Collecte de données biologiques à partir de sources multiples et hétérogènes. Vers une structure de médiation conviviale et orientée source

Marie-Dominique Devignes, Malika Smaïl, Nacer Boudjlida

Collecte de données biologiques à partir de sources multiples et hétérogènes Vers une structure de médiation conviviale et orientée source

Marie-Dominique Devignes[*,**] **, Malika Smaïl**[*]**, Nacer Boudjlida**[*]

[*]*LORIA, CNRS-UMR 7503 BP 239, F-54506 Vandœuvre-lès- Nancy Cedex*
{devignes, malika}@loria.fr

[**]*Genexpress, CNRS-FRE2376*
BP8, F-94801 Villejuif Cedex

Mots-Clés: sources de données, données génomiques, méta-données, structure de médiation, recherche d'information, qualité des données, authentification des données.

Introduction

La recherche sur les génomes nécessite souvent d'effectuer des requêtes multiples à travers des sources de données hétérogènes accessibles via le web. Les différences rencontrées dans les formats des données, les problèmes de nomenclature et le manque de documentation concernant la validité des données constituent des obstacles certains pour l'unification des données recueillies et la production d'une information utilisable. Cette situation dans laquelle l'utilisateur, dépassé, souhaite utiliser des outils informatiques appropriés pour interroger plusieurs sources s'inscrit dans la problématique du web sémantique. De fait les solutions aux problèmes rencontrés seront à trouver dans une représentation enrichie et structurée des données contenues dans les sources et dans une description également structurée et formalisée des sources elles-mêmes.

Après un rappel des principales spécificités des sources de données biologiques (section 1), nous décrivons brièvement deux études que nous avons effectuées en lien avec deux problèmes biologiques et portant sur la collecte de données selon des scénarios pré-établis d'interrogation d'une succession de sources (sections 2 et 3).

Nous évoquons ensuite, en section 4, les grandes lignes du projet de recherche qui fait suite et qui se veut être une généralisation des travaux précédents afin d'aider le biologiste, face à un problème assez complexe de collecte de données, à identifier des sources pertinentes, le guider dans la définition du scénario de collecte et mettre en œuvre ce scénario sur des données réelles. Nous explicitons dans la section 5 les points d'ancrage de ce projet dans la problématique du web sémantique ainsi que le premier domaine d'application envisagé.

1- Spécificités des sources de données biologiques et de leur intégration

Dans le domaine biologique l'accès à une multitude de sources de données est offerte à l'utilisateur à travers le web. Le manque de documentation sur les sources et leur multiplicité constituent souvent un handicap pour trouver où adresser au mieux les requêtes. Cependant cette multiplicité est un avantage quand il s'agit de compléter ou de vérifier une information, sachant que des données contradictoires peuvent exister dans des sources concurrentes.

Développées au gré des besoins et des expérimentations, les sources biologiques présentent des structures (à défaut de schémas) souvent hétérogènes. Ceci se traduit par exemple par des nomenclatures différentes pour le même type de données, liées à une absence de concertation entre

les différents organismes administrant les sources. Ces structures sont évolutives, en fonction de l'évolution des technologies. Elles peuvent également présenter des irrégularités (éléments incomplets, éléments annotés, typage variable de certains éléments). A ces problèmes s'ajoute celui de la maintenance des sources, dont le niveau, variable, conduit à des degrés de fiabilité différents d'une source à l'autre.

La collecte manuelle de données biologiques se révèle rapidement fastidieuse au vu de la diversité des sources et de la grande variété de situations rencontrées en fonction du problème posé. Une requête complexe doit être décomposée en requêtes élémentaires. Chacune de ces requêtes peut être soumise à une ou plusieurs sources de données. Les données collectées doivent ensuite être intégrées pour constituer une réponse complète à la requête de départ. De plus, la mise à jour fréquente des données dans les sources nécessite de réitérer la recherche en vue d'actualiser la réponse.

Une brève revue des réalisations dans le domaine de la collecte et de l'intégration de données biologiques hétérogènes montre qu'il en existe deux grands groupes (Markowitz, 1995). Dans le premier groupe, les auteurs créent une nouvelle base de données intégrée ou unifiée qui importe selon un schéma propre les données de diverses sources. Dans le deuxième groupe, il s'agit de systèmes multi-bases dans lesquels l'autonomie des bases est préservée, le couplage se situant au niveau de l'accès à chacune des ressources.

De nombreux exemples de bases de données intégrées existent tels que la GDB (Genome DataBase ; http://gdb.wehi.edu.au/gdb/) et Genecards™ (http://bioinformatics.weizmann.ac.il/cards/index.html). Cette stratégie peut aussi se décliner sous le mode de l'entrepôt de données qui permet de soumettre les données importées à des traitements et analyses. C'est le cas des ressources InterPro (Apweiler et al, 2001) ou BioMolQuest (Bukhman et al, 2001). La stratégie des bases de données intégrées a pour avantage la robustesse et la rapidité d'accès aux données. Les inconvénients liés à cette solution sont la lourdeur de l'investissement et de la maintenance nécessaires et surtout une certaine rigidité par rapport aux sources qui rend difficile l'ajout d'une nouvelle source ou le remplacement d'une source devenue obsolète.

Dans le cas des systèmes multi-bases, il importe de bien distinguer entre la navigation (« browsing ») à travers les sources par le moyen d'hyperliens sur le Web (cas de Swissprot, http://www.expasy.ch ; Bairoch et al, 2000) et l'interrogation (« querying ») couplée de plusieurs sources au moyen d'un langage de requête commun. La recherche par navigation est souvent avantageuse car il est plus facile de reconnaître quelque chose d'intéressant que de le décrire, mais elle peut conduire rapidement aux phénomènes de désorientation et de surcharge cognitive (Martin, 1996). Un système de couplage des bases de données bien connu des biologistes est le système SRS (« Sequence Retrieval System », http://www.infobiogen.fr/srs/ ; Etzold et al, 1996). Ce système permet à partir d'une interface commune de diriger les requêtes sur diverses bases de données indexées par le système et reliées entre elles par des liens faisant également l'objet d'une indexation. C'est à l'utilisateur de choisir la/les ressource(s) qu'il souhaite interroger.

Des travaux de recherche, assez rares du fait des obstacles liés aux spécificités évoquées ci-dessus, s'attaquent à la problématique de l'intégration de données biologiques (Baker et al. 1998 ; Moussouni et al. 1999 ; Eckman et al. 2001 ; Siepel et al. 2001).

2- Xmap : collecte de données liées à un problème de cartographie de gènes et de pathologies associées

Le projet Xmap (http://www.loria.fr/projets/Xmap/Index.htm) constitue la première étude et concerne la recherche de données de cartographie sur le génome humain, recherche dont l'un des objectifs est de corréler la position de nouveaux gènes avec celle des pathologies orphelines (Devignes et al. 2002b). Nous avons proposé un modèle de recherche d'information dans des bases de données hétérogènes accessibles sur le web, qui automatise un scénario pré-établi de navigation dans différentes sources (identifiées au préalable) et effectue la structuration des données récoltées dans le but de faciliter leur exploitation ultérieure. Une DTD XML spécifique a été conçue pour structurer et stocker les données collectées pendant une session de recherche. Pour chaque source interrogée un fichier de configuration comporte les éléments utiles à l'interaction avec cette source :

requête paramétrée, expressions régulières caractérisant le passage recherché dans le document retourné par la source ainsi que son contexte (chaîne de caractères précédant et/ou suivant le passage recherché). L'analyse de ce fichier de configuration permet la génération d'analyseurs lexicaux dont l'enchaînement selon le scénario défini conduit à la production d'un document XML contenant les données collectées lors d'une session. Le prototype Xmap est actuellement en test dans un laboratoire de génomique au CNRS à Villejuif.

Une historisation des différentes sessions dans une base de données relationnelle nous a permis en outre de constituer une sorte d'entrepôt de données (Strohmenger, 2001), lequel pourra faire l'objet de certaines analyses permettant d'exploiter de façon rationnelle la mémoire de sessions. Un accès aux données relationnelles en utilisant le langage SQL permet au biologiste de répondre à des questions ponctuelles (ex : nombre de gènes trouvés co-localisés avec des pathologies orphelines sur un chromosome donné). D'autres analyses de nature statistique seront également utiles afin de réaliser des synthèses sur les données collectées et permettre de répondre à des questions plus larges que le biologiste se pose.

3- Xprom : Application générique pour la collecte de données sur des gènes candidats de pathologies multi-factorielles

Le projet Xprom constitue la deuxième étude et porte sur la recherche de gènes candidats pour des pathologies multi-factorielles (obésité, cancers ..) (Norsa 2002, Devignes et al. 2002a). Cette étude a été l'occasion de construire une application générique capable de mettre en œuvre tout scénario décrit selon un modèle de scénario ; ce modèle étant défini sous forme d'une DTD XML. Selon ce modèle, un scénario est composé d'une suite d'étapes ; une étape est caractérisée par des données d'entrée, l'adresse de la ressource interrogée, la syntaxe de la requête et les données de sortie exprimées sous forme d'expressions régulières caractérisant le passage recherché dans le document retourné par la source. Les données collectées lors de la mise en œuvre d'un scénario particulier sont structurées selon une DTD calquée sur celle d'un scénario. Ces données peuvent ensuite faire l'objet d'une conversion dans une ou plusieurs structures plus signifiantes pour le biologiste, ces structures pouvant être basées sur des ontologies plus ou moins figées du domaine considéré.

4- Vers une structure de médiation conviviale pour des sources de données génomiques hétérogènes

Ces premières études nous ont permis de nous familiariser avec les caractéristiques des bases de données biologiques utilisées en génomique mais aussi d'aborder le problème de la collecte de données à partir de sources identifiées. Il est à présent opportun d'aller plus loin en abordant le problème de l'identification des sources pertinentes et de la définition d'un scénario de collecte à partir d'une question biologique quelconque. Pour cela, nous sommes amenés à généraliser notre approche par l'étude et la définition d'une structure de médiation conviviale pour des sources de données génomiques hétérogènes et accessibles via le web (Boudjlida et al. 2000, Boudjlida 2002).

Cela impliquera :

a. le choix d'un modèle de description dans lequel seront exprimées certaines méta-données sur les différentes sources ; cela nous amène à identifier différentes catégories de connaissances à associer aux sources telles que les connaissances relatives au contenu, à la structure, à la qualité d'une source (Berti 1999), aux modalités d'interaction avec une source, aux relations entre sources …

b. l'élaboration d'une fiche descriptive relative à chaque source puis la synthèse des fiches descriptives en une sorte d'annuaire des sources ;

c. la classification des sources en fonction de leur contenu, de leurs relations mais aussi de connaissances ontologiques spécifiques au(x) domaine(s) biologique(s) considéré(s) ;

d. l'élaboration d'une interface de navigation et d'interrogation de la fédération des sources. La classification et l'annuaire constituent une sorte de *carte spatio-sémantique* des sources et servent de pivot au processus d'interrogation et de navigation dans les sources. Cette interface permettra à un utilisateur biologiste d'interroger l'ensemble des bases de données "comme un tout". Le processus d'interrogation prendra en charge la sélection des sources pertinentes pour la satisfaction d'une requête utilisateur.

e. Lorsque l'utilisateur est confronté à un problème complexe nécessitant d'enchaîner l'interrogation de plusieurs sources, il est envisagé d'offrir une assistance pour la formalisation du scénario de collecte de donnée selon un modèle générique et de mettre à disposition un générateur automatique de collecteur de données dans un ensemble de sources hétérogènes selon un scénario défini (cf. section 3).

5- Bilan sur le projet proposé

L'approche que nous nous proposons d'adopter pour l'intégration des données génomiques est du type *Local As View* (Levy 1999, Halevy 2001), *i.e.*, chaque source sera décrite comme une vue sur le schéma du médiateur. Il nous reste, entre autres, à définir un langage de requête sur ce même schéma médiateur puis un mécanisme de reformulation d'une requête gloable en une succession de requêtes adressées aux différentes sources pertinentes ainsi qu'un mécanisme de définition de wrappers associés aux différentes sources. Un aspect intéressant du problème dans le domaine de la génomique concerne l'intégration de données homologues retournées par plusieurs sources avec gestion des contradictions et des redondances. Cette intégration devra exploiter la qualité relative des données et des sources dont elles sont issues afin de trier ces données et, dans la mesure du possible, aider à l'authentification de certaines données.

Les avantages majeurs de cette approche orientée source pour la médiation sont l'ouverture et le pouvoir d'expression. Elle permet, en effet, l'intégration de nouvelles sources à chaque fois que cela s'avère nécessaire et il est aisé d'attacher des contraintes, assez riches, relatives au contenu des sources ; ce qui est utile dans le cas de sources proches ou recouvrantes.

La validation de ces propositions, après la spécifiation d'une solution, pourra être appliqué aux travaux actuels sur l'annotation des génomes. Des banques de données primaires (telles que GenBank, PIR, Swissprot), ou secondaires, c'est-à-dire dérivées des premières (telles que Prosite, BLOCKS), des bases de données spécifiques de tel ou tel organisme (comme FlyBase pour la drosophile) ou encore des ressources unifiées (telles que UDB ou GeneCard du Weizmann Institute) pourront être considérées comme un premier jeu de sources de données à décrire et à classifier en vue de répondre à des questions d'annotation sur les gènes : que sait-on sur la fonction de tel gène ? quel gène correspond à tel gène dans tel autre organisme ?…

La complexité du domaine biologique rend nécessaire une approche ciblée mais représentative des types de difficultés rencontrées afin de parvenir à confronter nos propositions avec des solutions opérationnelles.

Références bibliographiques

Abiteboul S., Buneman P., « Data on the Web; From Relations to Semi-Structured Data and XML», Morgan Kaufmann Publishers, San Francisco (CA), 2000.

Apweiler R., Attwood T.K., Bairoch A., Bateman A., Birney E., Biswas M., Bucher P., Cerutti L., Corpet F. et al., «The InterPro database, an integrated documentation resource for protein families, domains and functional sites», *Nucleic Acids Research* 29:37-40, 2001.

Baker P.G., Brass A., Bechhofer S., Goble C.A., Paton N.W. , Stevens S., «TAMBIS – Transparent access to multiple biological information sources», *Proceedings of ISMB'98,* AAAI Press, p. 25-34, 1998.

Bairoch A. , Apweiler R., «The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000», *Nucleic Acids Research* 28:45-48, 2000.

Berti L., « Qualité de données multi-sources et recommandation multi-critère », Proceedings Inforsid 99, La Garde, p. 185-204, 1999.

Boudjlida N., Devignes M-D., Smaïl M., «Services for a Genomic Open Distributed Environment». Position Paper, XEWA Workshop, League City, TX, Décembre 2000.

Boudjlida. N., «A Mediator-Based Architecture for Capability Management». To appear in proceedings of IASTED Intern'l Conference on Software Engineering and Applications, SEA'20002, Cambridge, USA, November 2002.

Bukhman Y. and Skolnick J. , «BioMolQuest : integrated database-based retrieval of protein structural and functional information », *Bioinformatics* 17:468-478, 2001.

Devignes M-D., Schaaff A., Smaïl M., « Collecte et Intégration de données biologiques hétérogènes sur le web – Xmap, application dans le domaine de la cartographie du génome humain », Ingénierie et Systèmes d'Information, à paraître, 2002.

Devignes M-D., Norsa Y., Smaïl M., Collet P., Domendjoud L., «From putative ¨promoter sequence to genomic context : biological data collection on the web using a generic application (Xprom) », *Proceedings of First European Conference on Computational Biology (poster),* 2002.

Eckman B.A., Kosky A.S. , Laroco Jr L.A., «Extending traditional query-based integration approaches for functional characterization of post-genomic data», *Bioinformatics* 17:587-601, 2001.

Etzold T., Ulyanov A. , Argos P. , «SRS: information retrieval system for molecular biology data banks», *Methods in Enzymology*. 266:114-28, 1996. http://srs.ebi.ac.uk

Halevy A.Y., «Answering Queries using Views : a survey», *VLDB Journal*, 10:4270-294, 2001.

Levy A.Y., «Combining AI and DB for data integration», *Lecture Notes in AI "AI today : recent trends and developments"*, 1999.

Markowitz V.M., «Heterogeneous molecular biology databases», *J. Computational Biology* 2:537-538, 1995.

Martin P., « Exploitation de graphes conceptuels et de documents structurés et hypertextes pour l'acquisition de connaissances et la recherche d'informations », Thèse de doctorat de l'université de Nice - Sophia Antipolis, 1996.

Moussouni F., Paton N.W., Hayes A., et al., « Database Challenges for Genome Information in the Post-Sequencing Phase » , Proceedings of DEXA 1999, LNCS 1677, p. 540-549, 1999.

Norsa Y., «Réalisation d'une application générique de collecte de données» , rapport de licence professionnelle, Université de Nancy 2, 2002.

Siepel A., Farmer A., Tolopko A., Zhuang M., Mendes P., Beavis W., Sobral B. , «ISYS : a decentralized component-based approach to the integration of heterogeneous bioinformatics resources», *Bioinformatics* 17:83-94, 2001.

Strohmenger V., « Xmap-DB : une base de données relationnelle de sessions de recherche d'informations génomiques », Stage du DESS EGOIST, Université de Rouen, 2001.