An Agent-Based Enrichment System for Genetic Diversity Analyses

Giordano B. Soares-Souza[1], Guilherme P. G. Kingma[2], Eduardo Tarazona-Santos[1], Maíra R. Rodrigues[1]

[1]Instituto de Ciências Biológicas – Universidade Federal de Minas Gerais (UFMG)

[2]Instituto de Ciências Exatas e Informática – Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte – MG – Brazil

{jwojwo,guilherme.kingma}@gmail.com, {edutars,maira}@icb.ufmg.br

Abstract. *Different evolutionary forces produce genetic variants among individuals and populations. One of the most frequently studied variants is the Single Nucleotide Polymorphism (SNP). In some cases, population-specific genetic variants might be associated to diseases, resulting in a difference in the frequency of a disease among human populations. To achieve relevant genetic and biomedical knowledge, these variants need to be associated with other types of biological data, such as biochemical pathways, pharmacogenetics, and genome-wide associations. However, these data are fragmented in different databases. Therefore, it becomes necessary to develop tools to integrate distinct sources of biological data in order to enrich genetic diversity analyses. In this context, we propose an Agent-based enrichment system to integrate different sources of biologically relevant data in order to improve the genetic diversity analysis of native and Latin-American populations in particular.*

Resumo. *Diferentes fatores evolutivos produzem variantes genéticas entre indivíduos e populações. Uma das variantes mais estudadas é o Polimorfismo de Base Única (SNP). Em alguns casos, variantes específicas de determinadas populações podem estar associadas a doenças, resultando na diferença de frequência de uma doença entre populações humanas. Para alcançarem conhecimento genético e biomédico relevantes, essas variantes precisam ser associadas a outros tipos de informações biológicas, tais como vias bioquímicas, farmacogenética e estudos de associação por varredura genômica. No entanto, tais dados encontram-se fragmentados em diferentes bases de dados na Web. Com isso, torna-se necessário o desenvolvimento de ferramentas de integração de fontes distintas de dados biológicos, de forma a enriquecer as análises de diversidade genética. Nesse contexto, é proposto um sistema baseado em Agentes para enriquecimento de dados através da integração de diferentes fontes de dados biológicos relevantes, visando, em particular, melhorar as análises de diversidade genética de populações nativas e latino-americanas.*

1. Introduction

Although our genetic code is highly similar, different evolutionary forces produce genetic variants among individuals and populations. One of the most frequently studied variants is the Single Nucleotide Polymorphism (SNP). These polymorphisms are distributed across

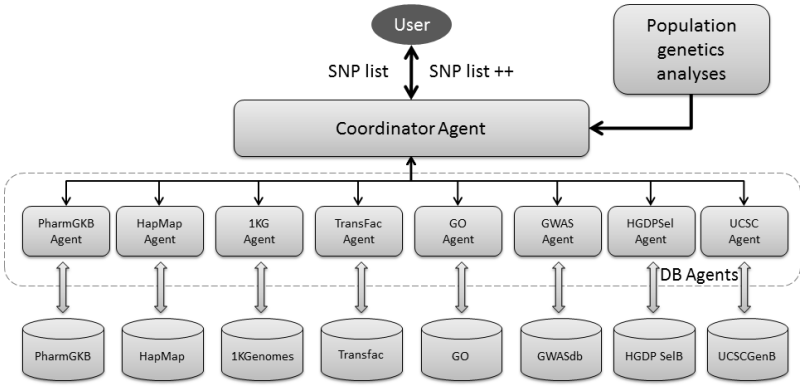

Figure 1. Multiagent enrichment system's overview.

the genome and can influence the expression of genes and, consequently, protein function. This influence is one of the causes of the different phenotypes in individuals from distinct populations, such as skin colour and hair texture. However, they sometimes lead to the loss of protein function and may cause the development of a disease. In addition to that, population-specific genetic variants might be associated to diseases or specific health-related conditions, leading to a difference in the frequency of a disease among human populations. This is the case, for example, of lactose intolerance, a condition that has higher occurrence in individuals from African and Asian populations. Therefore, the identification of variants with significant difference in frequency from one population to another is one of the main goals of genetic research and have a major biomedical interest [Myles et al. 2008].

In the field of genetic research, population genetics studies, in particular, have been used to explain human genetic diversity patterns in terms of population history (e.g., intercontinental migration history) and to understand the genetic bases of phenotypic adaptation (e.g., the genetic bases behind altitude adaptation of some Andine populations). Therefore, data generated by population genetics studies are the starting point to genetic diversity analyses of complex diseases. Some large-scale studies in this area include the HapMap project, SNP500Cancer and the on-going 1000 genomes project.

Despite the great amount of genetic diversity data available, the study of variants with biomedical interest requires their association with other types of biological data, such as biochemical pathways, pharmacogenetics, genome-wide association, and so on. Today, these heterogeneous data are fragmented in different databases. Therefore, it becomes necessary to develop tools to integrate different sources of biological data in order to enrich genetic diversity analyses and to move forward with genetic studies of particular diseases or populations. The agent-oriented paradigm has been advocated as a natural way to design and implement systems that are distributed and heterogeneous [Foster et al. 2004]. Therefore, given the heterogeneity of different types of genetic and biological data and their distribution over distinct sources over the Web, the development of a genetic analyses enrichment system points to an agent-based approach.

Moreover, there are several types of genetic diversity studies with requirements that vary from simple to more complex information integration processes. Therefore, each study will require a different enrichment process and, thus, a different combination

Database	Biological Information	Interface
PharmGKB	Diseases, Pharmacogenetics, Pathways	API to local scripts
HapMap	Populational genetic variation, Genomic annotation	Web service
1000Genomes	Populational genetic variation, Structural variation, Genomic annotation	Web service
Transfac	Transcription factors	API to local scripts, mySQL connection
Gene Ontolody	Biological categories, Term enrichment	API to local scripts, mySQL connection
GWASdb	Gene expression, Association studies, Evolution, AA substitutions, Genomic mapping	Web service
HGDP	iHS, XP-EHH, Fst	Web Service
UCSC	Nucleotide conservation	mySQL connection, DAS server, Local database

Table 1. Databases with relevant biological information.

of sources of information to be integrated. This requires that the computational entities in the system have the ability to make decisions about which process to follow, and to communicate in order to coordinate the enrichment process.

In this context, we propose an Agent-based enrichment system to integrate different sources of biologically relevant data in order to improve the genetic diversity analysis of Native and Latin-American populations, in particular, and to facilitate the understanding of the effects of these diversity patterns in the development of complex diseases in Amerindian populations.

2. MAS Architecture Overview

The user interacts with the Multiagent system by selecting one of the pre-defined genetic analyses enrichments. These enrichments vary in complexity. For example, simple analyses require the access to only one or two databases. This is the case of the basic genetic variation enrichment, which consists in finding information such as the gene, chromosomic region, chromosomic coordinate, and strand of a given SNP or list of SNPs, which can all be found in the dbSNP database. Complex enrichment analyses involve several databases. This is the case of false-positive enrichment analysis for case-control association studies, which requires the retrieval of information such as population variation and molecular function and localization of a SNP, its possible effects on gene expression and regulation, and its association to diseases, phenotypes and metabolic pathways.

The basic architecture of the Multiagente system for genetic analyses enrichment is shown in Figure 1. It is composed of two types of agents: a Coordinator agent and Database (DB) agents. The Coordinator agent has knowledge about a pre-defined number of population genetics and genetic epidemiology analyses. Each analysis requires specific types of biological data. According to a selected analysis, the Coordinator agent interacts with the corresponding DB agents to request information about a variant. Once all information is returned, the Coordinator agent combines them in a report. Agent communication will be supported by specific biological ontologies, such as the SNP-Ontology and Gene Ontology.

Currently, there are eight (8) DB agents, one for each biological database identified with relevant information for the genetic diversity analyses of Native and Latin-American populations. This is necessary because the selected databases are very heterogeneous in terms of both the type of biological information and their data access interface, as shown in Table 1. Therefore, each DB agent has the knowledge of the data access mode and data format accepted and returned by the database that it encapsulates.

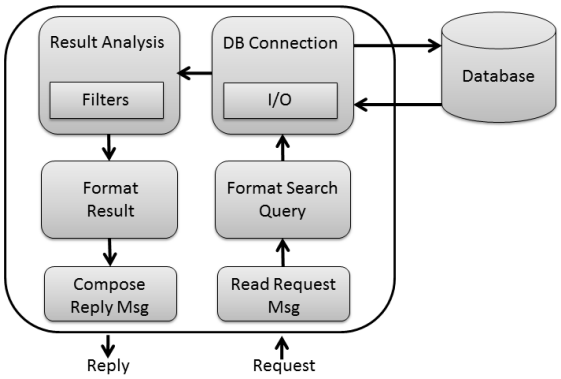

Figure 2. Internal architecture of a DB agent.

In order to search a biological database for specific information, a DB agent has the internal architecture shown in Figure 2. The `Read Requested Msg` module recognised the information that is being requested as compatible for search or not, depending on its knowledge of the database accepted biological data types. If the information is compatible, it is passed on to the `Format Search Query` module, which formats the query according to the type of data access interface and accepted input data. After that, the `DB Connection` module establishes the connection with the database in order to send the search query. Once the result arrives, it is passed on to the `Result Analysis` module, which filters out undesirable data or data with low statistical confidence or low quality values. Finally, the search results that passed the filtering process are formatted appropriately (module `Format Result`) and sent back to the Coordinator agent (`Compose Reply Msg` module). If the search query has a timeout, returns an error or no results, an error message is send to the Coordinator agent instead.

3. Conclusions

We propose an Agent-based enrichment system to improve genetic diversity analyses through the integration of different sources of biologically relevant data. This system will be used to analyse genetic diversity in Native and Latin-American populations and will be available to analyse variant data from other populations after the former is concluded. We believe that an agent-based approach is suited to implement such enrichment system given the heterogeneity of different types of genetic and biological data and their distribution over distinct sources over the Web. Moreover, Multiagent systems can cope with dynamic applications and, thus, can easily accommodate the incorporation of new data sources for complementary biological data that might be available in the future.

References

Foster, I., Jennings, N., and Kesselman, C. (2004). Brain meets brawn: Why grid and agents need each other. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 8–15. IEEE Computer Society.

Myles, S., Tang, K., Somel, M., Green, R., Kelso, J., and Stoneking, M. (2008). Identification and analysis of genomic regions with large between-population differentiation in humans. *Ann Hum Genet*, 72(Pt 1):99–100.