REPUBLIQUE ALGERIENNE DEMOCRATIQUE & POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE BADJI -MOKHTAR, ANNABA

FACULTE DES SCIENCES

DEPARTEMENT DE BIOCHIMIE

MEMOIRE POUR L'OBTENTION DU DIPLOME DE MAGISTER

OPTION : BIOCHIMIE APPLIQEE

THEME :

ETUDE DES RESEAUX TRANSCRIPTIONNELS ISSUS DES EXPERIMENTATIONS DES PUCES A ADN

AZEDDINE GHERS

Jury :
Président : **Pr. A. LADJAMA** (Université d'Annaba).
Directeur de thèse : **Pr. D. KIRANE-GACEMI** (Université d'Annaba).
Examinateurs : **Pr. N. BOUTEFNOUCHET** (Université d'Annaba).
& Dr. T. MERAD (Université d'Annaba).
Membre invité : **Mr. F. BENABAS** (Université d'Annaba).

ANNEE : 2007

REPUBLIQUE ALGERIENNE DEMOCRATIQUE & POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE BADJI -MOKHTAR, ANNABA

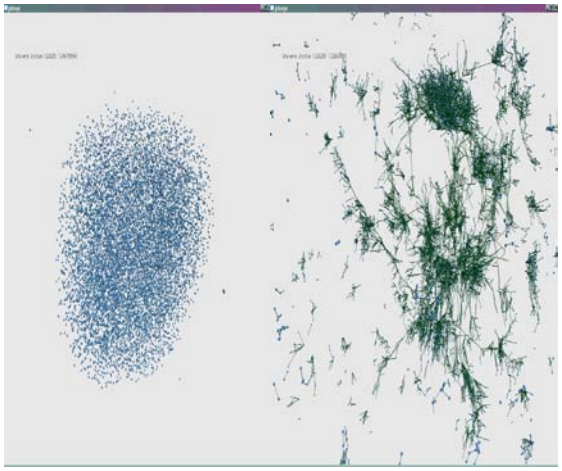
FACULTE DES SCIENCES

DEPARTEMENT DE BIOCHIMIE

MEMOIRE POUR L'OBTENTION DU DIPLOME DE MAGISTER

OPTION : BIOCHIMIE APPLIQEE

THEME



ETUDE DES RESEAUX TRANSCRIPTIONNELS

ISSUS DES EXPERIMENTATIONS DES PUCES A ADN

AZEDDINE GHERS

Jury :

Président : **Pr. A. LADJAMA** (Université d'Annaba).

Directeur de thèse : **Pr. D. KIRANE-GACEMI** (Université d'Annaba).

Examinateurs : **Pr. N. BOUTEFNOUCHET** (Université d'Annaba).

& Dr. T. MERAD (Université d'Annaba).

Membre invité : **Mr. F. BENABAS** (Université d'Annaba).

ANNEE : 2007

Résumé :

L’étude de la régulation de l'expression génique à l’échelle d’un génome complet, reste un des principaux problèmes de l'ère de la génomique fonctionnelle. Un des niveaux de régulation des gènes est constitué par l’action activatrice ou inhibitrice de facteurs de transcription, qui se fixent sur des sites spécifiques situés dans les régions promotrices des gènes. Les études de transcriptomes, c’est à dire la quantification de l'expression génique à grande échelle, dans des conditions biologiques variés et pour des instants multiples, pourrait nous donner une image affinée de la régulation génique à l'intérieur des cellules. Notre laboratoire a construit, à partir de milliers de biopuces Affymetrix relatives à plusieurs espèces et conditions biologiques, des réseaux transcriptionnels via la variation de la sur ou sous expression simultanée (corrélation positive, CORR) ou la sur ou sous expression inverse (corrélation négative, ANTI) de chaque couple de probesets.

Les principaux objectifs de cette étude : L’amélioration de nos réseaux transcriptionnels par filtrations des annotations et des assignations des jeux de sondes (probesets) déposés sur les différents modèles de puces. La mise en évidence des sous ensembles de gènes co-régulés par l’action de facteurs de transcriptions et/ou des sites consensus de fixation de facteurs de transcription.

Les principaux résultats : Les annotations des probesets nous ont montré que les sondes correspondantes à un probeset sont en fait localisées à des endroits différents. Nous avons classé les gènes pour lesquels un probeset a le plus grand nombre de sondes et nous avons constaté que certaines sondes sont parfois répétées un grand nombre de fois dans le génome à l'intérieur et à l'extérieur des gènes. Nous pensons que cette information qui n'avait jamais été rassemblée et exploitée auparavant permettra de mieux comprendre l’origine de la non reproductibilité des résultats que nous avons constatée entre des probesets qui représentent à priori le même gène. Une étude préliminaire a montré qu'il y avait un lien entre le score de corrélation entre deux gènes du réseau transcriptionnel et la présence dans leurs séquences promotrices de motifs de régulation communs.

Les principales conclusions et perspectives sont : La partie du travail effectuée a permis de rassembler les informations indispensables à la poursuite de ce travail. Il reste maintenant à approfondir et diversifier, grâce au matériel accumulé au cours de ce stage, les thèmes de recherches concernant la qualité des sondes et le rôle des motifs de régulation dans les séquences promotrices.

SUMMARY

The study of gene expression regulation on a complete genome scale remains one of the principal problems of the era of the genomic functional calculus. One of levels of gene regulation is consisted by the action of activation or inhibiting of factors of transcription, which are fixed on specific sites located in the areas promotrices of genes. Studies of transcriptoms, for example the quantification of the gene expression on a large scale, under biological conditions varied and for multiple moments, could give us a refined image of the gene regulation inside the cells. Our laboratory built, starting from thousands of Affymetrix biochips relative to several biological species and conditions, a transcriptional networks via the variation of on or under simultaneous expression (positive correlation, CORR) or the on or under expression reverses (negative correlation, ANTI) of each couple of probesets.

Principal objectives of this study: Improvement of our transcriptional network by filtrations of the annotations and the assignments of the sets of probes (probesets) deposited on the various models of chips. Description of gene subsets is co-controlled by the action of factors of transcriptions and/or of the sites consensus of fixing of factors of transcription.

Principal results: The annotations of the probesets showed us that the probes corresponding to a probeset in fact are located at different places.

We classified the genes for which a probeset has the greatest number of probes and we noted that certain probes are sometimes repeated a great number of times in the genome inside and outside genes. We think that this information which never had not been gathered and had been exploited before will make it possible to better include/understand the origin of the bad reproducibility of the results which we noted between probesets which represent a priori same gene. A preliminary study showed that there was a bond between the score of correlation between two genes of the transcriptional networks and the presence in their sequences promotrices of common reasons for regulation.

The principal conclusions and prospects are: The part of work carried out made it possible to gather essential information with the continuation of this work. It now remains to deepen and diversify, thanks to the material accumulated during this training course, the topics researches concerning the quality of the probes and the role of the reasons for regulation in the sequences promotrices.

ملخص:

إن دراسة نظام نسخ مورثات كل مادة الوارثة على مستوى خلية نسيج جسم كائن حي كاملة , يبقى واحد من المشاكل الرئيسية في عهد الجينوميك الوظيفية.
احد مستويات نظام نسخ المورثات مكون بالفعل المحرض او المثبط لعوامل (بروتينات نظامية) النسخ , التي تتمركز على مواضع خاصة تقع في الجهات القيادية للمورثات.
دراسات الترونسكريبتوم (نسخ كل مادة الوراثة ف-ي خلية ما) : أي حساب كمية **Les ARN m** المنسوخة من كل مورثات الخلية في شروط بيولوجية متنوعة و في أوقات عديدة , يمكن أن تعطينا صورة دقيقة عن نظام نسخ المورثات بداخل الخلايا.
لقد انشأ مخبرنا شبكات نسخية انطلاقا من آلاف الرقاقات الحيوية و الظروف البيولوجية , عن طريق التغير في الزيادة أو النقص في نسخ مورثتين بالتزامن (تزامن ايجابي .CORR) أو الزيادة أو النقص المعكوس (تزامن سلبي أو ANTI) لكل زوج من المسابير(**Les SONDES**) للمورثات عن الرقاقة الحيوية).

أهم أهداف الدراسة :

تحسين نوعية الشبكات النسجية للمورثات عن طريق التعليق بالشروحا ت و تحديد كل مجموعة مساير بالانتماء لمورثة وحيدة الموضوعة على مختلف أنواع الرقاقات.
تعيين المجموعات التحتية الخاصة بالمورثات المتزامنة النسخ أما عن طريق عوامل النسخ (بروتينات نسجية) أو مواضع تمركزها في الجهات القيادية للمورثات.

أهم النتائج :

شروح التعليقات حول مجموعات المياسير المتعلقة بالمورثة متموضعين في مواقع مختلفة لقد صنفنا المورثات اللواتي لهن اكبر عدد من المياسير و لقد لا حظنا أن بعض المياسير أحيانا مكررة و لمرات عديدة في جميع مادة الوراثة (جينوم). بداخل و خارج المورثات (الجينات) نعتقد أن هده المعلومة لم تجمع و تستغل من قبل سمحت لنا بفهم مصدر عدم مرد وديق النتائج الملاحظة بين مجموعات المياسير التي تمثل مبدئيا نفس المورثة .
دراسة تمهيدية بينت أن هناك علاقة بين حاصل التزامن بين مورثتين من شبكة النسخ ووجود مواقع مشتركة في المناطق القيادية لنظام نسخ المورثات.

أهم الخلاصات و الأبعاد :

الجزء المنجز حتى الآن من العمل مكن من جمع معلومات إلزامية لمواصلة العمل و بفضل الأدوات المكدسة خلال هدا التربص, يبقى الآن تعميق و تشعيب موضوعات البحث المتعلقة بنوعية المياسير ووظيفة الأمكنة الخاصة بتأثير عوامل نسخ المورثات في سلاسل القياديق للمورثات.

Listes des figures :

Liste des abréviations:

FDR : False Discovry Rate.

CORR: Corrélation positive de l'expression d'un couple de probesets.

ANTI: Corrélation négative de l'expression d'un couple de probesets.

ADN: Acide DésoxyriboNucléique.

ADNc: brin d'ADN complémentaire, qu'est élaboré à partir du brin d'ADN matrice.

ARNm: Acide RiboNucléique messager: transcrits du brin d'ADN matrice.

Sommaire:

Remerciements

Je remercie vivement **Pr. A. LADJAMA** pour l'honneur qu'il m'a fait en acceptant la présidence du jury de ce mémoire qu'il trouve ici l'expression de mon respect et ma gratitude.

Je suis heureux de remercie **Pr. D. KIRANE-GACEMI**, mon encadreur qui m'a toujours accueilli avec bienveillance et qui m'a consacré beaucoup de son temps. Je ne manquerai pas de lui présenter toute ma reconnaissance pour ses orientations et son aide précieuse.

Je remercie chaleureusement **Dr. T. MERRAD** & **Pr. N. BOUTEFNOUCHET** d'avoir bien voulu juger ce travail ; et également **CC. F. BENABAS** d'avoir répondu positivement à mon invitation.

Enfin, je ne peux pas suffisamment remercier **Dr. M. BELLIS** du Centre de Recherche en Biochimie Moléculaire du CNRS Montpellier, qui m'a fourni les informations nécessaires à l'accomplissement de ce mémoire.

Introduction Générale

Introduction Générale:

Tous les êtres vivants, animaux, végétaux et microorganismes, ont un point commun : ils sont formés de cellules (une seule pour les plus petits) contenant leur patrimoine héréditaire sous forme de chromosomes. Selon l'espèce, le nombre de chromosomes varie mais hormis chez certains virus, ils sont faits de molécules d'ADN **(figure 01)**. Vue de près, cette longue molécule d'ADN a la forme d'une échelle enroulée en spirale dont la formule chimique est simple. Les montants sont faits d'une succession de désoxyriboses (des sucres à 5 carbones) liés par des phosphates. Les barreaux sont constitués de 4 groupements chimiques (des bases) alignés deux par deux : adénine et thymine (A et T) ou guanine et cytosine (G et C). Ainsi, la seule variation tout au long de la double hélice d' ADN est la composition des barreaux successifs de l'échelle (A-T, T-A, G-C et C-G) qui se succèdent dans un ordre déterminé comme des lettres pour former des mots. L'ordre des barreaux, c'est-à-dire des bases de l'ADN, constitue bien un texte codé dont les mots sont des gènes, des ordres de fabrication et de fonctionnements de ce qui constitue un être vivant. Ces gènes, agencés dans un enchaînement précis en chromosomes, font de chacun de nous un être unique, ressemblant à ses parents mais différents de son voisin, d'une mouche à vinaigre ou d'un chou-fleur. Car c'est seulement avec quatre lettres, A, T, G et C, qu'est programmée toute la diversité du vivant **[1]**. L'information génétique d'une cellule est contenue dans son noyau sous forme de molécules d'ADN (chez les eucaryotes; les procaryotes ne possédant pas de noyau). La structure de cette macromolécule a été obtenue grâce aux avancés de la recherche en Biochimie Structurale. Son développement rapide en a fait une discipline reconnue en tant que telle, dont l'objectif est de décrire la vie au travers de la structure des macromolécules et de leurs interactions. Initialement centrée sur l'analyse et la comparaison des séquences nucléiques et protéiques, la bioinformatique intervient dans la plupart des disciplines biologiques. En effet, l'analyse des séquences biologiques est maintenant centrale pour un grand nombre d'études en biochimie et en biologie cellulaire. En outre dans la foulée des développements en génomique fonctionnelle, de nouvelles applications émergent, en biologie de développement, Neurobiologie, Immunologie...etc. Les expérimentations de la haute biotechnologie de nos jours reposent sur l'intervention pluridisciplinaire, c'est le cas de la ChIP to chip, où la biochimie coïncide avec l'immunogénétique et finisse avec l'informatique et la statistique pour résoudre des problèmes de génomique fonctionnelle **[2]**. C'est une approche récente permet l'investigation directe, à l'échelle génomique, des cibles géniques d'une protéine régulant la transcription en se liant à l'ADN **[3]**. Les protéines associées à l'ADN sont pontées

covalemment par un agent bivalent comme le formaldéhyde, qui pénètre dans les cellules vivantes. Les cellules sont tuées par ce traitement. L'ADN est rompu mécaniquement en fragments aux frontières aléatoires, dont la taille moyenne est contrôlée afin d'être de l'ordre de la longueur d'un gène. Les fragments d'ADN portant la protéine d'intérêt sont immuno-précipités par un anticorps dirigés spécifiquement contre elle **[4]**. La seconde phase fait appel à une dé-protéinisation, puis l'ADN peut être amplifié par PCR, ou identifié par hybridation à une micromatrice / puce. Idéalement cette micromatrice devrait comporter des sondes représentatives des régions géniques et intergéniques. Comme les régions régulatrices sont généralement intergéniques, on s'attend en principe que l'ADN précipité s'hybride avec certaines sondes intergéniques, et non aux sondes géniques. La première phase est appelée imunoprécipitation de la Chromatine, ou « Chromatin ImmunoPrecipitation » (ChIP), la seconde fait appel à une puce, ou (chip). La technique dans son ensemble est donc nommée « ChIP-chip » en anglais **[4]**.

L'étude de la régulation de l'expression génique à l'échelle d'un génome complet, reste un des principaux problèmes de l'ère de la génomique fonctionnelle. Un des niveaux de régulation des gènes est constitué par l'action activatrice ou inhibitrice de facteurs de transcription, qui se fixent sur des sites spécifiques situés dans les régions promotrices des gènes. Les études de transcriptomes, c'est à dire la quantification de l'expression génique à grande échelle, dans des conditions biologiques variées et à des instants multiples, pourrait nous donner une image affinée de la régulation génique à l'intérieur des cellules des différents tissus d'un organisme vivant. La disponibilité d'une technique performante et adaptée est l'une des premières exigences à satisfaire pour s'engager dans cette voie de recherche. Les puces à ADN répondent à ces exigences et se sont imposées comme la technologie phare dans cette thématique. L'utilisation intensive des puces à ADN au niveau mondial a aboutit à la constitution d'archives très importantes de résultats qui contiennent un ensemble énorme d'informations sur les transcriptomes. Pour une meilleure exploration de ces données, notre laboratoire: le CRBM du CNRS Montpellier, a construit des réseaux transcriptionnels à partir de milliers de biopuces Affymetrix relatives à plusieurs espèces. Nous allons, au cours de cette introduction, rappelez brièvement les notions essentielles relatives au transcriptome et aux biopuces, justifiez notre choix des puces Affymetrix, indiquez nos méthodes d'importations d'analyses et d'interprétations des données, et de construction des réseaux transcriptionnels. Enfin nous expliquerons l'objectif de notre étude, qui a pour but l'amélioration de nos réseaux transcriptionnels par filtrations des annotations et des assignations des jeux de sondes (probesets) déposés sur les différents modèles

de biopuces. Nous essayerons dans un deuxième temps d'expliquer l'observation, dans nos des réseaux transcriptionnels, de sous ensembles de gènes co-régulés par l'action de facteurs de transcriptions et/ou des sites consensus de fixation de facteurs de transcription. Cette approche nous permettra dans un premier temps de valider cette utilisation particulière des réseaux transcriptionnels en étudiant des voies de régulations déjà connues, et dans un deuxième temps de trouver de nouveaux gènes ainsi que de nouveaux facteurs pour des voies de régulation mal ou incomplètement connues.

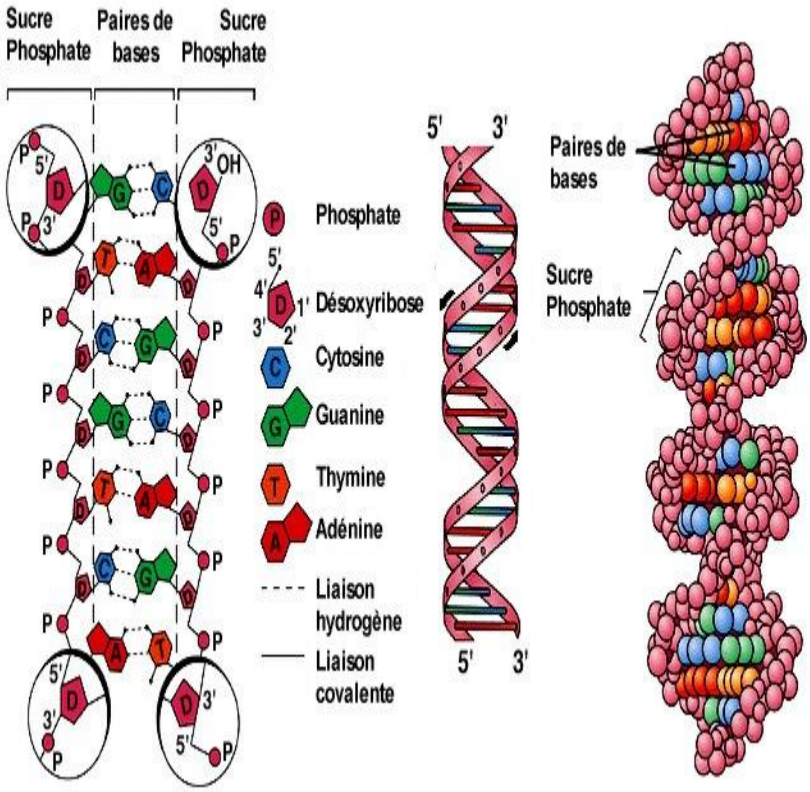


Fig. 01 : Configuration de la macromolécule biochimique de la l'ADN.

A. Le transcriptome:

L'étude de l'expression des gènes fait appel à deux approches : d'une part l'analyse du transcriptome constitué par l'ensemble des ARN messagers (ARNm) présents dans une cellule dans une situation donnée et d'autre part l'analyse du protéome représenté par les protéines que codent ces ARNm. Leur finalité commune est d'identifier et de quantifier les produits de l'expression des gènes d'une cellule ou d'un tissu à un instant et dans un environnement donné, dans un but de comparaison entre différents états biologiques.

Le transcriptome est défini comme la population d'ARNm présents dans les cellules.

Les ARNm sont produits dans le noyau par l'ARN polymérase II à partir d'une matrice d'ADN. Cette transcription est déclenchée par des facteurs de transcription capables d'activer spécifiquement certains gènes. Ces ARNm vont être exportés dans le cytoplasme pour être traduits en protéines. Ces protéines vont constituer d'une part des composants de la cellule (fibres d'actines, pores membranaire, pompe membranaires, etc.), d'autres part des molécules régulatrices (enzymes, facteurs de transcription, etc.).

La quantité d'une protéine et son activité sont régulées à différents niveaux (transcription, localisation des ARNm, traduction, maturation protéique, localisation et conformation de la protéine **(figure 02)**), mais souvent l'analyse du transcriptome nous donne une assez bonne vision du jeu de protéines présent dans la cellule. L'analyse du transcriptome étant pour des raisons techniques plus aisée que l'analyse du protéome, on dispose le plus souvent de la vision du transcriptome avant celle du protéome. L'approche du transcriptome est aujourd'hui rendue très accessible grâce à des méthodologies bien maîtrisées et au large spectre d'applications. L'analyse du transcriptome à grande échelle est possible grâce à la technique des puces à ADN ou microarrays.

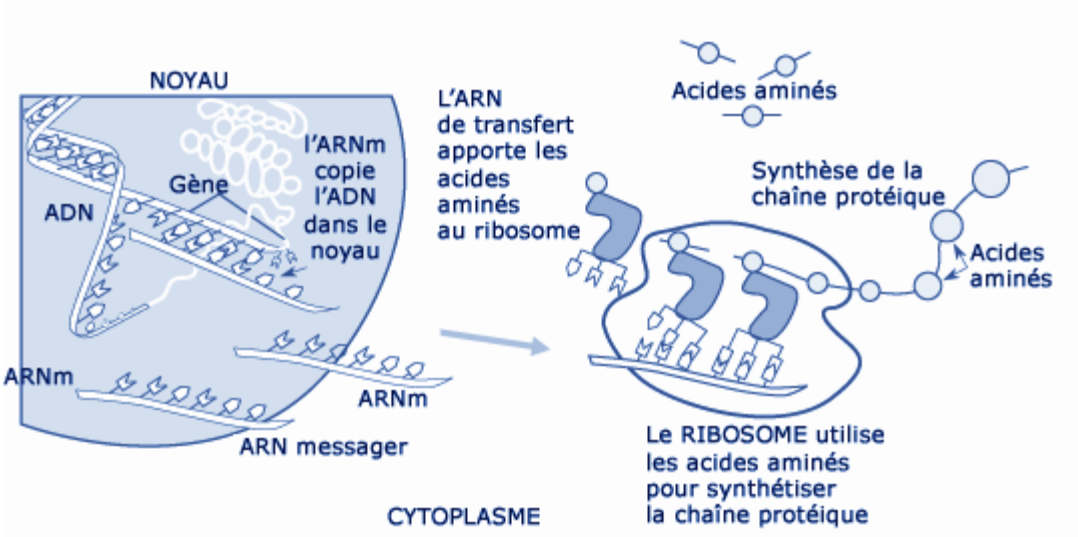


Fig. 02 : Les niveaux de régulation d'une protéine : (transcription, localisation des ARNm, traduction).

B. Les puces à ADN:

Récemment, une nouvelle technologie a permis aux chercheurs d'explorer rapidement des patrons d'expression de génomes entiers. Un microarray (ou puce à ADN) est une petite lame de verre d'un centimètre sur centimètre. La surface de cette lame est couverte de plus de 20 000 taches correspondant chacune à un oligonucléiotide (une courte séquence de nucléotides) différent. Des ADNc peuvent également être apposés sur la lame pour servir de sondes. D'autres supports, tels que membranes fines (macro-array) peuvent être utilisées à la place de lames de verre. Le point clé de ce type d'expérimentation réside dans le fait que chaque séquence d'ADN est immobilisée à la surface de la lame par une de ses extrémités. Les puces à ADN ne sont pas conceptuellement différentes des traditionnelles expériences d'hybridation telles le Southern Blot (hybridation d'ADN génomique avec une sonde d'ADN) ou le Northern Blot (hybridation d'ARNm avec une sonde d'ADN), sur une puce à ADN, chaque oligomère peut servir de sonde pour détecter un ADNc cible (ou ARNm). Ces oligomères peuvent être marqués par fluorescence, ce qui permet à la puce d'être analysée grâce à un scanner confocal ou une caméra CCD. La présence ou l'absence de la séquence complémentaire dans un échantillon d'ADN examiné sur une puce à ADN, détermine quelles positions sont « allumées » ou « éteintes » sur le support. Par conséquent, la présence ou l'absence d'environ 20 000 séquences dans un échantillon peut en théorie être démontrée expérimentalement avec une seule expérimentation sur une seule puce. Parmi d'autres avantages, les puces à ADN utilisent des sondes fluorescentes plutôt que des sondes radioactives utilisées dans les techniques traditionnelles peuvent être conçues de façon robotisée. Depuis la conception des puces jusqu'à la quantification des signaux, en passant par l'extraction de groupes de gènes ayant des profils d'expression associés, l'analyse des données de puce à ADN est difficile. Il est donc impensable de concevoir des puces à ADN et leur utilisation sans l'utilisation d'ordinateurs et de bases de données. Pour que les résultats expérimentations soient à la fois clairs et sans ambiguïté, chaque sonde d'ADN déposé doit être unique, de façon à ce qu'un seul gène de la cible puisse s'hybrider avec cette sonde. Si ce n'est pas le cas, la quantité de signal détecté pour chaque tache ne sera pas déterminée correctement. Les résultats d'expérimentations peuvent être difficiles à visualiser. Les expérimentations possèdent généralement au moins quatre dimensions (position X, position Y, intensité de fluorescence et durée). Une lecture directe des taches sur les images extraites de la grille n'est pas très informative. Des outils permettent d'extraire les fonctionnalités d'un ensemble de données d'ordre supérieur et de les présenter de manière intelligente sont par conséquent nécessaires.

Actuellement, la stratégie la plus utilisée pour l'analyse des données de puces à ADN est le regroupement ou classification (clustering) de profils d'expression **[5]**. Plusieurs méthodes de classification, telles que la classification hiérarchique ou les cartes auto-organisatrices (SOM pour self-organizing maps) fonctionnent plus ou moins bien selon les situations, mais le but générale de chacune de ces méthodes est le même. Plusieurs paquetages logiciels commerciaux, contenant des outils pour la visualisation et l'analyse de données d'expression, sont disponibles. Certains sont spécifiques à des équipements physiques ou à des configurations de grilles particulières. D'autres comme SpotFire et GeneSpring de Silicon Genetics sont plus universels. Ces paquetages logiciels sont souvent relativement coûteux, mais à ce stade du développement de la technologie des puces à ADN, ils sont rentables de par leur relative facilité d'utilisation **[6].**

Celles-ci sont utilisées pour quantifier l'expression des gènes dans une situation biologique donnée. L'analyse d'une masse suffisante de données d'expériences sur puces peut permettre d'identifier des familles et des réseaux fonctionnels de gènes mis en jeu sous l'effet du stimulus étudié. Ainsi, les puces à ADN nous permettent d'identifier les programmes d'expression génique mis en route dans un type cellulaire donné, après stimulation par un agent (facteur de croissance, cytokine, molécule médicamenteuse, etc.), dans certaines pathologies ou au cours du développement. Notre laboratoire est impliqué dans le domaine des biopuces depuis 1997. Une biopuce est dans sa définition la plus générale un assemblage d'un grand nombre de sites réactifs dans un très petit volume. Ce concept, qui a été élaboré progressivement au cours des années 1980-1990, est passé dans le domaine applicatif de la biologie à la fin des années 1990; depuis, l'utilisation de ce nouvel outil s'est considérablement développée et diversifiée. Les puces à ADN permettent de mesurer et de visualiser très rapidement les différences d'expression entre les gènes et ceci à l'échelle d'un génome complet. Si la mise en œuvre de la technique est assez compliquée, son principe est très simple **(figure 03)**.

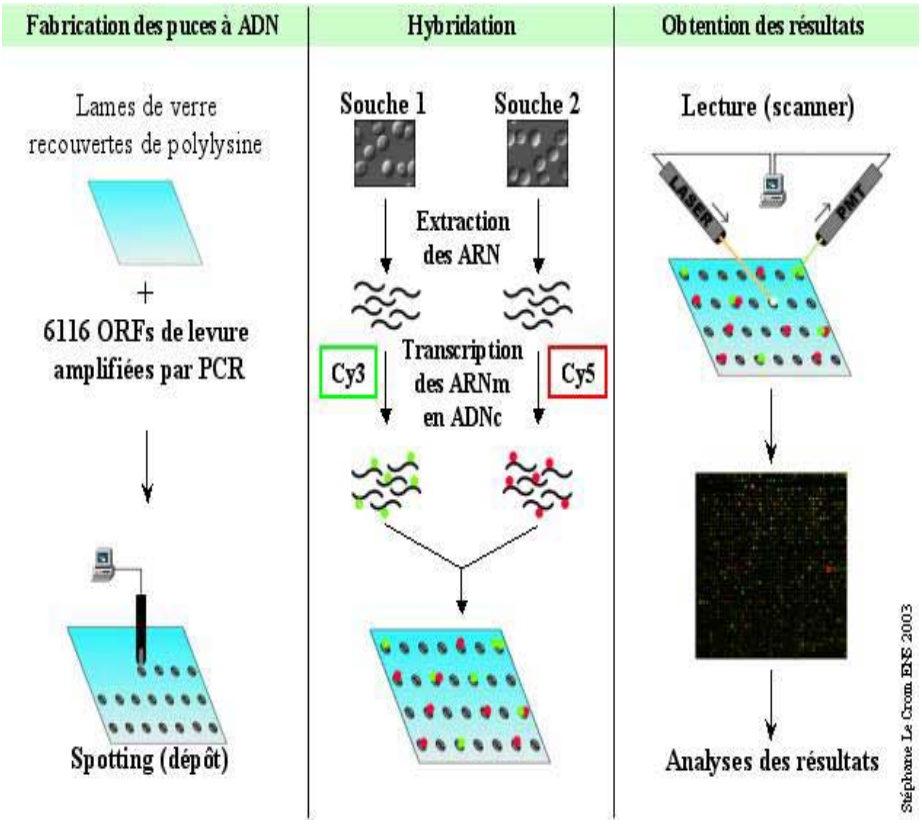


Fig. 03 : Principe de la puce à ADN.

En voici les principales étapes:

1. La fabrication de la puce :

Une puce à ADN est constituée d'un très grand nombre d'unités d'hybridations (de quelques milliers à plusieurs centaines de milliers) disposées côte à côte sur un substrat plan ou poreux et contenant chacune de l'ordre de un million de sondes identiques. Il existe deux types principaux de puces à ADN. Dans les puces à cDNA, les sondes sont des fragments d'ADN amplifiés par la technique de PCR et déposés sur une lame de microscope préalablement recouverte de polylysine. La polylysine a pour rôle d'assurer la fixation de l'ADN déposé via des interactions électrostatiques. La préparation de la lame est achevée en bloquant la polylysine n'ayant pas encore accroché d'ADN de façon à éviter une fixation non spécifique de la cible. Juste avant l'hybridation, on dénature l'ADN pour qu'il se trouve sous la forme simple brin sur la puce, ce qui lui permettra de s'accrocher au brin complémentaire contenu dans la cible. Les puces à oligonucléotides utilisent comme sonde des oligonucléotides (de 25 bases pour Affymetrix et de 60 à 80 bases pour Agilent), qui sont soit déposés avec une technique similaire à celle des puces à cDNA (puces fabriquées dans les plates-formes académiques), soit synthétisées in situ (puces commerciales Affymetrix, Agilent et NibleGene) **[7]**.

2. La préparation de la cible :

Les ARN sont extraits de la culture cellulaire ou du tissu dont on veut étudier l'expression. Les ARN messagers sont transformés en ADNc par transcription inverse qui est lui-même re-transcrit en ARN dans une étape finale pendant laquelle un marquage de la cible est effectué (soit marquage fluorescent pour les puces à cDNA, soit marquage indirect à la biotine pour les puces à oligonicléotides. Dans le cas des puces à cDNA on utilise une deuxième source d'ARN qui sert de contrôle et qui est marquée avec un autre fluorochrome **[7]**.

3. L'hybridation :

L'ADN marqué qui constitue la cible en solution est mis en contact avec les sondes d'ADN (simple brin) déposées sur la puce. La puce est alors incubée une nuit à 60 degrés dans des conditions de salinité adaptées (concentrations variables selon le génome : levure ou souris,…etc.) pour favoriser l'hybridation, c'est-à-dire le processus d'appariement entre les brins d'ADN complémentaires. Au cours de cette étape permissive, de nombreux évènements d'hybridation croisés, plus ou moins spécifiques ont lieu. On rajoute une étape de lavage à basse

force ionique qui provoque la séparation des brins les plus instables et améliorent grandement la spécificité du signal **[7]**.

4. La lecture :

Chaque spot est excité par un laser et on récupère la fluorescence émise via un photomultiplicateur couplé à un système de microscopie confocale.
On obtient alors une image dont le niveau de gris représente l'intensité de la fluorescence lue. Dans le cas de puces à cDNA, la lecture se fait successivement sur deux canaux chacun correspondant à la longueur d'onde d'émission d'un des deux fluorochromes utilisés pour distinguer le premier échantillon test du deuxième échantillon contrôle. On remplace les niveaux de gris de la première image, par des niveaux de vert et par des niveaux de rouge pour la seconde. On obtient en superposant ces deux images, une image dont la couleur indique le sens de variation du niveau d'expression : les spots vont du vert pur (forte induction du gène dans le test par rapport au contrôle) au rouge pur (forte répression du gène dans le test par rapport au contrôle) en passant par le jaune (pas de variation de l'expression entre les deux conditions) **[7]**.

5. La détection des variations statistiquement significatives :

Cette étape est essentielle car une des caractéristiques des puces à ADN est l'importance du bruit : lorsque l'on compare deux conditions biologiques identiques, une même sonde peut être mesurées avec des valeurs très différentes. Il faut donc mettre en œuvre des méthodes statistiques qui permettent au minimum d'assigner à toute variation une valeur p et au mieux d'estimer pour toute sélection un taux de false discovery rate (FDR), c'est-à-dire une estimation de la fraction de faux positifs présent dans la sélection effectuée. Nous utilisons pour cela une méthode développée au laboratoire que permet d'estimer outre ces deux quantités, la variation totale et la sensibilité d'une sélection (fraction de la variation totale sélectionnée) **[7]**.

6. Le regroupement en fonction des profils d'expression :

Lorsque l'on considère plusieurs conditions biologiques ou une cinétique, on peut ensuite essayer de regrouper des gènes ayant le même profil d'expression. Ce regroupement ou clustering peut se faire de proche en proche comme pour une phylogénie, ce qui consiste à calculer un critère de similitude entre les réponses et à rassembler les profils les plus similaires. On peut également faire appel à des techniques plus complexes comme l'analyse en composante principale ou les réseaux neuronaux. Au final on représente en général le résultat du clustering sous la forme d'une

matrice où chaque colonne correspond à une expérience et chaque ligne correspond à un gène. On normalise en général le signal par rapport à une condition de référence (par exemple le temps 0 dans une cinétique) et l'on représente ce ratio grâce à une échelle de couleur, par exemple du vert (gènes réprimés) au rouge (gènes induits) **[7]**.

7. Les puces à oligonucléotides:

Comme nous l'avons indiqué, il existe plusieurs types de puces à ADN.
Nous avons restreint notre étude aux données générées par des puces Affymetrix (http://www.affymetrix.com/index.affx) pour deux raisons : la première est qu'il s'agit de la seule technique monocanal. De ce fait chaque résultat est un résultat « absolu » qui peut être comparé à tout autre résultat. Au contraire, les résultats générés par les techniques double canal sont relatifs puisqu'ils dépendent du contrôle utilisé qui sera jamais le même dans des expérimentations provenant de laboratoires différents. Deuxièmement, il s'agit d'une solution industrielle ce qui implique une standardisation des modèles proposés ce qui s'avère bénéfique pour ce type d'études de masse: en effet, toutes les expérimentations effectuées à partir de cette technologie sont facilement comparables et utilisables pour construire un réseau transcriptionnel comprenant un très grand nombre de gènes. Les puces Affymetrix ont une structure bien particulière : un gène est représenté non pas par une seule sonde mais par une ensemble de couple de sondes (en général une vingtaines de couples) qui constituent un probeset. Chaque couple de sondes est constitué d'une sonde 'perfect match' qui est exactement complémentaire à la séquence connue de la cible, et d'une sonde 'mismatch' identique sauf pour la base du milieu qui est changée. Un algorithme propre à Affymetrix permet d'obtenir à partir de la quarantaine de signaux, une seule valeur représentant le 'niveau d'expression' du gène **[7]**.

C. Importation des données:

Nous avons décidé de traiter les données disponibles sur le serveur du NCBI: Gene Expression Omnibus (GEO : http://www.ncbi.nlm.nih.gov/geo/) qui est à présent le serveur le plus important. Il a affiché le 05 mai 2005: 38804 résultats groupés en 1258 expériences. Pour donner une idée de la dynamique d'accumulation, signalons que lorsque nous avons effectué notre première importation (Octobre 2003) nous étions à environ 1000 puces.
Nous avons développé un grand nombre d'outils (logiciels), afin d'importer de manière automatique les fichiers provenant de GEO directement dans Arrayon: outil d'analyse et de traitement des données de puces à ADN.

D. Les niveaux d'analyse et d'interprétation des données des puces à ADN:

En présence de données de transcriptome, on peut envisager plusieurs types d'analyse de complexité croissante selon le nombre de ces données.

1. Deux conditions biologiques => probesets sur ou sous exprimés :

C'est le niveau le plus simple, déjà décrit précédemment (détection de variations statistiquement significatives (I2). La sélection de probesets sur- ou sous exprimés se fait par estimation du paramètre de FDR. Il est bien évident qu'il existe une relation inverse entre le FDR et la sensibilité. Avec un FDR de 1% les résultats sont très fiables, mais l'on ne sélectionne qu'une fraction de la variation totale. On peut améliorer la sensibilité, mais au détriment de la fiabilité, en prenant par exemple un FDR de 5%. En réalité, c'est le type projeté d'utilisation des données qui permet de choisir le paramètre le mieux adapté. Ainsi si les données sont destinées à être vérifiées expérimentalement, l'on choisira de préférence un FDR de 1%. En revanche pour des analyses in silico, un FDR de 5% pourra très bien être retenu. Pour ce qui est des réseaux, nous avons observés une très grande robustesse de leur structure par rapport à la valeur du FDR **(figure 04)**.

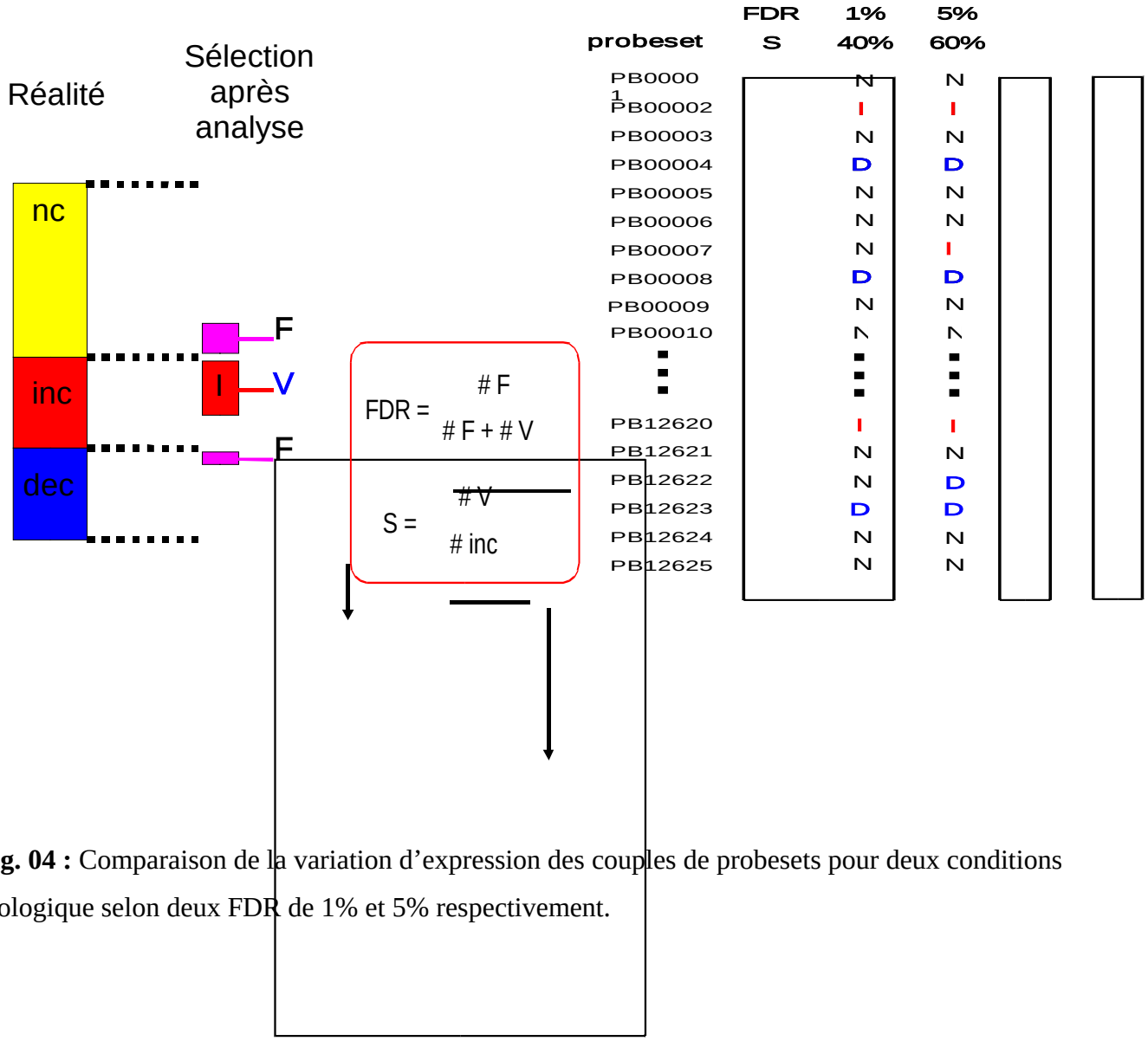


Fig. 04 : Comparaison de la variation d'expression des couples de probesets pour deux conditions biologique selon deux FDR de 1% et 5% respectivement.

2. Un petit nombre d'expérimentations => Question(s) biologique(s):

A ce stade, nous pourrons poser une ou plusieurs questions biologiques relatives à des conditions expérimentales plus nombreuses. Par exemple, on s'intéresse à la spermatogenèse et l'ovogenèse, avec l'ambition de mettre en évidence les gènes impliqués dans la recombinaison méiotique. Cette étape exige d'appliquer une méthode de clustering qui permette de répondre à la question biologique posée. La plupart des biologistes utilisent des méthodes de clustering géométriques qui consistent à représenter les probesets dans un espace multidimensionnel, à mesurer les distances entre tous les couples de probesets et à regrouper les probesets qui sont proches. .Nous avons développé au laboratoire, une méthode alternative basée sur une combinatoire systématique sur toutes les comparaisons effectuées, de tous les résultats possibles. Ainsi si trois comparaisons on été effectuée, nous classons les gènes dans au plus $3^3 = 27$ classes qui sont dénommées par une chaîne de symboles représentant la nature des variations successives. Par exemple la classe IDN indique que les gènes qui appartiennent
à cette classe sont surexprimés dans la première comparaison, sous exprimés dans la deuxième et invariant dans la troisième. Les vingt sept classes sont donc NNN, INN, NIN, NNI, IIN, INI, NII, III, DNN, NDN, NND, DDN, DND, NDD, DDD, IDD, DID, DDI, IID, IDI, DII, NID, NDI, IND, DNI, IDN, DIN. Ensuite ne sont retenues que les classes qui sont susceptibles de contenir les gènes recherchés. Ainsi pour la méiose, nous savons que les gènes impliqués dans le phénomène sont exprimés à partir du 14 ^{ième} jour post-partum chez le mâle et au 13 ^{ième} jour post-coitum chez la femelle.

3. Un grand nombre d'expérimentations => Régulation de la transcription:

L'approche expliquée précédemment, ne peut pas être utilisée telle quelle lorsque le nombre de conditions biologique est trop important. Nous l'avons cependant conservée et adapté pour traiter ce cas : si on a 150 conditions biologiques différentes, on effectuera toutes les comparaisons possibles, soit 150 * 149 / 2 = 11175 comparaisons. Ensuite on calculera, comme indiqué ci-après pour tout couple de probeset (sur la puce humaine de type U95, il y a environ 12 000 probesets, ce qui va faire de l'ordre de $72 * 10^6$ couples) un score de corrélation positive (CORR) et un score de corrélation négative (ANTI) **(figure 05)**.

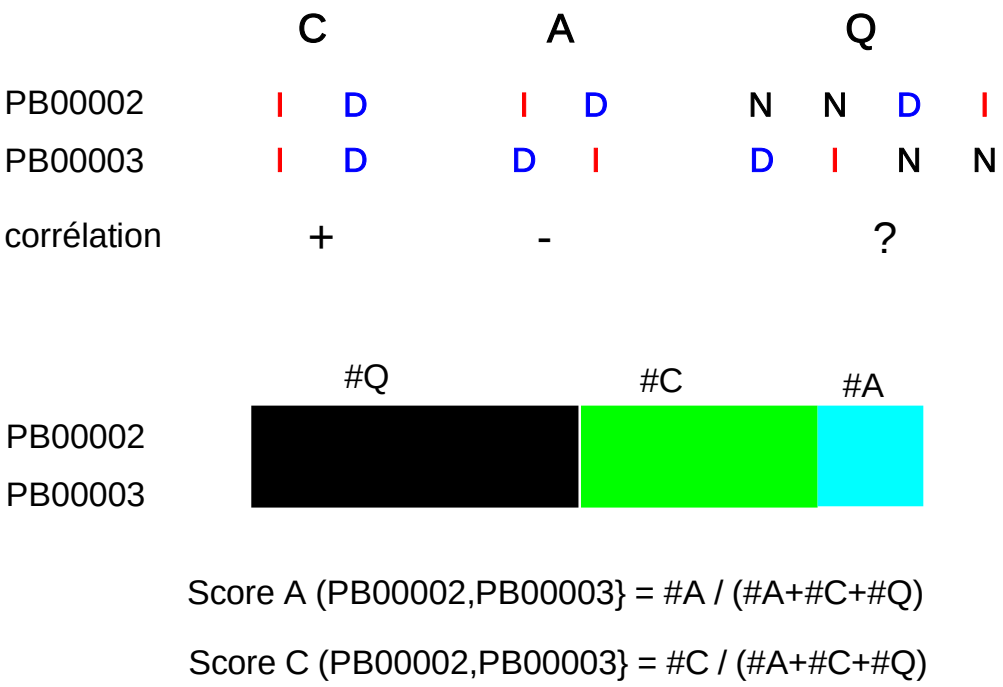


Fig. 05 : Stratégie de construction de réseaux transcriptionnels.

4. La stratégie de la construction:

La clef de voûte de notre stratégie de construction est la variation de la sur ou sous expression simultanée (corrélation positive, CORR) ou la sur ou sous expression inverse (corrélation négative, ANTI) de chaque couple de probesets; mais il existe encore un troisième cas que l'on le nomme questionnable (Q) qui englobe tout les autres cas qui ne relèvent, ni de la variation d'expression simultanée ni de la variation inverse. Les scores de la corrélation positive et négative sont calculés de la façon suivante:

CORR (probeset i, probeset j)= # CORR / (#CORR + #ANTI + #Q)

ANTI (probeset i, probeset j)= # ANTI / (#CORR + #ANTI + #Q)

E. Objectifs du travail

1. Calcul d'un index de qualité pour les probeset Affymetrix :

Les probesets des puces Affymetrix présentent parfois des problèmes d'annotation et/ou d'assignation des probesets. Pour mieux cerner ce problème, nous avons décidé de caractériser chaque probeset, relativement à la position des sondes qui le composent dans le génome. Nous utilisons, pour récupérer l'information, la base de données ENSEMBL: (http://www.ensembl.org/), qu'est une banque de données performante en matière de localisation, d'annotation et d'assignation des probesets. Afin de mieux comprendre comment la nature des sondes d'un probeset peut éventuellement influencer la qualité des résultats générés par la technologie Affymetrix, nous nous intéressons à un sous ensemble particulier de probesets. Il existe certains gènes qui sont représentés par plusieurs probesets sur les puces Affymetrix. Nous nous attendons à ce que ces points qui représentent un même gènes, soient très comparables les uns aux autres en ce qui concerne leur voisinage : ils devraient être corrélés et anti-corrélés aux même probesets. En fait pour une partie de ces probesets il n'en est rien, ce qui nous donne un moyen de détecter les groupes de probesets dont un au moins des membres pose un problème d'annotation ou de localisation. C'est en utilisant ces groupes que nous essayerons de comprendre ce qui peut expliquer ces anomalies du réseau transcriptionnel compte tenue des informations que nous aurons récoltées sur la localisation des sondes.

2. Etude des séquences promotrices

Nous projetons de rechercher toutes les séquences promotrices des gènes humains, murins et du rat, avec les sites consensus ou motifs de fixations de facteurs de transcription ou de protéines

régulatrices. Une fois que nous aurons ces informations en main nous essayerons d'expliquer les corrélations positives et/ou négatives existantes entre certains gènes, par l'action d'un facteur de transcription d'une protéine régulatrice.

Outils & Méthodes

Outils & Méthodes:

A. Outils:

a) Gestion des résultats de puces ADN

(1) Arrayon:

Outil de gestion, de représentation graphique, d'analyse statistique et de clustering des données de puces à ADN. Il est composé d'un ensemble de méthodes rigoureuses et puissantes permettant d'analyser à fond une expérimentation aussi complexe soit elle
Développé dans Superbase et MATLAB par le Dr. Michel Bellis **[8]**.

b) Gestionnaire de l'information :

(1) SQL (Structured Query Language) :

SQL (Structured Query Language, traduisez Langage de requêtes structuré) est un langage de définition de données (LDD, ou en anglais DDL Data Definition Language), un langage de manipulation de données (LMD, ou en anglais DML, Data Manipulation Language), et un langage de contrôle de données (LCD, ou en anglais DCL, Data Control Language), pour les bases de données relationnelles **[9]**. Deux environnements d'interrogation et de programmation ont été utilisés dans l'achèvement de ce travail:

(a) The MySQL database:

Administrator ; Query Browser; Server 4.1: MySQL Command Line Client, MySQL Server Instance Config Wizard; System Tray Monitor.

(b) SQLyog:

Il utilise la même connexion du MySQL, plus pratique pour l'interrogation **(figure 06)**.

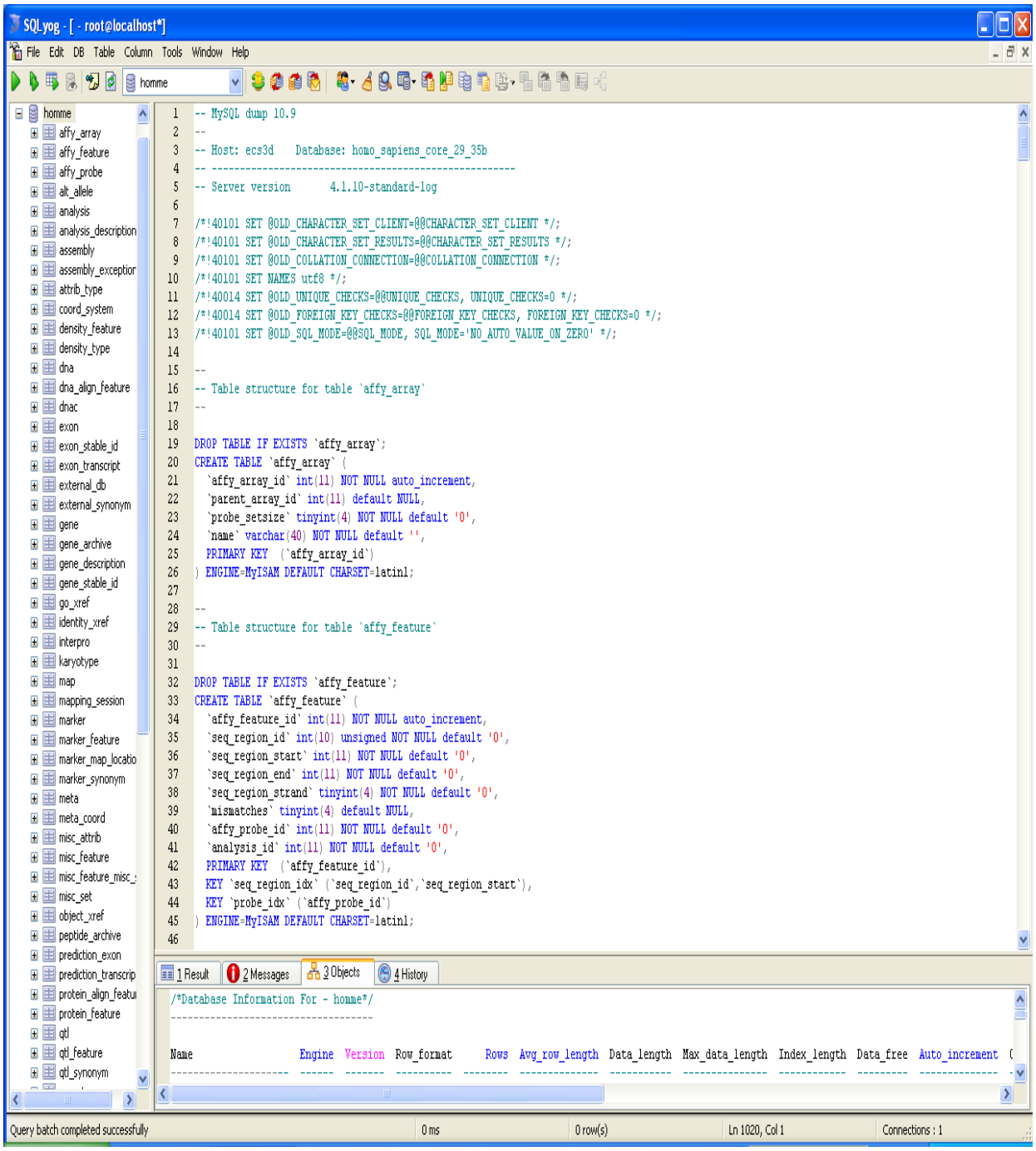


Fig. 06 : Etape d'exécution du script de création de la base de données relationnelle d'Homo sapiens via SQLyog, qu'est connecté à l'administrateur du MySQL via le login root.

(2) SRS:

C’est un système d'indexation de banques d'informations biologiques, il permet d'exploiter les liens qui existent entre les différentes banques. C'est un outil d'accès privilégié aux banques de séquences généralistes.

(3) SuperBase:

C'est une base de Données et un Outil de Développement RAD (Rapid Application Développement). Cette application dispose de toutes les fonctionnalités permettant d'utiliser les fonctions standard d’un SGDB ainsi qu’un langage de développement.

(4) PersonalBrain:

C'est un outil novateur de navigation et de classement d'idées et de fichiers basé sur une interface graphique représentant sous forme d'arbres les relations entre les différents objets.
(figure 07).

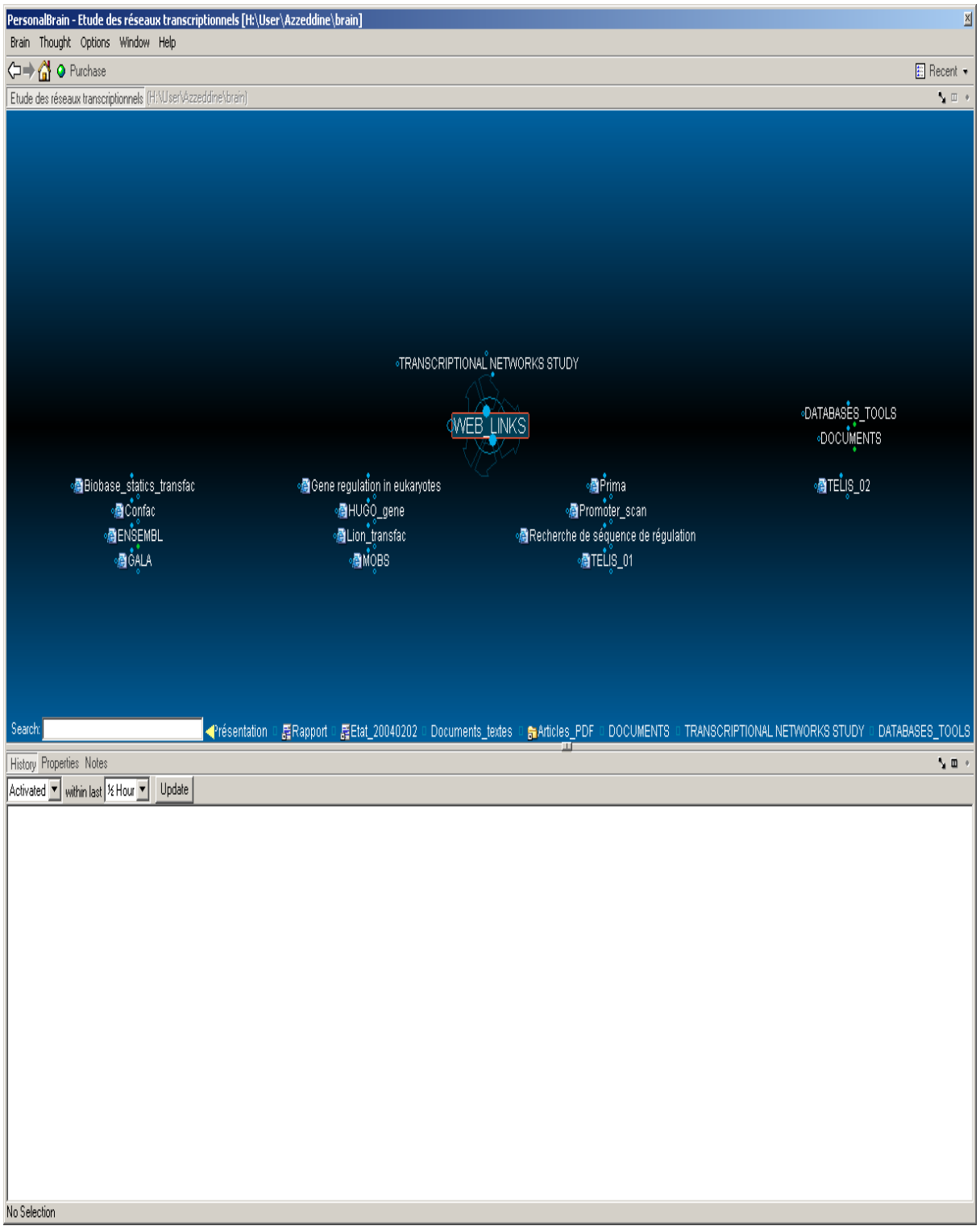


Fig. 07 : Mon « Brain » que j'ai développé à partir du logiciel PersonalBrain pour cette étude.

c) Outils de communication et de transfert :

(1) BlazeFtp:

BlazeFtp est un client FTP multisessions avec la gestion d'un cache et la possibilité de recherche hors ligne. Il est léger et très simple d'utilisation. Son fonctionnement ne nécessite que quelques minutes d'apprentissage.

d) Outils de recherche et de prédiction:

(1) Conserved Transcription Factor Binding Site Finder (CONFAC):

Outil de cherche des sites de fixation et des facteurs de transcription conservés dans les régions promotrices des gènes de gènes humains donnés et de l'homologue correspondant de souris **[10]**. Le processus implique quatre étapes. Voir le détail sur le site : (http://morenolab.whitehead.emory.edu/cgi-bin/confac/login.pl)

(2) The Transcription Element Listening System(TELIS):

C'est une base de données contenant les informations concernant la prévalence des motifs de fixation des facteurs de transcription se fixant sur les promoteurs de tout gène de l'homme, de la souris ou du rat existant sur les différents modèles de puces Affymetrix, Agilent ou autres **[11]**. Telis est un outil spécialisé dans la recherche des motifs de fixation des facteurs de transcription, il contient des nombres entiers indiquant les occurrences de chaque motif de fixation dans chaque promoteur;.Telis utilise une famille de matrices pour chaque facteur de transcription, et génère les données de fréquence après un balayage fait par PromoterScan qui utilise les séquences de nucléotides disponible via RefSeq du NCBI, et qui est conduit à une stringence donnée fixée par MatInspector qui utilise les valeurs de 80, 90 et 95 à travers une taille de promoteurs spécifiques de 300 ou 600 nucléotides en amont du site d'initiation de transcriptions ou une région qui débute de 1000 bases en amont du site d'initiation de transcription et se termine à 200 nucléotides en aval du site d'initiation de transcription. TELIS contient les données de 34622 gènes humains, 24384 gènes murins et 21053 gènes du rat **[11]**. Les motifs de fixation de facteurs de transcription sont définis par 108 matrices Position-Specific-Weight de la base de données JASPAR 2 (The High Quality transcription factor binding profile database), qui est non redondante ou par 192 matrices représentant tout les facteurs de transcription des vertébrés de la base de données TRANSFAC au choix des utilisateurs. Les motifs de fixation sont détectés par l'algorithme

MatInspector. TELIS est développé originalement pour cartographier le control des réseaux transcriptionnels, en conjonction avec PromoterStats (Outils Statistique). Il peut identifier des facteurs de transcription expliquant la dynamique de l'expression de gènes. Les motifs de fixation des facteurs de transcriptions sont importés par FTP à partir de TRANSFAC et de JASPAR. Pour les sites de transcription alternatifs les résultats multiples sont ajustés et arrondis de façon d'avoir un seul enregistrement avec les positions potentielles. La bases de données de TELIS est interrogée par l'identifiant de la base de données Hugo. Pour chacune des trois espèces, la correspondance entre nom des gènes et types de puces est fondamentale. TELIS offre des multiples possibilités d'analyses, des calculs statistiques associés : Analyse d'expression différentielle en utilisant les matrices de la base TRANSFAC : elle trouve les motifs de fixation des facteurs de transcription qui sont sur représentés dans les promoteurs des gènes présentant une variation d'expression **[11]**.

1. Analyse d'expression différentielle en utilisant les matrices de la base JASPAR 2 : Elle trouve les motifs de fixation des facteurs de transcription qui sont sur représentés dans les promoteurs de la variation d'expression de gènes.
2. Trouver le rang d'enregistrement dans la base de données : Elle renvoie le site de fixation du gène interrogé à partir de la base de donnée TELIS.
3. Analyses des fréquences: se fait par défaut par Z-test qui compare le nombre moyen de motif avec le nombre total des gènes
4. Analyses des incidences: c'est avec un test binomial afin de déterminer le motifs les plus représentés suite à la variation d'expression génique.
5. Telis utilise deux tests statistiques pour avoir des résultats statistiquement significatifs.
6. Pour estimer le taux des faux positifs relatifs à plusieurs motifs, nous utilisons le FDR au lieu du p – value relative à un motif particulier pour l'utilisation de cet outil, voici le site :(http://www.telis.ucla.edu/TELiSDifferentialExpression.htm).

(3) JASPAR 2 :

C'est une collection de facteurs de transcription se fixant sur l'ADN, modélisés par des matrices position spécifiques de poids (PSSMs). La différence principale avec les ressources semblables (TRANSFAC, TESS etc..) est la non redondance et la qualité.collecte des données à partir des régions expérimentalement déterminées dans des régions promotrices réelles ; cette distinction est clairement marquée dans l'annotation des profils.

(4) EnsMart (ENSEMBL MartView) :

C'est un outil d'interrogation de la banque de donnée Ensembl sur les puces à oligonucléotides Affymetrix de 16 différentes espèces, ainsi que les informations relatives à chaque interrogation dans les banques de données principales en bioinformatique et en génomique: RefSeq, EMBL, SwissProt, Affymetrix, Gene Ontology, OMIM, Interpro, et HUGO, etc.; ainsi qu'une comparaison multi espèces. Avec trois possibilité de recherche en Texte, Blast ou Mart.

(5) MatInspector:

C'est un outil qui utilise une grande bibliothèque de matrice pour les sites de fixation des facteurs de transcription pour localiser des motifs consensus dans les séquences promotrices des gènes. MatInspector assigne une estimation de qualité de match et permet ainsi la filtration et le choix qualité basée sur les matchs **[12]** suivant différentes astringences de 80, 90 et 95.

(6) TRANSFAT :

C'est un outil d'exploration de données désigné pour détecter la sous ou la sur représentation des facteurs de transcription humains par comparaison de deux ensembles de gènes en utilisant au choix : les identificateurs de gènes ENSEMBL ou les noms externes des gènes de l'homme. Le lien est disponible via le site : (http://babelomics.bioinfo.cnio.es/transfat/transfat.cgi)

e) Bases de Données utilisées (références):

(1) Gene Expression Omnibus (GEO):

Cest une base de données relatives à l'expression de gènes à haut débit. Elle compile les résultats des puces à ADN de différentes espèces. Les données de différentes puces sont disponibles sur le serveur FTP du site GEO: (ftp://ftp.ncbi.nih.gov/pub/geo/data/geo/)

(2) Ensembl Genome Browser (ENSEMBL)

C'est une base de données d'annotation automatique des génomes eucaryotes. Les annotations de toutes les biopuces Affymetrix sont misent à jours très souvent contrairement à celles de la base de Affymetrix **[13]**. Les détails sont disponibles via le site (www.ensembl.org).

(3) The mammalian promoter service (PromoSer)

C'est un site qui permet l'extraction des promoteurs des gènes de l'homme, de la souris et du rat. Les détails sont disponibles via le site (http://biowulf.bu.edu/zlab/promoser/).

(4) The Eukaryotic Promoter Database (EPD):

EPD est une base de données spécialisée d'annotation de la bibliothèque de données d'EMBL. Il

fournit des informations au sujet des instigateurs eucaryotiques disponibles dans la bibliothèque de données d'EMBL et est prévu pour aider les chercheurs expérimentaux, aussi bien que des analystes fonctionnels, dans la recherche sur les signaux eucaryotiques de transcription **[14]**. Les détails sont disponibles via le site (http://www.epd.isb-sib.ch/).

(5) NCBI Reference Sequence (RefSeq) :

C'est une banque généraliste de séquences nucléiques de référence pour l'annotation fonstionnelle du génome humain. Les détails du site sont disponible via le lien(http://www.ncbi.nih.gov/RefSeq/).

(6) The UniProt/Swiss-Prot Protein Knowledgebase (UniProt/Swiss-Prot):

C'est une base de donnée d'annotation et de classification automatique de séquences protéiques. (http://www.ebi.ac.uk/swissprot/).

(7) Transcription Regulatory Regions Database (TRRD):

C'est une base de données conçue pour l'accumulation des données expérimentales sur des régions de régulation étendues des gènes eucaryotiques **[15]**. Le lien à la base est disponible via : (http://wwwmgs.bionet.nsc.ru/mgs/dbases/trrd4/).

(8) The Transcription Factor Database (TRANSFAC):

C'est une banque spécialisée de séquences nucléiques et de facteurs de la cis régulation de transcription des eucaryotes. Sachant que dans les banques de données des sites de fixation des facteurs de transcription: TRRD, TRANSFAC et EPD, il y a une base de données communes TFSITE de TRANSFAC qui contient les résultats de 192 sites de fixation des facteurs de transcription se liant à des courts motifs de 5 à 25 nucléotides, ainsi que ces facteurs **[9]**.
Le lien à la base est disponible via : (http://transfac.gbf.de/TRANSFAC/).

(9) HUGO Gene Nomenclature Committee (HUGO):

C'est la base universelle de données de nomenclature des gènes humains. Les principes de nomenclature sont disponibles via le site : (http://www.john-libbey-eurotext.fr/fr/revues/bio_rech/abc/e-docs/00/00/C4/A1/article.md?type=text.html).

(10) PromoterScan:

C'est un serveur d'indexation de différents outils de prédiction, de recherche et de bases de données de données de la régulation de transcription. En plus c'est un outils de balayage des promoteurs en utilisant les séquences nucléotidiques de RefSeq de la base de données NCBI, le

niveau du balayage est spécifique à une stringence donnée calulée par MatInspector. Disponible via le lien : (http://www.bioinformatics.vg/biolinks/bioinformatics/Promoter%2520Scan.shtml).

(11) Pub Med :

C'est une banque de données de références de littératures. Site accessible via le lien: (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi).

(12) Affymetrix:

C'est une entreprise de fabrication des puces à oligonucléotides pour des multiples organismes vivants (http://www.affymetrix.com/index.affx).

(13)Superarray Biosciences Corporation :

Cette base de données contient les annotations des gènes pour des biopuces et de classification des gènes suivant 11 domaines (Apoptose, biomarqueurs, cancer, cycle cellulaire, biologie cellulaire et développement, maladies communes, cytokines et réponse inflammatoire, matrices cellulaires et adhésion de molécules, neuroscience, signal de transduction et toxicologie et métabolisme du drogue.

(14)GALA : Gene Alignement and Annotation Database :

Gala est banque de donnée de clustering des sites de fixation des facteurs de transcription ou des motifs. Elle incorpore les annotations génomiques avec des alignements de multiples espèces pour nous permettre des interrogations complexes sur des informations correspondantes publiquement disponibles.

B. Méthodes:

1. Filtration des probesets:

Nous essayons de résoudre le problème d'annotation et d'assignation des probesets Affymetrix en utilisant les informations de localisation des sondes d'ENSEMBL, qui sont mises à jours mensuellement.

Pour ce faire nous avons procédé en trois étapes, qui sont :

a) Importation des données et création de la base de données relationnelle:

Les tables SQL de toutes les espèces sont disponibles via (ftp://ftp.ensembl.org/pub/), et le script de téléchargement est disponible via:

http://www.ensembl.org/Docs/wiki/html/EnsemblDocs/InstallEnsemblOneFileAtATime.htm.
Parallèlement, nous avons utilisé Blazeftp pour l'importation des tables contenant les scripts des créations des bases de données relationnelles, pour l'homme, rat, souris. Nous avons installé un serveur SQL qui nous a permis de créer nos bases de données relationnelles, que nous avons les interrogées par le langage DQL (Data Query Language: langage de requête de données). Afin de compléter nos informations concernant les annotations et l'assignation des gènes nous avons utilisé EnsMart d'ENSEMBL, disponible via le lien: (http://www.ensembl.org/Multi/martview). L'utilisation du logiciel PersonalBrain s'est avérée utile pour manipuler aisément une telle masse d'informations.

b) Traitements et analyse des données:

Les données concernant la localisation des sondes et des exons sont exportées à partir de SQL dans un format « ASCC=II delimited » et importées dans Superbase. Sont également importées d'autres tables obtenues à partir d'EnsMart et qui contiennent des informations sur les identifiants Ensemble (EnsGeneId) : chromosome, bande cytogénétique, nom de gène (HUGO ou Autre). Dans cet exemple **(figure 08)**, Affymetrix a assigné le GeneSymbol TOP1 au probeset 1030_s_at. L'information donnée par Ensembl nous indique que les sondes correspondant à ce probeset sont en fait localisées en trois régions chromosomiques différentes. Nous avons classé FM04 en premier car c'est le gène pour lequel le probeset 1030_s_at a le plus grand nombre de sonde (11 contre 5 pour TOP1, cf. champ ProbeLoc Rank).
Nous avons aussi à disposition le nombre de fois qu'une sonde est présente à une localisation (champ ProbeLocRep). On observe également que la troisième localisation, 'no EnsG', n'est pas sur le bon brin de l'ADN (champ ProbeNoStrand). Toutes ces informations pourront être utilisées pour comprendre l'origine de la non reproductibilité des résultats entre des probesets qui représentent a priori le même gène.

ProbeSet	1030_s_at
GeneSymbol	TOP1
ProbeLocEns	{'ENSG00000076258','ENSG00000198900','no EnsG'}
ProbeLocSymbol	{'FMO4','TOP1','no EnsG'}
ProbeLocRank	{[1, 2, 3, 4, 6, 8, 9,11,12,15,16];[1, 7,10,13,14];[1, 5]}
ProbeLocType	{'protein_coding','protein_coding',''}
ProbeLocStatus	{'KNOWN','KNOWN',''}
ProbeLoc2Kb	[1,0,2]
ProbeLocRep	{[1,1,1,2,1,1,1,1,1,1,1];[6,3,3,1,3];[4,7]}
ProbeNoMism	{[1,1,1,2,1,1,1,1,1,1,1];[6,3,3,1,3];[0,0]}
ProbeOneMism	{[0,0,0,0,0,0,0,0,0,0,0];[0,0,0,0,0];[4,7]}
ProbeNoStrand	{[0,0,0,0,0,0,0,0,0,0,0];[0,0,0,0,0];[4,7]}
ProbeOneStrand	{[1,1,1,2,1,1,1,1,1,1,1];[6,3,3,1,3];[0,0]}
Symbol	FMO4
Chrom	1
Band	q24.3
GeneStart	168015031
GeneStop	168042880
EnsGeneId	ENSG00000076258
GeneType	protein_coding
GeneStatus	KNOWN

Fig. 08 : Exemple des problèmes d'annotations des probesets que présente Affymetrix.

2. Interprétation des modules transcriptionnels présents dans nos réseaux

Nous savons que la régulation de la transcription se fait par les facteurs de transcription qui se lient à des motifs ou séquences consensus présents dans les régions promotrices des gènes. Nous mettons à profit l'existence de nombreuses sources d'information, de nature distincte, concernant la présence de sites de fixation pour les facteurs de transcription en amont des gènes pour tester la relation qui existe éventuellement entre la présence de sous ensemble co–régulés, visibles dans nos réseaux transcriptionnel et la présence de motifs de reconnaissance de facteurs de transcription.

a) Utilisation des données de prédiction relatives à des ensembles de gènes co-régulés, impliqués dans les mêmes processus biologiques et qui ont des sites consensus communs :

(1) L'Algorithme « Ab Initio » :

Nous utilisons des ensembles de 100 jusqu' à 150 gènes qui ont un motif commun détecté par un algorithme puissant nommé : Ab Initio **[17]**. L'algorithme consiste à :

- Création d'un ensemble de gènes partageants le même motif découpé en 5 à 8 nucléotides, et le pattern d'expression des gènes sont comparé avec le pattern d'expression des facteurs de transcription connus, par orthologie entre homme, souris.
- Analyse statistique par des matrices d'alignement multiple: PAM1 et PAM10 des sites (motifs) de fixation sur représenté dans les régions promotrices évolutivement conservées des gènes orthologues.
- Deux filtres complémentaires sont utilisés pour affiner la co-régulation : partage des mêmes annotations de Gene Ontology et appartenance aux mêmes groupes de gènes co-régulés dans les résultats des ensembles d'expérimentation des microarrays **(figure 09)**.

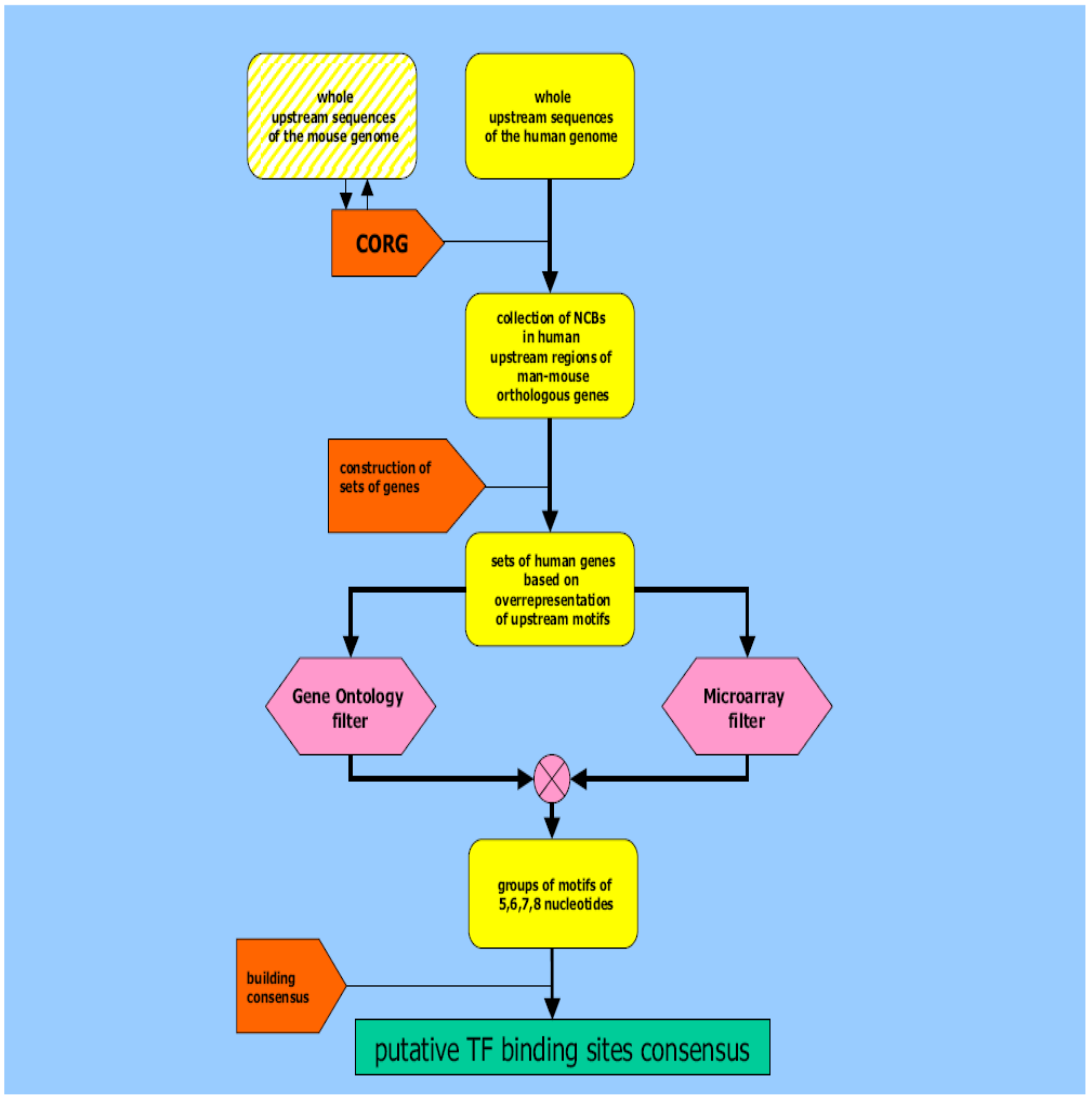


Figure 1
Flow-chart of the algorithm

Fig. 09: Organigramme général de l'algorithme Ab Initio, qui met en évidence des ensembles de gènes co-régulés et ayant un motif consensus en commun.

(2)L'outil de prédiction « CONFAC »:

Une fois qu'on trouvera un ensemble de gènes présents ensemble dans un modules avec des corrélation significatives, nous essayons de trouvés un facteurs de transcription putatifs, nouveau ou connus via l'outil de prédiction des facteurs de transcription pour des sites orthologues conservés entre l'homme et la souris. Nous utilisons des tables de deux colonnes, la première pour l'identificateur HUGO et la seconde pour l'identificateur RefSeq, nous faisons appel à l'EnsMart d'Ensembl pour avoir les noms Hugo et l'identificateur RefSeq des gènes pour pouvoir interroger CONFAC.

(3)La banque de données de sites et de motifs de fixations des facteurs de transcription GALA :

Sur cette base de données on a toutes les séquences promotrices des gènes humains avec des sites de fixation de facteurs de transcription ainsi que des facteurs connus d'une part, d'autre part, on connaît via ENSEMBL tout les débuts de gènes, la méthode consiste à rechercher des groupes gènes possédant des sites de fixation de facteurs connus afin de pouvoir interroger nos réseaux transcriptionnels avec ces informations.

b) Utilisation de l'outil Telis spécialisé dans le control de la dynamique de la régulation des réseaux transcriptionnels pour interpréter des modules de régulation:

Nous utilisons l'astringence et la taille de promoteurs dans la recherche des motifs de fixation des facteurs de transcription. Un promoteur de taille de 600 et une astringence de 90 conviennent bien dans certaines circonstances. Si on obtient un ratio faible, on peut réduire la taille du promoteur. Si nous augmentons l'astringence à 95, automatiquement nous pensons à élever la taille du promoteur à 1200 pour réduire le taux des résultats nuls.

Une bonne stratégie consiste à : débuter à moins 600 en amont du site de transcription et avec une astringence moyenne de 90 pour avoir un bon équilibre entre sensibilité/spécificité.

Pour augmenter la sensibilité, on va réduire la taille du promoteur 300/90, ou encore on va élever l'astringence à 95 et parallèlement pour la taille du promoteur, nous commençons à 1200 et on passe après à 600/95 et on termine par 300/95 (maximum de sensibilité). Sachant que plusieurs bons résultats disparaissent avec l'augmentation de l'astringence, à cet égard nous signalons que une faible astringence de 80 sera très utile parfois à cause des matrices imprécises et la mauvaise définition de l'astringence.

(1) Superarray & TELIS:

Nous avons des groupes de gènes sélectionnés expérimentalement et à partir de références bibliographiques pour être impliqués dans des processus biochimiques ou physiologiques bien spécifiés, comme la réparation d'ADN. Nous cherchons pour ces groupes de gènes des motifs communs ainsi que des facteurs de transcription afin d'avoir un groupes de régulation avec lequel nous interrogeons nos réseaux transcriptionnels pour trouver ces modules dedans.

(2) Modules des réseaux transcriptionnels & Telis :

Nous pouvons sélectionner à partir de nos réseaux des listes des probesets corrélées positivement ou négativement. Nous utilisons TELIS pour trouver un activateurs ou un répresseur commun pour ces probesets. Nous faisons appel à 'EnsMart (Ensembl) pour avoir les noms Hugo des gènes pour pouvoir interroger TELIS **[11]**.

Résultats & Discussions :

Résultats & Discussions :

1. Vérification des annotations de probesets proposées par Affymetrix:

Il existe sur les puces Affymetrix que nous avons utilisées pour construire nos réseaux, environ un millier de gènes qui sont représentés par deux ou plus prbesets, soit environ 4000 probesets. Nous avons constaté, en étudiant le score de corrélation entre probesets assignés par Affymetrix à un même gène, qu'il existait un ensemble important de gènes pour lesquels cette correlation était très faible, ce qui indique qu'au moins un des probesets à un problème d'assignation. Nous sommes donc intéressés à mettre au point un index de qualité des probesets, tirés des information de position des sondes qui le constituent. Ce travail a déjà été entrepris par site GeneAnnot **[18]**, mais les résultats produits par l'algorithme utilisé ne nous ont pas satisfait : nous n'avons observé aucune relation entre les scores GeneAnnot d'un couple de probesets assigné au même gène et leur valeur de corrélation observée dans le réseau. Les annotateurs d'Affymetrix ont toujours déclaré, quand cela était possible, un gène unique pour chaque probeset. Cependant, en réalité les séquences de probes à l'intérieur des probesets peuvent correspondre à plusieurs gènes ou être répétées un grand nombre de fois sur le génôme y compris en dehors de séquences codantes. Nous pensons que cette redondance, qui est parfois très importante et qui n'est pas utilisée dans le score GeneAnnot, pourrait permettre de comprendre l'origine de la non reproductibilité des résultats entre des probesets qui représentent à priori le même gène. En fait, l'équipe du docteur Vered Chalifa-Capsi, a proposé un algorithme d'annotation des jeux de sondes des puces Affymetrix, en se basant sur la puce humaine HG - U95, qu'était l'exemple de discussion de leur travail avec la possibilité d'application pour les autres puces **[18]**. Nous pensons que cet algorithme est criticable car en fait il passe dans un premier temps des probes aux transcrits puis seulement à la fin de ces derniers aux gènes. Nous avons choisi de passer en priorité par la séquence génomique par le schéma suivant, concernant les différentes tables importées à partir d'Ensembl :affy_array: table des puces humaines U133, U95 => affy_probe => affy_feature => seq_region => exon => gene. J'ai récupéré le maximum d'information à partir d'Ensembl, outre les informations de positions : description, type de séquence (protein, rna etc) et degré de confiance (connu, nouveau,...). Nous avons rajouté les noms HUGO avec les identificateurs RefSeq de la base de référence NCBI, SwissProt et les noms externes des gènes, pour arriver à avoir si possible un symbole de gènes pour chaque probeset. Le principe qui a conduit notre étude a été de classer les différentes localisations d'un probeset en fonction du nombre de sondes qui y sont présentes. L'exemple de la figure 08 montre le résultat de ce classement, sur le probeset

1030_s_at de la puce HG-U95 d'Affymetrix. Ce probeset est localisé par Ensembl à trois endroits, dont deux correspondent à des gènes. Nous avons classé en premier le gène FM04 pour lequel il existe 11 sondes homologue et en deuxième TOP1 qui n'en contient que 5. Nous conservons aussi l'information de la répétition de chaque sonde à une localisation (champ ProbeLocRep). On observe également que la troisième localisation, 'no EnsG', n'est pas sur le bon brin de l'ADN (champ ProbeNoStrand). Toutes ces informations pourront être utilisées pour comprendre l'origine de la non reproductibilité des résultats entre des probesets qui représentent a priori le même gène. Le format utilisé (présence d'accolade) nous permet d'importer facilement ces données dans Matlab sous forme de tableau de cellules (cell array) et de les manipuler convenablement.

2. Interprétation des sous-ensembles de gènes corrélés dans les réseaux transcriptionnels :

L'explication de la présence de sous ensembles de gènes corrélés par l'action de facteurs de transcriptions et/ou de sites consensus de fixation de facteurs de transcription est une approche très intéressante. Nous avons essayé plusieurs types de données, expérimentales ou prédites.
La base de donnée que nous avons retenue est TRANSFAC qui est la base la plus complète et qui sert de référence pour les autres bases du domaine. En ce qui concerne les données expérimentals, nos analyzes qui sont préliminaires et non exhaustives ne nous ont pas permis pour l'instant de trouver un exemple expliquant un de nos sous ensemble de gènes corrélés par l'action d'un facteur de transcription connu. Ce travail se poursuit actuellement.

3. Comparaison statistique de nos réseaux aux données prédites :

Deux grandes bases de données ont été utilisées :d'une part Ab Initio qui est une base prédictive et d'autre part GALA qui liste les gènes ayant en amont des séquences connues de fixation de facteurs de transcription.
Nous avons sélectionné dans notre réseaux transcriptionnel des couples de probesets pour lesquels la corrélation positive 'CORR) était supérieure à 30 ou 50 et la corrélation négative (ANTI) inférieure à la moitié de CORR. Nous avons ensuite regardé la distribution et la fréquence du nombre de motifs GALA en commun entre les deux probesets de chaque couple. Nous avons faits de même sur un tirage aléatoire de couples et nous observons que plus la valeur de CORR est élevée plus la fréquence ou le nombre de motif en commun s'éloigne de la distribution aléatoire **(figure 10)**.

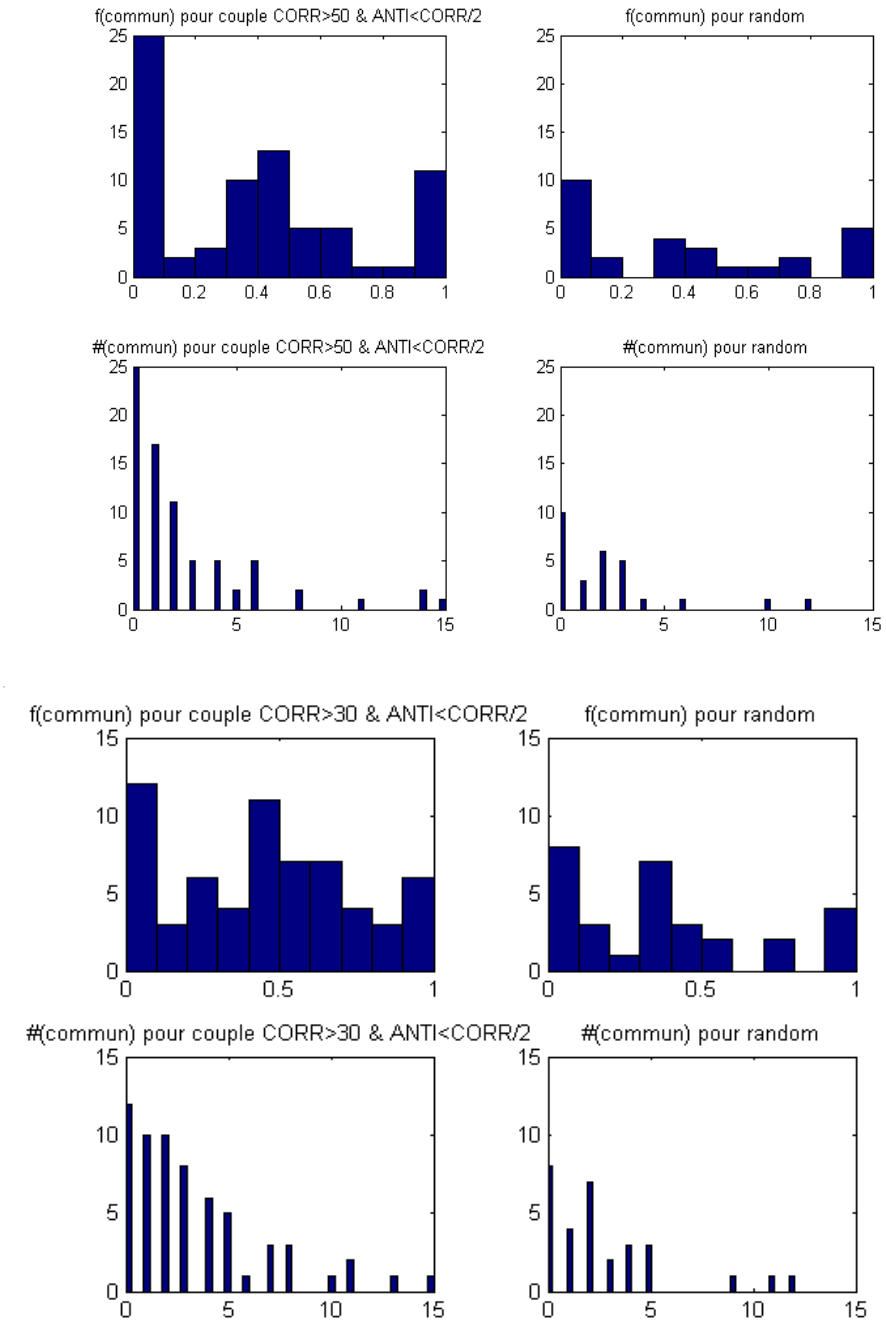


Fig. 10: Nombre (#) et fréquence (f) de motifs communs entre des couples de probesets sélectionnées au hasard ou sur un critère de corrélation positive.

Conclusions & Perspectives :

Conclusions & Perspectives :

Les objectifs que nous nous étions fixé pour ce stage ont été partiellement atteints. En particulier nous avons mis en place et fait fonctionner tous les outils nécessaires à l'importation et au traitement des données nécessaires au projet : données sur les séquences promotrices de différentes nature (expérimentales et prédictives), et données sur les positions des sondes. Concernant les sondes nous avons sélectionné et appliqué une méthode qui a aboutit à une caractérisation plus complète et pus rigoureuse des probesets par rapport à ce qui est actuellement disponible dans le domaine. La dernière partie qui concerne l'application à l'étude des réseaux transcriptionnels n'a pu être qu'abordée de manière exploratoire. Les premiers résultats que nous avons obtenus sur le lien entre la corrélation entre gènes dans le réseau transcriptionnel et la présence de motifs communs sont prometteurs et méritent d'être approfondis.

References bibliographiques:

[1] J.C. Revy, Communication, (1999), **"Ça m'intéresse": Patrimoine,** Edition de l'Association Française contre les Myopathies, 1:4.

[2] N. Nègre et al. (2006): **Mapping the distribution of chromatin proteins by ChIP on chip.** Methods In Enzymology, 1:26.

[3] F. Darel et al. (2002): **Bioinformatique: Génomique & post génomique**, Les éditions de l'Ecole Polytechnique: livre de 248 pages.

[4] M. Ghielmetti et al. (2006): **Gene expression profiling of the effects of intravenous immunoglobulin in human whole blood**. Molecular Immunology 43, 939:949.

[5] J. Hennetin et al. (2006): **Clustering methods for analyzing large datasets: gonad development, a study case.** Methods In Enzymology, 1:11.

[6] C. Gibas et al. (2002): **Introduction à la bioinformatique: concepts fondamentaux et outils et logiciels**, Edition O'REILLY: livre de 375 pages.

[7] http://transcriptome.ens.fr/sgdb/presentation/principle.php.fr: **Service de Génomique du Département de Biologie: Plate- forme transcriptome.** École normale supérieure 45, rue d'Ulm F-75230 Paris cedex 05.

[08] D. E. Martin et al. (2004): **Rank Difference Analysis of Microarrays (RDAM), a novel approach to statistical analysis of microarray expression profiling data**. BMC Bioinformatics, 5:148.

[09] R. R. Plew et al. (2003): **Le tout en poche, SQL, un apprentissage progressif pour apprendre à manipuler efficacement le SQL**. Campus Press, 402 pages.

[10] S. Karanam et al. (2004): **Confac : Automated application of comparative Genomic promoter analysis to DNA microarray**. Nucleic Acids Research, Vol. 32, 475:484.

[11] S. W. Cole1 et al. (2005): **Expression-based monitoring of transcription factor activity: the Telis database.** BIOINFORMATICS, Databases and ontologies, Vol. 21 n^{o}. 6, 803:810.

[12] K. Cartharius et al. (2005): **MatInspector and beyond: Promoter Analysis based on Transcription Factor Binding sites.** Bioinformatics Advance Access published.

[13] T. Hubbard et al. (2005): **Ensembl 2005**, Nucleic Acids Research, Vol. 33, Database issue 447: 453.

[14] R. Cavin et al. (1999): **The Eukaryotic Promoter Database (EPD): recent development**.

Nucleic Acids Research, Vol. 27, N° 1, 307:309.

[15] N. A. Kolchanov et al. (2002): **Transcription Regulatory Regions Database (TRRD)**: its status in 2002, Nucleic Acids Research, Vol 30, N° 1, 312: 317.

[16] T. Heinemeyer et al. (1998):**Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL.** Nucleic Acids Research, Vol. 26, No. 1, 362 : 367

[17] D. Corà et al. (2005): **Ab initio identification of putative human transcription factor binding sites by comparative genomics**. BMC Bioinformatics, 6:110.

[18] V. Chalifa-Caspi et al. (2004):**GeneAnnot: comprehensive two-way linking between oligonucleotide array probesets and GeneCards genes**. Bioinformatics, Vol. 20 no. 9, 1457:1458.