Modélisation d'un entrepôt de données dédié à l'analyse du transcriptome hépatique

Emilie Guérin[†] Fouzia Moussouni[†] Brice Courselaud[†] Olivier Loréal[†]

[†] Unité INSERM 522 – Rue Henri Le Guilloux, CHRU Pontchaillou - 35033 Rennes cedex

Courrier : {emilie.guerin, fouzia.moussouni, brice.courselaud, olivier.loreal}@rennes.inserm.fr

Résumé

Les techniques d'étude du transcriptome, dites à haut débit, génèrent un grand nombre d'informations qui nécessitent un traitement bioinformatique afin de les stocker, de les gérer et de les analyser. C'est dans ce cadre qu'a été initiée au sein de l'unité INSERM 522 la réalisation d'un entrepôt de données sur le transcriptome hépatique. Ce projet vise à regrouper au sein d'une même base de connaissances, des données complexes, variées et nombreuses sur les gènes du foie à des fins d'analyse. Ceci permettra de disposer d'un système de repérage de gènes cibles impliqués dans les pathologies du foie, afin de mettre au point des outils diagnostiques et thérapeutiques. Une étape primordiale à la réalisation de cet entrepôt est sa modélisation, aussi ai je été amenée à évaluer les besoins des utilisateurs par l'intermédiaire d'une enquête et à modéliser sous forme de diagrammes UML les cas d'utilisations. Cet article présente la démarche suivie, les cas d'utilisation de l'entrepôt et un exemple de scénario d'utilisation de l'entrepôt par un biologiste.

Mots-clés : *transcritpome, données en masse, fouille de données, entrepôt, modélisation*

1 Introduction

La génomique est une discipline en plein essor sur le plan scientifique et technologique. De nouvelles technologies dites à haut débit génèrent désormais une quantité considérable de données, faisant naître un besoin bioinformatique important afin de les gérer et de les analyser.

Le groupe fer de l'unité INSERM 522 est confronté à ce problème. Il tente de mettre en évidence des gènes impliqués dans les complications liées aux surcharges en fer dans l'organisme. Ces surcharges de différents types, favorisent en effet l'apparition de lésions de type cirrhose et cancer du foie au pronostic redoutable [16]. Ces complications pourraient être liées à la modulation de l'expression de certains gènes hépatiques. Le groupe s'intéresse donc aux variations du transcriptome hépatique (reflet de la nature et de la quantité des ARNm présents) en situation de surcharge en fer, afin, à terme, de mieux connaître le métabolisme du fer et de disposer d'outils diagnostiques et thérapeutiques. Depuis peu, le groupe a entrepris une étude globale du transcriptome, plus précisément, le laboratoire utilise depuis deux ans la technologie des macroarrays, qui permet d'étudier en parallèle l'expression de plus de 1200 gènes. D'autres études à haut débit sont en cours, notamment avec la réalisation d'une puce à ADN foie-spécifique (2 000 gènes)

Ces technologies, de par leur propriété à permettre l'étude d'un grand nombre de gènes en simultané, vont générer une masse considérable de données expérimentales sur les profils d'expression.

L'interprétation de ces données nécessite le recours à de nombreuses autres informations relatives à un gène et qui viennent se greffer à ces données d'expression. Ces données sont d'origines et de natures variées. Il peut notamment s'agir d'informations sur les séquences, les homologies, les motifs fonctionnels, les structures éventuelles, les interactions protéine-protéine, la distribution tissulaire, la localisation chromosomique. De plus, ces informations sont en perpétuel remaniement du fait de la croissance rapide de la connaissance que reflète en particulier l'augmentation du volume de la littérature scientifique.

C'est dans ce contexte que l'unité INSERM 522 a initié un projet bioinformatique qui consiste en la réalisation d'un environnement intégré, dédié à l'analyse des données considérables et spécifiquement obtenues par l'analyse du transcriptome hépatique. L'originalité de cette approche est d'opter pour une approche entrepôt de données (ou Data Warehouse) orienté objet, dans le domaine scientifique. Il s'agit d'une intégration des données d'intérêt sur le foie, d'origines diverses, dans lesquelles les utilisateurs puisent les informations à des fins d'analyse et de prise de décision à l'aide d'outils d'analyse et de restitution.

Cet entrepôt de données (fig.1), baptisé Gedaw (Gene Expression DAta Ware house), intègre à ce jour les données expérimentales du groupe fer de l'Unité 522 (par l'intermédiaire de la création d'une base de données relationnelle Transcriptome) et intégrera à terme des données expérimentales d'autres équipes, des données pertinentes provenant de grandes banques publiques (GenBank, EMBL…), des données bibliographiques et des données cliniques qui pourraient provenir du groupe "études cliniques en hépatologie " de l'Unité 522, ou de bases de connaissances médicales standard (telles que UMLS : Unified Medical Language System). Une fois

intégrées, ces données vont être analysées à l'aide de requêtes et d'outils dédiés, développés localement et/ou importés des sites Web.

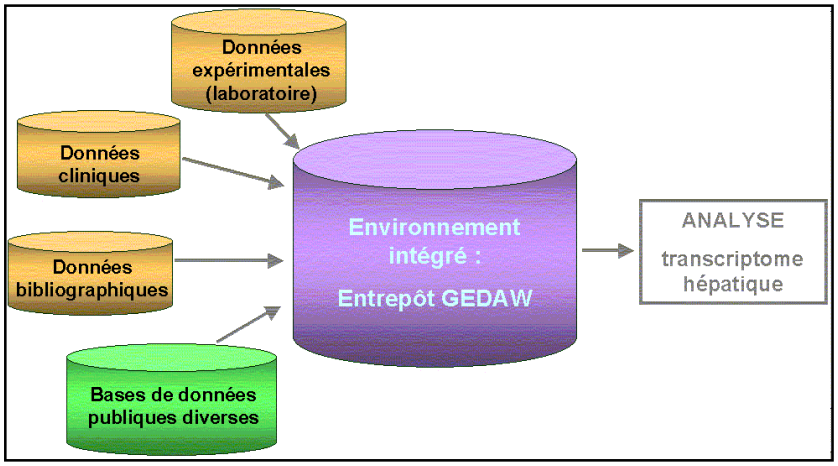

FIG.1 - L'entrepôt de données Gedaw (Gene Expression Data Ware house)

Les objectifs de la réalisation de cet entrepôt sont doubles :

1- Fournir aux chercheurs un outil complet et intégré facilitant ainsi l'accès aux données sur le transcriptome (analyse simple). Ceci correspond à une simple interrogation de la base de données.

2- Permettre par le biais de ce même outil, de fournir une aide à la décision, permettant d'orienter les recherches biologiques. Une fouille précise et complexe des données (ou Data Mining) préexistantes amènera à émettre des hypothèses qui vont guider la recherche sur le foie.

Cet article fait l'objet d'une modélisation de l'entrepôt Gedaw. Cette étape consiste à créer une représentation simplifiée du problème : le modèle. Grâce au modèle, il est possible de représenter la structure même de l'application, la manière dont elle va être réalisée, les acteurs en jeu et ses cas d'utilisation. C'est sur cette modélisation que se base tout le développement.

Aussi, pour la réalisation de cette modélisation, il faut s'imprégner du contexte biologique de l'Unité 522, prendre connaissance d'un certain nombre de concepts bioinformatiques et informatiques (entrepôt de données, concepts objet) et rencontrer un certain nombre de biologistes issus de diverses équipes afin de déterminer les besoins concernant l'analyse des données en masse générées par les nouvelles technologies.

2 La modélisation de Gedaw

2.1 Les enquêtes

Le projet de création de Gedaw a été initié dans le but de répondre aux attentes des biologistes en ce qui concerne l'analyse des données sur le transcriptome générées par des technologies haut débit. Aussi, la première étape dans la réalisation d'un tel outil est d'acquérir l'expertise puis de déterminer l'ensemble des besoins des biologistes.

Pour cela, des enquêtes ont été menées auprès des éventuels futurs utilisateurs par le biais d'un questionnaire afin de recueillir l'ensemble des cas d'utilisation d'un tel environnement [21].

Les fonctionnalités permises par Gedaw doivent en priorité répondre aux besoins des biologistes de l'unité INSERM 522. Cependant, d'autres équipes rennaises travaillent sur le transcriptome hépatique en collaboration avec l'Unité 522, c'est le cas de l'Unité INSERM 456, dans le cadre d'un projet de création d'une puce à ADN foie-spécifique. De plus, de nombreuses équipes travaillent sur le transcriptome d'une manière générale et envisagent de se lancer dans des études à grande échelle. Tous ces biologistes sont ou seront par la suite confrontés à la masse de données considérable générée par ces techniques et de ce fait seront amenés à faire appel à l'outil bioinformatique.

Nous avons été amenés à rencontrer un grand nombre de biologistes, préoccupés par l'étude du transcriptome et confrontés au mêmes difficultés, afin de déterminer au mieux et d'une manière globale quelles étaient leurs attentes en terme d'analyse des données générées sur le transcriptome.

Les enquêtes que nous avons effectué ont été menées auprès des personnes suivantes :

- Jean Léger de l'Unité INSERM U 533 à Nantes
- Nathalie Bdioui et Fabrice Morel de l'Unité INSERM U 456 à Rennes
- Francis Galibert, Jean Mosser et Frédérique Hublert de l'UMR 6061 à Rennes
- Madeleine Douaire et Christian Diot de l'INRA à Rennes. Département de génétique animale

- Christiane Guillouzo, Olivier Loréal et Brice courselaud de l'Unité INSERM U 522 à Rennes

Les enquêtes ont été menées auprès de ces personnes après accord par e-mail, à l'aide d'un questionnaire effectué par nous mêmes. Etant donnée la diversité des sujets d'étude des différents biologistes rencontrés, le questionnaire est resté relativement général, soulevant des questions de fond comme l'étude du transcriptome, les technologies d'étude employées et l'apport bioinformatique souhaité. En réalité, plus qu'un questionnaire, il s'agissait d'un support à la discussion. Le discours s'orientait selon les thématiques. En effet, il est plus difficile de réaliser un tel questionnaire dans le domaine de la biologie que dans le domaine de l'entreprise. La biologie est un vaste domaine d'étude qui présente encore à l'heure actuelle des points à préciser ou même à découvrir. De plus, certaines informations sont erronées et doivent pouvoir être remises en cause voire corrigées. A la fin de chaque entretien, nous devions pouvoir dégager des cas d'utilisation nouveaux ou déjà formulés. Il est apparu que les rencontres les plus intéressantes en terme de résultats, étaient celles avec des chercheurs ayant déjà entamé des études du transcriptome à grande échelle (type puces à ADN ou macroarrays), ils étaient alors réellement confrontés aux problèmes et concevaient relativement bien ce que pouvait leur apporter un outil comme Gedaw. Cependant, tous les entretiens furent enrichissants et bénéfiques pour l'étude.
Les cas d'utilisation les plus importants, dégagés de ces enquêtes sont décrits plus bas à la norme UML sous forme de diagrammes des cas d'utilisation et de diagrammes de séquence.

2.2 Les modèles dégagés des enquêtes

Toute la modélisation de l'entrepôt a été effectuée grâce aux différents diagrammes proposés par la methodologie UML. Nous avons choisi la méthodologie UML pour ses caractéristiques, c'est à dire son orientation objet et son dynamisme permettant une modélisation aisée des problèmes biologiques et bioinformatiques. La modélisation a permis d'envisager la structure de Gedaw mais surtout les cas d'utilisations directement dégagés des enquêtes menées.

2.2.1 Le diagramme de cas d'utilisation de l'entrepôt Gedaw

Le diagramme des cas d'utilisation (fig.2) représente donc l'ensemble des cas d'utilisation de Gedaw, les acteurs en jeu et les relations entre ces différents cas.
Notre système Gedaw présente pour l'instant deux acteurs que sont le bioinformaticien et le biologiste.
Le bioinformaticien interagit avec l'entrepôt pour intégrer, nettoyer et rafraîchir (mettre à jour) les données. Il intervient également pour réaliser l'interface de l'entrepôt et y rajouter une fonctionnalité lorsque les biologistes en émettent le souhait. Le rajout de fonctionnalité s'effectue après formatage de la requête.
Le biologiste peut interroger simplement la base de connaissances constituée de données agrégées sur le foie ou effectuer une analyse fouillée : clustering ou analyse des gènes. Nous verrons dans la description du scénario par un diagramme de séquence que clustering et analyse des gènes peuvent se succéder.
Ces différentes tâches peuvent également être réalisées par le bioinformaticien, mais ces liens ne sont pas représentés sur le diagramme, par souci de clarté.

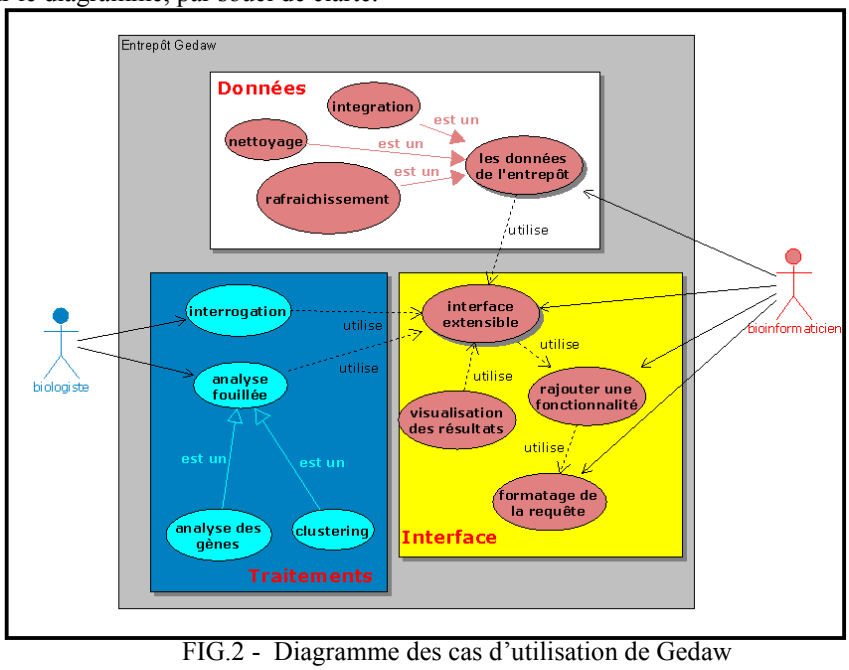

FIG.2 - Diagramme des cas d'utilisation de Gedaw

Il existe beaucoup de relations entre les différents cas d'utilisation. Par exemple, l'interrogation de la base de données dépend de (ou utilise) l'interface qui dépend elle même des données de l'entrepôt.

La réalisation de ce diagramme des cas d'utilisation a permis de distinguer 5 composantes importantes de l'entrepôt encore appelées domaines. Ces domaines sont les suivants : Expérience, Gène, Analyse des gènes, Clustering et Requête. Ces domaines ont été également modélisés par l'intermédiaire de diagrammes des classes, ils ne seront pas décrits dans l'article.

2.2.2 Définition des diagrammes de séquence

Le fait d'avoir représenté les cas d'utilisation d'un système apporte une connaissance sur l'interface d'un système vis à vis des utilisateurs du système. Toutefois, un cas d'utilisation nécessite pour être compris, une description plus détaillée.
Chaque cas peut être accompagné de plusieurs scénarios qui l'illustrent. Cette description de scénarios peut être donnée sous la forme de texte ou sous la forme de diagrammes de séquence. Ces diagrammes de séquence permettent de spécifier les échanges de données ou d'événements entre les composantes du système [8] [12]. A noter que les composantes du système ont préalablement été définies par les diagrammes de classes.
Les boîtes grises représentent des boîtes d'activation de messages. Les messages sont séquentiels et vont d'une classe à une autre. Les classes sont indiquées dans les boîtes rouges, ainsi que les domaines auxquels elles appartiennent. On ne peut concevoir ces diagrammes qu'après avoir défini les domaines par des diagrammes de classes.
Nous ne représenterons ici qu'un exemple de cas d'utilisation de l'entrepôt par un biologiste souhaitant effectuer une analyse fouillée. Plus particulièrement la requête est celle d'un clustering suivi d'un alignement multiple de séquences (fig.3).

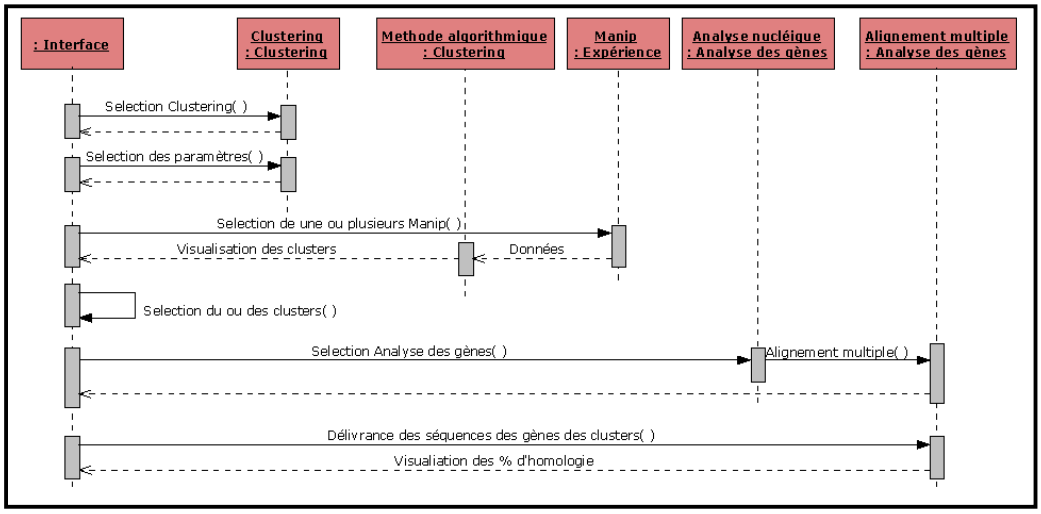

FIG.3 - Diagramme de séquence d'un scénario de clustering suivi d'un alignement multiple

Le biologiste, au niveau de l'entrepôt, sélectionne un outil de clustering, puis les paramètres qui lui sont associés. Il choisit ensuite sur quelles expériences il veut effectuer cette analyse. Ceci aboutit à la visualisation des clusters. Le chercheur peut continuer son analyse par un alignement multiple. Il sélectionne donc cette option après avoir choisi le ou les cluster(s) sur le(s)quel(s) réaliser cet alignement. Les séquences des gènes présents dans ce ou ces cluster(s) sont transmises à l'outil d'alignement et le biologiste peut ensuite visualiser les résultats sous forme de % d'homologie. Ces % d'homologie révéleront si les séquences appartenant à un même cluster sont proches (fort %) ou non (faible %).
Ceci est un exemple de requête, et l'analyse peut encore se poursuivre si le biologiste le souhaite, à la vue de résultats intéressants. Tout cela dans la limite des outils intégrés à l'entrepôt, qui évolueront cependant avec le temps. On peut notamment envisager une recherche de motif commun, donné ou non, une recherche de localisation chromosomique, une recherche bibliographique. Le chercheur peut donc pour ses requêtes, partir avec ou sans à priori.

3 Discussion

La modélisation de Gedaw a permis de totalement définir l'architecture de l'entrepôt et de ce fait la démarche à suivre pour sa conception. Nous pouvons notamment définir quelles données sont à intégrer au sein de l'entrepôt. A ce jour, l'entrepôt contient uniquement les données expérimentales du groupe fer. Nous sommes également en mesure d'extraire automatiquement, des séquences d'intérêt à partir de Genbank et de toute connaissance connexe, grâce à l'utilisation du concept d'ontologie et de la technologie XML.
La modélisation suggère également une intégration de données bibliographiques, de données cliniques, de données sur les séquences et la structure des protéines. Cependant, la sélection des données d'intérêt et le mode de rapatriement n'ont pas encore été évoqués. De plus, pour rendre performante l'extraction de nouvelles données sur le transcriptome, nous nous attachons au partage de la connaissance entre ontologies de domaines différents. Des extensions des concepts présents dans les ontologies MGED, UMLS et GO sont en cours afin de permettre une interconnexion plus explicite entre les gènes, leur modulation d'expression éventuelle et les pathologies pouvant y être associées.
Nous pouvons d'autre part lister un certain nombre d'outils d'analyse qui seront à intégrer, nous pensons en particulier aux outils de clustering, de comparaison de séquences, de recherche de motifs, de recherche de séquences répétées. Certains de ces outils sont commercialisés ou disponibles sur le Web. Nous tenterons au maximum d'utiliser au sein de Gedaw des outils déjà implémentés et flexibles, pouvant facilement être intégrés.

4 Références

[1] Attwood, T. (1999). Introduction to bioinformatics. *Cambridge Press ed*.

[2] Bouzeghoub, M., Gardarin, G., Valduriez, P. (1997). Les objets. *Eyrolles ed*.

[3] Clavel, G., Mirouze, N., Munerot, S., Pichon, E., Soukal, M., Tiffanneau, S. (2000). Java La synthèse. *Dunod ed*.

[4] Eisen, M.B., Brown, P.O. (1999). DNA Arrays for Analysis of Gene Expression. *Methods Enzymol* 303, 179-205.

[5] Franco, J.M. (1997). Le Data Warehouse. *Eyrolles ed*.

[6] Gardarin, G. (1999). Bases de données objet et relationnel. *Eyrolles ed*.

[7] Inmon, W.H., Hackathorn, R.D. (1994). Using the Data Warehouse. *Wiley-QED Publication*.

[8] Lai, M. (2000). UML La notation unifiée de modélisation objet. *Dunod ed*.

[9] Lainé, B. (2000). Contribution à la construction d'un entrepôt de données scientifique. *Rapport de stage de DESS TIMH*.

[10] Lemay, L., Cadenhead, R. (2001). Le magnum Java 2. *CampusPress ed*.

[11] Lukashin, A.V., Fuchs, R. (2001). Analysis of temporal gene expression profiles : clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics journal, 17, 405-414*.

[12] Morley, C., Leblanc, J.H.B. (2000). UML pour l'analyse d'un système d'information. Dunod ed.

[13] Moussouni, F., Paton, N.W., Hayes, A., Oliver, S., Goble, C., Brass, A. (1999). Database Challenges for Genome Information in the Post Sequencing Phase. *10th international conference on databases and expert systems applications. LNCS 1677, 540-549*.

[14] Moussouni, F., Berti, L. (2001). Ontologies and mapping rules for merging data from public databanks and gene expression experiments. *Colloque Ontologie et Biologie Moléculaire INRIA Grenoble.*

[15] Muller, P.A., Gaertner N. (2000). Modélisation Objet avec UML. *Eyrolles ed*.

[16] Niedereau, C., Fischer, R., Purschel, A., Stremmel, W., Haussinger, D., and Strohmeyer, G. (1996). Long-term survival in patients with hereditary hemochromatosis. *Gastroenterology 110, 1107-1119* .

[17] Paton, N.W., Khan, S., Hayes, A., Moussouni, F., Brass, A., Eilbeck, K., Goble, C.A., Hubbard, S., Oliver, S. (2000). Conceptual Modelling of Genomic Information. *Bioinformatics Journal* 16, 548-558. To appear in the Yearbook of Medical Informatics, Haux, R., and Kulikowski, C. (eds), 610-619, Schattauer, 2002.

[18] Pigeon C, Ilyin G, Courselaud B, Leroyer P, Turlin B, Brissot P, Loréal O. (2001). A new mouse liver specific gene, encoding a protein homologous to human antimicrobial peptide hepcidin, is overexpressed during iron overload. *J Biol Chem* 276, 7811-7819.

[19] Pigeon, C., Legrand, P., Leroyer, P., Bouriel, M., Turlin, B., Brissot, P., Loréal, O. (2001). Stearoyl coenzyme a desaturase 1 expression and activity are increased in the liver during iron overload. *Biochim Biophys Acta* 1535, 275-284.

[20] Rumbaugh et al. (1995). OMT Modélisation et Conception Orientée Objet. *Masson ed*.

[21] Stevens, R.D., Goble, C.A., Baker, P., Brass, A. (2000). A classification of tasks in bioinformatics. *Bioinformatics Journal to appear*.

[22] Turlin, B., Juguet, F., Moirand, R., Le Quilleuc, D., Loreal, O., Campion, J.P., Launois, B., Ramée, M.P., Brissot, P., Deugnier, Y. (1995). Increased liver iron stores in patients with hepatocellular carcinoma developed on a noncirrhotic liver. *Hepatology 22, 446-450* .