Extraction de données pharmacogénomiques à partir d'études cliniques : problématique

Adrien Coulet, Marie-Dominique Devignes, Malika Smaïl-Tabbone

Extraction de connaissances pharmacogénomiques à partir d'études cliniques : problématique

Adrien Coulet*, Marie Dominique Devignes, Malika Smaïl
LORIA, INRIA, Université Henri Poincare-Nancy 1, 54 506 Vandoeuvre-les-Nancy

Résumé

L'importance des variations individuelles dans les réactions aux médicaments devient un problème conséquent à la fois au niveau de la recherche pharmaceutique et au niveau médical. Notre projet de recherche vise à intégrer des données cliniques et génétiques issues d'études cliniques avec comme objectif d'en extraire une connaissance sur les relations existantes entre un génotype particulier et son action sur l'effet d'un médicament. Pour répondre à ce problème, nous cherchons des méthodes de fouille adaptées aux données biomédicales que nous souhaitons manipuler et capables d'intégrer les connaissances du domaine sous forme d'ontologie. Ce projet est l'objet d'une thèse qui a commencé en novembre 2004.

1. Introduction

Chaque être humain présente une version différente du génome. Ces différences résident principalement dans des variations ponctuelles (une paire de bases) disséminées le long des chromosomes : les "Single-Nucleotide Polymorphism" (SNP) [1]. Ces variations individuelles peuvent s'exprimer, entre autres, par des réactions différentes, selon les individus, à un même médicament. Etudier comment des variations dans les gènes humains entraînent des effets différents pour un médicament est l'objet de la pharmacogénomique [2].

2. Objectifs et problématique

Le projet dans lequel nous nous inscrivons consiste en la recherche de méthodologies adaptées au développement d'un système, qui permettrait l'étude quantitative et qualitative des relations entre traitement médical (données cliniques), évolution de la maladie et facteurs génétiques.

Un tel système et les connaissances auxquelles il donnerait accès trouvent, entre autres, des applications dans (1) les études cliniques, (2) la prescription médicale, (3) les études épidémiologiques et (4) les sciences fondamentales [2, 3].

De plus en plus de laboratoires pharmaceutiques intègrent des données génétiques dans les protocoles de leurs études cliniques. Ceci leur apporte une source d'information supplémentaire pour interpréter les éventuels effets indésirables d'un médicament sur un groupe d'individus.

*contact : adrien.coulet@loria.fr

Une meilleure connaissance des propriétés pharmacogénomiques d'une molécule devrait leur permettre de diminuer la probabilité qu'un nouveau médicament soit retiré du marché à moyen terme. Aussi, les dernières étapes des études pourraient se faire sur des populations plus réduites car génétiquement ciblées. Cependant, l'enjeu majeur pour ces industries reste le passage d'une recherche de molécules à large spectre, comme c'est le cas actuellement, au développement d'une médecine individualisée dans laquelle, pour une pathologie unique, deux individus recevraient deux médicaments différents [2, 4].

La base de connaissance gérant ces notions pourrait servir d'aide à la prescription médicale. En effet, si un patient nécessite un traitement dont les effets peuvent être dangereux s'il possède la mauvaise "version" d'un gène, son médecin pourra lui prescrire un diagnostic moléculaire (le séquençage d'une région d'un chromosome) qui déterminera si, oui ou non, il présente un polymorphisme à risque pour suivre le traitement en question. Par l'exploration locale du génome de son patient le médecin évitera une réaction intempestive à la thérapeutique et par-là participera à la diminution du nombre d'accident médicamenteux [5, 6].

Un tel système présente également un intérêt épidémiologique : il permettrait de consulter une somme d'information conséquente en proposant de recouper les résultats de plusieurs études cliniques. La somme d'information disponible pourrait par exemple permettre de constater qu'une réaction à un médicament est caractéristique d'un type de population (groupe ethnique) [7].

A un niveau plus global, le recoupement d'informations cliniques peut amener des connaissances plus fondamentales sur les relations génotype-phénotype. Déterminer qu'un gène est impliqué dans un réseau métabolique peut motiver une étude approfondie à son sujet et en faire une cible pour de futurs médicaments [3, 8].

La problématique principale du projet réside dans la détermination de relations entre un polymorphisme génétique et une réaction particulière à un médicament [9]. Ces relations sont complexes et en conséquence, difficiles à établir. En effet, elles représentent des phénomènes biologiques qui s'expriment à un niveau moléculaire. Les facteurs impliqués (gènes, protéines, principe actif d'un médicament…) interagissent simultanément avec divers mécanismes biologiques et leurs rôles varient en fonction du temps, du lieu, de l'individu — ce qui rend difficile le discernement de relations de cause à effet. Les informations cliniques et génétiques, mises à disposition à l'issue d'une étude clinique peuvent être enrichies par des banques de données biologiques externes (dbSNP, OMIM, KEGG, PubMed…). Il est difficile d'établir des relations systématiques entre toutes les données que nous avons à manipuler qui, en plus d'être de natures différentes, proviennent parfois de sources différentes. C'est pour cela qu'il est indispensable d'avoir recours à une méthodologie rigoureuse pour structurer convenablement les données en vue de la fouille [10, 11].

3. Caractéristiques des données

Les données à intégrer en vue de la fouille sont, de par leur nature particulière, des données complexes à manipuler [11].

3.1 Le volume des données

Les études cliniques se prêtent aux études statistiques d'autant mieux que les données

qu'elles comprennent sont nombreuses. Ces données peuvent, de plus, comprendre des fichiers volumineux (images médicales, signaux biologiques…). Indépendamment, les banques de données externes, connexes aux données génétiques des sujets étudiés sont des sources d'informations très riches dont la taille croit de manière exponentielle. En résumé, la quantité très importante des données à traiter est un critère à considérer dans le choix des méthodes d'analyse [10].

3.2 La qualité des données

Les données brutes sont de qualité très variable selon leur source : banques de données biologiques externes validées à la main ou non, mises à jour plus ou moins fréquemment, banques de données cliniques issues d'études plus ou moins précises (données manquantes ou insuffisantes…). Cette disparité est problématique pour l'intégration des données en un ensemble de qualité homogène.

3.3 La localisation des données

En général, les données cliniques sont locales, en revanche, les données biologiques, la littérature scientifique, les bases de connaissances pharmacogénomiques sont dans des banques de données externes, publiques et disponibles librement à des localisations éparses sur internet. Il peut aussi s'avérer intéressant de recouper le contenu de plusieurs études cliniques dont les sources peuvent être éloignées physiquement.

3.4 L'hétérogénéité des données

La nature de ces données est aussi très variable. Les données des études cliniques contiennent des informations générales sur le patient (âge, poids, taille, régime, pathologie…) ainsi que des données plus précises résultant de l'exploration biologique (dosages d'enzyme, génotype…). Aussi, on peut avoir une information disponible directement (un diagnostic, un chiffre significatif) ou une information qu'il faudra traiter préalablement avant d'en extraire une connaissance (image médicale, signal biologique). Le contenu des banques de données externes est également très hétéroclite puisque celles-ci peuvent renfermer, par exemple, une information textuelle (MEDLINE) [12], des séquences génomiques (GenBank), des voies métaboliques (KEGG).

3.5 L'évolutivité des données et les données temporelles

C'est une caractéristique majeure des données à intégrer. Les données des banques de données externes sont mises à jour régulièrement. Les données cliniques sur les patients d'une étude possèdent une dimension temporelle pour rendre compte de l'évolution de la pathologie et de l'effet du traitement au cours du temps.

Toutes ces caractéristiques propres aux données biomédicales que nous manipulerons en font des données complexes. Pour en extraire la connaissance qui nous intéresse, il nous faudra

trouver les méthodes de fouille de données complexes les plus appropriées à notre problématique.

4. Choix préliminaires et discussion

Ce projet scientifique fait l'objet d'une thèse qui a débuté en novembre 2004. Dans le cadre de cette thèse , les premiers éléments méthodologiques que nous avons identifiés sont les suivants.

La complexité des données et de leurs relations obligent à une méthodologie stricte. Ainsi, pour commencer à maîtriser les concepts à représenter, il est indispensable de structurer et d'organiser ces données [13, 14]. Une fois structurées, les données pourront être exploitées à partir d'une architecture basée sur un entrepôt de données [15, 16]. Ce dernier offrira la possibilité d'intégrer nos données complexes [17] et ainsi de leur appliquer des méthodes de fouille de données pour en dégager certains éléments de connaissance [18, 19]. Le processus de fouille pourra exploiter des ontologies de référence comme modèle du domaine [10, 11, 20, 21, 22].

La mise en œuvre de cette méthodologie commencera par des études de cas basées sur des scénarios précis avec comme objectif l'aide à la décision. L'analyse de ces premiers cas concrets nous aidera à formaliser les données en prenant en compte les ontologies du domaine. Ce travail permettra de mettre en œuvre la fouille qui, à son tour, produira de nouvelles connaissances qui viendront enrichir les ontologies du domaine. Dans ce contexte de fouille guidée par les connaissances du domaine, nous pourrons entre autres nous aider de la base de connaissances pharmacogénomiques PharmGKB (www.pharmGKB.org) [23]. PharmGKB est une ressource publique qui intègre des informations cliniques et génétiques en vue d'aider à comprendre comment les variations du génome humain contribuent aux variations de réactions à un même médicament.

Références

[1] TAILLON-MILLER, P., PIERNOT, E.E. and KWOK, P.Y. (1999): *Efficient approach to unique single-nucleotide polymorphism discovery.* Genome Res., 9(5), pp. 499-505.

[2] EVANS, W.E. and RELLING, M.V. (2004): *Moving towards individualized medicine with pharmacogenomics*. Nature, 429(6990), pp. 464-8.

[3] PIRAZZOLI, A. and R ECCHIA G. (2004): *Pharmacogenetics and pharmacogenomics: are they still promising?* Pharmacol Res., 49(4), pp. 357-61.

[4] RIOUX, P.P. (2000): *Clinical trials in pharmacogenetics and pharmacogenomics: methods and applications.* Am J Health Syst Pharm., 57(9), pp. 887-98.

[5] SEVERINO, G., DEL ZOMPO M. (2004): *Adverse drug reactions: role of pharmacogenomics.* Pharmacol Res., 49(4), pp. 363-73.

[6] ITO, R.K. and DEMERS, L.M. (2004): *Pharmacogenomics and pharmacogenetics: future role of molecular diagnostics in the clinical diagnostic laboratory*. Clin Chem., 50(9), pp. 1526-7.

[7] SANDERS, R. (1999) *Mining the Swedish clinical archives to develop pharmacogenomic tests.*Mol Diagn., 4(4), pp. 319-25.

[8] LINDPAINTNER, K. (2003): *Pharmacogenetics and pharmacogenomics in drug discovery and development: an overview.* Clin Chem Lab Med., 41(4), pp. 398-410.

[9] ALTMAN, R.B. (2003): *Genetic sequence data for pharmacogenomics.* Curr Opin Drug Discov Devel., 6(3), pp. 297-303.

[10] ALTMAN, R.B. and KLEIN, T.E. (2002): *Challenges for biomedical informatics and pharmacogenomics.* Annu Rev Pharmacol Toxicol., 42, pp. 113-33.

[11] OLIVER, D.E., RUBIN, D.L., STUART, J.M., HEWETT, M., KLEIN, T.E. and ALTMAN, R.B. (2002): *Ontology development for a pharmacogenetics knowledge base.* Pac Symp Biocomput., pp.65-76.

[12] RUBIN, D.L., CARRILLO, M., WOON, M., CONROY, J., KLEIN, T.E. and ALTMAN, R.B. (2004): *A resource to acquire and summarize pharmacogenetics knowledge in the literature.* Medinfo., 2004, pp. 793-7.

[13] XU, C., WINDEMUTH, A., LAN, Z.H., DUAN, D., WAN, C., HAWKES, K., ZHAI, J., N ALABALOU, S., ZHANG, S., WEI, M., J IANG, R. and JUDSON, R. (2002): *Application of Relational Database Management System and Object-Oriented Programming in Pharmacogenomics Based Drug Development Process.* World Multiconference on Systemics, Cybernetics and Informatics, 13, pp. 574-9.

[14] NADKARNI, P., SUN, K. and WIEPERT, M. (2002): *Designing and implementing special-purpose databases: lessons from the pharmacogenetic network.* Pharmacogenomics, 3(5), pp. 687-96.

[15] THALLINGER, G.G., TRAJANOSKI, S., STOCKER G. and TRAJANOSKI Z. (2002): *Information management systems for pharmacogenomics.* Pharmacogenomics, 3(5), pp. 651-67.

[16] HAN, J. (1997): *OLAP Mining : an Integration of OLAP with Data Mining*. In Proc. IFIP conference on Data Semantics (DS-97), Leysin, Switzerland, pp. 1-11.

[17] LEVY A.Y. (2000): *Logic-Based Techniques in Data Integration*. In Logic-Based Artificial Intelligence, Jack Minker (ed.): *Kluwer*.

[18] ROBSON, B.and MUSHLIN, R. (2003): *Clinical and pharmacogenomic data mining: 1. Generalized theory of expected information and application to the development of tools.* J Proteome Res., 2(3), pp. 283-302.

[19] ROBSON, B. (2004): *Clinical and pharmacogenomic data mining: 2. A simple method for the combination of information from associations and multivariances to facilitate analysis, decision, and design in clinical research and practice.* J Proteome Res., 3(4), pp. 697-711.

[20] RUBIN, D.L., SHAFA, F., OLIVER, D.E., HEWETT, M. and ALTMAN, R.B. (2002): *Representing genetic sequence data for pharmacogenomics: an evolutionary approach using ontological and relational models.*Bioinformatics. 18 Suppl 1, pp. S207-15.

[21] RUBIN, D.L., HEWETT, M., OLIVER, D.E., KLEIN, T.E., ALTMAN, R.B. (2002): *Automating data acquisition into ontologies from pharmacogenetics relational data sources using declarative object definitions and XML.* Pac Symp Biocomput., pp. 88-99.

[22] STEVENS, R., Wroe C., LORD, P. and GOBLE, C. (2004): *Ontologies in Bioinformatics.* In Handbook on Ontologies, S. Staab, R. Studer (eds.): *Springer*.

[23] HEWETT, M., OLIVER, D.E., RUBIN, D.L., EASTON, K.L., STUART, J.M., ALTMAN, R.B. and KLEIN, T.E. (2002): *PharmGKB: the Pharmacogenetics Knowledge Base.* Nucleic Acids Res., 30(1), pp. 163-5.