Conception d'outils bioinformatiques pour la modélisation de voies métaboliques et de leur régulation

Pierre-Yves Dupont

Ecole Doctorale des Sciences de
la Vie et de la Santé

Conception d'outils bioinformatiques pour la modélisation de voies métaboliques et de leur régulation

THÈSE

Présentée à l'Université d'Auvergne – Clermont 1 pour
l'obtention du grade de Docteur de l'Université

Informatique appliquée aux Sciences du Vivant

Soutenue le 15 Décembre 2011

Pierre-Yves DUPONT

Métabolisme Bioénergétique et Modélisation – Unité de Nutrition Humaine – INRA
Centre INRA de Clermont-Ferrand Theix, 63122 St Genès-Champanelle

Directeur de thèse
Dr Georges Stepien

Jury

Pr. Yves-Jean BIGNON,	Centre Jean Perrin, Clermont-Ferrand	Président du Jury
Pr. Jean-Pierre MAZAT,	Université de Bordeaux	Rapporteur
Dr. Jean-Paul ISSARTEL,	CNRS, Inserm U836, Grenoble	Rapporteur
Dr. Marc FERRARA,	INRA Clermont-Ferrand –Theix	Examinateur
Dr. Christophe HAUG,	Soluscience SA, Clermont-Ferrand	Examinateur

RÉSUMÉ

La biologie des systèmes actuelle s'appuie sur des techniques d'analyse biologique à haut débit comme la transcriptomique ou la métabolomique. Cependant, ces techniques haut débit ont leurs limites et peuvent générer des erreurs. En croisant les résultats de différentes techniques d'analyse biologique, nous espérons pallier à une partie de leurs limites. À cet effet, nous avons commencé à développer une plateforme de modélisation, MPSA (*Metabolic Pathways Software Analyzer*), permettant d'intégrer les données générées à des réseaux métaboliques. MPSA permet de représenter les graphes de voies métaboliques, d'effectuer des simulations basées sur la résolution de systèmes d'équations différentielles et d'étudier la structure des réseaux métaboliques par le calcul et la représentation des modes élémentaires. Nous avons développé différentes applications web permettant d'une part l'interprétation des résultats biologiques, en utilisant des bases de données et d'autre part leur export vers MPSA. La base de données centrale de ce développement est myKegg, incluant l'ensemble des voies métaboliques humaines de la base de données publique KEGG ainsi qu'une base de synonymes construite elle aussi à partir de KEGG. Cette base permet d'identifier des voies métaboliques et de les importer dans MPSA. Une base de données de métabolomique, BioNMR, a aussi été construite spécifiquement pour organiser les résultats générés à partir de spectres de RMN. Une autre application web, GeneProm, a été développée pour l'analyse de promoteurs de gènes ou promotologie. Un protocole d'étude a été mis au point et testé sur un groupe de 4 gènes codant pour les isoformes 1 à 4 de la protéine ANT, transporteur mitochondrial d'ATP, chacune ayant un rôle et un profil d'expression spécifique dans la bioénergétique cellulaire. L'étude par promotologie de ces 4 gènes a permis d'identifier des éléments de régulation spécifiques dans leurs séquences promotrices et d'identifier des gènes potentiellement co-régulés. Ces gènes peuvent ensuite être exportés vers notre plateforme MPSA. L'ensemble de ce développement sera inclus au projet de plateforme intégrative de l'Unité de Nutrition Humaine de l'INRA.

ABSTRACT

The actual systems biology relies on different high-throughput techniques of biological analysis like transcriptomics or metabolomics. However, these techniques may generate errors. By crossing the results from different analysis techniques, we hope to avoid at least a part of these limits. In this objective, we started to develop a modeling platform, MPSA (Metabolic Pathways Software Analyzer). MPSA allows integrating biological data on metabolic pathways. MPSA also ensure the display of metabolic pathways graphs, the simulation of the models based on ordinary differential equations systems solving and the study of network structures using elementary flux modes. We developed several web applications to allowing on one hand to interpret biological results by using databases, and on the other hand to export these data to MPSA. The main database of this work is myKegg. It includes all human KEGG metabolic pathways and a list of synonyms for human KEGG entries. This base allows to identify metabolic pathways from a list of biological compounds and to import them in MPSA. Another database, BioNMR, was developed to organize the data extracted from NMR spectra. A last web application named GeneProm was developed to analyze gene promoters. A promotology protocol was developed and tested on a set of four genes coding for the four ANT (adenine nucleotide translocator) protein isoforms. Each ANT isoform has a specific expression profile and role in cell bioenergetics. The promotology study of these four genes led us to construct specific regulatory models from identified regulatory elements in their promoter sequence. Potentially co-regulated genes were deduced from these models. Then they can be exported to our MPSA platform. This whole development will be included in the project of Integrative Biology platform in the INRA Human Nutrition Unit.

TABLE DES MATIÈRES

TABLE DES FIGURES

1 INTRODUCTION

1.1 CONTEXTE

L'encadrement de cette thèse a été partagé entre l'équipe Métabolisme Bioénergétique et Modélisation (MBM) de l'Unité de Nutrition Humaine (UNH) de l'INRA de Clermont-Ferrand Theix et la SA Soluscience. L'équipe de recherche MBM étant composée uniquement de biologistes, une collaboration avec Soluscience – société développant des systèmes d'informations pour des laboratoires de recherche et des applications web – a été initiée pour permettre un encadrement à la fois informatique et biologique de la thèse.

Les objectifs de cette thèse sont, d'une part de développer un logiciel permettant de chainer automatiquement des logiciels utilisés dans le cadre de l'étude de promoteurs et d'autre part de développer des outils d'intégration des données biologiques disponibles. Ces données sont issues d'analyses biologiques utilisant différentes techniques : métabolomique, transcriptomique et promotologie correspondant à différents niveaux d'étude d'une cellule présentés dans la figure 1.

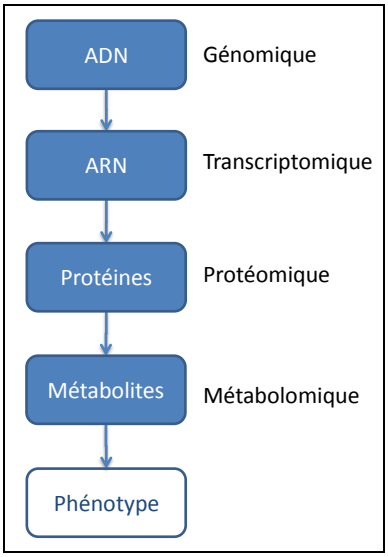


Figure 1 : Les différents niveaux d'étude d'une cellule

Métabolomique

La métabolomique est le champ de la biologie s'intéressant à l'étude des métabolites. Le terme de métabolome apparaît pour la première fois en 1998[86] dans une étude ou les auteurs réalisent la première étude de biologie des systèmes d'une cellule eucaryote (la levure) associant une analyse de la transcription des gènes (transcriptomique), des protéines (protéomique), du phénotype et des métabolites [86]. La métabolomique est l'étude de l'ensemble des métabolites présents à un temps donné. Cet ensemble de métabolites dépend du modèle biologique, de l'état physiologique, et de paramètres environnementaux. Le métabolome est le dernier d'expression des gènes et ainsi le plus caractéristique du phénotype. La métabolomique utilise des techniques d'analyse comme la Résonance

Magnétique Nucléaire (RMN) ou la Spectrométrie de Masse (MS) pour identifier les métabolites. Ces techniques de détection peuvent être couplées à des techniques de séparation comme la chromatographie en phase gazeuse (GC) ou la chromatographie en phase liquide à haute performance (HPLC).

Transcriptomique

La transcriptomique est l'étude du transcriptome, c'est-à-dire l'ensemble des molécules d'ARN (messager, transfert, ribosomique, micro, …) présentes à un instant donné dans une cellule. L'étude du transcriptome permet d'obtenir des informations sur la régulation de transcription des gènes. Le transcriptome est étudié en utilisant des puces à ADN, la RT-PCR ou le séquençage à haut débit.

Promotologie

La promotologie est l'étude des promoteurs des gènes. Les promoteurs sont des zones de l'ADN permettant de réguler la transcription des gènes. Ces promoteurs sont situés à proximité du gène qu'ils contrôlent, le plus souvent en amont du site d'initiation de transcription du gène. Des séquences régulatrices peuvent aussi être retrouvées dans les introns des gènes (parties non codantes d'un gène). Les promoteurs sont composés d'un ensemble de sites de fixation de facteurs de transcription (TFBS : *Transcription Factor Binding Site*). Ces TFBS sont de courtes séquences de 10 à 50 nucléotides environ reconnues par des protéines spécifiques, les facteurs de transcription. Ces facteurs de transcription reconnaissent les TFBS et s'y fixent. En fonction des facteurs de transcription fixés sur son promoteur, la transcription d'un gène pourra être activée ou réprimée.

1.2 PROBLÉMATIQUE

1.2.1 Besoins fonctionnels

Un des points forts de notre équipe de recherche est d'avoir les compétences nécessaires à l'étude de données de sources très différentes : métabolomique, transcriptomique, et promotologique. Il nous est nécessaire de trouver ou construire des outils bioinformatiques nécessaires à l'exploitation de ces différentes données en vue de comprendre les mécanismes biologiques mis en cause.

Une grande partie de nos données de métabolomique sont sous forme de spectres RMN HRMAS (*High Resolution Magic Angle Spinning*). Chaque spectre permet d'identifier une cinquantaine de métabolites dans un broyat cellulaire. Ces spectres sont dans un format propriétaire impossible à interpréter sans outil informatique. Ils doivent être calés horizontalement (la référence pour ce calage est le pic de la créatine à 3.035 ppm) et la ligne de base doit pouvoir être redessinée. Pour cela nous utilisons le logiciel Mnova. Les spectres que nous traitons sont des spectres RMN HRMAS à une ou deux dimensions. Nous étudions principalement des spectres du proton mais nous disposons aussi de spectres du carbone et du phosphore. Nous avons dû développer deux logiciels pour cette partie du travail : une application web couplée à une base de données permettant d'organiser les spectres de nos différentes expériences et d'en présenter les principales informations (BioNMR) ainsi qu'un logiciel en Python permettant la visualisation de spectres avec zooms et calculs d'aires sous les courbes (pyNMR). Nous avons aussi développé une procédure d'extraction des signaux pour le

logiciel Mnova permettant de trouver les signaux significatifs d'un spectre (rapport signal/bruit fixé par Mnova et non modifiable) pour les enregistrer dans BioNMR.

En ce qui concerne la promotologie, un premier projet (HumProm) avait été initié il y a plusieurs années (projet LifeGrid 2006-2008). HumProm consistait en un système d'information, mis au point par la société Soluscience, permettant d'analyser les résultats fournis par la base de données Genomatix que nous utilisons pour nos études des promoteurs. La technologie utilisée à l'époque dans ce système d'information était encore assez instable et nécessitait des installations lourdes sur les postes de travail. De plus, ce système d'information s'est avéré inutilisable à haut débit compte tenu de son interaction directe avec le site Genomatix. Enfin, ce système n'était disponible que sous Windows. Pour ces différentes raisons, nous avons initié la construction d'une nouvelle version de ce logiciel sous forme d'une application web appelée GeneProm.

Le développement principal porte sur un logiciel permettant de prédire des réseaux métaboliques à partir de données de différentes sources biologiques et de permettre une étude de ces réseaux. Nous ne disposons pas de données propres de transcriptomique (profils de puces à ADN) mais ce type de données étant apporté par d'autres équipes en collaboration, le logiciel développé devait pouvoir les prendre en compte. Ce développement consistera en un plugin pour le logiciel VANTED : MPSA. Ce logiciel devait permettre une représentation graphique des réseaux avec plusieurs styles de représentations possibles (*layout*). Les réseaux devait être prédis à partir de la base de données KEGG qui était très utilisée dans le laboratoire. Il était donc nécessaire de pouvoir représenter ces réseaux à partir des fichiers d'export de KEGG (les fichiers KGML, cf. 2.3.4.1). Ces réseaux devaient également pouvoir être exportés dans différents standards comme le SBML (cf. 2.3.3.3) ou le GML (langage de description des graphes) mais aussi dans un format XML permettant de communiquer avec un logiciel de simulation et d'étude des réseaux biologiques développé au Laboratoire d'Informatique (LIX) de l'École Polytechnique de Palaiseau [6].

1.2.2 Étude de l'existant

1.2.2.1 Genomatix

Genomatix est une suite de logiciels en ligne, nécessitant un abonnement payant, couplée à différentes bases de données dont une base de données de sites de liaison d'éléments de régulation ou matrices de régulation (TFBS : *Transcription Factors Biding Sites*) appelée MatBase. Les TFBS sont représentés sous forme de PWM (*Position Weight Martices*) [10]. L'outil MatInspector fourni par Genomatix permet de rechercher ces matrices sur des séquences nucléotidiques [97]. Les matrices contenues dans la base sont extraites de la littérature. Une matrice est composée d'une séquence *core* fortement conservée et de séquences *flanquantes* moins bien conservées. Les séquences *core* sont utilisées pour construire des familles de matrices. Une famille de matrices est composée de toutes les matrices ayant la même séquence core et des profils de conservation proches [13]. Genomatix définit aussi un « seuil optimal de conservation » des matrices [13]. Ils partent du constat que les matrices ayant des tailles et des profils de conservation très différents, il est impossible d'utiliser le même seuil de conservation pour toutes les matrices. Le seuil optimal de conservation est le seuil pour lequel une matrice donnée va être retrouvée au maximum trois fois pour 10 kpb dans un ensemble de séquences non régulatrices. L'utilisation de ce score permet de réduire le nombre de faux positifs et de faux négatifs [13]. Nous construisons sur Genomatix des modèles de promoteurs de gènes d'intérêt à partir d'éléments de régulation potentiels détectés par l'outil MatInspector. Le site nous permet la recherche de ces modèles soit sur l'intégralité d'un génome (GenBank pour l'homme), une liste de positions sur des clones est alors retournée, soit dans une banque de promoteurs ne contenant qu'environ un tiers des promoteurs humains. L'utilisateur doit analyser la liste des clones afin de retrouver des gènes

potentiels à proximité des occurrences du modèle. Le protocole complet que nous avons mis au point sur Genomatix est explicité au paragraphe 4.1.

1.2.2.2 HumProm

Le développement de HumProm a été initié en 2005 dans le cadre d'une collaboration avec la société Soluscience et nous l'avons achevé dans le cadre du programme Lifegrid en 2008. Il permettait de se connecter au site Genomatix et de réaliser le protocole complet de notre étude directement depuis le logiciel par envoi et interception de flux HTTP mimant les actions d'un utilisateur sur les pages du site. Compte tenu des flux élevés générés par ces analyses et les solutions de criblage haut débit proposée par Genomatix, ils ne nous ont pas autorisés à poursuivre ce protocole après 2008, ce qui a conditionné la réécriture du programme. Les algorithmes d'analyses des résultats Genomatix ont été récupérés pour être améliorés et réutilisés dans GeneProm. Par contre l'analyse Genomatix est maintenant réalisée via leur portail web et les résultats sont exportés pour être interprétés par GeneProm. L'intégralité du protocole d'analyse est détaillée dans la section 3.2.2.

1.2.2.3 Logiciels de représentation et d'étude des systèmes biologiques

Il existe de très nombreux logiciels d'étude des systèmes biologiques : en septembre 2011, 230 logiciels étaient enregistrés sur le site du SBML.org [169]. Il a donc été nécessaire de choisir différents critères de sélection :

- Logiciel disponible sous Windows qui est le système d'exploitation exclusivement utilisé dans l'équipe
- Représentation graphique des réseaux avec possibilité de modification
- Importation des graphes de réactions à partir de KEGG (lecture et écriture de fichiers au format KGML : format d'échange de la base de données KEGG) et édition de ces graphes
- Import et export des graphes sous des formats standards : SBML, GML, … Le SBML servant pour l'échange des données du modèle. Le GML est un langage générique de description des graphes utilisé pour l'import dans des logiciels de visualisation des graphes plus spécialisés
- Possibilité d'étendre les fonctionnalités du logiciel via un système de plugins. Cela sous-entend que le code source doit être disponible et que l'interface de programmation (API : *Application Programming Interface*) du logiciel doit être documentée.

En prenant en compte la disponibilité sous Windows et la disponibilité du code source, il ne reste que 113 utilisables. Les autres critères ne peuvent pas être filtrés en utilisant le tableau récapitulatif présenté sur le site du SBML [169]. Les logiciels restant doivent donc être filtrés manuellement. Parmi tous les logiciels testés, le plus intéressant était Cytoscape.

Cytoscape est un logiciel de visualisation de graphes biologiques codé en Java, disponible sous différentes plateformes dont Windows. Il dispose d'un système de développement de plugins actuellement bien documenté, mais ce n'était pas le cas en 2008. Il supporte différents formats de fichier en lecture et écriture : SIF (*Simple Interaction Format*), GML, BioPAX, PSI-MI, SBML, … Il est aussi possible de lui fournir des fichiers texte de type csv ou délimité par des tabulations. La visualisation des graphes est très bien développée et performante. Malheureusement il ne permettait pas la représentation des fichiers KGML de KEGG, ce qui est maintenant possible via un plugin. Ce logiciel est, à la base, tourné vers la représentation des graphes biologiques et leur manipulation, il

propose donc de très nombreuses options de représentations des différents éléments du graphe accessibles surtout aux utilisateurs avertis.

Le seul éditeur de fichiers KGML que nous avons trouvé était KGML-ed [63] Aucun système de plugin n'était cependant décrit. Ne trouvant aucun logiciel remplissant la contrainte de représentation des fichiers KGML, nous avons décidé d'essayer de construire un tel logiciel en nous basant sur les bibliothèques de représentation de graphes de la société Soluscience et d'inclure cet outil dans un système d'information. Ce logiciel a été présenté sous forme d'un poster à Cambridge à l'*international symposium on integrative bioinformatics* en avril 2010. À cette occasion, nous avons découvert le logiciel VANTED [52] (*Visualization and Analysis of Networks conTaining Experimental Data*) et le standard de représentation des systèmes biologiques SBGN (*Systems Biology Graphical Notation)*.

1.2.2.4 VANTED (Visualization and Analysis of Networks conTaining Experimental Data)

VANTED est un logiciel de représentation et d'analyse de voies métaboliques. Il est basé sur l'éditeur de graphes Gravisto [44] lui-même basé sur la bibliothèque Graffiti. VANTED intègre non seulement les bibliothèques de graphe et le système de plugins de Gravisto mais aussi le logiciel KGML-ed lui aussi basé sur Gravisto. Le logiciel VANTED permet donc, de représenter des voies métaboliques importées de KEGG et de développer des plugins permettant d'étendre les possibilités du logiciel. VANTED gère donc le KGML via KGML-ed. Il gère aussi de nombreux formats d'échanges de graphes comme le GML via Gravisto. L'interface de programmation utilisée pour le développement du plugin MPSA sera décrite dans le chapitre 3.3.2.

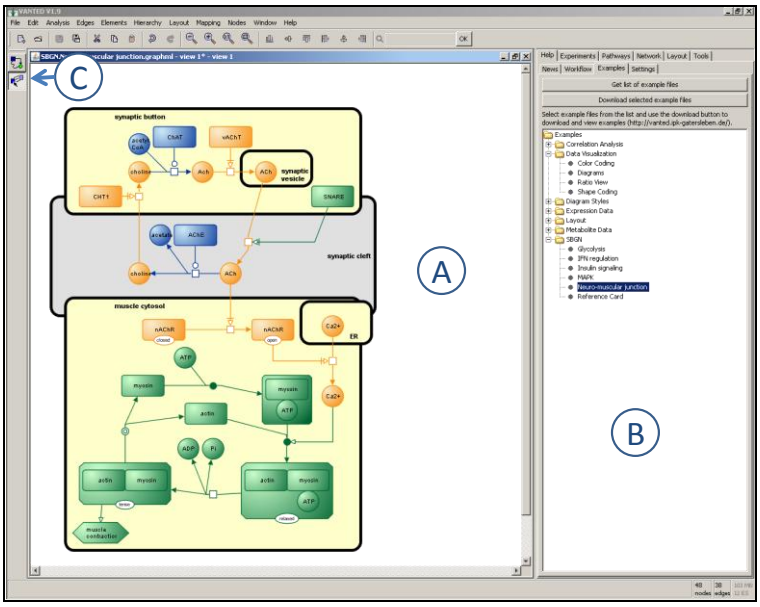


Figure 2 : Interface générale du logiciel VANTED
La partie A correspond à la fenêtre de manipulation des graphes. La partie B contient tous les outils du logiciel. Les boutons C permettent de passer d'un mode d'édition (ajout d'éléments au graphe) à un autre (modification des éléments du graphe). Le graphe représenté est un fichier d'exemple de représentation d'un graphe au format SBGN (cf. 2.3.3.2)

L'interface de VANTED se compose de différentes parties :

- Une fenêtre de visualisation des graphes (A). Il peut y avoir plusieurs fenêtres de graphe en cours.
- Un cadre (B) avec différents onglets permettant d'agir sur le graphe (changement des formes des nœuds, des arcs, etc.), les onglets permettant d'accéder aux interfaces avec les bases de données de voies métaboliques, l'onglet permettant de charger les résultats expérimentaux, l'onglet des outils statistiques, etc. Le système de plugins de VANTED permet d'ajouter de nouveaux onglets. Les onglets présents dans ce cadre varient en fonction des propriétés des graphes. Les onglets s'appliquent à la fenêtre de graphe active.
- À gauche, deux boutons (C) permettent de passer du mode « ajout de nœuds et d'arc » au mode « déplacement des nœuds et des arcs ».
- Une barre d'outil située juste au-dessus de la fenêtre (A) permet des opérations simples d'édition des graphes (copier, coller de nœuds), les opérations de zoom (accessibles via la molette de la souris) et l'alignement de nœuds sélectionnés (gauche, droite, en bas ou en haut de la fenêtre)
- La majorité des commandes accessibles dans la barre supérieure sont accessibles soit via les onglets de (B) soit via un menu contextuel de la souris. Le système de plugin permet d'ajouter des éléments dans cette barre.

Une fonctionnalité majeure de VANTED est de permettre d'afficher des données expérimentales sur des graphes métaboliques et de les interpréter.

Import de données

L'import de données correspond à l'import de données expérimentales dans VANTED. L'import de ces données peut se faire de manière indépendante de la création des graphes. Les données doivent être importées en utilisant une feuille de calcul Excel dont le format est fourni. La structure à respecter est précise et organisée pour décrire une expérience. Un schéma de cette structure est présenté dans la figure 3. *Remarque : VANTED a été développé en collaboration avec un laboratoire travaillant sur les plantes donc la plupart des illustrations tirées de la documentation du logiciel seront sur des expériences liées aux végétaux.*

Dans ces fichiers, l'élément le plus haut est l'expérience qui décrit la comparaison de différents génotypes, pour lesquels des mesures de concentrations de différents métabolites ont été faites pour différents temps. Toutes ces informations sont présentes dans la feuille Excel représentant cette expérience (figure 4).

L'expérience est décrite par une date de début d'expérience, des remarques (optionnelles), un nom (ce nom doit être unique si on importe plusieurs expériences dans VANTED), un expérimentateur. Le champ *Sequence-Name* correspond à un code qui peut être entré pour faire correspondre l'expérience à une entrée dans une base de données par exemple (figure 4).

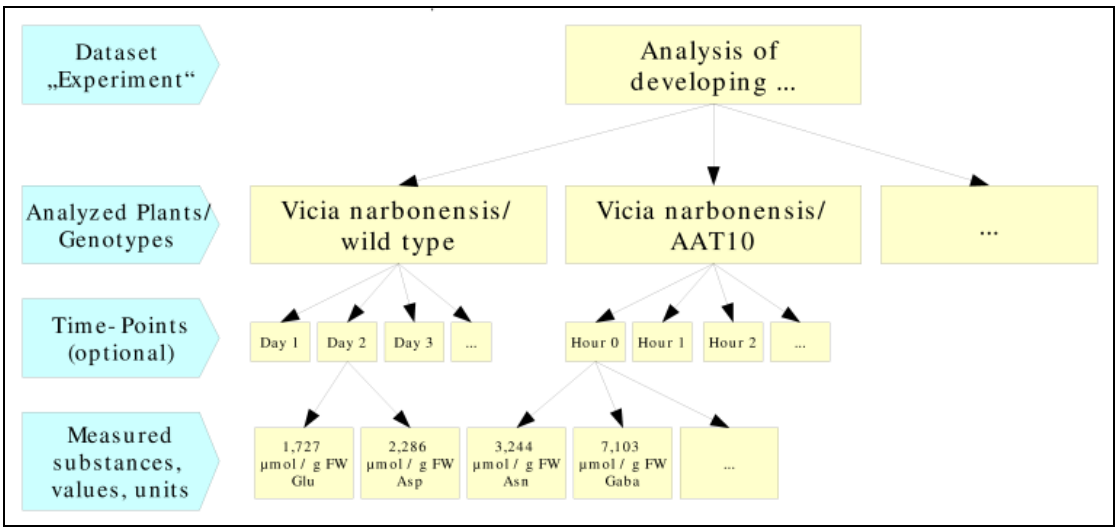


Figure 3 : Représentation hiérarchique de la structure d'un fichier d'import de résultats dans VANTED.
Tiré de [148]. Cette représentation hiérarchique décrit une expérience de comparaison du métabolisme de différents génotypes d'une même plante (Vicia narbonensis).

	A	B	C	D
1	**Input Template**			
2				
3	**Experiment**			
4	Start of Experiment (Dat	16/06/05		
5	Remark*			
6	Experiment Name (ID)	Analysis of developing vicia narbonensis ...		
7	Coordinator	Dr. Smith		
8	Sequence-Name*			

Figure 4 : Informations relatives à l'expérience reprises dans la feuille d'import des données dans VANTED.
Tiré de [148]

	A	B	C
⋮			
11	**Plants/Genotypes****	1	2
12	Species	Vicia narbonensis	Vicia narbonensis
13	Variety*		
14	Genotype	wild type	AAT10
15	Growth conditions*		
16	Treatment*		
17			

Figure 5 : Informations relatives aux organismes de l'expérience
Tiré de [148]

La feuille Excel est faite pour pouvoir comparer des génotypes de plantes entre eux. Les champs décrivant ces génotypes sont l'espèce, la variété (optionnelle), le génotype, les conditions de culture (optionnelles) et le traitement (figure 5). Cette formulation est suffisamment générique pour pouvoir être appliquée à n'importe quel organisme. Pour une expérience sur l'homme, les champs seraient :

Espèce : Homo sapiens

Génotype : si l'effet d'un même traitement est comparé sur des patients, les identifiants correspondraient aux différents patients. Si ce sont les effets de la dose sur des cellules cancéreuses en culture qui sont étudiés, le type de cellules cancéreuses peut être indiqué, etc.

Conditions de culture : sans objet pour des travaux sur des patients.

Traitement : nom de la molécule utilisée pour le traitement.

La première ligne de la description des génotypes (ligne 11 de la feuille de la figure 5) correspond à des identifiants dans la suite du fichier.

	A	B	C	D	E	F	G	H
⋮								
20	**Measurements**				**Substance**	Asp	Glu	Ser
21					**Meas.-Tool***	HPLC	HPLC	HPLC
22	**Plant/Genotype*****	**Replicate #**	**Time***	**Unit (Time)***	**Unit**	µmol / g FW	µmol / g FW	µmol / g FW
23	1	1	0	day		1,19	5,62	2,08
24	1	1	2	day		2,05	5,72	4,95
25	1	2	2	day		2,09	6,93	3,03
26	1	1	4	day		2,34	6,61	2,91
27	1	2	4	day		2,94	6,67	2,58
28	1	3	4	day		2,94	6,67	2,58

Figure 6 : Mesures
Tiré de [148]

Les mesures présentées dans la figure 6 se rapportent aux génotypes identifiés dans la partie précédente du fichier. Il est possible de tenir compte des réplicats (optionnel), des temps de mesures et des unités de temps (optionnel). *Remarque : la déclaration des unités de temps ne sert à rien dans VANTED puisqu'elles ne sont pas prises en compte dans les calculs, seul le chiffre est reconnu, donc pour le logiciel 30 secondes seront plus grandes qu'une minute.* Les mesures proprement dites commencent à la colonne F de la feuille de calcul. Les entêtes des colonnes doivent comporter : le nom du composé mesuré, la technique de mesure et l'unité. Ici encore, les unités ne sont pas utilisées dans le logiciel, il revient à l'utilisateur de rentrer des valeurs comparables entre elles. Il est en revanche possible de comparer les résultats des différentes techniques puisque l'utilisateur doit choisir les colonnes à comparer et qu'il ne peut y avoir qu'une seule technique par colonne.

Affichage de données sur des graphes

Les données importées dans le logiciel peuvent être placées sur les graphes de deux manières différentes. Si un graphe existe déjà le logiciel tentera d'afficher les données sur les nœuds déjà existant en comparant les noms des métabolites entrés dans la feuille Excel avec les noms des métabolites du graphe. S'il trouve des métabolites dans la feuille qui ne correspondent à aucun nœud dans le graphe, le logiciel créera un nouveau nœud. À l'utilisateur de placer correctement ce nœud correctement dans le graphe. Si aucun graphe n'a été créé, VANTED crée un graph *de novo* en créant

les nœuds correspondant aux métabolites et l'utilisateur devra les relier. Les valeurs expérimentales sont conservées lors d'un export au format GML.

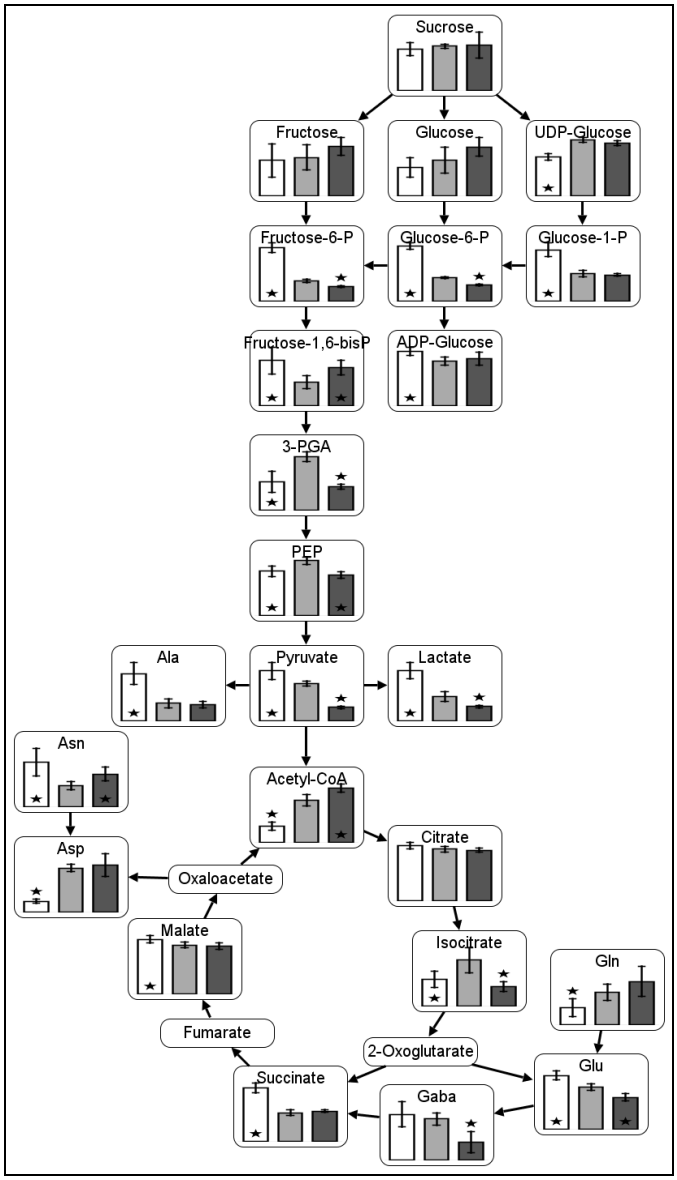


Figure 7 : Représentation de données d'expériences sur un graphe
Cette figure est tirée des exemples de représentation du logiciel VANTED. Elle représente des données comparant trois séries de données comparant le métabolisme (cycle de Krebs et glycolyse) de graines de soja en réponse à différentes concentrations d'un pesticide. Les trois séries de données sont représentées en utilisant des histogrammes. Un T-test a été effectué avec pour échantillon de référence la deuxième série. Les étoiles correspondent à une P-value inférieure à 5%.

Il est possible de représenter les données de différentes manières dans VANTED. En présence de réplicats, il est possible d'afficher la variance des mesures sur les histogrammes et sur les courbes. La figure 8 représente des données qui ne sont pas organisées sur un graphe, la figure 7 représente des données appliquées à un réseau métabolique.

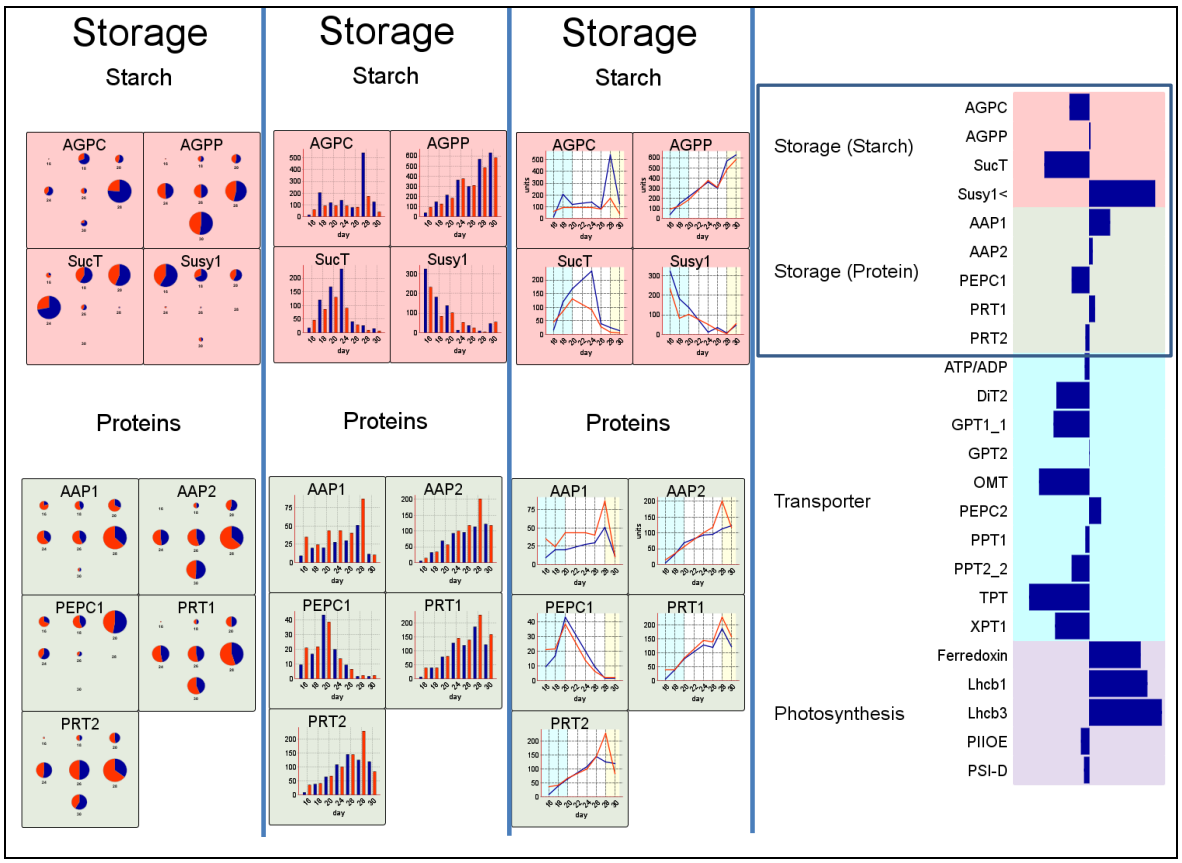


Figure 8 : Quatre types de représentations différentes des mêmes données dans VANTED
Tiré des exemples des fichiers d'exemples de VANTED. Les données correspondent à la comparaison de deux génotypes de fabacées (le type sauvage en bleu et une lignée en rouge où le gène codant pour le transporteur glucose-6-phosphate 1 est inhibé)[102]. Les quatre types de représentation sont : les diagrammes circulaires (ici en vue de dessus, mais peuvent être représentés en trois dimensions), les histogrammes (pouvant être représentés aussi en 3d) et les courbes. La représentation la plus à droite permet une comparaison de l'expression de gènes ou de la concentration de métabolites. Les données encadrées dans la partie de droite correspondent aux données des trois autres vues pour le jour 24 (quatrième série).

Analyses

Un des points forts de VANTED est de permettre de faire des analyses statistiques complexes de manière très simple. Comme montré dans la figure 7, un T-test peut être réalisé afin de trouver les séries variant par rapport à une série de référence pour un même composé. Les résultats d'un T-test ne sont visibles que sur les histogrammes. Cependant les analyses les plus intéressantes sont les analyses qui comparent les distributions de différents composés entre elles.

La première méthode est l'analyse de corrélation. Cette analyse peut être représentée par une matrice de nuages de points (*scatter plot matrix)*.

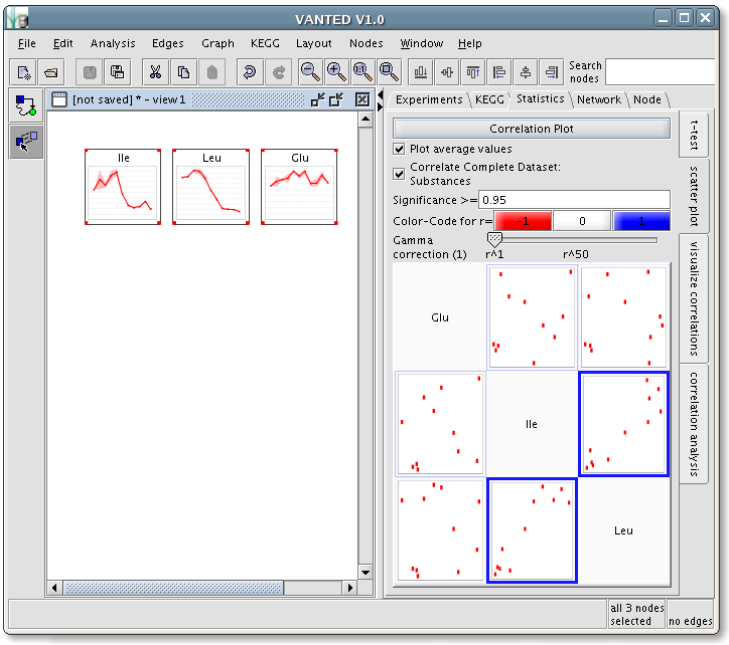


Figure 9 : Visualisation de corrélation par matrice de nuage de points
Les données comparées ici sont représentées dans la partie gauche de la figure, dans la partie correspondant à l'éditeur de graphe de VANTED. La partie statistique est ouverte à droite de la figure. La matrice de nuages de points permet la représentation de la corrélation des variations de concentrations des trois composés Isoleucine (ile), leucine (leu) et glutamate (glu). Les nuages de points encadrés en bleu correspondent aux séries les plus corrélées (coefficient de corrélation proche de 1), les cases non encadrées correspondent aux séries non corrélées (coefficient de corrélation proche de zéro), s'il y avait eu des séries inversement corrélées, elles auraient été encadrées en rouge.

Une telle matrice est une matrice carrée symétrique M de taille n telle que :

M_{ij} = nuage de point V_i en fonction V_j. V_i étant le vecteur des valeurs du composé de la ligne *i* et V_j le vecteur de valeurs du compos de la colonne *j*.

Le coefficient de corrélation des deux séries est alors calculé puis représenté par une échelle de couleurs : bleu pour une corrélation proche de 1, blanc pour une corrélation nulle et rouge pour une corrélation proche de -1.

Deux mesures de la corrélation ont été implémentées dans VANTED : la corrélation linéaire de Bravais-Pearson et la corrélation de Spearman. Pour une corrélation linéaire, l'étude de corrélation n'a de sens que si les séries sont linéaires. La valeur de la corrélation de Spearman sera forte si les deux séries sont monotones sans forcément être linéaires. Aucune corrélation ne sera trouvée pour des séries cycliques dans un cas comme dans l'autre. Nos données ne respectent pas ces contraintes et l'analyse par corrélation ne peut donner aucun résultat.

Toujours en utilisant les corrélations, il est possible de trouver les séries variant de la même manière qu'une série d'intérêt et de les afficher dans le graphe.

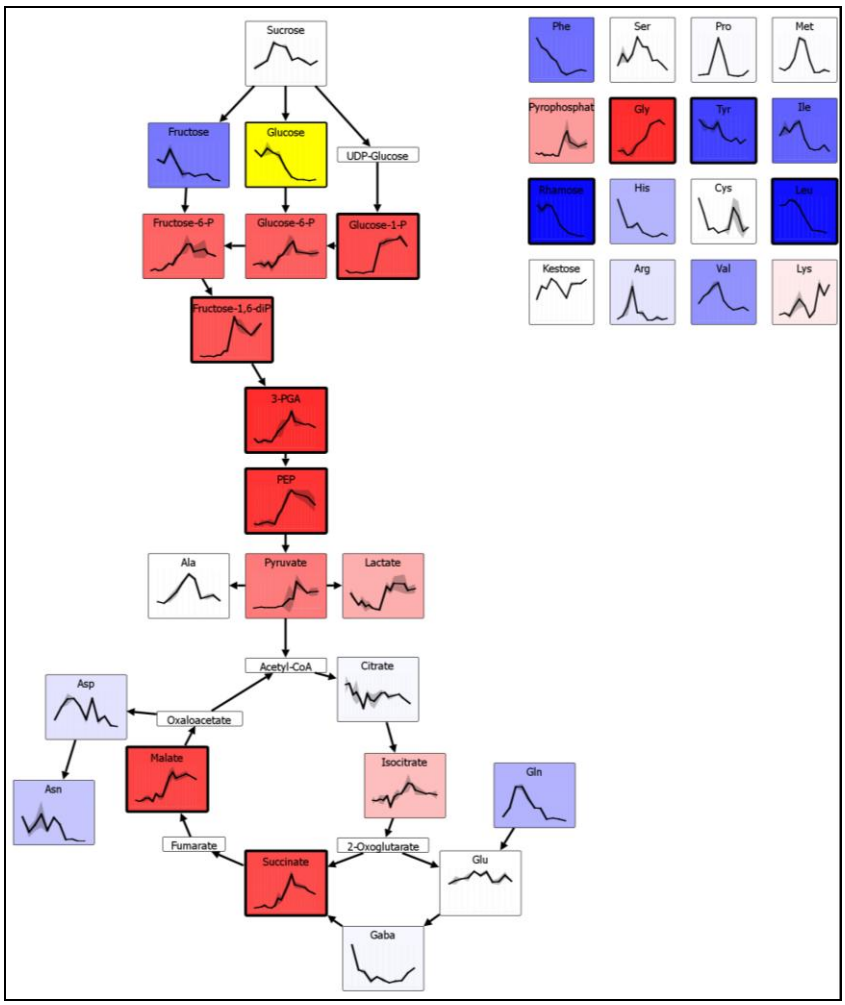


Figure 10 : Représentation de l'analyse de corrélation sur un graphe
Tiré des fichiers d'exemple du logiciel VANTED. Les couleurs représentent les coefficients de corrélation des différentes séries en fonction de la série en jaune (choisie par l'utilisateur). La voie métabolique est la même que dans la figure 7.

Dans ce type de représentation, le métabolite de référence est représenté en jaune. Plus le fond de la cellule représentant le métabolite est rouge plus le coefficient de corrélation est proche de -1 c'est-à-dire que la variation du composé est inversement corrélée à celle du composé de référence. Au contraire, les composés sur fond bleu varient de la même manière que le composé de référence. Dans la figure 10, la corrélation inverse entre le glucose (en jaune) et le glucose-6-phosphate indique que pour pouvoir produire du glucose-6-phosphate, il faut consommer du glucose. Le fait que la plupart des variations de concentrations de composés soient inversement proportionnelles à celle du glucose et du fructose montre bien que ces deux composés sont des composés de base pour la synthèse des autres métabolites du réseau. Les métabolites qui ne sont pas liés dans le réseau correspondent pour la plupart à des acides aminés qui sont synthétisés lors de réactions en parallèle du cycle de Krebs. Tous ces métabolites semblent être corrélés au glucose, mais on ne peut pas réellement faire de lien entre ces variations : les acides aminés sont consommés rapidement par l'organisme, ce qui explique la baisse de leur concentration au cours du temps.

VANTED permet aussi de représenter les corrélations entre plusieurs composés en même temps. Pour cela, il calcule les coefficients de corrélation entre tous ces composés et les relie par un

arc de couleur (la couleur reflétant le coefficient de corrélation trouvé). Cette représentation est peu claire donc elle ne fera l'objet d'aucune figure.

La méthode d'analyse la plus puissante implémentée dans VANTED est la méthode des cartes auto adaptatives : *self-organizing maps (SOM).* Les SOM sont une méthode de classification non supervisée qui est utilisée ici pour classer les composés dont les variations au cours du temps se ressemblent (cf. figure 11). Ce type d'analyse est plus intéressant que les analyses de corrélation car il n'est pas sensible à la distribution des données. Le paramètre le plus important pour réaliser cette classification est le nombre de *neurones* du modèle, c'est-à-dire le nombre maximal de groupes que l'on obtient à la fin de la classification. Cette classification est plus cohérente sur des expériences ayant de nombreux points à comparer comme les évolutions de concentrations de métabolites au cours du temps.

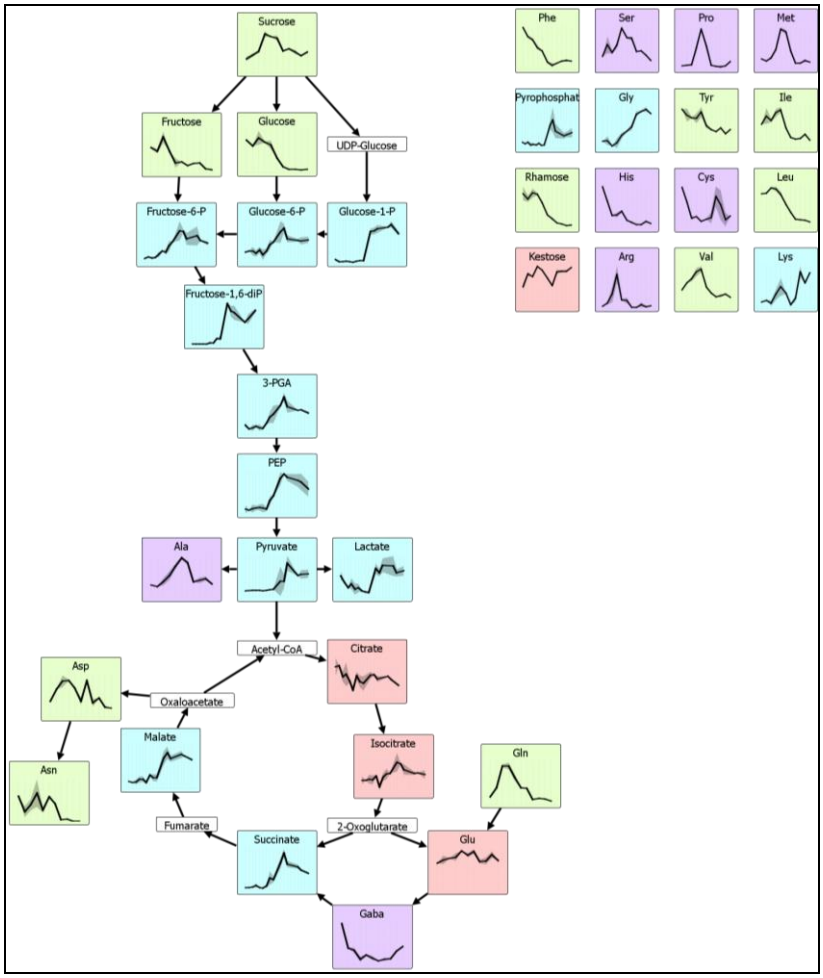


Figure 11 : Classification des métabolites par SOM
Le calcul de la SOM a été fait en imposant 4 neurones. Les quatre groupes trouvés sont représentés par des couleurs différentes directement sur le graphe.

En utilisant la classification des métabolites par SOM, on retrouve des résultats proches de la corrélation de la figure 10.

Corrélation de Spearman	**SOM**	
	En plus	**En moins**
Corrélation avec le glucose proche de 1 : glucose, fructose, aspartate, asparagine, glutamine, valine, histidine, leucine, isoleucine, tyrosine, phénylalanine et rhamose.	Sucrose	Histidine
Corrélation inversion avec le glucose (proche de -1) : Fructose-6-P, glucose-6-P, glucose-1-P, Fructose-1,6-bisP, 3PGA, PEP, pyruvate, lactate, malate, succinate, isocitrate, glycine, pyrophosphate.	Lysine	Isocitrate
Pas de corrélation avec le glucose (non classés): Sucrose, alanine, citrate, glutamate, GABA, serine, proline, methionine, cysteine, kestose, lysine.	Citrate, isocitrate, glutamate, kestose	
	GABA, alanine, arginine, histidine, cysteine, proline, methionine, serine	

Figure 12 : Comparaison des groupes trouvés par étude des coefficients de corrélation et par SOM.
À gauche sont notés les groupes obtenus par l'étude de la corrélation par rapport à la variation de la concentration du glucose : métabolites variant de la même manière que le glucose, métabolites variant de façon inverse par rapport au glucose et métabolites ne variant pas comme le glucose. À droite, dans le cas des deux premiers groupes, les métabolites présents dans les deux groupes ne sont pas marqués, les métabolites en plus dans le groupe SOM sont dans la première colonne, les métabolites présents dans le groupe de corrélation et pas dans celui de SOM sont dans la deuxième colonne. Pour les deux derniers groupes qui correspondent à la classification des métabolites n'ayant aucune corrélation avec le glucose, tous les métabolites sont notés.

La figure 12 montre que la classification effectuée par corrélation est retrouvée. Afin d'établir la classification par corrélation, il a fallu choisir un métabolite particulier dont la variation implique une perturbation globale du système. Cette classification aurait été totalement différente si le citrate avait été choisi comme métabolite de référence dans l'analyse par corrélation. Pour l'analyse par SOM, aucun métabolite de référence n'est choisi : ils sont regroupés entre eux par analyse de leur variation. Le seul paramètre choisi est le nombre de neurones, c'est-à-dire le nombre de groupes que l'on espère obtenir. Nous avons choisi quatre neurones afin de retrouver les mêmes groupes que dans la corrélation d'identifier une structure dans le groupe des métabolites ne corrélant pas avec le glucose.

Les concentrations de citrate, isocitrate, glutamate et kestose varient peu entre le début et la fin de l'expérience. Il semble donc logique de les regrouper ensembles. Le dernier groupe représente les métabolites qui n'ont pas pu être classés. Les métabolites regroupés au sein d'un même groupe peuvent être des métabolites impliqués dans les mêmes voies de synthèse (ou de dégradation) par exemple. Ici le groupe vert (du glucose) et le groupe bleu (du pyruvate) ont des comportements opposés (cf. figure 10) et pourtant ils sont impliqués dans une même voie métabolique, ce n'est donc pas parce que des métabolites sont impliqués dans des groupes différents qu'ils sont dans des voies différentes. Cela est d'autant plus vrai en métabolomique puisque les concentrations des métabolites sont reliées non seulement aux concentrations des autres métabolites mais aussi aux activités des enzymes et au taux d'expression de leurs gène. Ce type d'étude donne souvent des résultats plus simples à interpréter en transcriptomique [126,20]. On note aussi qu'une étude par SOM peut ne pas donner les mêmes résultats finaux avec les mêmes données et paramètres d'entrée car l'initialisation des neurones est aléatoire. Les valeurs les moins tranchées pourront être assignées à un groupe ou à un autre en fonction de la simulation.

Import de voies métaboliques KEGG

L'import d'une voie métabolique de KEGG peut se faire de deux manières différentes. La première, la plus simple consiste à ouvrir un fichier KGML directement récupéré sur la base de données KEGG. Seule cette manière fonctionne encore depuis que l'accès à une partie de la base de données KEGG est devenu payant. VANTED propose tout de même un accès direct au téléchargement d'une voie métabolique. Dans le panneau de droite du logiciel, dans la rubrique « Pathways » se trouve un onglet KEGG. L'utilisateur doit commencer par sélectionner l'organisme pour lequel il souhaite trouver une voie métabolique. S'il ne veut pas sélectionner d'organisme, il est invité à choisir « Reference Pathways (MAP) – map » dans le menu qui lui est proposé. Un champ de recherche est disponible. Cf. figure 13.

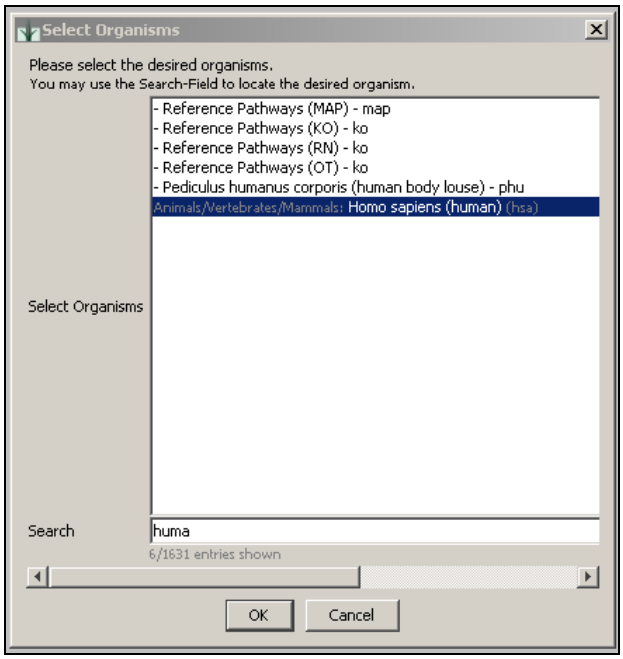


Figure 13 : Interface de choix des organismes pour la sélection des voies métaboliques
Interface correspondant à la recherche de l'espèce humaine parmi toutes les espèces proposées par KEGG.

Le logiciel propose ensuite plus de 200 voies métaboliques organisées par groupes de fonctions biologiques (base d'orthologie de KEGG : base KO). L'utilisateur peut alors trouver et charger la/les voies métaboliques qui l'intéressent.

2 BIOLOGIE DES SYSTÈMES

2.1 INTRODUCTION

La notion de réseaux est massivement utilisée en biologie depuis la seconde moitié du 20ème siècle. Elle est utilisée notamment pour représenter et mieux comprendre le fonctionnement des différents compartiments cellulaires. Ce processus a été amplifié par l'émergence de la génomique. Au 20ème siècle, une fonction est recherchée pour tous les gènes des génomes que séquencés. C'est l'avènement de la génomique, la transcriptomique et la protéomique. Au 21ème siècle, émergent les technologies « haut débit » permettant aux biologistes de voir les cellules comme des systèmes complexes et non plus comme des assemblages de compartiments cellulaires étudiés séparément. Le 21ème siècle est donc le siècle de la bioinformatique, de la modélisation mathématique et de la simulation par ordinateur [88]. La biologie intégrative développée depuis ces dernières années est nécessaire à l'étude de la physiologie tissulaire et à l'étude de nombreuses pathologies comme le cancer.

L'évolution des techniques d'analyse biologiques à haut débit et notamment des techniques de séquençage a permis le développement de la biologie des systèmes [46]. Le processus classique utilisé pour l'étude d'un modèle dans le cadre de la biologie des systèmes se déroule en quatre phases (cf. figure 14) :

1. Établissement de la liste des composés biologiques impliqués dans le modèle étudié
2. Recherche des interactions que ces composés ont entre eux permettant une reconstruction du réseau métabolique
3. Description mathématique ou informatique du réseau : de modélisation mathématique ou informatique
4. Ces différents modèles servent alors à analyser, interpréter, et prédire des résultats expérimentaux. Les prédictions conduisent à émettre des hypothèses devant être validées expérimentalement. L'expérimentation peut permettre de préciser le modèle initial.

La bioinformatique tient un rôle prépondérant dans la biologie des systèmes car elle seule est à même de pouvoir traiter les masses de données de la biologie des systèmes. La bioinformatique intervient à différents niveaux de la biologie des systèmes [132].

- Analyse des données. Les modèles informatiques sont très utilisés pour retrouver les mécanismes sous-jacents suggérés par les mesures expérimentales.
- Formulation d'hypothèses. Un modèle informatique est, par définition, une hypothèse tirée d'observations biologiques. La formulation de tels modèles et leur analyse permettent non seulement de proposer un fonctionnement pour un modèle biologique mais elles permettent aussi de prédire une évolution du métabolisme consécutive à des perturbations.
- Validation d'hypothèses. L'étude d'un modèle informatique permet de valider ou d'infirmer des hypothèses émises sur le modèle biologique (à condition que le modèle biologique et le modèle informatique soient cohérents). Souvent on formule des

hypothèses sur le modèle, on les vérifie, ce qui amène souvent à formuler de nouvelles hypothèses à vérifier, etc.

- Proposition d'expériences. L'émission de nouvelles hypothèses est souvent synonyme de nouvelles expériences biologiques à mener afin de valider ou non ces hypothèses.
- Génération de nouveaux modèles. Une fois différents modèles informatiques vérifiés, il est possible de les assembler pour créer de nouveaux modèles plus complexes.

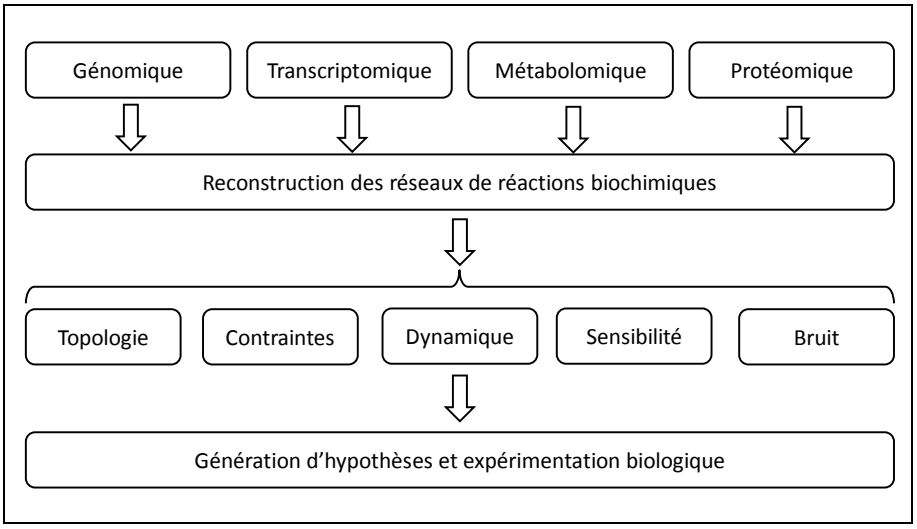


Figure 14 : Principales étapes mises en œuvre en biologie des systèmes.
La biologie des systèmes est basée sur la construction et l'étude de réseaux. Pour reconstituer ces réseaux, on fait appel à l'ensemble des « omiques ». Les réseaux reconstruits sont ensuite analysés par de nombreuses méthodes mathématiques dont quelques-unes sont citées ici. Cette analyse donne lieu à des hypothèses engendrant de nouvelles expérimentations biologiques. Ces expérimentations peuvent amener à modifier les réseaux construits et donc à faire une nouvelle étude.

2.2 ORIGINE DES DONNÉES BIOLOGIQUES

2.2.1 Transcriptomique

Les données de transcriptomique peuvent être utilisées de deux manières différentes. La première est de rechercher les voies métaboliques dans lesquelles les produits de gènes identifiés comme co-régulés sur une puce ADN sont impliqués. La seconde est comme point d'entrée à une étude de promotologie.

En effet, un ensemble de gènes co-exprimés est susceptible d'être régulé au niveau transcriptionnel par des éléments de régulation communs entre certains de ces gènes. Une étude de promotologie permettrait d'identifier certains de ces éléments régulateurs.

2.2.2 Promotologie

Comme décrite dans le paragraphe **0**, la promotologie est l'étude des promoteurs des gènes. Les données dont nous disposons sont principalement issues de deux bases de données : Genomatix décrite au paragraphe **1.2.2.1** et GeneProm (développé au cours de cette thèse). Les données que l'on peut extraire de ces deux bases concernent des gènes potentiellement co-régulés. Il serait intéressant de pouvoir relier les produits des gènes entre eux pour comprendre dans quelles voies métaboliques ces gènes sont impliqués et comment leurs produits interagissent entre eux. Ce processus est semblable à celui qui a été mis au point pour les études métabolomiques dans MPSA.

2.2.3 Métabolomique

Les données de métabolomique dont nous disposons au laboratoire sont issues pour la plupart de spectres RMN. Les données sont stockées sous deux formes : le spectre brut et une feuille Excel. Le spectre brut est au format propriétaire Bruker. Il contient toutes les données relatives à l'acquisition du spectre ainsi que le spectre en lui-même. Ce format de données ne peut cependant être lu que par un nombre restreint d'outils. Une partie des analyses des spectres sont faites directement via les outils d'analyse Bruker mais nous avons acquis une licence pour le logiciel Mnova produit par la société MestReLab permettant de lire ces spectres et de les représenter. Il permet en outre de calculer une ligne de base et de caler le spectre sur l'axe des abscisses (ppm) en utilisant des pics de référence.

La plupart des analyses se fait via une feuille Excel. Il est possible d'extraire les valeurs numériques des spectres pour les importer dans des logiciels de calcul. Cependant Excel ne permet pas d'importer autant de points, une étape de réduction de la résolution des spectres est nécessaire.

Pour retrouver les informations liées à l'expérience ayant conduit au spectre, il fallait faire référence à une feuille Excel regroupant les spectres par expérience, les modalités de l'expérience étant dans un cahier de laboratoire. Il était donc nécessaire de créer une base de données permettant de faire correspondre les données d'expérience, les fichiers de spectres et les feuilles Excel ayant servi à leur analyse. Toutes les données sont ainsi accessibles à chaque utilisateur. Cette base de données est la base BioNMR.

Via la plateforme d'exploration du métabolisme, nous sommes aussi à même de traiter des informations de spectrométrie de masse. L'étude du métabolisme via cette plateforme met en jeu différents types d'analyses : un couplage d'une chromatographie en phase liquide et une spectrométrie de masse (TOF : *Time Of Flight*) ou un couplage chromatographie liquide, RMN et spectrométrie de masse (*Metabolic Profiler*). La plateforme dispose de son propre système de stockage des données et fournit des résultats analysés ainsi que l'accès à des logiciels de traitement statistiques comme SIMCAP.

Les données de métabolomique reflètent l'état d'un organisme (ou d'un ensemble de cellules) à un instant donné. Pour interpréter des données de métabolomique, il est nécessaire de reconstruire le

réseau métabolique mis en jeu par le modèle biologique considéré. C'est le rôle du logiciel MPSA développé au cours de cette thèse.

2.3 LES RÉSEAUX BIOLOGIQUES

2.3.1 Définitions

Dans les sciences biologiques, les réseaux sont omniprésents. L'analyse de ces réseaux est donc particulièrement intéressante. Elle permet par exemple l'identification de cibles thérapeutiques, la détermination de fonctions de protéines ou de gènes ou de fournir un diagnostic précoce d'une pathologie. Différents types de réseaux ont été mis en évidence par la biologie des systèmes : les réseaux de signalisation, les réseaux de régulation transcriptionnelle, les réseaux d'interaction protéine-protéine et les réseaux métaboliques [91].

Les réseaux de signalisation

Les réseaux de signalisation sont constitués de l'ensemble des processus biochimiques impliqués dans la transduction d'un signal entre l'extérieur et l'intérieur d'une cellule et entre différents compartiments cellulaires. Les réseaux de signalisation sont impliqués dans la mise en œuvre de deux autres types de réseaux : régulation de transcription et métaboliques. Ce type de réseaux ne sera pas abordé plus en détails ici.

Les réseaux de régulation de transcription

L'expression, c'est-à-dire la transcription et la traduction du génome d'une cellule est soumise à une étroite régulation. Une cellule ne peut assurer son rôle que si cette régulation est fonctionnelle. Ainsi, seule une partie du génome cellulaire s'exprime à un temps donné pour un type cellulaire donné. Les gènes sont inhibés ou activés en fonction des différents signaux perçus par la cellule. Un gène est généralement sous contrôle de tout un ensemble de protéines. Ces protéines reconnaissent différentes séquences nucléotidiques localisées en général dans des zones non codantes en amont du gène. Ces séquence sont appelées éléments de régulation. L'ensemble de ces séquences constitue le promoteur du gène. *Remarque : l'état de condensation de la chromatine et la méthylation de l'ADN jouent aussi un rôle important dans le contrôle de l'expression des gènes*. Ces processus ne sont pas décrits dans les réseaux de régulation de transcription.

Les réseaux d'interaction protéine-protéine

Les réseaux d'interactions protéine-protéine décrivent comment une grande majorité des protéines interagissent entre elles pour assurer leur fonction. Les réseaux d'interactions protéines-protéines sont très utilisés pour prédire la fonction d'une protéine inconnue. Ces réseaux ne seront pas d'avantage décrits ici.

Les réseaux métaboliques

Les fonctions cellulaires sont assurées par des réseaux complexes de composés biochimiques interagissant entre eux. C'est l'ensemble de ces réactions biochimiques qui constitue les réseaux métaboliques. Une réaction biochimique fait en général intervenir un ou plusieurs substrats qui, en présence d'une enzyme réagissent pour former différents produits. Une réaction biochimique peut

impliquer tous les types de composés présents dans une cellule : glucides, lipides, protéines, petites molécules, etc. C'est ce type de réseaux qui va plus particulièrement nous intéresser ici.

Ces différents types de réseaux sont utilisés en biologie des systèmes pour mieux comprendre un fonctionnement observé expérimentalement. Les réseaux constituent alors une hypothèse de fonctionnement possible du système défini comme un modèle. Pour le *National Library of Medicine* un modèle est « *une représentation théorique qui simule le comportement ou l'activité des processus biologiques ou des maladies. Les modèles biologiques impliquent l'utilisation d'équations mathématiques, les ordinateurs et autres équipements électroniques* [163] ». La convention MIRIAM (*Minimum Information Requested In the Annotation of biochemical Models*) définit les informations nécessaires à la description de modèles biochimiques.

2.3.2 Convention de description d'un modèle : MIRIAM

MIRIAM est une convention apparue en 2005 [83]. Elle organise les informations nécessaires en trois catégories : les références du modèle, l'annotation du modèle et les liens vers des ressources externes [146].

Les références du modèle. Le modèle doit être décrit dans un format public, standardisé et lisible par ordinateur (SBML, CellML, BioPAX, … cf. paragraphe 2.3.3.3). L'écriture dans ce format doit être valide. Le modèle doit être associé à une description sous forme de publication scientifique par exemple. Si le modèle est composé de différentes parties, chacune des parties doit être décrite. La structure du modèle doit correspondre aux processus biologiques présentés dans la description. Le modèle doit être accompagné d'une simulation où les résultats et les valeurs initiales des paramètres doivent être définis. La simulation doit reproduire les résultats de la description.

Annotation du modèle. Le modèle doit avoir un nom. La publication de référence doit être citée (citation complète et non ambigüe). Les auteurs doivent être indiqués avec un moyen de les contacter. La date et l'heure de la dernière modification du modèle doivent être précisées. Les termes de distributions doivent être clairement énoncés.

Liens vers des ressources externes. L'annotation du modèle doit permettre de lier une information supplémentaire à tout élément du modèle. L'information doit être présentée en suivant le schéma : {type de donnée, identifiant, qualificatif}. Le type de donnée est fourni sous forme d'URI (*Uniform Resource Identifier :* chaîne de caractères identifiant une ressource sur un réseau), l'identifiant est celui impliqué dans le type de données. Le type de données et l'identifiant peuvent être combinés dans une même URI. Par exemple urn:miriam:uniprot:P62158 fait référence à la protéine P62158 (identifiant) dans la base de données UniProt (type de données). Le qualificatif est optionnel. Il ajoute des informations à la référence comme par exemple « homologue à … ». Ces URI doivent être standardisées pour être compréhensibles par tous. L'EBI (*European Bioinformatics Institute*), à l'origine de ce standard, a mis en place une API permettant de générer ces URI et de les comprendre [150].

```
1   <species metaid="heme" id="heme" compartment="comp" initialConcentration="0">
2       <annotation>
3           <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
4               xmlns:bqbiol="http://biomodels.net/biology-qualifiers/">
5               <rdf:Description rdf:about="#heme">
6                   <bqbiol:hasPart>
7                       <rdf:Bag>
8                           <rdf:li rdf:resource="urn:miriam:uniprot:P69905" />
9                           <rdf:li rdf:resource="urn:miriam:uniprot:P68871" />
10                          <rdf:li rdf:resource="urn:miriam:obo:chebi:CHEBI%3A17627" />
11                      </rdf:Bag>
12                  </bqbiol:hasPart>
13              </rdf:Description>
14          <rdf:RDF>
15      </annotation>
16  </species>
```

Figure 15 : Description de l'hémoglobine par des annotations MIRIAM inclus dans un fichier SBML
Pour plus d'informations sur le MIRIAM cf. 2.3.3.3. La ligne 1 définit le nom de l'entité décrite comme étant « heme *». La ligne 5 montre que la description s'applique sur l'objet «* heme *». Les lignes 6 à 11 décrivent l'objet comme un complexe de deux protéines décrites dans UniProt (P69905 : sous-unité alpha de l'hémoglobine humaine et P68871 : sous-unité beta de l'hémoglobine humaine) et un hème ferrique décrit dans CHEBI (identifiant CHEBI:17627).*

2.3.3 Représentation mathématique et informatique

2.3.3.1 Les graphes

En mathématique et en informatique les réseaux sont représentés par des structures complexes appelées « graphes ». Les graphes sont utilisés pour résoudre des problèmes dans tous les domaines impliquant des réseaux comme par exemple les réseaux sociaux, réseaux informatiques, Télécom…

Un graphe est un ensemble de points pouvant être reliés entre eux par un ou plusieurs liens. Les points sont appelés « sommets » ou « nœuds ». Les liens peuvent être non orientés (dans un graphe non orienté), on parle alors « d'arrêtes » ou orientés (on parle alors de graphe orienté), ils sont alors appelés aussi « arcs ». L'ensemble des arrêtes est souvent noté « E » (pour *edges* en anglais) et l'ensemble des sommets « V » (pour *vertices*). Le graphe est alors noté G(V,E). En biologie des systèmes, on crée des graphes de réaction. Les substrats et produits d'une réaction sont représentés par des nœuds et les réactions par des arrêtes (cf. Figure 16)

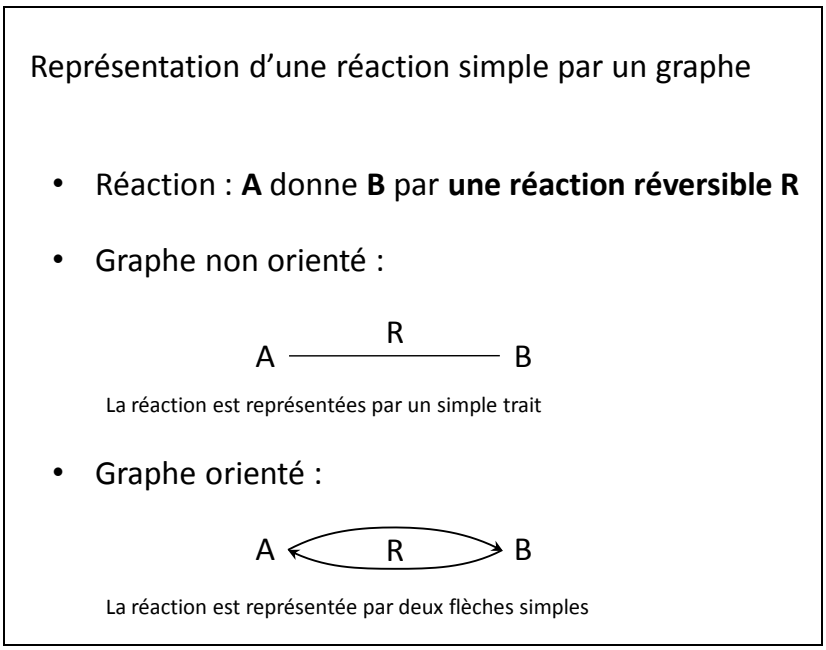
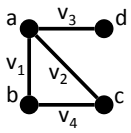


Figure 16 : Représentation d'une réaction simple par un graphe
Les composés A et B de la réaction (substrat et produit) sont représentés par des nœuds dans le graphe. Les réactions sont représentées par des flèches. Dans un graphe non orienté les arrêtes sont représentées soit par de simples traits soit par des double flèches, signifiant que l'on peut parcourir l'arrête dans les deux sens. Au contraire dans un graphe orienté, les arcs sont représentés par des flèches simples, on ne peut les parcourir que dans le sens indiqué par la flèche. Une réaction réversible pouvant se faire dans les deux sens, il est nécessaire de la représenter par deux flèches.

Un graphe G de sommets $x_1, \ldots, x_n$ peut être décrit par une matrice carrée M de dimension $n \times n$ (matrice ayant *n* lignes et *n* colonnes). Cette matrice appelée *matrice d'adjacence* du graphe cf. figure 17. Les éléments de M sont soit 0 soit 1. En notant a_{ij} l'élément situé au croisement de la ligne *i* et de la colonne *j*, M est définie par :

$a_{ij} = 1$ si et seulement si x_i et x_j sont voisins. Deux sommets étant voisins s'ils sont reliés par une arrête.



	a	b	c	d
a	0	1	1	1
b	1	0	1	0
c	1	1	0	0
d	1	0	0	0

Figure 17 : Matrice d'adjacence d'un graphe simple
À gauche un graphe simple à quatre sommets (a, b, c, d) et quatre arrêtes (v_1, v_2, v_3, v_4). À droite, la matrice d'adjacence de ce graphe. La matrice d'adjacence est diagonale, cela est figuré par la partie inférieure gauche de la matrice en grisé qui est exactement la même que la partie supérieure droite.

Il faut noter qu'une matrice d'adjacence peut contenir des éléments égaux à 1 sur sa diagonale si le graphe contient des boucles (arrête reliant un sommet à lui-même) et que c'est une matrice

diagonale. Dans un graphe dont les arrêtes sont pondérées (une valeur numérique est appliquée à l'arrête), on peut remplacer les 1 de la matrice par la valeur du poids de l'arrête.

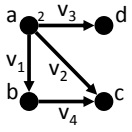
En biologie des systèmes, on utilise une structure proche de la matrice d'adjacence pour décrire les graphes de réaction : la matrice de stœchiométrie notée souvent S. Les métabolites sont représentés en lignes et les réactions en colonnes. Cette matrice indique quels sont les métabolites impliqués dans une réaction et s'ils sont substrats ou produits de cette réaction. La matrice est définie par :

$s_{ij} = 0$ si le métabolite i n'est pas impliqué dans la réaction j

$s_{ij} = - c_{ij}$ si le métabolite i est substrat de la réaction j avec un coefficient de stœchiométrie c_{ij}

$s_{ij} = c_{ij}$ si le métabolite i est produit de la réaction j avec un coefficient de stœchiométrie c_{ij}

V_1 : a → b
V_2 : a → c
V_3 : 2a → d
V_4 : b → c



	V_1	V_2	V_3	V_4
a	-1	-1	-2	0
b	1	0	0	-1
c	0	1	0	1
d	0	0	1	0

Figure 18 : Matrice de stœchiométrie d'un système simple
À gauche : un système de quatre équations simples. Au milieu : la représentation graphique de ce système. À droite : la matrice de stœchiométrie du système.

Si l'on considère le vecteur V contenant les flux des différentes réactions et le vecteur X contenant les concentrations des métabolites, il est possible de noter l'évolution des concentrations des métabolites en fonction des flux (cf. figure 19)

$$\frac{dx}{dt} = S\,V \quad \text{Avec :} \quad \frac{dx}{dt} = \left(\frac{d[a]}{dt}, \frac{d[b]}{dt}, \frac{d[c]}{dt}, \frac{d[d]}{dt}\right)$$

$$V = (v_1, v_2, v_3, v_4)$$

$$S = \begin{bmatrix} -1 & -1 & -2 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Alors :

$$\frac{d[a]}{dt} = -v_1 - v_2 - 2v_3$$

$$\frac{d[b]}{dt} = v_1 - v_4$$

$$\frac{d[c]}{dt} = v_2 + v_4$$

$$\frac{d[d]}{dt} = v_3$$

Figure 19 : Lien entre matrice de stœchiométrie, flux et concentrations.

L'inconvénient majeur de la matrice de stœchiométrie est qu'on ne peut pas faire la différence entre les réactions réversibles et les réactions irréversibles puisque les réactions ne sont décrites que dans un seul sens. Souvent pour définir une réaction réversible dans une telle matrice on la scinde en deux réactions irréversibles. Des logiciels comme Metatool [94] résolvent le problème en classant les réactions en deux groupes : irréversibles et réversibles et ils ne dupliquent pas les réactions réversibles.

La représentation graphique de réactions ayant plusieurs substrats et/ou plusieurs produits est beaucoup plus complexe que la représentation graphique de réactions simples. Il est nécessaire de représenter la réaction par un nœud. Une voie métabolique, qui peut être considérée comme une suite ordonnée de réactions biochimiques, est alors représentée par un graphe comprenant des nœuds représentant les composés biochimiques reliés à des nœuds représentant les réactions. Il n'existe pas de liens entre deux composés ou deux réactions. Ce type de graphe est appelé « graphe biparti ».

Définition : Un graphe biparti est un graphe $G(V_1,V_2,E)$ constitué de deux ensembles indépendants de nœuds V_1 et V_2 et d'un ensemble d'arrêtes E reliant deux nœuds de chacun des deux ensembles. Toutes les arrêtes « *e* » sont telles-que : e(x,y) avec ($x \in V_1$) et ($y \in V_2$) [111].

On pourrait représenter ces réactions complexes par un autre type de graphe : les « hypergraphes ». Les hypergraphes sont des graphes particuliers où une arrête ne relie pas deux sommets mais un nombre quelconque de sommets [133] (cf. figure 20).

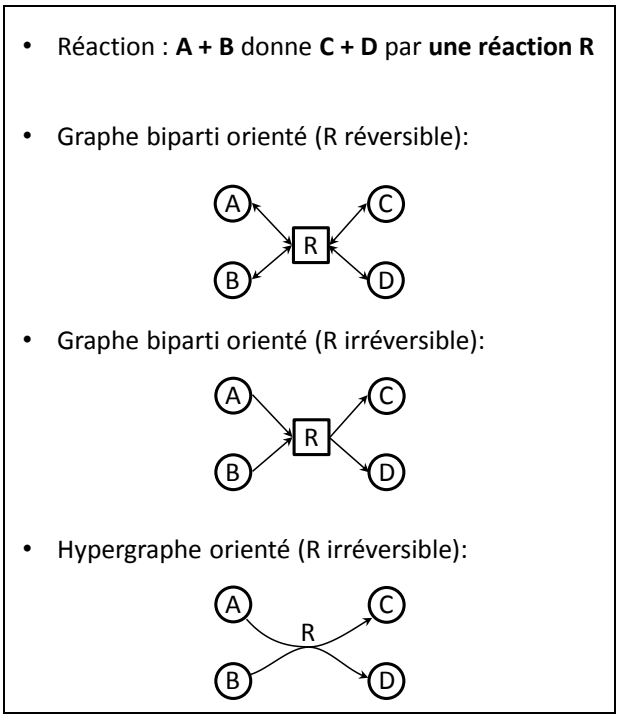


Figure 20 : Représentation d'une réaction biochimique ayant plusieurs substrats et produits par un graphe biparti ou un hypergraphe
La représentation de cette réaction est impossible par un graphe simple. On peut la représenter par un graphe biparti dont les deux ensembles de nœuds sont : les composés de la réaction d'une part (A, B, C et D) et le nœud représentant la réaction d'autre part. Dans la représentation sous forme d'hypergraphe, le nœud de la réaction disparait. Les substrats sont reliés aux produits par une unique arrête ayant deux nœuds en entrée et deux en sortie.

La représentation des réseaux biologiques par des graphes permet de faire une analyse de la structure du réseau. On parle d'analyse statique par opposition à une analyse dynamique du modèle

par la résolution de systèmes d'équations différentielles par exemple. En utilisant un graphe, on peut déterminer différents paramètres du réseau comme la connectivité des métabolites ou la robustesse d'un modèle face à une perturbation (suppression d'une réaction par exemple).

La représentation sous forme de graphe biparti est plus commode pour les graphes biologiques que la représentation sous forme d'hypergraphe. En effet, dans la représentation sous forme de graphe biparti, on dispose d'un nœud « réaction » sur lequel différents éléments biologiques pourront être appliqués. En effet, dans une cellule une réaction se fait presque exclusivement en présence d'une enzyme qui n'a pas encore été représentée. De plus une voie métabolique ne peut pas se limiter à une succession de réactions. Afin de pouvoir représenter ces différents processus, une convention de représentation a été développée : le SBGN (*Systems Biology Graphical Notation*).

2.3.3.2 SBGN (Systems biology Graphical Notation)

Le SBGN est un « langage visuel »[84] développé par un ensemble de biochimistes spécialistes de la modélisation des réseaux et par des informaticiens. Il existe sur internet de nombreuses bases de données de processus biologiques, chacune disposant d'un système de représentation propre [149,151,152,155,157,158,170,172] ou aucun système de représentation [106,171]. De plus de nombreuses représentations sont ambiguës.

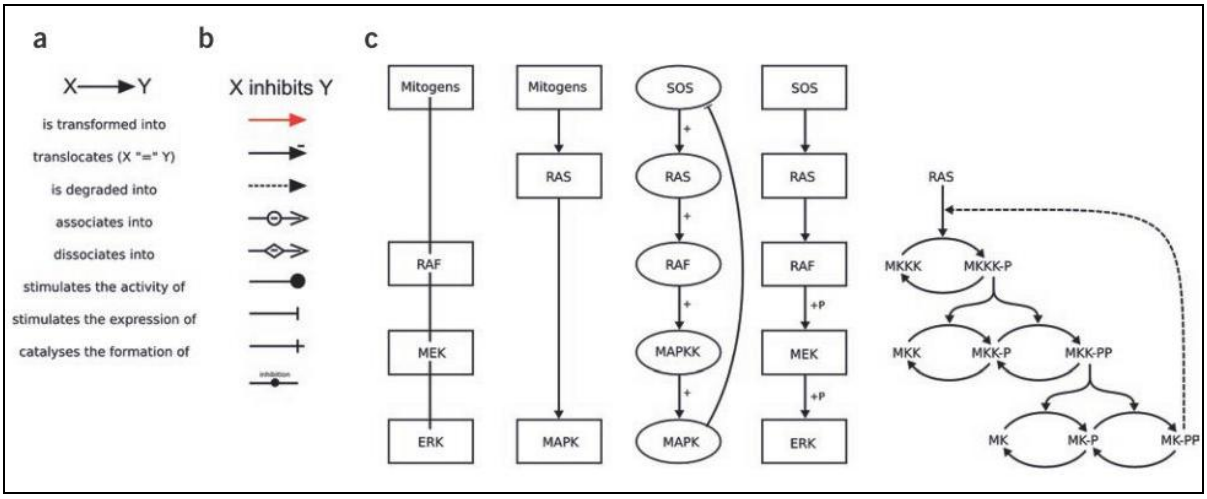


Figure 21 : Incohérence et ambiguïté des représentations actuelles non standardisées de réseaux biologiques
Figure tirée de l'article [84] (a) huit significations différentes associées au même symbole dans une représentation graphique du rôle de la cycline dans la régulation du cycle cellulaire [143]. (b) Neuf symboles différents dans la littérature pour représenter un même processus biologique : l'inhibition de la transcription. (c) Cinq représentations différentes de la cascade des MAP kinases dans la littérature scientifique, illustrant différents niveaux de connaissances biologiques et biochimiques. De gauche à droite: relations entre les gènes [173], influences globales des gènes du système [21], activations et inhibitions des gènes entre eux [22], processus biochimiques (+p : phosphorylation)[1], réactions biochimiques [24]. Dans le dernier schéma, un même type de flèche représente la catalyse et la production.

Afin de faciliter les interactions entre biologistes, il est nécessaire de mettre au point un système unifié de représentation des processus biologiques. Ce système doit être capable de représenter tous les processus biologiques connus sans ambiguïté et de manière la plus claire possible. C'est dans cette optique que le SBGN a été développé. Le SBGN est basé sur l'utilisation de « glyphes ». Ces glyphes sont les différentes représentations possibles d'un élément du graphe (arrêtes ou nœuds). Chaque

glyphe est un symbole appliqué à un élément du graphe qui fournit des informations sur le rôle de cet élément (cf. Figure 23).

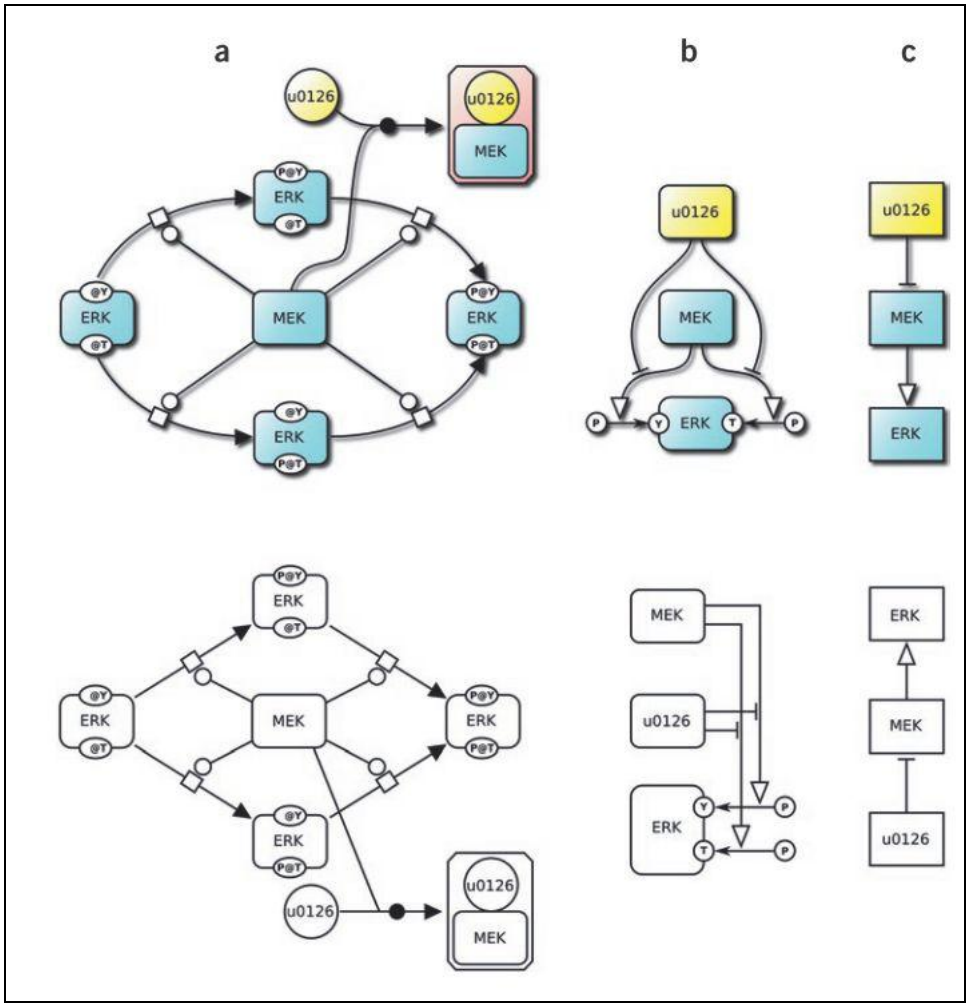


Figure 22 : Exemple des différentes représentations SBGN d'un même processus biologique : la phosphorylation d'une protéine catalysée par une enzyme et modulée par un inhibiteur.
Les représentations situées en bas de figure sont totalement équivalentes à celles situées en haut. En effet pour le SBGN, les couleurs des nœuds, l'épaisseur des liens ou la position des nœuds n'ont aucune importance. Seuls les glyphes sont pris en compte.

Le SBGN est constitué de trois niveaux représentés par trois types de diagrammes de plus en plus simples (cf. Figure 22): le diagramme de description de processus [80], le diagramme entité relation [85] et le diagramme de flux [79].

Le diagramme de description de processus

Un diagramme de description de processus (*process diagram*) est un diagramme représentant les processus moléculaires et les interactions ayant lieu entre des composés biochimiques. Ce diagramme décrit comment un composé se transforme en un autre composé. C'est ce type de diagramme, le plus décrit ici, qui est employé pour représenter les voies métaboliques. Ce type de diagramme permet de représenter les différents états d'un même composé (protéine phosphorylée vs. non phosphorylée par

exemple). Pour dessiner un diagramme de description de processus, six types principaux de glyphes sont disponibles : les glyphes servant à représenter les entités (composés biochimiques), les processus (réactions, associations, dissociations, …), les conteneurs (compartiments biologiques), les nœuds de référence (liens vers un sous réseau, annotations, …), les liens (consommation, production, activation, inhibition, …) et les opérateurs logiques (non utilisés ici, ils permettent d'insérer des conditions dans les graphes à la façon des algorigrammes).

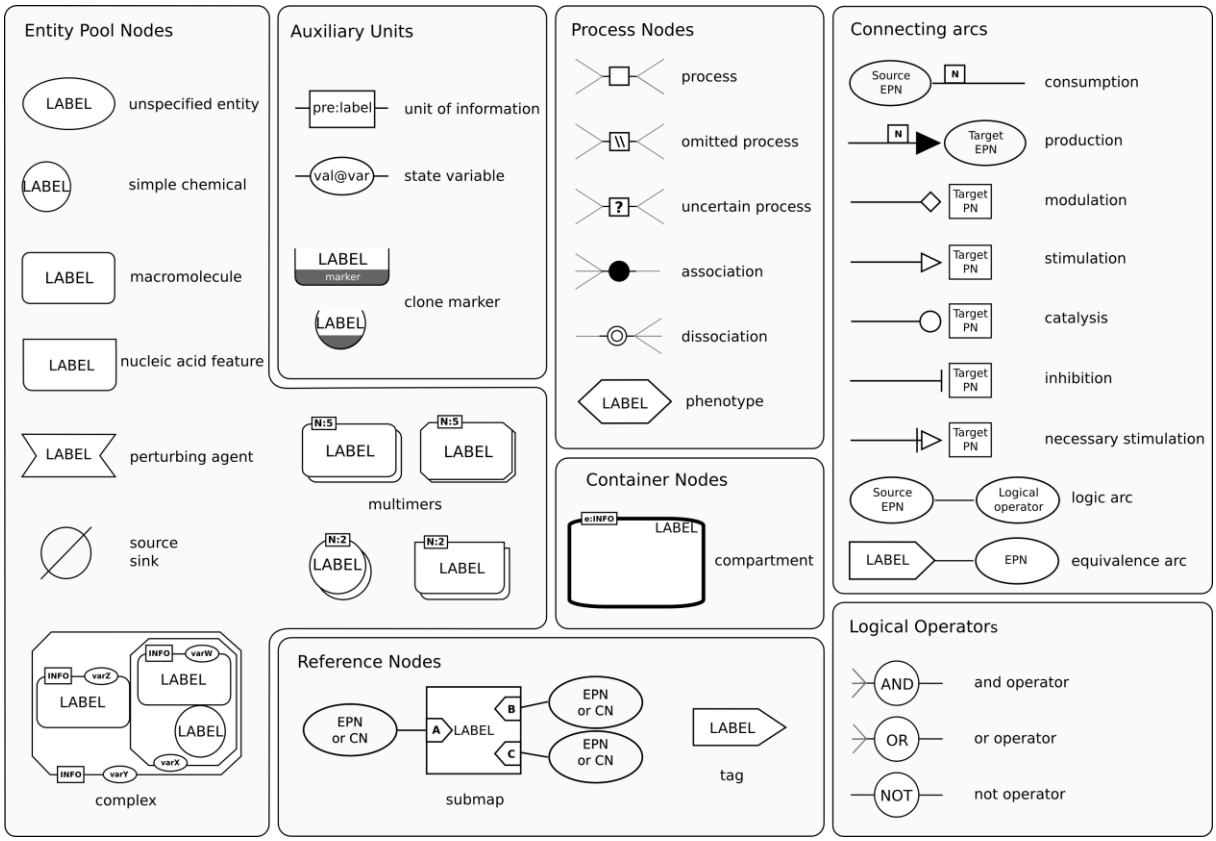


Figure 23 : Les différents glyphes utilisés dans le diagramme de description de processus

Dans la Figure 22 partie *a* décrivant la phosphorylation d'une protéine catalysée par une enzyme et modulée par un inhibiteur et représentée par un diagramme de description de processus, l'enzyme MEK catalyse quatre réactions différentes de phosphorylation de la protéine ERK sur la tyrosine (P@Y : phosphorylation sur la tyrosine notée Y dans l'alphabet IUPAC) et la thréonine (P@T : phosphorylation de la thréonine notée T). La formation d'un complexe MEK/u0126 est également observée. L'inhibition de MEK par u0126 est implicite dans la séquestration de MEK lors de la formation du complexe avec u0126 mais elle n'est jamais explicitement notée. Remarque : u0126 est un inhibiteur des kinases MEK1 et MEK2 qui inhibe la croissance des cellules cancéreuses [31] [36].

Le diagramme entité relation

Le diagramme d'entité relation ne met pas l'accent sur les processus comme le diagramme précédent mais sur les entités (les nœuds dans le graphe) et leurs relations. Chaque entité n'est représentée qu'une seule fois, le diagramme est donc souvent plus simple que le diagramme de description de processus. Ce diagramme sert à visualiser plus rapidement les différents effecteurs

susceptibles d'agir sur un composé donné. Les glyphes de processus sont absents de ce type de diagramme (cf. Figure 24). Les glyphes nécessaires sont répartis en trois groupes : les entités qui représentent aussi bien les composés biologiques que les états biologiques de ces composés ou les opérateurs logiques, les influences (stimulations, inhibition, …) et les déclarations (interactions, assignation qui va permettre de représenter des processus comme la phosphorylation, …). Les déclarations et les influences étant des relations entre les nœuds.

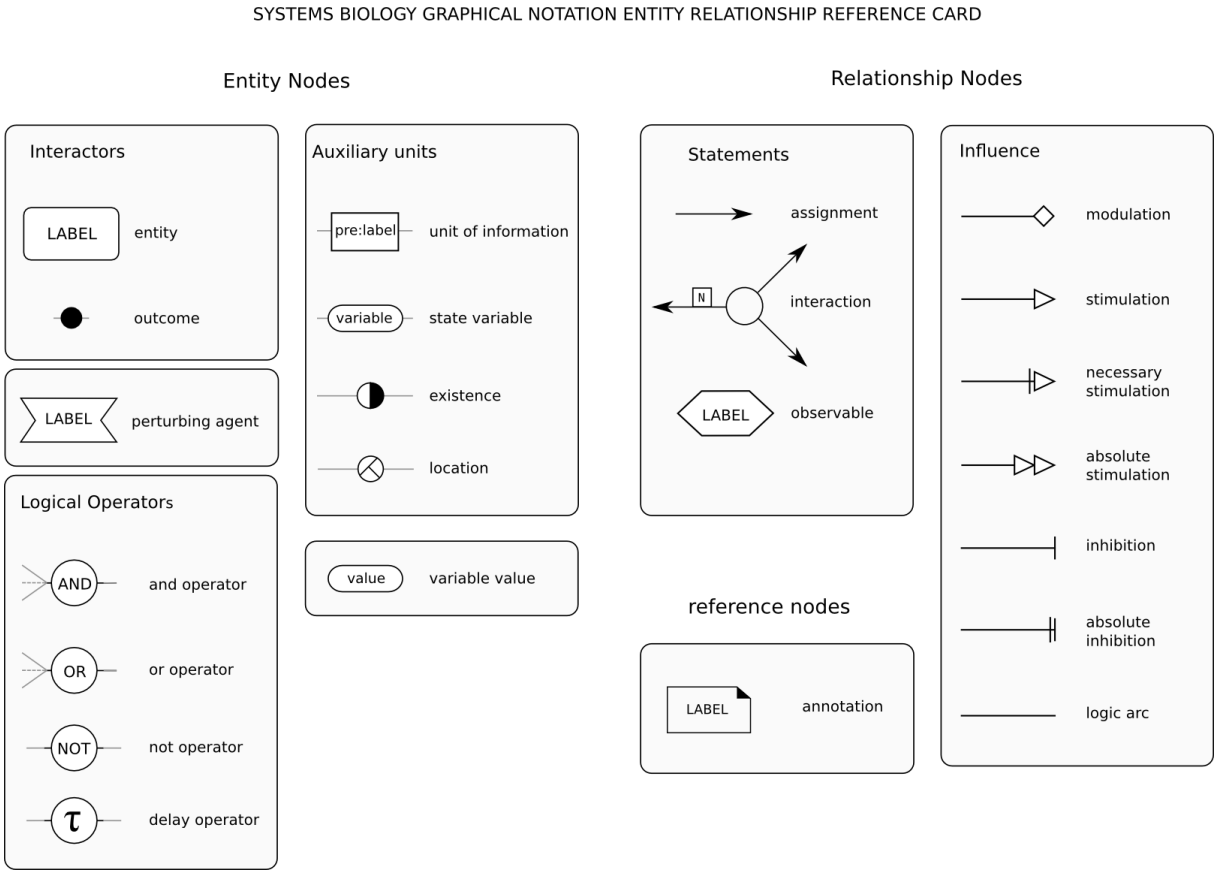


Figure 24 : Les différents glyphes utilisés dans le diagramme entité relation

Dans la Figure 22 partie *b*, il n'y a plus de différence de type entre la petite molécule u0126 et les protéines. Les phosphorylations sur la tyrosine et la thréonine d'ERK sont représentées cette fois par des assignations de phosphore (P) sur les acides aminés concernés. La catalyse par MEK est représentée par une stimulation. L'inhibition de u0126 sur MEK est explicite mais la formation du complexe u0126 / MEK est absente du diagramme.

Le diagramme de flux

Ce type de diagramme est utilisé pour avoir une vision globale d'un réseau. Il ne représente que les influences entre entités. Les entités ne sont présentes qu'une seule fois et elles sont directement connectées par des arcs modulateurs. Les différents états ne sont pas non plus représentés.

SYSTEMS BIOLOGY GRAPHICAL NOTATION ACTIVITY FLOW DIAGRAM REFERENCE CARD

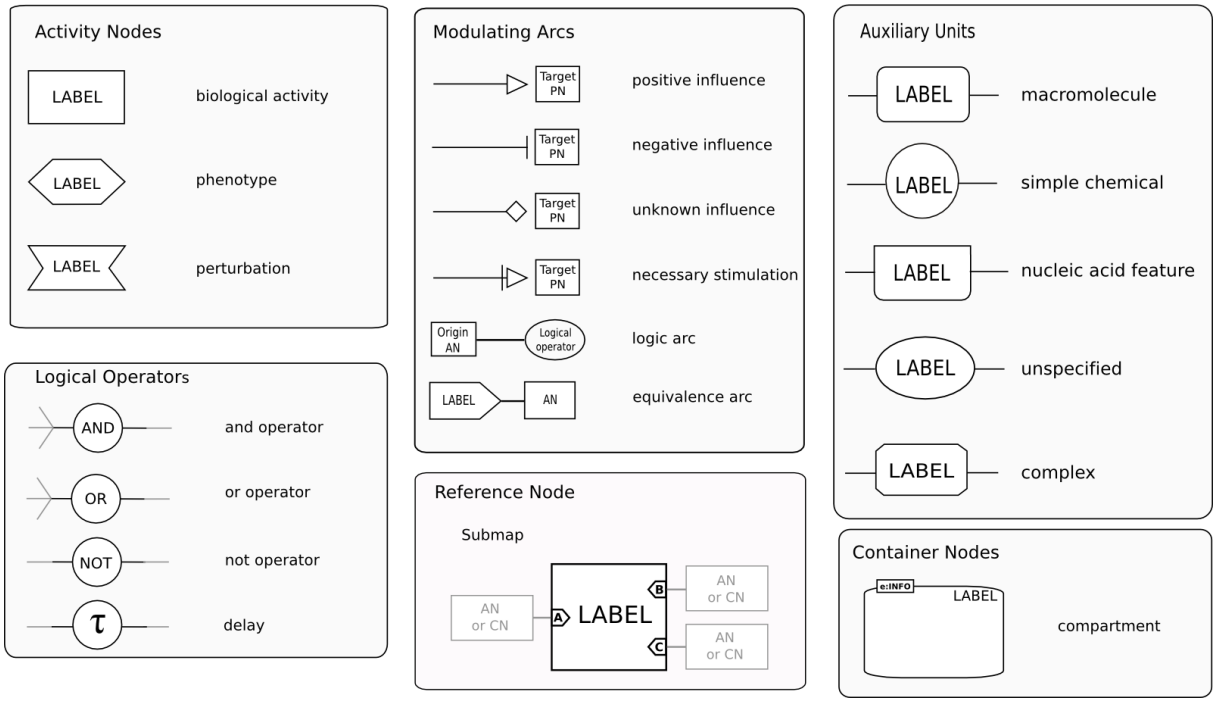


Figure 25 : Les différents glyphes utilisés dans le diagramme de flux

Dans la partie *c* de la Figure 22 la phosphorylation et la formation du complexe restent absentes. Seule l'inhibition d'u0126 sur MEK qui active ERK est représentée.

Développement et évolution du SBGN

De plus en plus de bases de données de voies métaboliques utilisent ce standard : MetaCrop [156], Reactome [164], Biomodels [151], Panther Patways [166]. En parallèle, de plus en plus de logiciels supportant le SBGN sont développés. En Octobre 2010 les développeurs du SBGN en comptaient déjà 19 [128], parmi eux : Cytoscape [118] via le plugin BiNoM [131], VANTED [36] de manière native ou via un plugin SBGN-ed [23] ou BioCham [12]. Une liste plus complète et régulièrement mise à jour peut être consultée sur le site internet du SBGN [168].

Depuis Octobre 2010 une librairie informatique est développée : LibSBGN [47]. Cette librairie a vu le jour afin d'homogénéiser la manière dont les programmes enregistrent les graphes au format SBGN en définissant un standard d'écriture : le SBGN-ML. Cette librairie permet de lire et d'écrire des fichiers SBGN-ML. Le SBGN-ML est une implémentation XML du SBGN. Pour l'instant seule l'écriture des diagrammes de description de processus est possible en SBGN-ML [160]. Un exemple simple de fichier SBGN-ML est disponible dans la présentation [128]. La libraire est disponible en Java et en C++. En Novembre 2010, seuls deux éditeurs de graphes biologiques supportaient le format SBGN-ML : VANTED via SBGN-ed [23] et PathVisio [48].

Le SBGN et le SBGN-ML permettent respectivement de représenter et d'enregistrer les graphes biologiques mais ils ne permettent pas d'enregistrer des modèles biologiques comportant des données mathématiques. En effet, dans un modèle biologique il est souvent nécessaire d'associer à une réaction donnée une loi biochimique et ses paramètres. Pour enregistrer et échanger de telles informations différents formats ont été développés dont le SBML (*Systems Biology Markup Language*).

2.3.3.3 SBML (Systems Biology Markup Language)

Définitions

Le SBML est né de la nécessité de produire un format standard d'échange de fichiers décrivant des modèles biologiques complexes. En biologie des systèmes, un modèle comprend non seulement des informations de graphe correspondant aux réseaux de réactions mais aussi des informations mathématiques issues d'expériences biologiques comme les fonctions des lois biochimiques, les valeurs de concentrations, etc. La première version du SBML a été publiée en 2003 [43] mais il continue d'évoluer puisque la dernière version du langage date de janvier 2010 [45]. Le SBML est basé sur le langage XML. Le SBML décrit un ensemble de concepts nécessaires à la description d'un modèle biologique et doit permettre de recalculer les résultats issus des simulations du modèle (on n'enregistre pas des résultats de simulations mais bien les équations mathématiques permettant de reproduire les simulations). Les différents éléments d'un fichier SBML doivent être décrits dans un ordre précis. Les éléments décrits ici sont définis à partir du niveau 2 du langage [162] (cf. figure 26).

1. **La définition de fonction** (*FunctionDefinition*) : définition une fonction pouvant être appliquée à une réaction. Une fonction doit être écrite en respectant le format MathML (écriture de formules mathématiques en XML).
2. **La définition d'unité** (UnitDefinition) : définition des unités employées dans le modèle en utilisant un ensemble d'unités de base prédéfinies en SBML. Cf. Figure 28.
3. **Le type de compartiment** (*CompartmentType*) : il peut être utile de décrire des types de compartiments permettant de construire un modèle avec plusieurs compartiments d'un même type (pour décrire par exemple un modèle comportant plusieurs mitochondries)
4. **Le compartiment** (*Compartment*) : c'est un conteneur d'un volume défini dont la composition chimique est connue et où les réactions biochimiques vont avoir lieu. Plusieurs compartiments différents peuvent être définis.
5. **L'espèce** (*Species*) sous-entendu espèces biochimique : c'est une substance chimique ou toute entité qui prendra part à une réaction biochimique : petites molécules, sucres, enzymes, …
6. **Le paramètre** (Parameter) : constante numérique. Un paramètre peut être propre à une réaction ou au modèle complet.
7. **La valeur initiale** (*InitialAssignment*) : sert par exemple à fixer la valeur de la concentration initiale d'un métabolite. Cette valeur initiale peut être calculée par une expression mathématique.
8. **La règle** (*Rule*) : une règle est une expression mathématique affectant les variables du modèle. Une règle ne s'applique pas obligatoirement à une réaction. Les règles peuvent être utilisées pour définir des comportements complexes (fonctions mathématiques différentes dépendant des concentrations des métabolites d'une réaction).
9. **La contrainte** (*Constraint*) : il peut être intéressant d'appliquer des contraintes à un modèle comme par exemple des limites de concentrations viables pour un métabolite donné.
10. **La réaction** (*Reaction*) : décrit la transformation, ou tout autre processus biochimique affectant une ou plusieurs espèces chimiques du modèle. Une réaction comprend notamment une liste de réactifs compris dans la liste des espèces, une liste de produits eux aussi définis dans la liste des espèces et une loi cinétique.
11. **Les évènements** (*Event*) : déclenche une action sur un évènement donné. Ce type de champ ne sera pas utilisé dans notre cas.

    Chacun de ces éléments peut être définit plusieurs fois, ils sont alors inclus dans des listes.

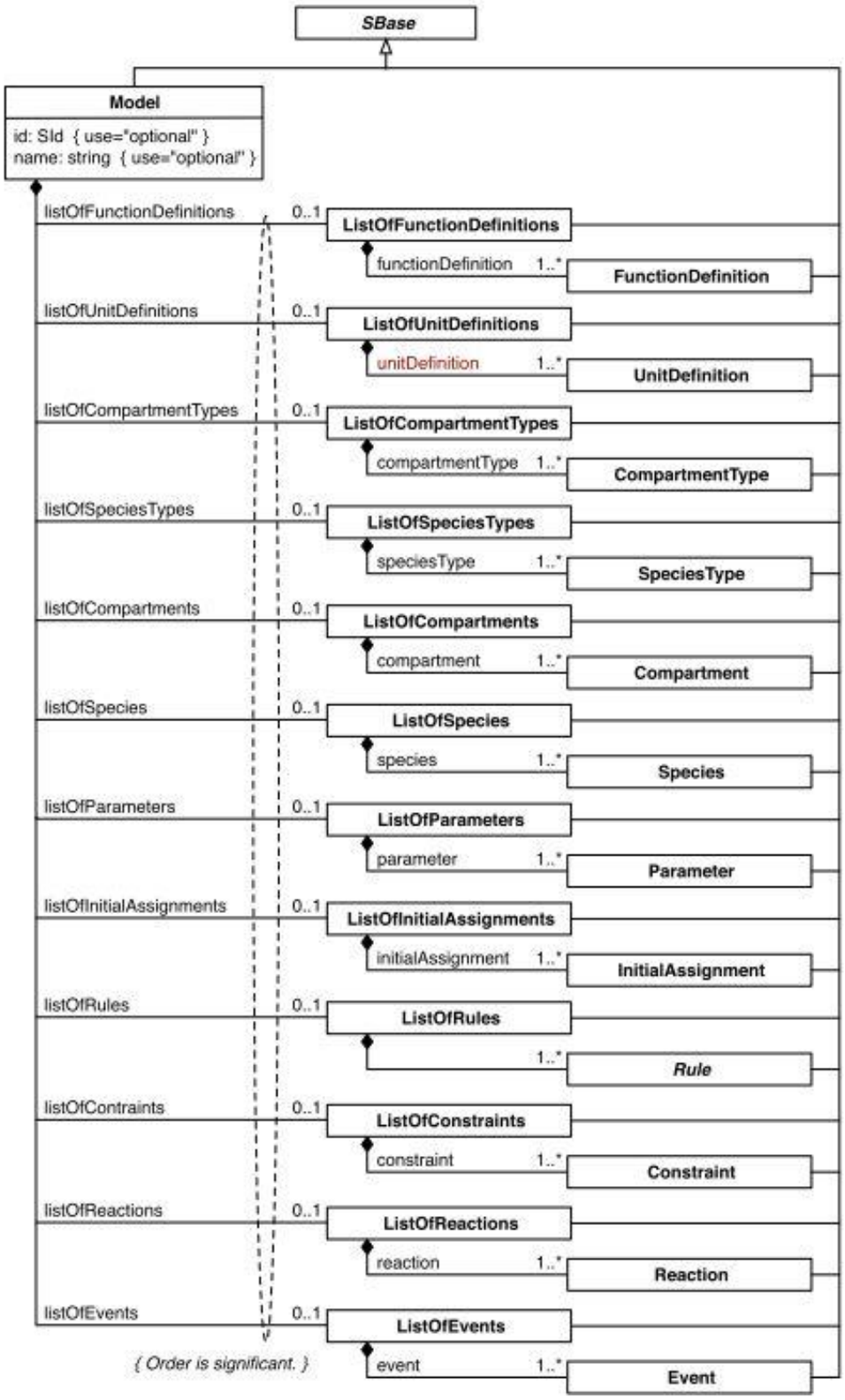


Figure 26 : Schéma global de la structure d'un fichier SBML.
Figure tirée de l'article [162][2].Un modèle biologique écrit au format SBML peut être décrits par 12 types d'éléments différents définis dans des listes : les fonctions, les unités, les types de compartiments, les types d'espèces chimiques, les compartiments, les espèces biochimiques, les paramètres, les valeurs initiales, les règles, les contraintes, les réactions et les évènements.

Exemple simple

Modèle biologique simple :

$$E + S \underset{k_{\text{off}}}{\overset{k_{\text{on}}}{\rightleftharpoons}} ES \xrightarrow{k_{\text{cat}}} E + P$$

Figure 27 : Modèle à écrire en SBML : Équation d'une réaction enzymatique suivant le modèle de Michaëlis-Menten.
Tiré de l'article [162]. E *représente l'enzyme,* S *le substrat,* P *le produit. Les paramètres* k_{on} *et* k_{off} *sont respectivement les constantes de vitesse d'association et de dissociation du complexe Enzyme/Substrat. Le paramètre kcat est la vitesse de production du produit à partir du complexe ES.*

Les vitesses de réaction sont définies par :

$$v_{eq} = V_{comp} \times \left(k_{on} \times [E] \times [S] - k_{off} \times [ES]\right) \quad (1)$$
$$v_{cat} = V_{comp} \times k_{cat} \times [ES] \quad (2)$$

Équation 1 : Vitesses de réaction du système
(1*) Vitesse de formation du complexe enzyme/substrat ES. (2) Vitesse de production du produit P*

Définition des unités nécessaires au modèle :

```
1   <listOfUnitDefinitions>
2       <unitDefinition id="per_second">
3           <listOfUnits>
4               <unit kind="second" exponent="-1" scale="0" multiplier="1"/>
5           </listOfUnits>
6       </unitDefinition>
7       <unitDefinition id="litre_per_mole_second">
8           <listOfUnits>
9               <unit kind="mole" exponent="-1" scale="0" multiplier="1"/>
10              <unit kind="litre" exponent="1" scale="0" multiplier="1"/>
11              <unit kind="second" exponent="-1" scale="0" multiplier="1"/>
12          </listOfUnits>
13      </unitDefinition>
14  </listOfUnitDefinitions>
```

Figure 28 : Définition de nouvelles unités en SBML

Deux unités sont décrites dans la figure 28: « par seconde » notée *per_second* et « litre par mole par seconde » notée *litre_per_mole_second*.

Ensuite il faut définir les compartiments et les espèces du modèle :

```
1   <listOfCompartments>
2       <compartment id="comp" size="1e-14" spatialDimensions="3" units="litre" constant="true"/>
3   </listOfCompartments>
4   <listOfSpecies>
5       <species compartment="comp" id="E" initialAmount="5e-21" substanceUnits="mole" constant="false"/>
6       <species compartment="comp" id="S" initialAmount="1e-20" substanceUnits="mole" constant="false"/>
7       <species compartment="comp" id="P" initialAmount="0" substanceUnits="mole" constant="false"/>
8       <species compartment="comp" id="ES" initialAmount="0" substanceUnits="mole" constant="false"/>
9   </listOfSpecies>
```

Figure 29 : Définition d'un compartiment et de quatre espèces biochimiques en SBML

Le compartiment « comp » en trois dimensions ayant un volume arbitraire constant de 10^{-14} litres (ligne 2). Les espèces sont définies aux lignes 5 à 8. Elles correspondent aux quatre composés E, S, P et ES nécessaires à la description du modèle dont les quantités pourront varier. La quantité initiale de E est de 5.10^{-21} et celle de S de 10^{-20} moles. Ces espèces sont définies dans le compartiment *comp* dont le volume est de 10^{-14} litres, les concentrations de E et S sont donc respectivement de 5.10^{-7} et 10^{-6} M.

Les réactions doivent ensuite être décrites. Tout d'abord la réaction de formation du complexe enzyme / substrat :

```
1   <reaction id="veq" reversible="true" fast="false">
2       <listOfReactants>
3           <speciesReference species="E" stoichiometry="1" />
4           <speciesReference species="S" stoichiometry="1" >
5       </listOfReactants>
6       <listOfProducts>
7           <speciesReference species="ES" stoichiometry="1" />
8       </listOfProducts>
9       <kineticLaw>
10          <math xmlns="http://www.w3.org/1998/Math/MathML">
11              <apply>
12                  <times/>
13                  <ci>comp</ci>
14                  <apply>
15                      <minus/>
16                      <apply>
17                          <times/>
18                          <ci>kon</ci>
19                          <ci>E</ci>
20                          <ci>S</ci>
21                      </apply>
22                      <apply>
23                          <times/>
24                          <ci>koff</ci>
25                          <ci>ES</ci>
26                      </apply>
27                  </apply>
28              </apply>
29          </math>
30          <listOfLocalParameters>
31              <localParameter id="kon" value="1000000" units="litre_per_mole_second"/>
32              <localParameter id="koff" value="0.2" units="per_second"/>
33          </listOfLocalParameters>
34      </kineticLaw>
35  </reaction>
```

Figure 30 : Description en MathML de la formation du complexe enzyme substrat dans le modèle de Michaelis-Menten

Les lignes 2 à 5 correspondent à la déclaration de la liste des substrats de la réaction : E et S avec des coefficients de stœchiométrie égaux à 1. De même, les lignes 6 à 8 correspondent à la déclaration de l'unique produit de la réaction : ES. La suite correspond à la définition de la loi cinétique de la réaction à l'équilibre. La loi proprement dite est définie en MathML de la ligne 10 à 29 (entre les balises *math*) et correspond à la formule (1) de l'équation 1. Les paramètres de l'équation k_{on} et k_{off}. Sont décrits de la ligne 30 à 33. Les paramètres *E*, *S*, *ES* et *comp* du système ne sont pas explicitement déclarés, ils correspondent respectivement aux concentrations de *E*, *S*, *ES* et au volume de *comp*.

De la même manière on décrit la deuxième réaction :

```
1   <reaction id="vcat" reversible="false" fast="false">
2       <listOfReactants>
3           <speciesReference species="ES" stoichiometry="1" constant="true"/>
4       </listOfReactants>
5       <listOfProducts>
6           <speciesReference species="E" stoichiometry="1" constant="true"/>
7           <speciesReference species="P" stoichiometry="1" constant="true"/>
8       </listOfProducts>
9       <kineticLaw>
10          <math xmlns="http://www.w3.org/1998/Math/MathML">
11              <apply>
12                  <times/>
13                  <ci>comp</ci>
14                  <ci>kcat</ci>
15                  <ci>ES</ci>
16              </apply>
17          </math>
18          <listOfLocalParameters>
19              <localParameter id="kcat" value="0.1" units="per_second"/>
20          </listOfLocalParameters>
21      </kineticLaw>
22  </reaction>
```

Figure 31 : Description en MathML de la réaction de formation du produit dans le modèle de Michaelis-Menten

Les lignes 2 à 8 correspondent à la déclaration des substrats (*ES*) et des produits (*E* et *P*). La loi cinétique décrite de la ligne 10 à 17 est celle de la formule (2) de l'équation 1. Enfin le paramètre k_{cat} est déclaré ligne 19.

Les éléments SBML présentés ici sont suffisants mais nécessaires à la description du modèle de la figure 27. Le document SBML présenté ici est décrit en respectant le niveau 2, version 4 du SBML mais le SBML est toujours en développement.

Les évolutions actuelles du SBML

Les révisions majeures du SBML sont appelées niveaux. Ces niveaux sont découpés en versions. La dernière version du langage est le niveau 3 - version 1 publiée en Octobre 2010 [45]. L'évolution majeure apportée par le niveau 3 est de rendre le SBML modulaire en utilisant un système de *packages* pouvant ajouter des fonctionnalités au langage SBML. Ces packages sont développés de manière indépendante du SBML proprement dit. Le cœur du langage SBML au niveau 3 reste très proche du niveau 2. Le système des packages a vraisemblablement été introduit afin de permettre à des laboratoires indépendants de développer leur propres extensions au SBML. De manière étonnante, le langage SBML qui se veut être une convention va utiliser de nouveaux éléments via l'utilisation des packages qui ne feront pas partie du SBML et qui ne seront potentiellement pas disponibles pour tous.

Alternatives au SBML

Une des critiques majeures faite au SBML est qu'il est impossible d'imbriquer les modèles entre eux (composition de modèles). Par exemple il est impossible d'inclure dans un modèle décrivant la chaîne respiratoire de la mitochondrie un modèle décrivant l'un des complexes de la chaîne.

Cette absence de composition de modèles est une des raisons qui a mené à la définition d'un autre standard : le CellML.

Le CellML

Le CellML [69] est un langage basé, comme le SBML, sur le langage XML. Il permet lui aussi de décrire des modèles biologiques. Les équations mathématiques sont aussi écrites en MathML et on retrouve le même système de gestion des unités qu'en SBML. Le SBML est très orienté vers la description de suites de réactions biochimiques alors que le CellML permet de représenter des modèles plus variés. Contrairement au SBML, le CellML permet de faire de la composition de modèles. En cela, il est beaucoup plus orienté vers la construction de modèles très complexes par la composition de modèles plus simples (cf. figure 32).

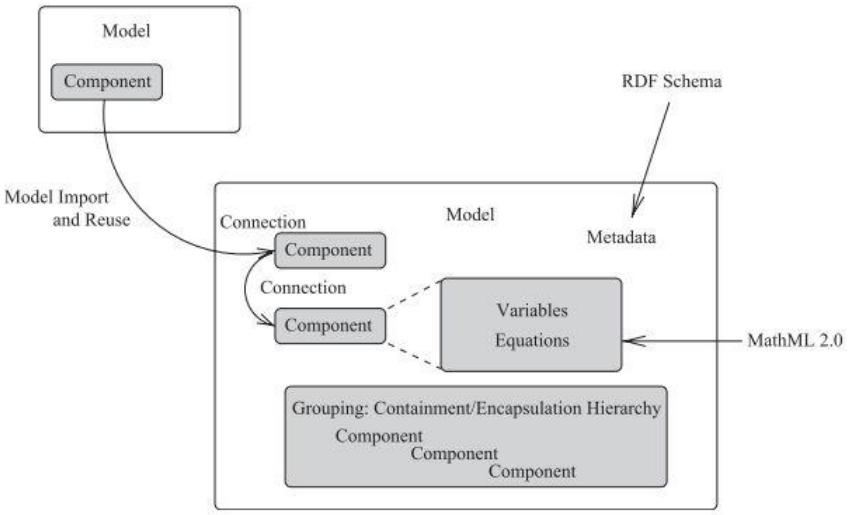


Figure 32 : Schéma général de la structure d'un modèle CellML
Figure tirée de l'article [69]. Un modèle CellML est constitué de composantes *(*components*) comprenant des variables utilisées dans des équations écrites en MathML. Ces variables représentent aussi bien les entités physiques du modèle que des paramètres mathématiques. Les composantes sont reliées entre elles par des* connections *permettant notamment d'utiliser les variables d'une composante dans une autre. Il est possible d'inclure des composantes d'un autre modèle. Les composantes peuvent être regroupées par encapsulation ou inclusion. L'encapsulation permet d'établir une hiérarchie entre les composantes et l'inclusion d'indiquer qu'une composante est incluse dans une autre (par exemple pour définir des réactions dans différents compartiments cellulaires). Les métadonnées d'un modèle sont décrites en respectant le standard RDF.*

Un modèle est constitué par des *composantes* (*components*) reliées entre elles par des *connections*. Une composante peut comprendre des variables, des équations mathématiques et des réactions. Les variables correspondent non seulement aux variables mathématiques du modèle comme k_{on} et k_{off} de la formule (1) de l'équation 1 mais aussi aux concentrations des métabolites.

Remarque : les métabolites ne sont jamais définis autrement que par l'utilisation des variables.

Les équations mathématiques sont, comme en SBML, écrites au format MathML. Les réactions sont utilisées pour la description de modèles représentant des voies métaboliques. Dans une réaction, peuvent être définis des réactifs, des produits, des catalyseurs et des inhibiteurs qui doivent être déclarés dans les variables de la composante. On peut aussi définir la stœchiométrie, le sens et la réversibilité de la réaction. On peut aussi ajouter une formule mathématique directement à une réaction mais cette pratique est découragée car elle est redondante avec la définition des équations mathématiques de la composante. La possibilité de déclarer des réactions permet de construire des modèles qualitatifs. Le système des variables ne décrivant pas obligatoirement des entités physiques permet de décrire une très large gamme de modèles pouvant être très éloignés des voies métaboliques. On peut décrire par exemple des systèmes de régulation géniques [30] ou des modèles d'étude de courants ioniques à travers des canaux [96].

Bien que le CellML palie à l'absence de possibilité de composition de modèle du SBML, il reprend une autre de ses lacunes : l'absence d'informations graphiques. Un modèle décrit en CellML peut même être bien plus complexe à représenter que le même modèle en SBML. En effet, si un modèle CellML ne comprend pas de champs « réaction », il devient impossible de le représenter sous forme de graphe puisqu'on ne peut pas faire la différence entre les entités physiques du modèle et les variables mathématiques. Les développeurs du CellML veulent même supprimer ce champ et conseillent de ne pas l'utiliser [147]. Bien que le SBML ne contienne aucune information graphique, le graphe sous-jacent au modèle décrit est possible à calculer grâce au système des réactions du SBML. Ce problème de représentation graphique des modèles intéresse particulièrement les biologistes puisque les modèles biologiques constituent des graphes très complexes à représenter [140] [139]. Dans certaines bases de données comme KEGG [158], les graphes représentant les voies métaboliques sont dessinés à la main. La représentation graphique des voies métaboliques est un problème commun à d'autres langages de description comme le BioPAX développé en parallèle du SBML et du CellML, pour décrire surtout des voies métaboliques.

BioPAX

Le BioPAX a été au départ créé pour décrire des modèles de voies métaboliques uniquement et a été étendu aux interactions protéine/protéine, aux réseaux de régulation géniques, etc.

La structure du BioPAX est très proche de la structure biologique d'une cellule. Le BioPAX définit une voie métabolique (*Pathway*) comme un ensemble d'*interactions* et/ou de voies métaboliques. Une voie métabolique définit un graphe. Une interaction est une relation entre au moins deux entités (*entities*). Une interaction peut décrire une interaction protéine-protéine, une réaction biochimique, … Les participants d'une interaction peuvent être des protéines, des complexes, de l'ADN, de l'ARN, des petites molécules, … Chacune de ces entités a une localisation précise dans la cellule. Pour chaque élément on peut associer des références vers des bases de données. Il est possible de retrouver des équivalences de beaucoup des champs du SBML en BioPAX (cf. figure 33).

Le SBML étant avant tout fait pour représenter un unique modèle (pouvant correspondre à une voie métabolique), la notion de voie métabolique dans un fichier SBML n'a aucun sens. Par contre on retrouve dans les deux formats la notion de réaction. En BioPAX, la réaction est plus finement décrite par la notion de « contrôle » (*control*) qui permet de préciser les modulateurs (inhibiteurs, activateurs), les enzymes, … Les notions « d'entité physique » (*PhysicalEntity*) en BioPAX et « d'espèce » (*species*) en SBML sont équivalentes. L'entité physique du BioPAX est un peu plus fine puisqu'il est possible de préciser si c'est un complexe, une protéine, … Enfin, on retrouve le même système de méta données en BioPAX (*openControlled Vocabulary*) et en SBML (*annotations*).

Le BioPAX ne permet pas de présenter les données mathématiques des modèles. Il ne permet que des représentations qualitatives des modèles. Par contre il peut être utilisé dans les modèles SBML et CellML comme métadonnées pour fournir des références croisées vers des bases de données biologiques comme on peut le faire avec les URI MIRIAM (cf. 2.3.2).

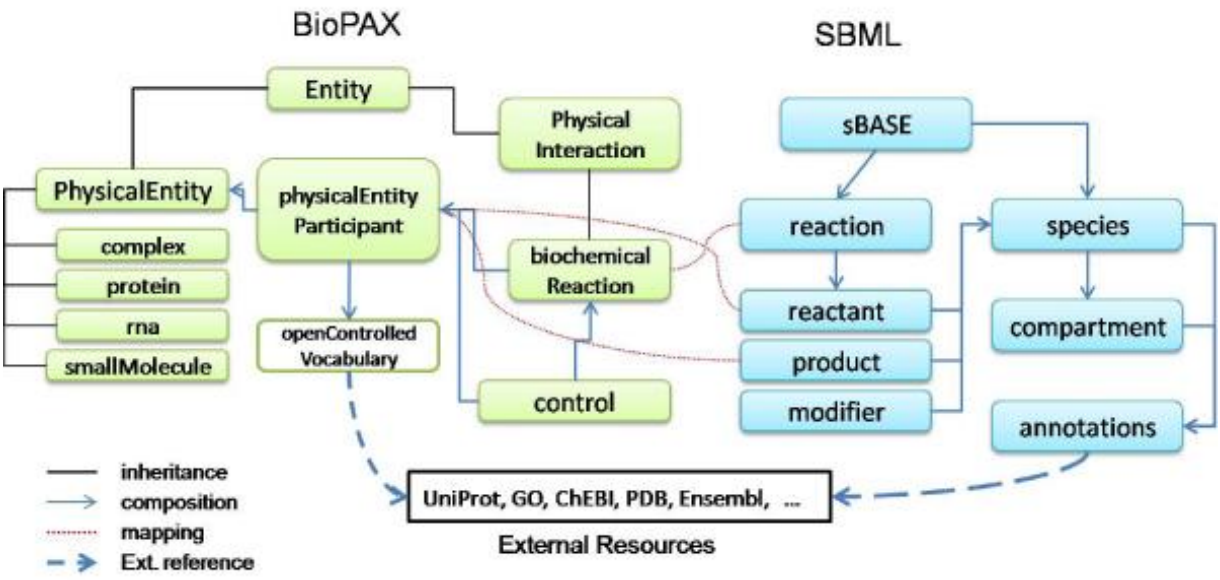


Figure 33 : Comparaison de la structure des fichiers BioPAX et SBML
Tiré de l'article [78]. À gauche, la structure globale d'un fichier BioPAX ; à droite celle d'un fichier SBML.

Le SBML reste tout de même le format le plus finalisé pour décrire des modèles biologiques ayant des données mathématiques associées. De plus, les développeurs du langage travaillent à la mise au point de la composition des modèles [119] et à l'ajout d'informations graphiques [38] [25]. Quel que soit le type de format utilisé pour décrire un modèle, étant donné la complexité de tels fichiers, il est impensable de construire ces modèles manuellement. Des programmes permettant de construire de manière graphique des modèles biologiques et de les enregistrer dans des formats standardisés sont obligatoires. Cependant, même en utilisant de tels programmes, la construction d'un modèle à la main peut s'avérer très longue et très complexe surtout si le modèle construit comporte de très nombreuses réactions ou si les informations disponibles ne sont pas suffisantes. De programmes permettant de reconstruire des réseaux métaboliques automatiquement sont aussi nécessaires.

2.3.4 Reconstruction des réseaux métaboliques

La reconstruction d'un réseau métabolique ne peut se faire qu'en associant des données de différentes sources afin de trouver la liste la plus exhaustive possible des réactions impliqué dans le modèle étudié et de leurs différents partenaires. On distingue quatre types principaux de sources de données (dans l'ordre d'importance) [21]:

- Biochimie

La preuve la plus évidente de l'existence d'une réaction biochimique est l'identification biochimique d'une enzyme, de sa fonction et donc des partenaires de la réaction.

- Génomique

L'attribution de fonction à des ORFs (*Open Reading Frame* ou cadre de lecture ouvert) par homologie de séquence permet de savoir si une enzyme peut bien être synthétisée. Par homologie avec des enzymes dont la fonction a été prouvée biochimiquement, on peut supposer qu'une réaction existe.

- Physiologie

La physiologie de la cellule peut conduire à des suppositions de réactions biochimiques. Il est ainsi possible de savoir, par exemple, si une cellule peut métaboliser *in vivo* un acide aminé donné. Par homologie avec des organismes connus un biologiste peut supposer l'existence d'une voie métabolique.

- Modélisation du métabolisme

Les modèles *in silico* du métabolisme permettent de valider certaines hypothèses émises et de proposer de nouvelles pistes expérimentales pour préciser le réseau construit.

Le travail de reconstruction a souvent déjà été mené, surtout sur les organismes modèles les plus étudiés comme l'homme ou la souris. Sur le web, de très nombreuses bases de données regroupent des modèles informatisés de ces organismes. La base de données Pathguide [4] recensait en 2005 190 bases de données de voies métaboliques. En 2011 elle en compte 328 dont 91 bases de données sur l'homme. Cela représente plus de 151000 voies métaboliques décrites (tous organismes confondus) [165]. Les bases de données les plus utilisées au laboratoire KEGG et EnsEMBL vont être détaillées par la suite.

2.3.4.1 KEGG (Kyoto Encyclopedia of Genes and Genomes)

Description

KEGG est une des bases les plus utilisées au laboratoire. KEGG est un ensemble de 13 bases de données (développé depuis 1995 par les laboratoires Kanehisa à Kyoto) organisant des données de biologie des systèmes, génomiques et chimiques. La base la plus importante est la base KEGG PATHWAY présentant des données de voies métaboliques. La représentation graphique des voies métaboliques dans KEGG est appelée une carte (*map*). Les voies métaboliques sont hiérarchisées en 7 grands groupes :

- *Global Map* : regroupant les voies métaboliques globales comme l'ensemble des voies métaboliques d'une cellule par exemple. Ces voies métaboliques sont classées à part car elles permettent surtout de visualiser la place des différentes voies métaboliques entre elles et de naviguer entre ces voies. Tous les composés ne sont pas représentés, seuls les noms des voies métaboliques sont présents. Le programme KEGG Atlas permet de naviguer dans ces très grandes cartes et de mettre en évidence des groupes de voies métaboliques (cf. figure 34).

- *Metabolism* : regroupe toutes les cartes reliées au métabolisme : métabolisme des sucres, métabolisme énergétique, métabolisme des lipides, …

- *Genetic Information Processing* : regroupe les cartes liées au traitement de l'information génétique : transcription, traduction, réparation de l'ADN, …

- *Environmental Information Processing* : regroupe les cartes liées au traitement des signaux extérieurs : transduction du signal, transports membranaires

- *Cellular Processes* : regroupe les cartes liées aux processus cellulaires n'entrant pas dans les catégories précédentes. Par exemple la motilité cellulaire, le cycle cellulaire, l'apoptose, …
- *Organismal Systems* : regroupe les cartes liées aux fonctions physiologiques de l'organisme comme le système immunitaire ou le système endocrinien
- *Human Diseases* : regroupe les cartes liées à des pathologies humaines
- *Drug Development* : regroupe les cartes de synthèse de certains médicaments.

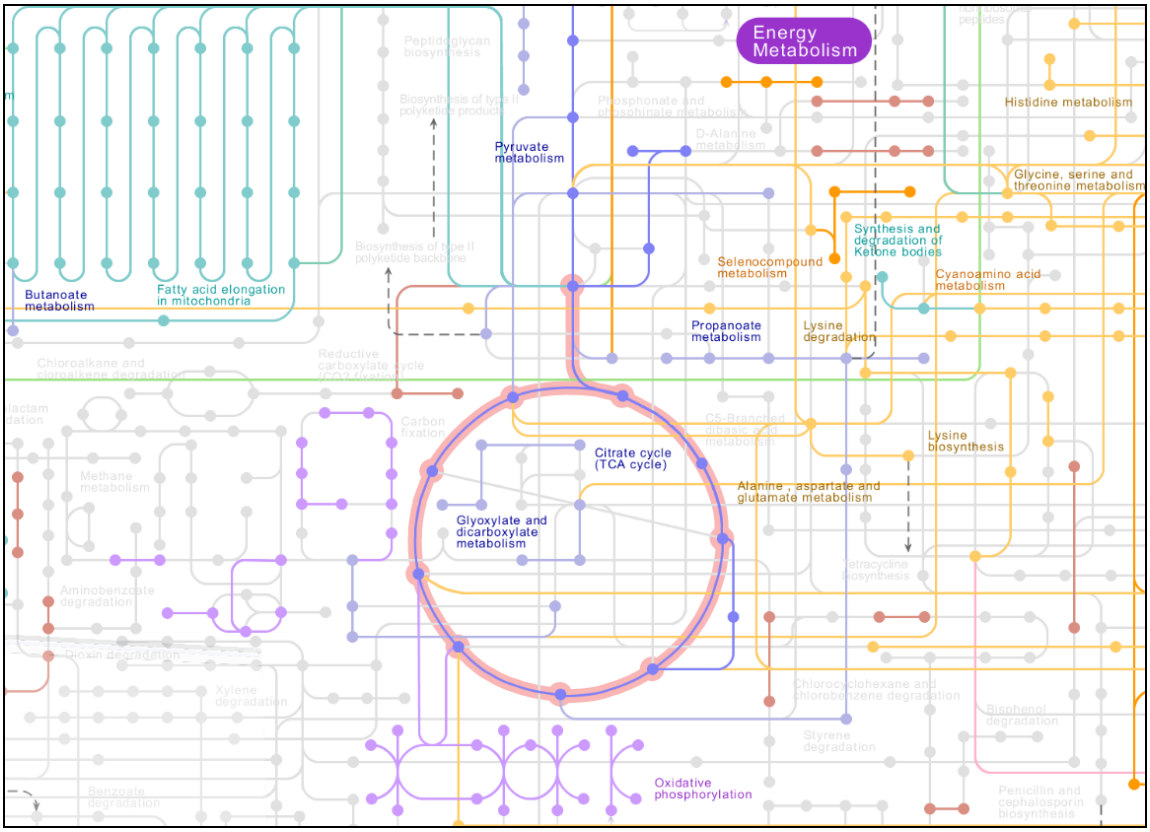
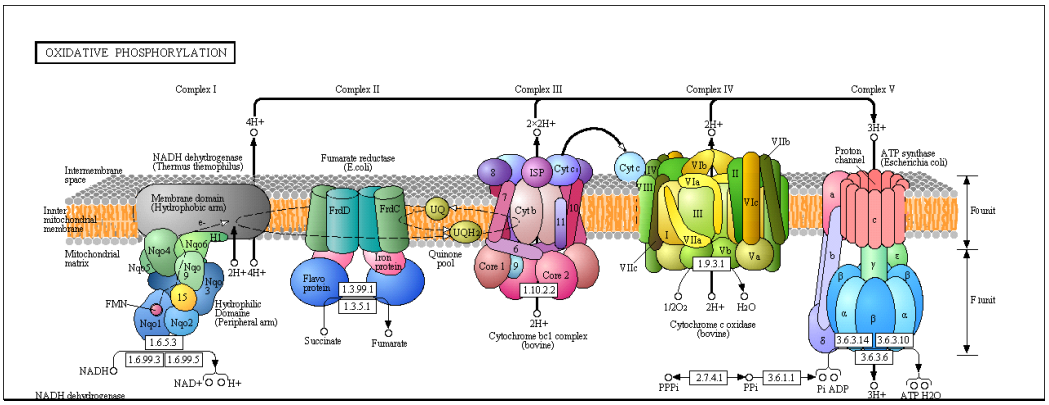


Figure 34 : Mise en évidence du cycle de Krebs dans le métabolisme global d'une cellule humaine par KEGG Atlas

Le but de ces cartes est de présenter des voies métaboliques de la manière la plus claire possible. Pour cela, les cartes sont toutes dessinées à la main. Les cartes présentées sont effectivement assez claires en regard de leur complexité. Cela est possible grâce à deux artifices de représentation : les petites molécules (ATP, eau, protons, …) sont très souvent non représentées et les composés présents plusieurs fois dans le graphe sont souvent dupliqués. Cela a deux conséquences : les cartes ne concordent pas avec la biologie puisque tous les participants ne sont pas représentés et la représentation des voies métaboliques donnée par ces cartes est faussée : il est impossible de visualiser simplement les composés les plus employés puisqu'ils sont dupliqués dans le graphe ou parfois même omis (eau, ATP, …). Dessiner les cartes à la main introduit aussi de grandes disparités de représentation entre certaines voies, on n'a donc pas vraiment de convention de représentation (cf. figure 35).



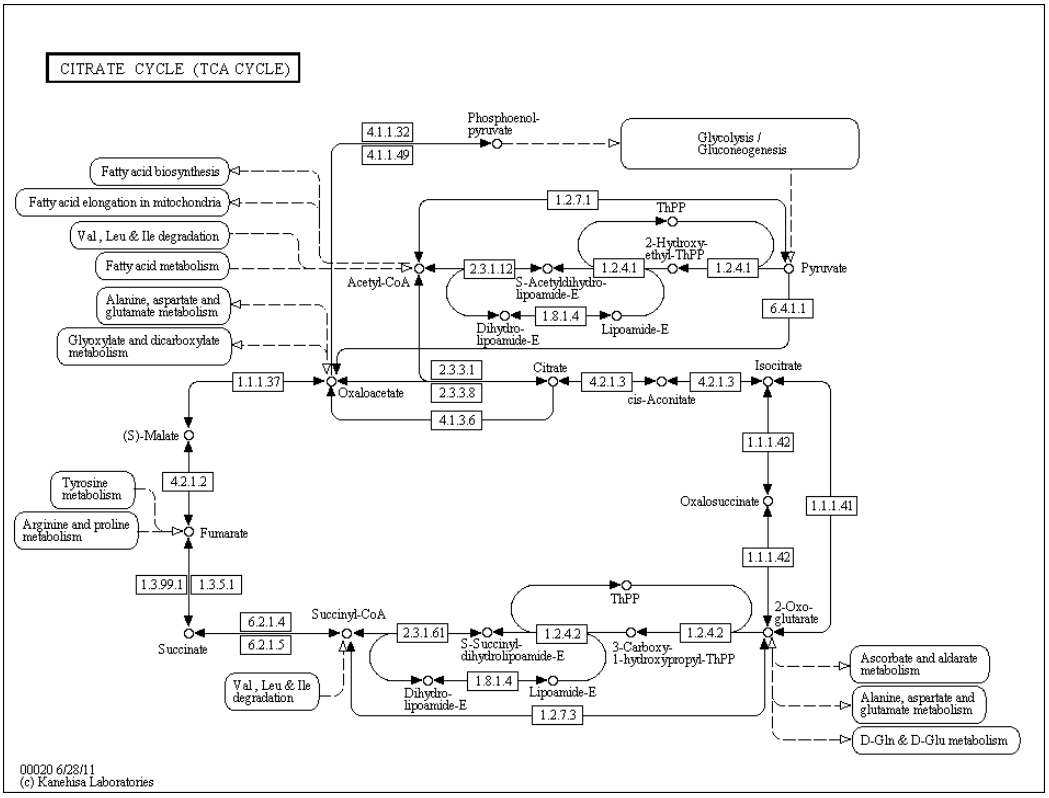


Figure 35 : Différences de représentations de deux voies métaboliques dans KEGG
Le haut de la figure représente une partie de la carte des phosphorylations oxydatives et la partie du bas le cycle de Krebs. La représentation du bas est la plus utilisée dans la base KEGG

KEGG exporte l'intégralité de ces cartes dans un format XML qui lui est propre : le KGML (cf. figure 36).

Le KGML

Le format KGML ne peut pas être considéré comme un format d'échange de fichier au même titre que le SBML. Le KGML ne sert qu'à décrire des voies métaboliques de KEGG comme le montre la déclaration d'une voie métabolique en KGML (cf. figure 37)

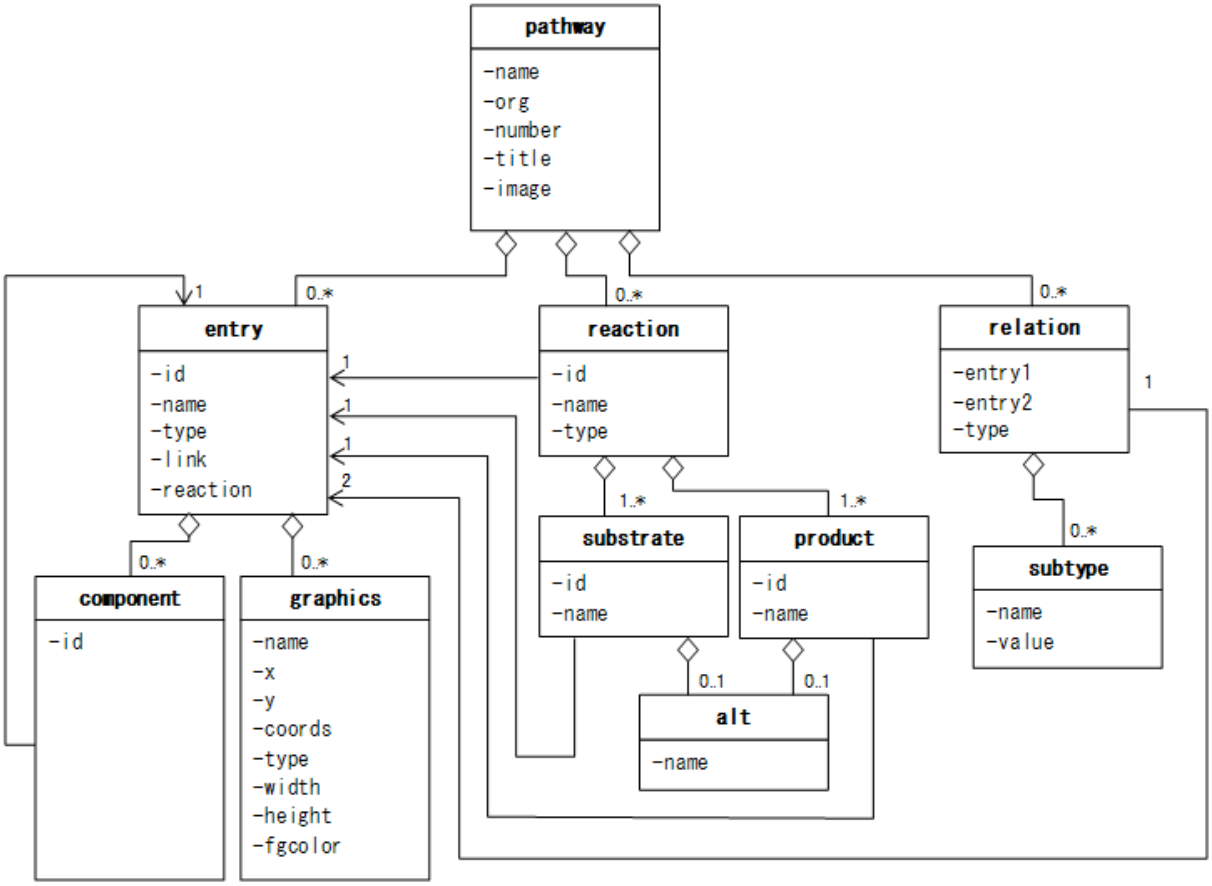


Figure 36 : Schéma de la structure d'un fichier KGML.
Tiré de [159]. Toutes les flèches signifient que l'élément en bout de flèche contient l'élément au départ de la flèche. Les flèches ayant une tête simple signifient que le nombre exact d'éléments contenus est connu. Pour les flèches à tête en forme de losange, la cardinalité maximale est indéterminée.

```
1    <pathway name="path:hsa00020" org="hsa" number="00020"
2             title="Citrate cycle (TCA cycle)"
3             image="http://www.genome.jp/kegg/pathway/hsa/hsa00020.png"
4             link="http://www.genome.jp/kegg-bin/show_pathway?hsa00020">
```

Figure 37 : Description d'une voie métabolique(le cycle de Krebs) en KGML.

L'élément central d'un fichier KGML est la *voie métabolique* (*Pathway*) cf. figure 37 . Une voie métabolique est caractérisée par :

- Un nom (*name*), identifiant unique de la voie métabolique pour un organisme donné dans KEGG. Il commence toujours par *path* indiquant que l'on décrit une voie métabolique. Il est ensuite composé du code de l'organisme (*org*) et de l'identifiant unique de la voie (*nombre*)

- Un organisme (*org*) : correspond à l'organisme vivant pour lequel cette voie est décrite. Il est déclaré par son code sur KEGG. Il correspond la plupart du temps à la première lettre du nom de genre et aux deux premières lettres du nom de l'espèce. Ici le code est « *hsa* » qui correspond à l'homme (*Homo sapiens*).

- Un nombre (*number*) : identifiant de la voie métabolique générale (hors du contexte de l'organisme). En effet, on peut récupérer une même voie métabolique (ayant le même identifiant) chez plusieurs organismes. Le *nom* variera mais pas le *nombre*. Par exemple, cette voie métabolique chez la souris est décrite par le *nombre* 00020 et par le *nom* path:mmu00020.

- Un titre (*title*) : correspond au titre de la voie métabolique sur KEGG

- Un lien vers la représentation graphique de la voie dans la base KEGG (*image*)

- Un lien vers une représentation graphique dynamique de la voie dans la base KEGG (*link*).

Les trois derniers éléments sont facultatifs puisqu'ils peuvent être déduits des trois premiers. Le fait que les trois premiers éléments de description fassent obligatoirement référence à des éléments dans KEGG montre bien que le KGML n'est pas un langage construit pour échanger des données de voies métaboliques. Une voie métabolique comprend des *entrées* (*entry*), des *réactions* et des *relations*. Les entrées correspondent à tous les nœuds possibles d'un graphe (composés, enzymes, …) cf. figure 38. Une entrée peut être composée d'entrées (complexe).

```
1    <entry id="44" name="hsa:4190 hsa:4191" type="gene" reaction="rn:R00342"
2          link="http://www.kegg.jp/dbget-bin/www_bget?hsa:4190+hsa:4191" >
3          <graphics name="MDH1, MDH-s, MDHA, MGC:1375, MOR2..." fgcolor="#000000" bgcolor="#BFFFBF"
4          type="rectangle" x="232" y="345" width="46" height="17" />
5    </entry>
```

Figure 38 : Description de l'entrée « malate déhydrogenase » dans la voie métabolique du cycle de Krebs.

Une entrée est définie par un identifiant unique dans le fichier KGML (*id*), un nom (*name*) composé des identifiants auxquels correspond l'entrée courante dans KEGG. Une entrée possède un *type* pouvant être une enzyme, un composé biochimique, … Ici l'entrée est le produit d'un gène (*gene*) c'est-à-dire une protéine. Il est impossible de faire la différence entre le gène et le produit d'un gène dans un fichier KGML. Une entrée peut aussi être de type *map* qui correspond à une autre voie métabolique. Des liens vers des voies métaboliques en KGML peuvent donc être inclus dans un modèle de voie métabolique, ce qui correspond un peu à de la composition de modèles. L'attribut réaction (*reaction*) correspond à la réaction que l'enzyme catalyse (il peut y avoir plusieurs identifiants dans ce champ, ce champ n'est présent que si l'entré est une enzyme). Les lignes 3 et 4 correspondent aux informations graphiques d'une entrée : nom, forme, couleurs, coordonnées et tailles.

Les réactions correspondent aux réactions biochimiques et associent des substrats et des produits.

```
1    <reaction id="44" name="rn:R00342" type="reversible">
2          <substrate id="64" name="cpd: C00149 "/>
3          <product id="61" name="cpd:C00036"/>
4    </reaction>
```

Figure 39 : Description d'une réaction dans un fichier KGML

Une réaction est caractérisée par un identifiant (*id*) unique dans le fichier KGML, par l'identifiant de la réaction dans KEGG (*name*) et par un type pouvant être réversible ou irréversible. Elle contient une liste de substrats et de produits ayant été au préalable définis parmi les entrées. Ils sont rappelés par leur identifiants. Les noms des substrats et des entrées n'ont pas d'utilité ici.

L'enzyme catalysant cette réaction est déterminée par l'identifiant de la réaction qui fait référence à un champ *reaction* d'une des entrées.

Enfin, les relations correspondent aux arcs du graphe. Une relation correspond à un lien entre trois points dans le graphe c'est-à-dire à deux arcs cf. figure 40.

```
1    <relation entry1="79" entry2="45" type="ECrel">
2        <subtype name="compound" value="59"/>
3    </relation>
```

Figure 40 : Description d'une relation dans un fichier KGML

La relation de la figure 40 fait intervenir trois entrées définies par leurs identifiants : 79, 45 et 59. La relation correspond ici à deux arcs : le premier entre l'entrée 79 et l'entrée 59 et le deuxième entre l'entrée 59 et l'entrée 45. Le type *ECRel* indique que l'on décrit une relation entre deux enzymes (79 et 45) « liées » par un composé (59). Il existe d'autres types de relation :

- PPRel : relation entre deux protéines.
- GERel : relation « Gene Expression » indiquant la relation entre un facteur de transcription et son gène cible.
- PCRel : relation entre une protéine et un composé biochimique.
- Maplink : relation vers une autre voie métabolique

Dans le cas des relations PPRel et GERel, la balise *subtype* ne correspond pas obligatoirement à un composé. Cela peut être une activation, une inhibition, une association, une dissociation, etc. Le type de la relation est alors indiqué dans la balise *subtype* dans l'attribut *name* et un symbole particulier est inscrit dans l'attribut *type* (« --> » pour une activation).

Le KGML associe donc des informations du modèle et des informations de graphe ce qui le rend souvent redondant. Ainsi les réactions sont représentées par des objets de type « réaction » mais aussi par des relations de type « ECrel ». De plus les informations graphiques ne permettent pas de recréer un graphe identique à celui présenté sur le site de KEGG. Puisqu'il correspond au graphe généré à partir des publications et non au graphe redessiné manuellement.

Les fichiers KGML étaient disponibles via un serveur FTP gratuit jusqu'en juillet 2011. Ils pouvaient donc être tous téléchargés facilement et rapidement et pouvaient servir à remplir une base de données de manière automatique. Depuis l'accès au serveur est devenu payant et les KGML ne peuvent plus être téléchargés gratuitement que via les pages web.

API KEGG

L'autre intérêt de KEGG est la mise à disposition d'une API. Une API (*Application Programming Interface*) est une interface fournie par un programme permettant à différents logiciels d'interagir entre eux. L'API KEGG est un programme permettant de faire des requêtes directement sur leurs serveurs via un programme informatique. L'avantage d'une telle structure est de pouvoir faire du chainage logiciel et de l'analyse à haut débit. L'API fournie est une API respectant le standard SOAP (*Simple Object Access Protocol*). SOAP est un protocole permettant le transfert, le plus souvent via http, de messages entre objets informatiques. Cette API est donc mise à disposition sous la forme de service web interrogeable dans de très nombreux langages de programmation.

L'exemple suivant de code écrit en Ruby utilisant l'API KEGG permet d'afficher la liste des identifiants et des descriptions des voies métaboliques humaines enregistrées dans KEGG.

```ruby
require 'bio'
serv = Bio::KEGG::API.new

list = serv.list_pathways("hsa")
list.each do |path|
  print path.entry_id, "\t", path.definition, "\n"
end
```

Figure 41 : Exemple d'utilisation de l'API KEGG
Petit programme en Ruby permettant d'afficher la liste des voies métaboliques humaines de KEGG et leur description.

Une des méthodes que nous utilisons le plus est la méthode *bget* qui permet d'obtenir des informations précises sur tout élément de KEGG :

```
print serv.bget("rn:R00220")

ENTRY       R00220                                  Reaction
NAME        L-serine ammonia-lyase
DEFINITION  L-Serine <=> Pyruvate + NH3
EQUATION    C00065 <=> C00022 + C00014
RPAIR       RP04290  C00022_C00065 main
            RP06120  C00014_C00065 leave
ENZYME      4.3.1.17    4.3.1.19
PATHWAY     rn00260         Glycine, serine and threonine metabolism
            rn01100         Metabolic pathways
            rn01110         Biosynthesis of secondary metabolites
ORTHOLOGY   K01752         L-serine dehydratase [EC:4.3.1.17]
            K01754         threonine dehydratase [EC:4.3.1.19]
///
```

Figure 42 : Exemple d'utilisation de la commande bget de l'API KEGG.
La première ligne montre la commande faisant appel à la méthode bget. Le reste de la figure correspond au résultat de cette commande.

La figure 42 montre le résultat de la commande *bget* sur la réaction *rn:R00220*. Le résultat de cette commande est une fiche présentant toutes les informations contenues dans KEGG à propos de cette réaction avec notamment le nom de l'enzyme catalysant la réaction et son équation. On voit bien dans cet exemple que les réactions sont confondues avec les enzymes, ce qui est biologiquement faux.

2.3.4.2 EnsEMBL

EnsEMBL est une base de données de génomique. Cette base de données offre aussi une API permettant d'annoter automatiquement des génomes. C'est un projet commun à l'EBI (*European Bioinformatics Institute*) et le *Sanger Institute*. EnsEMBL comprend plus de 60 génomes annotés dont une trentaine de mammifères. Des sites EnsEMBL propres aux plantes ou bactéries ont été développés [174–176]. Pour chaque génome analysé EnsEMBL tente d'annoter automatiquement les gènes reconnus en utilisant des programmes en perl libres d'accès. Pour faire cette annotation, il s'appuie sur une base de données de séquences nucléotidiques et protéiques. Pour être annotés les gènes sont alignés avec les séquences de la base de données en utilisant BLAT [55]. EnsEMBL constitue donc une immense base de données de séquences sur laquelle il est possible de travailler en utilisant leur API. L'API EnsEMBL a été très utilisée dans cette thèse. Son utilisation est décrite au paragraphe 3.2.2.

Une des limitations majeure de cette API est liée à son mode de mise à jour. Cette API constitue un très gros programme Perl qu'il faut installer en local pour pouvoir l'utiliser, au contraire de la plupart des API de bases de données en lignes qui sont directement interrogeables via le web. À chaque mise à jour correspond une nouvelle version de la base de données EnsEMBL. Normalement, si on reste sur une ancienne version de l'API on devrait interroger l'ancienne version de la base mais ce n'est pas toujours vrai. Par exemple, la position et la séquence des clones varie entre deux mises à jour car certaines données des clones sont prises dans les nouvelles versions de la base et d'autres dans les anciennes et qu'il y a une incohérence entre ces données. La richesse des outils et des données d'EnsEMBL est tellement grande qu'elle est totalement incontournable pour le travail que réalisé dans cette thèse.

2.3.5 Étude des réseaux métaboliques

Les réseaux métaboliques peuvent être étudiés en utilisant de nombreuses méthodes différentes. Certaines de ces méthodes font appel aux graphes mais pas toutes. Ce n'est par exemple pas le cas des systèmes d'équations différentielles.

2.3.5.1 Systèmes d'équations différentielles

La modélisation des réseaux métaboliques par des systèmes d'équations différentielle est assez ancienne et très répandue. Cette méthode permet une analyse dynamique des réseaux puisque les variables traitées sont continues. On modélise les variations de concentrations des métabolites du système.

Par exemple, le modèle de Michaëlis-Menten de la figure 27 en utilisant la loi d'action de masse pourrait être modélisé par le système d'équations différentielles de la figure 43.

L'étude de ce type de modèle se fait toujours à l'état stationnaire. Les variations des concentrations des métabolites sont alors globalement nulles. Cela implique des systèmes suffisamment stables, ce qui est souvent vrai pour des intervalles de temps courts. Il est donc impensable de modéliser ainsi la dynamique de systèmes comme la consommation d'ATP dans un muscle qui est extrêmement variable sur de courts intervalles de temps. Les molécules du système doivent aussi être en quantité suffisante pour pouvoir parler de concentrations. Ce type de modélisation est donc peu adapté à des systèmes ou les réactifs des réactions sont en quelques copies dans la cellule.

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_2} E + P$$

$$\begin{cases} \frac{d[S]}{dt} = -k_1\,[E][S] + k_{-1}\,[ES] \\ \frac{d[E]}{dt} = -k_1[E][S] + k_{-1}[ES] + k_2[ES] \\ \frac{d[ES]}{dt} = +k_1[E][S] - k_{-1}[ES] - k_2[ES] \\ \frac{d[P]}{dt} = +\,k_2[ES] \end{cases}$$

Figure 43 : Système d'équations différentielles décrivant le modèle de Michaëlis-Menten.
Ces équations représentent l'évolution de la concentration des différents métabolites au cours du temps. La première réaction correspond à la formation du complexe enzyme substrat. C'est une réaction réversible. La seconde réaction correspond à la formation du produit de la réaction. Elle est général considérée comme non réversible.

Dans ce modèle très simple, il y a seulement deux réactions et 7 paramètres : k_1, k_{-1}, k_2 et les concentrations de E, S, ES et P. Pour des modèles plus proches de la biologie, où les réactions ne vont plus répondre à la loi d'action de masse mais à des lois beaucoup complexes, le nombre de paramètres sera beaucoup plus grand. Comme ces paramètres sont souvent difficiles à estimer biologiquement. Il est donc intéressant de faire une prédiction de leurs valeurs.

Pour réaliser cette prédiction des paramètres du système, un logiciel de simulation en cours de développement au Laboratoire d'informatique de L'École Polytechnique sera utilisé [6,141].

2.3.5.2 Réseaux de Petri

L'utilisation des réseaux de Petri pour la modélisation des réseaux biologiques est beaucoup plus récente. Les réseaux de Petri permettent de faire une étude aussi bien dynamique que statique du réseau. Les réseaux de Petri, au contrairement aux systèmes d'équations différentielles impliquent des variables discrètes. Un réseau de Petri est défini par : des *places*, des *transitions*, des *arcs*, et des *jetons*. Les arcs relient les places et les transitions entre elles. Les jetons représentent les quantités des différentes espèces du système. Ils sont appliqués sur les places qui représentent les différents types d'entités du système. Les transitions permettent le passage des jetons d'une place à une autre. Le modèle biologique simple de la figure 27 est représenté dans la figure 44

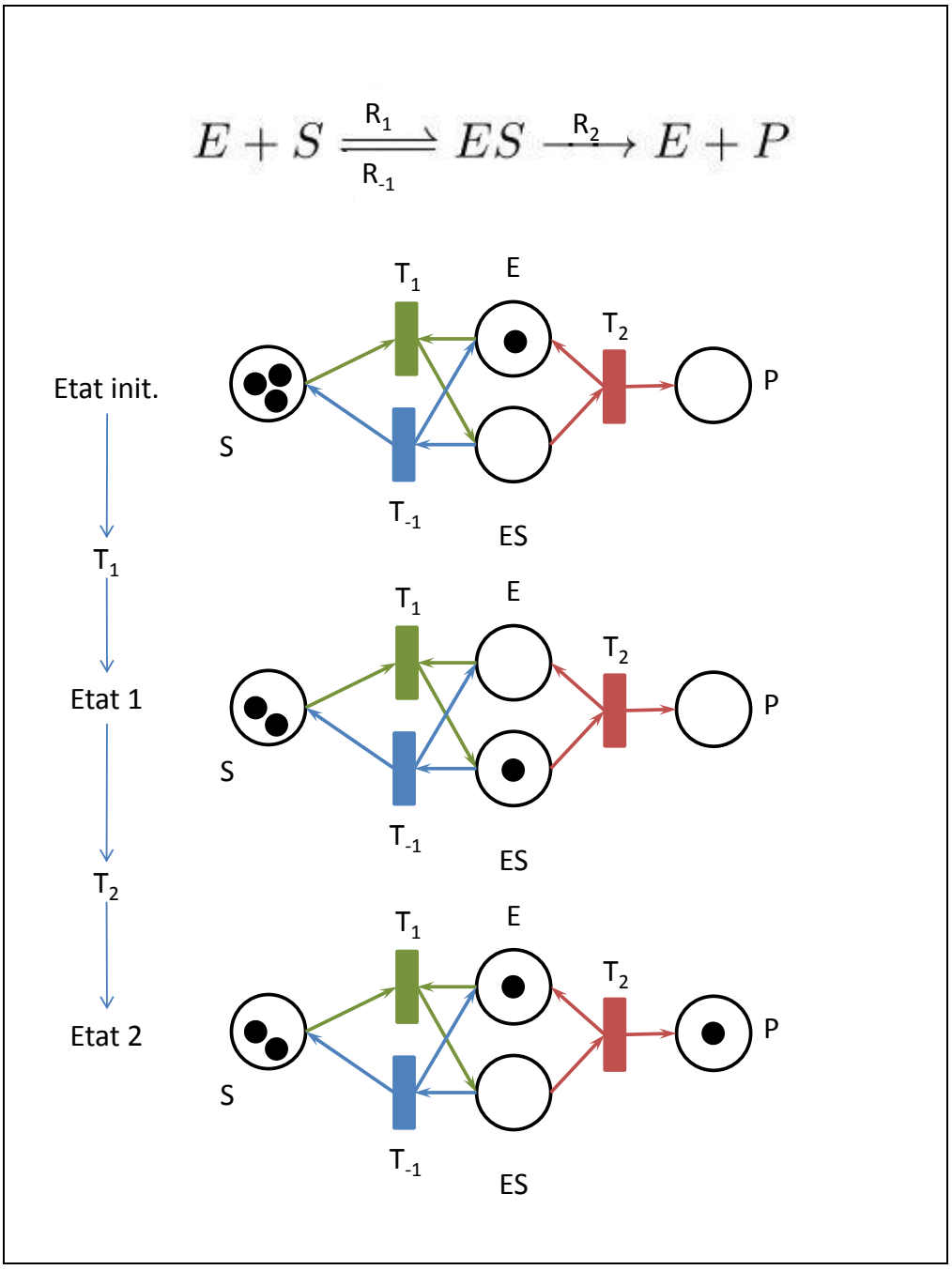


Figure 44 : Représentation d'un système enzymatique simple par un réseau de Petri.
Le modèle représenté dans cette figure est le modèle de Michaëlis-Menten (cf. figure 27). Le réseau de Petri correspondant est représenté par un graphe biparti dans lequel les places (E, S, ES et P) sont reliées aux transitions T_1, T_{-1}, et T_2. Les transitions représentent les réactions biologiques ayant le même indice. Les jetons sont représentés par des cercles noirs dans les places. À l'état initial, il y a dans le réseau trois substrats S et une enzyme E et seule la transition T_1 est possible: pour T_{-1} et T_2 ES est absent. La transition T_1 consomme un jeton de S et un jeton de E pour créer un jeton de ES et arriver dans l'état 1. Alors les transitions T_{-1} et T2 sont possibles. On choisit de jouer la transition T2 pour arriver dans l'état 2. Les couleurs appliquées aux transitions et aux arcs sont là pour plus de lisibilité mais ne reflètent pas de comportement particulier du réseau.

Dans un réseau de Petri, il faut dédoubler les réactions réversibles. Ce type de représentation peut donc devenir très complexe à comprendre si le réseau est de taille importante. Les transitions dans

le réseau peuvent être *jouées* arbitrairement ou en suivant une loi mathématique. Ce type de réseau a l'avantage de représenter le comportement d'entités en faible quantité dans le système et reflète le comportement individuel des molécules. Les réseaux de Petri permettent de modéliser des réseaux biochimiques, génétiques et d'interactions cellulaires. Il existe des réseaux de Petri colorés où les jetons sont porteurs d'information (souvent représentée par une couleur) qui permettent de savoir d'où vient un jeton à la fin d'une simulation, ou le niveau d'expression d'un gène par exemple.

2.3.5.3 Modes élémentaires et voies métaboliques extrêmes

Modes élémentaires

Un mode élémentaire (*elementary flux mode*) est un ensemble minimal de réactions (caractérisées par leurs enzymes) pouvant avoir lieu à l'état stationnaire. Pour calculer ces modes élémentaires, les métabolites sont classés en deux catégories : les métabolites *internes* et les métabolites *externes*. Les métabolites externes sont correspondent le plus souvent aux *puits* et aux *sources* du graphe. Dans un graphe, les nœuds puits sont des nœuds vers lesquels les arrêtes convergent mais d'où aucune arrête ne repart :

Dans un graphe G(V,E) un nœud *n* est *puits* si, pour tout nœud *m* différent de *n* il peut exister dans E une arrête (*m*,*n*) mais pas d'arrête (*n*,*m*).

Un nœud *source* est nœud n'ayant pas d'arrête entrante mais ayant des arrêtes sortantes :

Dans un graphe G(V,E) un nœud *n* est *source* si, pour tout nœud *m* différent de *n* il peut exister dans E une arrête (*n*,*m*) mais pas d'arrête (*m*,*n*).

Par définition, les métabolites externes ne seront pas équilibrés dans le système. Remarque : les métabolites externes ne sont pas obligatoirement des métabolites situés hors du système. Les métabolites internes sont tous les métabolites qui ne sont pas externes. Les métabolites internes doivent être équilibrés dans le système, c'est à dire que la formation de ces métabolites doit être équilibrée par leur consommation. Les réactions doivent elles aussi être séparées en deux groupes : les réactions réversibles et les réactions irréversibles. Une réaction est dite irréversible si le flux dans le sens inverse de la réaction est négligeable vis à vis de celui dans le sens de la réaction [113].

Un mode élémentaire doit répondre aux conditions suivantes (tirées de l'article [58] qui synthétise les articles [110,112–115] :

Notations : S est la matrice de stœchiométrie comprenant *m* métabolites (lignes) et *r* réactions (colonnes). *e* est un mode élémentaire représenté par un vecteur de taille *r* où chaque élément décrit le flux de réaction correspondante. Un mode élémentaire peut donc être vu comme un vecteur de flux. Une voie métabolique est l'ensemble des réactions effectives dans un mode élémentaire (toutes les réactions d'une voie métabolique doivent avoir des flux non nuls). Elle est notée *P(e)*.

Stabilité. **S e = 0.** Aucun des métabolites internes ne doit être produit ou consommé sur l'ensemble de mode élémentaire. Cela explique que les modes élémentaires relient souvent des métabolites externes ou forment des cycles.

Faisabilité. Un flux $e_i \geq 0$ si la réaction *i* est irréversible. Un mode élémentaire doit donc respecter les contraintes thermodynamiques du système.

Non décomposabilité. Il n'existe pas de vecteur *v* non nul et différent de *e* respectant les règles 1 et 2 tel que *P(v)* soit un sous ensemble de *P(e)*. Cela implique qu'il n'existe pas deux modes élémentaires identiques dans un même réseau. D'un point de vue biologique, toute suppression de

réaction dans un mode élémentaire implique la disparition de la voie métabolique correspondante car tout *knock out* d'une enzyme de cette voie inhibe la voie complète.

L'ensemble de ces trois règles implique que la totalité des états stables (ensemble des vecteurs de flux) d'un système est couvert par l'ensemble des modes élémentaires :

Équation 2

$V = \sum_j \alpha_j e_j$

V étant l'ensemble des vecteurs de flux faisables à l'état stable

$\alpha_j \geq 0$

e_j est le j-ème mode élémentaire.

Les bases du calcul des modes élémentaires dans les réseaux biologiques ont été apportées en 1994 [112]. La preuve que l'intégralité des modes élémentaires d'un réseau est trouvée par cet algorithme est apportée en 2002 [115] . Cet algorithme est encore utilisé dans les logiciels actuels de calcul des modes élémentaires.

Voies extrêmes

Les voies extrêmes répondent elles aussi aux trois lois définissant les modes élémentaires. Cela implique que les voies extrêmes sont un sous ensemble des modes élémentaires. Les voies extrêmes répondent à deux règles supplémentaires :

Reconfiguration du réseau. Chaque réaction doit être classée soit comme flux d'échange (entrées, sorties du système) soir comme réaction interne. Les réactions internes réversibles doivent être découpées en deux réactions irréversibles. On ne pourra donc pas avoir de flux négatif dans une voie extrême puisque toutes les réactions sont irréversibles. Les flux d'échanges peuvent être réversibles.

Indépendance systémique. L'ensemble des voies extrêmes d'un réseau l'ensemble *minimal* de modes élémentaires respectant l'équation 2.

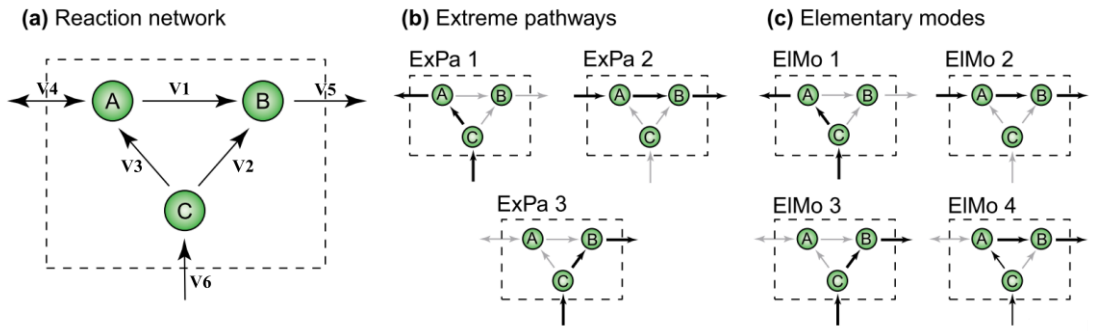


Figure 45 : Comparaison des voies extrêmes et des modes élémentaires.
Tirée de [90] fig1.

Comparaison voies extrêmes / modes élémentaires

L'article [90] fourni une comparaison des voies extrêmes et des modes élémentaires sur un système très simple.

La figure 45 présente le résultat du calcul des voies extrêmes (b) et des modes élémentaires (c) sur un réseau simple (a). Les métabolites externes ne sont pas figurés. Les métabolites A, B et C sont des métabolites internes. Pour les voies extrêmes, les réactions V4, V5 et V6 sont des flux d'échange, les autres sont des réactions internes. La réaction V4 est la seule réaction réversible du système. Trois voies extrêmes sont trouvées : V6 + V3 - V4 (ExPa 1), V4 + V1 + V5 (ExPa 2) et V6 + V2 + V5 (ExPa 3). Les trois premiers modes élémentaires correspondent exactement à ces trois voies extrêmes (les voies extrêmes sont bien un sous ensemble des modes élémentaires). On trouve le mode élémentaire V6 + V3 + V1 + V5 (ElMo 4) qui ne correspond pas à une voie extrême. Ce mode élémentaire peut être obtenu par la somme des deux premiers modes élémentaires :

ElMo1 + ElMo2 = V6 + V3 - V4 + V4 + V1 + V5 = V6 + V3 + V1 + V = ElMo 4

Cela implique que le mode 4 ne fait pas partie de l'ensemble minimal décrit dans la règle 5. Ce n'est donc pas une voie extrême.

L'obligation de reconfiguration du réseau pour calculer les voies extrêmes dès que sa topologie change peut s'avérer fastidieux et couteux en termes de calcul. On sait que le nombre de modes élémentaires dans un réseau peut très rapidement exploser**.** Par exemple un réseau de 63 métabolites et 83 réactions peut amener au calcul de plus de 49 millions de modes élémentaires [138]. De plus le nombre de voies extrêmes est toujours très proche de celui des modes élémentaires. Enfin, la suppression des voies non minimales n'a pas de sens biologique. En effet, sur cet exemple simple, on s'aperçoit que même si la réaction V2 est impossible on peut synthétiser du B à partir du C par le mode élémentaire 4, ce qu'il est impossible de retrouver par les voies extrêmes. Pour ces différentes raisons, nous avons préféré nous intéresser aux modes élémentaires sans considérer les voies extrêmes.

Logiciels de calcul de modes élémentaires

Les tailles importantes que peuvent prendre les réseaux biologiques posent deux contraintes au calcul des modes élémentaires : le temps de calcul qui peut être très long et le stockage des matrices nécessaires au calcul en mémoire vive. Souvent les programmes se focalisent sur le premier point et oublient le deuxième. Trois programmes majeurs permettent le calcul des modes élémentaires : Metatool [53,94], EfmTools [124] et ElmoComp [51,138]. Il existe aussi une librairie en Python qui permet ce calcul : ScrumPy [167]. Malheureusement, les auteurs ne garantissent pas son fonctionnement au-delà de 100 réactions et métabolites dans le système en raison d'une charge de la mémoire trop importante.

Metatool [94]. C'est une implémentation en C de l'algorithme de Schuster publié en 1999. Il a été amélioré de très nombreuses fois et les versions récentes sont construites pour fonctionner avec les logiciels Matlab ou Octave. Les logiciels d'étude des voies métaboliques YANA [116] et PySCeS [87] utilisent différentes versions de Metatool pour le calcul des modes élémentaires. Il faut fournir en entrée du logiciel la liste des réactions réversibles, des réactions irréversibles, les métabolites internes et les métabolites externes ainsi que les équations stœchiométriques du système. Des versions récentes de Metatool ont été incluses dans des packages MatLab.

EfmTools [124]. Cet algorithme est inclus dans un plugin pour le logiciel MatLab : CellNetAnalyzer [60] et ne peut pas être utilisé en dehors de CellNetAnalyzer. EfmTools est basé sur un algorithme très proche de celui de Metatool et devrait donner les mêmes résultats, ce que nous n'avons pas pu tester. CellNetAnalyzer est la version actuelle du programme FluxAnalyzer [61]. Ce programme est un des premiers à mettre en avant l'intérêt de l'étude des modes élémentaires dans un

réseau métabolique vis-à-vis de programmes très utilisés comme GEPASI [21–23] tournés vers l'étude cinétique des réseaux.

ElmoComp [51,138]. C'est une version parallélisée (dont les calculs peuvent être lancés sur plusieurs processeurs simultanément) de l'algorithme de recherche des modes élémentaires. Le calcul des modes élémentaires peut donc être distribué sur de nombreux ordinateurs en même temps ce qui peut améliorer la vitesse de calcul en fonction des ordinateurs utilisés. Sur 256 nœuds de calcul du super ordinateur Blue Gene, le calcul des 49.764.544 modes élémentaires du réseau de 63 métabolites et 83 réactions de [138] prend moins de trois heures. Ce type de calcul serait irréalisable sur une machine de bureau (si on suit le modèle expliqué dans [59], il faudrait plus 50 ans pour le calculer sur une machine standard). L'inconvénient majeur de cet algorithme est la nécessité de stocker en mémoire de nombreuses matrices et il nécessite donc une très grande place mémoire pour fonctionner.

Pour le calcul des modes élémentaires nous avons donc décidé d'utiliser le programme Metatool (v. 4.3) dont les entrées sorties sont simples et dont l'algorithme a été validé.

Intérêt des modes élémentaires

L'étude des modes élémentaires et des voies extrêmes permet de déduire certaines propriétés des réseaux biologiques.

La première de ces propriétés dérive directement de la définition des modes élémentaires : trouver les voies métaboliques dans un réseau biologiques. Elles nous permettent de savoir quelles sont les voies de synthèse ou de dégradation d'un métabolite par exemple.

Il est par exemple possible de déduire un milieu minimal pour des micro-organismes grâce aux voies extrêmes et aux modes élémentaires et que le milieu déduit est cohérent avec les résultats expérimentaux [108,109]. De la même manière, on peut savoir quels substrats sont à introduire dans le milieu pour permettre la synthèse d'un produit donné. Les auteurs identifient aussi des réactions qui ne sont présentes dans aucune des voies extrêmes trouvées. L'absence de ces réactions dans les voies extrêmes est vraisemblablement due au fait que les réactions utilisant leurs produits ou générant des substrats qui leur sont nécessaires ne sont pas présentes dans le réseau. Cette absence suggère que le réseau est incomplet.

Un autre intérêt de ce type d'étude est l'indentification de redondances dans le réseau, c'est-à-dire plusieurs modes élémentaires permettant la synthèse d'un même composé à partir d'un même substrat. Cela permet à l'organisme de continuer à synthétiser un composé même en l'absence de certaines réactions. On parle de plasticité ou de robustesse d'un réseau. Cela peut permettre d'expliquer la résistance de certains organismes à la délétion de certaines enzymes ou aux drogues [89,95,120]. Inversement, il est aussi possible de trouver des réactions importantes pour le réseau. La suppression d'une de ces réactions induit la suppression d'un ou plusieurs modes élémentaires du réseau. Pour les identifier, un bon moyen serait d'étudier la fréquence des réactions dans les modes élémentaires. Une réaction est vitale si elle est nécessaire à la synthèse d'au moins un composé vital.

Un des résultats fournis par Metatool, préalable à l'analyse des modes élémentaires, est la recherche de réactions opérant toujours ensembles dans un réseau. Biologiquement, identifier des réactions qui sont toujours effectuées ensembles permet de supposer que les gènes codant pour les enzymes de ces réactions peuvent être directement ou indirectement co-régulés [90,94].

Le calcul de ces modes qui est maintenant possible sur des réseaux de grande taille, voire même sur des génomes complets [34,108,109]. Le principal inconvénient des modes élémentaires réside dans leur interprétation et leur visualisation. Plusieurs millions de modes élémentaires peuvent être identifiés dans un réseau, et ces résultats ne peuvent pas être analysés manuellement. Il est donc nécessaire de trouver des méthodes d'analyse et de classification de ces modes élémentaires.

3 DÉVELOPPEMENT

Le développement informatique a constitué la majeure partie du travail de cette thèse. On peut séparer les développements effectués en deux parties : d'une part les applications web disponible sous la forme de sites internet et les applications dites « lourdes » nécessitant une installation sur les ordinateurs.

3.1 APPLICATIONS WEB VS. APPLICATIONS LOURDES

Le choix entre une application lourde et une application web se fait en fonction de l'utilisation du logiciel développé. Si une application nécessite une utilisation par plusieurs personnes avec partage des résultats, on préfèrera une application web. Si au contraire l'application nécessite des temps de réaction courts et des représentations graphiques poussées, on s'orientera plutôt vers des applications lourdes.

Ces règles ne sont pas immuables : par exemple les systèmes d'informations développés par la société Soluscience permettent d'échanger des informations via internet tout en restant une application lourde. Deux avantages principaux des applications lourdes contre les applications web : la plupart des applications lourdes sont utilisables sans accès internet et profitent au maximum des ressources matérielles de la machine sur laquelle elles sont installées. Les applications nécessitant des ressources graphiques importantes sont plutôt des applications lourdes. C'est par exemple le cas des éditeurs de graphes qui nécessitent des ressources graphiques et processeur importante, pour les algorithmes de dessin automatique de graphes notamment. Pourtant de nouvelles technologies comme WebGL apparaissent et pourraient permettre de développer ce type d'applications sous forme d'applications web utilisant les ressources locales via le navigateur web. Par rapport aux applications web, les applications lourdes sont plus complexes à déployer : il faut installer l'application sur toutes les machines et à mettre à jour.

Il ne faut pas confondre un site web et une application web. Un site web est un ensemble de pages liées entre elles. Une application web est un logiciel accessible et manipulable via un navigateur internet. Une application web, tout comme un site web, est placée sur un serveur auquel les utilisateurs de l'application se connectent pour l'utiliser. De nombreux utilisateurs peuvent donc utiliser une même application en même temps depuis n'importe où dans le monde. En général ils partagent une même base de données de résultats, ce qui facilite les échanges entre eux.

Pour nos développements, nous avons choisi de créer des applications web pour les logiciels ou l'échange de données était le plus important : GeneProm, BioNMR et myKegg. Pour le logiciel permettant de représenter des voies métaboliques sur des graphes, nous avons choisi de développer une application lourde sous forme d'un plugin pour le logiciel VANTED.

Remarque : le logiciel VANTED est disponible aussi sous forme « Java Web Start ». Ce type d'application ne peut pas être considéré comme une application web puisque pour s'exécuter il se télécharge en intégralité en local avant d'être exécuté.

3.2 APPLICATIONS WEB

Un part importante du développement d'une application est le choix des langages de programmation utilisés. Les applications web développées au cours de cette thèse sont écrites en Ruby en utilisant son *framework* Ruby on Rails. Le système de gestion de base de données choisi est MySQL. Le choix du système de gestion de base de données importe peu en Ruby on Rails puisqu'ils sont interchangeables.

3.2.1 Ruby et Ruby on Rails

Le Ruby est un langage de programmation *interprété*, orienté *objet*. Le côté interprété de Ruby est particulièrement intéressant quand on veut tester de petits morceaux de code puisqu'aucune étape de compilation n'est nécessaire et qu'il est possible de taper le code directement dans la console Ruby. Le concepteur du langage, Yukihiro Matsumoto, a voulu rassembler dans un même langage les points forts de Perl, Smalltalk, Ada et Lisp en voulant créer un langage *le plus naturel possible*. Russ Olsen affirme même que « si la programmation en Ruby est si agréable, c'est que le langage tente de s'effacer » [136]. Le paradigme de base de Ruby est que *tout est objet* (données et types) donc *toute fonction est une méthode* et *toute variable est une référence à un objet*. Toutefois cette structure n'impose pas au programmeur de devoir toujours déclarer ses objets et Ruby peut être utilisé comme un langage de script. Ainsi, le code de la figure 46 illustre le côté objet du langage où le chiffre « 5 » est considéré comme un objet de type *Numeric* sur lequel on appelle la méthode « *times* ». Ce code peut être exécuté directement dans un interpréteur Ruby.

De nombreuses bibliothèques sont disponibles pour Ruby, on parle de *gem*. L'installation de ces gems est effectuée via un système de gestion de gem fourni avec Ruby. Ce système permet aussi de gérer les mises à jour de toutes les gems installées sur un serveur. Il existe une gem *BioRuby* qui ne permet pour l'instant que de faire de la manipulation de séquence et de la conversion de format d'échanges de données biologiques, mais son développement est très actif et de nouvelles fonctionnalités apparaissent régulièrement [41]. Certaines fonctionnalités de l'API KEGG sont accessibles en utilisant cette gem. Un défaut de Ruby est sa lenteur due à son interpréteur. Différentes alternatives existent pour pallier à ce problème : l'inclusion des parties les plus lentes du code dans des librairies en C ou l'utilisation de JRuby qui est une implémentation de Ruby interprété par la machine virtuelle Java avec la possibilité d'inclure du code Java dans une application Ruby.

```
5.times { print "Hello world! " }
```

Figure 46 : Exemple de code Ruby
Ce code permet d'afficher cinq fois de suite le texte « Hello world ! » dans une console

Le langage Ruby est extrêmement flexible. Il est possible de modifier le cœur du système durant l'exécution d'un programme. On peut redéfinir toutes les classes du langage et surcharger toutes les méthodes, ainsi il est possible de rajouter ou de supprimer des fonctionnalités du langage. On ne peut pas faire d'héritage multiple en Ruby. Par contre les classes peuvent inclure plusieurs modules (un module

étant un groupe de méthodes). Cela permet de retrouver des comportements communs dans plusieurs classes sans dupliquer le code.

Ce qui a fait le succès de Ruby est son *framework* Ruby on Rails (également appelé Rails ou RoR). Un framework (ou cadre d'application) est, en programmation, un ensemble de composants logiciels structurels qui servent à créer les fondations ainsi que les grandes lignes de tout ou d'une partie d'un logiciel. Ruby on Rails permet la construction d'applications (surtout d'applications web) très rapidement car il impose une structure de développement répondant au *design pattern* (modèle de conception) Modèle Vue Contrôleur (MVC).

Les *design patterns* sont issus de la réflexion de Erich Gamma, Richard Helm, Ralf Johnson et John Vlissides (plus connus sous le nom de *Gang of Four* ou *GoF*) qui ont isolés des solutions optimales aux problèmes courants en programmation. Ils ont nommé, décrit et expliqué ces solutions et ont expliqués dans quel cadre les appliquer [134]. Ils ont identifié vingt-trois modèles de conception clés pour développer des programmes orientés objets. On retrouve ces exemples d'implémentation de ces *design patterns* dans de nombreux langages, dont le Ruby [136].

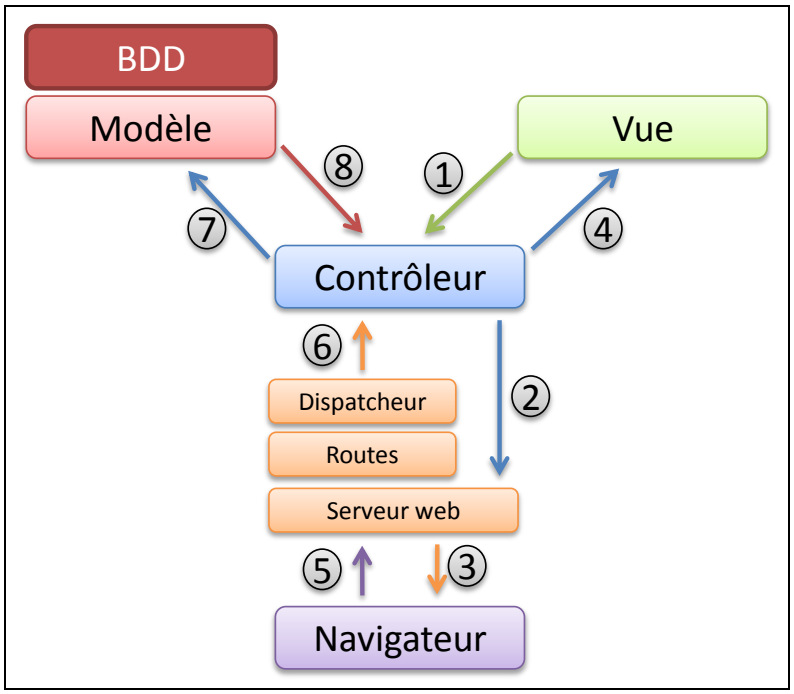


Figure 47 : Schéma du modèle MVC de RoR
Le schéma du MVC tel que décrit dans la littérature correspond aux cadres « Modèle », « Vue » et « Contrôleur ». La partie inférieure de la figure correspond à une partie propre à RoR et aux applications web. Le lien entre la base de données et le modèle est optionnel.

Le *design pattern* MVC tel qu'il est implémenté en Ruby on Rails est représenté dans la figure 47. Dans ce *design pattern*, le modèle regroupe tous les objets assurant la gestion des données. La structuration des données est déterminée par RoR en utilisant la structure de la base de données via un *mapping objet-relationnel* (cf 3.2.1.1). Le modèle représente les données de l'application, en RoR il est intimement lié à la base de données. La vue représente l'interface avec l'utilisateur et n'effectue aucun

traitement. Dans le MVC classique, il existe un lien entre la vue et le modèle. En RoR, les données du modèle à afficher par la vue sont fournies par le contrôleur. Le contrôleur gère l'interface entre le modèle et la vue (et donc l'utilisateur).

La vue est, par définition, la seule partie du programme accessible à l'utilisateur via le navigateur internet. La vue est fournie au navigateur via le serveur web (1, 2 et 3). La vue n'effectue aucun traitement, elle affiche les données fournies par le modèle via le contrôleur (4). Le navigateur envoie les actions de l'utilisateur au serveur web (5) qui les transmet au contrôleur (6). Les actions de l'utilisateur sont interprétées par le contrôleur qui va agir sur le modèle (7) et le cas échéant, modifier les données. Le contrôleur interroge le modèle (8) pour construire les vues demandées par l'utilisateur (4). À chaque vue correspond une route différente générée par le serveur web en fonction des actions de l'utilisateur. Le dispatcheur interprète ces routes pour envoyer les bonnes instructions au contrôleur.

L'utilisation du MVC permet non seulement de structurer le projet mais aussi évite d'écrire beaucoup de code « inutile », c'est-à-dire du code inutilement répété puisque la très grande majorité du moteur de l'application est inclus dans RoR. Cela permet de focaliser le développement sur des points plus importants comme la conception des applications et la rédaction de tests automatisés permettant de vérifier à tout moment le bon fonctionnement de l'application.

3.2.1.1 Mapping objet-relationnel

```
Ruby:   Author.create :lastname => "Durand", :forename => "Jean"

SQL:    INSERT INTO authors (lastname, forename) VALUES ('Durand', 'Jean');
_____________________________

Ruby:   a = Author.find_by_lastname("Durand")

SQL:    SELECT * FROM authors WHERE lastname = 'Durand' LIMIT 1;

Ruby:   a.forename

        => "Jean"
```

Figure 48 : Exemples d'utilisation de l'ORM Active Record de RoR pour interroger une base de données
Le premier exemple de code permet d'ajouter une nouvelle personne (Jean Durand) dans les auteurs de la base de données, la deuxième ligne correspond à la requête SQL qui a été générée pour insérer le nouvel enregistrement dans la base. Cette requête est invisible autant pour le programmeur que pour l'utilisateur. La deuxième instruction permet de trouver la première entrée de la table « authors » ayant pour nom de famille « Durand ». Dans les deux cas le code Ruby est bien plus simple à écrire et à comprendre que le code SQL auquel il correspond. La dernière instruction correspond à l'accès à l'attribut « forename » de l'objet retourné lors de la précédente instruction. Ces exemples sont tirés de l'application GeneProm.

Un ORM (*Object Relational Mapping* ou Mapping objet-relationnel) est une technique de programmation qui reproduit la structure d'une base de données relationnelle dans un langage objet en créant des correspondances entre les objets du programme et les tables de la base de données. L'implémentation en Ruby pour RoR répond aux standards dictés par le *design pattern* « active record » (enregistrement actif). Le nom de la couche l'ORM utilisé dans Rails est ActiveRecord. Ce type de

structure permet de s'affranchir, la plupart du temps, du SGBD pour ne travailler qu'avec des objets ce qui simplifie beaucoup le code (cf. figure 48).

C'est aussi ActiveRecord qui gère les liaisons entre les objets (et donc les enregistrements dans la base de données). Pour cela il déclare les relations entre les modèles dans les classes correspondant aux modèles. Pour cela, différentes relations sont disponibles : *has_many* (relation 1-n), *has_and_belongs_to_many* (relation n-n) et *has_one* (relation 1-1). La déclaration d'une relation *has_many* dans un modèle implique la déclaration d'une relation *belongs_to* dans son opposé. L'utilisation de ces relations est illustrée dans la figure 49.

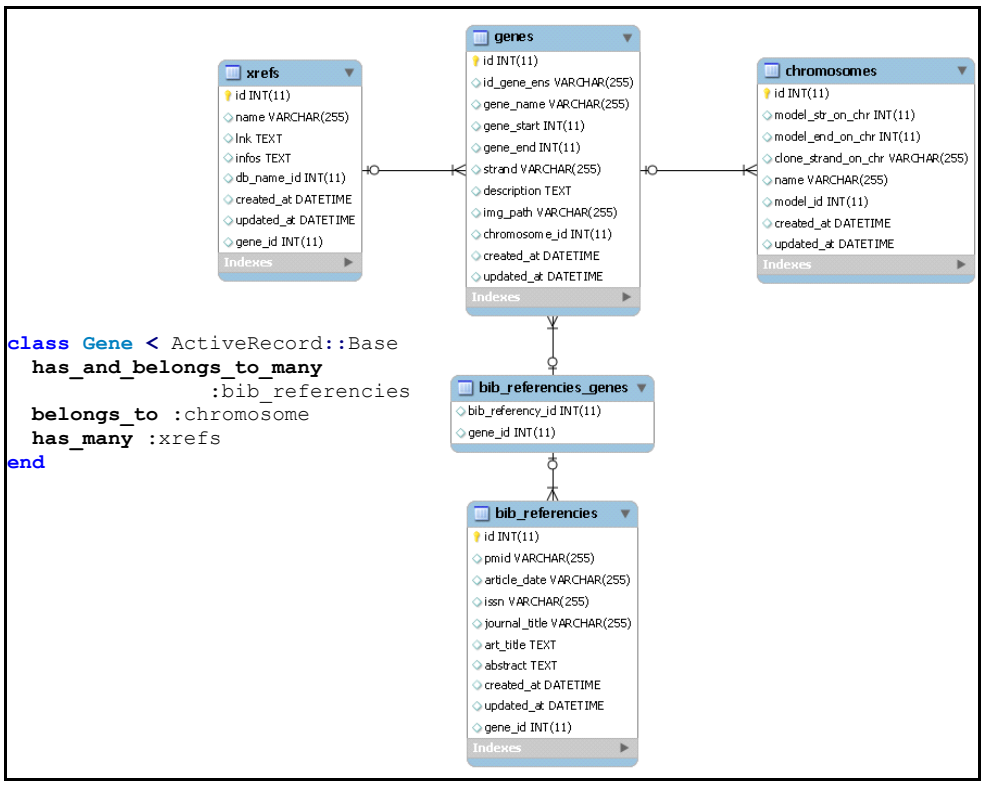


Figure 49 : Correspondance entre les liens ActiveRecord et les liens des tables de la base de données
Le code de la classe Gene contient l'ensemble des déclarations nécessaires à l'établissement des liens entre les tables de la base de données. La relation « has_and_belongs_to_many » correspond à une relation de cardinalité n-n entre les tables « genes » et « bib_referencies » permettant d'associer des références bibliographiques à des gènes. Un gène pouvant avoir plusieurs références associées et un article pouvant référencer plusieurs gènes. La relation « belongs_to :chromosome» déclare qu'un gène est sur un chromosome unique. Dans le modèle du chromosome, la déclaration inverse est présente c'est-à-dire : qu'un chromosome « has_many :genes ». La relation « has_many :xrefs » permet de déclarer les références croisées d'un gène dans d'autres bases de données. Ce schéma est tiré du schéma général de GeneProm. Il faut aussi noter que les attributs des objets (Gene) ne sont jamais explicitement décrits dans les modèles. Ils sont implicitement définis dans la base de données. Il est bien sûr possible de définir d'autres attributs ou de surcharger les attributs par défaut de la base dans les modèles.

ActiveRecord permet aussi de créer les modèles et les tables correspondantes ainsi que les modifications apportées au modèle de données (à la base de données) via un système appelé *migration*. Une migration dérive de la classe ActiveRecord::Migration. Elle définit deux méthodes : une méthode « up », qui permet d'ajouter une modification à la base de données, et une méthode « down », qui permet de retirer cette modification de la base de données. ActiveRecord fournit des méthodes qui permettent de définir des tâches de modification de la structure de la base de données : *create_table* (création d'une table avec des colonnes) et son contraire *drop_table*, *rename_table* (renommer une table), *add_column* (ajouter une colonne d'un type donné à une table donnée), et son contraire *remove_column*, …

```
1   class CreateGenes < ActiveRecord::Migration
2       def self.up
3           create_table :genes do |t|
4               t.string :id_gene_ens
5               t.string :gene_name
6               t.integer :gene_start
7               t.integer :gene_end
8               t.integer :d_gstr_mstr
9               t.integer :d_gstr_mend
10              t.string :strand
11              t.string :description
12              t.string :img_path
13              t.integer :chromosome_id
14
15              t.timestamps
16          end
17      end
18
19      def self.down
20          drop_table :genes
21      end
22  end
```

Figure 50 : Exemple de migration ActiveRecord
Deux méthodes sont définies dans une migration Activerecord : « up » (lignes 2-17) et « down » (19-21). La méthode « up » permet d'effectuer les changements voulus à la base de données, la méthode « down » permet de les annuler. Cette migration correspond à la création de la table « genes » du logiciel GeneProm. Les lignes 4-13 correspondent à la définition des colonnes de la table et de leur type. La ligne 15 permet d'ajouter deux colonnes supplémentaires donnant des informations sur la date de création et de modification de chaque enregistrement. Un autre champ est ajouté par défaut : le champ « id » qui correspond à la clé primaire de la table.

Un autre avantage majeur d'ActiveRecord est qu'il permet de changer de système de gestion de base de données de manière très simple puisqu'il suffit la plupart du temps de changer un mot dans un fichier de configuration. Les systèmes de gestion de base de données supportés dans les dernières versions du langage sont : MySQL, PostgreSQL, SQLite, SQL Server, Sybase, et Oracle.

En conclusion, le RoR et plus particulièrement son ORM ActiveRecord permettent un développement rapide et flexible d'applications web liées à des bases de données. Ce qui le rend particulièrement adapté aux techniques de développement agile. Tim Bray, directeur du département des technologies web chez Sun Microsystems jusqu'en 2010 a même déclaré que comparé au Java et au PHP le Ruby on Rails permettait de développer plus vite des applications plus faciles à maintenir que les deux autres.

3.2.1.2 Développement agile

Les méthodes agiles sont des méthodes de développement impliquent au maximum le demandeur pour permettre un produit final plus proche de ses attentes. Les méthodes agiles mettant l'accent sur l'adaptabilité du développement aux besoins des clients est particulièrement adapté au développement impliquant des phases de recherche importantes qui vont souvent bouleverser le fonctionnement de l'application et donc le code. Cela se traduit par un remaniement (*refactoring*) régulier du code.

Nous ne pouvons pas dire qu'on a véritablement appliqué l'intégralité des préceptes des méthodes agiles puisque la plupart du temps le client et le développeur constituaient une seule et même personne. Certaines règles ont tout de même été suivies pour le développement des applications web (dans une moindre mesure dans celui de MPSA). Beaucoup de ces règles sont issues de la méthode de développement agile *eXtreme Programming* :

- Présentation régulière des applications afin de vérifier que les fonctionnalités développées sont celles attendues et définition des changements à apporter ou des nouvelles fonctionnalités
- Remaniement du code pour qu'il réponde aux nouvelles demandes
- Écriture de tests automatiques permettant de s'assurer de la stabilité des nouvelles fonctionnalités en cas de nouveau remaniement.
- Utilisation d'un logiciel d'intégration continue couplé à un système de gestion de version. Un logiciel de gestion de version est un logiciel permettant de stocker un ensemble de fichiers sur un serveur en conservant la chronologie de toutes les modifications faites sur ces fichiers. Un système d'intégration continue est un système vérifiant régulièrement les mises à jour d'un logiciel et qui exécute les tests automatiques du logiciel pour toute nouvelle version. Cela permet de prouver que l'application fonctionne toujours de la même manière après un changement et en cas de bug, de savoir depuis quand le logiciel ne marche plus.
- Utilisation de métriques permettant de savoir quelles sont les parties testées et non testées du logiciel, les parties les plus complexes du code (vraisemblablement trop complexes, donc simplifiables et susceptibles de présenter des bugs) et de trouver le code « mort », c'est-à-dire le code jamais exécuté. Ces deux dernières mesures permettent d'écrire le code le plus simple possible.

3.2.2 GeneProm

GeneProm est un logiciel permettant d'analyser les résultats issus de Genomatix. L'algorithme de l'étude promotologique menée en utilisant Genomatix et GeneProm sera décrit dans le chapitre 4. Nous nous intéresserons ici plus particulièrement au logiciel GeneProm : ses fonctionnalités, sa structure. Comme dit en introduction, GeneProm est issu du système d'information HumProm, développé en collaboration avec Sébastien Duplant de la société Soluscience. Les raisons du passage d'une version

lourde en C++ à une l'application web GeneProm sont explicitées dans l'introduction de cette thèse. Nous nous intéresserons ici plus précisément à la structure du logiciel et à ses fonctionnalités.

HumProm permettait de réaliser l'étude Genomatix complète directement via son interface mais Genomatix a interdit cette pratique fin 2008. Il a donc fallu trouver un moyen d'exporter les résultats des études Genomatix et développer un nouveau logiciel permettant d'analyser ces résultats. Sur Genomatix nous construisons des modèles de promoteur pour des gènes d'intérêt et nous recherchons toutes les occurrences possibles de ce modèle sur le génome humain. Il nous faut récupérer deux informations : la structure du modèle recherché et les positions des occurrences du modèle trouvées. Après négociation avec les développeurs de Genomatix, nous avons conclu que les résultats de la recherche pouvaient uniquement être extraits sous forme de feuille Excel contenant la liste des clones sur lesquels ont été trouvés les modèles avec les coordonnées de début et de fin et le sens du modèle sur le clone (cf. figure 51). La structure du modèle est récupérée à partir d'une page HTML de Genomatix qu'il faut enregistrer et dans laquelle les développeurs ont inclus des champs cachés stables (ils se sont engagés à les laisser même en cas de remaniement du site et de ne pas changer la structure de ces champs) décrivant le modèle.

	A	B	C	D	E	F	G	H	I
1	**Seq. name**	**AC no.**	**Gene**	**GeneID**	**Model**	**Start pos.**	**End pos**	**Strand**	**Score**
2	AC087435	AC087435			ANT1_model_MGNfam_optbis_ATG	112868	112491	-	100
3	AC087435	AC087435			ANT1_model_MGNfam_optbis_ATG	112868	112396	-	100
4	AC087435	AC087435			ANT1_model_MGNfam_optbis_ATG	112868	112360	-	100
5	AC087435	AC087435			ANT1_model_MGNfam_optbis_ATG	112868	112336	-	100
6	AC087473	AC087473			ANT1_model_MGNfam_optbis_ATG	64415	64037	-	100
7	AC087473	AC087473			ANT1_model_MGNfam_optbis_ATG	64415	63897	-	100

Figure 51 : Extrait d'une feuille Excel d'export des résultats de Genomatix.
Les résultats de la recherche des modèles de promoteurs sur Genomatix sont exportés sous forme de feuille Excel. Chaque ligne du fichier correspond à la description d'une des positions trouvées du modèle. Dans notre cas, les deux premières colonnes sont toujours les mêmes et contiennent les numéros d'accession des clones GenBank (rev. 180 à la date d'écriture de ce manuscrit). Genomatix ne donne jamais d'information sur les gènes dans notre recherche donc les colonnes « Gene » et « GeneID » sont toujours vides. La colonne « Model » contient un nom, unique, donné au modèle recherché. Les colonnes « Start pos. », « End pos » et « Strand » correspondent respectivement aux positions de début et de fin et au sens du modèle sur le clone. La colonne « Score » correspond au pourcentage de matrices trouvées pour l'occurrence du modèle par rapport au nombre total de matrice dans le modèle à rechercher. Il est possible de fixer le seuil minimal de ce score. Ici toutes les matrices du modèle recherché doivent être retrouvées.

GeneProm va lire cette feuille Excel et, pour chaque position du modèle retournée par Genomatix, va rechercher s'il existe un gène à proximité du modèle trouvé et dans le même sens. Les résultats de GeneProm consistent donc en une liste de gènes pouvant être sous le contrôle du modèle de promoteur fourni. La recherche de ces gènes correspond à une succession de changements de référentiels de coordonnées et de calculs de distances et de positions permettant de vérifier que l'occurrence du modèle trouvée est bien dans une région pouvant correspondre à un promoteur et qu'elle est dans le même sens. Pour mieux comprendre le fonctionnement de l'algorithme, il est nécessaire de bien comprendre comment les données sont structurées dans le programme.

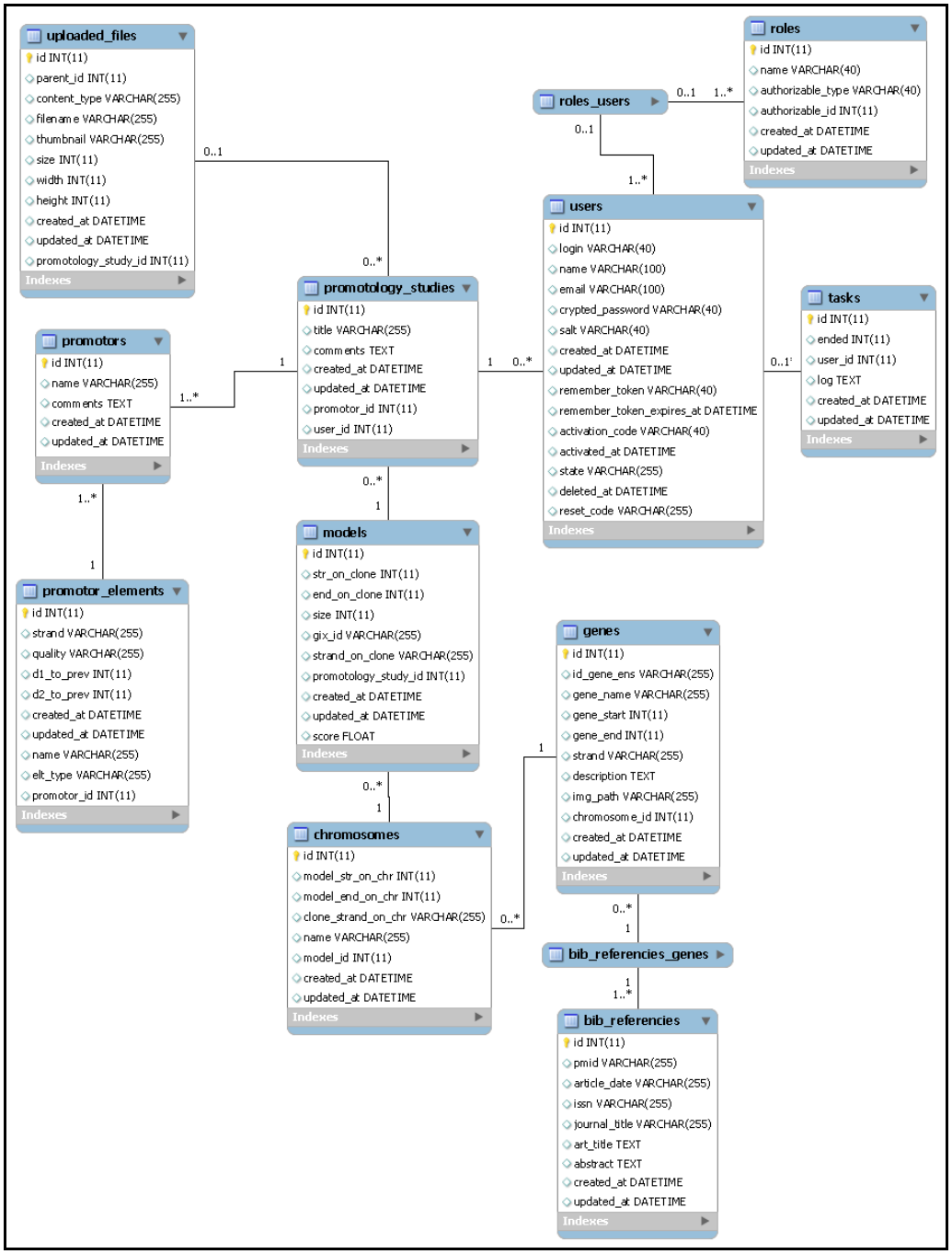


Figure 52 : Schéma entité-relation simplifié de la base de données du logiciel GeneProm
Ce schéma respecte le standard de représentation « Enhanced Entity-Relationship Model», noté modèle EER. Les cadres représentent les tables de la base de données. Les lignes dans les cadres représentent les colonnes des tables (nom et type). La clé primaire (colonne de chaque table assurant l'unicité de chaque enregistrement) est la colonne « id ». Les notations à chaque extrémité des relations correspondent aux cardinalités des liens. Par exemple, entre les tables « chromosomes » et « genes » : le symbole « 0.. » indique qu'on peut avoir identifié 0 à n gènes sur le chromosome, le symbole « 1 » indique qu'un gène appartient à un chromosome unique. Les tables dont les colonnes ne sont pas spécifiées comme « roles_users » sont des tables particulières appelées tables de jointure. Ces tables de jointures permettent de créer des relations n-n entre deux objets.*

3.2.2.1 Base de données

La base de données comporte 34 tables, mais pour plus de simplicité ne seront présentées ici que les tables essentielles à la compréhension de la structure du programme. Pour simplifier cette structure, la partie de la base de données servant à enregistrer les références externes (références des gènes vers d'autres bases de données) a été omise dans le diagramme entité–relation de la base de données présenté en figure 52.

La table centrale du modèle est la table correspondant à l'étude promotologique (*promotology_studies*). Une étude promotologique appartient à un utilisateur (*user*). La plupart des champs de la table utilisateurs correspondent aux champs nécessaires à l'identification par mot de passe sur le site web. Les utilisateurs ont des rôles différents. En fonction de leur rôle, les utilisateurs peuvent voir certaines pages et pas d'autres ou faire certaines actions. Par exemple, seuls les utilisateurs ayant le rôle « superadministrateur » peuvent supprimer des études promotologiques. À chaque fois qu'un utilisateur crée une étude promotologique, ou quand il fait certaines actions sur le site, une trace du déroulement de ces actions est conservé dans la table « task ». Cette table permet aussi de savoir si la création de l'étude promotologique est achevée ou pas. Les fichiers xls et html utilisés pour la création des études sont liés à celle-ci via la table « uploaded_files ». À chaque étude promotologique correspond un modèle de promoteur (tiré d'un fichier html fourni par l'utilisateur) enregistré dans la table « promotors ». Ce promoteur peut être commun à plusieurs études. Un promoteur est composé d'un ensemble d'éléments de régulation enregistrés dans la table « promotor_elements ». Les différentes occurrences du modèle fournies dans le fichier xls sont enregistrées dans la table « models ». Les coordonnées de ces modèles sont recherchées sur les chromosomes. La table « chromosomes » contient les enregistrements des modèles sur les chromosomes. Si le logiciel identifie un gène potentiellement sous le contrôle d'une des occurrences du modèle, il sera sauvegardé dans la table « genes ». De nombreuses tables correspondant à des références du gène dans d'autres bases de données sont liées à cette table. Ici ne sont figurées que les références bibliographiques des gènes dans la base de données *Medline*.

3.2.2.2 Algorithmes

L'algorithme de recherche des gènes potentiellement régulé par le modèle de promoteur se base sur l'interface de programmation (API) de la base de données EnsEMBL. L'utilisation de cette API impose de coder cet algorithme en Perl. La structure du modèle est construite en parallèle de cette recherche. Nous ne détaillerons ici que la partie concernant la recherche des gènes.

Comme vu dans la figure 51, les seules informations que fourni Genomatix sont les identifiants des clones sur lesquels sont les modèles, les coordonnées de ces modèles sur le clone et le sens du modèle sur le clone. La tâche principale à effectuer est de trouver les coordonnées du modèle sur le chromosome car c'est uniquement dans ce référentiel que l'on pourra rechercher les gènes potentiellement régulé par le modèle. L'algorithme de recherche va alors rechercher s'il existe un gène présent dans une zone de 2000 pb après le début du modèle (zone de recherche des gènes). La taille de cette zone de recherche correspond à une taille maximale qui permet de faire une présélection des gènes intéressants. Des filtres de distances supplémentaires sont disponibles sur les résultats d'une étude. Cette présélection est nécessaire car le temps de calcul d'une étude promotologique est relativement long (parfois plusieurs heures). Il est impensable de faire attendre l'utilisateur plusieurs heures à chaque fois qu'il change un paramètre du logiciel.

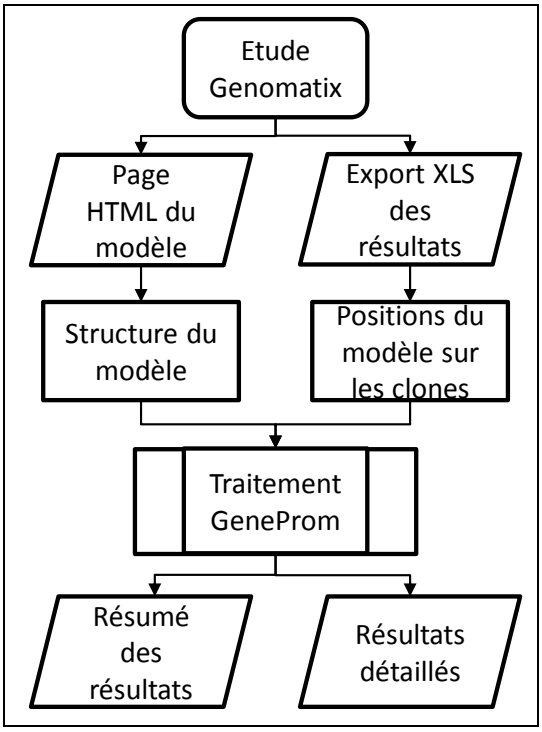


Figure 53 : Algorigramme de la procédure générale d'une étude promotologique sur GeneProm
En sortie d'étude Genomatix, deux fichiers permettent de créer une étude promotologique sur GeneProm : une description du modèle dans une page HTML et les positions des modèles trouvés sur le génome humain au format xls. Ces deux fichiers correspondent aux points d'entrée de la procédure de traitement de GeneProm. En fin d'analyse GeneProm permet d'afficher des pages de résultats résumées ou détaillées.

L'algorithme va dans un premier temps calculer les coordonnées de la zone de recherche des gènes à partir des 3 informations dont il dispose : le nom du clone, les coordonnées du modèle sur le clone et le sens du modèle sur le clone. Ces informations sont tirées de la feuille Excel qui est lue par un script Perl utilisant la bibliothèque *Spreadsheet::ParseExcel* qui permet de lire des feuilles Excel directement sans imposer à l'utilisateur de convertir ces feuilles en un format textuel. L'API EnsEMBL permet de trouver facilement le clone et de trouver ses coordonnées sur le chromosome :

```
1   my $clone = $slice_adaptor->fetch_by_region( 'clone', $clone_id );
2   my $projection = $clone->project("chromosome")->[0];
3   ( my $clone_start, my $clone_end, my $clone_on_chr ) = @{$projection};
```

Figure 54 : Code permettant de trouver un clone sur un chromosome en utilisant l'API EnsEMBL

La ligne 1 correspond à la recherche d'un clone ayant pour identifiant un chaine de caractère stockée dans la variable *$clone_id*. Ce clone est dans son propre référentiel, ses coordonnées de début et de fin correspondent respectivement à 0 et à sa taille. Les lignes 2 et 3 correspondent à la projection du clone dans le référentiel du chromosome et donc de trouver les coordonnées de ce clone sur le génome. Un

problème se pose avec l'API Perl d'EnsEMBL : les coordonnées de début et de fin des clones sont souvent fausses car EnsEMBL retire la partie 3' de la séquence qui chevauche le clone suivant. La partie 5' est aussi parfois tronquée. Ces modifications de séquences n'étaient pas présentes lors du développement de la première version de GeneProm. Les positions du modèle sur le clone renvoyées par Genomatix font référence à la séquence complète du clone. Nous avons donc conçu un script Perl permettant de retrouver les bonnes coordonnées du clone (et donc du modèle) en utilisant les séquences GenBank des clones qui sont complètes. Ce script fonctionne de la manière suivante :

- Récupération de la séquence du clone sur GenBank en utilisant l'API de l'EBI (ne fournit pas d'informations de coordonnées).
- Extraction d'une séquence courte en début et en fin de clone EnsEMBL (50nt) : les motifs de début et de fin du clone EnsEMBL. Les coordonnées (potentiellement erronées) de ces motifs sont connues.
- Les motifs sont alors recherchés sur la séquence GenBank du clone. Si les séquences GenBank et EnsEMBL sont les même, ces motifs correspondront respectivement aux 50 premiers et aux 50 derniers nucléotides de la séquence GenBank.
- Le nombre de bases entre le début du clone GenBank et le premier motif correspond au décalage 5' du clone et la distance entre le deuxième motif et la fin du clone GenBank correspond au décalage en 3'.
- À partir de ces deux décalages et du sens du clone sur le chromosome, les coordonnées réelles du clone sur le chromosome sont calculées.

Les coordonnées de la zone de recherche des gènes sont calculées à partir des informations de la feuille Excel et des coordonnées du clone sur le chromosome. Pour calculer ces coordonnées, il faut prendre en compte le sens du modèle sur le clone et du clone sur le chromosome. L'algorithme de la figure 55 permet de calculer ces coordonnées.

Il est ensuite possible de trouver les gènes dont la séquence est au moins partiellement comprise dans la zone de recherche des gènes en utilisant l'API EnsEMBL. Il est alors possible d'obtenir toutes les informations nécessaires sur ces gènes : identifiant EnsEMBL, nom, position (début, fin, sens) et description. Seuls les gènes étant dans le même sens que le modèle sont retenus. Les gènes dont le début (site d'initiation de transcription si disponible ou premier ATG du cadre de lecture sinon) n'est pas dans la zone de recherche des gènes sont écartés. Des schémas représentant le début du gène par rapport au modèle avec leur sens respectifs sont alors créés pour être par la suite affichés dans la section des résultats détaillés de l'étude sur GeneProm. Les résultats sont alors sérialisés dans un fichier XML qui est lu par GeneProm. Ce fichier contient toutes les informations nécessaires à la sauvegarde de l'étude dans la base de données.

```
1   si sens modèle / clone est +
2       si sens clone / chromosome est +
3           début_zrg = début_clone + début_modèle_gix
4           fin_zrg   = début_zrg + 2000
5       sinon
6           début_zrg = début_clone - début_modèle_gix
7           fin_zrg   = début_zrg - 2000
8   sinon
9       si sens clone / chromosome est +
10          début_zrg = début_clone + début_modèle_gix
11          fin_zrg   = début_zrg - 2000
12      sinon
13          début_zrg = début_clone - début_modèle_gix
14          fin_zrg   = début_zrg + 2000
```

Figure 55 : Code permettant de calculer les coordonnées génomiques de la zone de recherche des gènes
Les variables « début_zrg » et « fin_zrg » correspondent aux coordonnées de début et de fin de la zone de recherche des gènes. La variable « début_clone » correspond aux coordonnées du début du clone sur le chromosome. Enfin la variable « début_modèle_gix » correspond aux coordonnées du début du modèle sur le clone renvoyé par Genomatix.

Id	Name	Comments	Creation date		
262	ANT4HIFFbis	16 gènes dont 6 cohérents	2011/09/22	Abstract	Results
261	ANT4HIFFter	2 gènes dont 0 cohérent	2011/09/22	Abstract	Results
260	ANT4Prom1pintaATG	1 seul gène : ANT4	2011/09/22	Abstract	Results
259	ANT3phylopt	25 gènes dont 2 cohérents	2011/09/22	Abstract	Results
258	ANT4EBOXHIFFter	2 gènes dont 0 cohérents	2011/09/22	Abstract	Results
257	ANT4EBOXHIFF	2 gènes dont 2 cohérents	2011/09/22	Abstract	Results
256	ANT3phyloptter	4 gènes dont 0 cohérent	2011/09/22	Abstract	Results
255	ANT3phylopt2	2 gènes dont 0 cohérent	2011/09/22	Abstract	Results
254	ANT2260510part4	19 gènes positifs dont 10 cohérents	2011/09/22	Abstract	Results
253	ANT2260510part3	25 gènes positifs dont 10 cohérents	2011/09/21	Abstract	Results

Figure 56 : Présentation de la liste des études promotologiques de l'utilisateur courant
La colonne id correspond aux identifiants des études. Dans la deuxième colonne est affiché le nom que l'utilisateur a donné à l'étude. Dans la colonne suivante, sont affichés les commentaires liés à l'étude (si les commentaires sont trop longs ils sont abrégés sur cette page). La date de création de l'étude est ensuite rappelée. Via les boutons « abstract » et « results », l'utilisateur peut choisir de visionner une version abrégée des résultats de l'étude ou une page présentant la totalité des informations. La liste des études est présentée sur plusieurs pages, par groupes de dix, ordonnées par date de création décroissante.

3.2.2.3 Présentation des résultats

La présentation des résultats, et le lancement des études, se fait via une interface web. La première page que l'utilisateur rencontre en arrivant sur le site est la liste de ses propres études promotologiques (cf. figure 56).

Cette page permet d'afficher aussi les études en cours de création et les études dont la création n'a pas abouti à cause d'une erreur. Si l'utilisateur en a le droit, il peut effacer des études promotologiques depuis cette page ou aller sur l'interface d'administration des utilisateurs de GeneProm. Depuis cette page, l'utilisateur peut aller sur deux pages différentes de présentation des résultats : une page de résumé et une page de présentation des résultats complets. Sur ces deux pages, on retrouve un même entête qui rappelle le titre de l'étude, les commentaires (non abrégés) et la structure du modèle. On retrouve aussi une interface permettant de saisir deux paramètres de filtrage des résultats : une limite maximale de distance entre la fin du modèle et le début du gène (par défaut 500 nt) et une taille maximale de recouvrement entre le modèle et le début du gène (par défaut 200 nt). Cf. figure 57.

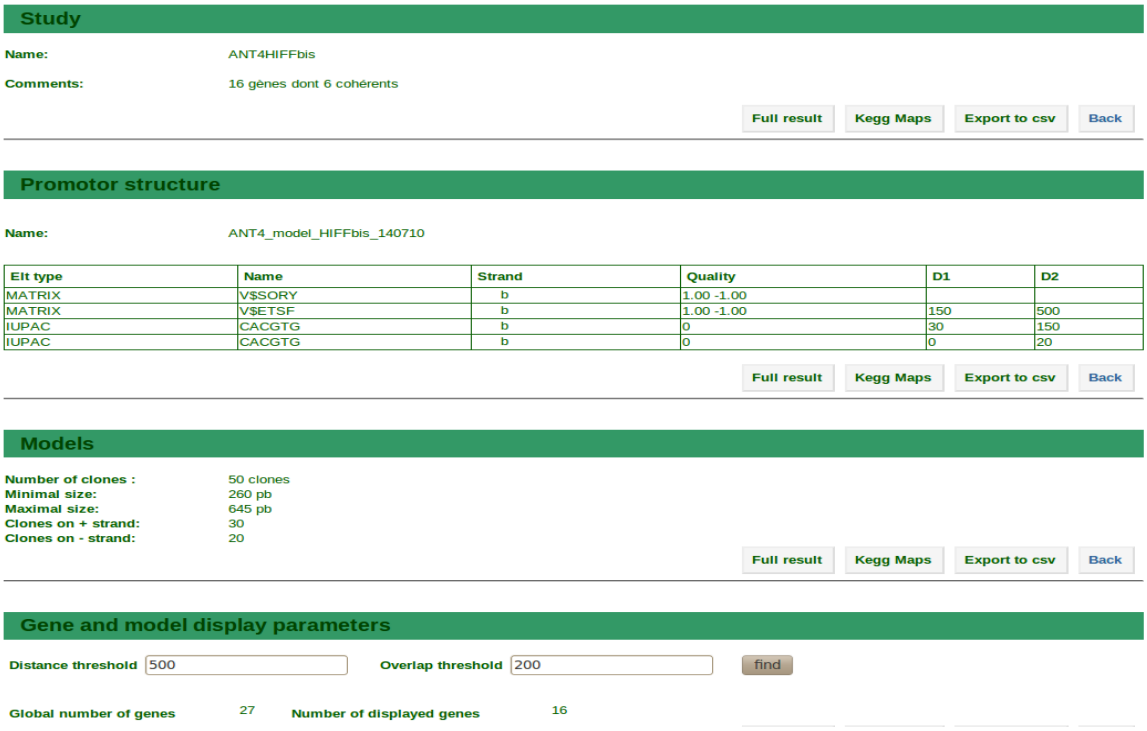


Figure 57 : Entête commun à toutes les pages de présentation des résultats de GeneProm
La première partie de l'entête correspond aux informations saisies pour l'étude : titre et commentaires, ensuite la structure du modèle de promoteur est présentée. La section suivante présente des statistiques simples sur les résultats de l'étude : nombre de clones positifs trouvés (clones porteurs d'un modèle pour lequel un gène a été identifié), la taille du modèle le plus petit et celle du plus grand (parmi les modèles positifs) et enfin le nombre de clones positifs en sens + et en sens -. La dernière section correspond à l'interface de saisie des paramètres de distance maximale et de recouvrement. Dans cette étude, 27 gènes positifs ont été trouvés mais seulement 16 répondent aux critères entrés.

La page de résumé des résultats correspond à un tableau contenant la liste des clones positifs, le début et la fin du modèle sur le clone, le sens du modèle sur le clone, la taille du modèle, la qualité du

modèle (pourcentage de matrice du modèle entré retrouvé dans le modèle courant) et le nom du gène potentiellement sous le contrôle du modèle. Cf. figure 58.

Name	Start	End	Strand	Size	Quality	Gene
AC048338	63662	64050	+	388	100.0	VPS33A
AC048338	63662	64050	+	388	100.0	RP11-512M8.5
AC023855	65267	64819	-	448	100.0	G6PC3
AC023855	65133	64819	-	314	100.0	G6PC3
AC020929	2255	2833	+	578	100.0	SPTBN4
AC012354	111836	112096	+	260	100.0	AC012354.6
AC009779	21667	21149	-	518	100.0	BLOC1S1
AC009779	21667	21149	-	518	100.0	RP11-644F5.10
AC009779	20851	21170	+	319	100.0	ITGA7
AC008772	102503	102037	-	466	100.0	CTD-2016O11.1
AL590543	9592	9992	+	400	100.0	C6orf211
AL590543	9444	9992	+	548	100.0	C6orf211
AL590543	9347	9992	+	645	100.0	C6orf211
AL354707	36005	36580	+	575	100.0	RP11-390F4.10
AL136221	88357	88913	+	556	100.0	LAMP1
AL136221	88357	88896	+	539	100.0	LAMP1
AL136221	88357	88879	+	522	100.0	LAMP1
AL050325	107760	107312	-	448	100.0	RSPO4
AL008723	112070	112379	+	309	100.0	CPSF1P1
AC124312	119415	118862	-	553	100.0	SNORD108
AC093591	33646	34055	+	409	100.0	SLC25A31
AC093591	33646	34041	+	395	100.0	SLC25A31
AC093591	33646	34027	+	381	100.0	SLC25A31
AC093591	33624	34055	+	431	100.0	SLC25A31
AC093591	33624	34041	+	417	100.0	SLC25A31
AC093591	33624	34027	+	403	100.0	SLC25A31
AC093591	33569	34055	+	486	100.0	SLC25A31
AC093591	33569	34041	+	472	100.0	SLC25A31
AC093591	33569	34027	+	458	100.0	SLC25A31
AC093591	33497	34055	+	558	100.0	SLC25A31
AC093591	33497	34041	+	544	100.0	SLC25A31
AC093591	33497	34027	+	530	100.0	SLC25A31

Figure 58 : Présentation résumée des résultats d'une étude promotologique
Les colonnes correspondent à : l'identifiant des clones, début du modèle sur les clones, fin du modèle sur les clones, sens du modèle sur les clones, taille du modèle, qualité du modèle, et nom du gène proche du modèle respectant les paramètres de filtre imposés par l'utilisateur.

Dans la page de résumé, les noms des gènes et des clones sont cliquables et permettent de se rendre directement au résultat détaillé. La figure 59 présente la représentation détaillée de deux gènes de l'étude : VPS33A et SLC25A31.

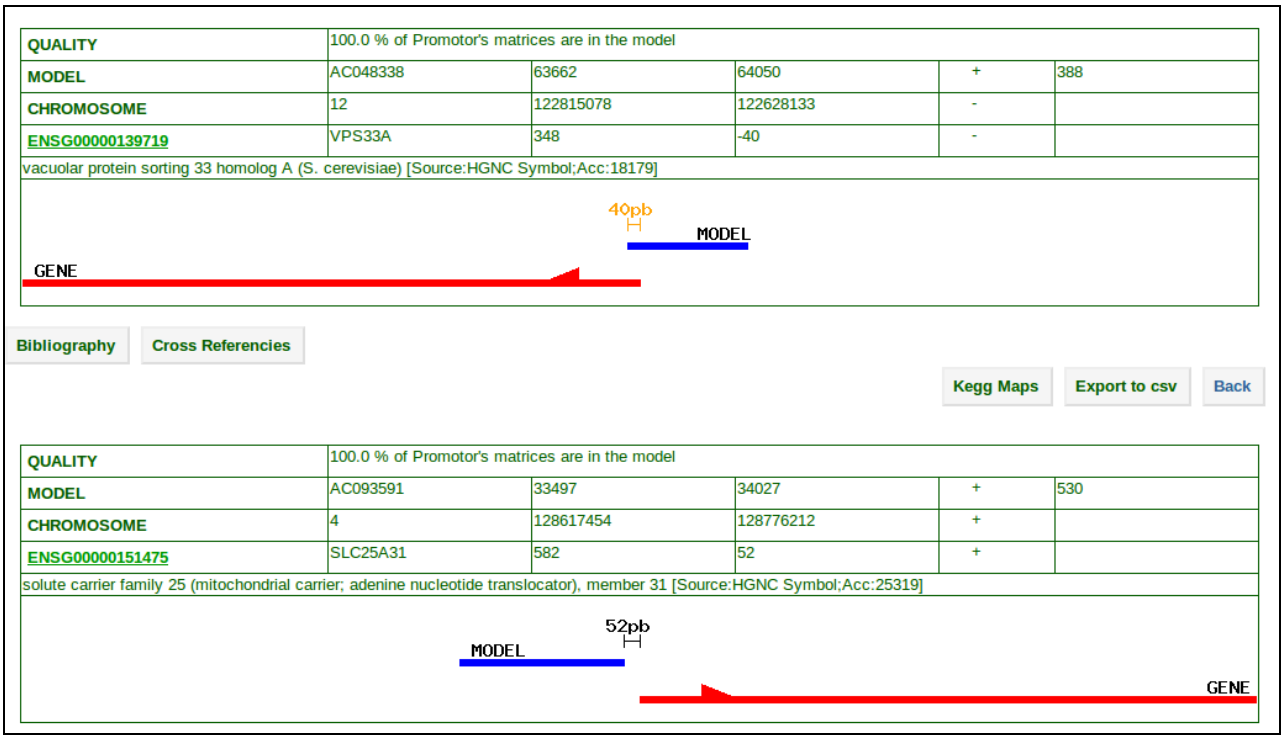


Figure 59 : Extrait de la page de représentation détaillée des résultats.
Chaque résultat est présenté sous forme d'un tableau suivi d'un schéma reprenant les informations de position du modèle par rapport au gène. Sous chaque tableau des liens vers les références du gène vers d'autres bases de données et vers des articles (s'ils ont été trouvés dans Medline) sont disponibles. Les liens « Kegg Maps » et « Export to csv » sont générique à l'étude globale. Ils correspondent respectivement à la liste des cartes KEGG ou les gènes de l'étude sont présents et à l'export de la liste des gènes au format csv interprétable par MPSA. Le bouton « Back » permet de retourner à la liste des études.

Les informations présentées dans les tableaux sont :

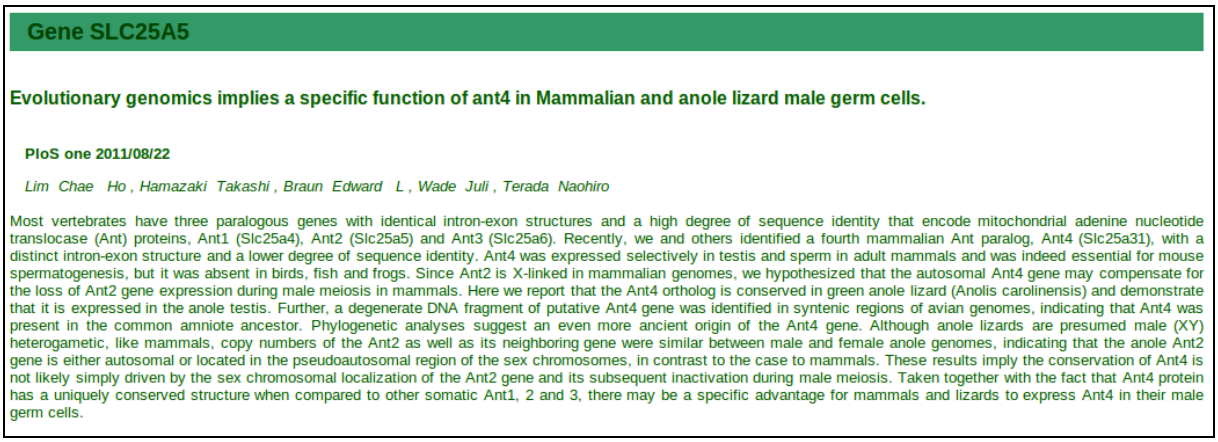
- Sur la première ligne : la qualité du modèle
- Sur la deuxième ligne : les informations du modèle : le nom du clone sur lequel il a été identifié, les coordonnées de début et de fin, le sens du modèle sur le clone et la taille du modèle
- Sur la troisième ligne : les informations relatives au clone sur le chromosome : le numéro du chromosome, le début et la fin du clone sur le chromosome et le sens du modèle sur le clone projeté sur le chromosome
- Sur la quatrième ligne : les informations relatives au gène : L'identifiant EnsEMBL du gène (lien vers la page EnsEMBL du gène), l'identifiant HGNC du gène [117], la distance entre le début du modèle et le début du gène, la distance entre la fin du modèle et le début du gène et le sens du gène sur le chromosome.
- Sur la cinquième ligne est figurée la description de la fonction du gène fournie par EnsEMBL (si disponible)

- Sur la sixième ligne le schéma de la position du modèle par rapport au gène est affiché. Le modèle est en bleu, le gène en rouge. Le sens du gène est figuré par une flèche sur le gène. Le sens du modèle est toujours le même que celui du gène. La taille du modèle est figurée à l'échelle (la figure complète représentant 4000pb). La distance entre la fin du modèle et le début du gène, qui est la distance examinée par les filtres, est aussi figurée à l'échelle. En cas de recouvrement du gène et du modèle, la distance est affichée en orange, sinon elle est en noir.

Pour chaque gène il est possible d'afficher les références vers d'autres bases de données, les principales étant : UniGene [177], Gene Ontology [3], GeneNames [117], KEGG [158], GeneCards [99], EMBL [178] et Uniprot [50,125]. Ces références sont obtenues en partie via l'API EnsEMBL et en partie via l'API KEGG.

Figure 60 : liste des références croisées trouvées pour le gène ANT2 (SLC25A5)
Le nom des bases de données est à gauche, les liens à droite amènent l'utilisateur directement aux pages correspondant au gène sur les différentes bases de données. Seules les bases de données pour lesquelles un enregistrement a été trouvé sont présentes sur cette page, la liste des bases de données varie donc d'un gène à l'autre.

Un autre type de lien vers une autre base de données est disponible : la base Medline. L'API du NCBI est utilisée pour construire ces liens vers les enregistrements PubMed correspondant.



Gene SLC25A5

Evolutionary genomics implies a specific function of ant4 in Mammalian and anole lizard male germ cells.

PloS one 2011/08/22

Lim Chae Ho , Hamazaki Takashi , Braun Edward L , Wade Juli , Terada Naohiro

Most vertebrates have three paralogous genes with identical intron-exon structures and a high degree of sequence identity that encode mitochondrial adenine nucleotide translocase (Ant) proteins, Ant1 (Slc25a4), Ant2 (Slc25a5) and Ant3 (Slc25a6). Recently, we and others identified a fourth mammalian Ant paralog, Ant4 (Slc25a31), with a distinct intron-exon structure and a lower degree of sequence identity. Ant4 was expressed selectively in testis and sperm in adult mammals and was indeed essential for mouse spermatogenesis, but it was absent in birds, fish and frogs. Since Ant2 is X-linked in mammalian genomes, we hypothesized that the autosomal Ant4 gene may compensate for the loss of Ant2 gene expression during male meiosis in mammals. Here we report that the Ant4 ortholog is conserved in green anole lizard (Anolis carolinensis) and demonstrate that it is expressed in the anole testis. Further, a degenerate DNA fragment of putative Ant4 gene was identified in syntenic regions of avian genomes, indicating that Ant4 was present in the common amniote ancestor. Phylogenetic analyses suggest an even more ancient origin of the Ant4 gene. Although anole lizards are presumed male (XY) heterogametic, like mammals, copy numbers of the Ant2 as well as its neighboring gene were similar between male and female anole genomes, indicating that the anole Ant2 gene is either autosomal or located in the pseudoautosomal region of the sex chromosomes, in contrast to the case to mammals. These results imply the conservation of Ant4 is not likely simply driven by the sex chromosomal localization of the Ant2 gene and its subsequent inactivation during male meiosis. Taken together with the fact that Ant4 protein has a uniquely conserved structure when compared to other somatic Ant1, 2 and 3, there may be a specific advantage for mammals and lizards to express Ant4 in their male germ cells.

Figure 61 : Extrait de la page de présentation des publications pour le gène ANT2 (SLC25A5)
Pour chaque publication le titre, les auteurs et le résumé sont fournis.

Le dernier lien fourni est un lien vers KEGG. Les produits des gènes de toute l'étude sont recherchés sur KEGG et situés sur les cartes. La liste des cartes KEGG est fournie pour l'étude avec les gènes impliqués dans chacune des cartes. Cela permet à l'utilisateur d'identifier rapidement les gènes impliqués dans les mêmes voies métaboliques. Malheureusement, peu des produits des gènes identifiés sont retrouvés sur KEGG. Peu de cartes sont donc trouvées. GeneProm fourni un lien vers les cartes KEGG sur lesquelles sont mis en évidence les gènes de l'étude.

3.2.3 myKegg

MyKegg est, comme GeneProm, une application web mais elle ne fournit pas d'interface html. Elle a été faite pour être interrogée par d'autres logiciels. L'interrogation de la base se fait via des requêtes http. À chaque action pouvant être effectuée par myKegg, correspond un URI particulière respectant l'architecture REST de l'application. L'application web répond sous la forme de fichiers XML devant être lus par le logiciel l'ayant interrogée. Par exemple l'URI : « mykegg.inra.fr/entries/find?name=galactose » renverra la liste des voies métaboliques où un composé ayant un nom proche de « galactose » est présent. *Remarque : le site mykegg.inra.fr n'existe pas, il correspond à une adresse fictive. Le logiciel sera déployé à l'INRA après la rédaction de ce manuscrit, mais il est pour l'instant sur des serveurs de test.* Le but de cette application est de fournir la liste des voies métaboliques où une liste de composés d'intérêt interviennent. La liste des composés étant donnée sous forme de liste de noms et non d'identifiants. Ce type de recherche pourrait être mené via l'API KEGG mais sa lenteur et le nombre limité de connexions la rendent impossible à utiliser pour faire de nombreuses requêtes. Elle a aussi été conçue de manière à pouvoir trouver rapidement quels sont les voisins d'un composé donné. Cette interrogation est récursive et peut se faire sur plusieurs niveaux, c'est-à-dire trouver les voisins des voisins, … Pour répondre à ce type de question la base de données contient la structure de graphe réelle des voies métaboliques que représentent les cartes KEGG.

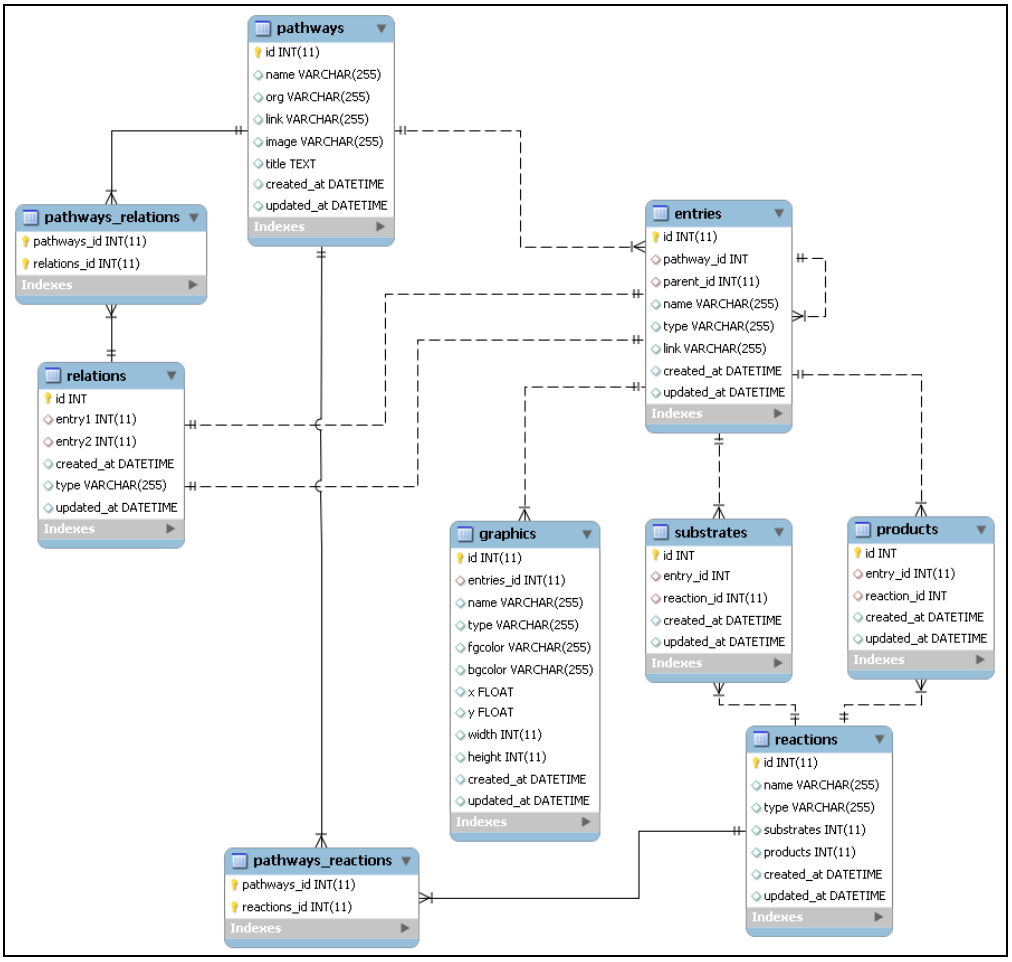


Figure 62 : Structure de la première version de la base de données myKegg
La structure de la première version de la première version de la base de données myKegg était très proche de la structure d'un graphe. La table « pathways » correspondant au graphe proprement dit, la table « entries » aux nœuds du graphe et la table « relations » aux arcs du graphe. La table « reaction » sert à trouver plus rapidement les réactions dans le graphe avec une mise en relation des substrats et des produits. Des informations concernant la représentation graphiques des nœuds sont stockées dans la table « graphics ».

Deux versions majeures de la base de données myKegg ont été développées. La première version correspondait à l'ensemble du contenu des fichiers KGML humains stockés sous forme de graphe. La structure de la base de données reflétait alors une structure de graphe. Elle contenait donc essentiellement des voies métaboliques qui contenaient des nœuds de différents types : réactions, gènes, métabolites, … avec leur nom et les liens entre ces nœuds. Elle contenait aussi les liens entre les voies métaboliques calculés à partir des informations des fichiers KGML. Cette base de données avait été faite pour fonctionner avec la version « Système d'information » de MPSA. Pour construire cette base de données, un parseur de KGML a été rédigé. Ce parseur permettant de trouver les informations de graphe dans les fichiers KGML et de les enregistrer en base. La base devait pouvoir fournir toutes les informations nécessaires à l'affichage d'un graphe complet sans passer par les fichiers KGML. Cette fonctionnalité n'avait plus de sens après le passage au développement du plugin MPSA pour VANTED puisque celui-ci représente déjà les fichiers KGML. Une nouvelle version de la base a donc été développée.

La deuxième version de la base de données contient toujours les informations de graphes, mais sa structure a complètement changé. La structure de la base est maintenant plus proche de la structure des fichiers KGML, et est donc très proche du schéma d'un fichier KGML présenté dans la figure 36 (chapitre 2.3.4.1). Cela rend la base de données beaucoup plus simple et beaucoup plus rapide à mettre à jour via les fichiers KGML tout en conservant les données de graphe. Deux autres modifications majeures ont été apportées à cette deuxième version : une ré-implémentation partielle de l'API KEGG et l'ajout d'une base de synonymes.

La base des synonymes est une fonctionnalité essentielle dans la recherche des composés intervenant dans une voie métabolique. En biologie, une même entité biologique, quel que soit le type d'entité (protéine, gène, petite molécule, acide gras, …) possède très souvent plusieurs noms. Un utilisateur cherchant les voies métaboliques ou le sucrose est impliqué doit trouver les mêmes résultats que s'il avait cherché les voies où le saccharose est impliqué puisque ce sont deux noms différents d'un même composé.

```
1   <?xml version="1.0" encoding="UTF-8" ?>
2   <names type="array">
3     <name>
4       <created-at type="datetime">2010-08-19T13:05:00Z</created-at>
5       <entry-id type="integer">381</entry-id>
6       <id type="integer">905</id>
7       <name>Saccharose</name>
8       <updated-at type="datetime">2010-08-19T13:05:00Z</updated-at>
9       <entry>
10        <created-at type="datetime">2010-08-19T13:04:58Z</created-at>
11        <ec-number nil="true"></ec-number>
12        <id type="integer">381</id>
13        <kegg-type>compound</kegg-type>
14        <updated-at type="datetime">2010-08-19T13:04:58Z</updated-at>
15        <pathways type="array">
16          <pathway>
17            <created-at type="datetime">2010-08-19T12:56:28Z</created-at>
18            <id type="integer">6</id>
19            <id-kegg>hsa00052</id-kegg>
20            <name>Galactose metabolism</name>
21            <organism-id type="integer">1</organism-id>
22            <updated-at type="datetime">2010-08-19T12:56:28Z</updated-at>
23          </pathway>
...
```

Figure 63 : Extrait de fichier XML décrivant les voies métaboliques ou le saccharose intervient

La figure 63 montre le fichier XML décrivant les voies métaboliques KEGG dans lequel le saccharose intervient. Les lignes 4 à 8 décrivent le composé trouvé dans myKegg (il peut y avoir plusieurs composés retournés). La ligne 6 correspond à l'identifiant dans myKegg et la ligne 7 au nom du composé trouvé dans la base de données. À ce composé correspond une entrée dans la base de données KEGG (« entry »). Celle-ci est décrite de la ligne 10 à 14 avec son type : « compound » et son identifiant « 381 ». Viennent ensuite les voies métaboliques où cette entrée KEGG est présente. Les lignes 16 à 23 décrivent la première de ces voies métaboliques (2 voies sont trouvées en tout). Cette voie est représentée par la

carte ayant pour identifiant « hsa00052 » qui décrit le métabolisme du galactose. L'autre voie décrite correspond au métabolisme de l'amidon et du saccharose.

La requête qui a été faite ici est faite sur un composé mais la même interrogation peut être faite avec une enzyme, un gène ou même une voie métabolique. La requête avec une voie métabolique donnera toutes les voies métaboliques liées à celle-ci. Par exemple avec le métabolisme du galactose trouvé précédemment, les voies retournées sont : les interconversions des pentoses et du glucuronate, le métabolisme du fructose et du mannose, et le métabolisme des osamines.

Pour pouvoir construire la base de synonymes, nous avons utilisé l'API KEGG et plus précisément la commande « bget » de l'API. Cette commande permet de renvoyer une fiche descriptive pour une entrée de la base de données KEGG. La Figure 42 du chapitre 2.3.4.1 donne un exemple de l'utilisation de la commande *bget*. La section « NAME » de la fiche *bget* donne une liste de noms connus pour une entrée donnée. Cette est donc idéale pour remplir une base de synonymes. Le problème est que cette commande doit être appelée sur toutes les entrées de la base de données, ou au moins sur tous les composés et les réactions (qui sont en fait des enzymes, comme expliqué dans le chapitre 2.3.4.1). Hors KEGG bloque l'accès à son API en cas de surcharge sur serveur. Pour résoudre ce problème, nous avons dû ré-implémenter la commande *bget* de KEGG dans notre base de données.

La commande *bget* de KEGG n'interroge pas leur base de données mais permet de fournir un accès à des fichiers stockés au format *bget*. Ces fichiers étant disponibles sur le ftp de KEGG, nous les avons récupérés. Ces fichiers ont ensuite été organisés en dossiers en fonction des types de fichiers auxquels ils réfèrent (cf figure 64).

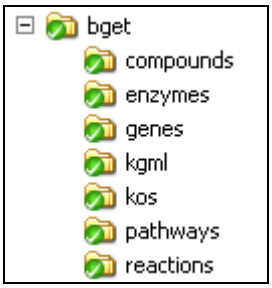


Figure 64 : organisation des fichiers bget
Les fichiers bget sont organisés en six dossiers : les composés (compounds), les enzymes (enzymes), les gènes (genes), les entrées de la base d'orthologie de KEGG (kos), les voies métaboliques (pathways) et les réactions (reactions). Un dossier contenant les fichiers KGML des voies métaboliques chez l'homme a aussi été créé.

Cet ensemble de dossiers peut être considéré comme une base de données de fichiers bget. Le nom des fichiers dans les dossiers correspond à l'identifiant KEGG de l'entrée décrite dans le fichier. La commande *bget* de l'API KEGG renvoie directement l'ensemble de la fiche, à la charge de l'utilisateur de la lire pour y trouver les informations souhaitées. L'implémentation de l'API *bget* réalisée intègre un parseur de fiches *bget* qui permet de récupérer les informations voulues.

```
1   bg = KeggService.bget "cpd:C00004"
2   bg.get_name
>> ["NADH", "DPNH", "Nicotinamide adenine dinucleotide"]

3   bg.get_db_links
>> {"3DMET"=>["B01127"],
    "ChEBI"=>["16908"],
    "KNApSAcK"=>["C00019343"],
    "NIKKAJI"=>["J213.546I"],
    "PDB-CCD"=>["NAI"],
    "PubChem"=>["3306"],
    "CAS"=>["58-68-4"]
    }
```

Figure 65 : Exemples d'utilisation de l'API bget de myKegg
Les lignes numérotées correspondent à des instructions, les lignes commençant par des chevrons correspondent aux résultats des instructions.

Dans la figure 65, la ligne 1 montre l'appel à la commande bget de myKegg. Ici l'utilisateur cherche des informations sur l'entrée KEGG ayant comme identifiant « cpd:C00004 ». L'identifiant commençant par « cpd », le logiciel comprends qu'il doit chercher dans le dossier « compounds » le fichier « C00004 ». La variable « bg » contient alors la fiche *bget* du composé. Il serait possible de l'afficher directement comme le ferait KEGG mais il est plus intéressant d'utiliser le parseur de fichiers *bget* pour retrouver des informations. La ligne 2 montre comment récupérer le contenu de la section « NAME » et renvoie un tableau contenant les noms du composé. Cette commande suffit pour la construction de la base de synonymes, mais le parseur permet de récupérer les informations de toutes les sections d'un fichier *bget (*36 sections de fichier *bget* différentes peuvent être analysées). Ainsi, la ligne 3 montre comment retrouver les identifiant du composé dans diverses bases de données. Le résultat est renvoyé sous forme de dictionnaire. En informatique, un dictionnaire est une structure associant à une *clé* unique une *valeur*. Ici la clé est située avant la flèche et correspond au nom de la base de données. Elle est associée à un tableau d'identifiants. Par exemple dans la base de données PubChem, le NADH a l'identifiant « 3306 ».

3.2.4 BioNMR

La base de données BioNMR a été développée pour répondre au besoin d'organisation des spectres RMN de l'équipe. Les contraintes pour le développement de cette base de données étaient liées à l'hétérogénéité des spectres RMN à stocker : le noyau observé (^{1}H, ^{13}C, …), spectres en 1 dimension ou 2 dimensions. Il fallait aussi pouvoir gérer les différentes expériences ayant menées à l'acquisition du spectre. La base de données devait enfin permettre l'identification automatique de pics en fonction d'une base de composés dont les déplacements chimiques sont connus. Enfin, un système pour inclure des algorithmes d'analyse statistique était souhaité pour permettre d'automatiser des études de *data minning* sur les spectres.

3.2.4.1 Base de données

En ce qui concerne la gestion des expériences (entourée en bleu dans le schéma de la figure 66), la table centrale est la table « experiments ». Il est possible de travailler sur des patients ou sur des animaux (table « organisms »). Une expérience est aussi caractérisée par une ou des pathologies et un ou des traitements. Les pathologies, traitements et expériences sont décrites par un simple champ texte, comme ils le seraient dans un cahier de laboratoire. Une expérience se rapporte à une série de spectres.

La partie de la base de données servant à la description des spectres est entourée en rouge dans la figure 66. Un spectre est caractérisé principalement par un type de RMN (c'est-à-dire le type de noyau observé) et une série de pics. Les pics sont représentés dans deux tables : « intensities » (ordonnée) et « shifts » (abscisses). La table intensité met en relation une valeur d'intensité avec le déplacement chimique (« shifts »). Un spectre 1d possède des pics ayant une seule coordonnée en abscisse : *shift_x*. Un spectre 2d possède des pics ayant deux coordonnées : *shift_x* et *shift_y*. Un déplacement chimique est caractérisé par une valeur de déplacement pour un type de noyau observé.

La partie servant à l'identification des pics est entourée en orange dans la figure 66. Un métabolite est caractérisé par un ou plusieurs noms et par des déplacements chimiques tirés de publications de référence [42,73]. Ainsi, on retrouve dans la base de données que l'alanine est caractérisée par des déplacements chimiques de 1.47 et 3.77 ppm en spectroscopie RMN HRMAS du proton et de 18.89 et 53.3 ppm pour la spectroscopie du carbone.

La partie concernant le lien vers des algorithmes de data minning n'est pas encore fonctionnelle. Le but de ces algorithmes est soit de classer les spectres en différents groupes en espérant pouvoir faire une correspondance avec l'état biologique de l'échantillon, soit de trouver les pics discriminants de deux états. Dans un premier temps une ANOVA (*ANalysis Of Variance*) est effectuée pour retenir les pics dont la variation est significative entre les deux états comparés (p-value < 0.05) puis un algorithme de *clustering* (création de groupes ou partition des données) est appliqué sur les spectres en ne gardant que ces pics. L'algorithme de clustering utilisé est « k-means » avec un nombre de groupes demandé de deux (nombre d'états biologiques identifiés). En appliquant ces algorithmes à une expérience de comparaison de tissus hépatique tumoral traité au CENU vs. non traité. Les deux groupes trouvés par l'algorithme sont bien les deux groupes biologiques et les métabolites variant sont en train d'être investigués. Ces algorithmes donnent donc des résultats intéressants mais ils ne peuvent pas encore être lancés via l'interface web.

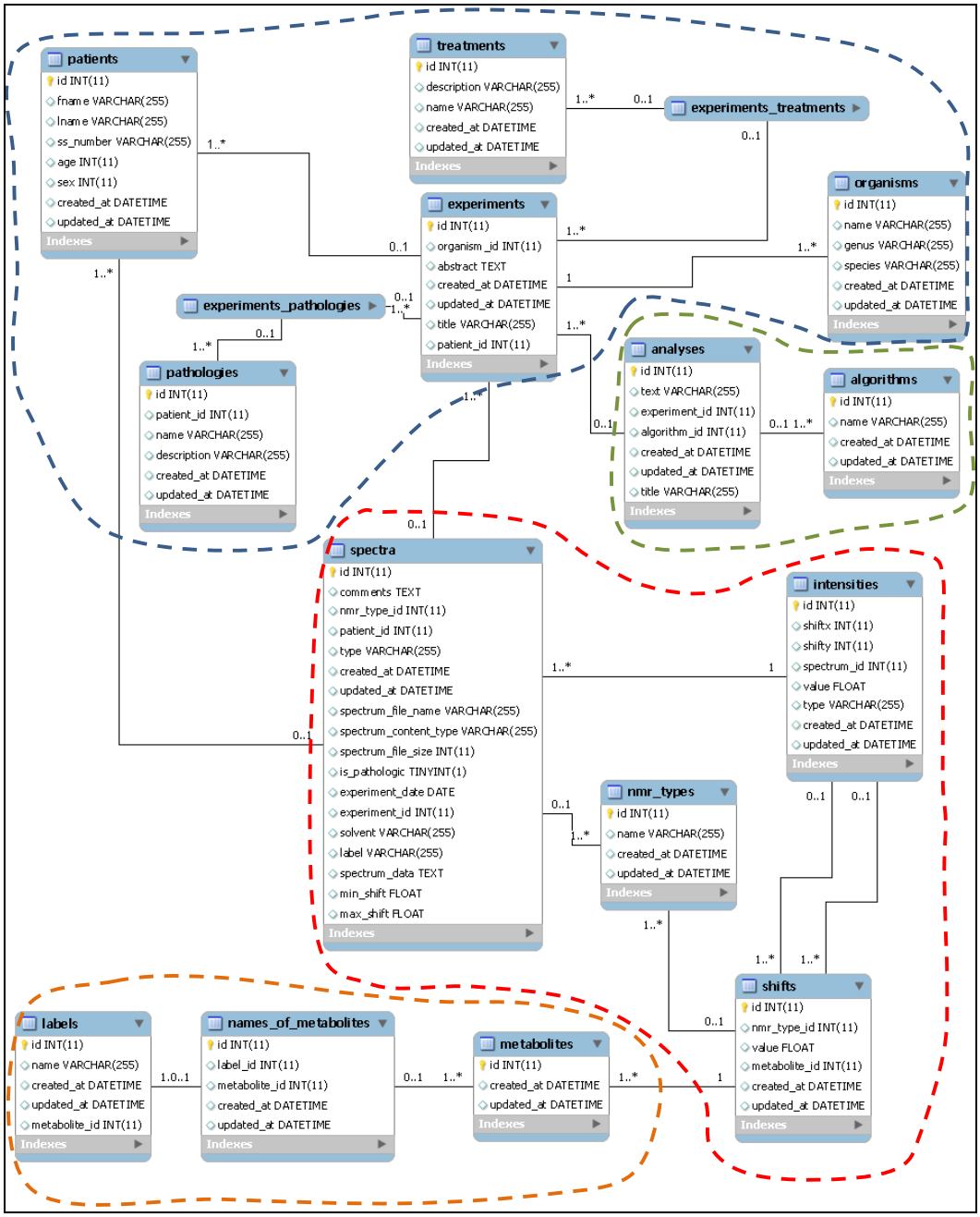


Figure 66 : Schéma simplifié de la base de données de BioNMR
Le schéma de la base de données BioNMR peut être découpé en quatre parties correspondant aux quatre besoins auxquels elle doit répondre. En bleu : la gestion des expériences, en rouge la gestion des spectres, en orange l'identification des composés en fonction des déplacements chimiques et en vert, les analyses statistiques.

3.2.4.2 Interfaces de visualisation

L'interface de saisie des nouveaux spectres (figure 67) permet de saisir des commentaires (*comments*) sur le spectre, de dire si c'est un spectre pathologique ou sain (*is pathologic*), la date de capture du spectre (*experiment date*), de charger les fichiers Bruker du spectre qui seront ainsi stockés sur le serveur (*spectrum file name*), de saisir les valeurs de pics (*spectrum* data)et de fixer les bornes inférieure et supérieure de la capture (*min shift* et *max shift*).

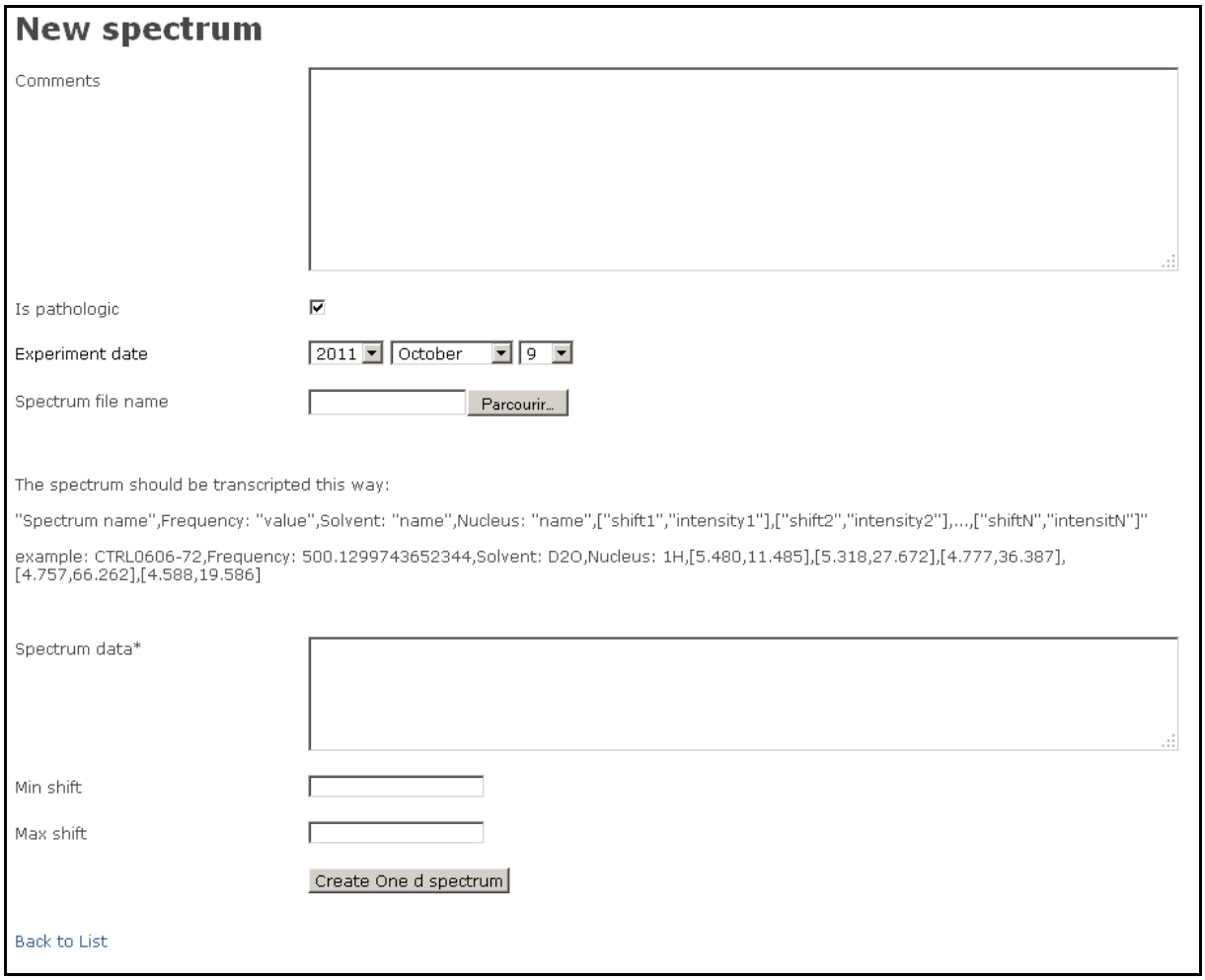


Figure 67: Interface de saisie d'un spectre RMN

Les données du spectre (les valeurs des pics) doivent être formatées correctement pour pouvoir être importées dans la base de données. L'utilisateur n'a pas à connaître ce format puisqu'il correspond au format utilisé par une procédure automatisée d'export des données de spectres implémentée dans le

logiciel Mnova. Cet export contient : le nom du spectre, la fréquence d'acquisition, le solvant utilisé, le noyau étudié, et une liste de pics. Les pics sont décrits de la manière suivante :

- un pic 1d est décrit par deux valeurs entre crochets : [shift, intensité]
- un pic 2d est décrit par trois valeurs entre crochets : [shift x, shift y, intensité]

Le choix des pics exporté n'est pas laissé au libre choix de l'utilisateur dans Mnova. Le logiciel calcul exporte les pics significatifs, c'est-à-dire les pics pour lesquels le rapport signal sur bruit est « suffisant » la valeur seuil du rapport et le calcul dépendent de l'intensité du bruit dans une portion du spectre située autour de chaque pic. Les développeurs de Mnova n'ont donné aucune explication supplémentaire sur ce calcul.

Une fois saisi, le spectre peut être visualisé. Cf. figure 68.

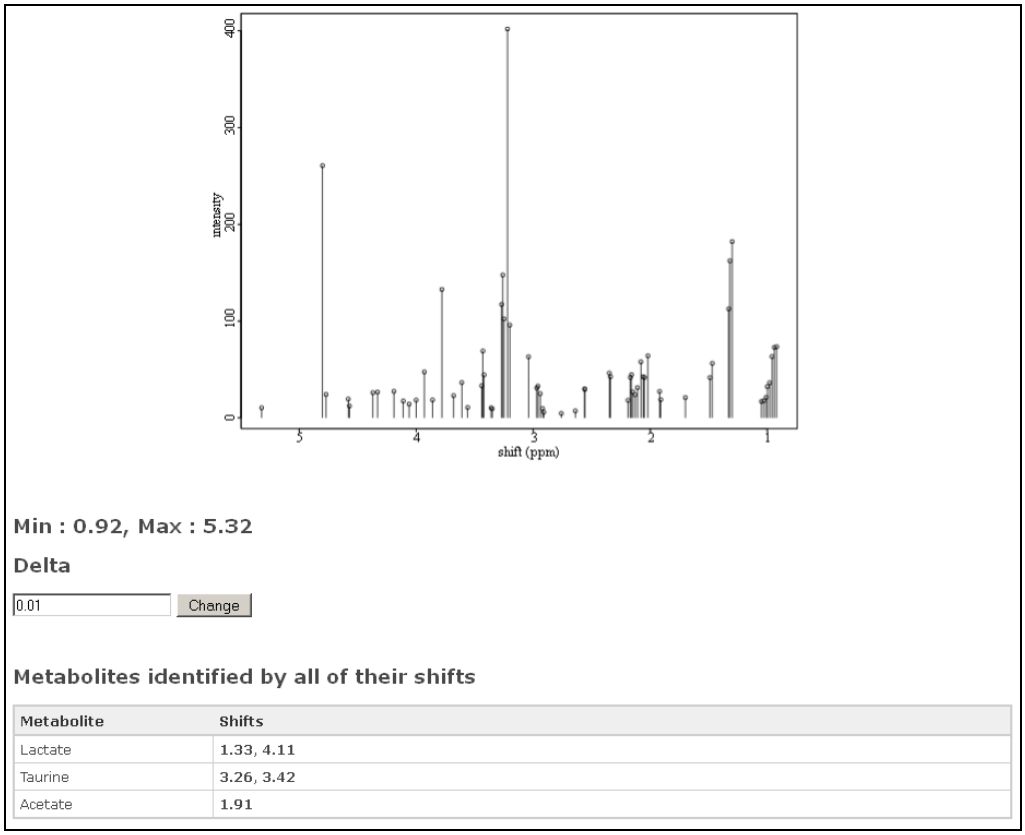


Figure 68 : représentation graphique d'un spectre 1d et des métabolites identifiés.

Un graphique est généré en R à partir des données fournies pour les spectre 1d. Il représente l'intensité des pics entrés en fonction du déplacement chimique. Les valeurs « Min » et « Max » correspondent aux bornes inférieure et supérieure de capture des déplacements chimiques. Si l'utilisateur les a saisies dans l'interface précédente, ce sont les bornes saisies, sinon ce sont les valeurs minimales et maximales de déplacement chimique du spectre. Le paramètre « Delta » correspond à la marge autorisée

pour l'attribution d'un métabolite à un pic. La liste des métabolites identifiés par tous leurs pics est ensuite fournie. Seuls les pics présents entre les bornes fournies sont considérés pour faire l'attribution. Le paramètre Delta n'est pas pris en compte ici. Les trois seuls métabolites identifiés de manière sûre sur ce spectre sont donc le lactate, la taurine et l'acétate.

Sous cette liste des métabolites identifiés de manière sûre vient la liste des métabolites pour lesquels on a trouvé une partie des pics dans le spectre : les métabolites potentiels. Cf. figure 69.

Metabolites identified by some of their shifts

Metabolite	Shifts
Fatty acids	1.29 (*1.3*), **1.3**, 1.36, 1.59, **2.05**, 2.25, 2.82, **5.32**
Choline	3.19 (*3.2*), 3.52, **4.06**
Phenylalanine	3.12, 3.3, **4.0**
Cr	3.03 (*3.04*), **3.93**
alpha-Glucose	3.4, 3.69 (*3.68*), **3.78**, 3.83, 5.22
Valine	**0.98**, 1.04 (*1.03*, *1.05*), 2.28, **3.61**
Glycerol	**3.56**, 3.64, 3.79 (*3.78*)
beta-Glucose	**3.42**, 3.46, 3.49, 3.72, 3.9, 4.64
Arginine	1.67, 1.72, **3.22**, 3.87 (*3.86*)
PCho	**3.2**, 3.57 (*3.56*), 4.18 (*4.19*)
Asparagine	2.86, **2.96**, 3.99 (*4.0*)
Glutamate	2.04 (*2.05*), **2.11**, **2.34**, 3.75
Methionine	2.12 (*2.11*, *2.13*), **2.19**, 2.63 (*2.64*), 3.85 (*3.86*)
Proline	2.0, **2.06**, 2.36 (*2.35*), 3.34 (*3.35*), **3.42**, 4.12 (*4.11*)
Alanine	**1.47**, 3.77 (*3.78*)
Isoleucine	**0.94**, **1.0**, 1.25, 1.46 (*1.47*), 1.97, 3.67 (*3.68*)
Leucine	0.95 (*0.94*, *0.96*), **0.96**, 1.71 (*1.7*), 3.74

Peaks

Metabolite	Shift	Intensity
Fatty acids	5.32	10.46
Unknown metabolite	4.8	260.72
Unknown metabolite	4.77	24.25
Unknown metabolite	4.58	19.51

Figure 69 : Liste des métabolites potentiels du spectre

Dans le tableau des métabolites potentiels sont affichés les métabolites identifiés par au moins un de leur pics. Tous les pics caractéristiques du métabolite dans la zone d'étude du spectre sont affichés. Les valeurs en gras correspondent aux pics retrouvés dans le spectre, les valeurs entre parenthèses correspondent à des pics retrouvés dans le spectre pour lesquels on retrouve une valeur de déplacement chimique d'un métabolite à +/- Delta. Ainsi, la choline est caractérisée par trois pics dans la zone 0.92 – 5.32 ppm (pour un spectre RMN HRMAS du proton) : 3.19, 3.52 et 4.06. Le pic 4.06 est retrouvé dans le spectre. Le spectre présente aussi un pic à 3.20 ppm proche du pic à 3.19 de la choline. Sous ce tableau vient la liste complète des pics du spectre.

Pour chaque spectre et pour chaque expérience, il est possible d'exporter la liste des composés identifiés dans un format csv interprétable par MPSA pour trouver les voies métaboliques dans lesquelles les composés identifiés sont impliqués.

3.2.4.3 pyNMR

En parallèle de cette base de données, a été développé un programme Python permettant de visualiser les spectres sauvegardés au format csv impossible à importer dans Mnova. Ce programme n'est pas une base de données ni une application web mais il permet de visualiser des données de RMN, comme BioNMR, il est donc présenté avec cette application web. Le but de ce programme était de permettre à un utilisateur de charger un fichier csv et de le visualiser. Les opérations de zoom devaient être possibles ainsi que le calcul d'aire sous la courbe. Il a été développé dans le cadre d'une étude où deux séries de spectres avaient été faites sur des échantillons de mélanomes traités au CENU. Pour chaque échantillon un spectre proton et un spectre proton + carbone était réalisé. Le spectre du carbone pur était très bruité et la ligne de base complexe à dessiner, nous voulions donc voir si une soustraction du spectre proton au spectre proton-carbone donnait des résultats intéressants. Ce type de soustraction de spectres RMN a aussi été utilisé dans le cadre d'une étude comparative du métabolisme de tissus hépatique tumoral vs. Tissus hépatique sain (d'un même patient). Le spectre du tissu sain était soustrait au spectre du tissu tumoral pour identifier les différences de métabolisme entre les deux états. Cette étude a permis d'identifier les métabolites dont la concentration varie de manière significative entre les deux états. Aucune application de RMN ne permettait de faire cela donc les spectres étaient exportés au format csv et travaillés sous Excel. Mais Excel ne permet pas de faire facilement des zooms et des calculs d'aire sous la courbe. L'application pyNMR a donc été développée.

Le spectre du proton est le spectre de référence. Il est considéré comme correctement callé en abscisses. Via la boite à outils du logiciel, on peut déplacer le spectre du proton-carbone par rapport au spectre du proton pour minimiser la différence entre les deux. Ce déplacement peut se faire manuellement avec trois niveaux de finesse : 0.0005, 0.001 ou 0.01 ppm vers la gauche ou la droite. Un déplacement plus rapide peut être effectué avec un curseur. Il est aussi possible de minimiser la différence entre les deux spectres automatiquement. Pour cela le logiciel crée une courbe de la somme des différences entre les deux spectres sur plus ou moins 2 ppm et retient la valeur optimale pour laquelle la différence est minimale. Dans la fenêtre du bas, un clic gauche affiche les coordonnées du point cliqué sur la courbe la plus proche. Deux clics droits successifs en deux points différents d'une même courbe permettent d'afficher la valeur de l'aire comprise entre la courbe et la droite passant par les deux points cliqués.

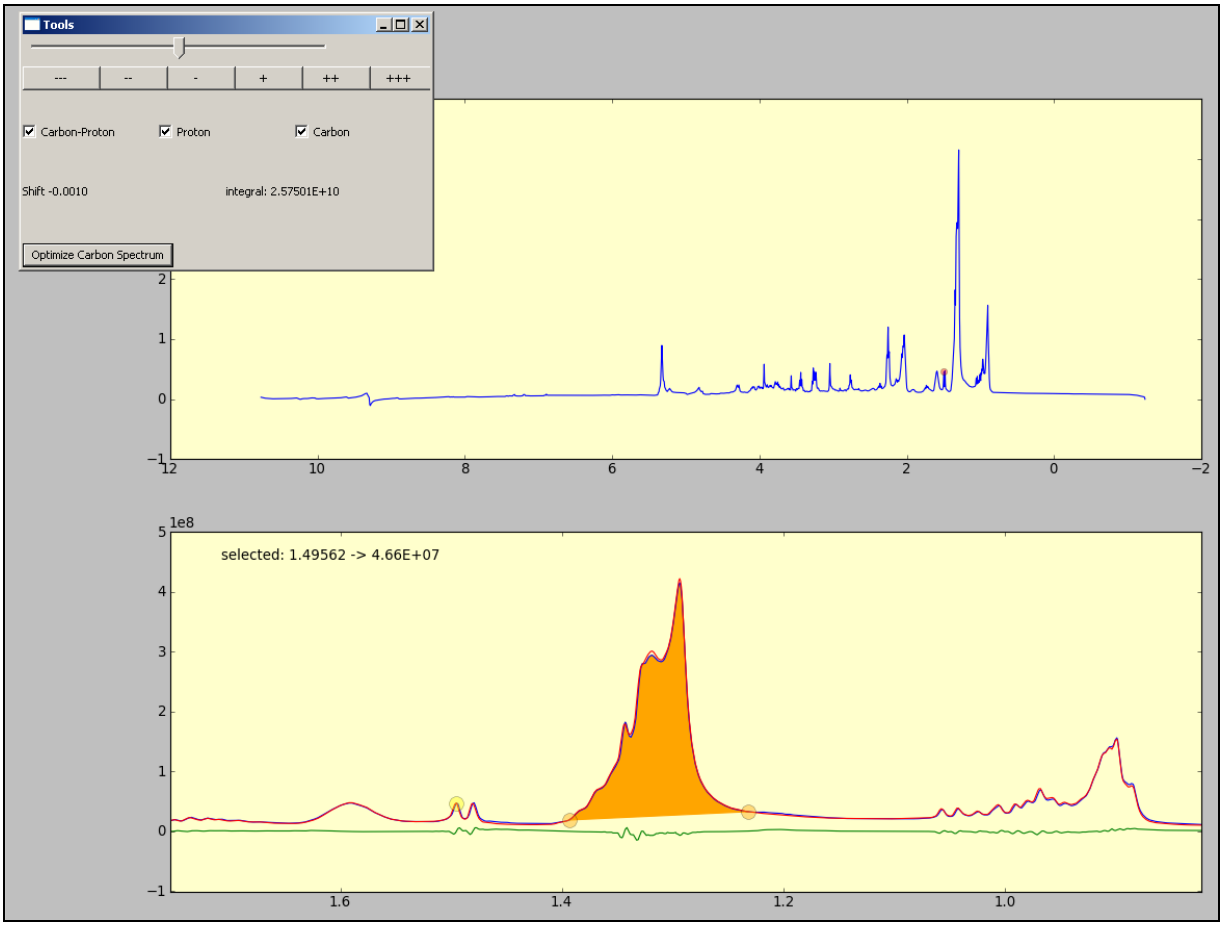


Figure 70 : Interface du logiciel pyNMR
L'interface du logiciel se découpe en trois parties : une boite à outils (en haut à gauche), une vue globale du spectre sur laquelle s'effectuent les zooms, les parties zoomées sont affichées en dessous. Le spectre proton-carbone est affiché en rouge, le spectre du proton en bleu et le spectre du carbone (issu de la soustraction des deux précédents) en vert.

3.3 MPSA : PLUGIN DE VANTED

Les expériences biologiques fournissent des informations très ponctuelles. Une expérience de métabolomique fournit une liste de composés intéressants, c'est-à-dire les composés dont les concentrations varient entre deux états comparés. Pour donner du sens à ces résultats, il est nécessaire de relier les métabolites entre eux pour trouver quelles sont les voies métaboliques impliquées dans la variation de concentration des composés. Pour réaliser cette opération, les chercheurs de l'équipe se basaient sur leurs connaissances et sur la base de données KEGG. Ce processus était très long et nécessitait une grande expertise. Hors une grande partie des informations était disponible en ligne sur la base de données KEGG et donc possibles à récupérer automatiquement. Le premier but de ce programme est donc de pouvoir inférer des voies métaboliques à partir d'une liste de composés : petite molécules, protéines, gènes, etc.

Les données de réseau étaient ensuite utilisées par des informaticiens à l'École Polytechnique pour réaliser des simulations de ces réseaux en construisant puis en résolvant des systèmes d'équations différentielles. Les équations différentielles sont déduites des lois enzymatiques des réactions du système. Seul un expert biologiste peut dire quelle est la loi biochimique qui s'applique à une réaction enzymatique. Cette donnée doit donc pouvoir être saisie de manière simple par l'utilisateur dans le logiciel. Il en va de même pour la concentration des métabolites. La partie construction du système d'équation différentielles et simulation doit se faire sans intervention de l'utilisateur. Un logiciel permettant de construire et de résoudre ces systèmes d'équations différentielles est en cours de développement à Polytechnique. Ce logiciel permet en outre de prédire les paramètres du système inconnus de l'utilisateur, ce qui est le cas de beaucoup de constantes dans les lois biochimiques.

Comme annoncé dans le chapitre sur la biologie des systèmes, il est aussi important de pouvoir étudier la structure du réseau. Un des moyens d'étude est le calcul des modes élémentaires. Une interface permettant de lancer le calcul et de visualiser les modes doit aussi être proposée.

3.3.1 Avant MPSA

Une première version du logiciel a été codée en C++ en utilisant un éditeur de graphes développé par la société Soluscience avec laquelle nous avons collaboré pendant cette thèse. Cet éditeur de graphe a été développé à l'origine pour créer un moteur de programmation graphique *SDView*. Cette première version a permis d'identifier différents problèmes liés à la représentation de graphes KEGG notamment. En utilisant les fichiers KGML nous espérions pouvoir représenter des voies métaboliques ressemblant aux cartes sur KEGG, ce qui n'est jamais le cas. Nous pensions utiliser l'API KEGG pour retrouver les composés impliqués dans les voies mais ses limitations en termes de trafic l'ont vite rendue inutilisable. C'est avec cette première version de l'application que nous avons vu la nécessité de créer la base de données myKegg qui remplaçait une partie des services de KEGG. Il nous est aussi vite apparut que pour une représentation de voies métaboliques, la bibliothèque de graphes utilisée ne suffisait pas autant en termes de représentation des nœuds (beaucoup de formes auraient dû être ajoutées) qu'en termes de positionnement automatisé des nœuds (*layout*).

Par contre c'est dans cette maquette qu'ont été conçues toutes les interfaces de saisie des paramètres des lois et des concentrations. Le format d'échange XML entre le logiciel et le simulateur développé à l'École Polytechnique a aussi été mis au point avec cette version du logiciel. Nous avons aussi gardé les algorithmes de simplification des graphes qui seront expliqués dans le paragraphe traitant du fonctionnement de MPSA.

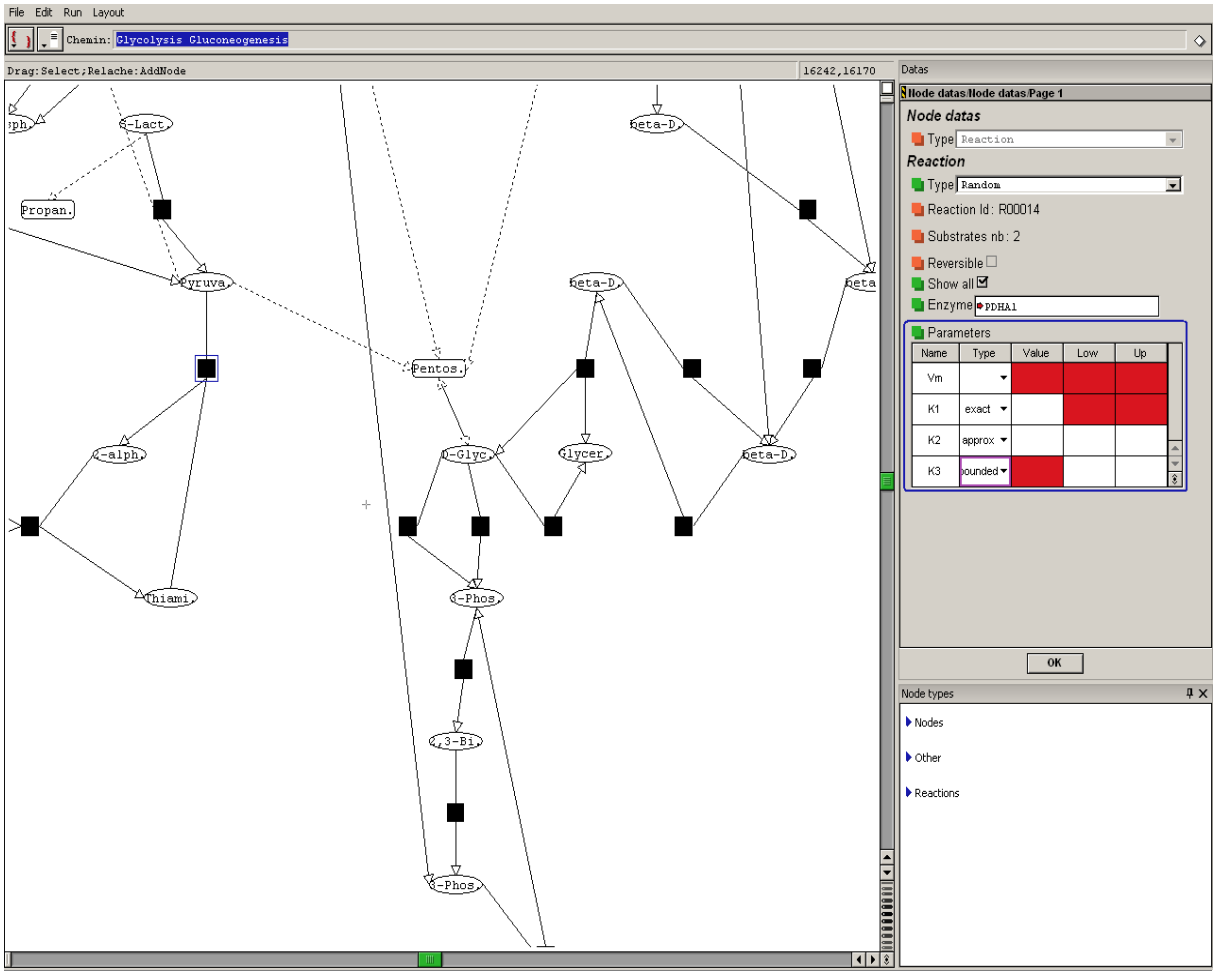


Figure 71 : Représentation d'une partie de la voie du cycle de Krebs dans la première version de MPSA

La partie de représentation du graphe est à gauche. Les carrés noirs représentent les réactions. Les composés sont figurés par des ovales. Les boites à coins ronds sont les voies métaboliques auxquelles est reliée la voie courante. La partie de droite représente l'interface de saisie des paramètres de la réaction sélectionnée (encadrée dans le graphe).

Cette version du logiciel a été abandonnée pour deux raisons : d'une part l'incompatibilité des licences de distribution du code écrit par Soluscience et le contrat de collaboration établi entre l'INRA et Soluscience et d'autre part pour pouvoir profiter des avantages du logiciel VANTED. La majeure partie du travail effectué dans cette version de MPSA était lié au moteur de rendu des graphes et notamment à la représentation de graphes KGML. Ce travail était déjà fait dans la partie KGML-ed de VANTED. Une autre partie importante du développement concernait myKegg. MyKegg étant une application web indépendante de MPSA, seules les connexions à la base de données étaient à ré-implémenter. Il fallait aussi ré-implémenter les interfaces de saisie des paramètres des lois et des concentrations.

3.3.2 MPSA

Comme expliqué dans l'introduction de ce chapitre, la version finale de MPSA est un plugin pour le logiciel VANTED. Le système de développement de plugins intégré à VANTED impose une certaine structure dans le développement. Dans VANTED, un plugin est appelé « Addon ». Chaque addon doit être défini dans un fichier XML lu par VANTED permettant de savoir quel est le fichier à charger pour que l'ensemble du plugin soit intégré. Le développement de MPSA a nécessité l'inclusion d'interfaces graphiques ajoutées dans les onglets de la boite à outil, ce sont des « InspectorTab ». Ces onglets contiennent une référence au graphe courant et des actions sont déclenchées à chaque évènement dans la fenêtre du graphe : ajout de nœuds, de liens, … Les actions et les calculs à effectuer sur le graphe sont ajoutés sous forme « d'Algorithme ». Ces algorithmes connaissent le graphe courant, mais pas les actions sur le graphe. Il est possible d'ajouter des conditions et des paramètres pour le déclenchement de ces algorithmes. Aucune documentation n'est fournie pour l'API de VANTED, seul un plugin d'exemple utilisant une grande partie des fonctionnalités de l'API est fourni.

Pour pouvoir fonctionner, MPSA nécessite un autre plugin créé par l'équipe de développeurs de VANTED : SBGN-ed [23]. Ce plugin permet d'éditer des graphes biologiques respectant le standard de représentation SBGN décrit au paragraphe 2.3.3.2. Ce plugin permet de créer des graphes en utilisant les glyphes SBGN, de représenter des cartes KGML en SBGN et de trouver les erreurs de représentation d'un graphe SBGN.

Le logiciel développé reprend toutes les fonctionnalités de la version en C++ et répond aux attentes exprimées précédemment. Via VANTED il est maintenant possible de charger directement un fichier KGML. La connexion avec myKegg a été développée et permet de trouver les voies métaboliques dans lesquelles une liste de composés d'intérêt intervient. Les interfaces de saisie des lois biochimiques, des paramètres de ces lois et des concentrations des métabolites ont aussi été développées pour pouvoir être exportées vers le solveur développé par l'École Polytechnique. Enfin, il est possible de calculer et de visualiser les modes élémentaires d'un réseau. Les fonctionnalités de VANTED peuvent être séparées en deux groupes : la reconstruction et la simulation des réseaux et l'étude de la structure des réseaux par les modes élémentaires.

3.3.2.1 Reconstruction et simulation de réseaux

Format d'entrée simple

Le point de départ de la reconstruction d'un réseau est une liste d'entités biologiques : métabolites, gènes, protéines, petites molécules, … MPSA a été conçu de manière à ce que cette liste puisse être fournie de manière très simple : un simple fichier csv. Ce fichier csv contient sur chaque ligne le nom d'une entité. Ce fichier peut être écrit très simplement par l'utilisateur dans un simple éditeur de texte ou il peut être généré par une application externe, comme c'est le cas pour GeneProm et BioNMR.

```
glutamate
glucose
phospho choline
phosphatidyl choline
choline
ethanolamine
phospho ethanolamine
phosphatidyl ethanolamine
glycerophosphocholine
glycerophosphoethanolamine
```

Figure 72 : fichier csv destiné à l'import dans MPSA
Ce fichier csv est issu d'une expérience de métabolomique.

Ce fichier est lu par MPSA qui se connecte à myKegg pour rechercher chacun des composés. Pour chaque composé myKegg va renvoyer une liste de composés dont le nom est proche de celui entré (cette recherche des chaines de caractère approchante utilise le logiciel Sphinx [179–181]).

Reconstruction d'un graphe à partir de données biologiques

Pour chaque composé entré dans le fichier csv, myKegg va proposer une liste de composés présents dans la base de données. Si le nom fourni dans le csv est retrouvé de manière exacte dans la base (insensible à la casse), un seul composé est retourné. L'utilisateur doit ensuite choisir pour chaque composé le composé qui lui correspond dans la liste fournie. Les choix de l'utilisateur sont retenus dans une liste qui lui est propre, sauvegardée dans son profil Windows (cf. figure 73).

Une fois que l'utilisateur aura sélectionné tous les composés, il pourra charger les voies métaboliques où ils sont présents et les charger dans le logiciel VANTED. Les voies métaboliques seront ordonnées par nombre de métabolites de la liste présents. Ceux-ci sont mis en évidence dans la liste pour permettre à l'utilisateur de sélectionner les voies métaboliques les plus pertinentes (cf. figure 74).

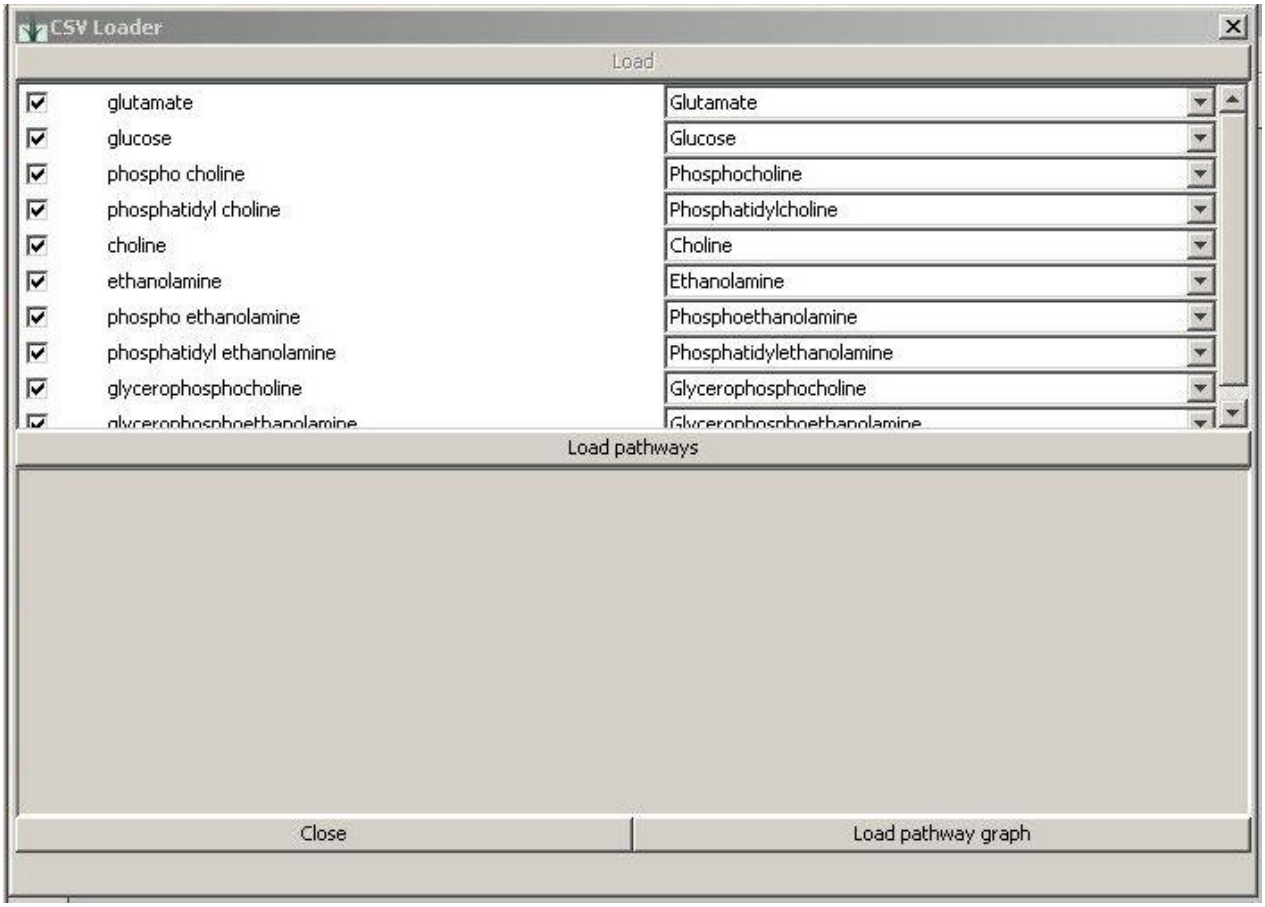


Figure 73 : interface permettant à l'utilisateur de choisir les composés de myKegg correspondant aux composés du csv.
La colonne de gauche correspond aux composés entrés dans le fichier csv. Pour chacun de ces composés, une liste de composés est fournie, sauf si le nom du composé est retrouvé de manière exacte dans myKegg. Par exemple, pour le glutamate, il n'y a que « Glutamate » proposé, alors que pour « phosphatidyl choline », 4 composés sont retournés dont la « Phosphatidylcholine » que l'utilisateur a ici choisi.

L'utilisateur peut alors charger la voie qui l'intéresse dans MPSA. Elle est récupérée sur KEGG en utilisant VANTED. Remarque : cette fonctionnalité a été temporairement désactivée en raison de changement de droits d'accès dans KEGG, une solution utilisant myKegg sera développée. Lors du chargement, le graphe est automatique transformée en SBGN puis simplifié.

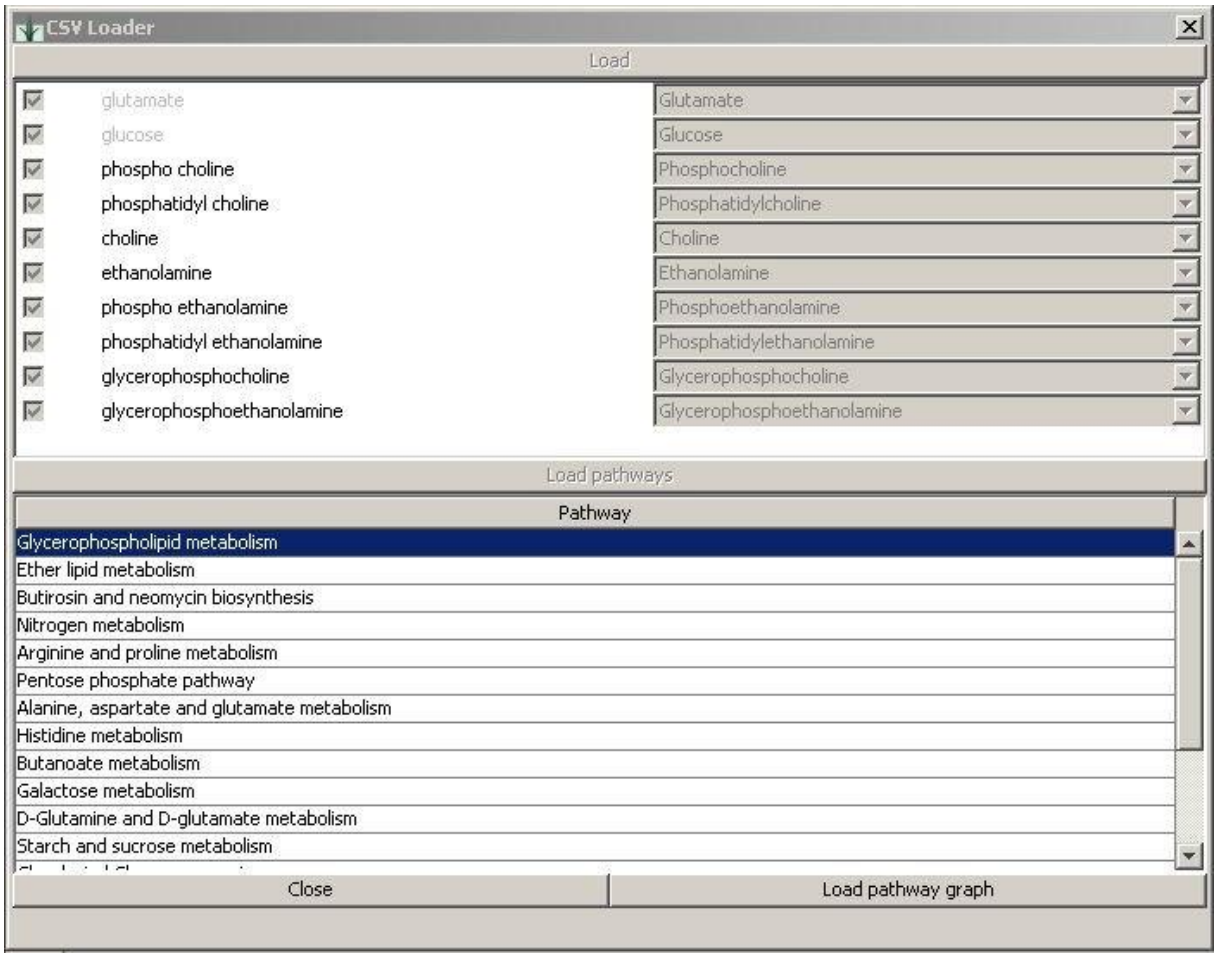


Figure 74 : Liste des voies métaboliques ou les composés sélectionnés par l'utilisateur sont présents
La liste des composés entrée permet d'identifier 16 voies métaboliques où au moins un des composés est impliqué. La voie la plus intéressante est la première de liste puisque c'est cette voie qui contient le plus grand nombre de composés de la liste. Quand l'utilisateur sélectionne cette voie, les métabolites phosphocholine, phosphatidylcholine, choline, éthanolamine, phosphoéthanolamine, phosphatidylethanolamine, glycérophosphocholine, glycérophosphoéthanolamine sont mis en évidence dans la liste des métabolites entrés pour montrer qu'ils participent à cette voie.

Simplification des graphes et export de données vers solveur Lix

La simplification est une étape importante pour la simulation par système d'équations différentielles. En effet, pour construire les ces systèmes d'équations différentielles, le logiciel utilise les lois biochimiques. Dans ces lois, le nombre de paramètres est important, on obtient des systèmes dont le nombre d'équations est inférieur au nombre de paramètres. Par exemple, la loi de Michaelis Menten comprend trois paramètres : le Vmax, le Km et la concentration du substrat. Il est impossible de résoudre de manière exacte ces systèmes mais il est possible de faire des estimations des paramètres. La simplification des graphes permet de réduire le nombre de réactions et donc le nombre de paramètre à

estimer. L'algorithme de simplification des graphes remplace une suite de réactions répondant à certains critères par une « super réaction » (cf. figure 75). Les critères auxquels doivent répondre ces réactions sont :

L'utilisateur ne doit pas avoir interdit la suppression de la réaction

La réaction ne doit pas avoir de substrat ou de produit parmi la liste des composés fournie par l'utilisateur. Ces composés sont par définition des composés pour lesquels des informations sont disponibles puisqu'ils ont servi à construire le réseau

La réaction ne doit pas être un point d'entrée ou de sortie du graphe

Deux réactions ayant les mêmes substrats et les mêmes produits peuvent être simplifiées par une seule réaction

Deux réactions successives supprimées sont remplacées par une « super réaction ». La suppression des réactions peut entrainer la suppression de métabolites dans le graphe. Par exemple, la suite de réaction A→B→C sera simplifiée en A→C. Cela n'est possible que si le métabolite supprimé répond à certains critères :

Le métabolite ne doit pas être consommé (ou produit) par plusieurs réactions. Si c'est le cas, le métabolite correspond au début d'une branche dans le réseau, s'il est supprimé la structure du réseau est perdue.

L'utilisateur ne doit pas avoir interdit la suppression du métabolite

Tant qu'il reste des réactions ou des métabolites à supprimer, l'algorithme se poursuit pour obtenir un graphe le plus simple possible. La figure 75 illustre cette simplification sur un exemple simple. Des essais de simplification sur des voies métaboliques issues de KEGG montrent que relativement peu de nœuds sont supprimés dans le graphe car il y a beaucoup de composés « branchés » et donc impossibles à supprimer.

Les réseaux simplifiés peuvent être ensuite exportés dans un format XML pour être intégrés au simulateur développé par l'École Polytechnique. Cet export est fonctionnel mais le simulateur est encore en développement donc les résultats des simulations ne peuvent pas encore être affichés dans VANTED. Ces graphes peuvent aussi être sauvegardés au format SBML avec prise en charge des paramètres et des lois. Le simulateur développé à Polytechnique permettra de prédire les paramètres manquants du système. MPSA pourra alors afficher ces paramètres ainsi que les vitesses de réactions, fournissant à l'utilisateur des informations précieuses sur le fonctionnement du modèle.

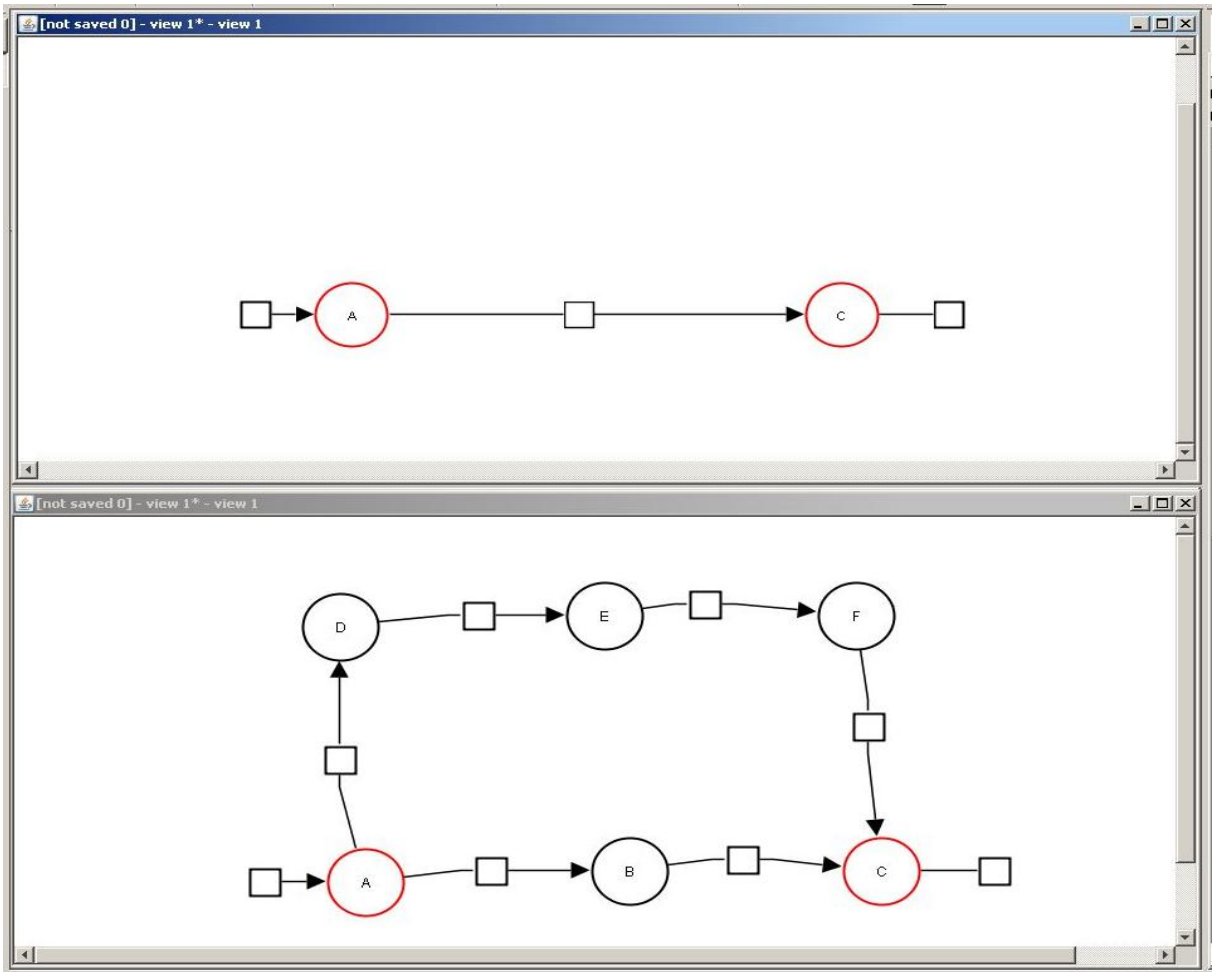


Figure 75 : Simplification de réseau
Le réseau simplifié est figuré en haut, le réseau de base en bas. La série de réaction A→D→E→F→C est simplifiée en A→C puis A→B→C est simplifié en A→C aussi. Une seule des deux réactions A→C est gardée, ce qui donne le graphe simplifié. Les réactions qui n'ont pas de substrat sont des points d'entrée du réseau. Les réactions qui n'ont pas de produit sont des points de sortie.

Système de plugin pour les lois

Il existe une très grande variété de lois biochimiques, il est donc impossible de toutes les proposer à un utilisateur du logiciel. Un système permettant de les rajouter facilement a donc été pensé. Dans le programme une loi est définie par différents valeurs : son nom, ses paramètres, sa formule et dans quelles conditions elle est applicable. Les paramètres de la loi sont séparés en deux types : les paramètres que le biologiste peut avoir éventuellement mesuré et les paramètres vraisemblablement inconnus (notés K_1, K_2, …). Dans le cadre d'une équation de Michaëlis-Menten, tous les paramètres sont potentiellement mesurables en faisant des études de cinétique enzymatique in vitro. Mais cette équation n'est applicable qu'a des réactions n'ayant qu'un seul substrat. D'autres lois doivent donc être utilisées dans le cadre de réactions ayant plusieurs substrats. Prenons l'exemple du mécanisme « ping-pong » de réaction enzymatique présenté dans la figure 76.

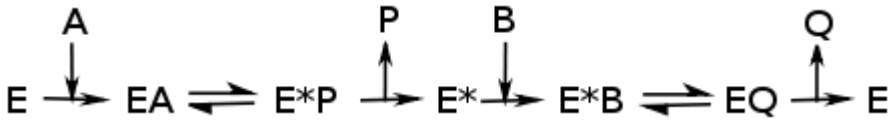


Figure 76 : Mécanisme ping-pong de catalyse enzymatique

Cette suite de réaction correspond à la catalyse par une enzyme E de la réaction A+B→P+Q. La première étape correspond à la fixation de A sur l'enzyme, ce qui amène à la création du produit P. L'enzyme est alors libre et dans la bonne configuration pour fixer B et permettre la création du produit Q.

L'équation régissant ce mécanisme est :

$$\frac{V_{max}\,[\mathrm{A}][\mathrm{B}]}{\mathrm{K}_1[\mathrm{A}][\mathrm{B}] + [\mathrm{A})[\mathrm{B}]}$$

Dans cette équation, le paramètre V_{max} est possible à mesurer, en revanche le paramètre K_1 est beaucoup plus dur à estimer. Cela est d'autant plus vrai si on multiplie le nombre de substrats de la réaction.

Pour définir de nouvelles équations, il faut définir deux nouveaux objets java : un objet décrivant de manière simple le nom, les paramètres « simples » comme V_{max} et « complexes » comme K_1. La seule différence dans le logiciel entre ces deux types de paramètres est que les paramètres complexes sont cachés à l'utilisateur par défaut. Cet objet doit aussi définir les conditions dans lesquelles cette loi peut s'appliquer. Par exemple le mécanisme ping-pong n'a de sens que s'il y a plus d'un substrat à la réaction. Le deuxième objet permet de décrire comment doit être écrite la réaction. Celle-ci est ensuite transformée au format MathML par le logiciel.

Ce système d'ajout de lois nécessite de recompiler le plugin pour chaque ajout de loi et de connaître certaines bases de fonctionnement du logiciel. Par défaut, six mécanismes de réactions enzymatiques et quatre types d'inhibition ont été implémentés. Ce sont les mécanismes trouvés dans le livre [137].

3.3.2.2 Etude des modes élémentaires

Le calcul des modes élémentaires se fait via le logiciel Metatool. Il faut donc, dans MPSA générer à partir du graphe le fichier d'entrée de Metatool. Ce fichier comporte : la liste des réactions réversibles, des réactions irréversibles, les métabolites internes et les métabolites externes et enfin les équations des réactions du système. Toutes ces informations sont disponibles dans le graphe dessiné ou chargé par l'utilisateur. Pour être correctement transformé au format d'entrée de Metatool, les graphes doivent être représentés en utilisant le plugin SBGN-ed ou avoir été convertis au format SBGN par les outils que ce plugin intègre. Les métabolites externes doivent être saisis par l'utilisateur pendant qu'il construit son graphe en ajoutant « _ext » à la fin du nom du métabolite. Le programme Metatool est alors exécuté et une fenêtre de retour des actions et des erreurs potentielles est affichée. Il est aussi possible de créer automatiquement un graphe depuis un fichier d'entrée Metatool et de charger les résultats de Metatool sur celui-ci pour le visualiser.

Le fichier de sortie de Metatool est alors analysé pour présenter une partie des résultats qu'il contient. La liste des modes élémentaires est fournie à l'utilisateur. Pour chaque mode élémentaire sont affichés : les métabolites externes d'entrée et de sortie (s'ils existent) ou si ce sont des cycles. Dans la liste affichée, l'utilisateur peut cliquer sur un mode élémentaire pour le visualiser sur le graphe. Les réactions peuvent être colorées en fonction du nombre de modes élémentaires passant par elles. La définition d'un mode élémentaire impose que si une réaction de ce mode est supprimée, le mode complet est invalidé puisqu'il ne peut plus être équilibré. Donc pouvoir identifier les réactions qui sont le plus souvent présentes dans les modes élémentaires permet d'identifier des points clés du réseau. Il est aussi possible de trouver facilement l'ensemble des modes élémentaires qui impliquent un ensemble de composés en sélectionnant ces composés dans une liste.

Afin de tester MPSA, nous avons construit un réseau reprenant trois voies métaboliques majeures de la mitochondrie ; les oxydations phosphorylantes, le cycle de Krebs et le cycle de l'urée. Ce réseau est décrit dans la thèse de Sabine Pérès [142]. Le réseau comporte : 20 réactions enzymatiques dont 8 sont irréversibles, la chaine respiratoire est représentée par ses bilans sous forme de deux réactions irréversibles : NADH + 10 H → NADH + 10 H_ext et $FADH_2$ + 6 H → FAD + 6 H_ext, et 17 réactions correspondant à des transports. Ce réseau comporte 29 métabolites internes et 20 métabolites externes. Metatool trouve 4637 modes élémentaires. La taille moyenne des modes élémentaires est de 15 réactions. C'est un réseau très complexe à représenter car il comprend des métabolites reliés à de très nombreuses réactions comme le proton, la molécule d'eau ou le phosphate. La représentation graphique des modes élémentaires dans ce cas n'est pas très pertinente car le réseau est complexe à visualiser. Par contre la coloration des réactions en fonction du nombre de réactions impliquées donne des informations intéressantes (cf. figure 77).

Par cette simple représentation, on s'aperçoit que les réactions les plus utilisées dans les modes élémentaires sont des réactions clés du métabolisme de la mitochondrie : ADP + Pi- + 3 H_ext ⇔ ATP + 3 H correspondant à la réaction de l'ATP synthase (utilisée dans 58 % des modes élémentaires du système), la réaction et les réactions du cycle de Krebs, toutes impliquées dans plus de la moitié des modes élémentaires. Si la réaction d'ATP synthase est supprimée, le réseau complet devient impossible à équilibrer, ce qui montre bien son importance dans le métabolisme de la mitochondrie.

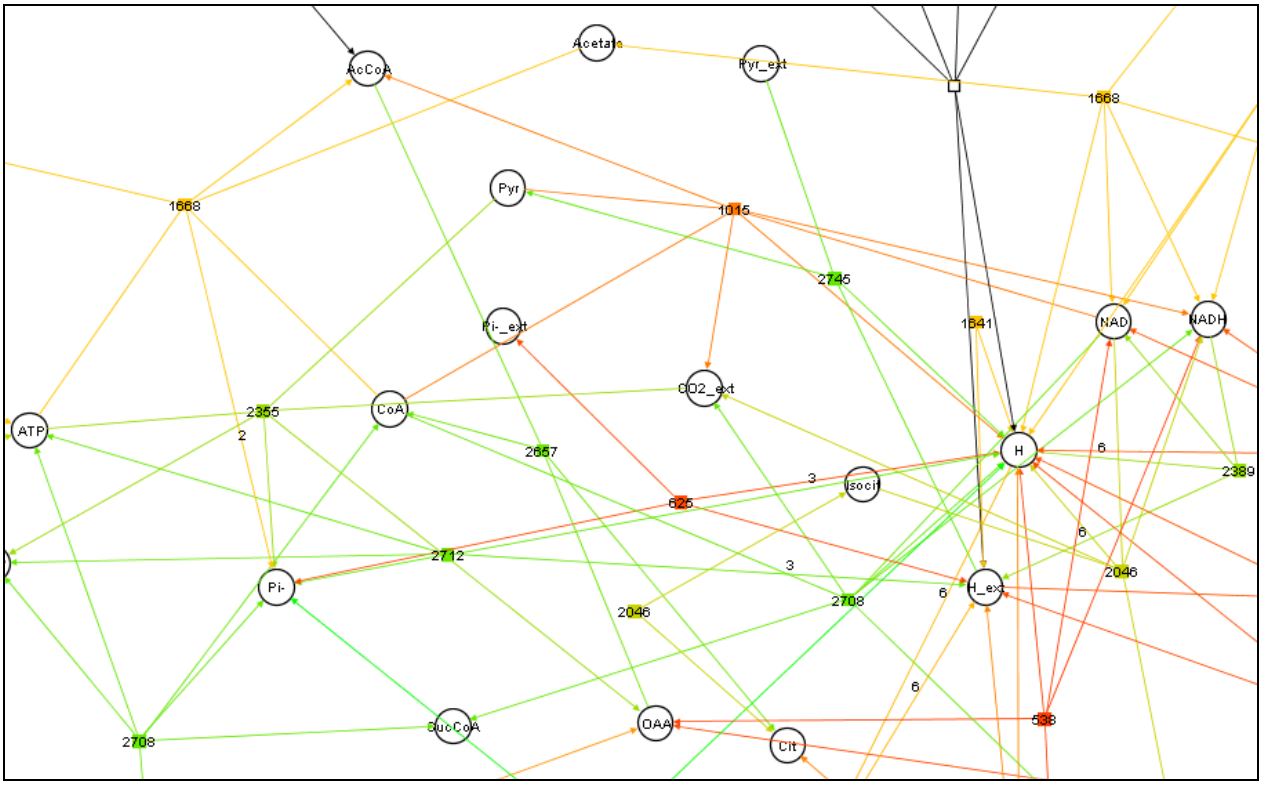


Figure 77 : Représentation graphique de l'importance des réactions dans les modes élémentaires dans le métabolisme de la mitochondrie de la levure
Ce graphe est issu d'une partie du métabolisme de la mitochondrie de la levure. C'est un graphe très complexe à représenter notamment en raison de la présence de l'ion hydrogène auquel sont liées 18 des 37 réactions du système. Les réactions sont figurées par des carrés. Ces carrés sont colorés en fonction du nombre de modes élémentaires dans lesquels les réactions sont impliquées. Plus la réaction est verte plus elle est impliquée dans beaucoup de modes élémentaires. Les chiffres appliqués sur les réactions représentent le nombre de modes élémentaires où cette réaction est impliquée.

Nous avons aussi étudié les modes élémentaires impliqués dans un modèle très simple construit avec l'École Polytechnique. Ce modèle décrit le métabolisme des glycérophospholipides. Il comprend 18 réactions dont 6 sont réversibles et correspondent aux entrées/sorties du modèle et 14 métabolites dont 8 internes. Ce modèle a été décrit dans [6]. 48 modes élémentaires sont trouvés d'une taille moyenne de 4 réactions. Quand on représente les réactions en fonction du nombre de modes élémentaires dans lesquelles elles sont représentées (cf. figure 78), la réaction la plus représentée quand on fait abstraction des réactions d'entrées et sorties du système est la réaction catalysée par l'enzyme clé du modèle : la PEMT qui permet la transformation de la phosphatidyléthanolamine en phosphatidylcholine. C'est une enzyme surtout active dans le foie qui est essentielle au maintien de la concentration en phosphatidylcholine dans la cellule. La voie majoritaire de synthèse de la phosphatidylcholine provient de la choline présente dans le sang, qui dans ce modèle, est convertie en phosphocholine puis en phosphatidylcholine. On retrouve les réactions de cette voie dans les réactions importantes du système dans la figure 78.

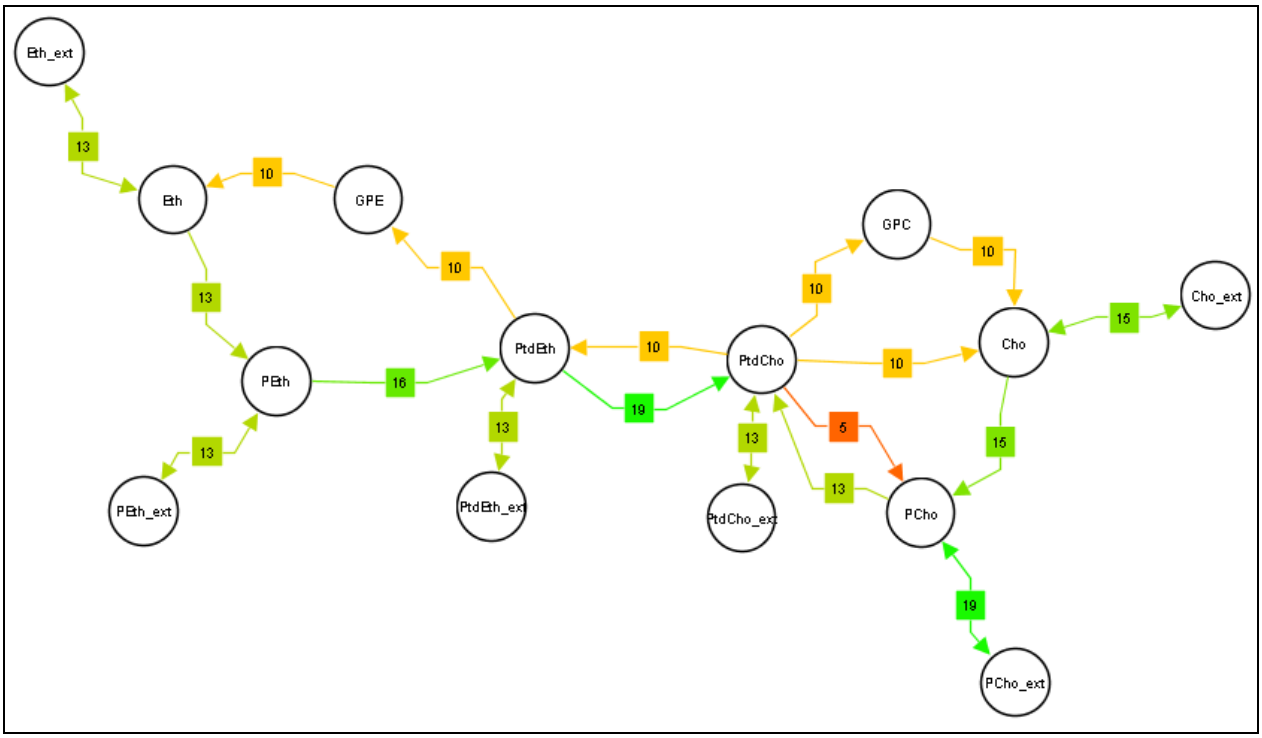


Figure 78 : Importance des réactions dans les modes élémentaires dans un modèle du métabolisme des glycérophospholipides dans le foie
Sur cette figure, la réaction PtdEth → PtdCho catalysée par la PEMT est impliquée dans 40 % des modes élémentaires.Les réactions menant à la synthèse de la phosphatidylcholine depuis la choline sont impliquées dans plus de 27 % des modes élémentaires. Même chose pour les réactions conduisant à la synthèse de la phosphatidylethanolamine depuis l'éthanolamine. Liste des abréviations de la figure : Eth : éthanolamine, GPE : glycérophosphoétanolamine, Peth : phosphoétanomanie, PtdEth : phosphatidylethanolamine, PtdCho : phosphatidylcholine, GPC : glycérophosphocholine, Cho : choline, PCho : phsphocholine. Les métabolites dont le nom finit par « _ext » correspondent aux métabolites externes du système.

MPSA permet donc à la fois une création des systèmes via sont interaction avec Kegg et la base de données myKegg. Le format simple d'import des données le rend utilisable pour des expériences de biologie très variées : transcriptomique, métabolomique, promotologie. Lorsque la connexion avec le simulateur en cours de développement à l'École polytechnique sera effective, il permettra d'afficher des informations sur les vitesses de réactions en utilisant un système de prédiction des paramètres inconnus des réactions. Cela permettra d'obtenir des informations sur les vitesses de réactions du système. MPSA permet aussi d'afficher de calculer et d'afficher les modes élémentaires du modèle. Une fonctionnalité particulièrement prometteuse est l'affichage de l'importance des réactions dans les modes élémentaires qui permet de trouver rapidement des réactions qui sont biologiquement essentielles.

3.4 DISCUSSION

Les différents logiciels développés ici constituent les bases d'une plateforme permettant de croiser des données de différentes techniques d'analyse biologique. Le mode simple d'import des données biologiques, via un fichier csv, soit rédigé manuellement soit généré par des applications, permet d'importer des données très diverses. Ce fichier contient une liste de noms de composés biologiques : gènes, protéines ou métabolites. Cette liste est interprétée par MPSA et myKegg pour permettre de trouver des voies métaboliques dans lesquelles ces composés sont impliqués. Les données de transcriptomique ne peuvent être importées que via un fichier csv. Les données de métabolomique peuvent être importées soit via un fichier csv rédigé manuellement soit via la base de données BioNMR qui le génère pour une expérience ou un spectre donné. La promotologie, utilisant à la fois les bases de données Genomatix et GeneProm, permet de trouver une liste de gènes potentiellement co-régulés. La liste de ces gènes peut être exportée depuis GeneProm vers MPSA.

Différents développements et améliorations sont à prévoir pour compléter les programmes construits. La première amélioration à apporter concerne la base de données KEGG. Cette base de données comporte d'importantes limites pour certaines évoquées dans le paragraphe 2.3.4.1. La première de ces limites est que la base de données KEGG a été construite pour être une base de données de voies métaboliques. Les métabolites et les enzymes sont simples à identifier. La plupart des gènes codant pour les enzymes sont aussi enregistrés. Par contre il y a très peu d'informations concernant les régulations de ces gènes. La base de données KEGG est une base de données très figée puisqu'elle ne peut être modifiée (correction d'erreur, ajouts de voies ou de réactions) que par les membres de leurs laboratoires. Des bases comme celles du projet MetaCyc [14,103,170], contenant 9 fois plus de bases de données et deux fois plus de réactions différentes, totalement libres d'accès et laissant la possibilité de créer ses propre bases de données, seraient peut-être mieux adaptées de ce point de vue. Une autre limite majeure concerne l'API de la base de données. Celle-ci s'avère impossible à utiliser à haut débit contrairement à ce qui est annoncé sur les pages d'aide de la base de données. Cette inutilité de l'API nous a contraints à développer une base de données simplement pour accéder aux données normalement récupérables via l'API. Cette inutilité de l'API était compensée par la mise à disposition de l'intégralité des enregistrements de la base sous forme de fichiers texte (KGML et Bget) sur un serveur ftp libre d'accès jusqu'en juillet 2011. La base de données myKegg a été construite uniquement en utilisant ces fichiers. Depuis cette date l'accès à ces fichiers n'est possible qu'en souscrivant à un abonnement annuel onéreux. Il sera donc nécessaire de trouver d'autres bases de données pour remplacer en partie KEGG et compléter les informations contenues dans myKegg. Des bases de données comme MetaCyc, Reactome [22,74,164] ou BIGG [106,171] pourraient être avantageusement utilisées. L'import dans la base de données myKegg peut être réalisé via la lecture d'un fichier BioPax pour MetaCyc, SBML ou BioPax pour Reactome et SBML pour BIGG. Depuis que l'accès au ftp de KEGG est devenu payant, les mises à jour de myKegg (utilisant les fichiers du ftp) sont devenues impossibles. Depuis cette même mise à jour, la recherche et le chargement des voies métaboliques depuis VANTED est devenu impossible. Les développeurs de VANTED sont en cours de négociation avec les propriétaires de KEGG pour pallier à ce problème (retrouvé aussi dans tous les logiciels basés sur KEGG comme ChipInspector de Genomatix). MyKegg possède des copies de ces fichiers KGML, il serait peut-être intéressant de développer un chargement de ces fichiers en attendant qu'un accord soit trouvé entre KEGG et les développeurs de VANTED.

En ce qui concerne la base de données BioNMR, la partie servant à gérer les expériences est partiellement à revoir : autoriser la comparaison de plusieurs pathologies et/ou traitement dans une expérience, lier les spectres à une pathologie et un traitement. La partie d'exploitation des données est à réaliser en utilisant dans un premier temps les algorithmes déjà implémentés. Ce développement

correspond à l'implémentation d'une interface de lancement des scripts R avec les paramètres nécessaires, une interface de suivi du déroulement du calcul R et une interface de présentation des résultats.

Le développement effectué sur le logiciel MPSA constitue une première étape d'un logiciel permettant réellement une intégration des données de métabolomique, transcriptomique et de promotologie. De nombreux développements sont encore à effectuer sur ce dernier. Différentes pistes sont envisagées ici. La première piste est l'utilisation des systèmes de *clustering* (SOM et étude de corrélation) de VANTED pour la prédiction de voies préférentiellement utilisées.

Un autre point important à développer est une utilisation plus poussée des modes élémentaires. Il serait intéressant de développer une interface permettant de représenter les réactions qui sont toujours présentes ensembles dans les modes élémentaires (indiqué dans les fichiers de résultat de Metatool). Il pourrait être aussi intéressant de coupler cette recherche des réactions toujours effectuées ensembles à l'algorithme de simplification des graphes. Il serait aussi souhaitable de classer les modes élémentaires et de visualiser cette classification sur le graphe. Il existe divers algorithmes de classification des modes élémentaires comme ACoM [93,92]. Les calculs Metatool (et ACoM) sont des calculs parfois très long pouvant nécessiter des ressources informatiques importantes. Il serait donc avantageux de placer ces exécutables sur un serveur puissant et de développer une interface d'interrogation web entre ce serveur et MPSA.

Il est parfois peu pertinent de se limiter à des voies métaboliques comme le fait KEGG. Il pourrait être intéressant de reconstruire les voies métaboliques en recherchant les voisins des métabolites entrés (les voisins étant dans ce cas tous les métabolites des réactions dans lesquelles le métabolite d'intérêt est impliqué). En exécutant cet algorithme de manière itérative sur les voisins, il est possible de créer un graphe dans lequel les métabolites entrés, ou une partie de ces métabolites, seront reliés. La recherche des voisins est déjà implémentée dans la base de données. Il faudrait donc développer l'interface d'interrogation de la base de données en utilisant les outils déjà implémentés dans MPSA et l'algorithme de *layout* automatique pour la représentation des résultats basé sur des algorithmes comme les *force-based layouts* (algorithmes de dessin basé sur les forces) déjà implémentés dans VANTED.

Pour ce qui concerne la promotologie, deux problèmes ont été identifiés. GeneProm est très dépendant des mises à jour d'EnsEMBL qui ne sont jamais annoncées. L'autre problème concerne une redondance de certaines fonctionnalités entre GeneProm et myKegg. Pour les mises à jour, il faudrait tester régulièrement leur présence et les effectuer le cas échéant. Ensuite, il faut effectuer une étude test pour vérifier si les résultats sont exacts en comparant les résultats de l'étude de test aux résultats de cette même étude réalisée préalablement et vérifiés. Les fonctionnalités redondantes entre myKegg et GeneProm concernent les relations vers les bases de données externes. Il serait nécessaire de supprimer le code de GeneProm et faire appel aux fonctionnalités de myKegg (mieux écrites et plus poussées) pour générer les liens. Genomatix semble développer les mêmes fonctionnalités que celles implémentées dans GeneProm. La seule fonctionnalité manquante reste la localisation des modèles de promoteurs en amont de gènes ne présentant pas de zone promotrice connue dans Genomatix. Si cette fonctionnalité est développée à l'avenir par Genomatix, le maintien de l'interaction Genomatix/GeneProm deviendra vraisemblablement inutile.

L'ensemble des outils développés constitue une première étape essentielle au développement d'une plateforme logicielle dédiée au croisement de données biologiques, à leur représentation sur des graphes et à la modélisation statique ou dynamique de réseaux biologiques.

4 PROMOTOLOGIE

Le génome des métazoaires est composé de régions codantes encadrées de régions non-codantes contenant des éléments de régulation, courtes séquences nucléotidiques, qui permettent de programmer la lecture du code génétique à un temps donné et dans une cellule spécifique [18]. Ces séquences (ou motifs) nucléotidiques sont généralement localisées dans la zone promotrice proximale du gène (de –1000 au site d'initiation de transcription). Les protéines de régulation ou facteurs de transcription, reconnaissant ces motifs nucléotidiques, agissent en synergie avec des ARN spécifiques (micro ARN : miRNA et ARN interférents : RNAi) reconnaissant des séquences localisées en 3' [15,129]. Ces séquences de régulation et les gènes qu'elles régissent définissent des espaces fonctionnels dans le génome et jouent ainsi un rôle majeur dans le développement, l'adaptation à un environnement, la réponse à des éléments nutritionnels et la pathogenèse.

Ces motifs de régulation constituent des sites de fixation de facteurs de transcription (TFBS : *Transcription Factor Binding Sites*) et sont de courtes séquences nucléotidiques (de 10 à 30 nt), difficiles à différentier de séquences aléatoires non fonctionnelles du génome [37]. La plupart des sites seraient localisés sur la séquence correspondant aux 300 nt en amont du site de début de transcription (TSS : *Transcription Start Site*) [56].

En dépit d'avancées remarquables en génomique ayant permis d'identifier la plupart des gènes du génome humain, la connaissance de leur régulation transcriptionnelle reste élémentaire. La promotologie a pour but l'étude de mécanismes de cette régulation du génome. Elle consiste en une analyse des séquences promotrices des gènes : la structure de ces séquences promotrices est encore mal connue et serait composée d'une combinaison de plusieurs motifs de régulation (chacun de l'ordre de dix à cinquante nucléotides) appelée module cis-régulateur (CRM : *Cis-Regulatory Module*), capable d'être reconnue par un ou plusieurs facteurs de régulation permettant l'activation ou la répression de la transcription. La promotologie s'appuie sur un ensemble d'outils bioinformatiques permettant l'étude de ces éléments de régulation.

Trois banques de données majeures proposent une collection des séquences de TFBS et de CRM issues de la littérature.

- TRANSFAC, base de données ayant un accès gratuit et un accès professionnel payant permettant d'accéder à un plus grand nombre de données. Dans sa version gratuite la base comprend 7915 TFBS pour 6133 facteurs de transcription, la version professionnelle permet l'accès à 14.490 TFBS et 30.118 sites (en juillet 2011) [144,145]. La majorité des TFBS sont représentés par la séquence fournie par la littérature mais très peu de matrices représentant l'état dégénéré (différences en fonction des gènes régulés) du TFBS sont disponibles (398 pour l'accès gratuit et 1422 pour l'accès professionnel). Ces matrices contiennent les informations les plus intéressantes des TFBS.

- PReMod, base de donnée répertoriant 118 000 CRM dans le génome humain [7,33,161]. Cette base de données a été construite à partir de la base Transfac et d'un algorithme de découverte de TFBS. La très grande majorité de ces TFBS sont donc issus d'un algorithme de prédiction et n'ont jamais été validés biologiquement.

- Genomatix, base de données couplée à un ensemble d'outils de recherche de motifs et d'annotation de génomes [153]. L'accès à Genomatix est payant. Elle propose une collection de 61.329 TFBS (7018 chez l'homme) et 21.203 facteurs de transcription dont 8495 sont associés à un TFBS. 1326 matrices sont disponibles organisées en 407

familles. La majorité des TFBS et des matrices de Genomatix sont issus de la littérature et ont été validés biologiquement.

4.1 GENOMATIX

Etant donné le nombre important de TFBS et de matrices disponibles sur Genomatix, notamment chez l'homme, et la diversité des outils présents sur leur site, nous avons choisi de mener nos études via cette base de données. Genomatix propose plusieurs logiciels et bases de données permettant une analyse méthodique de séquences promotrices du génome humain et d'une sélection d'espèces animales et végétales. Le plus intéressant est la base de données MatBase et l'outil de recherche de matrice couplé à cette base [13,97]. MatBase rassemble de nombreuses séquences de régulation issues de la bibliographie et leur traitement en matrices de régulation.

Ces matrices sont issues d'un alignement de courtes séquences de régulation (de l'ordre de 15 à 30 nucléotides) identifiées chez plusieurs organismes. Cet alignement permet de proposer une séquence *core* (généralement 4 nucléotides) très bien conservée dans les différentes espèces et une séquence flanquante plus variable [35] (cf. figure 79). Les modèles sont constitués d'un ensemble de matrices séparées par des séquences aléatoires plus ou moins longues (figure 79).

D'autres logiciels proposés par Genomatix (menu GEMS Launcher) permettent d'utiliser cette banque de matrices afin d'analyser la régulation d'un gène donné ou de retrouver un ensemble de gènes régulés de la même manière [62] :

- *FastM* : permet de construire des modèles de régulation constitués d'une combinaison de plusieurs matrices de régulation ou de matrices combinées à des séquences nucléotidiques propres à l'utilisateur.
- *FrameWorker* : permet de rechercher des combinaisons de matrices (défini comme un modèle de régulation) partagées par plusieurs séquences promotrices [13].
- *ModelInspector*: permet de rechercher un modèle, défini par l'utilisateur, soit sur une banque de clones couvrant l'ensemble d'un génome, soit sur une banque de séquences promotrices.
- *Search for phylogenetically conserved promoter models* (dans un menu *Complex Patterns*) : permet de rechercher un modèle (proposé par Genomatix ou l'utilisateur) conservé dans une sélection de plusieurs organismes. Cette fonctionnalité a été ajoutée pendant l'été 2011.
- *PromoterInspector*: logiciel permettant de détecter des régions promotrices sur des séquences. [107]

Les trois derniers outils correspondent à la recherche de matrices connues ou saisies par l'utilisateur sur des séquences. Tous ces outils sont basés sur MatInspector qui est le programme clé de Genomatix.

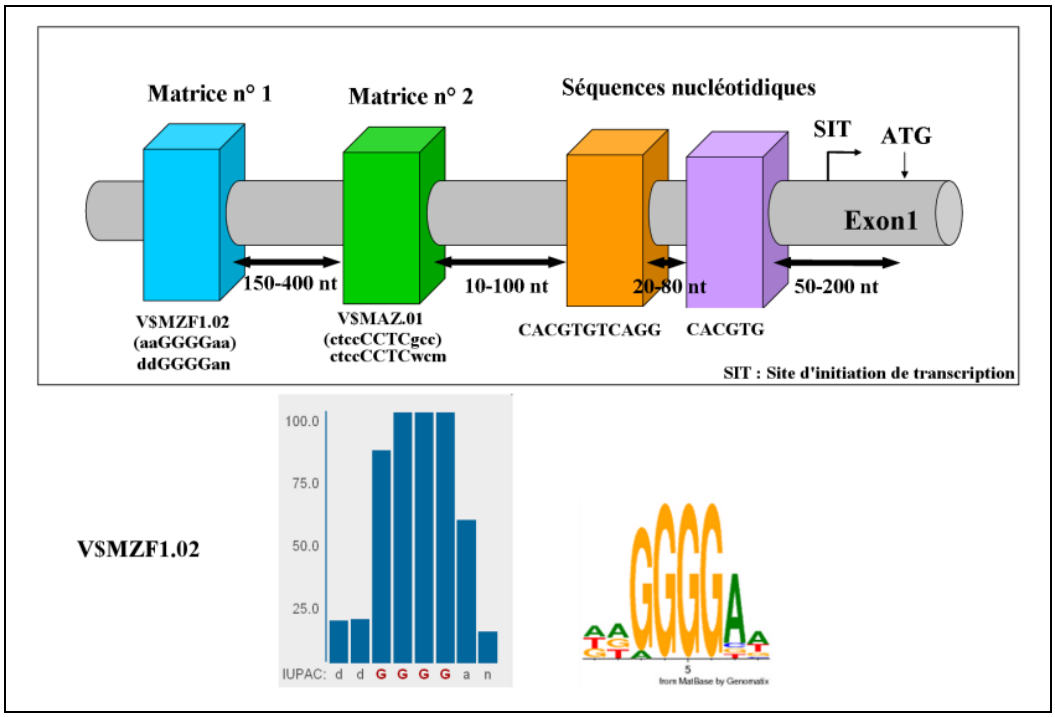


Figure 79 : Matrice et modèle Genomatix
La partie du haut représente un modèle Genomatix. Celui-ci est composé de deux matrices notées V$MZF1.02 et V$MAZ.01 et de deux séquences entrées par l'utilisateur : CACGTGTCAGG et CACGTG et de l'ATG du gène. Les matrices sont identifiées sur des alignements phylogénétiques de séquences de régulation. Elles sont extrêmement dégénérées. Elles comportent deux types de séquence : la séquence « core » très bien conservées en phylogénie et la séquence flanquante plus variable mais globalement mieux conservée que le reste du promoteur. Les éléments du modèle sont séparés par une distance dont les bornes inférieures et supérieures doivent être fixées par l'utilisateur. Dans le nom des matrices, « V$ » indique que c'est une matrice de vertébré, puis vient le nom de la famille de la matrice et enfin le numéro de la matrice dans la famille. Dans la partie inférieure de la figure, sont figurées deux représentations différente d'une même matrice (V$MZF1.02). La séquence (respectant le code IUPAC) est en abscisses. Dans la représentation de gauche la fréquence des bases est visualisée sous forme d'histogramme, dans celle de droite, la fréquence de chaque base est indiquée par la taille de la lettre.

La fonctionnalité ModelInspector est très intéressante car elle permet l'identification d'un ensemble de gènes disposant dans leur séquence promotrice d'une combinaison de séquences régulatrices proches. Ces gènes sont susceptibles d'être co-régulés par la même combinaison de facteurs de régulation et ainsi participer à une même fonction biologique. Cependant, deux restrictions majeures sont à mentionner :

- La banque de séquences promotrices ne couvre qu'une fraction des gènes humains. À ce jour, Genomatix, par sa banque de données Eldorado, propose environ 237 000 ARN messagers primaires possibles chez l'homme, 131 704 régions promotrices identifiées. 55 715 de ces régions ont un site de début de transcription identifié et 21 726 de ces dernières seulement sont utilisables à ce jour par le logiciel ModelInspector (base de données ElDorado 08-2011). Ainsi 83.5 % environ des séquences promotrices de Genomatix correspondent soit des gènes encore non identifiés (protéine seule connue), soit des gènes dont la zone promotrice n'est pas bien caractérisée, ou soit des gènes dont la zone promotrice est incorrectement séquencée.

- Bien que Genomatix intègre une base de données du génome humain, il est impossible de rechercher un modèle de régulation sur l'ensemble du génome humain car cette

banque de données ne comprend pas la totalité du génome mais uniquement les séquences géniques, les séquences promotrices connues et les séquences identifiées par PromoterInspector comme séquences promotrices potentielles. La recherche de modèles sur la totalité du génome par ModelInspector se fait sur les clones de la base de données EMBL-Bank (*European Molecular Biology Laboratory Nucleotide Sequence Database, Roswell Park Cancer Institute Human BAC Library*). Cette base de données a été construite par une collaboration [154] entre GenBank, base de données de séquences génétiques du NIH (Institut national de la santé américaine), l'ENA (*European Nucleotide Archive*) et la banque de données japonaise DDBJ (*DNA Database of Japan*). ModelInspector permet la localisation d'un modèle sur la séquence d'un clone. Cependant, aucune information génique n'est disponible pour les clones, donc ModelInspector ne propose pas de gènes, connus ou supposés, à la proximité immédiate du modèle localisé. Afin de pallier à ce manque de fonctionnalité, nous avons développé cette localisation dans le logiciel GeneProm décrit au paragraphe 3.2.2.

4.2 PUBLICATIONS

Les premières études que nous avions réalisées sur Genomatix concernaient un ensemble de séquences de régions promotrices de gènes connus pour être co-exprimés, soit grâce à des expériences de transcriptomique, soit par la littérature. FrameWorker permettait alors de trouver des matrices communes à plusieurs de ces gènes. Ces matrices étaient alors utilisées pour construire les modèles de régulation. Cependant, même si un ensemble de gènes sont co-induits ou co-réprimés sur une biopuce ou par des analyses de RT-PCR quantitative, rien ne permet d'affirmer qu'ils partagent un même mécanisme de régulation transcriptionnelle impliquant des TFBS communs. Ainsi, peu de matrices communes étaient retrouvées simultanément sur plusieurs gènes et les modèles devenaient très complexes à assembler.

Nous avons donc cherché un moyen d'identifier des matrices sur le promoteur d'un seul gène. Compte tenu de la conservation des séquences promotrices au cours de l'évolution [123], un nouveau protocole basé sur des alignements phylogénétiques a été établi. Ce protocole a été évalué lors d'une étude de la régulation différentielle des gènes codant pour les isoformes de la protéine ANT (*Adenine Nucléotide Translocator*) et notamment du gène ANT4 [29].

Isoforme	Symbole HGNC	Expression	Elements de régulation connus
ANT1	*SLC25A4* (4q35.1)	Cœur et muscle squelettique	OXBOX [68]
ANT2	*SLC25A5* (Xq24)	Cellules en prolifération	GRBOX [39], Oct1, SV40, AP2, Sp1 [64,70]
ANT3	*SLC25A6* (Xp22.33)	Ubiquitaire	Sp1, NF-kB, GAS-like [5]
ANT4	*SLC25A31* (4q28.1)	Spermatozoïdes	E2F6 [54]

Figure 80 : Expression différentielle des quatre isoformes du gène *ANT*
Pour chaque isoforme, le symbole officiel HGNC du gène est fourni ainsi que son profil d'expression. La liste des éléments de régulation déjà identifiés dans la littérature est aussi donnée.

Les gènes *ANT* sont des gènes nucléaires codant pour un transporteur des nucléotides adényliques (échange ADP / ATP) à travers la membrane interne de la mitochondrie. C'est la voie majeure d'import et d'export de l'ATP et de l'ADP à travers la membrane mitochondriale. Quatre gènes ANT codent pour quatre isoformes : ANT1, ANT2, ANT3 et ANT4. Chaque isoforme possèderait des paramètres cinétiques spécifiques, ce qui permet à la cellule d'adapter sa production d'énergie à son environnement [121]. Les isoformes ANT1 et ANT3 permettent d'exporter l'ATP

produit dans mitochondrie en échange d'ADP tandis que l'isoforme ANT2 a le rôle inverse [16] : importer l'ATP cytosolique dans la mitochondrie en conditions glycolytiques pour permettre le maintien du gradient de potentiel transmembranaire mitochondrial essentiel à la survie de la cellule.

Une première étude a été réalisée pour analyser la structure du promoteur du gène *ANT4* et vient d'être publiée [29] (paragraphe 4.2.1). Une seconde étude a été conduite récemment en utilisant le même protocole pour l'analyse comparée de la régulation des quatre gènes des isoformes de la protéine ANT (manuscrit en préparation).

4.2.1 Analyse informatique de la régulation transcriptionnelle du gène ANT4 et son rôle dans la spermatogenèse.

L'isoforme ANT4 a été identifiée récemment chez l'homme. Elle est exprimée dans le spermatozoïde et dans les testicules [28]. Cette isoforme est apparue chez les mammifères et est essentielle à la spermatogenèse [9]. La séquence de la protéine est très proche de celle des autres isoformes du gène (66 à 68 % d'identité). Une des particularités du peptide est la présence de séquences additionnelles en N-terminal (13 acides aminés) et C-terminal (8 acides aminés) probablement impliquées dans l'adressage de la protéine vers le flagelle du spermatozoïde [57]. Le rôle supposé de cette isoforme est de compenser l'absence de l'isoforme ANT2, présente sur le chromosome X, dans le spermatozoïde male [9]. Le protocole d'étude de la régulation du gène ANT4 est résumé dans la figure 81.

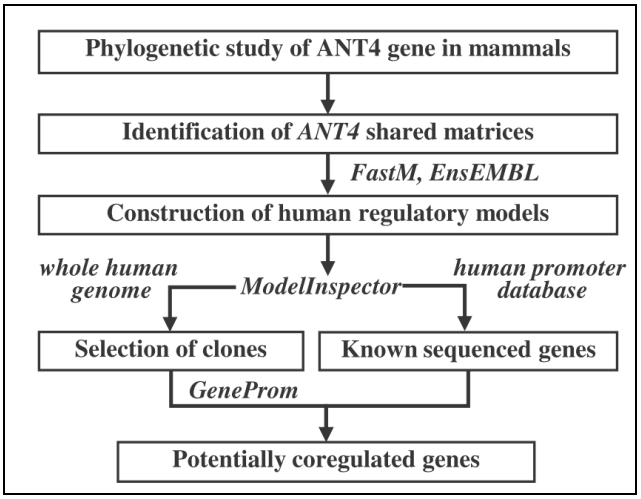


Figure 81 : Schéma du protocole de l'étude promotologique du gène ANT4
Tiré de [29]. Les cinq étapes de l'étude sont : 1) étude phylogénétique des gènes ANT4 de certains mammifères permettant de filtrer les séquences inutilisables 2) identification de matrices de régulation partagées par toutes les séquences alignées dans la première étape (via FrameWorker et MatInspector sur Genomatix) 3) construction des modèles de régulation en combinant les matrices identifiées précédemment 4) localisation du modèle sur le génome humain ou sur une base de séquences de promoteurs 5) identification de gènes potentiellement co-régulés à proximité du modèle.

La première étape du protocole correspond à une étude phylogénétique du gène ANT4. Pour cela, toutes les séquences des gènes ANT4 des mammifères de la base de données EnsEMBL sont extraites et analysées. Ainsi, 30 séquences dont celle de l'homme sont extraites et alignées. Les

séquences extraites comportent la séquence complète du gène et une séquence de 1500 nucléotides en amont du site d'initiation de transcription (zone contenant le promoteur). Toutes les séquences comportant plus de 5% de bases indéterminées dans la zone contenant le promoteur sont écartées. Seules 12 séquences répondent à ce critère. Les séquences sont ensuite alignées an utilisant l'algorithme Clustal 2.0 puis un arbre phylogénétique est construit en utilisant deux méthodes: *neighbour-joining* [105] et *minimal evolution* [135] avec une évaluation du *boostrap* (score évaluant la résistance de la topologie d'un arbre à une perturbation de l'alignement, peut être considéré comme un score de confiance) des branches dans les deux cas. Les deux techniques fournissent le même arbre phylogénétique qui peut donc être considéré comme valide. Les arbres permettent de vérifier qu'il n'y a pas de gène très différent des 29 autres, ce qui pourrait correspondre à une autre isoforme mal annotée ou un gène ayant évolué très rapidement chez un organisme. Une phylogénie séparée d'une zone incluant le promoteur est aussi effectuée pour identifier des promoteurs qui seraient très différents du promoteur de l'ANT4 chez l'homme. La position du site d'initiation de la transcription que prédit EnsEMBL est aussi vérifiée par rapport à celle de l'homme. Après cette étude, seules sept séquences sur les 30 entrées sont conservées. Ces séquences sont celle de : l'homme, du chimpanzé, du macaque, du chien, du bœuf, du dauphin et de la souris. Ce nombre de séquences apparait faible. Cependant, comme l'ANT4 n'est présente que chez les mammifères et que les séquences retenues comprennent l'homme et la souris avec plusieurs intermédiaires, la diversité phylogénétique du groupe reste suffisante.

Les 7 zones promotrices (1500 nucléotides en amont du TSS) sont soumises à l'outil FrameWorker afin d'identifier les matrices de régulation partagées par l'ensemble de ces séquences . Un faible nombre de matrices sont identifiées : V$SMAD3.01 (reconnue par le facteur de transcription Smad3 impliqué dans les voies de signalisation du TGF-beta), V$MZF1.02 (reconnue par la protéine MZF1 : *myeloid zinc finger protein*), et une séquence identifiée sur les alignements : CACGTGTCAGG retrouvée en plusieurs copies (3 chez l'homme) dans les 33 nucléotides en amont du TSS. Dans cette séquence, deux familles de matrices (V$EBOX et V$HIFF) sont retrouvées, toutes deux sont impliquées dans la glycolyse. Les modèles construits sont constitués de la combinaison de matrices et de séquences : V$SMAD3.01 / V$MZF1.02 / CACGTGTCAGG / TSS. Les distances entre les éléments du modèle sont fixées de manière à permettre une variation de plus ou moins 50% par rapport aux distances observées chez l'homme. Les paramètres dans ce modèle sont : le nombre de matrices présentes, le nombre de copies de la séquence entrée manuellement et la stringence de la recherche des matrices.

Les modèles construits sont alors recherchés soit sur la totalité du génome soit sur la banque de promoteurs de Genomatix (ElDorado). Si le modèle est recherché sur l'ensemble du génome, les positions du modèle sont retournées sur des clones GenBank et les gènes à proximité du modèle sont recherchés par GeneProm. Si le modèle est recherché sur ElDorado, une liste de gènes est directement fournie. Les gènes trouvés pour l'ensemble des modèles créés sont présentés dans la figure 82. Dans cette figure ne sont présentés que les gènes pour lesquels une protéine est connue. Plus de la moitié des gènes identifiés ne répondent pas à cette condition.

Gene, ID	Encoded protein (Ensembl)	Protein function (NCBI, GeneCards)
AMN	Amnionless homolog	Extraembryonic visceral endoderm layer
APEX1	APEX nuclease (multifunctional DNA repair enzyme) 1	Repair of apurinic/apyrimidinic sites **in testis** (Raffoul et al., 2004)
BAX*	BCL2-associated X protein	Mutagenesis regulation in **spermatogenesis** (Xu et al., 2010)
CDK4*	Cyclin-dependent kinase 4	Cell cycle G1 phase progression in male reproduction (Buchold et al., 2007)
FLJ32713 fis	Unknown (TESTI2000756)	Unknown (**expressed in testis**)
HSPBAP1	HSPB (heat shock 27 kDa) associated protein 1	Regulating stress response
IGF2BP3	Insulin-like growth factor 2 mRNA binding protein 3	RNA synthesis/metabolism, major foetal growth factor (Nielsen et al., 1999)
IGF2R*	Insulin-like growth factor 2 receptor	Receptor for insulin-like growth factor 2 (IGF2) and mannose 6-phosphate
KAT5*	K(lysine) acetyltransferase 5	Chromatin remodelling with an abundant **spermatid** protein (Reynard et al., 2009)
LAMP1*	Lysosomal-associated membrane protein 1	Binds amelogenin, differentially expressed in **spermiogenesis** (Guttman et al., 2004)
RMND1*	Required for meiotic nuclear division 1 homolog	Unknown
RPUSD4*	RNA pseudouridylate synthase domain-containing protein 4	Unknown, **expressed in prostate**
SLC25A31	Solute carrier family 25, adenine nucleotide translocator ANT4	Mitochondrial ATP/ADP carrier in **spermatozoid** (Dolce et al., 2005)
SLC2A4	Solute carrier family 2, member 4 (GLUT4)	Facilitated glucose transporter, **detected in human testis** (Angulo et al., 1998)
SOHLH1*	**Spermatogenesis** and oogenesis specific basic helix-loop-helix 1	Germ cell-specific, oogenesis regulator and **male germ cells** (Matson et al.)
TDRD1*	Tudor domain containing 1	Essential for **spermiogenesis** (Yabuta et al., 2011; Wang et al., 2001)
THAP8*	O-sialoglycoprotein endopeptidase (**TESTI2004929**)	Unknown
TKTL1	Transketolase-like 1	Important role in transketolase activity, **testis** expressed (Coy et al., 1996)
TMEM184A	Transmembrane protein 184A = Sdmg1	**Male-specific** expression in embryonic gonads (Svingen et al., 2007)
UBE2B*	Ubiquitin-conjugating enzyme E2B (RAD6 homolog)	Post-replicative DNA damage repair **in spermatogenesis** (van der Laan et al., 2004)
SUN1	Chr. 7 unc-84 homolog A	Nuclear anchorage/migration, expression of **meiotic reproductive genes** (Chi et al., 2009)

Figure 82 : Ensemble des gènes trouvés par l'étude promotologique du gène ANT4
Tiré de [29]. Cette figure présente l'ensemble des gènes, dont la protéine est connue, pour lesquels une occurrence du modèle a été trouvée dans la zone promotrice. Les gènes marqués par une étoile correspondent aux gènes identifiés à la fois par la recherche GeneProm et par la recherche ElDorado. Les mots en gras dans la description correspondent aux fonctions liées à la spermatogenèse, ou définissant un rôle dans les testicules ou la prostate.

Parmi les 21 gènes identifiés, 16 ont une fonction liée à la spermatogenèse ou s'expriment dans le testicule ou la prostate (80 % des gènes identifiés). Cette étude a donc permis de caractériser un ensemble de TFBS vraisemblablement impliqués dans la régulation de l'ANT4. La séquence CACGTG a aussi été caractérisée. Cette séquence est présente en plusieurs copies dans le promoteur d'ANT4 et dans la famille de matrices V$HIFF impliquées dans la réponse à l'hypoxie, ce qui est cohérent avec le métabolisme essentiellement glycolytique du spermatozoïde. La protéine ANT4 est connue pour être exprimée au cours de la spermatogenèse et nos résultats fournissent une première validation du modèle de promoteur proposé puisque de nombreux gènes identifiés sont impliqués dans des voies liées à la spermatogenèse.

L'ANT2 permet l'import de l'ATP vers l'intérieur de la mitochondrie dans des conditions de métabolisme glycolytiques. Etant donné le caractère glycolytique du métabolisme du spermatozoïde, il est logique de penser que la protéine ANT4 possède le même rôle que la protéine ANT2 exprimée dans ces mêmes conditions glycolytiques et absente du spermatozoïde mâle (ANT4 codée par le chromosome X). L'ANT4 serait donc apparue au cours de l'évolution pour compenser l'absence de l'ANT2 dans ces spermatozoïdes. Si on prend comme hypothèse que la cinétique de l'ANT4 est moins rapide que celle de l'ANT2, une différence de concentration d'ATP cytosolique entre le spermatozoïde mâle exprimant uniquement ANT4 et le spermatozoïde femelle exprimant ANT2 et ANT4 peut être déduite. Ainsi, le spermatozoïde femelle contiendrait plus d'ATP mitochondrial et moins d'ATP cytosolique que le mâle. Comme c'est de la concentration d'ATP cytosolique que dépend la vitesse de déplacement du spermatozoïde, le spermatozoïde mâle serait donc plus mobile. Par contre, une plus grande concentration d'ATP mitochondrial permet au spermatozoïde femelle d'avoir une durée de vie supérieure dans l'oviducte [40], et ainsi de favoriser une fécondation dans une phase plus tardive de l'ovulation que celle d'un spermatozoïde mâle.

Lors de notre étude, nous avons identifié une seconde région promotrice dans le gène ANT4 en utilisant le logiciel PromoterInspector de Genomatix. Ce promoteur est situé dans le deuxième intron, en amont immédiat du troisième exon. Aucun transcrit du gène ANT4 ne correspond à ce promoteur. Afin de déterminer si ce promoteur pouvait être biologiquement valide, des études complémentaires ont été menées.

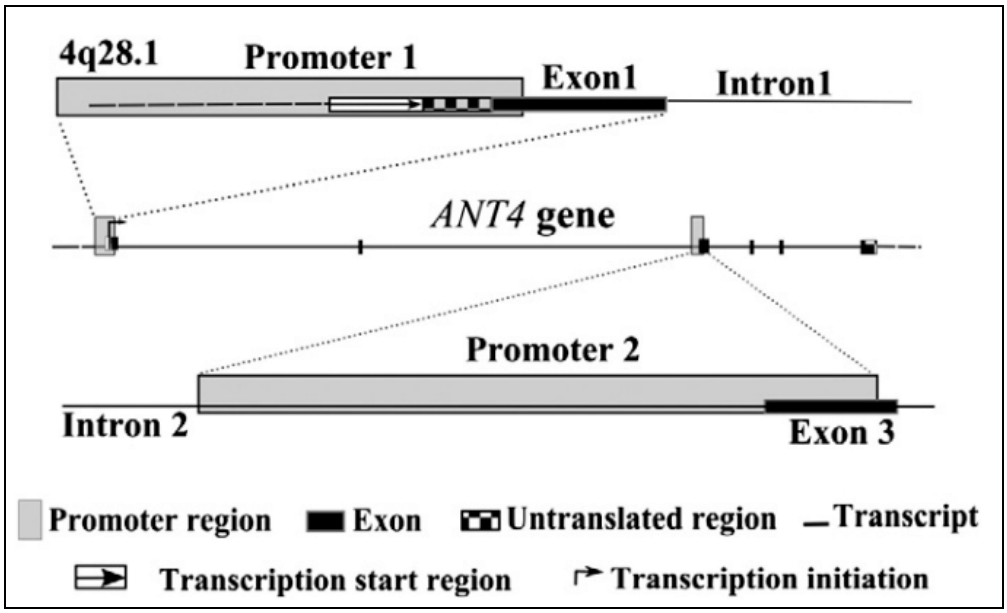


Figure 83 : Représentation schématique de la structure du gène ANT4
Tiré de [29]. Le promoteur connu est situé en amont du gène. Un deuxième promoteur a été identifié par PromoterInspector (Genomatix) à la fin du deuxième intron.

Une étude d'identité de séquence entre différentes parties de la séquence du gène ANT4 et de son promoteur a été effectuée entre l'homme et la souris. Les pourcentages d'identité de séquences sont présentés dans la figure 84.

	Human	Chimpanzee	Monkey	Dog	Ox	Dolphin	Mouse
H. Distal DNA	ID	**80.9**	39.9	31.3	31.3	28.7	27.3
M. Distal DNA	27.3	22.1	18.0	28.0	30.4	23.1	ID
H. Promoter 1	ID	**94.8**	**84.4**	52.6	45.1	61.7	35.0
M. Promoter 1	35.0	35.2	35.0	34.1	36.4	37.7	ID
H. Exon 1	ID	**98.7**	**96.1**	**84.0**	**84.8**	**81.9**	78.7
M. Exon 1	78.7	79.5	78.7	75.4	75.8	74.5	ID
H. Intron 1	ID	**96**.6	**90.4**	38.7	54.5	60.8	27.7
M. Intron 1	27.7	26.2	27.0	30.3	27.6	31.2	ID
H. Promoter 2	ID	**98.8**	**93.2**	56.4	55.1	60.6	43.1
M. Promoter 2	43.1	43.2	43.1	47.4	41.0	45.5	ID
H. Exon 3	ID	**100**	**99.1**	**94.0**	**95.7**	**97.4**	**90.6**
M. Exon 3	**90.6**	**90.6**	**91.5**	**88.1**	**91.5**	**90.6**	ID

Figure 84 : Pourcentage d'identité de séquences entre différents éléments de la séquence du gène ANT4

Cette étude d'identité a été effectuée pour comparer la ressemblance entre le promoteur identifié et d'autres éléments de la séquence. L'identité de séquence entre deux séquences non codantes pris à 10.000 nucléotides en amont du TSS sur l'alignement est de 22.1 % entre l'homme et la souris. L'identité de séquence des 500 premières bases du promoteur connu (en amont du premier exon) est

de 35.2 %, de 79.5 % au niveau du premier exon, , et de 26.2 % au niveau du premier intron (500 dernières bases) de. L'identité entre les séquences du second promoteur identifié par PromoterInspector est de 43.2 %. La conservation de cette région est donc bien supérieure à celle d'une région intronique classique ou même du promoteur connu. Il semblerait donc qu'une pression de sélection se soit opérée à cet endroit du gène. Afin de vérifier la validité de ce promoteur proposé par PromoterInspector, nous avons réalisé une étude de promotologie par le protocole décrit précédemment.

Les matrices et motifs nucléotidiques identifiées dans cette région sont : V$NKX25.02 (reconnue par le facteur à homéo-domaine Hkx-2.5/Csx), V$DLX1.01 (reconnue par des facteurs de transcription n'ayant pas d'homéo-domaine en partie distale) et V$HNF3B.02 (reconnue par facteur nucléaire hépatique 3 beta FOXA2), la TATAbox et un ATG. Différents modèles sont construits et les gènes à proximité des occurrences du modèle sont recherchés. Seuls 15 gènes connus sont trouvés, parmi eux, 5 seulement codent pour une protéine connue et aucune fonction commune à ces protéines n'a pu être identifiée. L'analyse n'a donc pas conduit à des résultats concluants dans le cas de ce second promoteur, vraisemblablement sans réelle fonction biologique. Le manuscrit complet de cet article est en annexe A.

4.2.2 Analyse promotologique de la régulation transcriptionnelle des quatre isoformes du gène *ANT*

Le même protocole que celui décrit pour l'étude de la régulation du gène *ANT* a été appliqué à l'étude de la régulation de la transcription des quatre isoformes du gène ANT. L'objectif de cette étude était de construire des modèles de régulation permettant d'identifier des gènes potentiellement co-régulés cohérents avec le rôle et l'expression de chacune des isoformes. Pour réaliser cette étude nous avons, comme précédemment, réalisé une étude phylogénétique des ANTS chez les mammifères. Nous avons tout d'abord réalisé une étude complète des quatre isoformes (cf. figure 85) puis une étude de chaque isoforme séparément. Au terme de cette étude, 8 séquences pour ANT1, 10 pour ANT2, 4 pour ANT3 et 7 pour ANT4 ont été retenues. Pour chaque isoforme, une série de matrices a été identifiée :

ANT1 (SLC25A4)

La plupart des matrices ou des séquences IUPAC trouvées pour les modèle ANT1 sont directement ou indirectement impliqués dans la croissance des cellules musculaires ou dans leur différenciation : V$MEF2 (reconnue par le facteur amplificateur myocyte-spécifique) est impliqué dans un mécanisme, conservé de la drosophile aux mammifères, permettant de réguler l'expression génique dans les muscles et probablement d'autres tissus [32,43]. La famille de matrices V$GATA, comprenant des facteurs de transcription comme GATA4 contrôlant la prolifération des cardiomyocytes par une régulation coordonnée de nombreux gènes du cycle cellulaire [101]**.** GATA4 régule notamment l'expression génique de PGC-1 alpha dans le muscle squelettique [49]. V$NRF1 est reconnue par le facteur respiratoire nucléaire 1, bZIP, qui contrôle l'expression de gènes nucléaires codant pour des protéines mitochondriales comme MEF2 (Myocyte Enhancer Factor 2A) [98]. V$NRF2.01, reconnue par le facteur nucléaire respiratoire 2 (ETS1) active la transcription de PDGF-A dans les cellules musculaires lisses [8]. V$KLFS est reconnu par les facteurs de transcription Krueppel-like, régulateurs essentiels de la différenciation cellulaire impliqués dans le développement du muscle squelettique et du muscle lisse [122]. V$STAT (transducteur de signal et activateur de la transcription) est impliqué dans une cascade de signalisation qui joue un rôle crucial dans la régulation de la myogenèse [127]. Le motif OXBOX est un élément positif connu pour être impliqué dans la transcription du gène ANT1 [68].

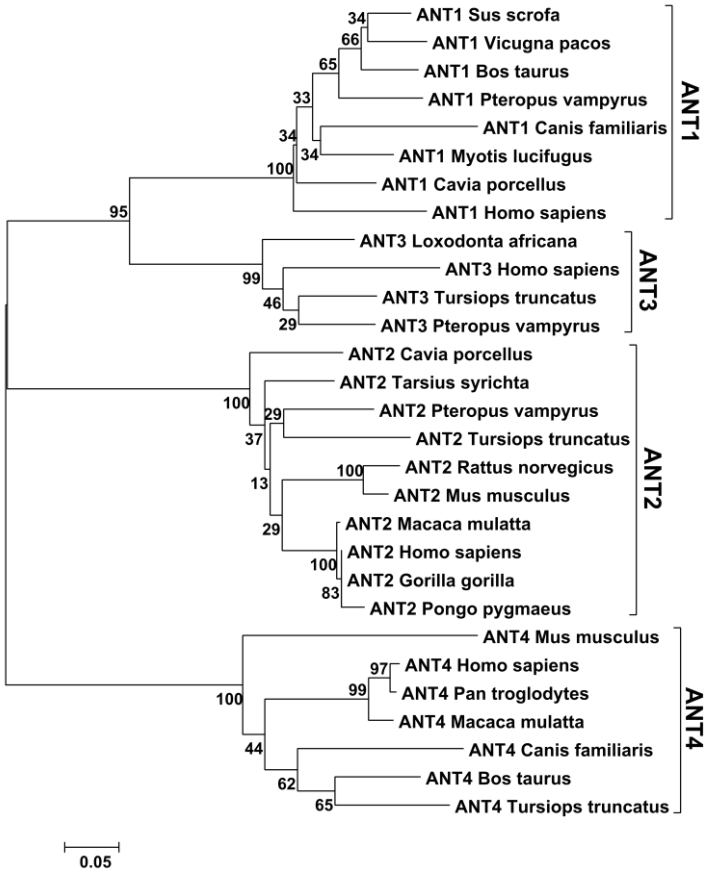


Figure 85 : Arbre phylogénétique montrant les relations phylogénétiques entre les gènes codant pour les quatre isoformes de l'ANT
Cet arbre a été reconstruit en utilisant la technique du neighbor-joining. Les scores appliqués sur les branches sont des valeurs de bootsrap exprimées en pourcentages. Les boostrap très élevés au niveau des branches séparant les différentes isoformes montrent que l'annotation de celles-ci est valide. L'échelle en bas à gauche correspond à une distance phylogénétique exprimée en nombre de substitution par site.

ANT2 (SLC25A5)

La plupart des matrices ou des séquences IUPAC trouvées pour les modèles ANT2 sont impliquées dans la croissance et la prolifération cellulaire : les gènes paralogues HOX 1 à 8 codent pour des facteurs de régulation se fixant sur V$HOXF et contrôlent le développement normal, des processus cellulaires primaires impliqués dans la carcinogenèse [19]. V$MZF1.01 (reconnue par des protéines en doigt de zinc myéloïde MZF1) est impliquée dans la prolifération cellulaire et le cancer [81]. V$EGR1.02 (EGR1, *Early Growth* Response *1*) est impliquée dans la régulation de l'homéostasie des cellules souches hématopoïétiques en contrôlant leur prolifération et leur migration [27]. V$SP1.01 et V$SP1.03 (*Stimulating Protein 1*), est connue pour réguler de manière coordonnée la lipogénèse et la prolifération des cellules cancéreuses [72]. La famille de matrices V$CEBP (CCAAT / *Enhancer Binding Proteins*) est connue pour jouer un rôle dans la croissance et la prolifération cellulaire [71]. La matrice V$MYT1 reconnue par le facteur de transcription Myt1, protéine de liaison à l'ADN à doigts de zinc qui régule la prolifération cellulaire et la différenciation des oligodendrocytes [82]. V$PARF (famille PAR/bZIP), comprenant la protéine à domaine PAR 1, PDP1, (régulateur du développement des larves chez la drosophile, de la mitose et de l'endoreplication) joue un rôle

essentiel dans la coordination de la croissance et de la réplication de l'ADN [100]. Les facteurs reconnaissant V$GATA permettent de contrôler le développement de nombreux tissus et sont impliqués dans différents mécanismes de la cancérogenèse [130].

ANT3 (SLC25A6)

Plusieurs fonctions cellulaires essentielles sont proposées pour les facteurs de transcription du gène ANT3 identifiés par cette analyse : V$CTCF (familles de des gènes CTCF et BORIS), V$CHRE (éléments de réponse aux glucides), V$RXRF (sites de reconnaissance des hétérodimères RXR) et V$RORA (*v-ERB and RAR-related orphan receptor alpha*). Cependant, contrairement aux trois gènes codant pour les trois autres isoformes d'ANT, la stringence des paramètres de recherche des modèles ANT3 sur le génome était trop faible pour conduire à des résultats concluants. Très peu de gènes sont identifiés comme potentiellement co-régulés avec ANT3 dans cette analyse.

ANT4 (SLC25A31)

La plupart des matrices ou des séquences IUPAC identifiées pour les modèle ANT4 sont impliquées dans le développement testiculaire, la spermatogenèse ou le métabolisme glycolytique. Les différentes matrices sont indiquées dans le paragraphe 4.2.

Pour chacun des gènes des 4 isoformes, des modèles sont construits en utilisant les matrices identifiées. Les distances entre les matrices sont fixées de manière à faire varier de plus ou moins 50 % de la distance observée chez l'homme. Les modèles sont alors recherchés sur l'ensemble du génome humain comme dans l'étude précédente. Les gènes identifiés à proximité des modèles montrent pour la plupart des fonctions cohérentes avec les fonctions et les profils d'expression connus des isoformes du gène *ANT*.

Les modèles construits pour le gène *ANT1* permettent l'identification de 28 gènes dont 11 ont des fonctions inconnues. 13 des 17 gènes restant sont spécifiquement exprimés dans le muscle ou sont impliqués dans le métabolisme bioénergétique. Avec les modèles construits pour ANT2, 56 gènes sont identifiés dont 23 codent pour une protéine dont la fonction est inconnue. Parmi les 33 gènes restant 17 sont impliqués dans des voies métaboliques diverses, non liées directement au métabolisme bioénergétique. 13 des 16 gènes restants sont impliqués soit dans la production d'énergie soit dans la prolifération cellulaire. Pour le gène *ANT3*, très peu de matrices ont pu être identifiées et les modèles construit n'ont permis de proposer que 12 gènes dont la moitié présente une fonction inconnue. Aucune fonction ubiquitaire connue n'a pu être trouvée pour les 6 gènes restants. Pour *ANT4*, 35 gènes ont été identifiés (les analyses du premier article ont été refaites après mise à jour des différentes bases de données utilisées, ce qui explique la différence du nombre de gènes identifiés). Parmi ces 35 gènes, 11 n'ont pas de fonction connue et 17 des 24 gènes restants sont spécifiquement exprimés dans les testicules, la prostate et au cours de la spermatogenèse. Les gènes identifiés sont décrits dans le paragraphe suivant. Une liste complète est fournie en annexe B.

4.3 DISCUSSION

Le protocole d'étude de promotologie que nous avons mis en place permet, via l'utilisation de la base de données Genomatix et de l'application web GeneProm que nous avons développée, de définir des éléments de régulation potentiels dans les promoteurs de gènes d'intérêt et d'identifier des gènes potentiellement co-régulés. Cette analyse a été conçue pour à terme prédire une signature précise d'une évolution métabolique cellulaire, qu'elle soit la résultante d'une évolution physiologique, d'une pathologie, ou d'une réponse à un traitement spécifique.

Ce protocole a été appliqué avec succès sur un ensemble de gènes codant pour les quatre isoformes d'une protéine, l'ANT (*Adenine Nucleotide Translocator*), chacune dotée d'un rôle précis

dans le métabolisme cellulaire et chacune liée à une spécificité cellulaire précise. Trois de ces protéines sont transcriptionnellement régulées par un mécanisme spécifique et la quatrième est l'isoforme ubiquiste exprimée de manière constitutive dans toutes les cellules. L'application de notre analyse promotologique à ce groupe de 4 gènes constitue ainsi une validation performante de notre stratégie.

ANT1. Parmi les 17 gènes identifiés contenant dans leur zone promotrice proximale un modèle spécifiquement construit à partir du promoteur du gène ANT1, 13 ont un lien direct avec le métabolisme de la cellule musculaire ou avec la bioénergétique mitochondriale. En particulier, 6 de ces gènes codent pour des protéines appartenant à 3 des complexes des phosphorylations oxydatives mitochondriales : NADH déshydrogénase (complexe I), cytochrome oxydase (complexe IV) et ATP synthase (complexe V) [66], complexes permettant la synthèse de l'ATP mitochondrial. De plus un autre gène, COQ7 intervient dans la synthèse de ces complexes [67]. D'autres gènes identifiés codent pour des protéines impliquées dans des voies majeures du métabolisme musculaire tel qu'ANO1 dans le transport du Ca++ [24], Myosine IF dans la contraction musculaire [104], SOD3 [26] et SLC35C2 [65] dans la réaction cellulaire aux ROS ou à l'hypoxie.

ANT2. La plupart des 33 gènes portant un modèle issu du gène ANT2 codent pour des protéines intervenant dans des voies liées à la division et prolifération cellulaire (AURKC, BTG1, FGL1, GDF15, NPPC, PHIP). Plusieurs autres gènes identifiés codent pour des protéines de signalisation telles que CDKN2AIP, GDD45B ou HIF1-alpha. Cette dernière protéine est connue pour induire la transcription du gène HKII, gène spécifiquement induit dans les conditions d'un métabolisme glycolytique et impliqué, comme ANT2, dans le transport de l'ATP glycolytique au travers de la membrane interne mitochondriale [17,39,121].

ANT3. À stringence similaire à celles utilisées pour la recherche des modèles construits pour les 3 autres isoformes, aucun gène portant de modèle construit à partir du promoteur du gène ANT3 n'a pu être identifié. Des analyses avec des stringences inférieures révèlent une dizaine de gènes dont la moitié (6) codent pour des protéines dont le rôle est encore inconnu, les 6 autres n'ayant pas entre eux de lien fonctionnel. De plus, l'analyse de la zone promotrice du gène ANT3 par l'outil PromoterInspector (Genomatix) montre la présence d'un nombre de matrices de régulation très faible par rapport aux gènes des 3 autres isoformes régulées.

ANT4. Les modèles réalisés à partir du gène ANT4 ont permis d'identifier 17 gènes à fonction connue et tous ont un rôle dans la spermatogenèse. Nos travaux antérieurs sur le promoteur de ce gène ANT4 avaient déjà permis d'identifier une fraction de ces gènes co-régulés [29]. La nouvelle version de notre logiciel GeneProm nous a permis d'identifier 5 nouveaux gènes également liés directement à la spermatogenèse. Parmi ces 5 gènes, deux, les glucose-6-phosphatases 2 et 3, exprimées spécifiquement dans les testicules [11], ont la fonction de produire du glucose à partir de glucose-6-phosphate avec ainsi génération d'ATP, fonction qui est tout à fait en accord non seulement avec le métabolisme exclusivement glycolytique du spermatozoïde mais également avec l'expression et le rôle spécifique de l'isoforme 4 dans la bioénergétique du spermatozoïde. Ainsi une partie de l'ATP produit par les glucose-6-phosphatases 2 et 3 pourrait être importée par l'ANT4 (gène localisé sur le chromosome 4) dans la mitochondrie pour pallier à l'absence de l'isoforme ANT2 (gène localisé sur le chromosome X dans une zone non transcrite au cours de la spermatogenèse [29].

Ainsi, cette analyse *in silico* permet de conduire à des conclusions très intéressantes sur les relations entre la régulation transcriptionnelle et la fonction de protéines dans un réseau métabolique cellulaire. Un groupe de gènes codant pour des isoformes d'une protéine exprimées de manière tissu spécifique s'est avéré être un matériel de travail très intéressant. L'analyse phylogénétique de promoteurs permet également d'identifier des matrices possédant une fonction majeure dans la régulation d'un gène. Cependant, l'utilisation exclusive de matrices de régulation a ses limites : elle nécessite de multiples analyses en fonction de différents paramètres de stringence et de position de chacune des matrices par rapport aux autres. La présence dans une zone promotrice de séquences nucléotidiques, décrites et validées, comme les séquences OXBOX [68] et GRBOX [39], connues

pour intervenir dans la régulation des gènes ANT1 (OXBOX) et ANT2 (GRBOX), est un facteur déterminant dans la construction d'un modèle performant, c'est-à-dire permettant d'identifier plusieurs gènes potentiellement co-régulés. Dans notre étude, la présence d'un gène codant pour une isoforme ubiquiste, l'ANT3, et l'absence d'identification de gènes co-régulés avec les modèles construits à partir de sa zone promotrice permettent de valider notre stratégie. De plus, cette analyse de gène ubiquiste non régulé servira de référence pour nos travaux ultérieurs en fournissant une échelle précise de similarité des matrices de modèles. Ainsi, une stratégie promotologique reposant à la fois sur une analyse phylogénétique concluante et sur des séquences régulatrices déjà validées permet une identification performante de gènes co-régulés.

Par ailleurs, notre stratégie est complémentaire à celle d'autres techniques utilisées dans l'étude de l'expression génique. Par exemple, pour notre modèle d'isoformes ANT, les analyses par puces à ADN sont encore peu adaptées : compte tenu de leur séquence très proche, il n'existe pas de puces commerciales capables de distinguer chacune des quatre isoformes humaines. Par ailleurs, le caractère très hydrophobe de ces protéines ne permet pas de les séparer et les identifier sur électrophorèse 2D et il n'existe pas d'anticorps spécifique de chacune de ces isoformes.

En conclusion, notre stratégie *in silico* sur un groupe de quatre isoformes, connues pour leurs fonctions spécifiques dans la bioénergétique cellulaire, nous a permis de mettre au point une analyse performante des zones promotrices de gènes. Notre analyse permet d'identifier un ensemble de gènes co-régulés, impliqués dans une même fonction cellulaire. Elle devrait apporter, en association avec la transcriptomique et la métabolomique, une aide majeure dans l'élaboration de réseaux métaboliques cellulaires et dans l'étude de leur régulation.

5 CONCLUSION ET PERSPECTIVES

Le travail de développement effectué au cours de cette thèse est résumé dans la figure 86

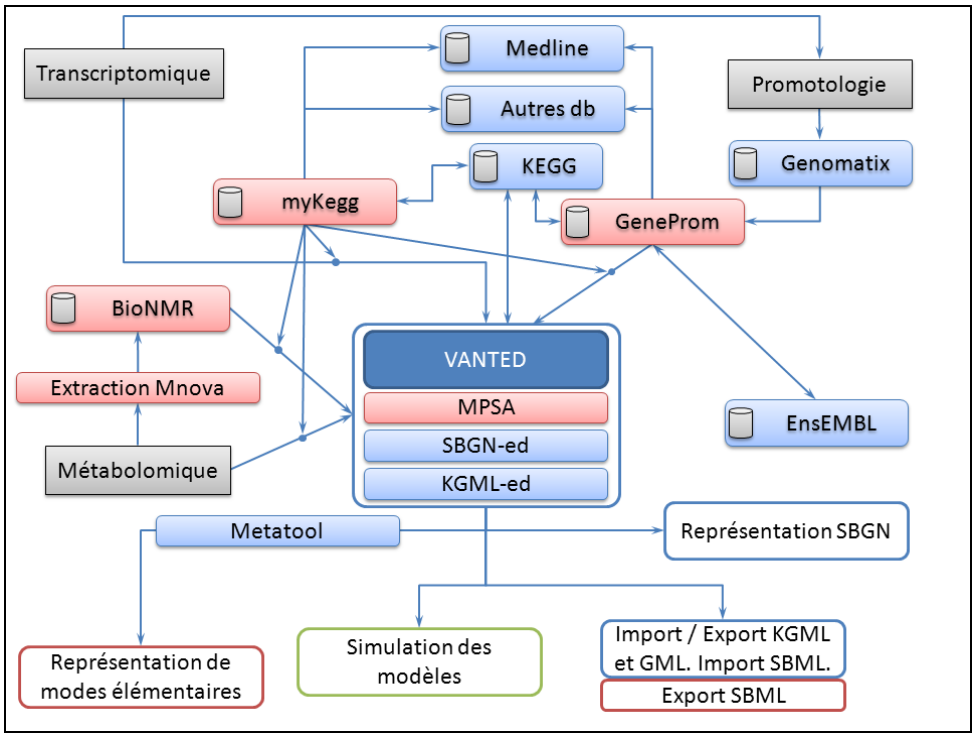


Figure 86 : Diagramme des outils bioinformatiques utilisés et développés pendant la thèse
Les techniques d'analyse biologiques sont représentées dans des cadres sur fond gris. Les logiciels sont représentés dans des cadres à fond plein (rouge si développés pendant la thèse, bleu si simplement utilisés). Les logiciels couplés à des bases de données présentent un cylindre à gauche de leur nom. Les fonctionnalités principales du logiciel VANTED sont représentés dans des cadres sur fond blanc (encadrés en rouge si la fonctionnalité a été développée pendant la thèse, en bleu sinon). La partie simulation est encadrée en vert car elle est en cours de développement en collaboration avec l'École Polytechnique.

Le logiciel central de cette thèse est le plugin MPSA (*Metabolic Pathways Software Analyzer*) pour VANTED. Ce plugin permet d'importer des données de différentes techniques d'analyse biologique en se connectant à la base de données myKegg. Cette base de données constitue l'application web centrale du développement effectué pendant cette thèse car elle permet une interconnexion entre les différents logiciels et base de données développés. Les logiciels utilisés vont dépendre de l'expérience biologique. L'utilisateur a toujours le choix d'utiliser le logiciel VANTED brut qui impose de dessiner un modèle de son réseau avant importation des données quantitatives.

MyKegg est une base de données permettant de trouver, à partir du nom d'une entité biologique (gène, protéine, métabolite,), les voies métaboliques dans lesquelles cette entité est impliquée. MyKegg reprend pour cela une partie de la structure de la base de données Kegg et ré-implémente la partie *bget* de son API. Elle permet donc de faire correspondre une liste de noms en fichiers KGML lisibles par VANTED. MyKegg permet aussi de trouver, pour chaque entrée de la base (gène,

métabolite, protéine…), des références dans d'autres bases de données de bioinformatique comme celles du NCBI ou de l'EMBL.

Aucun outil propre au traitement des données de transcriptomique n'a été développé car les seules données de transcriptomique ont été utilisées exclusivement pour initier des études de promotologie. L'utilisateur peut tout de même extraire manuellement les noms des gènes d'intérêt dans son expérience de transcriptomique au format csv et ainsi importer ses données dans le plugin MPSA. Cela lui permet de vérifier si ses gènes sont impliqués dans des voies de signalisation similaires.

Pour les expériences de métabolomique, l'utilisateur peut extraire manuellement les données de ses expériences pour fournir une liste de composés au format csv et les importer dans MPSA en utilisant myKegg. Il peut aussi, dans le cas d'une expérience de RMN utiliser la méthode d'extraction des pics implémentée dans Mnova pour enregistrer ses spectres RMN dans BioNMR. BioNMR est une base de données permettant d'organiser des expériences de RMN et d'enregistrer les spectres RMN de ces expériences. BioNMR possède un outil d'extraction des données au format csv qui permet d'importer les données dans MPSA en utilisant encore une fois myKegg.

Pour les études de promotologie, l'utilisateur utilise la base de données Genomatix et exporte les résultats vers l'application web GeneProm. Cette analyse permet de trouver une liste de gènes potentiellement régulés par le même mécanisme qu'un gène d'intérêt. Pour cela GeneProm se connecte à la base de données EnsEMBL. Il est possible d'exporter la liste de ces gènes depuis GeneProm au format csv pour être importée dans MPSA. Pour chaque gène trouvé, GeneProm donne une liste de références sur d'autres bases de données ainsi que des références bibliographiques. Le protocole d'étude de promotologie mis en place au cours de cette thèse permet de trouver des éléments de régulation potentiels dans les promoteurs de gènes d'intérêt et d'en déduire des modèles de régulation. Il permet aussi de découvrir un ensemble de gènes potentiellement co-régulés avec le gène d'intérêt. Ce protocole a été appliqué avec succès à l'étude de la régulation des gènes codant pour les isoformes de la protéine ANT ayant des profils d'expression différents. *ANT1* est exprimée dans la cellule musculaire, *ANT2* dans les cellules en prolifération, *ANT3* est l'isoforme ubiquitaire et *ANT4* est exprimée dans les spermatozoïdes et lors de la spermatogenèse. Ce protocole a permis de mettre en évidence, pour chaque isoforme, un ensemble de matrices de régulation dont le rôle connu est cohérent avec celui de chaque isoforme ANT. La majorité des gènes identifiés en utilisant les modèles basés sur le promoteur d'*ANT1* sont impliqués dans le métabolisme de la cellule musculaire ou dans la bioénergétique mitochondriale plus particulièrement dans les complexes I et IV de la chaîne respiratoire. Les gènes identifiés avec les modèles issus d'*ANT2* sont, pour la plupart, exprimés dans les cellules en prolifération. Très peu de matrices de régulation et de gènes sont retrouvés pour le gène *ANT3*, ce qui est cohérent avec son expression ubiquiste. Pour le gène *ANT4*, la majorité des gènes retrouvés sont liés à la spermatogenèse.

L'import de données dans MPSA permet d'étudier les réseaux dans lesquels les composés sont impliqués. Il est possible d'étudier les réseaux de deux manières : d'une part d'analyser leur structure via l'étude des modes élémentaires et d'autre part de réaliser des simulations de la dynamique de ces réseaux en fonction d'équations cinétiques. Il est aussi possible d'exporter ces réseaux dans différents formats standards permettant de les étudier par d'autres logiciels plus spécialisés : le GML pour l'export vers des logiciels de représentation et d'étude des graphes et le SBML pour lequel MPSA permet l'export (en plus de l'import prévu dans VANTED) pour des logiciels de biologie des systèmes.

L'ensemble du développement et des études effectué au cours de cette thèse ont été validés par des applications sur des thèmes de recherche poursuivis dans l'Unité de Nutrition Humaine. Ce travail pourrait être intégré à la plateforme de biologie intégrative en projet au sein de l'unité. Les fonctionnalités de cette plateforme permettront le croisement de données issues d'expérimentations sur différents modèles biologiques, réalisées par différentes techniques d'analyse biologique.

6 BIBLIOGRAPHIE

[**1**] « KEGG Pathway hsa04010 ». Nature Publishing Group, 15/7/2009.

[**2**] « Systems Biology Markup Language (SBML) Level 2: Structures and Facilities for Model Definitions : Nature Precedings ». .

[**3**] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, et G. Sherlock, « Gene ontology: tool for the unification of biology. The Gene Ontology Consortium », *Nature Genetics*, vol. 25, n°. 1, p. 25-29, mai 2000.

[**4**] G. D. Bader, M. P. Cary, et C. Sander, « Pathguide: a pathway resource list », *Nucleic Acids Research*, vol. 34, p. D504-506, janvier 2006.

[**5**] P. Barath, K. Luciakova, Z. Hodny, R. Li, et B. D. Nelson, « The growth-dependent expression of the adenine nucleotide translocase-2 (ANT2) gene is regulated at the level of transcription and is a marker of cell proliferation », *Experimental Cell Research*, vol. 248, n°. 2, p. 583-588, mai 1999.

[**6**] M. Behzadi, A. Demidem, D. Morvan, L. Schwartz, G. Stepien, et J.-M. Steyaert, « A model of phospholipid biosynthesis in tumor in response to an anticancer agent in vivo », *Journal of Integrative Bioinformatics*, vol. 7, n°. 3, 2010.

[**7**] M. Blanchette, A. R. Bataille, X. Chen, C. Poitras, J. Laganière, C. Lefèbvre, G. Deblois, V. Giguère, V. Ferretti, D. Bergeron, B. Coulombe, et F. Robert, « Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression », *Genome Research*, vol. 16, n°. 5, p. 656-668, mai 2006.

[**8**] M. R. Bonello, Y. V. Bobryshev, et L. M. Khachigian, « Peroxide-inducible Ets-1 mediates platelet-derived growth factor receptor-alpha gene transcription in vascular smooth muscle cells », *The American Journal of Pathology*, vol. 167, n°. 4, p. 1149-1159, octobre 2005.

[**9**] J. V. Brower, N. Rodic, T. Seki, M. Jorgensen, N. Fliess, A. T. Yachnis, J. R. McCarrey, S. P. Oh, et N. Terada, « Evolutionarily conserved mammalian adenine nucleotide translocase 4 is essential for spermatogenesis », *J. Biol. Chem.*, p. M704386200, août 2007.

[**10**] P. Bucher, « Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences », *Journal of Molecular Biology*, vol. 212, n°. 4, p. 563-578, avril 1990.

[**11**] A. Burchell, S. L. Watkins, et R. Hume, « Human fetal testis endoplasmic reticulum glucose-6-phosphatase enzyme protein », *Biology of Reproduction*, vol. 55, n°. 2, p. 298-303, août 1996.

[**12**] L. Calzone, F. Fages, et S. Soliman, « BIOCHAM: an environment for modeling biological systems and formalizing experimental knowledge », *Bioinformatics (Oxford, England)*, vol. 22, n°. 14, p. 1805-1807, juillet 2006.

[**13**] K. Cartharius, K. Frech, K. Grote, B. Klocke, M. Haltmeier, A. Klingenhoff, M. Frisch, M. Bayerlein, et T. Werner, « MatInspector and beyond: promoter analysis based on transcription factor binding sites », *Bioinformatics*, vol. 21, n°. 13, p. 2933-2942, juillet 2005.

[**14**] R. Caspi, T. Altman, J. M. Dale, K. Dreher, C. A. Fulcher, F. Gilham, P. Kaipa, A. S. Karthikeyan, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, S. Paley, L. Popescu, A. Pujar, A. G. Shearer, P. Zhang, et P. D. Karp, « The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases », *Nucleic Acids Research*, vol. 38, p. D473-D479, octobre 2009.

[**15**] C.-Y. Chen, S.-T. Chen, C.-S. Fuh, H.-F. Juan, et H.-C. Huang, « Coregulation of transcription factors and microRNAs in human transcriptional regulatory network », *BMC Bioinformatics*, vol. 12 Suppl 1, p. S41, 2011.

[**16**] A. Chevrollier, D. Loiseau, B. Chabi, G. Renier, O. Douay, Y. Malthièry, et G. Stepien, « ANT2 isoform required for cancer cell glycolysis », *Journal of Bioenergetics and Biomembranes*, vol. 37, n°. 5, p. 307-16, octobre 2005.

[**17**] A. Chevrollier, D. Loiseau, B. Chabi, G. Renier, O. Douay, Y. Malthièry, et G. Stepien, « ANT2 isoform required for cancer cell glycolysis », *Journal of Bioenergetics and Biomembranes*, vol. 37, n°. 5, p. 307-316, octobre 2005.

[**18**] V. S. Chopra, « Chromosomal organization at the level of gene complexes », *Cellular and Molecular Life Sciences: CMLS*, vol. 68, n°. 6, p. 977-990, mars 2011.

[**19**] C. Cillo, G. Schiavo, M. Cantile, M. P. Bihl, P. Sorrentino, V. Carafa, M. D' Armiento, M. Roncalli, S. Sansano, R. Vecchione, L. Tornillo, L. Mori, G. De Libero, J. Zucman-Rossi, et L. Terracciano, « The HOX gene network in hepatocellular carcinoma », *International Journal of Cancer. Journal International Du Cancer*, vol. 129, n°. 11, p. 2577-2587, décembre 2011.

[**20**] D. G. Covell, A. Wallqvist, A. A. Rabow, et N. Thanki, « Molecular Classification of Cancer: Unsupervised Self-Organizing Map Analysis of Gene Expression Microarray Data », *Molecular Cancer Therapeutics*, vol. 2, n°. 3, p. 317 -332, mars 2003.

[**21**] M. W. Covert, C. H. Schilling, I. Famili, J. S. Edwards, I. I. Goryanin, E. Selkov, et B. O. Palsson, « Metabolic modeling of microbial strains in silico », *Trends in Biochemical Sciences*, vol. 26, n°. 3, p. 179-186, mars 2001.

[**22**] D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, S. Jupe, I. Kalatskaya, S. Mahajan, B. May, N. Ndegwa, E. Schmidt, V. Shamovsky, C. Yung, E. Birney, H. Hermjakob, P. D'Eustachio, et L. Stein, « Reactome: a database of reactions, pathways and biological processes », *Nucleic Acids Research*, vol. 39, n°. Database issue, p. D691-697, janvier 2011.

[**23**] T. Czauderna, C. Klukas, et F. Schreiber, « Editing, validating and translating of SBGN maps », *Bioinformatics*, vol. 26, n°. 18, p. 2340 -2341, 2010.

[**24**] A. J. Davis, A. S. Forrest, T. A. Jepps, M. L. Valencik, M. Wiwchar, C. A. Singer, W. R. Sones, I. A. Greenwood, et N. Leblanc, « Expression profile and protein translation of TMEM16A in murine smooth muscle », *American Journal of Physiology. Cell Physiology*, vol. 299, n°. 5, p. C948-959, novembre 2010.

[**25**] A. Deckard, F. T. Bergmann, et H. M. Sauro, « Supporting the SBML layout extension », *Bioinformatics*, vol. 22, n°. 23, p. 2966 -2967, décembre 2006.

[**26**] E. D. van Deel, Z. Lu, X. Xu, G. Zhu, X. Hu, T. D. Oury, R. J. Bache, D. J. Duncker, et Y. Chen, « Extracellular superoxide dismutase protects the heart against oxidative stress and hypertrophy after myocardial infarction », *Free Radical Biology & Medicine*, vol. 44, n°. 7, p. 1305-1313, avril 2008.

[**27**] J. T. DeLigio et D. A. R. Zorio, « Early growth response 1 (EGR1): a gene with as many names as biological functions », *Cancer Biology & Therapy*, vol. 8, n°. 20, p. 1889-1892, octobre 2009.

[**28**] V. Dolce, P. Scarcia, D. Iacopetta, et F. Palmieri, « A fourth ADP/ATP carrier isoform in man: identification, bacterial expression, functional characterization and tissue distribution », *FEBS Letters*, vol. 579, n°. 3, p. 633-637, janvier 2005.

[**29**] P.-Y. Dupont et G. Stepien, « Computational analysis of the transcriptional regulation of the adenine nucleotide translocator isoform 4 gene and its role in spermatozoid glycolytic metabolism », *Gene*, vol. 487, n°. 1, p. 38-45, novembre 2011.

[**30**] M. B. Elowitz et S. Leibler, « A synthetic oscillatory network of transcriptional regulators », *Nature*, vol. 403, n°. 6767, p. 335-338, janvier 2000.

[**31**] M. F. Favata, K. Y. Horiuchi, E. J. Manos, A. J. Daulerio, D. A. Stradley, W. S. Feeser, D. E. Van Dyk, W. J. Pitts, R. A. Earl, F. Hobbs, R. A. Copeland, R. L. Magolda, P. A. Scherle, et J. M. Trzaskos, « Identification of a novel inhibitor of mitogen-activated protein kinase kinase », *The Journal of Biological Chemistry*, vol. 273, n°. 29, p. 18623-18632, juillet 1998.

[**32**] H. Feng, T. Cheng, J. H. Steer, D. A. Joyce, N. J. Pavlos, C. Leong, J. Kular, J. Liu, X. Feng, M. H. Zheng, et J. Xu, « Myocyte enhancer factor 2 and microphthalmia-associated transcription factor cooperate with NFATc1 to transactivate the V-ATPase d2 promoter during RANKL-induced osteoclastogenesis », *The Journal of Biological Chemistry*, vol. 284, n°. 21, p. 14667-14676, mai 2009.

[**33**] V. Ferretti, C. Poitras, D. Bergeron, B. Coulombe, F. Robert, et M. Blanchette, « PReMod: a database of genome-wide mammalian cis-regulatory module predictions », *Nucleic Acids Research*, vol. 35, n°. Database issue, p. D122-126, janvier 2007.

[**34**] L. F. de Figueiredo, A. Podhorski, A. Rubio, C. Kaleta, J. E. Beasley, S. Schuster, et F. J. Planes, « Computing the shortest elementary flux modes in genome-scale metabolic networks », *Bioinformatics*, vol. 25, n°. 23, p. 3158 -3165, décembre 2009.

[**35**] G. B. Fogel, D. G. Weekes, G. Varga, E. R. Dow, A. M. Craven, H. B. Harlow, E. W. Su, J. E. Onyia, et C. Su, « A statistical analysis of the TRANSFAC database », *Bio Systems*, vol. 81, n°. 2, p. 137-54, août 2005.

[**36**] H. Fukazawa, K. Noguchi, Y. Murakami, et Y. Uehara, « Mitogen-activated Protein/Extracellular Signal-regulated Kinase Kinase (MEK) Inhibitors Restore Anoikis Sensitivity in Human Breast Cancer Cell Lines with a Constitutively Activated Extracellular- regulated Kinase (ERK) Pathway », *Molecular Cancer Therapeutics*, vol. 1, n°. 5, p. 303 -309, mars 2002.

[**37**] D. J. Gaffney, R. Blekhman, et J. Majewski, « Selective constraints in experimentally defined primate regulatory regions », *PLoS Genetics*, vol. 4, n°. 8, p. e1000157, 2008.

[**38**] R. Gauges, U. Rost, S. Sahle, et K. Wegner, « A model diagram layout extension for SBML », *Bioinformatics*, vol. 22, n°. 15, p. 1879 -1885, 2006.

[**39**] S. Giraud, C. Bonod-Bidaud, M. Wesolowski-Louvel, et G. Stepien, « Expression of human ANT2 gene in highly proliferative cells: GRBOX, a new transcriptional element, is involved in the regulation of glycolytic ATP import into mitochondria », *Journal of Molecular Biology*, vol. 281, n°. 3, p. 409-418, août 1998.

[**40**] S. Giraud, C. Bonod-Bidaud, M. Wesolowski-Louvel, et G. Stepien, « Expression of human ANT2 gene in highly proliferative cells: GRBOX, a new transcriptional element, is involved in the regulation of glycolytic ATP import into mitochondria », *Journal of Molecular Biology*, vol. 281, n°. 3, p. 409-418, août 1998.

[**41**] N. Goto, P. Prins, M. Nakao, R. Bonnal, J. Aerts, et T. Katayama, « BioRuby: bioinformatics software for the Ruby programming language », *Bioinformatics*, vol. 26, n°. 20, p. 2617 -2619, octobre 2010.

[**42**] V. Govindaraju, K. Young, et A. A. Maudsley, « Proton NMR chemical shifts and coupling constants for brain metabolites », *NMR in Biomedicine*, vol. 13, n°. 3, p. 129-153, mai 2000.

[**43**] L. Guerrero, R. Marco-Ferreres, A. L. Serrano, J. J. Arredondo, et M. Cervera, « Secondary enhancers synergise with primary enhancers to guarantee fine-tuned muscle gene expression », *Developmental Biology*, vol. 337, n°. 1, p. 16-28, janvier 2010.

[**44**] P. Holleis, T. Zimmermann, et D. Gmach, « Drawing graphs within graphs », *Journal of Graph Algorithms and Applications*, vol. 9, n°. 1, p. 7–18, 2005.

[**45**] M. Hucka, M. Hucka, F. Bergmann, S. Hoops, S. Keating, S. Sahle, et D. Wilkinson, « The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 1 Core (Release 1 Candidate) », *Nature Precedings*, janvier 2010.

[**46**] T. Ideker, T. Galitski, et L. Hood, « A new approach to decoding life: systems biology », *Annual Review of Genomics and Human Genetics*, vol. 2, p. 343-372, 2001.

[**47**] M. van Iersel, M. van Iersel, S. Boyd, F. Bergmann, S. Moodie, F. Schreiber, T. Czauderna, E. Demir, N. Le Novère, A. Sorokin, H. Mi, A. Luna, U. Dogrusoz, Y. Matsuoka, A. Funahashi, H. Kitano, M. Aladjem, M. Blinov, et A. Villéger, « LibSBGN: Electronic Processing of SBGN maps », *Nature Precedings*, octobre 2010.

[**48**] M. van Iersel, T. Kelder, A. Pico, K. Hanspers, S. Coort, B. Conklin, et C. Evelo, « Presenting and exploring biological pathways with PathVisio », *BMC Bioinformatics*, vol. 9, n°. 1, p. 399, 2008.

[**49**] I. Irrcher, V. Ljubicic, A. F. Kirwan, et D. A. Hood, « AMP-activated protein kinase-regulated activation of the PGC-1alpha promoter in skeletal muscle cells », *PloS One*, vol. 3, n°. 10, p. e3614, 2008.

[**50**] E. Jain, A. Bairoch, S. Duvaud, I. Phan, N. Redaschi, B. Suzek, M. Martin, P. McGarvey, et E. Gasteiger, « Infrastructure for the life sciences: design and implementation of the UniProt website », *BMC Bioinformatics*, vol. 10, n°. 1, p. 136, 2009.

[**51**] D. Jevremovic, C. T. Trinh, F. Srienc, C. P. Sosa, et D. Boley, « Parallelization of Nullspace Algorithm for the computation of metabolic pathways », *Parallel Computing*, vol. In Press, Accepted Manuscript, 08:14:02.

[**52**] B. Junker, C. Klukas, et F. Schreiber, « VANTED: A system for advanced data analysis and visualization in the context of biological networks », *BMC Bioinformatics*, vol. 7, n°. 1, p. 109, 2006.

[**53**] A. von Kamp et S. Schuster, « Metatool 5.0: fast and flexible elementary modes analysis », *Bioinformatics (Oxford, England)*, vol. 22, n°. 15, p. 1930-1931, août 2006.

[**54**] S. M. Kehoe, M. Oka, K. E. Hankowski, N. Reichert, S. Garcia, J. R. McCarrey, S. Gaubatz, et N. Terada, « A conserved E2F6-binding element in murine meiosis-specific gene promoters », *Biology of Reproduction*, vol. 79, n°. 5, p. 921-930, novembre 2008.

[**55**] W. J. Kent, « BLAT--the BLAST-like alignment tool », *Genome Research*, vol. 12, n°. 4, p. 656-664, avril 2002.

[**56**] N.-K. Kim, K. Tharakaraman, L. Mariño-Ramírez, et J. L. Spouge, « Finding sequence motifs with Bayesian models incorporating positional information: an application to transcription factor binding sites », *BMC Bioinformatics*, vol. 9, p. 262, 2008.

[**57**] Y.-H. Kim, G. Haidl, M. Schaefer, U. Egner, A. Mandal, et J. C. Herr, « Compartmentalization of a unique ADP/ATP carrier protein SFEC (Sperm Flagellar Energy Carrier, AAC4) with glycolytic enzymes in the fibrous sheath of the human sperm flagellar principal piece », *Developmental Biology*, vol. 302, n°. 2, p. 463-476, février 2007.

[**58**] S. Klamt et J. Stelling, « Two approaches for metabolic pathway analysis? », *Trends in Biotechnology*, vol. 21, n°. 2, p. 64-69, février 2003.

[**59**] S. Klamt et J. Stelling, « Combinatorial complexity of pathway analysis in metabolic networks », *Molecular Biology Reports*, vol. 29, n°. 1-2, p. 233-236, 2002.

[**60**] S. Klamt, J. Saez-Rodriguez, et E. Gilles, « Structural and functional analysis of cellular networks with CellNetAnalyzer », *BMC Systems Biology*, vol. 1, n°. 1, p. 2, 2007.

[**61**] S. Klamt, J. Stelling, M. Ginkel, et E. D. Gilles, « FluxAnalyzer: exploring structure, pathways, and flux distributions in metabolic networks on interactive flux maps », *Bioinformatics*, vol. 19, n°. 2, p. 261 - 269, janvier 2003.

[**62**] A. Klingenhoff, K. Frech, K. Quandt, et T. Werner, « Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity », *Bioinformatics (Oxford, England)*, vol. 15, n°. 3, p. 180-186, mars 1999.

[**63**] C. Klukas et F. Schreiber, « Dynamic exploration and editing of KEGG pathway diagrams », *Bioinformatics*, vol. 23, n°. 3, p. 344-350, février 2007.

[**64**] D. H. Ku, J. Kagan, S. T. Chen, C. D. Chang, R. Baserga, et J. Wurzel, « The human fibroblast adenine nucleotide translocator gene. Molecular cloning and sequence », *The Journal of Biological Chemistry*, vol. 265, n°. 27, p. 16060-16063, septembre 1990.

[**65**] R. E. Leach, Z. M. Duniec-Dmuchowski, G. Pesole, T. S. Tanaka, M. S. H. Ko, D. R. Armant, et S. A. Krawetz, « Identification, molecular characterization, and tissue expression of OVCOV1 », *Mammalian Genome: Official Journal of the International Mammalian Genome Society*, vol. 13, n°. 11, p. 619-624, novembre 2002.

[**66**] G. Lenaz et M. L. Genova, « Structural and functional organization of the mitochondrial respiratory chain: a dynamic super-assembly », *The International Journal of Biochemistry & Cell Biology*, vol. 41, n°. 10, p. 1750-1772, octobre 2009.

[**67**] F. Levavasseur, H. Miyadera, J. Sirois, M. L. Tremblay, K. Kita, E. Shoubridge, et S. Hekimi, « Ubiquinone is necessary for mouse embryonic development but is not essential for mitochondrial respiration », *The Journal of Biological Chemistry*, vol. 276, n°. 49, p. 46160-46164, décembre 2001.

[**68**] K. Li, J. A. Hodge, et D. C. Wallace, « OXBOX, a positive transcriptional element of the heart-skeletal muscle ADP/ATP translocator gene », *The Journal of Biological Chemistry*, vol. 265, n°. 33, p. 20585-20588, novembre 1990.

[**69**] C. M. Lloyd, M. D. B. Halstead, et P. F. Nielsen, « CellML: its future, present and past », *Progress in Biophysics and Molecular Biology*, vol. 85, n°. 2-3, p. 433-450, juin.

[**70**] K. Luciakova, P. Barath, D. Poliakova, A. Persson, et B. D. Nelson, « Repression of the human adenine nucleotide translocase-2 gene in growth-arrested human diploid cells: the role of nuclear factor-1 », *The Journal of Biological Chemistry*, vol. 278, n°. 33, p. 30624-30633, août 2003.

[**71**] G.-D. Lu, C. H.-W. Leung, B. Yan, C. M.-Y. Tan, S. Y. Low, M. O. Aung, M. Salto-Tellez, S. G. Lim, et S. C. Hooi, « C/EBPalpha is up-regulated in a subset of hepatocellular carcinomas and plays a role in cell growth and proliferation », *Gastroenterology*, vol. 139, n°. 2, p. 632-643, 643.e1-4, août 2010.

[**72**] S. Lu et M. C. Archer, « Sp1 coordinately regulates de novo lipogenesis and proliferation in cancer cells », *International Journal of Cancer. Journal International Du Cancer*, vol. 126, n°. 2, p. 416-425, janvier 2010.

[**73**] M. C. Martínez-Bisbal, L. Martí-Bonmatí, J. Piquer, A. Revert, P. Ferrer, J. L. Llácer, M. Piotto, O. Assemat, et B. Celda, « 1H and 13C HR-MAS spectroscopy of intact biopsy samples ex vivo and in vivo 1H MRS study of human high grade gliomas », *NMR in Biomedicine*, vol. 17, n°. 4, p. 191-205, juin 2004.

[**74**] L. Matthews, G. Gopinath, M. Gillespie, M. Caudy, D. Croft, B. de Bono, P. Garapati, J. Hemish, H. Hermjakob, B. Jassal, A. Kanapin, S. Lewis, S. Mahajan, B. May, E. Schmidt, I. Vastrik, G. Wu, E. Birney, L. Stein, et P. D'Eustachio, « Reactome knowledgebase of human biological pathways and processes », *Nucleic Acids Research*, vol. 37, n°. Database issue, p. D619-622, janvier 2009.

[**75**] P. Mendes, « Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3 », *Trends in Biochemical Sciences*, vol. 22, n°. 9, p. 361-363, septembre 1997.

[**76**] P. Mendes, « GEPASI: a software package for modelling the dynamics, steady states and control of biochemical and other systems », *Computer Applications in the Biosciences: CABIOS*, vol. 9, n°. 5, p. 563-571, octobre 1993.

[**77**] P. Mendes et D. Kell, « Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation », *Bioinformatics (Oxford, England)*, vol. 14, n°. 10, p. 869-883, 1998.

[**78**] M. Mesiti, E. Jimenez-Ruiz, I. Sanz, R. Berlanga-Llavori, P. Perlasca, G. Valentini, et D. Manset, « XML-based approaches for the integration of heterogeneous bio-molecular data », *BMC Bioinformatics*, vol. 10, n°. 12, p. S7, 2009.

[**79**] H. Mi, F. Schreiber, N. Le Novére, S. Moodie, et A. Sorokin, « Systems Biology Graphical Notation: Activity Flow language Level 1 », *Nature Precedings*, septembre 2009.

[**80**] S. Moodie, N. Le Novere, A. Sorokin, H. Mi, et F. Schreiber, « Systems Biology Graphical Notation: Process Description language Level 1 », *Nature Precedings*, septembre 2009.

[**81**] G. Mudduluru, P. Vajkoczy, et H. Allgayer, « Myeloid zinc finger 1 induces migration, invasion, and in vivo metastasis through Axl gene expression in solid cancer », *Molecular Cancer Research: MCR*, vol. 8, n°. 2, p. 159-169, février 2010.

[**82**] J. A. Nielsen, J. A. Berndt, L. D. Hudson, et R. C. Armstrong, « Myelin transcription factor 1 (Myt1) modulates the proliferation and differentiation of oligodendrocyte lineage cells », *Molecular and Cellular Neurosciences*, vol. 25, n°. 1, p. 111-123, janvier 2004.

[**83**] N. L. Novere, A. Finney, M. Hucka, U. S. Bhalla, F. Campagne, J. Collado-Vides, E. J. Crampin, M. Halstead, E. Klipp, P. Mendes, P. Nielsen, H. Sauro, B. Shapiro, J. L. Snoep, H. D. Spence, et B. L. Wanner, « Minimum information requested in the annotation of biochemical models (MIRIAM) », *Nat Biotech*, vol. 23, n°. 12, p. 1509-1515, décembre 2005.

[**84**] N. L. Novere, M. Hucka, H. Mi, S. Moodie, F. Schreiber, A. Sorokin, E. Demir, K. Wegner, M. I. Aladjem, S. M. Wimalaratne, F. T. Bergman, R. Gauges, P. Ghazal, H. Kawaji, L. Li, Y. Matsuoka, A. Villeger, S. E. Boyd, L. Calzone, M. Courtot, U. Dogrusoz, T. C. Freeman, A. Funahashi, S. Ghosh, A. Jouraku, S. Kim, F. Kolpakov, A. Luna, S. Sahle, E. Schmidt, S. Watterson, G. Wu, I. Goryanin, D. B. Kell, C. Sander, H. Sauro, J. L. Snoep, K. Kohn, et H. Kitano, « The Systems Biology Graphical Notation », *Nat Biotech*, vol. 27, n°. 8, p. 735-741, 2009.

[**85**] N. Le Novere, N. Le Novere, E. Demir, H. Mi, S. Moodie, et A. Villeger, « Systems Biology Graphical Notation: Entity Relationship language Level 1 (Version 1.2) », *Nature Precedings*, avril 2011.

[**86**] S. G. Oliver, M. K. Winson, D. B. Kell, et F. Baganz, « Systematic functional analysis of the yeast genome », *Trends in Biotechnology*, vol. 16, n°. 9, p. 373-378, septembre 1998.

[**87**] B. G. Olivier, J. M. Rohwer, et J.-H. S. Hofmeyr, « Modelling cellular systems with PySCeS », *Bioinformatics*, vol. 21, n°. 4, p. 560-561, février 2005.

[**88**] B. Palsson, « The challenges of in silico biology », *Nat Biotech*, vol. 18, n°. 11, p. 1147-1150, novembre 2000.

[**89**] J. A. Papin, N. D. Price, J. S. Edwards, et B. Ø. Palsson B, « The genome-scale metabolic extreme pathway structure in Haemophilus influenzae shows significant network redundancy », *Journal of Theoretical Biology*, vol. 215, n°. 1, p. 67-82, mars 2002.

[**90**] J. A. Papin, N. D. Price, S. J. Wiback, D. A. Fell, et B. O. Palsson, « Metabolic pathways in the post-genome era », *Trends in Biochemical Sciences*, vol. 28, n°. 5, p. 250-258, mai 2003.

[**91**] G. A. Pavlopoulos, M. Secrier, C. N. Moschopoulos, T. G. Soldatos, S. Kossida, J. Aerts, R. Schneider, et P. G. Bagos, « Using graph theory to analyze biological networks », *BioData Mining*, vol. 4, n°. 1, p. 10, 2011.

[**92**] S. Pérès, M. Beurton-Aimar, et J. P. Mazat, « Pathway classification of TCA cycle », *Systems Biology*, vol. 153, n°. 5, p. 369-371, septembre 2006.

[**93**] S. Pérès, F. Vallée, M. Beurton-Aimar, et J. P. Mazat, « ACoM: A classification method for elementary flux modes based on motif finding », *Bio Systems*, vol. 103, n°. 3, p. 410-419, mars 2011.

[**94**] T. Pfeiffer, I. Sánchez-Valdenebro, J. C. Nuño, F. Montero, et S. Schuster, « METATOOL: for studying metabolic networks », *Bioinformatics (Oxford, England)*, vol. 15, n°. 3, p. 251-257, mars 1999.

[**95**] N. D. Price, J. A. Papin, et B. Ø. Palsson, « Determination of Redundancy and Systems Properties of the Metabolic Network of Helicobacter pylori Using Genome-Scale Extreme Pathway Analysis », *Genome Research*, vol. 12, n°. 5, p. 760-769, mai 2002.

[**96**] L. K. Purvis et R. J. Butera, « Ionic current model of a hypoglossal motoneuron », *Journal of Neurophysiology*, vol. 93, n°. 2, p. 723-733, février 2005.

[**97**] K. Quandt, K. Frech, H. Karas, E. Wingender, et T. Werner, « MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. », *Nucleic Acids Research*, vol. 23, n°. 23, p. 4878–4884, décembre 1995.

[**98**] B. Ramachandran, G. Yu, et T. Gulick, « Nuclear respiratory factor 1 controls myocyte enhancer factor 2A transcription to provide a mechanism for coordinate expression of respiratory chain subunits », *The Journal of Biological Chemistry*, vol. 283, n°. 18, p. 11935-11946, mai 2008.

[**99**] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, et D. Lancet, « GeneCards: integrating information about genes, proteins and diseases », *Trends in Genetics: TIG*, vol. 13, n°. 4, p. 163, avril 1997.

[**100**] K. L. Reddy, M. K. Rovani, A. Wohlwill, A. Katzen, et R. V. Storti, « The Drosophila Par domain protein I gene, Pdp1, is a regulator of larval growth, mitosis and endoreplication », *Developmental Biology*, vol. 289, n°. 1, p. 100-114, janvier 2006.

[**101**] A. Rojas, S. W. Kong, P. Agarwal, B. Gilliss, W. T. Pu, et B. L. Black, « GATA4 is a direct transcriptional activator of cyclin D2 and Cdk4 and is required for cardiomyocyte proliferation in anterior heart field-derived myocardium », *Molecular and Cellular Biology*, vol. 28, n°. 17, p. 5420-5431, septembre 2008.

[**102**] H. Rolletschek, T. H. Nguyen, R. E. Häusler, T. Rutten, C. Göbel, I. Feussner, R. Radchuk, A. Tewes, B. Claus, C. Klukas, U. Linemann, H. Weber, U. Wobus, et L. Borisjuk, « Antisense inhibition of the plastidial glucose-6-phosphate/phosphate translocator in Vicia seeds shifts cellular differentiation and promotes protein storage », *The Plant Journal*, vol. 51, n°. 3, p. 468-484, août 2007.

[**103**] P. Romero, J. Wagg, M. Green, D. Kaiser, M. Krummenacker, et P. Karp, « Computational prediction of human metabolic pathways from the complete human genome. », *Genome biology*, vol. 6, n°. 1, p. R2, 2005.

[**104**] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, et M. Vidal, « Towards a proteome-scale map of the human protein-protein interaction network », *Nature*, vol. 437, n°. 7062, p. 1173-1178, octobre 2005.

[**105**] N. Saitou et M. Nei, « The neighbor-joining method: a new method for reconstructing phylogenetic trees », *Molecular Biology and Evolution*, vol. 4, n°. 4, p. 406-425, juillet 1987.

[**106**] J. Schellenberger, J. Park, T. Conrad, et B. Palsson, « BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions », *BMC Bioinformatics*, vol. 11, n°. 1, p. 213, 2010.

[**107**] M. Scherf, A. Klingenhoff, et T. Werner, « Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach », *Journal of Molecular Biology*, vol. 297, n°. 3, p. 599-606, mars 2000.

[**108**] C. H. Schilling et B. O. Palsson, « Assessment of the metabolic capabilities of Haemophilus influenzae Rd through a genome-scale pathway analysis », *Journal of Theoretical Biology*, vol. 203, n°. 3, p. 249-283, avril 2000.

[**109**] C. H. Schilling, M. W. Covert, I. Famili, G. M. Church, J. S. Edwards, et B. O. Palsson, « Genome-scale metabolic model of Helicobacter pylori 26695 », *Journal of Bacteriology*, vol. 184, n°. 16, p. 4582-4593, août 2002.

[**110**] C. H. Schilling, D. Letscher, et B. O. Palsson, « Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective », *Journal of Theoretical Biology*, vol. 203, n°. 3, p. 229-248, avril 2000.

[**111**] H.-J. Schulz, M. John, A. Unger, et H. Shumann, « Visual Analysis of Bipartite Biological Networks », *Eurographics Workshop on Visual Computing for Biomedicine*, 2008.

[**112**] S. Schuster et C. Hilgetag, « On elementary flux modes in biochemical reaction systems at steady state », *Journal of Biological Systems*, vol. 2, n°. 2, p. 165-182, 1994.

[**113**] S. Schuster, T. Dandekar, et D. A. Fell, « Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering », *Trends in Biotechnology*, vol. 17, n°. 2, p. 53-60, février 1999.

[**114**] S. Schuster, D. A. Fell, et T. Dandekar, « A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks », *Nature Biotechnology*, vol. 18, n°. 3, p. 326-332, mars 2000.

[**115**] S. Schuster, C. Hilgetag, J. H. Woods, et D. A. Fell, « Reaction routes in biochemical reaction systems: algebraic properties, validated calculation procedure and example from nucleotide metabolism », *Journal of Mathematical Biology*, vol. 45, n°. 2, p. 153-181, août 2002.

[**116**] R. Schwarz, P. Musch, A. von Kamp, B. Engels, H. Schirmer, S. Schuster, et T. Dandekar, « YANA - a software tool for analyzing flux modes, gene-expression and enzyme activities », *BMC Bioinformatics*, vol. 6, p. 135, 2005.

[**117**] R. L. Seal, S. M. Gordon, M. J. Lush, M. W. Wright, et E. A. Bruford, « genenames.org: the HGNC resources in 2011 », *Nucleic Acids Research*, vol. 39, n°. Database issue, p. D514-519, janvier 2011.

[**118**] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, et T. Ideker, « Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks », *Genome Research*, vol. 13, n°. 11, p. 2498 -2504, novembre 2003.

[**119**] L. P. Smith et M. Hucka, « Hierrachical Model Composition », *SBML issue tracking system*, n°. 2404771, avril 2011.

[**120**] J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, et E. D. Gilles, « Metabolic network structure determines key aspects of functionality and regulation », *Nature*, vol. 420, n°. 6912, p. 190-193, novembre 2002.

[**121**] G. Stepien, A. Torroni, A. B. Chung, J. A. Hodge, et D. C. Wallace, « Differential expression of adenine nucleotide translocator isoforms in mammalian tissues and during muscle cell differentiation », *The Journal of Biological Chemistry*, vol. 267, n°. 21, p. 14592-14597, juillet 1992.

[**122**] S. K. Swamynathan, « Krüppel-like factors: three fingers in control », *Human Genomics*, vol. 4, n°. 4, p. 263-270, avril 2010.

[**123**] A. Tanay, I. Gat-Viks, et R. Shamir, « A Global View of the Selection Forces in the Evolution of Yeast Cis-Regulation », *Genome Research*, vol. 14, n°. 5, p. 829-834, mai 2004.

[**124**] M. Terzer et J. Stelling, « Large-scale computation of elementary flux modes with bit pattern trees », *Bioinformatics*, vol. 24, n°. 19, p. 2229 -2235, octobre 2008.

[**125**] The UniProt Consortium, « Ongoing and future developments at the Universal Protein Resource », *Nucleic Acids Research*, vol. 39, p. D214-D219, novembre 2010.

[**126**] P. Törönen, M. Kolehmainen, G. Wong, et E. Castrén, « Analysis of gene expression data using self-organizing maps », *FEBS Letters*, vol. 451, n°. 2, p. 142-146, mai 1999.

[**127**] M. K. Trenerry, P. A. Della Gatta, et D. Cameron-Smith, « JAK/STAT signaling and human in vitro myogenesis », *BMC Physiology*, vol. 11, p. 6, 2011.

[**128**] A. Villeger et A. Villeger, « LibSBGN: current status and future plans », *Nature Precedings*, novembre 2010.

[**129**] X. Xie, J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander, et M. Kellis, « Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals », *Nature*, vol. 434, n°. 7031, p. 338-345, mars 2005.

[**130**] R. Zheng et G. A. Blobel, « GATA Transcription Factors and Cancer », *Genes & Cancer*, vol. 1, n°. 12, p. 1178-1188, décembre 2010.

[**131**] A. Zinovyev, E. Viara, L. Calzone, et E. Barillot, « BiNoM: a Cytoscape plugin for manipulating and analyzing biological networks », *Bioinformatics*, vol. 24, n°. 6, p. 876 -877, mars 2008.

[**132**] D. A. Beard et H. Qian, *Chemical biophysics: quantitative analysis of cellular systems*. Cambridge University Press, 2008.

[**133**] C. Berge, *Hypergraphes : combinatoire des ensembles finis*. 1987.

[**134**] E. Gamma, R. Helm, R. Johnson, et J. Vlissides, *Design Patterns: Elements of Reusable Object-Oriented Software*, 1er éd. Addison Wesley, 1994.

[**135**] M. Nei et S. Kumar, *Molecular Evolution and Phylogenetics*. OUP USA, 2000.

[**136**] R. Olsen, *Design patterns in Ruby*. Upper Saddle River NJ : Addison-Wesley, 2008.

[**137**] Sauro, *Enzyme Kinetics for Systems Biology*. Future Skill Software, 2011.

[**138**] D. Jevremovic, D. Boley, et C. P. Sosa, « Divide-and-Conquer Approach to the Parallel Computation of Elementary Flux Modes in Metabolic Networks », *2011 IEEE International Symposium on Parallel and Distributed Processing Workshops and Phd Forum (IPDPSW)*, 2011, p. 502-511.

[**139**] P. D. Karp et S. Paley, « Automated drawing of metabolic pathways », *Proceedings of the Third International Conferences on Bioinformatics and Genome Research*, 1994.

[**140**] P. D. Karp et S. M. Paley, « Representations of metabolic knowledge: pathways », *Proceedings of International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 1994, vol. 2, p. 203-211.

[**141**] M. Behzadi, « A Mathematical Model of Phospholipid Biosynthesis », Ecole Polytechnique, 2011.

[**142**] S. Pérès, « Analyse de la structure des réseaux métaboliques : application au métabolisme énergétique mitochondrial », Thèse de Doctorat, Université de Bordeaux II, 2005.

[**143**] Abcam, « Abcam - antibodies and reagents supplier, find any antibody ». En ligne. Adresse : http://www.abcam.com/ps/pdf/nuclearsignal/cell_cycle.pdf. [Accès le: 18/10/2011].

[**144**] Biobase, « Biological Databases for Gene Expression, Pathway & NGS Analysis ». En ligne. Adresse : http://www.biobase-international.com/. [Accès le: 12/10/2011].

[**145**] Biobase, « Gene Regulation ». En ligne. Adresse : http://www.gene-regulation.com/pub/databases/transfac/doc/relationsSM.html. [Accès le: 12/10/2011].

[**146**] Biomodels, « BioModels.net: MIRIAM ». En ligne. Adresse : http://biomodels.net/miriam/. [Accès le: 11/8/2011].

[**147**] CellML, « CellML primer — CellML ». En ligne. Adresse : http://www.cellml.org/getting-started/cellml-primer. [Accès le: 10/8/2011].

[**148**] K. Christian, « VANTED - Template 1 ». En ligne. Adresse : http://vanted.ipk-gatersleben.de/index.php?file=doc8.html. [Accès le: 03/10/2011].

[**149**] Consortium for functional glycomics, « Glycosylation Pathways ». En ligne. Adresse : http://web.mit.edu/glycomics/gt/gtdb.shtml. [Accès le: 08/8/2011].

[**150**] EBI, « MIRIAM Web Service Documention ». En ligne. Adresse : http://www.ebi.ac.uk/miriam/main/mdb?section=ws_help. [Accès le: 11/8/2011].

[**151**] EMBL-EBI, « BioModels Database », *BioModels Database*. En ligne. Adresse : http://www.ebi.ac.uk/biomodels-main/. [Accès le: 08/8/2011].

[**152**] GeneNetWorks, « GeneNet database and GeneNet viewer ». En ligne. Adresse : http://wwwmgs.bionet.nsc.ru/mgs/gnw/genenet/. [Accès le: 08/8/2011].

[**153**] Genomatix, « Genomatix - Personalized Medicine - Relevance for scientists ». En ligne. Adresse : http://www.genomatix.de/en/index.html. [Accès le: 12/10/2011].

[**154**] INSDC, « INSDC ». En ligne. Adresse : http://www.insdc.org/index.html. [Accès le: 12/10/2011].

[**155**] invitrogen, « LINNEA Pathways ». En ligne. Adresse : http://escience.invitrogen.com/ipath/index.jsp. [Accès le: 08/8/2011].

[**156**] IPK Gatersleben, « MetaCrop Home ». En ligne. Adresse : http://pgrc-35.ipk-gatersleben.de/pls/htmldb_pgrc/f?p=112:1:4413919539968980. [Accès le: 08/8/2011].

[**157**] JWS Online, « JWS Online Cellular Systems Modelling: Home ». En ligne. Adresse : http://jjj.biochem.sun.ac.za/index.html. [Accès le: 08/8/2011].

[**158**] Kanehisa Laboratories, « KEGG: Kyoto Encyclopedia of Genes and Genomes ». En ligne. Adresse : http://www.genome.jp/kegg/. [Accès le: 08/8/2011].

[**159**] KEGG, « KGML Document ». En ligne. Adresse : http://www.genome.jp/kegg/xml/docs/. [Accès le: 16/8/2011].

[**160**] LibSBGN, « LibSBGN - Exchange Format ». En ligne. Adresse : http://sourceforge.net/apps/mediawiki/libsbgn/index.php?title=Exchange_Format. [Accès le: 08/8/2011].

[**161**] Mc Gill University, « Home - PReMod ». En ligne. Adresse : http://genomequebec.mcgill.ca/PReMod/welcome;jsessionid=DE564AC2523C079BEF25FC8BD4B6E16B. [Accès le: 12/10/2011].

[**162**] Michael Hucka, Andrew M. Finney, Stefan Hoops, Sarah M. Keating, et Nicolas Le Novere, « Systems Biology Markup Language (SBML) Level 2: Structures and Facilities for Model Definitions », 26/10/2007. En ligne. Adresse : http://precedings.nature.com/documents/58/version/2. [Accès le: 08/8/2011].

[**163**] National Library of Medicine - Medical Subject Headings, « Models, Biological », *National Library of Medicine*. En ligne. Adresse : http://www.nlm.nih.gov/cgi/mesh/2008/MB_cgi?term=%20models%2C%20biological. [Accès le: 11/8/2011].

[**164**] NIH EMBL-EBI, « Reactome ». En ligne. Adresse : http://www.reactome.org/ReactomeGWT/entrypoint.html. [Accès le: 08/8/2011].

[**165**] Pathguide, « Pathguide: Statistical Summary of Pathguide Resources ». En ligne. Adresse : http://www.pathguide.org/statistics.php. [Accès le: 11/8/2011].

[**166**] Paul Thomas, « PANTHER - Pathways ». En ligne. Adresse : http://www.pantherdb.org/pathway/. [Accès le: 08/8/2011].

[**167**] M. G. Poolman, « IEEE Xplore - ScrumPy: metabolic modelling with Python », 07:32:12. En ligne. Adresse : http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1678088. [Accès le: 04/5/2011].

[**168**] sbgn.org, « SBGN Softwares - sbgn.org ». En ligne. Adresse : http://www.sbgn.org/SBGN_Software. [Accès le: 08/8/2011].

[**169**] SBML.org, « SBML Software Guide/SBML Software Matrix - SBML.org ». En ligne. Adresse : http://sbml.org/SBML_Software_Guide/SBML_Software_Matrix. [Accès le: 27/9/2011].

[**170**] SRI International, « MetaCyc Encyclopedia of Metabolic Pathways ». En ligne. Adresse : http://www.metacyc.org/. [Accès le: 08/8/2011].

[**171**] Systems Biology Research Group, University of California, San Diego., « BIGG Database », *BIGG Database*. En ligne. Adresse : http://bigg.ucsd.edu/bigg/home.pl. [Accès le: 08/8/2011].

[**172**] WikiPathways, « WikiPathways ». En ligne. Adresse : http://wikipathways.org/index.php/WikiPathways. [Accès le: 08/8/2011].

[**173**] « MAP kinases », *Riken BioResource Center DNA Bank. Gene set bank.*, 26/7/2011. En ligne. Adresse : http://www.brc.riken.go.jp/lab/dna/en/GENESETBANK/index.html.

[**174**] « Ensembl Genome Browser Genomes ». En ligne. Adresse : http://www.ensemblgenomes.org/. [Accès le: 17/10/2011].

[**175**] « Ensembl Genomes Browser Bacteria ». En ligne. Adresse : http://bacteria.ensembl.org/index.html. [Accès le: 17/10/2011].

[**176**] « Ensembl Genomes Browser Plants ». En ligne. Adresse : http://plants.ensembl.org/index.html. [Accès le: 17/10/2011].

[**177**] « UniGene Home ». En ligne. Adresse : http://www.ncbi.nlm.nih.gov/unigene/. [Accès le: 08/10/2011].

[**178**] « European Bioinformatics Institute | Homepage | EBI ». En ligne. Adresse : http://www.ebi.ac.uk/. [Accès le: 08/10/2011].

[**179**] « Slash Dot Dash » Blog Archive » Rails searching with Sphinx ». En ligne. Adresse : http://www.slashdotdash.net/2007/08/06/rails-searching-with-sphinx/. [Accès le: 10/10/2011].

[**180**] « PHP Search Engine Showdown - O'Reilly Media ». En ligne. Adresse : http://onlamp.com/pub/a/php/2006/02/16/search-engine-showdown.html?page=2. [Accès le: 10/10/2011].

[**181**] « Beast acts_as_sphinx - O'Reilly Ruby ». En ligne. Adresse : http://www.oreillynet.com/ruby/blog/2008/03/beast_acts_as_sphinx.html. [Accès le: 10/10/2011].

7 GLOSSAIRE

Active record	Active record est un patron de conception (cf. Design pattern) permettant de simplifier la lecture et l'écriture de données dans une base de données. Active record impose que les objets interrogeant la base de données aient une structure équivalente à celle des tables de la base.
Ada	Langage de programmation orienté objet. Ce langage a été nommé ainsi en l'honneur d'Ada Lovelace (1815-1852), à l'origine de la notion d'algorithme
Adenine Nucléotide Translocator	Transporteur des nucléotides adényliques permettant l'échange ATP/ADP à travers la membrane mitochondriale. Il existe quatre isoformes de la protéine ANT chez l'homme ayant des profils d'expression et des cinétiques différentes. ANT1 est exprimée dans le cœur et les muscles squelettiques, ANT2 est exprimée dans les cellules en prolifération, ANT3 est une isoforme ubiquiste et ANT4 est exprimée dans le spermatozoïde. Les ANT1 et 3 permettent l'export de l'ATP vers le cytosol et les ANT2 et 4 permettent son import.
Analyse de corrélation	L'analyse de corrélation est basée sur le calcul de coefficients de corrélations entre deux séries de données. Le coefficient de corrélation est une mesure évaluant si deux séries varient ensemble. Sa valeur est indépendante des unités dans lesquelles les deux variables de mesure sont exprimées et doit être comprise entre -1 et +1 inclus.
ANT	cf. Adenine Nucléotide Translocator
API	Application Programming Interface ou interface de programmation. C'est l'interface fournie par un programme informatique (ensemble de d'objets et de méthodes en programmation orientée objet). Elle permet l'interaction de différents programmes entre eux.
Arc	En théorie des graphes, un arc est un lien orienté entre deux sommets
Arrête	En théorie des graphes, une arrête est un lien non orienté entre deux sommets
Attribut informatique	cf. Programmation orientée objet
Bibliothèque de programmation	En programmation, une bibliothèque est un ensemble de fonctions permettant d'étendre les fonctionnalités d'un logiciel
BiNoM	Biological Network Manager. C'est un plugin pour le logiciel Cytoscape permettant la manipulation de réseaux biologiques en utilisant les standards SBML, SBGN et BioPAX.
BioNMR	BioNMR est une base de données de spectres de résonnance magnétique nucléaire produits lors d'expérience de métabolomique
BioPAX	Standard de description de voies métaboliques
BLAT	BLAST-Like Alignment Tool (BLAST : Basic Local Alignment Search Tool). Programme permettant d'identifier des similarités entre des séquences nucléotidiques ou protéiques. Programme équivalent à BLAST mais beaucoup plus rapide.
Bootstrap	En phylogénie, le boostrap est la méthode la plus souvent utilisée pour tester la fiabilité des branches internes d'un arbre phylogénétique. Évalue la résistance d'une topologie à une perturbation.
C++	Langage de programmation orienté objet
CellML	Standard de description de voies métaboliques basé sur le format XML
CHEBI	Chemical Entities of Biological Interest. Base de données de structures moléculaires. Comprend essentiellement des composés biochimiques de petite taille.

Classe informatique	En informatique, une classe déclare les propriétés d'un ensemble d'objets. Elle déclare l'ensemble des attributs et des méthodes d'une catégorie d'objets. Un objet correspond à une instance d'une classe (création d'un objet possédant les propriétés d'une classe)
Corrélation de Spearman	Le coefficient de corrélation de Spearman permet de comparer deux séries non linéaires. Ce coefficient sera proche de 1 si les deux séries sont monotones dans le même sens
Corrélation linéaire de Bravais-Pearson	Le coefficient de corrélation linéaire de Bravais-Pearson permet d'évaluer la corrélation de deux séries linéaires
CRM	Module cis-régulateur. Combinaison de motifs de régulation pouvant être reconnue par un ou plusieurs facteurs de transcription
CSV	Comma-Separated Values. Le format CSV est un format de fichier représentant des tableaux de valeurs dont les colonnes sont délimitées par des virgules et les lignes par des retours à la ligne
Design Pattern	Patron de conception. Architecture logicielle permettant de répondre de la manière la plus efficace possible à un problème donné
Développement agile	Méthode de développement informatique permettant la création rapide de programmes impliquant au maximum le demandeur dans le processus de développement
EBI	European Bioinformatics Institute
EFM	cf. Mode élémentaire
EfmTools	Plugin du logiciel CellNetAnalyzer permettant de calculer les modes élémentaires d'un réseau
ElmoComp	Programme parallélisé basé sur l'algorithme du logiciel Metatool permettant le calcul des modes élémentaires.
EMBL	European Molecular Biology Laboratory
EnsEMBL	Base de données de génomique développée par l'EMBL et l'EBI. La base de données est couplée à un système d'annotation automatisée de génomes. Elle propose une API permettant de construire des pipelines bioinformatiques.
Famille de matrices	Dans la base de données Genomatix, une famille de matrices est un ensemble de matrices de régulation ayant la même séquence core. (cf Séquence core)
FastM	Outil de la base de données Genomatix permettant de construire des modèles de régulation par combinaison de matrices et/ou de séquences IUPAC
framework	Cadre d'application. Un framework est, en programmation, un ensemble de composants logiciels structurels qui servent à créer les fondations ainsi que les grandes lignes de tout ou d'une partie d'un logiciel
FrameWorker	Outil Genomatix permettant de rechercher des combinaisons de matrices partagées par plusieurs séquences promotrices
GC	Gas Chromatography (chromatographie en phase gazeuse). Chromatographie permettant de séparer les composés gazeux ou pouvant être vaporisés
GEMS Launcher	GEMS Launcher est le programme de Genomatix gérant les accès aux différents outils proposés
Gènes homologues	Les gènes homologues sont des gènes présents chez différentes espèces provenant d'un même ancêtre commun
Gènes paralogues	Des gènes paralogues sont des gènes homologues ayant subi chez une ou plusieurs espèces un évènement de duplication.
Glyphe	Pour SBGN, un glyphe est la forme des nœuds et des arrêtes d'un graphe. Chaque glyphe fournit des informations précises sur le rôle de l'élément qu'il représente
GML	Graph Modelling Language. Format de fichier de description de graphes
Graphe	En théorie des graphes, un graphe est un ensemble de points (nœuds) reliés par des liens (arrêtes)

Graphe biparti	Un graphe biparti est un graphe constitué de deux ensembles de nœuds dont les arrêtes relient deux nœuds de chacun des deux ensembles
GRBOX	Glycolysis Regulated Box. Séquence de régulation identifié sur le promoteur du gène *ANT2* et supposée permettre son induction dans des conditions de métabolisme glycolytique
Héritage	L'héritage est un des concepts de base de la programmation objet. Il permet d'établir une hiérarchie entre les classes. Une classe fille possède les mêmes caractéristiques que sa classe mère. Ceci permet de factoriser le code commun à plusieurs classes
HPLC	High Pressure Liquid Chromatography ou chromatographie en phase liquide à haute pression. Technique de séparation de composés en fonction de leur hydrophobicité.
HRMAS	High Resolution Magic Angle Spinning. Technique d'acquisition RMN ou l'échantillon analysé a une rotation selon un axe incliné de 54,74° (angle magique) par rapport au champ magnétique. Ceci augmente la résolution du spectre.
HTTP	Hypertext Transfer Protocol. C'est le protocole de communication client-serveur développé pour le web.
Hypergraphe	Graphe dont les arrêtes peuvent relier un nombre quelconque de sommets
IUPAC	International Union of Pure and Applied Chemistry. Nomenclature internationale des composés chimiques. Pour les bases nucléotidiques : A (Adénine), T (Thymine), G (Guanine), C (Cytosine), U (Uracile), R (A ou G), Y (C ou T), S (G ou C), W (A ou T), K (G ou T), M (A ou C), B (C, G ou T), D (A, G ou T), H (A, C ou T), V (A, C ou G), N (A, T, G ou C)
KEGG	Kyoto Encyclopedia of Genes and Genomes. Base de données de voies métaboliques et de génomique
Kestose	Trisaccharide de la famille des fructo-oligosaccharides utilisé comme édulcorant
KGML	KeGg Markup Language. Format XML de description des voies métaboliques KEGG.
KGML-ed	KGML editor. Plugin du logiciel VANTED permettant de lire et de représenter les fichiers KGML
Langage à typage dynamique	Dans un langage à typage dynamique, le type d'une variable n'est jamais explicitement défini. Il est entièrement géré par le langage de programmation et peut changer au cours de l'exécution du programme. On parle aussi de « typage à la canard » (duck typing) en référence à une phrase couramment expliquée pour expliquer son fonctionnement : « Si je vois un animal qui vole comme un canard, cancane comme un canard, et nage comme un canard, alors j'appelle cet oiseau un canard »
Langage à typage statique	Dans un langage à typage statique, le type de chaque variable doit être déclaré en même temps que celle-ci et ne peut pas être modifié
Layout	Un algorithme de layout est un algorithme permettant de placer automatiquement les nœuds et les arrêtes d'un graphe pour le représenter
LibSBGN	Bibliothèque de programmation permettant de sauvegarder des graphes biologiques au format SBGN-ML
Mapping Objet-Relationnel	Technique de programmation créant l'illusion d'une base de données dans un programme orienté objet à partir d'une base de données relationnelle. On parle de correspondance entre le monde objet et le monde relationnel
MatBase	Base de données de matrices de régulation de Genomatix
MathML	Représentation au format XML de formules mathématiques
MatInspector	Outil Genomatix permettant la recherche de matrices de régulation sur des séquences nucléotidiques
Matrice d'adjacence	Représentation d'un graphe sous forme de matrice carrée indiquant pour tous les couples de sommets du graphe s'il existe une arrête entre eux

Matrice de régulation	Dans la base de données Genomatix, une matrice de régulation est la représentation sous forme de matrice d'une séquence de régulation. Les colonnes correspondent aux sites de la séquence et les lignes à la fréquence des différentes bases (A, T, G et C) pour les différentes séquences reconnues par un même facteur de transcription
Matrice de stœchiométrie	Type de matrice utilisé pour la représentation de graphes de réaction. Les métabolites sont disposés en ligne et les réactions en colonne. Dans chaque case, sont indiqués les coefficients de stœchiométrie des métabolites dans les réactions. Ces coefficients sont négatifs dans le cas d'une consommation et positifs pour une synthèse
Métabolite externe	Les métabolites externes d'un système sont les métabolites qui ne sont pas équilibrés dans le système. Ils sont définis pour le calcul des modes élémentaires.
Métabolite interne	Les métabolites internes d'un système sont tous les métabolites du système qui ne sont pas externes
Métabolomique	La métabolomique est l'étude des métabolites d'un système à un moment donné
MetaCrop	Metabolic pathways of Crop Plants database. Base de données de voies métaboliques de céréales
Méthode informatique	cf. Programmation orientée objet
MIRIAM	Minimum Information Requested In the Annotation of biochemical Models. Convention regroupant toutes les informations nécessaires à la description d'un modèle
Mode élémentaire	Ensemble minimal de réactions pouvant avoir lieu à l'état stationnaire
ModelInspector	Outil Genomatix permettant de rechercher un modèle de régulation (ensemble de matrices de régulation) soit sur une banque de clones soit sur la base de données de séquences promotrices d'ElDorado
Modélisation	La modélisation est la création et l'étude d'un modèle. Un modèle est un concept ou un objet considéré comme représentatif d'un autre. La modélisation consiste à la fois à simplifier la réalité, en éliminant les détails difficiles à reproduire, et à obtenir un résultat plus net, en se concentrant sur les seuls traits jugés importants
Module cis-régulateur	Structure des séquences promotrices composée d'une combinaison de plusieurs motifs de régulation de dix à cinquante nucléotides
MPSA	Metabolic Pathways Software Analyzer
MVC	Model View Controller (Modèle Vue Contrôleur). Design Pattern imposant d'organiser les classes en trois groupes : modèles, vues et contrôleurs. Le Modèle correspond aux données, la vue à une représentation de ces données et le Contrôleur assure le dialogue entre la Vue et le Modèle.
myKegg	Base de données reprenant d'une part l'ensemble des voies métaboliques humaines et d'autre part une base de données de synonymes. Cette base de données a été bâtie en utilisant essentiellement les données de KEGG
MySQL	Système de gestion de base de données le plus utilisé dans le monde à l'heure actuelle
NCBI	National Center for Biotechnology Information
Neighbour-Joining	Méthode de reconstruction d'arbres phylogénétiques basée sur des mesures de distances résultantes d'un alignement de séquences nucléotidiques
Nœud	En théorie des graphes, un nœud est un point du graphe. Les nœuds sont reliés par des arcs ou des arrêtes.
Objet informatique	cf. Programmation orientée objet et classe informatique
ORF	Open Reading Frame (Cadre ouvert de lecture). Séquence d'ADN débutant par un codon d'initiation et se terminant par un codon stop.
ORM	Object-Relational Mapping, cf. Mapping objet-relationnel
OXBOX	Oxidative Box, séquence nucléotidique identifiée sur le promoteur du gène ANT1 humain par retard de migration (gel shift) et supposée être impliquée dans l'expression muscle spécifique de cette isoforme

Patron de conception	cf. Design pattern
PCR	Polymerase Chain Reaction (Réaction en chaîne par polymérase)
Perl	Langage de programmation interprété, utilisé essentiellement sous forme de scripts. Particulièrement adapté à la manipulation de fichiers texte
Places	Dans un réseau de Petri, les places correspondent à un des deux types de nœuds (avec les transitions). En biologie des systèmes, ils correspondent aux entités biologiques (protéines, métabolites, gènes, etc.)
Plugin	Un plugin est un programme qui permet d'étendre les fonctionnalités d'un autre programme
Programmation orientée objet	Souvent notée POO. Paradigme de programmation informatique consistant à découper un programme en briques logicielles appelées objets. Un objet peut représenter un concept, une idée ou tout objet physique. Il possède un ensemble de fonctions, appelées méthodes, et de données appelées attributs. Les objets sont liés entre eux par différents liens comme par exemple l'héritage (Cf. Héritage) ou l'inclusion.
Programme parallélisé	Programme informatique dont le calcul est distribué sur plusieurs processeurs pour gagner en efficacité
Promotologie	Étude des séquences promotrices des gènes
PSI-MI	Proteomics Standards Initiative Molecular Interaction. Format de type XML permettant de décrire des interactions protéine-protéine
Puits	Dans un graphe orienté, un puits est un nœud vers lequel les arrêtes pointent mais d'où aucune arrête ne repart. Dans un graphe métabolique, un puits correspond à un métabolite synthétisé mais jamais consommé
PWM	Position Weight Matrix. Cf. Matrice de régulation
pyNMR	Logiciel en Python permettant la visualisation et la manipulation de spectres enregistrés au format CSV
Python	Langage de programmation orienté objet à typage dynamique. Langage interprété multiplateforme permettant la rédaction de scripts aussi bien que de programmes complexes
Refactoring	Remaniement du code. En informatique, le refactoring désigne l'action de réécriture d'un code, souvent dans le même langage, pour l'améliorer (ajout de fonctionnalité, optimisation de mémoire ou d'exécution, etc.)
Réseau de Petri	Modèle mathématique permettant de représenter des systèmes complexes et d'en étudier le comportement. Un réseau de Petri est représenté par un graphe biparti dont les deux types de nœuds sont les places et les transitions. Dans la représentation d'un réseau métabolique, les places représentent les métabolites et les transitions les réactions. Les places peuvent contenir des jetons représentant les ressources disponibles. Les transitions peuvent porter des lois permettant de régir la distribution des jetons
Résonance Magnétique Nucléaire	La spectroscopie de Résonance Magnétique Nucléaire est une technique d'analyse exploitant les propriétés magnétiques de certains noyaux atomiques. En biologie, les RMN du proton (1H), du carbone (13C) et du phosphore (31P) sont les plus utilisées. La RMN est applicable à tout noyau possédant un spin non nul. Le signal mesuré est appelé FID (Free Induction Decay) correspondant à la superposition des sinusoïdes produites par la résonance des atomes. Une transformée de Fourrier est appliquée pour convertir le FID en spectre de fréquence. Sur un spectre, chaque composé biochimique est caractérisé par un ensemble de pics d'ordonnées (déplacement chimique) connues.
Rhamose	Hexose de type aldose noté aussi rhamnose, isodulcitol ou 6-deoxy-L-mannose
RMN	cf. Résonance Magnétique Nucléaire
RoR	cf. Ruby on Rails
Ruby	Langage de programmation interprété fortement orienté objet

Ruby on Rails	Framework Ruby suivant le design pattern MVC surtout utilisé pour la création d'applications web
SBGN	Systems Biology Graphical Notation. Système de représentation graphique de graphes biologiques. Il est basé sur des nœuds dont la forme porte une information (glyphes)
SBGN-ed	Plugin de VANTED permettant la représentation et la sauvegarde de graphes biologiques au format SBGN.
SBGN-ML	Format XML de sauvegarde des graphes au format SBGN. Décrit dans la Lib-SBGN
SBML	Systems Biology Markup Language. Format XML de sauvegarde de modèles de biologie des systèmes : voies métaboliques et lois biochimiques des réactions
Script	En informatique, un script est un programme court lu par un interpréteur (et donc écrit dans un langage interprété) permettant d'automatiser des tâches informatiques plus ou moins complexes et répétitives.
Séquence core	Une matrice de régulation est composée de deux types de séquences nucléotidiques : séquences core hautement conservées entre les différents promoteurs et en phylogénie et les séquences flanquantes, moins conservées
Séquence flanquante	Cf. séquence core
Série monotone	En mathématiques, une série (ou une fonction) monotone est une série dont le sens de variation de change pas sur un intervalle donné
SGBD	Cf. Système de gestion de base de données
Smalltalk	Langage de programmation multiplateforme orienté objet à typage dynamique
Source	En théorie des graphes, un nœud source est le contraire d'un nœud puits. Dans un graphe orienté, les arrêtes ne peuvent que partir du nœud source et jamais y arriver. Dans un graphe de réaction, une source correspond à un métabolite consommé et jamais synthétisé
Spectre	Cf. Résonance Magnétique Nucléaire
Spectrométrie de Masse	Technique d'analyse permettant d'identifier des molécules par mesure de leur rapport masse / charge. En biologie, les grosses molécules comme les protéines sont fragmentées et le rapport masse / charge des fragments est mesuré pour identifier la molécule complète.
Système de gestion de base de données	Souvent noté SGBD (ou DBMS pour DataBase Management System). Logiciel permettant l'interrogation et gestion des données dans une base de données. Deux SGBD ont été utilisés ici : mySQL et SQLite
Tests automatiques	En informatique, un test (ou test automatique) est un programme permettant de vérifier partiellement les fonctionnalités d'un logiciel. L'ensemble du code du logiciel testé permet de définir la couverture des tests (code coverage)
TFBS	Transcription Factor Binding Site (site de fixation de facteurs de transcription)
Transcriptomique	La transcriptomique est l'étude de l'ensemble des ARN transcrits à un instant donné dans une cellule. Elle est basée sur l'utilisation de puces à ADN, de la PCR quantitative ou du séquençage à haut débit
Transitions	cf. réseau de Petri
TSS	Transcription start site (site de début de transcription)
Type informatique	En informatique, le type d'une variable correspond au type de valeurs qu'elle peut prendre. Il existe des types prédéfinis comme : le booléen, le nombre entier, la chaine de caractères, le réel … Mais il est possible de déterminer de nouveaux types
UniGene	Base de données du NCBI de transcriptomique et de génomique
Uniprot	Base de données de protéomique (séquences et fonctions des protéines). Issue de la fusion des bases de données de protéomique de trois instituts : EBI (European Bioinformatics Institute), SIB (Swiss Institute of Bioinformatics) et PIR (Protein Information Resource)

URI	Uniform Resource Identifier : chaîne de caractères identifiant une ressource sur un réseau
VANTED	Visualization and Analysis of Networks conTaining Experimental Data. Logiciel de représentation et d'analyse de voies métaboliques
Voie extrême	Sous ensemble des modes élémentaires d'un système
Voie Métabolique	Ensemble de réactions liées par leurs substrats et leurs produits. Elle comprend un ensemble de substrats de base et conduit à la formation d'un ensemble de produits avec potentiellement plusieurs étapes intermédiaires entre eux. On considère aussi souvent une voie métabolique comme un ensemble de réactions impliquées dans une même fonction cellulaire
WebGL	Spécification d'affichage 3d pour les navigateurs web permettant d'utiliser le standard openGL à partir de code JavaScript inclus dans une page web
XML	Extensible Markup Language (langage de balisage extensible). Langage informatique permettant la description d'informations encadrées ou portées par des balises. Une balise est caractérisée par un nom unique encadré par des chevrons. Chaque balise doit être ouverte puis fermée. Les balises délimitent ainsi des unités sémantiques dans un fichier

8 INDEX

H

I

J

K

L

M

O

P

R

S

T

U

V

W

X

9 ANNEXES

A Computational analysis of the transcriptional regulation of the adenine nucleotide translocator isoform 4 gene and its role in spermatozoid glycolytic metabolism

B Computational identification of transcriptionally co-regulated genes, validation with the four ANT isoform genes

Annexe A : Computational analysis of the transcriptional regulation of the adenine nucleotide translocator isoform 4 gene and its role in spermatozoid glycolytic metabolism

Gene 487 (2011) 38–45

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Computational analysis of the transcriptional regulation of the adenine nucleotide translocator isoform 4 gene and its role in spermatozoid glycolytic metabolism

Pierre-Yves Dupont, Georges Stepien *

INRA, UMR 1019, UNH, F-63122 St Genès-Champanelle; Clermont Université, Université d'Auvergne, Unité de Nutrition Humaine, BP 10448, F-63000 Clermont-Ferrand, France

ARTICLE INFO

Article history:
Accepted 14 July 2011
Available online 31 July 2011

Received by A.J. van Wijnen

Keywords:
Transcriptional regulation
Promoter analysis
Adenine nucleotide translocator
Spermatozoid bioenergetics

ABSTRACT

Computational phylogenetic analysis coupled to promoter sequence alignment was used to understand mechanisms of transcriptional regulation and to identify potentially coregulated genes. Our strategy was validated on the human *ANT4* gene which encodes the fourth isoform of the mitochondrial adenine nucleotide translocator specifically expressed during spermatogenesis. The movement of sperm flagella is driven mainly by ATP generated by glycolytic pathways, and the specific induction of the mitochondrial ANT4 protein presented an interesting puzzle. We analysed the sequences of the promoters, introns and exons of 30 mammalian *ANT4* genes and constructed regulatory models. The whole human genome and promoter database were screened for genes that were potentially regulated by the generated models. 80% of the identified co-regulated genes encoded proteins with specific roles in spermatogenesis and with functions linked to male reproduction. Our *in silico* study enabled us to precise the specific role of the ANT4 isoform in spermatozoid bioenergetics.

1. Introduction

The ANT (adenine nucleotide translocator) protein, which is also known by the generic name of ADP/ATP carrier (AAC), is encoded by the nuclear genome. This protein is located within the mitochondrial internal membrane and permits the electrogenic exchange of ATP and ADP nucleotides between the mitochondrial matrix space and the cytoplasm. ATP could be also transported in electroneutral exchange for phosphate by the mitochondrial ATP-Mg/Pi transporter (Fiermonte et al., 2004). The fact that several isoforms of the ANT protein are expressed from different genes from yeast to human underscores the importance of this protein. Each of these isoforms displays specific kinetic parameters, which enable cells to adapt energy production to the specific metabolic parameters required for their cellular or tissue environments (Table 1) (Stepien et al., 1992).

The kinetic properties of the ANT1 isoform encoded by the *SLC25A4* gene, which is specifically expressed in human muscle tissue, promote a fast and massive export of mitochondrial ATP that is essential for muscle contraction. The ANT2 isoform (*SLC25A5* gene) is expressed preferentially in growing and proliferative cells with primarily glycolytic metabolism, such as embryonic and transformed cells. The ANT2 isoform displays kinetic properties that enable it to carry out opposite transport, exchanging mitochondrial ADP for cytosolic ATP generated by cytosolic glycolytic phosphorylation (Stepien et al., 1992). ANT3 (*SLC25A6* gene) is the ubiquitous isoform of ANT, which is constitutively expressed independently of any specific transcriptional control and is thus expressed without any tissue or cell type specificity. The last isoform, ANT4 (*SLC25A31* gene), was recently identified in humans, and it is expressed mainly in the testicle (Dolce et al., 2005). This isoform appears in mammals and is essential during spermatogenesis (Brower et al., 2007). The peptide sequence of this isoform is very similar (66–68% of identity) to that of other ANT isoforms. The main characteristic of this isoform is the presence of additional peptides, specifically the N− (13 amino acids) and C- (8 amino acids) terminal sequences, which the other three isoforms lack. These extra peptides could be related to the specific localisation of this isoform to the sperm flagellum (Kim et al., 2007). The proposed hypothesis for the role of this isoform is that it appears to compensate for the loss of function of the *ANT2* gene (encoded by the X chromosome) during male meiosis (Brower et al., 2007).

In this study, we investigated the mechanisms of transcriptional regulation of this new ANT4 isoform through analysis of nucleotide sequences upstream of the supposed sites of transcription initiation. Two distinct promoter regions are proposed: the first is upstream of the transcription initiation site of the first exon and the second is at the end of intron 2, upstream of exon 3 (Fig. 1). The first promoter allows the expression of the experimentally verified 5′ complete transcript and the second one is associated to an annotated transcript in several mammals but without confirmation for 5′ completeness (Genomatix). The nucleotide sequences of these promoter regions

Abbreviations: AAC, ADP/ATP carrier; ANC, adenine nucleotide carrier; ANT, adenine nucleotide translocator; SLC25A31, solute carrier family 25 member 31; TS, transcription initiation site.

* Corresponding author. Tel.: +33 473624458; fax: +33 473624755.
E-mail address: georges.stepien@clermont.inra.fr (G. Stepien).

0378-1119/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2011.07.024

Table 1
Nomenclature of the *ANT* genes and their expression in human, mouse and yeast.

Human isoform	HGNC symbol (locus)	Ensembl gene ID	Expression	Regulatory factors	Corresponding isoform
				Mouse	Yeast
ANT1 or ANC1	SLC25A4 (4q35.1)	ENSG00000151729	Heart and skeletal muscles	OXBOX (Li et al., 1990)	Ant1	AAC1
ANT2 or ANC3	*SLC25A5* (Xq24)	ENSG00000005022	Proliferative cells	GRBOX (Giraud et al., 1998); Oct1, SV40, AP2, Sp1 (Ku et al., 1990; Luciakova et al., 2003)	Ant2	
ANT3 or ANC2	*SLC25A6* (Xp22.33)	ENSG00000169100	Ubiquitous	Sp1, NF-kB, GAS-like (Barath et al., 1999a)	–	AAC2
ANT4	*SLC25A31* (4q28.1)	ENSG00000151475	Spermatozoids	E2F6 (Kehoe et al., 2008)	Ant4	–

ANT (Ant): adenine nucleotide translocator; ANC: adenine nucleotide carrier; AAC: ATP/ADP carrier; HGNC: HUGO Gene Nomenclature Committee; SLC25 solute carrier gene family (Palmieri, 2004).

from several mammalian species were compared to follow the phylogeny of specific sequences of transcriptional regulation, as those promoter sequences preserved throughout evolution might be of major importance to the survival of the organism (Tanay et al., 2004). We used a combination of software and databases (some available online, such as Genomatix and EnsEMBL, or those built in our laboratory, such as GeneProm). The program package was validated by its use in the analysis of the transcriptional regulation of the sex-specific isoform of ANT4. Interestingly, our results lead us to propose a hypothesis as to the consequences of over-expression of this ANT4 isoform on the specific bioenergetic properties of spermatozoids carrying either the X or Y chromosomes.

2. Materials and methods

2.1. Process of bioinformatics study

An outline of the bioinformatic pipeline implemented for analysis of *ANT* sequences is illustrated in Fig. 2. Our study was performed in five steps. We began by performing a short phylogenetic study of the *ANT4* genes from various mammalian species. This step enabled us to check database annotations and eliminate sequences which were impossible to align to the human ANT4 gene. The application of the Genomatix tools to our selected sequences gave us a list of regulatory elements (either matrices or nucleotide strings) identified in the aligned sequences. We organised these elements into regulatory models which we screened across the whole human genome. The results were then either filtered with software we developed (GeneProm) or retrieved from the Genomatix human promoter database. Thus, we obtained a list of genes that were potentially co-regulated by similar transcriptional models. GeneProm software consists of a web application coupled to a database and will be available as soon as our Web server will be capable to manage a large number of users.

2.2. Phylogenetic study of the ANT4 gene in mammals

Nucleotide sequences of the *ANT4* isoform genes (*SLC25A31*) were screened in 30 mammalian species including Human (Table 2). Sequences were imported from the EnsEMBL database. These sequences included the whole gene sequence and an additional 1500 nt sequence upstream of the Transcription Start Site (TSS). A phylogeny of the retained coding sequences was carried out to validate the functional annotations of these genes. The alignment algorithm used was CLUSTAL 2.0, and it was used with the default settings. The phylogeny was rebuilt using 2 different methods: neighbour-joining and minimum evolution (both with bootstrap evaluation). Both techniques provided concordant results. This phylogeny was carried out to verify the annotations of the sequences in the databases and the concordance between genes encoding a specific protein or different isoforms. After considering the phylogenetic results of our study, we decided to keep only a small number of mammalian sequences that did not contain undetermined bases and were aligned from the consensus human sequence by using CLUSTAL 2.0 software. Additionally, the transcription initiation sites of all the

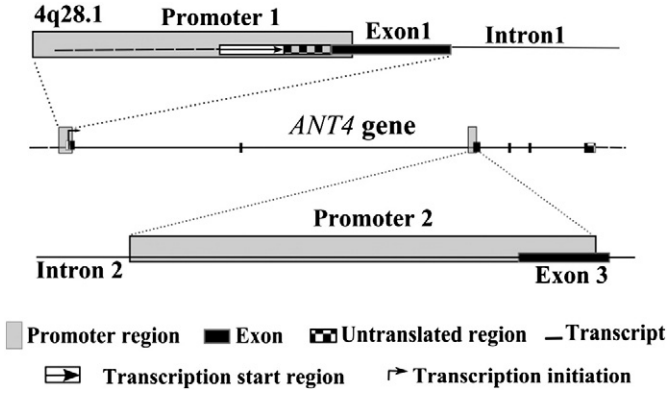


Fig. 1. Schematic representation of the structure of the *ANT4 (SLC25A31)* gene. 5′ non-coding region is represented by a dashed line, the transcript start sequence and the transcription initiation site are indicated by arrows, the untranslated regions are indicated by checkerboards, the promoter regions are represented by a grey box, the exons are shown as black rectangles, and the introns are indicated by a plain line. The top of the figure represents the known promoter regions at the beginning of the gene. The global structure of the ANT4 gene is represented in the middle. The promoter detected by the PromoterInspector algorithm is shown at the bottom.

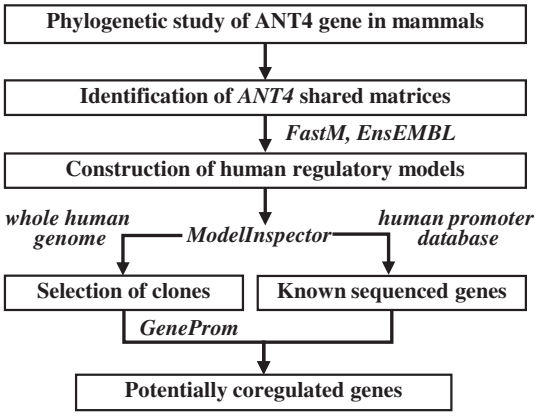


Fig. 2. The five steps of our bioinformatics study. 1) Phylogenetic study of ANT4 genes in various mammalian species to filter unusable sequences. 2) Identification of ANT4 regulation matrices shared by all aligned sequences (via Genomatix FrameWorker and MatInspector tools). 3) Construction of regulation models by combining the previously selected matrices. 4) Location of the model in the whole human genome or in the ElDorado promoter database. 5) Identification of potentially co-regulated genes (with the given model in the promoter).

Table 2
ANT4 gene sequences in mammals.

Species	Common name	Selected sequence
Bos taurus	**Ox**	**ENSBTAG00000012826**
Canis familiaris	**Dog**	**ENSCAFG00000003924**
Cavia porcellus	Guinea pig	ENSCPOG00000001977
Choloepus hoffmanni	Sloth	ENSCHOG00000005592
Dasypus novemcinctus	Armadillo	ENSDNOG00000009200
Dipodomys ordii	Kangaroo rat	ENSDORG00000000368
Echinops telfairi	Hedgehog	ENSETEG00000019595
Felis catus	Cat	ENSFCAG00000007128
Gorilla gorilla	Gorilla	ENSGGOG00000013994
Homo sapiens	**Human**	**ENSG00000151475**
Loxodonta africana	Elephant	*ENSLAFG00000000407*
Macaca mulatta	**Monkey**	**ENSMMUG00000015243**
Microcebus murinus	Lemur	*ENSMICG00000016947*
Monodelphis domestica	Opossum	*ENSMODG00000012128*
Mus musculus	**Mouse**	**ENSMUSG00000069041**
Myotis lucifugus	Microbat	ENSMLUG00000003042
Ochotona princeps	Pika	*ENSOPRG00000004154*
Oryctolagus cuniculus	Rabbit	*ENSOCUG00000015617*
Otolemur garnettii	Galago	*ENSOGAG00000005752*
Pan troglodytes	**Chimpanzee**	**ENSPTRG00000016432**
Pongo pygmaeus	Orangutan	*ENSPPYG00000015055*
Procavia capensis	Hyrax	*ENSPCAG00000003722*
Pteropus vamyrus	Megabat	ENSPVAG00000002926
Sorex araneus	Shrew	*ENSSARG00000009084*
Spermophilus tridecemlineatus	Squirrel	*ENSSTOG00000014908*
Sus scrofa	Pig	*ENSSSCG00000009078*
Tarsius syrichta	Tarsier	*ENSTSYG00000001361*
Tupaia belangeri	Tree shrew	*ENSTBEG00000000518*
Tursiops truncatus	**Dolphin**	**ENSTTRG00000012978**
Vicugna paosc	Alpaca	ENSVPAG00000007214

30 mammalian *ANT4* gene sequences extracted from EnsEMBL data base were aligned (CLUSTALW software) and 7 sequences were selected (in bold) according to concordant DNA sequences. These selected sequences included the 6 exons and 5 introns of the *ANT4* gene sequence. The latin names of species are in italics.

retained sequences were compatible with the known human initiation site.

2.3. Selection of ANT4 consistent mammalian promoter sequences

The alignment of 5′ sequences (1500 nt upstream of the ATG site) was performed using the CLUSTAL 2.0 software. Similar alignments have been obtained using MUSCLE and T-Coffee. Ten sequences of the *ANT4* gene (*SLC25A31*) were aligned. Additionally, the sequences of the three other gene isoforms were also aligned to verify their annotation in the EnsEMBL database. Thus, the list of 7 selected species takes into account the sequencing quality of this genomic region and the availability of the sequences of the three other isoforms (Table 2). Moreover, a fragment of same length (500 nucleotides) of different regions (distal DNA, proximal promoter, exons 1 and 3) of the seven gene sequences was aligned to quantify the percentage of identity between human and mouse *ANT4* sequences with corresponding sequences from the 5 other mammals.

2.4. Identification of transcriptional matrices shared by the ANT4 sequences

The FrameWorker tool of the Genomatix software package was used in this analysis (Genomatix). We identified transcriptional regulatory sequences named regulatory matrices, which were shared by a set of gene promoter sequences. Matrices were identified from the Genomatix database produced by the alignment of all regulatory elements identified to date in all vertebrates. This database includes 727 vertebrate matrices classified into 170 families (noted V$) and 16 additional general matrices from higher organisms, classified into ten families (noted O$). Additionally, the FrameWorker tool enables identification of the regulatory models (combination of several matrices) shared by different promoter sequences. The MatInspector tool simultaneously allows identification of the regulatory matrices in a promoter sequence and analysis of different parameters, permitting selection of a particular matrix according to bibliographical data related to the corresponding gene.

The ubiquitously expressed ANT isoform (ANT3) was selected as our control gene. A 1500 nt sequence upstream the ANT3 TSS was analysed by both the MatInspector and FrameWorker tools (study not presented). This ANT3 analysis allowed us to determine the level threshold for matrix similarity, which should be further used as the background for ANT4 promoter analysis.

2.5. Construction of regulatory models and their screening in primate genomes and in human promoter database

The set of selected matrices identified by the FrameWorker and MatInspector analysis was then combined with short nucleotide sequences (strings) identified from the alignment of the promoter sequences from the 7 selected species and with the common promoter CAAT and TATA sequences. The nucleotide distances between different matrices and IUPAC strings were then narrowed between +/− 50% as compared to distances in human ANT4 promoter sequence. The resulting models (example in Fig. 4 and all of them are presented in a supplementary file), which generally included from 3 to 6 matrices and nucleotide strings and are screened searched with different stringencies in two databases: 1) The full human genome (GenBank Release 175) using the ModelInspector Genomatix tool (Frech et al., 1997). A list of clones (contigs) of the EMBL database containing the studied model was obtained in this analysis with the exact positions of the required models and the orientation of each clone; 2) A database of promoter sequences from various mammals (Genomatix Eldorado 2–2010), including a proposed list of known genes. The different stringency parameters of this study (ModelInspector parameters) were as follows: maximum number of mismatches allowed for nucleotide strings, matrix similarity (Cartharius et al., 2005), threshold (number of matrices or strings present in the sequence *vs*. the number of matrices or strings to find), the research of individual or matrix families, global and core matrix similarities, matrix sense in relation to the clone, and minimal and maximal distances between two matrices or nucleotide strings.

2.6. Identification and selection of known, supposed or unknown genes containing a model in their proximal promoter regions

The list of EMBL clones proposed by ModelInspector was exported into our own GeneProm software, which allowed us to position each model based on its chromosomal location to identify genes present in the immediate proximity (or in partial overlap) of the model considered. The filters used in this analysis were: 1) distance from the beginning of the model sequence to the transcription start site of the gene (1000 nt by default); 2) the value of the superposition between the model and the beginning of the gene (100 nt by default); 3) the respective orientations of the model, the identified gene, the EMBL contig and the chromosomal DNA strand. These filters permitted selection of genes directly downstream of the researched model from the same strand of DNA with the same orientation. Thus, the list obtained contained sequences located by default at a maximum distance of 1000 nt from the 5′ end of the first matrix or string of the model considered.

2.7. Analysis and confirmation of co-regulated genes and their link with metabolic pathways

The list of the genes obtained from the two analyses (ModelInspector on the GenBank Release and on the Genomatix Human Promoters) was manually analysed: an exhaustive bibliography was obtained for each gene (EnsEMBL, GeneCards, and PubMed) to

identify the primary arguments on the function of each corresponding protein and their direct or indirect links with the reference protein. GeneProm software was then paired by a script with the KEGG database, allowing identification of the metabolic networks involved in the protein encoded from each gene identified in the previous step. A complementary bibliographic analysis, relating the identified proteins to suggested metabolic pathways, was carried to identify the key proteins involved in these pathways and those pathways that were potentially co-regulated with the initially analysed genes.

3. Results

3.1. Alignment and selection of ANT4 promoter sequences from mammalian species

The mammalian species shown in Table 2 were screened for the *ANT4* gene sequence. Manual sorting of these sequences was carried out by withdrawing the sequences containing either a high number of undefined bases or annotation errors at the transcription initiation sites (identified by the alignment of the sequences, including the first exon, in each species). Twelve gene sequences were selected according to their full length sequences (1500 nt upstream of the gene sequence and their exon 1 sequence) without any unknown nucleotide. The 12 sequences were aligned using CLUSTAL software and 7 of them were selected for: 1) the correct overlap of their upstream sequence, 2) highly conserved sequences in the first exon and 3) the clear identification of the genes of the three other ANT isoforms.

Five mammalian sequences with probable gaps or incomplete gene sequences were discarded. Interestingly, the *ANT4* gene might contain a second potential promoter sequence, as detected by the Genomatix PromoterInspector algorithm (Scherf et al., 2000), localised in the second intron, adjacent to the 5′ end of exon 3 (based on comparative genomics), and including CAAT, 3 serial TATAAA and ATG sequences. However, no corresponding transcript has been identified (Genomatix). The nucleotide sequence of this specific supposed promoter upstream the third exon sequence was extracted from the 7 selected mammalian genomes.

3.2. Comparison of the ANT4 gene sequences and validation of the two promoter regions

The 7 mammalian *ANT4* nucleotide sequences selected in Table 2 were aligned, and the percentages of identity between the different sequence regions described in Fig. 1 were computed: 500 nucleotides of the proximal part of the promoter sequences 1 and 2 sequences upstream from the ATG codon; full sequence of corresponding exons 1 and 3; intron sequence (the first 500 nucleotides of intron 1 sequence); and same length of a distal sequence out of the ANT4 promoter region and without identified or supposed coding sequence (500 nt from 10 000 nt upstream of exon 1). Table 3 shows the percentages of identity between the human and mouse sequences and the sequences of the five other species.

In our study, the mouse sequence is the most different sequence from the human gene. The percent identity between these species is 78.7% in exon 1, as compared to 90.6% in exon 3. The percentage identity between non-coding distal sequences is 27.3%. At the level of promoters 1 and 2, the percentage of identity is 35% and 43%, respectively, whereas the level of identity of intron 1 is only 27.7%. Thus, the conservation of the putative promoter 2 seems more significant compared to that of a simple intron sequence. Moreover, the promoter 2 sequence presents a percentage of identity that is higher than promoter 1. It contains also the canonical CAAT and TATAAA sequences. However, the functionality of promoter 2 is questionable, as indicated by the absence of a validated specific transcript in the 7 mammals tested in this study.

Table 3
Percentage of identity between human and mouse *ANT4* sequences with corresponding sequences from other mammals.

	Human	Chimpanzee	Monkey	Dog	Ox	Dolphin	Mouse
H. Distal DNA	ID	**80.9**	39.9	31.3	31.3	28.7	27.3
M. Distal DNA	27.3	22.1	18.0	28.0	30.4	23.1	ID
H. Promoter 1	ID	**94.8**	**84.4**	52.6	45.1	61.7	35.0
M. Promoter 1	35.0	35.2	35.0	34.1	36.4	37.7	ID
H. Exon 1	ID	**98.7**	**96.1**	**84.0**	**84.8**	**81.9**	78.7
M. Exon 1	78.7	79.5	78.7	75.4	75.8	74.5	ID
H. Intron 1	ID	**96**.6	**90.4**	38.7	54.5	60.8	27.7
M. Intron 1	27.7	26.2	27.0	30.3	27.6	31.2	ID
H. Promoter 2	ID	**98.8**	**93.2**	56.4	55.1	60.6	43.1
M. Promoter 2	43.1	43.2	43.1	47.4	41.0	45.5	ID
H. Exon 3	ID	**100**	**99.1**	**94.0**	**95.7**	**97.4**	**90.6**
M. Exon 3	**90.6**	**90.6**	**91.5**	**88.1**	**91.5**	**90.6**	ID

Distal DNA, 500 nt sequence 10,000 nt upstream of exon 1; Promoters 1 and 2, proximal part of promoters 1 and 2, 500 nt sequences upstream of the ATG; Exons 1 and 3, full exon sequences; Intron 1, first 500 nucleotides of intron 1. Percentages of identity higher than 80 % are in bold. H., human; M., mouse.

3.3. Identification of transcriptional matrices shared by ANT4 sequences

3.3.1. ANT4 promoter 1 (upstream exon 1)

Using the FrameWorker tool from Genomatix, we found several matrices shared by the selected mammalian promoter sequences: V$SMAD3.01 (bound by the Smad3 transcription factor involved in TGF-beta signalling), V$MZF1.02 (recognised by the myeloid zinc finger protein MZF1) and a specific nucleotide string CACGTGTCAGG, which is present in several copies (three in human) spaced 3 nucleotides apart, with the final string located 33 nt upstream of the transcription start nucleotide (TS) (Fig. 3). This final nucleotide sequence string was found in several matrices from the V$EBOX family, including V$USF.02 (an upstream stimulating factor), and from the V$HIFF family including V$HRE.01 (hypoxia response elements, binding sites for HIFF1alpha/ARNT heterodimers), both of which are involved in the glycolytic pathway.

A complementary analysis of the promoter region of ANT4 genes from three additional mammals (guinea pig, microbat, and megabat), initially discarded because of incomplete sequences, showed the presence of the same combination of regulatory matrices (result not shown). Several human models were constructed, including V$SMAD3.01/V$MZF1.02/CACGTGTCAGG using different sets of parameters (number of matrices and of nucleotide string copies, distance between matrices and nucleotide strings, and stringency of the each sequence).

An example of model is shown in Fig. 4. Either the full human genome or the Genomatix promoter database was screened to localise the models (Genomatix ModelInspector software) and a set of genes, which represent those genes potentially regulated by the models, were identified (GeneProm software) (Table 4). Only genes with a model located next to their transcription start sites were selected. Moreover, many locations (about half of the results) in clones without identified genes were not retained. Our results identified only known genes encoding proteins with known or unknown functions. Many of our results reveal localisation of models in EMBL clones without any adjacent coding sequence or to sequences supposedly encoding proteins. Amongst the 21 selected sequences, (including ANT4), 16 are directly or indirectly related to spermatogenesis or with a role in testis or in prostate tissues (corresponding genes or proteins referenced in Table 4).

3.3.2. ANT4 promoter 2 (upstream exon 3)

Using a similar FrameWorker analysis from Genomatix, we searched the matrices shared by the selected mammalian sequences

ANT4 promoter1 and exon1 sequences

AAATTTAAGGGGAACATCACTTGAATCACACAACAGGTAAGACTGTCTTTTTTAAGATGCCAAGAAAGGTCAAAG
GAACCATCTCTGGTCAAAAAAAACGTAACGGATTAAAGCGATTAAGAAAGTGAAGGGGGTAACACTAGGAACTAA
AACTAAATTCTGTTAATCTCACACCGCTGGTTTACTGCTCTGGCCTAGTCTCAGCTACTAAATTTCTTCACACAG
AACTGGCTCTCCTCGGCCCCTCCCCTCGCCTTTCCTGCTCAATGCTCACCGCCTCCGGACCCCTCCCTCATCAGA
AAGCCCAGGCTCCGCTCGTAGAAGTGCGCAGGCGTCACCGCGCATCCAGGAGCCACGTGTCAGGAGTCACGTGT
CAGGTGTCACGTGTCAGGCGTCACGTGGCTGGAGGCCGTTGGAGCGCCTGCGCAGCTTTCCGCACGCGCCTCG
CCGGCGCGCGGCTCTCTCAGCGTCCCAAGAGCCACTTTCTCGCCAGTACGATGCTGCAGCGGTTTTCCGGTTTTC
CGCTTCCCTTCATCGTAGCTCCCGTACTCATTTTTAGCCACTGCTGCCGGTTTTTATATCCTTCTCCATCATGCA
TCGTGAGCCTGCGAAAAAGAAGGCAGAAAAGCGGCTG.....

ANT4 promoter2 and exon3 sequences

TTCACCCTTCACAGAAAAAGGGCACAACTAAAGCCTCCATATGTGCCTGACATTGTAATCATAGCAAGCACTATT
ATTAAGAATTATGTAGCTATTAAGGAAAAAAAAAAAGTATACAGCGTAAAGAGATGATTATGCGACCTCTTTTGC
TAATTTGGGCTGCTTTGCTGGAGAACATTAAAATATCCTGGAATTTCAGTATGTCAGCATCTAAACTGTGTCTT
CCAGACTGCATGAAACAAAAGCTATATGGCTATAAAAATAAAATATTTTGGATAGCATGGTATGCATTATTATAG
ATGTTTACTTTATAAAAGTGGCTTATAAATCTGGTGTTTTAACATTTATAAAAATAAGAATTTACCAGGTATT
TTAAACAGTTAGATTCTGTGGTTTAATAACCACTTTTAAATTAATATGTTTCAGTTCTGGAGGTGGTTTTTGGCA
AACCTGGCTTCTGGTGGAGCTGCTGGGGCAACATCCTTATGTGTAGTATATCCTCTAGATTTTGCCCGAACCCGA
TTAGGTGTCGATATTGGAAAAG

Fig. 3. Location of the matrices identified by Genomatix. The matrices and the nucleotide strings are framed; the exons are represented on a grey background. The top of the figure represents the known promoter regions with potential regulation matrices. The bottom of the figure represents the potential promoter identified by the PromoterInspector algorithm. The potential start of the protein (ATG in the exon 3), three potential TATAAA boxes at the end of the second intron and some potential regulation matrices can be seen in this diagram.

from the second supposed *ANT4* promoter. This second promoter was identified by the Genomatix analysis (ElDorado 02–2010 database), but to date no transcripts from this promoter have been identified in humans. Identification of this promoter sequence was based on comparative genomics (Locus: SLC25A31/GXL_33815, promoter belongs to Promoter Set 2 of the Genomatix Homology Group Hg13644). Moreover, relevant transcripts from this promoter are proposed by Eldorado in several mammals (monkey, mouse, rat, dog, ox and opossum). However, this proposed promoter sequence overlaps the coding sequence of *ANT4* in the first 98 nt of exon 3.

The constructed models included several matrices: NKX25.02 (bound by homeodomain factor Nkx-2.5/Csx) (Chen and Schwartz, 1995), DLX1.01 (bound by distal-less homeodomain transcription factors), and HNF3B.02 (hepatic nuclear factor 3beta or FOXA2) (Rada-Iglesias et al., 2005) found upstream of the 5′ end of exon 3, the classical TATAAA, and the ATG sequences. Analyses with different sets of parameters (number of matrices, distance between matrices and nucleotide strings, sequence stringency) identified a set of genes supposedly co-regulated with *ANT4*. Amongst the 15 positive results, 5 correspond to proteins with no known function, and the 10 other are linked to unrelated functions.

4. Discussion

The specific transcriptional regulation of each of the four adenine nucleotide translocator isoforms is an interesting example of multi-isoform gene regulation. The metabolic and physiological consequences of these molecular regulatory mechanisms play a major role in the evolution of cellular metabolic pathways. Each of the four isoforms plays a very precise role in cellular bioenergetics: ANT1 provides mitochondrial ATP for muscle fibre contraction (Stepien et al., 1992), ANT2 allows the maintenance of intra-mitochondrial functions in glycolytic conditions (Barath et al., 1999b; Chevrollier et al., 2005), ANT3 is the constitutively expressed ubiquitous isoform that can be integrated in the mitochondrial membrane when no other isoform is produced (Stepien et al., 1992), and ANT4, which was recently identified (Dolce et al., 2005), is believed to be the isoform that overcomes the absence of the ANT2 isoform (the gene for which is located on the X chromosome) in the Y spermatozoid. Consequently, ANT4 plays a very specific role during impregnation. In our work, we studied the very specific function of the ANT4 isoform through precise analysis of its transcriptional regulation mechanisms. This approach enabled us to propose an ANT4 bioenergetics function and to identify other proteins involved in spermatogenesis and controlled by the same transcriptional regulation mechanism.

4.1. Appearance of the ANT4 gene during evolution

At the stage of evolution where meiotic sex chromosome inactivation arose (*i.e.*, the point of divergence of the eutherian and metatherian lineages) (Turner et al., 2006), the ANT2 isoform, which is essential to cellular metabolism and is encoded on the X chromosome, was replaced by a new isoform with the same function (*i.e.*, uptake of glycolytic ATP towards the mitochondrial matrix) (Brower et al., 2007). This glycolytic ATP requirement is supported by two arguments: the primary glycolytic trend of spermatozoid metabolism and the reverse kinetic of ATP transport (export of ATP produced by the mitochondria). These features are incompatible with

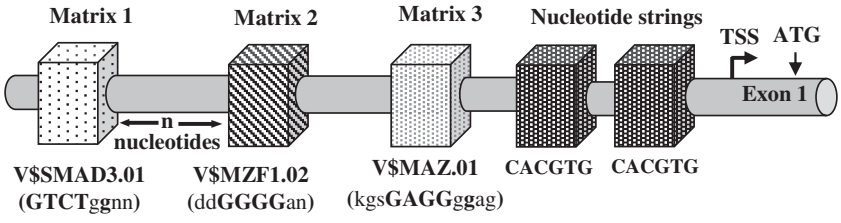


Fig. 4. Schematic representation of a regulatory model. The boxes represent matrices and nucleotide strings included in one of the models constructed for the promoter 1. A matrix is composed of two kinds of sub-sequences: the core (highly conserved sequence in the middle of the matrix, uppercase) and the flanking sequences (less conserved sequences, lowercase) (Genomatix). d: A, T, U, or G (not C); k: T, U, or G; s: C or G. The transcription start site (TSS) is represented by an arrow. The distance between matrices and strings can vary between the lower and upper bounds (given by the user).

Table 4
Screening of *ANT4* co-regulated genes at the level of the first promoter.

Gene, ID	Encoded protein (Ensembl)	Protein function (NCBI, GeneCards)
AMN	Amnionless homolog	Extraembryonic visceral endoderm layer
APEX1	APEX nuclease (multifunctional DNA repair enzyme) 1	Repair of apurinic/apyrimidinic sites **in testis** (Raffoul et al., 2004)
BAX*	BCL2-associated X protein	Mutagenesis regulation in **spermatogenesis** (Xu et al., 2010)
CDK4*	Cyclin-dependent kinase 4	Cell cycle G1 phase progression in male reproduction (Buchold et al., 2007)
FLJ32713 fis	Unknown (TESTI2000756)	Unknown (**expressed in testis**)
HSPBAP1	HSPB (heat shock 27 kDa) associated protein 1	Regulating stress response
IGF2BP3	Insulin-like growth factor 2 mRNA binding protein 3	RNA synthesis/metabolism, major foetal growth factor (Nielsen et al., 1999)
IGF2R*	Insulin-like growth factor 2 receptor	Receptor for insulin-like growth factor 2 (IGF2) and mannose 6-phosphate
KAT5*	K(lysine) acetyltransferase 5	Chromatin remodelling with an abundant **spermatid** protein (Reynard et al., 2009)
LAMP1*	Lysosomal-associated membrane protein 1	Binds amelogenin, differentially expressed in **spermiogenesis** (Guttman et al., 2004)
RMND1*	Required for meiotic nuclear division 1 homolog	Unknown
RPUSD4*	RNA pseudouridylate synthase domain-containing protein 4	Unknown, **expressed in prostate**
SLC25A31	Solute carrier family 25, adenine nucleotide translocator ANT4	Mitochondrial ATP/ADP carrier in **spermatozoid** (Dolce et al., 2005)
SLC2A4	Solute carrier family 2, member 4 (GLUT4)	Facilitated glucose transporter, **detected in human testis** (Angulo et al., 1998)
SOHLH1*	**Spermatogenesis** and oogenesis specific basic helix-loop-helix 1	Germ cell-specific, oogenesis regulator and **male germ cells** (Matson et al.)
TDRD1*	Tudor domain containing 1	Essential for **spermiogenesis** (Yabuta et al., 2011; Wang et al., 2001)
THAP8*	O-sialoglycoprotein endopeptidase (**TESTI2004929**)	Unknown
TKTL1	Transketolase-like 1	Important role in transketolase activity, **testis** expressed (Coy et al., 1996)
TMEM184A	Transmembrane protein 184A = Sdmg1	**Male-specific** expression in embryonic gonads (Svingen et al., 2007)
UBE2B*	Ubiquitin-conjugating enzyme E2B (RAD6 homolog)	Post-replicative DNA damage repair **in spermatogenesis** (van der Laan et al., 2004)
SUN1	Chr. 7 unc-84 homolog A	Nuclear anchorage/migration, expression of **meiotic reproductive genes** (Chi et al., 2009)

Several constructed models of the *ANT4* promoter region 1 were screened as described in Fig. 2 either on the full chromosomal human sequences or the human promoter library (results with an asterisk). Proteins with a function in testis or prostate are shown in bold characters.

the expression of the ubiquitous ANT3 isoform. Indeed, the gene encoding the ANT3 isoform is located on the human Xp22 chromosome, a region that is highly conserved between the X and Y chromosomes and is known to escape sex chromosome inactivation. The process of evolution did not require maintenance of a functional copy of this gene on the Y chromosome.

Our first strategy was to align the promoter sequences of the *ANT4* genes from 29 mammalian species. The alignment of 7 of these sequences enabled us to build several models of transcriptional regulation, which we simultaneously searched for in the whole human genome and in a promoter sequence database (Promoters of all genes from ElDorado 2–2010) (Genomatix). Considering the structure of the models (combinations of short sequences) in our study, the probability of finding these models in the full genomic sequence has a probability proportional to the stringency of the selected parameters. This probability implies that a fraction of the genes (or coding sequences) selected by such an analysis can be false positives with a simple similarity of several parts of their promoter sequence. An additional complication of such an analysis is the supposed presence of a second promoter sequence in the *ANT4* gene located immediately upstream of exon 3. This promoter was identified based on comparative genomics (PromoterInspector algorithms) (Genomatix). No specific transcript starting from the third exon has been detected in humans (Fig. 1). However, this second promoter was also predicted in several other mammalian species (monkey, chimpanzee, mouse, rat, dog, ox, horse, and opossum). In all cases, however, no transcript resulting from this second promoter was confirmed by our own expression analysis (Stepien et al., 1992) or by database analysis. Our analysis only shows that the sequence of this promoter area is evolutionarily preserved better than the sequence of the main *ANT4* promoter. A physiological role of this second promoter remains uncertain.

4.2. Transcriptional regulation of the ANT4 gene

Our promoter analysis (promotology) enabled us to identify a list of 20 genes that are potentially regulated by the regulatory models similar to those identified in the *ANT4* gene promoter. Specifically, our analysis identified a particular proximal sequence string (CACGTG) that was repeated several times in most mammalian *ANT4* promoters (including human). This specific oligonucleotide is a member of the V $HIFF transcriptional factor family (Genomatix), which are known for their role in the response to hypoxia and the switch towards a glycolytic metabolism (Denko, 2008). Amongst the 20 identified genes, 16 (80%) are associated with male reproduction. The ANT4 protein is known to be specifically expressed during spermatogenesis and our results provide a first validation to our computational analysis of promoter sequences: several genes involved in spermatogenesis share with the ANT4 gene particular strings in their promoter and some of these strings (matrices from the V$EBOX and V$HIFF families) are known to be specifically involved in the glycolytic pathway. The last identified genes encode for proteins with different functions: cell cycle progression (CDK4, (Buchold et al., 2007)), cell growth (IGF2BP3 and IGF2R, (Nielsen et al., 1999)), or meiosis (RMND1) which could be associated with spermatogenesis. The presence of the previously proposed E2F6-binding element (Kehoe et al., 2008), which is found in the promoters of genes involved in mouse meiosis, was not used in our analysis: This E2F6-binding element is localised downstream the two proposed transcription start sites in positions 128 651 532 (EnsEMBL) and 128 651 555 (Eldorado) and thus could not be considered as a valid regulatory sequence. Future work will involve a thorough analysis of the expression (microarrays, databases, and protein profiles) and the involvement of these target genes in metabolic pathways. This line of study will investigate the possible specific roles of each of these proteins. Moreover, the promoter sequences of these genes will be carefully analysed to identify any possible second level of transcription regulation that allows for the creation of new models. In this way, it should be possible to identify additional co-regulated genes encoding proteins that take part in the same biological functions.

Several models of regulatory matrices were generated for promoter 2, upstream of the exon 3 coding sequence. Our analysis allowed us to identify 14 genes or coding sequences possibly controlled by the same transcription factors as the *ANT4* gene. The 10 proteins with known function are involved in very different cellular pathways, and none of these proteins are involved in spermatogenesis. Moreover, the absence of an experimentally validated *ANT4* transcript containing only exons 3 to 6 is not in favour of the expression of such a truncated protein for which no biological role can be proposed.

4.3. ANT isoforms and the sex of the offspring

Although several authors have affirmed that a normal ATP supply through oxidative phosphorylation is essential for male germ cell meiosis (Brower et al., 2007), this idea could be refuted by several arguments. First, taking into account the absence of the ANT2 isoform (*ANT2* gene on the X chromosome) in the male spermatozoid, ANT4 would be presumed to carry out the same biological function (transporting cytosolic glycolytic ATP into the mitochondria (Chevrollier et al., 2005)) and, consequently, would be presumed to be essential for glycolytic metabolism with an arrest of mitochondrial oxidative phosphorylation. Second, the bioenergetics of the spermatozoid at the time of fecundation supposes a mainly glycolytic metabolism resulting from the anaerobic environment of the female genital tract. In this case, spermatozoid motility would rest primarily on glycolytic ATP with the maintenance of intra-mitochondrial metabolic pathways essential to cell survival. Thus, ANT4 would be required for this survival of the Y spermatozoid and not for its motility. This result is confirmed in *Ant4*-deficient female mice, which do not present a significant decrease in fertility (Brower et al., 2009). There is no meiotic sex chromosome inactivation in female mammals, and the role of ANT4 can be compensated for by ANT2 (which may have the same kinetic properties). The consequence of this proposal is interesting: the kinetics of progression in the genital tract and the respective survival of the X and Y spermatozoids could be directly related to the expressed ANT isoform in the spermatozoids (ANT4 in Y spermatozoid and ANT4 + ANT2 in the X).

It is likely that the ANT2 and 4 proteins are regulated differently and thus expressed at different level. Moreover, it is almost impossible that the kinetic parameters of the two isoforms are identical. Presumably, the ANT2 isoform (already present in yeast) has higher performing kinetic properties than ANT4, which appeared after the divergence of the eutherian (placental) and metatherian (marsupial) lineages (Turner et al., 2006). Thus, X and Y spermatozoid bioenergetics should be slightly different: the relatively weaker ATP uptake into the mitochondria by ANT4 should lead to greater cytosolic glycolytic ATP concentrations in the Y spermatozoid. This higher cytosolic ATP concentration would presumably lead to higher motility of the Y spermatozoid. However, as glycolytic metabolism is dependent on intra-mitochondrial functions requiring imported ATP (Giraud et al., 1998), the survival capability of the Y spermatozoids could be slightly lower than that of the X spermatozoids that exhibit higher ATP uptake and, consequently, better performing mitochondria. Such differences should be low because they rely on slightly different kinetic parameters between the two isoforms and a fraction of the ANT pool in X and Y spermatozoids. As a result, this survival and motility difference could be linked to the ovulation time for a parental selection of the gender of their children.

5. Conclusions

Our bioinformatics analysis allowed us to clarify the structure and the organisation of the promoter of the *ANT4* gene. The second promoter located inside an intron sequence was not validated in this study. We propose a set of genes that are specifically co-expressed during spermatogenesis and are likely to be transcriptionally regulated in a manner similar to the *ANT4* gene. Further analysis of the regulation of these genes will lead to the modelling of bioenergetic metabolic pathways and their specific regulation in spermatozoids. This study enabled us to suggest a bioenergetic explanation that accounts for the differences in the motility and survival of the X and Y spermatozoids. Lastly, this type of computational analysis that associates phylogeny and promoter analysis should be applicable to any human or animal model.

Acknowledgments

Thanks are due to the Soluscience company for helpful expert assistance in informatics. This work was supported by the Cancéropôle Lyon Auvergne Rhône-Alpes (CLARA), "Nutrition Métabolisme et Cancer" ProCan axis, to G.S; and the Conseil Régional Auvergne to GS (LifeGrid funds) and P.-Y.D ("Innovation Région" funds).

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.gene.2011.07.024.

References

Angulo, C., et al., 1998. Hexose transporter expression and function in mammalian spermatozoa: cellular localization and transport of hexoses and vitamin C. J. Cell. Biochem. 71, 189–203.
Barath, P., Albert-Fournier, B., Luciakova, K., Nelson, B.D., 1999a. Characterization of a silencer element and purification of a silencer protein that negatively regulates the human adenine nucleotide translocator 2 promoter. J. Biol. Chem. 274, 3378–3384.
Barath, P., Luciakova, K., Hodny, Z., Li, R., Nelson, B.D., 1999b. The growth-dependent expression of the adenine nucleotide translocase-2 (ANT2) gene is regulated at the level of transcription and is a marker of cell proliferation. Exp. Cell Res. 248, 583–588.
Brower, J.V., et al., 2007. Evolutionarily conserved mammalian adenine nucleotide translocase 4 is essential for spermatogenesis. J. Biol. Chem. 282, 29658–29666.
Brower, J.V., Lim, C.H., Jorgensen, M., Oh, S.P., Terada, N., 2009. Adenine nucleotide translocase 4 deficiency leads to early meiotic arrest of murine male germ cells. Reproduction 138, 463–470.
Buchold, G.M., Magyar, P.L., O'Brien, D.A., 2007. Mice lacking cyclin-dependent kinase inhibitor p19Ink4d show strain-specific effects on male reproduction. Mol. Reprod. Dev. 74, 1008–1020.
Cartharius, K., et al., 2005. MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics 21, 2933–2942.
Chen, C.Y., Schwartz, R.J., 1995. Identification of novel DNA binding targets and regulatory domains of a murine tinman homeodomain factor, nkx-2.5. J. Biol. Chem. 270, 15628–15633.
Chevrollier, A., et al., 2005. ANT2 isoform required for cancer cell glycolysis. J. Bioenerg. Biomembr. 37, 307–316.
Chi, Y.H., et al., 2009. Requirement for Sun1 in the expression of meiotic reproductive genes and piRNA. Development 136, 965–973.
Coy, J.F., et al., 1996. Molecular cloning of tissue-specific transcripts of a transketolase-related gene: implications for the evolution of new vertebrate genes. Genomics 32, 309–316.
Denko, N.C., 2008. Hypoxia, HIF1 and glucose metabolism in the solid tumour. Nat. Rev. Cancer 8, 705–713.
Dolce, V., Scarcia, P., Iacopetta, D., Palmieri, F., 2005. A fourth ADP/ATP carrier isoform in man: identification, bacterial expression, functional characterization and tissue distribution. FEBS Lett. 579, 633–637.
Fiermonte, G., De Leonardis, F., Todisco, S., Palmieri, L., Lasorsa, F.M., Palmieri, F., 2004. Identification of the mitochondrial ATP-Mg/Pi transporter. Bacterial expression, reconstitution, functional characterization, and tissue distribution. J. Biol. Chem. 279, 30722–30730.
Frech, K., Danescu-Mayer, J., Werner, T., 1997. A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. J. Mol. Biol. 270, 674–687.
Giraud, S., Bonod-Bidaud, C., Wesolowski-Louvel, M., Stepien, G., 1998. Expression of human ANT2 gene in highly proliferative cells: GRBOX, a new transcriptional element, is involved in the regulation of glycolytic ATP import into mitochondria. J. Mol. Biol. 281, 409–418.
Guttman, J.A., Takai, Y., Vogl, A.W., 2004. Evidence that tubulobulbar complexes in the seminiferous epithelium are involved with internalization of adhesion junctions. Biol. Reprod. 71, 548–559.
Kehoe, S.M., et al., 2008. A conserved E2F6-binding element in murine meiosis-specific gene promoters. Biol. Reprod. 79, 921–930.
Kim, Y.H., Haidl, G., Schaefer, M., Egner, U., Mandal, A., Herr, J.C., 2007. Compartmentalization of a unique ADP/ATP carrier protein SFEC (Sperm Flagellar Energy Carrier, AAC4) with glycolytic enzymes in the fibrous sheath of the human sperm flagellar principal piece. Dev. Biol. 302, 463–476.
Ku, D.H., Kagan, J., Chen, S.T., Chang, C.D., Baserga, R., Wurzel, J., 1990. The human fibroblast adenine nucleotide translocator gene. Molecular cloning and sequence. J. Biol. Chem. 265, 16060–16063.
Li, K., Hodge, J.A., Wallace, D.C., 1990. OXBOX, a positive transcriptional element of the heart-skeletal muscle ADP/ATP translocator gene. J. Biol. Chem. 265, 20585–20588.
Luciakova, K., Barath, P., Poliakova, D., Persson, A., Nelson, B.D., 2003. Repression of the human adenine nucleotide translocase-2 gene in growth-arrested human diploid cells: the role of nuclear factor-1. J. Biol. Chem. 278, 30624–30633.
Matson, C.K., Murphy, M.W., Griswold, M.D., Yoshida, S., Bardwell, V.J. and Zarkower, D. The mammalian doublesex homolog DMRT1 is a transcriptional gatekeeper that controls the mitosis versus meiosis decision in male germ cells. Dev Cell 19, pp. 612–24.

Nielsen, J., Christiansen, J., Lykke-Andersen, J., Johnsen, A.H., Wewer, U.M., Nielsen, F.C., 1999. A family of insulin-like growth factor II mRNA-binding proteins represses translation in late development. Mol. Cell. Biol. 19, 1262–1270.
Palmieri, F., 2004. The mitochondrial transporter family (SLC25): physiological and pathological implications. Pflugers Arch. 447, 689–709.
Rada-Iglesias, A., et al., 2005. Binding sites for metabolic disease related transcription factors inferred at base pair resolution by chromatin immunoprecipitation and genomic microarrays. Hum. Mol. Genet. 14, 3435–3447.
Raffoul, J.J., Cabelof, D.C., Nakamura, J., Meira, L.B., Friedberg, E.C., Heydari, A.R., 2004. Apurinic/apyrimidinic endonuclease (APE/REF-1) haploinsufficient mice display tissue-specific differences in DNA polymerase beta-dependent base excision repair. J. Biol. Chem. 279, 18425–18433.
Reynard, L.N., Cocquet, J., Burgoyne, P.S., 2009. The multi-copy mouse gene Sycp3-like Y-linked (Sly) encodes an abundant spermatid protein that interacts with a histone acetyltransferase and an acrosomal protein. Biol. Reprod. 81, 250–257.
Scherf, M., Klingenhoff, A., Werner, T., 2000. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. J. Mol. Biol. 297, 599–606.
Stepien, G., Torroni, A., Chung, A.B., Hodge, J.A., Wallace, D.C., 1992. Differential expression of adenine nucleotide translocator isoforms in mammalian tissues and during muscle cell differentiation. J. Biol. Chem. 267, 14592–14597.
Svingen, T., et al., 2007. Sex-specific expression of a novel gene Tmem184a during mouse testis differentiation. Reproduction 133, 983–989.
Tanay, A., Gat-Viks, I., Shamir, R., 2004. A global view of the selection forces in the evolution of yeast cis-regulation. Genome Res. 14, 829–834.
Turner, J.M., Mahadevaiah, S.K., Ellis, P.J., Mitchell, M.J., Burgoyne, P.S., 2006. Pachytene asynapsis drives meiotic sex chromosome inactivation and leads to substantial postmeiotic repression in spermatids. Dev. Cell 10, 521–529.
van der Laan, R., et al., 2004. Ubiquitin ligase Rad18Sc localizes to the XY body and to other chromosomal regions that are unpaired and transcriptionally silenced during male meiotic prophase. J. Cell Sci. 117, 5023–5033.
Wang, P.J., McCarrey, J.R., Yang, F., Page, D.C., 2001. An abundance of X-linked genes expressed in spermatogonia. Nat. Genet. 27, 422–426.
Xu, G., Vogel, K.S., McMahan, C.A., Herbert, D.C., Walter, C.A., 2010. BAX and tumor suppressor TRP53 are important in regulating mutagenesis in spermatogenic cells in mice. Biol. Reprod. 83, 979–987.
Yabuta, Y., Ohta, H., Abe, T., Kurimoto, K., Chuma, S., Saitou, M., 2011. TDRD5 is required for retrotransposon silencing, chromatoid body assembly, and spermiogenesis in mice. J. Cell Biol. 192, 781–795.

DATABASES

EnsEMBL: Ensembl genome Browser EBI [http://www.ensembl.org].GeneCards: Gene Database Weizmann Institute of Science, [http://www.genecards.org].Genomatix: Personalized Medicine - Relevance for scientists [http://genomatix.de].KEGG: Kyoto Encyclopedia of Genes and Genomes [http://www.genome.jp/keg].NCBI: National Center for Biotechnology Information [http://www.ncbi.nlm.nih.gov].

Annexe B : Computational identification of transcriptionally co-regulated genes, validation with the four ANT isoform genes

Computational identification of transcriptionally co-regulated genes, validation with the four ANT isoform genes

Pierre-Yves Dupont[a] and Georges Stepien[a*]

[a]INRA, UMR 1019, UNH, F-63122 St Genès-Champanelle; Clermont Université, Université d'Auvergne, Unité de Nutrition Humaine, BP 10448, F-63000 Clermont-Ferrand

*To whom correspondence should be addressed. Tel: +33 473624458; Fax: +33 473624755; Email: georges.stepien@clermont.inra.fr

Abbreviations

AAC: ADP/ATP carrier; ANC: adenine nucleotide carrier; ANT: adenine nucleotide translocator; SLC25: solute carrier family 25; TSS: transcription start site

Abstract

The analysis of the promoters of genes is an essential step towards the comprehension of the mechanisms of transcriptional regulation required during the evolution of biological networks under the effects of physiological processes, nutritional intake or pathologies. In higher eukaryotes, transcriptional regulation implies the recruitment of a whole set of regulatory proteins which bind on combinations of nucleotidic motifs. We developed a computational phylogenetic analysis, the promotology, combining several programs allowing to build regulatory models and to carry out a crossed research on several databanks (Genomatix, Ensembl) of genes likely to be coregulated. This strategy was tested on a set of four human genes encoding the isoforms 1 to 4 of the mitochondrial ATP/ATP carrier ANT (*Adenine Nucleotide Translocator*). Each isoform has a specific tissue expression profile linked to its role in the cellular bioenergetic. The study by promotology of these four genes allowed to construct, from their promoter sequence and from the evolution of these sequences in the mammals phylogeny, combinations of specific regulatory elements. These models were screened on the full human genome, on databases of promoter sequences from Human and several mammalian species. For each three transcriptionally regulated *ANT: ANT1, 2 and 4* genes, a set of co-regulated genes was identified. Most of them have a cellular function and specificity in agreement with those of the corresponding *ANT* gene. An example of the implication of such a transcriptional co-regulation on the cellular metabolism is described for each ANT isoform.

INTRODUCTION

The promotology

The metazoan genome is composed of coding sequences flanked by not-coding regions containing regulatory elements. These elements consist in short nucleotidic sequences, which allow to program the gene expression at a given time and in a specific cell (1). In spite of significant progresses in genomics, which led to the identification of most genes of the human genome, the knowledge of their transcriptionnal regulation remains unclear. The promotology lean on a set of bioinformatic tools allowing the study of mechanisms of this genome regulation. It relies on an analysis of gene promoter sequences which would include a combination of several regulatory elements (from ten to fifty nucleotide length) named cis-regulatory modules (CRM). Modules are recognized by regulatory factors allowing activation or repression of gene transcription. These regulatory elements form transcription factor binding sites (TFBS). TFBS are difficult to differentiate from nonfunctional random genomic sequences (2). Most of the elements would be localised on the first 300 nucleotides upstream of the transcription start site (TSS) (3). These regulatory sequences and the genes that they govern define functional spaces in the genome. Thus, they play a major role in the development, the environmental adaptation, the response to nutritional uptake and the pathogenesis.

Four genes code for four isoforms of the human adenine nucleotide translocator (ANT)

The ANT, also called by the generic term "ADP/ATP carrier" (AAC), is a protein encoded from the nuclear genome, inserted in the mitochondrial internal membrane. It allows the exchange of the ATP and ADP adenylic nucleotides between the mitochondrial matrix and the cytoplasm. Such a function is of primary importance since the ANT would be the main protein of the mitochondrial internal membrane able to convey this energy.

The importance of this ANT protein is stressed by the fact that there exists, from yeasts to Humans, four isoforms with different kinetic properties, encoded from four independent genes, each with a specific expression depending on the nature of the tissue, cell type, developmental stage and status of cell proliferation. This allows to adapt the energy production to the metabolic parameters linked to the cell environment and cycle (Table I) (4). The peptidic sequences of these four isoforms are very close (96% of homology): they differ only by several amino acids which would be involved in ATP and ADP interaction sites.

Table I. Nomenclature of the four *ANT* genes and their expression in human, mouse and yeast

Human isoform	**HGNC symbol** (locus) **Ensembl gene ID**	**Expression**	**Regulatory factors**	**Corresponding isoform**	
				mouse	yeast
ANT1 or ANC1	*SLC25A4* (4q35.1) ENSG00000151729	heart and skeletal muscles	OXBOX (5)	Ant1	AAC1
ANT2 or ANC3	*SLC25A5* (Xq24) ENSG00000005022	proliferative cells	GRBOX (6); Oct1, SV40, AP2, Sp1 (7) (8)	Ant2	AAC3
ANT3 or ANC2	*SLC25A6* (Xp22.33) ENSG00000169100	ubiquitous	Sp1, NF-kB, GAS-like (9)	-	AAC2
ANT4	*SLC25A31* (4q28.1) ENSG00000151475	spermatozoids	E2F6 (10)	Ant4	-

ANT (Ant): adenine nucleotide translocator; ANC: adenine nucleotide carrier; AAC: ATP / ADP carrier; HGNC (11); SLC25 solute carrier gene family (12).

The specific transcriptional regulation of each of the four adenine nucleotide translocator isoforms (ANT) is an interesting example of multi-isoform gene regulation. The metabolic and physiological consequences of these molecular regulatory mechanisms play a major role in the evolution of cellular metabolic pathways. Each of the four isoforms is known to play a specific role in cellular bioenergetics: ANT1 provides mitochondrial ATP for heart and skeletal muscle contraction (13). The kinetic properties of this ANT1 isoform allow a rapid and massive mitochondrial ATP export required for the muscle contraction. The second isoform, ANT2, is not or weakly expressed in human tissues. It allows the maintenance of intra-mitochondrial functions in glycolytic conditions required in proliferative cells (14); (15). ANT2 is known to do an opposite transport as that of ANT1, transport of glycolytic ATP import towards the mitochondrial matrix (13). We identified a specific regulatory sequence in the promoter region of the human ANT2 gene: the GRBOX element (Glycolysis Regulated Box), localised -1196 to -1184 (CATTGTTATGATT) upstream of the TSS (6). A part of this sequence is present in the yeast corresponding *AAC3* gene, encoding an isoform specifically expressed in anaerobic conditions. This sequence is recognized by the yeast ROX1 protein (16), a regulatory protein known to repress several other genes under the oxygen effect (17) (18); (19). ANT3 is the constitutively expressed ubiquitous isoform that is allowed to be integrated in the mitochondrial membrane when no other isoform is produced

(13). In rodents, the *Ant3* gene was lost during evolution. Probably, the rodent physiology would not require, such as in Human, two isoforms with different kinetics (ANT1 and ANT3). This assumption would be supported by the disappearance, in the rodents, of the OXBOX regulatory element from the *ANT1* promoter (5), which would determine a muscle specific expression of this isoform. The last isoform, ANT4 (*SLC25A31* gene), was recently identified in humans, and it is expressed mainly in the testicle (20). This isoform appears in mammals and is essential during spermatogenesis (21). Its peptide sequence is very similar (66-68 % of identity) to that of other ANT isoforms. The main characteristic of this isoform is the presence of additional peptides, specifically the N- (13 amino acids) and C- (8 amino acids) terminal sequences, which the other three isoforms lack. These extra peptides could be related to the specific localisation of this isoform to the sperm flagellum (22). The proposed hypothesis for the role of this isoform is that it appears to compensate for the loss of function of the *ANT2* gene (encoded by the X chromosome) during male meiosis (21).

In this computational analysis, we investigated and compared the mechanisms of transcriptional regulation of the four ANT isoforms through analysis of nucleotide sequences upstream of the supposed sites of transcription initiation. The nucleotide sequences of these promoter regions from several mammalian species were compared to follow the phylogeny of specific sequences of transcriptional regulation. Such promoter sequences preserved throughout evolution might be of major importance to the survival of the organism (23). This study is based on a combination of software and databases (some available on-line, such as Genomatix (24) and EnsEMBL (25), some available in our laboratory, such as GeneProm).

Materials and methods

Process of bioinformatics study

An outline of the bioinformatic pipeline implemented for *ANT* sequences analysis is illustrated in Figure 1. Our study was performed in five steps. We began by a short phylogenetic study of the four *ANT* genes from various mammalian species. This step enabled us to check database annotations and eliminate sequences which were impossible to align to the four human ANT genes. The application of the Genomatix tool MatInspector (24) to the selected sequences gave us a list of regulatory elements identified in the aligned sequences. These elements are represented either by matrices (resulting from combination of short DNA sequences) or nucleotide strings (IUPAC sequences). We organised these elements into regulatory models which we screened by ModelInspector (Genomatix) across the whole human genome (GenBank clones) or in a human promoter database Eldorado (Genomatix). Genes were identified from their localisation on GenBank clones with GeneProm software we developed. Genes

retrieved from the Genomatix human promoter database were filtered with a search for phylogenetically conserved models (Genomatix). Thus, we obtained a list of genes that were potentially co-regulated by similar transcriptional models. GeneProm software consists in a web application that will be available as soon as our web server will be capable to manage many users (sources are available on demand).

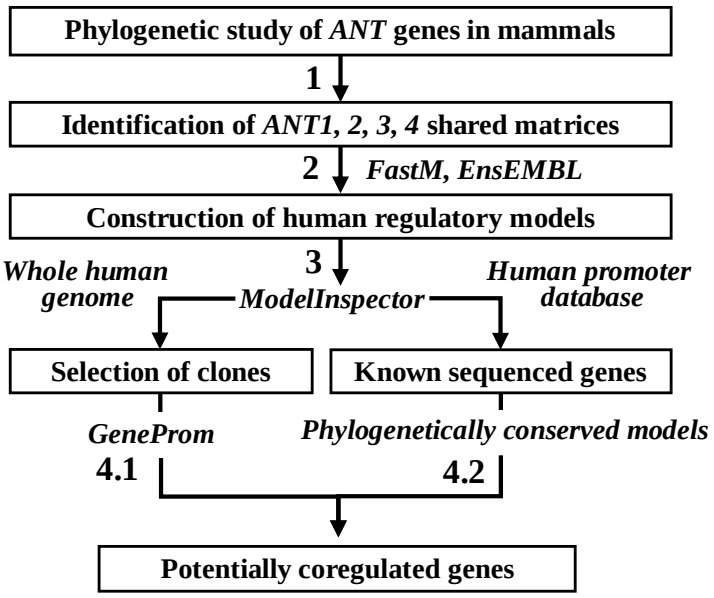


Fig. 1. The five steps of the bioinformatics study
1 - Phylogenetic alignment of ANT genes in various mammalian species to filter unusable sequences; 2 - Identification of ANT regulation matrices shared by all aligned sequences from each ANT isoform gene (via Genomatix FrameWorker and MatInspector tools); 3 - Construction of regulation models by combining the previously selected matrices and location of the model in the whole human genome or in the ElDorado promoter database; 4.1 - Identification of potentially co-regulated genes by GeneProm and 4.2 - validation of genes from the human promoter database by phylogeny (Genomatix).

Phylogenetic study of the *ANT4* gene in mammals

Nucleotide sequences of the four *ANT* isoform genes were screened in 30 mammalian species including Human. Sequences were imported from the EnsEMBL database (25). These sequences included the 500 first bases of the gene including exon 1 and an additional 1500 nt sequence upstream of the TSS.

Manual sorting of these sequences was carried out by withdrawing the sequences containing either a high number of undefined bases (more than 10% in the promoter region) or annotation errors at the transcription initiation sites (identified by the alignment of the sequences, including the first exon for each specie). A phylogeny reconstruction of the retained coding sequences was carried out to validate the functional annotations of these genes. The alignment algorithm used was CLUSTAL 2.0 with the default settings. The phylogeny was rebuilt using two different methods: the neighbour-joining method with minimum evolution and UPGMA (Unweighted Pair Group Method with Arithmetic Mean) both with bootstrap evaluation. Both techniques provided concordant results. This phylogeny was carried out to verify the annotations of the sequences in the databases and that the promoter sequences can be compared.

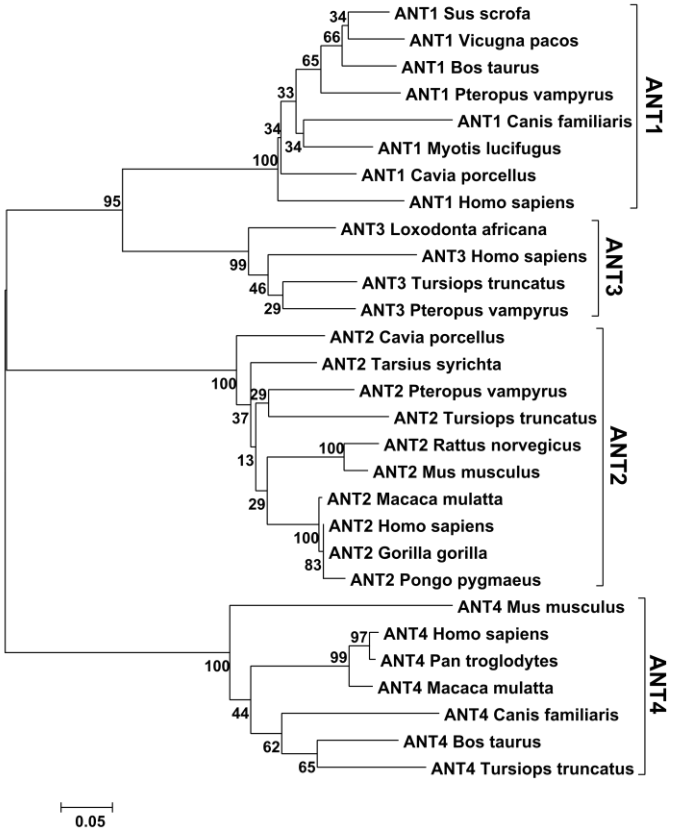


Fig. 2. Evolutionary relationships of taxa.
The evolutionary history was inferred using the Neighbor-Joining method (26). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (100 replicates) is shown next to the branches (27). The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances (number of base substitutions per site) were computed using the Maximum Composite Likelihood method (28). The analysis involved 29 nucleotide sequences. All positions containing gaps and missing data were eliminated. Evolutionary analyses were conducted in MEGA5 (29)

After considering the phylogenetic results of our study, we decided to keep only a small number (29) of mammalian sequences that did not contain undetermined bases in their promoter region and were not too divergent from the human sequence (Figure 2 and supplemental Table 1). Additionally, the transcription initiation sites of all the retained sequences were compatible with the known human initiation site. The alignment of 5' sequences (1500 nt upstream of the ATG site) was performed using the CLUSTAL 2.0 software. Similar alignments have been obtained using MUSCLE. The sequences of the four *ANT* genes (SLC25A4, SLC25A5, SLC25A6, SLC25A31) were aligned to verify their annotation in the EnsEMBL database. Thus, the list of 15 selected species takes into account the sequencing quality of this genomic region (Figure 2 and supplemental Table 1).

Identification of transcriptional matrices shared by each *ANT* sequence

The FrameWorker tool of the Genomatix software package was used in this part of the analysis (24). This tool identified transcriptional regulatory sequences from the bibliography, called regulatory matrices, which were shared by a set of gene promoter sequences. Matrices were identified from the Genomatix database produced by the alignment of all regulatory elements identified to date in all vertebrates. The database includes 727 vertebrate matrices classified into 170 families (noted V$) and 16 additional general

matrices from higher organisms, classified into ten families (noted O$). Additionally, the FrameWorker tool enables identification of the regulatory models shared by different promoter sequences. The MatInspector tool simultaneously allows the identification of the regulatory matrices in a promoter sequence and the analysis of different parameters.

Construction of regulatory models and screening in mammalian genomes

The nucleotide distances between the identified different matrices of the most relevant models were then bound between minimal and maximal values. Previously identified and potentially interesting nucleotide sequences in the studied promoter area were then combined with the set of selected matrices. The resulting new model, which generally included from 4 to 6 matrices or nucleotide strings (Table II), could be searched with the selected stringencies in three databases:

1 - The full human genome (GenBank Release 180) using the ModelInspector Genomatix tool (30). A list of clones (contigs) containing the studied model was obtained in this analysis with the exact positions of the identified models and their orientation on each clone;

2 - The Eldorado 08-2011 database of human promoter sequences (24) including a set of about 120.000 promoter sequences associated with transcripts (predicted by eukaryotic pol II promoter regions)

3 - The Eldorado 08-2011 database of a set of mammalian promoter sequences. The "Search for phylogenetically conserved promoter models" tool allows to search for models which are conserved in orthologous promoter sequences of several mammalian species. This gives evidence for the functionality of promoter models through their preservation during evolution.

The different stringency parameters of this study (ModelInspector parameters) were as follows: maximum number of mismatches allowed for either matrix or nucleotide string; threshold (number of matrices or strings present in the sequence *vs.* the number of matrices or strings to find); research of individual or matrix families; global and core matrix similarities; matrix sense in relation to the clone; and minimal and maximal distances between two matrices or nucleotide strings.

Identification and selection of known, supposed or unknown genes containing a model in their proximal promoter regions

The list of GenBank clones proposed by ModelInspector was exported into our GeneProm software. GeneProm allowed us to find chromosomal location of models and to identify genes present in the immediate proximity (or in partial overlap) of the considered model (Figure 3).

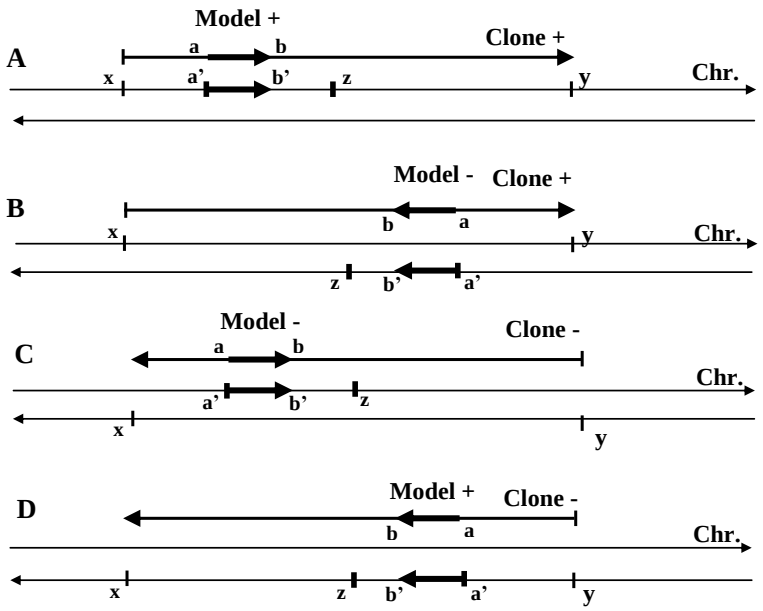


Fig. 3. GeneProm filters for the respective orientations of the model, the identified gene, the EMBL contig and the chromosomal DNA strand.

2000 nt was the maximal distance between the 5' end of the model and the TSS of a gene. **A** (model + / clone + / chromosome +): **a'** = x + a and **z** = x + a + 2000; **B** (-/+/-): **a'** = x + a and **z** = x + a – 2000; **C** (- /- / +): **a'** = y - a and **z** = y – a + 2000; **D** (+/-/-): **a'** = y - a and **z** = y – a – 2000.

The filters used in this analysis were: 1 - the respective orientations of the model, the EMBL contig, the chromosomal DNA strand and the identified gene; 2 – the maximal distance from the 3' end of the model to the transcription start site of the gene (500 nt by default); 3 - the maximal length of the overlapped sequence between the model and the 5' end sequence of the identified gene (200 nt by default). These filters permitted selection of genes directly downstream of the researched model from the same strand of DNA with the same orientation.

Analysis and confirmation of co-regulated genes and their link to known metabolic pathways

The list of the genes obtained from the two analyses (ModelInspector on GenBank and on Eldorado database) was manually analysed: an exhaustive bibliography was obtained for each gene (HUGO Gene Nomenclature Committee, HGNC) to identify the function of each corresponding protein and their direct or indirect links with the ANT protein. GeneProm software was then paired by a script with the KEGG database (31), allowing identification of the metabolic networks involved in the protein encoded from each gene identified in the previous step. A complementary bibliographic analysis (HGNC and PubMed) was carried to relate the identified proteins to suggested metabolic pathways. Key proteins involved in these pathways could be identified.

Results

Alignment and selection of ANT promoter sequences from mammalian species

The mammalian species were screened for the homologous *ANT* gene sequences. 30 mammalian species were screened for the four *ANT* genes and *ANTs* sequences were imported: 1500 nt upstream of the gene sequence and the first 500 nt of the gene including exon 1. The selected species correspond to the 30 mammalian sequences available when the sequences were searched on EnsEMBL. The mammalian sequences from each *ANT* gene were aligned using CLUSTAL software and several sequences were selected according for the correct overlap of their 5' upstream first exon sequences. Mammalian sequences with gaps, incomplete sequences or with more than 10% unknown nucleotides were discarded. 8 sequences (including Human) were obtained for the *ANT1* gene, 10 for the *ANT2* gene, 4 for the *ANT3* gene and 7 for the *ANT4* gene (Figure 2).

Identification of transcriptional matrices shared by the ANT1, 2, 3 and 4 gene sequences

Using the FrameWorker tool from Genomatix, we found several matrices shared by the selected mammalian promoter sequences of each ANT gene. Human models were constructed (Table II) using different sets of parameters (number of matrices and of nucleotide string copies, distance between matrices and nucleotide strings, and stringency of each sequence). An example of model is shown in the promoter sequence of each isoform (Figure 4). The distance between matrices and nucleotide strings was by default set to twice the distance given for the human corresponding ANT promoter.

- ANT1 (SLC25A4): Most involved matrices and IUPAC strings from the ANT1 models are directly or indirectly involved in muscle cell growth and differentiation: V$MEF2 (bound by the myocyte-specific enhancer factor) is involved in a mechanism, conserved from flies to mammals, to fine-tune gene expression in each muscle and probably other tissues (32). V$GATA, such as GATA4 element binding factors, controls cardiomyocyte proliferation by regulating numerous genes involved in the cell cycle (33). GATA4 also controls a regulatory pathway that regulates PGC-1alpha gene expression in skeletal muscle (34). V$NRF1 is bound by nuclear respiratory factor 1, bZIP transcription factor that acts on nuclear genes encoding mitochondrial proteins through MEF2 (myocyte enhancer factor 2A) (35). V$NRF2.01, bound by the nuclear respiratory factor 2 (ETS1 factors), activates the PDGF-A transcription in smooth muscle cells (36). V$KLFS Krueppel like transcription factors are critical regulators of cell differentiation such as in skeletal and smooth muscle development (37). V$STAT (signal transducer and activator of transcription) is involved in a signalling cascade with a crucial role in regulating myogenesis (38). The OXBOX motif is a positive transcriptional element of the heart-skeletal muscle ADP/ATP translocator gene (5).

>***SLC25A4*** chromosome:GRCh37:4:186063885:186064524:1
CTTAGTATCTATTTGAGAGGACTGCCATTTATTCTCGGGAGCGCACGGCTCTAAAGAGGCCCATATCCAGGCAGTGAGCTCTGGTGGGGGGCGCCTTTA
GATGCAAGAAGGAGGAAACAGCTCGAAATCCCTGGGCCTGAGCGCGGCCCGTGCAGGCCGGAGGGTCAAGAACTCTCCACCGGCGGCAGCGGCCCGGTG
TCTGCCCCGGCTTCGCCCCGGCCTAAGGCTGCCTGTGCTATAAATACGCGGCCCACATGCCGCGGTGACACGGTGTTCCCTGGGCTCGGCGGGACAGAT
AACATGAATGTGCCCTTTAAACGTCCCAAGTTGCAGGGACAGCCCCCGGCCCAGCCTCGCTCCCGGAAGCGCCTTCGCCCCCGATGCCCTCTGCAGCTG
GGAGGAGGGGGCGCCCCGCACCTGCCCAGCCAATGCGCGGCGCGAGCGCCGGCCGCGACCCGCCTCCTCTCGCGAGAGCCCGGCGGGGATATAAGGGGG
AGCTGCGGGCCAGGCGGCGGCCCCCTAGCGTCGCGCAGGGTCGGGGACTGCGCGGCGGTGCCAGGCCGGGCGTGGGCGAGAGCACGAACGGGCTGCCTG
CGGGCTGAGAGCGTCGAGCTGTCACCATGGGTGATCACGCTTGGAG

>***SLC25A5*** chromosome:GRCh37:X:118601203:118602499:1
CATTGTT……AAGGGTTCTTTTGGCATCCCTGTAGCAAATGCGTCTCACCCTAGTCCTGGTCCTGCTCCAAGGGTTTTTGTCCAGGCACATCGTGACC
TCACCCTTCCTCCCCTCTCCGAGGCCTCTCTCAGGGTCCAGCGTTCAAGTCCCGGGTGTTCTCTGGACCCGCCCCTTCCTCTCGCCGGGTCAGGTGCCG
AGGGCGAGCACGGGCGGCGCGGAGAGCAGTCCCGGCCCGCCCTCCACGACTCCTCCTCCTGCGAGCTGGGCCACTCAGCGGCTCTGCCGCCCAGCGCGC
CGGGGCCCAGACCCCGCCCCGGCCCCGCCTCCGACGCCTGCCGCTCC**AGCTCCGGCTCCCCCTATATAAATCGGC**CATTTGCTTCGCTCCGCCCCGCAG
CGCCGGAGTCAAAGCCGGTTCCCGGCCCAGTCCCGTCCTGCAGCAGTCTGCCTCCTCTTTCAACATGACAGATGCCGCTGTGTC

>***SLC25A6*** chromosome:GRCh37:X:1511577:1511976:-1
AGCAGGGAAGGAAGAGACGGGGAGAAAAGGGCGGGCGGGAGAGAGCCGGGAAGGGAGAGGAGGCGGGAGGAGGGAGGCGGGAGGGAGGGAGG
AGGGAGGCGGGAGGGAGGGAGGGAGGGAGGGAGGAAGGGCGGAGTGGGAGGTCGCGAAGGAGGACCCCGGCCGCAGCCCCCGCCGGGCAGACACGGGCA
CAGACACGGGGAAATAGGGGAGCCCGGGCGGGTGGGCGCGGCCCACGGAACCCCCAGCCCCGGCGCGGAGGGCACCGCGGGGTCAAAGAACGGAGGCCC
AGAGAGATCCCCCCCCCACCAACCTAGGCCGGGCGCGGAGGGCGGCGCGGAGTCACGGATCGGCAGCGGGACTGAGCCGCGCCCACGTGGAAGCCGCCG

>***SLC25A31*** chromosome:GRCh37:4:128650863:128651663:1
ATCAATTTGAAAAACAAAACAAAACACCCCTCATCCCCTTGAATGTGCAGTCTGGATTGAGAACCACTGCTCTGGATATTATCAATTTAAATACAATAT
GAAAAAGAGGAACATGTATTATGCTATGGGGGCACAATCATCAAAACGACACAGAAAGTGACTAGACAAAGGATCCCTGCTTAATTAAAGGAACAGATT
AAAGCAATTTCCAAGGGAATTTGAAGCATTAACAGGCACATAGGGTGAAATTTAAGGGGAACATCACTTGAATCACACAACAGGTAAGACTGTCTTTTT
TAAGATGCCAAGAAAGGTCAAAGGAACCATCTCTGGTCAAAAAAAACGTAACGGATTAAAGCGATTAAGAAAGTGAAGGGGGTAACACTAGGAACTAAA
ACTAAATTCTGTTAATCTCACACCGCTGGTTTACTGCTCTGGCCTAGTCTCAGCTACTAAATTTCTTCACACAGAACTGGCTCTCCTCGGCCCCTCCCC
TCGCCTTTCCTGCTCAATGCTCACCGCCTCCGGACCCCTCCCTCATCAGAAAGCCCAGGCTCCGCTCGTAGAAGTGCGCAGGCGTCACCGCGCATCCAG
GAGCCACGTGTCAGGAGTCACGTGTCAGGTGTCACGTGTCAGGCGTCACGTGGCTGGAGGCCGTTGGAGCGCCTGCGCAGCTTTTCCGCACGCGCCTCG

Fig. 4. Examples of human models constructed from the 4 ANT genes sequences. Part of the promoter sequences and the beginning of the first exon (highlighted in grey) are shown for each isoform. Short sequences are in white characters (highlighted in black) are matrix sequences, nucleotide strings or TSS (transcription start site). **ANT1 model**: GGCTCTAAA (OXBOX) / GATAA (V$GATA1.04) / TATAA (O$VTATA.01) / G (TSS) /ATG; **ANT2 model**: GGCTCTAAA (OXBOX) / GATAA (V$GATA1.04) / TATAA (O$VTATA.01) / G (TSS) / ATG; **ANT3 model**: GAGGG (V$CTCF) / CACGGG (V$CHRE) / ACCTA (V$RORA) / C (TSS); **ANT4 model**: GTCTGG (V$SMAD3.01) / GGGGA **(**V$MZF1.02) / CCCCTC (V$MAZ.01) / CACGTG (V$EBOX/HIFF) / G (TSS)/ ATG.

- ANT2 (SLC25A5): Most involved matrices and IUPAC strings from the ANT2 models were found involved in cell growth and proliferation: V$HOXF, Paralog HOX genes 1 to 8, control normal development, primary cellular processes and their deregulation are involved in carcinogenesis (39). V$MZF1.01 (myeloid zinc finger protein MZF1) is involved in cell proliferation and its abnormal expression results in cancer development (40). V$EGR1.02 (EGR1, early growth response 1) is involved in regulating homeostasis of hematopoietic stem cells (HSC) by coordinating proliferation and migration (41). V$SP1.01 and .03 (stimulating protein 1) are known to regulate lipogenesis and proliferation in cancer cells (42). The V$CEBP (ccaat/Enhancer Binding Proteins) family is known to play a role in cell growth and proliferation (43). V$MYT1 Myelin transcription factor 1 (Myt1) is a zinc finger DNA-binding protein that regulates cell proliferation and differentiation of oligodendrocyte lineage cells (44).

V$PARF (PAR/bZIP family), including the Drosophila PAR domain protein I gene, Pdp1 (regulator of larval growth, mitosis and endoreplication) plays a critical role in coordinating growth and DNA replication (45). V$GATA factors control the development of diverse tissues and they are involved in different mechanisms of carcinogenesis apart from their normal functions (46).

Table II. *ANT* gene models in mammals

	Motif 1	**Motif 2**	**Motif 3**	**Motif 4**	**Motif 5**	**Motif 6**
ANT1	V$MEF2 (TAWAAATA)	V$GATA* (GATAA)	V$NRF1 (GCGCabgcgc)	(TSS)	ATG	-
	V$GATA (GATAA)	V$NRF2.01 (c**GGAA**g)	CAAT	TATAA	(TSS)	ATG
	OXBOX (GGCTCTAAA)	V$GATA1.04* (GATAa)	CAAT	O$VTATA.01 (ta**TAAA**)	(TSS)	ATG
	V$MEF2* (TAWAAATA)	V$GREF* (GTGTTCT)	V$GATA* (GATAA)	V$ETSF* (CGGAAG)	(TSS)	-
ANT2	GRBOX (ATTGTT)	V$MZF1.01 (**GGGG**a)	V$EGR1.02 (gngg**GGGC**g)	V$SP1.03 (**GGGC**gg)	TATAAA	(TSS)
	GRBOX CATTGTT	V$MZF1.01 (**GGGG**a)	V$EGR1.02 (gngg**GGGC**g)	V$SP1.01 (**GGGC**ggg)	O$VTATA.01 (ta**TAAA**)	(TSS)
	V$HOXF* (TAATTA)	V$CEBP* (TTGTGMAA)	V$MYT1* (AAAGTTT)	V$PARF* (TTTGCAA)	V$GATA* (GATAA)	(TSS)
ANT3	V$CTCF* (CCCTC)	V$CHRE* (CACGng)	V$RXRF* (AAGTTCA)	V$RORA* (TAGGT**)**	(TSS)	
	V$CTCF* (CCCTC)	V$CHRE* (CACGng)	V$RXRF* (AAGTTCA)	V$RORA* (TAGGT**)**	ATG	(TSS)
ANT4	V$SMAD3.01* (**GTCT**gg)	V$MZF1.02* (**GGGG**)	V$EBOX/HIFF* (CACGTG)	V$EBOX/HIFF* (CACGTG)	(TSS)	-
	V$SMAD3.01* (**GTCT**gg)	V$MZF1.02* (**GGGG**)	V$MAZ.01* (**GAGG**gg)	V$EBOX/HIFF* (CACGTG)	V$EBOX/HIFF* CACGTG	(TSS)
	V$HMGA.01* (tn**AAT**)	V$ETS2.01* (c**AGGA**a)	V$EBOX/HIFF* (CACGTG)	V$EBOX/HIFF* (CACGTG)	(TSS)	-
	V$SORY* (AACAAT)	V$ETSF* (CGGAAG)	V$EBOX/HIFF* (CACGTG)	V$EBOX/HIFF* (CACGTG)	(TSS)	-

Matrix families (i.e. V$MEF2) or matrices (i.e. V$NRF2.01) with an asterisk were identified from phylogenetic analyses. Only nucleotides (IUPAC) with high information content are presented (the matrix exhibits a high conservation at this position). For matrices, nucleotides in bold capital letters denote the core sequence used by MatInspector (defined as the highest, usually four, conserved consecutive positions) (47).

- ***ANT3 (SLC25A6):*** Multiple major functions are proposed for the transcriptional factors found by phylogenetic analysis to be involved in the ANT3 gene regulation: V$CTCF (CTCF and BORIS gene family), V$CHRE (carbohydrate response elements), V$RXRF (RXR heterodimer binding sites) and V$RORA (v-ERB and RAR-related orphan receptor alpha). However, contrary to the three genes encoding the three other ANT isoforms, the stringency of the ANT3 models required to identify coregulated genes was too low to lead to conclusive results (i.e. the "Screening of co-regulated genes" part).

- ***ANT4 (SLC25A31):*** Most involved matrices and IUPAC strings from the ANT4 models are involved in testis development, spermatogenesis or the glycolytic energetic metabolism: V$SMAD3.01 (bound by the Smad3 transcription factor) determines androgen responsiveness and is involved in testis development (48). V$MZF1.02 (recognised by the myeloid zinc finger protein MZF1) and a specific nucleotide string CACGTGTCAGG, which is present in several copies (three in human) spaced 3 nucleotides apart, with the final string located 33 nt upstream of the transcription start nucleotide (TSS) (Figure 3). This final nucleotide sequence is shared by several matrices from the V$EBOX family, including V$USF.02 (an upstream stimulating factor), V$HIFF and V$HRE.01 (hypoxia response elements, binding sites for HIF1-alpha/ARNT heterodimers). Such matrices are involved in the glycolytic pathway and in the regulation of the hypoxia response (49), as it is suggested during spermatogenesis (50).

Screening of co-regulated genes with each ANT gene using ModelInspector (Genomatix), Search for phylogenetically conserved promoter models (Genomatix) and ***GeneProm***

Constructed models of each *ANT* gene promoter were screened as described in Figure 1 either on the full chromosomal human sequences or the human promoter database (results with an asterisk) using ModelInspector **(**Genomatix) and GeneProm. Only genes with a model located next to their transcription start sites were selected. Moreover, several locations (one third of the results) in clones without identified genes were not retained. Encoded protein data were obtained from the EnsEMBL database (25) and data on protein function were obtained from HGNC.

- Results with ANT1 models: genes revealed from screening with the ANT1 promoter models are listed in Table III. 28 genes were identified by ModelInspector / GeneProm analyses (supplemental Table 2). 11 of these genes have an unknown or supposed function. 13 from the remaining 17 genes are specifically expressed in muscle or directly involved and bioenergetic metabolism.

- Results with ANT2 models: 56 genes were identified by ModelInspector / GeneProm analyses (supplemental Table 3). 23 of these genes have an unknown or only supposed function and 17 have a proposed function which seems not directly linked to each others. 13 from the remaining 16 genes (Table IV) are expressed in conditions directly related to cell proliferation and glycolytic metabolism.

- Results with ANT3 models: the stringency required for the ANT3 models to identify coregulated genes was too low to lead to conclusive results. Among the few (10) identified with the ModelInspector and GenePrompt analyses, 5 genes have unknown function and the 5 remaining have heterogeneous functions not linked to bioenergetic pathways (supplemental Table 4).

Table III: ***Genes coregulated with the ANT1 gene***

Gene, ID	Encoded protein (Ensembl)	Protein function (HGNC)
ANO1 ENSG00000131620	Anoctamin-1	Calcium-activated chloride channel, higher levels **skeletal muscle**.
ARRDC3 ENSG00000113369	Arrestin domain-containing protein 3	Associated with plasma membrane, highly expressed in **skeletal muscle**
ATP5B* ENSG00000110955	ATP synthase, beta polypeptide	H+ transporting, **mitochondrial** F1 complex
ATP5D ENSG00000099624	ATP synthase subunit delta, mitochondrial	**Mitochondrial** membrane ATP synthase
ATP13A4* ENSG00000127249	Probable cation-transporting ATPase 13A4	ATP + H_2O = ADP + phosphate ; Expressed in heart, **skeletal muscles**
CBY1* ENSG00000100211	Protein chibby homolog 1	Expressed at higher levels in heart, **skeletal muscle**
COQ7* ENSG00000167186	Ubiquinone biosynthesis protein COQ7 homolog	Involved in ubiquinone biosynthesis. expressed in heart and **skeletal muscle**
COX6B2* ENSG00000160471	Cytochrome c oxidase subunit VIb isoform 2	Connects the two **COX** monomers into the physiological dimeric form
COX7B* ENSG00000131174	cytochrome c oxidase subunit VIIb	one of the nuclear-coded polypeptide chains of **cytochrome c oxidase**
MCAM* ENSG00000076706	Melanoma cell adhesion molecule	Appears to be limited to vascular smooth **muscle** in normal adult tissues
MYO1F* ENSG00000142347	Myosin IF	**Myosins** are actin-based motor molecules with ATPase activity
NDUFA9 ENSG00000139180	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 9	subunit of **mitochondrial** respiratory chain NADH dehydrogenase
NDUFS1* ENSG00000023228	NADH-ubiquinone oxidoreductase 75 kDa subunit, mitochondrial	subunit of **mitochondrial** respiratory chain NADH dehydrogenase
SLC35C2* ENSG00000080189	Solute carrier family 35 member C2	May play an important role in the cellular response to tissue hypoxia.
SOD3* ENSG00000109610	Superoxide dismutase 3, extracellular	Protect the extracellular space from toxic effect of reactive oxygen
TMC4* ENSG00000167608	Transmembrane channel-like protein 4	May function as ion channels, transporters or modulators of such
WASF1* ENSG00000112290	WAS protein family, member 1	Signal from tyrosine kinase receptors / smallGTPases to actin cytoskeleton

Several constructed models of the *ANT1* promoter region were screened as described in Figure 1 either on the full chromosomal human sequences or the human promoter library (results with an asterisk). Protein function linked to muscle bioenergetic metabolism are shown in bold characters. The full set of results is shown in an additional extra file.

Table IV. ***Genes coregulated with the ANT2 gene***

Gene, ID	Encoded protein (Ensembl)	Protein function (HGNC)
AURKC ENSG00000105146	Serine/threonine-protein kinase 13	Organizing microtubules during **mitosis** and chromosome segregation
BAZ1A* ENSG00000198604	Bromodomain adjacent to zinc finger domain protein 1A	Component of ACF complex, **ATP-dependent** chromatin remodelling complex
BTG1 ENSG00000133639	B-cell translocation gene 1 protein	Associated with the **early G1 phase** of the cell cycle
CDKN2AIP ENSG00000168564	CDKN2A-interacting protein	Activates **p53**/TP53 by CDKN2A-dependent and independent pathways
CKB ENSG00000166165	Creatine kinase B-type	Catalyzes the **transfer of phosphate** between ATP and various phosphogens
FGF5 ENSG00000138675	Fibroblast growth factor 5	Functions as an inhibitor of hair elongation
FGL1* ENSG00000104760	Fibrinogen-like protein 1	Hepatocyte **mitogenic** activity
GADD45B ENSG00000099860	Growth arrest and DNA damage-inducible protein	Involved in the regulation of **growth and apoptosis**. Activation of MTK1/MEKK4
GDF15 ENSG00000130513	Growth/differentiation factor 15	Transforming **growth factor** beta receptor signaling pathway
HIF1A ENSG00000100644	Hypoxia-inducible factor 1-alpha	Master transcriptional regulator of the **adaptive response to hypoxia**
MAF ENSG00000178573	Transcription factor Maf	Transcriptional activator or repressor in embryonic lens fiber cell development.
MORF4L1 ENSG00000185787	Mortality factor 4-like protein 1	NuA4 histone acetyltransferase complex, involved in transcriptional activation
NPPC ENSG00000163273	C-type natriuretic peptide	Regulation of chondrocytes **proliferation** and differentiation
PHIP* ENSG00000146247	PH-interacting protein	Stimulates **cell proliferation** through regulation of cyclin transcription
SLC2A3 ENSG00000059804	Solute carrier family, facilitated glucose transporter member 3	Facilitative **glucose transporter**. Probably a neuronal glucose transporter
SMCHD1* ENSG00000101596	Structural maintenance of chromosomes hinge protein	ATP binding

Several constructed models of the *ANT2* promoter region were screened as described in Figure1 either on the full chromosomal human sequences or the human promoter library (results with an asterisk). Function involved in cell proliferation is shown in bold characters. The full set of results is shown in the additional Table 2.

- Results with ANT4 models: 35 genes were identified by ModelInspector / GeneProm analyses (supplemental Table 5). 11 of these genes have an unknown or a supposed function. 17 from the remaining 24 genes are specifically expressed in the testis, during spermatogenesis, or involved in the prostate metabolism (Table V).

Table V. ***Genes coregulated with the ANT4 gene***

Gene, ID	Encoded protein (Ensembl)	Protein function (HGNC)
APEX1* ENSG00000100823	APEX nuclease (multifunctional DNA repair enzyme) 1	repair of apurinic / apyrimidinic sites **in testis**
BAX* ENSG00000087088	BCL2-associated X protein	Mutagenesis regulation in **spermatogenesis**
CDK4* ENSG00000135446	cyclin-dependent kinase 4	cell cycle G1 phase progression in **male reproduction**
***FNDC3A**§* ENSG00000102531	Fibronectin type-III domain-containing protein 3A	Mediates spermatid-Sertoli adhesion during **spermatogenesis**
G6PC2* ENSG00000152254	Glucose-6-phosphatase 2	Glucose production through glycogenolysis and gluconeogenesis **in testis**
***G6PC3**§* ENSG00000141349	Glucose-6-phosphatase 3	Hydrolyzes glucose-6-phosphate to glucose in **testis** endoplasmic reticulum
KAT5* ENSG00000172977	K(lysine) acetyltransferase 5	chromatin remodelling with an abundant **spermatid** protein
***KLHL12**§* ENSG00000117153	Kelch-like protein 12	Adapter for the ubiquitin-protein E3 ligase complex, highly expressed **in testis**
LAMP1* ENSG00000185896	lysosomal-associated membrane protein 1	binds amelogenin, differentially expressed in **spermiogenesis**
RPUSD4* ENSG00000165526	RNA pseudouridylate synthase domain-containing protein 4	Unknown, **expressed in prostate**
SLC2A4 ENSG00000181856	solute carrier family 2, member 4 (GLUT4)	facilitated glucose transporter, **detected in human testis**
SOHLH1* ENSG00000165643	**spermatogenesis** and oogenesis specific basic helix-loop-helix 1	germ cell-specific, oogenesis regulator and **male germ cells**
SUN1 ENSG00000164828	chr. 7 unc-84 homolog A	nuclear anchorage/migration, expression of **meiotic reproductive genes**
TDRD1* ENSG00000095627	tudor domain containing 1	essential for **spermiogenesis**
TKTL1 ENSG00000007350	transketolase-like 1	important role in transketolase activity, **testis** expressed
TMEM184A ENSG00000164855	transmembrane protein 184A = Sdmg1	**male-specific** expression in embryonic gonads
UBE2B* ENSG00000119048	ubiquitin-conjugating enzyme E2B (RAD6 homolog)	post-replicative DNA damage repair **in spermatogenesis**

Several constructed models of the *ANT4* promoter region were screened as described in Figure 1 either on the full chromosomal human sequences or the human promoter library (results with an asterisk). Function involved in spermatogenesis or in testis or prostate metabolism is shown in bold characters. [§]Genes not identified in the previous ModelInspector / GeneProm analysis (51). The full set of results is shown in an additional Table 4.

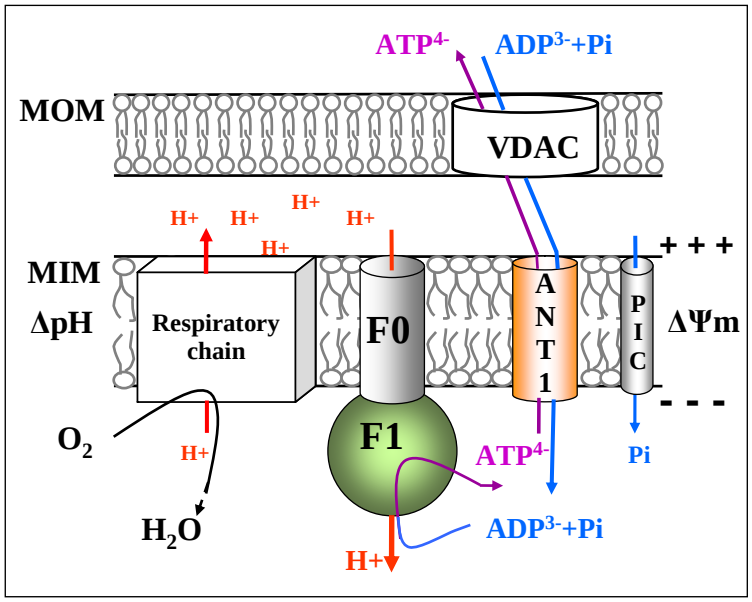
Discussion

We built a pipeline of bioinformatic analyses allowing the study of the transcriptional regulation of a set of genes and the prediction of co-regulated genes with each of them. Co-regulated genes could encode proteins involved in the same metabolic network including a whole set of different pathways. This analysis was designed for, in the longer term, predicting a precise signature of a cellular metabolic change, which is the consequence of physiological conditions, disease, or a response to a specific pharmacological or nutritional treatment. This pipeline allows the analysis of the structure of the promoter region of a gene and the construction of regulatory models composed by a combination of several small nucleotide sequences specifically linked to the gene function. The strategy of the promotology analysis that we adopted relies on the crossing of three complementary analyses: 1 – screening of genes located next to the constructed models on the full human genome (combination of the ModelInspector tools and GeneProm software) (24); (51); 2 – screening of these models on a database of human promoter sequences by ModelInspector / Human Promoters (24); 3 - screening of selected models in several mammalian species (Search for phylogenetically conserved promoter models), (24). The crossing of these three analyses allowed to identify, with higher stringency, a limited set of genes controlled by the same model of promoter sequence.

This bioinformatic protocol was tested on a set of genes encoding four isoforms of the ANT protein (adenine nucleotide translocator), each of them having a specific role in a specific cellular type. Three of these four proteins (ANT1, ANT2 and ANT4) are controlled at the transcriptional level by a specific mechanism. The fourth (ANT3) is the ubiquitous isoform constitutively expressed in all cells. The implementation of our promotology analysis on this set of four *ANT* genes accounted for a powerful validation of our strategy:

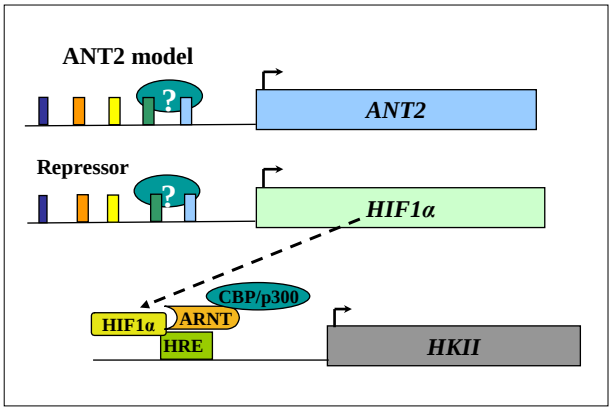
ANT1: the gene encoding the isoform specifically expressed in muscle tissues (13) enabled us to build four models of promoter sequence (Table II). These models are found in promoters of 17 genes highly expressed in muscle tissues and / or involved in the bioenergetic metabolism (Table IV). Among these 17 identified genes, 13 have a direct connection with the muscle cell metabolism or the mitochondrial ATP synthesis. In particular, 6 of these genes encode proteins included to 3 of the mitochondrial complexes of oxidative phosphorylation: NADH dehydrogenase (complex I), cytochrome oxidase (complex IV) and ATP synthase (complex V) (52) (Figure 5). Moreover, another gene carrying a model, COQ7, is involved in the synthesis of these complexes (53). Other identified genes encode proteins involved in major pathways of the muscular metabolism such as ANO1/TMEM16A in the calcium transport (54), MYO1F in the muscle contraction (55), SOD3 (56) and SLC35C2 (57) in the cell response to ROS or to the oxygen pressure.

Fig. 5. Schematic representation of the role of ANT1 isoform in muscle cell oxidative phosphorylation. ADP^{3-} and inorganic phosphate (Pi) are transported across the mitochondrial inner membrane (MIM) into the mitochondrial matrix by the mitochondrial ANT and phosphate carrier (PiC), respectively. F1F0-ATPase combines Pi and ADP to form ATP, which is then exchanged for ADP across the MIM by ANT1 across MOM (mitochondrial outer membrane). The whole reaction is driven by a proton gradient maintained mainly by the respiratory chain. Six of the genes identified from our promotology analysis encode proteins included in the oxidative phosphorylation (respiratory chain and F0-F1 ATP synthase proteins).



ANT2: Most of the 16 genes carrying a model resulting from the ANT2 gene encode proteins intervening in pathways related to the cell division and proliferation (AURKC, BTG1, FGL1, GDF15, NPPC, PHIP) (Table V). Several other identified genes encode signalling proteins such as CDKNÀIP, GDD45B or HIF1-alpha. This last HIF1-alpha protein is known to induce the transcription of the HKII gene under metabolism glycolytic conditions (58) (59). As ANT2, HKII is involved in the uptake of the glycolytic ATP through the mitochondrial internal membrane. This indirect of HKII / ANT2 co-regulation is coherent with their complementary roles in glycolytic conditions (Figure 6) (4,6,60).

Fig. 6. Schematic representation of the proposed indirect HKII / ANT2 co-regulation in glycolytic conditions. In glycolytic conditions, the mitochondrial hexokinase isoform, HK II, generates ATP from cytoplasmic glucose 6-P (G-6P). The ATP^{4-} is then imported into mitochondria by the ANT2 isoform, contributing to the maintenance of the mitochondrial membrane potential (ΔΨm). The HKII gene transcription is induced by the HIF1-alpha protein.



ANT3: No gene carrying models built from the ANT3 gene promoter could be identified if stringency parameters similar to that used for three other isoforms were selected. Analyses with lower stringencies (lower number of matrices in models and lower similarity of matrices or IUPAC strings as compared to the three other ANT genes) reveal up to 10 genes. Half encode proteins with an unknown function; the five others have no apparent functional link between them (additional Table 4). Moreover, the analysis of the *ANT3* promoter region by the PromoterInspector tool (24) proposed few regulation matrices as compared with genes of the three other regulated isoforms (result not shown).

ANT4: The models constructed from the ANT4 gene promoter allowed to identify 17 genes with known function and all have a role in spermatogenenesis (Table IV). Our previous work on the promoter of this ANT4 gene had already led to identify a part of these co-regulated genes (51). The new version of our GeneProm software enabled us to identify five new genes also directly related to spermatogenesis. Among these five genes specifically expressed in testicles (61), two encode the glucose-6-phosphatases 2 and 3 enzymes, producing glucose from glucose-6-phosphate, leading to ATP synthesis (52). This function is totally in agreement with the exclusively glycolytic metabolism of spermatozoids (21). Moreover, this glycolytic ATP production is also consistent with the expression and the specific role of the ANT4 isoform in spermatozoid bioenergetics: part of ATP produced by glucose-6-phosphatases 2 and 3 could be imported into mitochondria by ANT4 (gene located on chromosome 4) to compensate for the absence of the ANT2 isoform (gene located in a area of chromosome X not transcribed during spermatogenesis) (Figure 7) (51).

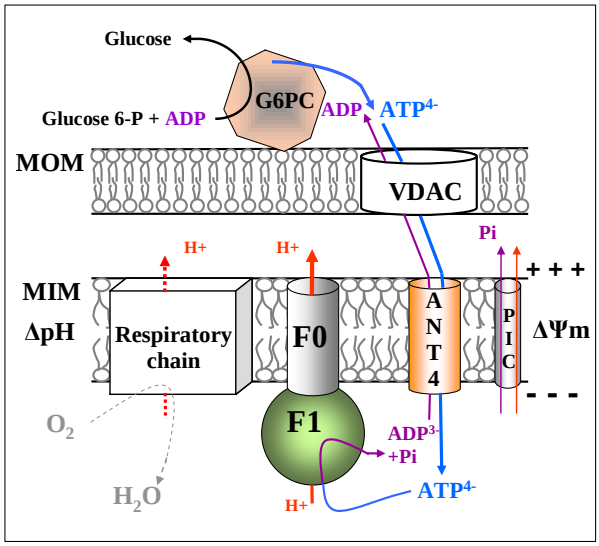


Fig. 7. Schematic representation of the role of ANT4 isoform in bioenergetics during spermatogenenesis. The glucose-6-phosphatases 2 and 3 (G6PC) generates ATP from glucose 6-P (G-6P) produced by the cytoplasmic hexokinase, HK. The ATP^{4-} is imported into mitochondria across the MOM (mitochondrial outer membrane) through the voltage-dependent anion channel (VDAC), and then across the MIM (the mitochondrial inner membrane) by the ANT4 isoform. ATP^{4-} contributes to the maintenance of the mitochondrial membrane potential ($\Delta\Psi m$) in spermatozoid mitochondria. The hydrolysis of imported glycolytic ATP^{4-} by the F1 component of the ATP synthase leads to 1 - the release of ADP^{3-} in mitochondria with the gain of a negative charge on the matrix side; and 2 - the ejection of a proton into the intermembrane space through the F0 component.

Thus, this *in silico* analysis allows to lead to very interesting conclusions on the relation between transcriptional regulatory pathways and protein function in a cellular metabolic network. A set of genes encoding protein isoforms expressed with tissue specificity turned out to be a clear validation of our strategy. The phylogenetic comparison of promoters allowed the identification of nucleotidic matrices with major roles in gene regulation. However, the exclusive use of regulatory matrices has limits: it requires multiple analyses with different parameters of stringency and layout of each matrix compared with the others. The presence, in a promoter region, of described and validated nucleotide sequences, such as OXBOX (5) and GRBOX (6), known to intervene in the regulation of the *ANT1* and *ANT2* genes, respectively, is a crucial argument in the construction of a powerful regulatory model.

In our study, the presence of a gene encoding a ubiquitous isoform, ANT3, and the failing to identify co-regulated genes with models built from its promoter region allows to validate our strategy. Moreover, this analysis of an unregulated ubiquitous gene becomes a standard for further works by providing a precise scale of similarity of matrices in models. Thus, a strategy of promotology, based simultaneously on a conclusive phylogenetic analysis and on already validated regulatory nucleotide sequences, allows a powerful identification of co-regulated genes.

Our work on this set of isoforms also showed that transcriptional regulation is a major mechanism of cellular specificity. The structure of the promoter sequence directly upstream of the transcription start site itself allows the identification of co-regulated genes. This suggests that, at least for the regulation of bioenergetic pathways described in this work, other supposed regulatory mechanisms including microRNA or the messenger RNA stability, would intervene in other cellular functions. Moreover, our strategy allows overcoming the insufficiencies of other techniques used for the study of the gene expression: taking into account their very close coding sequence of the four ANT isoforms, no currently commercial microarrays, able to quantify simultaneously and specifically the four transcripts, are available. In addition, the very hydrophobic properties of these proteins do not allow their identification by 2D electrophoresis and no specific antibody of each isoform is available.

In conclusion, our computational strategy on a set of four isoforms, known for their specific functions in cell bioenergetics, enabled us to develop a powerful analysis of gene promoter sequences. Our analyses enabled to identify a whole set of co-regulated genes, involved in the same cellular function. This study validates the exclusive role of the proximal promoter region in tissue specificity and should bring, with transcriptomics and metabolomics, a precious help in development of cellular metabolic networks and the study of their regulatory pathways.

References

1. Chopra, V.S. (2011) Chromosomal organization at the level of gene complexes. *Cell Mol Life Sci*, **68**, 977-990.
2. Gaffney, D.J., Blekhman, R. and Majewski, J. (2008) Selective constraints in experimentally defined primate regulatory regions. *PLoS Genet*, **4**, e1000157.
3. Kim, N.K., Tharakaraman, K., Marino-Ramirez, L. and Spouge, J.L. (2008) Finding sequence motifs with Bayesian models incorporating positional information: an application to transcription factor binding sites. *BMC Bioinformatics*, **9**, 262.
4. Chevrollier, A., Loiseau, D., Reynier, P. and Stepien, G. (2010) Adenine nucleotide translocase 2 is a key mitochondrial protein in cancer metabolism. *Biochim Biophys Acta*, **1807**, 562-567.
5. Li, K., Hodge, J.A. and Wallace, D.C. (1990) OXBOX, a positive transcriptional element of the heart-skeletal muscle ADP/ATP translocator gene. *J Biol Chem*, **265**, 20585-20588.
6. Giraud, S., Bonod-Bidaud, C., Wesolowski-Louvel, M. and Stepien, G. (1998) Expression of human ANT2 gene in highly proliferative cells: GRBOX, a new transcriptional element, is involved in the regulation of glycolytic ATP import into mitochondria. *J Mol Biol*, **281**, 409-418.
7. Ku, D.H., Kagan, J., Chen, S.T., Chang, C.D., Baserga, R. and Wurzel, J. (1990) The human fibroblast adenine nucleotide translocator gene. Molecular cloning and sequence. *J Biol Chem*, **265**, 16060-16063.
8. Luciakova, K., Barath, P., Poliakova, D., Persson, A. and Nelson, B.D. (2003) Repression of the human adenine nucleotide translocase-2 gene in growth-arrested human diploid cells: the role of nuclear factor-1. *J Biol Chem*, **278**, 30624-30633.
9. Barath, P., Albert-Fournier, B., Luciakova, K. and Nelson, B.D. (1999) Characterization of a silencer element and purification of a silencer protein that negatively regulates the human adenine nucleotide translocator 2 promoter. *J Biol Chem*, **274**, 3378-3384.
10. Kehoe, S.M., Oka, M., Hankowski, K.E., Reichert, N., Garcia, S., McCarrey, J.R., Gaubatz, S. and Terada, N. (2008) A conserved E2F6-binding element in murine meiosis-specific gene promoters. *Biol Reprod*, **79**, 921-930.
11. HGNC. HUGO Gene Nomenclature Committee: http://www.genenames.org/.
12. Palmieri, F. (2004) The mitochondrial transporter family (SLC25): physiological and pathological implications. *Pflugers Arch*, **447**, 689-709.
13. Stepien, G., Torroni, A., Chung, A.B., Hodge, J.A. and Wallace, D.C. (1992) Differential expression of adenine nucleotide translocator isoforms in mammalian tissues and during muscle cell differentiation. *J Biol Chem*, **267**, 14592-14597.
14. Barath, P., Luciakova, K., Hodny, Z., Li, R. and Nelson, B.D. (1999) The growth-dependent expression of the adenine nucleotide translocase-2 (ANT2) gene is regulated at the level of transcription and is a marker of cell proliferation. *Exp Cell Res*, **248**, 583-588.
15. Chevrollier, A., Loiseau, D., Chabi, B., Renier, G., Douay, O., Malthiery, Y. and Stepien, G. (2005) ANT2 isoform required for cancer cell glycolysis. *J Bioenerg Biomembr*, **37**, 307-316.
16. Sabova, L., Zeman, I., Supek, F. and Kolarov, J. (1993) Transcriptional control of AAC3 gene encoding mitochondrial ADP/ATP translocator in Saccharomyces cerevisiae by oxygen, heme and ROX1 factor. *Eur J Biochem*, **213**, 547-553.
17. Lowry, C.V. and Zitomer, R.S. (1988) ROX1 encodes a heme-induced repression factor regulating ANB1 and CYC7 of Saccharomyces cerevisiae. *Mol Cell Biol*, **8**, 4651-4658.
18. Zitomer, R.S. and Lowry, C.V. (1992) Regulation of gene expression by oxygen in Saccharomyces cerevisiae. *Microbiol Rev*, **56**, 1-11.
19. Zitomer, R.S., Sellers, J.W., McCarter, D.W., Hastings, G.A., Wick, P. and Lowry, C.V. (1987) Elements involved in oxygen regulation of the Saccharomyces cerevisiae CYC7 gene. *Mol Cell Biol*, **7**, 2212-2220.

20. Dolce, V., Scarcia, P., Iacopetta, D. and Palmieri, F. (2005) A fourth ADP/ATP carrier isoform in man: identification, bacterial expression, functional characterization and tissue distribution. *FEBS Lett*, **579**, 633-637.
21. Brower, J.V., Rodic, N., Seki, T., Jorgensen, M., Fliess, N., Yachnis, A.T., McCarrey, J.R., Oh, S.P. and Terada, N. (2007) Evolutionarily conserved mammalian adenine nucleotide translocase 4 is essential for spermatogenesis. *J Biol Chem*, **282**, 29658-29666.
22. Kim, Y.H., Haidl, G., Schaefer, M., Egner, U., Mandal, A. and Herr, J.C. (2007) Compartmentalization of a unique ADP/ATP carrier protein SFEC (Sperm Flagellar Energy Carrier, AAC4) with glycolytic enzymes in the fibrous sheath of the human sperm flagellar principal piece. *Dev Biol*, **302**, 463-476.
23. Tanay, A., Gat-Viks, I. and Shamir, R. (2004) A global view of the selection forces in the evolution of yeast cis-regulation. *Genome Res*, **14**, 829-834.
24. Genomatix: Personalized Medicine - Relevance for scientists September 2010 [http://genomatix.de]
25. EnsEMBL: Ensembl genome Browser EBI September 2010 [http://www.ensembl.org]
26. Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, **4**, 406-425.
27. Zharkikh, A. and Li, W.H. (1995) Estimation of confidence in phylogeny: the complete-and-partial bootstrap technique. *Mol Phylogenet Evol*, **4**, 44-63.
28. Lindsay, B.G. (1988) Composite Likelihood method. *Contemporary Mathematics* **80**, 221-239.
29. Kumar, S., Tamura, K. and Nei, M. (2004) MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform*, **5**, 150-163.
30. Frech, K., Danescu-Mayer, J. and Werner, T. (1997) A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. *J Mol Biol*, **270**, 674-687.
31. KEGG. Kyoto Encyclopedia of Genes and Genomes Kanehisa Laboratories. September 2010 [http://www.genome.jp/keg
32. Feng, H., Cheng, T., Steer, J.H., Joyce, D.A., Pavlos, N.J., Leong, C., Kular, J., Liu, J., Feng, X., Zheng, M.H. *et al.* (2009) Myocyte enhancer factor 2 and microphthalmia-associated transcription factor cooperate with NFATc1 to transactivate the V-ATPase d2 promoter during RANKL-induced osteoclastogenesis. *J Biol Chem*, **284**, 14667-14676.
33. Rojas, A., Kong, S.W., Agarwal, P., Gilliss, B., Pu, W.T. and Black, B.L. (2008) GATA4 is a direct transcriptional activator of cyclin D2 and Cdk4 and is required for cardiomyocyte proliferation in anterior heart field-derived myocardium. *Mol Cell Biol*, **28**, 5420-5431.
34. Irrcher, I., Ljubicic, V., Kirwan, A.F. and Hood, D.A. (2008) AMP-activated protein kinase-regulated activation of the PGC-1alpha promoter in skeletal muscle cells. *PLoS One*, **3**, e3614.
35. Ramachandran, B., Yu, G. and Gulick, T. (2008) Nuclear respiratory factor 1 controls myocyte enhancer factor 2A transcription to provide a mechanism for coordinate expression of respiratory chain subunits. *J Biol Chem*, **283**, 11935-11946.
36. Bonello, M.R., Bobryshev, Y.V. and Khachigian, L.M. (2005) Peroxide-inducible Ets-1 mediates platelet-derived growth factor receptor-alpha gene transcription in vascular smooth muscle cells. *Am J Pathol*, **167**, 1149-1159.
37. Swamynathan, S.K. (2010) Kruppel-like factors: three fingers in control. *Hum Genomics*, **4**, 263-270.
38. Trenerry, M.K., Della Gatta, P.A. and Cameron-Smith, D. (2011) JAK/STAT signaling and human in vitro myogenesis. *BMC Physiol*, **11**, 6.
39. Cillo, C., Schiavo, G., Cantile, M., Bihl, M.P., Sorrentino, P., Carafa, V., M, D.A., Roncalli, M., Sansano, S., Vecchione, R. *et al.* (2011) The HOX gene network in hepatocellular carcinoma. *Int J Cancer*, **129**, 2577-2587.

40. Mudduluru, G., Vajkoczy, P. and Allgayer, H. (2010) Myeloid zinc finger 1 induces migration, invasion, and in vivo metastasis through Axl gene expression in solid cancer. *Mol Cancer Res*, **8**, 159-169.
41. DeLigio, J.T. and Zorio, D.A. (2009) Early growth response 1 (EGR1): a gene with as many names as biological functions. *Cancer Biol Ther*, **8**, 1889-1892.
42. Lu, S. and Archer, M.C. (2010) Sp1 coordinately regulates de novo lipogenesis and proliferation in cancer cells. *Int J Cancer*, **126**, 416-425.
43. Lu, G.D., Leung, C.H., Yan, B., Tan, C.M., Low, S.Y., Aung, M.O., Salto-Tellez, M., Lim, S.G. and Hooi, S.C. (2010) C/EBPalpha is up-regulated in a subset of hepatocellular carcinomas and plays a role in cell growth and proliferation. *Gastroenterology*, **139**, 632-643, 643 e631-634.
44. Nielsen, J.A., Berndt, J.A., Hudson, L.D. and Armstrong, R.C. (2004) Myelin transcription factor 1 (Myt1) modulates the proliferation and differentiation of oligodendrocyte lineage cells. *Mol Cell Neurosci*, **25**, 111-123.
45. Reddy, K.L., Rovani, M.K., Wohlwill, A., Katzen, A. and Storti, R.V. (2006) The Drosophila Par domain protein I gene, Pdp1, is a regulator of larval growth, mitosis and endoreplication. *Dev Biol*, **289**, 100-114.
46. Zheng, R. and Blobel, G.A. (2010) GATA Transcription Factors and Cancer. *Genes Cancer*, **1**, 1178-1188.
47. Quandt, K., Frech, K., Karas, H., Wingender, E. and Werner, T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res*, **23**, 4878-4884.
48. Itman, C., Wong, C., Hunyadi, B., Ernst, M., Jans, D.A. and Loveland, K.L. (2011) Smad3 dosage determines androgen responsiveness and sets the pace of postnatal testis development. *Endocrinology*, **152**, 2076-2089.
49. Heikkila, M., Pasanen, A., Kivirikko, K.I. and Myllyharju, J. (2011) Roles of the human hypoxia-inducible factor (HIF)-3alpha variants in the hypoxia response. *Cell Mol Life Sci*.
50. Marti, H.H., Katschinski, D.M., Wagner, K.F., Schaffer, L., Stier, B. and Wenger, R.H. (2002) Isoform-specific expression of hypoxia-inducible factor-1alpha during the late stages of mouse spermiogenesis. *Mol Endocrinol*, **16**, 234-243.
51. Dupont, P.Y. and Stepien, G. (2011) Computational analysis of the transcriptional regulation of the adenine nucleotide translocator isoform 4 gene and its role in spermatozoid glycolytic metabolism. *Gene*, **487**, 38-45.
52. Lenaz, G. and Genova, M.L. (2009) Structural and functional organization of the mitochondrial respiratory chain: a dynamic super-assembly. *Int J Biochem Cell Biol*, **41**, 1750-1772.
53. Levavasseur, F., Miyadera, H., Sirois, J., Tremblay, M.L., Kita, K., Shoubridge, E. and Hekimi, S. (2001) Ubiquinone is necessary for mouse embryonic development but is not essential for mitochondrial respiration. *J Biol Chem*, **276**, 46160-46164.
54. Davis, A.J., Forrest, A.S., Jepps, T.A., Valencik, M.L., Wiwchar, M., Singer, C.A., Sones, W.R., Greenwood, I.A. and Leblanc, N. (2010) Expression profile and protein translation of TMEM16A in murine smooth muscle. *Am J Physiol Cell Physiol*, **299**, C948-959.
55. Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N. *et al.* (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173-1178.
56. van Deel, E.D., Lu, Z., Xu, X., Zhu, G., Hu, X., Oury, T.D., Bache, R.J., Duncker, D.J. and Chen, Y. (2008) Extracellular superoxide dismutase protects the heart against oxidative stress and hypertrophy after myocardial infarction. *Free Radic Biol Med*, **44**, 1305-1313.
57. Leach, R.E., Duniec-Dmuchowski, Z.M., Pesole, G., Tanaka, T.S., Ko, M.S., Armant, D.R. and Krawetz, S.A. (2002) Identification, molecular characterization, and tissue expression of OVCOV1. *Mamm Genome*, **13**, 619-624.

58. Mathupala, S.P., Rempel, A. and Pedersen, P.L. (2001) Glucose catabolism in cancer cells: identification and characterization of a marked activation response of the type II hexokinase gene to hypoxic conditions. *J Biol Chem*, **276**, 43407-43412.
59. Pedersen, P.L., Mathupala, S., Rempel, A., Geschwind, J.F. and Ko, Y.H. (2002) Mitochondrial bound type II hexokinase: a key player in the growth and survival of many cancers and an ideal prospect for therapeutic intervention. *Biochim Biophys Acta*, **1555**, 14-20.
60. Chevrollier, A., Loiseau, D. and Stepien, G. (2005) [What is the specific role of ANT2 in cancer cells?]. *Med Sci (Paris)*, **21**, 156-161.
61. Burchell, A., Watkins, S.L. and Hume, R. (1996) Human fetal testis endoplasmic reticulum glucose-6-phosphatase enzyme protein. *Biol Reprod*, **55**, 298-303.

Acknowledgements

Thanks are due to the Soluscience company for helpful expert assistance in informatics. This work was supported by the Cancéropôle Lyon Auvergne Rhône-Alpes (CLARA), "Nutrition Métabolisme et Cancer" ProCan axis, to G. S; the Conseil Régional Auvergne to GS (LifeGrid funds) and P.-Y. D ("Innovation Région" funds) and FEDER (Fonds Européen de Développement Régional) to P.-Y. D.

Supplemental Tables

Supplemental Table 1. *ANT* gene sequences in mammals. Mammalian *ANT* gene sequences selected for the 4 ANT isofoms in 24 mammals including Human. Sequences extracted from EnsEMBL data base {EnsEMBL, #46} Sequences in bold are mammalian sequences that did not contain undetermined bases in the promoter and were not too divergent from the human sequence..

Species **(common name)**	**Selected sequences for genes of ANT isoforms**
Homo sapiens (Human)	***ANT1 (ENSG00000151729); ANT2 (ENSG00000005022); ANT3 (ENSG00000169100); ANT4 (ENSG00000151475)***
Bos Taurus (ox)	***ANT1 (ENSBTAT00000017580);*** *ANT2 (ENSBTAG00000046037);* ***ANT4 (ENSBTAG00000012826)***
Canis familiaris (dog)	***ANT1 (ENSCAFG00000007596); ANT2 (ENSCAFG00000018384);*** *ANT3 (ENSCAFG00000010987);* ***ANT4 (ENSCAFG00000003924)***
Cavia porcellus (guinea pig)	***ANT1 (ENSCPOG00000005275)****; ANT2 (ENSCPOG00000009202)*
Echinops telfairi (hedgehog)	*ANT1 (ENSEEUG00000001021)*
Felis catus (cat)	*ANT1 (ENSFCAG00000007057); ANT2 (ENSFCAG00000005481); ANT3 (ENSFCAG00000001211)*
Gorilla gorilla (gorilla)	***ANT2 (ENSGGOG00000014279)***
Loxodonta africana (elephant)	***ANT3 (ENSLAFG00000001584)***
Macaca mulatta (macaque)	***ANT2 (ENSMMUG00000022663);*** *ANT3 (ENSMMUG00000006899);* ***ANT4 (ENSMMUG00000015243)***
Mus musculus (mouse)	***ANT2 (ENSMUSG00000016319) ; ANT4 (ENSMUSG00000069041)***
Myotis lucifugus (microbat)	***ANT1 (ENSMLUG00000003712)***
Nomascus leucogenys (gibbon)	*ANT1 (ENSNLEG00000011142)*
Otolemur garnettii (galago)	*ANT4 (ENSOGAG00000005752)*
Pan troglodytes (chimpanzee)	*ANT3 (ENSPTRG00000029304);* ***ANT4 (ENSPTRG00000016432)***
Pongo pygmaeus (orangutan)	*ANT2 (ENSPPYG00000020669); ANT3 (ENSPPYG00000019151)*
Procavia capensis (hyrax)	*ANT3 (ENSPCAG00000014264)*
Pteropus vamyrus (megabat)	***ANT1 (ENSPVAG00000015279); ANT2 (ENSPVAG00000014924) ANT3 (ENSPVAG00000004356)****; ANT4 (ENSPVAG00000002926)*
Rattus Norvegicus (rat)	***ANT2 (ENSRNOG00000039980)***
Sus scrofa (pig)	***ANT1 (ENSSSCG00000015790)***
Tarsius syrichta (tarsier)	*ANT1 (ENSTSYG00000004233);* ***ANT2 (ENSTSYG00000002246)***
Tupaia belangeri (tree shrew)	*ANT1 (ENSTBEG00000011502)*
Tursiops truncatus (dolphin)	***ANT2 (ENSTTRG00000001007); ANT3 (ENSTTRG00000015399); ANT4 (ENSTTRG00000012978)***
Vicugna paosc (alpaca)	***ANT1 (ENSVPAG00000008737);*** *ANT2 (ENSVPAG00000000644)*

Supplemental Table 2: Genes coregulated with the *ANT1* gene. The full set of results obtained from the analysis with all constructed models of the *ANT1* promoter regions were screened as described in Figure 1 either on the full chromosomal human sequences or the human promoter library (results with an asterisk). Protein function linked to muscle bioenergetic metabolism is shown in bold characters.

Gene, ID	Encoded protein (Ensembl)	Protein function (Uniprot)
ANO1 ENSG00000131620	Anoctamin-1	calcium-activated chloride channel, higher levels in liver and **skeletal muscle**.
ARRDC3 ENSG00000113369	Arrestin domain-containing protein 3	Associated with plasma membrane, highly expressed in **skeletal muscle**
ATP5B ENSG00000110955*	ATP synthase, beta polypeptide	H+ transporting, **mitochondrial** F1 complex
ATP5D ENSG00000099624	ATP synthase subunit delta, mitochondrial	**Mitochondrial** membrane ATP synthase
ATP9A ENSG00000054793*	Probable phospholipid-transporting ATPase IIA	ATP + H_2O + phospholipid(In) = ADP + phosphate + phospholipid(Out)
ATP13A4 ENSG00000127249*	Probable cation-transporting ATPase 13A4	ATP + H_2O = ADP + phosphate ; Expressed in heart, placenta, liver, **skeletal muscles**
CBY1 ENSG00000100211*	Protein chibby homolog 1	Expressed at higher levels in heart, **skeletal muscle**
COQ7 ENSG00000167186*	Ubiquinone biosynthesis protein COQ7 homolog	Involved in ubiquinone biosynthesis. expressed in heart and **skeletal muscle**
COX6B2 ENSG00000160471*	Cytochrome c oxidase subunit VIb isoform 2	Connects the two COX monomers into the physiological dimeric form
COX7B ENSG00000131174*	cytochrome c oxidase subunit VIIb	one of the nuclear-coded polypeptide chains of **cytochrome c oxidase**
FGF12 ENSG00000114279*	Fibroblast growth factor 12	Probably involved in nervous system development and function
FN1 ENSG00000115414*	Fibronectin 1	Fibronectins are involved in cell adhesion, **cell motility**, opsonization
HK3 ENSG00000160883*	Hexokinase 3	Carbohydrate metabolism; hexose metabolism.
LDHAL6B ENSG00000171989	L-lactate dehydrogenase A-like 6B	(S)-lactate + NAD^+ = pyruvate + NADH Higher expression level in adult testis
MCAM ENSG00000076706*	Melanoma cell adhesion molecule	Appears to be limited to vascular smooth **muscle** in normal adult tissues
MPP1 ENSG00000130830	Membrane protein, palmitoylated 1	Regulates neutrophil polarization by regulating AKT1 phosphorylation
MYO1F ENSG00000142347*	Myosin IF	**Myosins** are actin-based motor molecules with ATPase activity
NDUFA9 ENSG00000139180	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 9	subunit of the **mitochondrial** membrane respiratory chain NADH dehydrogenase
NDUFS1 ENSG00000023228*	NADH-ubiquinone oxidoreductase 75 kDa subunit, mitochondrial	subunit of the **mitochondrial** membrane respiratory chain NADH dehydrogenase
RNF121 ENSG00000137522*	RING finger protein 121	Membrane; Multi-pass membrane protein
SERPINA3 ENSG00000196136	Serpin peptidase inhibitor clade A, member 3 (antiproteinase, antitrypsin)	Could inhibit neutrophil cathepsin G and mast cell chymase
SH2D7 ENSG00000183476	SH2 domain-containing protein 7	unknown
SIDT1 ENSG00000072858*	SID1 transmembrane family, member 1	unknown

*SLC9A5** *ENSG00000135740*	Sodium/hydrogen exchanger 5	Involved in pH regulation to eliminate acids generated by **active metabolism**
*SOD3** *ENSG00000109610*	Superoxide dismutase 3, extracellular	Protect the extracellular space from toxic effect of reactive oxygen intermediates
*TMC4** *ENSG00000167608*	Transmembrane channel-like protein 4	May function as ion channels, transporters or modulators of such
*WASF1** *ENSG00000112290*	WAS protein family, member 1	Signal transmission from tyrosine kinase re ceptors/smallGTPases to actin cytoskeleton
*WDR64** *ENSG00000162843*	WD repeat domain 64	unknown

Supplemental Table 3: Genes coregulated with the *ANT2* gene. The full set of results obtained from the analysis with all constructed models of the *ANT2* promoter regions were screened as described in Figure 1 either on the full chromosomal human sequences or the human promoter library (results with an asterisk). Protein function linked to cell proliferation and glycolytic metabolism is shown in bold characters.

Gene, ID	**Encoded protein (Ensembl)**	**Protein function (Uniprot)**
AURKC *ENSG00000105146*	Serine/threonine-protein kinase 13	Organizing microtubules during **mitosis** and chromosome segregation
AQP2 *ENSG00000167580*	Aquaporin-2	Forms a water-specific channel that provides the renal plasma membranes
BAT2 *ENSG00000231370*	Protein PRRC2A	May play a role in the regulation of pre-mRNA splicing
*BAZ1A** *ENSG00000198604*	Bromodomain adjacent to zinc finger domain protein 1A	Component of the ACF complex, an **ATP-dependent** chromatin remodeling complex
BTG1 *ENSG00000133639*	B-cell translocation gene 1 protein	Anti-proliferative protein, associated with the **early G1 phase** of the cell cycle
CCNI2 *ENSG00000205089*	Cyclin-I2	Regulation of cyclin-dependent protein kinase activity
CD177 *ENSG00000204936*	CD177 antigen	Highly expressed in normal bone marrow and in granulocytes of patients with polycythemia vera
CDKN2AIP *ENSG00000168564*	CDKN2A-interacting protein	Activates **p53**/TP53 by CDKN2A-dependent and independent pathways
CEBPB *ENSG00000172216*	CCAAT/enhancer-binding protein beta	Transcriptional activator of genes involved in immune and inflammatory responses
CKB *ENSG00000166165*	Creatine kinase B-type	Catalyzes the **transfer of phosphate** between ATP and various phosphogens
COL1A1 *ENSG00000108821*	Collagen alpha-1(I) chain	Type I collagen is a member of group I collagen
CX3CL1 *ENSG00000006210*	Fractalkine	May play a role in regulating leukocyte adhesion and migration
CYP26C1 *ENSG00000187553*	Cytochrome P450 26C1	Plays a role in retinoic acid metabolism
*DDIT4L** *ENSG00000145358*	DNA damage-inducible transcript 4-like protein	Inhibits cell growth by regulating the TOR signaling pathway
DEFB136 *ENSG00000205884*	Beta-defensin 136	Has antibacterial activity
DKK1 *ENSG00000107984*	Dickkopf-related protein 1	Antagonizes canonical Wnt signaling by inhibiting LRP5/6 interaction with Wnt
*EDN1**	Endothelin-1	Endothelium-derived vasoconstrictor peptides

ENSG00000078401		
*ENPP5** *ENSG00000112796*	Ectonucleotide pyrophosphatase /phosphodiesterase family member 5	May play a role in neuronal cell communication
FAM25A *ENSG00000188100*	Protein FAM25	Unknown
FBXO25 *ENSG00000147364*	F-box only protein 25	Component of the SKP1-CUL1-F-box protein-type E3 ubiquitin ligase complex
FGF5 *ENSG00000138675*	Fibroblast growth factor 5	Functions as an inhibitor of hair elongation
*FGL1** *ENSG00000104760*	Fibrinogen-like protein 1	Hepatocyte **mitogenic** activity
FOXD1 *ENSG00000183900*	Forkhead box protein D1	Transcription factor required for formation of positional identity in the developing retina
FSTL3 *ENSG00000070404*	Follistatin-related protein 3	binding and antagonizing protein for members of the TGF-beta family
GADD45B *ENSG00000099860*	Growth arrest and DNA damage-inducible protein GADD45 beta	**Regulation of growth** and apoptosis, activation of stress-responsive MTK1/MEKK4 MAPKKK
GDF15 *ENSG00000130513*	Growth/differentiation factor 15	Transforming **growth factor** beta receptor signaling pathway
*GSTM3** *ENSG00000134202*	Glutathione S-transferase Mu 3	May govern uptake and detoxification of both endogenous compounds and xenobiotics
HAP1 *ENSG00000173805*	Huntingtin-associated protein 1	Associates specifically with huntingtin, Predominantly expressed in brain
HES3 *ENSG00000173673*	Transcription factor HES-3	Transcriptional repressor of genes that require a bHLH protein for their transcription
HIF1A *ENSG00000100644*	Hypoxia-inducible factor 1-alpha	Functions as a master transcriptional regulator of the **adaptive response to hypoxia**
*IL23A** *ENSG00000110944*	Interleukin-23 subunit alpha	IL-23 may constitute with IL-17 an acute response to infection in peripheral tissues
INSIG1 *ENSG00000186480*	Insulin-induced gene 1 protein	Mediates feedback control of cholesterol synthesis by controlling SCAP and HMGCR
IQCF5 *ENSG00000214681*	IQ domain-containing protein F5	Unknown
KCNJ8 *ENSG00000121361*	ATP-sensitive inward rectifier potassium channel 8	This potassium channel is controlled by G proteins
KRTAP10-9 *ENSG00000221837*	Keratin-associated protein 10-9	Interfilamentous matrix, consisting of hair keratin-associated proteins (KRTAP)
LCT *ENSG00000115850*	Lactase-phlorizin hydrolase	Splits lactose in the small intestine
*LIN28B** *ENSG00000187772*	Protein lin-28 homolog B	Suppressor of miRNA biogenesis by binding the let-7 miRNA precursor
*LY96** *ENSG00000154589*	Lymphocyte antigen 96	Cooperates with TLR4 in the innate immune response to bacterial lipopolysaccharide
MAF *ENSG00000178573*	Transcription factor Maf	Transcriptional activator or repressor in embryonic lens fiber cell development.
MORF4L1 *ENSG00000185787*	Mortality factor 4-like protein 1	NuA4 histone acetyltransferase omponent involved in transcriptional activation
MT1X *ENSG00000187193*	Metallothionein-1X	High content of cysteine residues that bind various heavy metals
NPPC *ENSG00000163273*	C-type natriuretic peptide	Regulation of cartilaginous growth plate chondrocytes **proliferation** and differentiation
NRARP *ENSG00000198435*	Notch-regulated ankyrin repeat-containing protein	May play a role in the formation of somites
NR2F2 *ENSG00000185551*	COUP transcription factor 2	Ligand-activated transcription factor. Activated by of 9-cis- and all-trans-retinoic acid
PGAM2 *ENSG00000164708*	Phosphoglycerate mutase 2	Interconversion of phosphoglycerate with 2,3-bisphosphoglycerate
*PHIP** *ENSG00000146247*	PH-interacting protein	Stimulates cell proliferation through regulation of cyclin transcription

PPY ENSG00000108849	Pancreatic prohormone	Regulator of pancreatic and gastrointestinal functions
RARRES1 ENSG00000118849	Retinoic acid receptor responder protein 1	Negative regulation of cell proliferation
RPS28 ENSG00000233927	40S ribosomal protein S28	Ribonucleoprotein
SDR9C7 ENSG00000170426	Short-chain dehydrogenase / reductase family 9C member 7	Weak conversion of all-trans-retinal to all-trans-retinol in the presence of NADH
S100A9 ENSG00000163220	Protein S100-A9	Calcium-binding protein. Has antimicrobial activity towards bacteria and fungi
*SLC4A2** *ENSG00000164889*	Anion exchange protein 2	Plasma membrane anion exchange protein of wide distribution
*SLC2A3** *ENSG00000059804*	Solute carrier family 2, facilitated glucose transporter member 3	Facilitative **glucose transporter**. Probably a neuronal glucose transporter
*SMCHD1** *ENSG00000101596*	Structural maintenance of chromosomes flexible hinge domain	**ATP binding**
*TBX6** *ENSG00000149922*	T-box transcription factor TBX6	Probable transcriptional regulator involved in developmental processes
TFRC ENSG00000072274	Transferrin receptor protein 1	Iron uptake via receptor-mediated endocytosis of ligand-occupied transferrin receptor

Supplemental Table 4: Genes coregulated with the *ANT3* gene. The full set of results obtained from the analysis with all constructed models of the *ANT3* promoter regions were screened as describeed in Figure 1 either on the full chromosomal human sequences or the human promoter library (results with an asterisk).

Gene, ID	Encoded protein (Ensembl)	Protein function (Uniprot)
APC2 ENSG00000115266	Adenomatous polyposis coli protein 2	Promotes rapid degradation of CTNNB1 and may function as a tumor suppressor
CSAG1 ENSG00000198930	Putative chondrosarcoma-associated gene 1 protein	Unknown
EFNA2 ENSG00000099617	Ephrin-A2	Unknown, binds to the receptor tyrosine kinases EPHA3, EPHA4 and EPHA5
IFT88 ENSG00000032742	Intraflagellar transport protein 88 homolog	Involved in primary cilium biogenesis
LIPC ENSG00000166035	Hepatic triacylglycerol lipase	Hepatic lipase hydrolyses phospholipids, glycerides, and acyl-CoA thioesters
MYEOV ENSG00000172927	Myeloma-overexpressed gene protein	Unknown, Overexpressed in tumor cells lines with a t(11;14)(q13;q32) translocation
NR1D2 ENSG00000174738	Nuclear receptor subfamily 1 group D member 2	Acts as a potent competitive repressor of ROR alpha function
RNF166 ENSG00000158717	RING finger protein 166	Unknown
STMN3 ENSG00000197457	Stathmin-3	Unknown, neuron specific
TTI1 ENSG00000101407	TEL2-interacting protein 1 homolog	Promotes assembly, stabilizes the activity of mTORC1 and mTORC2 complexes

Supplemental Table 5: Genes coregulated with the *ANT4* gene. The full set of results obtained from the analysis with all constructed models of the *ANT4* promoter regions were screened as described in Figure 1 either on the full chromosomal human sequences or the human promoter library (results with an asterisk). Protein function linked to spermatogenesis or involved in the prostate metabolism is shown in bold.

Gene, ID	Encoded protein (Ensembl)	Protein function (NCBI, GeneCards)
AAAS* ENSG00000094914	Aladin	Plays a role in the normal development of the peripheral and central nervous system
AMN ENSG00000166126	amnionless homolog	extraembryonic visceral endoderm layer
AMDHD2* ENSG00000162066	Putative N-acetylglucosamine-6-phosphate deacetylase	N-acetyl-D-glucosamine 6-phosphate + H_2O = D-glucosamine 6-phosphate + acetate
APEX1* ENSG00000100823	APEX nuclease (multifunctional DNA repair enzyme) 1	repair of apurinic / apyrimidinic sites **in testis**
BAX* ENSG00000087088	BCL2-associated X protein	Mutagenesis regulation in **spermatogenesis**
CD82* ENSG00000085117	CD82 antigen	Associates with CD4 or CD8 and delivers signals for the TCR/CD3 pathway
CDK4* ENSG00000135446	cyclin-dependent kinase 4	cell cycle G1 phase progression in male reproduction
CDYL2* ENSG00000166446	Chromodomain Y-like protein 2	Unknown
CLPB* ENSG00000162129	Caseinolytic peptidase B protein homolog	May function as a regulatory ATPase and be related to secretion/protein trafficking process
DNAJC13* ENSG00000138246	DnaJ homolog subfamily C member 13	Chaperone
ERP29* ENSG00000089248	Endoplasmic reticulum resident protein 29	Important role in the processing of secretory proteins within the endoplasmic reticulum
FIP1L1 ENSG00000145216	Pre-mRNA 3'-end-processing factor FIP1	Contributes to poly(A) site recognition and stimulates poly(A) addition
FLJ32713 fis	Unknown (TESTI2000756)	unknown (**expressed in testis**)
FNDC3A* ENSG00000102531	Fibronectin type-III domain-containing protein 3A	Mediates spermatid-Sertoli adhesion during **spermatogenesis**
G6PC2* ENSG00000152254	Glucose-6-phosphatase 2	Glucose production through glycogenolysis and gluconeogenesis, expressed **in testis**
G6PC3* ENSG00000141349	Glucose-6-phosphatase 3	Hydrolyzes glucose-6-phosphate to glucose in the endoplasmic reticulum, expressed **in testis**
HSPBAP1* ENSG00000169087	HSPB (heat shock 27kDa) associated protein 1	regulating stress response
IGF2R* ENSG00000197081	insulin-like growth factor 2 receptor	receptor for insulin-like growth factor 2 (IGF2) and mannose 6-phosphate
KAT5* ENSG00000172977	K(lysine) acetyltransferase 5	chromatin remodelling with an abundant **spermatid** protein
KLHL12* ENSG00000117153	Kelch-like protein 12	Subs.-specific adapter for ubiquitin-protein E3 ligase complex, highly expressed **in testis**
LAMP1* ENSG00000185896	lysosomal-associated membrane protein 1	binds amelogenin, differentially expressed in **spermiogenesis**
LIMCH1* ENSG00000064042	LIM and calponin homology domains-containing protein 1	Unknown
MAP7D2* ENSG00000184368	MAP7 domain-containing protein 2	Unknown
OSGEP* ENSG00000092094	Probable tRNA threonylcarbamoyl adenosine biosynthesis protein	Required for the formation of a threonylcarbamoyl group on adenosine

RMND1* ENSG00000155906	required for meiotic nuclear division 1 homolog	unknown
RPUSD4* ENSG00000165526	RNA pseudouridylate synthase domain-containing protein 4	Unknown, **expressed in prostate**
SLC25A31 ENSG00000151475	solute carrier family 25, adenine nucleotide translocator ANT4	mitochondrial ATP/ADP carrier in **spermatozoid**
SLC2A4 ENSG00000181856	solute carrier family 2, member 4 (GLUT4)	facilitated glucose transporter, **detected in human testis**
SOHLH1* ENSG00000165643	**spermatogenesis** and oogenesis specific basic helix-loop-helix 1	germ cell-specific, oogenesis regulator and **male germ cells**
TDRD1* ENSG00000095627	tudor domain containing 1	essential for **spermiogenesis**
THAP8* ENSG00000161277	O-sialoglycoprotein endopeptidase (**TESTI2004929)**	unknown
TKTL1 ENSG00000007350	transketolase-like 1	important role in transketolase activity, **testis** expressed
TMEM184A ENSG00000164855	transmembrane protein 184A = Sdmg1	**male-specific** expression in embryonic gonads
SOHLH1*		
SUN1 ENSG00000164828	chr. 7 unc-84 homolog A	nuclear anchorage/migration, expression of **meiotic reproductive genes**
UBE2B* ENSG00000119048	ubiquitin-conjugating enzyme E2B (RAD6 homolog)	post-replicative DNA damage repair **in spermatogenesis**