SNPRanker: a tool for identification and scoring of SNPs associated to target genes

Andrea Calabria[1,2,3], Ettore Mosca[1,2], Federica Viti[1], Ivan Merelli[1] and Luciano Milanesi[*,1]

[1]Institute for Biomedical Technologies, National Research Council, via F.lli Cervi 93, 20090, Segrate (MI), Italy, `http://www.itb.cnr.it`

[2]Dept. of Informatics, Systems and Communication, University of Milano Bicocca, Viale Sarca, 336, 20126 Milano, Italy, `http://www.disco.unimib.it`

[3]Dept. of Medicine, Surgery and Dentistry, University of Milano, Via Di Rudini 8, 20142 Milano, Italy, `http://www.unimi.it`

Summary

The identification of genes and SNPs involved in human diseases remains a challenge. Many public resources, databases and applications, collect biological data and perform annotations, increasing the global biological knowledge. The need of SNPs prioritization is emerging with the development of new high-throughput genotyping technologies, which allow to develop customized disease-oriented chips. Therefore, given a list of genes related to a specific biological process or disease as input, a crucial issue is finding the most relevant SNPs to analyse. The selection of these SNPs may rely on the relevant a-priori knowledge of *biomolecular features* characterising all the annotated SNPs and genes of the provided list. The bioinformatics approach described here allows to retrieve a ranked list of significant SNPs from a set of input genes, such as candidate genes associated with a specific disease. The system enriches the genes set by including other genes, associated to the original ones by ontological similarity evaluation. The proposed method relies on the integration of data from public resources in a vertical perspective (from genomics to systems biology data), the evaluation of features from biomolecular knowledge, the computation of partial scores for SNPs and finally their ranking, relying on their global score. Our approach has been implemented into a web based tool called *SNPRanker*, which is accessible through at the URL `http://www.itb.cnr.it/snpranker` . An interesting application of the presented system is the prioritisation of SNPs related to genes involved in specific pathologies, in order to produce custom arrays.

1 Introduction

In the recent past years, genotyping technologies knew a great development and studies about genotype markers are increasing in importance [1, 2]. Among them, the evaluation of Single Nucleotide Polimorphisms, also known as SNPs, is revealing very promising. SNPs are nowadays widely exploited for Genome Wide Association Studies (GWAS) [13, 4, 5], identification of Copy Number Variations (CNV) [6], observations of Population Stratification [7] and so

[*] E-Mail Addresses: {andrea.calabria; ettore.mosca; federica.viti; ivan.merelli; luciano.milanesi}@itb.cnr.it

doi:10.2390/biecoll-jib-2010-138

forth. Since SNPs represent established genomic differences, their knowledge can be exploited to characterize each subject from others on study by correlating specific phenotype with a corresponding genomic pattern.
The total number of SNPs in the whole human genome exceeds 12 millions and each SNP is related to different genomic properties depending on its position on the DNA strand: i.e. a SNP can be located within inter-gene regions or within intra-gene ones. Nowadays chip technologies allow to analyze up to one million SNPs for each patient due to chemical and physical limits that affect probe density. To overcome this limit, together with technological improvements, researchers are trying to define reasonable strategies to filter the initial amount of 12 millions SNPs.
The first approach to optimize the SNPs probeset relies on the concept of Linkage Disequilibrium [8] (LD). LD mapping exploits a statistical similarity measure between adjacent SNPs and computes how much two SNPs are related each other, thus defining what is the genetic information improvement using both or just one of them. LD mapping is thus used to optimize the information contained into 1 million SNPs arrays. The second method able to reduce the number of SNP probes within a chip regards the possibility to create disease-oriented chip. This approach not only allows adapting the analysis to specific genetic studies but even to produce smaller arrays. In fact, this aspect relies on a crucial topic: which strategy can be followed to select the subset of SNPs suitable to create a specific disease-oriented chip.

The presented work is related to SNPs' probeset identification for producing genotyping arrays dedicated to pathologies, starting from genes or biological processes involved in such diseases. The tool scores different biomolecular features for SNPs associated to a set of genes, that is given as input and that can be expanded through an ontology-based engine. Once the SNPs final scores are computed, the system provides a ranked list of the most significant SNPs associated to the input set of genes.
Moreover *SNPRanker* can be used for gene enrichment through the identification of the ontological similarity between the input genes and whole set of human genes. This allows to extend the initial gene list, by including also genes presenting a similar biological function (according to the considered ontology) thus potentially involved in the same disease.

The paper is organized as follows. Section 2 describes the related works; Section 3 lists the functional elements the system is composed by; Section 4 describes the activities pipeline performed by the system; Section 5 overviews the system's capabilities and presents a use case; Section 6 provides conclusions and future works.

2 Related works

The analysis of genomic variations is usually performed following two main approaches: statistical methods or machine learning techniques. The main difference is that while the exploitation of statistical methods requires a data model involving a set of a-priori hypotheses and parameters values, the use of machine learning approaches do not need a-priori evaluations, since models and rules are derived directly from a training set of data and the system is trained to fit a general model which will be then adopted for all other data. An alternative solution is offered by data mining approaches where users can visualize, plot and reorder data without fixed models and, if the system supports customizations, they can also validate their own new models on data and infer new knowledge.

Statistical methods are widely used in epidemiology and got many positive results in application studies [9, 10]. Nevertheless, statistical approaches are often computationally intensive, especially when dealing with large amount of data produced by high-throughput techniques, and this approach is often impracticable for most of the research laboratories. Therefore, many scientists found machine learning approaches very attractive: so far many studies exist exploiting machine learning approaches and are providing encouraging results [11, 12].
Machine learning methods are among the most promising approaches due to the flexibility and adaptation to data. When exploiting *supervised* methods, the training set must be carefully selected, since the model is created on it: the training set must therefore embed all peculiarities of the considered data type, thus allowing the model a-priori knowledge. In our case data are single nucleotide genomic variations (SNPs) and the a-priori knowledge is represented by features that characterize each SNP or the related gene and protein.
Data mining methods are mostly employed in business fields, for example for intelligent customer support and business analyses. In the genetic context a few significant works have been produced, for example in [13] the authors mine SNPs from families; we need a tool for scoring SNPs based on a priori information and where users can infer knowledge by setting parameters, like data mining facilities usually support.
Only a few applications exploit machine learning in genotyping context. An example, concerning the genes ranking, is the so called *gene prioritization* [14], a method that, given a set of training genes, considers a number of features from them, which represent the a-priori knowledge available from multiple data sources. Given a set of test genes, the cited system computes features values and ranks test genes with respect to their similarities, achieving the prioritization and highlighting the most important genes, with respect to the selected features.

No methods are available in literature to achieve, in SNPs context, results similar to gene prioritization through data mining and machine learning approach. The paper presents a new method for evaluating SNPs by features scoring.

3 System's functional components

The core of the designed system can be decoupled into four levels:

- *gene list enrichment*: exploitation of gene ontological annotation to enrich the initial list of genes provided as input;
- *data integration*: creation of a database for the integration of biomolecular knowledge retrieved from public sources;
- *features set*: choice of the features characterizing each SNP;
- *evaluation function*: definition of the function that provides a final score for each SNP.

Following subsections present in details the levels mentioned above and the related characteristics of the implemented prototype.

3.1 Data integration

The first step concerns the design of a database for integrating genes and genes products information, in order to provide a solid knowledge base on which the whole SNP-scoring system relies on. For this purpose a systems biology oriented and ontology-based database has been developed, which integrates data from a wide range of public resources.

It exploits a MySQL server and relies on a data warehouse approach, which consists in collecting and transforming heterogeneous data from different sources, to allow their integration and accessibility. This approach is typical of data integration models and differs from data integrity models, often used for normalized databases which are widely used to maintain primary resources.

The database is gene-centric and considers at the moment only human genes which are annotated, among other features, by symbol, description, aliases and sequences. Data about SNP are downloaded from dbSNP [15], which allows to integrate data about chromosomal and *contig* position, heterozygosity, alleles and function of the related DNA portion. Moreover, gene products have been collected as list of mRNAs sequences and related protein isoforms according to the NCBI RefSeq annotations (NCBI Nucleotide [15]). Data about proteins include all the identifiers suitable to download the related sequences, functional domains [16, 17] and structural models from the Protein Data Bank [18].

The systems biology perspective leads to consider data such as the list of the biochemical pathways (KEGG [19] and Reactome [20]) where human gene products are involved, and information about protein-protein interactions (PPIs), collected from BioGRID [21], which complement knowledge about pathways and enable crucial network based analysis.

The peculiarity of the developed database is represented by the multi-level approach to data integration, which enables a more comprehensive view of the examined process or disease: therefore, it should lead to a better selection of the set of SNPs to be included in a disease-oriented custom chip.

Finally, in order to provide a standard framework for data integration and a reliable engine for SNPs selection, the database has been built on a strong ontology layer. Whenever available, data have been annotated using ontological terms: Gene Ontology [22] for genes and KEGG Pathway ontology (derived from the hierarchical organization of KEGG pathways) for pathways are just some of the hierarchically structured vocabularies that underlie the infrastructure. Additionally, ontology structures allow to improve the performance of statistical and analytical evaluations by means of the graphs that undergo the hierarchically structured vocabularies and that shed light on the relationships between biological components.

3.2 Features set

The term “features” indicates a set of characteristics, related to each SNP, the machine learning approach relies on, in a direct way or through the gene and the gene product knowledge. The set of features chosen to characterize each SNP represents an a-priori knowledge of the scientific problem.

Since the described approach considers as input a list of genes, the system must consider biomolecular elements as related to genes in a *vertical* perspective. Therefore information is considered in both a top-down view, from proteins interactions to sequences of nucleotides, and a bottom-up perspective, starting from genes characterization and climbing up to processes and complex biological systems.

The features set includes genomic (Minor Allele Frequencies, localization on the DNA sequence), proteomic (InterPro [17] domains), interactomic (hub proteins), phenotypic information (essential genes).
In the following, a complete list of the considered features will be described.

Hub proteins The evaluation of the number of protein-protein interactions (PPIs) established by a protein, i.e. the degree of the protein in the PPIs network, is an important aspect in assessing the biological relevance of a SNP occurring within or close to the gene that encodes for that protein. Indeed, it is known that *hub* proteins play a crucial role for the cell functioning and, in fact, are often encoded by essential genes. The database used in this work integrates PPIs data from HPRD [23] and BioGRID. By means of this information it is possible to score a SNP (x) associated to a gene with the following characteristic function, which considers the number of PPIs k_x established by the protein encoded in the gene:

$$f_1(x) = \begin{cases} 1, & k_x \geq \alpha \\ 0, & k_x < \alpha \end{cases}$$

where α is the number of PPIs requested to be considered as hub. By default $\alpha = 20$, according to [24]: through the web interface the user is able to modify this value.

Protein domain A SNP can create a missense or even a frame shift in the coding sequence, thus causing changes in the protein amino acid sequence, which can have a deep impact on the protein function according to the region where the modification occurs. In fact, if a change is localized within a domain, thus being functionally important, its effect on the biological function is potentially more relevant than if it is placed within a inter-domain region.
Information related to the domain localization of the SNPs can be obtaining by linking InterPro Domain Architecture (IDA) data (which report the localization of the protein domain according to the amino acid position within the protein chain) with the knowledge concerning the amino acid position (which indicates the position of the amino acid modified by the SNP). This information can be easily accessed from the integrative database.
The score of this feature is provided by a characteristic function, which applies greater values to SNPs occurring into protein domains (D) encoding regions.

$$f_2(x) = \begin{cases} 1, & x \in D \\ 0, & x \notin D \end{cases}$$

Minor Allele Frequencies The Minor Allele Frequency (MAF) represents the frequency of the less frequent allele of a SNP in a specific population: it defines how much an allele (and thus its SNP) is relevant for a population. MAF score can even be employed to measure the "penetrance" of a disease in a population, in case the minor allele is more diffused in the affected phenotype than in the unaffected one in the control population.
Since we are designing a general system for different studies and applications, all populations considered in the HapMap project [25] have been included, with their MAF values. The user is required to choose the MAF of interest.
For this feature, the scoring function is expressed as the original MAF value $m_x \in \Re | (0 \leq m_x < 0.5)$ obtained from the HapMap project:

$$f_3(x) = m_x.$$

doi:10.2390/biecoll-jib-2010-138

When the "1000 Genomes Project" [26] will provide more accurate and updated values about MAF scores, we will update these values.

Localization on the DNA sequence According to the annotation provided by dbSNP [27], functional relationship between SNPs (and possibly alleles) and genes are defined. Relying on this annotation the following categories have been considered $C = \{$ "unknown", "coding-synonymous", "intron", "near-gene 3"', "near-gene 5"', "nonsense", "missense", "frameshift", "untranslated 3"', "untranslated 5"'$\}$.
According to the user's specific analysis the whole set C or subsets of it can be considered. Considering the vector $\mathbf{c} = (c_1, ..., c_10)$, whose elements $c_i \in 0, 1$ indicate the selection or the exclusion of the elements of C, the scoring function can be formalized as follows:

$$f_4(x) = \mathbf{c}^T \times \mathbf{v}$$

where $\mathbf{v}$ is a weight vector.

Essential genes Another important feature for SNP scoring is the kind of gene where the polymorphism occurs. It is known that some genes are *essential* to support cellular life, i.e. if their products are not correctly produced the cell hardly survives. The knowledge base used in this work includes data providing such information for *homo sapiens* [28]. Considering the set of human genes G and the subset of essential genes $E \subset G$, the scoring function is the following characteristic function:

$$f_5(x) = \begin{cases} 1, & x \in E \\ 0, & x \in G - E \end{cases}$$

3.3 Core scoring function

In order to obtain a significant score for each SNP, features' values must be processed through a *core scoring function*. This engine allows to compute the final SNP value as a real number, considering genes and genes' products information embedded in the defined set of features.
Given the a-priori knowledge embedded in the described set of features, the user can interact with this information in order to better adapt it to his scientific studies: this is possible by associating each feature to a *weight* that represents the importance of that feature for the calculation of the final SNP score. The default values for all the elements in the vector $\vec{w}$ of the features appears in the tool web page: this model assigns the same importance to each feature, without assuming any specific perspective while performing the analysis. The scoring function g maps the values returned by the features scoring functions $f_1, \ldots, f_5$ and the weights vector $\mathbf{w} \in \mathbb{R}^5$ to a single final value, which is used to calculate the final SNPs ranked list:

$$g : \mathbb{B} \times \mathbb{B} \times \mathbb{R} \times \mathbb{R} \times \mathbb{B} \times \mathbb{R}^5 \to \mathbb{R}$$

where $\mathbb{B} = \{0, 1\}$. The scoring function g has been defined in two forms; as the sum 1 or the product 2 (which determines a more restrictive SNPs selection) of the values returned by f_1, f_2, f_3, f_4, f_5 according to the weights $\mathbf{w}$

$$g : (f_1, f_2, f_3, f_4, f_5, \mathbf{w}) \mapsto w_1 f_1 + w_2 f_2 + w_3 f_3 + w_4 f_4 + w_5 f_5 \tag{1}$$

$$g : (f_1, f_2, f_3, f_4, f_5, \mathbf{w}) \mapsto w_1 f_1 \times w_2 f_2 \times w_3 f_3 \times w_4 f_4 \times w_5 f_5 \tag{2}$$

3.4 Gene list enrichment

Ontologies are controlled vocabularies hierarchically organized. Their exploitation allows not only the use of standardized and recognized descriptive terms, but even to infer relations among objects that are annotated through ontologies. In the implemented system the ontology layer is exploited, other than for annotation aims, to enrich the list of genes provided to the system as input. From the input genes list g_1 the system generates the list $g_2 \supseteq g_1$, which includes also the genes that present a high semantic similarity with the genes in g_1. Four main methods exist in literature to carry on this task: three methods [29, 30, 31] determine the semantic similarities of two terms based on their distances to the closest common ancestor term and/or the annotation statistics of their common ancestor terms.

A crucial drawback of these methods is that the distances to the closest common ancestor term cannot accurately represent the semantic difference of two terms: if two terms sharing the same parent are near the root of the ontology, thus being more general and less informative, they should have larger semantic difference than two terms having the same parent and being far away from the root of the ontology. Moreover, measuring the semantic similarity of two ontological terms based only on the number of common ancestor terms cannot discern the semantic contributions of the ancestor terms to these two specific terms. The fourth method [32] evaluates these limits and provides an alternative solution for measuring ontological terms similarity: the measure is based on the graph of the considered ontology.

Within the proposed system, the [32] and the [31] strategies have been implemented, thus providing the user the possibility to choose the preferred method.

4 System design

The schema of the designed system is presented in Figure 1. The system guarantees the flexibility and suitability to users scientific applications by allowing several parameters customizations that enable features set up values to better test hypotheses.

User can access the system, provide the input list, modify system options and retrieve the final results through a PHP and JavaScript based web interface.

Available options will be widely explained in next paragraphs and some screen-shots of the developed web site will be shown.

4.1 System input

The system takes as input an arbitrary list of genes. The dataset can arise from experimental sources (genes of interest originated from a laboratory experiment or from bibliographic research related to a specific scientific aspect): it can be provided as a list of comma separated standard gene symbols. Alternatively, the set of genes can be retrieved by considering the Gene Ontology for a specific biological process of interest, thus obtaining all genes annotated with the defined GO Biological Process term. Process selection is supported by an auto-completion JavaScript function.

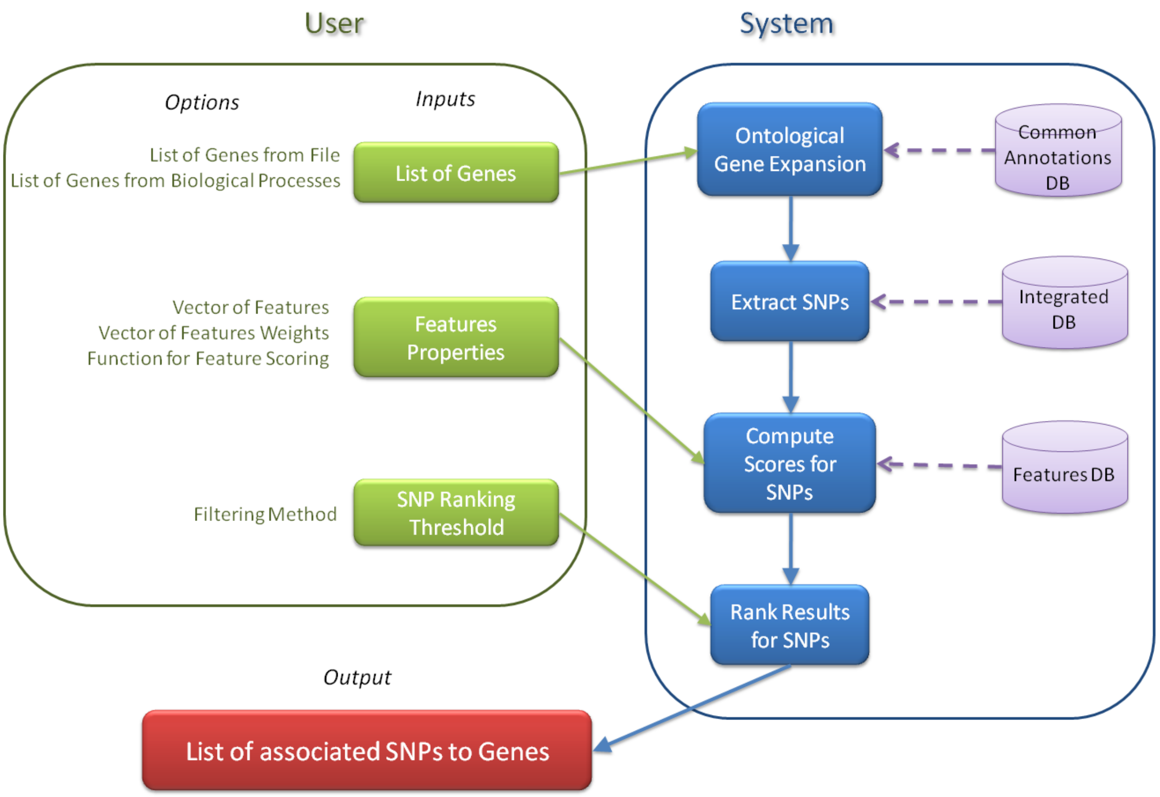


Figure 1: General schema of the system

4.2 Ontology-based expansion

In order to promote the identification of new SNPs relations and to define a custom "model" for scoring SNPs that better allows data mining and new hypotheses formulation about SNPs influence, especially on genetic pathologies, a crucial function is available within the system, which enriches the input list g_1 by adding new genes that are biologically related with them. This step is performed through the exploitation of ontology similarity measures. Depending on the interests of the user, for each gene in g_1 the system retrieves a number of genes with the highest semantic similarity according to a specific ontology, such as the Gene Ontology, exploiting Wang similarity measure [32] and Schlicker one [31]. Similarity score is retrieved by a pre-calculated matrix, which provides an affinity measure for all couples of genes. Since the matrix creation is computationally intensive, a pre-filtering phase has been applied in order to select couples of genes which present at least one common ontology term.

4.3 SNPs extraction

The whole system relies on considering and evaluating data and metadata concerning the biomolecular building blocks. In order to obtain the SNPs associated to the enriched list of input genes, the system performs multiple SQL queries on the database. The output of this step is the list of SNPs associated to the input genes, scored but unranked. Since genomic coordinates and thus the definitions of genes can change during assembly upgrades, we keep track of historical data and users can specify the dbSNP release on which query the system.

4.4 Score computation

The function exploited to obtain the final score associated to each SNP relying on a-priori knowledge has been introduced in section 3.3. From the application point of view, the user can widely interact with the system to perform this step. First of all the user can specify which features to adopt for SNPs scoring: actually he could be interested to consider some characteristics while neglecting others. The default condition is considering all the listed gene and gene products features. Moreover, user can freely assign different weights to each feature, according to the scientific challenge that has to be faced: in specific context some properties might be more valuable than others. By default all features have a weight assigned by authors on the basis of most common genotyping studies. User can even interact with some thresholds considered in the system. He can freely choose: whether applying the ontological expansion or not and what similarity score threshold to consider; the minimum number of interactions valid to consider a protein as a hub protein; whether to exploit a sum-based or a multiplication-based scoring function, to provide respectively similar importance to all the chosen features in the computation of the SNPs score or to select just those SNPs that are significant in a specific biological context.

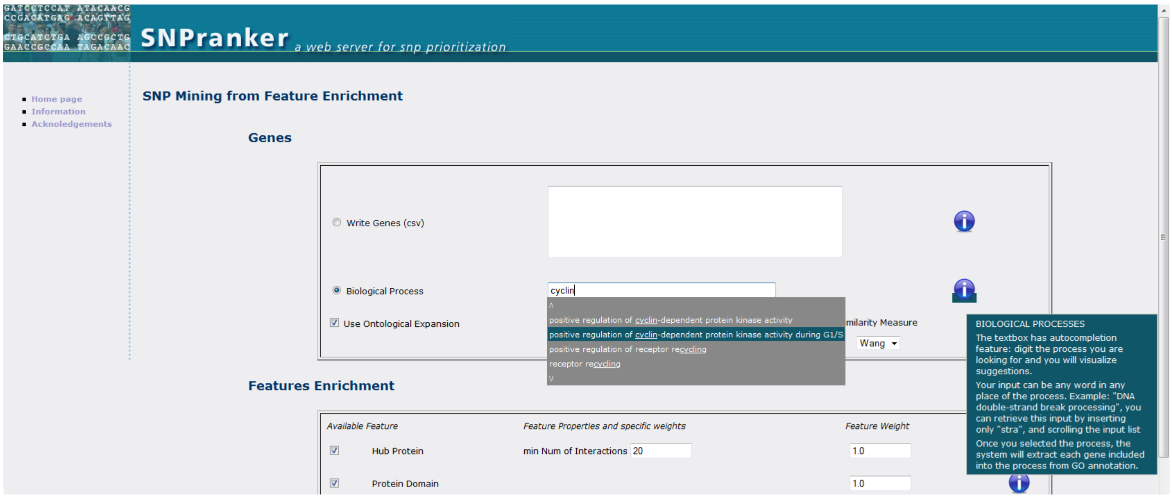
4.5 SNPs ranking and Filtering

Once all scores have been computed, the last step consists in ranking all SNPs relying on their score value. Due to the great amount of SNPs potentially reported as output, the user can decide to cut the list. In fact, before running the processing, user can select the percentile where he wants the result list to be cut. Available options are “Percentage” (followed by the corresponding threshold) or “No filters”. Final results have to be written in a file and then user can download file through the web link.

5 Results and Discussion

SNPRanker is available at the web link `http://www.itb.cnr.it/snpranker`. The developed system is aimed to support SNPs analysis, particularly interesting for helping SNPs/disease association studies. The input of the system can be either a predefined list of genes, typically whose evidences have been found related to the same pathology, or a set of genes associated to a particular biological process, as shown in Figure 2. The ontological expansion is an important tool for studying SNPs related to pathologies, since it allows to extend the analysis to SNPs that could potentially be involved in a pathology onset but that have not being highlighted by more traditional approaches. In fact, this tool permits to increase the number of SNPs in analysis even considering those belonging to genes that present similar semantic annotations with the genes initially considered. For instance, in Table 1, we show the top ranked genes showing the highest semantic similarity with the gene *CCND1* encoding for the *cyclin D1*, which controls the cell cycle process. The semantic similarity was calculated by means of the Wang’s method [32] considering the Gene Ontology Biological Process. The method successfully identifies the genes annotated similarly to CCND1.

The setting of SNPs feature weights values has been thought as a support for population genetics studies. Actually, this property allows to assign different levels of importance to diverse biomolecular aspects. For instance, depending on the aim of the specific study, it is possible



SNP Mining from Feature Enrichment

Genes

○ Write Genes (csv)

● Biological Process: cyclin

☑ Use Ontological Expansion

Similarity Measure: Wang

positive regulation of cyclin-dependent protein kinase activity
positive regulation of cyclin-dependent protein kinase activity during G1/S
positive regulation of receptor recycling
receptor recycling

BIOLOGICAL PROCESSES
The textbox has autocompletion feature: digit the process you are looking for and you will visualize suggestions.
Your input can be any word in any place of the process. Example: "DNA double-strand break processing", you can retrieve this input by inserting only "stra", and scrolling the input list
Once you selected the process, the system will extract each gene included into the process from GO annotation.

Features Enrichment

Available Feature		Feature Properties and specific weights	Feature Weight
☑	Hub Protein	min Num of Interactions 20	1.0
☑	Protein Domain		1.0

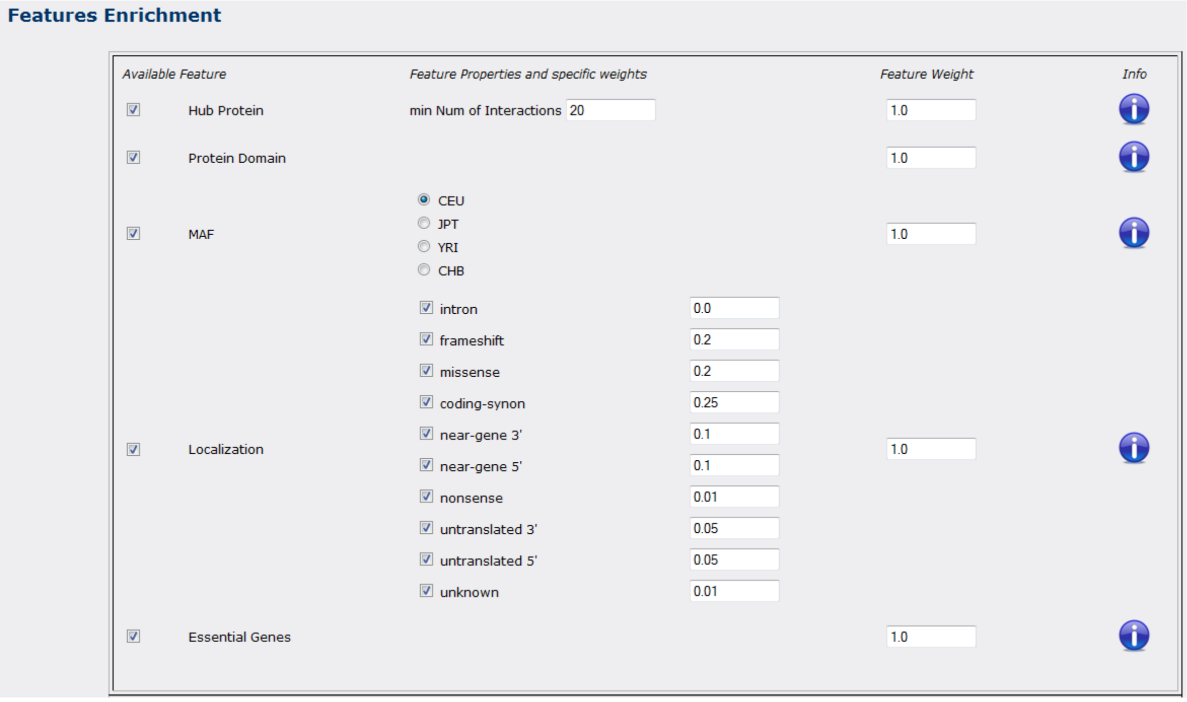
Figure 2: Screenshot from the web site. Input gene list definition.

EG ID	EG Symbol	semantic similarity score
894	CCND2	0.81
896	CCND3	0.81
8941	CDK5R2	0.72
56647	BCCIP	0.72
28984	C13orf15	0.72

Table 1: Wang's method [32] semantic similarity scores using the GO Biological Processes: the top 5 ranked genes are listed.

to assign a high relevance to SNPs associated to hub proteins or SNPs occurring in regulatory regions such as the 5' near gene region. An overview of these possibilities is provided in Figure 3. Finally, the retrieval of the scored SNPs ranked list is specifically aimed to support evaluation of genetic diseases.

Starting from the queried gene CCND1 and its semantically more similar genes (CCND2, CCND3, BCCIP, CDK5R, C13orf15) a list of ranked SNPs has been obtained exploiting the developed core scoring function. The whole list includes more than 1500 SNPs characterized by diverse features and thus assuming different degree of importance. Considering just genes with a final score ≥ 0.01 the list can be reduced to 499 genes. Part of the results obtained from CCND1 gene is reported in Figure 4. At the top of the ranked list there are 4 SNPs belonging to CCND1 (which codes for an hub protein), and placed on functional protein domains. Information about essential genes in not exploited in the described example, since no genes from the input list (nor in the original version neither in the enriched one) are labeled as essential for life. An interesting parameter is represented by DNA localization. Within this feature crucial information related to the position of the SNP on the DNA chain is included: different forms of localization can be considered and a weight can be provided to each of them, according to the specific use case. In particular, referring to the considered example, while performing evaluations about DNA localization of listed SNPs it results that most of them are placed on intronic regions (around 60%) on DNA: this is obvious considering the high percentage of intronic regions on DNA strand compared to the esonic areas. According to the considered weight values SNPs present on introns are localized at lower positions within the ranked list, while at the



Features Enrichment

Available Feature	Feature Properties and specific weights	Feature Weight	Info
☑ Hub Protein	min Num of Interactions 20	1.0	ⓘ
☑ Protein Domain		1.0	ⓘ
☑ MAF	● CEU ○ JPT ○ YRI ○ CHB	1.0	ⓘ
☑ Localization	☑ intron 0.0 ☑ frameshift 0.2 ☑ missense 0.2 ☑ coding-synon 0.25 ☑ near-gene 3' 0.1 ☑ near-gene 5' 0.1 ☑ nonsense 0.01 ☑ untranslated 3' 0.05 ☑ untranslated 5' 0.05 ☑ unknown 0.01	1.0	ⓘ
☑ Essential Genes		1.0	ⓘ

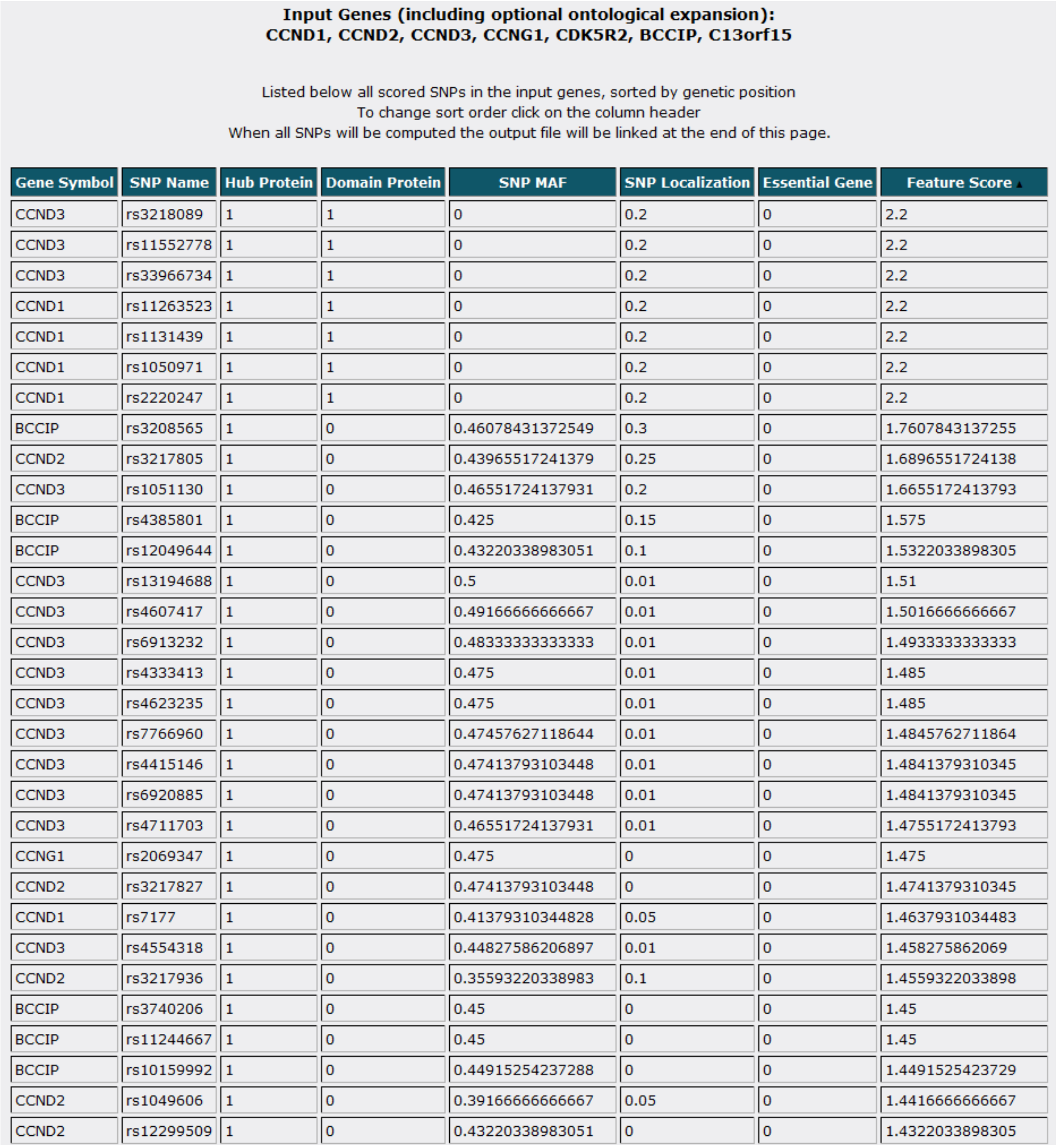
Figure 3: Screenshot from the web site. Features selection and weights values setting.

top of it many “frameshift” and “missense” are concentrated. The latter localization types are only around the 3% of the whole set but SNPs in these positions obtained higher scores. Descending the ranked list, SNPs occur that are localised in UTR regions and near the gene. The “unknown” value is important for covering other DNA regions.

6 Conclusions and Future works

The presented work concerns a system aimed at supporting SNPs based pathologies and biological pathways studies, especially in the context of clinical genotyping chips single-disease oriented. It relies on the identification of a set of crucial features characterizing each SNP related to a list of input genes. This represents the a-priori knowledge and allows to assign a final score, that can be modulated by the user according to the considered scientific aspect, to each SNP. A ranked list of SNPs is retrieved. The system can be exploited to identify the most important SNPs in population genetics studies.

Although SNPs scoring features have been carefully selected by performing an overview of all the biomolecular available knowledge, the system still lacks a solid reference to disease direct knowledge. The availability of even sparse data associating specific SNPs to defined pathologies can in fact be relevant for a more precise evaluation of SNPs importance in disease onset. Future developments of the presented system will consist in the identification and integration of well established disease-oriented databases, such as OMIM [33] which contains manually curated data about clinical evidences of genes/pathologies correlations, and other sources of *non-mendelian* diseases.



Input Genes (including optional ontological expansion):
CCND1, CCND2, CCND3, CCNG1, CDK5R2, BCCIP, C13orf15

Listed below all scored SNPs in the input genes, sorted by genetic position
To change sort order click on the column header
When all SNPs will be computed the output file will be linked at the end of this page.

Gene Symbol	SNP Name	Hub Protein	Domain Protein	SNP MAF	SNP Localization	Essential Gene	Feature Score ▴
CCND3	rs3218089	1	1	0	0.2	0	2.2
CCND3	rs11552778	1	1	0	0.2	0	2.2
CCND3	rs33966734	1	1	0	0.2	0	2.2
CCND1	rs11263523	1	1	0	0.2	0	2.2
CCND1	rs1131439	1	1	0	0.2	0	2.2
CCND1	rs1050971	1	1	0	0.2	0	2.2
CCND1	rs2220247	1	1	0	0.2	0	2.2
BCCIP	rs3208565	1	0	0.46078431372549	0.3	0	1.7607843137255
CCND2	rs3217805	1	0	0.43965517241379	0.25	0	1.6896551724138
CCND3	rs1051130	1	0	0.46551724137931	0.2	0	1.6655172413793
BCCIP	rs4385801	1	0	0.425	0.15	0	1.575
BCCIP	rs12049644	1	0	0.43220338983051	0.1	0	1.5322033898305
CCND3	rs13194688	1	0	0.5	0.01	0	1.51
CCND3	rs4607417	1	0	0.49166666666667	0.01	0	1.5016666666667
CCND3	rs6913232	1	0	0.48333333333333	0.01	0	1.4933333333333
CCND3	rs4333413	1	0	0.475	0.01	0	1.485
CCND3	rs4623235	1	0	0.475	0.01	0	1.485
CCND3	rs7766960	1	0	0.47457627118644	0.01	0	1.4845762711864
CCND3	rs4415146	1	0	0.47413793103448	0.01	0	1.4841379310345
CCND3	rs6920885	1	0	0.47413793103448	0.01	0	1.4841379310345
CCND3	rs4711703	1	0	0.46551724137931	0.01	0	1.4755172413793
CCNG1	rs2069347	1	0	0.475	0	0	1.475
CCND2	rs3217827	1	0	0.47413793103448	0	0	1.4741379310345
CCND1	rs7177	1	0	0.41379310344828	0.05	0	1.4637931034483
CCND3	rs4554318	1	0	0.44827586206897	0.01	0	1.458275862069
CCND2	rs3217936	1	0	0.35593220338983	0.1	0	1.4559322033898
BCCIP	rs3740206	1	0	0.45	0	0	1.45
BCCIP	rs11244667	1	0	0.45	0	0	1.45
BCCIP	rs10159992	1	0	0.44915254237288	0	0	1.4491525423729
CCND2	rs1049606	1	0	0.39166666666667	0.05	0	1.4416666666667
CCND2	rs12299509	1	0	0.43220338983051	0	0	1.4322033898305

Figure 4: Screenshot from the web site. Results page showing SNPs ranked list obtained from CCND1 example, discussed in the text. SNPs are sorted by features score. Ontological expansion is performed with similarity threshold set to 0.7; chosen features weights are shown in Figure 3.

Acknowledgements

This work has been designed to solve a problem on the project Hypergenes. This work has been supported by the Hypergenes European Network for Genetic-Epidemiological Studies (HEALTH-F4-2007-20150), NET2DRUG, EGEE-III, BBMRI, EDGE European projects and by the MIUR FIRB LITBIO (RBLA0332RH), ITALBIONET (RBPR05ZK2Z), BIOPOPGEN (RBIN064YAT), CNR-BIOINFORMATICS initiatives.

References

[1] Paul I W de Bakker, Roman Yelensky, Itsik Pe'er, Stacey B Gabriel, Mark J Daly, and David Altshuler. Efficiency and power in genetic association studies. *Nat Genet*, 37(11):1217–1223, Nov 2005.

[2] David B Goldstein and Gianpiero L Cavalleri. Genomics: understanding human diversity. *Nature*, 437(7063):1241–1242, Oct 2005.

[3] Heping Zhang, Lei Liu, Xueqin Wang, and Jeffrey R Gruen. Guideline for data analysis of genomewide association studies. *Cancer Genomics Proteomics*, 4(1):27–34, 2007.

[4] P. C. Sham, S. S. Cherny, and S. Purcell. Application of genome-wide snp data for uncovering pairwise relationships and quantitative trait loci. *Genetica*, 136(2):237–243, Jun 2009.

[5] William Paul Hanage and David Michael Aanensen. Methods for data analysis. *Methods Mol Biol*, 551:287–304, 2009.

[6] Gloria W C Tam, Richard Redon, Nigel P Carter, and Seth G N Grant. The role of dna copy number variation in schizophrenia. *Biol Psychiatry*, 66(11):1005–1012, Dec 2009.

[7] Hemant K Tiwari, Jill Barnholtz-Sloan, Nathan Wineinger, Miguel A Padilla, Laura K Vaughan, and David B Allison. Review and evaluation of methods correcting for population stratification with a focus on underlying statistical principles. *Hum Hered*, 66(2):67–86, 2008.

[8] David Altshuler, Mark J Daly, and Eric S Lander. Genetic mapping in human disease. *Science*, 322(5903):881–888, Nov 2008.

[9] Claire Infante-Rivard, Lucia Mirea, and Shelley B Bull. Combining case-control and case-trio data from the same population in genetic association analyses: overview of approaches and illustration with a candidate gene study. *Am J Epidemiol*, 170(5):657–664, Sep 2009.

[10] Peter J Taub and Emily Westheimer. Biostatistics. *Plast Reconstr Surg*, 124(2):200e–208e, Aug 2009.

[11] Jianlin Cheng and Pierre Baldi. A machine learning information retrieval approach to protein fold recognition. *Bioinformatics*, 22(12):1456–1463, Jun 2006.

[12] L. Hamel, N. Nahar, M. S. Poptsova, O. Zhaxybayeva, and J. P. Gogarten. Unsupervised learning in detection of gene transfer. *J Biomed Biotechnol*, 2008:472719, 2008.

[13] Zhang Fan, Li Xia, Krivosheev Ivan, Viktorovich, Gong Binsheng, Du Lei and Li Chunquan. A Heuristic Approach for Target SNP Mining Based on Genome-Wide IBD Profile. *ICNC '07 Proceedings* IEEE Computer Society, 2007:227–232.

[14] Stein Aerts, Diether Lambrechts et al. Gene prioritization through genomic data fusion. *Nat Biotechnol*, 24(5):537–544, May 2006.

[15] Benson DA Bryant SH et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 37(Database issue):D5–D15, 2009.

[16] UniProt Consortium. The universal protein resource (uniprot). *Nucleic Acids Res*, 37(Database issue):D169–D174, 2009.

[17] http://www.ebi.ac.uk/interpro/

[18] Nakamura H Markley JL Berman H, Henrick K. The worldwide protein data bank (wwpdb): ensuring a single, uniform archive of pdb data. *Nucleic Acids Res*, 35(Database issue):D301–D303, 2007.

[19] Kanehisa M Aoki-Kinoshita KF. Gene annotation and pathway mapping in kegg. *Methods Mol Biol*, 396:71–91, 2007.

[20] Gillespie M Caudy M et al. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res*, 37(Database issue):D619–D622, 2009.

[21] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic Acids Res*, 34(Database issue):D535–D539, Jan 2006.

[22] The Gene Ontology Consortium. The gene ontology's reference genome project: a unified framework for functional annotation across species. *PLoS Comput Biol*, 5(7), 2009.

[23] Prasad, T. S. K. et al. Human Protein Reference Database - 2009 Update. *Nucleic Acids Research*, 37, D767-72, 2009.

[24] Aragues R, Sali A, Bonet J, Marti-Renom MA, Oliva B. Characterization of protein hubs by inferring interacting motifs from protein interactions. *PLoS Comput Bio*, 3: e178, 2007.

[25] The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449, 851-861. 2007

[26] Nayanah Siva. 1000 genomes project. *Nat Biotechnol*, 26(3):256, Mar 2008.

[27] E. M. Smigielski, K. Sirotkin, M. Ward, and S. T. Sherry. dbsnp: a database of single nucleotide polymorphisms. *Nucleic Acids Res*, 28(1):352–355, Jan 2000.

[28] Ren Zhang, Hong-Yu Ou, and Chun-Ting Zhang. Deg: a database of essential genes. *Nucleic Acids Res*, 32(Database issue):D271–D272, Jan 2004.

[29] Resnik Philip. Semantic similarity in a taxonomy: An Information-Based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95-130, 1999.

[30] Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of 10th Interna-tional Conference on Research In Computational Linguistics*, 1997.

[31] Andreas Schlicker, Francisco S Domingues, Jörg Rahnenführer, and Thomas Lengauer. A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*, 7:302, 2006.

[32] James Z Wang, Zhidian Du, Rapeeporn Payattakool, Philip S Yu, and Chin-Fu Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281, May 2007.

[33] A. Hamosh, A. F. Scott, J. Amberger, D. Valle, and V. A. McKusick. Online mendelian inheritance in man (omim). *Hum Mutat*, 15(1):57–61, 2000.