Université d'Evry Val d'Essonne

Anné Universitaire 2008 / 2009

Master 2 Génie Biologique et Informatique

Gabriel Chandesris

Systèmes d'intégration de données en biologie

Entrepôts de données et systèmes de médiations

Étude sur BioWareHouse et BioMediator

Dans le cadre du cours de : `Fariza Tahi`

Cours systèmes d'intégration - M2 GBI 2008-2009

Table des matières

1 Introduction

La problématique du grand nombre de données disponibles en biologie a amené à créer un certain nombre de banques de données généralistes ou spécialisées, dans des formats souvent différents et des systèmes peu compatibles. Différentes approches de systèmes d'intégrations de données ont été étudiées ces dernières années afin d'assurer un accès transparent aux banques de données et à leur contenu de façon uniforme.

Ces systèmes sont les entrepôts de données et les systèmes de médiation : les premiers sont utiles pour un stockage préalable à l'accès aux données, les seconds pour un accès transparent distant aux données accessibles en ligne.

---

2 Deux approches : entrepôt et médiation, exemples

2.1 L'entrepôt : l'information centralisée et normalisée

Les entrepôts de données sont un type de systèmes d'intégrations de données centralisées. Il s'agit d'un stockage de données consolidées et rassemblées en un même endroit, à partir de plusieurs sources. Un tel système possède un certain nombre d'avantages comme le recoupement et la non-redondance des données, ainsi qu'un accès rapide à l'ensemble des données, même avec une requête complexe.

De plus, l'existence d'un schéma global à l'ensemble de l'entrepôt de données y assure une cohérence, notamment lors de l'intégration de sources de données ayant des schémas différents, données externes publiques et / ou données internes. L'application de ce schéma global, ainsi que la récupération des données à partir des sources, implique une mise à jour régulière à partir des sources afin d'éviter l'obsolescence de l'entrepôt de données et de son contenu.

2.2 Un entrepôt de données : BioWarehouse

BioWarehouse [7] a été conçu et développé comme un système de construction et de gestion d'entrepôts de données, afin de permettre l'interopérabilité de bases de données bioinformatiques disparates.

2.2.1 Les sources de l'entrepôt

Les sources définies à la conception de BioWarehouse sont recensées dans le tableau de la figure 1. Le système PublicHouse, donné à titre d'exemple par les concepteurs de BioWarehouse, intègre dans son entrepôt les bases de données accessibles publiquement.

BioCyc Dbs	BioPAX format
Comprehensive Microbial Resource (CMR)	ENZYME DB
GenBank (bacteria only)	Eco2dbase
Kyoto Encyclopedia of Genes and Genomes (KEGG)	Gene Ontology
MAGE-ML format	MetaCyc Ontology
UniProt (Swiss-Prot and TrEMBL)	Taxonomy DB

FIG. 1 – Sources de données intégrée à la conception de BioWarehouse

2.2.2 Extraction, Transformation et Chargement

L'extraction des données s'effectue selon la lecture des bases définies et le chargement de données est fait dans la base de BioWareHouse selon le schéma global de l'entrepôt (conversion des sources en un schéma relationnel et selon la sémantique de BioWarehouse), il n'y a pas de traitement des redondances. Les enregistrements précédents sont conservés.

Chaque module de chargement (*loader*) est spécifique à la source correspondante, ces modules sont implémentés généralement en C ou en Java. Un module est chargé de la récupération des données sources, de leur extraction et de leur conversion selon le schéma de l'entrepôt. Le chargement des données dans la base s'effectue sans traitement autre que le respect de la sémantique et du schéma global. Des améliorations et des ajouts de module de chargement sont possibles.

2.2.3 Gestion de l'entrepôt

L'implémentation de BioWarehouse est prévue pour être utilisée selon un schéma relationnel et pouvant être utilisé avec des bases relationnelles libres comme MySQL ou commerciales comme ORACLE. La version accessible en ligne *PublicHouse* utilise une base MySQL.

Le schéma d'intégration de BioWarehouse (Figure 2) est défini de façon globale dans un fichier XML en deux parties. La première partie, appelée «CORE» définit l'ensemble des données(Figure 2), la seconde partie appelée «MAGE» est une extension pour gérer les annotations d'expressions géniques. Les tables du schéma relationnel sont définies à partir de schémas fréquemments rencontrés en biologie avec une unification des termes utilisés (utilisation d'ontologies) : ceci permet une intégration de données de sources diverses chargées à partir de différents modules.

Différents concepts biologiques sont définis au sein du schéma de BioWarehouse, notamment les taxons, les sources biologiques, les acides nucléiques, les gènes, les protéines, les micromolécules, les voies biochimiques... Diverses relations ont été établies entre ces concepts afin de pouvoir les regrouper lors d'une requête.

2.2.4 Utilisation et requêtage

L'accès à la base s'effectue par des requêtes SQL. Des utilitaires ont été créés spécifiquement pour BioWarehouse, sous la forme de classes Java pouvant être utilisées dans des logiciels clients d'accès à BioWarehouse.

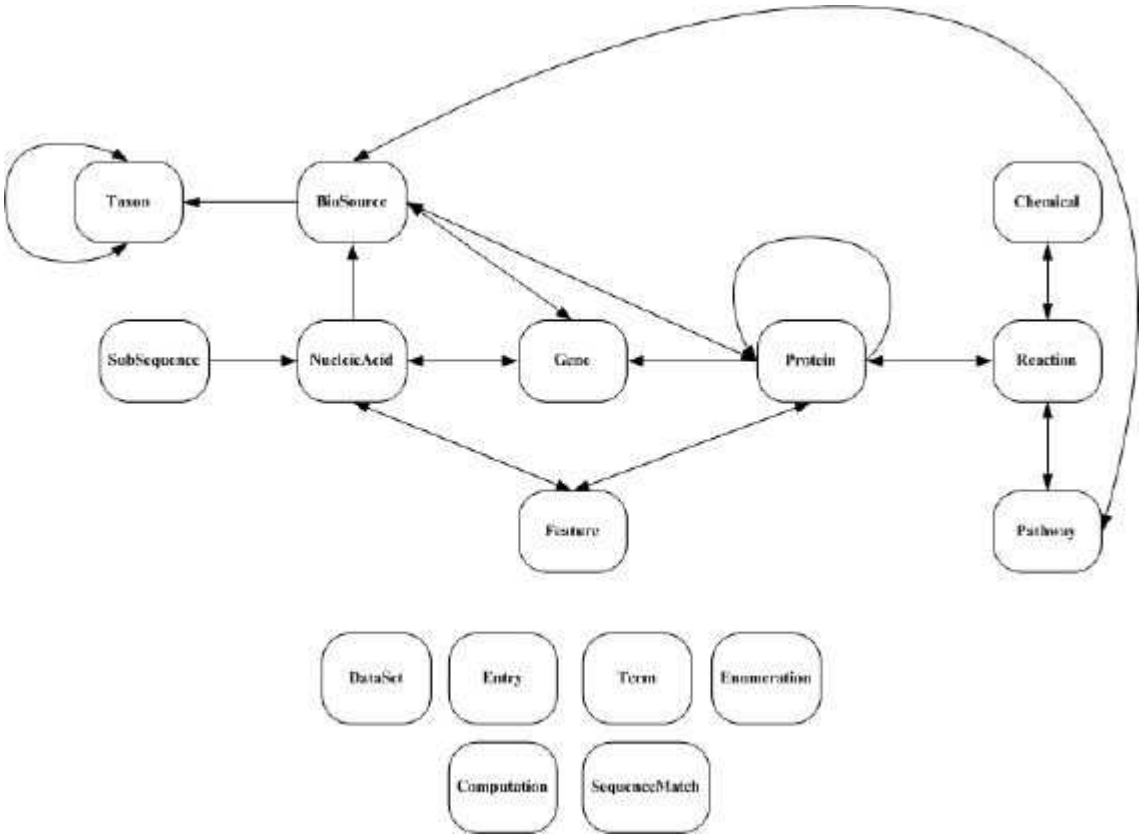

FIG. 2 – Schéma global de BioWarehouse

Une documentation libre contient des notes de version, un guide de démarrage, la description des schémas d'intégration et d'autres éléments nécessaires pour démarrer BioWarehouse. Des exemples d'utilisation de BioWarehouse et sa documentation sont disponibles à l'adresse
<http ://biowarehouse.ai.sri.com/repos/doc/index.html>.

2.2.5 Disponibilité

BioWarehouse est disponible en téléchargement et en utilisation à distance (accès à la base MySQL ou interface web *PublicHouse*, sous réserve d'enregistrement comme utilisateur), des informations techniques détaillées sur l'accès à distance sont disponibles à l'adresse
<http ://biowarehouse.ai.sri.com/PublicHouseOverview.html>.

2.3 La médiation : accès distant transparent

Les systèmes de médiations sont des interfaces permettant un accès transparent aux données, sans que celles-ci soient centralisées, contrairement aux entrepôts de données. Ceci permet un accès aux dernières mises à jour des données sources, mais seulement pour les sources et bases de données accessibles, que ce soit par accès distant ou un adaptateur (*wrapper*) de transformation de la requête et de récupération des données.

La limite de ce type de système d'intégration de données est la représentation des données, afin de permettre une requête générale qui est ensuite appliquée de façon spécifique aux différentes sources, et ensuite de rassembler ces données uniformément. Pour cela un médiateur transforme la requête dans un langage pivot et la transmet à différents adaptateurs, chargés de lancer la requête ou une partie de celle-ci sur les différentes sources. Ces adaptateurs récupèrent ensuite les données des sources, les traduisent conformément au langage pivot pour les transmettre au médiateur chargé de rassembler ces données de résultat.

Les recherches dans le domaine des systèmes de médiation ont permis de définir certaines de leurs propriétés spécifiques :

1 Permettre l'utilisation à un grand nombre d'utilisateurs.
2 Être flexible et accepter des requêtes diverses.
3 Permettre à différents schémas centralisés (de médiation) pour interpréter ces requêtes.
4 Être suffisemment modulaire pour évoluer avec les sources de données et permettre une utilisation croissante de ces sources.

Différents sytèmes de médiations ont été développés dans la domaine de la biologie et de la médecine, afin de permettre de fédérer les données :

- Object Protocol Model (OPM) [1], pour l'encapsulation des sources et la traduction des requêtes, ainsi que l'intégration de schémas.
- Kleisli [2], développé plus spécifiquement pour la biologie moléculaire, l'absence de schéma global permet d'intégrer un grand nombre de sources, mais oblige à connaitre les spécificités des sources.
- TAMBIS [10] est un projet d'utilisation d'une ontologie sur une base de connaissance, qui est utilisée comme schéma d'ensemble d'éléments potentiels de requêtes, ces éléments pouvant être combinés pour des requêtes complexes.

2.4 Un système de médiation : BioMediator

BioMediator [3], a été développé à l'université de Washington et est disponible à l'adresse<*http ://www.biomediator.org/*>. Son architecture (Figure 3) comprend six composants principaux pour permettre l'intégration de données sources, structurées ou semi-structurées. Le fonctionnement de cette architecture peut être décrite de la façon suivante :

A **Formulation de la requête** : Query / PQL (Path Based Query Language)
B **Traduction de la requête** : Query reformuler to XQuery
C **Base de connaissance** : Source Knowledge Base (*Protégé*)
D **Traduction de la requête** : Query Engine to URL (**Qexo**)
E **MetaWrapper** (*Semantic Transition step*)
F **Adaptateurs** / ***Wrapper(s)*** (*Syntactic translation Step*) and link to Public Data Sources

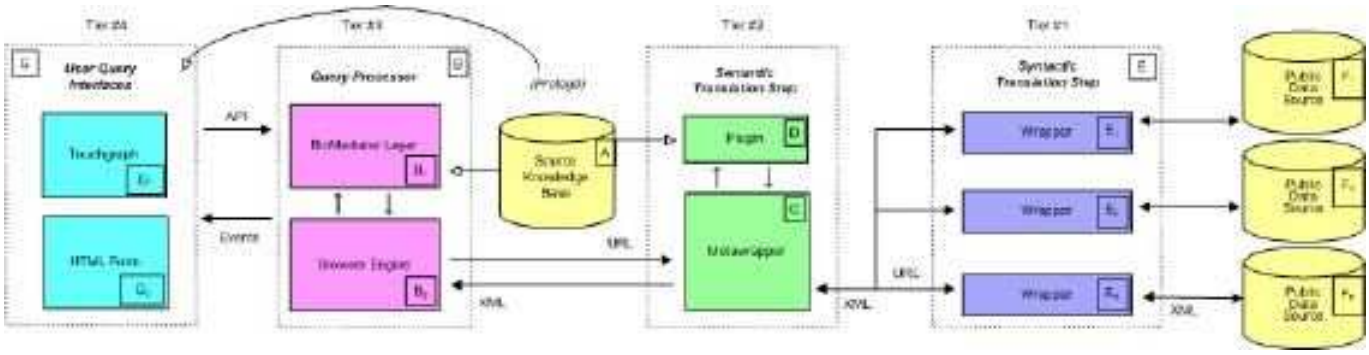

FIG. 3 – Architecture de BioMediator [3]

De manière générale, BioMediator est décrit par ses concepteurs comme un assemblage d'une interface de requêtage utilisateur, un système de traitment de la requête, un système de traduction sémantique et un ensembles d'interfaces avec les sources, ces dernières portant le nom d'adaptateurs ou *Wrappers*. À ces quatres éléments s'ajoute le système de requêtage via une interface dédiée pour l'utilisateur.

Les **ajouts et avantages** par rapport à d'autres systèmes de médiation :
- L'architecture est conçue pour être modulaire et facilement adaptable.
- Le langage de requête *PQL (Path based Query Language)* est suffisemment souple pour relier les éléments entre eux en plus d'une valeur désirée.
- Le système utilisé dans la base de connaissance *Protégé* est un bon format de représentation des données, facile à utiliser pour faire évoluer le schéma.

Des **améliorations sont possibles** afin d'étendre le projet BioMediator :
- Utilisation d'un système en langage naturel (Natural Language Processing) (données disponibles directement seulement à partir de données structurées ou semi-structurées).
- Intégration d'outils d'analyses (BLAST par exemple) car seules les données d'annotation déjà présentes sont utilisées.
- Adaptation d'une interface plus adaptée aux biologistes (diagrammes, création de requêtes par l'exemple).
- Adaptation plus pertinente des adaptateurs aux différentes sources, des modifications sont nécessaires malgré l'approche généraliste du système BioMediator.

2.4.1 Les sources du système de médiation

Aucune source n'est formellement donnée, même à titre d'exemple, pour l'utilisation de BioMediator, mais des adaptateurs ont été notamment développés pour OMIM (parties *Disease* et *Gene*), Entrez (parties *Gene*, *Nucleotid* et *Protein*) et Swissprot.

2.4.2 Médiateur

BioMediator utilise d'un schéma de médiation (Figure 4) annoté pour décrire l'ensemble des données (et permettre la création de requêtes avec ajout éventuel d'informations. La base de connaissance est modifiable via l'interface pour correspondre aux besoins des utilisateurs.

Le médiateur proprement dit est constitué des constituants B et C (Figure 3), le système Qexo et le MetaWrapper, ce qui constitue des éléments de traduction de la requête et de récupération des résultats. Les métadonnées contenues dans la base de connaissances *Protégé* (élément A) sont utilisées dans la traduction de la requête par le MetaWrapper (éléments C et D). L'intérêt du médiateur est de traduire la requête entrante au format PQL en un ensemble de requêtes au format XQuery et destinées aux différents adaptateurs.

2.4.3 Adaptateur (*Wrapper*)

Les adaptateurs sont généralisés au maximum de données fournies par la source correspondante, ce qui limite leur modification à l'extension des données de la source. Ces modules spécifiques aux différentes sources effectuent le lien et le requêtage auprès des sources, ainsi que l'analyse des réponses obtenues.

L'interrogation et la réponse d'une source donnée s'effectuent dans son propre format, l'adaptateur converti la requête dans le format de la source, et la réponse dans un format XML correspondant à la base de connaissances et au schéma de médiation du système. C'est cette réponse au format XML qui est ensuite retransmise au médiateur, pour y être regroupée avec les autres réponses des autres adaptateurs.

2.4.4 Métadonnées

La base de connaissances *Protégé* (élément A, Figure 3) constitue un schéma de médiation et de correspondances entre données, tel qu'il peut être visualisé dans la figure 4. L'ensemble des métadonnées utilise notamment un système d'ontologies pour permettre la traduction sémantique dans le médiateur. L'intérêt d'un schéma de médiation est de relier les concepts communs aux différents sources externes au système de médiation, ainsi que les relations entre ces concepts.

Pour chaque source de données, les concepts contenus sont récupérés ainsi que leurs relations, de plus le concept principal autour duquel est organisé la source pour permettre de résoudres les relations ambiguës dans le schéma des sources. Une telle utilisation du schéma des sources amène à concevoir BioMediator comme un système de médiation *LaV (Local as View)*.

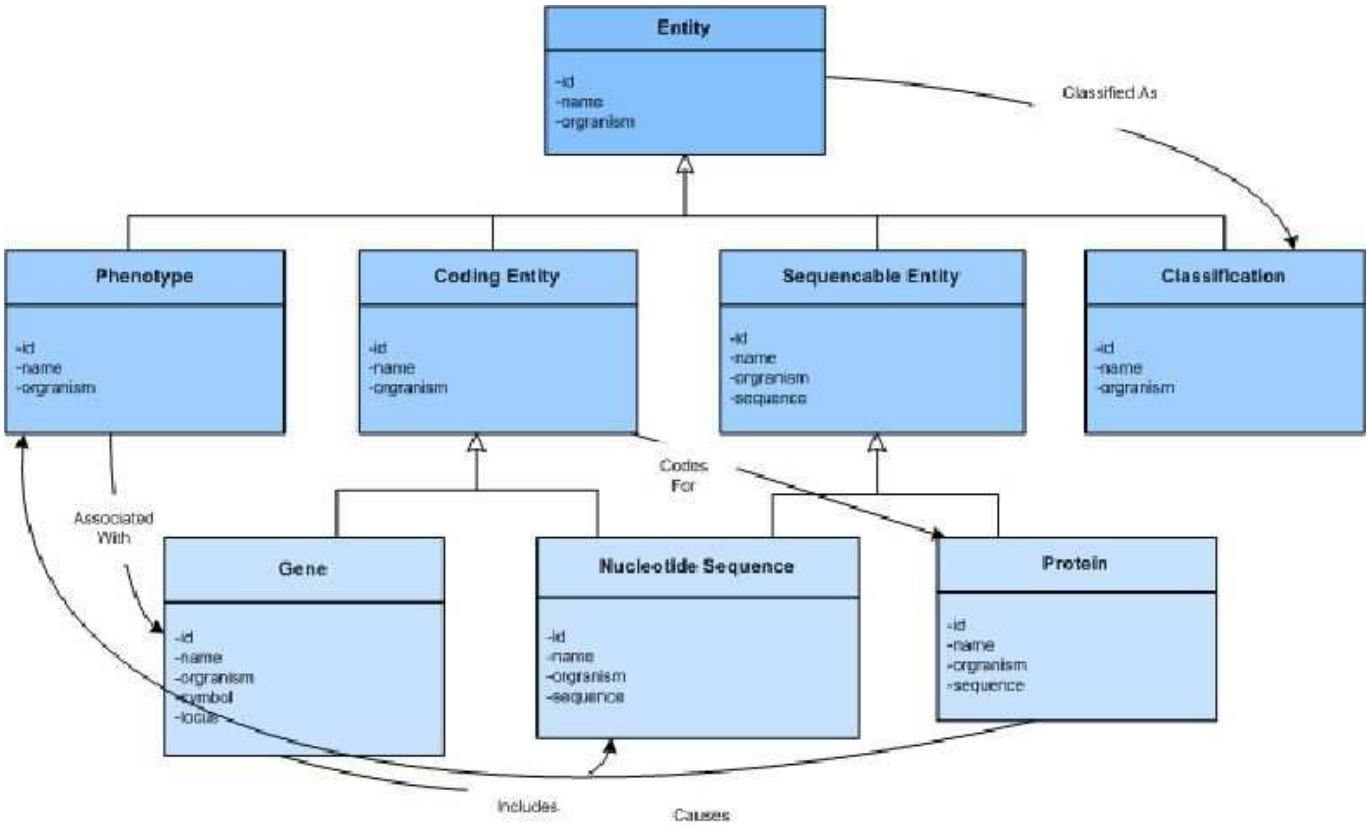

FIG. 4 – Schéma de médiation de BioMediator [3]

2.4.5 Utilisation et requêtage

L'objectif principal d'utilisation de BioMediator est un requétage en langage naturel, le système construit alors un ensemble de schéma de données à valider par le chercheur pour sa requête via le système de médiation. Actuellement l'interface de BioMediaor est conçue autour des schéma de données construits par l'utilisateur avec PQL. Cette requête est ensuite reprise, traduite et utilisée par le médiateur.

Les résultats obtenus par les adaptateurs et transmis au médiateurs pour être regroupés sont ensuite traduits de façon graphique et affichés sous froem de graphes et d'arbres à l'utilisateur, ces arbres et graphes donnant une vue d'ensemble des éléments de réponse avec leurs liens. L'ensemble du processus entre la requête et la réponse étant totalement transparent pour l'utilisateur.

2.4.6 Disponibilité

BioMediator est décrit précisément, documenté et téléchargeable à l'adresse <http ://www.biomediator.org/>. Son code source est également consultable. Ce logiciel n'est plus développé depuis mars 2007, mais le site web et son contenu restent encore accessible actuellement (octobre 2008).

---

3 Conclusion et perspectives

Les **systèmes de médiations** permettent une certaine souplesse dans une adaptation et un accès à différentes sources de données non centralisées, la principale difficulté étant la définition d'un schéma local ou global de recherche et de recollement des données selon une médiation efficace et des adaptateurs corrects. Le traitement des éventuelles erreurs et redondances devant être effectué de façon dynamique lors de la recherche de données et du rendu de résultats : l'inconvénient majeur de ce type de système étant l'adaptation à l'évolution des systèmes de données sources et non de leur contenu.

Le mode de fonctionnement des systèmes de médiation est une interface au sein de systèmes hétérogènes souvent très différents et peu normalisés entre eux, et dont il est difficile de maîtriser les mécanismes de mises à jour. L'usage d'un tel système d'intégration de données peut se révéler utile pour un système de veille informative autour de thématiques diverses, d'actualité, de recherche autour de sources externes, mais se limite souvent aux sources du domaine public.

Quant aux **entrepôts de données**, l'intérêt est essentiellement de pouvoir stocker de façon centralisée des données disparates selon un schéma d'ensemble unifié et organisé pour permettre une correction des données (comme la correction des erreurs, la suppression des redondance...) et un usage approfondi des données, afin de permettre une analyse et une fouille des données efficaces dans un cadre de rechrche ou de prise de décisions. Le principal inconvénient de ce type de système est la mise à jour régulière nécessaire de son contenu par rapport aux données sources.

Un entrepôt de données est utile dans un cadre de recherche, de recoupement et / ou d'analyses de données autour de problématiques précises ou d'un ensemble de données à organiser de la même façon. De plus, ce type de système permet un usage concerté de données externes et internes au sein d'un organisme public ou d'une entreprise privée.

4 Bibliographie et médiagraphie

Références

[1] I.A. Chen et V.M. Markowitz : An overview of the Object-Protocol Model (OPM) and OPM data management tools. *Information Systems*, 20(5):393–418, 1995.

[2] Su Yun Chung et Limsoon Wong : Kleisli : a new tool for data integration in biology. *Trends biotechnologies*, 17(9):351–355, September 1999.

[3] Loren Donelson, Peter Tarczy-Hornoch, Peter Mork, Cindy Dolan, Joyce A. Mitchell, M. Barriera et Hao Mei : The BioMediator system as a data integration tool to answer diverse biologic queries. *Studies in health technology and informatics*, 107(2):768–772, 2004.

[4] Robert C. Gentleman, Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J. Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y. H. Yang et Jianhua Zhang : Bioconductor : open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, September 2004.

[5] Peter D. Karp, Thomas J. Lee et Valerie Wagner : BioWarehouse : Relational integration of eleven bioinformatics databases and formats. *Data Integration in the Life Sciences*, 5109:5–7, June 2008.

[6] Markus Krummenacker, Suzanne Paley, Lukas Mueller, Thomas Yan et Peter D. Karp : Querying and computing with BioCyc databases. *Bioinformatics*, 21(16):3454–35455, August 2005.

[7] Thomas J. Lee, Yannick Pouliot, Valerie Wagner, Priyanka Gupta, David W. J. Stringer-Calvert, Jessica D. Tenenbaum et Peter D. Karp : BioWarehouse : a bioinformatics database warehouse toolkit. *BMC Bioinformatics*, 7:170, March 2006.

[8] B. Louis, Peter Mork, R. Shaker, N. Kolker et Peter Tarczy-Hornoch : Integration of data for gene annotation using the biomediator system. *AMIA Annual Symposium proceedings*, 2005:1036, 2005.

[9] Hao Mei, Peter Tarczy-Hornoch, Peter Mork, A. J. Rossini, R. Shaker et Loren Donelson : Expression array annotation using the BioMediator biological data integration system and the bioconductor analytic platform. *AMIA Annual Symposium proceedings*, 2003:445–449, 2003.

[10] Robert Stevens, Patricia Baker, Sean Bechhofer, Gary Ng, Alex Jacoby, Norman W. Paton, Carole A. Goble et Andy Brass : TAMBIS : Transparent Access to Multiple Bioinformatics Information Sources. *Bioinformatics*, 16(2):184–185, 2000.