31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Search of phenotype related candidate genes using Gene Ontology-based semantic similarity and protein interaction information: Application to Brugada syndrome.

Raimon Massanet, Joan-Josep Gallardo-Chacón, Pere Caminal and Alexandre Perera
{raimon.massanet, joan.josep.gallardo, pere.caminal, alexandre.perera}@upc.edu

Abstract— **This work presents a methodology for finding phenotype candidate genes starting from a set of known related genes. This is accomplished by automatically mining and organizing the available scientific literature using Gene Ontology-based semantic similarity. As a case study, *Brugada* syndrome related genes have been used as input in order to obtain a list of other possible candidate genes related with this disease. *Brugada* anomaly produces a typical alteration in the Electrocardiogram and carriers of the disease show an increased probability of sudden death. Results show a set of semantically coherent proteins that are shown to be related with synaptic transmission and muscle contraction physiological processes.**

I. INTRODUCTION

Recent years have seen a drastic increase of the publicly available genetic and proteomic information. There are two main entry points for accessing gene and protein information. In the first place, UniProtKB[1], which is a collaboration between the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR). Secondly, Entrez[2], which is a cross-database search tool offered by the National Center for Biotechnology Information (NCBI) from the US. This portal facilitates access to many available resources. In particular, Entrez Gene is a central source of information on genes, and transcripts. From all the publicly available sources of information on protein-protein interactions (PPI), three of them are specially popular: *IntAct* contains binary and complex interactions extracted from the literature [3], *BioGRID* contains protein and genetic interactions derived from reported studies and curated according to the literature [4], finally *Reactome* contains information on chemical reactions involving proteins [5], and uses a particular exhaustive data scheme that allows for the storage of reactions of very different complexities. On the other hand, from protein databases it is possible to retrieve semantic annotations made by researchers. This semantic information could include data on where a protein can be found, what biochemical function it performs or in what biological process it is involved. These annotations used to be written in natural language so it was difficult to analyze them in a programmatic way. This yielded to further efforts to organize this semantic annotations, lead by the Gene Ontology (GO) project [6]. GO provides a common framework for researchers to annotate proteins and genes with predefined semantic terms. An ontology is a controlled set of terms and relations among them. The *isA* relationship establishes a hierarchy in which a child term refines the meaning of its ancestor term. A parent term can be seen as a generalization of the meaning of all its child terms. GO organizes its terms in three ontologies, each of which contains a different sort of semantic information about a gene or a protein. The *Cellular component* ontology defines possible locations of proteins or expression of genes (e.g. *Cellular membrane* or *Nucleus*). The *Molecular function* ontology defines biochemical functions that a protein, or potentially any molecule, can fulfill (e.g. *Catalytic activity* or *Transporter activity*). Finally, the *Biological process* ontology defines cell processes in which a gene or a protein can be involved (e.g. *Blood coagulation* or *Signal transduction*). The Gene Ontology Annotation (GOA) is a database that contains protein to GO terms annotations[7]. Resnik applied information theory concepts to ontologies [8] defining a semantic similarity using the information content (IC) measure and the *most informative common ancestor* between two terms in an ontology, which have been further extended by other authors (e.g. Jiang and Conrath [9]). The measures cited above define similarity between two terms of an ontology based on the appearance frequency of every term in a knowledge corpus. However, proteins and genes tend to be annotated to sets of terms. Thus, a definition of similarity between sets of terms of an ontology was needed. The Gene Ontology, seen from the *isA* relationship point of view can be seen as a directed acyclic graph (DAG) where nodes are terms and edges go from parents to children nodes. Recently, new measures have been proposed to compare sets of terms of an ontology based on the graph structure induced by them. Falcon and Gentleman proposed an union-intersection based measure called *simUI*, which counts the nodes on the intersection of the two induced graphs and divides it by the number of nodes on the union [10]. This approach was used by Pesquita et al to define *simGIC*, which computes the IC sum of the nodes in the intersection and normalizes it to the IC sum of the nodes in the union [11]. Chagoyen et al used previously proposed measures to define the concept of *coherence* of a protein set [12]. In addition, Ovaska

This work was partially funded by CICYT TEC2007-63637/TCM and Ramón y Cajal program from Ministerio de Educación y Ciencia.

Raimon Massanet, Pere Caminal and Alexandre Perera are with Biomedical Engineering Research Center (CREB), Technical University of Catalonia (UPC). http://www.creb.upc.es, http://www.upc.edu.

Joan-Josep Gallardo-Chacón is with CIBER de Bioingeniería,Biomateriales y Nanomedicina. http://www.ciber-bbn.es/

978-1-4244-3296-7/09/$25.00 ©2009 IEEE

et al used ontology-based semantic similarity measures for clustering genes resulting from microarray experiments [13]. According to the authors, this methodology should simplify the analysis of the large number of genes yielded by the experiment. Other authors have employed a similar approach to the prediction of disease candidate genes by means of PPI information [14]. When studying possible genes related to a phenotype, researchers often select target genes based on *a priori* knowledge of gene functions. This can be challenging when the number of candidate genes is very large. Semantic similarity offers a fast and automatic way of comparing *a priori* knowledge of genes.

This manuscript proposes a methodology in order to obtain a set of genes that are candidate to be related with a given phenotype. Given this specific phenotype, the set of already known proteins related to this trait are found. The common set of proteins related with this initial set is obtained through mining of PPI databases. This total set of proteins is organized through a semantic measure and a clustering algorithm in order to obtain a subset of candidates.

II. MATERIALS AND METHODS

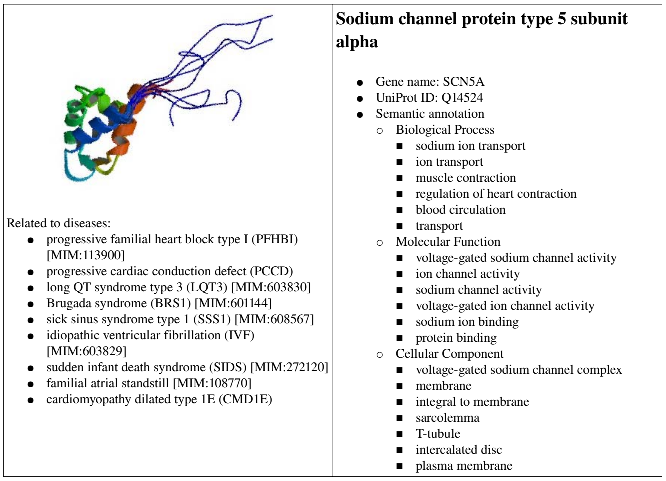
The methodology has been applied to two genes known to be related with *Brugada* syndrome phenotype: *SCN5A* and *CACNA1C*. Fig.1 shows further details on the *SCN5A* gene. Defects in *SCN5A* have been related to *Brugada syndrome type* 1 (*BRS1*) [MIM:#601144] and defects in *CACNA1C* have shown relation with *Brugada syndrome type 3 (BRS3)* [MIM:#611875]. Starting from a known set of genes related with a phenotype, the minimum connected subset of the human interactome that contains this set is generated. This is done by growing the map of interactions with new interactants at every iteration until there is only one fully connected component in the graph. Semantic annotations were retrieved for all nodes of the graph, and the extended set of annotations was computed. Extending the set of annotations reflects the fact that if a protein is annotated to a term t then it must also be annotated to all ancestors of t. Then, semantic similarity was computed between all pairs of nodes. To do so IC measure was computed , $IC(t) = -log\ p(t)$, where $p(t)$ is the probability of appearance of term t in the extended set of annotations[8]. Semantic similarity was computed using *simGIC* measure [11]:

$$simGIC(p_1, p_2) = \frac{\sum_{t \in \{ann(p_1) \cap ann(p_2)\}} IC(t)}{\sum_{t \in \{ann(p_1) \cup ann(p_2)\}} IC(t)} \quad (1)$$

where $ann(p)$ is the expanded set of semantic annotations for gene p. From (1) it is obvious that semantic similarity is normalized to 1 for any values of information content. For this reason, IC has not been normalized. Semantic dissimilarity is then computed as:

$$distGIC(p_1, p_2) = 1 - simGIC(p_1, p_2) \quad (2)$$

Semantic similarity defines a kernel in which genes organize. The goal is to retrieve other genes which are within the minimum semantic hypersphere, given that kernel, that includes all initial genes. The *pam*[15] clustering algorithm is employed to partition the data into semantically similar groups. The number of clusters k is defined as the maximum k so that all initial genes are contained in the same cluster. Thus, partitioning the data into $k+1$ groups split the group of initial genes apart (e.g. *SCN5A* and *CACNA1C*). This yielded a set of semantically coherent genes. The overall process is graphically described in Fig. 2.



Sodium channel protein type 5 subunit alpha

- Gene name: SCN5A
- UniProt ID: Q14524
- Semantic annotation
  - Biological Process
    - sodium ion transport
    - ion transport
    - muscle contraction
    - regulation of heart contraction
    - blood circulation
    - transport
  - Molecular Function
    - voltage-gated sodium channel activity
    - ion channel activity
    - sodium channel activity
    - voltage-gated ion channel activity
    - sodium ion binding
    - protein binding
  - Cellular Component
    - voltage-gated sodium channel complex
    - membrane
    - integral to membrane
    - sarcolemma
    - T-tubule
    - intercalated disc
    - plasma membrane

Related to diseases:

- progressive familial heart block type I (PFHBI) [MIM:113900]
- progressive cardiac conduction defect (PCCD)
- long QT syndrome type 3 (LQT3) [MIM:603830]
- Brugada syndrome (BRS1) [MIM:601144]
- sick sinus syndrome type 1 (SSS1) [MIM:608567]
- idiopathic ventricular fibrillation (IVF) [MIM:603829]
- sudden infant death syndrome (SIDS) [MIM:272120]
- familial atrial standstill [MIM:108770]
- cardiomyopathy dilated type 1E (CMD1E)

Fig. 1. Description of protein produced by gene SCN5A.

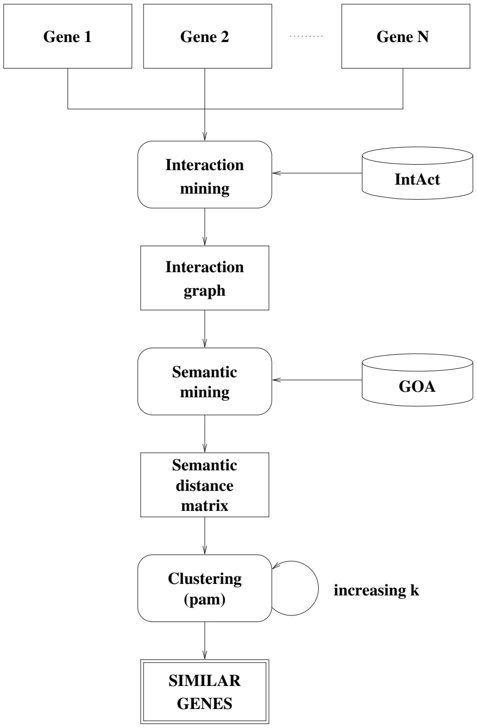


Fig. 2. Graphical description of the proposed methodology.

III. RESULTS

Coronary diseases are polygenic and highly dependent on many environmental factors. The SCN5A gene is part of a sodium channel which mediates the voltage-dependent sodium permeability of excitable membranes. When it is de-

TABLE I

PROTEINS THAT BELONG TO THE SAME CLUSTER AS THE SET OF INPUT PROTEINS.

UniProt ID	Gene ID	Gene name	Known relations with disease
Q14643	ITPR1	Inositol 1,4,5-trisphosphate receptor type 1	Spinocerebellar ataxia type 15 (SCA15) (SCA15) [MIM:606658]
P54284	CACNB3	Voltage-dependent L-type calcium channel subunit beta-3	
Q14500	KCNJ12	ATP-sensitive inward rectifier potassium channel 12	
P48050	KCNJ4	Inward rectifier potassium channel 4	
P21817	RYR1	Ryanodine receptor 1	Malignant hyperthermia susceptibility type 1 (MHS1) [MIM:145600] — Central core disease of muscle (CCD) [MIM:117000] — Multiminicore disease with external ophthalmoplegia (MMDO) [MIM:255320]
O00555	CACNA1A	Voltage-dependent P/Q-type calcium channel subunit alpha-1A	Spinocerebellar ataxia type 6 (SCA6) [MIM:183086] — Familial hemiplegic migraine (FHM) [MIM:141500] — Episodic ataxia type 2 (EA2) [MIM:108500]
Q15413	RYR3	Ryanodine receptor 3	
Q14573	ITPR3	Inositol 1,4,5-trisphosphate receptor type 3	
P28472	GABRB3	Gamma-aminobutyric acid receptor subunit beta-3	Chronic insomnia — Childhood absence epilepsy type 5 (ECA5) [MIM:612269]
P36543	ATP6V1E1	V-type proton ATPase subunit E 1	
O00305	CACNB4	Voltage-dependent L-type calcium channel subunit beta-4	Idiopathic generalized epilepsy (IGE) [MIM:600669] — Juvenile myoclonic epilepsy (EJM) [MIM:606904]
P48995	TRPC1	Short transient receptor potential channel 1	
Q9Y2W7	KCNIP3	Calsenilin	
Q9UBN4	TRPC4	Short transient receptor potential channel 4	
P39086	GRIK1	Glutamate receptor, ionotropic kainate 1	
Q9UL62	TRPC5	Short transient receptor potential channel 5	
Q13507	TRPC3	Short transient receptor potential channel 3	
Q9Y210	TRPC6	Short transient receptor potential channel 6	Focal segmental glomerulosclerosis 2 (FSGS2) [MIM:603965]
Q9UHC3	ACCN3	Amiloride-sensitive cation channel 3	
P78348	ACCN2	Amiloride-sensitive cation channel 2, neuronal	
Q9HCX4	TRPC7	Short transient receptor potential channel 7	
P35498	SCN1A	Sodium channel protein type 1 subunit alpha	Generalized epilepsy with febrile seizures plus type 2 (GEFS+2) [MIM:604233] — Severe myoclonic epilepsy in infancy (SMEI) [MIM:607208] — Intractable childhood epilepsy with generalized tonic-clonic seizures (ICEGTC) [MIM:607208] — Familial hemiplegic migraine 3 (FHM3) [MIM:609634] — Familial febrile convulsions type 3 (FEB3) [MIM:604403]
Q6PIL6	KCNIP4	Kv channel-interacting protein 4	

fective, it modifies the normal Electrocardiogram (ECG) and for some individuals could cause syncope and sudden death. On the other hand, mutations in subunits of calcium channel (CACNA1C) have also been reported for Brugada syndrome related phenotypes. Table I shows the proteins included in the same cluster than the previous proteins, which are able to modify the ECG. These proteins are the most semantically similar to original set (*SCN5A* and *CACNA1C*) in an interaction network. Most proteins in the table are related with voltage dependent calcium channels which are necessary for synaptic transmission. The exocytosis of neurotransmitter vesicles is produced by calcium entrance. Moreover, once the transmission has finished, cell homeostatic equilibrium must be recovered releasing the excess of cations. Any defect in neurotransmission regulation could be the cause of lacks of coordination, cerebral disorders or metabolic problems. Because of this, it is a process strongly controlled in all the stages of cell regulation. From the set of proteins selected *Q9Y2W7* (*Calsenilin*) modulates the transcription of proteins by binding elements of genes in response to calcium concentrations in cells. *Calsenilin* modifies channels density, inactivation kinetics and rate of recovery from inactivation. Other proteins in Table I are directly part of cation channels and some of them have been previously reported to be related with neuromuscular diseases. Calcium channels *Q14643* and *O00555* have been found to be associated with *ataxia* (i.e. lack of coordination of muscle movements). Furthermore, defects in genes coding for *O00305* and *P28472* are associated with neurologic symptoms such as *epilepsy*, *insomnia* and *familiar hemiplegic migraine*, respectively. Ion channels are proteins in membrane which allow the pass of ions in order to maintain a charge concentration gradient between the intracellular and extracellular space. The current conduction generates the action potential necessary for excitable cell functions. An increase of calcium in cytosol is necessary for muscular contraction. Alterations in *P21817* functions interfere in correct sarcoplasmatic reticulum calcium release and in the T-tubules depolarization necessary for muscle contraction. Different myopathies have been related with mutations in this gen. In addition, other proteins connected with the permeability for ions (*P48050*, *P21817* or *Q9UHC3*) are necessary for many processes dependent on osmotic equilibrium. The overall set of obtained proteins is therefore coherent from a biological point of view.

From a computational perspective, obtaining the minimum connecting graph requires checking at every iteration whether the graph is connected. This can be done in time $O(|V| + |E|)$, where V is the set of nodes and E is the set of edges. However, for very large graphs it is time consuming. In our case $|V| = 665$ and $|E| = 858$. Calculating semantic similarity between all pairs of genes is the most time and memory consuming task, where $\frac{|V|^2}{2}$ values have to be computed. Each calculation involves two searches in a pre-calculated annotation list, several searches in a pre-calculated IC table, two sums — the intersection and the union — and a division. A search in the annotation list can be done in the worst case in $O(|V|)$. Each search in the IC table will take $O(|T|)$ where T is the set of all different semantic terms. The number of searches will depend on the average number of annotations per gene, a. Thus, the final cost of this step is:

$$cost = O(\frac{|V|^2}{2} \cdot |V| \cdot a \cdot |T|) = O(|V|^3 \cdot a \cdot |T|) \qquad (3)$$

considering constant sum and division time. Although polynomial, this cost is very non-linear, and it can be very time expensive for high values of $|V|$ and $|T|$. In this study, $a \simeq 14$ and $|T| = 2157$. Finally, if the starting set of genes is semantically very heterogeneous, calculating the semantic hypersphere that contains them might yield a very large portion of the genes in the graph, which would not be of much interest.

IV. CONCLUSIONS AND FUTURE WORK

From two genes related with a hereditary anomaly that affects heart contraction and cardiac conduction (altering the patient Electrocardiogram), the proposed method is able to obtain a set of genes that have a clear relation with synaptic transmition and muscle contraction. Moreover, it has been reported that mutations in some of them produce different physiological disorders related with these physiological processes. The method is based on mining protein-protein interaction databases and organizing the set of related proteins through a semantic measure defined from their GO terms. The proposed methodology provides for an automatic way of mining existing curated literature and exploiting it for phenotype related candidate genes generation. This method can effectively be used for finding new candidates related to a disease but also for confirming candidate genes obtained from experimental data, (e.g. microarray datasets of association or linkage analysis).

V. ACKNOWLEDGMENTS

CIBER de Bioingeniería,Biomateriales y Nanomedicina is an initiative of the ISCIII.

REFERENCES

[1] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. S. Yeh, "Uniprot: the universal protein knowledgebase." *Nucleic Acids Res*, vol. 32, no. Database issue, January 2004.

[2] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez gene: gene-centered information at ncbi." *Nucleic Acids Res*, vol. 35, no. Database issue, January 2007.

[3] H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, and R. Apweiler, "Intact: an open source molecular interaction database." *Nucleic Acids Res*, vol. 32, no. Database issue, January 2004.

[4] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "Biogrid: a general repository for interaction datasets," *Nucleic acids research*, vol. 34, no. suppl. 1, pp. D535–539, January 1 2006.

[5] L. Matthews, G. Gopinath, M. Gillespie, M. Caudy, D. Croft, B. de Bono, P. Garapati, J. Hemish, H. Hermjakob, B. Jassal, A. Kanapin, S. Lewis, S. Mahajan, B. May, E. Schmidt, I. Vastrik, G. Wu, E. Birney, L. Stein, and P. D'Eustachio, "Reactome knowledgebase of human biological pathways and processes," *Nucleic acids research*, vol. 37, no. suppl. 1, pp. D619–622, January 1 2009.

[6] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. the gene ontology consortium," *Nature genetics*, vol. 25, no. 1, pp. 25–29, May 2000.

[7] D. Barrell, E. Dimmer, R. P. Huntley, D. Binns, C. O'Donovan, and R. Apweiler, "The goa database in 2009–an integrated gene ontology annotation resource," *Nucl.Acids Res.*, p. gkn803, October 2008.

[8] P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language," *Journal of Artificial Intelligence Research*, vol. 11, pp. 95–130, 1999.

[9] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *eprint arXiv:cmp-lg/9709008*, September 1997, p. 9008.

[10] S. Falcon and R. Gentleman, "Using gostats to test gene lists for go term association." *Bioinformatics*, vol. 23, no. 2, pp. 257–258, 2007.

[11] C. Pesquita, D. Faria, H. Bastos, A. E. Ferreira, A. O. Falcao, and F. M. Couto, "Metrics for go based protein semantic similarity: a systematic evaluation," *BMC Bioinformatics*, vol. 9, p. S4, 2008, pT: C; CT: 10th Bio-Ontologies-Special-Interest-Group Workshop; CY: JUL 20, 2007; CL: Vienna, AUSTRIA; SU: Suppl. 5.

[12] M. Chagoyen, J. M. Carazo, and A. Pascual-Montano, "Assessment of protein set coherence using functional annotations," *BMC Bioinformatics*, vol. 9, p. 444, Oct 20 2008.

[13] K. Ovaska, M. Laakso, and S. Hautaniemi, "Fast gene ontology based clustering for microarray experiments," *BioData mining*, vol. 1, no. 1, p. 11, Nov 21 2008.

[14] J. Chen, B. J. Aronow, and A. G. Jegga, "Disease candidate gene identification and prioritization using protein interaction networks," *BMC Bioinformatics*, vol. 10, p. 73, Feb 27 2009.

[15] M. Maechler, P. Rousseeuw, A. Struyf, and M. Hubert, "Cluster analysis basics and extensions," 2005, [1].

[1] Rousseeuw et al provided the S original which has been ported to R by Kurt Hornik and has since been enhanced by Martin Maechler: speed improvements, silhouette() functionality, bug fixes, etc. See the 'Changelog' file (in the package source)