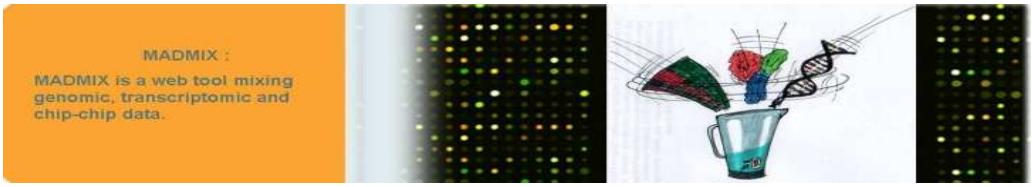
UNIVERSITÉ DE NANTES

Master Professionnel – Bioinformatique

Septembre 2007

Roxane LEGAIE

MADMIX :
Outil d'intégration de données de Génomique, de Transcriptomique et de ChIP-chip



INSERM U533 – Institut du Thorax

Maître de stage : Dr Rémi HOULGATTE

L'institut du thorax

genopole® OUEST

Inserm U533
Institut national de la santé et de la recherche médicale

REMERCIEMENTS

Je tiens à remercier très sincèrement le Dr Rémi HOULGATTE pour m'avoir renouvelé sa confiance en m'accueillant à nouveau au sein de son équipe, pour ses précieux conseils, sa simplicité et son écoute.

Un grand merci à toute l'équipe de Bioinformatique, Audrey Bihouée, Solenne Carat, Emeric Dubois, Mahatsangy Raharijaona et Raluca Teusan, pour leur accueil très chaleureux, leur complicité et leur soutien.

Merci aussi à l'ensemble de l'équipe *Génomique et Biothérapies*, et notamment à Daniel Baron, Catherine Chevalier, Richard Danger et Sylvia Quemener, pour leurs conseils et leur amitié.

A toute l'équipe de l'unité INSERM U533, merci pour leur accueil et leur sympathie.

Enfin, un merci tout particulier aux personnes extérieures à la bioinformatique pour leur soutien et leur confiance tout au long de mes études et en particulier au cours de ce stage.

SOMMAIRE

ABREVIATIONS ET DEFINITIONS

Analyseur syntaxique : (*parser* en anglais) est un programme informatique qui consiste à exhiber la structure d'un texte.

CSS : *Cascading Style Sheet*. Langage informatique utilisé pour décrire la présentation d'un document structuré écrit en HTML ou XML

DOM : *Document Object Model*. Recommandation du W3C qui permet de construire une arborescence de la structure d'un document et de ses éléments. Ce langage est utilisé pour créer et modifier facilement des documents XML.

Ensembl : Système logiciel qui produit et maintient l'annotation automatique sur les génomes eucaryotes.
(www.ensembl.org)

GEO : *Gene Expression Omnibus*. Base de données publique qui archive et distribue librement les données de puces à ADN et de ChIP-chip soumises par la communauté scientifique.
(www.ncbi.nlm.nih.gov/geo)

GO : *Gene Ontology.* Consortium international créé afin d'établir un vocabulaire structuré et contrôlé pour qualifier les gènes. Un gène y est classé selon 3 axes : les fonctions moléculaires, les processus biologiques et les composants cellulaires.
(www.geneontology.org)

HTML : *Hypertext Markup Language.* Langage informatique de balisage conçu pour écrire les pages Web.

Oligonucléotide : Petit segment d'ADN simple brin (représentant le plus souvent un gène).

PCR : *Polymerase Chain Reaction*. Méthode de biologie moléculaire permettant d'amplifier le nombre de copies d'une séquence spécifique d'ADN.

PHP : *Hypertext PreProcessor*. Langage de scripts libre principalement utilisé pour être exécuté par un serveur web.

QuickGO : Navigateur permettant l'étude et l'observation des données de Gene Ontology.
(www.ebi.ac.uk/ego)

SGBD : *Système de Gestion de Base de Données.* Système qui héberge généralement plusieurs bases de données, qui sont destinées à des logiciels ou des thématiques différentes.

W3C : *World Wide Web Consortium.* Consortium fondé en 1994 pour promouvoir la compatibilité des technologies du World Wide Web telles que HTML, XHTML, XML, CSS, PNG, etc.

XML : *Extensible Markup Language.* Langage informatique de balisage générique. Il permet de créer des fichiers dont la structure est définissable et validable par un schéma.

I. INTRODUCTION

J'ai effectué mon stage de Master Professionnel en Bioinformatique à l'Institut du Thorax de Nantes, sur la Plateforme Transcriptome de l'unité INSERM U533. J'y ai rejoint l'équipe de Bioinformatique dirigée par le Dr Rémi HOULGATTE.

Ce stage de plusieurs mois au sein même d'une équipe de bioinformatique m'a permis d'appliquer les notions théoriques vues au cours de ma formation et d'acquérir une expérience déterminante pour mon avenir professionnel, tant au niveau biologique qu'informatique.

A. Contexte biologique

L'acide désoxyribonucléique (ADN) est une molécule présente dans toutes nos cellules. Elle est le support de l'hérédité et constitue le génome des êtres vivants. L'ADN est composé de séquences de nucléotides formant des gènes codant pour des protéines indispensables à la vie cellulaire.

La transcription est un mécanisme biologique permettant de passer d'un gène à une protéine. Chaque gène code pour une molécule d'ARN (Acide RiboNucléique). Celle-ci subit une étape d'épissage permettant d'obtenir différents transcrits qui correspondent chacun à une ou plusieurs protéines. La transcription est régulée par des protéines particulières appelées « Facteurs de transcription » (FT). L'action de ces FT sur les promoteurs de leurs gènes cibles constitue un réseau de régulation.

Dans un grand nombre de pathologies, la transcription de certains gènes est altérée, ce qui entraîne un dérèglement du fonctionnement cellulaire. L'ensemble des variations du niveau d'expression de ces gènes constitue une véritable signature transcriptionnelle, spécifique de la pathologie étudiée [1].

Le mode de fonctionnement des gènes, leur localisation génomique et surtout leurs voies de régulation apparaissent donc comme des points clés pour la compréhension des mécanismes mis en jeu dans les pathologies. La connaissance du génome humain notamment par le séquençage complet de ce dernier [2,3] a été déterminante dans l'avancée de la recherche.

Il est en effet possible aujourd'hui d'étudier par des techniques à grande échelle (dites à « haut débit »), les variations du transcriptome et les mécanismes de régulation transcriptionnelle, à l'échelle sans précédent du génome.

Ainsi, l'expression de tous les gènes d'un échantillon biologique (sain ou pathologique) peut être mesurée simultanément et rapidement par les puces à ADN (ou « microarrays »). Il s'agit de supports solides miniaturisés (lame de verre ou membrane de nylon) sur lesquels sont déposés des milliers de sondes génomiques (oligonucléotides ou produits de PCR), correspondant chacune à un gène (figure 1).

L'hybridation de la puce avec l'ARN reverse-transcrit et marqué (par un fluorochrome ou par radioactivité) d'un échantillon biologique permet de détecter et de quantifier l'ensemble des transcrits qu'il contient en une seule expérience [4].

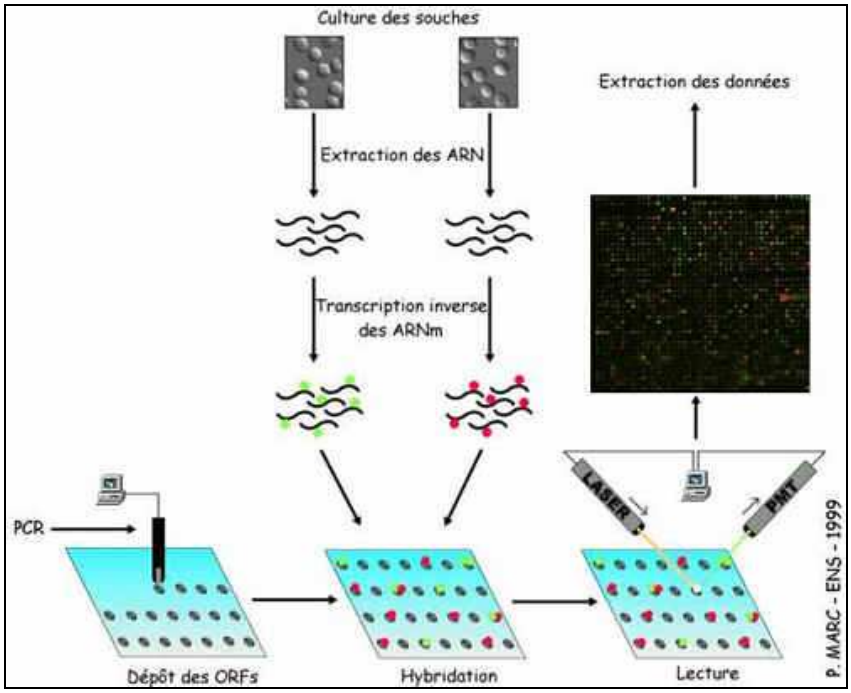


Figure 1 : Principe de la puce à ADN sur lame de verre

Après acquisition numérique des mesures d'hybridation, normalisation et filtrage, la comparaison des profils d'expression obtenus sur des échantillons « sains » et « malades », permet de mettre en évidence les signatures transcriptionnelles caractéristiques de la pathologie étudiée.

D'autre part, pour un échantillon biologique donné, il est également possible d'identifier tous les gènes cibles d'un même FT par la technique du ChIP-Chip (Chromatine ImmunoPrecipitation on Chip) [5,6].

En effet, les FT sont des protéines qui se fixent sur les régions proximales des gènes. Ces facteurs interagissent en complexes pour activer ou réprimer l'expression des gènes. La compréhension des systèmes de régulation sous-jacents à une pathologie passe par l'identification des gènes régulés par ces FT.

La technique du ChIP-Chip consiste à immunoprécipiter la chromatine de l'échantillon biologique, au moyen d'un anticorps dirigé contre le FT étudié (figure 2). L'avantage de cette technique est qu'elle permet d'observer les interactions dans les conditions physiologiques.

Suite à une purification, il est possible de récupérer les régions d'ADN génomique sur lesquelles s'était fixé le FT. Ces séquences sont ensuite hybridées, après amplification et marquage, sur une puce pan-génomique (« Tiling Array ») le plus souvent.

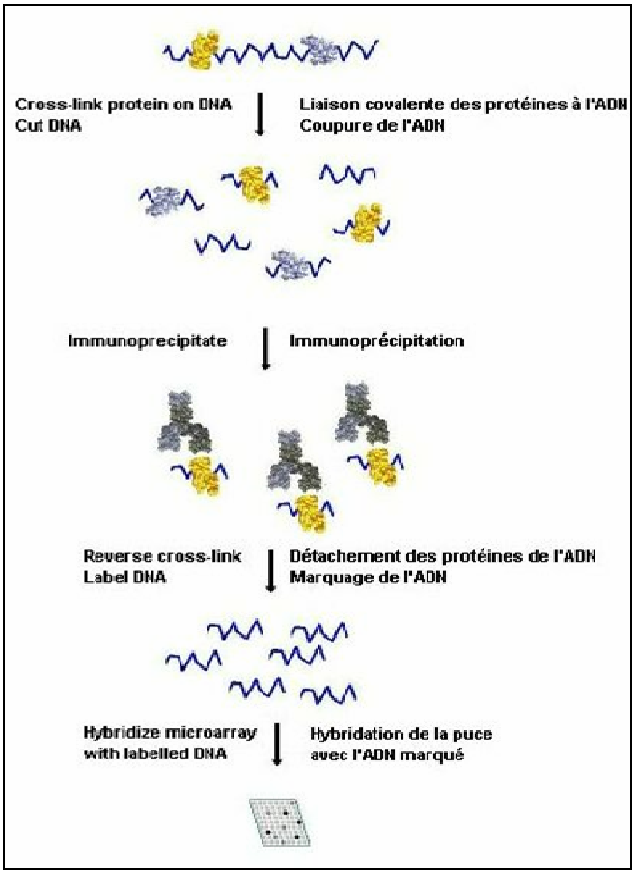


Figure 2 : Principe du ChIP-chip

On obtient après révélation un signal pour chaque sonde de la puce. On peut alors distinguer deux groupes de séquences : les dites positives en ChIP, qui possèdent un site de fixation pour ce FT, et les négatives, où le FT n'était pas fixé au moment de l'expérience. La distinction des deux types de séquences se fait à l'aide d'un seuil statistique.

L'objectif fondamental de l'équipe *Génomique et Biothérapies* (dont fait partie la thématique Bioinformatique) est de mettre en évidence ces signatures transcriptionnelles pour permettre le diagnostic et le pronostic de cancers [7] ou de pathologies cardiovasculaires [8]. Découvrir les réseaux de régulations transcriptionnelles sous-jacents à ces signatures permettra de découvrir de nouvelles cibles thérapeutiques raisonnées en distinguant les causes des conséquences.

Dans ce contexte, l'objectif de mon stage a été de développer un outil permettant d'intégrer facilement et rapidement des données de Génomique (localisation génomique des gènes), de Transcriptome (pour l'étude des niveaux d'expression de gènes) et de ChIP-chip (pour l'identification des gènes régulés par tel(s) FT). L'intégration de ces données permettra de modéliser les mécanismes de régulation transcriptionnelle sous-jacents aux pathologies étudiées par l'équipe.

B. Contexte bioinformatique:

Les puces à ADN génèrent donc de grandes quantités de données qu'il convient d'archiver, de traiter, d'analyser et d'annoter. Pour cela l'équipe de Bioinformatique a développé un ensemble d'outils appelé MADTOOLS (MicroArray Data Tools), disponible à cette url : www.madtools.org .

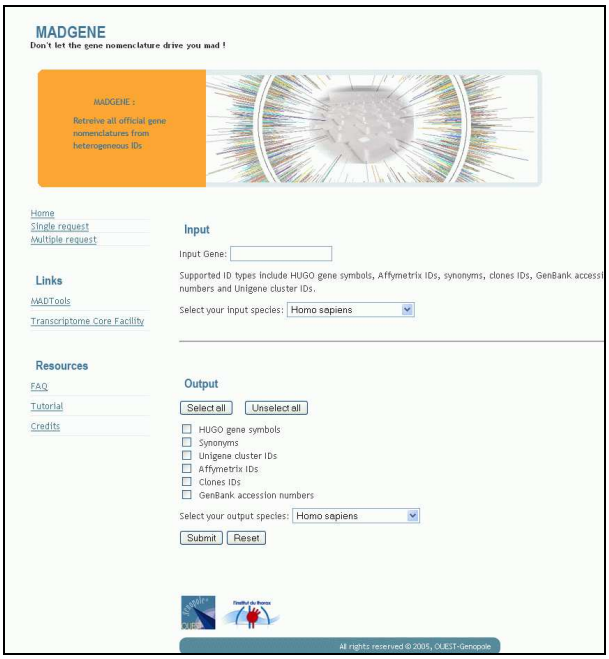


Figure 3 :
Capture d'écran de l'outil MADGENE

MADGENE :

Une base de données permettant la conversion d'identifiants de gènes de manière automatique (symbole HUGO [9], synonymes, Unigene cluster ID [10], Affymetrix ID, Clone ID, Accession Number, identifiants Ensembl [11], RefSeq…).

Elle permet également de passer d'une espèce à une autre grâce aux gènes orthologues (figure 3). Cet outil est géré par Audrey Bihouée.

Figure 4 :
Capture d'écran de l'outil MADSENSE

MADSENSE :

Une base de connaissances qui intègre des informations biologiques et bibliographiques obtenues à partir de différentes banques de données publiques (figure 4).

L'information est organisée sous la forme de cartes d'identité (ou « GeneCards ») qui aident à la compréhension des fonctions et des relations entre les gènes humains. Cet outil a été développé par Raluca Teusan.

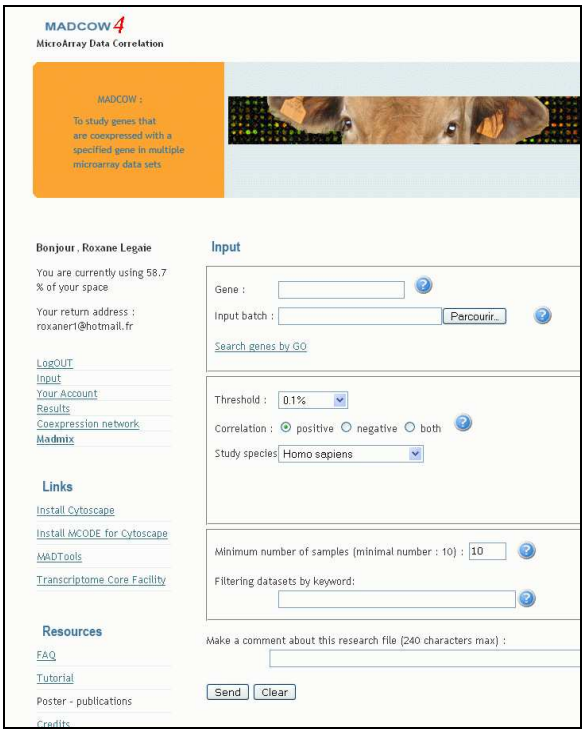


Figure 5 :
Capture d'écran de l'outil MADCOW

MADCOW :

Un logiciel permettant d'identifier les gènes coexprimés avec un gène d'intérêt à travers les données publiques de puces à ADN, selon la méthodologie décrite dans Lee *et al.* [12] (figure 5).

Plusieurs études ont montré que les gènes d'expression corrélée participent à une même fonction biologique [12,13].

Les résultats peuvent être filtrés, comparés et annotés par des statistiques sur les termes de Gene Ontology [14]. L'interface propose également la réalisation de réseaux de coexpression visualisables grâce à l'outil Cytoscape [15]. MadCow a été mis au point par Emeric Dubois.

Figure 6 :
Capture d'écran de l'outil MADMUSCLE

MADMUSCLE :

Une base de données spécialisée contenant des expériences publiques de puce à ADN analysées sur le muscle sain ou malade (figure 6).

Elle fournit des listes de gènes annotées ayant un profil d'expression similaire. Elle contient un outil permettant de comparer une liste quelconque à celles contenues dans la base Madmuscle. Cet outil a été créé par Daniel Baron**.**

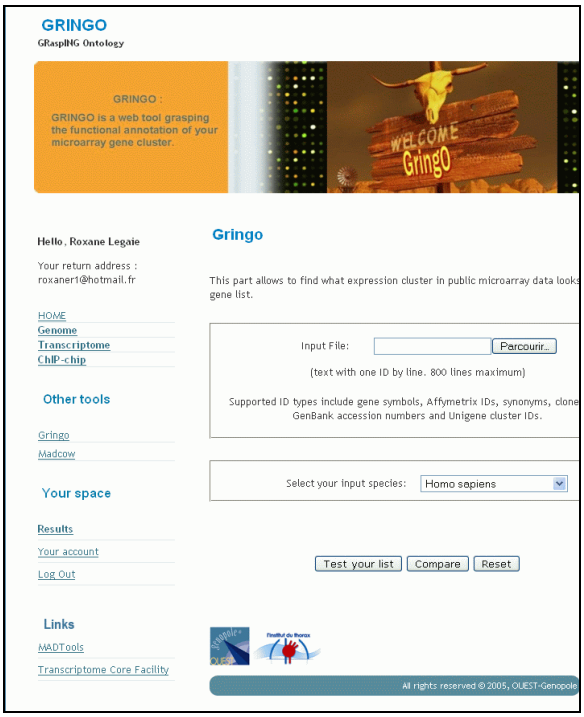


Figure 7 :
Capture d'écran de l'outil GRINGO

GRINGO (GraspING Ontologies) :

Un outil qui permet de construire des ontologies non biaisées à partir de données expérimentales, en se basant sur la ré-analyse de l'ensemble des jeux de données publiques. Il permet en outre de comparer en terme de fonctions biologiques, des données hétérogènes issues de technologies différentes, ou provenant d'espèces différentes
(figure 7).

Il permet enfin de mettre en évidence des fonctions biologiques inconnues, par l'existence de groupes de gènes d'expression coordonnée. Cet outil a été développé par Emeric Dubois.

C. Objectifs du stage

L'objet de mon stage a été de développer un outil innovant permettant d'intégrer à la fois des données de Génomique, de Transcriptomique et de ChIP-chip. L'outil, que j'ai choisi d'appeler « MADMIX », doit permettre de « mixer » ces trois types de données.

Il devra également s'intégrer aux MADTOOLS déjà existants (notamment MADCOW et GRINGO). D'autre part il devra sortir en résultat des fichiers formatés de façon à permettre des opérations communes à ces outils comme l'intersection, l'union et le filtrage. Enfin l'interface devra également être conviviale et identique à celle des MADTOOLS déjà disponibles via Internet.

L'utilisateur pourra ainsi effectuer des requêtes complexes permises par aucun autre outil disponible actuellement, du type : « Quels sont les gènes fortement exprimés dans le muscle atteint de myopathie de Duchenne, ayant un ilot CpG à moins de 1kb (kilobases) de leur promoteur et étant positifs en Chip pour tel FT ?».

MADMIX aidera ainsi le biologiste à tirer des conclusions biologiques à partir de données publiques et à comprendre des mécanismes de régulation mis en jeu dans un tissu ou une pathologie d'intérêt.

II. MATERIEL ET METHODES

L'outil que j'ai conçu repose sur une architecture 3-tiers : l'application PHP lancée sur le serveur web fait le lien entre les requêtes soumises par l'utilisateur et le serveur de base de données (figure 8).

- Serveur de base de données (premier tiers) :

L'ensemble des données nécessaires à l'outil est stockée sur un serveur MySQL (version 4.1.2). Il s'agit d'un SGBD gratuit (« open source »), qui a su démontrer sa haute fiabilité et sa simplicité de mise en œuvre. De plus ce système montre une très bonne intégration dans l'environnement Apache/PHP.

- Serveur web (deuxième tiers) :

Le serveur web est Apache (version 2.0.52) sous environnement Unix. Il comprend entre autres l'environnement PHP (version 4.3.9). L'ensemble de l'application est écrit dans ce langage de programmation. En effet, ce langage est très portable, capable de gérer une architecture multi-tiers et il permet de générer du code HTML interprétable par un navigateur Internet.

- Machine cliente (troisième tiers) :

La machine cliente reçoit des pages interprétées et codées au format HTML par le serveur. La présentation s'appuie sur la technologie CSS, recommandée par le consortium W3C.

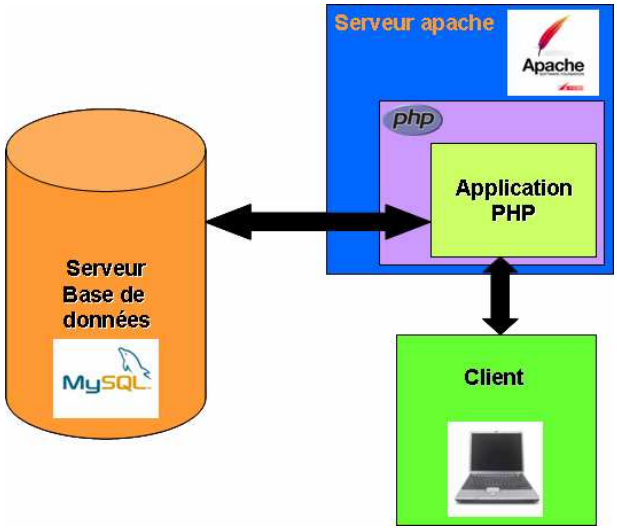


Figure 8 : Architecture multi-tiers de l'outil MADMIX

MADMIX a été divisé en trois modules : Génome, Transcriptome et ChIP-chip, pour permettre différents types de requêtes. Il est ainsi possible d'effectuer une requête simple n'interrogeant que la partie Génome par exemple, ou bien une requête très complexe faisant appel aux trois modules. De plus cela facilitera la maintenance et l'évolution de l'outil.

J'ai donc créé trois bases de données indépendantes. De même l'interface propose trois formulaires qui vont chacun interroger la base correspondante.

A. Stockage et Analyse des données

Après avoir créé les bases de données, j'ai utilisé le langage Perl (module DBI – « *DataBase Interface* ») afin de faire une analyse syntaxique des fichiers de données publiques et de remplir les tables.

1. Module Génome

Le but de ce module est d'obtenir une cartographie des objets du génome. Pour cela j'ai complété une base de données (BD) préexistante (MADGENE). On appellera la BD complète « *genome* ».

Un même gène peut être référencé par de nombreux identifiants (Unigene, Symbol, Name…). C'est pourquoi une ingénieure de l'équipe a développé MADGENE. La base de données sous-jacente à cet outil répertorie pour chaque gène tous ses identifiants connus. Elle est multi-espèces et permet donc de retrouver les orthologues d'un gène donné.

La dernière version de cette base contient également les informations concernant les gènes, transcrits, exons et promoteurs proximaux (1 à 3 kb en amont du départ de transcription) de chaque gène humain enregistré (figure 9).

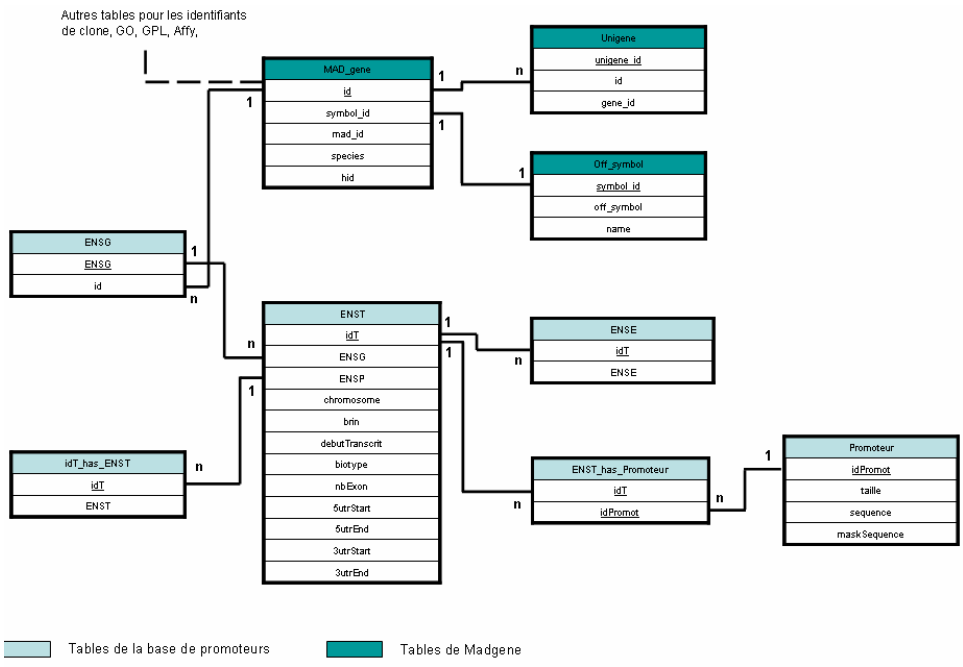


Figure 9 : Structure de la base de données MADGENE (dernière version)

Pour la mise au point de mon outil, j'ai complété cette BD par la position dans le génome (chromosome, position de début, position de fin et brin) de ces différents éléments (figure 10).

Pour cela, j'ai interrogé le site Ensembl [11] (version 43_36^{e}) via son API (*Application Programming Interface)* Perl à l'aide de ses identifiants : ENSG (*ENSemblGene*) pour les gènes, ENST (*ENSemblTranscrit*) pour les transcrits, ENSE (*ENSemblExon*) pour les exons.

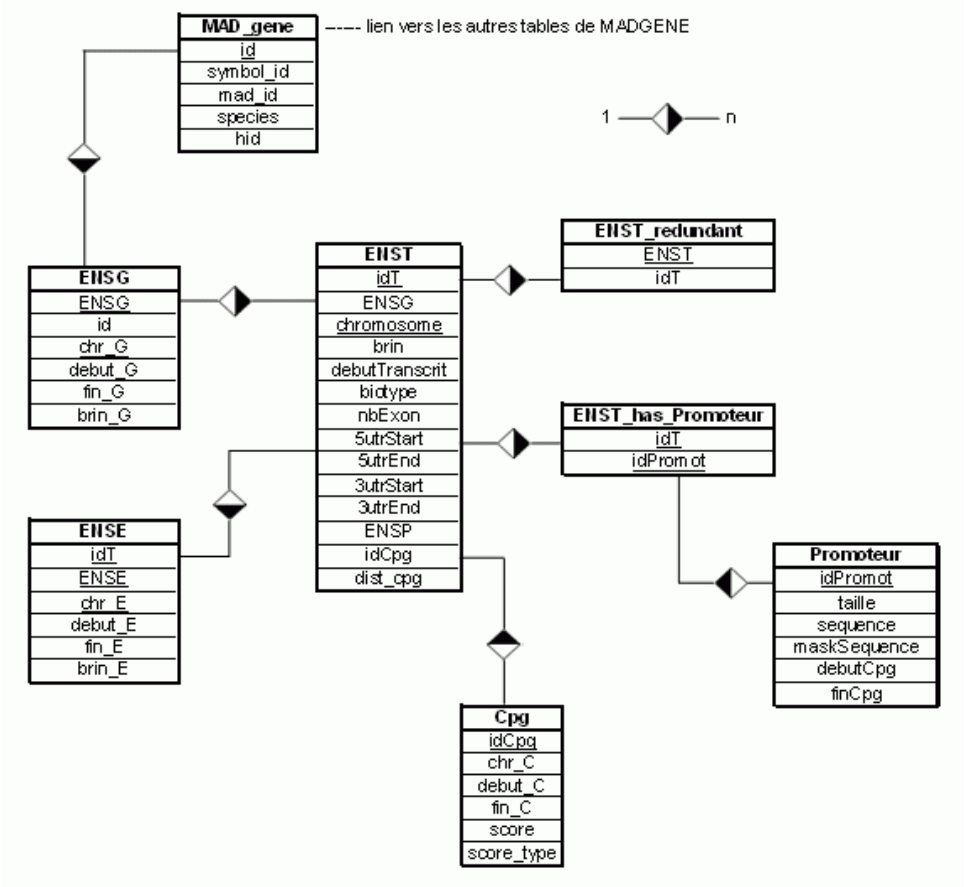


Figure 10 : Structure de la base de données *genome*

La table *ENSG* répertorie pour l'ensemble des gènes, leur position sur le génome : le chromosome ('chr_G'), la position de début ('debut_G'), la position de fin ('fin_G') et le brin ('brin_G'). En effet, l'ADN étant double brin, il existe des gènes positionnés sur le brin dit « sens » (valeur 1) et d'autres sur le brin dit « anti-sens » (valeur -1). C'est pourquoi il est nécessaire de préciser sur quel brin se trouve le gène.

La clé primaire est constituée de deux attributs : 'ENSG' et le 'chr_G'. En effet, il existe des gènes présents à la fois sur le chromosome X et sur le chromosome Y. Ensembl leur attribut le même identifiant. Les positions d'un même gène sur X et sur Y n'étant pas les mêmes, il était nécessaire de les différencier par leur chromosome.

L'attribut 'id' permet de faire le lien avec les tables de la BD originale MADGENE.

La table *ENST* répertoriait déjà les positions des transcrits à l'aide des attributs 'debutTranscrit', '5utrStart', etc.

Certains transcrits sont répertoriés dans Ensembl sous plusieurs identifiants 'ENST' alors qu'ils présentent les mêmes caractéristiques (position génomique, nombre d'exons, longueur, séquence ...). Il s'agit en fait du même transcrit mais il se différencie par certains identifiants, ce qui ne présente aucun intérêt pour nous (redondance). Pour conserver une base cohérente, la table *ENST* auto-incrémente un identifiant unique 'idT' pour chaque transcrit. La table *ENST_redundant* répertorie le ou les identifiants 'ENST' correspondant.

La clé primaire est ici aussi constituée de deux attributs : 'idT' et 'chr_T'. Ceci pour les mêmes raisons que pour la table *ENSG*.

A chaque transcrit sont associés trois promoteurs : à 1kb, 2kb et 3kb en amont de son départ de transcription. La table *ENST_has_Promoteur* fait cette association entre un transcrit 'idT' et un promoteur 'idPromot'. La table *Promoteur* donne des informations sur la séquence et la taille de celui-ci.

Je n'ai pas modifié ces tables car les promoteurs ne sont pas des régions localisées de façon fiable sur le génome. Ce ne sont que des régions supposées promotrices. Il est donc très difficile de définir précisément la localisation du promoteur d'un transcrit. Le plus souvent il se situe à proximité en amont du transcrit, mais il peut aussi se trouver très éloigné et en aval.

La table *ENSE* répertorie les positions des différents exons qui composent un transcrit. Elle présente une clé primaire de trois attributs ('idT', 'ENSE' et 'chr_E') car un même exon peut appartenir à plusieurs transcrits ('idT').

Par ailleurs, j'ai ajouté une table pour répertorier les ilots CpG et leur position sur le génome. En effet, on connait aujourd'hui l'importance des ilots CpG dans la régulation de l'expression génique. Les dinucléotides symétriques CpG sont les sites majeurs de méthylation de la cytosine (un des nucléotides constituant l'ADN). Le phénomène qui existe chez tous les organismes multicellulaires n'a pas encore reçu une justification claire. Chez les mammifères les CpG ont été progressivement perdus, sauf dans des séquences courtes enrichies en CpG, appelées ilots CpG. Environ la moitié des gènes humains contiennent des îlots CpG dans leur région promotrice, le niveau de méthylation étant un élément régulateur majeur de leur expression. Leur présence témoigne d'une bonne localisation du gène sur le génome.

Pour cela j'ai tout d'abord récupéré les positions de tous les ilots CpG répertoriés par Ensembl. Puis, à chaque transcrit ('idT') enregistré dans la base, un programme associe l'ilot CpG ('idCpg' de la table *Cpg*) le plus proche de son départ de transcription, ainsi que la distance correspondante ('dist_cpg'). Cette association aurait pu être faite pour chaque gène mais chacun possédant plusieurs transcrits avec des positions différentes, il était préférable de les associer aux transcrits.

La BD *genome* devra être complétée par les positions génomiques des éléments de génomes d'autres espèces par la suite.

2. Module Transcriptome

Le module Transcriptome doit permettre de définir le niveau d'expression d'un gène dans une condition donnée. Cette condition est définie par l'échantillon analysé (tissu, pathologie…) au cours d'une expérience de puce à ADN donnée (espèce étudiée, puce utilisée…).

Pour cela il a d'abord fallu définir un moyen d'estimer le niveau d'expression d'un gène à partir de la valeur du signal mesuré lors de l'analyse de la puce à ADN. Ensuite j'ai créé une BD *transcriptome* pour stocker ces informations de façon à optimiser la rapidité des requêtes qui seront ensuite lancées via l'interface web.

a. Calcul des rangs

Pour estimer le niveau d'expression d'un gène, nous avons choisi d'utiliser les « rangs ». Il s'agit de la position relative du gène dans l'échantillon en fonction de l'intensité du spot mesurée.

Pour chaque échantillon d'une expérience, les gènes sont classés par ordre croissant en fonction de la valeur du signal qui leur est attribuée dans le fichier brut. Le rang est calculé à l'aide de cette équation :

$$\text{Rang} = \frac{\text{position du gène dans l'échantillon}}{\text{nombre de mesures de l'échantillon}} \times 100$$

Le gène avec la plus forte intensité dans l'échantillon aura le rang 100. Ainsi, les gènes fortement exprimés présenteront un rang élevé, tandis que les faiblement exprimés auront un rang proche de zéro.

b. Base de données de Transcriptome

Par souci de faisabilité dans les temps et de mise au point de l'outil, les jeux de données sur lesquels j'ai créé la BD *transcriptome* ont été réduit aux expériences faites sur des échantillons de Muscle. Soit les 41 expériences de MADMUSCLE disponibles via le FTP (File Transfert Protocole) du site GEO [16]. Celui-ci répertorie l'ensemble des données d'expériences de puces à ADN dans des fichiers formatés, relativement faciles à analyser de façon automatique.

La base sera par la suite complétée par l'équipe avec les expériences de leur choix à l'aide d'un pipeline de programmes que j'ai conçu et qui permet de récupérer et de traiter les données directement à partir des fichiers bruts (figure 11).

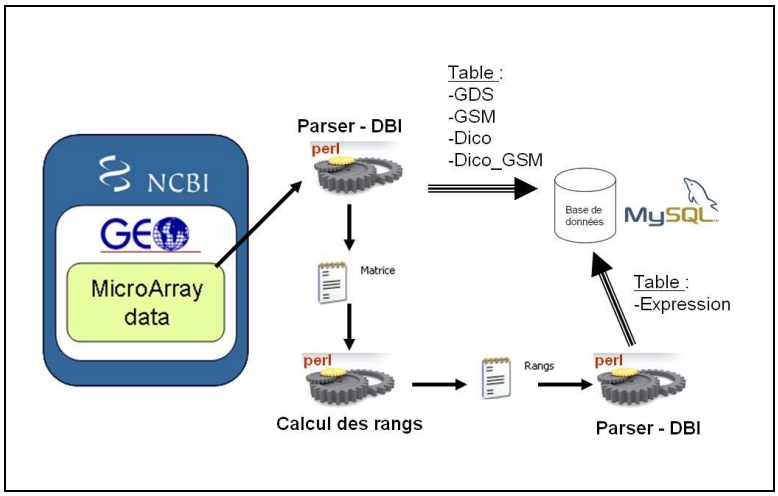


Figure 11 : Schéma du pipeline de traitement des données de Transcriptome

Un analyseur syntaxique (en Perl) récupère les informations concernant l'expérience, les différents échantillons testés et les valeurs du signal mesuré. Ensuite un programme traite ces valeurs et calcule les rangs. Enfin un dernier programme entre les informations d'intérêt dans les tables correspondantes de la BD *transcriptome* (figure 12) :

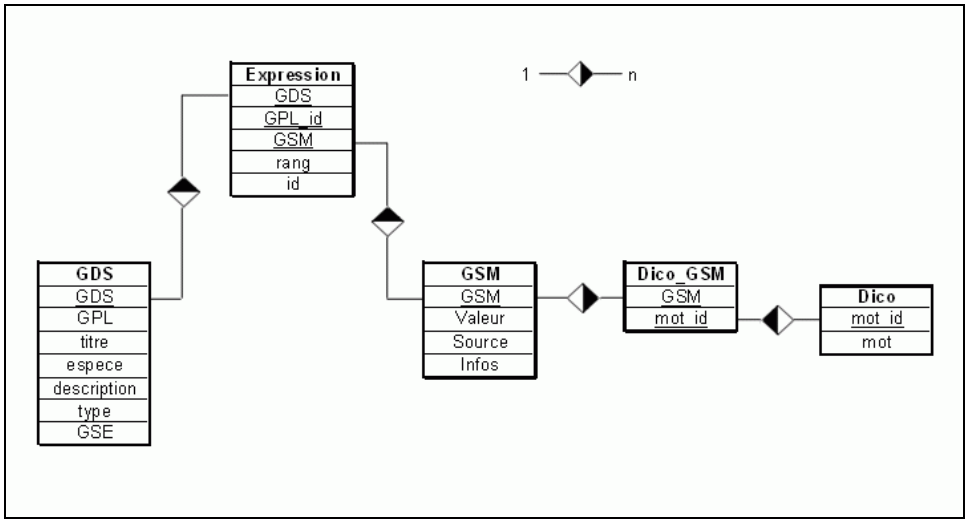


Figure 12 : Structure de la base de données *transcriptome*

La table *GDS* stocke les informations sur l'expérience :
- 'GDS' (GEO Data Set) : l'identifiant unique de l'expérience dans GEO
- 'GPL' (GEO PLatform) : l'identifiant unique GEO de la puce utilisée pour l'hybridation
- 'titre' : le titre de l'expérience lorsqu'il est disponible
- 'espece' : l'espèce à laquelle appartiennent les échantillons testés lors de l'expérience
- 'description' : une brève description de l'expérience
- le type de données : 'count' pour les expériences à un canal de fluorescence, 'log ratio' pour celles utilisant deux canaux. Le traitement de ces deux types de données ne sera pas le même.
- 'GSE' (GEO SEries) : la série d'expérience à laquelle appartient le GDS

La table *GSM* stocke les informations sur les échantillons testés :
- 'GSM' (GEO SaMple) : l'identifiant unique de l'échantillon dans GEO
- Les attributs 'Valeur', 'Source' et 'Infos', sont donnés arbitrairement par l'auteur de l'expérience. Ils permettent de définir les caractéristiques de l'échantillon.

Les tables *Dico* et *Dico_GSM* :

Pour optimiser les sélections d'échantillons qu'effectuera par la suite l'application web, un « dictionnaire » a été créé à partir des ces données.

Ainsi, à chaque ajout d'un nouvel échantillon dans la table *GSM*, la table *Dico* est mise à jour. Cette dernière répertorie l'ensemble des mots présents dans les attributs non-clés de la table *GSM*. Chacun ayant un identifiant 'mot_id' unique.

Enfin la table *Dico_GSM* associe à chaque 'GSM' le ou les mots-clés qui le caractérisent via le 'mot_id'.

La table *Expression* est le cœur de la BD *transcriptome* :

Elle stocke pour chaque gène ('GPL_id') ayant été testé dans un échantillon donné ('GSM') lors d'une expérience donnée ('GDS'), l'estimation de son expression ('rang').

Enfin, lors du remplissage de cette table, une requête permet également de récupérer pour chaque gène l'identifiant MADGENE 'id' correspondant. Ceci sera très utile lors des croisements de résultats entre les différents modules de MADMIX.

3. Module ChIP-chip

L'objectif du module ChIP-chip est de déterminer les gènes positifs à un certain seuil statistique pour un FT donné ayant été testé au cours d'une expérience (espèce étudiée, puce utilisée…) sur un type d'échantillon précis (tissu, pathologie…).

Pour cela il a fallu définir une stratégie statistique permettant d'estimer à partir de quelle valeur de signal mesuré un gène doit être considéré positif.

J'ai ensuite créé la BD *chip_chip* pour stocker ces informations de façon à optimiser la rapidité des requêtes qui seront ensuite lancées via l'interface.

a. Calcul des positifs

Tout d'abord, pour chaque échantillon d'une expérience, les valeurs brutes du signal mesuré pour l'ensemble des gènes testés sont normalisées par Lowess (« LOcally Weighted robust Scatterplot Smoothing ») [17]. Il s'agit d'un programme en R qui est lancé automatiquement sur les données. Cette normalisation permet de corriger les effets locaux non-linéaires (bruit de fond et saturation). Puis les valeurs normalisées sont passées en log ratio (le log2 du ratio du signal ChIP sur le signal Input). Cette transformation logarithmique permet :
- de donner un même poids à une augmentation ou une diminution de facteur 2
- d'avoir une distribution des valeurs qui se rapproche d'une Gaussienne
- d'éviter de favoriser les valeurs extrêmes
- de transformer les effets multiplicatifs en effets additifs

Ensuite, pour déterminer quels gènes sont positifs en ChIP dans l'échantillon, nous avons défini cette stratégie (figure 13) :

A partir des valeurs obtenues précédemment, on trace la courbe de distribution afin de vérifier qu'elle suit quasiment une gaussienne et on vérifie que la médiane de ces valeurs est bien centrée sur zéro (autant de valeurs négatives que positives).

Ensuite, les gènes sont triés de façon croissante en fonction de la valeur de leur log-ratio. A l'aide du nombre de valeur négatives (ex:10000 log-ratio < 0) et du seuil voulu (ex:5%), on récupère la valeur négative correspondant à la position ainsi mesurée (ex:10000*5%=500ème valeur en partant de la plus petite soit -4.3 par exemple).

Cette valeur est passée en positif et correspond alors à la valeur minimale positive au seuil (ex:4.3). Tous les signaux supérieurs à la valeur minimale positive au seuil (ex:>4.3) sont considérés comme significatifs au seuil donné.

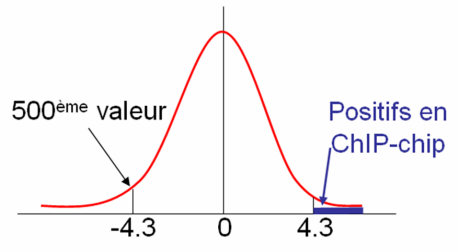


Figure 13 : Courbe illustrant la stratégie de calcul des positifs en ChIP-chip

On obtient alors les gènes positifs en ChIP pour le FT testé sur un échantillon donné au cours d'une expérience à un seuil statistique choisi.

Les seuils les plus pertinents que nous avons retenus sont : 0.01%, 0.005%, 0.003% et 0.001%. Ces seuils permettent de réduire au maximum le nombre de faux-positifs. Ils seront pré-calculés par le programme avant l'insertion des enregistrements dans la BD *chip_chip*.

b. Base de données de ChIP-chip

Pour les mêmes raisons de faisabilité et pour être cohérent avec le module Transcriptome, j'ai créé la BD *chip_chip* à partir de données d'expériences faites sur le muscle. Malheureusement celles-ci sont encore assez peu nombreuses, notamment dans les données publiques. Actuellement la table contient donc seulement 6 expériences. Celles-ci ont été ajoutées à partir des données de GEO ou des fichiers envoyés par les auteurs d'expériences de ChIP-chip directement.

De même que pour le Transcriptome, la base sera par la suite complétée par l'équipe avec les expériences de leur choix à l'aide d'un autre pipeline de programmes que j'ai conçu et qui permet de récupérer les données à partir des fichiers bruts de GEO (figure 14).

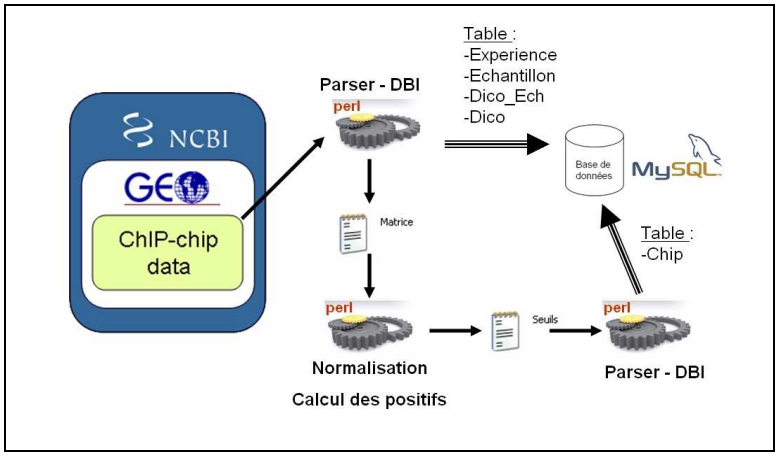


Figure 14 : Schéma du pipeline de traitement des données de ChIP-chip

Un analyseur syntaxique Perl récupère les informations concernant l'expérience, les différents échantillons testés et les valeurs du signal mesuré. Ensuite un programme traite ces valeurs, les normalise et calcule les positifs. Enfin un dernier programme entre les informations d'intérêt dans les tables correspondantes de la BD *chip_chip* (figure 15) :

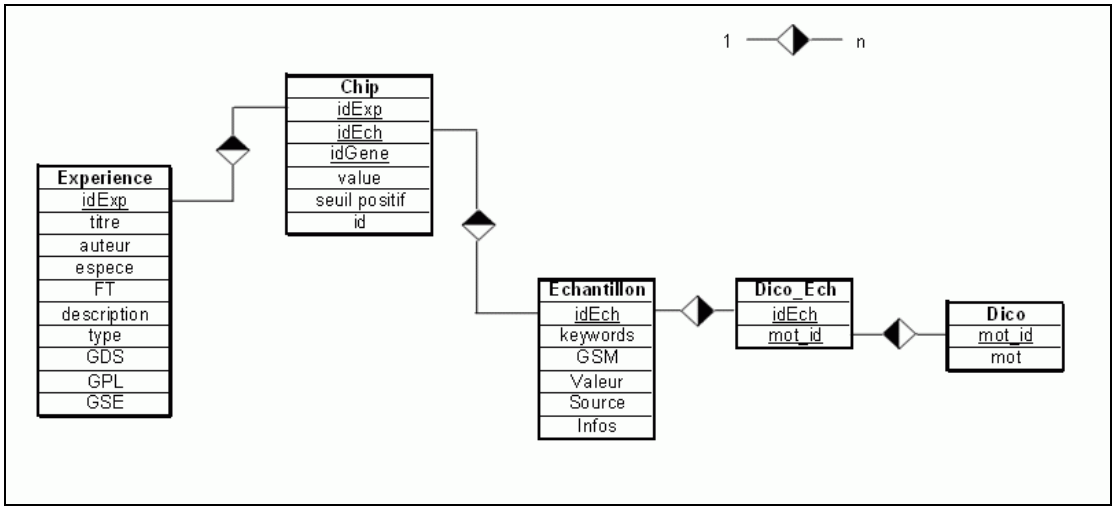


Figure 15 : Structure de la base de données *chip_chip*

La table *Experience* apporte des informations sur l'expérience réalisée : le titre éventuel, l'auteur, l'espèce, la description, le type et lorsqu'ils sont disponibles, les identifiants des GDS, GPL et GSE correspondants.

Un identifiant unique et auto-incrémenté 'idExp' a été créé car toutes les expériences ne possèdent pas d'identifiant 'GDS' (données non-publiques).

L'attribut 'FT' correspondant au facteur de transcription étudié est rempli lors de l'ajout de l'expérience dans la table. Il participera à la sélection des expériences lors des requêtes effectuées sur l'interface web.

La table *Echantillon* présente les mêmes attributs que la table *GSM* de la BD *transcriptome*. Cette fois un identifiant unique et auto-incrémenté 'idEch' a été créé car tous les échantillons ne possèdent pas d'identifiant 'GSM' (données non-publiques).

Enfin, un attribut 'keywords' est rempli manuellement à chaque ajout d'échantillon dans la base. Ces mots-clés seront déterminants pour la sélection d'échantillons lors des requêtes utilisateurs.

Un « dictionnaire » a tout de même été créé (tables *Dico* et *Dico_Ech*) pour assurer une sélection pertinente en cas d'absence de mots-clés (attribut 'keywords'). De plus cela autorise une sélection des échantillons beaucoup plus vaste.

La table *Chip* est le centre de la BD *chip_chip* : Elle stocke pour chaque gène ('idGene') ayant été testé dans un échantillon ('idEch') lors d'une expérience donnée ('idExp'), la valeur du log-ratio calculé ('value'), le seuil minimum auquel il est considéré comme positif ('seuil') et enfin l'identifiant de MADGENE correspondant ('id').
Seuls les gènes considérés positifs sont ajoutés dans la table *Chip*.

Les données n'étant pas toutes publiques, il existe des identifiants de gène (attribut 'idGene') absents de MADGENE. Il en résultait des enregistrements de la table *Chip* ayant l'attribut 'id' égal à zéro. De même, certaines expériences ne donnent pas d'identifiant pour le gène testé mais juste la position génomique de la sonde d'ADN utilisée.

J'ai donc mis au point un programme permettant de récupérer l'identifiant 'id' du gène correspondant à l'aide de ma BD *genome* qui contient les positions des gènes humains:
- Pour les identifiants absents de MADGENE, le programme récupère la position génomique de la sonde à l'aide du fichier brut correspondant au GPL (la puce).
- A partir de la position génomique de la sonde, une fonction recherche le gène le plus proche : Tout d'abord, il recherche s'il existe un gène incluant la sonde. S'il n'y en a pas, il cherche le gène le plus proche en aval (car souvent il s'agit de sondes pour les promoteurs situés en amont des gènes) ou en amont.

Ceci ne peut être appliqué qu'à des enregistrements correspondant à des expériences faites chez l'humain pour le moment.

Chacune des BD a été conçue de façon à optimiser le temps de réponse des requêtes effectuées par l'utilisateur sur l'interface web que j'ai développée par la suite.

B. Modules de requêtes et Interface Web

L'interface doit permettre d'interroger les trois modules de façon indépendante, rapide et pertinente.

De plus, elle doit permettre à l'utilisateur de conserver ses résultats et de les faire interagir (intersection, union…). Ces opérations relationnelles doivent pouvoir être faites entre résultats provenant des différents modules de MADMIX mais aussi entre résultats provenant d'outils différents : MADCOW et GRINGO notamment.

Enfin, l'interface se doit d'être conviviale et semblable à celle des MADTOOLS afin d'obtenir un site complet et homogène.

L'interface web a été développée en PHP et HTML. L'application sous-jacente interroge les bases de données à l'aide de requêtes MySQL.

1. Module Génome

Ce formulaire doit permettre à l'utilisateur de trouver les gènes ayant un ilot CpG à moins de tant de kb de leur départ de transcription et/ou localisés sur le chromosome de son choix (figure 16).

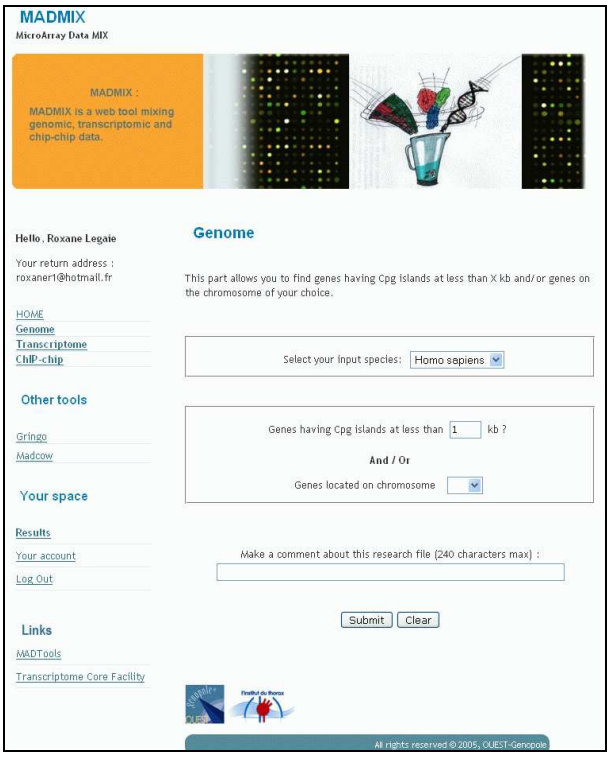


Figure 16 :
Capture d'écran du formulaire 'Genome'

La sélection de l'espèce sur laquelle l'utilisateur souhaite travailler se fait à l'aide d'un menu déroulant. Pour le moment, seule l'espèce humaine (*Homo sapiens*) est disponible. Mais lorsque la BD *genome* sera mise à jour, on pourra travailler sur diverses espèces telles que la souris (*Mus musculus*), le rat (*Rattus norvegicus*), la drosophile (*Drosophila melanogaster*), le chien (*Canis familiaris*), la poule (*Gallus gallus*), le nématode (*Caenorhabditis elegans*), etc.

Une zone de texte permet de choisir la distance (en kb) maximale entre le gène et l'ilot le plus proche connu. Ceci afin de laisser une plus grande liberté de requête à l'utilisateur.

L'application interroge la BD *genome* et récupère les gènes (identifiant MADGENE 'id' à partir de la table *ENSG*) correspondants aux transcrits ayant un ilot CpG à une distance ('dist_cpg') inférieure à celle entrée dans la zone de texte.

La sélection du chromosome se fait à l'aide d'un menu déroulant (chromosomes humains pour le moment) ce qui permet de mieux contrôler l'entrée du formulaire.
Une simple requête permet de sélectionner les gènes présents (attribut 'id') sur le chromosome voulu à l'aide de l'attribut 'chr_G' de la table *ENSG*.

Si l'utilisateur renseigne ces deux zones, l'application prendra en compte à la fois la distance et la localisation chromosomique (figure 17). Le résultat final sera l'intersection de ces deux listes de gènes ('id'). Sinon elle n'effectuera la recherche que sur la zone renseignée.

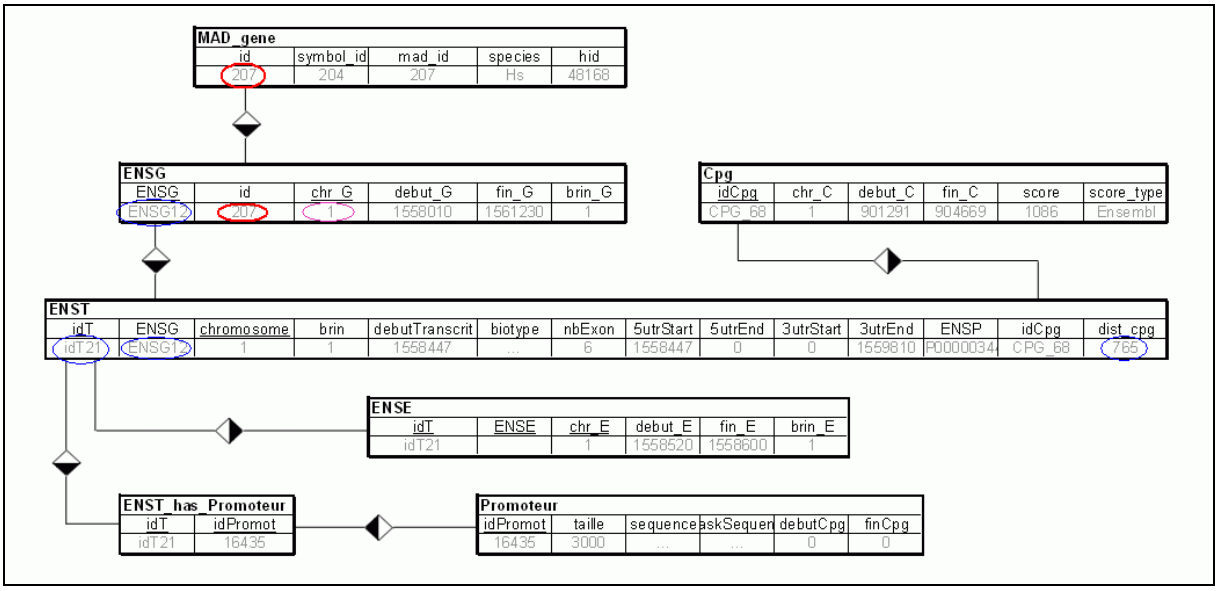


Figure 17 : Schéma du module de requête Genome
En bleu la sélection par la distance aux ilots CpG. En rose la sélection par le chromosome.
En rouge, le lien avec l'identifiant unique 'id' de MADGENE

Enfin il est proposé à l'utilisateur de faire un commentaire sur sa recherche. Celui-ci l'aidera à retrouver son résultat dans son espace de travail par la suite.

2. *Module Transcriptome*

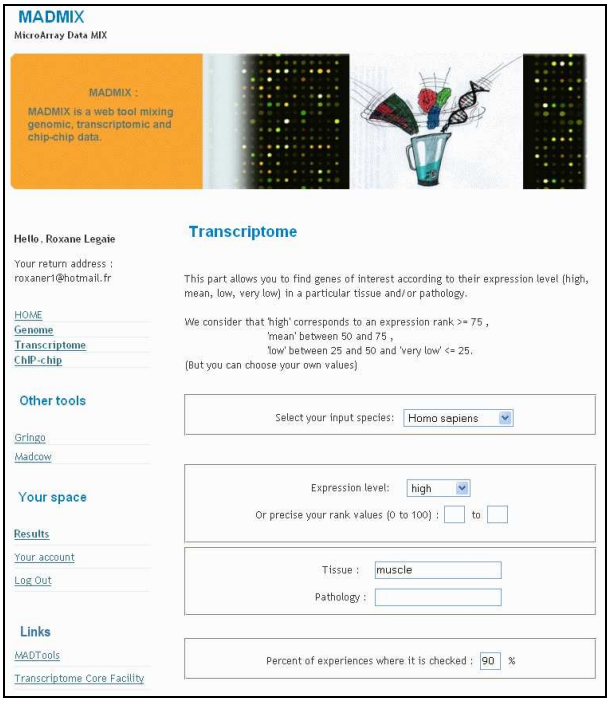


Figure 18 :
Capture d'écran du formulaire 'Transcriptome'

Cette partie doit permettre de sélectionner des gènes selon leur niveau d'expression mesuré dans un tissu et/ou une pathologie donnée (figure 18). Ces derniers sont caractérisés par une expérience et un échantillon.

L'application sélectionnera les expériences ('GDS') correspondant à l'espèce choisie par l'utilisateur. Celle-ci se fait à l'aide d'un menu déroulant dynamique qui interroge en temps réel la BD *transcriptome* afin de proposer seulement les espèces des expériences disponibles au moment de la requête.

Le niveau d'expression qui intéresse l'utilisateur est évalué à l'aide des rangs mémorisés dans la table Expression de la BD t*ranscriptome* (attribut 'rang').

Il peut être choisi sous deux formes :
- un menu déroulant permet de sélectionner un des quatre niveaux d'expression que nous avons définis : on considère un niveau élevé pour un rang supérieur ou égal à 75, un niveau moyen de 50 à 75, un faible de 25 à 50 et un très faible en dessous de 25.
- deux zones de texte permettent sinon à l'utilisateur de définir lui-même son intervalle de rangs. S'il est intéressé par un niveau très élevé par exemple, il pourra choisir des rangs supérieurs à 90.

Le troisième champ permet de sélectionner les échantillons d'intérêt en fonction du tissu et/ou de la pathologie étudiée. Pour permettre une sélection non-restrictive, il s'agit de deux zones de texte où l'utilisateur peut entrer des mots-clés séparés par des virgules.

L'application recherchera ces mots-clés dans le dictionnaire (tables *Dico* et *Dico_GSM*) et fera le lien avec les échantillons concernés (attribut 'GSM').

Ainsi, seuls les 'id' des enregistrements de la table *Expression* concernant ces 'GDS', ces 'GSM' et ces 'rangs' seront retenus pour le résultat (figure 19).

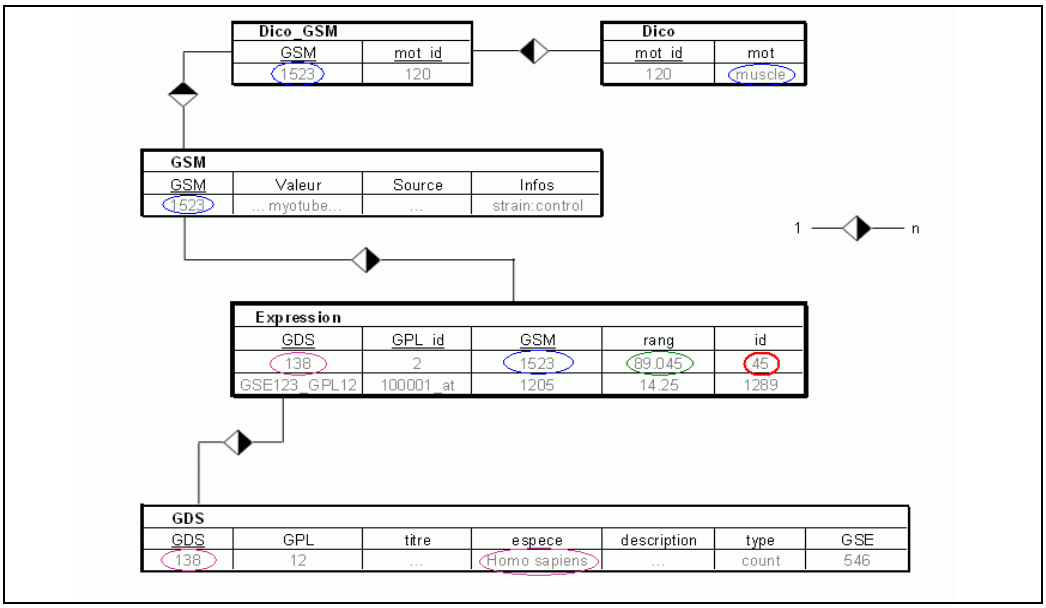


Figure 19 : Schéma du module de requête Transcriptome
En bleu la sélection des échantillons. En rose la sélection des expériences. En vert, la vérification du rang. En rouge, le lien avec l'identifiant unique 'id' de MADGENE

Enfin, il est utile de préciser un pourcentage d'expériences dans lesquelles ce résultat est vérifié afin de s'assurer de sa fiabilité. Seuls les gènes ayant ce niveau d'expression dans cette espèce, ce tissu et/ou pathologie et pour ce pourcentage d'expériences seront retenus dans le résultat final.

C’est pourquoi une zone de texte (avec une valeur par défaut à 50%) autorise l’utilisateur à définir ce pourcentage. Ceci lui permet de s’assurer de la qualité du résultat de sa requête.

De même que pour la partie génome, il est proposé à l’utilisateur de faire un commentaire sur sa requête.

3. Module ChIP-chip

Cette partie doit permettre de trouver les gènes régulés par un Facteur de Transcription, dans un tissu et/ou pathologie, et à un seuil donné (figure 20).

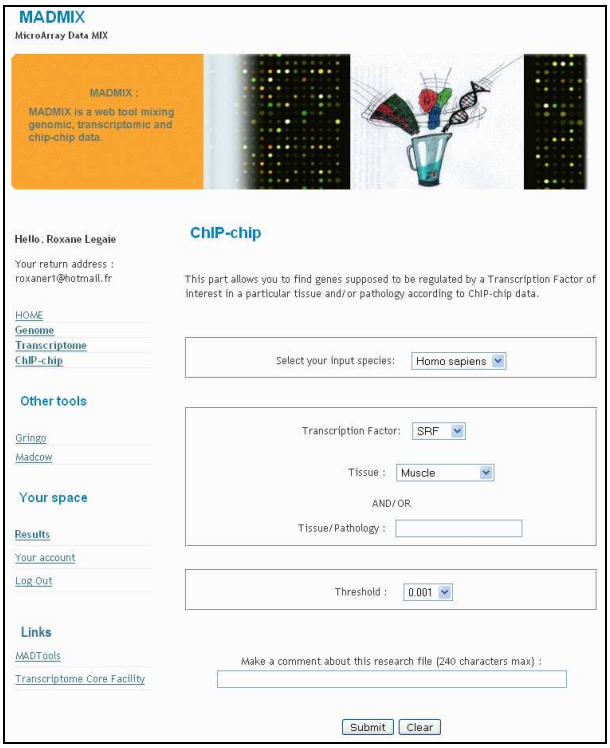


Figure 20 :
Capture d'écran du formulaire 'ChIP-chip'

De même que pour la partie Transcriptome, la sélection de l’espèce se fait à l’aide d’un menu déroulant dynamique qui interroge en temps réel la BD *chip_chip* (attribut ‘espece’ de la table *Experience*) afin de proposer seulement les espèces des expériences disponibles au moment de la requête. L’application listera les expériences (‘idExp’) correspondant à l’espèce voulue.

La sélection du facteur de transcription d’intérêt se fait également à l’aide d’un menu déroulant dynamique qui interroge en temps réel la table *Experience* (attribut ‘FT’).

Le choix des échantillons est une étape cruciale pour la pertinence des résultats de la requête utilisateur. J’ai donc choisi de proposer deux types de sélection :
- Un menu déroulant dynamique : il interroge la table *Echantillon* pour récupérer les ‘keywords’ entrés lors de l’ajout d’une expérience (et donc des échantillons) dans la base.
- Un champ texte : Il utilise le dictionnaire par le même procédé que pour la partie Transcriptome.

Ces deux champs permettent donc de sélectionner les échantillons (‘idEch’) concernés par la requête. Si les deux sont renseignés, les échantillons sélectionnés sont ceux qui contiennent à la fois le mot-clé demandé et les mots entrés dans le champ texte.

Ensuite, un menu déroulant permet de sélectionner le seuil à partir duquel les gènes retenus doivent être considérés positifs. Les quatre seuils pertinents calculés lors de l'analyse des données sont proposés : 0.01%, 0.005%, 0.003% et 0.001%. Le seuil le moins stringent est celui par défaut (0.01%).

L'ensemble de ces champs permet de sélectionner les enregistrements de la table *Chip*. Seuls les 'id' de ceux correspondant à l'une des expériences listées ('idExp'), à l'un des échantillons sélectionnés ('idEch'), au FT demandé ('FT') et ayant l'attribut 'seuil positif' inférieur ou égal à celui sélectionné seront retenus (figure 21).

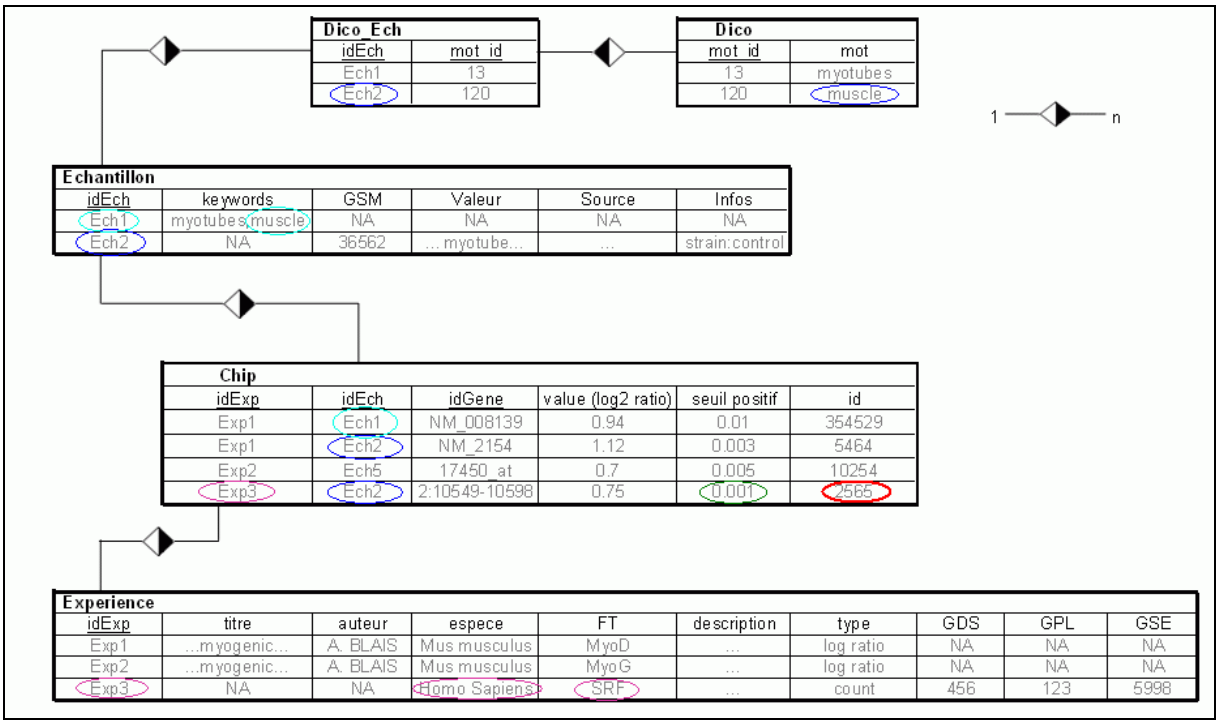


Figure 21 : Schéma du module de requête ChIP-chip
En bleu la sélection des échantillons : mots-clés en bleu clair, dictionnaire en bleu foncé.
La sélection des expériences via l'espèce en rose et via le facteur de transcription en violet.
En vert, la vérification du seuil. En rouge, le lien avec l'identifiant unique 'id' de MADGENE

Enfin, de même que pour les autres parties, il est proposé à l'utilisateur de faire un commentaire sur sa requête.

4. Association aux MADTOOLS

MADMIX ayant pour vocation d'être ajouté par la suite aux MADTOOLS déjà disponibles, un espace de travail commun à MADMIX, MADCOW et GRINGO a dû été créé.

De plus, les résultats en sortie de MADMIX doivent être stockés et compatibles avec les résultats de ces autres outils afin d'en permettre l'intersection et l'union notamment.

a. Compte utilisateur

Pour réaliser cet espace de travail commun, il a été nécessaire de créer un compte utilisateur. Celui-ci s'avère indispensable pour répondre aux attentes de l'outil : la nécessité de stocker les résultats et la volonté d'interagir sur ces derniers.

La page d'accueil de l'outil propose donc en premier lieu à l'utilisateur de s'identifier à l'aide de ses coordonnées professionnelles. Ces informations sont stockées dans une BD commune aux différents outils. Un login et un mot de passe sont créés et envoyés par un mail de confirmation d'inscription.

Lors de l'identification de l'utilisateur à l'aide de son login et de son mot de passe, un cookie contenant ces informations est créé pour une durée limitée. Afin de garantir la confidentialité, son contenu est crypté par un moyen irréversible. La méthode de cryptage utilisée est le md5.

b. Fichiers résultats

MADCOW et GRINGO sortent en résultat des listes de gènes. Chaque liste est stockée sous forme de fichier XML afin de pouvoir récupérer rapidement les informations et de faire facilement des opérations relationnelles (intersection, union…) sur ces listes.

La même méthode a donc été utilisée pour MADMIX, ce qui permet d'effectuer ce type d'opération entre fichiers de résultats d'outils différents.

Chaque requête effectuée sur l'un ou l'autre des modules de MADMIX donne en résultat une liste d'identifiants MADGENE de gènes ('id'). Ceux-ci sont transformés à l'aide d'une requête simple sur la BD *genome* en identifiants 'Symbol' et 'Name'.

Figure 22 : Structure type d'un fichier de résultat en sortie de MADMIX au format XML

Ces informations, ainsi que la requête d'entrée, et des informations supplémentaires en fonction de l'outil utilisé, sont mémorisées dans un fichier XML créé en utilisant le langage DOM. Ce fichier est stocké dans le répertoire de travail de l'utilisateur : un dossier unique sur le serveur créé lors de l'enregistrement de l'utilisateur.

De plus le document XML présente une structure similaire pour tous les outils. La balise 'tool' au début du fichier permettra de connaitre l'outil à partir duquel le résultat a été obtenu lors de la visualisation du résultat : 'Madmix_genome', 'Madmix_transcriptome' ou 'Madmix_chipchip' (figure 22).

L'ensemble des résultats, obtenus à partir des différents outils, est visible à partir de la même page web de l'espace utilisateur : la page 'Results'. Celle-ci permet à l'utilisateur de regarder le contenu de ses fichiers résultats, de les filtrer à l'aide d'ontologies, de les supprimer ou de les télécharger. Elle propose également l'intersection et l'union des différents fichiers.

Pour la visualisation, un simple analyseur syntaxique XML permet de récupérer pour chaque gène de la liste résultat, les valeurs des balises 'symbol' et 'name'. Ensuite l'application affiche le nombre de gènes présents dans la liste et pour chacun, son symbole et son nom.

Le filtrage permet de réduire la liste de gènes à l'aide des ontologies GO [14]. L'utilisateur peut soit sélectionner seulement les gènes qui correspondent à une ontologie voulue (ex : 'Transcription Factor Activity'), soit à l'inverse les supprimer.

Pour cela, une BD nommée *go* (déjà présente sur le serveur) répertorie pour chaque gène, les ontologies auxquelles il appartient. Une simple requête permet alors de faire le tri. La nouvelle liste de gènes ainsi formée peut alors être sauvegardée dans l'espace de travail sous la forme d'un fichier XML. La balise 'tool' aura pour valeur 'Madmix_filter'.

L'intersection et l'union sont possibles entre plusieurs fichiers sélectionnés, quelque soit leur outil de provenance.

Un programme récupère les listes d'identifiants de gènes (balises 'madid') de chacun des fichiers sélectionnés puis en cherche l'intersection ou l'union. Pour l'intersection, le nombre d'occurrences de chaque gène est également calculé.

La liste d'identifiants de gènes qui en résulte est transformée en identifiants 'symbol' et 'name' afin de créer un nouveau fichier XML. Celui-ci sera également stocké dans le répertoire de travail utilisateur. La balise 'tool' aura pour valeur 'Madmix_intersection' ou 'Madmix_union' en fonction de l'opération qui a été effectuée.

c. Interface

Afin d'obtenir une interface conviviale et homogène avec celle de MADTOOL, la structure HTML de MADCOW a été utilisée lors de la création de MADMIX. De même, j'ai adapté la feuille de style CSS utilisée pour MADCOW pour la lui appliquer.

III. RESULTATS

A. Les bases de données

La particularité de mon outil est de permettre de « mixer » des données de génomique, de transcriptomique et de ChIP-chip.

Pour cela, la BD *genome* recense les positions génomiques de l'ensemble des objets du génome Humain : gènes, exons, transcrits et ilots CpG associés. Celle-ci devra être complétée au fur et à mesure avec les autres espèces par l'équipe, et mise à jour régulièrement avec les nouvelles versions d'Ensembl.

La BD *transcriptome* stocke les résultats, sous forme de rangs, des expériences de puces à ADN effectuées sur le muscle. Elle permet d'obtenir pour un gène, dans un échantillon testé au cours d'une expérience donnée, une évaluation de son niveau d'expression. Elle stocke également les informations concernant les expériences ('GDS') et les échantillons ('GSM'). Elle devra être complétée avec des expériences faites sur d'autres tissus que le muscle à l'aide du pipeline de programmes que j'ai conçu.

La BD *chip_chip* gère les résultats, après normalisation et analyse, d'expériences de ChIP-chip faites sur du muscle. Elle permet de récupérer les gènes considérés positifs pour un FT donné à un seuil donné, dans un échantillon donné au cours d'une expérience donnée. Elle stocke elle aussi les informations concernant les expériences et les échantillons testés. Elle sera également mise-à-jour à l'aide d'un autre pipeline que j'ai créé.

Le lien entre ces trois BD est l'identifiant unique de MADGENE : 'id'. Ainsi, quelle que soit l'information recherchée et la BD interrogée, le résultat se rapporte à un enregistrement unique dans la BD de MADGENE. Les résultats sont alors cohérents et compatibles pour permettre leur « mixage ». De plus, cela permet d'obtenir des informations supplémentaires sur un gène donné : ses différents identifiants et ses orthologues.

B. L'interface

L'objectif final de MADMIX est d'être par la suite intégré aux MADTOOLS. C'est pourquoi l'interface a été conçue de façon cohérente et homogène avec ces derniers.

De plus, elle permet une interaction entre les résultats obtenus à partir de chacun des modules de MADMIX mais aussi avec ceux obtenus à partir de GRINGO et MADCOW.

1. *L'accès*

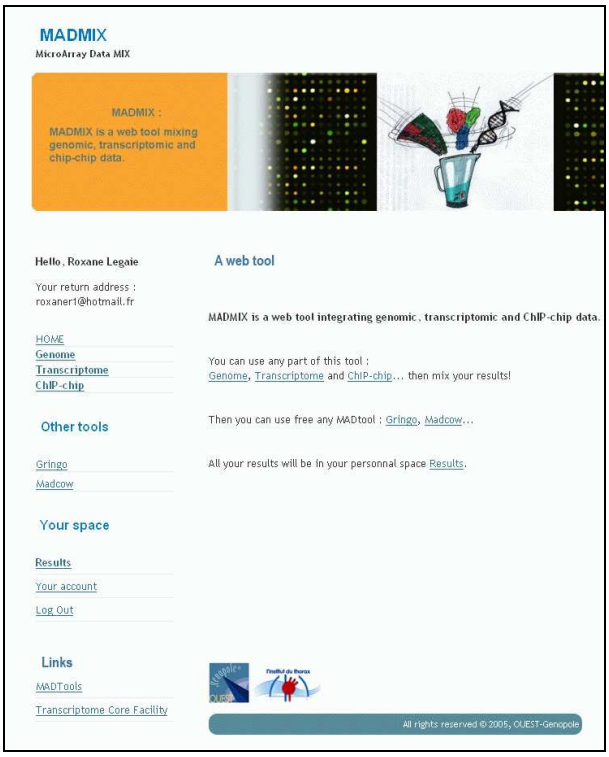


Figure 23 :
Capture d'écran du formulaire d'accès

Tout d'abord, afin de permettre à l'utilisateur de passer de l'un à l'autre de ces outils sans avoir à se reconnecter, le compte utilisateur est commun à MADMIX, GRINGO et MADCOW.

Le menu de gauche propose l'accès aux trois modules de MADMIX ainsi qu'aux autres MADTOOLS (figure 23).

2. *Les modules de MADMIX*

La page d'accès au module 'Genome' permet à l'utilisateur de rechercher les gènes présents sur le chromosome de son choix et/ou ayant un ilot CpG à une certaine distance. La requête utilisateur est récupérée à l'aide du formulaire et est transmise à une application qui se charge d'effectuer la ou les requêtes MySQL sur la BD g*enome*.

Le formulaire du module 'Transcriptome' permet de récupérer les gènes présentant tel niveau d'expression dans tel tissu et/ou pathologie. De la même façon, la requête est traitée par l'application PHP sous-jacente qui va interroger la BD *transcriptome*.

Enfin la partie 'Chip-chip' interroge la BD *chip_chip* pour permettre à l'utilisateur de trouver les gènes cibles à un seuil donné, dans un tissu et/ou une pathologie, du FT d'intérêt.

3. L'espace de travail

L'innovation majeure apportée par le regroupement d'outils sous un seul espace de travail est la possibilité pour l'utilisateur de formuler tous types de requêtes, de la plus simple à la plus complexe : associer des informations de génomique, de transcriptomique et de ChIP-chip (MADMIX), avec celles d'annotations fonctionnelles (GRINGO) et de corrélations inter-géniques (MADCOW). De plus, il sera possible d'accueillir de nouveaux outils présentant le même type de résultats par la suite.

Les résultats sont donc stockés dans un même répertoire et sont visibles à partir d'une seule et même page : 'Results'. Elle se présente sous la forme d'un tableau qui propose plusieurs accès à ces fichiers (figure 24).

Figure 24 :
Capture d'écran de l'espace de travail 'Results'

- *La visualisation (icône View)*

Ce lien permet d'observer la liste de gènes correspondant au résultat de la requête utilisateur. Pour chaque gène de la liste sont affichés son symbole et son nom. De plus, l'interface propose :

- un lien à partir du symbole vers la carte d'identité du gène (MADSENSE)
- un lien vers MADGENE afin d'obtenir les autres identifiants disponibles (Other Ids available) pour ce gène, ainsi que ses orthologues si l'on choisit une espèce différente que celle de la requête première
- un lien vers les ontologies (GO) dans lesquelles il est recensé. Chaque GO_id étant lui-même un lien vers le site QuickGO [18] afin de pouvoir notamment observer où se situe cette ontologie dans l'arborescence des ontologies.

Si cette liste parait intéressante à l'utilisateur, il a la possibilité de la sauvegarder au format texte : bouton 'Save in text format'. Le fichier texte est mémorisé dans le même répertoire de travail et apparait à la fin du tableau des fichiers résultats.

L'utilisateur peut alors le télécharger et interroger avec les outils GRINGO ou MADCOW.

Enfin, le lien 'GO Statistics' permet de déterminer quelle est l'ontologie surreprésentée dans la liste de gènes afin d'avoir une idée des processus biologiques auxquels ils participent et des composants cellulaires auxquels ils appartiennent.

Pour cela un programme, créé par un ingénieur de l'équipe, effectue une annotation fonctionnelle par méthode statistique [19] : il évalue la représentation des différents termes d'ontologies de la liste de gènes par rapport au reste des gènes du génome de la même espèce en s'appuyant sur les dénominations présentes dans la BD *go*.

- *La requête (colonnes Input et Date)*

Ce sont des rappels sur les arguments d'entrée de la requête et la date où elle a été effectuée afin d'aider l'utilisateur à retrouver ses fichiers de résultats rapidement.

- *Le commentaire (colonne Comment)*

Cela permet de visualiser le commentaire laissé par l'utilisateur lors de sa requête ou apporte diverses informations sur les fichiers concernés.

- *Le filtrage (icône Filter)*

Ce bouton permet de filtrer le fichier de résultat selon les ontologies. La case 'select' permet de ne conserver que les gènes de la liste appartenant à l'ontologie voulue. La case 'cut' propose elle de les supprimer de la liste. Cette nouvelle liste de gènes peut être sauvegardée à son tour.

- *La suppression et le téléchargement (icône Delete et Download)*

Ces boutons permettent de supprimer ou de télécharger le fichier de résultat voulu.

- *L'intersection et l'union*

En cochant les fichiers d'intérêt à l'aide des cases 'Select', l'utilisateur peut ensuite en demander l'intersection ou l'union. Il en résulte un nouveau fichier dans l'espace de travail sous l'intitulé 'Intersection' ou 'Union' selon l'opération demandée.

Celui-ci peut alors être traité comme tous les autres fichiers résultats : visualisation, sélection, filtrage, etc.

4. *Exemple d'utilisation*

MADMIX permet d'effectuer des requêtes très complexes du type : « Trouver les gènes ayant un ilot CpG à moins de 1 kb, fortement exprimés dans le muscle et positifs en ChIP-chip pour le facteur de transcription SRF ». Voici un exemple d'utilisation de MADMIX pour une telle requête utilisateur :

- *Interrogation du module Genome*

A l'aide de la page d'accès 'Genome' de l'interface, on sélectionne l'espèce (Homo sapiens) et la distance minimale à un ilot CpG (1 kb). Afin d'effectuer une recherche sur tout le génome humain, aucun chromosome particulier n'est sélectionné (figure 16).

L'envoi de la requête renvoie sur une page web indiquant qu'il existe 8975 gènes correspondant. Le fichier résultat est stocké dans l'espace utilisateur, accessible via le lien 'Results' sous l'intitulé 'genome'. L'icône 'View' permet de le visualiser (figure 25).

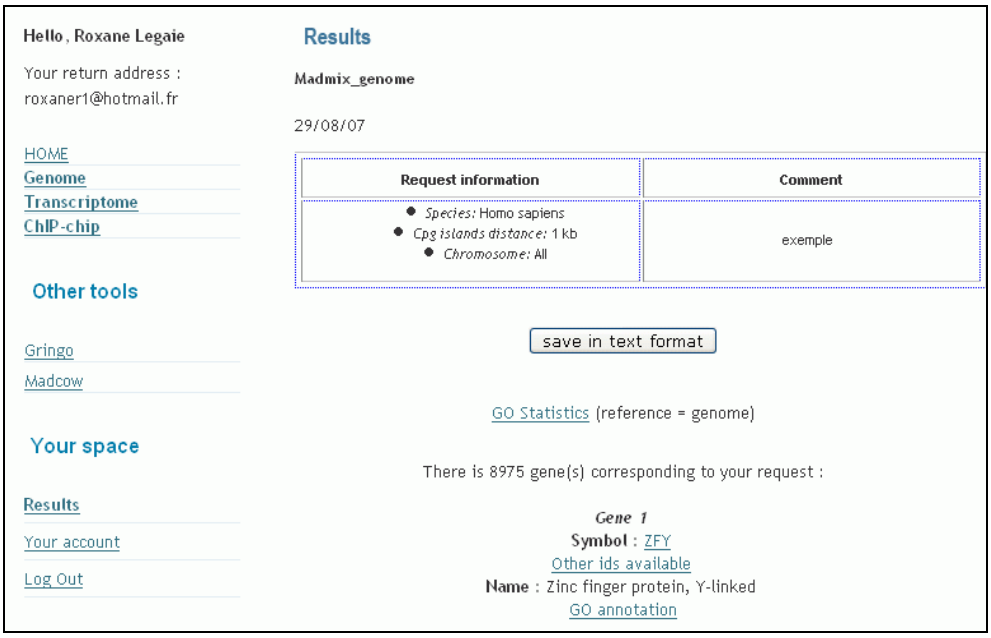


Figure 25 :
Capture d'écran de la visualisation du résultat d'une requête effectuée sur le module 'Genome'

- *Interrogation du module Transcriptome*

Sur la page d'accès au module 'Transcriptome', on sélectionne l'espèce (Homo sapiens) et le niveau d'expression (élevé – *high*). Le tissu d'intérêt (muscle) est entré dans le champ texte prévu à cet effet, sans préciser de pathologie particulière. Pour plus de fiabilité des résultats, il est possible de demander un pourcentage très élevé (90%) d'expériences où est vérifiée la requête (figure 18).

La page résultante indique qu'il existe 2118 gènes correspondant à cette requête. Le fichier résultat est stocké dans l'espace utilisateur sous l'intitulé 'transcriptome' (figure 26).

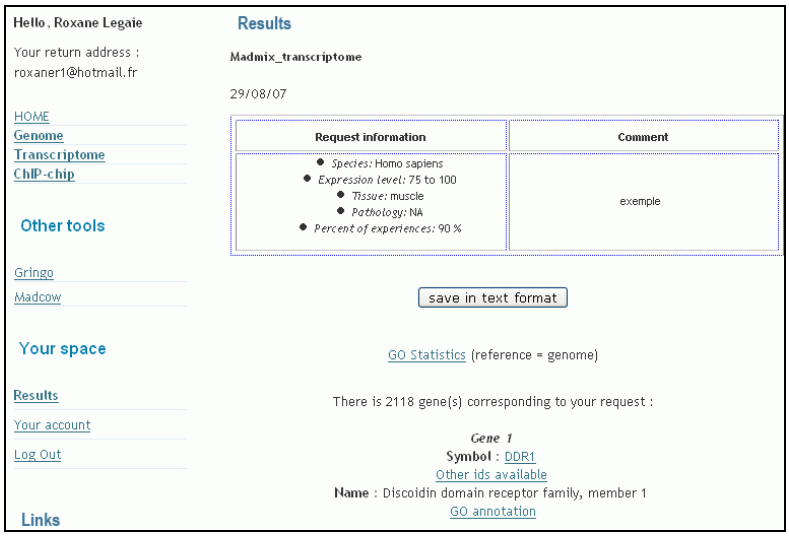


Figure 26 :
Capture d'écran de la visualisation du résultat d'une requête effectuée sur le module 'Transcriptome'

- *Interrogation du module ChIP-chip*

Enfin, le module 'ChIP-chip' est interrogé (figure 20). L'espèce est sélectionnée (toujours Homo sapiens afin d'obtenir des résultats comparables) ainsi que le FT d'intérêt (SRF). Les menus déroulants permettent de sélectionner le tissu (muscle) et le seuil (0.001%).

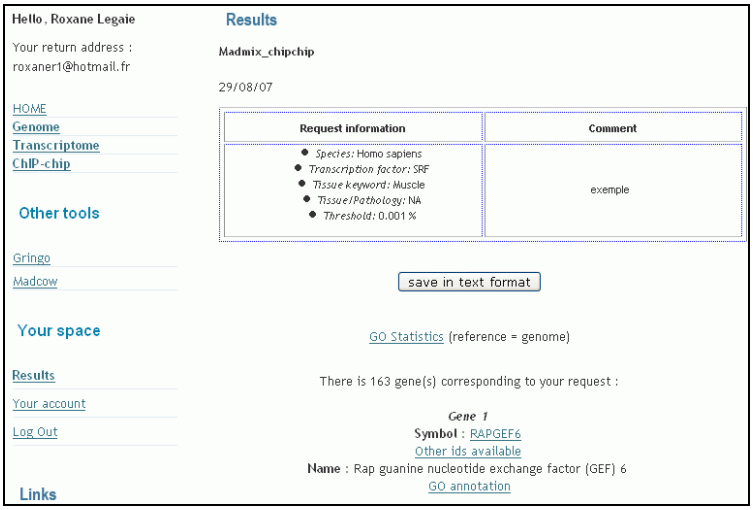


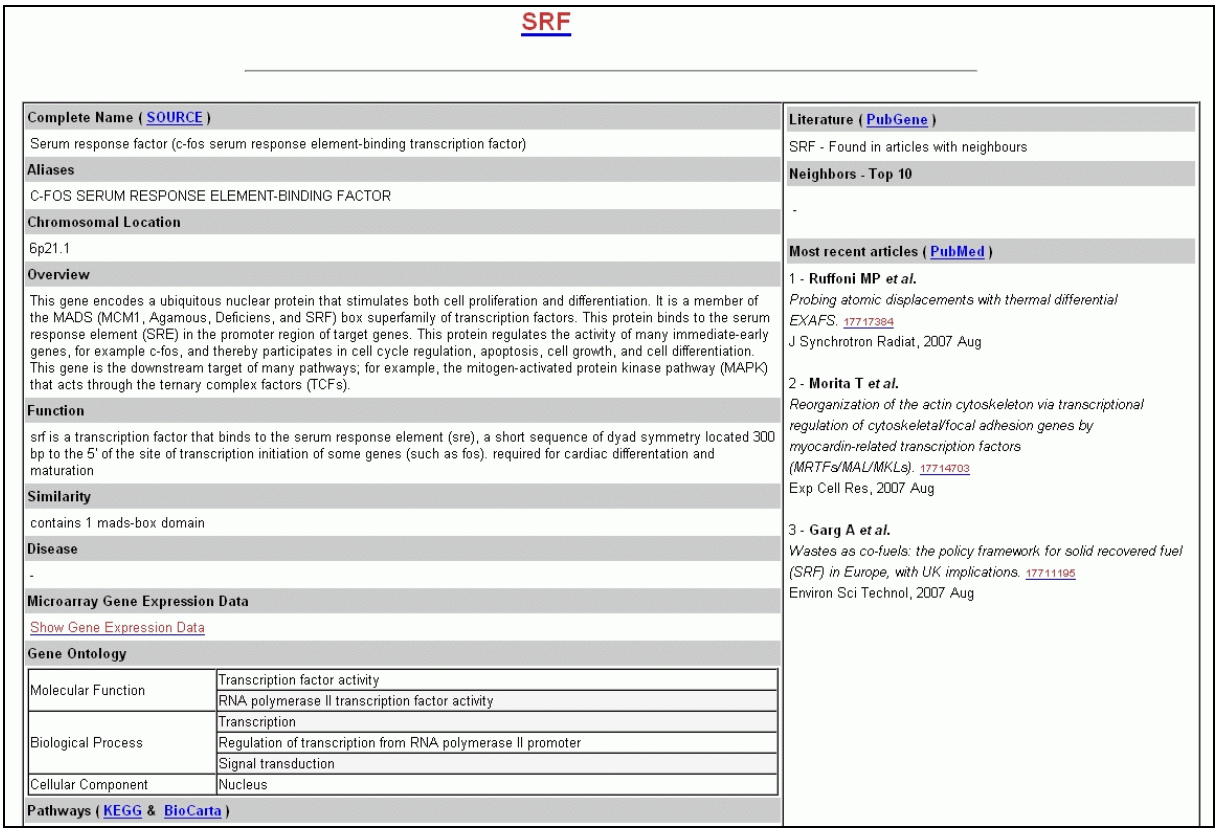
Figure 27 :
Capture d'écran de la visualisation du résultat d'une requête effectuée sur le module 'ChIP-chip'

La page suivante indique qu'il existe 163 gènes correspondants à cette requête. Le fichier résultat est stocké dans l'espace utilisateur sous l'intitulé 'chipchip'(figure 27).

A l'aide de l'icône 'Filter', on peut procéder ensuite au filtrage de la liste des 8975 gènes contenue dans le fichier résultat 'genome' afin de ne conserver par exemple que les FT. En position 12 de cette nouvelle liste, le FT 'SRF' sélectionné pour la requête ChIP-chip apparait.

- Lien avec MADSENSE

En cliquant sur le lien de son symbole, une nouvelle fenêtre affiche sa « carte d'identité » proposée par l'outil MADSENSE. On y trouve notamment des informations sur sa localisation génomique, sa fonction et les ontologies (Gene Ontology) auxquelles il appartient (figure 28).



SRF

Complete Name (SOURCE)

Serum response factor (c-fos serum response element-binding transcription factor)

Aliases

C-FOS SERUM RESPONSE ELEMENT-BINDING FACTOR

Chromosomal Location

6p21.1

Overview

This gene encodes a ubiquitous nuclear protein that stimulates both cell proliferation and differentiation. It is a member of the MADS (MCM1, Agamous, Deficiens, and SRF) box superfamily of transcription factors. This protein binds to the serum response element (SRE) in the promoter region of target genes. This protein regulates the activity of many immediate-early genes, for example c-fos, and thereby participates in cell cycle regulation, apoptosis, cell growth, and cell differentiation. This gene is the downstream target of many pathways; for example, the mitogen-activated protein kinase pathway (MAPK) that acts through the ternary complex factors (TCFs).

Function

srf is a transcription factor that binds to the serum response element (sre), a short sequence of dyad symmetry located 300 bp to the 5' of the site of transcription initiation of some genes (such as fos). required for cardiac differentation and maturation

Similarity

contains 1 mads-box domain

Disease

-

Microarray Gene Expression Data

Show Gene Expression Data

Gene Ontology

Molecular Function	Transcription factor activity
	RNA polymerase II transcription factor activity
Biological Process	Transcription
	Regulation of transcription from RNA polymerase II promoter
	Signal transduction
Cellular Component	Nucleus

Pathways (KEGG & BioCarta)

Literature (PubGene)

SRF - Found in articles with neighbours

Neighbors - Top 10

-

Most recent articles (PubMed)

1 - **Ruffoni MP *et al.***
Probing atomic displacements with thermal differential EXAFS. 17717384
J Synchrotron Radiat, 2007 Aug

2 - **Morita T *et al.***
Reorganization of the actin cytoskeleton via transcriptional regulation of cytoskeletal/focal adhesion genes by myocardin-related transcription factors (MRTFs/MAL/MKLs). 17714703
Exp Cell Res, 2007 Aug

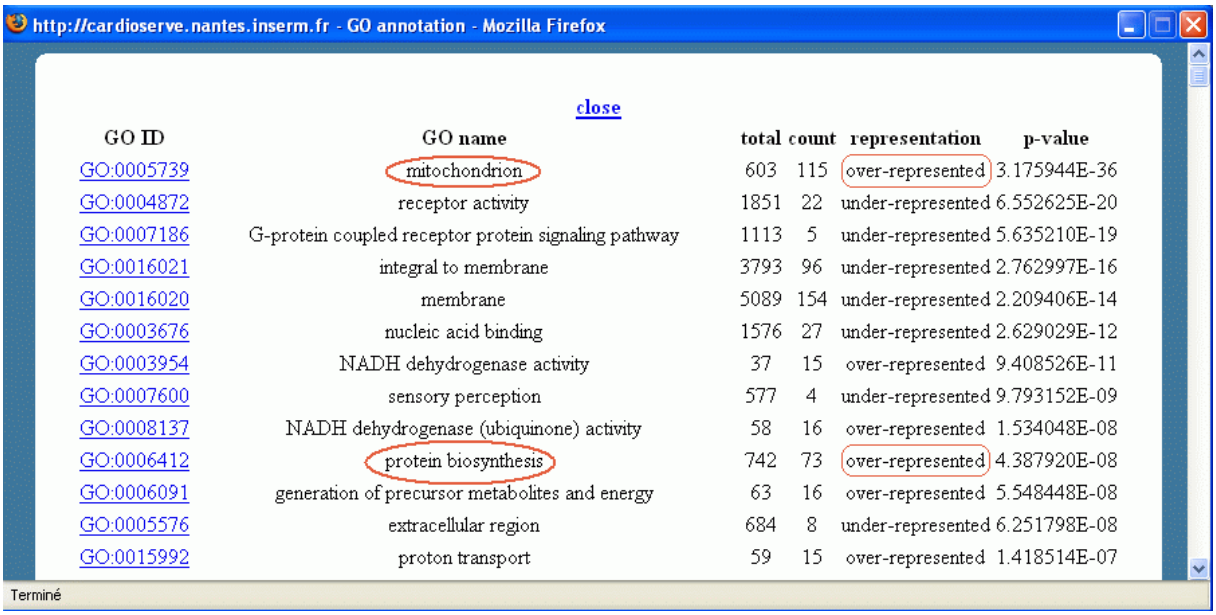
3 - **Garg A *et al.***
Wastes as co-fuels: the policy framework for solid recovered fuel (SRF) in Europe, with UK implications. 17711195
Environ Sci Technol, 2007 Aug

Figure 28 :
Capture d'écran de la carte d'identité de SRF obtenue avec MADSENSE

D'autre part, il est possible de demander l'intersection des trois fichiers résultats obtenus lors de la requête sur MADMIX. Il en résulte un nouveau fichier présentant une liste de 1251 gènes.

- Analyse des ontologies de gènes

Le lien 'GO statistics' permet d'observer les ontologies surreprésentées dans cette nouvelle liste. Ceci permet d'avoir une idée des processus biologiques auxquels participent les gènes correspondants à la requête globale (figure 29).



GO ID	GO name	total	count	representation	p-value
GO:0005739	mitochondrion	603	115	over-represented	3.175944E-36
GO:0004872	receptor activity	1851	22	under-represented	6.552625E-20
GO:0007186	G-protein coupled receptor protein signaling pathway	1113	5	under-represented	5.635210E-19
GO:0016021	integral to membrane	3793	96	under-represented	2.762997E-16
GO:0016020	membrane	5089	154	under-represented	2.209406E-14
GO:0003676	nucleic acid binding	1576	27	under-represented	2.629029E-12
GO:0003954	NADH dehydrogenase activity	37	15	over-represented	9.408526E-11
GO:0007600	sensory perception	577	4	under-represented	9.793152E-09
GO:0008137	NADH dehydrogenase (ubiquinone) activity	58	16	over-represented	1.534048E-08
GO:0006412	protein biosynthesis	742	73	over-represented	4.387920E-08
GO:0006091	generation of precursor metabolites and energy	63	16	over-represented	5.548448E-08
GO:0005576	extracellular region	684	8	under-represented	6.251798E-08
GO:0015992	proton transport	59	15	over-represented	1.418514E-07

Figure 29 : Observation des ontologies statistiquement surreprésentées

L'ontologie surreprésentée avec la plus faible p-value est « mitochodrion ». En cliquant sur le lien de l'identifiant GO correspondant (GO:0005739), on peut observer l'arborescence des ontologies sur le site QuickGO. Il s'agit en fait de l'ontologie 'Composant cellulaire'. Cela signifie que la majorité des gènes de la liste sont localisés dans les mitochondries.

L'ontologie « protein biosynthesis » de la catégorie 'Biological process' est également surreprésentée.

- Lien avec GRINGO

Cette liste de gènes résultant de l'intersection des résultats des trois requêtes peut être sauvegardée au format texte afin d'interroger les autres outils disponibles.

Après avoir téléchargé ce fichier texte, l'accès à l'outil GRINGO se fait directement à l'aide du lien situé dans le menu gauche de l'interface, sans avoir à se reconnecter. GRINGO va permettre d'annoter la liste de gènes en la comparant à des données de puces à ADN classées hiérarchiquement et annotées (figure 30).

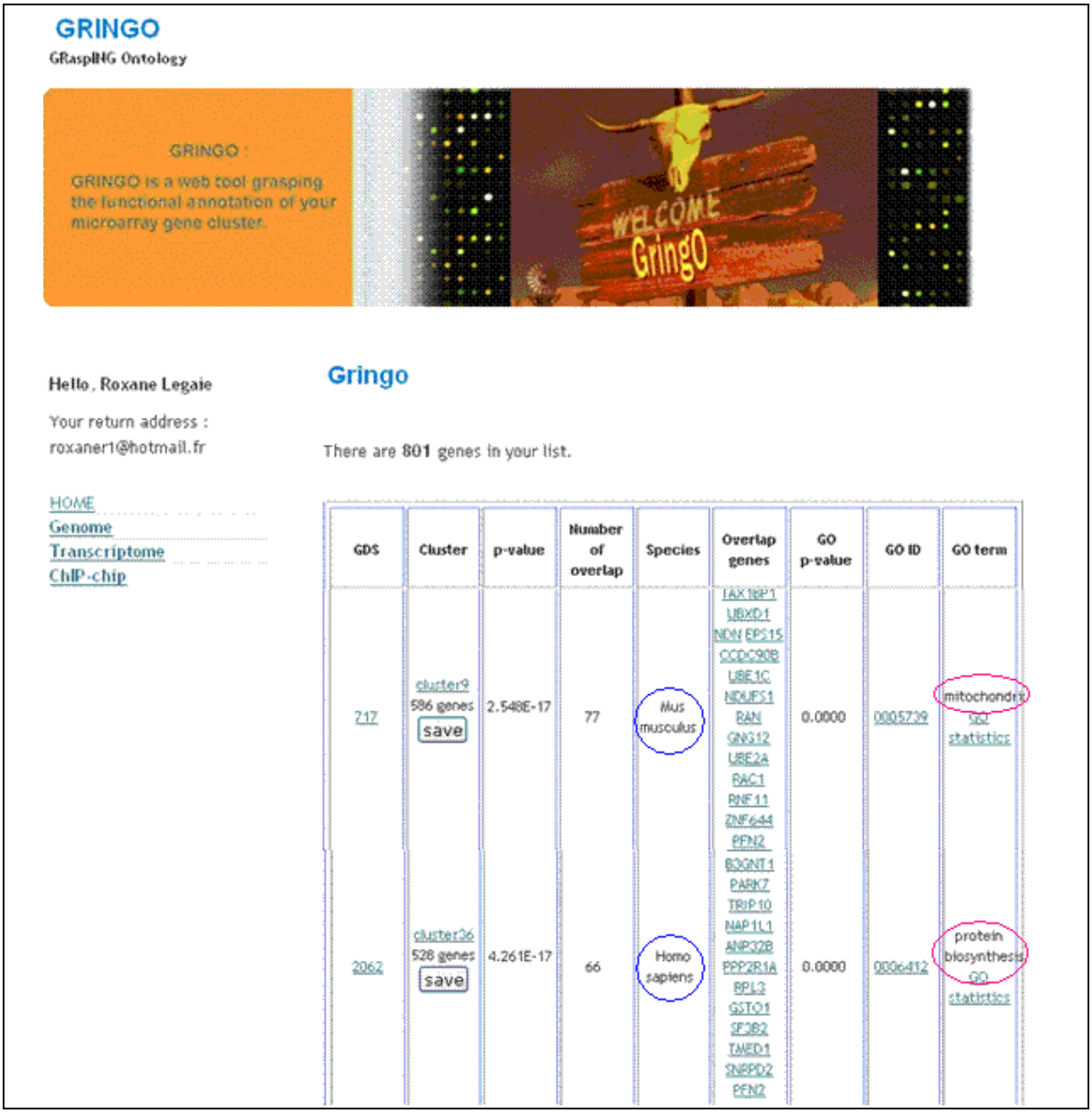


Figure 30 :
Capture d'écran du résultat de GRINGO
En rose, les annotations trouvées. En bleu, les espèces concernées

On observe notamment des groupes de gènes d'expression corrélée (clusters) tirés d'expériences faites chez la souris (*Mus musculus*), annotés « mitchodrion » et « protein biosynthesis », fortement similaires à la liste de gènes initiale. Il est possible de sauvegarder ces clusters dans l'espace de travail et de les traiter comme tous les autres fichiers résultats.

De même des requêtes sur MADCOW peuvent être effectuées à partir des fichiers résultats au format texte de MADMIX ou GRINGO.

L'association de ces trois outils permet à l'utilisateur d'effectuer des requêtes très complexes et très variées.

IV. DISCUSSION

L'objectif de mon stage était de créer un outil pour ajouter de nouvelles fonctionnalités aux bases de données existantes dans mon laboratoire d'accueil. Il s'agissait particulièrement d'intégrer des requêtes de positionnement sur le génome, et des requêtes sur les données expérimentales de puces à ADN et de ChIP-chip. La finalité étant de pouvoir faire des requêtes du type « Sélectionner tous les gènes exprimés dans le muscle, ayant le facteur SRF fixé sur leur promoteur, et situé à moins de 1kb d'un îlot CpG ». Cet outil devant intégrer des données de natures diverses, je l'ai appelé MADMIX.

Son développement a soulevé plusieurs problèmes : la nature hétérogène des données de puces à ADN et de ChIP-chip, la caractérisation de l'expression d'un gène dans un tissu donné, l'évaluation statistique des séquences positives en ChIP pour un FT donné, etc.

Nous avons trouvé des solutions à chacune de ces questions, mais il serait intéressant de tester d'autres stratégies alternatives.

Toutefois, malgré ces difficultés rencontrées, MADMIX est actuellement fonctionnel au laboratoire, et permet de faire les requêtes que nous avions prévues. Le retour des utilisateurs biologistes nous fournira probablement des pistes d'améliorations.

Les principaux problèmes rencontrés et les améliorations à leur apporter :

Pour le module 'Genome', les positions des différents éléments ont été obtenues à l'aide d'Ensembl car la BD préexistante que j'ai complétée avait été créée à l'aide de ces données. En interrogeant une autre base telle que celle de l'UCSC [20], le nombre et les positions des ilots CpG sont légèrement différents. Ceci s'explique par la méthode de détection des ilots qui est propre à chaque site.

Pour plus de cohérence j'ai choisi de conserver les positions données par Ensembl, mais il pourrait être judicieux de confronter les différentes sources de données disponibles.

Le premier obstacle à la création de la BD *transcriptome* a été l'hétérogénéité des données publiques. En effet, plusieurs sites recensent des données de puces à ADN, dont GEO et Array-Express [21]. Les fichiers de GEO, régulièrement utilisés par l'équipe, sont bien formatés et permettent une analyse syntaxique automatique des données brutes. Alors que le site Array-Express propose des fichiers complexes, lourds et parfois mal annotés.

Il a donc été décidé de ne traiter que des données provenant de GEO pour l'objet de mon stage. Toutefois, si certaines expériences ne sont disponibles que sur le site d'Array-Express, il serait utile de trouver un moyen de les analyser de façon automatique.

La sélection des échantillons et expériences d'intérêts pour les requêtes sur le module Transcriptome est un réel problème auquel nous avons trouvé une solution temporaire.
En effet, la description des échantillons et des expériences est librement faite par l'auteur. Il n'existe pas de système de mots-clés permettant de connaitre clairement les tissus correspondant à l'échantillon.

Le dictionnaire permet une sélection rapide des échantillons d'intérêt mais sa fiabilité est limitée. Actuellement, chercher un gène fortement exprimé dans le muscle revient à chercher un gène fortement exprimé aussi bien dans un échantillon de muscle sain que dans un échantillon de muscle atteint de myopathie de Duchenne par exemple.

Pour éviter cela, il faudrait proposer à l'utilisateur de donner également les tissus et/ou pathologies auxquel(le)s ne doivent pas appartenir les échantillons sélectionnés. Il devrait alors indiquer dans sa requête le tissu d'intérêt et préciser de ne pas sélectionner les échantillons de toutes les pathologies qui lui sont associées (ex : muscle sauf myopathie, dystrophie…) et qui ne l'intéressent pas. Ce qui n'est bien-sur pas concevable. D'autant que l'outil inclura bientôt des données d'expériences faites sur de nombreux types tissulaires.

Une solution intermédiaire serait que l'application fasse une première sélection à l'aide du dictionnaire, puis que l'interface propose à l'utilisateur de sélectionner les échantillons et expériences qui l'intéressent à l'aide de leur descriptif (facilement accessible à l'aide des tables *GDS* et *GSM* de la BD *transcriptome*). Toutefois, lorsque des centaines d'expériences seraient sélectionnées par l'application, il serait lourd pour l'utilisateur d'avoir à en faire le tri.

La meilleure solution reste l'annotation manuelle par l'équipe de chacune des expériences de Transcriptome (41 disponibles sur GEO pour le muscle, mais plus de 2000 au total). Chaque expérience et chaque échantillon serait alors référencé par des mots-clés que l'on pourrait proposer dans un menu déroulant dynamique. C'est probablement vers cette solution que l'équipe se dirigera.

Enfin il serait intéressant d'apporter une fonction complémentaire au module Transcriptome. En effet, actuellement celui-ci permet de trouver les gènes présentant un niveau d'expression particulier. Mais l'intérêt principal des puces à ADN étant d'établir des profils d'expression de gènes, il serait intéressant que ce module propose également de trouver les gènes présentant une variation de leur niveau d'expression entre deux conditions données. On pourrait alors chercher par exemple les gènes surexprimés dans des échantillons de muscle atteint de myopathie de Duchenne versus des échantillons de muscle sain.

Ceci pourrait venir d'un module d'analyse automatique des données de puces à ADN (MADPRO : MicroArray Data Processing) en cours de développement dans l'équipe, qui détecterait les gènes différentiellement exprimés dans chaque expérience.

Le problème des données hétérogènes a également été rencontré pour le module 'ChIP-chip'. Peu de données étant actuellement publiques, nous avons contacté des auteurs afin d'obtenir leurs données brutes. Ces données fournies par les auteurs eux-mêmes ont dues être analysées et traitées au cas par cas avant de lancer le pipeline de programmes dessus.

Par ailleurs, certaines puces utilisées pour le ChIP-chip sont constituées de promoteurs de gènes. Il est alors facile d'attribuer un signal positif d'une sonde à un gène particulier. Mais d'autres puces sont constituées de sondes réparties régulièrement sur l'ensemble du génome (« Tiling Array »). Dans ce cas l'attribution d'une sonde à un gène est une étape délicate.

Nous avons choisi la solution la plus évidente : associer la sonde au gène connu le plus proche. Mais cette stratégie soulève des questions de pertinence lorsque le gène est très loin (100 kb) ou que la sonde est située entre deux gènes.

Les problèmes soulevés par la sélection des échantillons sont apparus également pour les expériences de ChIP-chip. Compte-tenu du faible nombre de données disponibles et du fait que celles-ci seront ajoutées au fur et à mesure par l'équipe, j'ai pu mettre en place le système de mots-clés ('keywords') proposé pour le Transcriptome.

Toutefois, la sélection par mots-clés limite l'étendue de la sélection et peu parfois biaisée celle-ci. En effet, les échantillons notés 'smooth muscle' (muscle lisse) par exemple ne seraient pas sélectionnés par l'application si l'utilisateur sélectionne juste 'muscle' dans le menu déroulant. Pour éviter cette erreur, le système du dictionnaire a été ajouté au système de mots-clés, ce qui semble être un bon compromis.

Par ailleurs, l'utilisation d'un tel menu déroulant dépend fortement du bon remplissage de l'attribut 'keywords' (table *Echantillon*) lors de l'ajout d'une nouvelle expérience dans la base.

Le ChIP-chip donne parfois des signaux très proches du bruit de fond de l'expérience. Il a donc été assez difficile de mettre au point une méthode statistique de détection des signaux positifs. La solution que nous avons choisi est simple mais pourrait être améliorée.

L'idée est que l'on a les signaux de l'ADN avant et après immunoprécipitation avec le FT. Ces signaux sont très similaires mais se distinguent par un enrichissement (et jamais un appauvrissement) en ADN correspondant aux sites de fixation du FT sur la chromatine. Nous avons donc calculé le log du rapport ADN après immunoprécipitation sur ADN avant immunoprécipitation. Nous avons utilisé les distributions des valeurs négatives (absence d'hybridation) pour déterminer les seuils de significativité, et nous avons appliqué ces seuils aux valeurs positives en changeant le signe des seuils.

La difficulté rencontrée lors de la création de l'interface web a été de proposer à l'utilisateur une interface identique à celle des MADTOOLS afin de permettre une navigation agréable et une utilisation facile.

Pour cela j'ai dû apprendre le langage CSS, récupérer les feuilles de style utilisées pour MADCOW notamment et les modifier afin de les appliquer à mon outil.

La structure HTML de mon interface à également été réalisée en fonction de ces feuilles de styles. J'ai ainsi obtenu une interface tout à fait semblable à celle des MADTOOLS.

La rapidité d'exécution d'une application sous-jacente à un site web est une qualité importante. Les bases de données des différents modules ont été créées afin d'optimiser les requêtes utilisateurs. Toutefois le module Transcriptome est un peu plus long en exécution que les autres. Ceci s'explique par la taille de la BD (15 millions d'enregistrements).

Lorsque celle-ci sera complétée par des expériences supplémentaires, il sera surement nécessaire de mettre au point un système de file d'attente pour les requêtes. Ce système est déjà utilisé pour MADCOW et GRINGO. L'utilisateur doit alors attendre de recevoir un mail de confirmation pour aller voir le fichier de résultat dans son espace utilisateur.

Ce problème sera également rencontré pour le module ChIP-chip lorsque le nombre d'expériences disponibles sera devenu important.

Malgré tous ces problèmes, MADMIX s'est révélé très pertinent pour répondre à une requête complexe du type « Trouver les gènes ayant un ilot CpG à moins de 1 kb, fortement exprimés dans le muscle et positifs en ChIP-chip pour le facteur de transcription SRF ».

En effet, les résultats obtenus sont en accord avec ce que l'on connaît de SRF : le facteur de transcription SRF est effectivement connu pour participer activement à la biosynthèse protéique et notamment celle des protéines mitochondriales, dans les tissus musculaires [22].

V. CONCLUSION

Ce stage aura été une expérience professionnelle et personnelle très enrichissante. Tout d'abord, j'ai pu participer à la prise de décision lors de la conception de l'outil. Ensuite, au cours de sa réalisation, j'ai eu l'occasion d'appliquer et d'améliorer mes compétences en Informatique : la programmation, la création et la gestion de bases de données, ainsi que la création d'un site web. J'ai également pu approfondir mes connaissances en Bioinformatique et notamment en analyse statistique de données. Enfin, MADMIX résulte d'un travail en équipe important, en particulier pour l'intégration de celui-ci aux MADTOOLS.

MADMIX est un outil bioinformatique actuellement disponible via Internet à l'url http://cardioserve.nantes.inserm.fr/developpement/roxane/. Celui-ci n'étant pas encore public, merci de ne pas communiquer cette adresse.

L'innovation apportée par MADMIX repose sur la mixité des informations : diverses sources (Ensembl, GEO, données non-publiques), diverses technologies (Séquençage de génomes, puces à ADN, ChIP-chip) et divers sous-domaines (Génomique, Transcriptomique et Régulation génique).

Les résultats obtenus sont concluants et montrent que l'on peut étudier des processus biologiques complexes en mixant des données de sources et de technologies différentes.

On peut, par exemple, croiser des données de Transcriptome et de ChIP-chip afin de définir le type de régulation (activation ou répression) subit par un gène. En effet, si l'on observe que tel gène régulé par tel FT présente un niveau d'expression très élevé dans une pathologie donnée, on peut supposer une activation de son expression par le biais du FT mis en cause. Cette observation ouvre alors de nouvelles voies de recherche thérapeutiques et apporte des informations complémentaires sur la pathologie étudiée.

Une des perspectives de l'équipe est de compléter MADMIX avec différentes fonctions : déterminer les positions des promoteurs des gènes appartenant à une liste résultat et effectuer de la découverte de motifs sur ces régions promotrices. Ceci afin de mettre à jour des complexes de FT impliqués dans les réseaux de régulation et de vérifier par la suite ces hypothèses à l'aide des techniques de ChIP-chip en cours de mise au point sur la plateforme.

VI. REFERENCES BIBLIOGRAPHIQUES

[1] Bertucci F. et al. Profils d'expression génique et puces à ADN dans le cancer du sein : intérêt pronostique. Bulletin du cancer 2002; 89: 6.

[2] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 2001; 409: 860-921.

[3] Human Genome Project
[http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml]

[4] Bertucci F. et al. Puces à ADN : technologie et applications. Bulletin du cancer 2001;88:3.

[5] Kirmizis and Farnham. Experimental approaches that aid in the identification of transcription factor target genes. Experimental Biology and Medicine 2004; 229(8): 705.

[6] Weinmann A.S. et al. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. Genes Dev 2002; 16: 235–244.

[7] Ballester B. et al. Gene expression profiling identifies molecular subgroups among nodal peripheral T-cell lymphomas. Oncogene 2005; 9;25(10): 1560-70.

[8] Steenman M. et al. Distinct molecular portraits of human failing hearts identified by dedicated cDNA microarrays. Eur J Heart Fail. 2005; 7(2): 157-65.

[9] Hester M. et al. Guidelines for Human Gene Nomenclature. Genomics 2002; 79: 464-470. [http://www.gene.ucl.ac.uk/nomenclature/]

[10] Boguski et al. ESTablishing a human transcript map. Nature Genetics 1995; 10: 369-371.

[11] Hubbard T.J.P et al. Ensembl 2007. Nucleic Acids Research 2007; Database issue doi:10.1093/nar/gkl996.

[12] Lee H.K. et al. Coexpression Analysis of Human Genes Across Many Microarray Data Sets. Genome Research 2004; 14: 1085-1094.

[13] Eisen M.B. et al. Cluster analysis and display of genome-wide expression patterns. Proceeding of the National Academy of Sciences 1998; 95: 14863-14868.

[14] Ashburner M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics 2000; 25: 25-29.

[15] Shannon P. et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research 2003; 13: 2498-2504.

[16] Edgar R. et al. Gene Expression Omnibus: NCBI gene expression and hybridation array data repository. Nucleic Acids Research 2002; 30(1): 207-210.

[17] Yang Y.H. et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Research 2002; 30: e15.
[18] Camon E. et al. The Gene Ontology Annotation (GOA) Project: Implementation of GO in SWISS-PROT, TrEMBL and Interpro. Genome Research 2003; 13(4): 662-672.
[19] Zeeberg B.R. et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biology 2003; 4: R28.
[20] Karolchik D. et al. The UCSC Genome Browser Database. Nucleic Acids Research 2003; 31(1): 51-54.
[21] Brazma A. et al. ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Research 2003; 1; 31(1): 68-71.
[22] Irrcher I. et al. Regulation of Egr-1, SRF, and Sp1 mRNA expression in contracting skeletal muscle cells. J Appl Physiol 2004; 97: 2207–2213.