BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 2 2009, pages 230–236
doi:10.1093/bioinformatics/btn612

Genetics and population analysis

Gene–disease relationship discovery based on model-driven data integration and database view definition

S. Yilmaz[1], P. Jonveaux[1], C. Bicep[2], L. Pierron[2], M. Smaïl-Tabbone[2] and M. D. Devignes[2,*]

[1]Laboratory for Human Genetics, Nancy Medical Faculty, rue du Morvan, 54500 Vandoeuvre-les-Nancy cedex and [2]LORIA UMR7503, CNRS, INRIA, Nancy-Université, BP239, 54506 Vandoeuvre-les-Nancy cedex, France

Received on July 1, 2008; revised on November 20, 2008; accepted on November 21, 2008

Advance Access publication November 27, 2008

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Computational methods are widely used to discover gene–disease relationships hidden in vast masses of available genomic and post-genomic data. In most current methods, a similarity measure is calculated between gene annotations and known disease genes or disease descriptions. However, more explicit gene–disease relationships are required for better insights into the molecular bases of diseases, especially for complex multi-gene diseases.

Results: Explicit relationships between genes and diseases are formulated as candidate gene definitions that may include intermediary genes, e.g. orthologous or interacting genes. These definitions guide data modelling in our database approach for gene–disease relationship discovery and are expressed as views which ultimately lead to the retrieval of documented sets of candidate genes. A system called ACGR (Approach for Candidate Gene Retrieval) has been implemented and tested with three case studies including a rare orphan gene disease.

Availability: The ACGR sources are freely available at http://bioinfo.loria.fr/projects/acgr/acgr-software/. See especially the file 'disease_description' and the folders 'Xcollect_scenarios' and 'ACGR_views'.

Contact: devignes@loria.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Understanding the molecular basis of a disease ultimately means correlating disease symptoms with altered gene function(s) thus highlighting gene–disease relationships. Identifying the genes responsible for human diseases is a first step towards this goal. More than 6100 disease phenotypes are described in the OMIM (Online Mendelian Inheritance in Man) database (DB). Among these phenotypes, more than 2400 have at least one known molecular basis (entries prefixed with #). Thus, about 3700 disease phenotypes described in the OMIM DB are not yet associated with any responsible gene. These disease phenotypes are particularly challenging since they include rare syndromes for which limited experimental data are available and complex multi-genic disorders involving various causative and susceptibility genes (Botstein and Risch, 2003).

Integrative genomics approaches are becoming indispensable tools for discovering new gene–disease relationships. These approaches rely on efficient exploitation of functional genomics data sources (Giallourakis *et al*., 2005) and take advantage of numerous computer-based systems that have been developed in the last 5 years. These systems can be classified into three main groups. First, generalist systems predict disease genes based on their properties or interactions (Adie *et al*., 2005; Calvo *et al*., 2007; Lopez-Bigas and Ouzounis, 2004; Lopez-Bigas *et al*., 2006; Oti *et al*., 2006; Tu *et al*., 2006; Xu and Li, 2006). Consistent features are thus detected among approximately 1600 disease genes listed in the OMIM morbid map and used for these studies. Indeed, disease genes tend to be longer, are composed of more exons, show a higher degree of interspecies conservation, and are involved in more interactions than other genes. However, these approaches are unable to establish the correspondence between a given disease and a set of genes.

The second group of systems apply strategies relying on the hypothesis that similar diseases are most likely caused by similar genes. These strategies are often called prioritization methods since they aim to rank a given list of genes with respect to their probability to cause a disease (Adie *et al*., 2006; Aerts *et al*., 2006; Freudenberg and Propping, 2002; George *et al*., 2006; Perez-Iratxeta *et al*., 2002, 2005; Rossi *et al*., 2006; Turner *et al*., 2003). Additionally, alternative strategies based on the same similarity hypothesis aim to characterize user-defined groups of genes (Barillot *et al*., 2004; Chiang *et al*., 2006; Masseroli *et al*., 2004, 2005; Sun *et al*., 2006). In order to find additional responsible genes, prioritization methods are often applied to a single disease whose associated chromosomal loci are known. A pool of statistical methods is then used to compute similarity measures dealing with various gene features. Such gene features are particularly well covered in the endeavour system (Aerts *et al*., 2006), e.g. sequence similarity, domain composition, tissue expression, Gene Ontology (GO) annotation, interspecies conservation, protein–protein interactions, involved pathways and *cis*-regulatory elements. However, this type of prioritization strategy requires at least one well-known gene to be used as a reference candidate gene.

*To whom correspondence should be addressed.

Finally, a third group of methods gathers integrated systems that help users to formulate complex multi-criteria queries to retrieve appropriate collections of relevant genes. For instance, the GeneSeeker system (van Driel *et al*., 2005) and the GeneSorter functionality proposed by UCSC Genome Browser (Kent *et al*., 2005) allow experts to test various hypotheses on criteria that can link genes to diseases. An example is found in Tiffin *et al*. (2005), who developed a strategy to identify genes expressed in tissue affected by a disease. Hence, candidate genes are selected if their corresponding annotations with respect to a controlled vocabulary (i.e. eVOC, which is used in Ensembl EST annotation) match the disease annotation. Relevant eVOC annotations for the studied diseases were derived from PubMed abstracts using text-mining techniques.

The Approach for Candidate Gene Retrieval (ACGR) presented in this article is inspired from this last group of methods. Indeed, we propose four steps to guide the discovery of gene–disease relationships. First, several precise definitions of candidate genes are formulated. Next, these definitions are used to design a relational data model and to populate a dedicated DB with relevant data extracted from various internet resources. Finally, to retrieve sets of candidate genes, DB views that express candidate gene definitions are created. Available experimental data can be included in the disease gene definitions and thus exploited together with public annotation data. The approach presented here is tested with three case studies, including a rare orphan gene syndrome.

2 SYSTEMS AND METHODS

2.1 Explicit gene–disease relationships

The definition of a candidate gene provided by the Webster Medical Dictionary is ‘any gene thought likely to cause a disease’. This definition implies that a candidate gene is a gene which is somehow related to a disease. However, specific gene–disease relationships that exist between candidate genes and studied diseases can be articulated in more useful ways by considering information that is available in various public DBs as well as wet-lab datasets.

The most obvious relationship between candidate genes and disease, hereafter called ‘is_co-localized_with’ (denoted by L), expresses the inferred relationship between the localization of a candidate gene and a chromosomal region linked to a given disease. This principle embodied within this statement has guided positional cloning for a long time. The precision of disease localization on chromosomes is highly variable depending on available data. Thanks to recent techniques such as array-CGH (Shaw-Smith *et al*., 2004; Vermeesch *et al*., 2007; Vissers *et al*., 2005), available localization data can be refined using experimental data.

Another direct relationship is tissue or developmental co-expression of both genes and disease features. This relationship has been used in various prioritization methods (Tiffin *et al*., 2005). A variant of this relationship called ‘is_dysregulated_in’ (denoted by D) considers the dysregulation (over-expression or repression) of candidate genes in transcriptomic studies involving patient samples.

Functional annotation of genes is improving in most available DBs and can be connected to disease descriptions. Hence a relationship called ‘has_similar_functional_annotation_with’ (denoted by F) is defined on the basis of a similarity measure between functional annotations of a gene and a disease.

One key aspect of our approach is that the relationship between a candidate gene and a disease may also involve an intermediate gene which satisfies some relationship with the disease. Here, we explore two types of intermediate genes, namely orthologous and interacting genes. It is noteworthy that the co-localization relationship L only applies to the candidate gene itself; whereas, both dysregulation D and functional similarity F relationships apply to intermediate genes as well. Complex definitions are then constructed in the form: ‘a candidate gene is a gene that is co-localized with the disease and is orthologous to a gene that has similar functional annotation with the disease’ and ‘a candidate gene is a gene that is co-localized with the disease and that interacts with a gene that is dysregulated in patients affected by the disease’. The former definition assumes the existence of two relationships, namely L and F, which connect the disease with the candidate gene and with one of its orthologs in a model organism, respectively. The latter definition assumes the existence of two relationships, namely L and D, which connect the disease with the candidate gene and with one of its interaction partners, respectively. Further complex definitions can be formulated similarly, such as ‘a candidate gene is a gene that is co-localized with the disease and that interacts with a gene which is in turn orthologous to a gene having similar functional annotation with the disease’. Retrieving sets of candidate genes which match such complex definitions from masses of biological data are the challenge taken up by the ACGR approach described in this article.

2.2 Relevance of functional gene–disease relationships

In order to assess the relevance of discovered gene–disease relationships, we introduce a measure quantifying the functional similarity relationship F between a gene and a disease. However, to date, no common vocabulary is available to describe functional features of both diseases and genes, hence impeding any straight-forward comparison of disease and gene functional annotations. Current prioritization methods quantify the functional similarity between test genes and training genes based on their GO annotations (Khatri and Draghici, 2005). Ideally the disease functional features should be described with GO vocabulary so that the similarity between gene and disease can be obtained by calculating the similarity between their GO annotations. In practice such disease annotation is performed by an expert of the disease.

This procedure for assessing the relevance of gene–disease relationship presents three main advantages. First, an initial set of training genes is no longer required. Second, available knowledge about the disease is included in disease description. Finally, the rich GO annotations that are available for genes from model organisms will be propagated to human genes thanks to candidate gene definitions involving intermediate orthologous genes.

2.3 Overall presentation of the ACGR approach

The following five steps conceptually describe the proposed *in silico* methodology for candidate gene retrieval. (i) Our system takes as input a functional description of a disease, established by an expert using the GO vocabulary (see Section 3.2), as well as available experimental datasets. The system then collects data from various public DBs. (ii) It first retrieves genes sharing GO annotations with the input disease from either human or model organisms. (iii) Next, relevant annotations of these genes are added, including cytogenetic localization, functional annotation, interacting genes and human orthologs of genes from model organisms. (iv) All retrieved genes are then assigned similarity values that are calculated on the basis of their annotation similarity with the input disease. (v) Finally, sets of candidate genes along with relevant annotation data are built that correspond to various candidate gene definitions.

Our system’s architecture is centred on a DB which is controlled by a DataBase Management System (DBMS). There are three main features of a DBMS that make it attractive to use: centralized data management, data independence and data integration. This contrasts with conventional data processing systems in which each application program has direct access to the data it manipulates. In a DBMS, all data are integrated thereby reducing redundancies and inconsistencies and making data management more efficient. Finally, the existence of a domain data model ensures global data coherence.

The most commonly used conceptual framework for a DBMS is the three-level architecture suggested by the ANSI/SPARC committee (ANSI/X3/SPARC, 1975). The three levels are considered as three different views on the data: (i) the external level or individual user view; (ii) the conceptual level or community user view; and (iii) the internal level or storage view. This three-level DB architecture allows a clear separation of the information meaning (conceptual view) from the physical data structure layer. A DB system that can separate these modelling levels is likely to be flexible and adaptable. The external level is a restricted view on the data, and the same DB may provide a number of different views for different categories of users or needs. In our approach, the candidate gene definitions proposed in Section 2.1 constitute external views on data collected about genes and diseases. The conceptual level determines the data model of the domain of interest, and includes all the information that will be represented in the DB. Finally, the physical model will be replaced here with the so-called 'logical model' (Teorey *et al*., 2006) because the latter is independent of any particular commercial DBMS.

3 ALGORITHM

3.1 DB design

The detailed definitions and relationships presented in Section 2.1 lead to a specification of the various types of data relevant for the retrieval of candidate genes. The resulting conceptual data model is presented in Figure 1 in a common entity–relationship (ER) format.

Queries corresponding to any candidate gene definition (Section 2.1) can be addressed to a DB constructed according to the model shown in Figure 1. For example, the definition of a candidate gene as 'a gene that is co-localized (L gene–disease relationship) with a disease and that is orthologous to a gene that has similar functional annotation with that disease (F gene–disease relationship)' can be represented using the 'Gene', 'Disease', 'GO_term' and 'Ranking_Tool' entities that are linked by the 'Is_Orthologous_To' and 'Is_Ranked As' relationships. The 'Has-Value_in' relationship expresses the D gene–disease relationship as a ratio between experimental values measured, for a given gene and a given experiment, in samples from diseased versus healthy patients. The relational logical data model presented in Table 1 is derived from this conceptual model.

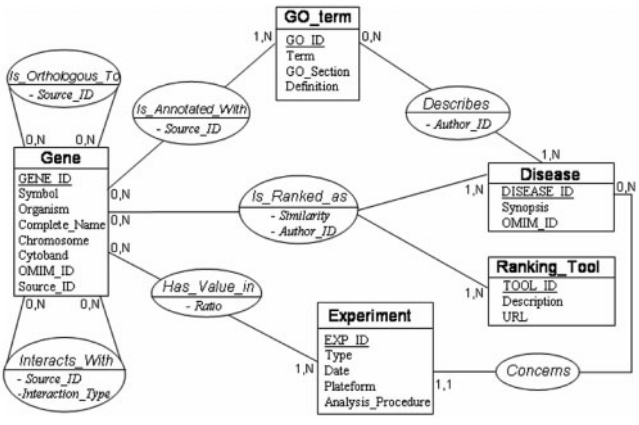

Fig. 1. Conceptual data model for the ACGR DB. Entity types are represented as boxes and relationship types as ellipses. Participation of an entity in a relationship is quantified as the minimal and maximal number of times each occurrence of the entity can participate in the relationship. Note that *Cytoband* in the *Gene* entity is an abbreviation for 'cytogenetic band'.

3.2 Populating the DB

On the basis of the relational data model, it is possible to specify the initialization steps of ACGR DB. Entering a disease description consists of inserting one row of data, hereafter called a tuple, into the 'Disease' table and several tuples in the 'GO_Term' and 'Disease_GO_Term' tables. To this aim, an expert of the studied disease has to carefully (i) extract from her knowledge and from OMIM the phenotypes which characterize the disease, (ii) associate keywords to these phenotypes and (iii) retrieve the most relevant GO terms corresponding to these keywords. The 'Author_ID' attribute is useful to distinguish different descriptions of the same disease. When available, experimental data are entered by inserting one tuple into the 'Experiment' table for each performed experiment, and several tuples into the 'Gene' and 'Gene_In_Experiment' tables, representing all signature genes and their dysregulation ratios. Finally, the system retrieves from public DBs all human, mouse and fly genes that are annotated by at least one GO term associated with the studied disease. Only gene identifiers are inserted into the 'Gene' table at the initialization stage.

The data collection process consists of first retrieving identifiers of human orthologs for mouse and fly genes and then retrieving all required annotations for all gene identifiers present in the 'Gene' table. In particular, interacting genes are retrieved and inserted into the 'Interaction' table. Identifiers for interacting genes which are not present in the 'Gene' table are then added and undergo their own data collection process. Nevertheless at this stage, interaction partners are omitted to prevent an explosion of relationships.

The specification of data wrappers implies selecting appropriate DBs (see Section 4) and mapping the relevant fields onto the ACGR relational data model. Specific wrappers have been designed to plug in external ranking tools for calculating functional similarity values between genes and diseases. Such wrappers will insert tuples into the 'Gene_Disease_Similarity' table, i.e. one tuple per gene and per ranking tool.

3.3 Building sets of candidate genes

In order to express the candidate gene definitions, views are defined in Standard Query Language (SQL) at the logical level of our conceptual framework. A view associates an SQL query with a view name leading to the creation of a virtual table. We have selected four basic definitions leading to the four views described below.

Table 1. Relational logical data model for ACGR DB

Table name	Attribute set
Gene	**Gene_ID**, Symbol, Organism, Complete_name, Chromosome, Cytoband, OMIM_ID, Source_ID
GO_Term	**GO_ID**, Term, GO_section, Definition
Gene_GO_Term	***Gene_ID*, *GO_ID***, Source_ID
Orthology	***Gene_ID1, Gene_ID2***, Source_ID
Interaction	***Gene_ID1, Gene_ID2***, Source_ID, Interaction_Type
Disease	**Disease_ID**, Synopsis, OMIM_ID
Disease_GO_Term	***Disease_ID, GO_ID*, Author_ID**
Ranking_Tool	**Tool_ID**, Description, URL
Gene_Disease_Similarity	***Gene_ID*, *Disease_ID*, Author_ID, *Tool_ID***,Similarity
Experiment	**Exp_ID**, Type, Date, Platform, Analysis_procedure, *Disease_ID*
Gene_In_Experiment	**Gene_ID, Exp_ID**, Ratio

It consists of a set of abbreviated table schemas. Each table contains a set of attributes including a primary key (in bold face) and one or more foreign keys (in italics).

The corresponding SQL queries can be found in the Supplementary Material. For the sake of readability, the datasets produced upon view execution are called 'Datasets'.

Dataset1: genes ranked according to their functional similarity with disease description. This first view retrieves the gene symbol, species, cytogenetic localization and similarity of all ACGR DB genes, sorted by decreasing similarity value. Human, mouse and fly genes are thus collated according to their similarity with disease description. Mouse genes are often ranked better than their human orthologs because of the richer annotation in the model organism. The higher a gene is ranked in Dataset1, the stronger is the functional relationship with the disease.

Dataset2: human orthologs of model organism genes listed in Dataset1. This second view displays all features of Dataset1 for genes retrieved from model organisms (here, mouse and fly) together with the gene symbol, cytogenetic localization and similarity of their human orthologs. Good ranking of a mouse gene can pull its human ortholog to the top of Dataset2 when it was formerly at the bottom of Dataset1 because of poor GO annotation in human. This behaviour is observed, in the *CHD7* gene of CHARGE syndrome, for example (see subsequently).

Dataset3: *genes interacting with the genes listed in Dataset1*. For each gene in Dataset1, the symbol, cytogenetic localization and similarity of the genes reported as interacting with it (mostly via the gene products but other types of interactions are not excluded) are displayed. The source of information concerning these interactions is also displayed. Only intra-species interactions are listed here. Genes that display proper cytogenetic localization but poor similarity values may reveal good disease candidates because of interactions with well-ranked genes mapped elsewhere in the genome.

Dataset4: human orthologs of model organism genes listed in Dataset3. Dataset4 is intended to display candidate genes which are human orthologs of model organism genes that interact with well-ranked genes.

When experimental data are available, it can be included into each of the views described above, thereby producing four supplementary views: from Dataset1Exp to Dataset4Exp. An example of this is presented below in the case study on AICARDI syndrome.

Further queries on the basic ACGR views can then provide customized lists of candidate genes. Indeed, creating sets of annotated candidate genes as SQL views allow biologists to benefit from the numerous advantages of this powerful approach. First, writing new queries is simplified. Second, the views are automatically refreshed whenever the DB is updated. Finally, defining views contributes to the integrity and security of the DB because end-users may be given tuned privileges on views rather than on the underlying data tables.

4 IMPLEMENTATION

The technical implementation choices described in this work are not mandatory since other techniques are conceivable depending on the target deployment environment. For example, here wrappers for retrieving and integrating data from various data sources have been implemented as scenarios of the Xcollect software (Devignes *et al*., 2005). Xcollect scenarios are configured to formulate queries automatically, send them to a remote web resource, parse the returned document and store the desired data in an XML document. Capturing the date of last DB update is included in each scenario to help track data quality. The specific Xcollect scenarios used here are available in the Supplementary Material.

In this work, data sources were selected according to their updating frequencies, annotation quality and coverage. Thus, GO terms corresponding to keywords describing the disease were retrieved from AMIGO DB; genes annotated with selected GO terms were retrieved from Entrez-Gene at NCBI as well as all gene annotations. Symbols of orthologous genes were retrieved from Entrez-HomoloGene.

The storage of the collected data in the ACGR DB was performed with the help of XSL transformations designed to convert each Xcollect session document into appropriate SQL commands. Besides Xcollect wrappers, we developed a wrapper to invoke the GO-Family program available in the GOToolBox (Martin *et al*., 2004). The program was modified slightly because a list of GO terms rather than a list of reference gene symbols is required as well as the list of genes to be ranked. Briefly, the program fetches all GO terms annotating a candidate gene as well as their parent terms. It also fetches all parents of the disease-specific GO terms. Then it calculates a similarity percentage taking into account identical and non-identical terms between the set of GO terms associated with the candidate gene and the set of disease-specific GO terms.

The EasyPHP package was used for data management and user interface development. This package includes a web server (Apache), a DBMS (mySQL) and a script language (PHP). The corresponding programs along with a user guide are available at http://bioinfo.loria.fr/projects/acgr/acgr-software/.

5 RESULTS AND DISCUSSION

5.1 Three case studies

The ACGR approach was initially motivated by the need to analyse results obtained for AICARDI syndrome (OMIM %304050) which is currently being investigated experimentally (Yilmaz *et al*., 2007). To date, no responsible gene is known for this disease. Two other rare syndromes, CHARGE (OMIM #214800) and GOLTZ (OMIM #305600), were selected from the literature. The genes responsible for these two syndromes have recently been reported (Grzeschik *et al*., 2007; Vissers *et al*., 2004; Wang *et al*., 2007b), but this information is not included in the annotations collected in the ACGR DB. It is therefore relevant to test the ACGR approach on these recently elucidated diseases.

5.2 Populating the DB

Table 2 shows for the three case studies the correspondence between disease phenotypes and Biological Process GO terms. Phenotypes were selected from OMIM notices regarding diagnoses. Keywords (data not shown, see Supplementary Material) were chosen to characterize each phenotype. For a given keyword, GO terms were selected at the relevant level of the GO hierarchy. A GO term is included when all its children are relevant. In the case of AICARDI syndrome, a third phenotype (infantile spasms) is frequently observed but does not correspond to any specific GO term. According to the clinicians, this phenotype is covered by the 'Forebrain development' GO term.

Experimental data were inserted into the DB for the AICARDI syndrome as explained in Section 3.2. These data concern 300 genes which ANOVA analysis of several transcriptomic experiments found to be dysregulated (Yilmaz, 2007). For these genes the ratio attribute was set to 1; whereas, it was set to 0 for any other gene.

Table 3 summarizes the contents of the ACGR DB for the three case studies. The *#GO* column displays the number of GO terms specific to the disease. The *#fly*, *#mouse* and *#human* columns show the number of genes annotated by at least one of these GO terms for each organism. The '#dysregulated' column indicates the number of experimentally determined human dysregulated genes stored in the DB. The last column gives the total number of genes after the inclusion of other orthologous and interacting genes.

5.3 Building sets of annotated candidate genes

Dataset1 to Dataset4 were constructed for each case study as described in Section 3.3 to enable queries reflecting expert hypotheses about candidate genes to be formulated. The complete tables are available as Supplementary Material. Table 4 displays the first three tuples from CHARGE Dataset2. The human *CHD7* gene that is responsible for this disease (Vissers *et al*., 2004) appears in second position as orthologous to the mouse Chd7 gene which has a high similarity to disease description (48%). It is worth noting that the low similarity of the human *CHD7* gene annotation to CHARGE GO terms (4%) relegates it to the bottom of Dataset1. Selecting human genes from chromosome 8 in CHARGE Dataset2 yields the *CHD7* gene as the first-ranked candidate gene.

The CHARGE case study shows that the ACGR approach would have been able to designate the *CHD7* gene as the best candidate gene in the group of nine genes identified by the authors at 8q12 thus prioritizing its sequencing. It is worth noting that although the association of *CHD7* with CHARGE syndrome was established 3 years ago, the GO annotation of this gene does not reflect this association.

Table 5 shows the first six tuples from GOLTZ Dataset4. Despite its low similarity to disease description (7%), the responsible human *PORCN* gene appears at the fifth position in GOLTZ Dataset4 that contains 51 lines and as the first candidate gene located on chromosome X. This is due to the fact that the mouse Porcn gene is reported as interacting with the mouse Wnt7a gene which has good similarity to the disease description. Hence the ACGR approach could have pointed to the *PORCN* gene even before the localization refinement of the disease provided by the CGH array experiment (Grzeschik *et al*., 2007; Wang *et al*., 2007b).

In the case of AICARDI syndrome, Dataset1Exp to Dataset4Exp were produced including transcriptomic data. A first query on Dataset1Exp retrieved 71 genes located on human chromosome X. Table 6 displays the first four genes of this list. The best-ranked *PLXNA3* gene seems to be an interesting candidate. Its annotation is rather similar to the AICARDI GO terms (56%). However, to date, it has not been associated with any human disease. The following *ARX* and *SOX3* genes, namely MRX54 (OMIM #300419)

Table 2. List of GO terms defined by the clinicians on August 31, 2007

Syndrome	Phenotype	GO term (Biological Process hierarchy)
CHARGE	Coloboma	Camera-type eye morphogenesis [47]
	Choanal atresia	Nose development [2]
		Embryonic cranial skeleton morphogenesis [16]
	Ear abnormality	Ear development [155]
	Deafness	Sensory perception of sound [203]
	Heart anomaly	Heart morphogenesis [69]
GOLTZ	Skin defects	Skin development [22]
	Digital anomalies	Embryonic digit morphogenesis [28]
	Skeletal defects	Embryonic skeletal morphogenesis [25]
AICARDI	Corpus callosum agenesis	Forebrain development [191]
		Corpus callosum development [0]
		Corpus callosum morphogenesis [0]
		Neuron migration [139]
		Neural plate development [117]
	Chorioretinal lacunae	Camera-type eye morphogenesis [47]

The number of genes annotated by a GO term is indicated in brackets.

Table 3. Numbers of genes stored in the ACGR DB for the three case studies

Disease	#GO	#fly	#mouse	#human	#dysregulated	#genes
CHARGE	6	29	172	223	0	1410
GOLTZ	3	0	55	272	0	1583
AICARDI	6	2	182	166	300	2218

Table 4. The three top-ranked tuples from CHARGE Dataset2

Symbol	Organism	Cyto-band	Sim	Orthol_ Symbol	Orthol_ Cytoband	Orthol_ Sim
Tmie	mouse	9 64.0 cM	62	*TMIE*	3p21	62
Chd7	mouse	4 1.0 cM	48	*CHD7*	8q12.2	4
Gjb6	mouse	14 22.5 cM	48	*GJB6*	13q12	45

The column *Sim* refers to the values taken by the *Similarity* attribute of the *Gene_Disease_Similarity* table. These values are expressed as percentages owing to the particular ranking tool that was used. The *Orthol_Symbol, Orthol_Cytoband* and *Orthol_Sim* columns display values for the human orthologs of the considered mouse genes.

Table 5. The six top-ranked tuples from GOLTZ Dataset4

Symbol	Organism	Cytoband	Sim	Interac_ Symbol	Source	Interac_ Cytoband	Interac_Sim	Orthol_ Symbol	Orthol_ Cytoband	Orthol_ Sim
Gna12	mouse	5 82.0 cM	35	Ppp5c	BIND	7 4.0 cM	4	*PPP5C*	19q13.3	4
Col5a2	mouse	1 C1	32	Smad2	BIND	18 48.0 cM	15	*SMAD2*	18q21.1	17
Col5a2	mouse	1 C1	32	Smad7	BIND	18 unknown	5	*SMAD7*	18q21.1	5
Col5a2	mouse	1 C1	32	Samd3	BIND	9 unknown	15	*SMAD3*	15q22.33	16
Wnt7a	mouse	6 39.5 cM	31	Porcn	BIND	X 2.15 cM	5	***PORCN***	**Xp11.23**	7
Lgals3	mouse	14 C1	28	Sufu	BIND	19 47.0 cM	17	*SUFU*	10q24.32	19

Columns *Interac_Symbol* to *Interac_Sim* columns refer to the interacting genes of considered mouse genes. Columns *Orthol_Symbol* to *Orthol_Sim* are described in Table 4.

Table 6. The four top-ranked human genes localized on chromosome X from AICARDI Dataset1Exp

Symbol	Organism	Sim	Cytoband	Ratio
PLXNA3	Human	56	Xq28	0
ARX	Human	40	Xp21	0
SOX3	Human	34	Xq27.1	0
DCX	Human	26	Xq22.3-q23	0

Table 7. The four top-ranked human tuples from AICARDI Dataset3Exp

Symbol	Cytoband	Sim	Ratio	Interac_Symbol	Interac_Cytoband	Interac_Sim	Interac_Ratio	Source
DLX5	7q22	50	1	*MAGED1*	Xp11.23	3	0	HPRD
UBE3A	15q11-q13	22	1	*UBQLN2*	Xp11.23-p11.1	8	0	HPRD
CXCL10	4q21	21	1	*CXCR3*	Xq13	10	0	HPRD
IGF1	12q22-q23	21	1	*IGSF1*	Xq25	6	0	BIND

The columns *Symbol* to *Ratio* refer to dysregulated genes, and the columns *Interac_Symbol* to *Interac_Ratio* refer to the interacting candidate genes. The *Source* column indicates the database where the interaction is documented.

and MRGH (OMIM #300123), are both responsible for diseases involving mental retardation. The next *DCX* gene is a good internal control since it is responsible for X-linked lissencephaly (LISX, OMIM #300067), a disease-like AICARDI syndrome involving agenesis of the corpus callosum and multiple heterotopia.

Further queries were applied to AICARDI Dataset3Exp to explore possible interactions between dysregulated genes and candidate genes. Table 7 shows four candidate genes ('Interac_Symbol' column) from Dataset3Exp, located on chromosome X and interacting with the four best-ranked dysregulated genes ('Symbol' column). The *MAGED1* gene interacts with the *DLX5* gene which is dysregulated in our transcriptomic experiments and its GO annotation displays 50% similarity with the AICARDI-specific GO terms. The interaction between these two gene products is based on *in vivo* experiments (Masuda *et al*., 2001).

5.4 Discussion

Overall, the ACGR approach has received enthusiastic feedback from experimentalists. Indeed conducted experiments yielded very satisfying results in the CHARGE and GOLTZ case studies. We have shown that in both cases responsible genes related to the disease are found at the first rank position when chromosome localization is taken into account. Thus, the ACGR approach would have been useful at the time of the discovery of these responsible genes to avoid unnecessary sequencing. In the case of AICARDI syndrome, the ACGR approach provided several meaningful and promising candidate genes that are currently being analysed further. For instance, the *MAGED1* gene displays several features associated with disease genes (Tu *et al*., 2006). It is a 99.3 kb long gene due to a large intron (91 kb) separating the first exon from the 12 other exons that are grouped over the remaining 8 kb. Interestingly, two of the retrieved candidate genes (*MAGED1* and *UBQLN2*) are located in the same cytogenetic band (Xp11.23), which is known to be correlated with several neuro-psychiatric disorders. It should be noted that for this disease, the small number of recruited patients hampers the application of purely experimental protocols. In addition to the presented case studies, ongoing investigations indicate that the approach presented here may facilitate future endeavours to identify susceptibility genes for complex diseases.

The robustness and flexibility of our approach makes it possible to explore various alternative approaches or strategies, including varying the ranking procedure and the selection of primary data sources. For example, data about interaction networks could be retrieved from the protein complexes curated by Lage *et al*. (2007). The GO-Family algorithm used for gene ranking in this study could be replaced by any other similarity measurement between GO terms (Lord *et al*., 2003; Wang *et al*., 2007a; Zhang *et al*., 2006). The similarity between eVOC terms annotating both gene expression and affected tissues could be used to assess 'is_co-expressed' relationships (Tiffin *et al*., 2005), for example.

A possible limitation of the current work may be the low number of case studies analysed. Since an expert of each studied disease has to be involved in the first step of the approach, this clearly hampers automated large-scale evaluation. Moreover, it should be stressed that success in retrieving at a good rank the gene responsible for a disease strongly depends on both user's expertise and the quality of available data.

Nevertheless, the results presented here clearly demonstrate the explicit querying capabilities of the ACGR system and the originality of this approach for providing explanations on why a certain gene is related to a disease.

ACKNOWLEDGEMENTS

We thank Sylvain Lambermont for his contribution at early stage of the work, Dr Leheup for helping in selecting disease-specific GO terms, Amine Rouhane-Hacène and Dave Ritchie for careful reading of the manuscript. S.Y. was supported by the AAL (Amis d'Anne-Lorène) association and Région Lorraine.

Funding: Contrat de Plan Etat-Région Lorraine (PRST Intelligence Logicielle).

Conflict of Interest: none declared.

REFERENCES

Adie,E.A. *et al*. (2005) Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, **6**, 55.

Adie,E.A. *et al*. (2006) SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics*, **22**, 773–774.

Aerts,S. *et al*. (2006) Gene prioritization through genomic data fusion. *Nat. Biotechnol.*, **24**, 537–544.

ANSI/X3/SPARC (1975) *Study Group on Data Base Management Systems, Interim Report, FDT 7 No. 2*. ACM, New York.

Barillot,R. *et al*. (2004) New strategy for the representation and the integration of biomolecular knowledge at a cellular scale. *Nucleic Acids Res.,* **32**, 3581–3589.

Botstein,D. and Risch,N. (2003) Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nat. Genet.*, **33(Suppl)**, 228–236.

Calvo,B. *et al*. (2007) A partially supervised classification approach to dominant and recessive human disease gene prediction. *Comput. Methods Programs Biomed.*, **85**, 229–237.

Chiang,J.H. *et al*. (2006) GeneLibrarian: an effective gene-information summarization and visualization system. *BMC Bioinformatics*, **7**, 392.

Devignes,M.D. *et al*. (2005) User-designed web services to support heterogeneous biological data retrieval. NETTAB workshop on Workflows management: new abilities for the biological information overflow, available at http://www.nettab.org/2005/progr.html (last accessed date December 8, 2008).

Freudenberg,J. and Propping,P. (2002) A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, **18** (Suppl. 2), S110–S115.

George,R.A. *et al*. (2006) Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res.*, **34**, e130.

Giallourakis,C. *et al*. (2005) Disease gene discovery through integrative genomics. *Annu. Rev. Genomics Hum. Genet.*, **6**, 381–406.

Grzeschik,K.H. *et al*. (2007) Deficiency of PORCN, a regulator of Wnt signaling, is associated with focal dermal hypoplasia. *Nat. Genet.*, **39**, 833–835.

Kent,W.J. *et al*. (2005) Exploring relationships and mining data with the UCSC Gene Sorter. *Genome Res.*, **15**, 737–741.

Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.

Lage,K. *et al*. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.,* **285**, 309–316.

Lopez-Bigas,N. and Ouzounis,C.A. (2004) Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res.*, **32**, 3108–3114.

Lopez-Bigas,N. *et al*. (2006) Highly consistent patterns for inherited human diseases at the molecular level. *Bioinformatics*, **22**, 269–277.

Lord,P.W. *et al*. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275–1283.

Martin,D. *et al*. (2004) GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol.*, **5**, R101.

Masseroli,M. *et al*. (2004) GFINDer: Genome Function INtegrated Discoverer through dynamic annotation, statistical analysis, and mining. *Nucleic Acids Res.*, **32**, W293–W300.

Masseroli,M. *et al*. (2005) GFINDer: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists. *Nucleic Acids Res.*, **33**, W717–W723.

Masuda,Y. *et al*. (2001) Dlxin-1, a novel protein that binds Dlx5 and regulates its transcriptional function. *J. Biol. Chem.*, **276**, 5331–5338.

Oti,M. *et al*. (2006) Predicting disease genes using protein–protein interactions. *J. Med. Genet.*, **43**, 691–698.

Perez-Iratxeta,C. *et al*. (2002) Association of genes to genetically inherited diseases using data mining. *Nat. Genet.*, **31**, 316–319.

Perez-Iratxeta,C. *et al*. (2005) G2D: a tool for mining genes associated with disease. *BMC Genetics*, **6**, 45.

Rossi,S. *et al*. (2006) TOM: a web-based integrated approach for identification of candidate disease genes. *Nucleic Acids Res.*, **34**, W285–W292.

Shaw-Smith,C. *et al*. (2004) Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. *J. Med. Genet.*, **41**, 241–248.

Sun,H. *et al*. (2006) GOFFA: Gene Ontology for functional analysis – a FDA Gene Ontology tool for analysis of genomic and proteomic data. *BMC Bioinformatics*, **7** (Suppl. 2), S23.

Teorey,T.J. *et al*. (2006) In Cerra,D.D. (ed.) *Database Modeling and Design: Logical Design.* Morgan Kaufmann Publishers, San Francisco.

Tiffin,N. *et al*. (2005) Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res.*, **33**,1544–1552.

Tu,Z. *et al*. (2006) Further understanding human disease genes by comparing with housekeeping genes and other genes. *BMC Genomics*, **7**, 31.

Turner,F.S. *et al*. (2003) POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol.*, **4**, R75.

van Driel,M.A. *et al*. (2005) GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. *Nucleic Acids Res.*, **33**, W758–W761.

Vermeesch,J.R. *et al*. (2007) Guidelines for molecular karyotyping in constitutional genetic diagnosis. *Eur. J. Hum. Genet*., **15**, 1105–1114.

Vissers,L.E. *et al*. (2004) Mutations in a new member of the chromodomain gene family cause CHARGE syndrome. *Nat. Genet.*, **36**, 955–957.

Vissers,L.E. *et al*. (2005) Identification of disease genes by whole genome CGH arrays. *Hum. Mol. Genet.*, **14**, R215–R223.

Wang,J.Z. *et al*. (2007a) A new method to measure the semantic similarity of GO terms. *Bioinformatics*, **23**, 1274–1281.

Wang,X. *et al*. (2007b) Mutations in X-linked PORCN, a putative regulator of Wnt signaling, cause focal dermal hypoplasia. *Nat. Genet.*, **39**, 836–838.

Xu,J. and Li,Y. (2006) Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics,* **22**, 2800–2805.

Yilmaz,S. (2007) Searching Candidate Genes for AICARDI Syndrome : Combining Experimental Approach and Bioinformatics. PhD thesis, Université Henri Poincaré, Nancy 1.

Yilmaz,S. *et al*. (2007) Screening of subtle copy number changes in Aicardi syndrome patients with a high resolution X chromosome array-CGH. *Eur. J. Med. Genet*., **50**, 386–391.

Zhang,P. *et al*. (2006) Gene functional similarity search tool (GFSST). *BMC Bioinformatics*, **7**, 135.