CONSERVATOIRE NATIONAL DES ARTS ET MÉTIERS

Centre Régional Languedoc-Roussillon

Spécialité : INFORMATIQUE

MÉMOIRE

ANNEXES

Création d'une application intégrée pour la gestion et l'analyse de données protéomiques

---

Soutenu le 5 juillet 2005

par

Cédric BOUTTES

JURY

Président : Jean-Yves RANCHIN, Professeur, CNAM
Membres : Isabelle MOUGENOT, Maître de conférence, Université Montpellier II
Marc NANARD, Professeur, CNAM
Michel ROSSIGNOL, Directeur de Recherche, INRA
Thierry HOTELIER, Ingénieur, ENSA Montpellier

LISTE DES ANNEXES

ANNEXE 1 : Analyse fonctionnelle par cas d'utilisation

SOMMAIRE

1 Analyse du cas d'utilisation *« Saisie des données »*

a) Spécification des besoins par écrit

D'après le cahier des charges, la saisie des données doit se faire dans un classeur Excel nommé aussi format d'échange, les données saisies étant ensuite importées dans la base à l'aide d'un programme informatique.

Il faut dès lors tenir compte du fait qu'une procédure de saisie doit être organisée au sein des versions locales de ProteomIs installées dans les différents laboratoires intéressés et également au sein du module GnpProt installé sur le serveur de Génoplante Info.

C'est une partie des données saisies dans les versions locales qui viendront alimenter la base intégrative du module GnpProt mise à disposition sur le serveur de Génoplante. Pour les utilisateurs ne disposant pas de version locale, les données seront envoyées directement à Génoplante par l'intermédiaire du format d'échange.

b) Cas d'utilisation détaillée

Sur le **diagramme 1** page suivante est présenté le diagramme de cas d'utilisation *« Mise à jour des données »*.

Sur ce diagramme nous avons deux packages.

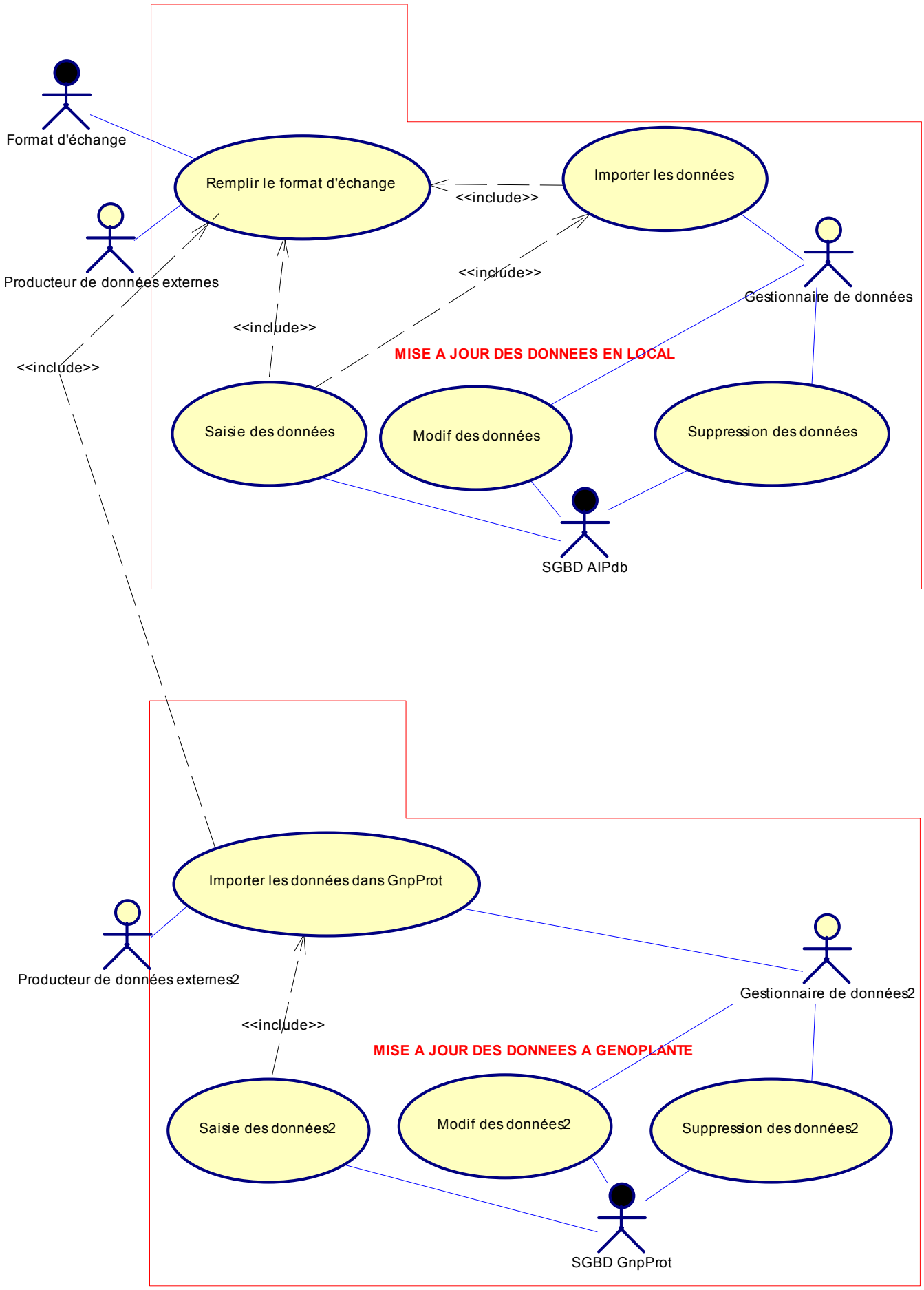
Le premier package intitulé *« Remplir le format d'échange »* présente essentiellement le cas d'utilisation *« Remplir le format d'échange »*.

Ce cas d'utilisation est relié de manière optionnelle par le cas d'utilisation *« Importer les données dans la base locale »*.

En effet ce dernier cas d'utilisation est optionnel puisque tous les utilisateurs ne disposent pas d'une base locale. Le format d'échange peut simplement servir à alimenter la base intégrée GnpProt installée sur le serveur de Génoplante Info.

Nous avons ensuite trois cas d'utilisations : *« Saisie des données »*, *« Modification des données »* et *« Suppression des données »*. La saisie de nouvelles données passe par l'importation des données contenues dans le format d'échange. L'importation des données est activée par le gestionnaire des données ; le producteur de données ne faisant que remplir le format d'échange.

Diagramme 1 : Diagramme de cas d'utilisation : Mise à jour des données



c) Aspects dynamiques *(faisant intervenir les acteurs)*

Tout comme l'action d'importer de nouvelles données, l'action de modifier et supprimer des données est effectuée également par le gestionnaire des données sur simple demande du producteur de données. Le gestionnaire des données effectuera alors ces modifications et suppressions de données de manière manuelle à l'aide du SGBD d'après les informations qui lui auront été parvenues dans la demande du producteur de données.
Nous allons maintenant illustrer les aspects dynamiques des cas d'utilisation en utilisant les diagrammes de séquences de la notation UML. Ces digrammes de séquences vont montrer le déroulement temporel d'une interaction entre les acteurs identifiés dans les cas d'utilisation, ceci dans le but de réaliser une fonction du système.
Sur le **diagramme 2** est présenté le diagramme de séquence *« Saisie des données »*.
Sur le **diagramme 3** est présenté le diagramme de séquence *« Modification des données »*.
Enfin sur le **diagramme 4** est présenté le diagramme de séquence *« Suppression des données »*.
A ce stade de l'analyse nous considèrerons le programme informatique comme une boîte noire représentée par l'acteur *programme*. Le comportement du programme est en fait décrit vu de l'extérieur, sans préjuger de *comment* il sera réalisé.

Diagramme 2 : Diagramme de séquence : Saisie des données

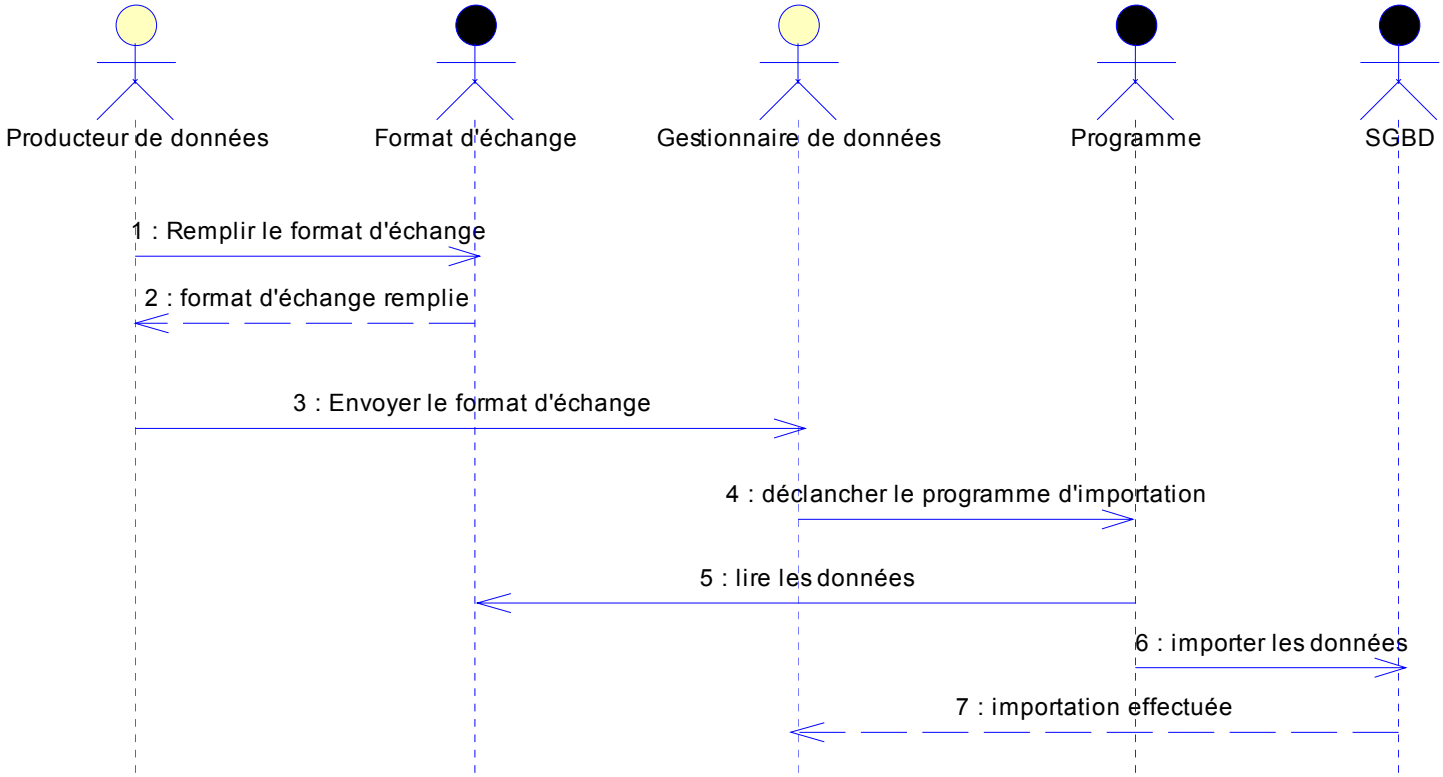


Diagramme 3: Diagramme de séquence : Modification des données

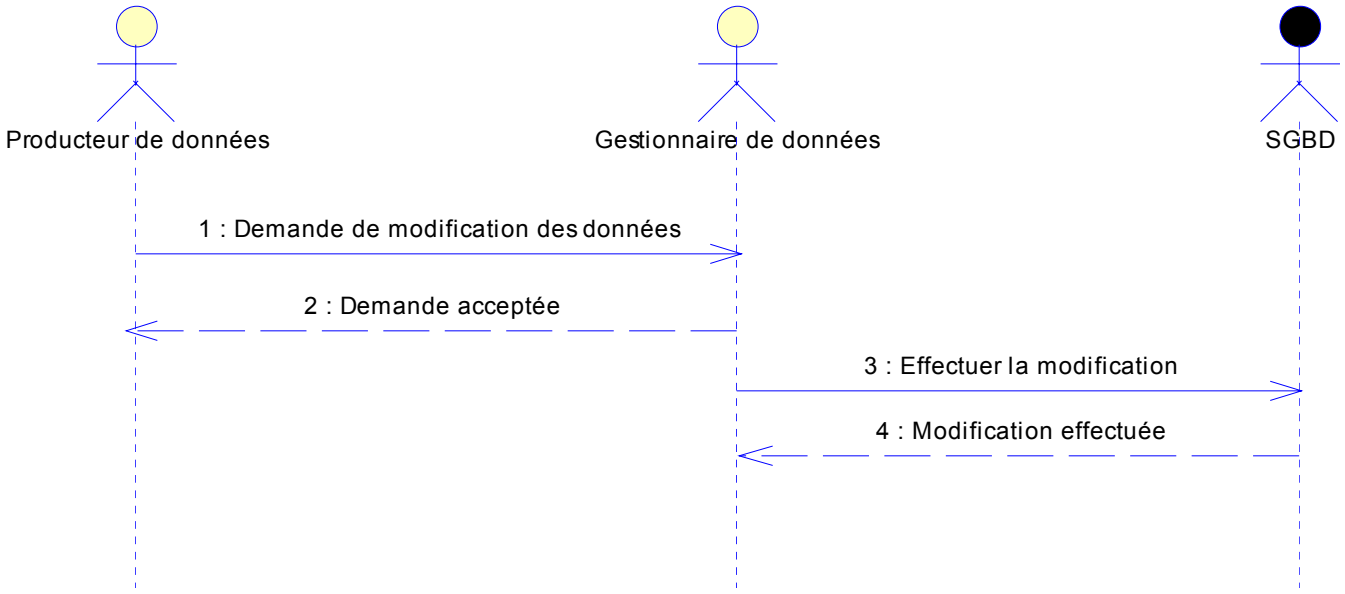
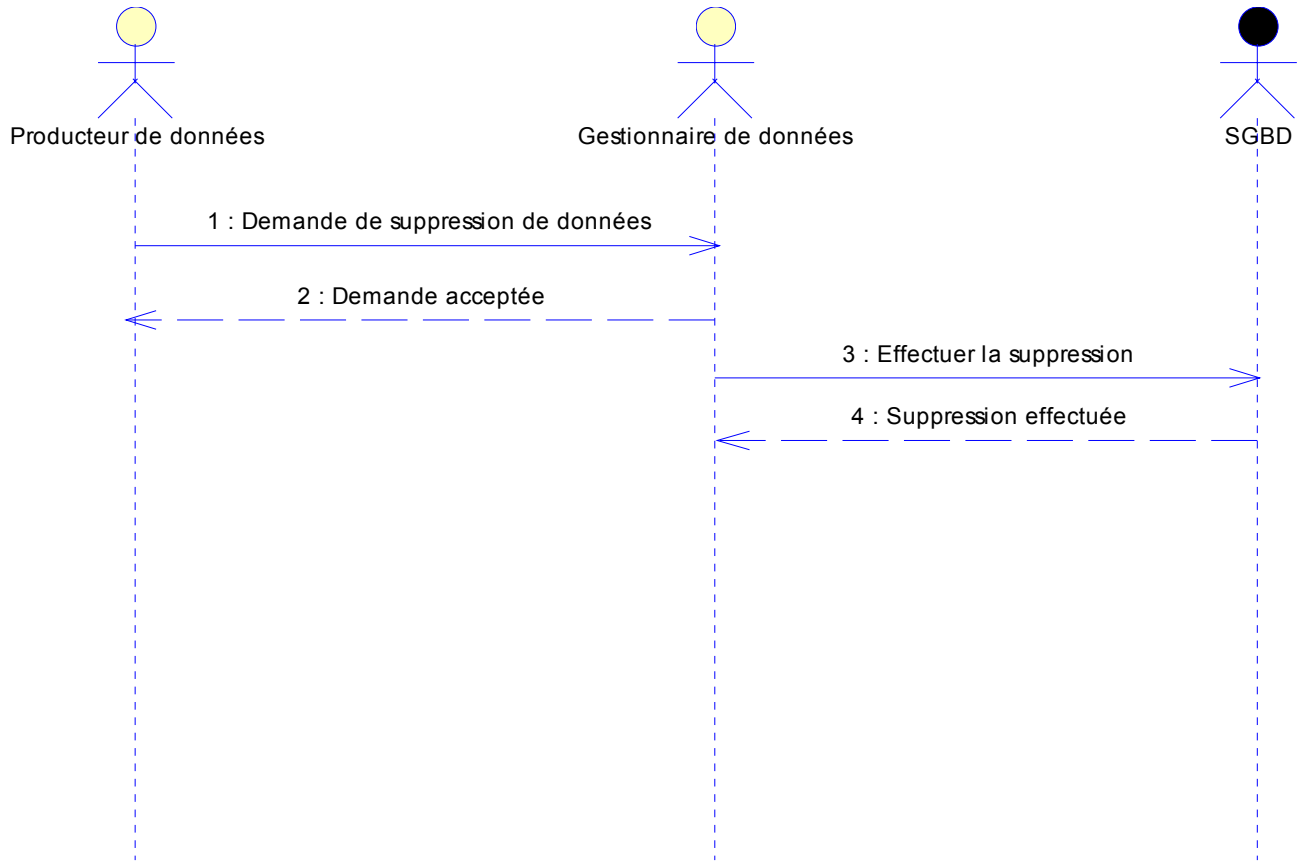


Diagramme 4 : Diagramme de séquence : Suppression des données



2 Analyse du cas d'utilisation *« Interrogation et visualisation des données »*

a) Spécification des besoins par écrit

D'après la spécification des besoins, nous avons vu qu'il était question de développer des interfaces d'interrogation et de visualisation des données contenues dans la base.
A partir de ces interfaces, des liens doivent permettre de retrouver les informations correspondantes sur les protéines dans les banques de données publiques et banques de données spécialisées sur Arabidopsis.

Enfin avec la version intégrée GnpProt, il doit être possible de rechercher des informations complémentaires dans le système d'information de Génoplante GpiIS.

Une option supplémentaire est la possibilité pour l'utilisateur de pouvoir exporter les données contenues dans certaines interfaces de visualisation. L'exportation doit pouvoir se faire à la fois au format CSV et dans Excel afin que les biologistes puissent facilement retravailler ces données.

b) Cas d'utilisation détaillée

Dans cette partie, nous avons détaillé le cas d'utilisation « Consulter les données » (**diagramme 5**).

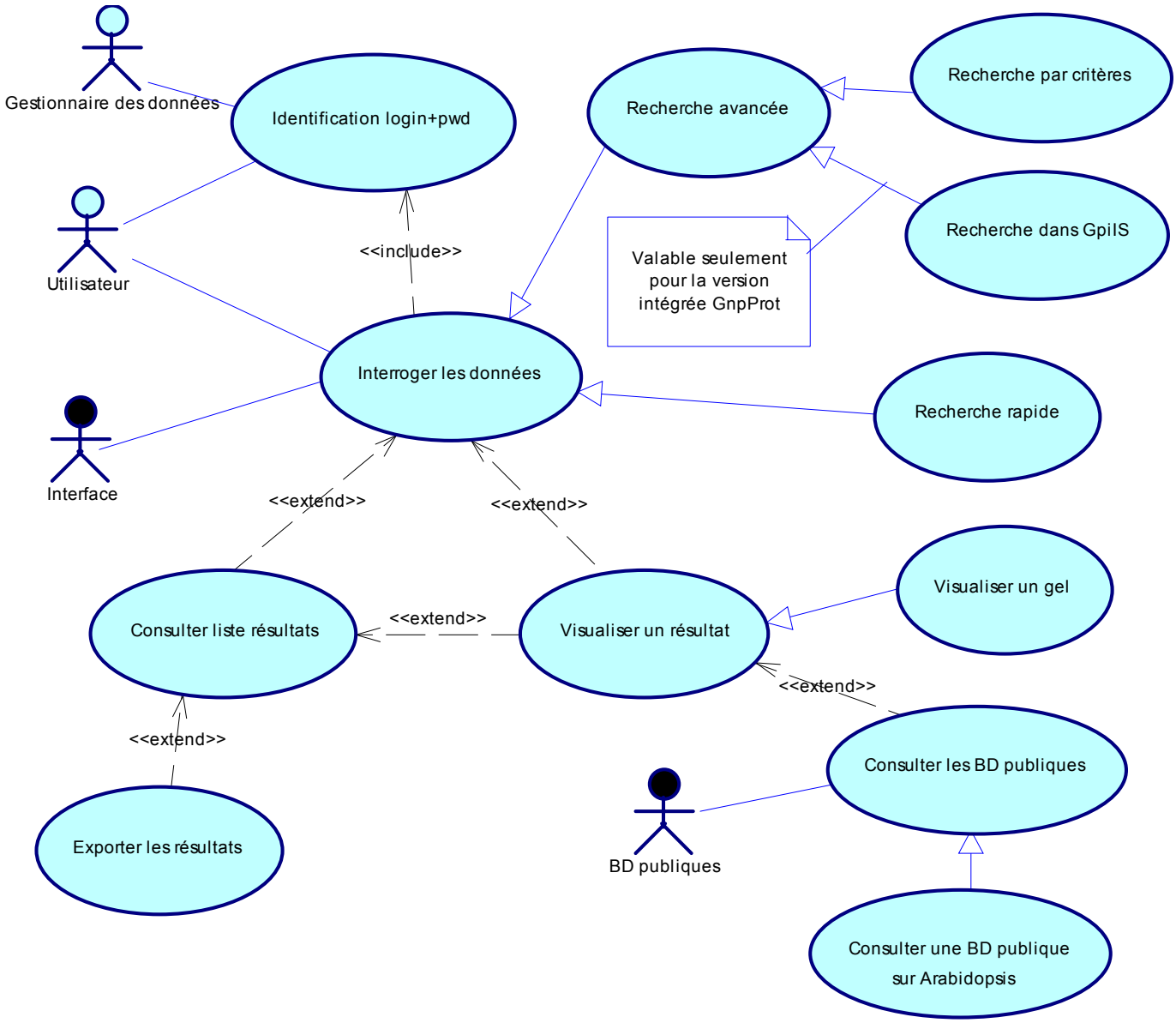


Diagramme 5 : Cas d'utilisation : Consulter les données

Nous avons ensuite jugé utile de détailler le cas d'utilisation « Visualiser un gel » (**diagramme 6**) qui est une spécialisation du cas d'utilisation « Visualiser un résultat ».

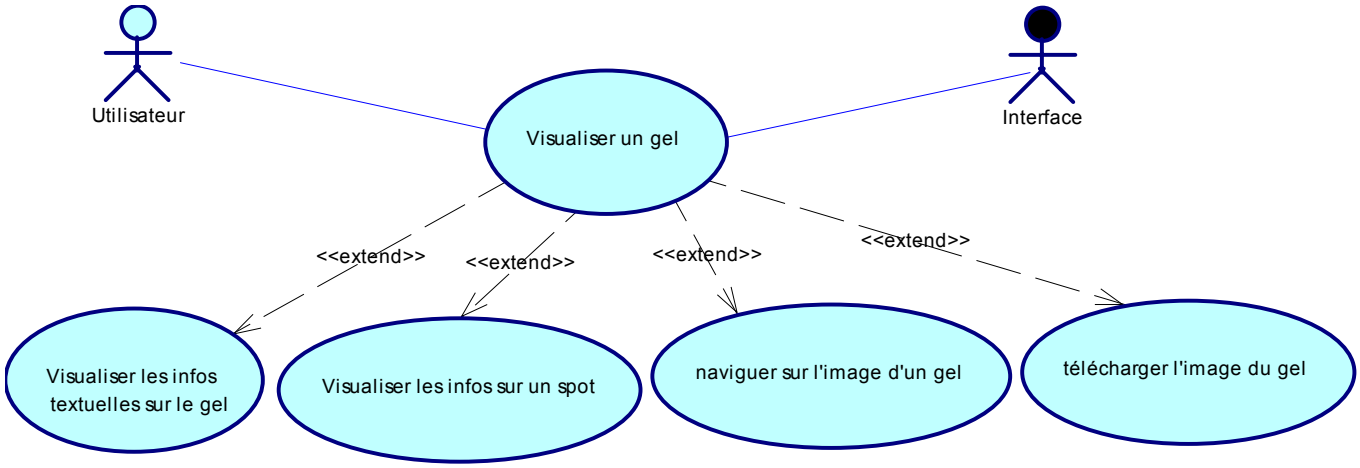
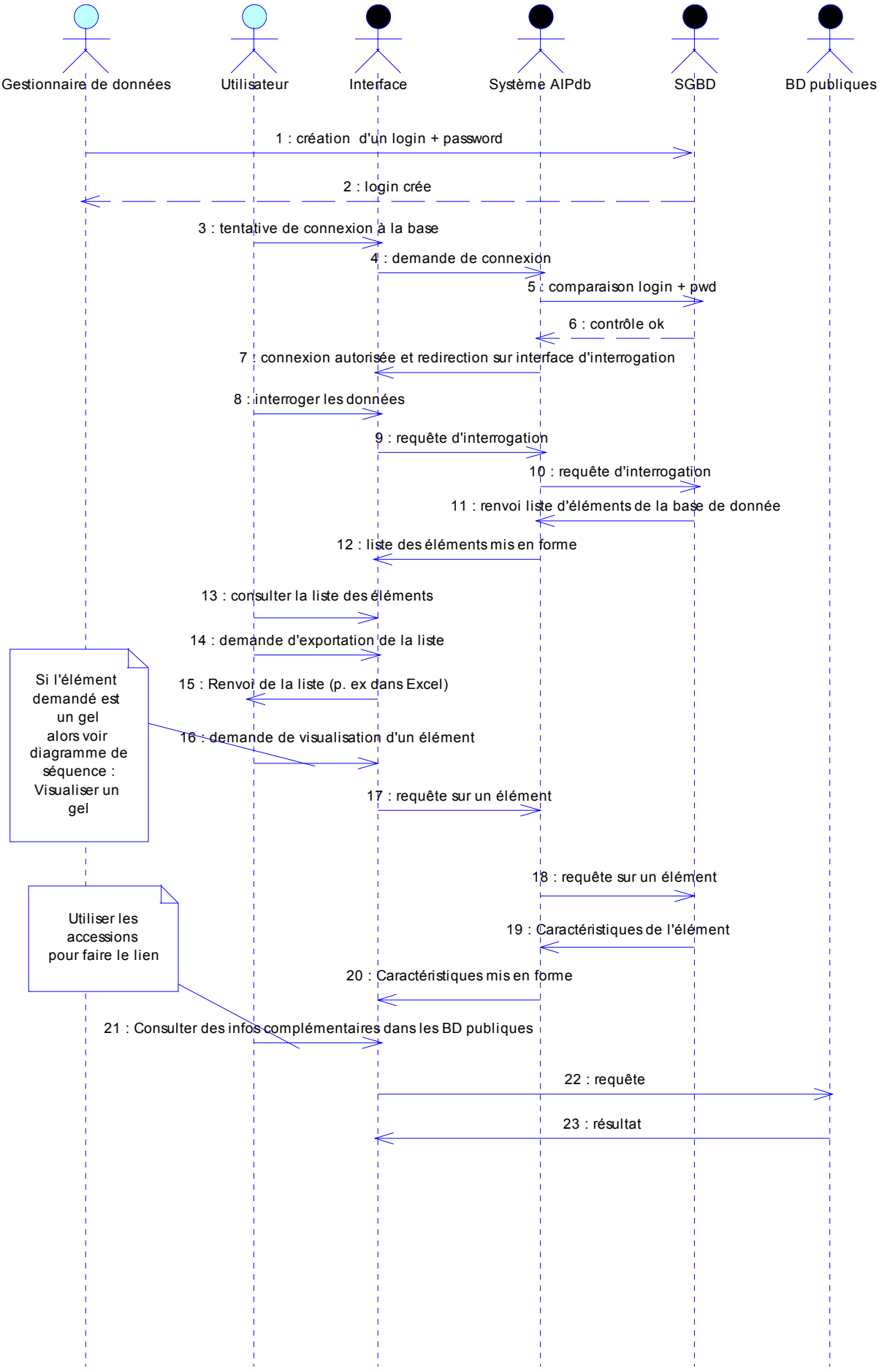


Diagramme 6 : Cas d'utilisation : Visualiser un gel

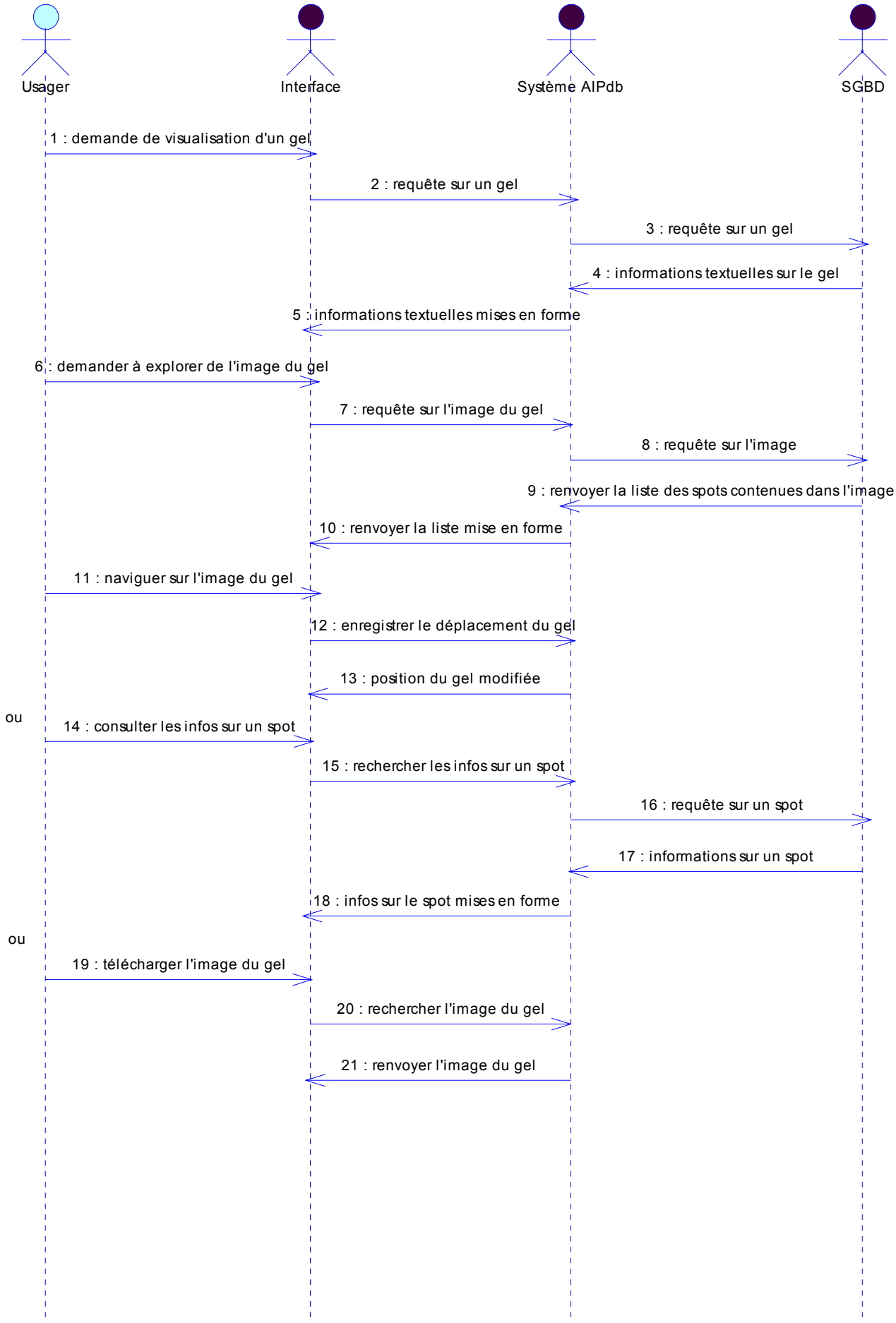
Ceci va nous aider à préciser les contours de l'interface de visualisation des gels qui est une des interfaces la plus importante et la plus complexe.
A ce stade il est déjà possible de préciser qu'il faudra pouvoir naviguer facilement sur le gel et pouvoir accéder facilement aux informations sur les spots identifiés sur le gel.

c) Aspects dynamiques *(faisant intervenir les acteurs)*

Les cas d'utilisation décrivent les interactions des acteurs avec le système ProteomIs que nous voulons spécifier et concevoir. Lors de ces interactions, les acteurs génèrent des messages qui affectent le système informatique et font appel généralement à une réponse de celui-ci. Nous allons isoler ces messages et les représenter graphiquement sur les diagrammes de séquences UML. Sur le **diagramme 7** (page suivante) est représenté le diagramme de séquence *« Consulter les données »*. Sur ce diagramme, est décrit de manière dynamique toute la procédure permettant d'interroger et visualiser les données de ProteomIs. La séquence de messages déclenchée par le message *Demande de visualisation d'un gel* fait l'objet du diagramme de séquence *« Visualiser un gel »* présenté dans le **diagramme 8** (page suivante).



<u>Diagramme 7 :</u> Diagramme de séquence « Consulter les données »



<u>Diagramme 8 :</u> Diagramme de séquence « Visualiser un gel »

3 Analyse du cas d'utilisation *« Analyse bioinformatique des données »*

Nous avons vu au début de l'analyse du projet ProteomIs/GnpProt que le cas d'utilisation *« Analyse bioinformatique des données »* constituait en fait un package contenant lui-même plusieurs cas d'utilisation (voir mémoire partie **6.2.2 Spécification des exigences d'après les cas d'utilisation**). Nous allons approfondir leur analyse dans chacune des sous partie suivante.

3.1 Création de groupes non redondants de protéines (« clustering »)

a) Spécification des besoins par écrit

La priorité au niveau de l'analyse bioinformatique des données est de créer des groupes non redondants de protéines (« clustering »). La solution retenue pour ce travail a été discutée partie **5.2.1** du mémoire. Elle consiste à effectuer un regroupement automatique des protéines similaires contenues dans la base de donnée ProteomIs/GnpProt en se basant sur les équivalences entre numéros d'accessions. Un identifiant unique devra être attribué à ces groupements.

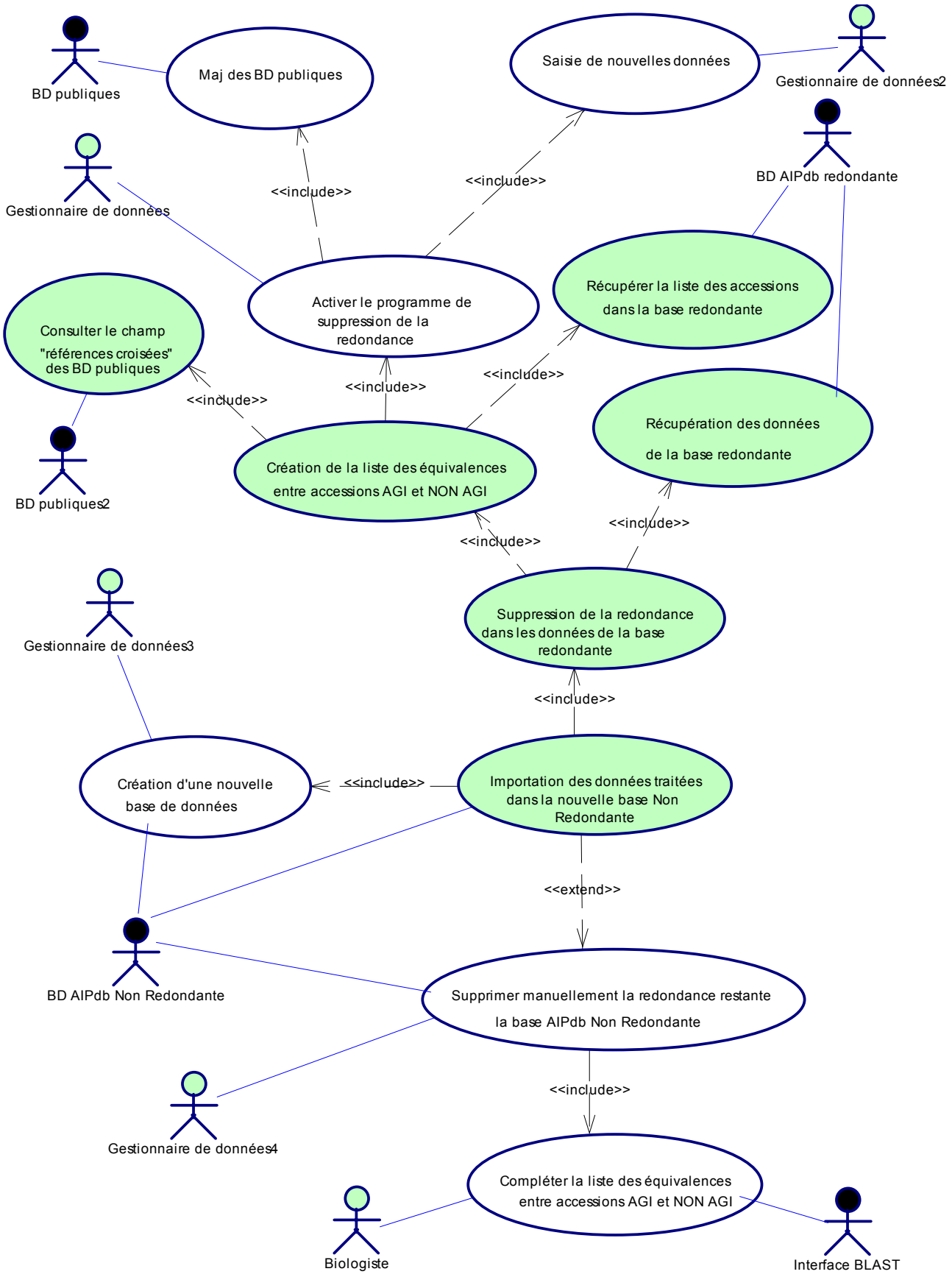
b) Cas d'utilisation détaillé

Le diagramme de cas d'utilisation : *« Suppression de la redondance dans la base ProteomIs (« clustering ») »* (**diagramme 10** page suivante) décrit toute la procédure permettant de faire migrer les données d'une version redondante de la base ProteomIs vers une version non redondante de cette base.

Les différentes phases de la procédure seront automatisées au travers d'un programme informatique. L'acteur programme et ses différentes relations ne sont pas représentés dans la cas d'utilisation pour des raisons de clarté. Cet acteur est cependant à l'origine de tous les cas d'utilisation colorés en vert.
Le gestionnaire de données est responsable de l'exécution de ce programme que l'on appelera programme de clustering.

Le gestionnaire de données devra également lancer l'exécution du programme après chacune des fois où la base ProteomIs aura été mis à jour avec les nouvelles données des producteurs de données importées à partir du format d'échange.

En fait lorsque la base passe dans l'état non redondant, elle devient à nouveau une base redondante à chacune des mises à jour par les données du format d'échange.

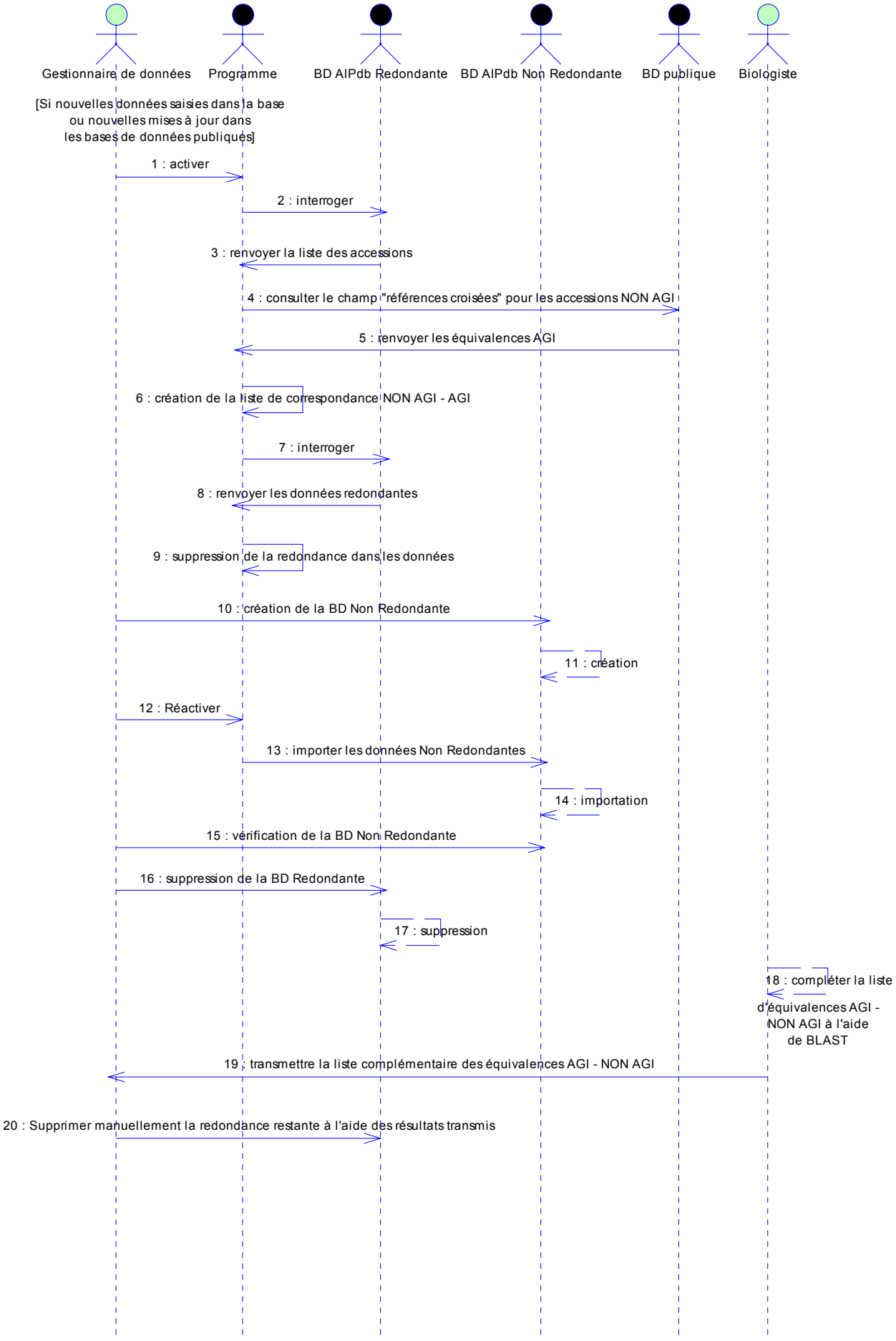


<u>Diagramme 10 :</u> Diagramme de cas d'utilisation : Suppression de la redondance dans la base ProteomIs (« clustering »)

c) Aspects dynamiques et diagramme d'état

Le diagramme de séquence *« Suppression de la redondance dans la base ProteomIs (« clustering ») »* (**diagramme 11**) représente la chronologie des différentes tâches décrites dans le diagramme de cas d'utilisation *« Suppression de la redondance dans la base ProteomIs (« clustering ») »*.

La périodicité pour lancer l'exécution du programme devra être synchronisée avec le rythme des mises à jour dans les bases de données publiques. En effet lors de ces mises à jour les champs « références croisées » sont enrichis par les gestionnaires de ces bases. Ceci permettrait éventuellement au programme de clustering de créer de nouvelles associations entre accessions NON AGI et AGI alors qu'elles n'avaient pas été trouvé en interrogeant la précédente version des bases de données publiques.



<u>Diagramme 11 :</u> Diagramme de séquence : Suppression de la redondance dans la base AIPdb (« clustering »)

Il faut noter qu'une nouvelle instance vide de la base ProteomIs est crée dans le SGBD afin d'accueillir les nouvelles données non redondantes. L'ancienne base ProteomIs redondante sera supprimée après que le gestionnaire de données ait vérifié que les données dans la base non redondante sont correctes. Ainsi même lors de la migration les utilisateurs disposeront toujours d'une base de donnée ProteomIs fonctionnelle. Ce changement d'état récurrent est décrit dans le diagramme d'états-transitions du **diagramme 12** ci-dessous.

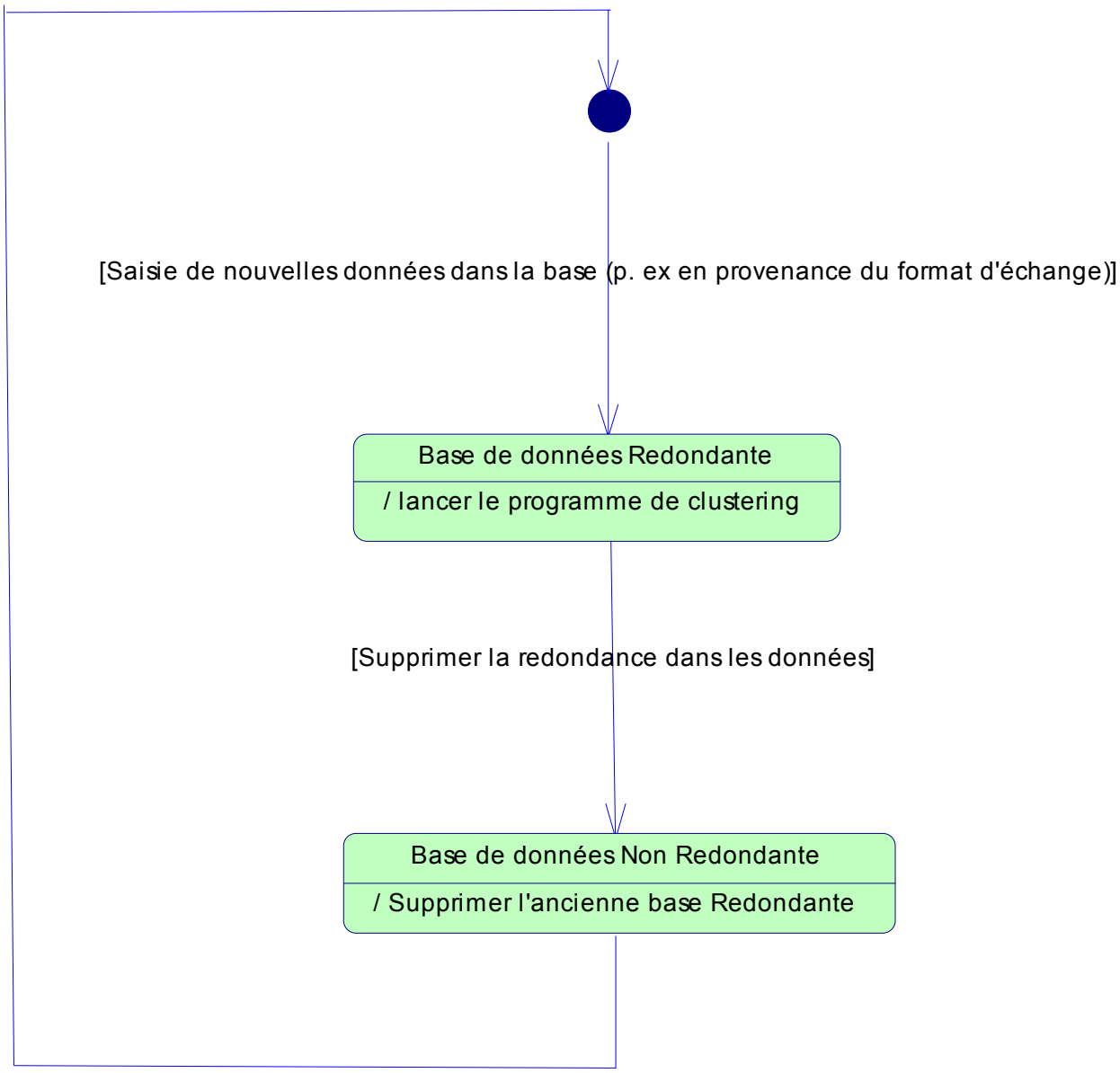


Diagramme 12 : Diagramme d'états : Suppression de la redondance dans la base ProteomIs (« clustering »)

3.2 Importation et comparaison de séquences

a) Spécification des besoins par écrit

Le programme de clustering décrit précédemment nous a permis de normaliser l'ensemble des numéros d'accession des protéines de ProteomIs au format AGI. Pour chacun de ces numéros d'accession , l'objectif va maintenant être d'importer la séquence correspondante au format FASTA dans la base de donnée ProteomIs. Les séquences ainsi rapatriées en local pourront être plus accessibles plus rapidement pour des traitements ultérieurs (comparaison de séquences, recherche de motifs).

Ces séquences peuvent être récupérées dans la banque de données TAIR via le site FTP : ftp://ftp.arabidopsis.org/home/tair/. Le format FASTA des séquences est un des plus répandus dans le monde de la bioinformatique. Il s'agit d'un format textuel très simple qui comprend une ligne d'en-tête identifiable par le symbole « > » qui précède les informations caractéristiques de la séquence. Sur cette ligne, se trouve notamment le numéro d'accession AGI de la séquence qui permet de l'identifier de manière unique. Ce numéros d'accession nous est extrêmement utile puisqu'il va notamment nous permettre d'établir des correspondances avec les séquences contenues dans ProteomIs.

Les lignes suivantes sont dévolues à l'enchaînement protéique de la séquence. Un exemple de séquence protéique de TAIR au format FASTA est donné dans le **document 1** ci-dessous.

```
>At1g01010.1 68414.m00001 no apical meristem (NAM) family protein contains
Pfam PF02365: No apical meristem (NAM) domain; similar to NAC domain protein
NAM GB: AAD17313 GI:4325282 from [Arabidopsis thaliana]
MEDQVGFGFRPNDEELVGHYLRNKIEGNTSRDVEVAISEVNICSYDPWNLRFQSKYKSRD
AMWYFFSRRENNKGNRQSRTTVSGKWKLTGESVEVKDQWGFCSEGFRGKIGHKRVLVFLD
GRYPDKTKSDWVIHEFHYDLLPEHQRTYVICRLEYKGDDADILSAYAIDPTPAFVPNMTS
SAGSVVNQSRQRNSGSYNTYSEYDSANHGQQFNENSNIMQQQPLQGSFNPLLEYDFANHG
GQWLSDYIDLQQQVPYLAPYENESEMIWKHVIEENFEFLVDERTSMQQHYSDHRPKKPVS
GVLPDDSSDTETGSMIFEDTSSSTDSVGSSDEPGHTRIDDIPSLNIIEPLHNYKAQEQPK
QQSKEKVISSQKSECEWKMAEDSIKIPPSTNTVKQSWIVLENAQWNYLKNMIIGVLLFIS
VISWIILVG*
```

document 1 : Une séquence de TAIR au format FASTA

Ce sont ces séquences qui seront importées dans la base ProteomIs pour mettre à jour la table séquence prévue à cette effet (voir mémoire partie **6.3.2 Le modèle de données**).

Une collection de fichiers de séquences correspondant au jeu de séquences importé dans la base ProteomIs va ensuite servir de banque personnelle de confrontation en association avec l'outil de recherche de similarité BLAST.

Cette collection de fichiers devra être prétraitée au préalable par des techniques d'indexation. En effet l'outil BLAST utilise des structures à base d'index pour rendre les recherche de similarité quasi-instantanée. Nous conservons la banque de données à la fois sous sa forme originelle (collection de fichier FASTA) qui nous sera précieuse pour d'autres traitements (recherche de motifs par exemple) et sous sa forme indexée (uniquement pour BLAST).

b) Cas d'utilisation détaillée

Le cas d'utilisation *« Importation des séquences »* (**diagramme 14**) décrit la mise à jour des séquences dans ProteomIs et la création du dépôt de séquences au format FASTA. Les cas d'utilisation en vert correspondent aux tâches effectuées par le programme d'importation des séquences.
Le cas d'utilisation *« Comparaison de séquence avec BLAST »* (**diagramme 15**) décrit l'utilisation du logiciel BLAST. Les cas d'utilisation en vert correspondent aux tâches effectuées par le programme qui interface BLAST et affiche les résultats du BLAST dans l'interface

Une interface doit permettre à l'utilisateur de saisir la séquence et les bons paramètres pour lancer l'exécution de BLAST en ligne de commande par l'intermédiaire d'un programme. Les résultats doivent également être accessibles par l'utilisateur à travers une interface.

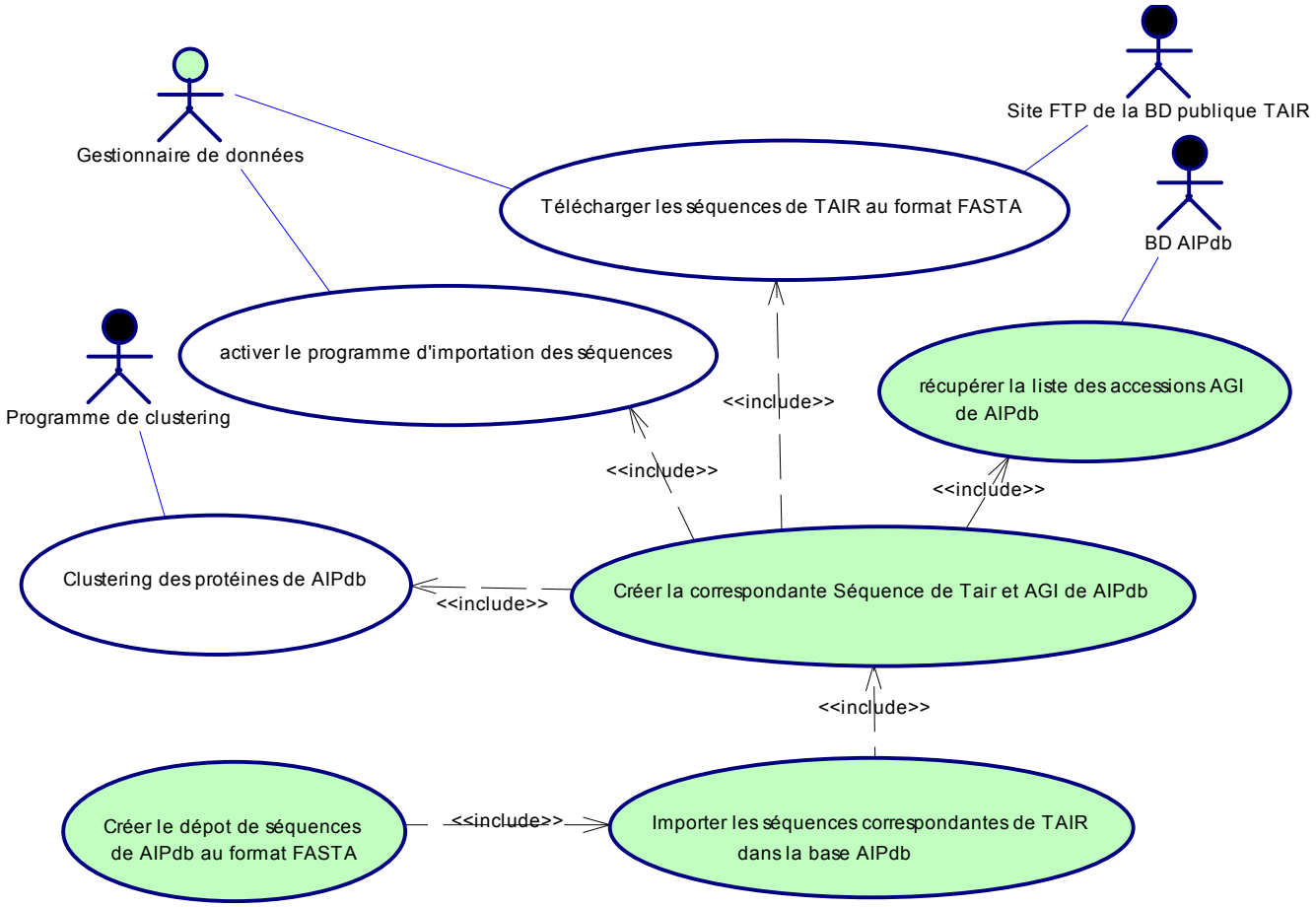
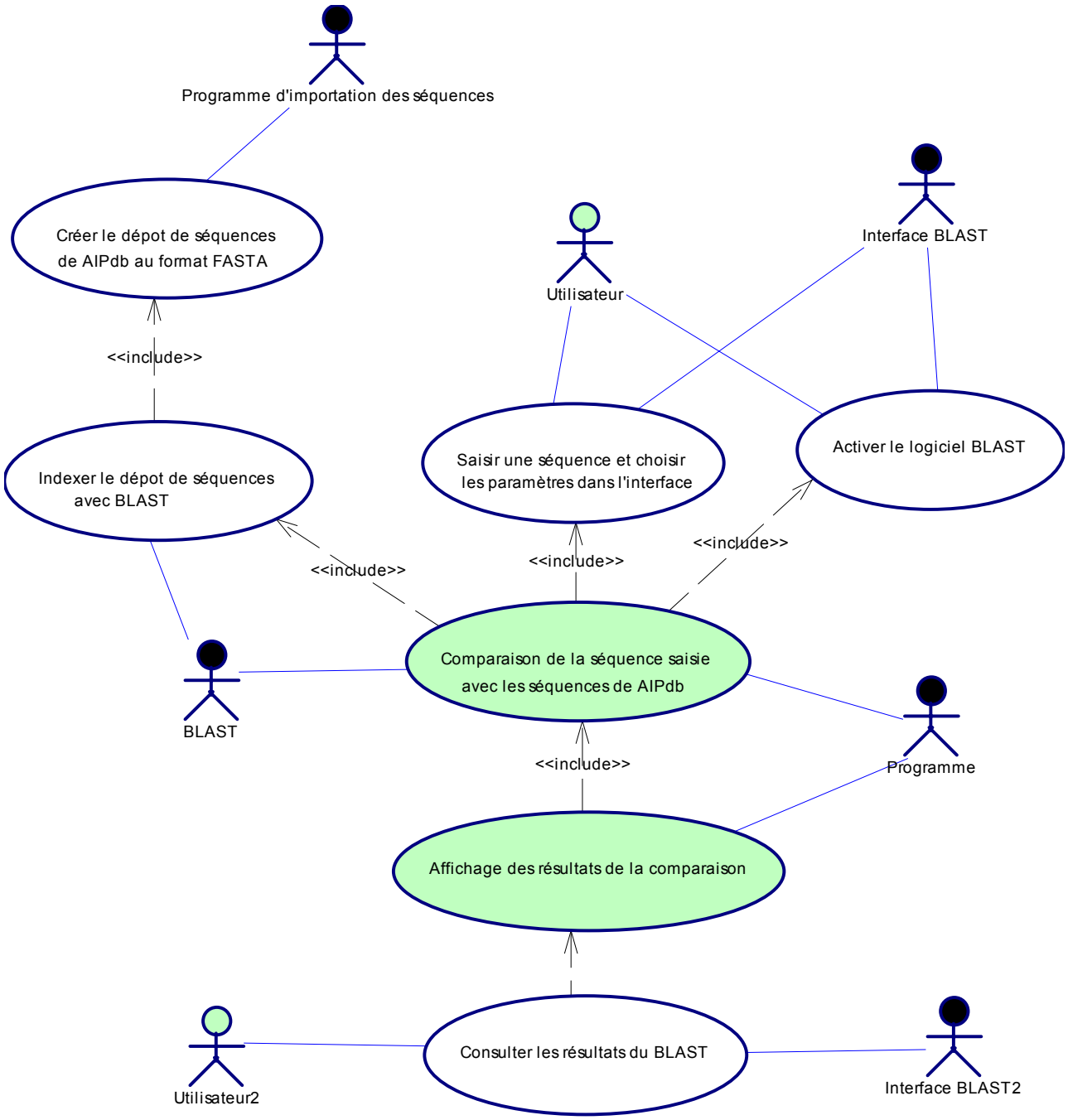


Diagramme 14: Diagramme de cas d'utilisation : Importation des séquences

<u>Diagramme 15 :</u> Diagramme de cas d'utilisation : Comparaison de séquences avec BLAST



c) Aspects dynamiques

Le diagramme de séquence *« Importation des séquences »* (**diagramme 16**) décrit la chronologie des évènements dans le cas d'utilisation *« Importation des séquences »*.

Le diagramme de séquence *« Comparaison de séquence avec BLAST »* (**diagramme 17**) correspond lui au cas d'utilisation *« Comparaison de séquence avec BLAST »*.

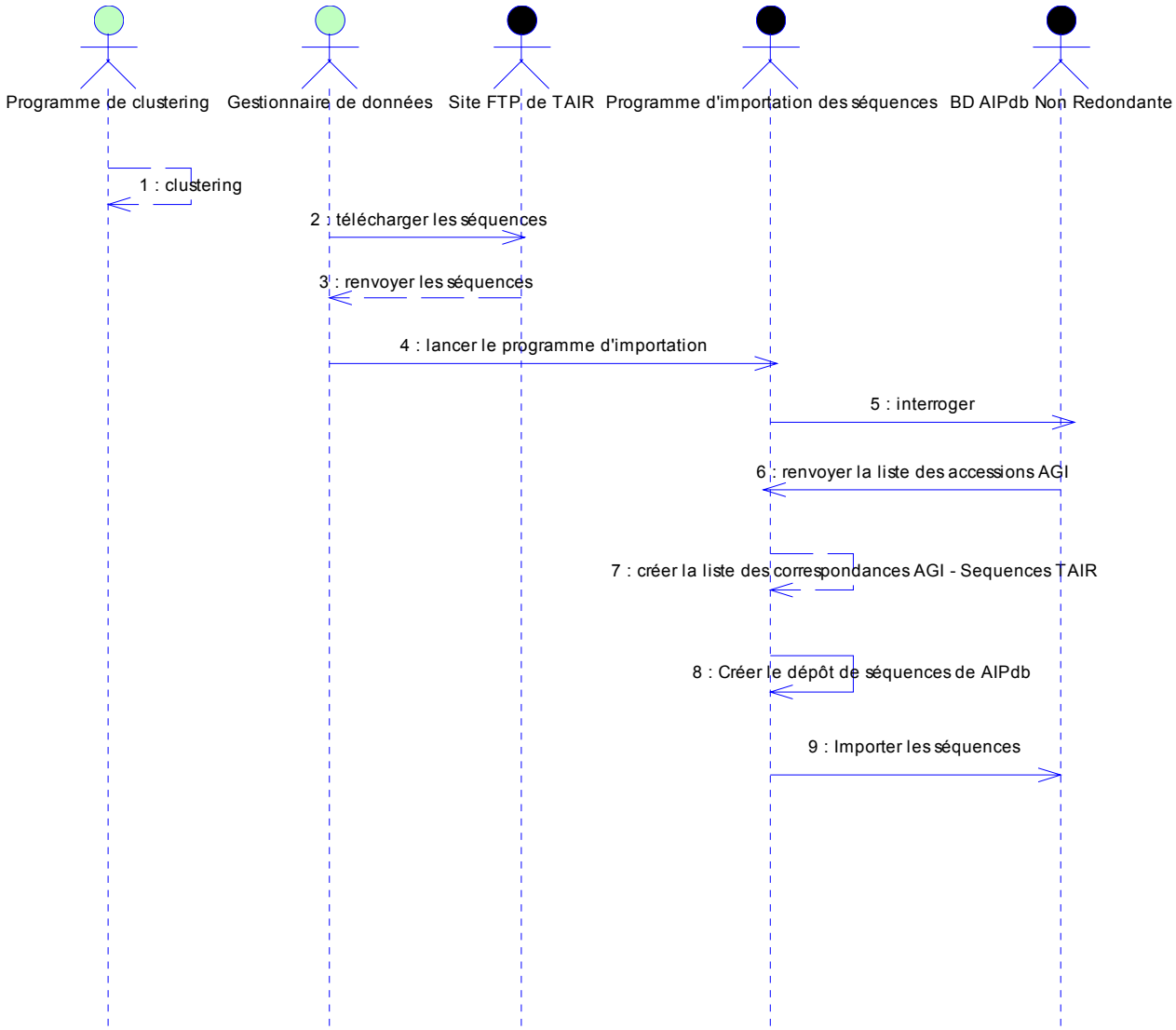
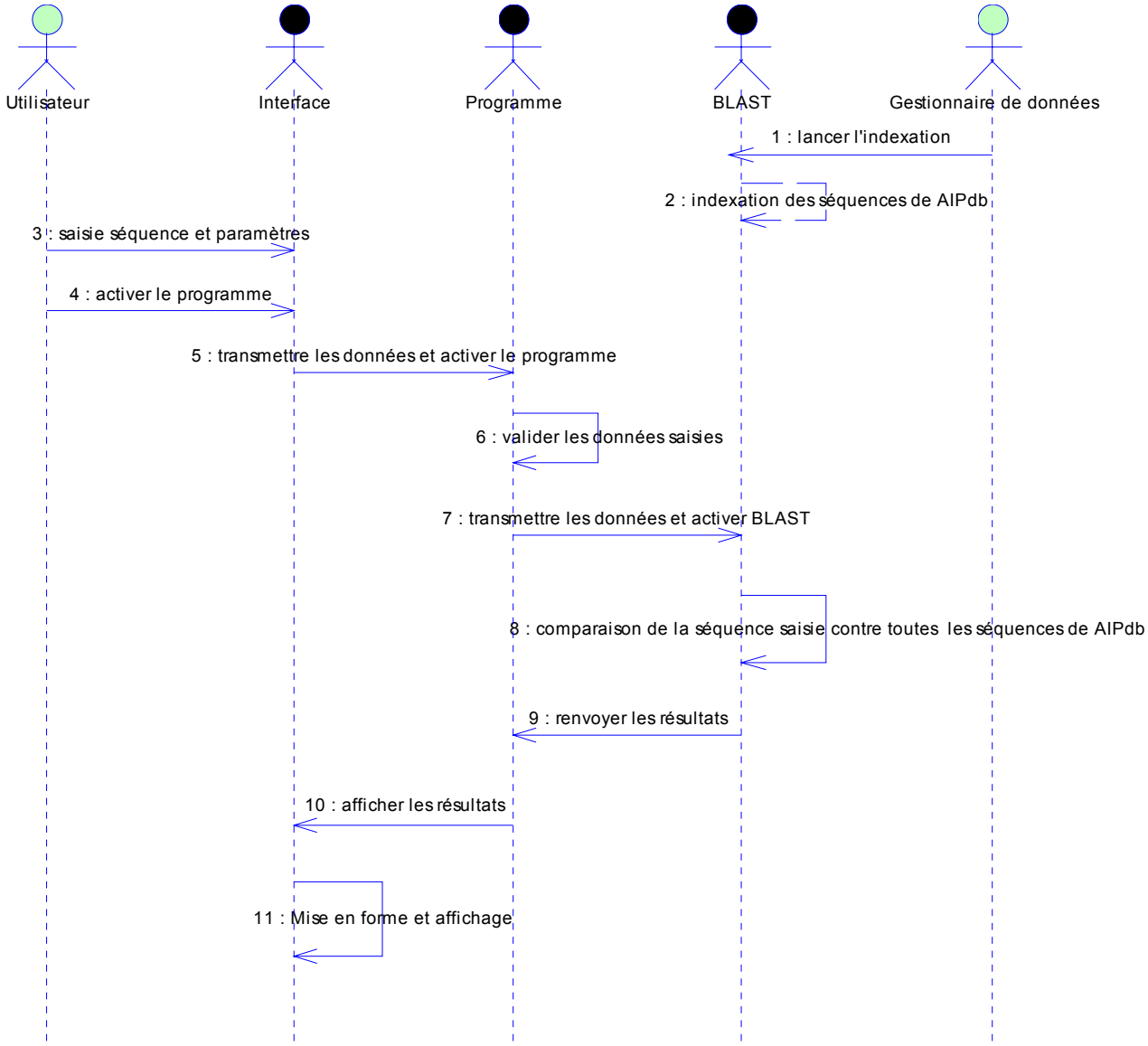


Diagramme 16 : Diagramme de séquence : Importation des séquences

Diagramme 17 : Diagramme de séquences : Comparaison de séquences avec BLAST



3.3 Recherche de motifs

a) Spécification des besoins par écrit

D'après le cahier des charges, la priorité est de se concentrer sur la recherche de motifs correspondant aux sites de phosphorylation.
Il s'agit ici de rappeler et détailler la solution retenue en **5.2.3** afin de préparer le travail de modélisation à l'aide d'UML.

Nous avons vu en **5.2.3** que l'objectif était de construire un programme capable d'automatiser le travail des logiciels MSDigest et NetPhos sur toutes les séquences de ProteomIs. Le programme doit donc être capable d'accepter en entrée un grand nombre de séquences ainsi que des résultats expérimentaux sur ces séquences afin de construire un tableau de synthèse permettant de comparer résultats expérimentaux et résultats d'analyses bioinformatiques.

Cet outil devra être intégré au sein de l'application ProteomIs et il devra être possible de le lancer manuellement ou automatiquement à intervalle régulier pour analyser toutes les données de séquences. Enfin les résultats d'analyse devront au final être stockés dans ProteomIs.

Pour faciliter l'accès des données au biologiste, les résultats devront pouvoir être consultables à travers une interface. On choisira de développer dans un premier temps des interfaces fournissant les résultats des analyses bioinformatiques sous la forme de tableaux.
On prévoira également une possibilité d'exporter les résultats dans un fichier sous différents formats (CSV, Excel, XML …). Dans un deuxième temps et de manière complémentaire, je choisirai de développer une interface graphique permettant de visualiser de manière schématique la position des acides aminés phosphorylés dans la séquence (voir partie **8.2 Perspectives** du mémoire).

Parallèlement la chaîne de traitements devra pouvoir être utilisée manuellement à travers un formulaire autorisant la saisie ou l'importation à partir d'un fichier de séquences propriétaires. Une option dans ce formulaire devra permettre également d'importer les données expérimentales pour les intégrer aux résultats d'analyses. Ces résultats devront être accessibles dans le même format que celui énoncé précédemment.

b) Cas d'utilisation détaillée

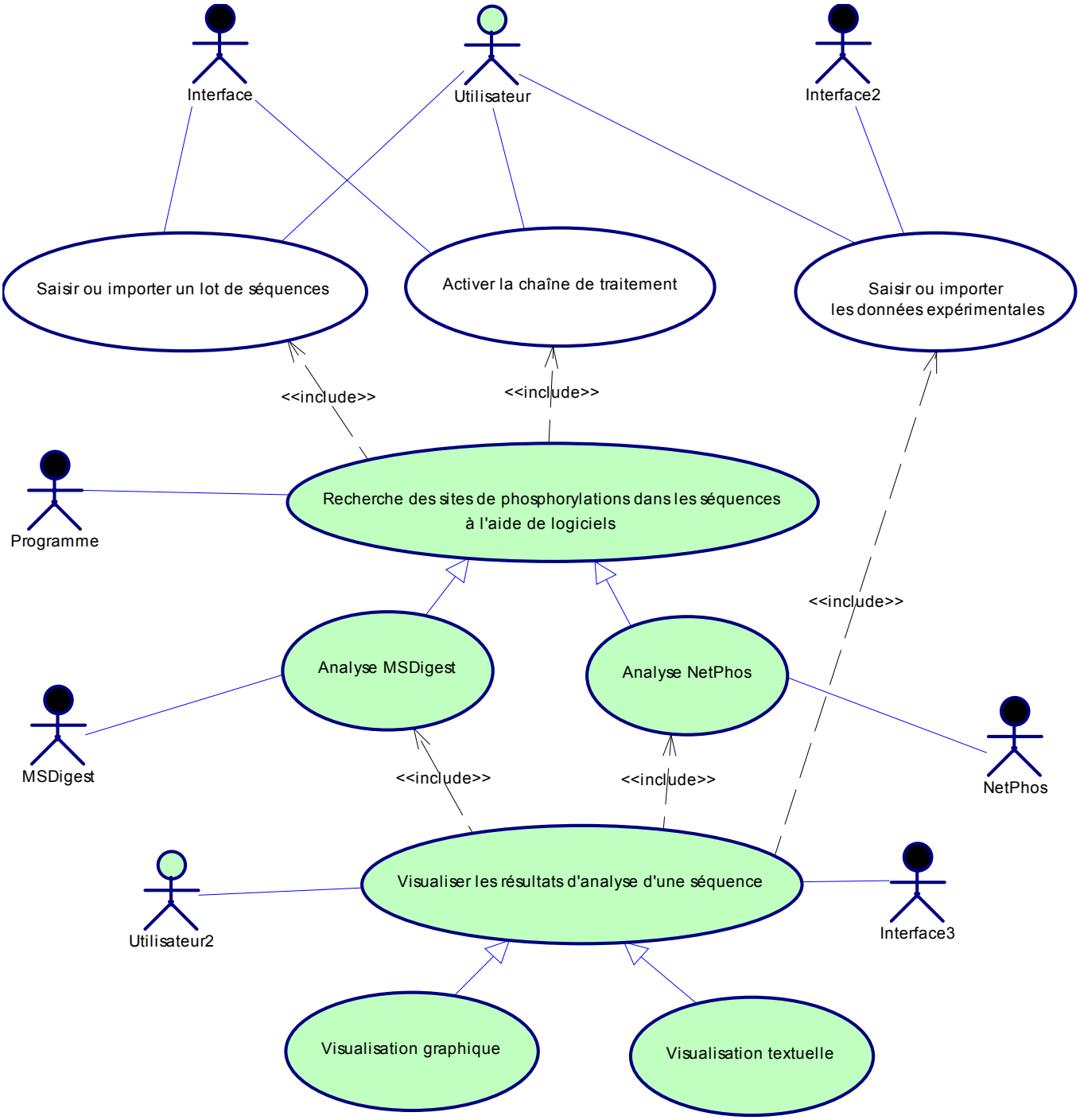
Sur le **diagramme 18** est présenté le diagramme de cas d'utilisation *« Recherche des motifs de phosphorylations dans les séquences contenues dans ProteomIs »*.
Le travail de la chaîne de traitement proprement dite est associé à l'acteur « Programme » et aux cas d'utilisation colorés en vert. Le programme ayant aussi un rôle d'affichage les cas d'utilisations qui traitent de la visualisation sont également en vert.

C'est le gestionnaire des données qui activera périodiquement en ligne de commande la chaîne de traitements sur les séquences contenues dans la base, en fonction des mises à jour.

Le travail du programme « chaîne de traitements » se termine par l'importation des séquences dans la base ProteomIs.
Parallèlement le gestionnaire de données est chargé d'alimenter le programme avec les données expérimentales que peut lui fournir le biologiste sur les données de séquences.

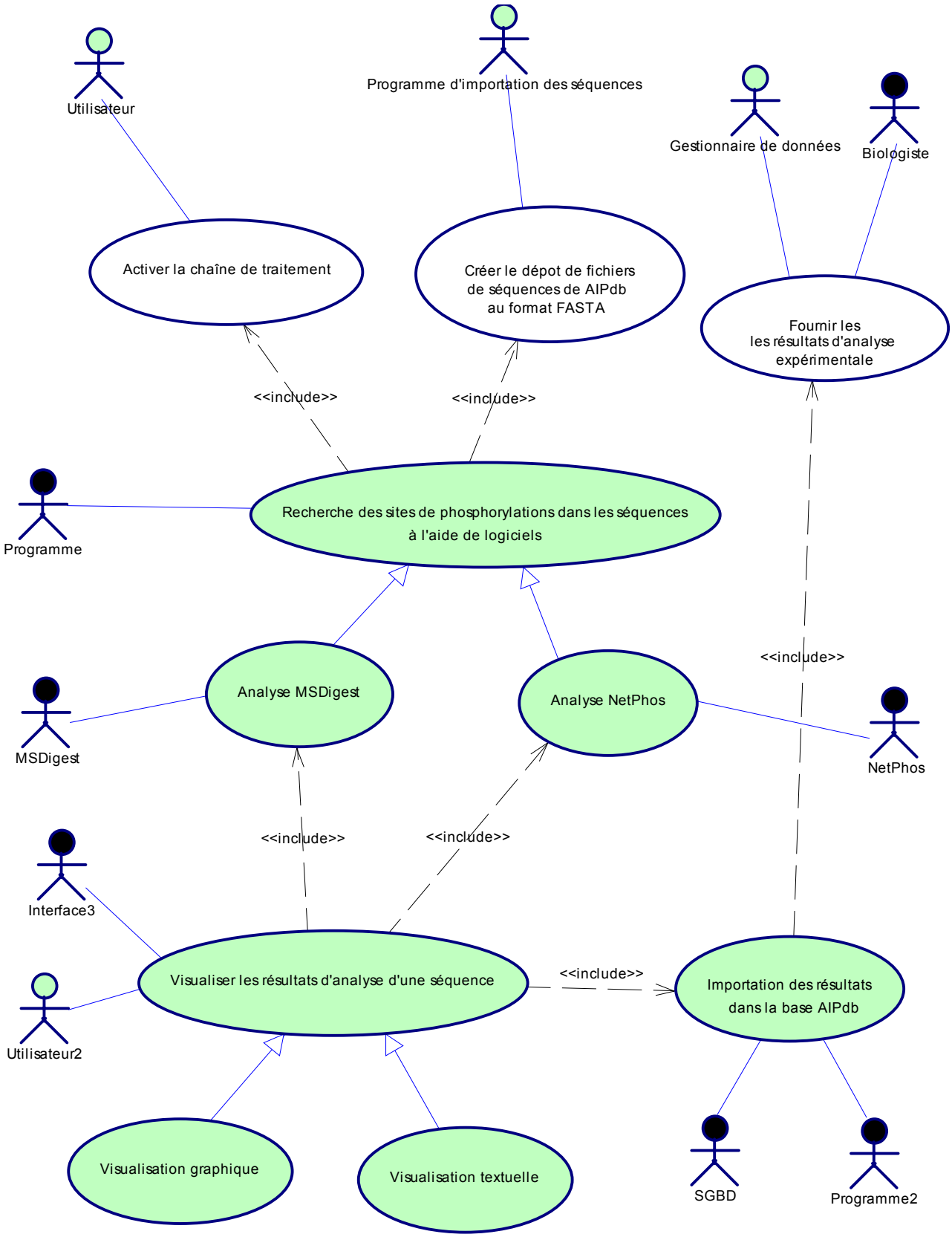
Sur le **diagramme 19** est présenté ensuite le diagramme de cas d'utilisation *« Recherche des motifs de phosphorylations à partir d'un lot de séquences fournies par l'utilisateur »*.

C'est la procédure permettant à un utilisateur d'utiliser la chaîne de traitements directement sur ces propres de données sans passer par la base et en les saisissant dans un formulaire.



<u>Diagramme 18 :</u> Diagramme de cas d'utilisation : Recherche manuelle des motifs de phosphorylation à partir d'un lot de séquences fournies par l'utilisateur

<u>Diagramme 19 :</u> Diagramme de cas d'utilisation : Recherche des motifs de phosphorylation dans toutes les séquences contenues dans la base ProteomIs



c) Aspects dynamiques

Nous avons représenté ici l'aspect chronologique des traitements décrits dans les cas d'utilisation précédents.

Le diagramme de séquence *« Recherche des motifs de phosphorylations dans les séquences contenues dans ProteomIs »* (**diagramme 20**) correspond au cas d'utilisation *« Recherche des motifs de phosphorylations dans les séquences contenues dans ProteomIs »*.

Le diagramme de séquence *« Recherche des motifs de phosphorylations à partir d'un lot de séquence fournit par l'utilisateur »* (**diagramme 21**) correspond au cas d'utilisation *« Recherche des motifs de phosphorylations à partir d'un lot de séquences fourni par l'utilisateur »*.

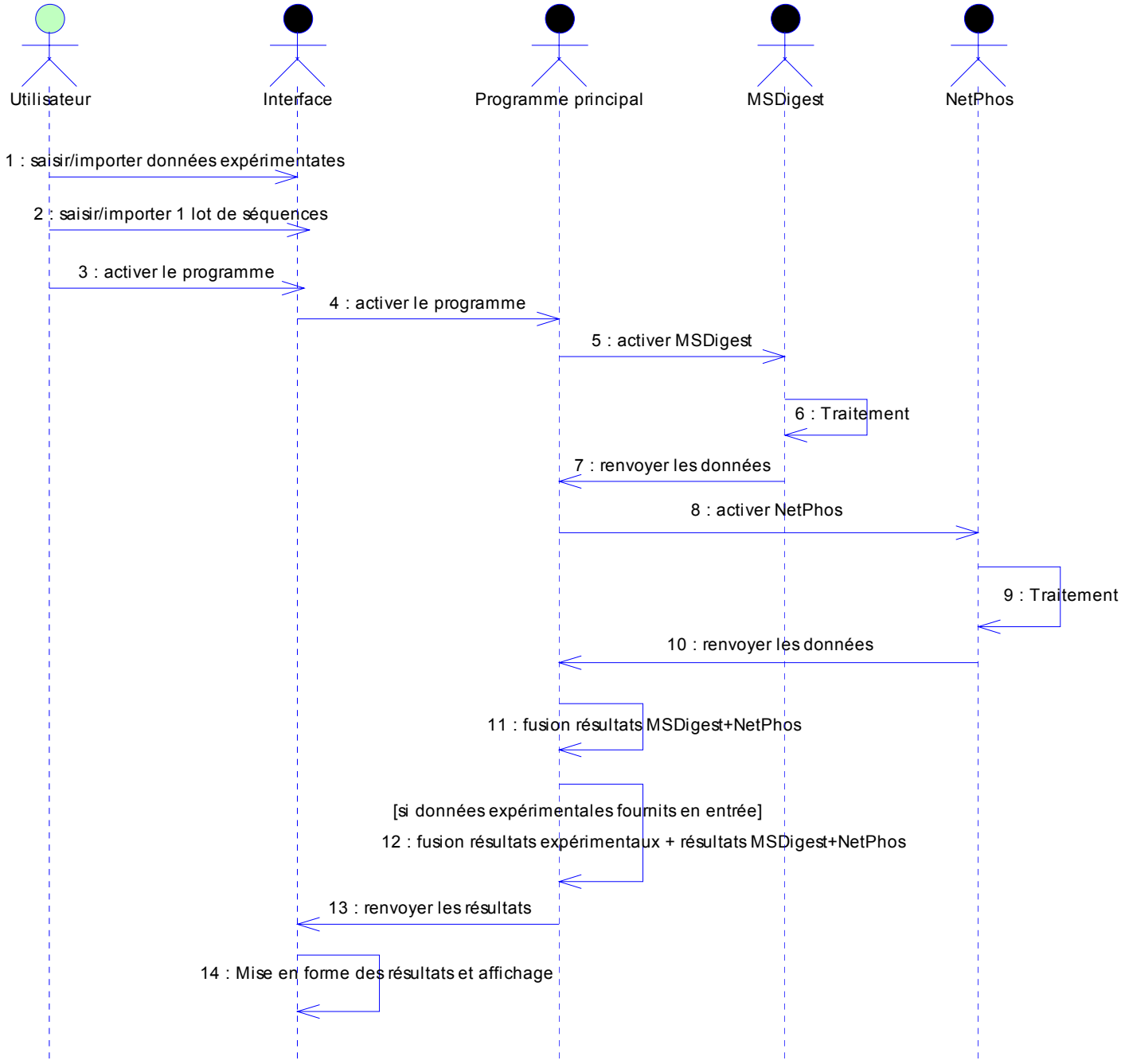
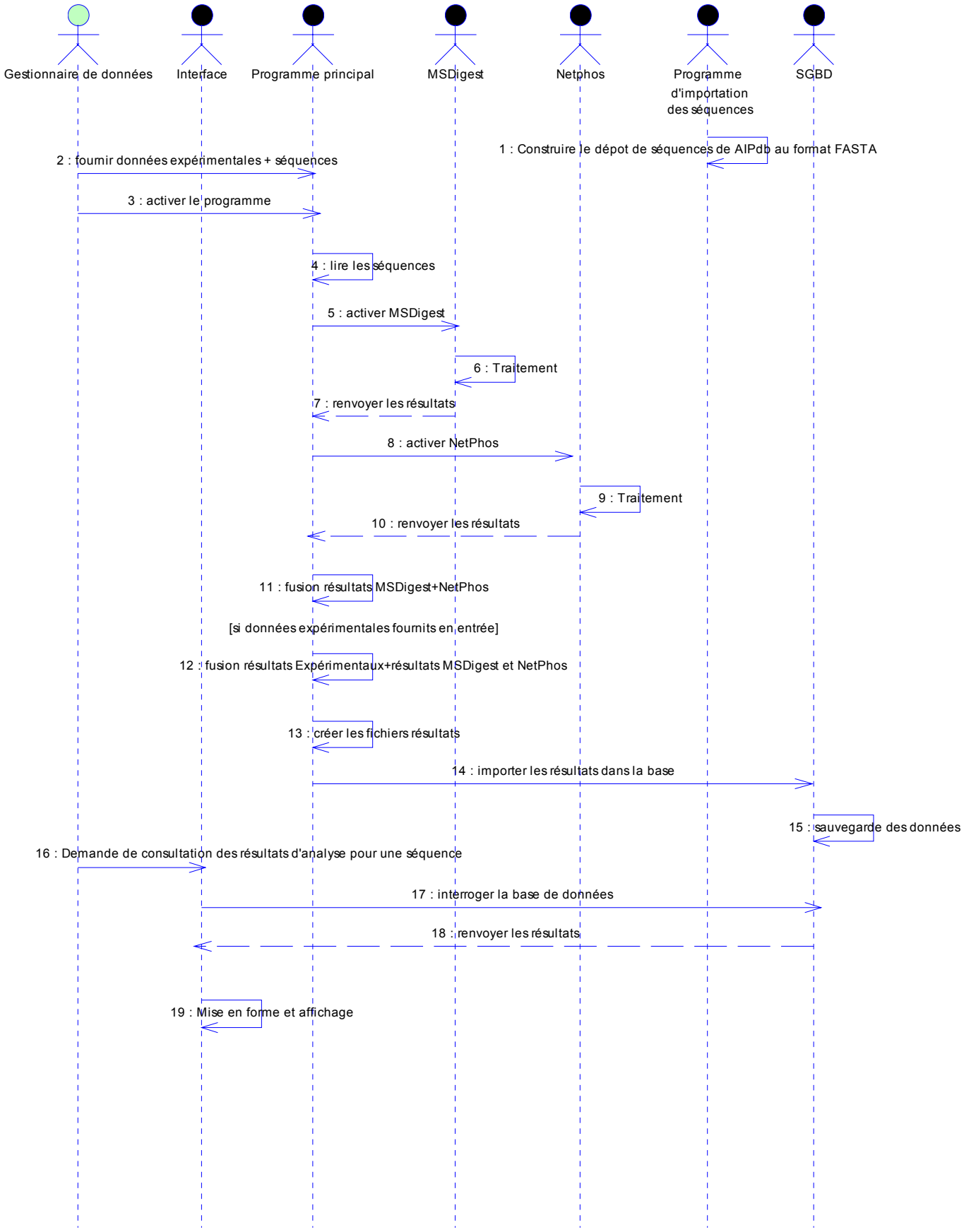


Diagramme 20 : Diagramme de séquence : Recherche automatique des motifs de phosphorylation à partir d'un lot de séquences fourni par l'utilisateur

<u>Diagramme 21 :</u> Diagramme de séquence : Recherche des motifs de phosphorylation dans les séquences contenues dans ProteomIs



ANNEXE 2 : Phase de maquettage (complément)

Interface d'interrogation :

Description :
Sur cette interface des options de recherche avancées devront être mis à la disposition de l'utilisateur.
Maquette :
Sur le **document 1** page suivante est présenté la maquette de l'interface d'interrogation.
L'interface est composée de 3 parties :

- La barre de menu
- La barre de recherche rapide
- Le corps du documents

La barre de menu contient :

- un lien intitulé « Home » qui nous permet de revenir sur l'interface d'accueil.
- un lien intitulé « Search » (flèche jaune 1) qui permet de rediriger directement sur l'interface d'interrogation : l'Interrogation tools.
- un lien intitulé « BLAST » qui dirige vers le formulaire permettant de faire une recherche avec l'outil BLAST. Cette interface est présentée dans la partie **7.3.2** de ce mémoire.
- un lien intitulé « Motifs » qui dirige vers le formulaire permettant de faire une recherche avec l'outil de recherche de motifs. Cette interface sera présentée dans la partie **7.3.3** de ce mémoire.
- un lien intitulé « GpiIS » (uniquement pour la version intégrée de ProteomIs) qui pourrait rediriger vers une interface permettant de faire des recherches sur tout le système GpiIS et non plus uniquement sur le module protéomique GnpProt du système.

  Cette interface ne sera pas présentée dans ce mémoire car son implémentation n'a pas encore été effectuée. Ce travail devrait démarrer en collaboration avec Génoplante dans le courant de l'année (voir mémoire partie **8.2 Perspectives**).
- un lien sur la documentation utilisateur de l'application.

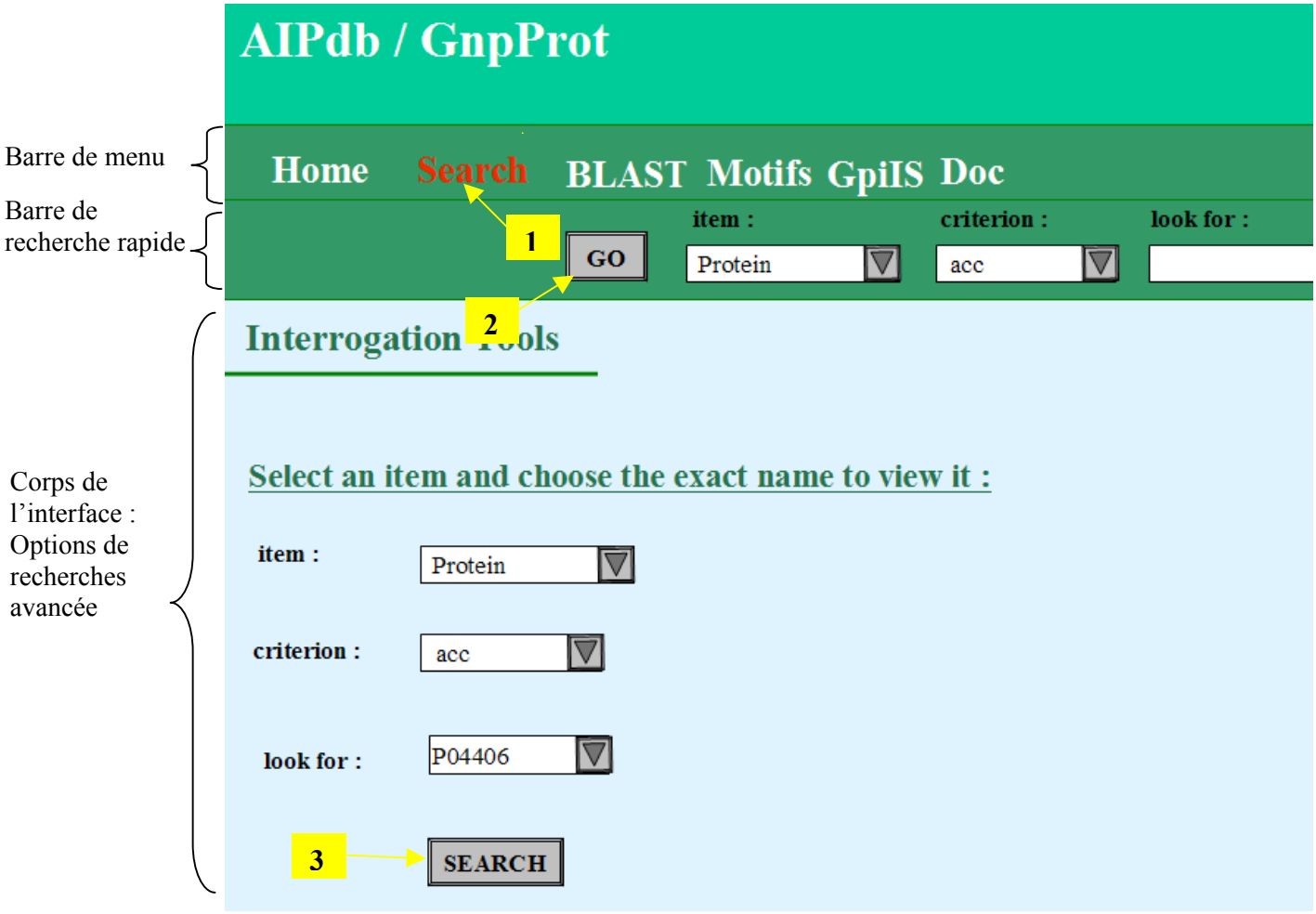
La barre de recherche rapide permet ensuite de réaliser une interrogation par mots clés.
Cette interrogation se fait grâce à 2 listes déroulantes et une zone de saisie :

- la liste Item : permet de sélectionner le type d'objet recherché (exemple : un projet, un contact, une expérience, une protéine, etc …)
- la liste Criterion : permet de sélectionner le critère sur lequel on va pouvoir sélectionner les occurrences de cet objet (exemple : pour une protéine cela peut-être son numéro d'accession, son nom, sa fonction …).
- la zone de saisie (flèche jaune 2) : on peut saisir ici tout ou une partie du nom de l'objet recherché. On obtiendra alors la liste des objets contenant tout ou partie de ce nom.

Dans le corps de l'interface sont présentés les options de recherche avancées du système ProteomIs. Pour l'instant une seule option a été présentée permettant d'effectuer une combinaison importante d'interrogations. Cette option demeurant cependant très complète, elle s'est avérée pour l'instant suffisante pour les utilisateurs.
C'est en fait une recherche par critère constituée cette fois de trois listes déroulantes au lieu de deux comme dans le cas de la recherche rapide :

- la liste Item : permet de sélectionner le type d'objet recherché
- la liste Criterion : permet de sélectionner le critère sur lequel on va pouvoir sélectionner les occurrences de cet objet
- la liste Look for : permet de sélectionner au final parmi les occurrences disponibles pour le critère sélectionné

<u>Document 1 :</u> Maquette PowerPoint « Interface d'interrogation »



Interface Liste de résultats :

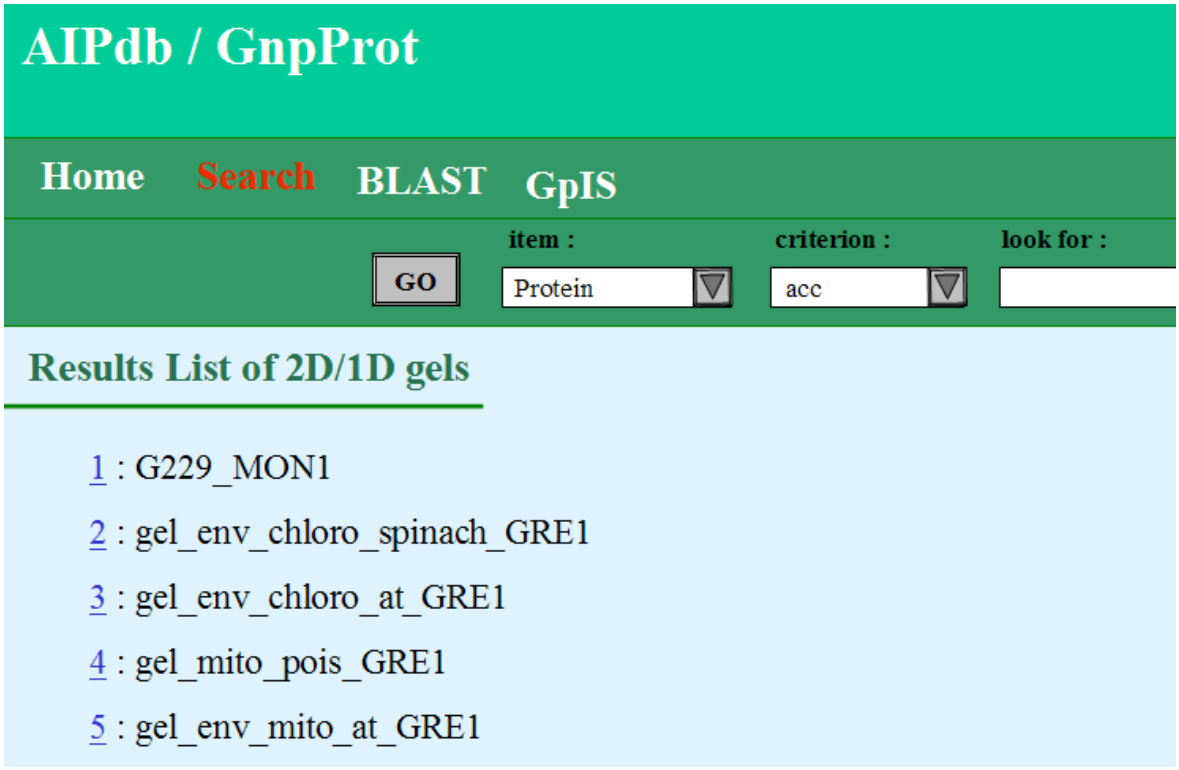
<u>Description :</u>

C'est une interface présentant le résultat d'une requête d'interrogation sous la forme d'une liste lorsque le nombre de résultats renvoyés par la requête est > 1.

<u>Maquette :</u>

Voici pour exemple sur le **<u>document 2</u>** la maquette de la liste de résultats obtenus à partir de la requête : « Rechercher tous les gels disponibles dans ProteomIs ».

<u>Document 2 :</u> Maquette PowerPoint « Interface Liste des résultats ».



Chacun des identifiants des gels contenus dans la liste offre un lien sur l'interface de visualisation détaillée du gel correspondant. Si la liste des résultats est > 50 une option permettant de faire défiler l'affichage des résultats par groupe d'éléments devra être développée. Ceci permettra à l'utilisateur de parcourir la liste des éléments avec plus de précision qu'avec l'aide d'un ascenseur.

<u>Aspects préliminaires de présentation :</u>

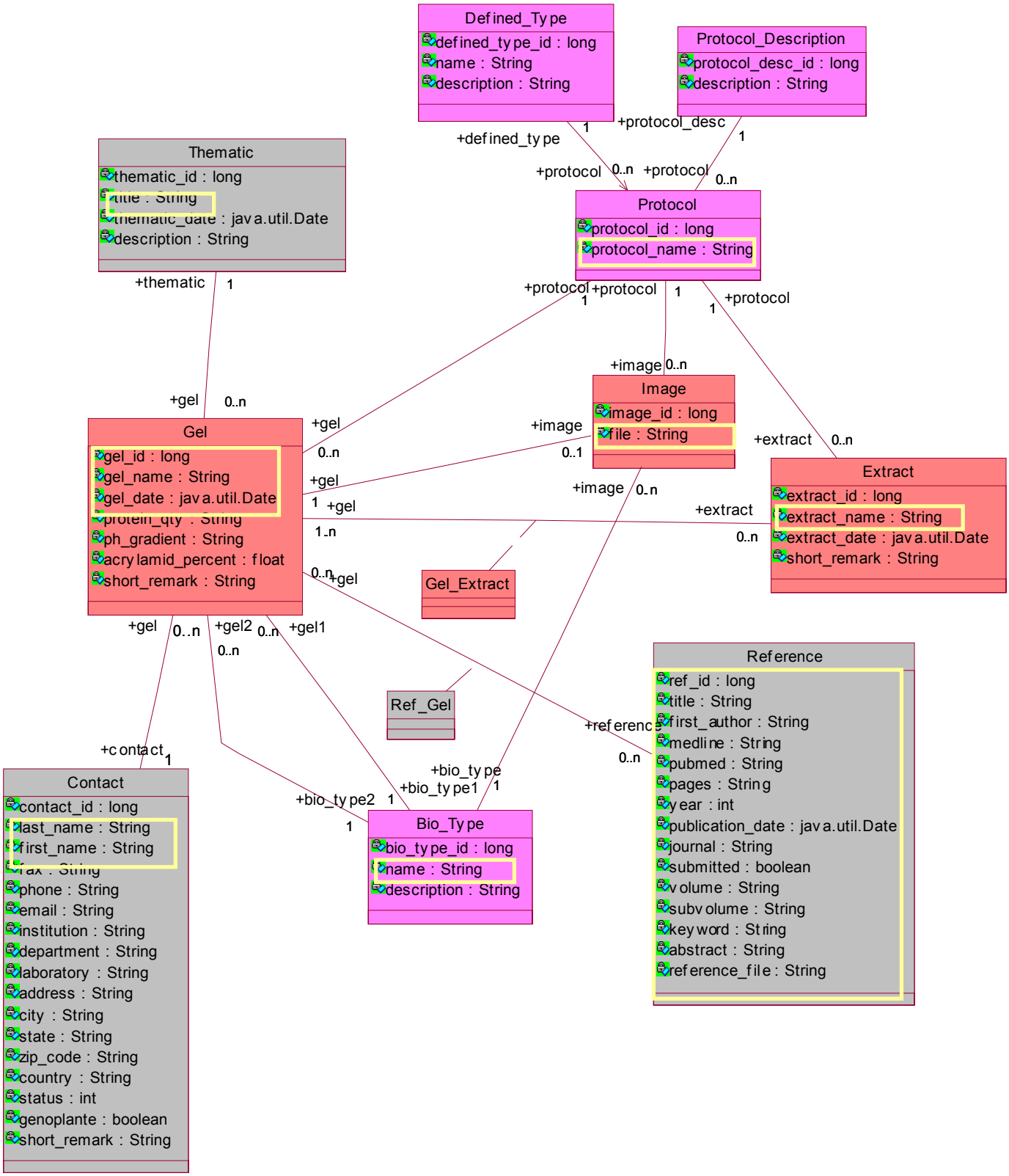
On remarquera que déjà dans la phase de maquettage on s'attache à respecter une certaine harmonie dans l'organisation visuelle et fonctionnelle entre les différentes interfaces de l'application. Ceci afin de susciter déjà de la part de l'utilisateur quelques réactions à ce niveau avant la phase d'implémentation. A ce stade de la conception, on peut déjà prévoir que les interfaces doivent être présentées de façon cohérente conformément à une charte graphique. La charte graphique permet d'uniformiser la présentation visuelle des interfaces dans toute l'application en définissant des choix fixe de styles, de polices de caractères, de couleurs, etc. Au niveau de l'harmonisation en terme de fonctionnalité on va par exemple choisir de conserver le bandeau de menu et la barre de recherche rapide sur toutes les interfaces de visualisation de l'application.

<u>Interfaces de visualisation des informations sur un Gel :</u>

Il s'agit de l'interface permettant de visualiser sous forme textuelle les informations sur un gel.

<u>Inventaire des données à visualiser :</u>

On a ici représenté sur le **<u>diagramme 3</u>** l'ensemble des données du Modèle Conceptuel de Données permettant d'apporter des informations sur un gel. La première étape a consisté à préciser l'ensemble des données devant être visualisées par l'interface de navigation des gels. Ces données correspondent à celles encadrées en jaune dans le diagramme.



<u>Diagramme 3 :</u> Domaine du MCD couvert par l'interface de visualisation des informations textuelles sur un gel

Maquette :

Sur le **document 4** et le **document 5** (page suivante) est présenté en deux parties la maquette de l'interface de visualisation correspondant au gel 2D nommé G229_MON1.

Document 4 : Maquette Powerpoint « Interface de visualisation des informations textuelles d'un gel »

AIPdb / GnpProt

Home Search BLAST GpIS

GO item : Protein criterion : acc look for :

Gel details

General Informations

Gel Id : 1

Gel Name : G229_MON1

General Description

Date : 2001-10-17

Gel type : 2D

Short remark : 18cm

Gel Protocol : Analytical 2D gel

Gel coloration type : bleu de Coomassie

Contact : Dupont Jean

Thematic : Root proteome characterization_MON1

Image

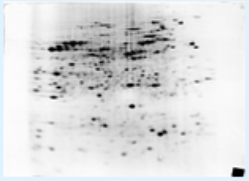
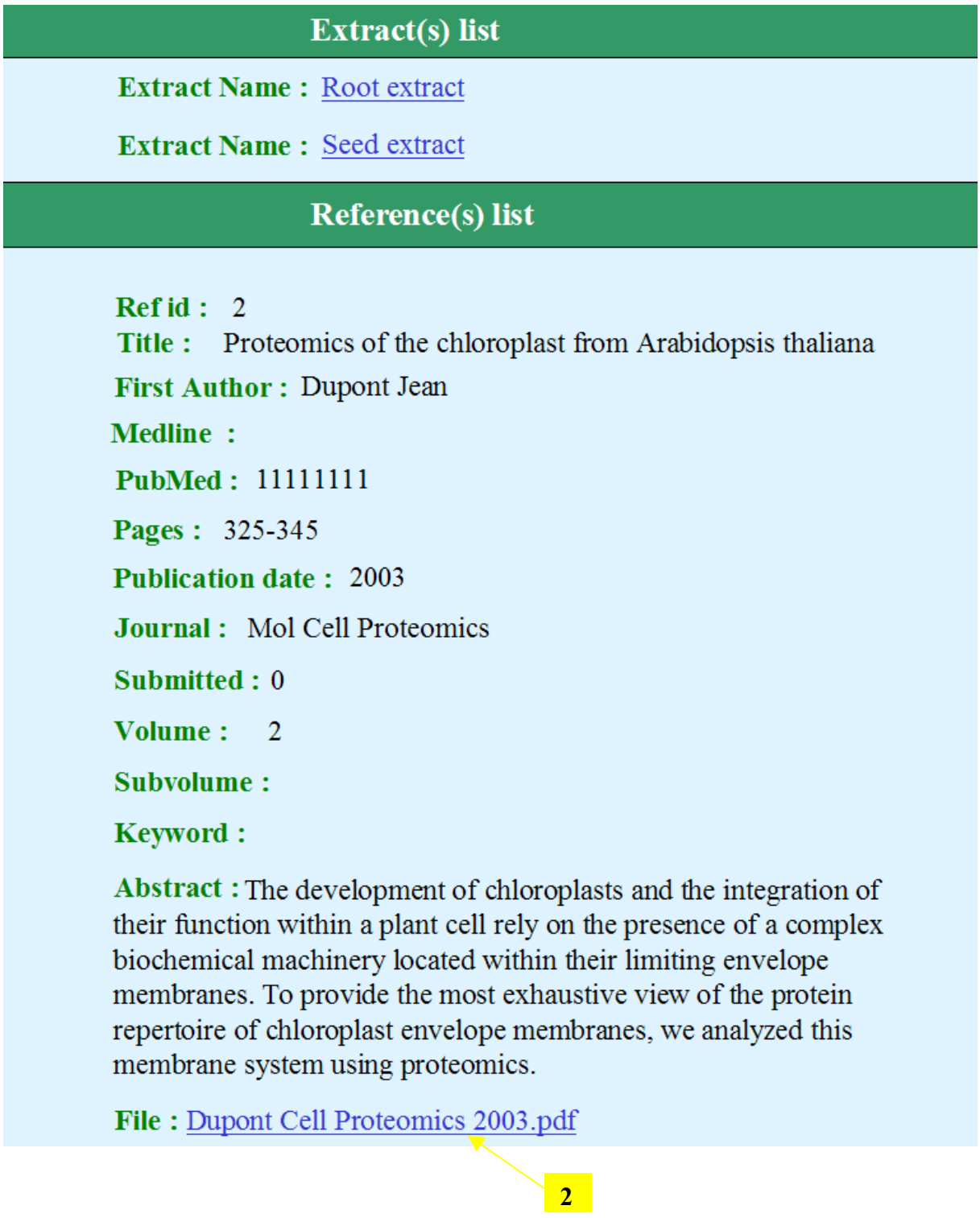


Image Format : jpg

Image protocol : CCB Image Analysis

Document 5 : **Maquette Powerpoint « Interface de visualisation d'un gel » (Suite)**



Toutes les interfaces de visualisation seront conçues sur ce modèle. Encore une fois l'harmonie du graphisme avec les maquettes précédente est respectée et les bandeaux « menus » et « recherche rapide » sont présents.

Le corps du document contenant l'information est ensuite divisé en sections qui correspondent à un type de données bien distinct. Les éléments de la rubrique « General informations » sont des informations qui permettent d'identifier l'élément étudié : ici l'identifiant dans la base de données (qui est un numéros incrémenté par le SGBD) et le nom du Gel. Les éléments de la rubrique « General Description » sont des informations qui décrivent l'objet.

Lorsqu'un type d'information ne peut pas tenir sur une ligne (exemple l'image du gel) on prévoit de lui réserver une section (section Image). C'est également le cas lorsqu'un type de données est représenté plusieurs fois. Dans notre maquette on a le cas des extraits (section Extract(s) List) et des publications (section Reference(s) List) associés à un gel puisque un gel peut être associé à un ou plusieurs extraits et un gel peut être associé à une ou plusieurs publications.

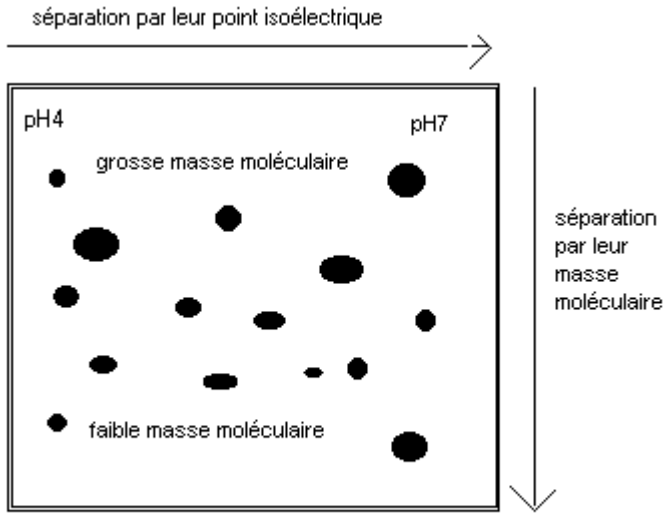
Fichiers téléchargeables :

Dans la section « Rerence(s) List » un lien (flèche numéro 2) est disponible vers le fichier pdf de la publication permettant de télécharger le fichier correspondant.
Maintenant que l'on a présenté le principe de conception de la maquette pour une interface de visualisation, nous nous contenterons de faire un inventaire des autres interfaces en décrivant simplement leur contenu très succinctement.

ANNEXE 3 : Démarche expérimentale en protéomique utilisant le couplage des techniques d'électrophorèse bidimensionnelle et de spectrométrie de Masse Maldi tof

I – Première étape : gel 2d d'électrophorèse

L'électrophorèse bidimensionnelle est, à l'heure actuelle, la technique la plus résolutive de séparation des mélanges complexes de protéines. Elle permet de séparer les protéines en 2 dimensions selon 2 propriétés physico-chimiques différentes. La première séparation se fait dans une première dimension par focalisation isoélectrique (IEF) selon leur point isoélectrique (pI) et la seconde se fait dans la direction perpendiculaire par rapport à leur masse moléculaire.

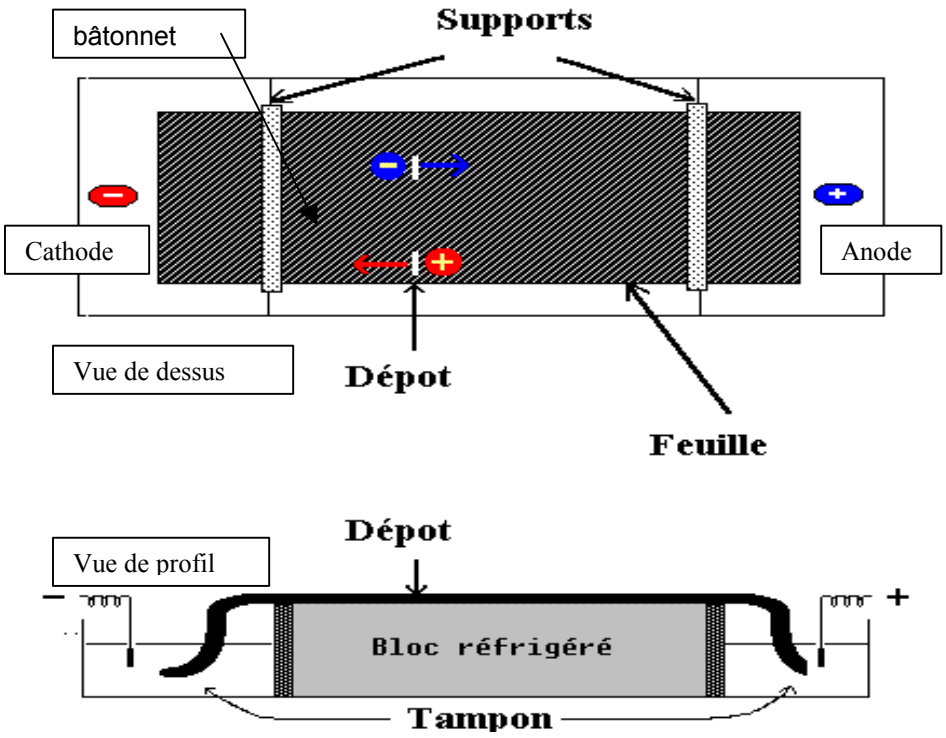


A. *Préparation des extraits protéiques*

Les protéines sont extraites à partir d'un échantillon de tissus organiques (p.ex 1,5 g de racine d'Arabidopsis …).
Pour cela, ces tissus organiques sont broyées dans l'azote liquide. Les protéines sont ensuite précipitées et sédimentées par centrifugation. Puis elles sont finalement solubilisées dans une solution tampon qui servira à l'électrofocalisation

B. *Première dimension*

L'électrofocalisation est une méthode qui sépare les protéines selon leur point isoélectrique (pI). Les protéines sont des molécules amphotères, elles peuvent être chargées positivement, négativement ou ne pas avoir de charge selon le pH de la solution dans laquelle elles se trouvent. La charge nette d'une protéine est la somme des charges négatives et positives de ses extrémités et de ses chaînes latérales qui la composent. Les protéines chargées positivement migrent, à travers le gradient de pH, vers la cathode en perdant, au fur et à mesure, leurs charges positives pour arriver à une charge nette nulle quand elles atteignent leur pI.



Séparation 1ere dimension

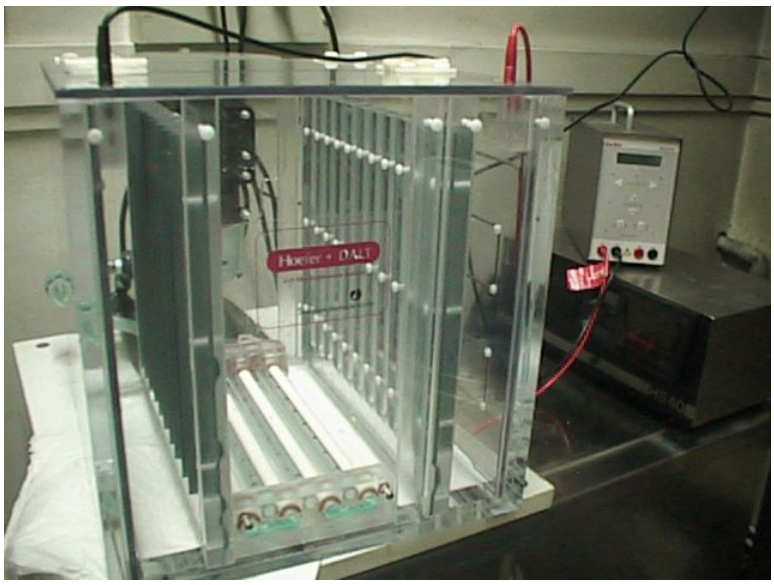
Ainsi, on retrouve les protéines éparpillées sur le bâtonnet d'une taille variant de 12 à 24 cm selon pHi. Cette étape se réalise sur le type d'appareil si dessous :

Séparation 1ere dimension

C. Seconde dimension

Dans cette deuxième dimension, les protéines vont être séparées selon leur masse moléculaire. Les protéines qui se trouvent sur la bandelette vont maintenant pénétrer dans le gel par électro-élution et être séparés en fonction de leur masse. On réalise donc une électrophorèse en gel de polyacrylamide en présence de SDS. La migration n'est pas déterminée par les charges électriques des protéines mais par leur poids moléculaire. En effet, le SDS est un agent anionique qui se fixe sur les protéines, masquant la charge propre de celles-ci. Le gel de polyacrylamide sert de tamis moléculaire.

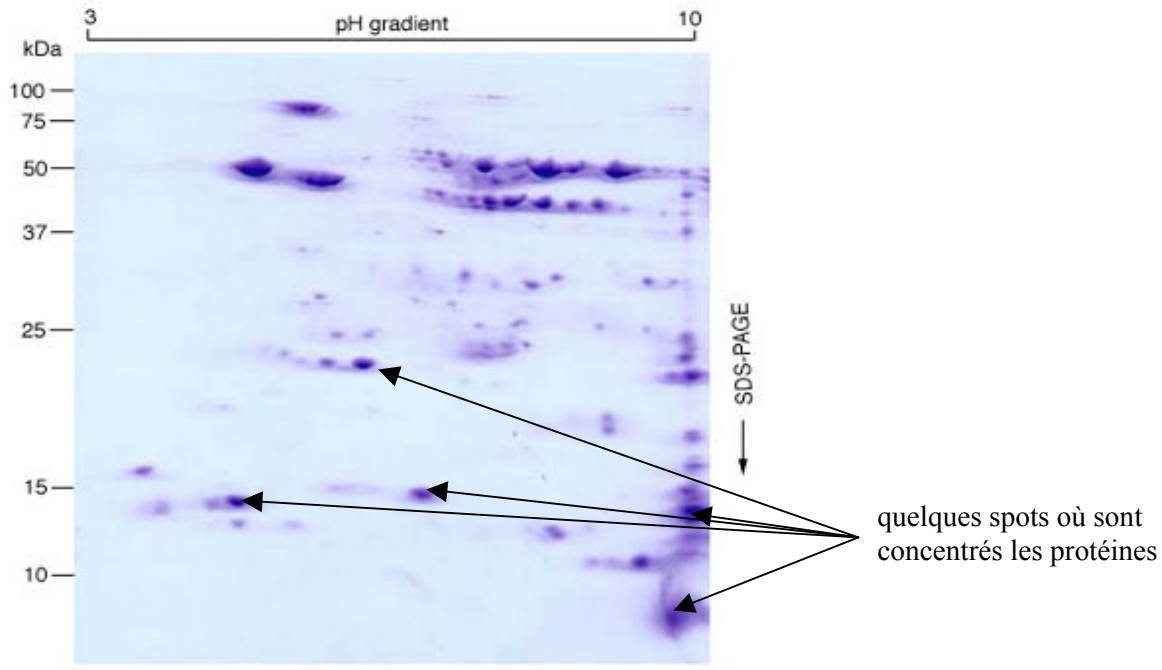
Nous plaçons la bandelette au-dessus du gel. Elle est stabilisée en contact avec le gel en versant de l'agarose. On peut également déposer de chaque côté de la bandelette des marqueurs de poids moléculaire mélangés avec l'agarose. Ainsi, grâce à ces étalons, on pourra avoir une idée approximative de la masse de la protéine que l'on prélèvera.



Cuve électrophorèse

Cette manipulation se réalise sur ce type d'appareil.

Les protéines ainsi séparées sur cette plaque sont ensuite colorées au bleu de Coomassie. L'image du gel est finalement numérisée par un scanner de haute résolution et transformés en données informatiques (fichier image).

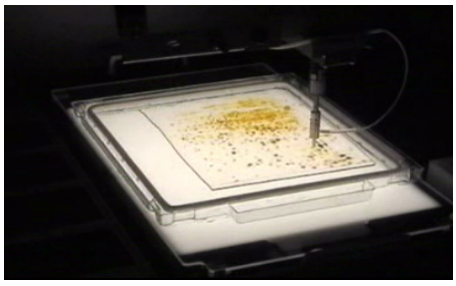


Gel 2D d'électrophorèse

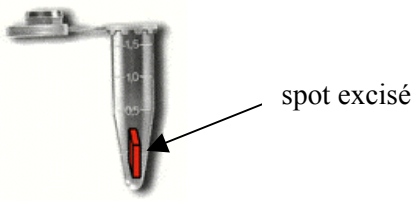
Un logiciel d'analyse d'image spécifique (Mélanie) permet de localiser les spots représentés par des tâches sombres sur le gel. C'est dans ces spots que sont concentrés les protéines.

II – Deuxième étape : découpe des spots

Les spots qui nous intéressent du gel de référence sont soigneusement excisés en morceaux d'1 ou 2 mm^3 (environnement sans poussières ni empreintes de doigts afin d'éviter toutes intrusions de contaminants) et placés dans un micro tube (eppendorf) très propre ou ils seront analysés ultérieurement par spectrométrie de masse afin de déterminer la nature des protéines qu'ils contiennent



Excision des spots sur le gel



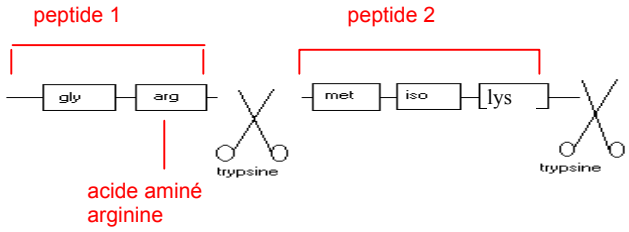
tube eppendorf

III - Troisième étape : digestion trypsique et extraction des protéines

A Digestion des protéines par la trypsine

Nous introduisons donc dans le tube, contenant le morceau de gel excisé, 8µl de solution de trypsine.
La digestion des protéines par la trypsine permet de transformer la protéine contenue dans le gel en plusieurs peptides qui la composent.
En effet, la trypsine coupe les liaisons peptidiques après deux acides aminés (a.a) connus (la lysine et l'arginine).

De manière schématique, la trypsine agit de cette manière :



Cela va nous permettre d'obtenir une combinaison (empreinte) de peptides caractéristiques de la protéine.

B Extraction des peptides

Cette étape va nous permettre de sortir les peptides du gel.

C Dessalage

Le but de cette étape est d'éliminer les sels et autres contaminants introduits dans la solution contenant les peptides. Cela est possible grâce à une phase chromatographique située à l'extrémité d'un cône que l'on peut greffer sur une pipette. On utilise pour cela des ZipTip C_{18}^{TM} (Millipore, Bedford, MA)

ZipTip C_{18}^{TM} (Millipore, Bedford, MA)

IV - Quatrième étape : analyse par spectrométrie de masse

La spectrométrie de masse est une méthode qui permet l'ionisation des biomolécules et qui est en train de devenir un outil puissant pour l'analyse des protéines.

La méthode qui est utilisé en routine à l'unité de recherche 1199 de Montpellier est la méthode MALDI-TOF (Matrix Assisted Laser Désorption Ionisation Time Of Flight). Ce spectromètre de masse est un appareil qui permet de mesurer la masse des peptides. Le peptide (neutre) doit être ionisé dans la source de l'appareil. Nous accélèrerons ensuite ces peptides en appliquant une forte différence de potentiel (d.d.p) entre deux plaques. Un détecteur mesure le temps de vol de chaque peptide. Ce temps de vol est proportionnel au rapport masse sur charge (m/z) du peptide. Le détecteur est relié à une station de travail pour l'archivage informatique et le traitement des spectres de masses obtenus

SPECTROMETRE DE MASSE MALDI-TOF (Bruker, Biflex III)

Il existe cependant une autre méthode appelée Electrospray et qui permet d'ioniser les peptides et protéines et qui peut être associé à un analyseur de type « MS/MS » à des fins de séquençage de peptides. Dans cette annexe, nous nous intéressons à la détermination de protéines en connaissant l'enchaînement des peptides de la protéine issus de la digestion par la trypsine.

Voici résumé les différentes étapes de l'analyse par spectrométrie de masse :

A Dépôt sur plaque

La solution de peptides récupérée dans la troisième étape va être déposée et séchée sur une plaque métallique qui va être ensuite placée dans le spectromètre de masse MALDI-TOF

Dépôt sur plaque

B Analyse au MALDI-TOF

Les peptides vont ensuite être décollés de la plaque par un rayon laser et ionisés.
Un champ électrique envoie les peptides ionisés dans un tube de vol jusqu'au détecteur. Le champ électrique est crée par une forte différence de potentiel (d.d.p) de 19kV entre deux plaques.

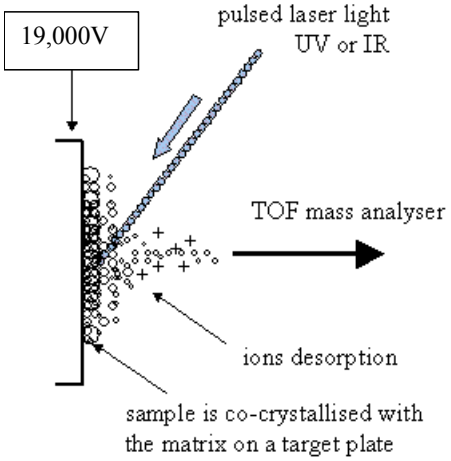
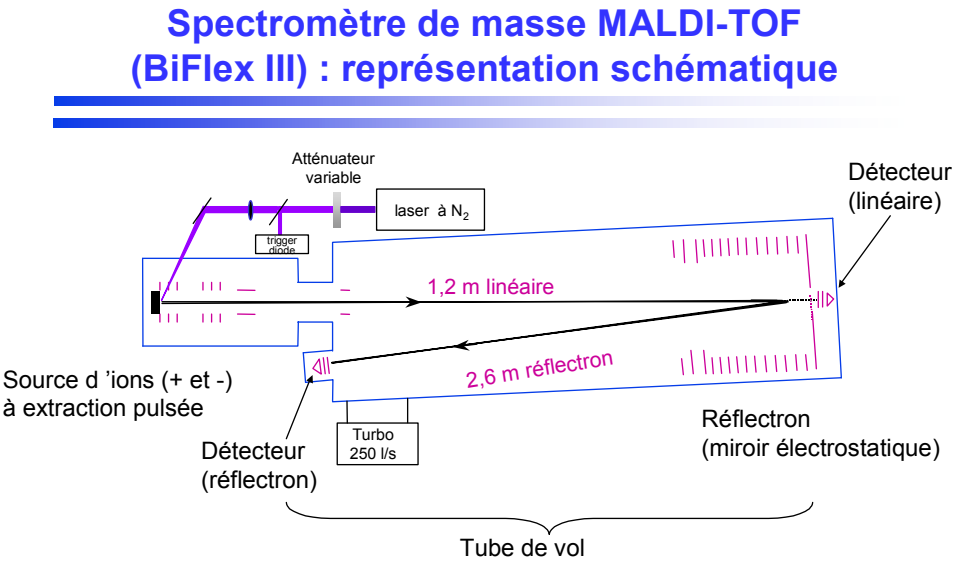


Schéma du principe de l'ablation, désorption et ionisation laser assistée par la matrice (MALDI)

Cette tension permet d'envoyer les ions chargés vers le détecteur étant donné que les ions sont chargés positivement (tension de même signe).
La masse sur la charge (m/z) est reliée au temps de vol. En effet, les peptides ayant une masse importante mettrons plus de temps à atteindre le détecteur que les peptides ayant une faible masse.

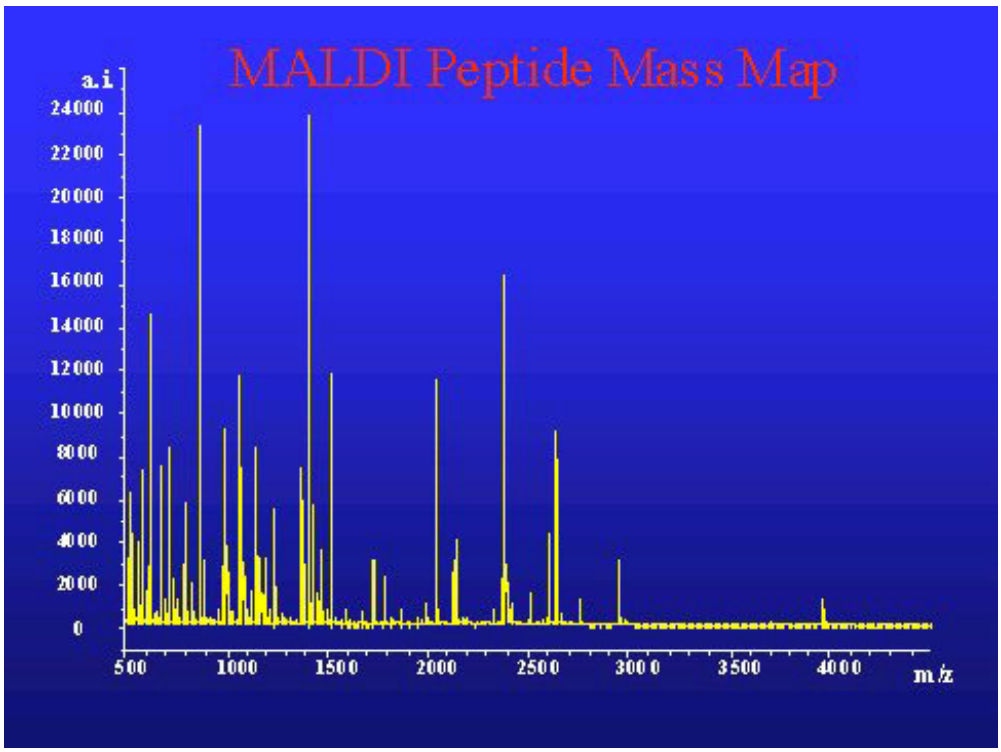


MALDI : Matrix Assisted Laser Desorption Ionization
TOF : Time Of Flight

Nicolas SOMMERER

Le schéma détaillé ci-dessus résume brièvement la manière de fonctionner du MALDI-TOF.

Le détecteur transmet l'information à un ordinateur et nous permet d'obtenir un spectre comme celui page suivante. Sous chaque pic se dissimule un peptide avec une masse bien précise.

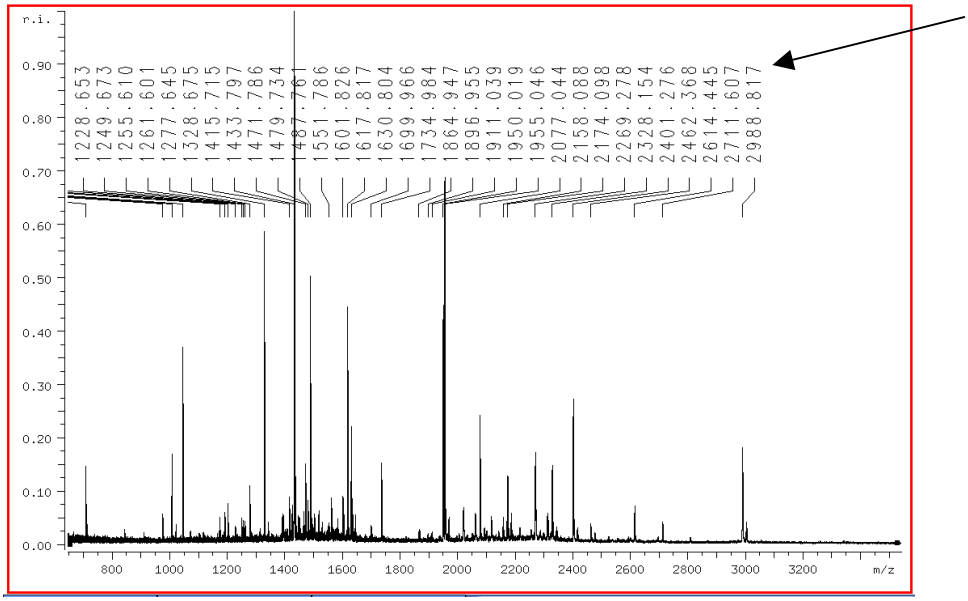


spectre MALDI-TOF-MS

V - Cinquième étape : retraitement

A Annotation

Les différents pics obtenus sont ensuite analysés et annotés de manière à obtenir la masse exacte des peptides en étalonnant les masses obtenues à l'aide de pics connus correspondant à l'auto-hydrolyse de la trypsine.



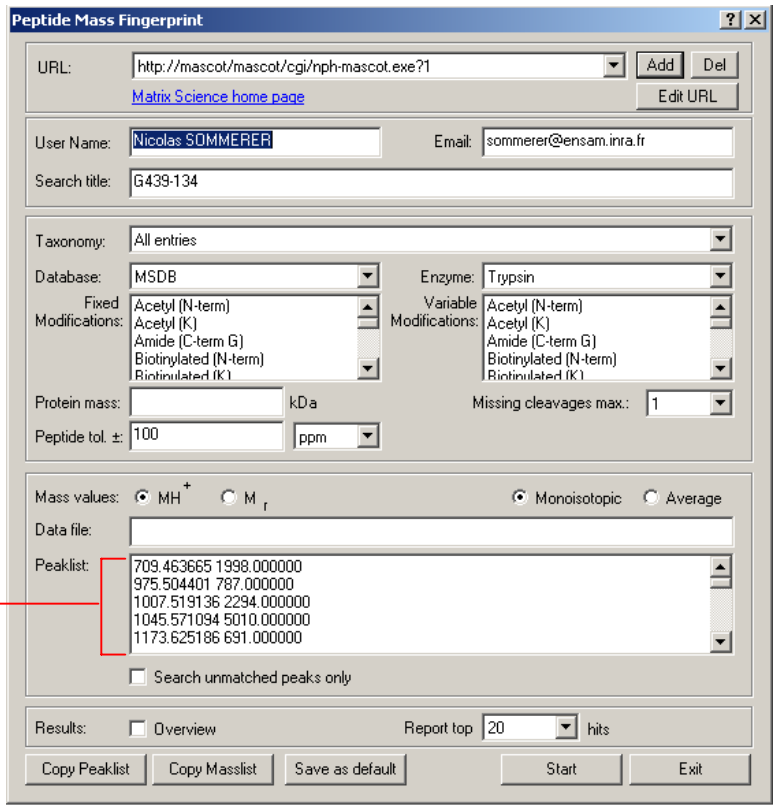
Masse de chaque peptide

Spectre annoté

Le but est ensuite de sélectionner tous les pics d'intérêts de notre spectre puisque sous chaque pic se dissimule un peptide avec une masse bien précise. C'est cet ensemble de masses de peptidiques qui va constituer une **empreinte peptidique massique** caractéristique de la protéine analysée.

B Recherche de la protéine dans les bases de données à l'aide du logiciel Mascot

La liste des masses sélectionnées va être entrée dans le logiciel Mascot d'interrogation de banque de donnée qui va nous permettre d'identifier la nature de la protéine analysée.



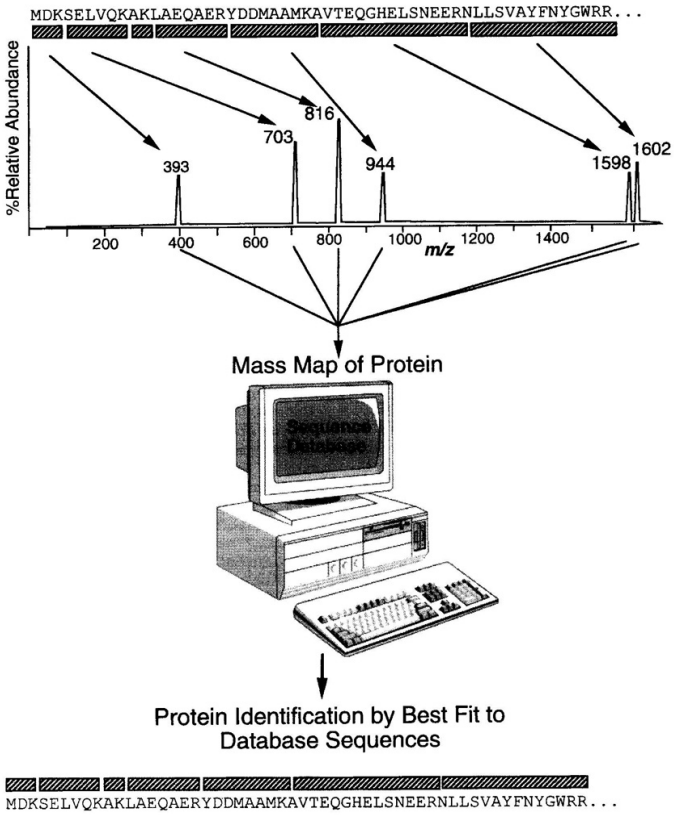
Liste des masses sélectionnée

L'interface d'interrogation du logiciel Mascot

Les banques de données (e.g NCBInr) interrogée par Mascot comprennent un grand nombre de protéines dont la séquence est parfaitement connue. Les informations associées à ces protéines correspondent notamment à l'ensemble des masses peptiques théoriques prédites après une digestion virtuelle par la trypsine rendue possible par le fait que cette enzyme, on l'a vu précédemment possède des sites de coupures très spécifiques (après la lysine ou l'argine).

Ainsi par comparaison l'empreinte peptidique massique expérimentale de notre protéine peut être comparé grâce à Mascot à l'ensemble des empreintes peptidiques massiques théoriques des protéines de la banque de données.

Le schéma ci-dessous résume brièvement la méthode d'analyse par Mascot :



C Résultats

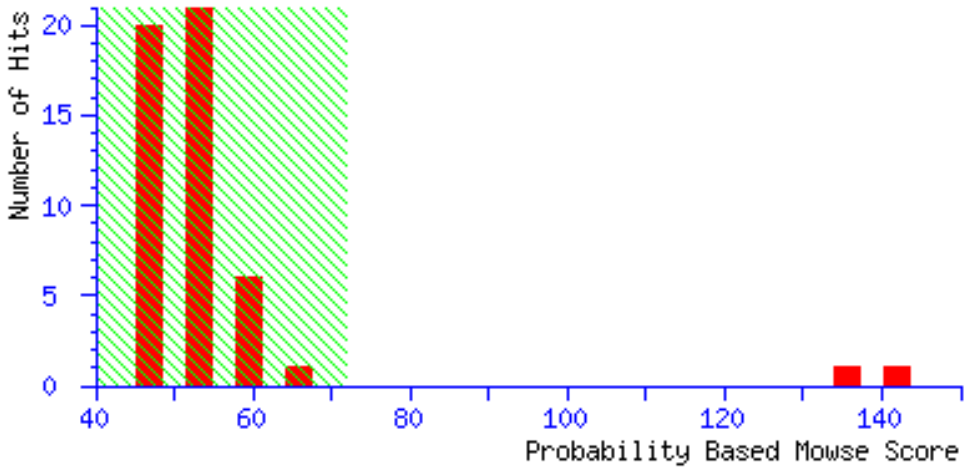
Les résultats donnés par Mascot se présentent sous forme d'une liste ou plusieurs protéines proposées auxquelles le logiciel attribue un score. Même si le plus souvent la protéine sélectionnée est celle ayant le score le plus élevée, c'est le biologiste qui décide de l'identité finale de la protéine analysée.

Ci-dessous est présenté pour exemple une fiche de résultats Mascot.
Dans ce cas, nous remarquons que la protéine ayant le meilleur score est une une aconitate hydratase numérotée 665.

Mascot Search Results

User : SOMMERER
Email : sommerer@ensam.inra.fr
Search title : 665
Database : MSDB 20020122 (852010 sequences; 265451329 residues)
Timestamp : 20 Jun 2002 at 12:42:23 GMT
Top Score : 142 for T04820, *aconitate hydratase (EC 4.2.1.3)*
F10M23.310 - *Arabidopsis thaliana*

Probabilité : Les scores au-dessus de 72 sont significatif.



Index

Accession	Mass	Score	Description
1. T04820	98891	142	aconitate hydratase (EC 4.2.1.3) F10M23.310 - Arabidopsis thaliana
2. Q94A28	108311	135	AT4G26970/F10M23_310.- Arabidopsis thaliana (Mouse-ear cress).
3. Q97AW0	42495	65	TVG0703837 PROTEIN.- Thermoplasma volcanium.
4. AAD28718	155010	59	AF112359 NID: - Schmidtea mediterranea
5. DNAK_HELPY	66880	58	Chaperone protein dnaK (Heat shock protein 70) Helicobacter p
6. E64533	67011	58	dnaK-type molecular chaperone - Helicobacter pylori (strain 26695)
7. AAF44560	34276	57	AF198100 NID: - Fowlpox virus
8. S73091	123038	57	hypothetical protein c0624 - Sulfolobus solfataricus
9. BAA01749	44001	57	SYORPOD1 NID: - Synechococcus sp.
10. I40351	25588	56	excinuclease ABC chain A - Brucella abortus (fragment)

Results List

1. T04820 **Mass:** 98891 **Score:** 142
aconitate hydratase (EC 4.2.1.3) F10M23.310 - Arabidopsis thaliana

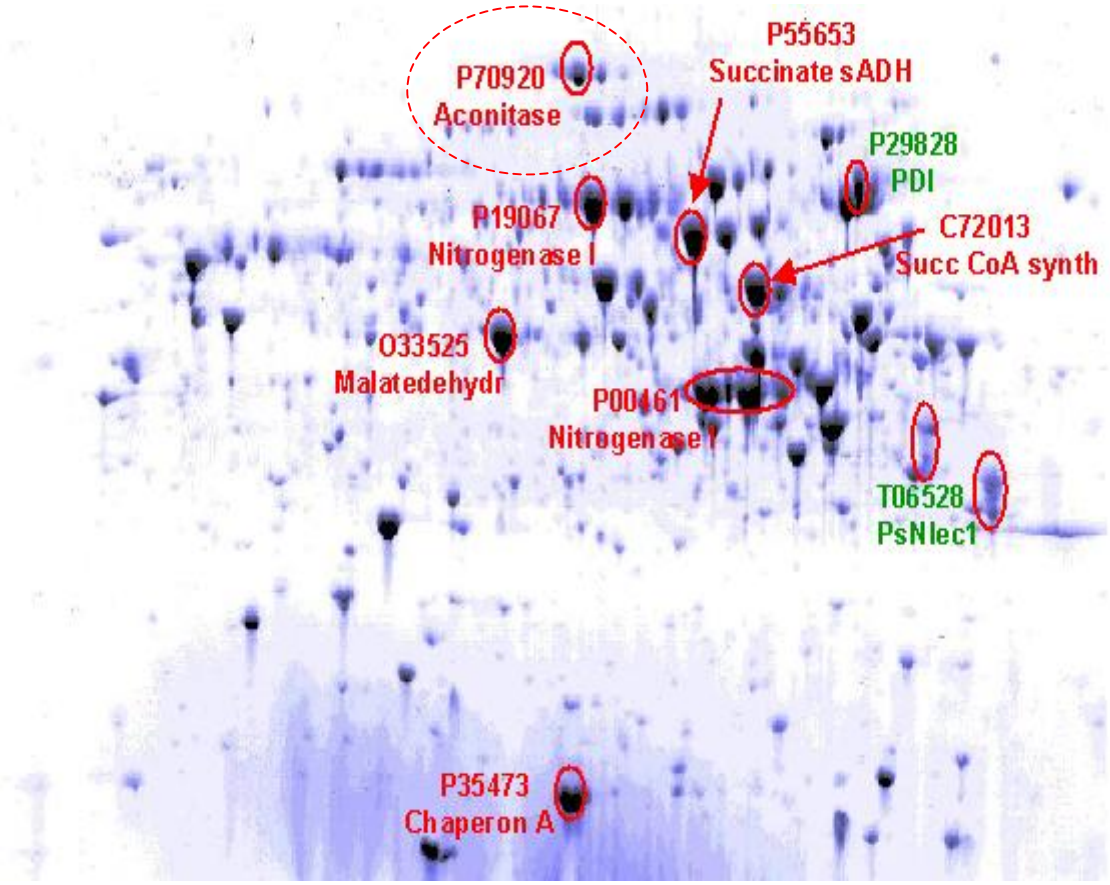
Observed	Mr(expt)	Mr(calc)	Delta	Start	End	Miss	Peptide
778.49	777.49	777.41	0.07	735 -	741	0	GTFANIR
914.55	913.55	913.56	-0.01	54 -	61	0	ILLESAIR
958.47	957.47	957.53	-0.06	849 -	856	0	YTVHLPTK
1074.61	1073.60	1073.62	-0.02	45 -	53	1	IDKLPFSVR
1158.66	1157.65	1157.66	-0.00	86 -	95	0	QVEIAFKPAR
1206.57	1205.56	1205.59	-0.03	76 -	85	0	ILDWENTSTK
1253.59	1252.59	1252.61	-0.02	428 -	438	0	FSYNGQPAEIK
1430.69	1429.68	1429.65	0.03	713 -	725	0	GVISEDFNSYGSR
1534.78	1533.77	1533.78	-0.01	749 -	763	0	GEVGPNTVHIPTGEK
1555.77	1554.77	1554.73	0.04	834 -	848	0	AGEDAETLGLTGHER
1829.07	1828.06	1828.03	0.03	266 -	282	0	EGVTATDLVLTVTQILR
1889.02	1888.02	1888.04	-0.03	746 -	763	1	LLKGEVGPNTVHIPTGEK
1957.01	1956.00	1956.13	-0.13	266 -	283	1	EGVTATDLVLTVTQILRK
1965.91	1964.91	1964.98	-0.08	422 -	438	1	QEEVVKFSYNGQPAEIK
2026.02	2025.01	2024.96	0.04	775 -	794	0	TAEQDTIILAGAEYGSGSSR
2070.26	2069.25	2069.21	0.04	264 -	282	1	LKEGVTATDLVLTVTQILR
2470.36	2469.35	2469.22	0.13	882 -	902	0	FDTEVELAYYDHGGILPYVIR

No match to: 712.39, 733.30, 832.34, 945.33, 1025.47, 1090.61, 1141.61, 1179.62, 1184.66, 1292.62, 1307.69, 1475.78, 1507.84, 1513.73, 1545.72, 1605.85, 1778.93, 1856.81, 1870.83, 1928.88, 2089.15, 2384.10, 2753.53, 2870.68, 3035.81, 3134.80, 3186.94

On peut observer sur la fiche le détail de la comparaison entre les masses expérimentales (**Mr(expt)**) des peptides de la protéine et les masses théoriques calculées (**Mr(calc)**) présente dans la base interrogée par le logiciel Mascot.

D Cartographie des gels

Quand l'ensemble des spots d'un gel a ainsi été analysés, la cartographie d'un gel de référence se présente comme sous la forme çi-dessous.
Le spot contenant la protéine Aconitase précédemment identifiée a été reporté sur notre gel et entourée en rouge pointillée.

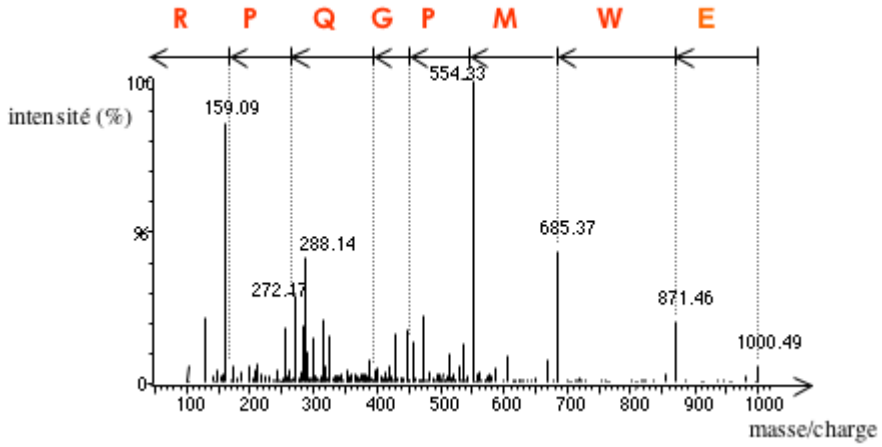


Cela permet par exemple aux biologistes d'établir la voie métabolique mise en place par l'organisme étudiée en comparaison avec d'autres gels établies dans d'autres conditions.

ANNEXE 4 : La spectrométrie de masse MS/MS

Dans la technique de spectrométrie de masse, dite « en tandem » (ou MS/MS) chacun des peptides obtenue par la technique de séparation par électrophorèse (**annexe 5**) est ensuite lui-même fragmenté, et ses produits de fragmentation analysés. Cette technique apporte des informations concernant la séquence de la protéine étudiée, car la fragmentation des peptides a lieu sur les liaisons chimiques entre les acides aminés et les « libère » de la chaîne peptidique, un par un ou en sous-fragments très courts.

Exemple de spectre MS/MS.



En abscisse, la masse des ions (plus exactement, le rapport masse/charge) ; en ordonnée, le pourcentage des ions possédant une masse donnée. En pratique, seul l'espacement entre les pics est interprété, pas leur hauteur. En interprétant ces espacements, il est possible de reconstituer la séquence peptidique. Sur cet exemple, la lecture du spectre de droite à gauche permet de reconstituer la séquence EWMPGQPR (chacun des 20 acides aminés est ici désigné par un code d'une lettre).
Les spectromètres de masse actuels permettent d'enregistrer, de manière entièrement automatique, plusieurs milliers de spectres MS/MS en quelques heures. C'est leur interprétation qui pose alors problème. Le rythme d'obtention de cette masse de données dépasse bien sûr largement celui d'une analyse « manuelle ». Quant au dépouillement automatique des spectres, il a encore beaucoup de progrès à faire.

ANNEXE 5 Présentation des unités de recherche en protéomique de Toulouse, Grenoble et Nantes et des sociétés Biogemma et Bayer CropScience

➢ UMR 5546 du pôle de biotechnologie Végétale, Agrobiopole de Toulouse :

Responsable : Rafael Pont-Lezica
Interlocuteur direct : Dr Elisabeth JAMET
Adresse :
UMR 5546 CNRS/Université P. Sabatier
Equipe Paroi et développement
Pole de Biotechnologie Vegetale
24 chemin de Borderouge
BP17 AUZEVILLE
31326 CASTANET TOLOSAN

Activités de recherches orientées protéomiques :

La partie protéomique des activités de recherche de cette équipe porte sur le protéome de la paroi végétale avec un programme de protéomique systématique (inventaire des protéines du compartiment pariétal) et un programme de protéomique fonctionnelle (en fonction de situations physiologiques). L'organisme retenu pour cette étude soutenue par Génoplante est là encore Arabidopsis thaliana. Le projet utilise les installations de la **plate-forme de protéomique de Toulouse** dont Bernard Monsarrat est le responsable.

Liens internet :
L'UMR 5546 : http://www.smcv.ups-tlse.fr/
Les thématiques scientifiques de l'UMR 5546 :
http://www.smcv.ups-tlse.fr/root/equipes/paroi/equipe.php
La plate-forme de protéomique de Toulouse :
http://genopole-toulouse.prd.fr/layout.php?page=home2&id=19&lang=fr

➢ Plate-forme protéomique de l'URPVI (Unité de Recherches sur les Protéines Végétales et leurs Interactions) de l'INRA de Nantes (URPVI)

Interlocuteur direct : Dominique Tessier
Adresse :
Unité de Recherche sur les Protéines végétales et leurs Interactions - INRA
rue de la géraudière, BP 1627 44316 Nantes Cedex 03
Objectifs scientifiques : Séparation et identification de protéines végétales
Moyens humains : 3 ingénieurs, 1 assistant-ingénieur, 1 technicien
Techniques :
- Système d'électrophorèse 2D
- Analyse d'images (Mélanie 3)
- Plateforme spectrométrie de masse : QTOF Ultima Global (Micromass/Waters) et Trappe d'ions LCQ Advantage (Thermofinnigan)

Projets de recherche :

- Etude de la composition protéique des graines de mutants d'*Arabidopsis thaliana* affectés dans la maturation de la graine avec pour objectif la caractérisation de gènes impliqués dans le contrôle de la qualité des graines chez *Arabidopsis thaliana* et le clonage de gènes homologues chez le colza.
- Etude du protéome de la graine de blé en voie de développement.

- Etude de familles de protéines par des méthodes bioinformatiques (identification, recherche et inférence de motifs, analyse des structures).

Lien internet : http://www.nantes.inra.fr/centre/unites-recherche/urpvi/index.html#proteomique

➢ **UMR 5019 UJF-CNRS-CEA Laboratoire de Physiologie Cellulaire Végétale (Interactions Plastes-Cytoplasme-Mitochondries)**

Directeur : Jacques JOYARD
Interlocuteurs direct : Daphné BERNY, Myriam FERRO, Norbert ROLLAND
Adresse :
UMR5168 Laboratoire de Physiologie Cellulaire Végétale CEA Grenoble
Bâtiment C2 17 rue des Martyrs 38054 GRENOBLE CEDEX 9

Thématique scientifique :
- Propriétés spécifiques des mitochondries végétales:
Le complexe de la glycine décarboxylase; Biosynthèse des cofacteurs associés; Protéome.
- Structure et fonctions de l'enveloppe des plastes: Identifications de nouveaux transporteurs; Biogenèse des plastes; enveloppe des apicoplastes de Plasmodium et Toxoplasma.
- Cytosquelette microtubulaire de la cellule végétale: Caractérisation de protéines impliquées dans la régulation de l'assemblage des microtubules au cours du cycle cellulaire. Métabolisme et Stress: Impact de carences et de stress environnementaux sur le métabolisme de la cellule végétale et de la plante entière (Ecophysiologie des plantes alpines); Structure, dynamique et contrôle du métabolisme.
Pour réaliser ses expériences en protéomiques ce laboratoire utilise les installations de la **plate-forme de protéomique du CEA de Grenoble** dont le responsable est Jérôme Garin.

Liens internet :
La plate-forme de protéomique du CEA de Grenoble :
http://www-dsv.cea.fr/content/cea/science_vivant/proteomique.htm
Les thématiques scientifiques du laboratoire :
http://www.ujf-grenoble.fr/BIO/biovege/3cycle.htm

➢ **BAYER CropScience**

Directeur du site d'Evry : Evelyne JAMES
Coordonnées :
1, rue Pierre Fontaine
91058 EVRY
Date d'installation : 01/07/2000
Thématique principale : Agro-industries.
Domaine d'activité sur Evry :
BioInformatique et génomique végétale.

Descriptif d'activité :
Bayer CropScience est l'un des leaders mondiaux de la protection et de la production des cultures, et développe également une activité de premier plan en hygiène et en santé publique.
Elle recherche, développe et commercialise des solutions innovantes répondant à la fois aux besoins des agriculteurs d'aujourd'hui - cultures saines, amélioration des rendements, des plantes et de la qualité de l'alimentation - et aux impératifs constants du respect de l'environnement, selon les principes du développement durable. Bayer CropScience conduit à Evry des recherches en BioInformatique pour l'ensemble des activités Bioscience du groupe Bayer CropScience.

Lien internet : http://www.bayercropscience.com

➢ **Biogemma**

Directeur du site :
Georges FREYSSINET
Coordonnées
2, rue Gaston Crémieux
CP 5707 - 91057 EVRY Cedex
Siège Social et administration
5, rue Saint Germain L'Auxerrois
75001 PARIS
Date de création : 01/07/1997
Date d'installation : 01/01/1999
Thématique principale : Agro-industrie - semences.
Domaine d'activité : Analyse des génomes de plantes de grande culture.
Descriptif d'activité :
La plate-forme de génomique végétale et de bioinformatique d'Evry a été créée en 1999 pour développer les technologies d'analyse des génomes, en particulier dans le domaine de la production des banques d'ADN et dans les technologies de mesure de l'expression des gènes (puces à ADN). Plusieurs puces sont disponibles sur les espèces suivantes : *Arabidopsis*, riz, maïs, blé et colza. En parallèle, une plate-forme de bioinformatique a été installée. Elle stocke les bases de données publiques et privées et développe différents logiciels d'exploitation et d'analyse. Cette base bioinformatique est accessible à tous les chercheurs de Biogemma et à ses actionnaires. Les activités de la plate-forme sont fortement intégrées dans les programmes de Génoplante.
Les résultats et les outils développés par la plateforme d'Evry et Génoplante doivent aider Biogemma et ses actionnaires à accélérer la mise au point de nouvelles variétés végétales. Par ailleurs, cette plate-forme est ouverte pour des collaborations avec des équipes intéressées par les technologies qui y sont mises en oeuvre.
Discipline : Biotechnologie végétale.
Thèmes de recherche : Génomique, biologie moléculaire et bioinformatique.
Applications/utilisation : Agriculture, semences et génome végétal.
Lien internet : http www.biogemma.com

ANNEXE 6 : GpiIS : un système de bases de données intégrées source : génoplante, biogemma

Génoplante programme:
BioInformatics
Project N°: Bi2001077

https://genoplante.infobiogen.fr
http://genoplante-info..infobiogen.fr
http://www.biogemma.com

Towards an integrated system around plant genomes

GnpIS: Towards an integrated system around Plant genomes

Delphine Samson[1], Fabrice Legeai[1], Farid Chetouani[1], Emmanuelle Karsenty[1], Jean-Baptiste Veyrieras[1], Stéphane Rouille[1], Delphine Grando[1], Erik Kimmel[1], Isabelle Luyten[1], Guillaume Albini[1], Sophie Creno[1], Frederic Sapet[1], Michael Alaux[1], David Mignon[1], Aymeric Duclert[1] (coordinator)
Virginie Chataigner[2], Déa Giardella[2], Sébastien Frade[8], Olivier Gigonzac[9], Guillaume Kerboul[2], Christophe Ketterlé[2], Marianne Liauzu[2], Christelle Pelletier[2], Eric Perche[8], Nicolas Sajot[2], Denis Scala[2], Bruce Thomas[2], Aurélie Defferrard[8], Evelyne James[8] (coordinator)
Séverine Gagnot[3], Joelle Amselem[1], Cédric Bouttes[5], Johann Joets[6], Marie-Henriette Flament[7,2]
[1]Génoplante-info (GPI) Evry, France, [2]Biogemma (BIB) EVRY, [3]INRA-CNRS URGV Evry, France, [5]ENSAM URP Montpellier, France [6]INRA UMR Génétique Végétale du Moulon, France, [7]INRA Clermont ASP, France, [8]Bayer CropScience (BCS) BioScience (BioInformatique) Evry, France, [9]Currently Aventis Pharma

Short Summary:
Génoplante-info (GPI) information system is a web based system composed of several applications (in Java and Perl) built above a relational database (Oracle) that includes integrated schemas for sequence data (EST, ,mRNA), map data (genetics maps, QTL maps), expression data, proteomic data and SNP data. The goal of this system is to set up a complete and integrated bioinformatic web environment for the analysis of genetic and genomic data.

Development in progress

1) GnpSeq:

Sequences database summary (EST+mRNA+FST for Gnp species of interest) :
2 456 466 (Total number of Sequences in Sept 04)

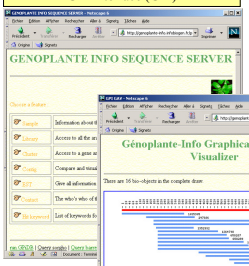
	Public	Génoplante	Gnp Public DbEST Submission (good quality sequences only)	Nbr of Contigs and Annotation V1	V2	V3	V4
Arabidopsis thaliana	240 578	1404	*1281*	29140	On GPI public site	/	/
Brassica napus	3 611	47 467	34 011	13 083	14257	/	/
Helianthus annuus	47124	21 828	12 886	6 632	21764	Autumn 2004	/
Medicago truncatula	198 397	/	/	36 935	On GPI public site	/	/
Oryza sativa	408 906	40 000	0	48 826	67183	Running…	/
Pisum sativum	913	6099	2 816	1 470	1 550	2157 (with 4345 Gnp EST)	/
Sorghum bicolor	84 777	/	/	23 965	On GPI public site	/	/
Triticum sp.*	615 506	111496	78 351	46 510	73 009	106860	/
Vitis vinifera	137 838	14 264	14 264	Autumn 2004	/	/	/
Zea mays	279 666	152 763	124 801	40 072	59 961	53033	58574 (On GPI public site) On private: running
Zinnia elegans	9 835	490	0	197	254	/	/

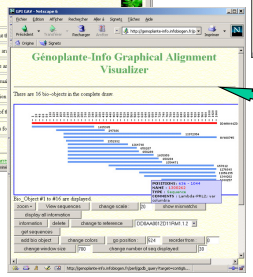
Data providers
EST, mRNA, genomic

Sequence Data Submission files (text)

Data insertion programs

GPI interface (GPI)





ADEL (GPI)

Project | Taxon | Sample | Library | Sequence | Contig | Hit | Keyword

2) GnpMap: Mapping database summary :

2) Working group: GPI, INRA Moulon, CIRAD, Rhobioinf/Biogemma, INRA Clermont, INRA URGV, INRA Versailles
Data Submitters: INRA Moulon, INRA Clermont, CIRAD/IRD

Species	Maps	Loci (framework+non framework+ projected)	Loci Genoplante	QTLs Genoplante
Zea mays	LRDG_031103	497	0	12 Maps
	LRE_031103	320	0	
	LRG_031103	220	0	
	IBM2_Neighbors	5733	0	+ 3 maps in progress
	IBM_031124	3419	2397	
	LHRE_020527	922	248	
	REFMAP_031127	9120	3232	
Wheat	ITMI_040618	268	…	*185 QTL/RER*
	RER_040618	243	…	*In progress*
	ARE_040618	194	…	
	CtCs_040630	683	…	Other: waiting for data…
	Deletion maps: in progress	*15729*	…	
Pea	RIL2_031013	202	0	Waiting…
Rice	EGRAM	364	0	
	RGP in progress	*2275*	0	Wainting…

Waiting for data :
-Rape maps, and QTL maps
-Sunflower maps, QTL maps
-Wheat QTL maps

Prospects:
-Physical maps integration (colza, rice, grape, wheat)
-Synteny data integration: maize/rice/wheat
- GnpMap publication in preparation
-Installation on GPI public si

ActionMap
INRA Moulon

Mapping
Data Submission files

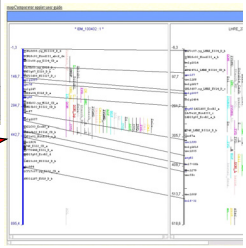
BioMercator
INRA Moulon
MC-QTL : meta-analysis tool

NOMENCLATURE GROUP
The working group has defined a controlled vocabulary for traits to define standards for QTL-related information.

GnpMap: text query interface

QTL AND MAP PROJECTION TOOL
A tool to allow to position on the same map informations originating from different maps, using the shared markers as anchor points

GnpMap: Map Comparator (GPI)



Project | Taxon | QTL | Marker | Locus | Map | Chromosom Bin, Bin deletion | Assignment

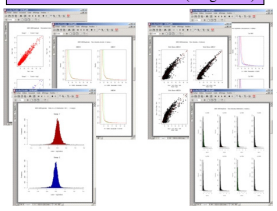
GnpDB, an integrated database

3) GnpArray

Work group:
Rhobioinf / Biogemma
Génoplante-Info/ INRA URGI
INRA Biocell/Versailles
Bayer CropScience

- Database schema : done
- Web interface: done
- Data insertion: in progress (Arabidopsis, maize, wheat, rape, rice), available in 2005 for Arabidopsis
- Perspectives: normalised values and gene list integration

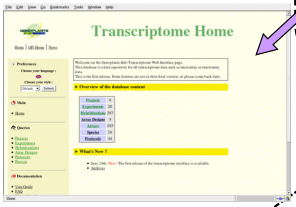
eNorm
Data control and normalization(Biogemma).



Platform Data Submission files (text)

Project | Taxon | Sample | Extract | Hybrid | Array | Spot | Expression

Query interface



Interfiltres
INRA Biologie cellulaire /Versailles

Integration with other Gnp modules: in progress
INRA Montpellier: Publication in preparation

4) GnpProt:(AFPDB)

INRA Montpellier / ENSAM
Coll. GPI / INRA URGI
INRA Nantes

Project | Taxon | Sample | Extract | Gel | Liquid Chromat | MS Analysis | Immuno Analysis | Spot | Protein

Query interface

Gel viewer application

Available expression data in GnpArray

Project Code	Platform submitter	GP 1 or 2
AF1999018	BIB	GP1
AF1999018	BIB	GP1
AF1999064	BIB	GP1
AF2001001	URGV	GP2
AF2001032	URGV	GP2
AF2001019031	URGV	GP2
CI2000005	INRA Versailles	GP1
NO2001027	URGV	GP2

Available proteomic data in GnpProt

Project Code	Submitter	Proteins/Peptides Lists
NO19993663	CEA Grenoble	53
NO19993663	INRA Montpellier	158
NO20011027	CEA Grenoble	163
NO20011027	UMR 5546 Toulouse	104

Other proteomic data in files

CP2P1	INRA Estrées-Mons	409 peptides (86 spots)
ZMS1P5	INRA Moulon	223

5) GnpSNP

Work group:
GPI/INRA URGI
INRAVersailles/CNG

Database diagram: done
GOPT4 data inserted into GnpSnp
Database interface in progress

Waiting for new maize and wheat data in GnpSnp format

Data summary

Project	Species	Submitter	Sequences	Others
Data inserted in GPiDB				
GOP-PEAA	Pea	INRA Versailles	15	-
GOP-PEAA	Pea	INRA Dijon	478	1972 features
ZMS3P1	Maize	INRA Lusignan	115	1568 features
SNP data already available				
GOP-T4	Rape	INRA Versailles	525	597 primers
ZMS3P1	Maize	INRA Moulon	34	-

6) OryzaTagLine (Rice mutant phenotypes)

CIRAD Montpellier
Coll: GPI / INRA URGI

Available on private plateforms sites
Integration with other Gnp modules: in preparation
Publication in preparation
Public version on OTL available on GPI public site

Dépôt APP (Genoplante Valor):
GpiDB Sequence: Deposit number IDDN.FR.001.21009.000.D.P.2002.000.10300
GnpMap , GnpArray, OryzaTagLine

Publication: GénoPlante-Info (GPI): a collection of databases and bioinformatics resources for plant genomics. Nucleic Acids Research, 2003, Vol.31, No.1, 179-182.

GPI Private Web site: https://genoplante.infobiogen.fr
GPI Public Web site: http://genoplante-info.infobiogen.fr
GPI Contact: genoplante-info@infobiogen.fr
Genoplante Web site: http://www.genoplante.com
Biogemma : http://www.biogemma.com
Bayer CropScience: http://www.bayercropscience.com

ANNEXE 7 : Les différents types de numéros d'accessions (ou accessions)

Un numéro d'accession est un identifiant unique qui caractérise un gène ou une protéine dans une base de données biologique.

Chaque base de données biologique utilise son propre format pour ses numéros d'accessions.

Pour illustrer ces différents formats on peut citer :

- ✓ Les banques de données GenBank/EMBL/DDBJ

  Pour les séquences protéiques l'accession est constitué de :

  2 lettres suivies de 6 chiffres **AY123456**

- ✓ La banque de données protéiques GenPept

  Banque qui contient les traductions des séquences GenBank/EMBL/DDBJ qui ont une séquence codante.

  L'accession est constitué de 3 lettres suivies de 5 chiffres **AAA12345**

- ✓ La banque de données protéiques PIR

  L'accession est constitué d'une lettre suivie de 5 chiffres **S30148**

- ✓ La banque de données protéiques SwissProt

  L'accession est constituée de :

  1 lettre soit O,P,Q

  1 chiffre

  3 caractères alphanumériques [A-Z,0-9]

  1 chiffre **P12345** ou **Q9JJS7**

  On trouve également les locus SwissProt

  4 lettres suivies _ARATH **ACOC_ARATH**

- ✓ La banque de données sur Arabidopsis thaliana TAIR

  At (Arabidopsis thaliana)

  numéro du chromosome : 1,2,3,4,5, M mitochondrial, C chloroplastique

  g (gène) puis 5 chiffres **At4g22690**

Les accessions de ce type sont des accessions AGI *(Arabidopsis Genome Identifier)* **[G15]** et sont spécifiques du génome d'Arabidopsis thaliana.

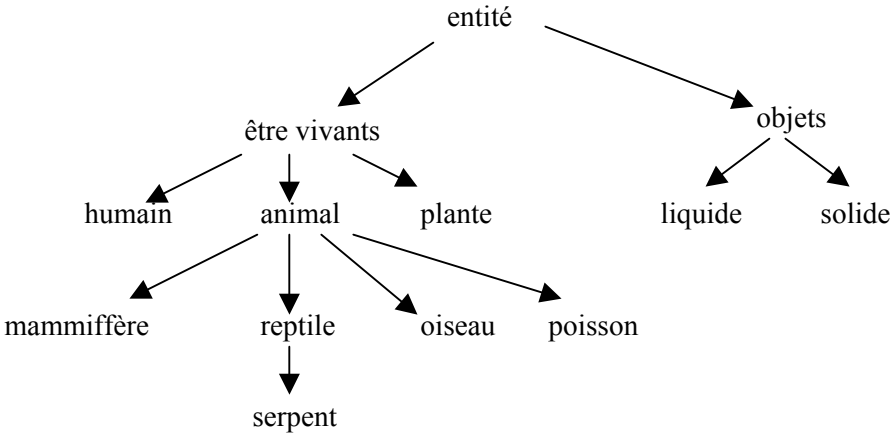
ANNEXE 8 : Présentation des ontologies

La définition d'une ontologie est la suivante : une ontologie est un catalogue sémantique, dont les descriptions sont à la fois concises, non ambiguës, et qui se doit d'être exploitable par un logiciel (description formelle) comme par un opérateur humain (description littéraire).

Les ontologies ont des structures hiérarchiques censées représenter les choses du monde. Il y a deux approches différentes:

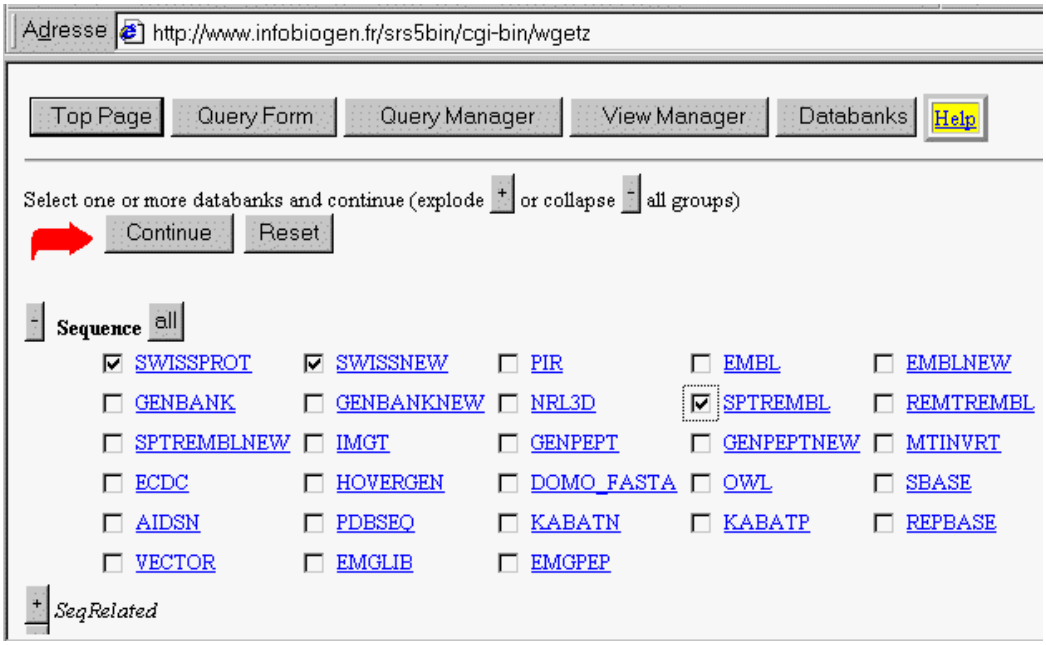
- Trouver une taxinomie des concepts les plus généraux. Cette ontologie de niveau supérieur vise à englober la totalité des concepts, sans pourtant entrer dans les détails des concepts spécifiques.(Guarino, 1997)
- Trouver une hiérarchie de concepts d'un domaine limité, par exemple de la législation, de la production de voitures, de la maladie de cancer... Cette ontologie de domaine peut contenir des représentations de concepts spécifiques très détaillées.

Il s'agit donc en fait d'unités de sens assez généraux qui servent à regrouper des termes voisins plus spécialisés qui ont la même signification ou dont le sens est proche. Par exemple sur la figure ci-dessous ont constate que les termes *homme*, *animal* et *plante* peuvent être rattaché à l'unité de sens *être vivant*.

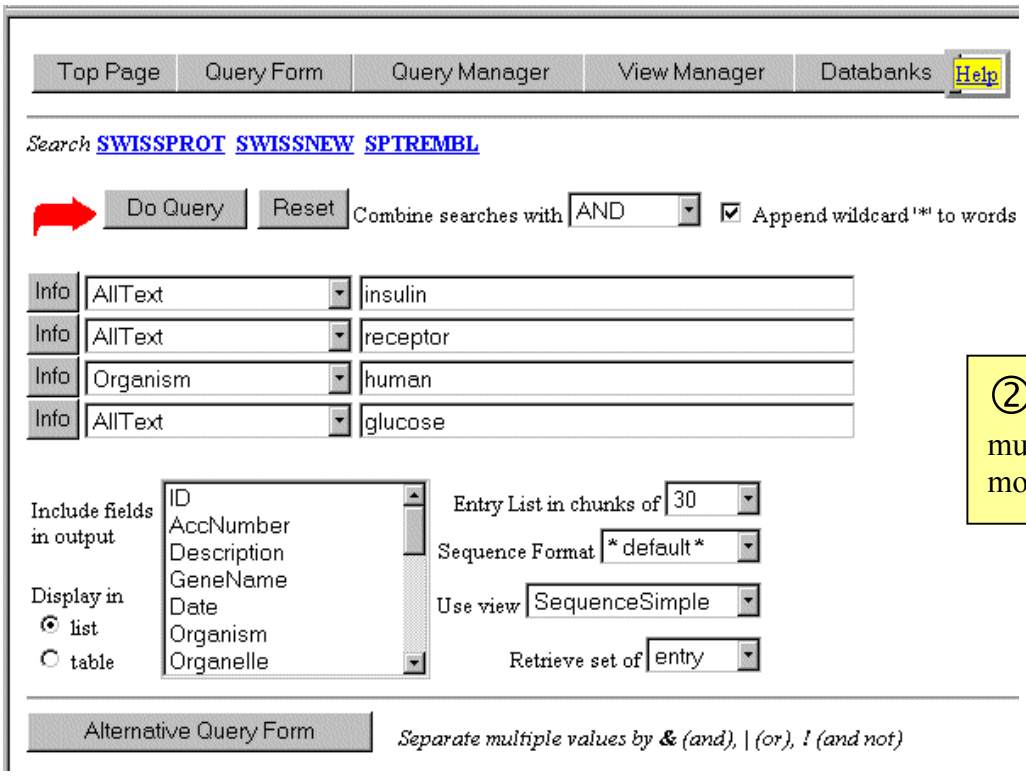


Exemple d'ontologie

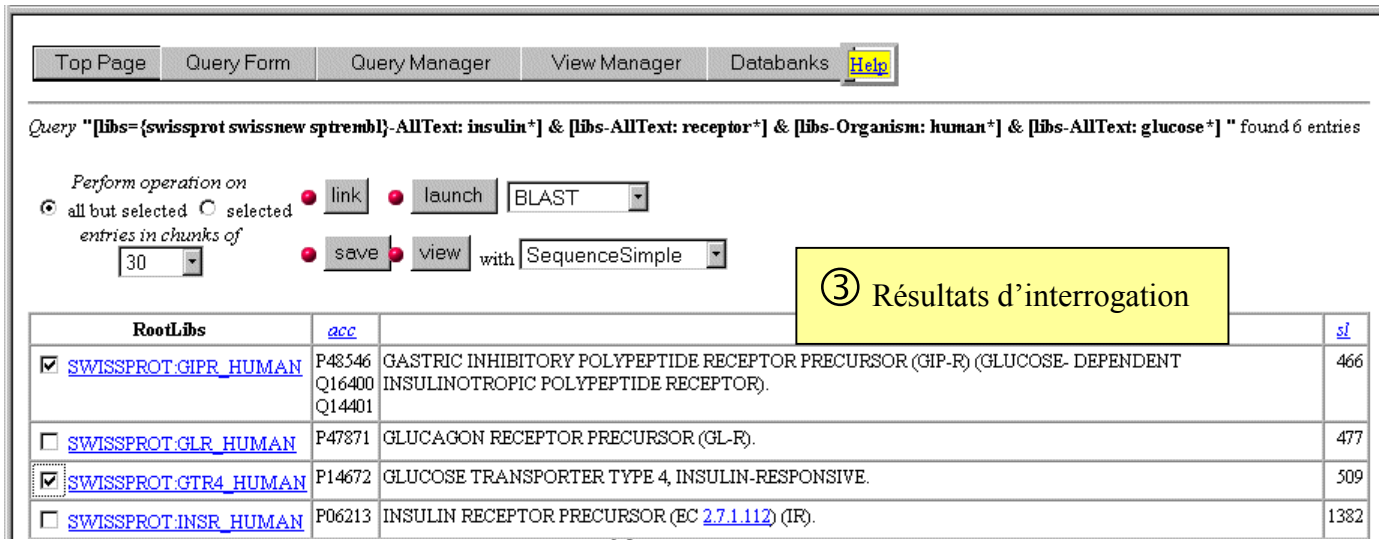
ANNEXE 9 : Exemple d'interrogation SRS sur INFOBIOGEN utilisant la banque SWISSPROT



① Sélection des bases à interroger



② Recherche multi-critères par mots clés



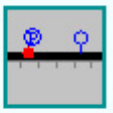
③ Résultats d'interrogation

ANNEXE 10 : Liste des masses des acides aminés

Nom	Lettre	Masse (en Dalton)
Alanine	A	71.037114
Cysteine	C	103.00919
Aspartic Acid	D	115.02694
Glutamic Acid	E	129.04259
Phenylalanine	F	147.06841
Glycine	G	57.021464
Histidine	H	137.05891
Isoleucine	I	113.08406
Lysine	K	128.09496
Leucine	L	113.08406
Methionine	M	131.04048
Asparagine	N	114.04293
Proline	P	97.052764
Glutamine	Q	128.05858
Arginine	R	156.10111
Serine	S	87.032029
Threonine	T	101.04768
Valine	V	99.068414
Tryptophan	W	186.07931
Tyrosine	Y	163.06333

ANNEXE 11 : Liens sur les banques de données sur Arabidopsis

- **Links for AGI accession number At3g23990 :**



AFPDB Protein Sequence Viewer (Under Construction) At3g23990

Liens hypertextes vers des banques de données spécialisées sur Arabidopsis

GO Annotations At3g23990

At3g23990

At3g23990

At3g23990

At3g23990

At3g23990

- **These links have possible no results for AGI accession number At3g23990 :**

At3g23990 *If FlagDB++ does not start automatically please read the Inst*

Search in Swiss-Prot and TrEMBL for : At3g23990

Search in UniProt (joining the information contained in Swiss-Prot, T

At3g23990

ANNEXE 12 : Liens sur l'application FlagDB++ de Génoplante à partir de ProteomIs

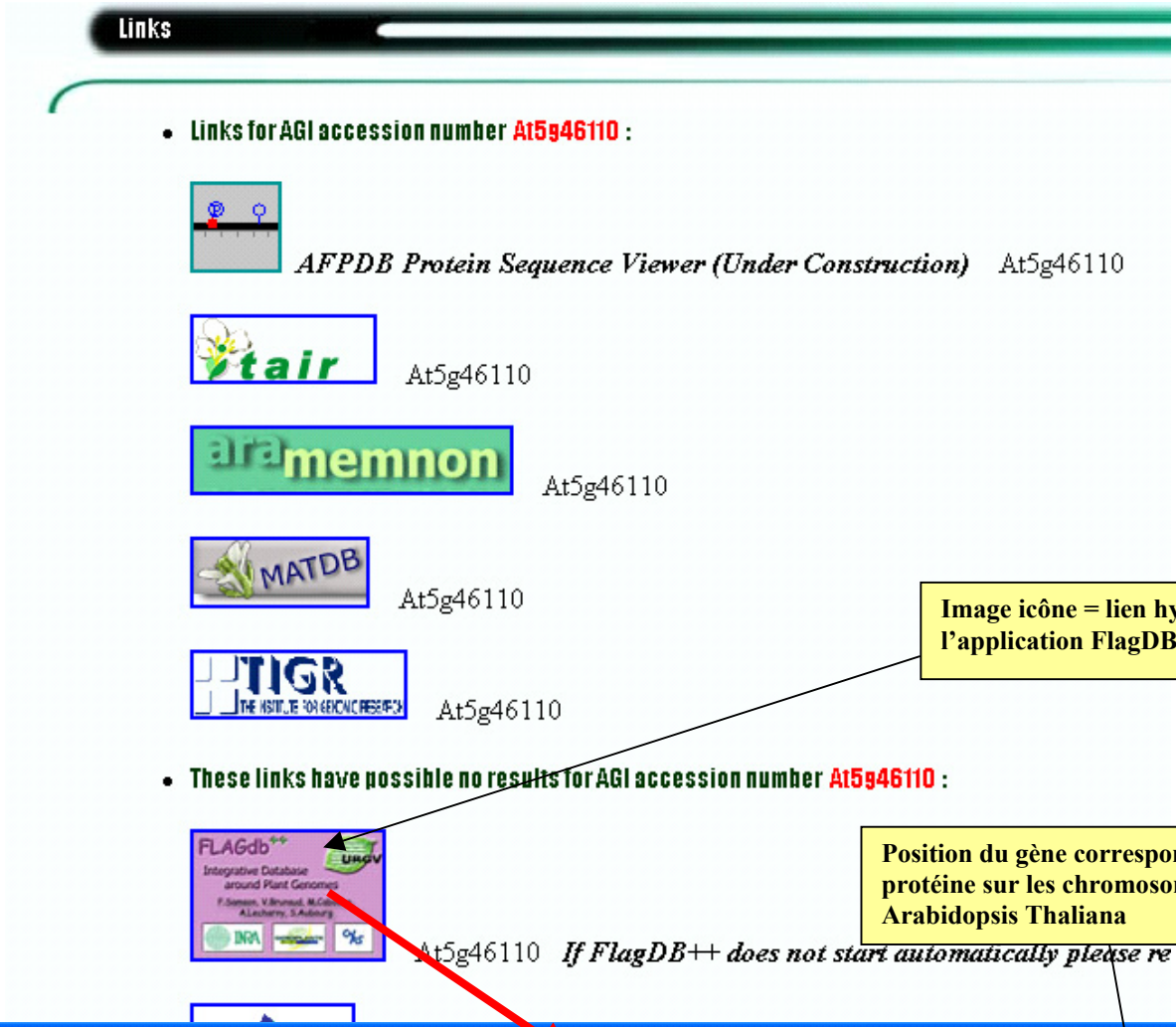
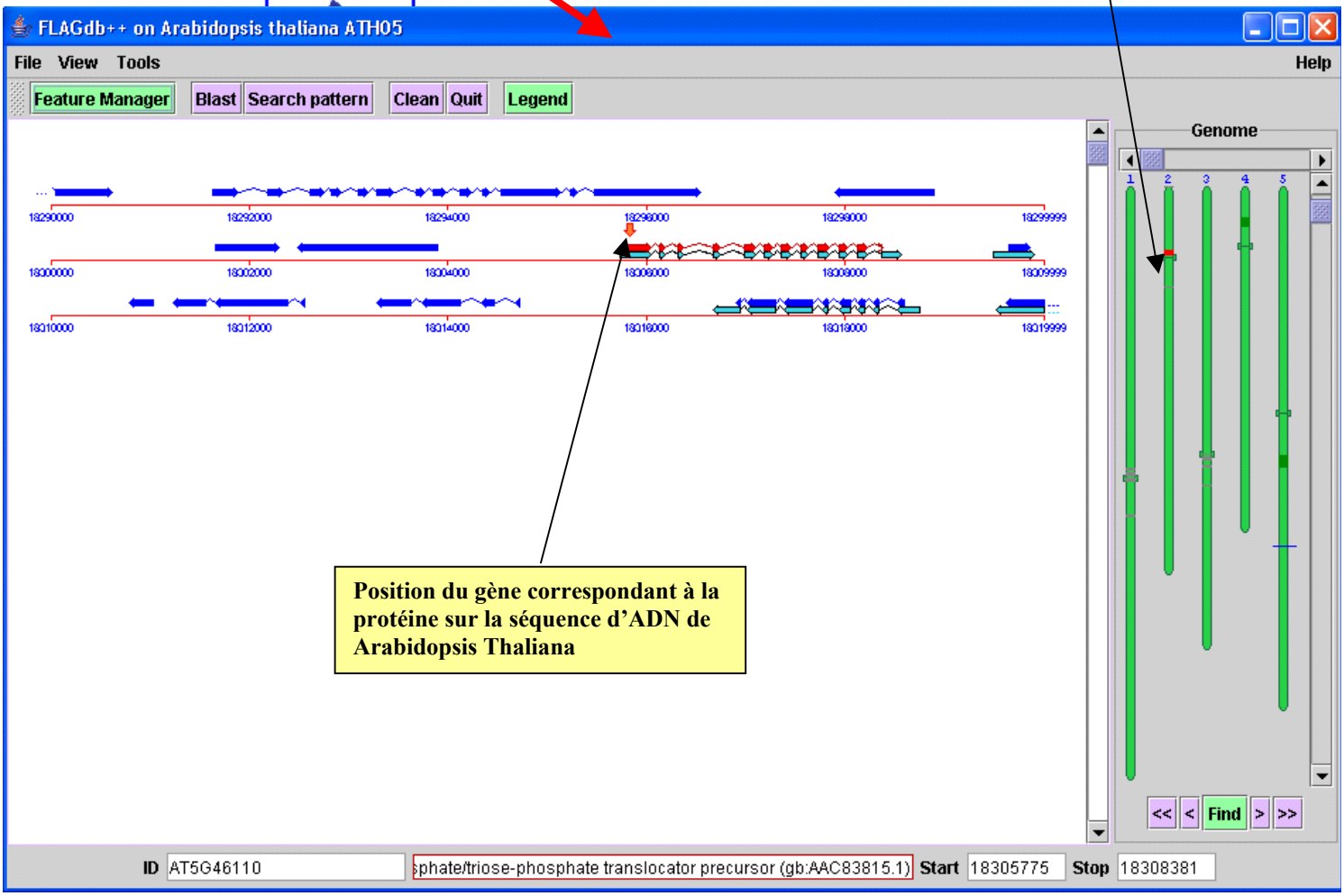


Image icône = lien hypertexte sur l'application FlagDB++

Position du gène correspondant à la protéine sur les chromosomes de Arabidopsis Thaliana



Position du gène correspondant à la protéine sur la séquence d'ADN de Arabidopsis Thaliana

ANNEXE 13 : Dictionnaire des données

Dans ce dictionnaire des données une liste déroulante permet d'accéder plus facilement aux données des différentes tables.

L'ensemble des tables est décrit sur 3 feuilles qui regroupent chacune un thème :

- Feuille *« échantillons, extraits »* : regroupe les tables concernant la gestion des échantillons, extraits, traitements associés …

- Feuille *« gels, LC, ms, prot »* : regroupe les tables concernant les résultats d'expériences (gels, LC), techniques d'identifications (spectrométrie de masse, immuno) et résultats d'interrogations (protéines) …

- Feuille *« projets, doc, publi »* : regroupe les tables concernant les données administratives (projets, thématiques, documents, publications …).

Liste déroulante

Gestion des gels, LC, MS, protéines identifiées

Sélectionnez l'entité que vous souhaitez visualiser

(Tous)
(10 premiers…)
(Personnalisé…)
ACC
BIOTYPE
DB
GEL
HARDWARE
IMAGE
IMMUNO_ANALY
LC
MS_ANALYSIS
MS_ANALYSIS_M
MS_ANALYSIS_R
PROTEIN
PROTEIN_CONTA
Protein_immuno
Protein_MSAnaly
PTM
SOFTWARE

Légende :
A=Automatique
S=Saisi
I=Importé

GEL — Gel 1D, 2D, Western (séparation électrophorétique)

Name	Definition	Example	Description	Structure	Table constraint	Source	Length	Field constraint
gelid				numeric	not null	A		integer>0
gel_name		ApexGel		alphanumeric	not null	S		
gel_date				date	not null	S		
short_remark	commentaires			alphanumeric		S		
protein_qty				alphanumeric		S		
ph_gradient				numeric		S		
acrylamid_percent				numeric		S		
coloration_type_id				numeric	not null	BIOTYPE.id		
gel_type_id				numeric	not null	BIOTYPE.id		integer>0
imagid				numeric		IMAGE.id		integer>0
thematicid				numeric	not null	PROJECT.id		integer>0
contactid				numeric	not null	CONTACT.id		integer>0
protocolid				numeric	not null	PROTOCOLE.id		integer>0

IMAGE — Image du gel

Name	Definition	Example	Description	Structure	Table constraint	Source	Length	Field constraint
imagid				numeric	not null	A		integer>0
file		1D ou 2D		binary	not null	I		
format_type_id	type de format	tiff, jpeg		alphanumeric	not null	BIOTYPE.id		integer>0
protocolid				numeric	not null	PROTOCOL.id		integer>0
softid				numeric	not null	SOFTWARE.id		integer>0

SOFTWARE — Software utilisé pour analyser le gel, ou pour l'interrogation

Name	Definition	Example	Description	Structure	Table constraint	Source	Length	Field constraint
software_id				numeric	not null	A		integer>0

ANNEXE 14 : Questionnaire

Fichier Edition Affichage Insertion Format Outils Données Fenêtre

	A	B	C	D	E	F
1						
2		**Pour avancer la réflexion, pouvez-vous, pour la liste ci-dessous :**		Grille proposée :		
3		**1)** classer les propositions par ordre de priorité pour votre projet *(cf* grille ci-contre)		**A = nécessaire**		
4		**2)** les commenter / amender		**B = souhaitable**		
5		**3)** compléter la liste		**C = *sans* préférence**		
6						
7	**I)**	**Recensement des propositions sur les types de données à archiver, selon enquête**			**Classement**	**Commentaires**
8						
9		**Données**	**Protocoles**	**Ex**		
10						
11		Nom de la protéine			A	
12		Accession			A	
13		locus			A	Quelle nomenclature est-elle employée ? Position sur les BAC ou At...
14		localisation sub-cellulaire			A	
15		visualisation des caractéristiques physico-chimiques		pI, PM, TMS, modifications post-traductionnelles, ...	A	Nombre d'amino-acides
16		conditions particulières d'expression		organe(s), stade de développement, réponse à stress, ...	A	
17		Auteur(s)				S'agit-il de la personne qui entre les données dans la base ?
18		Projet(s)			A	Intitulé général
19		Références associées		publis sur le(s) projet(s)		Cela concerne-t-il aussi des références existantes le cas échéant ?
20		date de création de la fiche			A	
21		date de dernière modification de la fiche			A	
22						
23		plante		espèce	A	
24		genotype		ecotype, cultivar, mutant, ...	A	
25		stade de développement			A	
26		organe/partie d'organe/tissu	protocoles prélèvement		A	
27		fraction sub-cellulaire	protocoles fractionnement	organite, membrane, paroi, ...	A	
28		résultats de caractérisation des fractions	protocoles caractérisation	western, ...	A	Champ non obligatoire
29						
30		conditions de culture	protocoles culture	hydroponique, suspension cellulaire, ...	A	
31		traitement particulier	protocoles traitement	carence, ...	A	
32						
33		obtention des protéines	protocoles extraction, fractionnement, purification	extraction solvant/micelles/NaOH, ...		
34				chromatographie d'affinité, exclusion, ...		
35						
36		quantification	protocole de quantification		C	Champ non obligatoire
37						
38		images de gels	protocoles de coloration	gels 1D/2D .tif	B	
39		coordonnées des spots/bandes		fichier .txt	B	Prévoir les analyses 1-D et 2-D
40		nomenclature des spots/bandes dans le gel		*cf* fichier .txt ci-dessus	B	
41						
42		données MS	type d'analyse	MALDI, LC-MS/MS	A	Prévoir PSD
43			protocoles de MS	matrice, débit, calibration, ...	B	
44			spectres MS	fichiers .dat	B	
45		interrogation	protocoles d'interrogation	Profound, Mascot, ...	B	
46			banque interrogée, version		B	
47			résultats/listes de pics	fichiers .html	C	
48						
49						
50		*autres données ... (préciser)*	*... / ...*	*... / ...*		Pour les protéines adressées à un compartiment, prévoir la possibilité
51						d'indiquer la taille du peptide d'adressage, le PM, le PI et le nombre
52						d'amino-acides de la protéine mature

ANNEXE 15 : Questionnaire adressé aux différents laboratoires partenaires pour la définition des grandes lignes du projet

QUESTIONS / REPONSES DU LABORATOIRE UMR 5019 de Grenoble (et de ses partenaires) :

1. Laboratoire / référence du projet Génoplante / responsable

Labo 1. Laboratoire de Chimie des protéines, ERIT-M INSERM/CEA, CEA-Grenoble, resp. J. Garin
Labo 2. Laboratoire de Physiologie cellulaire végétale, UMR 5019 CNRS/CEA/UJF, CEA-Grenoble, resp. J. Joyard
Labo 4. Institut des Sciences du végétal, UPR 2355 CNRS, Gif sur Yvette, Resp. H. Barbier-Brygoo
Labo 4. Laboratoire de KUNTZ, UMR 5019 CNRS/UJF, Grenoble, resp. M. Kuntz

2. Interlocuteur privilégié (pour les phases d'analyse / mise en place / utilisation de la banque de données) ?

Norbert Rolland & Myriam Ferro

3. Types de données protéomiques générées

gels 1D (tous labos) et 2D (labo 2)
cartes peptidiques massiques, spectres MS/MS, chromatogrammes associés à la LC-MS/MS, séquences peptidiques
liste de « gènes », localisation subcellulaire des protéines

4. Spectromètres utilisés

Maldi-Tof autoflex (Bruker)
Q-Tof 1 et Q-tof Ultima (Micromass)
Nano LC Ultimate (LC Packing)

5. Logiciels utilisés pour interrogations

Mascot, Proteinlinks Global server
Protein prospector, Blast P, Tblast N, BlastComp (développé en local)
HMMTOP, ChloroP, Predotar….

6. Format des données de sortie

Séquences peptidiques en données Fasta
liste de protéines sous format html
fichiers texte et Excel pour résultats validés

7. Volume total / débit / fréquence estimés

1 échantillon biologique différent par mois, soit entre 30 et 50 protéines/mois

8. Types de données expérimentales associées, et à archiver

images gels 1D (tous labos) et 2D (labo 2)
localisation subcellulaire des protéines
nécessité de pouvoir visualiser les caractéristiques physicochimiques des protéines

9. Types de données biologiques associées, et à archiver

nature du matériel biologique (fraction subcellulaire, type de membranes…)
caractérisation des fractions utilisées (pureté, western…)
traitements particuliers :

- conditions de culture du matériel biologique
- conditions de fractionnement des protéines (extraction chloroforme/méthanol, NaOH…, conditions de chromatographie…)

10. Autres types de donnée à archiver

Quantification (méthodologie associée)

11. De quoi jugez-vous nécessaire de disposer en local

Station(s) de travail sous Unix
Accès (sécurisé ou pas) à la base de donnée
Page de soumission on line

12. Desiderata en termes d'interfaces

Interfaçage avec les bases de données locales (base de donnée de la Génopole)

13. Liens souhaités

Swiss Prot, autres projets Génoplante, AMPL (J. Ward), etc…
Implémentation de FLAGDb avec la validation des protéines hypothétiques et intégration des informations de localisation subcellulaire

14. Critères d'interrogation de la base souhaités (tous types de données)

par protéine
par localisation subcellulaire
par fonction
par caractéristique physicochimique (TM, pI…)

15. Analyses bioinformatiques complémentaires prévues et fonctionnalités de retraitement souhaitées (classification de données d'expression, ...)

Création d'une banque de donnée protéique avec annotation spécifiées par les utilisateurs (pI, TM, Res/TM, localisation subcellulaire prédite…)

16. Recoupements bioinformatiques prévus avec d'autres données Génoplante (données transcriptomiques, ...)

Eventuellement pour la protéomique différentielle

17. Ressources souhaitées dans la page web

Cf. point 15

18. Environnement informatique disponible

windows/mac/unix

19. Présence d'une personne-ressource en informatique dans le labo

Non

ANNEXE 16 : Les initiatives de normalisation

La protéomique associée à des approches technologiques dites à haut débit (électrophorèse bi-dimensionnelle, spectrométrie de masse …) conduit à la production de données de nature diverses qui sont intrinsèquement liées au processus d'acquisition et de traitement (protocoles, conditions expérimentales, méthodes d'analyse…)

Si la gestion interne des données est du ressort des centres de production à travers des outils comme les LIMS, un autre aspect concerne la mise à disposition de ces données au sein de la communauté scientifique. Ainsi on l'a vu dans la partie précédente de ce mémoire de nombreuses bases de données existent pour les données de gels 2D et, dans une moindre mesure, de spectrométrie de masse. Néanmoins, la constitution de ces bases de données ne répond que partiellement aux besoins liés à une logique de dissémination et d'échange des données de protéomique, dans la mesure où ce n'est pas leur vocation première.

De ce fait, la question se pose aujourd'hui de définir des standards de compatibilité dans une logique d'intégration, de collecte, de publication ou d'utilisation par des tiers **[a3] [a4]**. La relative jeunesse de la protéomique et son évolution constante rendent particulièrement difficile la définition de données clés dans la somme de résultats produits. Par exemple, selon les outils de production et les différents logiciels utilisés, la diversité des formats et la manière dont les données sont structurées rendent difficiles la comparaison et l'échange des données produites.

Par ailleurs, tout comme les études du transcriptome, ce domaine d'activité produit des données qui sont interprétables en fonction du contexte qui a permis de les générer.

Une représentation standard des méthodes utilisées et des données générées par les expériences de protéomique, analogue au projet MIAME[1] a déjà été proposée. C'est le modèle PEDRo **[a1]** *(Proteomics Experiment Data Repository)*. Cette représentation basée sur un formalisme UML permet de représenter les connaissances du domaine (**document 1**) des données expérimentales en protéomique (génération et traitement d'échantillons biologiques, expériences de spectrométrie de masse, résultatsd'analyse *in silico*). Un diagramme de classe a été défini pour fournir un modèle conceptuel servant de base à une implémentation sous forme de relationnel et de schéma XML *(PEML : Proteomic Experiment Markup Language)* et relationnel **[a2]** visant à standardiser le vocabulaire utilisé.

Cependant si le modèle PEDRo fournit un modèle de schéma standard pour la conception d'une base de donnée complète en protéomique, il ne fournit pas la définition d'un format standard pour l'échange de ces données. C'est l'objectif du projet MIAPE *(Minimum Information About a Proteomics Experiment)* que de spécifier l'ensemble des informations minimales qui doivent accompagner tout jeu de données protéomiques, en particulier lors d'une publication, afin qu'il puisse être analysé ou ré analysé dans un contexte différent de son contexte d'obtention. MIAPE est l'une des initiatives de normes de HUPO-PSI **[a3]** *(HUman Proteome Organization - Proteomics Standard Initiative)* et dérive du travail sur PEDRo.

---

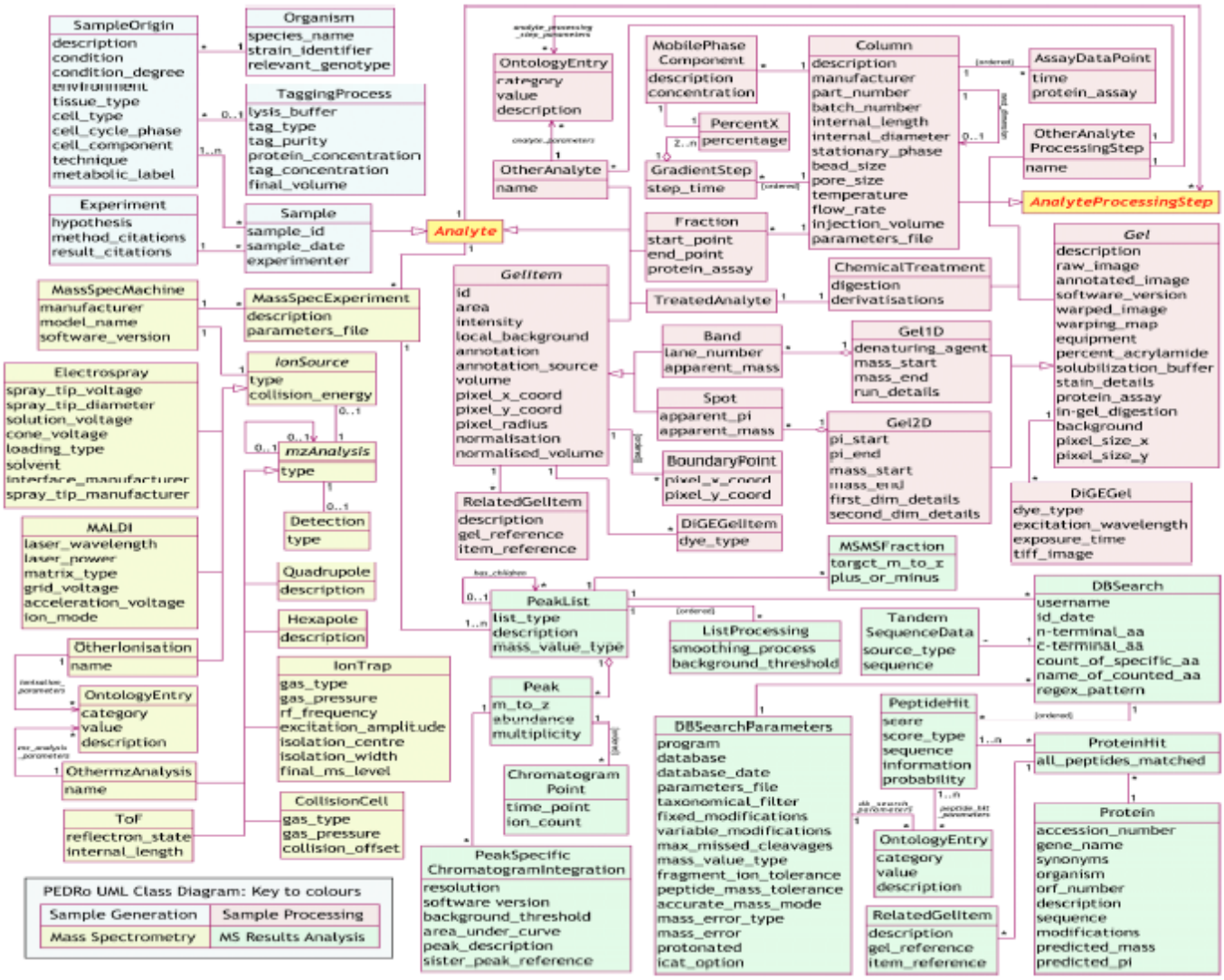
[1]MIAME (Minimal Information About Microarray Experiment) (Brazma et al, 2001) propose une description des informations minimales sur une expérience dans le domaine du transcriptome. Cette norme décrit l'information nécessaire pour s'assurer que les données ainsi stockées puissent être facilement interprétées et que les résultats puissent être indépendamment vérifiés.

Une nouvelle version standard de modèle objet pour la protéomique est également en cours d'élaboration au sein du groupe PSI : c'est le modèle PSI-OM *(PSI – Object Model)* **[i1]**. Pour préciser la sémantique de ce modèle, il faut fournir une documentation explicite qui accompagne ce modèle : c'est le PSI-ML *(PSI-Markup Language)* qui implémente en XML le modèle PSI-OM.

Au niveau des acteurs français de la protéomique la proposition MIAPE a été considérée par l'ensemble des acteurs français de la protéomique lors d'une journée Inter-Génopoles le 01/06/2004 **[i2]**. L'avis général est qu'il est nécessaire que la communauté nationale s'implique dans son élaboration et son affinement.

Il a été proposé lors de cette journée de constituer un groupe de travail sur cette problématique des formats d'échange afin d'être en mesure de participer de façon crédible et active au travail du groupe qui élabore le standard MIAPE.

Je participe moi-même à ce groupe de travail. A terme ProteomIs devrait être compatible avec les normes du MIAPE. Pour que le schéma de ProteomIs soit compatible il faudra que toutes les types de données spécifiées dans le modèle UML de ProteomIs soient contenues dans celles du MIAPE. En effet un des objectif de ProteomIs étant de favoriser l'échange des résultats de données protéomiques au sein de la communauté scientifique, cet aspect de standardisation reste essentiel.



<u>Document 1 :</u> Le modèle standard objet Pedro [a9]

Références :

[i1] Journée "Quelle informatique et bioinformatique pour la protéomique ?", Grenoble, 1er juin 2004 http://www-helix.inrialpes.fr/article615.html#15

[i2] Le modèle PSI-OM : http://psidev.sourceforge.net/gps/misc/PSI-OM_0.12.gif

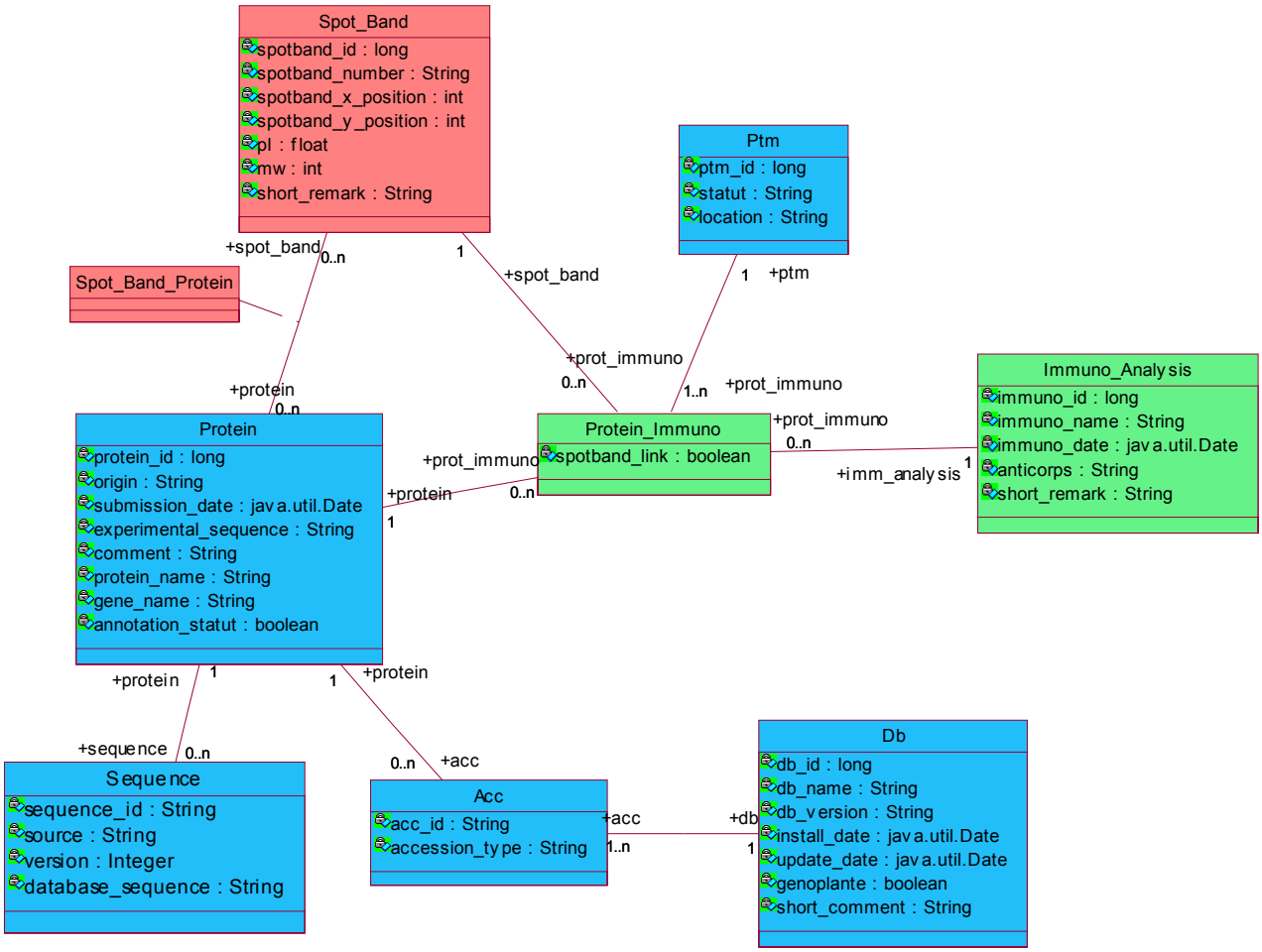
[a9] A systematic approach to modelling, capturing, and disseminating proteomics experimental data Chris F. Taylor[1, 2], Norman W. Paton[2], Kevin L. Garwood[2] Nature Biotechnology **21**, 247 - 254 (2003)

[a10] PEDRo: a database for storing, searching and disseminating experimental proteomics data. BMC Genomics. 2004 Sep 17;5(1):68. Garwood K, McLaughlin T, Garwood C PMID: 15377392

[a3] The HUPO PSI's molecular interaction format - a community standard for the represnetation of protein interaction data. Hermjakob, H. et al. Nat. Biotechnol. 22 : 177-183

[a4] Improved tools for biological sequence comparison Pearson, W. R. and Lipman, D. J. (1988). Proc. Natl. Acad. Sci. USA **85**2444-2448

ANNEXE 17 : Diagramme de classe « identification des protéines par immuno-détection »



ANNEXE 18 : Le Modèle Logique de Données (MLD)

Modèle Logique de Données :

➢ **Paquetage « Echantillons et extraits »**

treatment(**#treatment_id**,treatment_name,treatment_type,time_elapsed_val,compound,measurement,description,pk_treatment);

taxon(**#taxon_id**,scientific_name,taxon_id_ncbi,url_tax_ncbi,url_tax_usda,common_name,domain,kingdom,phylum,division,tax_order,family,genus,species,synonyms,short_remark);

dev_stage(**#dev_stage_id**,dev_stage_name,short_remark);

sample_source
(**#sample_source_id**,sample_source_name,cultivar,growth_media,growth_condition,type_media,genotype,short_remark,***#contact_id,#project_id,#taxon_id***);

sample **(#sample_id**,sample_name,harvest_date,age,short_remark**,*#sample_source_id***);

dev_stage_sample **(#sample_id,#dev_stage_id**);

exp_condition **(#exp_condition_id**,exp_condition_name,short_remark);

exp_cond_sample_source **(#exp_condition_id,#sample_source_id**);

exp_cond_treatment **(#treatment_id**,#exp_condition_id,treatment_date);

extract (**#extract_id**,extract_name,extract_date,short_remark,***#protocol_id,#molecule_type_id***);

sample_extract (**#sample_id,#extract_id**);

sample_treated (**treatment_id,sample_id**,treatment_date);

tissue (**#tissue_id**,tissue_name,short_remark);

tissue_sample (**#sample_id,#tissue_id**);

➢ **Paquetage « Séparation des protéines »**

image (**#image_id**,file,***#protocol_id,#format_type_id***);

gel
(**#gel_id**,gel_name,gel_date,protein_qty,ph_gradient,acrylamid_percent,short_remark,***#thematic_id,#contact_id,#protocol_id,#bio_type_id,#coloration_type_id,#image_id***);

gel_extract (***#extract_id,#gel_id***);

immuno_analysis **(#immuno_id**,immuno_name,immuno_date,anticorps,
short_remark,**#contact_id,#protocol_id,#bio_type_id, #image_id**);

lc (**#lc_id**,lc_name,lc_date,short_remark,***#thematic_id,#contact_id,#protocol_id,#bio_type_id***);

lc_extract (***#lc_id,#extract_id***);

spot_band
(**#spotband_id**,spotband_numeric,spotband_x_position,spotband_y_position,pi,mw,short_remark,***#gel_id***);

spot_attribute (**#spot_attr_id**,value,short_remark,***#spotband_id,#bio_type_id***);

spot_band_protein (**#spotband_id,#protein_id**);

➢ **Paquetage « Identification des protéines »**

db(**#db_id**,db_name,db_version,install_date,date,update_date,date,genoplante,short_comment);

protein
(**#protein_id**,protein_name,gene_name,annotation_statut,origin,submission_date,experimental_sequence,
comment,***#taxon_id,#subcellular_location_typid***);

acc (**#acc_id,#protein_id,#db_id**,accession_type);

ms_analysis
(**#ms_analysis_id**,ms_analysis_name,ms_analysis_date,ms_analysis_file,short_remark,***#lc_id,#spotband_id,#contact_id,#protocol_id,#bio_type_id***);

ms_analysis_result
(**#ms_analysis_result_id**,ms_analysis_res_date,ms_analysis_res_file,ko,description,***contact_id,protocol_id***);

ms_analysis_ms_result ***(#ms_analysis_id,#ms_analysis_result_id***);

protein_contact (**#protein_id,#contact_id**,last_modif_date);

ptm (**#ptm_id**,statut,location,***#bio_type_id,#predicted_subcell_loc_typid,#software_id***);

protein_immuno (**#spotband_id,#immuno_id,#protein_id,#ptm_id**,spotband_link);

protein_msanalysisresult (**#ms_analysis_result_id,#protein_id,#ptm_id**,spotband_link);

sequence (**#sequence_id**,source,version,database_sequence,***#protein_id***);

➢ **Paquetage « Protocoles »**

bio_type(**#bio_type_id**,name,description);

defined_type (**#defined_type_id**,name,description);

protocol_description (**#protocol_desc_id**,description);

software(**#software_id**,software_name,version,description,***#bio_type_id***);

hardware (**#hardware_id**,hardware_name,model,make,description,***#bio_type_id***);

protocol
(**#protocol_id**,protocol_name,***#protocol_desc_id,#defined_type_id,#software_id,#hardware_id***);

molecule_type (**#molecule_type_id**,molecule_type_name,description,strand);

➢ **Paquetage « Données administratives »**

contact
(**#contact_id**,last_name,first_name,fax,phone,email,institution,department,laboratory,address,city,state,zip_code,country,status,genoplante,short_remark);

project
(**#project_id**,project_code,project_name,status,title,genoplante,creation_date,closure_date,short_remark);

thematic **(#thematic_id**,title,thematic_date,description,***#project_id***);

documents **(#documents_id**,title,file,short_remark,***#thematic_id,#contact_id,#bio_type_id***);

project_bioinfo (**#project_id,#contact_id**);

project_coordinator (**#project_id,#contact_id**);

project_partner (***#project_id,#contact_id***);

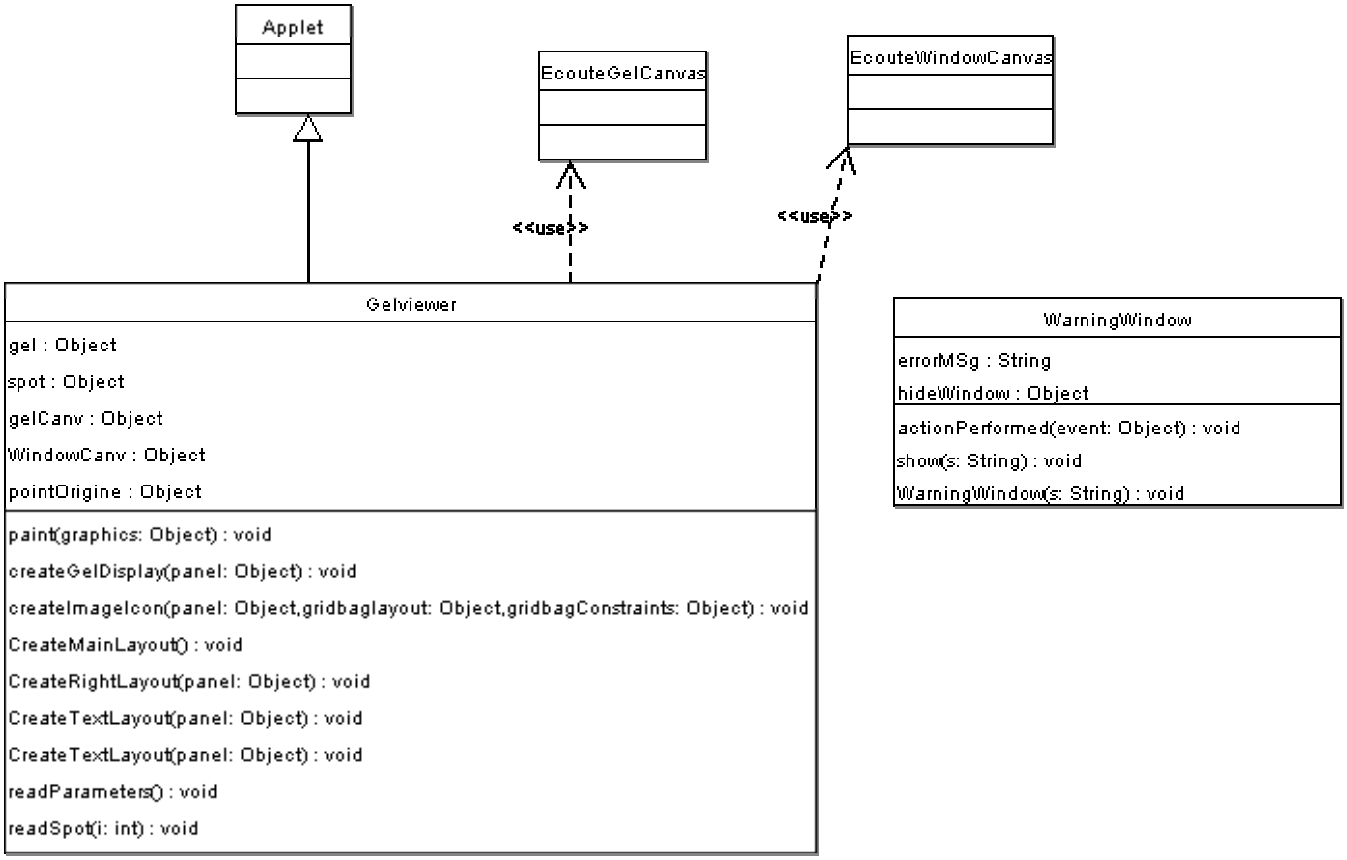
reference
(**#ref_id**,title,first_author,medline,pubmed,pages,publication_date,journal,submitted,volume,subvolume,keyword,abstract,reference_file);

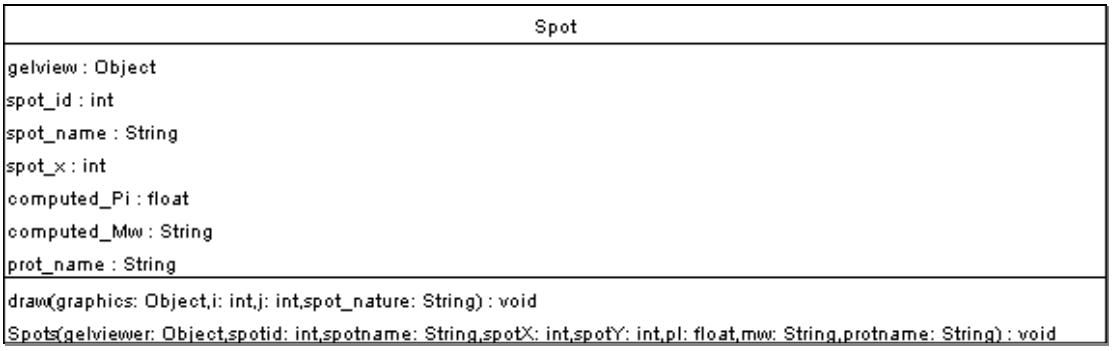
ref_gel (**#gel_id,#ref_id**);

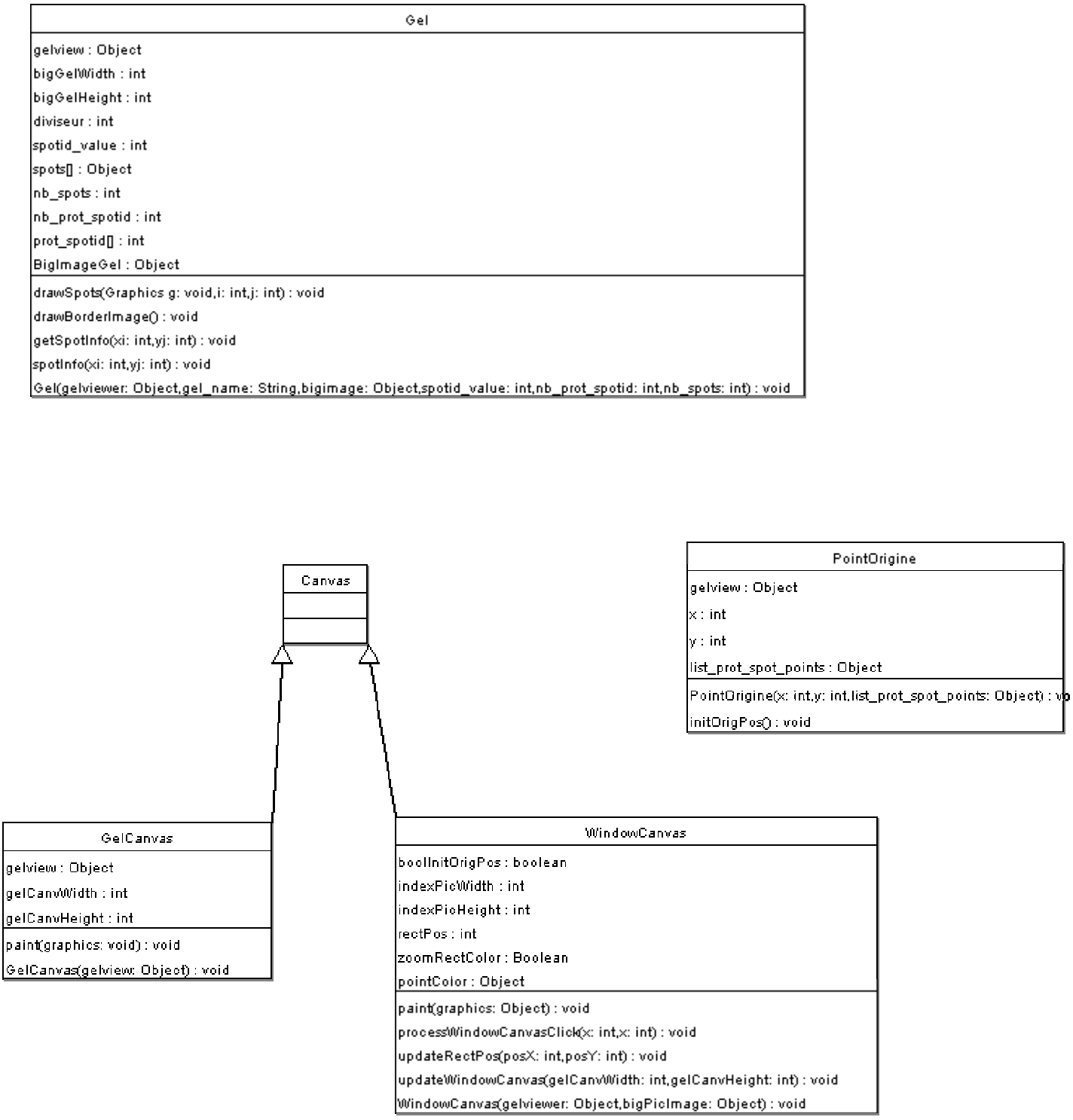
ref_lc (**#lc_id,#ref_id**);

thematic_contact (**#thematic_id,#contact_id**);

ANNEXE 19 : Diagramme de classe de l'applet de visualisation de gels






Gel
gelview : Object
bigGelWidth : int
bigGelHeight : int
diviseur : int
spotid_value : int
spots[] : Object
nb_spots : int
nb_prot_spotid : int
prot_spotid[] : int
BigImageGel : Object
drawSpots(Graphics g: void,i: int,j: int) : void
drawBorderImage() : void
getSpotInfo(xi: int,yj: int) : void
spotInfo(xi: int,yj: int) : void
Gel(gelviewer: Object,gel_name: String,bigimage: Object,spotid_value: int,nb_prot_spotid: int,nb_spots: int) : void
Canvas
PointOrigine
gelview : Object
x : int
y : int
list_prot_spot_points : Object
PointOrigine(x: int,y: int,list_prot_spot_points: Object) : vo
initOrigPos() : void
GelCanvas
gelview : Object
gelCanvWidth : int
gelCanvHeight : int
paint(graphics: void) : void
GelCanvas(gelview: Object) : void
WindowCanvas
boolInitOrigPos : boolean
indexPicWidth : int
indexPicHeight : int
rectPos : int
zoomRectColor : Boolean
pointColor : Object
paint(graphics: Object) : void
processWindowCanvasClick(x: int,x: int) : void
updateRectPos(posX: int,posY: int) : void
updateWindowCanvas(gelCanvWidth: int,gelCanvHeight: int) : void
WindowCanvas(gelviewer: Object,bigPicImage: Object) : void

ANNEXE 20 : Liste de référence sur des frameworks MVC2

Barracuda

Ce framework qui se veut très complet est développé par la
communauté Open Source Enhydra http://www.enhydra.org/
Il comporte :

- Un modèle événementiel complet,
- Un modèle et une librairie de composants MVC pour l'interface utilisateur
- Un système de validation des formulaires et de mappings avec des objets Java.
- Un contrôleur MVC2
- Une librairie Javascript.

Hammock

Ce framework n'existe plus. Il permettait de développer une application web de la même façon qu'une application Swing. Pour cela, Hammock utilisait le même modèle : système d'événements identique, présence de composants IHM 'Panel', et de layout 'FlowLayout ' ou 'TableLayout' etc. Ce système fonctionnait sans JSP. Signalons qu'il n'était pas Open Source mais développé par OOP.

Tapestry

Ce framework est axé sur les composants : la création d'une page se fait par le développement de composants qui vont définir cette page. Il propose des concepts intéressants comme la séparation de l'état de la page et de la page elle-même afin de disposer d'un pool de pages ou encore des possibilités d'internationalisation via des templates. Tapestry facilite la gestion des erreurs et de la charge.
Tapestry a d'abord été développé par Primix et est maintenant disponible en Open source.
Lien internet : http://jakarta.apache.org/tapestry/

Webwork

Ce framework a été délaissé au profit de Tapestry.
Il utilisait le concept du Pull Hierarchical Model View Controller. Il était similaire techniquement au framework Struts dont la présentation suit, mais son API est plus réduite.
Il pouvait être utilisé avec d'autres technologies que les servlets.

Struts

Struts est un projet Open Source développé par la communauté Jakarta d'Apache. Il a débuté en mai 2000 sous la direction de Craig R Mc Clanahan, qui participe également au développement de Tomcat.
Aujourd'hui, Struts est géré par plusieurs committers. Sa mailing-list comporte un millier de personnes. C'est un projet très actif.
Lien internet : http://jakarta.apache.org/struts/

ANNEXE 21 : Comparaison Servlet/JSP

Nous comparons ici un programme écrit à l'aide d'une Servlet et le même programme écrit à l'aide d'une JSP

Exemple de programmation à l'aide de Servlet :

Voici sur le **document 1** un exemple simple de servlet dont le seul but est de compter le nombre de fois que l'on y accède et d'afficher le résultats sur le navigateur du client.

```
/*Servlet SimpleCounter.java */
import java.io.*;
import javax.servlet.*;
import javax.servlet.http.*;

public class SimpleCounter extends HttpServlet {

    int count = 0;

    protected void doGet(HttpServletRequest request, HttpServletResponse re
    throws ServletException, IOException {

        response.setContentType("text/html");
        PrintWriter out = response.getWriter();

        /*incrémentation du compteur*/
        count++;

        /* impression de la page HTML ici */
        out.println("<html>");
        out.println("<head>");
        out.println("<title>Exemple de compteur à l'aide d'une Servet</title>
        out.println("</head>");
        out.println("<body>");

        out.println("Valeur du compteur lue depuis la servlet : "+ count);

        out.println("</body>");
        out.println("</html>");
        out.close();
    }
}
```

Document 1 : Servlet SimpleCounter.java

Ce processus est possible car la Servlet instanciée une première fois reste résidente en mémoire et ainsi la valeur du compteur peut être incrémentée.

Le problème avec cette approche est que la construction de la page doit être faite par la Servlet, ce qui veut dire qu'un infographiste qui voudrait modifier l'apparence de la page renvoyée devrait modifier le code Java de la Servlet et la recompiler. En effet, pour génerer le code HTML avec une Servlet, il faut utiliser la méthode *println*. De plus ce code permettant de mettre en forme le document est mélangé au code réservé aux traitements (*count*++) ce qui ne facilite pas la maintenance des programmes.

Exemple de programmation à l'aide de Java Server Pages :

L'utilisation des Java Server Pages est une réponse au problème précédent tout en conservant les avantages propres à l'utilisation des Servlets. Voici sur le **document 2** le même programme écrit sous la forme d'une JSP utilisant un JavaBean.

```
<%@page contentType="text/html"%>

<html>
<head><title>Exemple de compteur à l'aide d'une JSP</title></head>
<body>

<%--Instanciation du bean Compteur avec l'identifiant 'compteur'--%>
<jsp:useBean id="compteur" scope="session" class="chapitre2.Compteur" />
<p>

<%--Lecture de la propriété 'compteur' du bean --%>
<%--à l'aide de l'action standard jsp:getProperty--%>
<jsp:getProperty name="compteur"  property="compteur" /> --%>

</p>
</body>
</html>
```

```
/*JavaBean Compteur.java */

public class Compteur {

    int compteur = 0;

    public int getCompteur() {
        compteur++;
        return compteur;
    }

    public void setCompteur(int compteur) {
        this.compteur=compteur;
    }

}
```

Document 2 : JSP *compteur.jsp* qui compte et affiche dans un navigateur le nombre de fois que l'on y a accédé et utilisant le bean Compteur : *Compteur.java*

Le bean *Compteur.java* possède une seule propriété, de type *int*, nommée *compteur*, qui compte le nombre d'accès à la propriété. Il contient également les accesseurs nécessaires pour utiliser cette propriété. Ensuite la JSP *compteur.jsp* utilise une action standard <*jsp :useBean*> qui recherche une instance du JavaBean *Compteur* et la crée en cas de besoin, avec la portée (*scope=session* **[G7]**) et l'identificateur indiquée (*id=compteur*). La valeur de l'attribut id est la clé associée à l'instance de l'objet dans la portée spécifiée. L'action standard <jsp :getProperty> lit ensuite la valeur de la propriété getCompteur du JavaBean, la convertit en une chaîne de caractère et la renvoie au client. Ainsi à travers cet exemple on a vérifié l'avantage, par rapport à la solution Servlet précédente, qui est que par l'utilisation des actions standards JSP et de JavaBean il est clairement possible de séparer le code Java du code HTML servant à créer la page JSP. Le code Java peut ainsi être réutilisé par d'autres développeurs. Ensuite cela facilite aussi la maintenance mais permet également la spécialisation lors d'un développement : des designers, graphistes pouvant se concentrer sur la partie graphique tandis que les développeurs, intégrateurs se travaillent sur la couche traitement avec le développement de JavaBean.

ANNEXE 22 : ORM utilisé par Génoplante et développé par la société SYSRA

Réalisation d'un mappage objet-relationnel

Ce paquetage contient un ensemble de classes permettant de réaliser un mapping objet relationnel.

Classes principales :

DbObject

Chaque classe faisant partie d'un mappage objet relationnel doit étendre DbObject.
On procède par exemple comme suit avec la table PERSON :

SQL :

```
Create table PERSON
(
name varchar(32) PRIMARY KEY
)
```

Classe java :

```
/** Squelette de classe incomplet */
public classe Person extends DbObject
      public String getName()
      private String name;
}
```

DbFactory :

Cette classe abstraite permet en l'étendant de réaliser des fabriques pour des instances de classes faisant partie d'un mappage objet/relationnel.
Elle fournit notamment un cache et des méthodes pour gérer ce cache.

Si on reprend l'exemple précédent :

```
/** Squelette de classe incomplet */
public class PersonFactory extends DbFactory
{
      public static Person get (DbKey PersonId);
      public static Person[] getPred(String Predicate);
}
```

Les methods suivantes doivent être implémentées par les fabriques concrètes :
- get() : permet de récupérer une instance de la classe désirée en donnant une clé primaire.
- getPred() : permet de récupérer les instances de la classe désirée vérifiant un prédicat.

DbKey :

Une clé primaire ou secondaire dans une table.
Pour des raisons de commodités, les clés sont conservées sous forme de chaînes de caractères.

MultiValueKey :

Une clé primaire ou secondaire sur plusieurs colonnes dans une table.

Classes secondaires :

Cache :

Cette classe permet de réaliser des caches d'objets.
Ces objets doivent implémenter la classe Cacheable pour pouvoir être conservée dans le cache.
Si vous réalisez un mappage objet/relationnel, vous n'avez pas besoin d'utiliser cette classe directement.
Les classes DbFactory et DbObject s'en chargent déjà. Néanmoins cette classe est suffisamment générique pour être utilisée dans un autre contexte.

Cacheable :

Voir ci-dessus.

JdbcConnectionPool :

Cette classe permet de réaliser un pool de connexions à une base de données.

Classes générées :

Pour chaque table, 3 classes Java sont générées :

- une classe représentant le contenu d'une table. Une instance de cette classe correspond à une ligne de la table.
- Une classe fabrique : comme son nom l'indique elle sert à construire des instances de la classe précédente. Ceci est réalisé à l'aide de requêtes SQL.
- Une classe Collection contenant un ensemble d'instances de la première classe.

Ainsi pour la table PERSON sont générées les classes Java Person, PersonFactory et PersonCollection :

Exemple :

```
Create table PERSON
(
      name varchar(32) PRIMARY KEY
)
```

La classe :

```
Public class Person extends DbObject
{
      public String getName();
      private String name;
}
```

A noter que cette classe hérite de DbObject;
Les instances de cette classe sont immutables.

La Fabrique :

```
Public class PersonFactory extends DbFactory
{
      public static Person get (DbKey PersonId);
      public static PersonCollection getPred(String Predicate);
}
```

A noter que cette classe hérite de la classe DbFactory.
Elle contient un cache d'objets dont la taille doit être fixée pour chaque fabrique avec la méthode initCache().
Les méthodes get() et getPred() retournent des instances de Person correspondant a une clé de la table PERSON ou un predicat. Ces instances peuvent provenir du cache ou être crées si elles n'existaient pas.

Précautions : la génération des classes permet d'automatiser 90 % de la réalisation du mappage. En effet les fabriques peuvent se révéler incorrectes suivant le cas.
En particulier :

- le nom des tables relationnelles utilisées dans les requêtes peuvent être différents
- les clés primaires multiples ne sont pas prises en compte
- le type des clés primaires est considéré comme étant un entier

ANNEXE 23 : Liste des feuilles du format d'échange

Feuille Gel :
Description :
Contient toutes les informations relatives à un gel notamment ses extraits d'origine, les publications relatives et l'email du préparateur.
Fichier d'accompagnement : fournir l'image du gel au format jpg ou gif.

Feuille Spot :
Description :
Contient la liste des spots identifiés pour un ou plusieurs gel.

Feuille Contact :
Description :
Regroupe toutes les personnes participant à un projet et étant susceptibles d'avoir participé à la réalisation d'un gel, rédigé un protocole, réalisé l'analyse, réalisé l'interrogation pour un spot, validé l'annotation du spot ….

Feuille Thématique :
Description :
Regroupe les différents thèmes d'études mise en place au sein des différents projets

Feuille DocThematic :
Description :
Cette feuille recense les différents documents à archiver dans la base qui concerne directement une thématique autre que les images de gel, les fichiers protocoles et de spectro qui sont archivés par ailleurs. Cela peut être par exemple : un cluster, une cinétique
Fichier d'accompagnement : fournir le document. On autorise tout fichier qui peut être ouvert par l'intermédiaire d'un navigateur (exemple : document word, pdf, texte ou image au format gif ou jpg …).

Feuille Protocole :
Description :
Regroupe tous les protocoles associés aux LC, Gel, Analyse, Interrogations, Image, Extract, Sample.
Fichier d'accompagnement :
Fournir le fichier expliquant le protocole utilisé.

Feuille Taxon :
Description :
Contient la liste des Taxon utilisés dans la base.
Tables alimentées :

Feuille Sample_Source :
Description :
C'est la liste des échantillons d'origine à partir desquels on a obtenu un ou plusieurs extraits

Feuille Extract_Sample :
Description :
Ceci est la liste des extraits de tissu obtenu à partir d'un échantillon d'origine

Feuille Treatment :
Description :
Contient la liste des traitements effectués sur un échantillon d'origine (Sample_Source).

Feuille References :
Description :

Contient toutes les publications qui concernent directement un ou plusieurs gels ou expérience de LC.

Feuille LC :
Description :
Contient toutes les informations relatives à une expérience de Chromatographie en phase Liquide (LC, HPLC) notamment ses extraits d'origine, les publications relatives et l'email du préparateur.

Feuille Immuno_analyse :
Description :
Ceci est la liste des analyses immuno avec le spot d'origine (identifié par spot_number-gel_name) ou l'expérience de LC relative.

Feuille Immuno_Protein :
Description :
Ceci est la liste des protéines identifiées par analyse immuno. Il y a en plus l'information sur le niveau de maturation de la protéine au moment de son identification.

Feuille MS_analyse :
Description :
Ceci est la liste des analyses spectro avec le spot d'origine (identifié par spot_number-gel_name) ou l'expérience de LC relative.

Feuille MS_analysis_result :
Description :
Ceci est la liste des interrogations de spectro avec la manip de spectro d'origine.

Feuille Maturation_Protein :
Description :
Ceci est la liste des protéines identifiées avec le ou les fichiers d'interrogations sur lequels elles sont identifiées. Il y a en plus l'information sur le niveau de maturation de la protéine au moment de son identification.

Feuille Protein :
Description :
Contient la liste des protéines identifiées par analyse de spectro ou immuno suite à un gel ou une expérience de LC.

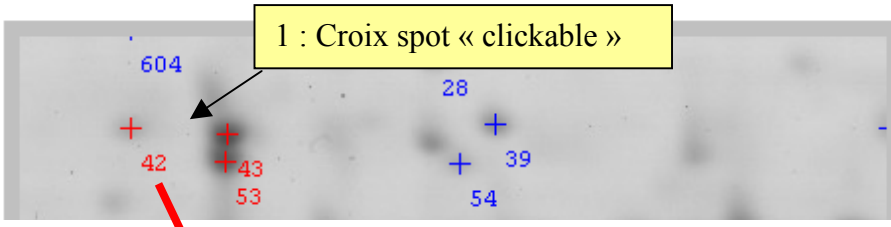
Feuille List :
A ce stade il reste encore un certain nombre de tables non alimentées par le format d'échange.
Ces tables sont les suivantes : Hardware, Software, Db, Tissue, Dv_stage_sample, Bio_type, Defined_Type, Molecule_type. Le contenu de ces tables contenant peu de champs sera détaillé dans une seule feuille nommée feuille List qui servira en même temps de feuille de référence pour remplir les autres Feuilles.

Display of the gel 2D : G229_MON1

ANNEXE 24 : Page Web de visualisation « Fiche Spot » éditée à partir de l'interface de navigation dans l'image d'un gel

Full Display

1 : Croix spot « clickable »



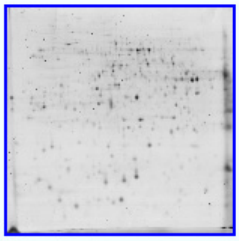
Display of Gel : G229_MON1

Detail spot information

General Informations

- **Spot Id :** 19
- **Spot number :** 42
- **X position :** 1001
- **Y position :** 331
- **Pi :** 5.31
- **Mw :** 94
- **Short_remark :**

Related Gel



- **Gel 2D :** G229_MON1

Ms_analysis 1

- **Ms analysis id :** 19
- **Ms analysis name :** G229-042
- **Date :** 2002-02-14
- **Ms analysis file :** G229-042-PL
- **Short remark :**
- **Ms analysis type :** PMF
- **Ms analysis Protocol :** Peptide Mass Fingerprint
- **Contact :** Sommerer Nicolas

Ms_analysis 1 result

- **Ms analysis result id :** 19
- **Date :** 2002-02-17
- **Ms analysis result file :** G229-042-MCT.htm
- **ko :** yes
- **Description :**
- **Ms analysis result Protocol :** Mascot search PMF
- **Contact :** Sommerer Nicolas

Protein :

- **Acc id :** S20876
- **Protein name :** chaperonin hsp60 precursor
- **Database name :** NCBI

Index

	Accession	Mass	Score	Description
1.	S20876	61312	83	chaperonin hsp60 precursor
2.	BAB03017	61242	83	AP001297 NID: - Arabidops
3.	AAG50729	70036	50	AC079041 NID: - Arabidops
4.	AAG50788	70378	50	AC074309 NID: - Arabidops
5.	Q9FNE4	114170	42	GENOMIC DNA, CHROMOSOME 5,
6.	Q9ZU67	47297	42	PUTATIVE MEMBRANE-ASSOCIAT
7.	Q9M8W6	24083	40	PUTATIVE SRF-TYPE TRANSCRI
8.	Q9FJJ1	53392	40	GENOMIC DNA, CHROMOSOME 5,
9.	Q9LJI7	53683	40	PROTEIN KINASE.- Arabidops
10.	AAG50784	59566	40	AC074309 NID: - Arabidops
11.	T49091	124794	39	gamma response I protein -
12.	S42547	20743	39	glucose-1-phosphate adenyl
13.	CAA51778	20800	39	ATRNAAPL2 NID: - Arabidop
14.	AAC98029	32755	37	F5O8 NID: - Arabidopsis t
15.	T00509	68563	37	hypothetical protein T20D1
16.	T47497	35339	35	hypothetical protein F9K21
17.	Q9LNA2	28040	35	F5O11.17.- Arabidopsis tha
18.	T01504	78927	34	hypothetical protein T10M1
19.	T06310	91895	34	hypothetical protein F11C1
20.	Q9M0Z8	42744	33	HYPOTHETICAL 42.8 KDA PROT

Results List

1. S20876 Mass: 61312 Score: 83
chaperonin hsp60 precursor - Arabidopsis thaliana

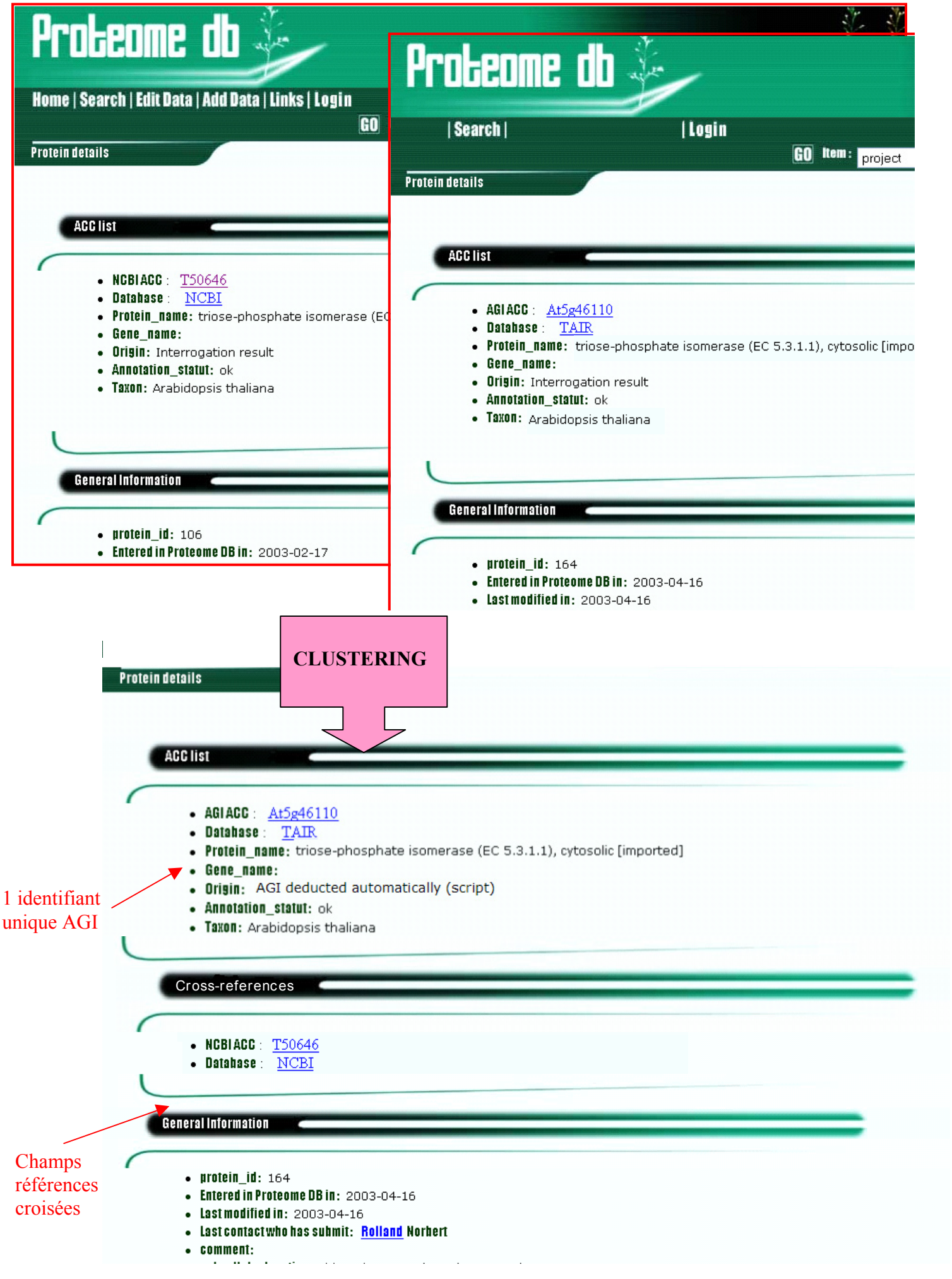
Observed	Mr(expt)	Mr(calc)	Delta	Start	End	Miss
941.59	940.58	940.61	-0.03	266 -	273	1
1235.55	1234.54	1234.58	-0.04	136 -	147	0
1327.66	1326.65	1326.69	-0.04	67 -	78	0
1418.66	1417.65	1417.69	-0.03	228 -	239	0
1532.77	1531.77	1531.79	-0.02	455 -	467	1
1661.77	1660.76	1660.79	-0.03	499 -	513	0
1927.06	1926.05	1926.06	-0.01	435 -	454	0

No match to: 678.49, 1179.57, 1475.75, 1594.77, 1791.67, 1908.96

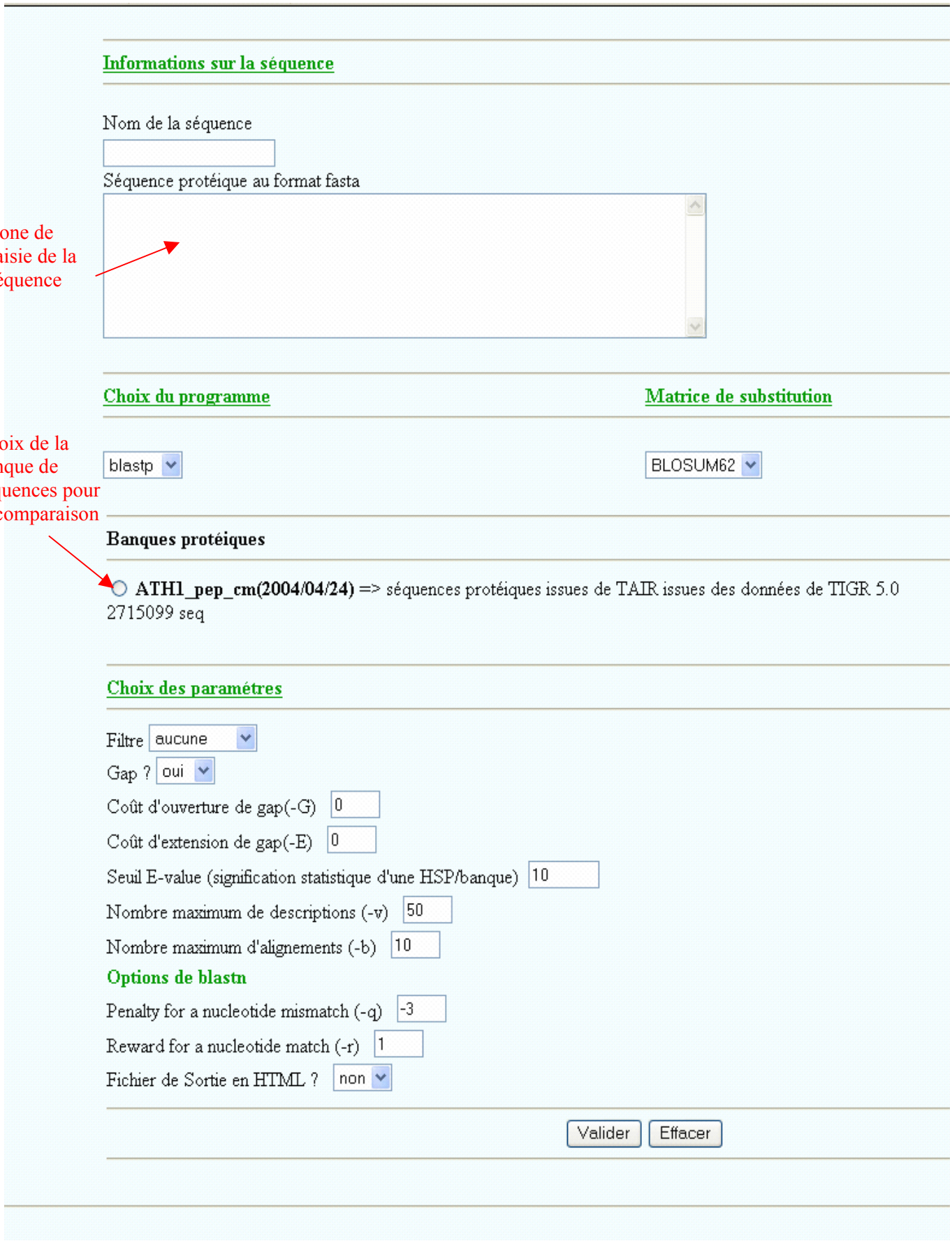
2. BAB03017 Mass: 61242 Score: 83

2 : Fiche Mascot

ANNEXE 25 : Maquette des interfaces protéines T50646 et At5g46110 avant et après clustering



ANNEXE 26 : Interface du logiciel BLAST dans ProteomIs



ANNEXE 27 Scripts de création des tables temporaires des tables temp_spotband_list et temp_protein_list

1) Script 1 de création des tables de la table temp_spotband_list :

```
/* -- DROP TABLE "temp_spotband_list"; --*/

CREATE TABLE "temp_spotband_list" (

    "spotband_id" numeric(38,0) NOT NULL,
    "spotband_numeric" varchar(50) NOT NULL,
    "spotband_x_position" numeric(10,0),
    "spotband_y_position" numeric(10,0),
    "pi" real,
    "mw" varchar(15) NOT NULL,
    "gel_id" numeric(38,0) NOT NULL,
    "gel_name" varchar(100) NOT NULL,
    "nb_prot" numeric(38,0) NOT NULL,
    "protein_id" numeric(38,0) NOT NULL,
    "acc_id" varchar(20) NOT NULL,
    "protein_name" varchar(200) NOT NULL,
    "acc_type_name" varchar(50) NOT NULL,

    CONSTRAINT "pk_temp_spotband_list" PRIMARY KEY ("spotband_id", "protein_id", "acc_id",
"protein_name")

);

CREATE  INDEX "idx_spotband_id_temp_spotband_list" ON "temp_spotband_list" ("spotband_id");
CREATE  INDEX "idx_gel_id_temp_spotband_list" ON "temp_spotband_list" ("gel_id");

INSERT INTO temp_spotband_list (spotband_id, spotband_numeric, spotband_x_position,
spotband_y_position, pi, mw, gel_id, gel_name, nb_prot, protein_id, acc_id, protein_name,
acc_type_name)

select spot_band.spotband_id,spot_band.spotband_numeric, spot_band.spotband_x_position,
spot_band.spotband_y_position, spot_band.pi, spot_band.mw, gel.gel_id,gel.gel_name, nb_prot,
protein.protein_id, acc.acc_id,acc.protein_name,bio_type.name  from temp_spotband_list_reduce,
spot_band, gel, spot_band_protein, acc, bio_type
        where
        spot_band.gel_id=gel.gel_id and spot_band.spotband_id=temp_spotband_list_reduce.spotband_id
        and spot_band.spotband_id=spot_band_protein.spotband_id and
spot_band_protein.protein_id=protein.protein_id
        and protein.protein_id=acc.protein_id and acc.bio_type_id=bio_type.bio_type_id

UNION

select spot_band.spotband_id,spot_band.spotband_numeric, spot_band.spotband_x_position,
spot_band.spotband_y_position, spot_band.pi, spot_band.mw, gel.gel_id,gel.gel_name, '0' as nb_prot, '-
100' as protein_id, '-100' as acc_id, 'none' as protein_name, 'none' as name from spot_band, gel
where spot_band.gel_id=gel.gel_id

and spot_band.spotband_id NOT IN (

select spot_band.spotband_id  from spot_band, gel, spot_band_protein, acc, bio_type
```

```
        where spot_band.gel_id=gel.gel_id and spot_band.spotband_id=spot_band_protein.spotband_id
and spot_band_protein.protein_id=protein.protein_id
        and protein.protein_id=acc.protein_id and acc.bio_type_id=bio_type.bio_type_id )

order by gel_id,spotband_id;
```

2) Script 2 de création des tables de la table temp_protein_list :

```
/*-- DROP TABLE "temp_protein_list"; --*/

CREATE TABLE "temp_protein_list" (

    "protein_id" numeric(38,0) NOT NULL,
    "subcellular_location_typid" numeric(38,0) NOT NULL,
    "subcell_location" varchar(100) NOT NULL,
    "nb_acc" numeric(38,0) NOT NULL,
    "acc_id" varchar(20) NOT NULL,
    "acc_type_name" varchar(50) NOT NULL,
    "protein_name" varchar(200) NOT NULL,

    CONSTRAINT "pk_temp_protein_list" PRIMARY KEY ("protein_id", "acc_id", "protein_name")

);

CREATE  INDEX "idx_protein_id_temp_protein_list" ON "temp_protein_list" ("protein_id");
CREATE  INDEX "idx_acc_id_temp_protein_list" ON "temp_protein_list" ("acc_id");

INSERT INTO temp_protein_list (protein_id, subcellular_location_typid, subcell_location, nb_acc,
acc_id, acc_type_name, protein_name)

select protein.protein_id, protein.subcellular_location_typid, subcell_type.name as subcell_location,
nb_acc, acc.acc_id, bio_type.name, acc.protein_name
from protein, bio_type as subcell_type, acc, bio_type, temp_protein_list_reduce
    where
    protein.protein_id=temp_protein_list_reduce.protein_id
    and protein.subcellular_location_typid=subcell_type.bio_type_id
    and  protein.protein_id=acc.protein_id and acc.bio_type_id=bio_type.bio_type_id

UNION

select protein.protein_id, protein.subcellular_location_typid, subcell_type.name as subcell_location, '0' as
nb_acc, '-100' as acc_id, 'none' as name, 'none' as protein_name
from protein, bio_type as subcell_type where
protein.subcellular_location_typid=subcell_type.bio_type_id

and  protein.protein_id NOT IN (

select protein.protein_id
from protein, bio_type as subcell_type, acc, bio_type where
protein.subcellular_location_typid=subcell_type.bio_type_id
and  protein.protein_id=acc.protein_id and acc.bio_type_id=bio_type.bio_type_id)

order by subcellular_location_typid;
```