insertion est exécutée par la procédure **Annote_comm_prot**. La table **Inter_comm_prot** fait l'association entre le commentaire et la protéine choisie.

3.6 Programmes de relance automatisée des échantillons

La possibilité de relancer un échantillon est une fonction importante du système de la plate-forme protéomique. Lorsqu'un échantillon n'obtient aucune identification, un commentaire particulier, par exemple « échantillon à relancer », peut lui être assigné entraînant la re-soumission de cet échantillon au moteur Mascot avec ses paramètres originaux. Lors de la mise à jour des banques protéiques locales, il sera alors possible de ré-

analyser systématiquement ces échantillons. Cette fonction sera éventuellement automatisée. L'usager a aussi le choix de re-soumettre les échantillons de son choix, peu importe sa validation. Le choix de relance par date ou par intervalles de 50 échantillons est aussi possible. La procédure **Choix_ech** offre la sélection de ces différentes options. Cette procédure calcule le nombre d'échantillons non identifiés et, parmi ceux-ci, le nombre de ceux qui sont aptes à être re-soumis. En effet, certains échantillons ne possèdent pas des spectres de masse d'une qualité suffisante pour permettre l'obtention de meilleurs résultats. La liste des échantillons à re-soumettre peut être raffinée, au besoin, au moyen de cases à cocher situées devant chaque fichier d'échantillons. La procédure exécutant la relance des fichiers d'échantillons s'appelle **Relance_ech**. La relance constitue un avantage important du système protéomique, puisque l'usager n'a pas à refaire une analyse manuelle des tous les échantillons non identifiés, surtout si chaque échantillon requiert des paramètres différents de recherche.

La procédure **Relance_ech** exécute une boucle. Elle utilise les numéros de résultats de chaque fichier pour récupérer le nom des fichiers de résultats. Elle déclenche par la suite un script Perl nommé **Make_input_txt.pl** via la fonction **Run_perl**. Ce script ouvre le fichier de résultats et re-génère le fichier de paramètres original de format MIME. Ce fichier de paramètres est sauvegardé dans un sous-répertoire de Mascot. Puis, la procédure appelle le script Perl **Relance.pl** qui se charge de fournir au programme principal de Mascot, **nph-mascot.exe,** le fichier de paramètres en question. De plus, **Relance.pl** crée un fichier nommé **out.tmp** contenant le nom du fichier d'entrée original, le type d'instrument qui a produit le fichier original et le nom du nouveau fichier de résultats Mascot. Finalement, la procédure appelle le script **Sommaire2.pl** qui ouvre le fichier **out.tmp** et en extrait les informations afin de stocker les nouvelles identifications de Mascot dans la base Murin. La boucle se termine lorsque tous les fichiers à relancer ont été traités.

3.7 Création d'interfaces de consultation

3.7.1 Interfaces HTML

Les interfaces de consultation générées en HTML servent à interroger la base Murin via un fureteur web. Les postes de travail PC, UNIX et Linux peuvent utiliser les fureteurs Netscape, Internet Explorer, Mozilla, ou tout autre fureteur compatible. Pour les postes de travail Macintoch, Safari est le seul fureteur pour le moment utilisable, car il est le seul à exécuter des scripts Java 1.4.1. Cette librairie Java est incluse dans l'installation complète de ces fureteurs dont les versions remontent à moins de deux ans. L'utilisation de navigateurs contourne l'installation de logiciels clients pour l'interrogation de la base Murin. Ceci permet la réutilisation des outils gratuits et disponibles sur l'Internet en plus d'éviter les inconvénients reliés aux programmes trop spécifiques. L'affichage à l'écran des données utilise le format HTML de version 4.0 et plus. Ce format est compatible avec tous les fureteurs web. La présentation de l'information est accompagnée de tableaux, de graphiques et autres effets visuels. Les usagers éditent leurs données, en ajoutent ou en suppriment, selon les possibilités du système, via les interfaces de consultations.

3.7.2 Applets de visualisation JAVA

Les applets JAVA sont des programmes autonomes s'exécutant sur le poste de travail de l'usager et affichant via la page HTML. Ces programmes permettent de générer des graphiques et de les rendre interactifs. Ces applets apportent un aspect plus ergonomique lors de l'interrogation de la base lorsque des informations peuvent être schématisées. Trois applets ont été développés pour la consultation. Les applets ont été conçues et écrites en Java par M. Hugo Laliberté.

Le spectre de masse

Le premier applet, nommé **GraphMaldi**, sert à afficher les spectres de masse. Le spectre s'illustre au moyen d'un diagramme contenant les masses d'un échantillon sur

l'axe des abscisses, et leur intensités relative sur l'axe des ordonnées. Ces données proviennent de la table Masse_maldi et sont fournies en paramètre à l'application JAVA via la procédure **Annonce_graphe**. L'applet permet de visualiser rapidement les masses théoriques de l'identification des masses expérimentales obtenues. Les masses théoriques de chaque identification protéique sont sauvegardées dans la table **Info_masse**. Il suffit de comparer les masses expérimentales et théoriques entre elles et d'identifier celles qui ont une concordance. Le code Java bâtit le spectre à l'aide d'une liste complexe de paramètres ordonnés qu'il a reçue de la procédure PL/SQL **Annonce_graphe** (annexe B). Plusieurs options sont mises à la disposition de l'usager telles l'affichage des valeurs des masses expérimentales, la révélation des masses expérimentales concordant avec les masses théoriques et la modification de la taille du spectre.

Les images de gels

L'applet, **ImageGel2D**, affichant les images de gels est appelée soit par la procédure **Annonce_gels** ou **Annonce_expérience,** selon le contexte. Cette applet affiche les images de gels regroupés en traitements d'une expérience et localise soit toutes les taches prélevées, soit celles choisies. Chacune des taches possède un numéro et un lien qui mène vers une page d'informations sur les comparaisons d'intensités des taches entre les traitements. L'affichage de toutes les images gels d'une même expérience permet la comparaison des différences d'intensités qualitatives entre deux traitements donnés. L'usager peut choisir parmi les traitements, celui qu'il considère comme étant un contrôle. **Visionne_un_gel** n'affiche à l'écran qu'une seule image de gel indiquant la localisation du prélèvement des taches. L'affichage d'une seule image permet de découvrir d'autres taches potentiellement intéressantes pour un prélèvement, ou simplement de faire un examen visuel des taches prélevées. Des cercles de couleurs s'affichent autour des taches prélevées. Les couleurs diffèrent proportionnellement au rapport de la moyenne de l'intensité des taches entre deux traitements donnés. Une option d'agrandissement affiche les détails de l'image et une recherche par le numéro de tache, ou de l'échantillon, permet de centrer les gels vers la tache recherchée.

La procédure **Annonce_expérience** envoie une liste complexe de paramètres ordonnées à l'applet **ImageGel2D** et concerne les différents traitements de l'expérience, les numéros de taches ainsi que leurs coordonnées et leur intensités, le nom des gels et leurs images (voir annexe C). La procédure **Visionne_un_gel** envoie le même type de paramètres que la procédure précédente, mais pour un gel uniquement. La procédure **Annonce_gels** envoie le même type de paramètres que la procédure **Annonce_expérience**, sauf qu'elle ne localise que le numéro sélectionné de la tache, soit celui correspondant à l'échantillon sélectionné par l'usager à partir de la procédure **Analyse_no_interne**. Les données utilisées par ces procédures sont stockées dans les tables Gel, Spot, Num_spot, Échantillon, Description_gel et Organelle du schéma Protéomique. Les informations relatives aux traitements sont tirées des tables du schéma Invivo, dont Chirurgie, Traitement, Temps, Groupe et Protocole.

L'utilisation des applets pour l'affichage des images de gels a dû être abandonnée lors de l'ajout de pare-feu à notre réseau. L'applet est un programme chargé sur un poste de travail distant et génère l'affichage à partir des paramètres reçus. Pour afficher les images, il doit faire une connexion à la base de données afin d'aller chercher ces images. Puisque le pare-feu empêchait l'applet de communiquer avec la base, l'affichage ne pouvait se faire correctement. L'usage d'un servlet était donc nécessaire pour le chargement des images afin de détourner la connexion à la base en provenance de l'applet. Le servlet est un module qui s'exécute sur le serveur d'application et non sur le poste client. Comme le serveur d'application n'a pas à traverser de pare-feu, il se connecte directement à la base. Par conséquent, l'applet ne communique plus directement avec la base de données. Du côté client, l'Applet est chargé dans le fureteur du client. L'applet établit une connexion sécurisée de type https avec les servlet et envoie une requête au module iAS Oracle. OC4J, qui est un serveur d'application web, appartenant à iAS. Le Servlet se connecte à la base de données Murin via un interface de communication JDBC (Java DataBase Connection). Il obtient ainsi l'image du gel de la base et l'envoie à l'applet par la connexion https. La figure 3.3 illustre en bref les couches des différents modules utilisés pour le chargement complet des applet/servlets.

Le diagramme de comparaison des profils d'intensités des taches

Cet applet, nommé BarreApplet, permet d'afficher dans un diagramme à barres verticales, pour un numéro de tache d'une expérience, la moyenne des intensités et son erreur type pour chaque traitement. L'échelle graduée du diagramme en abscisses est une valeur relative d'un rapport de moyenne d'intensité de deux traitements choisis par l'usager. La valeur du traitement de référence est fixée par défaut à 100%. Par défaut, les colonnes du diagramme sont de couleur bleu. Si l'échantillon d'un traitement a obtenu une identification protéique, la colonne est de couleur rouge. Ce diagramme est accompagné d'un tableau contenant les informations relatives au numéro de la tache appartenant à chaque gel de l'expérience. Ces informations incluent la quantification normalisée, la valeur p, seuil de signification du test de t, lorsque celui-ci est calculé sur un traitement, et un lien menant aux informations d'un échantillon, si la tache a été prélevée. La procédure **Stat_echant** génère la liste des paramètres pour l'application JAVA et inclut les informations sur la moyenne des intensités des taches de chaque traitement.

3.8 Utilisation d'un système de code à barres

L'utilisation d'un système de code à barres facilite l'identification des contenants utilisés en laboratoire au moyen d'étiquettes codées. Les gels, les sacs qui contiennent les gels pour l'entreposage et les plaques de 96 puits, sont identifiés à l'aide de ces codes à barres. Les procédures **Imprime_lot_etiquette** et **Imprime_etiquette** sont des procédures qui fournissent les paramètres nécessaires à l'impression des étiquettes. Ces paramètres sont le format d'étiquette, le nombre d'impressions, la date, une brève description du contenant et du type de tissu s'il y a lieu. La procédure appelle un CGI (Common Gateway Interface) Perl, nommé **barcode.pl**, qui se charge de créer un fichier en format d'impression. Toutes les étiquettes peuvent être lues par un lecteur optique installé sur le poste de travail. Le système facilite le suivi des échantillons tout en évitant les erreurs de saisie des numéros d'échantillons.

3.9 Les bases publiques de la plate-forme de bioinformatique

Les bases publiques de la plate-forme de bioinformatique regroupent certaines références à des annotations disponibles dans les bases biologiques sur Internet. Ces référencess concernent la séquence des gènes, des ARN, des protéines ainsi que des EST. Les principales banques de données contenant les informations sur les protéines sont (Swiss-Prot, NCBI, PIR, Trembl, etc.). Les informations sur les gènes et les ARN se retrouvent dans les banques de gènes (Unigene, NCBI, ENSEMBL, UCSC), les banques de EST (NCBI, Genest), les banques de Ditags du SAGE, celles des sondes Affimetrix et des sentiers de signalisation ou métabolique (Pathway). Les tables permettent d'établir des liens sur les références des annotations existant sur Internet et de les maintenir à jour. Grâce à elles, les associations entre les résultats des différentes plates-formes de recherche du CHUL travaillant sur un même projet peuvent être intégrées.

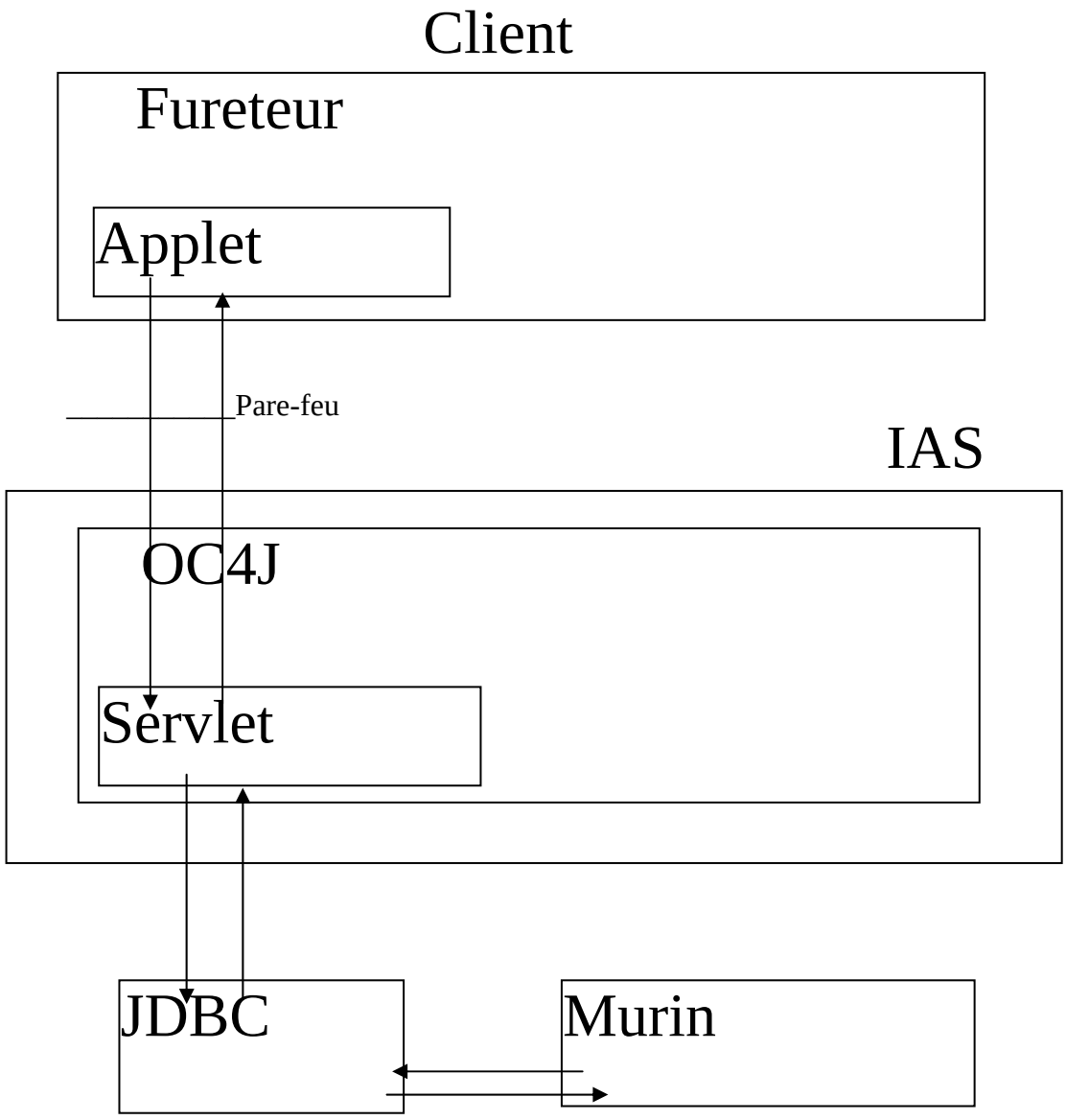


Figure 3.3 Schéma illustrant la communication entre les modules pour afficher les images des gels.

CHAPITRE IV

EXPLOITATION DU SYSTÈME ET DE SES RETOMBÉES

La section précédente décrivait l'architecture et l'implantation du système de la base protéomique. Cette section décrit l'utilisation de ce système. Les nombreuses pages d'accès à la base, de format HTML, ont été regroupées selon leur contenu et insérées dans un menu. La figure 4.1 illustre les sections du menu principal du système protéomique : Murin contient les liens pour accéder et quitter la base Murin, Ajout contient des liens vers des formulaires qui insèrent différentes informations telles les expériences, les contrôles, les Commentaires , etc. Pour valider les échantillons, les pages sont insérées dans le menu Validation. La consultation des résultats et des informations relatives aux gels, ces pages sont accessibles dans le menu Consultation. Pour l'ajout de paramètres, vérifier l'archive des connexions et des dernières identifications, accéder à des statistiques, et relancer des échantillons à Mascot, le menu Administration contient des liens vers ces pages. Finalement, le menu Aide affiche une liste de liens menant vers les manuels d'utilisation des fonctions de la base protéomique. Chacune de ces sections et sous-sections est élaborée dans ce chapitre et l'ordre de présentation est en fonction de la chronologie de la manipulation des données : la cueillette et la saisie de l'information, l'analyse et la consultation des données.

Les menus

Figure 4.1 Le menu principal de la plate-forme protéomique

Voici, en détail, les éléments du menu qui permettent d'accéder aux différentes pages du système.

- **Murin**
  - ⇒ **Retour au menu principal**
  - ⇒ **Quitter**
- **Ajout**
  - ⇒ **Gels**
    - → **Ajouter une expérience**
    - → **Éditer une expérience**
    - → **Test de gels**
    - → **Ajouter un sac**
    - → **Description de gels**
  - ⇒ **Plaques**
  - ⇒ **Séquences**
  - ⇒ **Digestions**
  - ⇒ **Redirections**
  - ⇒ **Échantillons**
    - → **Ajouter des contrôles**
    - → **Ajouter des RD**
  - ⇒ **Commentaires**
    - → **Ajouter commentaire pour protéine**
    - → **Ajouter commentaire pour échantillon**
    - → **Ajouter commentaire pour plaque**
  - ⇒ **Projets**
- **Validation**

- ⇒ **Par plaque**
- ⇒ **Par numéro d'échantillon**
- ⇒ **Par numéro de tache ou de gel**
- ⇒ **Par gel**
- ⇒ **Par profil d'expression**
- ⇒ **Par expérience**
- ⇒ **Des ressoumissions**

➢ **Consultation**

- ⇒ **Des échantillons validés**
- ⇒ **Liste des gels**

➢ **Administration**

- ⇒ **Information instrument**
- ⇒ **Dernières identifications**
- ⇒ **Liste des paramètres**
- ⇒ **Journal des connexions**
- ⇒ **Statistiques**
- ⇒ **Relance des échantillons**

➢ **Aide**

- ⇒ **Gels-2d**
  - → **Ajouter des gels**
  - → **Éditer des gels**
  - → **Sacs**
  - → **Description des gels**
  - → **Imprimer des étiquettes de gels**
- ⇒ **Contrôles**
  - → **Contrôles et RD**
- ⇒ **Plaques**
- ⇒ **Redirections**
- ⇒ **Commentaires**
- ⇒ **Analyse des échantillons**
  - → **Outil d'analyse**

→ **Identification des protéines**

→ **Visualiser les échantillons**

⇒ **Relance d'échantillons**

⇒ **Consultation des identifications**

4.1 Entrée manuelle des données

4.1.1 Les gels

4.1.1.1 Entrée d'une nouvelle expérience

L'ajout d'une nouvelle expérience consiste à regrouper des gels de différents traitements, ainsi que leurs propriétés, dans un même ensemble (voir le formulaire d'ajout de la figure 4.2). Seuls les gels faisant partie d'une même expérience seront comparés entre eux. Il est impératif de remplir tous les champs du formulaire pour un traitement donné. La taille des gels sur l'axe des ordonnées est une valeur par défaut établie par le logiciel PDquest pour tous les gels de l'expérience. Elle se retrouve déjà dans un champ du formulaire mais elle est modifiable. Le champ « Traitement1 » spécifie le nom du traitement à entrer, qui est soit un contrôle, soit un traitement. Ce traitement est identifié par le numéro de Génome Québec. Ce champ contient le numéro de protocole du type "GQ..." ou le numéro de code à barres venant du lecteur optique. Ce code donne l'information sur le type de tissu, le sexe et le type de chirurgie. Les noms des gels appartenant à un traitement donné figurent dans le champ gels et sont délimités par le caractère "-". Le numéro du traitement et celui des gels associés sont obligatoires dans le formulaire de saisie d'un traitement. Le champ "Autre description" permet aux gels d'ajouter une information particulière ne se retrouvant pas dans la liste des fractions cellulaires. Ce champ informe, par exemple, que les gels appartiennent à un chercheur particulier ou que les échantillons ont été centrifugés à une vitesse particulière. Le pH de la bande utilisée pour la migration des échantillons sur le gel à deux dimensions est saisi dans le champ « pH-bande ». Les gels sont décrits de façon plus globale au moyen d'une liste déroulante de descriptions. De plus, les quantités de tissus et les fractions cellulaires sont aussi sélectionnées parmi les listes disponibles du formulaire. Par défaut, il est possible d'entrer, pour une expérience donnée, jusqu'à 5 traitements différents. La liste

des champs est donc toujours affichée en 5 copies. L'ordre des saisies de traitements est sans importance.

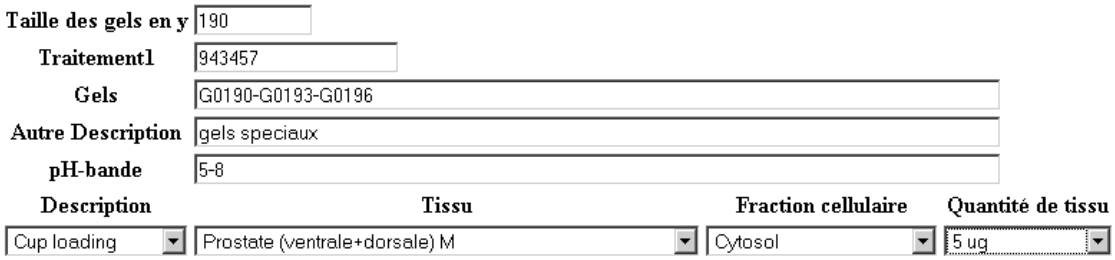


Figure 4.2 Écran de saisie pour l'ajout d'une expérience

4.1.1.2 Entrée de gels test

Les gels tests sont des gels qui ne sont associés à aucune expérience. Ces gels tests sont normalement créés pour apporter des modifications aux protocoles expérimentaux utilisés et en vérifier les effets. Ces gels ont un numéro GQ ou n'en ont pas. Les champs de l'interface de saisie sont les mêmes que pour les gels associés à une expérience (voir figure 4.2) sauf l'ajout du champ sur la taille des gels en abscisses. Ce champ n'est pas présent dans le formulaire d'ajout d'une expérience, puisqu'il est inconnu au moment de la préparation des gels. Par défaut, cette valeur est à 175.0 mm. Un gel test déjà inséré dans la base peut être associé par la suite à une expérience existante. Cette démarche est décrite dans la section « Editer les gels ».

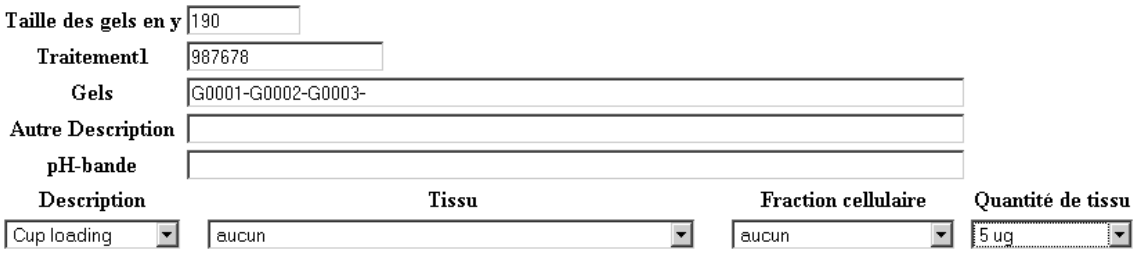


Figure 4.3 Interface d'ajout de gels test.

L'édition d'une expérience sert à apporter les modifications suivantes:

4.1.1.3 Modifier les informations des gels sur une expérience existante

La modification des gels d'une expérience se fait dans un formulaire accessible via le menu « Ajout->Gels->Éditer une expérience ». Pour ce faire, les noms de chaque gel à modifier doivent être entrés dans le champ gel. Le champ de saisie du traitement n'est pas obligatoire, puisqu'un traitement n'est pas spécifique à un gel donné et que l'édition de gels n'est pas spécifique à ceux provenant de traitements. Si des champs sont laissés vides, aucune information n'est modifiée concernant ces paramètres. Par contre, si le champ « taillex » est utilisé, il est nécessaire d'associer chaque taille à son gel, tel qu'illustré au formulaire de la figure suivante (4.4). Dans les listes déroulantes "Description", "Fraction cellulaire" et "Quantité de tissu", la valeur est sélectionnée par défaut à "inchangé". Si cette valeur n'est pas modifiée, aucun changement n'est apporté.

4.1.1.4 Modifier les informations de gels non associés à une expérience existante (gels tests)

La modification des gels tests se fait sur le même formulaire que la section précédente. Afin de modifier ce type de gel, la case "cocher ici si ce sont des gels test" doit être cochée afin d'indiquer au système de ne pas tenir compte des informations se rapportant aux expériences. Dans les listes déroulantes "Description","Fraction cellulaire" et "Quantite de tissu", le choix par défaut est "inchangé". Si ce choix n'est pas modifié, les valeures demeurent inchangées.

4.1.1.5 Associer des gels existants à une nouvelle expérience

Dans la situation où des gels tests doivent être associés à une nouvelle expérience, c'est-à-dire lorsque des gels doivent être regroupés dans un ensemble, la procédure d'édition utilise le même formulaire que la section précédente. La case à cocher "cocher ici pour associer les gels à une nouvelle expérience" doit être cochée. Les gels sont regroupés par traitement en prenant soin d'associer chaque groupe de gels au traitement correspondant ou bien de lire la bonne étiquette à l'aide du lecteur optique. Le nom du traitement devrait être inscrit dans le champ du traitement. Mais s’il ne l’est pas, le nom d'un gel appartenant déjà à l'expérience visée doit être inscrit en dernier dans le champ "gels". Le système va se charger de rapatrier le traitement appartenant à ce gel. Tous les champs ne sont pas requis pour associer les gels cibles à une nouvelle expérience, puisque ces informations peuvent déjà avoir été fournies par un usager lors de l'insertion de l'expérience dans la base.

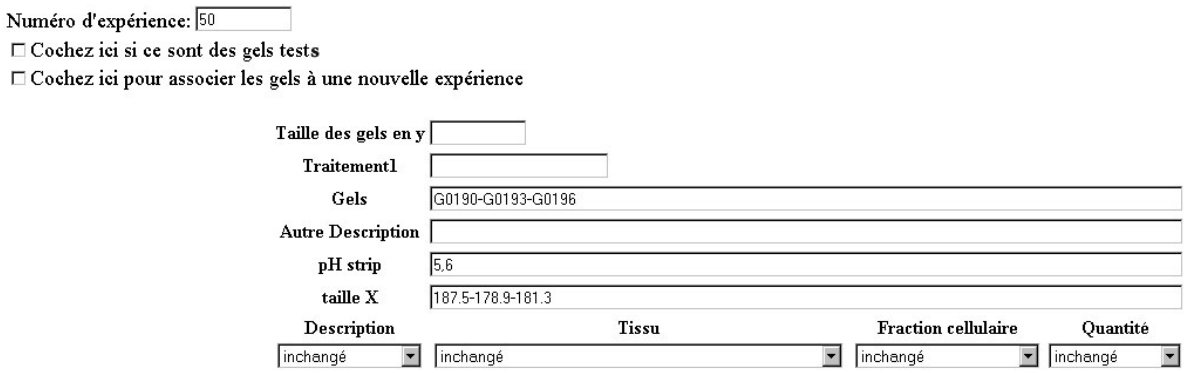


Figure 4.4 Interface d'édition des gels.

4.1.1.6 Associer des gels existants à une expérience existante

L'association de gels existants à une nouvelle expérience s'effectue dans le même formulaire illustré à la figure précédente. Lorsque des gels doivent être associés à une expérience existante, le numéro d'expérience est inscrit dans le champ de saisie "Numéro d'expérience". Ce numéro se trouve dans la liste des gels située au menu « Consultation->Liste des gels ». Le type de traitement pour chaque groupe de gels doit être saisi dans le formulaire. Si le traitement ne peut pas être inséré, le nom du gel doit apparaître en dernier dans le champ « Gels ». Tous les champs ne sont pas requis pour associer les gels cibles à une nouvelle expérience, puisque ces informations peuvent déjà avoir été fournies par l'usager lors de la saisie de l'expérience dans la base.

4.1.1.7 Modifier les informations sur un gel

L'interface utilisée pour modifier un gel à la fois est accessible via le menu « Consultation->Liste des gels ». Il suffit de cliquer sur le bouton « Détail » sur la ligne du tableau correspondant au gel à éditer. L'ouverture du formulaire d'édition est illustrée à la figure 4.5. Le bouton « Modifier » effectue la mise à jour des champs du formulaire dans la base.

4.1.1.8 Ajouter un Sac

L'information concernant l'entreposage des gels dans des sacs de plastique (section 3.3) est emmagasinée dans la base. Cette information facilite le repérage rapide des gels destinés au repiquage. Le repiquage est un nouveau prélèvement d'échantillons sur les gels.

Le formulaire d'ajout des sacs est disponible via le menu « Ajout-> Gels>Ajouter un sac ». Lors de l'ajout d'un nouveau sac, le système lui assigne un numéro d'identification. Un sac peut contenir jusqu'à 12 gels, et donc 12 cases sont disponibles pour enregistrer le nom de chacun des gels. Il n'est pas obligatoire que le sac contienne tous les 12 gels pour être validé. Le champ « destination » informe sur l'entreposage du sac. Il est aussi possible d'ajouter de nouveaux gels à un sac déjà existant, comme décrit dans la section suivante.

Lorsqu'une modification doit être faite sur un sac, le bouton « détail » du tableau des sacs de la figure 4.6 affiche une page intermédiaire dans laquelle il faut cliquer sur le lien "Modifier le sac". Le formulaire illustré à la figure 4.7 affiche le numéro du sac et les numéros des gels contenus dans ce formulaire. Les suppressions ou les insertions de gels sont validées en cliquant sur le lien "Insérer les gels".

L'impression d'étiquettes (voir Figure 4.8) se fait soit à l'insertion ou à la modification d'un sac, en cliquant sur le bouton "Détail" de la liste des sacs (voir figure 4.6), et puis sur le bouton "Étiquettes". Il suffit d'ajuster le nombre de copies désirées et une courte description devant être imprimée. L'impression se fera en cliquant sur le bouton "Imprimer". Il est possible d'imprimer ultérieurement les étiquettes des sacs déjà présents dans la base.

GEL G0555

NOM: LEHTAM01
DATE: 29-JAN-03
SAC: 63
TRAITEMENT: GQ024M03D01P1
TAILLEX: 175
TAILLEY: 190
DESCRIPTION:
PH: 7,10
Experience:
Description: Cup loading
Fraction cellulaire: Microsomes
Quantite de tissu: aucun
Tissu: Prostate (ventrale+dorsale) M
Modifier

Figure 4.5 Formulaire affichant les informations d'un gel.

28	G0209	G0210	G0211	G0212	G0213	G0214	G0215	G0216	G0217	G0218	-	-	DETAILS
29	G0201	G0202	G0203	G0204	G0205	G0206	G0207	G0208	-	-	-	-	DETAILS
30	G0183	G0185	G0186	G0187	G0188	G0190	G0191	G0192	G0194	-	-	-	DETAILS
31	G0184	G0193	G0195	G0196	G0197	G0198	G0199	G0200	-	-	-	-	DETAILS

Figure 4.6 Page HTML affichant la liste des sacs.

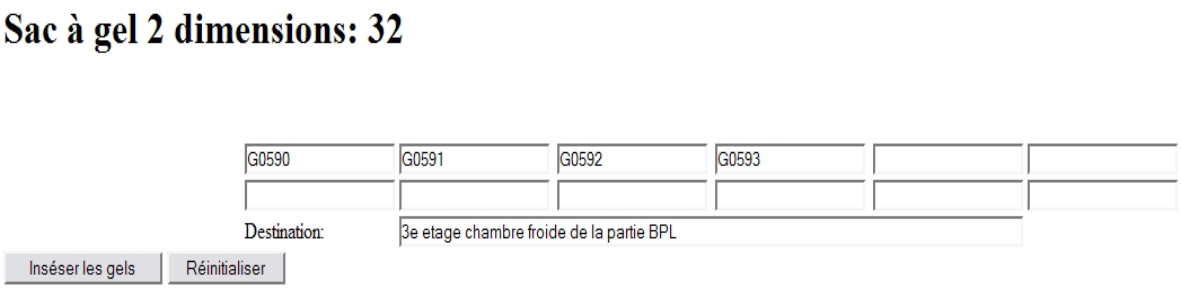


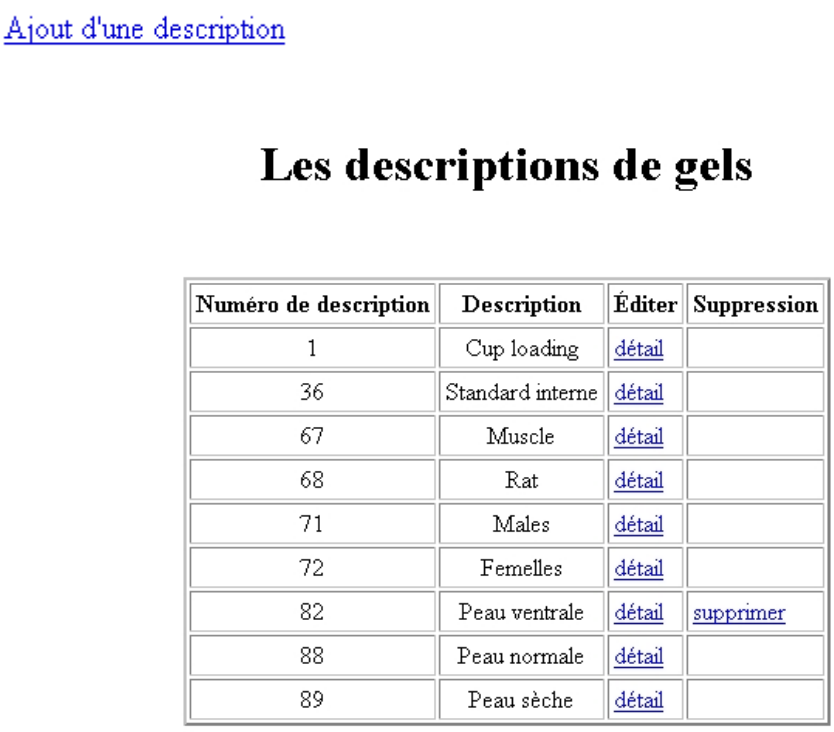
Figure 4.7 Formulaire d'ajout de gels pour un sac.

Figure 4.8 Formulaire d'impression d'étiquettes sur les sacs.

4.1.1.9 Ajout et visualisation des descriptions de gels

Les gels sont accompagnés d'informations sur le type de gel, la quantité de protéines, le type de tissu et la fraction cellulaire. Ces descriptions sont tirées des tables **Description_gels**, **Quantité**, **Tissu** et **Organelle**. Ces descriptions sont accessibles via le lien «Ajout->Gels->Description de gels » et sont affichées dans des tableaux. La figure 4.9 illustre un tableau de descriptions de gels provenant de la table **Description_gels**. Le numéro d'identification de la description, la description elle-même, un lien pour l'édition et un lien pour la suppression sont contenus dans le tableau. Une description ne peut pas être supprimée lorsqu'elle sert à décrire un gel de la base. Le lien de suppression ne s'affiche pas.

Pour modifier une description, on clique sur le lien « Ajout d'une description », tel qu'illustré dans la figure d'ajout et d'édition de descriptions (4.10), qui s'affiche par le lien "Ajout d'une description" à la page de la liste des descriptions (4.9). On inscrit la description dans le champ prévu et on indique le type de description en cochant l'une des trois options au-dessus du champ de saisie. En cliquant sur "Enregistrer", l'insertion s'effectue et les tableaux de descriptions s'affichent à nouveau. Pour éditer toute donnée, il faut cliquer sur le lien "détail" de la colonne « Éditer », puis cliquer sur "Enregistrer" sur le formulaire illustré dans la figure 4.10.


Ajout d'une description

Les descriptions de gels

Numéro de description	Description	Éditer	Suppression
1	Cup loading	détail	
36	Standard interne	détail	
67	Muscle	détail	
68	Rat	détail	
71	Males	détail	
72	Femelles	détail	
82	Peau ventrale	détail	supprimer
88	Peau normale	détail	
89	Peau sèche	détail	

Figure 4.9 Liste des descriptions de gels.

Figure 4.10 Ajout et édition des descriptions.

4.1.1.10 Imprimer des étiquettes de gels

L'impression d'étiquettes de gels se fait après la saisie ou l'édition de gels. Le menu « Consultation-> Liste des gels » affiche le formulaire illustré à la figure 4.11. Les noms des gels à imprimer sont saisis dans le champ « Imprimer des étiquettes pour les gels ». Il suffit ensuite de cliquer sur le bouton « Étiquettes ».

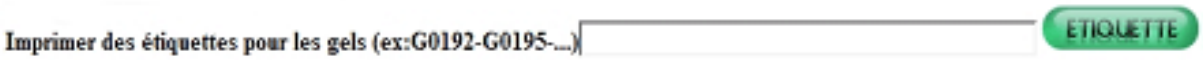


Figure 4.11 Formulaire d'impression d'étiquettes de gels.

4.1.2 Les échantillons de type contrôles

4.1.2.1 Ajout des contrôles

Les contrôles sont des échantillons servant surtout à vérifier la calibration des spectromètres de masse ou à vérifier des protocoles expérimentaux.

L'ajout de contrôles de type RD ou de type cAAA099A01 se fait de la même façon, mais sur des formulaires différents. La nomenclature du dernier type de contrôle est la

suivante : « c » pour contrôle, « AAA » spécifie sa composition, « 099 » correspond au numéro de la plaque et « A01 » informe sur la position sur la plaque. Dans le menu "Ajout->échantillons", deux liens : "Ajouter des contrôles" et "Ajouter des RD", tel qu'illustré sur la figure 4.12, permettent d'accéder à ces formulaires. Dans les deux cas, il faut sélectionner la composition du contrôle ou l'écrire dans le champ « Autre description », si ce choix n'existe pas. En positionnant le curseur de la souris sur une des compositions du contrôle, celle-ci s'affiche en mode détaillé dans une fenêtre bleue. On sélectionne ensuite une quantité ou on inscrit une nouvelle quantité dans le champ « Autre quantité ». Chaque nouvelle insertion apparaît dans la liste déroulante au prochain ajout. On inscrit finalement le numéro de plaque, sa position et on sélectionne la provenance de l'échantillon. Le bouton "Ajouter" valide l'ajout.

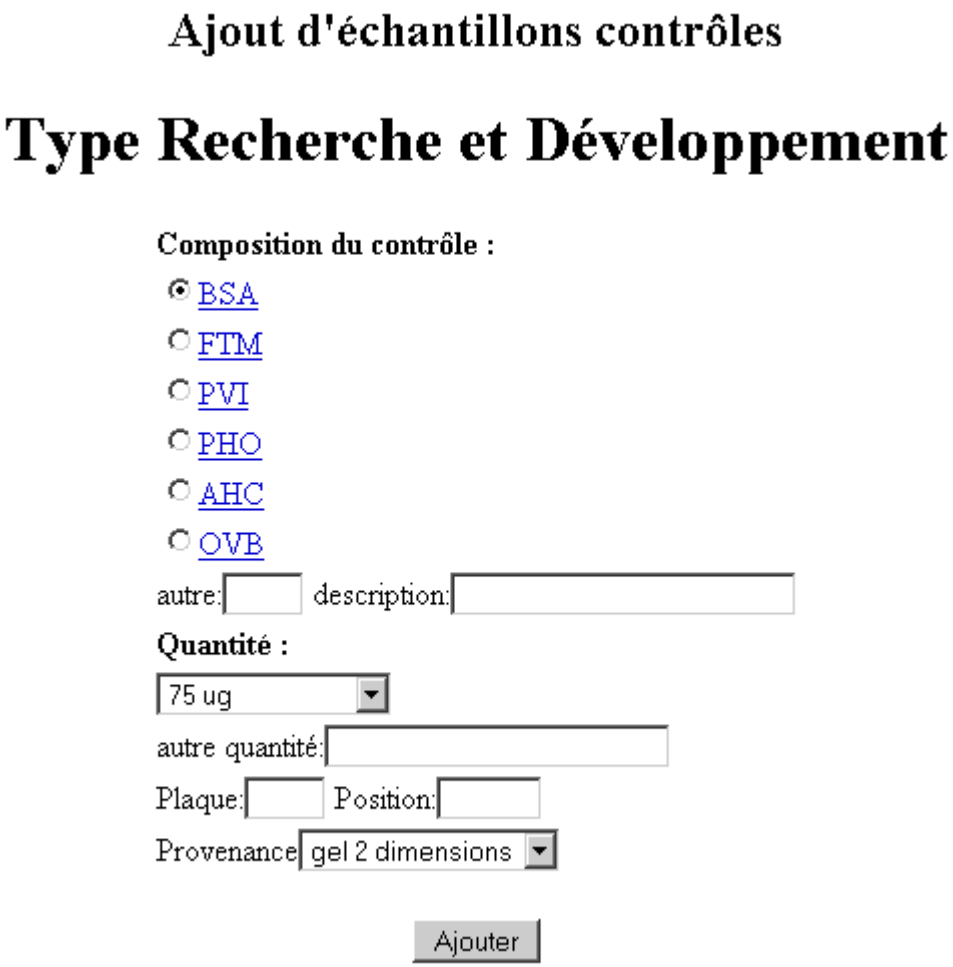


Figure 4.12 Formulaire d'ajout d'échantillon de type contrôles.

4.1.3 Les Plaques

4.1.3.1 Ajouter une plaque

La description des plaques dans la base assure un suivi de celles-ci et permet d'en connaître le contenu. Le tableau de la figure 4.13 illustre la liste des plaques disponibles de la base Murin et son accès se fait via le lien « Ajout->Plaques ». Ces plaques correspondent aux plaques de 96 puits utilisées après la digestion tryptique et contiennent les échantillons à être analysés par les spectromètres de masse. Dans ce tableau, on retrouve le numéro d'identification unique de la plaque, son titre, une description facultative et spécifique, le protocole, l'appartenance de la plaque à un protéome particulier et un lien menant vers l'information de la plaque et son édition.

Note: Le lien suivant vous conduira vers le prochain numéro de plaque. Si vous voulez annoter une nouvelle plaque alors cliquez sur ce lien:Annoter une plaque

1 à 100
101 à 200

Numéro de plaque	Titre	Commentaire	Protocole	Appartenance	Édition et Étiquette
101	TOV112D Ovarian Epithelial Carcinoma P1-P96		plaques 96 puits Plaques U-bottom : Costar Plaques PCR : ABgene 2.5 X moins de trypsine Digestion 1 heure à 58°C Absorbtion de la trypsine pendant 45 à 60 minutes à 4°C Extraction toute la nuit dans la solution d'extraction #1 Trypsine buffer fait a la main	Chercheurs	DETAILS
102	ingel 1581-1594		plaques 96 puits Méthode markers, spot, lavage Plaques U-bottom : Costar Plaques PCR : ABgene 2.5 X moins de trypsine Absorbtion de la trypsine pendant 45 à 60 minutes à 4°C digestion o.n. 37°C	Chercheurs	DETAILS
103	chercheurs 1595-		plaques 96 puits Plaques U-bottom : Costar Plaques PCR : ABgene 2.5 X moins de trypsine Digestion 1 heure à 58°C Absorbtion de la trypsine pendant 45 à 60 minutes à 4°C Méthode: Digestion-S_5.7_voyager_markers_spot_lavage_3trans	Chercheurs	DETAILS

Figure 4.13 Page HTML de la liste des plaques.

L'ajout d'une nouvelle plaque, dont le formulaire est illustré à la figure 4.14, se fait en cliquant sur le lien "Ajout d'une plaque" situé au haut de la liste des plaques de la figure 4.13. Ce lien mène alors vers la page du formulaire. Le numéro de plaque est non

modifiable. On peut y insérer le titre de la plaque dans le champ « Description », une description plus spécifique dans le champ « commentaire », ainsi que plusieurs autres propriétés plus standards en cochant les parties de protocoles appropriées. Afin de visualiser l'information concernant un commentaire de plaque, le curseur de la souris doit être positionné sur le commentaire de couleur bleu.

Figure 4.14 Ajout d'une nouvelle plaque.

4.1.3.2 Éditer une plaque

L'édition de plaque se fait au moyen du formulaire illustré à la figure 4.15. Ce formulaire s'affiche en cliquant sur le bouton « détail », de la plaque à éditer, situé dans le tableau de la liste des plaques (figure 4.13). Également, il est important de connaître la provenance des échantillons de la plaque. Par exemple, une plaque peut contenir des

échantillons provenant de chercheurs ou de projets locaux, comme celui de Génome Québec. Les champs « chercheurs » et « protéome variable » sont sélectionnés dans la liste. Tous les champs sont alors modifiables, excepté le numéro de la plaque, et ils seront mis à jour dans la base en cliquant sur le bouton « Modifier ».

4.1.3.3 Imprimer des étiquettes de plaques

L'impression d'étiquettes se fait en cliquant sur le bouton "detail" dans le tableau des plaques, et puis sur "étiquettes" de la page suivante. Le formulaire d'impression est illustré à la figure 4.16. Il suffit d'ajuster le nombre de copies désiré et d'inscrire une courte description destinée à apparaître sur l'étiquette. L'impression s'exécute en cliquant sur le bouton "Imprimer". Il est tout à fait possible d'imprimer des étiquettes de toutes les plaques présentes dans la base de données.

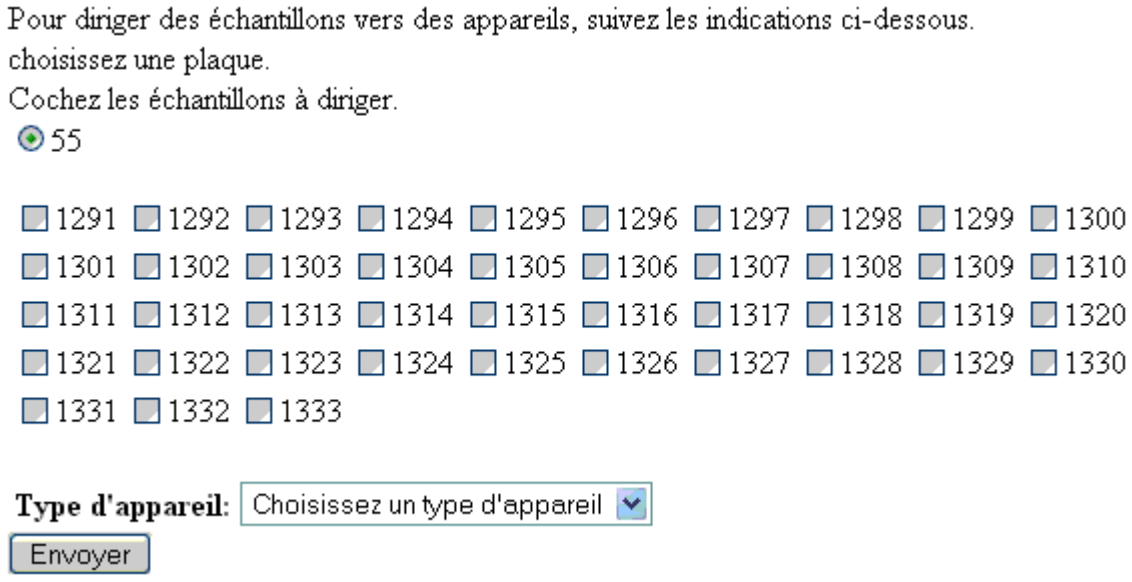


Figure 4.15 Formulaire d'édition d'une plaque.

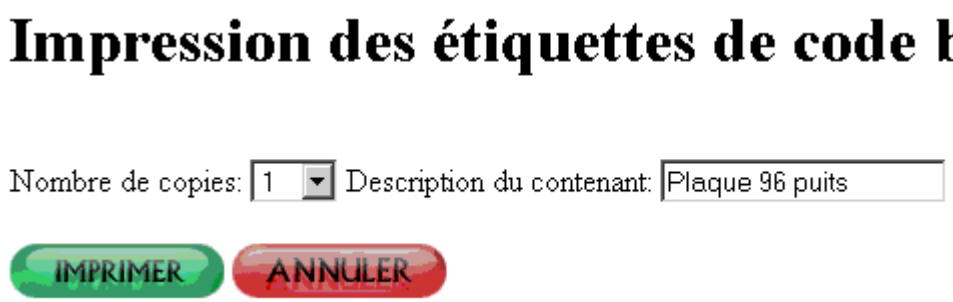


Figure 4.16 Formulaire d'impression d'étiquettes d'une plaque.

4.1.4 Les redirections

4.1.4.1 Afficher redirections des échantillons

Généralement, les échantillons de la plate-forme protéomique sont d'abord envoyés vers l'instrument MALDI-TOF, pour la production de spectres de masse. Toutefois, ces échantillons peuvent être re-dirigés vers d'autres appareils, comme le MS/MS (LCQ). Une séquence LCQ, telle que décrite précédemment, est une suite de paramètres propres à chaque échantillon et utilisés par le logiciel de l'appareil LCQ pour produire les spectres de masse. L'interface illustrée à la figure 4.17 permet de créer pour une plaque donnée une séquence qui sera utilisée par le spectromètre de masse en tandem. Le formulaire est disponible par le menu « Ajout->Redirections ». Dans ce formulaire, on sélectionne la plaque pour laquelle une séquence doit être générée; on choisit le spectromètre de masse pour lequel la séquence doit être générée; les champs "chemin" et "méthode instrument" permettent de changer les valeurs des champs "path" et "instrument method" dans le tableau de séquences.

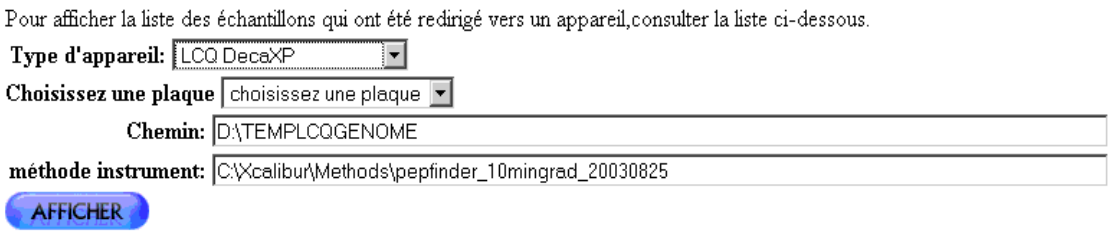


Figure 4.17 Sélection d'une plaque pour afficher les redirections.

La figure 4.18 illustre une séquence de la plaque 55. Si un échantillon doit être enlevé de la liste, il suffit de le supprimer en cliquant sur le lien "supprimer" correspondant à la ligne de l'échantillon. Pour qu'une séquence soit importée dans le logiciel Xcalibur, elle doit être sauvegardée dans un fichier de format texte. Le lien "Afficher en format texte", au haut du tableau, affiche à nouveau le tableau en format texte (Figure 4.19). Ce format doit être sauvegardé en allant dans le menu de la fenêtre du navigateur "fichier ->sauvegarder sous" et en l'identifiant d'un nom descriptif.

Redirection de la plaque 127

Afficher en format texte

File Name	Sample ID	Path	Instrument Method	Position	Inj Vol	Suppression
GQ2517	G08211105	D:\TEMPLCQGENOME	C:\Xcalibur\Methods\pepfinder_10mingrad_20030825	127:A01	5.0	Supprimer
GQ2518	G08216006	D:\TEMPLCQGENOME	C:\Xcalibur\Methods\pepfinder_10mingrad_20030825	127:B01	5.0	Supprimer
GQ2519	G08217407	D:\TEMPLCQGENOME	C:\Xcalibur\Methods\pepfinder_10mingrad_20030825	127:C01	5.0	Supprimer
GQ2520	G08217203	D:\TEMPLCQGENOME	C:\Xcalibur\Methods\pepfinder_10mingrad_20030825	127:D01	5.0	Supprimer
GQ2522	G08213301	D:\TEMPLCQGENOME	C:\Xcalibur\Methods\pepfinder_10mingrad_20030825	127:F01	5.0	Supprimer

Figure 4.18 Page HTML du tableau des redirections de la plaque sélectionnée.

```
Bracket Type=4
File Name/tSample ID/tPath/tInstrument/tMethod/tPosition/tInj Vol
2517    G08211105       D:\TEMPLCQGENOME        C:\Xcalibur\Methods\pepfinder_10mingrad_20030825        127:A01 5.0
2518    G08216006       D:\TEMPLCQGENOME        C:\Xcalibur\Methods\pepfinder_10mingrad_20030825        127:B01 5.0
2519    G08217407       D:\TEMPLCQGENOME        C:\Xcalibur\Methods\pepfinder_10mingrad_20030825        127:C01 5.0
2520    G08217203       D:\TEMPLCQGENOME        C:\Xcalibur\Methods\pepfinder_10mingrad_20030825        127:D01 5.0
2522    G08213301       D:\TEMPLCQGENOME        C:\Xcalibur\Methods\pepfinder_10mingrad_20030825        127:F01 5.0
2523    G08212107       D:\TEMPLCQGENOME        C:\Xcalibur\Methods\pepfinder_10mingrad_20030825        127:G01 5.0
2524    G08217105       D:\TEMPLCQGENOME        C:\Xcalibur\Methods\pepfinder_10mingrad_20030825        127:H01 5.0
2525    G08214503       D:\TEMPLCQGENOME        C:\Xcalibur\Methods\pepfinder_10mingrad_20030825        127:A02 5.0
2526    G08215501       D:\TEMPLCQGENOME        C:\Xcalibur\Methods\pepfinder_10mingrad_20030825        127:B02 5.0
2527    G08218002       D:\TEMPLCQGENOME        C:\Xcalibur\Methods\pepfinder_10mingrad_20030825        127:C02 5.0
2530    G08217303       D:\TEMPLCQGENOME        C:\Xcalibur\Methods\pepfinder_10mingrad_20030825        127:F02 5.0
2531    G08214301       D:\TEMPLCQGENOME        C:\Xcalibur\Methods\pepfinder_10mingrad_20030825        127:G02 5.0
2532    G08217807       D:\TEMPLCQGENOME        C:\Xcalibur\Methods\pepfinder_10mingrad_20030825        127:H02 5.0
```

Figure 4.19 Tableau d'une séquence de format texte.

4.1.4.2 Principe des redirections

Il existe deux façons de rediriger un échantillon. L'une est utilisée lors de la validation des échantillons et est expliquée dans la section 4.3 "Validation des résultats". La deuxième permet de rediriger une plaque dans son ensemble ou en partie via le formulaire de la section « Redirection immédiate » du menu « Ajout->Redirections » (figure 4.20). On sélectionne la plaque contenant les échantillons à rediriger et on clique sur le bouton « continuer ».

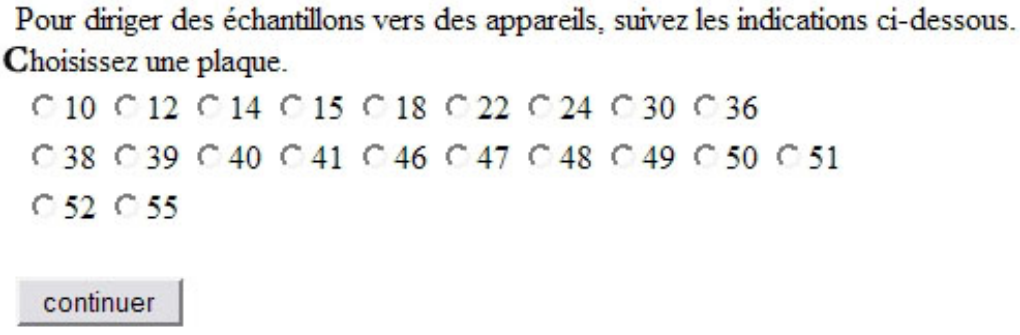


Figure 4.20 Formulaire de sélection d'une plaque à rediriger.

4.1.4.3 Redirection partielle

La figure 4.21 représente un exemple d'affichage des échantillons appartenant à la plaque sélectionnée. On coche les échantillons à rediriger vers le spectromètre de masse choisi dans la liste déroulante.

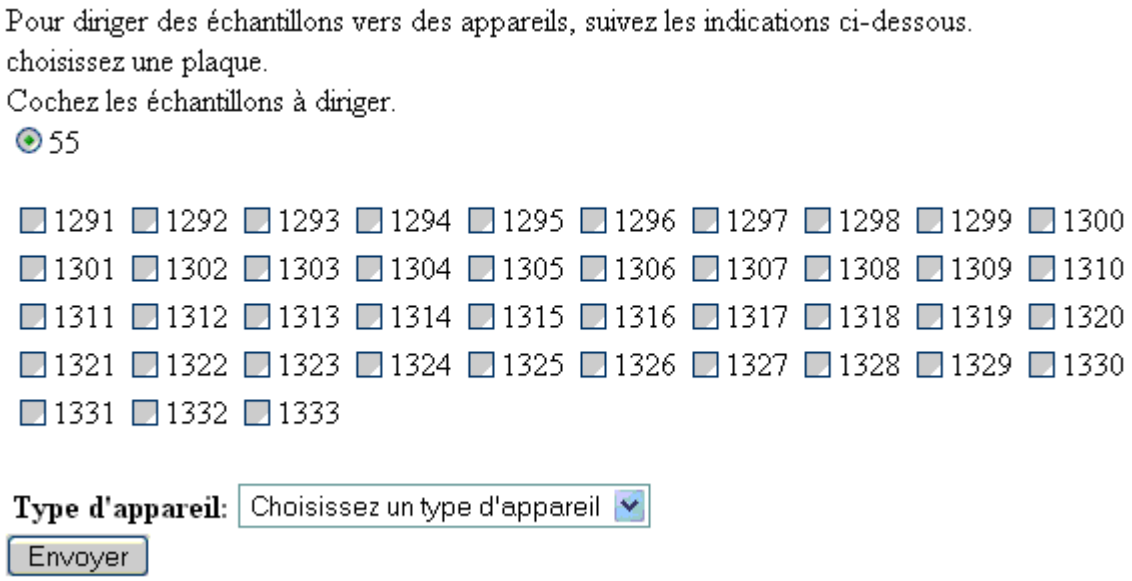


Figure 4.21 Formulaire de redirection partielle ou complète de la plaque.

4.1.5 Les commentaires

Les commentaires sont des outils permettant d'apporter des précisions sur les échantillons, les identifications protéiques et sur les plaques. Le même commentaire peut être assigné à plusieurs échantillons ou plaques. Les différents titres de commentaires peuvent être modifiés par les usagers. Les commentaires sont décrits ci-dessous.

LCQ DecaXP

commentaire échantillon		
Contamination	éditer	supprimer
Pas d'identification (LCQ)	éditer	supprimer
Carry-over	éditer	supprimer
commentaire protéine		
presence de contaminants	éditer	supprimer
Score identique pour la proteine homologue de la souris	éditer	supprimer
Bonne identification (plusieurs peptides de score moyen)	éditer	supprimer
Présence de kératines humaines	éditer	supprimer
Voir Mascot pour meilleure description	éditer	supprimer
Contamination	éditer	supprimer
Voir Sequest	éditer	supprimer
Excellente identification (LCQ)	éditer	supprimer
Seulement un peptide avec un bon score	éditer	supprimer
Corroboré avec Sequest	éditer	supprimer

Figure 4.22 Page HTML de la liste des commentaires d'échantillons et de protéines.

4.1.5.1 Commentaire concernant les échantillons

Les commentaires sont ajoutés lors de la validation des identifications protéiques. Ces commentaires qualifient la qualité de l'identification, fournissent des informations sur le spectre de masse ou expliquent la cause d'absence d'identification. Les commentaires décrivent l'échantillon dans son ensemble sans référer spécifiquement à aux résultats obtenus lors d’une recherche d'identification protéique.

4.1.5.2 Commentaire concernant les protéines

Les commentaires de protéines sont ajoutés lors de l'identification d'une ou des protéines. Ces commentaires diffèrent de ceux concernant les échantillons, car ils tiennent compte des résultats de l'identification de la protéine. Par exemple, l'identification de deux protéines dans un échantillon, dont l'une est un contaminant, doit être caractérisée par un commentaire concernant la protéine. Un tableau des commentaires pour les échantillons et les protéines est présenté à la figure 4.22.

4.1.5.3 Commentaire concernant les plaques

Un commentaire de plaques décrit le protocole utilisé dans la préparation des échantillons. Il est possible d'ajouter des commentaires spécifiques en plus des commentaires standardisés.

4.1.5.4 Ajout des différents commentaires

L'ajout de commentaires concernant les échantillons et les identifications protéiques se fait au moyen du formulaire illustré à la figure 4.23. Le formulaire servant à l'ajout des commentaires sur les plaques est sensiblement le même que le précédent à l'exception du choix de l'appareil. Le stockage du nouveau commentaire est effectué en cliquant sur le bouton "Enregistrer".

Figure 4.23 Formulaire d'ajout de commentaire concernant les échantillons et les protéines.

4.1.5.5 Édition des différents commentaires

L'édition ou la suppression des commentaires est possible via le lien « éditer les commentaires » des menus « Ajout->Commentaire->Commentaire pour échantillon », « Commentaire->Commentaire pour protéine » et « Commentaire->Commentaire pour plaque ». La page affiche un tableau, tel qu'illustré à la figure 4.22, sur les commentaires. La suppression d'un commentaire est impossible s'il est déjà utilisé dans la description d'échantillons, de plaques, ou d'identification protéiques. Lors d'une tentative de suppression d'un commentaire utilisé, un message d'avertissement s'affiche et la suppression est annulée. L'activation du lien "éditer" apparaissant dans la figure 4.22, affiche le formulaire illustré à la figure 4.24. Les informations déjà insérées dans la base sont alors « éditables » et mises à jour par le bouton "Enregistrer".

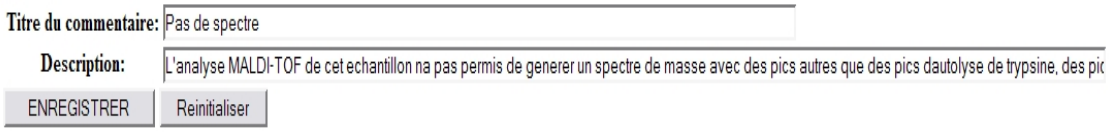


Figure 4.24 Formulaire d'édition d'un commentaire.

4.2 Analyses des données

4.2.1 Station de travail MALDI, LCQ, Proteomix

Les fichiers de spectres de masse sont soumis au moteur de recherche Mascot de deux façons. Il serait possible de soumettre de façon automatique les fichiers de spectres au démon de Mascot. Dans ce cas, chaque nouveau spectre de masse généré par un spectromètre serait transmis directement au moteur de recherche par son démon. Cependant, cette méthode n'a pas été choisie, puisque la qualité du spectre de masse n'est pas toujours acceptable pour être traitée par le moteur de recherche. Telle que mentionné précédemment, la qualité d'un spectre est établie en fonction du nombre de pics obtenus et de leur quantité. L'administrateur de la plate-forme désire plutôt vérifier la qualité de chaque spectre de masse

avant de les soumettre au moteur de recherche. Pour ce faire, nous avons utilisé la méthode qui suit.

4.2.2 Création d'une tâche démon

Le logiciel Mascot Daemon doit être préalablement installé à la station de travail, correctement configuré à l'ODBC installé sur la station. À l'ouverture du démon, celui-ci est fonctionnel aussitôt que sa connexion avec la base de données est opérationnelle. Pour lancer une recherche, il faut d'abord créer une séquence ou tâche « task ». Dans les différents menus principaux, illustrés à la figure 4.25, l'onglet « Task editor » est sélectionné. La tâche doit posséder un nom, une sélection de fichiers à envoyer et un fichier de paramètres. Ce fichier de paramètres est créé dans l'onglet "Parameter editor" où sont paramétrés la banque de protéines, les modifications peptidiques, la tolérance, et autres paramètres. Ces paramètres sont sauvegardés dans un fichier utilisable ultérieurement. Finalement, lorsque tous les choix sont terminés, on clique sur le bouton "RUN".

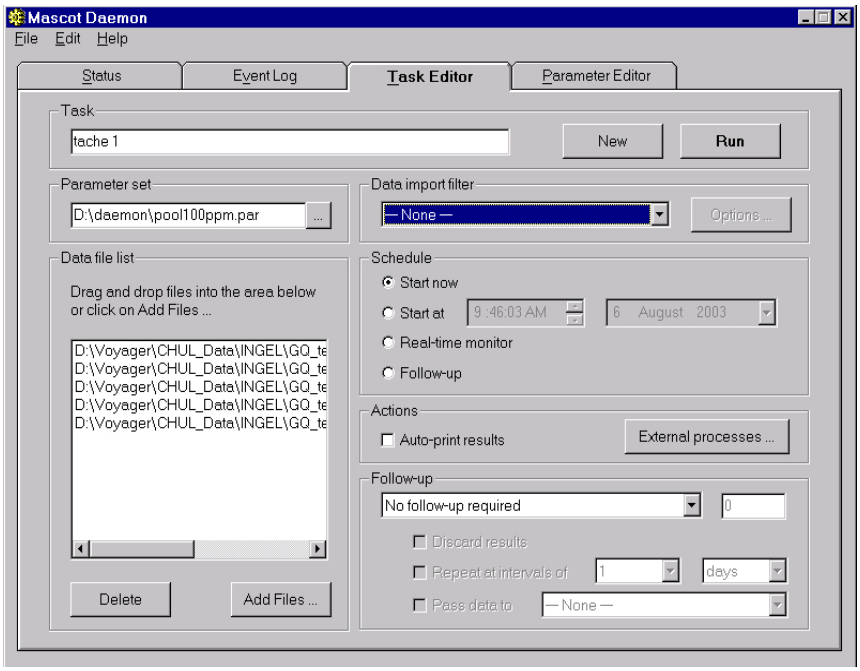


Figure 4.25 Formulaire de tâches du démon Mascot.

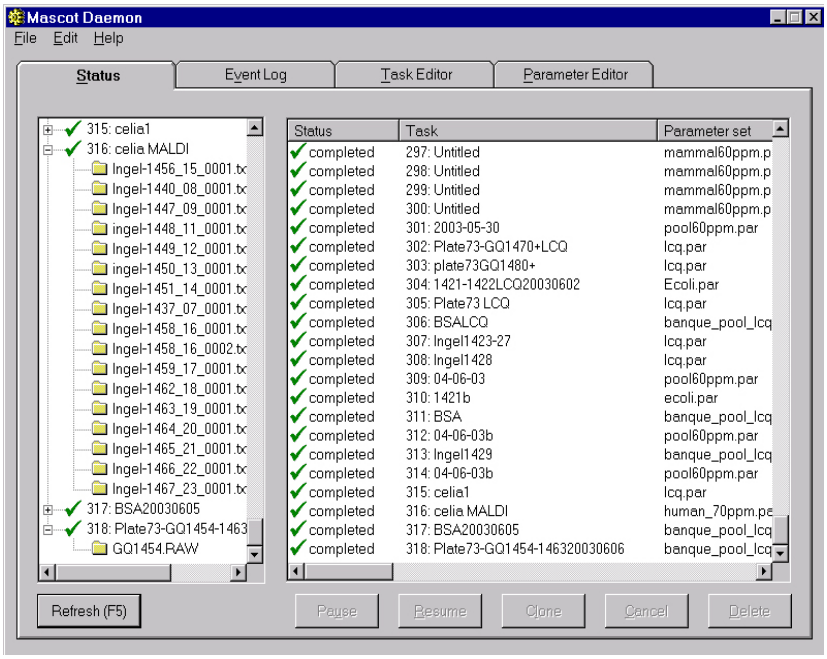


Figure 4.26 Journal de bord des tâches du démon Mascot.

4.2.3 Journal de bord des transactions de recherche

Lors du lancement d'une tâche, le journal de bord du démon fournit des indications sur la progression du processus de recherche. L'attribut "completed" apparaît (figure 4.26) vis-à-vis le numéro de la tâche lorsque la recherche est terminée. Le fichier ou résultat de la recherche est alors disponible dans la base Murin. Le menu "Envent log" donne accès au journal du démon.

4.2.4 Journal de soumission des tâches

Le journal de bord illustré à la figure 4.26 maintient à jour toutes les tâches lancées au moteur Mascot et qui sont contenues dans sa base de données. Pour chaque fichier soumis au moteur de recherche, un lien Internet est disponible dans la fenêtre des informations de la recherche afin de visualiser les résultats. Ce lien ouvre une page Mascot affichant les

résultats d'identifications protéiques, tel qu'illustré à la figure 4.27. Cette page peut être aussi accessible via la base Murin.

4.2.5 Résultats de la recherche

La consultation des données générées par Mascot (Figure 4.27) se fait dans la base Murin. Les fichiers de résultats Mascot sont des outils servant à valider les identifications protéiques. Pour accéder à la page d'affichage de la figure 4.27, il faut d'abord accéder à l'une des pages disponibles sous le menu « Validation » (figure 4.28). Par exemple, on saisit un échantillon ou on choisit une plaque et on clique sur le bouton « Afficher ». Une page s'ouvre et affiche un tableau sommaire des échantillons choisis (figure 4.29). On clique le nom d'un échantillon et ensuite sur le lien "consulter les identifications Mascot" au bas de la page. Les informations générales des identifications protéiques, sauvegardées dans la base par le démon, sont affichées (figure 4.27). Cependant, seule l'identification protéique la plus probable est présentée à l'écran. Il arrive que les bonnes identifications de protéines ne soient pas à la première position de la liste protéique. Afin de valider les résultats, nous avons dû modifier des scripts Perl fournis avec Mascot pour permettre l'insertion des autres identifications protéiques de la liste dans la base.

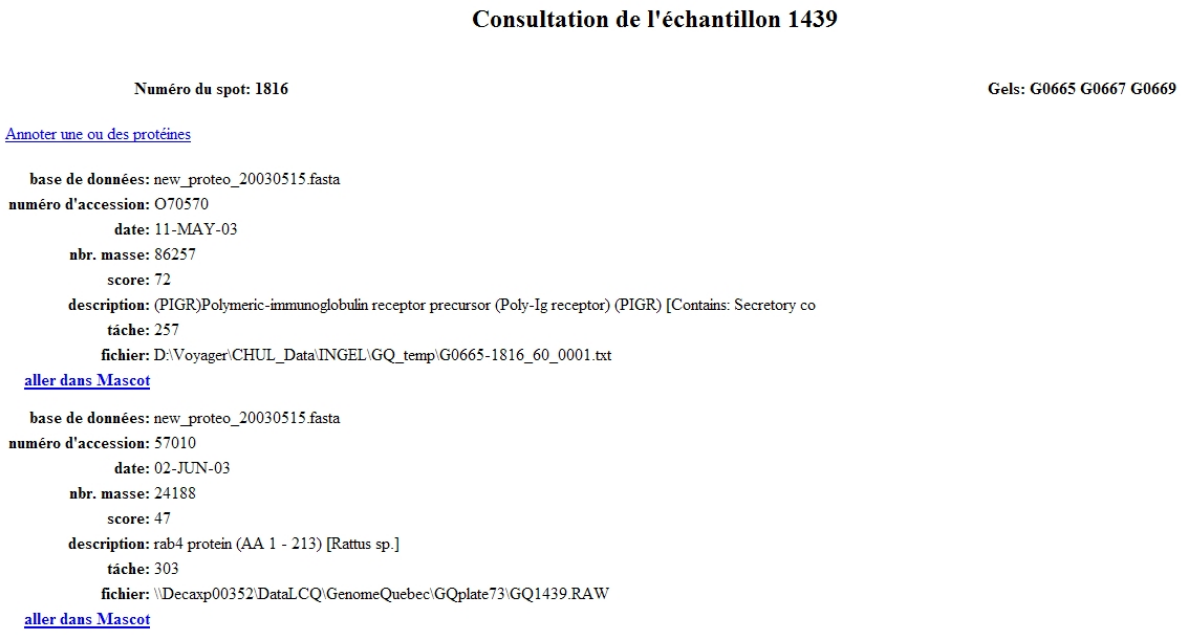


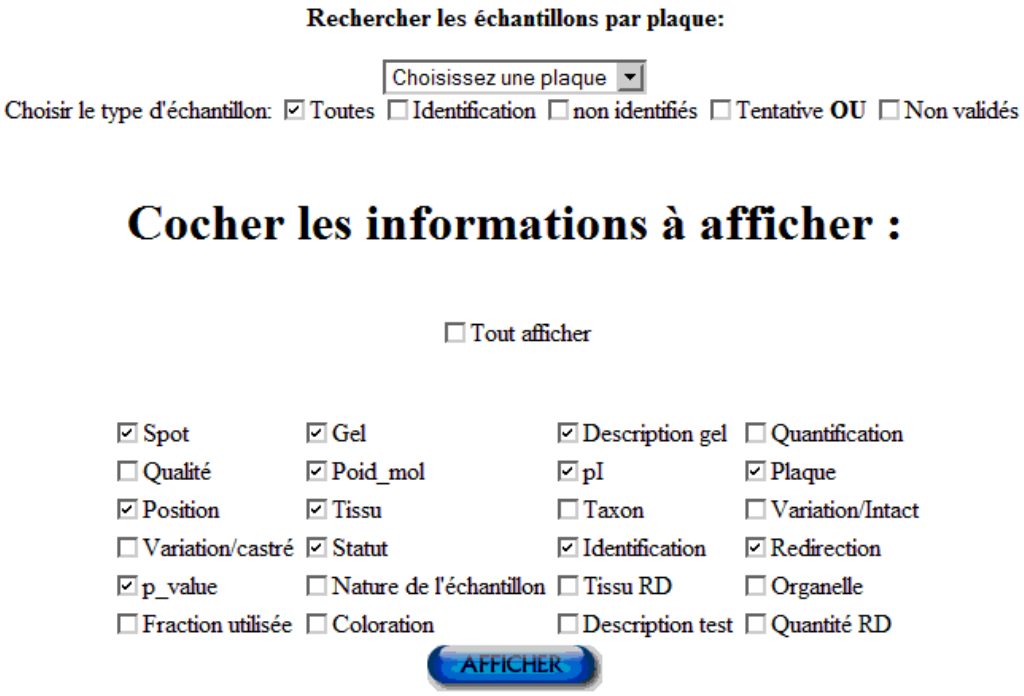
Figure 4.27 Page HTML Mascot affichant les résultats des identifications protéiques.

4.3 Validation des résultats

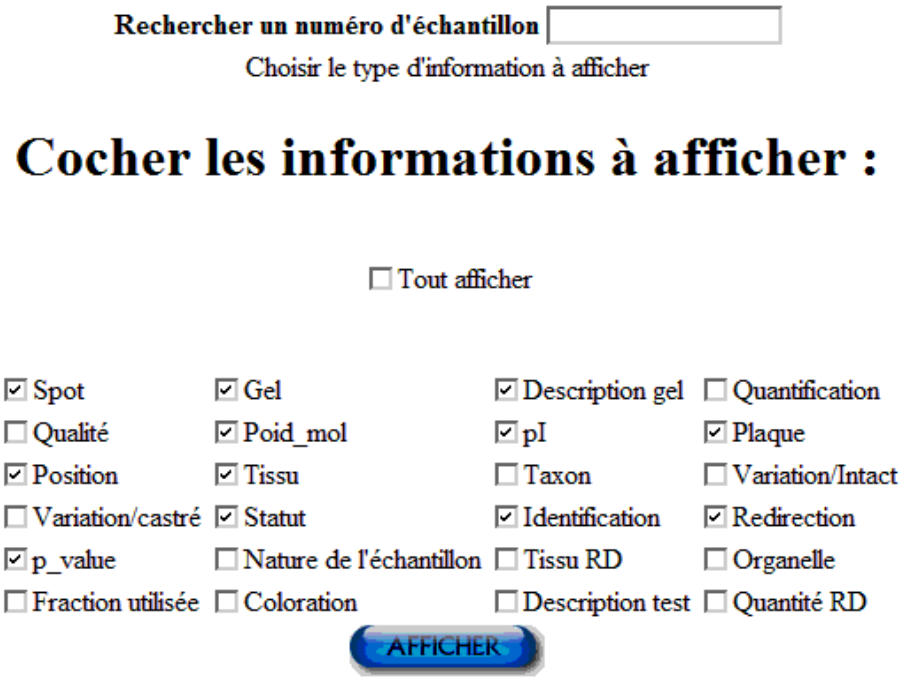
L'interface de consultation des échantillons regroupe les échantillons dans un tableau afin de mieux les valider (figure 4.29). Il y a deux façons d'accéder l'information d'un échantillon. La première, soit la plus utilisée, affiche les échantillons d'une plaque en la sélectionnant dans la liste à la page du menu « Validation->Par plaque » (figure 4.28). Les cases à cocher déterminent les colonnes du tableau à être incluses dans l'affichage final. Les échantillons peuvent être triés selon leur identification. La deuxième façon est d'accéder aux informations d'échantillon par l'affichage des images de gels. Toutes les taches de gels prélevées ont un lien de référence aux informations sur les profils d'expression et sur les informations de chaque échantillon. Cette dernière façon n'est pas souvent utilisée, car elle demande beaucoup plus de temps que la première dû aux temps de téléchargement des images et des pages d'informations qui s'affichent avant d'accéder è celle de la validation. Cependant, une visualisation des échantillons sur les gels est préférable lors d'une analyse plus approfondie (voir section 4.4.2). Les formulaires de saisie au bas de la figure ci-dessous

(4.28 a,b,c,d et e) paramètrent l'affichage des informations relatives à une série échantillons ou aux échantillons appartenant à un gel, à une tache, ou à un profil d'expression.

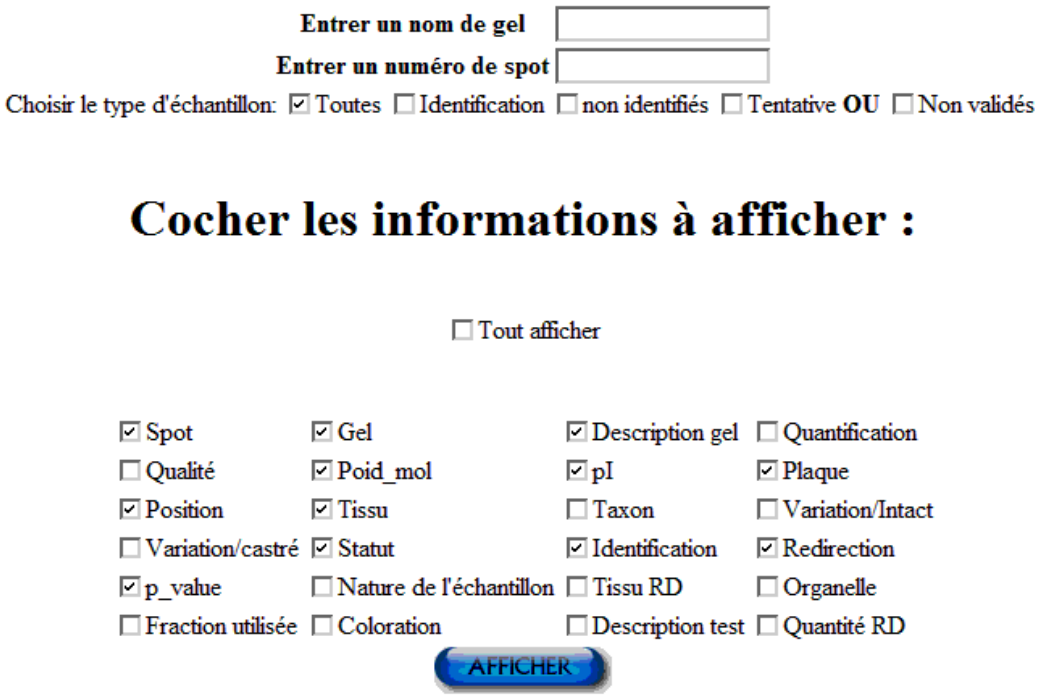
a)



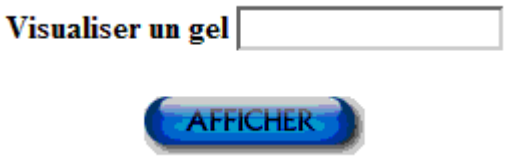
b)



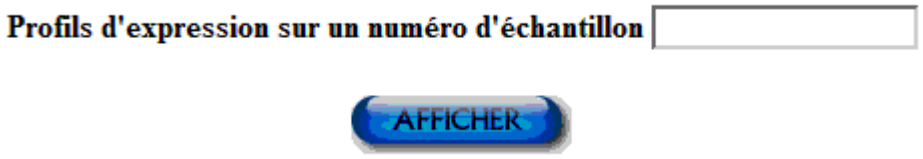
c)


Entrer un nom de gel
Entrer un numéro de spot
Choisir le type d'échantillon: Toutes Identification non identifiés Tentative OU Non validés
Cocher les informations à afficher :
Tout afficher
Spot
Gel
Description gel
Quantification
Qualité
Poid_mol
pI
Plaque
Position
Tissu
Taxon
Variation/Intact
Variation/castré
Statut
Identification
Redirection
p_value
Nature de l'échantillon
Tissu RD
Organelle
Fraction utilisée
Coloration
Description test
Quantité RD
AFFICHER

d)


Visualiser un gel
AFFICHER

e)


Profils d'expression sur un numéro d'échantillon
AFFICHER

f)

Figure 4.28 Formulaires de sélection des informations. Chacune des sous figures représente une façon d'accéder aux informations de la base.

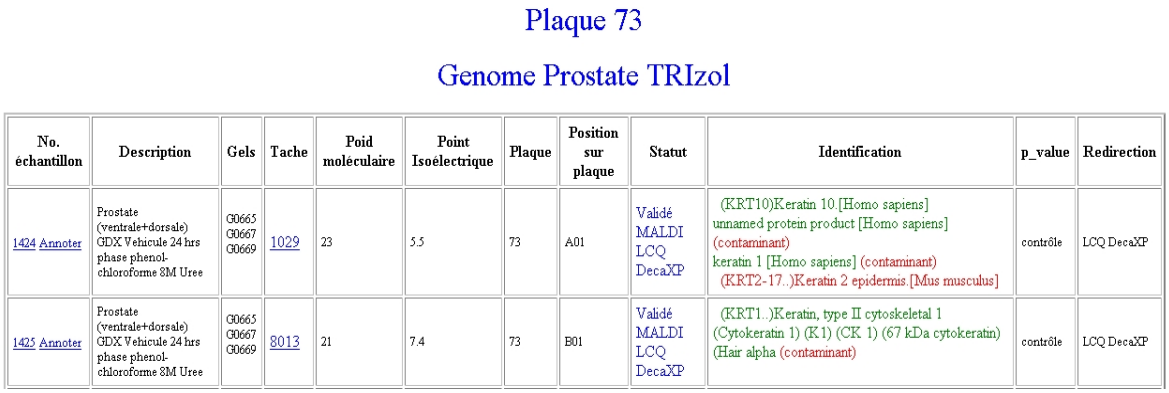


Figure 4.29 Tableau des échantillons.

4.3.1 Résultats de la recherche

La figure 4.29 illustre un tableau sommaire des analyses d'échantillons. Le tableau contient une partie des informations affichables résultant de la sélection des paramètres de la figure précédente 4.28. Il contient les informations suivantes :

- *Échantillon*

Chaque identifiant d'échantillon constitue un lien menant à des informations plus exhaustives que celles présentées dans le tableau (figure 4.30). De plus, le lien « Annoter » est un raccourci qui permet d'accéder directement à la page des annotations (figure 4.31).

- *Description*

La description du gel est l'information sur les gels fournie par un administrateur (ou opérateur) de la plate-forme des gels 2D. Ces descriptions concernent le type de tissu, la fraction cellulaire, la quantité et tout autre description spécifique. Dans le câdre du projet ATLAS, le tissu correspond au numéro GQ attribué par la plate-forme In vivo de chaque expérience. Ce descripteur suit chaque échantillon de chaque plate-forme faisant partie du projet. Cette colonne indique aussi le traitement appliqué, le temps de traitement et le sexe de l'animal.

- *Gels*

On y retrouve le numéro des gels d'où a été extrait l'échantillon.

- *Tache*

La tache est une zone définie sur le gel indiquant la présence de matière protéique. Cette matière est révélée au moyen d'un rayon ultra-violet suite à une coloration à l'argent des protéines. Un numéro unique par gel est attribué à la tache par le logiciel PDquest (tel que décrit à la section 3.1.1). L'échantillon peut être généré en utilisant la même tache sur différents gels. Or, le numéro est identique pour l'ensemble des gels appartenant à un même traitement d'une expérience. Le lien sur la tache mène à une page contenant les images de gels de l'expérience et les endroits où les taches ont été prélevées (figure 4.38).

- *Quantité*

La quantité réfère à l'intensité de la tache détectée par le logiciel PDquest. Plus l'intensité est élevée, plus la tache est opaque, plus elle renferme de la matière protéique. Cette quantité est

utilisée par le logiciel pour déterminer les coordonnées des taches sur le gel à être prélevées par le couteau à gels.

- *Qualité*

La qualité est fonction de l'uniformité da la tache selon la distribution d'une courbe de Gauss. Plus le pourcentage est élevé, plus la tache est uniforme et distincte. Par ailleurs, un faible pourcentage indique soit la présence d'un chevauchement de taches, soit un faible rendement de migration sur le gel.

- *Poids moléculaire*

Poids moléculaire expérimental calculé par le logiciel Pdquest de version 5 et plus.

- *Point isoélectrique*

Point isoélectrique expérimental attribué par le logiciel Pdquest de version 5 et plus.

- *Plaque*

Le numéro de la plaque à 96 puits utilisée par la plate-forme protéomique et contenant les échantillons.

- *Position sur plaque*

La position de chaque échantillon sur la plaque. La plaque est numérotée en rangées de A à H, et en colonnes de 1 à 12.

- *Source*

La source, champ absent sur la figure 4.29, identifie l'espèce d'où provient l'échantillon. Par défaut, tous les échantillons du projet ATLAS proviennent de la souris. Pour les autres échantillons, l'espèce doit être spécifiée.

- *Ratio /Intact*

Moyenne des valeurs de densité optique (D.O.) de la tache d'un traitement par rapport à la moyenne des D.O. de la tache du contrôle intact.

- *Ratio /Castré*

Moyenne des quantités de la tache d'un traitement par rapport à la moyenne des quantités de la tache du contrôle castré.

- *Statut*

La colonne « statut » précise le cheminement de l'échantillon. Ainsi, l'attribut "NON TRAITÉ" indique que l'échantillon présent dans la base n'a pas été soumis au moteur de recherche Mascot. Le libellé "Soumis à Mascot mais non validé", affiché en rouge, signifie que Mascot a fait l'analyse de cet échantillon, mais son identification n'a pas été confirmée. Dans le cas d'une validation, l'information est indiquée en bleu et l'instrument d'où provient l'échantillon est spécifié.

- *Identification*

Dans la colonne "Identification", le nom de la protéine apparaît en vert, si cette identification a été confirmée et est définitive, ou en rouge, s'il s'agit plutôt d'une présomption d'identification. Lorsqu'aucune identification n'a été trouvée, l'attribut "non" est indiqué.

- *Valeur statistique du seuil de signification*

La valeur de P est le « p-value » sur la variation de D.O. de la tache d'un traitement par rapport au sujet contrôle castré. La valeur du seuil de signification est obtenue par un test de t de Student et indique si la valeur du traitement est significative par rapport au contrôle.

- *Redirection*

La redirection sert à indiquer si un échantillon a été redirigé vers un second spectromètre de masse. Le champ contient le nom de l'appareil vers lequel l'échantillon a été envoyé et mentionne si l'identification protéique de ce dernier a été validée par la suite. Par défaut, tout échantillon est d'abord envoyé vers le MALDI-TOF. Par conséquent, l'échantillon redirigé est envoyé vers un autre type de spectromètre. Si tel est le cas, le nom de l'instrument est affiché en rouge. Lorsque l'identification protéique provenant de cet instrument est validé, la table Redirection est mise à jour et le nom de l'instrument apparaît par la suite en noir.

4.3.2 Information sur l'échantillon

Lorsqu'un échantillon est validé, ses informations sont affichées de la façon suivante (figure 4.30) : Chaque identification protéique est associée à un appareil de spectrométrie de masse. Les informations sont affichées dans deux tableaux. Le premier tableau affiche les informations globales sur l'échantillon : son identification, le nom du fichier de son spectre de masse, son poids moléculaire, ses numéros d'accession de la protéine associée, son statut de confirmation, etc. Le nom de la protéine dans ce premier tableau est inscrit tel qu'il apparaît dans la banque publique de protéines. Suivent les commentaires au sujet de l’échantillon. En positionnant le curseur de la souris sur le commentaire, on obtient la description détaillée de ce commentaire. Le deuxième tableau contient toutes les masses responsables de l'identification de la protéine. La masse expérimentale, la masse théorique, la différence entre ces deux masses et la séquence en acides aminés sont des champs servant à la validation de l'identification. Sous ce dernier tableau est affiché un hyper-lien pointant sur la page de Mascot contenant les résultats de la recherche de l'échantillon. Le lien au bas de la page « Annoter des protéines » (non illustré sur la figure 4.30) mène vers un écran de saisie de l'annotation des protéines identifiées par Mascot (figure 4.31).

MALDI

Commentaire d'échantillon:

Numéro de protéine	Accession GeneBank	Poids molculaire
714921	-	33899
Numéro Swiss prot	**Numéro gi**	**Numéro PIR**
Q64374	-	-
Score	**Date**	**Banque protéique**
46.1	27-AUG-02	proteo_nr_20020827.fasta
Confirmation	**Taxonomie**	**Origine**
oui	Mammalia (mammals)	MALDI
Identificateur		**Provenance de la base**
RON		sp

Description: (RGN..)Senescence marker protein-30 (SMP-30) (Regucalcin) (RC).[Mus musculus]

Commentaire de protéine:
Excellente identification
Graphique

Excellente identification
L'analyse MALDI-TOF de cet échantillon a generer un spectre de masse. Ce spectre a permis a l'outil de recherche dans les bases de donnes (Mascot) d'identifier une proteine et ceci avec un bon score de probabilite. De plus, tous, ou presque tous les pics de ce spectre ont pu etre attribues a la proteine identifiee.

masse	Debut	Fin	Delta	Sites non coupés	Nbr. ion concordant	Ion1		Séquence
1056.5563	265	274	-.01053	0	0	0	-	DGLNAEGLLR
1174.573	102	112	-.021338	1	0	0	-	FNDGKVDPAGR
1194.5603	214	223	-.019574	0	0	0	-	LWVACYNGGR
1231.5945	42	51	.005162	0	0	0	-	WDTVSNQVQR
1282.7244	52	64	-.029851	0	0	0	-	VAVDAPVSSVALR
1870.877	113	129	-.07607	0	0	0	- oxydation M	YFAGTMAEETAPAVLER

Consulter dans Mascot

Figure 4.30 Information sur un échantillon.

4.3.3 Validation d'un échantillon

La figure 4.31 illustre la page de validation d'un échantillon. Cette page est activée via le menu « Validation ». On choisit alors un type de recherche parmi les 6 premiers choix disponibles dans le formulaire illustré à la figure 4.28. On sélectionne une plaque ou on saisit l'information appropriée dans les champs des formulaires a, b et c. En cliquant le bouton « Afficher », puis sur le lien « Annoter », le formulaire de validation s'affiche avec tous les résultats protéiques de toutes les recherches effectuées par le logiciel Mascot sur cet échantillon. Chacune des sections des listes protéiques est identifiée par le URL (Uniform Resource Locator) du fichier de résultat, le nom du fichier de résultat donné par Mascot, sa date, le nom d'origine du fichier de spectre de masse, le numéro de résultat du fichier et la banque de protéines utilisée et identifiée de couleur rouge. Pour chaque liste de protéines, un lien "Aller dans Mascot" conduit à la page des résultats de Mascot. Chaque entrée protéique

est accompagnée de son numéro d'accession, sa masse théorique, son pointage (score), la date de soumission du fichier, son numéro de résultat dans la base et le numéro de la tâche du démon Mascot. Lorsqu'un fichier provient de l'instrument MALDI-TOF, son spectre de masse est visualisé en cliquant sur le bouton « Graphique » vis-à-vis chaque identification protéique. Les pics responsables de l'identification (voir la section 4.4) sont également mis en évidence. La validation d'une identification protéique se fait de la façon suivante :

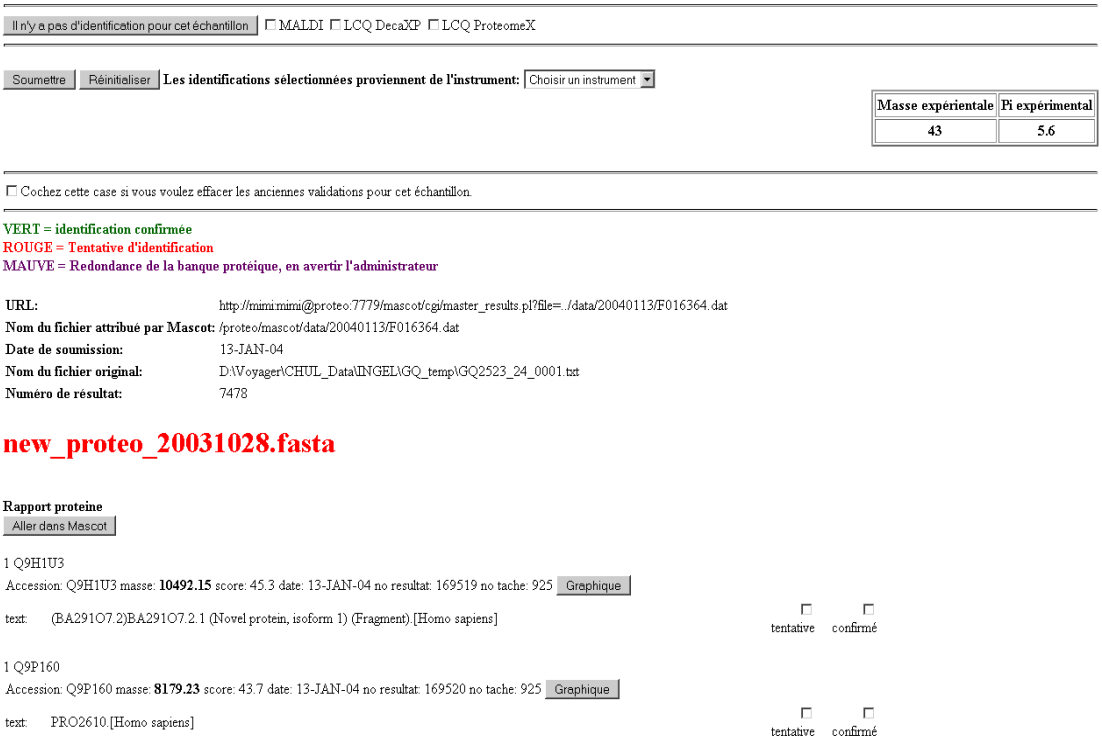


Figure 4.31 Validation des échantillons.

4.3.3.1 Protéine(s) identifiée(s)

Pour confirmer une identification, on sélectionne la ou les protéines par le type de confirmation, soit confirmée ou présomptive, affiché dans la liste des protéines (figure 4.31). Dans le formulaire de validation, on sélectionne, dans la liste déroulante située à côté du tableau de masse, le type d'instrument accompagnant la validation et on clique sur le bouton "Soumettre".

4.3.3.2 Ajout d'information supplémentaire (Identification)

Suite à la validation d'une identification de protéine (figure 4.32), une brève description de la protéine nouvellement ajoutée est affichée. Un commentaire, concernant soit l'échantillon, soit la protéine doit être sélectionnée dans la liste de choix possibles. Les informations complètes de chaque commentaire sont disponibles en positionnant le curseur de la souris sur le titre du commentaire. Ces commentaires sont sélectionnés à l'aide des cases à cocher. Au besoin, l'usager redirige l'échantillon vers un autre appareil au moyen du formulaire "Redirection facultative" (non illustré) présent dans cette page. Cette redirection informe que l'échantillon doit être réanalysé par l'appareil ciblé. L'ajout des commentaires et de la redirection s'exécutent en cliquant sur le bouton "Continuer".

4.3.3.3 Protéine(s) non identifiée(s)

La non identification de protéines pour un échantillon doit être confirmée par un usager. Une confirmation de l'usager est nécessaire pour indiquer au système qu'un échantillon n'a pas été identifié. Pour ce faire, l'usager accède à l'écran illustré à la figure 4.31. Pour valider, l'usager sélectionne l'appareil d'où le spectre de masse a été généré et clique sur le bouton « Aucune identification ».

4.3.3.4 Ajout d'information supplémentaire (Protéine non identifiée)

Suite à la validation d'un échantillon, l'ajout d'un commentaire concernant l'échantillon n'est pas nécessaire, sauf dans le cas d'une identification. L'usager redirige, s'il le désire, l'échantillon non identifié.

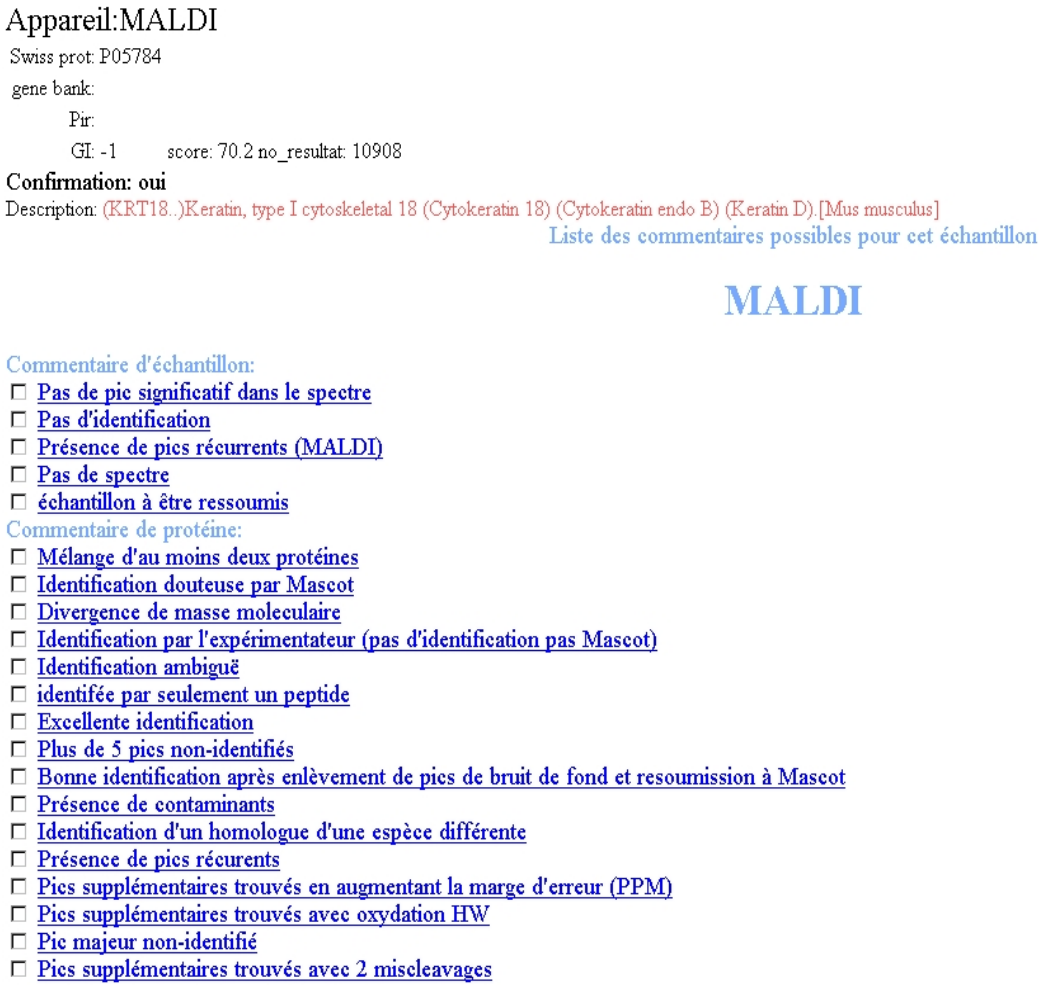


Figure 4.32 Ajout des commentaires de l'échantillon (Identification).

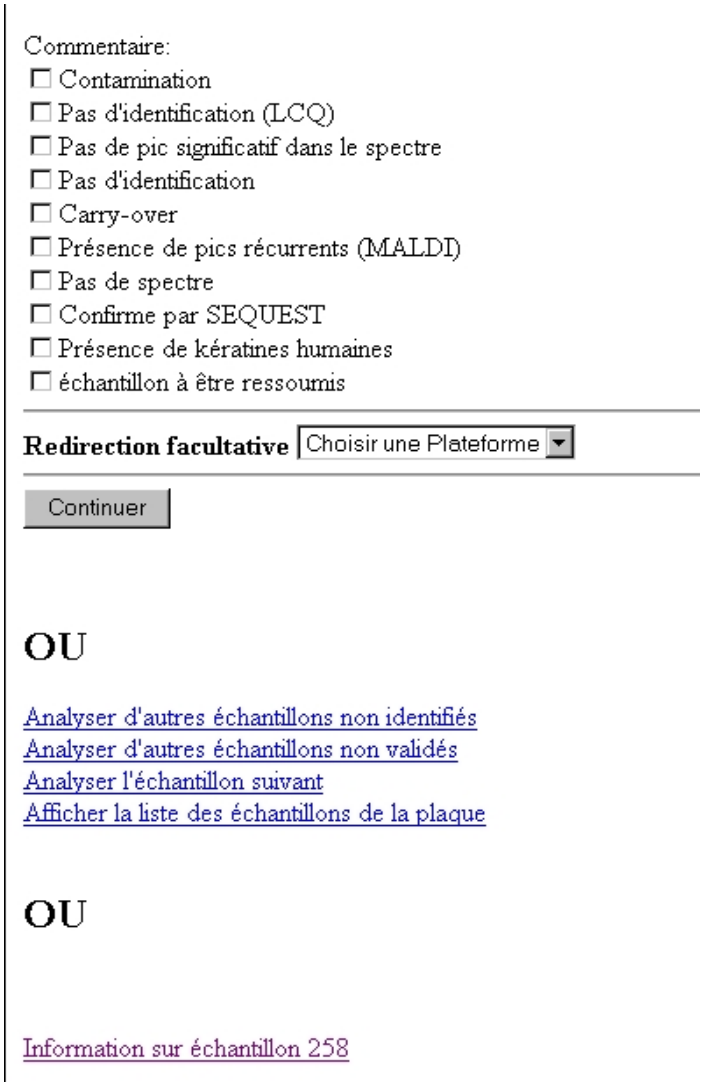


Figure 4.33 Ajout des commentaires et redirection facultative de l'échantillon (pas d'identification).

4.3.3.5 Différents liens utiles

Après l'insertion des commentaires (figure 4.32 et 4.33), une page affiche la confirmation des insertions et une liste de liens (figure 4.34). Le lien "Analyser d'autres échantillons non identifiés" affiche la liste des échantillons de la plaque n'ayant aucune

identification, alors que le lien "Analyser d'autres échantillons non validés" affiche les échantillons de la plaque traités par Mascot, mais dont leurs identifications protéiques n'ont pas encore été validées. Le lien « Analyser l'échantillon suivant » affiche les informations sur l'échantillon suivant le dernier échantillon accédé. Finalement, « Afficher la liste des échantillons de la plaque » affiche tous les échantillons de cette plaque.

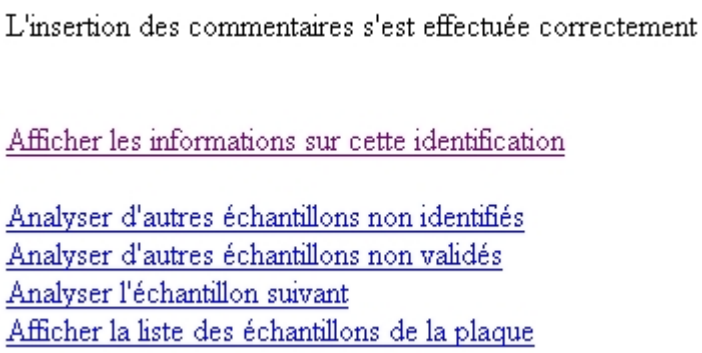


Figure 4.34 Liens utiles post validation.

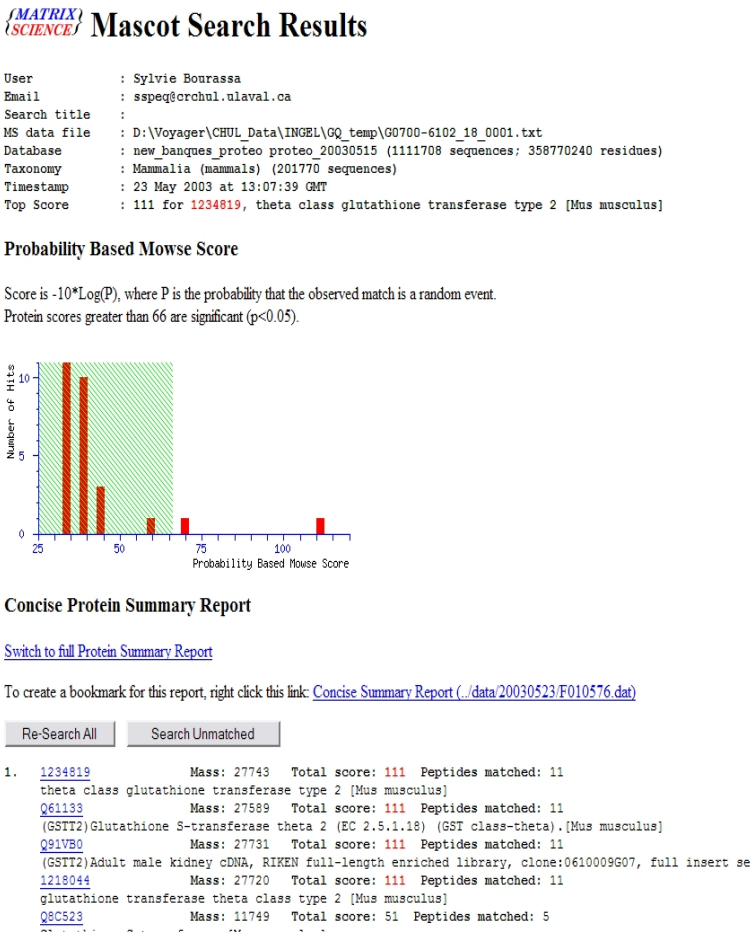
4.4 Visualisation des données

Le moteur de recherche Mascot analyse les fichiers de spectre de masse qui lui sont soumis et produit un fichier de résultats pouvant être lu par un navigateur. Des liens sont disponibles dans les écrans de validation ou de consultation des données pour accéder au contenu du fichier Mascot (figure 4.35). En haut de la page apparaissent les informations telles que le nom du fichier de spectre de masse, l'usager qui a soumis la recherche, etc. Le graphique représente la probabilité que l'identification soit aléatoire. Ainsi, les identifications au-delà de la zone carrelée ont une valeur où le seuil de signification est de 0.05 ou moins. Sous ce graphique est affichée une liste contenant les identifications protéiques. Plus d'une identification protéique peuvent avoir la même probabilité d'identification. Cette situation peut provenir d'un mélange protéique. Chaque identification est accompagnée d'un lien menant aux informations suivantes : le pourcentage de couverture de la séquence protéique, le point isoélectrique calculé et les masses théoriques et expérimentales concordantes. Ces

informations sont enregistrées dans la base de données Murin, et sont accessibles dans les deux tableaux associés à l'identification validée (voir figure 4.30).

4.4.1 Spectres de masse

Les spectres de masse des échantillons analysés par le MALDI-TOF sont affichables lors de la validation ou de la consultation grâce à un applet Java. L'applet est accessible via un lien situé sur la page affichant la liste de toutes les entrées protéique de l'échantillon (figure 4.31). Ce graphique (figure 4.36) identifie les masses expérimentales, en rouge, ayant permis l'identification de la protéine. Le nom de l'identification figure également sur le graphique. Le spectre est agrandi en sélectionnant le pourcentage d'agrandissement dans la liste déroulante. Il affiche aussi la valeur des masses et des intensités. La visualisation de ce spectre aide considérablement à la validation des identifications des protéines.



MATRIX SCIENCE **Mascot Search Results**

User : Sylvie Bourassa
Email : sspeq@crchul.ulaval.ca
Search title :
MS data file : D:\Voyager\CHUL_Data\INGEL\GQ_temp\G0700-6102_18_0001.txt
Database : new_banques_proteo proteo_20030515 (1111708 sequences; 358770240 residues)
Taxonomy : Mammalia (mammals) (201770 sequences)
Timestamp : 23 May 2003 at 13:07:39 GMT
Top Score : 111 for 1234819, theta class glutathione transferase type 2 [Mus musculus]

Probability Based Mowse Score

Score is -10*Log(P), where P is the probability that the observed match is a random event.
Protein scores greater than 66 are significant (p<0.05).

Concise Protein Summary Report

Switch to full Protein Summary Report

To create a bookmark for this report, right click this link: Concise Summary Report (../data/20030523/F010576.dat)

Re-Search All | Search Unmatched

1. 1234819 Mass: 27743 Total score: 111 Peptides matched: 11
theta class glutathione transferase type 2 [Mus musculus]
Q61133 Mass: 27589 Total score: 111 Peptides matched: 11
(GSTT2)Glutathione S-transferase theta 2 (EC 2.5.1.18) (GST class-theta).[Mus musculus]
Q91VB0 Mass: 27731 Total score: 111 Peptides matched: 11
(GSTT2)Adult male kidney cDNA, RIKEN full-length enriched library, clone:0610009G07, full insert se
1218044 Mass: 27720 Total score: 111 Peptides matched: 11
glutathione transferase theta class type 2 [Mus musculus]
Q8C523 Mass: 11749 Total score: 51 Peptides matched: 5

Figure 4.35 Page des résultats de recherche de Mascot.

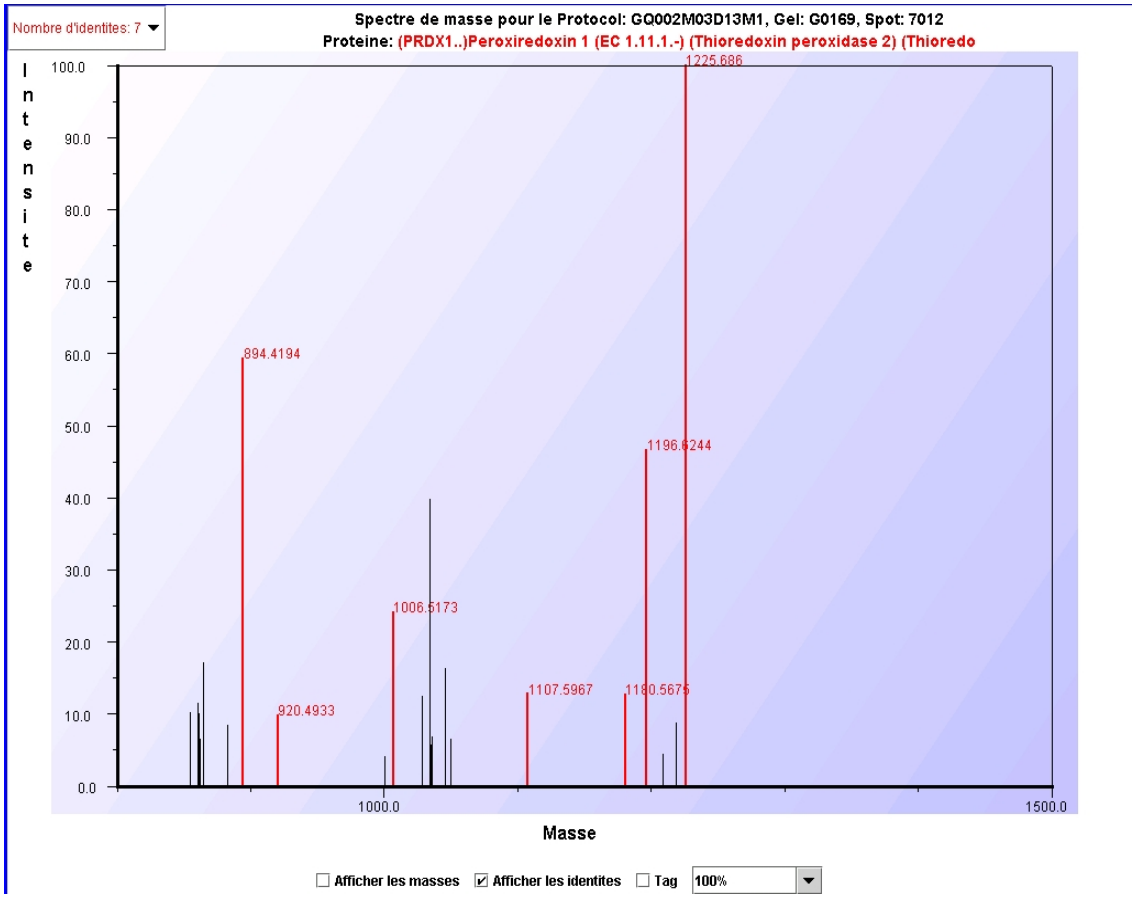


Figure 4.36 Applet JAVA du spectre de masse d'un échantillon.

4.4.2 Les gels

Dans le formulaire généré par l'item du menu « Validation->Par gel », et illustré à la figure 4.28 d, le champ "Visualiser un gel" affiche l'image de ce gel (figure 4.37). La sélection d'une tache ou du numéro d'échantillon se fait au moyen de la saisie dans les champs « Échantillon » et « Spots ». L'image du gel est agrandie à différentes résolutions via la liste de sélection « Zoom ». La figure 4.37 illustre l'image du gel G0189 contenant tous les prélèvements ayant servi à générer les échantillons.

La visualisation simultanée de tous les gels permet de vérifier la qualité des gels et celle de la migration des protéines. Cette visualisation est générée par l'interface « annonce_experience », accessible au menu "Validation->Par expérience", en cliquant sur une expérience dans la section « Visualiser les gels par expérience » de la figure 4.28. Cette interface est très similaire à l'écran « Visualiser un gel » présenté précédemment (figure 4.37 et 4.38). Les gels sont regroupés par traitement. Les cercles autour des taches peuvent avoir des couleurs différentes. La couleur est fonction du ratio de la moyenne de l'intensité des taches situées au même endroit sur les gels provenant de traitements différents. Le traitement de référence est sélectionné en cliquant sur le nom du traitement choisi. Ainsi, d'après la légende au bas du graphique, une tache de couleur rouge indique que la moyenne de l'intensité de cet échantillon est d'au moins 5 fois supérieure à celle du traitement de référence. Si une tache est sélectionnée, à partir de la liste des taches, tous les échantillons de l'expérience associés à ce même numéro de tache sont sélectionnés Cette étape permet de comparer rapidement les intensités des taches entre les traitements. Dans l'exemple de la figue 4.38, l'échantillon 338 possèderait une expression protéique plus élevée que l'échantillon 413. En pointant sur la zone 318, des moyennes d'intensités sont affichées. L'agrandissement des images gels, à partir de la liste des valeurs de grandeur, affiche les détails des taches. Certains noms de gels sont inscrits en rouge et d'autres, en noir. Seuls ceux inscrits en rouge ont été choisis pour représenter le traitement. Les valeurs d'intensités et les numéros de taches ont alors été normalisés à partir de ces gels.

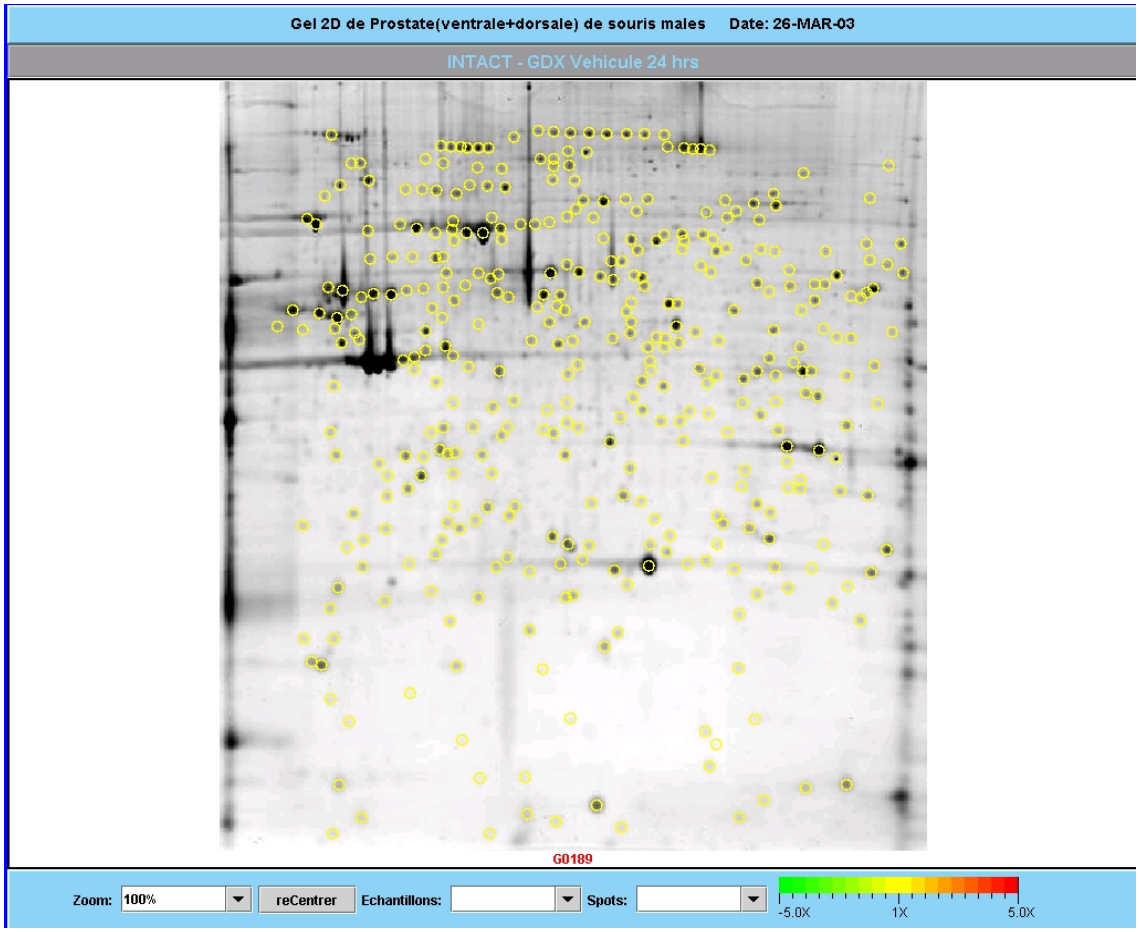


Figure 4.37 Applet JAVA affichant un image gel et localisant des zones de prélèvement.

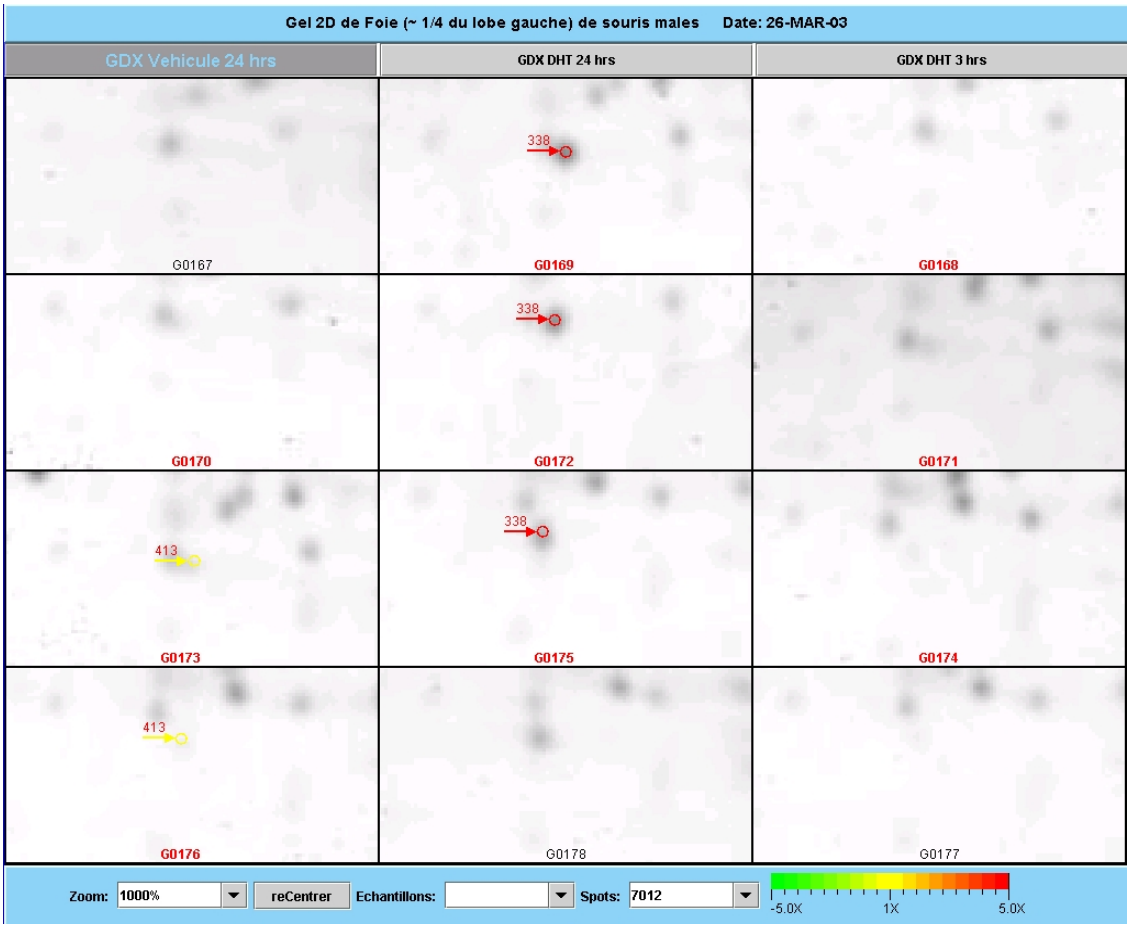


Figure 4.38 Applet JAVA sur la visualisation des gels d'une expérience.

4.4.3 Les Profils d'expression

En cliquant sur le lien de la tache (figure 4.37 et 4.38), la page de comparaison des profils d'expression s'affiche (figure 4.39). Pour chaque numéro de tache, les échantillons sont affichés par traitement avec leurs informations. La valeur du seuil de signification est calculée entre un traitement et son contrôle. Cette valeur est obtenue par un test de t de Student tel que mentionné précédemment. Les lignes en rouge, dans le tableau illustré dans

la figure 4.39, indiquent que ces taches ont été analysées par spectrométrie de masse. Le graphique illustre un histogramme représentant la variation de l'intensité de la tache pour chaque traitement. La ligne verticale sur chaque colonne représente l'erreur type. Les colonnes de couleur rouge sont des traitements pour lesquels la protéine a été identifiée, alors que les colonnes bleues démontrent une absence d'identification. Le traitement sélectionné sert arbitrairement de contrôle pour le pourcentage du ratio des intensités. En cliquant sur le nom d'un traitement, celui-ci devient le traitement contrôle.

Échantillon	Tache	Gel	Quantification normalisée	Moyenne	Valeur p
INTACT - GDX Vehicule 24 hrs					
-	2303	G0286	3897.1665	4577.17	-
-	2303	G0287	3343.6099	-	-
-	2303	G0288	6490.7197	-	-
GDX Vehicule 24 hrs					
-	2303	G0289	6336.6855	5895.65	-
-	2303	G0290	4464.8672	-	-
581	2303	G0291	6885.4067	-	-
GDX DHT 24 hrs					
-	2303	G0294	15458.3115	12476.45	-
-	2303	G0295	8888.6904	-	-
575	2303	G0296	13082.335	-	.061

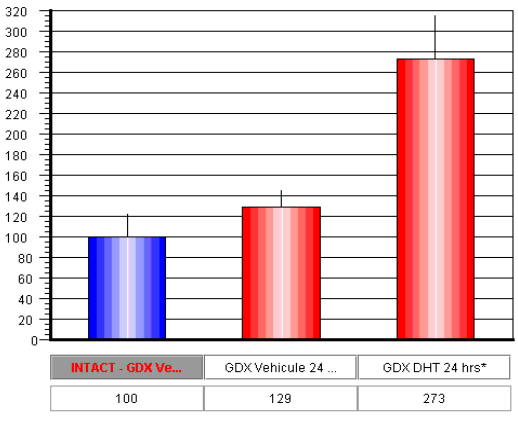


Figure 4.39 Applet JAVA sur la comparaison des profils d'expression de l'expérience.

4.5 Relance d'échantillons

La relance d'échantillons est un outil pour resoumettre avec les paramètres originaux des fichiers crées par mascot et provenant d'échantillons n'ayant généré aucune identification. Puisque les banques de protéines sont régulièrement mises à jour, il est intéressant de resoumettre ces échantillons en utilisant des banques protéiques actualisées. Dans le menu « Administration->Relance des échantillons », on choisit l'appareil vers lequel les resoumissions seront destinées (figure 4.40). Chaque identification protéique provenant

d'un spectromètre de masse doit être validée. Or, une identification validée par plus d'un spectromètre de masse augmente la fiabilité du résultat.

4.5.1 Sélection d'un bloc échantillons

Tous les échantillons non identifiés ne peuvent être resoumis (figure 4.41). Si le spectre de masse ne possède pas assez de pics significatifs permettant une identification, il est exclus de la liste d'échantillons disponibles pour la relance, mais peuvent toutefois être sélectionnés manuellement pour celle-ci via un champ de saisie. Le commentaire "Pas de spectre" accompagne ce type d'échantillon. Les échantillons disponibles sont regroupés par lot de 50 pour la relance. La resoumission d'échantillons peut être aussi faite par date de la première soumision des fichiers.

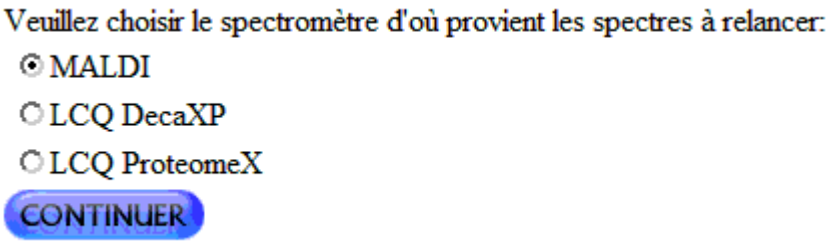


Figure 4.40 formulaire de sélection des échantillons.

4.5.2 Sélection des fichiers à resoumettre

Suite à la sélection des échantillons, le tableau suivant (figure 4.42) affiche tous les fichiers rattachés aux échantillons à resoumettre. Plusieurs fichiers de spectres de masse peuvent être rattachés à un échantillon. Cette réalité peut faire suite à une recalibration de l'appareil ou tout autre changement de paramètres lors de la génération de ces spectres. Les cases à cocher sont disponibles pour sélectionner les fichiers à être soumis de nouveau au

moteur de recherche. Une fois le choix complété, la relance est soumise en cliquant sur le bouton "Continuer".

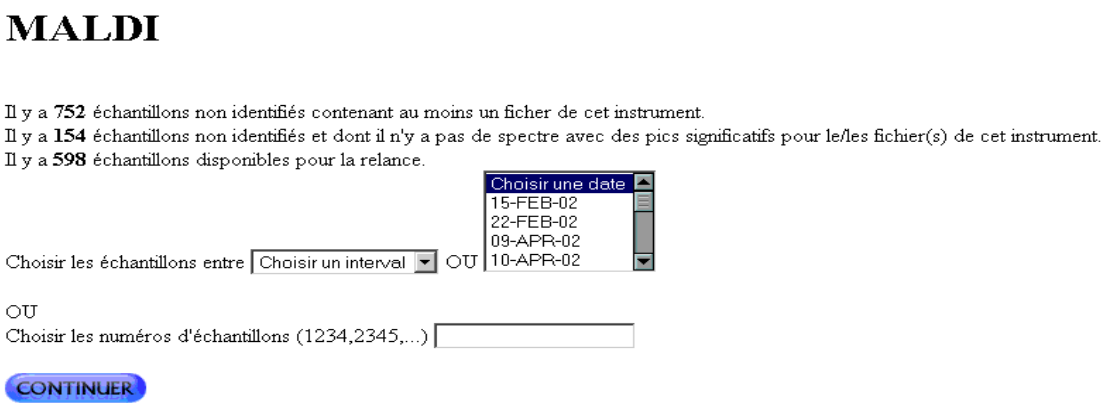


Figure 4.41 Formulaire de sélection des fichiers à resoumettre.

Sélectionner au besoin les fichiers à relancer

Numéro d'échantillon	Numéro de résultat	URL	Fichier	Cases de sélection
1552	3948	http://mimimimi@tahaa:7779/mascot/cgi/master_results.pl?file=../data/20030529/F010793.dat	D:\Voyager\CHUL_Data\INGEL\GQ_temp\G0372-2722_45_0001.txt	☑
1553	3947	http://mimimimi@tahaa:7779/mascot/cgi/master_results.pl?file=../data/20030529/F010792.dat	D:\Voyager\CHUL_Data\INGEL\GQ_temp\G0372-1420_46_0001.txt	☑
1554	3946	http://mimimimi@tahaa:7779/mascot/cgi/master_results.pl?file=../data/20030529/F010791.dat	D:\Voyager\CHUL_Data\INGEL\GQ_temp\G0372-5712_47_0001.txt	☑

Figure 4.42 Raffinement de la sélection des fichiers.

4.5.3 Validation des resoumissions

La validation des identifications des spectres relancés est simple (figure 4.43) et accessible par l'item du lien du menu « Validation-> Des resoumissions ». Le lien ouvre une page affichant la liste des instruments de spectromètres de masse. Le choix d'un des instruments affiche une page contenant un tableau de tous les échantillons relancés et non validés pour cet appareil. Ce tableau affiche l'avant dernier et le dernier résultat Mowse obtenu. Ce résultat est calculé par la formule :

Résultat de Mowse = $-10*Log_{10}(P)$

P : probabilité de Mowse

La valeur de P est la probabilité que l'identification protéique soit obtenue par hasard. La valeur de P est transformée en appliquant le Log sur la valeur de P et en multipliant ce résultat par un facteur de –10. Le produit de cette équation est appelé résultat de Mowse et c'est la valeur de ce résultat qui est retrouvée dans les rapports d'identifications protéiques. Une valeur de probabilité de 10^{-20} équivaut à une valeur de Mowse de 200. Cette transformation a été suggérée par David Perkins (Perkins *et. al.* 1999). effectuée car l'interprétation de la valeur de la probabilité P n'est pas nécessairement évidente. En partie parce que l'ordre de grandeur de cette valeur peut se situer entre 10^{-5} et 10^{-30}, et que plus la valeur de cette probabilité est faible, moins il y a de chance d'obtenir le résultat d'une identification protéique par hasard.

Le lien sur chaque échantillon de la figure 4.43 affiche les informations telles qu'illustrées à la figure 4.30. La validation se fait de la même façon que celle mentionnée dans la section 4.3.3.

Validation des resoumissions

Échantillon	Avant-dernier résultat Mowse	Dernier résultat Mowse	SUPPRIMER
333	88	88	☐
329	161	161	☐
308	88	89	☐
309	68	68	☐
325	33	27	☐
317	47	47	☐
429	36	36	☐
458	68	52	☐
512	58	50	☐
530	63	63	☐
541	62	67	☐
542	40	40	☐
543	61	66	☐

Figure 4.43 Tableau des échantillons resoumis

4.6 Consultation des identifications

Les sections 4.2 à 4.5 expliquaient des parties réservées aux administrateurs de la partie protéomique de la base Murin. Les usagers sans statut d'administrateur n'ont accès qu'à la partie de consultation des données et ne peuvent accéder à la validation de l'identification des protéines.

4.6.1 Interface de recherche

La consultation des données de la base est accessible par le lien « Consultation->Des échantillons validés » du menu principal. Cette page affiche le choix du projet parmi la liste du menu déroulant à l'utilisateur. L'usager de la base ne voit que les projets qui lui sont attribués. Un formulaire (figure 4.44) s'affiche automatiquement après la sélection du projet. Ce formulaire présente différents critères de recherche pour visualiser les données relatives aux identifications protéiques.

Un formulaire de consultation a plusieurs paramètres de recherche. L'usager peut, par exemple, choisir un ou plusieurs tissus en les sélectionnant simultanément dans la liste de tissus de la plate-forme. Dans certains cas, il est intéressant de sortir les protéines identifiées plus d'une fois. Le menu déroulant « Identification » permet de générer la liste des identifications protéiques selon le nombre de fois qu'elles ont été validées.

L'affichage des résultats est basé sur le même principe que le tableau d'analyse des résultats (figure 4.28 et 4.29). Comme le montre la figure 4.45, les cases à cocher affichent les colonnes d'informations dans le tableau des résultats (figure 4.46). Si l'usager désire faire afficher le nom d'une protéine, son numéro d'accession ou encore son numéro d'échantillon, il le fait par une saisie dans un des champs de la figure 4.45 et clique ensuite sur le bouton « Afficher ».

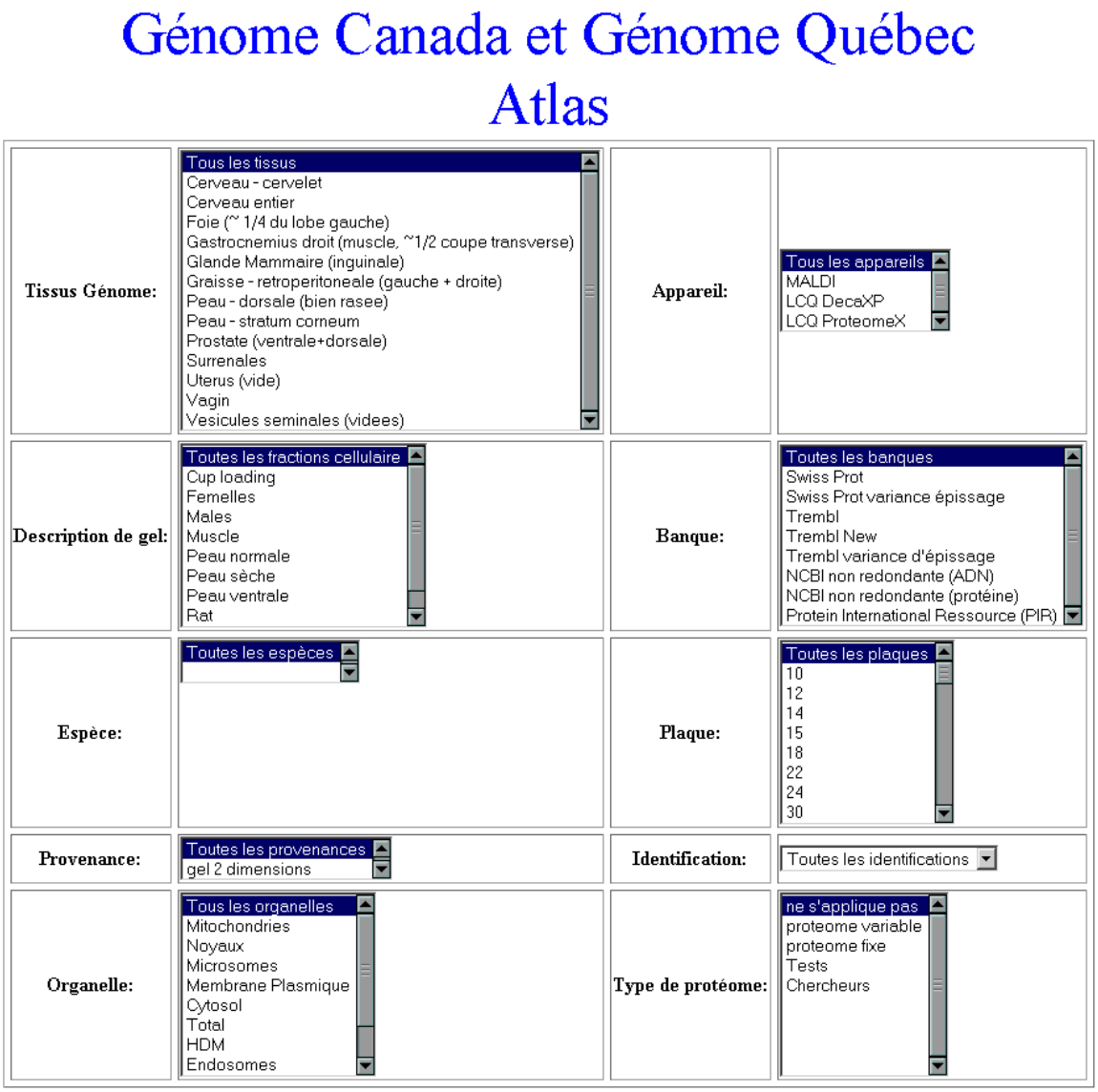


Figure 4.44 Formulaire de recherche pour le projet Génome Canada.

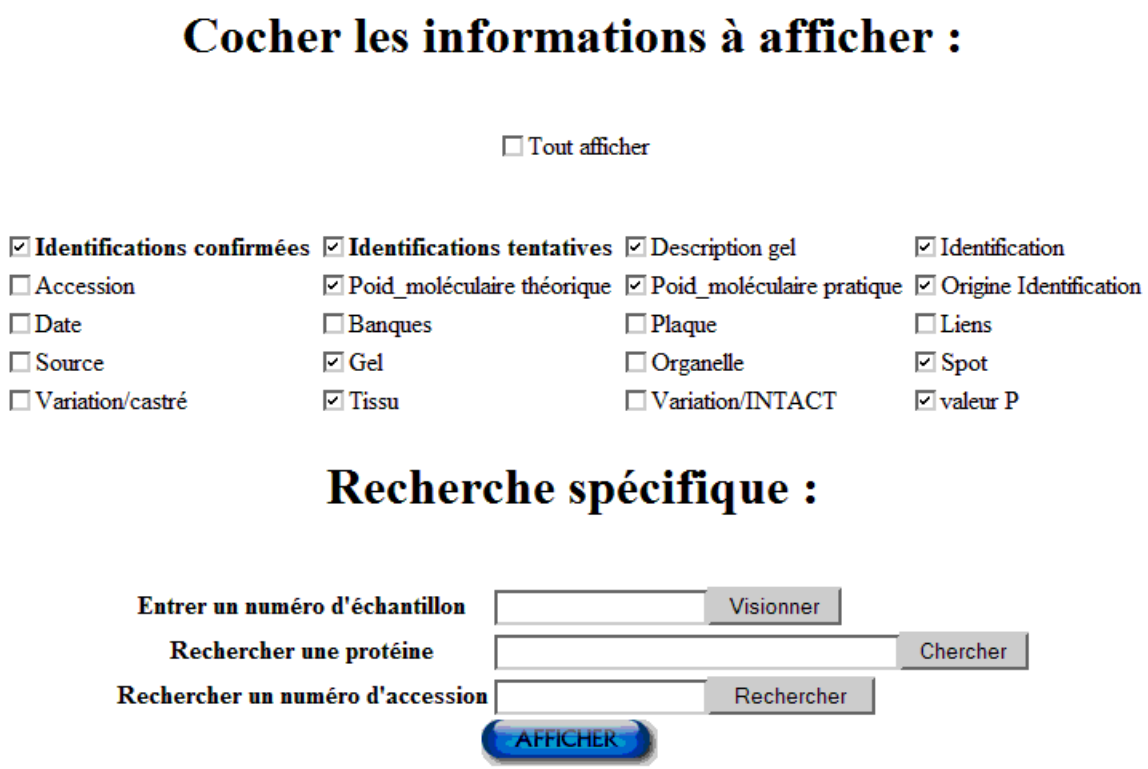
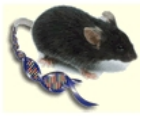


Figure 4.45 Formulaire de sélection des champs à visualiser.

4.6.2 Résultats de la recherche

Le bouton « Tableau pour impression », situé en haut de la page, redimensionne la page dans un format compatible à l'impression. C'est le tableau illustré à la figure 4.46. Une option de tri existe pour la plupart des colonnes du tableau (non illustré). Par exemple, le tri des valeurs de seuil de signification affiche en ordre croissant les identifications. Chaque nom de protéine contient un lien menant vers une page affichant d'autres informations plus générales. Ces informations résident sur les sites publics tels NCBI, Swiss-prot, PIR et UNIGEN. La couleur du nom de la protéine est fonction du type de validation. Les identifications finales sont de couleur verte, et les présomptives, de couleur rouge. La modification du contenu du tableau s'effectue en paramétrant le tableau à l'aide des formulaires précédents illustrés aux figures 4.44 et 4.45



Projet ATLAS

Tableau des identifications protéiques de la plate-forme proteomique du CHUL

65 Champs trouvés

# exp.	Gel	spot	Traitement	Description	poids mol. pratique	poids mol. théo.	Identification	Appareil	Valeur P
1456	G0666 G0668 G0670	1026	24 hrs	Prostate	22	19910.82	modifier 2 [Mus musculus]	MALDI	.026
1424	G0665 G0667 G0669	1029	24 hrs	Prostate	23	59719.96	unnamed protein product [Homo sapiens]	MALDI	-
1424	G0665 G0667 G0669	1029	24 hrs	Prostate	23	66149.05	keratin 1 [Homo sapiens]	MALDI	-
1424	G0665 G0667 G0669	1029	24 hrs	Prostate	23	57384.13	(KRT10)Keratin 10.[Homo sapiens]	LCQ DecaXP	-
1450	G0665 G0667 G0669	1302	24 hrs	Prostate	55	47222.45	(TXNDC4..)1110001E24Rik protein (RIKEN cDNA 1110001E24 gene).[Mus musculus]	MALDI	-
1483	G0666 G0668 G0670	1302	24 hrs	Prostate	55	47222.45	(TXNDC4..)1110001E24Rik protein (RIKEN cDNA 1110001E24 gene).[Mus musculus]	MALDI	.07
1443	G0665 G0667 G0669	1715	24 hrs	Prostate	94	86356.07	polymeric immunoglobulin receptor [Mus musculus]	MALDI	-

Figure 4.46 Affichage du tableau de consultation.

4.6.3 Informations sur la protéine et son profil d'expression

La figure 4.47 illustre les informations relatives à la protéine. L'histogramme est le même que celui décrit à la section 4.4.3. Différents hyper-liens pointent vers d'autres sites web. Par exemple, le numéro gi de NCBI, 28317, de la figure 4.47, affiche les informations du site NCBI sur la protéine. D'autres items pointent vers les banques Swiss prot, PIR et ClusTr, lorsque celles-ci sont existantes pour cette protéine.

Organe:	Prostate (ventrale+dorsale)
Fraction:	phase phenol-chloroforme 8M Uree Prostate
nom de la protéine:	unnamed protein product [Homo sapiens]
Nom du gène:	**non disponible**
Poids moléculaire:	59719.96
Point Isoélectrique:	5.5
Numéro d'accession:	
NCBI #:	28317
second # identification:	X14487
Famille:	CluSTr

Numéros d'identification	Hybridation	RT-PCR	SAGE	Biopuces
X14487				

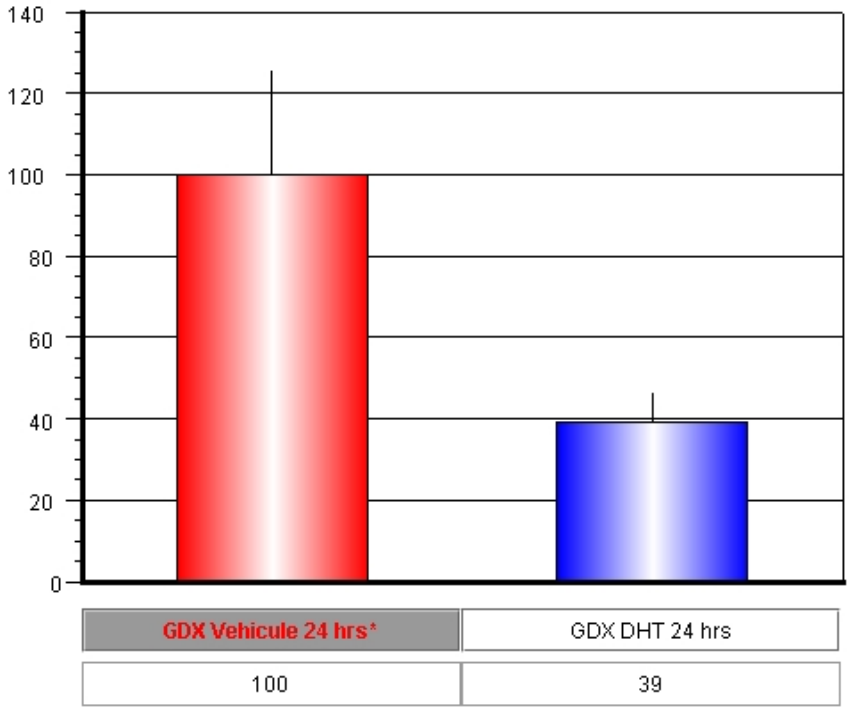


Figure 4.47 Page d'information sur un résultat.

4.7 Les problèmes de redondance des banques de protéines

Plusieurs banques publiques de protéines disponibles sur Internet sont généralement mises à jour sur une base hebdomadaire. Les tables publiques de la base de données Murin doivent donc elles aussi être mises à jour. Pour effectuer cette mise à jour, des scripts en Perl et en PL/SQL téléchargent les fichiers de banques protéiques en format FASTA, exécutent les insertions des nouveaux enregistrements et font la maintenance des indexes protéiques des tables. La principale difficulté dans la maintenance provient de la possibilité de redondance dans les tables publiques. Il peut s'avérer que deux informations identifiant une même protéine portent un identifiant différent, pour une même banque suite à sa mise à jour. L'association entre l'ancien identifiant et le nouveau doit être faite en téléchargeant un autre fichier contenant l'historique des identifiants, car seul l'identifiant le plus récent est présent dans la banque de format FASTA. Il devient alors nécessaire de conserver en archive les anciens identifiants de chaque protéine afin d'être en mesure de faire la correspondance entre les anciens identifiants et le nouveau de chaque protéine.

La mise à jour de la banque protéique Swiss-Prot peut être une cause de la redondance dans la base de données protéomique. Cette banque est divisée en trois sections soit : TrEMBL new, TrEMBL et puis Swiss-Prot. Les protéines nouvellement identifiées sont entreposées dans la section TrEMBLl new. À ce stade, chaque protéine est transcrite à partir des séquences nucléotidiques du site EMBL. Ces informations protéiques possèdent déjà un numéro « Genebank » (gb) et un numéro d'accession Swiss-Prot. Après un certain temps, ces annotations sont transférées dans TrEMBL et restent dans cette banque en attente d’une classification finale. Les informations protéiques sont annotées manuellement dans cette banque par l'ajout de références bibliographiques et la comparaison de la séquence avec d'autres séquences protéiques de la banque. Le numéro attribué à une information protéique sera modifié s'il y a présence de redondance dans la banque. Un nouveau numéro d'accession va être attribué à cette information protéique. Le changement de ce numéro d'accession est parfois nécessaire pour assurer la cohérence de la banque de protéines. Par exemple, lorsque deux entrées protéiques sont identiques, un numéro d'accession est attribué pour la nouvelle information protéique et les numéros d'accession initiaux sont conservés dans l'historique de cette information protéique. Les anciens numéros d'accession se retrouvent dans un fichier à

part de format non FASTA et contenant les informations protéiques de la banque Swiss-Prot. Finalement, toutes les protéines annotées manuellement sont transférées dans la banque Swiss-Prot. À ce stade, le numéro d'accession Swiss-Prot est soit officialisé ou modifié. La base doit tenir compte des modifications possibles de numéros d'accession de la banque Swiss-prot.

La présence de plusieurs identifiants pour une même annotation se retrouve aussi dans la banque NCBI. Cette banque fournit à ces protéines des numéros d'identification appelés GI. Dans cette banque, une entrée peut posséder plusieurs numéros GI en plus du numéro « Genebank ». Les anciens numéros GI sont mis en archive dans des fichiers de format non FASTA qui sont téléchargés pour maintenir la cohérence dans la base locale.

Avec plus d'un million d'entrées dans les tables dites publiques, la présence de protéines redondantes est une réalité qui n'est pas facile à gérer. Le défi consiste à maintenir la cohérence entre les anciennes entrées protéiques de la table Protéine avec les nouvelles lors de la mise à jour des tables publiques. Deux approches sont employées pour cette maintenance. La première façon est d’insérer un nouvel enregistrement dans la table Protéine et de référencer, s'il y a présence de redondance, l'ancien enregistrement vers le nouveau. La référence de l'ancien enregistrement vers la nouveau s'effectue en insérant la clé unique de l’attribut **proteine_id** du nouvel enregistrement dans l'attribut **lien_proteine_id** de l'ancien enregistrement. Cette méthode est similaire au concept d'une liste chaînée. Toutefois, cette approche risque de faire augmenter rapidement le nombre d'enregistrements dans la table.

Proteine							
Proteine_id	Gi	Description	Pir_id	Accession_sp	banque	Date	Lien_proteine_id
37667575	5454443	ATP synthase	-	Q61646	Swiss prot	17 jan 2004	
2665354	345435	ATP synthase	-	P46788	Swiss prot	20 avr 2003	37667575

Figure 4.48 La première approche de la gestion de redondance des enregistrements protéiques. L'ancienne insertion de l'information protéique est référencée vers l'information actualisée par le champ lien_proteine_id.

Une deuxième approche consiste à conserver toutes les insertions des informations protéiques dans une table nommée Protéine_archive et de conserver les informations les plus récentes dans la table Protéine. La clé primaire de la table Protéine fait la référence dans le ou les tuples de la table Protéine_archive correspondant à la même information protéique. Or, si une information protéique de la table Protéine est maintenant identifiée par un nouvel identifiant, celui-ci vient remplacer l'identifiant actuel de la table Protéine et une insertion du nouvel enregistrement se fait dans la table Protéine_archive.

La deuxième approche prévaut sur la première afin de minimiser la quantité de tuples de la table Protéine et l'insertion d'information redondante. La première approche doit être utilisée dans la situation où les enregistrements sont déjà présents dans la table Proteine. Ces enregistrements ne peuvent être effacés, car ils risqueraient d'invalider des références faites sur eux entre les tables du schéma protéomique et d'autres tables externes. La date d'insertion ou de mise à jour fait partie d'un enregistrement dans la table Proteine.

Il arrive que deux enregistrements référant à une même protéine se retrouvent dans la base. L'une d'elles possède un numéro GI, et l'autre, un numéro d'accession Swiss-Prot. Cette redondance ne peut être éliminée, car ces deux enregistrements proviennent de banques protéiques différentes et qu'il n'existe pas d'information permettant de faire le lien entre elles. Sachant qu'une protéine de la table Proteine peut posséder un numéro GI, un numéro d'accession Swiss-Prot, un identifiant PIR et un identifiant « Genebank », les scripts de mise

à jour doivent vérifier à partir de ces numéros les possibilités de redondance afin de mettre à jour les tables faisant les relations entre les différents identifiants d'un même enregistrement protéique.

Protéine_archive			
Prot_accession_sp	Prot_pir_id	Prot_gi	Proteine_id
P48788		345435	37667575
		5454443	37667575
Q64374			37667575

Proteine							
Proteine_id	Gi	Description	Pir_id	Accession_sp	banque	Date	Lien_proteine_id
37667575	5454443	ATP synthase	-	Q61646	Swiss prot	17 jan 2004	-

Figure 4.49 La deuxième approche de la gestion de redondance des enregistrements protéiques. Cet exemple utilise les tables Proteine_archive et Proteine. Les informations dans ces champs sont fictives. La table Protéine_archive contient tous les numéros d'accession se référant à une insertion de la table Proteine. Cette insertion contient les informations actualisées sur la protéine.

4.8 Les améliorations à apporter à la plate-forme protéomique

Le système de la plate-forme protéomique est un système convivial, il est aussi très dynamique et en continuelle évolution. Toutefois, l'architecture de base de la cueillette des échantillons et de leur validation, avec plus de trois milles échantillons traités, est bien implantée et ne change pas. Les modifications à apporter au système se situent plutôt à la périphérie de la structure. Par exemple, les modifications au niveau de la structure de la base font référence à l'ajout de nouvelles tables pour contenir les informations d'un nouvel appareil de mesure, l'ajout et la suppression d'attributs dans une table, ou les changements de relation entre les tables. Plusieurs facteurs contribuent à ces changements. Entre autre, les besoins des utilisateurs du système changent graduellement et d'autres s'ajoutent. Les modifications au niveau de la forme concernent la présentation des résultats à l'écran des utilisateurs. Ces modifications sont l'ajout d'une information manquante ou la création de structures pour la présentation de données telles que des tableaux, des graphiques, etc.

L'élaboration du système protéomique s'est effectué en exploitant les outils informatiques connus et les connaissances acquises tout au long du projet. L'expérience de travail au début et à la fin du projet s'est accrue et si la conception des tables avait été à refaire, celle-ci aurait été quelque peu différente pour mieux accommoder l'ajout de nouveaux blocs d'informations. Les tables originales ont été construites en fonction de l'idée de départ, et que celle-ci se modifie à mesure que le projet prend de l'ampleur. Donc, toute modification ultérieure de tables existantes doit être prise avec soin, car les données déjà contenues dans ces tables risquent d'être affectées. Quant à l'ajout de nouvelles tables, leur conception doit toujours tenir compte de la possibilité de modifier ces tables sans affecter le reste de l'architecture de la base déjà établie.

La conception du modèle relationnel de la base constitue le coeur de l'architecture du système. Une faiblesse dans cette conception est le manque de clarté dans l'identification de certaines tables et attributs. Par exemple, la table Provenance contient les informations sur la provenance de la création d'un échantillon. Celui-ci peut provenir des gels à une ou deux dimensions, du Protéomix, des chercheurs ou d'autres origines non répertoriées. Le terme « provenance » prend tout son sens, mais n'est tout simplement pas assez descriptif pour

déduire rapidement le contenu de cette table. Au niveau des variables, dans la table info_masse, on peut y lire la variable miss_cliveage. Cette variable entrepose le nombre de digestions tryptiques manquantes dans une séquence peptidique correspondant à une masse du spectre MS ou MS/MS. En lisant cette description, nous comprenons mieux l'utilité de cette variable. Il est donc nécessaire de documenter les attributs qui ne sont pas suffisamment significatif, et d'utiliser des noms appropriés pour les futurs attributs.

L'élaboration de fonctionnalités au système protéomique implique la production de pages de codes informatiques accompagnées d'une documentation expliquant la logique de conception de chaque code et comment faire une utilisation efficace des fonctions qu'elles codent. Toutefois, la production rapide de pages de codes sources afin d'accélérer l'utilisation des nouvelles fonctionnalités, telles celles générées en Perl et en PL/SQL, vient compromettre le temps consacré à la documentation complète et claire de ces pages de codes. Cela nécessite de consulter ultérieurement ces pages pour compléter leur documentation, ce qui requiert parfois un certain temps pour saisir à nouveau la logique du code. Au moment de la rédaction, 40% du code est correctement documenté. Cette documentation doit être complétée afin de rendre utilisable et compréhensible l'ensemble du code à d'autres programmeurs.

Conclusion

Résumé

La conception d'un système automatisé en protéomique ne consiste pas simplement à monter l'architecture, mais à développer une vue d'ensemble des différents domaines qui gravitent autour du système, comme la biologie et l'informatique, d'où l'expansion de la bioinformatique. Le système de cueillette et de gestion des données de la plate-forme de protéomique du CHUL a permis d'organiser les données en fonction de leurs inter-relations et de leurs rôles. De plus, le système d'identification et d'analyse a été conçu principalement en fonction des besoins de la plate-forme. En plus d'avoir permis de réduire considérablement le temps relié à la validation des données, le système a pu gérer leurs cueillettes automatiques, leurs traitements et leurs sauvegardes. Son ergonomie a permis aux utilisateurs de l'exploiter quotidiennement et efficacement.

Perspectives futures

Le perfectionnement du système de protéomique s'accroît chaque jour afin de répondre aux nouveaux besoins des chercheurs et de rester productif dans le domaine de la recherche. Par conséquent, la versatilité du système est un facteur important lors de la conception. Mascot est le logiciel de recherche qui a été intégré pour identifier les échantillons. Toutefois, l'utilisation d'un seul algorithme n'est pas conseillée, puisque ce dernier peut avoir des limitations devant certaines exceptions de la biologie. Par exemple, l'apomucine sous-maxilaire porcine contient 80% de séquences répétées et son traitement par Mascot pourrait donner une probabilité faussement élevée. Ce genre de cas doit être étudié avec soin. Un autre logiciel de recherche, celui de Sequest, est aussi généralement très performant et donne pratiquement les mêmes résultats que le logiciel précédent. Mais puisqu'il a été construit avec une architecture différente, il pourrait apporter un résultat plus intéressant dans la situation où Mascot a donné un résultat erroné et vice versa. Il est prévu que le logiciel Sequest soit ajouté dans le présent système afin de permettre une double identification. La versatilité est donc un motif important dans la conception du système protéomique. De plus, les exigences actuelles de la publication des résultats requièrent des identifications protéiques venant des logiciels d’identification ainsi que les spectres de masse ayant servis aux analyses.

Le système actuel va éventuellement bénéficier d'un outil statistique supplémentaire pour la validation des résultats. Il s'agit d'un module indépendant, PeptideProphet et ProteinProphet, qui travaille de paire avec les algorithmes Sequest et Mascot (Keller et al. 2002). Cet outil permettra d'attribuer une probabilité statistique sur chaque peptide servant à identifier une protéine. Cette probabilité permet de déceler les identifications peptidiques erronées et ainsi raffiner les identifications, tout en réduisant davantage le temps d'analyse.

L’annotation grandissante de séquences protéiques permet d’établir une classification de protéines. En utilisant l’algorithme d’alignement de Smith et waterman (Smith *et al.* 1981), différentes familles de protéines sont établies en fonction de la similarité des domaines de leur séquence peptidique. Des banques protéiques telles CluSTr regroupent ces familles de protéines et peuvent être utilisées pour des études phylogénétiques. La comparaison des

protéines identifiées par la plate-forme protéomique à ces banques permettrait d'établir de nouvelles relations entre les fonctions des gènes qui sont modulés de façon concertée (Kriventseva *et al.* 2001, Maleszka *et al.* 2001).

Une des tâches importantes dans la réalisation du projet ATLAS est l'intégration des données provenant des différentes plates-formes de recherche du CHUL soit : les biopuces, l'hybridation *in situ*, le SAGE (Serial Analysis and Gene Expression), les sentiers de signalisations et la protéomique. Un profil d'expression d'une protéine, obtenu en protéomique par le traitement de la DHT (dihydroxy-testosterone), peut, par exemple, être corrélé avec le profil d'expression de son ARN messager (ARNm). Cette corrélation peut être faite avec différents tissus.

L'intégration des données de protéomique avec les données de génomique demande une attention particulière, surtout lorsqu'il y a une discordance entre les profils d'expression d'une protéine, observés en protéomique, et son ARNm, observé dans une autre plate-forme. Dans un tissu, il arrive que le taux d'expression d'une protéine ne soit pas modifié suite à un traitement à la DHT, alors que l'on observe un changement du taux d'expression à la hausse de son ARNm. Cette discordance pourrait, par exemple, être due à une erreur de manipulation lors de la réalisation de l'expérience. Par ailleurs, la technique des gels 2D utilisée pour détecter une différence d'expression différentielle en protéomique est plutôt qualitative, car les mesures sont obtenues par une différence de densité optique des taches entre plusieurs gels et cette différence peut varier d'une expérience à l'autre. Une technique plus sensible pour mesurer l'expression différentielle en spectrométrie de masse, le ICAT (Isotope-Coded Affinity Tag), pourrait être une alternative intéressante. Cette approche consiste à digérer deux échantillons de protéines ou deux complexes protéiques purifiés. Après la digestion, les peptides contenant des résidus de cystéine sont marqués par liaison covalente d'un isotope de leucine. Chaque cystéine d'un des échantillons est marquée par un isotope de leucine à 8 atomes d'hydrogène, alors que l'autre est marqué par des leucines contenant 8 atomes de deutérium. Ces échantillons sont ensuite mélangés et analysés par spectrométrie de masse. La différence de masse entre deux peptides de même séquence portant des isotopes différents permet de quantifier l'abondance des protéines présentes dans chacun des échantillons (Radish *et. Al.* 2003, Zhou *et. Al.* 2002).

Les protocoles expérimentaux du projet de Génome Canada sont conçus pour ne faire varier qu'un facteur, soit celui d'un stéroïde tel le DHT. Il y a cependant d'autres facteurs qui peuvent varier sans que ces protocoles ne puissent les maîtriser. Un facteur environnemental, comme celui du stress de l'animal juste avant le sacrifice, pourrait provoquer une répression de la traduction de la protéine en activant un gène codant pour une protéine inhibitrice. À l'opposé, le profil d'expression d'une protéine pourrait être modulé à la hausse, alors que le profil d'expression de son ARNm est modulé à la baisse. Plusieurs hypothèses tentent d'expliquer ce phénomène autre qu'une erreur de manipulation ou de lecture de l'instrument. Certains ARNm sont instables dans leur milieu et sont dégradés rapidement. La demi-vie[5] de ces ARNm est par conséquent relativement courte. L'absence de DHT induirait l'expression de ces ARNm, mais sa présence ne les réprimerait pas. En se dégradant, on observerait une diminution du profil d'expression de l'ARNm et on observerait une légère augmentation de l'expression de la protéine. Une autre hypothèse, l'ARNm aurait une plus grande stabilité et serait dégradé par la présence de la DHT, mais cette dernière n'aurait pas d'effet sur la dégradation de la protéine. Il pourrait aussi s'avérer que le gène codant pour l'ARNm soit régulé par l'activation d'un autre gène, lui-même régulé par la DHT. Pour une seule protéine, le profil de son expression peut être différent selon le tissu, tout comme le profil d'expression de son ARNm, et caractérisé par une régulation soit au niveau de la transcription, soit au niveau de la traduction. La stabilité d'une protéine peut aussi être modifiée sous différentes conditions physiologiques (Anderson *et al.* 2001).

L'intégration des données de la génomique avec celles de la protéomique n'est pas toujours simple. Les différents facteurs, mentionnés précédemment, peuvent influencer l'interprétation des résultats d'une expression différentielle des gènes d'intérêt. Par conséquent, la compréhension des mécanismes de régulation génique et protéique des gènes est nécessaire. Toutefois, l'annotation des gènes, des transcrits, et des protéines dans les banques de données publiques facilite l'établissement de relation entre une protéine et son transcrit ou une protéine et son gène. Les références de ces annotations peuvent entraîner l'élaboration des sentiers métaboliques ou de signalisation hypothétique. Ces sentiers contribueront à la compréhension des profils d'expressions des protéines et leurs

[5] La demi-vie est le temps nécessaire pour que la concentration d'une substance dans un milieu déterminé soit réduite de moitié dans ce milieu.

interactions. L'intégration de données entre diverses plates-formes aidera à mettre en évidence ces mécanismes.

Puisque les besoins des plates-formes de recherche en protéomique se ressemblent, l'accessibilité du système développé au CHUL à d'autres laboratoires est envisageable. Certaines contraintes sont cependant requises pour maximiser la portabilité et la compatibilité à travers les différents laboratoires et ce, tant au niveau du matériel informatique que des instruments de mesure.

Conclusion générale

La protéomique évolue rapidement et la génomique génère d'importantes quantités d'informations sur la séquence des génomes. Ces séquences, pour la plupart non identifiées, devront être caractérisées et annotées afin d'associer le gène à la protéine. Pour répondre à l'énorme quantité de données, il devient nécessaire d'établir un système qui puisse permettre d'identifier des protéines provenant de différents échantillons, et aussi de caractériser de nouvelles protéines dans un système à haut débit. L'identification de ces protéines sera une première étape avant d'entrer dans la définition de leurs rôles, leurs interactions protéine-protéine, leurs niveaux d'expression selon différentes conditions, leur localisation cellulaire et tissulaire, les prédictions de leurs conformations tridimentionnelle, la conception et la fabrication des activateurs, inhibiteurs pour traitements médicaux, etc. Ce projet nécessite donc l'accès aux dernières connaissances en protéomique, en génomique et en bioinformatique dans le but d'offrir une recherche de qualité et des plus perfectionnées.

Bibliographie

Ouvrages cités

Aebersold, R., Mann, M., Mass spectrometry-based proteomics., 2003, *nature*, vol 422,198-207.

Apweiler R., Biswas M., Fleishmann W., Kanapin A., Karavidopoulou Y., Kersey, P., Kriventseva E. V., Mittard V., Mulder N., Phan I., Zdobnov E., Proteome analysis database: online application of InterPro and CluSTr for the functional classification of proteins in whole genomes. 2001, *Nuc. ac. Res*., vol 29 no 1, 44-48.

Bairoch, A., Apweiler, R., The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, *Nucleic Acids Research*, 2000 , vol.28, No. 1, 45-48.

Baevis, R., Fenyo, D., Informatics and data management in proteomics, *Trends in biotechnology*, 2002, vol. 20 no 12, s53-s38.

Baxevanis, A. D. The molecular biology database collection: an updated compilation of biological database ressources, *Nucleic Acids Research*, 2001 ,vol.29, No. 1, 1-10.

Benson, D. A., Boguski, M. S., Ouellette, F. Genbank, *Nucleic Acids Research*, 1998, vol.26, No. 1, 1-7.

Blackstock W., Mann M. Proteomics: A trends Guide, *Review*, july 2000 ,1-51.

Buckingham, S., Programmed for success, *Nature*, sept 2003, vol. 425,209-216.

Chait, B; Zhang, W. Profound : an expert system for protein identification using mass spectrometric peptide mapping information, *Anal. Chem.*, 2000, 72: 2482-2489.

Chait, B., Fenyo, D., Proteomics: the new trend in tools, *Genome technology*, 2001, 36-44

Choudhary, J., Blackstock, W., Creasy, D., Cottrell, J. Interrogating the human genome using uninterpreted mass spectrometry data, *Proteomics*, 2001, 1: 651-667.

Clauser, K. R., Baker, P., Burlingame, A. L., Role of accurate mass measurement (± 10 ppm) in protein identification strategies emplying MS or MS/MS and database searching, Anal. Chem., 1999, 71:2871-2882.

Claverie, J.-M., Audic, S., Abergel, C., La bioinformatique: une discipline stratégique pour l'analyse et la valorisation des génomes, 2000, CNRS-AVENTIS UMR 1889 , Marseille, http://igs-server.cnrs-mrs.fr/jcnrs.html.

Craig, R., Beavis, R. C., A method for reducing the time required to match protein sequences with tandem mass spectra, 2003, rapid communications in mass spectrometry, 17:2310-2316.

Cook, N., The proteomic revolution, 2002, *Genomics and Proteomics*, 11-14

Eng, J. K., McCormack, A. L., Yates, J. R., An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database, 1994, J. Am. Soc. Mass spectrum., 5:976-989.

Gavin, A.-C., Bösche, M., Krause, R., Grandi, *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes, 2002, *Nature*, vol. 415, 141-147.

Goodman, N., Text mining: help, *Genome technology*, 2003, 44-47.

Jung, E., Veuthey, A.-L., Gasteiger, E., Bairoch, A. Annotation of glycoproteins in the SWISS-PROT database, *Proteomics*, 2001, 1: 262-268.

Kearney, P., Thibault, P., Bioinformatics meets proteomics bridging the gap between mass spectrometry data analysis and cell biology, 2003, *journ. Of bioinformatics and computational biology*, vol 1, no 1, 183-200.

Keller, A., Nesvizhskii, A.I., Kolker, E., Aebersold, R., Empirical statistical model to estimate the accuracy of peptide identification made by MS/MS and database search, 2002, *Anal. Chem.*, vol 74, 5383-5392.

Kriventseva, E.V., Fleischmann, W., Zdobnov, E.V., Apweiler, R., CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins, *Nucleic acids research*, 2001, vol29, no 1, 33-36.

Kuster, B., Mortensen, P., Andersen , J., Mann, M. Mass spectrometry allows direct identification of proteins in large genomes, *Proteomics*, 2001, 1: 641-650.

Lester, P., Hubbard, S.J., Comparative bioinformatic analysis of complete proteomes and protein parameters for cross-species identification in proteomics, 2002, *Proteomics*, vol. 2, 1392-1405.

Maleska, R., Gabor. Miklos, G. Protein functions and biological contexts, *Proteomics*, 2001, 1: 169-178.

Pappin, D.J.C., Hojrup, P., Bleasby, A.J. Rapid identification of proteins by peptide-mass fingerprinting, *Cur. Biol., 1993*, vol 3 no 6 327- 332.

Pappin, D. N., Perkins, D J., Creasy, D. M., Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data, *Electrophoresis*, 1999, 20: 3551-3567.

Patterson, S.D., Data analysis-the Achilles heel of proteomics, *Nature Biotechnology*, 2003, vol. 21, 221-222.

Radish, J., Yi, E. C., Leslie, D. M., Purvine, S. O., Goodlett, D. R., Eng, J., Aebersold, R., The study of macromolecular complexes by quantitative proteomics, *Nature Genetics*, 2003, vol. 33, 349-355.

Sanseau, P. Impact of human genome sequencing for *in silico* target discovery, *DDT*, vol 6, no. 6: 316-323.

Smith, T.F., Waterman, M.S., Identification of common molecular subsequences, *J. Mol. Biol.*, 1981, vol 147, 195-197.

Stajich, J.E., Block, D, Boulez, K., Brenner, *et al.*, The Bioperl Toolkit: Perl modules for the Life Sciences, 2002, *Genome Research*, vol 12, 1611-1618.

Stein, L., Creating a bioinformatics nation, 2002, *Nature*, vol 417, 119-120

Taylor, C.F., Paton, N.W., Garwood, K.L., Kirby, P.D., *et al.*, A systematic approach to modeling, capturing, and disseminating proteomics experimental data, 2003, *Nature Biotechnology*, vol 21, 247-254.

The sequence of the Human genome, Science, vol 291, 2001, 1304-1351.

Tyers, M., Mann, M., From genomics to proteomics, 2003, *Nature*, vol422, 193-197.

Wickelgren, I., Spinning junk into gold, *Science*, 2003, vol 300, 1646-1649.

Wilkins, M. R., Williams, K. L., Cross-species protein identification using amino acid composition, peptide mass fingerprinting, isoelectric point and molecular mass: a theoretical evaluation, 1997, J. Theor. Biol. ,186:7-15.

Wu, C. H., L. Yey, L.-S., Huang, H., Arminski, *et al.*, The Proteine Information ressource, 2003, *Nucleic Acids Research*, vol. 31, no. 1, *345-347*

Zhou, H., Ranish, J. A., Watts, J. D., Aebersold, R., Quantitative proteome analysis by solid-phase isotope tagging and mass spectrometry, *Nature Biotechnology*, 2002, vol. 19, 512-515.

Ouvrages de références

Deitel, H. M., Deitel, P. J., Java Comment programmer, 4e édition, Repentigny 2002, (106)

Flanagan, D., JavaScript: the definitive guide, 4e édition, Sebastopol 2002, (1-2)

Gamache, A., Architecture, modèles et langages de données, volume 1, Sainte-Foy 2001,(1,171-174,229-231,273-275)

Gamache, A., Architecture, modèles et langages de données, volume 2, Sainte-Foy 2001,(1,23,63,125-126,157,202-203)

Soutou,C., *Objet-relationnel sous Oracle 8*, Paris, 1999, (119 - 171)

Si Albir, S. *UML in a nutshell*, Sebastopol 1998, (1 - 3)

Références de sites web

« Banques de données protéiques non redondantes ». Site Suisse Embnet , [En ligne]. ftp://ftp.ch.embnet.org/pub/databases/nr_prot/ (page consultée de 2001-2004)

¨ « Le progrès du séquencage du génome humain ». *Site de National Center of Biotechnology Information*, [En ligne].
http://www.ncbi.nlm.nih.gov/genome/seq/HsHome.shtml (page consultée de 2001-2004)

"La banque de données protéiques PDB". *Site de Protein Data Bank* , [En ligne]. http://www.rcsb.org/pdb/ (page consultée en 2002-2003)

« Une liste de liens de la plupart des banques de données protéiques sur Internet » *Site de Protein Data Bank* , [En ligne]. http://www.rcsb.org/pdb/links.html# (page consultée en 2002-2003)

« La banque génomique de la souris » . Site du MGD (Mouse Genomic database),*[En ligne]*. *http://www.informatics.jax.org* (page consultée en 2002-2004)

Binns, Kathleen L., « Présentation vidéo du fonctionnement du MS/MS ». The Pawson Lab.*Site du Laboratoire de Pawson*, [En ligne]. http://www.mshri.on.ca/pawson/MS/MS.html (page consultée en 2002-2004*)*

« Le grand dictionnaire terminologique ». *Site de l'office québécois de la langue française*, [En ligne].http://www.granddictionnaire.com (page consultée en 2003-2004*)*

« Banque de données protéique de Swiss prot ». *Site du centre protéomique ExPASy*, [En ligne]. http://us.expasy.org/. (page consultée en 2001-2004)

« La banque de protéines PIR ». *Site du Protein Information ressource (PIR)*, [En ligne]. http://pir.georgetown.edu/home.shtml (page consultée en 2001-2004*)*

« La banque unifiée de Swiss prot, PIR et TrEMBL ». *Site de l'universal protein resource UniProt*, [En ligne]. http://expasy.uniprot.org/index.shtml (page consultée en 2004*)*

ANNEXE A

Contenu du fichier de paramètres de Format MIME pour l'exécution du moteur de recherche Mascot.

```
-----------------------------127121488018957
Content-Disposition: form-data; name="accession"


-----------------------------127121488018957
Content-Disposition: form-data; name="ltol"


-----------------------------127121488018957
Content-Disposition: form-data; name="reptype"

Peptide
-----------------------------127121488018957
Content-Disposition: form-data; name="frag"


-----------------------------127121488018957
Content-Disposition: form-data; name="it_mods"

Carbamidomethyl (C),Oxidation (M)
-----------------------------127121488018957
Content-Disposition: form-data; name="search"

PMF
-----------------------------127121488018957
Content-Disposition: form-data; name="pfa"

1
-----------------------------127121488018957
Content-Disposition: form-data; name="formver"

1.01
-----------------------------127121488018957
Content-Disposition: form-data; name="iytol"


-----------------------------127121488018957
Content-Disposition: form-data; name="user10"


-----------------------------127121488018957
Content-Disposition: form-data; name="segt"


-----------------------------127121488018957
Content-Disposition: form-data; name="user11"


-----------------------------127121488018957
Content-Disposition: form-data; name="user12"
```

----------------------------127121488018957
Content-Disposition: form-data; name="cle"

Trypsin/P
----------------------------127121488018957
Content-Disposition: form-data; name="nm"

----------------------------127121488018957
Content-Disposition: form-data; name="icat"

----------------------------127121488018957
Content-Disposition: form-data; name="tol"

60
----------------------------127121488018957
Content-Disposition: form-data; name="seg"

----------------------------127121488018957
Content-Disposition: form-data; name="iy2tol"

----------------------------127121488018957
Content-Disposition: form-data; name="taxonomy"

. . . . . Mammalia (mammals)
----------------------------127121488018957
Content-Disposition: form-data; name="segtu"

----------------------------127121488018957
Content-Disposition: form-data; name="subcluster"

----------------------------127121488018957
Content-Disposition: form-data; name="tolu"

ppm
----------------------------127121488018957
Content-Disposition: form-data; name="overview"

----------------------------127121488018957
Content-Disposition: form-data; name="license"

Licensed to: CHUL Research Center, (1 processor).
----------------------------127121488018957
Content-Disposition: form-data; name="username"

----------------------------127121488018957
Content-Disposition: form-data; name="iystol"

----------------------------127121488018957
Content-Disposition: form-data; name="ia2tol"

----------------------------127121488018957
Content-Disposition: form-data; name="ib2tol"

----------------------------127121488018957
Content-Disposition: form-data; name="iastol"

----------------------------127121488018957
Content-Disposition: form-data; name="two"

----------------------------127121488018957
Content-Disposition: form-data; name="ibstol"

----------------------------127121488018957
Content-Disposition: form-data; name="user00"

----------------------------127121488018957
Content-Disposition: form-data; name="iatol"

----------------------------127121488018957
Content-Disposition: form-data; name="format"

Mascot generic
----------------------------127121488018957
Content-Disposition: form-data; name="itolu"

Da
----------------------------127121488018957
Content-Disposition: form-data; name="user01"

----------------------------127121488018957
Content-Disposition: form-data; name="mods"

----------------------------127121488018957
Content-Disposition: form-data; name="peak"

----------------------------127121488018957
Content-Disposition: form-data; name="user02"

----------------------------127121488018957
Content-Disposition: form-data; name="user03"

----------------------------127121488018957
Content-Disposition: form-data; name="ibtol"

----------------------------127121488018957

Content-Disposition: form-data; name="user04"

-----------------------------127121488018957
Content-Disposition: form-data; name="que"

-----------------------------127121488018957
Content-Disposition: form-data; name="user05"

-----------------------------127121488018957
Content-Disposition: form-data; name="precursor"

-----------------------------127121488018957
Content-Disposition: form-data; name="user06"

-----------------------------127121488018957
Content-Disposition: form-data; name="user07"

-----------------------------127121488018957
Content-Disposition: form-data; name="user08"

-----------------------------127121488018957
Content-Disposition: form-data; name="mp"

-----------------------------127121488018957
Content-Disposition: form-data; name="db"

new_banques_proteo
-----------------------------127121488018957
Content-Disposition: form-data; name="useremail"

-----------------------------127121488018957
Content-Disposition: form-data; name="user09"

-----------------------------127121488018957
Content-Disposition: form-data; name="intermediate"

../data/20020409/F002449.dat
-----------------------------127121488018957
Content-Disposition: form-data; name="ith"

-----------------------------127121488018957
Content-Disposition: form-data; name="itol"

0.5
-----------------------------127121488018957
Content-Disposition: form-data; name="charge"

1+

----------127121488018957
Content-Disposition: form-data; name="com"

----------127121488018957
Content-Disposition: form-data; name="mass"

Monoisotopic
----------127121488018957
Content-Disposition: form-data; name="file"

D:\Voyager\CHUL Data\INGEL\GQ\G0170-2405_74_0001.txt
----------127121488018957
Content-Disposition: form-data; name="report"

20
----------127121488018957
Content-Disposition: form-data; name="QUE"

861.125000
867.144900
877.116300
912.474800
928.468500
944.463800
974.560500
998.570200
1041.561000
1110.580900
1158.060800
1218.645000
1256.701300
1260.714100
1272.700500
1288.707400
1353.666800
1514.794600
1605.801000
1646.854700
----------127121488018957

ANNEXE B

Code source d'une liste de paramètres pour la génération d'un spectre de masse en Java.

```
<html>
<head>
<meta HTTP-EQUIV=Refresh CONTENT="1200;URL=invivo.quitter">
<title>graphique</title>
</head>
<body bgcolor="BLUE">
<p align="center"><applet code =" graphMaldi/GraphJApplet.class"
codebase="/proteomique/" archive="GraphJApplet.jar"
pluginspage="http://java.sun.com/products/plugin/1.1.2/plugin-
install.html" width =" 1024" height =" 768">
<param NAME="gq" VALUE="GQ002M03D13M1">
<param NAME="gel" VALUE="G0169">
<param NAME="spot" VALUE="7012">
<param NAME="proteine" VALUE="(PRDX1..)Peroxiredoxin 1 (EC 1.11.1.-)
(Thioredoxin peroxidase 2) (Thioredo">
<param NAME="masse1" VALUE="855.0627">
<param NAME="intensite1" VALUE="10.35">
<param NAME="masse2" VALUE="860.3944">
<param NAME="intensite2" VALUE="11.58">
<param NAME="masse3" VALUE="861.0446">

<param NAME="intensite3" VALUE="10.24">
<param NAME="masse4" VALUE="862.4059">
<param NAME="intensite4" VALUE="6.64">
<param NAME="masse5" VALUE="864.4727">
<param NAME="intensite5" VALUE="17.23">
<param NAME="masse6" VALUE="882.5272">
<param NAME="intensite6" VALUE="8.62">
<param NAME="masse7" VALUE="894.4194">
<param NAME="intensite7" VALUE="59.26">
<param NAME="masse8" VALUE="920.4933">
<param NAME="intensite8" VALUE="9.84">
<param NAME="masse9" VALUE="1000.5094">
<param NAME="intensite9" VALUE="4.24">
<param NAME="masse10" VALUE="1006.5173">
<param NAME="intensite10" VALUE="24.05">
<param NAME="masse11" VALUE="1028.5127">
<param NAME="intensite11" VALUE="12.69">

<param NAME="masse12" VALUE="1034.1049">
<param NAME="intensite12" VALUE="40.01">
<param NAME="masse13" VALUE="1035.1171">
<param NAME="intensite13" VALUE="5.86">
<param NAME="masse14" VALUE="1036.1047">
<param NAME="intensite14" VALUE="7">
<param NAME="masse15" VALUE="1045.5564">
<param NAME="intensite15" VALUE="16.55">
<param NAME="masse16" VALUE="1050.0582">
<param NAME="intensite16" VALUE="6.63">
<param NAME="masse17" VALUE="1107.5967">
<param NAME="intensite17" VALUE="12.76">
```

```
<param NAME="masse18" VALUE="1180.5675">
<param NAME="intensite18" VALUE="12.69">
<param NAME="masse19" VALUE="1196.6244">
<param NAME="intensite19" VALUE="46.65">
<param NAME="masse20" VALUE="1208.6685">

<param NAME="intensite20" VALUE="4.62">
<param NAME="masse21" VALUE="1218.6482">
<param NAME="intensite21" VALUE="8.92">
<param NAME="masse22" VALUE="1225.686">
<param NAME="intensite22" VALUE="100">
<param NAME="match7" VALUE="894.4194">
<param NAME="match8" VALUE="920.4933">
<param NAME="match10" VALUE="1006.5173">
<param NAME="match17" VALUE="1107.5967">
<param NAME="match18" VALUE="1180.5675">
<param NAME="match19" VALUE="1196.6244">
<param NAME="match22" VALUE="1225.686">
<param NAME="message" VALUE="ok">
</applet></p>
</body>
</html>
```

ANNEXE C

Code source d'une liste de paramètres requise pour l'affichage des images gel d'une expérience et la localisation de leurs taches prélevées.

```
<html>
<head>
<meta HTTP-EQUIV=Refresh CONTENT="600;URL=invivo.quitter">
<title>experiences</title>
</head>
<body TEXT="#000000"  BGCOLOR="BLUE">
<p align="center"><applet code =" imageGel2D/Gel2DApplet.class" codebase="/proteomique/"
archive="ImageGel2D.jar"  pluginspage="http://java.sun.com/products/plugin/1.4.1/plugin-install.html" width
=" 1024" height =" 850">
<param NAME="titre" VALUE="Gel 2D de souris">
<param NAME="date" VALUE="12-JAN-04" >
<param NAME="spot1" VALUE="8">
<param NAME="spot2" VALUE="1407">
<param NAME="spot3" VALUE="2004">
<param NAME="spot4" VALUE="2101">
<param NAME="spot5" VALUE="2104">
<param NAME="spot6" VALUE="2106">
<param NAME="spot7" VALUE="2302">

<param NAME="spot8" VALUE="2502">
<param NAME="spot9" VALUE="2609">
<param NAME="spot10" VALUE="2612">
<param NAME="spot11" VALUE="4001">
<param NAME="spot12" VALUE="4005">
<param NAME="spot13" VALUE="4610">
<param NAME="spot14" VALUE="4701">
<param NAME="spot15" VALUE="4702">
<param NAME="spot16" VALUE="4704">
<param NAME="spot17" VALUE="5101">
<param NAME="spot18" VALUE="5103">
<param NAME="spot19" VALUE="5105">
<param NAME="spot20" VALUE="5201">
<param NAME="spot21" VALUE="5202">
<param NAME="spot22" VALUE="5203">
<param NAME="spot23" VALUE="5301">
<param NAME="spot24" VALUE="6204">

<param NAME="spot25" VALUE="7001">
<param NAME="spot26" VALUE="7102">
<param NAME="spot27" VALUE="7103">
<param NAME="spot28" VALUE="7203">
<param NAME="spot29" VALUE="7702">
<param NAME="spot30" VALUE="8002">
<param NAME="e1" VALUE="740">
<param NAME="eurl1"
                    VALUE="proteomique.stat_echant?lechant=740#flag">
<param NAME="espot1" VALUE="4001">
<param NAME="e2" VALUE="741">
<param NAME="eurl2"
                    VALUE="proteomique.stat_echant?lechant=741#flag">
<param NAME="espot2" VALUE="2106">
```

```
<param NAME="e3" VALUE="742">
<param NAME="eurl3"
                VALUE="proteomique.stat_echant?lechant=742#flag">
<param NAME="espot3" VALUE="7702">
<param NAME="e4" VALUE="743">
<param NAME="eurl4"
                VALUE="proteomique.stat_echant?lechant=743#flag">

<param NAME="espot4" VALUE="2104">
<param NAME="e5" VALUE="744">
<param NAME="eurl5"
                VALUE="proteomique.stat_echant?lechant=744#flag">
<param NAME="espot5" VALUE="5202">
<param NAME="e6" VALUE="745">
<param NAME="eurl6"
                VALUE="proteomique.stat_echant?lechant=745#flag">
<param NAME="espot6" VALUE="5203">
<param NAME="e7" VALUE="746">
<param NAME="eurl7"
                VALUE="proteomique.stat_echant?lechant=746#flag">
<param NAME="espot7" VALUE="2101">
<param NAME="e8" VALUE="747">
<param NAME="eurl8"
                VALUE="proteomique.stat_echant?lechant=747#flag">
<param NAME="espot8" VALUE="2302">
<param NAME="e9" VALUE="748">
<param NAME="eurl9"
                VALUE="proteomique.stat_echant?lechant=748#flag">
<param NAME="espot9" VALUE="4610">
<param NAME="e10" VALUE="749">

<param NAME="eurl10"
                VALUE="proteomique.stat_echant?lechant=749#flag">
<param NAME="espot10" VALUE="4005">
<param NAME="e11" VALUE="750">
<param NAME="eurl11"
                VALUE="proteomique.stat_echant?lechant=750#flag">
<param NAME="espot11" VALUE="2612">
<param NAME="e12" VALUE="751">
<param NAME="eurl12"
                VALUE="proteomique.stat_echant?lechant=751#flag">
<param NAME="espot12" VALUE="2609">
<param NAME="e13" VALUE="752">
<param NAME="eurl13"
                VALUE="proteomique.stat_echant?lechant=752#flag">
<param NAME="espot13" VALUE="8002">
<param NAME="e14" VALUE="753">
<param NAME="eurl14"
                VALUE="proteomique.stat_echant?lechant=753#flag">
<param NAME="espot14" VALUE="5103">
<param NAME="e15" VALUE="754">
<param NAME="eurl15"
                VALUE="proteomique.stat_echant?lechant=754#flag">
<param NAME="espot15" VALUE="5105">

<param NAME="e16" VALUE="755">
<param NAME="eurl16"
                VALUE="proteomique.stat_echant?lechant=755#flag">
<param NAME="espot16" VALUE="2004">
```

```
<param NAME="e17" VALUE="756">
<param NAME="eurl17"
                    VALUE="proteomique.stat_echant?lechant=756#flag">
<param NAME="espot17" VALUE="5101">
<param NAME="e18" VALUE="757">
<param NAME="eurl18"
                    VALUE="proteomique.stat_echant?lechant=757#flag">
<param NAME="espot18" VALUE="7001">
<param NAME="e19" VALUE="758">
<param NAME="eurl19"
                    VALUE="proteomique.stat_echant?lechant=758#flag">
<param NAME="espot19" VALUE="8">
<param NAME="e20" VALUE="759">
<param NAME="eurl20"
                    VALUE="proteomique.stat_echant?lechant=759#flag">
<param NAME="espot20" VALUE="7203">
<param NAME="e21" VALUE="760">
<param NAME="eurl21"
                    VALUE="proteomique.stat_echant?lechant=760#flag">

<param NAME="espot21" VALUE="5301">
<param NAME="e22" VALUE="761">
<param NAME="eurl22"
                    VALUE="proteomique.stat_echant?lechant=761#flag">
<param NAME="espot22" VALUE="5201">
<param NAME="e23" VALUE="762">
<param NAME="eurl23"
                    VALUE="proteomique.stat_echant?lechant=762#flag">
<param NAME="espot23" VALUE="4701">
<param NAME="e24" VALUE="763">
<param NAME="eurl24"
                    VALUE="proteomique.stat_echant?lechant=763#flag">
<param NAME="espot24" VALUE="6204">
<param NAME="e25" VALUE="764">
<param NAME="eurl25"
                    VALUE="proteomique.stat_echant?lechant=764#flag">
<param NAME="espot25" VALUE="7103">
<param NAME="e26" VALUE="765">
<param NAME="eurl26"
                    VALUE="proteomique.stat_echant?lechant=765#flag">
<param NAME="espot26" VALUE="7102">
<param NAME="e27" VALUE="766">

<param NAME="eurl27"
                    VALUE="proteomique.stat_echant?lechant=766#flag">
<param NAME="espot27" VALUE="4702">
<param NAME="e28" VALUE="767">
<param NAME="eurl28"
                    VALUE="proteomique.stat_echant?lechant=767#flag">
<param NAME="espot28" VALUE="2502">
<param NAME="e29" VALUE="768">
<param NAME="eurl29"
                    VALUE="proteomique.stat_echant?lechant=768#flag">
<param NAME="espot29" VALUE="4704">
<param NAME="e30" VALUE="769">
<param NAME="eurl30"
                    VALUE="proteomique.stat_echant?lechant=769#flag">
<param NAME="espot30" VALUE="1407">
<param NAME="nomtrt1" VALUE="colonne 1">
```

<param NAME="spot1.1" VALUE="8">
<param NAME="qm1.1" VALUE="39905.09">
<param NAME="std_dev1.1" VALUE="34860.17">
<param NAME="spot1.2" VALUE="2004">
<param NAME="qm1.2" VALUE="10330.06">

<param NAME="std_dev1.2" VALUE="96.6">
<param NAME="spot1.3" VALUE="2609">
<param NAME="qm1.3" VALUE="7904.13">
<param NAME="std_dev1.3" VALUE="1303.87">
<param NAME="spot1.4" VALUE="2612">
<param NAME="qm1.4" VALUE="2906.66">
<param NAME="std_dev1.4" VALUE="70.23">
<param NAME="spot1.5" VALUE="5101">
<param NAME="qm1.5" VALUE="8260.06">
<param NAME="std_dev1.5" VALUE="7549.83">
<param NAME="spot1.6" VALUE="5103">
<param NAME="qm1.6" VALUE="9455.83">
<param NAME="std_dev1.6" VALUE="601.86">
<param NAME="spot1.7" VALUE="5105">
<param NAME="qm1.7" VALUE="9166.04">
<param NAME="std_dev1.7" VALUE="277.35">
<param NAME="spot1.8" VALUE="7001">

<param NAME="qm1.8" VALUE="13740.8">
<param NAME="std_dev1.8" VALUE="7100.08">
<param NAME="spot1.9" VALUE="7203">
<param NAME="qm1.9" VALUE="65396.82">
<param NAME="std_dev1.9" VALUE="52462.47">
<param NAME="spot1.10" VALUE="8002">
<param NAME="qm1.10" VALUE="5739.21">
<param NAME="std_dev1.10" VALUE="3500.37">
<param NAME="gel1.1" VALUE="G0402 Muscle Rat">
<param NAME="match_set1.1" VALUE="false">
<param NAME="fichier1.1" VALUE="/proteo/ImageGel?noGel=G0402">
<param NAME="tx1.1" VALUE="175">
<param NAME="ty1.1" VALUE="190">
<param NAME="gel1.2" VALUE="G0403 Muscle Rat">
<param NAME="match_set1.2" VALUE="true">
<param NAME="fichier1.2" VALUE="/proteo/ImageGel?noGel=G0403">
<param NAME="tx1.2" VALUE="181.2">

<param NAME="ty1.2" VALUE="190">
<param NAME="gel1.3" VALUE="G0408 Muscle Rat">
<param NAME="match_set1.3" VALUE="true">
<param NAME="fichier1.3" VALUE="/proteo/ImageGel?noGel=G0408">
<param NAME="tx1.3" VALUE="178.8">
<param NAME="ty1.3" VALUE="190">
<param NAME="s1.3.1" VALUE="8">
<param NAME="sx1.3.1" VALUE="5.6804">
<param NAME="sy1.3.1" VALUE="16.0502">
<param NAME="e1.3.1" VALUE="758">
<param NAME="s1.3.2" VALUE="2004">
<param NAME="sx1.3.2" VALUE="49.9506">
<param NAME="sy1.3.2" VALUE="45.184">
<param NAME="e1.3.2" VALUE="755">
<param NAME="s1.3.3" VALUE="2609">
<param NAME="sx1.3.3" VALUE="52.506">
<param NAME="sy1.3.3" VALUE="152.2553">

<param NAME="e1.3.3" VALUE="751">
<param NAME="s1.3.4" VALUE="2612">
<param NAME="sx1.3.4" VALUE="57.5896">
<param NAME="sy1.3.4" VALUE="151.78">
<param NAME="e1.3.4" VALUE="750">
<param NAME="s1.3.5" VALUE="5101">
<param NAME="sx1.3.5" VALUE="100.854">
<param NAME="sy1.3.5" VALUE="97.6937">
<param NAME="e1.3.5" VALUE="756">
<param NAME="s1.3.6" VALUE="5103">
<param NAME="sx1.3.6" VALUE="110.1338">
<param NAME="sy1.3.6" VALUE="94.5198">
<param NAME="e1.3.6" VALUE="753">
<param NAME="s1.3.7" VALUE="5105">
<param NAME="sx1.3.7" VALUE="117.7999">
<param NAME="sy1.3.7" VALUE="98.384">
<param NAME="e1.3.7" VALUE="754">

<param NAME="s1.3.8" VALUE="7001">
<param NAME="sx1.3.8" VALUE="149.0858">
<param NAME="sy1.3.8" VALUE="50.9165">
<param NAME="e1.3.8" VALUE="757">
<param NAME="s1.3.9" VALUE="7203">
<param NAME="sx1.3.9" VALUE="149.6105">
<param NAME="sy1.3.9" VALUE="113.2357">
<param NAME="e1.3.9" VALUE="759">
<param NAME="s1.3.10" VALUE="8002">
<param NAME="sx1.3.10" VALUE="153.4569">
<param NAME="sy1.3.10" VALUE="39.0005">
<param NAME="e1.3.10" VALUE="752">
<param NAME="gel1.4" VALUE="G0409 Muscle Rat">
<param NAME="match_set1.4" VALUE="false">
<param NAME="fichier1.4" VALUE="/proteo/ImageGel?noGel=G0409">
<param NAME="tx1.4" VALUE="175">
<param NAME="ty1.4" VALUE="190">

<param NAME="nomtrt2" VALUE="colonne 2">
<param NAME="spot2.1" VALUE="2101">
<param NAME="qm2.1" VALUE="3836.61">
<param NAME="std_dev2.1" VALUE="4008.68">
<param NAME="spot2.2" VALUE="2104">
<param NAME="qm2.2" VALUE="1286.53">
<param NAME="std_dev2.2" VALUE="631.88">
<param NAME="spot2.3" VALUE="2106">
<param NAME="qm2.3" VALUE="798.98">
<param NAME="std_dev2.3" VALUE="800.48">
<param NAME="spot2.4" VALUE="2302">
<param NAME="qm2.4" VALUE="23841.78">
<param NAME="std_dev2.4" VALUE="13190.84">
<param NAME="spot2.5" VALUE="4001">
<param NAME="qm2.5" VALUE="2268.51">
<param NAME="std_dev2.5" VALUE="1132.5">
<param NAME="spot2.6" VALUE="4005">

<param NAME="qm2.6" VALUE="18911.34">
<param NAME="std_dev2.6" VALUE="2576.84">
<param NAME="spot2.7" VALUE="4610">
<param NAME="qm2.7" VALUE="20646.65">

<param NAME="std_dev2.7" VALUE="5209.96">
<param NAME="spot2.8" VALUE="5202">
<param NAME="qm2.8" VALUE="2462.62">
<param NAME="std_dev2.8" VALUE="1402.87">
<param NAME="spot2.9" VALUE="5203">
<param NAME="qm2.9" VALUE="4189.04">
<param NAME="std_dev2.9" VALUE="4458.66">
<param NAME="spot2.10" VALUE="7702">
<param NAME="qm2.10" VALUE="935.69">
<param NAME="std_dev2.10" VALUE="950.75">
<param NAME="gel2.1" VALUE="G0404 Muscle Rat">
<param NAME="match_set2.1" VALUE="false">
<param NAME="fichier2.1" VALUE="/proteo/ImageGel?noGel=G0404">

<param NAME="tx2.1" VALUE="175">
<param NAME="ty2.1" VALUE="190">
<param NAME="gel2.2" VALUE="G0405 Muscle Rat">
<param NAME="match_set2.2" VALUE="true">
<param NAME="fichier2.2" VALUE="/proteo/ImageGel?noGel=G0405">
<param NAME="tx2.2" VALUE="180.1">
<param NAME="ty2.2" VALUE="190">
<param NAME="s2.2.1" VALUE="2101">
<param NAME="sx2.2.1" VALUE="53.9399">
<param NAME="sy2.2.1" VALUE="87.9914">
<param NAME="e2.2.1" VALUE="746">
<param NAME="s2.2.2" VALUE="2104">
<param NAME="sx2.2.2" VALUE="51.9303">
<param NAME="sy2.2.2" VALUE="76.9262">
<param NAME="e2.2.2" VALUE="743">
<param NAME="s2.2.3" VALUE="2106">
<param NAME="sx2.2.3" VALUE="44.8246">

<param NAME="sy2.2.3" VALUE="72.2897">
<param NAME="e2.2.3" VALUE="741">
<param NAME="s2.2.4" VALUE="2302">
<param NAME="sx2.2.4" VALUE="49.1478">
<param NAME="sy2.2.4" VALUE="122.1119">
<param NAME="e2.2.4" VALUE="747">
<param NAME="s2.2.5" VALUE="4001">
<param NAME="sx2.2.5" VALUE="82.9487">
<param NAME="sy2.2.5" VALUE="55.6679">
<param NAME="e2.2.5" VALUE="740">
<param NAME="s2.2.6" VALUE="4005">
<param NAME="sx2.2.6" VALUE="97.9305">
<param NAME="sy2.2.6" VALUE="48.4591">
<param NAME="e2.2.6" VALUE="749">
<param NAME="s2.2.7" VALUE="4610">
<param NAME="sx2.2.7" VALUE="94.2291">
<param NAME="sy2.2.7" VALUE="154.5752">

<param NAME="e2.2.7" VALUE="748">
<param NAME="s2.2.8" VALUE="5202">
<param NAME="sx2.2.8" VALUE="98.6549">
<param NAME="sy2.2.8" VALUE="112.832">
<param NAME="e2.2.8" VALUE="744">
<param NAME="s2.2.9" VALUE="5203">
<param NAME="sx2.2.9" VALUE="102.6032">
<param NAME="sy2.2.9" VALUE="101.6786">
<param NAME="e2.2.9" VALUE="745">

<param NAME="s2.2.10" VALUE="7702">
<param NAME="sx2.2.10" VALUE="146.7621">
<param NAME="sy2.2.10" VALUE="167.4022">
<param NAME="e2.2.10" VALUE="742">
<param NAME="gel2.3" VALUE="G0410 Muscle Rat">
<param NAME="match_set2.3" VALUE="true">
<param NAME="fichier2.3" VALUE="/proteo/ImageGel?noGel=G0410">
<param NAME="tx2.3" VALUE="181.6">

<param NAME="ty2.3" VALUE="190">
<param NAME="gel2.4" VALUE="G0411 Muscle Rat">
<param NAME="match_set2.4" VALUE="false">
<param NAME="fichier2.4" VALUE="/proteo/ImageGel?noGel=G0411">
<param NAME="tx2.4" VALUE="175">
<param NAME="ty2.4" VALUE="190">
<param NAME="nomtrt3" VALUE="colonne 3">
<param NAME="spot3.1" VALUE="1407">
<param NAME="qm3.1" VALUE="219393.98">
<param NAME="std_dev3.1" VALUE="40084.65">
<param NAME="spot3.2" VALUE="2502">
<param NAME="qm3.2" VALUE="12069.82">
<param NAME="std_dev3.2" VALUE="13045.45">
<param NAME="spot3.3" VALUE="4701">
<param NAME="qm3.3" VALUE="3121.53">
<param NAME="std_dev3.3" VALUE="1774.95">
<param NAME="spot3.4" VALUE="4702">

<param NAME="qm3.4" VALUE="10242.15">
<param NAME="std_dev3.4" VALUE="881.17">
<param NAME="spot3.5" VALUE="4704">
<param NAME="qm3.5" VALUE="20075.73">
<param NAME="std_dev3.5" VALUE="2385.29">
<param NAME="spot3.6" VALUE="5201">
<param NAME="qm3.6" VALUE="3777.55">
<param NAME="std_dev3.6" VALUE="1379.74">
<param NAME="spot3.7" VALUE="5301">
<param NAME="qm3.7" VALUE="2807.11">
<param NAME="std_dev3.7" VALUE="1258.64">
<param NAME="spot3.8" VALUE="6204">
<param NAME="qm3.8" VALUE="5421.81">
<param NAME="std_dev3.8" VALUE="1720.73">
<param NAME="spot3.9" VALUE="7102">
<param NAME="qm3.9" VALUE="8130">
<param NAME="std_dev3.9" VALUE="1618.05">

<param NAME="spot3.10" VALUE="7103">
<param NAME="qm3.10" VALUE="4145.71">
<param NAME="std_dev3.10" VALUE="4580.92">
<param NAME="gel3.1" VALUE="G0406 Muscle Rat">
<param NAME="match_set3.1" VALUE="false">
<param NAME="fichier3.1" VALUE="/proteo/ImageGel?noGel=G0406">
<param NAME="tx3.1" VALUE="175">
<param NAME="ty3.1" VALUE="190">
<param NAME="gel3.2" VALUE="G0407 Muscle Rat">
<param NAME="match_set3.2" VALUE="true">
<param NAME="fichier3.2" VALUE="/proteo/ImageGel?noGel=G0407">
<param NAME="tx3.2" VALUE="189.2">
<param NAME="ty3.2" VALUE="190">
<param NAME="gel3.3" VALUE="G0412 Muscle Rat">

<param NAME="match_set3.3" VALUE="true">
<param NAME="fichier3.3" VALUE="/proteo/ImageGel?noGel=G0412">
<param NAME="tx3.3" VALUE="177.1">

<param NAME="ty3.3" VALUE="190">
<param NAME="s3.3.1" VALUE="1407">
<param NAME="sx3.3.1" VALUE="40.3139">
<param NAME="sy3.3.1" VALUE="120.0713">
<param NAME="e3.3.1" VALUE="769">
<param NAME="s3.3.2" VALUE="2502">
<param NAME="sx3.3.2" VALUE="46.5119">
<param NAME="sy3.3.2" VALUE="129.4216">
<param NAME="e3.3.2" VALUE="767">
<param NAME="s3.3.3" VALUE="4701">
<param NAME="sx3.3.3" VALUE="81.503">
<param NAME="sy3.3.3" VALUE="153.5846">
<param NAME="e3.3.3" VALUE="762">
<param NAME="s3.3.4" VALUE="4702">
<param NAME="sx3.3.4" VALUE="86.3686">
<param NAME="sy3.3.4" VALUE="153.3208">
<param NAME="e3.3.4" VALUE="766">

<param NAME="s3.3.5" VALUE="4704">
<param NAME="sx3.3.5" VALUE="91.8833">
<param NAME="sy3.3.5" VALUE="152.6452">
<param NAME="e3.3.5" VALUE="768">
<param NAME="s3.3.6" VALUE="5201">
<param NAME="sx3.3.6" VALUE="99.6159">
<param NAME="sy3.3.6" VALUE="101.0378">
<param NAME="e3.3.6" VALUE="761">
<param NAME="s3.3.7" VALUE="5301">
<param NAME="sx3.3.7" VALUE="105.5751">
<param NAME="sy3.3.7" VALUE="108.5299">
<param NAME="e3.3.7" VALUE="760">
<param NAME="s3.3.8" VALUE="6204">
<param NAME="sx3.3.8" VALUE="139.5627">
<param NAME="sy3.3.8" VALUE="70.188">
<param NAME="e3.3.8" VALUE="763">
<param NAME="s3.3.9" VALUE="7102">

<param NAME="sx3.3.9" VALUE="151.8519">
<param NAME="sy3.3.9" VALUE="41.3632">
<param NAME="e3.3.9" VALUE="765">
<param NAME="s3.3.10" VALUE="7103">
<param NAME="sx3.3.10" VALUE="160.6648">
<param NAME="sy3.3.10" VALUE="61.5009">
<param NAME="e3.3.10" VALUE="764">
<param NAME="gel3.4" VALUE="G0413 Muscle Rat">
<param NAME="match_set3.4" VALUE="false">
<param NAME="fichier3.4" VALUE="/proteo/ImageGel?noGel=G0413">
<param NAME="tx3.4" VALUE="175">
<param NAME="ty3.4" VALUE="190">
</applet></p>
</body>
</html>

ANNEXE D

Schéma relationnel de la base de données (ci-inclus).

Proteomique

Schéma relationnel de la base de données protéomique