Weighting Indirect Relations to Elucidate the Direct Association of SNP-Disease by Use of SPARQL Queries

Remzi Çelebi[1], Özgür Gümüş[1], Yeşim Aydın Son[2]
[1] Department of Computer Engineering, Ege University
İzmir, Turkey
{remzi.celebi, ozgur.gumus}@ege.edu.tr

[2] Department of Health Informatics, Middle East Technical University
Ankara, Turkey
yesim@metu.edu.tr

Abstract. One of the current issues in the bioinformatics domain is to identify genomic variations underlying the complex diseases. There are millions of genetic variations as well as environmental factors that may cause human diseases. Semantic web interlinks diverse data that may reveal many hidden relations and can be utilized for personalized medicine. This requires discovering relationships between phenotypes and genotypes, to answer how the genotype of an individual affects his/her health. Additionally, through identification of genomic variations based on an individual's genotype we can predict the response to a selected drug therapy and accordingly suggest treatment or drug regimes. A personalized medicine knowledgebase can interlink genotypic variations and its possible somatic changes that effects drug targets to pick best treatment and drug regimens for individuals. Such a knowledgebase may help to identify the factors that best explain the association between genotype and phenotype. We've used SPARQL queries to weight factors which link the genotype and phenotype via indirect relationships, and the paths of relationships. A personalized medicine knowledgebase build with the presented approach can interlink genotypic variations and its possible somatic changes that effects drug targets to pick best treatment and drug regimens for individuals, and may help to identify the factors that best explain the association between genotype and phenotype.

Keywords: SPARQL, SNP, personalized medicine.

1 Introduction

Semantic web[1] interlinks diverse data that may reveal many hidden relations and can be utilized in personalized medicine, which requires discovering relationships between phenotypes and genotypes, to answer how the genotype of an individual affects his/her health and accordingly suggest treatment or drug regimes. Through identification of genomic variations based on an individual genotype we can predict the response to a selected drug therapy. A personalized medicine knowledgebase can interlink genotypic variations and its possible somatic changes that effects drug targets to pick best treatment and drug regimens for individuals.

Single nucleotide polymorphisms (SNPs) are the most common form of genetic variations and they can represent an individual's genetic variability in greatest detail. However, an associated SNP is likely part of a larger region of linkage disequilibrium. This makes it difficult to precisely identify the causal SNPs for different phenotypes. In addition to SNPs, individual genes of the region being studied and biological pathways they are involved in should be considered while investigating relations between genotypes to phenotypes.

We have used SPARQL query language to semantically retrieve and manipulate biological data in RDF. An integrated multiple datasets from different sources is used to build a network of disease, pathway, gene, SNP and LD-SNP (linkage disequilibrium of SNP). Relation between resources is presented in Figure 1. With integration of these resources, distinguishing secondary knowledge that uses indirect relations rather than a direct one in the emergent linked data can be utilized for weighting and prioritizing possible disease related SNPs. Also, how much each factor contributes to the association of SNP-disease can be revealed by using all integrated information related with the association.

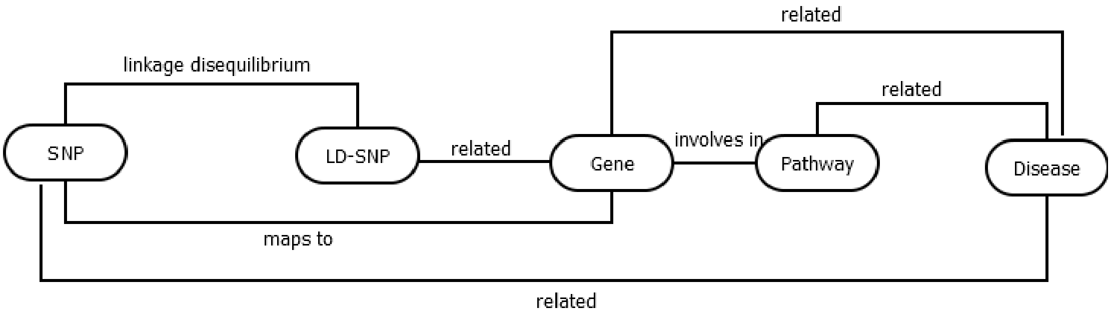

Figure 1: Relation between resources

2 Method

2.1 Datasets

The datasets used to build our knowledgebase have been gathered from multiple data sources. Some of them were already available in RDF format. CTD dataset is used for Disease-Gene-Pathway association. For Gene-Disease information, OMIM and for Pathway-Disease association PharmGKB datasets are used. These datasets are publicly accessible through Bio2RDF project (bio2rdf.org). Other resources required data preprocessing to be converted into RDF. SNP related information are extracted from dbSNP and converted to RDF by a Python script. A subset of SNPs in the dbSNP is used in order to lower the number of SNPs to a manageable level. SNPs listed in Ilumina, Affymetrix platforms and disease associated SNPs defined in OMIM, PharmGKB databases are selected lowering the number of SNPs to be processed from approximately 50 million to 4.3 million. Additionally linkage disequilibrium information between SNP pairs is provided through Hapmap project (hapmap.org). Regression ratios above 0.75 are considered meaningful and collected for the linkage disequilibrium between any two SNPs.

TABLE 1: List of relation paths from SNP to disease and the statistics of the corresponding the SPARQL query

RELATION PATHS	# of MATCHES	# of RETRIEVED	PRECISON	RECALL
SNP - Gene - Disease	11	1757	11/ 1757	11/ 13
SNP - Gene - Pathway -Disease	8	398217	8/ 398217	8/13
SNP - LD-SNP - Gene - Disease	4	850	4/ 850	4/ 13
SNP - LD-SNP -Gene - Pathway - Disease	2	192306	2/ 192306	2/13
SNP - LD-SNP - Disease	0	60	0/60	0/13

2.2 Weighting semantic paths

In information retrieval context, precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved. SPARQL is a query language for semantic data. SPARQL query is mostly used to retrieve and manipulate RDF data but it can also be used for information retrieval. One can utilize a SPARQL query as information retrieval method and can measure the performance by calculating precision and recall values. So, how well a SPARQL query reveals the observed relation can be evaluated based on the secondary knowledge provided.

a) Query-1

```
PREFIX dbsnp_voc: <http://bio2rdf.org/dbsnp_vocabulary:>
PREFIX ctd_voc: <http://bio2rdf.org/ctd_vocabulary:>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT  distinct ?rsid
WHERE {
<http://bio2rdf.org/pharmgkb:PA444370> rdfs:seeAlso ?disease .
?disease rdf:type ctd_voc:Disease .
?disease ctd_voc:pathway ?pathway .
?gene ctd:voc:pathway ?pathway .
?rsid dbsnp_voc :geneId ?gene .
}
```

b) Query-2

```
PREFIX ctd_voc: <http://bio2rdf.org/ctd_vocabulary:>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dbsnp_voc: <http://bio2rdf.org/dbsnp_vocabulary:>

SELECT distinct ?rsid
WHERE {
<http://bio2rdf.org/pharmgkb:PA444370> rdfs:seeAlso ?disease .
?assoc rdf:type ctd_voc:Gene-Disease-Association .
?assoc ctd_voc:disease ?disease .
?assoc ctd_voc:gene ?gene .
?rsid dbsnp_voc:geneId ?gene .
}
```

Figure 2: Use case

2.3 Queries

We have defined a set of SPARQL queries which examine SNP-disease association by use of different paths of relation. Heart failure is considered as a case study. In Figure-2, two of defined queries which use different relation paths are given. Query-1 reveals the indirect SNP-disease relations by following "SNP - Gene - Pathway –Disease" and Query-2 finds the indirect relations using relation path of "SNP - Gene - Disease ".

3 Results

In PharmGKB dataset, Heart Failure is associated to 13 SNPs and some of these SNPs can be found by the relation path of "SNP - Gene - Pathway -Disease". 8 of 13 SNPs can be retrieved by this path of relationship and unique 398217 SNPs retrieved from total 586758 SNPs as result of the query. Precision and recall of this query can be seen in Table 1. Similarly, when “SNP - Gene -Disease” path is used, less matched SNPs are retrieved but recall value is much better than previous query. All possible relation paths and precision-recall values are listed Table 2.

TABLE 2: List of SNPs and its match by SPARQL Queries. Letter abbreviations; S:SNP, LD:LD-SNP, G:Gene, P:Pathway, D:Disease (1 means “match” , 0 means “no match”)

	MATCH via	**MATCH via**	**MATCH via**	**MATCH via**	**MATCH via**
SNP ID	**S - G- D**	**S- G- P- D**	**S – LD- G -D**	**S -LD- G -P- D**	**S - LD- D**
rs1042713	1	1	1	1	0
rs1801252	1	1	0	0	0
rs1042714	1	1	1	1	0
rs1801253	1	1	0	0	0
rs1799752	1	1	0	0	0
rs1800566	1	1	0	0	0
rs1800888	1	1	0	0	0
rs1001179	1	1	0	0	0
rs4880	1	0	1	0	0
rs1056892	1	0	1	0	0
rs877087	1	0	0	0	0
rs17098707	0	0	0	0	0
rs2207418	0	0	0	0	0
# of MATCHES	11	8	4	2	0
# of RETRIEVED	1757	398217	850	192306	60
PRECISON	11/ 1757	8 / 398217	4/850	2/192306	0
RECALL	11 /13	8 / 13	4/13	2/13	0

4 Conclusion

Here possible semantic pathways are presented to link SNPs and their associated diseases through available biological databases and the overall performance is compared to manually curated information from PharmGKB. The weighting paths of relationship may be helpful to better define underlying factors SNPs’ biological link with diseases and molecular etiology of diseases. In the example presented here, searching the disease related genes and mapping the SNPs on it provided the best performance. Even though there are number of limitations about our current knowledge of SNP disease associations, in all scenarios there were high number of false positives which points out that additional approaches for the filtering is needed. Also, the paths including LD-SNP information presents the lowest number of hits, but the study needs to be repeated with larger data sets and different disease groups to validate these findings. Additionally we suggest that, integrating more descriptive data in our knowledgebase such as protein-protein interaction (PPI), gene expression profiles, and evolutionary conservation information, would be helpful to explain effects of indirect relations to SNP-disease association.

References

1. Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific american, 284(5), 28-37.
2. Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. International Journal on Semantic Web and Information Systems (IJSWIS), 5(3), 1-22
3. Williams, A. J., Harland, L., Groth, P., Pettifer, S., Chichester, C., Willighagen, E. L., et al., "Open PHACTS: semantic interoperability for drug discovery." Drug discovery today 17.21 (2012): 1188.
4. Wild, D. J., Ding, Y., Sheth, A. P., Harland, L., Gifford, E. M., & Lajiness, M. S., "Systems chemical biology and the Semantic Web: what they mean for the future of drug discovery research." Drug discovery today 17.9 (2012): 469-474.