Université des Sciences et Technologies de Lille

D.E.S.S. de Bioinformatique

Spot Manager : Un nouvel outil d'aide à la gestion des spots identifiés par spectrométrie de masse

Pierre LAURENCE

Responsable de stage : Dr Florence PINET

INSERM U508
INSTITUT PASTEUR DE LILLE

Mai-septembre 2004

Remerciements

Je tiens à remercier Monsieur le Professeur Philippe Amouyel de m'avoir accueilli au sein de son unité.

Je remercie le Docteur Florence Pinet pour m'avoir laissé une autonomie dans le déroulement de ce projet, et surtout pour m'avoir fait confiance dans mes choix.

Je remercie particulièrement Annabelle Dupont, Olivia Beseme et Philippe Ratajczak pour leur intérêt, leurs remarques et leurs idées.

Enfin, un grand merci à tous les autres membres de l'unité INSERM 508, les chercheurs, les techniciens, et les nombreux stagiaires que j'ai pu rencontrer, pour leur sympathie et la bonne ambiance de travail.

Table des matières

INTRODUCTION....4

1 CADRE ET PROJET DU STAGE....5

1.1 L'unité INSERM 508....5
1.2 L'analyse protéomique....6
1.3 Le logiciel Spot Manager....8

2 OUTILS ET TECHNOLOGIES UTILISES....11

2.1 Les logiciels bioinformatiques....11
2.1.1 Les logiciels d'analyses 2D....11
2.1.2 Les logiciels d'identification de protéines....13
2.2 Les bases de données publiques....14
2.3 Les choix technologiques informatiques....15
2.3.1 Le langage de programmation....15
2.3.2 Le SGBD....16
2.3.3 Le format de fichier XML....16

3 PRESENTATION DE SPOT MANAGER....17

3.1 L'architecture générale....17
3.2 Vue d'ensemble de Spot Manager....18
3.3 L'insertion de nouveaux spots....20
3.4 Le site WEB de visualisation de gel....21

BIBLIOGRAPHIE....24

Introduction

L' objectif de ce rapport est de présenter le travail que j'ai réalisé au cours de cinq mois de stage. Ce document n'est pas une documentation technique et encore moins une documentation pour l'utilisateur de Spot Manager, le logiciel que j'ai conçu. C'est plutôt une vision d'ensemble qui permet de comprendre le déroulement de sa conception ainsi que d'apprécier ses caractéristiques.

Conscient de m'adresser à un public mixte (biochimistes, biologistes et bioinformaticiens), j'ai voulu garder un langage et un discours ouvert à tous. J'espère donc que chacun trouvera son compte et comprendra ainsi l'intérêt de mon stage.

Tout d'abord, je vais préciser mon cadre de travail. Cette partie décrit l'unité de recherche qui m'a accueilli, les bases de l'analyse protéomique et finalement comment Spot Manager se propose d'intervenir dans ce processus. Ensuite, je présenterai les outils qui m'ont été nécessaires pour sa réalisation. Enfin, je montrerai concrètement les fonctionnalités du travail réalisé. En guise de conclusion, j'indiquerai rapidement comment j'ai vécu ce stage ainsi que les perspectives de mon travail.

1 Cadre et projet du stage

1.1 L'unité INSERM 508

J'ai réalisé mon stage à l'**Institut Pasteur de Lille** au sein du laboratoire de l'unité **INSERM 508**. Celle-ci étudie les déterminants génétiques et environnementaux des maladies neurodégénératives et cardiovasculaires.

Cinq thèmes de recherches sont étudiés pour le moment :

- Epidémiologie descriptive et analytique des maladies cardiovasculaires
- Analyse fonctionnelle des interactions gène-environnement dans les maladies cardio-vasculaires
- Altération des fonctions cognitives. Liens avec les facteurs de susceptibilité génétique des maladies cardio-vasculaires
- Facteurs de susceptibilité génétique des maladies neurodégénératives
- Pharmacogénétique des maladies chroniques

Les différents thèmes sont abordés principalement sous l'angle de l'**épidémiologie**, de la **génétique** et de la **protéomique**.

J'ai travaillé dans l'équipe impliquée dans la recherche de facteurs de risque chez les personnes atteintes d'anévrysme de l'aorte abdominale par les techniques d'analyses protéomiques.

Les développements récents des techniques de biologie moléculaire à haut débit et de protéomique, essentiellement mises en place dans le cadre de la **Génopole Nord-Pas-de-Calais**, ont amené l'équipe à développer des stratégies de recherche de nouveaux facteurs de susceptibilités génétiques. Par des techniques d'étude différentielle des protéines et des ARN messagers, l'unité cherche à mettre en évidence de nouveaux gènes impliqués dans les interactions gène-environnement. L'étude débute ces approches par la mise en place d'un modèle fondé sur des macrophages issus de patients qui présentent des anévrysmes de l'aorte abdominale comparés à ceux des sujets sans anévrysme. L'identification d'une ou plusieurs protéines, connues ou non, grâce à une **analyse protéomique comparative**, permettra de caractériser des séquences génomiques à partir desquelles l'identification de nouveaux facteurs de susceptibilité génétique devient possible. La finalité des recherches apporterait de nouvelles perspectives de prévention ainsi que de prises en charge thérapeutiques.

Actuellement, l'**indexation des macrophages humains** est finie et a été publiée [1]. Jusqu'à présent, aucune cartographie en gel 2-D de macrophages n'avait été disponible.

1.2 L'analyse protéomique

A la base de l'analyse protéomique [2], on trouve principalement une technique : l'**électrophorèse en gel bidimensionnel** ou **2D**. Celle-ci permet d'établir une carte (un **gel**) qui contient un très grand nombre de protéines. Après une étape de préparation des échantillons (l'extraction), l'électrophorèse sépare les protéines selon deux dimensions. La première est le point isoélectrique (**pI**), la seconde la masse moléculaire (**Mr**). La figure 1.1 illustre schématiquement ces deux dimensions.

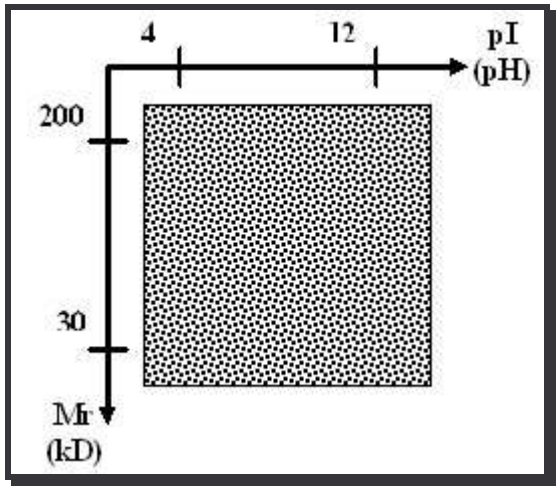

FIGURE. 1.1 – *Electrophorèse 2D schématique*

Grâce à une coloration (au nitrate d'argent par exemple), la visualisation des **spots** (« taches ») de protéines devient possible (voir FIGURE 1.2). La comparaison de plusieurs gels (typiquement deux groupes : témoins et malades) grâce à un **logiciel d'analyse 2D** (couplé à un scanner d'image haut de gamme) permet de repérer des protéines d'intérêt. Celles-ci présentent des différences significatives dans leur niveau d'expression selon les groupes de gels.

Ensuite, les spots des protéines d'intérêt sont excisés du gel puis fragmentée en peptides grâce à une **digestion** par une enzyme (typiquement la Trypsine qui coupe après tous les acides aminés basiques Arg et Lys). Une autre technique est mise à contribution : la **spectrométrie de masse**.

Le principe de la spectrométrie de masse MALDI-TOF (*Matrix Assisted Laser Desorption and Ionisation*) consiste à irradier un échantillon co-cristallisé dans une matrice qui absorbe à la longueur d'onde du rayonnement laser. Les ions mono-chargés ainsi formés sont accélérés puis évoluent selon leur masse dans une zone de vide avant d'être détectés par un analyseur. Les masses des peptides mesurés (voir FIGURE 1.3) par MALDI-TOF constituent ainsi l'empreinte peptidique des protéines du mélange.

Chaque protéine possède une **empreinte peptidique** qui lui est propre du fait de l'enchaînement particulier des acides aminés qui lui confère, pour une méthode de digestion donnée, une combinaison unique de masses peptidiques. Cette empreinte peptidique est comparée aux empreintes peptidiques issues de la digestion théorique (*in silico*) des protéines répertoriées dans les banques de données (voir FIGure 1.4) grâce à des logiciels spécialisés. Finalement, l'identification des protéines est fondée sur la reconnaissance d'un **nombre maximal** de peptides issus de l'empreinte peptidique avec les candidates proposées.

L'analyse par électrophorèse 2D est actuellement la technique la plus puissante pour séparer des protéines. En une seule étape, on a la possibilité d'étudier **plusieurs milliers** de protéines. La technologie est compatible avec la grande majorité des types de polypeptides. Les gels avec des gammes de pI restreintes permettent une très **grande résolution**. En plus, on a la possibilité d'observer des protéines faiblement exprimées. Ces caractéristiques font de l'analyse protéomique une **méthode sans a priori**.

Alors que les scientifiques veulent comparer plusieurs gels entre eux, il faut savoir que la qualité et la reproductibilité des gels 2D sont parfois **médiocres** [3]. Or, une faible résolution, un niveau de bruit de fond élevé ou des distorsions avancées entravent fortement l'analyse. Les avancées techniques tendent à réduire ces inconvénients. En effet, les algorithmes des logiciels d'analyse de gel 2D ont été significativement améliorés.

Malgré ces avancées, aucun logiciel d'analyse ne peut résoudre les problèmes d'ordre expérimental. Le désir d'obtenir des informations à partir de gels de faible qualité continue à être illusoire. Avant de commencer une étude d'image de gel 2D, il apparaît essentiel d'**optimiser** toutes les étapes dans la production des gels pour assurer les meilleures qualité et reproductibilité possibles. Cela passe par un protocole expérimental bien précis pour la préparation des échantillons, une utilisation précise et correcte des conditions expérimentales pendant la séparation 2D et la coloration, une utilisation de gels plus grands ou avec une gamme de pI plus restreinte. Les paramètres d'acquisition d'image jouent aussi un rôle très important (haute résolution illustrée par le détail de la FIGure 1.2).

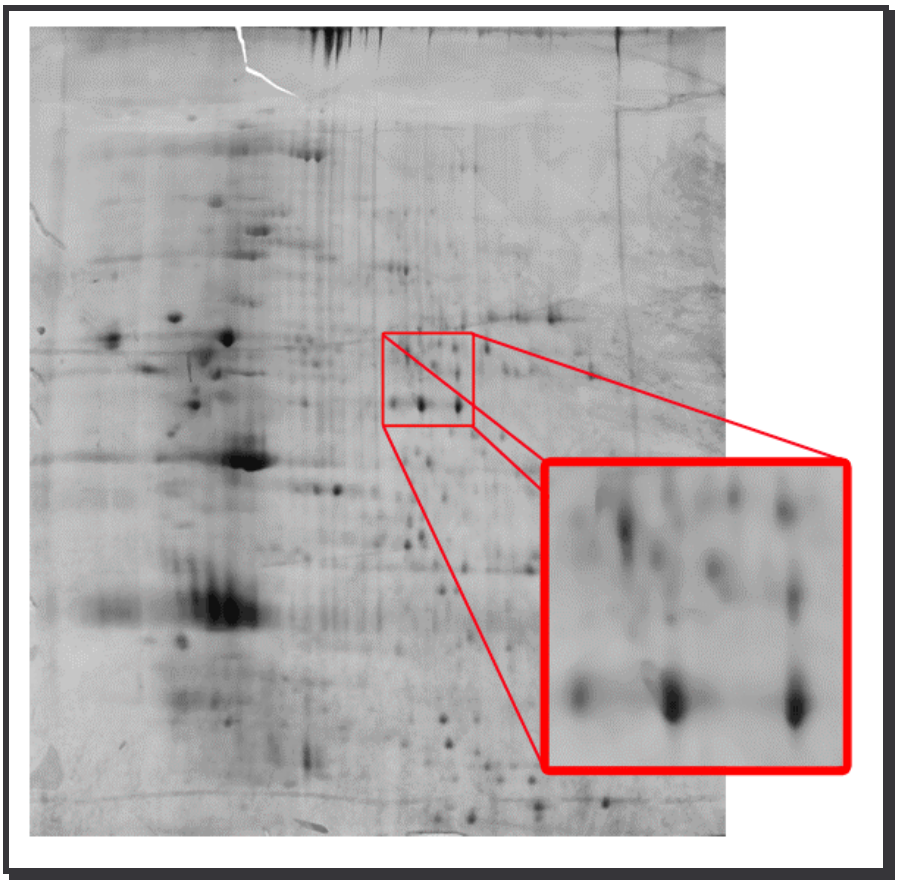

FIGURE 1.2 – *Image d'un gel d'électrophorèse 2D (macrophages humains). Détail d'une zone de 200 x 200 pixels*

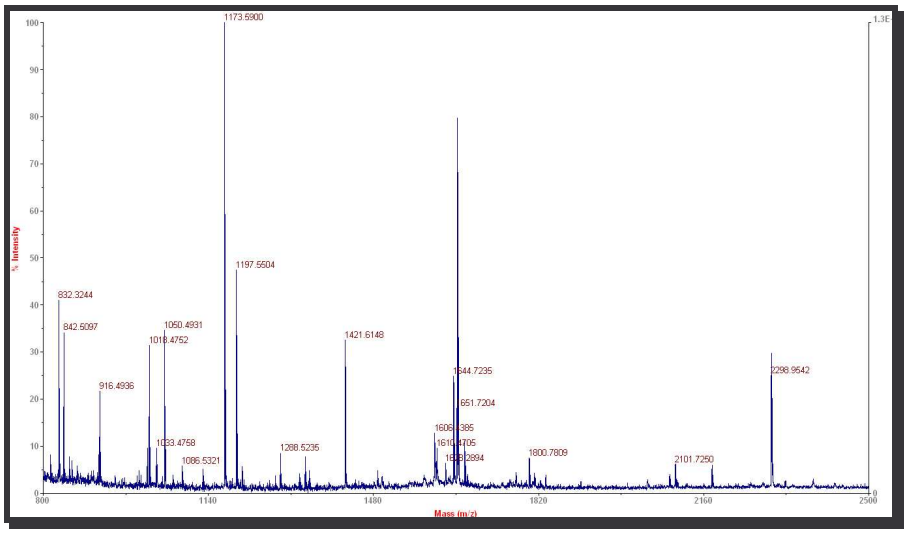

FIGURE. 1.3 – *Exemple de pics de masses issus de la spectrométrie de masse (Aldolase C. de rat)*

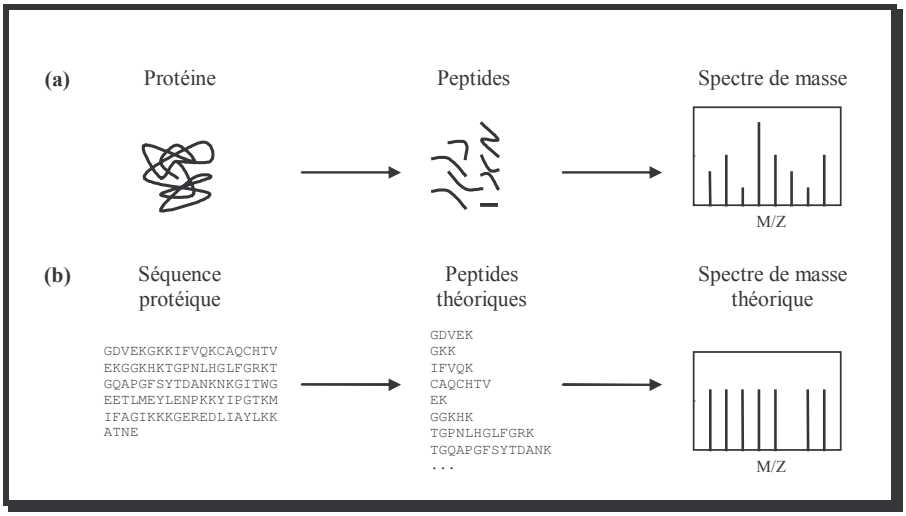

FIGURE 1.4 – *Identification de protéine par comparaison d'empreintes peptidiques :*
(a) Digestion expérimentale avec une enzyme et mesure des masses des peptides par spectrométrie de masse
(b) Dans une banque de protéine, chaque séquence est digérée théoriquement. Les masses des peptides obtenues sont calculées et un spectre de masse théorique est construit.

1.3 Le logiciel Spot Manager

L'objectif majeur du stage était de fournir une **interface WEB** pour mettre à la disposition de la communauté scientifique les informations relatives à la carte des macrophages humains. Le site devait posséder au moins les fonctionnalités essentielles suivantes :

- Affichage des informations relatives à un gel
- Affichage des informations relatives aux spots
- Visualisation du gel et des spots identifiés

Bien sur, il fallait aussi garder à l'esprit qu'à terme, d'autres gels pourraient être intégrés au site et que leurs insertions devraient être aisées.

Tout d'abord, il apparaît indispensable d'établir une **base de données** qui stocke les gels numérisés ainsi que les différentes informations souhaitées. L'étape préliminaire consistait justement à établir précisément quelles seraient ces dites informations (voir TABLEAU. 1.1).

TABLEAU 1.1 – *Les données à gérer*

Spot	Gel
➢ Localisation sur le gel ➢ Mr, pI expérimentaux ➢ Protéine associée : o nom o accession dans les banques protéiques o fonction o tissu o taux de couverture o nombre de peptides matchés o Mr, pI théoriques	➢ Image numérisée ➢ Titre du projet ➢ Type de cellules ➢ Type d'extraction ➢ Type de matrice ➢ Auteur ➢ Expérimentateur ➢ Date de création, de publication

Pendant cette démarche, j'ai observé comment ces données étaient obtenues. A la sortie de spectromètre de masse, on obtient des fichiers qui contiennent les listes de masses qui représentent les empreintes peptidiques respectives. Ensuite, un biologiste identifie manuellement les protéines associées précisément. Tout d'abord, ces listes de masses sont soumises aux logiciels d'identification de protéines via interface WEB. Les résultats des requêtes sont sauvegardés. Les résultats sont analysés pour dégager une identification commune associée à une empreinte peptidique. Ensuite, chaque protéine identifiée voit ses caractéristiques complétées grâce à l'usage des informations dans les banques de données publiques. Un fichier Excel, qui comporte toutes ces informations, est ainsi complété au fur et à mesure.

C'est un travail qui m'a paru long, fastidieux et répétitif. Les formulaires à remplir pour l'identification de protéines requièrent toujours les mêmes paramètres. Ensuite, le nombre de fichiers à sauvegarder devient très vite conséquent et, sans organisation particulière, la gestion devient assez difficile. Enfin, quand les informations sur les protéines se trouvent déjà dans des banques, celles-ci ne nécessitent alors pas l'aide de la bibliographie et se limitent à des simples « copier-coller ».

L'**automatisation** de l'acquisition et la **fédération** de ces informations m'ont rapidement semblé réalisables. Finalement, le concept de Spot Manager était né !

Spot Manager assure ainsi plusieurs rôles :

- Automatisation pour :
  - l'identification des spots.
  - la récupération et la mise à jour de nombreuses informations pour les proteines.
  - la mise à jour du site WEB.
- Fédération des données :
  - suivi de l'évolution au cours du temps de la cartographie d'un gel (introduction de la notion de « projet »).
  - toutes les informations accessibles par la même interface.
  - le site WEB est juste le reflet de plusieurs projets.

La figure de la page suivante (FIGURE 1.5) résume le processus de l'analyse protéomique et illustre la position de Spot Manager au sein de celui-ci.

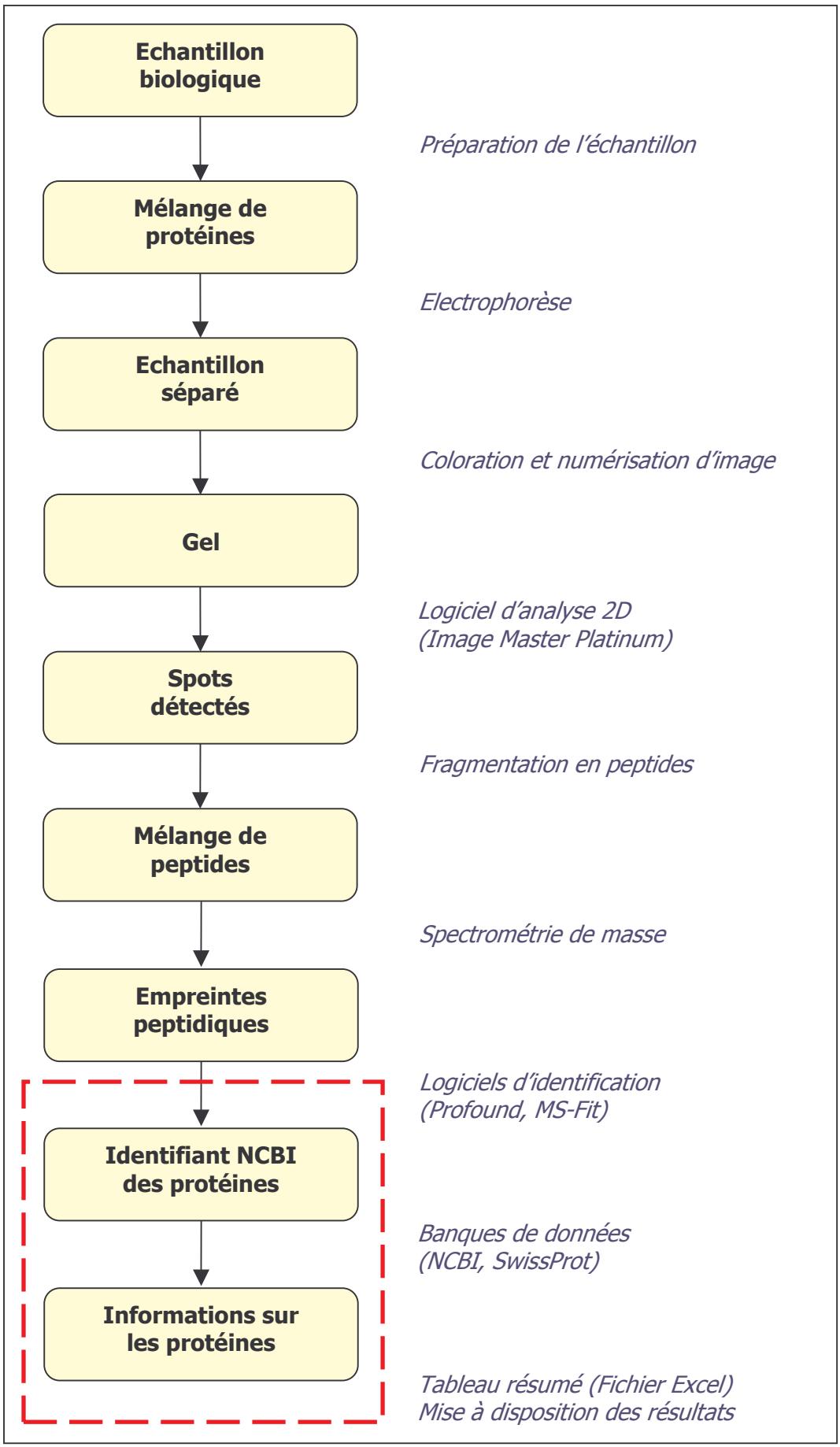

FIGURE 1.5 – *Processus d'analyse protéomique – en pointillés l'automatisation par Spot Manager*

2 Outils et technologies utilisés

2.1 Les logiciels bioinformatiques

2.1.1 Les logiciels d'analyses 2D

Une analyse de gel 2D comporte typiquement la série d'étapes suivantes :

⇒ *Acquisition des données* :
La résolution avec laquelle est scannée le gel est primordiale et influence le degré de détail visible sur l'image. Si la résolution est trop basse, les spots ne peuvent pas être distingués individuellement. A contrario, une résolution trop haute entraîne une taille des fichiers images plus importante et ralentira l'analyse de gel.
Au laboratoire, les gels sont scannés avec une résolution de 300 dpi en niveau de gris. De plus, l'ordinateur réservé à l'analyse contient une grande quantité de mémoire vive (1 go).

⇒ *Calibration des gels* :
Cette étape est indispensable pour compenser les différences d'images entre les gels.

⇒ *Détection des spots* (voir FIGURE 2.1):
Une détection automatique et efficace des spots par un logiciel est nécessaire pour identifier un maximum de protéines. Celle-ci doit minimiser le nombre d'artefacts détectés incorrectement comme des spots et surtout maximiser le nombre de spots détectés. L'algorithme de détection doit être un bon compromis entre la sensibilité et la spécificité des spots. L'édition manuelle des spots est parfois nécessaire.

⇒ *Quantification des spots* (voir FIGURE 2.2):
Une fois les spots détectés, un algorithme calcule la quantité (intensité, aire et volume) de protéine présente pour chaque spot.

⇒ *Matching des gels* (voir FIGURE 2.3):
Une fois les spots détectés sur plusieurs gels, le match des gels est possible : c'est à dire que l'on peut trouver une correspondance entre les spots des différents gels. Cette étape est faite automatiquement par le logiciel après avoir placé quelques landmarks (un petit nombre de correspondances de spots faites manuellement). De plus, on peut créer un gel référence qui contient tous les spots qui apparaissent dans la majorité des gels : ce gel servira d'image pour la carte de référence.

Les gels 2D de l'unité sont analysés avec le logiciel **Image Master 2D Platinum** (version 5.0). Ce logiciel est basé à partir de Melanie 5 [4] et a été développé par une équipe de chercheurs du SIB (*Swiss Institute of Bioinformatics*) en collaboration avec Geneva Bioinformatics (*GeneBio*) S.A. et *Amersham Biosciences*.

Ce logiciel est rapidement abordable grâce à son interface utilisateur agréable, sa documentation et son didacticiel. Je me suis rapidement formé aux fonctionnalités proposées par celui-ci. Une d'entre elle a d'ailleurs retenu particulièrement mon attention : l'**exportation de données**. Il permet à Spot Manager de récupérer des données essentielles liés aux spots : les coordonnées de ceux-ci sur le gel (issues de la détection), ainsi que leurs données expérimentales (issus de la quantification). Sans oublier que ce logiciel est aussi un moyen de récupérer l'image du gel.

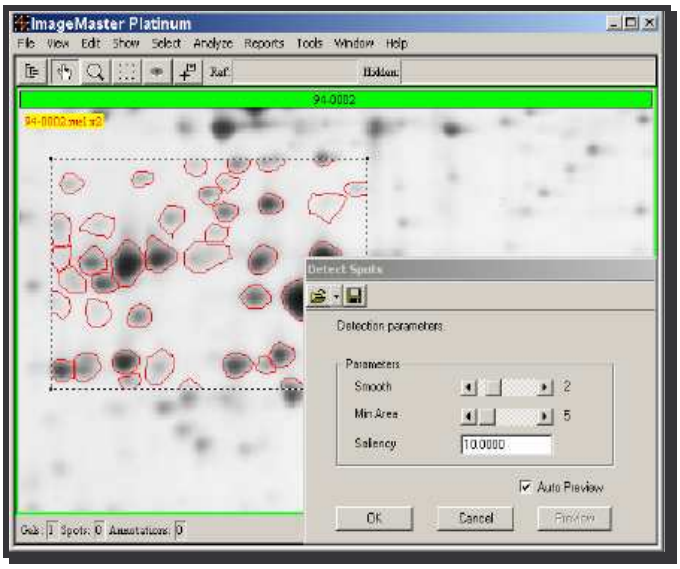

FIGURE 2.1 – *Détection de spots (sous Image Master 2D Platinum)*

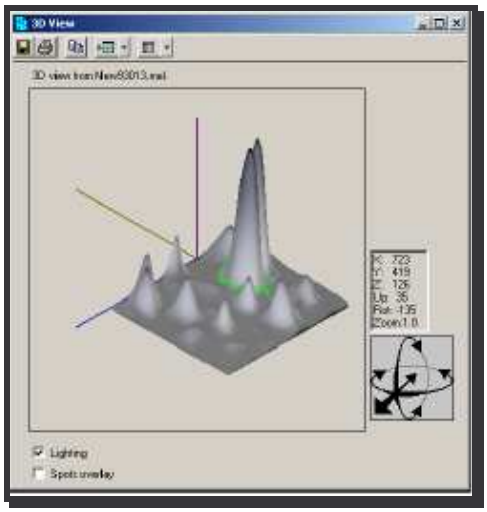

FIGURE 2.2 – *Visualisation 3D d'un spot - intensité, aire et volume issus de la quantification (sous Image Master 2D Platinum)*

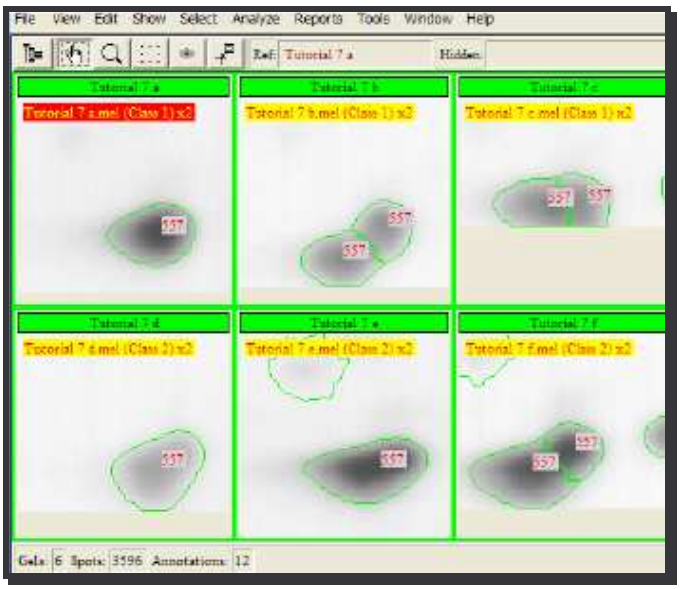

FIGURE 2.3 – *Match d'un spot avec d'autres spots de plusieurs gels (sous Image Master 2D Platinum)*

2.1.2 Les logiciels d'identification de protéines

La spectrométrie de masse réalisée, l'identification de protéines correspond à l'association d'une empreinte peptidique (une liste de masses) avec une protéine. Cette association est faite par des logiciels spécialisés [5]. A une liste de masses, ceux-ci fournissent une liste de protéines candidates triées dans un ordre particulier grâce à une fonction de score (de la plus à la moins probable). Cette liste n'est pas exactement la même selon les logiciels utilisés. En effet, les stratégies d'identification [6] ne sont pas forcément les mêmes. Afin d'avoir des résultats probants, les identifications faites au laboratoire sont effectuées grâce à **deux** logiciels disponibles sur le WEB (**MsFit** et **Profound**). Les deux listes de candidates sont alors comparées : si à une empreinte peptidique correspond une même protéine candidate dans les deux listes alors elle est une bonne candidate et est retenue. Mon objectif était donc de reproduire cette méthode d'identification de manière **automatique**.

Profound [7] est un système expert[1] pour l'identification de protéines. Il utilise la théorie de Bayés pour trier les candidates selon leurs probabilités d'occurrence. Cette approche permet d'inclure différents types d'informations de façon naturelle et augmente la sensibilité et la sélectivité de l'algorithme. Le serveur WEB qui héberge Profound (Université Rockefeller de New York) est rapide et les requêtes avec les paramètres utilisés en routine ont été totalement automatisées relativement rapidement.

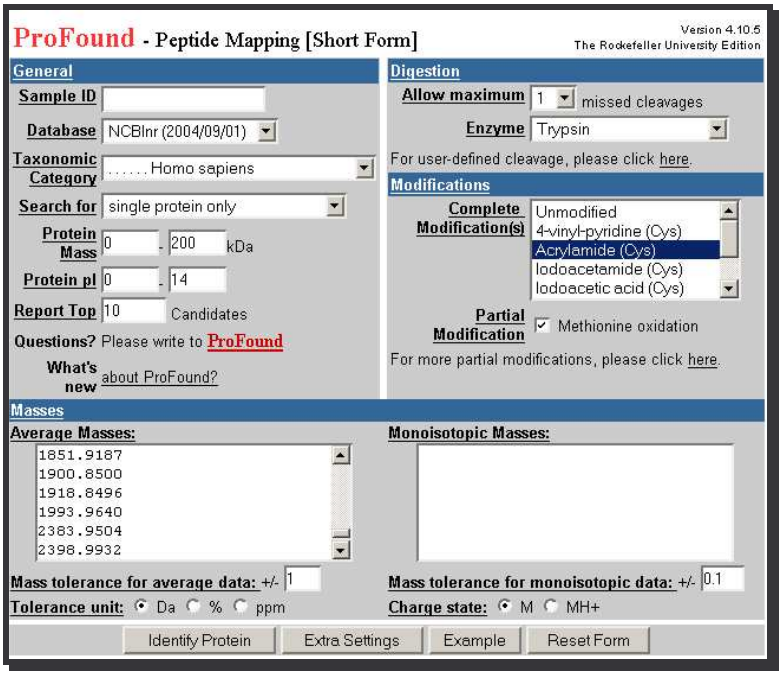

FIGURE 2.4 – *Interface WEB de Profound*

Pour le cas de MsFit, la réalisation du même travail a posé un problème. Les données envoyées au serveur sont encapsulées dans « une boite noire ». J'ai envoyé un mail au responsable du serveur (Université de Californie de San Francisco) pour savoir comment résoudre cette difficulté. Celui-ci m'a répondu que leur serveur ne pouvait subvenir à mes besoins d'automatisation. Par contre, il m'a indiqué que l'achat d'une licence MsFit résoudrait parfaitement mes problèmes… C'est vers une autre alternative que je me suis tourné.

[1] Système expert : logiciel qui exploite dans un domaine particulier des connaissances explicites et organisées et pouvant se substituer à un expert humain.

Au cours de ma maîtrise informatique, j'ai réalisé pour l'UMR CNRS 8009 de Chimie Organique et Macromoléculaire de l'Université des Sciences et Technologies de Lille, un logiciel d'identification de protéines par empreinte peptidique: **ASCQ-PROT**. Ce logiciel se différencie des autres par certaines caractéristiques. Parmi celles-ci, il est **libre de droits** (avec le code source accessible à tous). Ensuite, son utilisation est locale (indépendante du WEB). Et enfin, sa fonction de score (qui permet le classement des protéines candidates) est **paramétrable** et adaptable aux besoins.

Son principe de fonctionnement est un peu moins sophistiqué que Profound. ASCQ-PROT effectue les digestions enzymatiques théoriques des séquences dans une banque de protéines et les compare aux données issues de la spectrométrie de masse. Les candidates sont triées selon une fonction de score basée sur le taux de couverture de la protéine, le nombre de peptides matchés, la précision associée…

Bien qu'il ait fait l'objet d'un abstract lors des 20èmes journées françaises de spectrométrie de masse (septembre 2003), et plus récemment d'une présentation lors d'une journée organisée par Rhône-Alpes Génopole (juin 2004), ce logiciel n'avait jusque là jamais été autant testé. En conditions réelles, j'ai pu améliorer ses paramètres d'utilisation.

2.2 Les bases de données publiques

Historiquement au moins, la motivation du développement des banques protéiques généralistes n'est pas seulement, comme dans le cas des grandes banques nucléotidiques, d'établir une collection complète des séquences disponibles, mais aussi de constituer une ressource de connaissances centrée sur les protéines et leurs propriétés biologiques. Les critères de qualité qu'elles essaient d'appliquer sont notamment la **non redondance**, l'homogénéité de l'information, et la valeur scientifique des **annotations**. En plus du travail de collecte, stockage, gestion et diffusion des données, les auteurs effectuent un travail colossal de nettoyage, tri, documentation. La redondance y est définie, identifiée et éliminée. Des annotations, c'est-à-dire des informations telles que la localisation cellulaire, des modifications post-traductionnelles, des caractéristiques fonctionnelles, des références à d'autres banques de données, extraites de la littérature ou issues d'analyses bioinformatiques sont ajoutées.

Le processus d'identification de protéines renvoie à un numéro d'accession de la banque **nrprot** du NCBI (National Center for Biotechnology Information). Celle-ci est une banque américaine de protéines qui a pour objectif d'être exhaustive, mais qui sacrifie pour l'atteindre tous les critères de qualité évoqués précédemment.

Néanmoins, d'un autre coté, on trouve **SwissProt** une banque de protéines réputée pour la qualité des informations qu'on y trouve. SwissProt est actuellement développée en collaboration par le SIB (Swiss Institute of Bioinformatics) et l'EBI (European Bioinformatics Institute). C'est certainement la banque de séquences protéiques la moins redondante et la mieux annotée. Les séquences sont soumises directement à SwissProt par leurs auteurs, ou (majoritairement maintenant) extraites des séquences de la banque nucléotidique EMBL. Le processus d'intégration d'une nouvelle séquence dans SwissProt comprend plusieurs étapes (passage par une banque intermédiaire TrEMBL entre autre) visant à contrôler la pertinence de son entrée, limiter la redondance interne à la banque (et de ce fait une entrée SwissProt peut contenir plusieurs séquences, identiques ou « presque »); valider et enrichir l'information biologique associée à la séquence. En particulier, un grand soin est apporté à indiquer la nature, expérimentale ou bioinformatique, des informations fonctionnelles, ainsi que le niveau de confiance qui leur est accordée.

Pour résoudre l'objectif d'automatisation de Spot Manager, il fallait trouver un moyen pour obtenir les informations de SwissProt, quand elles existent, à partir d'un numéro d'accession de nrprot.

La banque **PIR-NREF**, maintenue par PIR (Protein Information Resource) pour la National Biomedical Research Foundation de l'université Georgetown de Washington, est une banque, à la manière de nrprot, assez peu annotée mais exhaustive. Celle-ci contient de manière non redondante les séquences de protéines de nombreuses autres banques. De plus, PIR établit une interconnexion entre ces banques. La correspondance SwissProt/TrEMBL et nrprot est alors apparu possible grâce à cet outil.

Sur le serveur ExPASy (Expert Protein Analysis System) du SIB dédié à l'analyse des proteines, il existe une page (**WORLD-2DPAGE**) qui contient les références des bases de données de gels 2D. Aussi, il propose un outil de recherche de proteines (2D-HUNT) parmi ces pages. A terme, le SIB veut fédérer toutes les ressources de pages 2D [8]. Le site WEB réalisé dans ce stage se devait de respecter les critères pour être fédéré dans la WORLD-2DPAGE.

2.3 Les choix technologiques informatiques

2.3.1 Le langage de programmation

Un bon choix du langage de programmation est nécessaire pour un développement efficace. J'ai choisi d'utiliser la technologie de SUN MicroSystem, le langage **JAVA**. Ce choix n'a pas été fait au hasard ; ce langage a des spécifications bien particulières et intéressantes dans ce contexte.

Tout d'abord, l'avantage le plus important de Java est probablement sa **portabilité**, puisqu'il peut tourner sur n'importe quelle machine disposant d'un interpréteur Java. C'est un langage indépendant du système d'exploitation. Le fer de lance de Java est "Write once, run everywhere", ou "Ecrire une fois, utiliser partout". Spot Manager bien que développé dans un environnement Windows 2000 peut donc fonctionner sans aucune modification sous Linux.

Ensuite, Java est un langage **orienté objet**, c'est à dire qu'on ne va pas manipuler des fonctions et des procédures mais des objets qui vont s'échanger des messages. Le principal avantage est que l'on peut réaliser une programmation modulaire : tous les objets peuvent être mis au point séparément. Cela permet aussi une réutilisation des objets dans d'autres projets, d'autres contextes. A titre indicatif, Spot Manager comporte une cinquantaine d'objets.

Un autre atout de Java est la possibilité de faire des **interfaces graphiques** agréables (JAVA-SWING). Spot Manager s'adresse à un public non-informaticien. Par conséquent, cette fonctionnalité était indispensable.

Le dernier argument en faveur de mon choix est l'existence de serveurs de pages WEB dynamiques[1] basés sur le langage Java (le serveur Apache Tomcat par exemple). Ceux-ci s'appuient sur les **JSP** (Java Server Page) de SUN. La technologie des JSP est un des moyens les plus performants pour générer des pages WEB dynamiques et reprend tous les avantages du JAVA. Ainsi, pouvoir réutiliser les objets créés pour Spot Manager dans le site WEB de visualisation de gel a été un gain de temps indéniable.

1 Page WEB dynamique : page WEB qui s'adapte en fonction des requêtes de l'utilisateur.
Exemple : une page qui affiche un gel est la même pour tous les gels (présentation, boutons) mais s'adapte au gel sélectionné par l'utilisateur (affichage de l'image appropriée).

2.3.2 Le Système de Gestion de Base de Données

Le site WEB nécessite une base de données pour stocker toutes les informations relatives aux gels. Il fallait donc choisir un SGBD (Système de Gestion de Base de Données). J'ai choisi d'utiliser **MySQL**.

MySQL est le serveur de base de données libre de droits le plus utilisé dans le monde. Son architecture logicielle le rend extrêmement rapide et facile à personnaliser. Ses principaux avantages sont la rapidité, la robustesse et la facilité d'utilisation.

De plus, les possibilités d'hébergement du site WEB justifient aussi ce choix. Si il est hébergé sur le serveur de la Génopole de Lille, MySQL est déjà en place. Si il est hébergé ailleurs, le déploiement de MySQL est bien plus facile (sous Windows comme sous Linux) que ses concurrents. En particulier, on peut penser à son concurrent direct PostgreSQL plus fonctionnel mais non disponible en natif pour Windows; il utilise une couche d'émulation.

Néanmoins, MySQL n'est pas exempt de défauts. Certaines fonctionnalités manquantes entraînent une surcharge de travail au niveau de la programmation. Malgré tout, choisir MySQL assure un **déploiement rapide** du site, et cela me semblait primordial et prioritaire.

2.3.3 Le format de fichier XML

Le **XML** (eXtensible Markup Language) a été mis au point et validé par le consortium W3C (World Wide Web Consortium). XML est un ensemble de règles, de lignes directrices, de conventions pour la conception de formats de fichier texte qui permettent de **structurer les données**. Il facilite la réalisation de fichiers qui ne soient pas ambigus, et qui évitent les pièges courants, tels que la non-extensibilité, l'absence de prise en charge de l'internationalisation et la dépendance par rapport aux plateformes.

Il est pratiquement impossible d'effectuer des recherches dans différentes bases de données incompatibles. Le XML, en revanche, permet de combiner facilement des données structurées qui proviennent de différentes sources. Cette technologie est devenue la solution simple pour **stocker** ou **échanger** des informations entre diverses applications.

Au cours de ce projet, l'usage du XML a été **intensif** et particulièrement **utile** pour :

- la sauvegarde d'un projet par Spot Manager.
- la sauvegarde des fichiers de configuration de Spot Manager.
- la communication d'informations entre Image Master 2D Platinum et Spot Manager.
- la mise à jour de la banque de protéines nécessaire à ASCQ-PROT.

3 Présentation de Spot Manager

3.1 L'architecture générale

On peut résumer l'architecture de Spot Manager et la mise en place de toutes les technologies présentées dans les parties précédentes par la figure suivante :

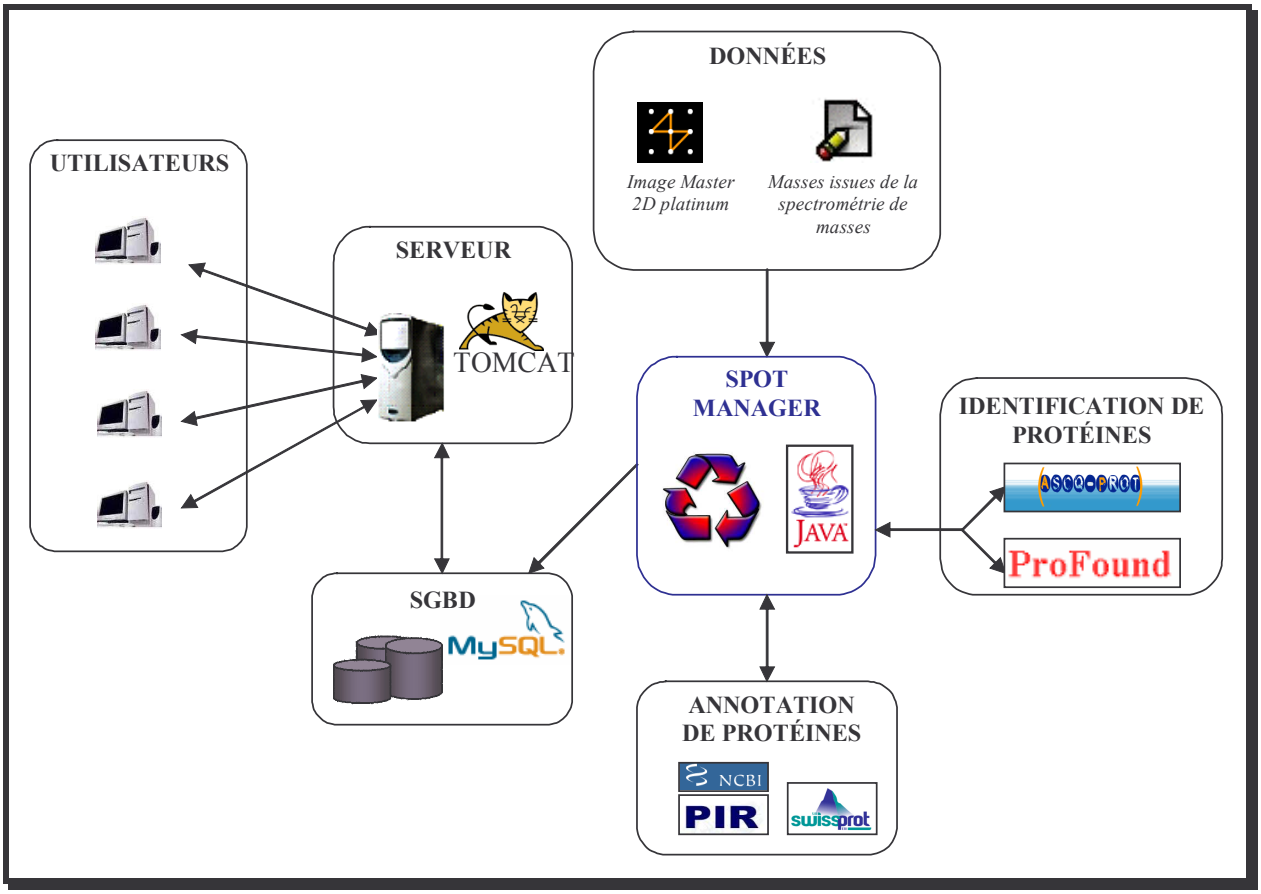

FIGURE 3.1 – *Architecture générale de Spot Manager*

L'utilisateur crée à partir des données d'image Master 2D Platinum (image et propriétés d'un gel), un nouveau projet sous Spot Manager.

Au fur et à mesure de l'avancement du projet, il peut rajouter des spots par l'intermédiaire de fichiers de masses issues de la spectrométrie. Spot Manager tentera d'identifier la protéine associée et de l'annoter automatiquement. Il propose aussi des outils d'aide à l'édition et à la gestion de ces protéines.

A tout moment, l'utilisateur peut, s'il le souhaite, insérer son projet dans la base de données. Le serveur de pages dynamiques se connecte au SGBD pour créer les pages suivant les requêtes des utilisateurs clients connectés. Ainsi, le site WEB est le reflet de plusieurs projets et est mis à jour automatiquement.

3.2 Vue d'ensemble de Spot Manager

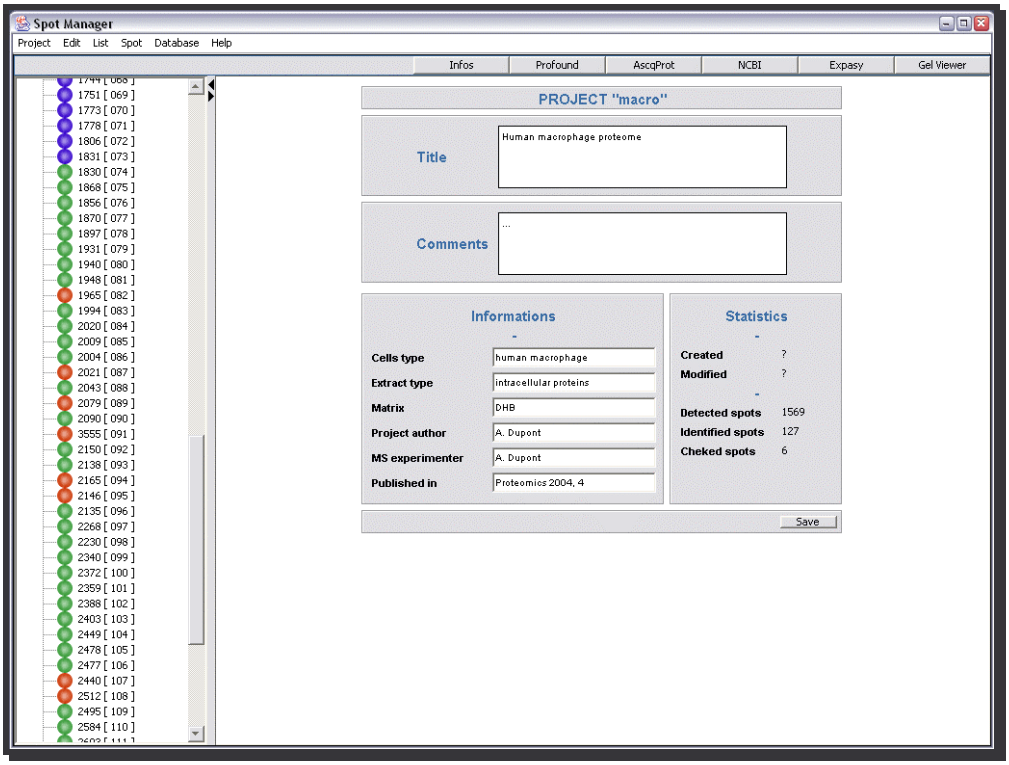

FIGURE 3.2 – *Interface de Spot Manager*

Un des buts de Spot Manager est la fédération de données. C'est pourquoi, son interface utilisateur se veut **agréable**, **complète** et **pratique**. Toutes les informations sont accessibles via la même interface. De plus, la plupart des fonctions du logiciel ont un raccourci clavier.

La figure 3.2 montre un aperçu global de cette interface. Outre le menu classique à toutes les applications fenêtrées, la fenêtre principale est composée de trois éléments principaux:

- Le **panneau de droite** : c'est la partie de l'interface qui sert à afficher des données. (ici, par exemple, on aperçoit les propriétés du projet en cours). La lecture et la modification des données se font dans cette partie de la fenêtre.
- La **partie de gauche** : elle contient la liste des **spots identifiés** ou en cours d'identification du projet chargé dans Spot Manager. Chaque petit cercle de couleur correspond à un spot. Un code de couleur a été mis en place : Vert, l'identification du spot est probable ; rouge peu probable et bleu vérifiée par l'utilisateur. Un simple coup d'œil permet de voir la progression du projet. Cette liste peut être triée selon un critère particulier (numéro de spot, Mr, pI…).
- La **barre de boutons en haut** : elle permet d'avoir accès, pour chaque spot, aux données qui lui sont associées : les annotations de la protéine associée (voir Figure 3.3), les pages de résultats de Profound et d'ASCQ-PROT, les informations des sites WEB correspondants (NCBI et SwissProt), ainsi que sa position sur le gel (voir FIGURE 3.4). Une action sur un de ces boutons met à jour la partie de droite. L'accès aux différentes informations est ainsi très rapide.

Il existe bien d'autres « petites » fonctionnalités dans Spot Manager…. Parmi elles, on trouve notamment : la copie dans le clipboard (presse-papier) des masses associées à un spot, l'export de la liste de spot et de leurs caractéristiques dans un tableau Excel, l'insertion d'un projet dans la base de données, la mise à jour de la banque pour ASCQ-PROT…

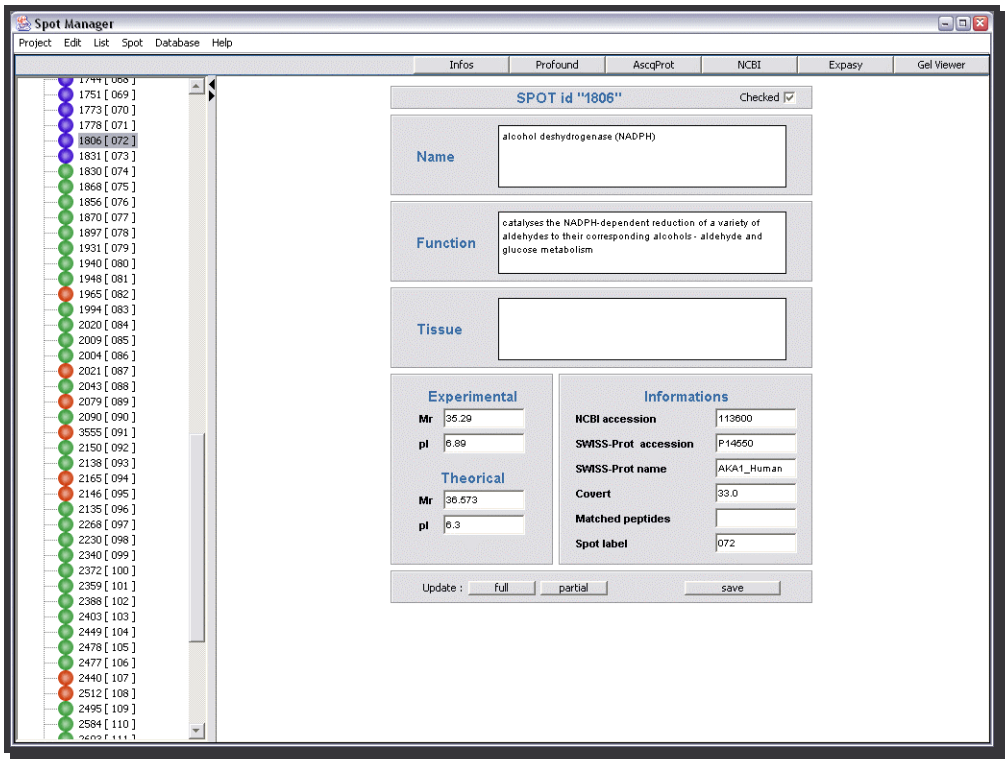

FIGURE 3.3 – *Interface de visualisation et d'édition de spot sous Spot Manager*

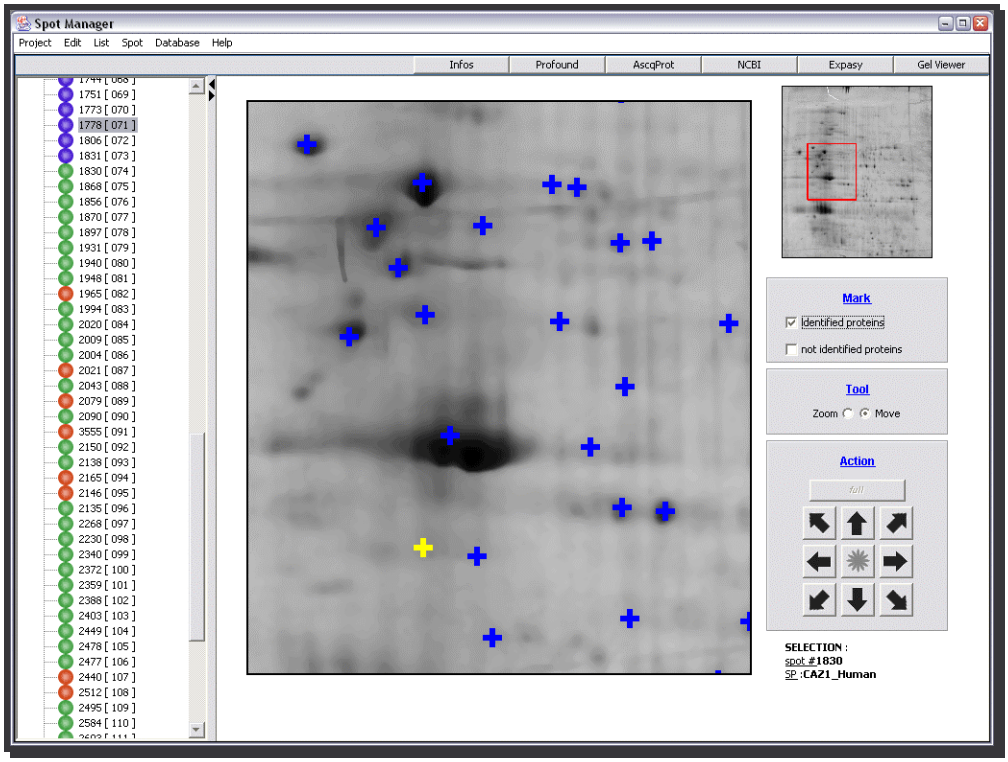

FIGURE 3.4 – *Visualisation de gel sous Spot Manager*

3.3 L'insertion de nouveaux spots

L'utilisateur peut rajouter des spots et cela de deux façons différentes : un par un ou **plusieurs d'un seul coup**. Spot Manager peut soit ajouter un spot grâce à un fichier, soit grâce à un répertoire contenant plusieurs fichiers (voir FIGURE 3.5).

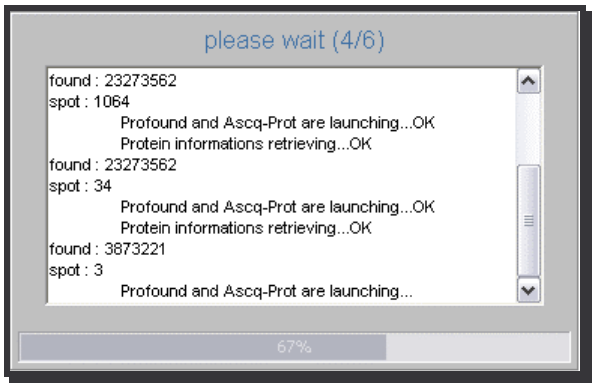

FIGURE 3.5 – *Fenêtre sous Spot Manager qui indique l'avancement de l'identification de plusieurs fichiers de masses*

Néanmoins, pour faire le lien entre les données et les fichiers de masse, il faut instaurer un certain **protocole**. Ces fichiers doivent impérativement contenir dans leur nom, l'identifiant de spot qu'Image Master 2D Platinum lui a attribué lors de la détection de spot. C'est la seule contrainte que j'ai dû imposer.

Une fois ce prérequis, Spot Manager tente d'identifier les spots automatiquement. Au départ ce processus était loin d'être satisfaisant puisqu'il ne trouvait que des kératines humaines. Ces protéines sont ce qu'on appelle des **contaminants**. Au cours des différentes manipulations pour l'élaboration du gel, les expérimentateurs peuvent contaminer les résultats par leurs cheveux, par des bouts de peau. Pour remédier à ce problème, une première idée a été trouvée : supprimer les pics de masses qui correspondent aux kératines. Cette méthode m'a été déconseillée par un massiste que j'ai rencontré à l'Institut Biologique de Lille. Les mêmes pics se retrouvent dans d'autres protéines (risques de mauvais résultats par la suite). Au final, j'utilise une autre solution plus radicale : les kératines humaines sont **supprimées** de ma banques de protéines…

De la même manière que quand on le réalise à la main, le résultat n'est pas pour autant toujours parfait. Mais, le code de couleurs instauré dans l'interface permet de localiser très rapidement les erreurs. De plus, tous les fichiers de résultats sont sauvegardés… En cas de doute, l'utilisateur peut parcourir et modifier les informations aisément. A titre indicatif, un biologiste met environ deux semaines pour une cinquantaine de spots à identifier et à annoter. Spot Manager met environ **quarante secondes par spot**. Les différents tests du logiciel ont montré qu'environ **70% des spots étaient correctement identifiés**… et 30% d'entre eux complètement annotés (en fait, ce dernier chiffre dépend uniquement de la présence ou non de la protéine dans la banque Swiss-Prot).

3.4 Le site WEB de visualisation de gel

Le site WEB a été une autre grande partie du travail réalisé. Celui-ci se veut **simple** et **ergonomique**. Il n'utilise qu'une seule fenêtre sans frame (fenêtre dans la fenêtre). Une certaine charte graphique s'est imposée : classique, sobre et aux couleurs de l'Institut Pasteur. A tout moment, grâce à un lien sur chaque page, on peut retourner à la page d'accueil du site (voir FIGURE 3.6) qui propose la liste des fonctionnalités accessibles à l'utilisateur. Parmi elles, on trouve :

- La **recherche** d'une ou d'une liste de protéine(s) dans la base de données grâce :
  - à un identifiant Swiss-PROT
  - à un ou plusieurs mots
  - à une gamme de pI et de Mr
  - à une localisation sur le gel
- L'**affichage** d'éléments de la base de données :
  - la liste des protéines contenues dans un gel particulier
  - la liste des projets contenus dans la base de données
  - les informations relatives à un projet.

Toutes les pages proposent si possible des **liens entre elles**. Par exemple, la recherche d'une protéine grâce à un mot mène à une page qui affiche une liste de protéines correspondantes. Un lien sur chaque protéine permet l'affichage de la page qui affiche les informations de l'entrée correspondante (voir FIGURE 3.7). Parmi ces informations, on trouve différents liens vers les pages des banques publiques (SwissProt, NCBI) ainsi qu'un lien vers sa localisation sur le gel (voir FIGURE 3.8).

La plupart des pages sont des JSP. Pour autant, celles-ci contiennent vraiment **très peu de code JAVA**. Elles se contentent d'utiliser des objets JAVA. De plus, ces objets utilisés sont ceux de Spot Manager…

Les images de gels sont **générées** en fonctions des requêtes des utilisateurs (affichage des spots identifiés et/ou des spots non identifiés, zoom, et emplacement sur le gel). Un système de tampon a été mis en place : amélioration du temps de réponse du serveur et prévention de la saturation de l'espace disque alloué.

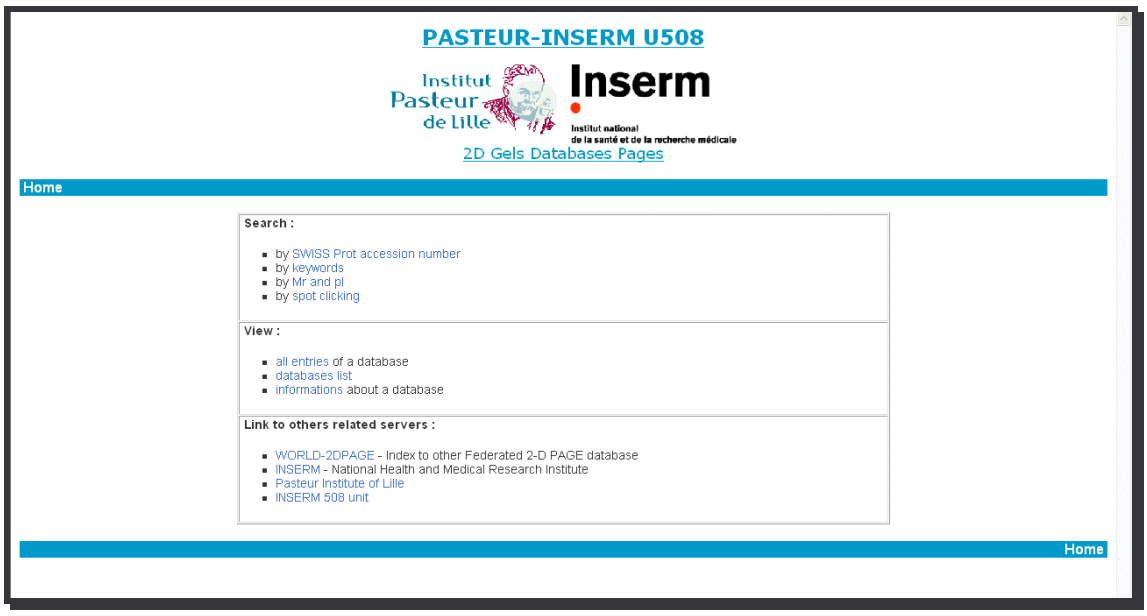

FIGURE 3.6 – *Page WEB d'accueil du site de gels 2D*

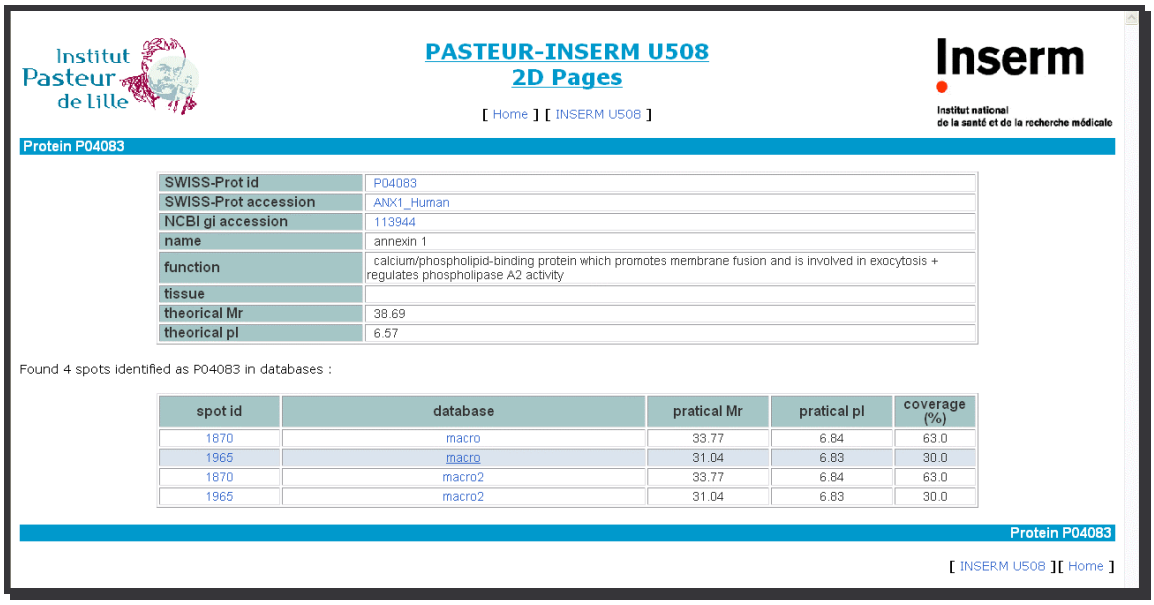

FIGURE 3.7 – *Page WEB d'informations relatives à une protéine*

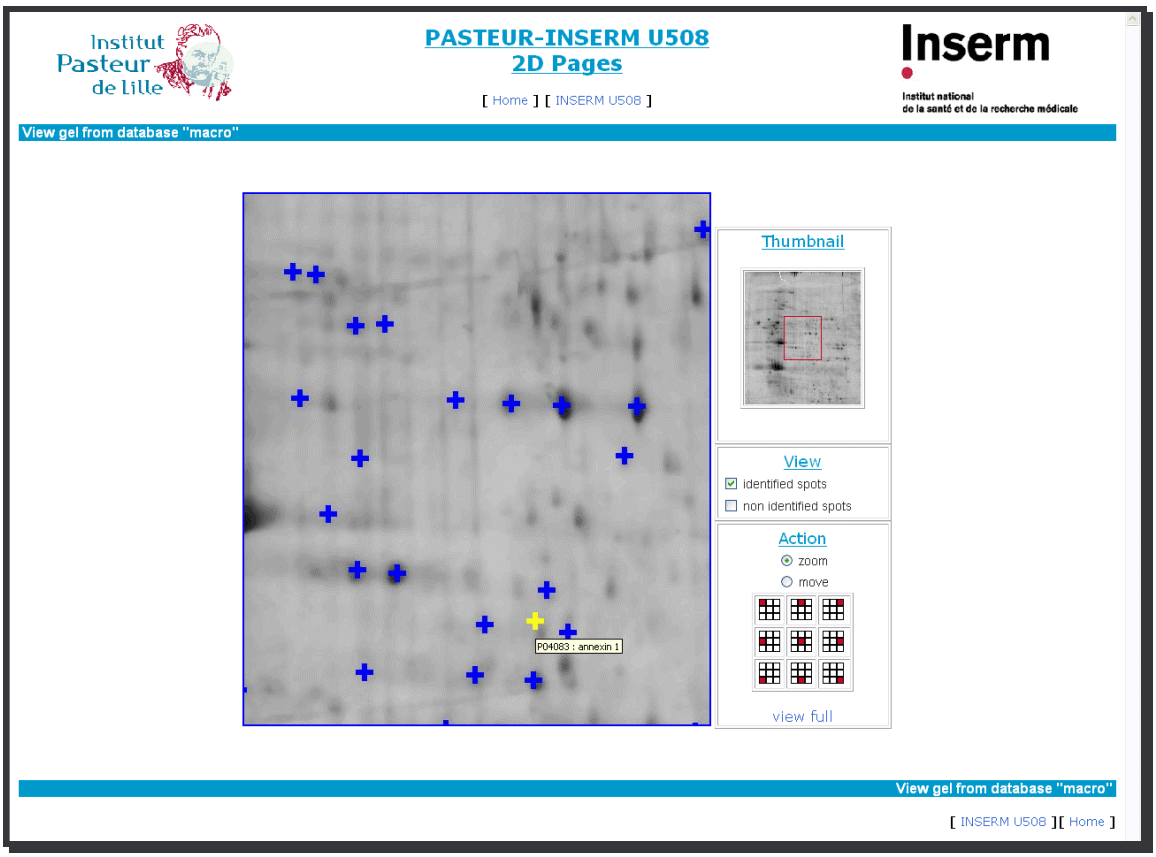

FIGURE 3.8 – *Page WEB de visualisation de gel*

Conclusion et perspectives

La vision d'un informaticien dans un laboratoire telle que l'unité INSERM 508 apporte indéniablement une valeur ajoutée. Conscient de cela, le laboratoire m'a laissé une grande autonomie et une grande liberté d'action. Cette confiance tout au long du stage a été très motivante. Ainsi, en plus de la demande initiale du site WEB, je pense avoir réussi à cerner et analyser des besoins et à y avoir répondu.

Du point de vue technique, ce stage a été la mise en application d'un panel très large de mes compétences: analyse, conception, programmation, base de données, site WEB et notions de protéomique. En plus, il m'a permis de renforcer nettement les acquis de mon stage de maîtrise et m'a donné l'occasion de mettre en pratique, dans des conditions réelles, mon logiciel ASCQ-PROT.

Les perspectives de mon travail sont nombreuses. Tout d'abord, le site WEB doit être installé sur un serveur dédié, mis en ligne et rendu disponible à la communauté. Ensuite, Spot Manager doit être soumis et testé par ses utilisateurs. Aussi, même si pour l'instant, il est réservé à l'usage propre du laboratoire, une grande amélioration serait de le rendre plus polyvalent, c'est-à-dire de le rendre beaucoup plus paramétrable. Exemple : pour l'instant, il est exclusivement réservé à l'espèce humaine, trop spécifique. Une autre perspective de travail serait d'améliorer l'identification de protéine en ajoutant des fonctionnalités à ASCQ-PROT.

Et finalement, ces perspectives, je pourrai les explorer moi-même puisque mon stage se prolonge par un CDD de 3 mois au sein de l'unité U508 de l'INSERM.

Bibliographie

1: Dupont A, Tokarski C, Dekeyzer O, Guihot AL, Amouyel P, Rolando C, Pinet F. Two-dimensional maps and databases of the human macrophage proteome and secretome. Proteomics. 2004 Jun;4(6):1761-78.

2: Chambers G, Lawrie L, Cash P, Murray GI. Proteomics: a new approach to the study of disease.J Pathol. 2000 Nov;192(3):280-8. Review.

3: Choe LH, Lee KH. Quantitative and qualitative measure of intralaboratory two-dimensional protein gel reproducibility and the effects of sample preparation, sample load, and image analysis.Electrophoresis. 2003 Oct;24(19-20):3500-7.

4: Appel RD, Hochstrasser DF, Funk M, Vargas JR, Pellegrini C, Muller AF, Scherrer JR. The MELANIE project: from a biopsy to automatic protein map interpretation by computer.Electrophoresis. 1991 Oct;12(10):722-35.

5: Chakravarti DN, Chakravarti B, Moutsatsos I. Informatic tools for proteome profiling.Biotechniques. 2002 Mar;Suppl:4-10, 12-5. Review.

6: Chamrad DC, Korting G, Stuhler K, Meyer HE, Klose J, Bluggel M. Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data. Proteomics. 2004 Mar;4(3):619-28.

7: Zhang W, Chait BT. ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. Anal Chem. 2000 Jun 1;72(11):2482-9.

8: Appel RD, Bairoch A, Sanchez JC, Vargas JR, Golaz O, Pasquali C, Hochstrasser DF. Federated two-dimensional electrophoresis database: a simple means of publishing two-dimensional electrophoresis data.Electrophoresis. 1996 Mar;17(3):540-6.