A Semantic-Based Similarity Measure for Human Druggable Target Proteins

Eduardo C. dos Santos, Marcelo M. Santoro
Marcos A. dos Santos, Julio C. D. Lopes
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
edu@edusantos.eti.br, santoro@icb.ufmg.br
marcos@dcc.ufmg.br, jlopes.ufmg@gmail.com

Bráulio R. G. M. Couto
Centro Universitário de Belo Horizonte UNIBH
Belo Horizonte, Brazil
coutobraulio@hotmail.com

***Abstract*—The target identification is the first step on drug discovery pipeline. Thus, techniques which address the selection of potential "druggable targets" and potential "therapeutical targets" are very relevant to the discovery of new drugs and therapies. Nowadays, public databases with drug target indication provide target similarity searching based on BLAST. We demonstrate that the current protein annotation terms may be used on the development of semantic-based measures to provide target similarity searching. This approach allows to predict target similarities based on known signatures of a given protein even without the knowledge of the whole sequence. Our method produces a semantic ordering of the drug targets and provides a tool for latent information retrieving and for clustering analysis. New candidates may be compared against the known targets in a reduced space vector defined by singular value decomposition.**

Keywords-drug targets; SVD; semantic similarity; clustering.

I. INTRODUCTION

The drug discovery pipeline has the target identification and validation as the very first phases. Recently, known drug usage has been optimized addressing it to different targets [1], [2], [3]. The purpose is to use the known chemical properties and response of compounds with acceptable ADMET (administration, distribution, metabolism, excretion and toxicity) properties on developing new therapies (repurposing approved drugs) and/or on developing new lead compounds on a information-driven rational approach. The study of target similarity may be also helpful for predicting promiscuous binding sites and some kind of side-effects.

Public resources with drug target indication (as TTD [4] and DrugBank [5]) provide target similarity searching based on BLAST algorithm. But it is known that there are also important correlations (structural similarity and off-target similarity) even for low-similar sequences. Known signatures of targets (as annotated on GO, InterPro, Pfam, PROSITE and other resources) may be used for predicting correlations among different targets and/or among different target subsets. Indeed, 130 InterPro entries were identified on "druggable genome" searching [6]. It was also shown that Pfam annotation may be used for the same purpose [7].

The objective of this study was to evaluate whether semantic similarity across protein annotation terms can be used as an alternative to sequence alignment for predicting target similarities.

In general, semantics is the study of meaning. Semantic similarity is a concept whereby a metric is assigned to terms or documents in a set of terms or documents according to the likeness of their meaning in a pragmatic approach (i.e., considering how the context contributes to meaning). Broadly speaking, "two objects are semantically similar if they are related to similar objects" [8]. A semantic similarity measure may reveal new correlations, which are not possible by strictly direct queries onto relational databases. It is called *latent information retrieving*. Furthermore, semantic-based similarities may be determined over data hold in the form of annotation, which are more suitable for humans, and may be used to knowledge discovery exploring scientific data resources. Indeed, the use of semantic-based similarities across the Gene Ontology (GO) has been evaluated in the literature [9], [10].

Firstly, a protein drug target was represented by a binary column vector with m rows, each one representing the presence or absence of one InterPro signature in the sequence. A database with n protein drug targets is represented by a mxn binary matrix A, that is submitted to singular value decomposition [11] in order to develop a similarity measure among human protein drug targets.

The methodology can be expanded to incorporate different kinds of descriptors (e.g., MeSH terms) to discover more specific drug target relationships.

A. Singular Value Decomposition

The Singular Value Decomposition (SVD) establishes non-obvious but relevant relationships among clustered entities [11], [12], [13]. The rationale behind SVD is that a mxn matrix A can be represented by a set of derived matrices [13], which allows by a numerically different data representation without loss of semantic meaning.

Let A be any mxn matrix of ranking r. Then there exist a mxm matrix U_f, a nxn matrix V and a mxn matrix S for which:

$$A = U_f S V^T, \tag{1}$$

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-260-8

where:

- U_f is an mxm orthogonal matrix, which columns are the eigenvectors of the matrix AA^T;
- S is a mxn diagonal matrix with the **singular values** of A along its main diagonal in decreasing order;
- V is an nxn orthogonal matrix, which columns are the eigenvectors of the matrix A^TA.

These dimensions are for what is called the **full SVD**. Since all the elements of S below the n^{th} row are zero, partitioning the matrix U it can be taken the so called **thin SVD**[12]:

$$A = USV^T, \tag{2}$$

where:

- U is an mxn orthogonal matrix;
- S is a nxn diagonal matrix with the **singular values** of A along its main diagonal in decreasing order;
- V is an nxn orthogonal matrix.

Taking only the k most significant singular values of A, where $k < r$, the matrix A can be approximated by a low-dimensional matrix (A_k) given by:

$$A \approx A_k = U_k S_k V_k^T = \sum_{e=1}^{k} u_e s_e v_e^T, \tag{3}$$

where u_e and v_e are, respectively, the $column$ vector of U and the row vector of V both related to the e^{th} singular value in the decreasing order and k is the index of the highest relevant singular value.

The data approximation depends on how many singular values are used [14]. In this case, the k number of singular values is also the rank of the matrix A_k. The technique allows information extraction with less data. It is possible to compress/decompress data within a non-exponential execution time, and it make viable complex analysis across large amount of data [14]. A data set represented by a smaller number of singular values than the full size original data set has a tendency to group together certain data items that would not be grouped if the original data set is used [13]. This could explain why clusters derived from SVD can expose non-trivial relationships among the original data set items [15].

There are different methods to determine the $rank\ k$ of A_k. One of them is by the *scree test* [16].

A new entity represented by a column vector q, which is equivalent to ones of the original matrix A, may be compared with the entities represented in A in the smaller-dimensional space by a simple and low computing cost method. First, obtain the equivalent vector q_k in the reduced space vector. This can be made, as proposed by Lars Eldén [11], by computing:

$$q_k = q^T U_k. \tag{4}$$

Then apply some similarity metric (e.g., cosine measure or Euclidean distance) to compare q_k with the row vectors of $V_k S_k$. Thus, it is not necessary to compute the SVD factorization every time that a new target is introduced. It is only necessary to recompute the SVD factorization with the new query vector if it can not be represented by a combination of the vectors of the base. Otherwise, the new vector q_k may be incorporated to the matrix V_k.

B. Similarity measures

To assess the similarity between two entities, it can be used some similarity measure and evaluate its significance. There are different measures which may be tested, as the Euclidean distance; the cosine similarity; etc. In this paper, entities were represented in the low-dimensional space produced by SVD factorization and, after that, it was applied a cosine-based similarity measure, which is calculated as:

$$sim(c_i, c_j) = \cos(\alpha_{ij}) = \frac{c_i c'_j}{\sqrt{c_i c'_i}\sqrt{c_j c'_j}}, \tag{5}$$

where:
- c_i corresponds to the *i-th* row in $V_k S_k$ and;
- c'_i is the transpose of the vector row c_i.

II. Material and methods

A matrix with 1906 binary vectors was constructed, which represent protein drug targets retrieved from public databases (TTD [4], DrugBank [5] and KEGG-Drug [17]). Each protein representing vector is a set of 2700 binary descriptors. Each of these descriptors represent an InterPro annotation. It was used InterPro annotations of the following types: Family (F), Domain (D), Region (R), Active Site (A) and Binding Site (B). On considering every site-related annotation it was observed if the signature has occurred or not on a region of the sequence for which exists some annotation of F, D or G type. 365 of the 1906 targets were extracted randomly for training and validating purpose and the remaining 1541 were used to generate a representative vector space using SVD.

SVD factorization was applied to A and $k = 320$ factors were selected by *scree test* to determine the low-rank approximation A_k (Fig. 1).

The factorization provided a reduced dimensionality space in which relationships among the drug targets could be established. The similarity between any pair of drug targets was calculated as the cosine of the angle between the respective target representing vectors on the reduced space. Thus, the similarity measure of a pair of targets is equivalent to the dot product between the respective rows of the matrix $V_k S_k$ given by the (3).

The similarity relationships were analyzed by using clustering techniques implemented in the software named Multi-Experiment Viewer (MeV) [18], a freely available software

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-260-8

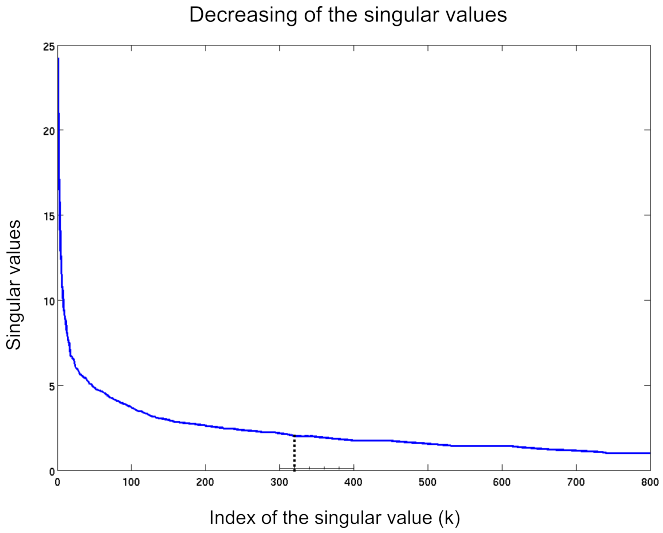

Figure 1. Singular values of A (as obtained by SVD factorization). The first 320 singular values (and respective orthogonal vectors) were selected by the *scree test*.

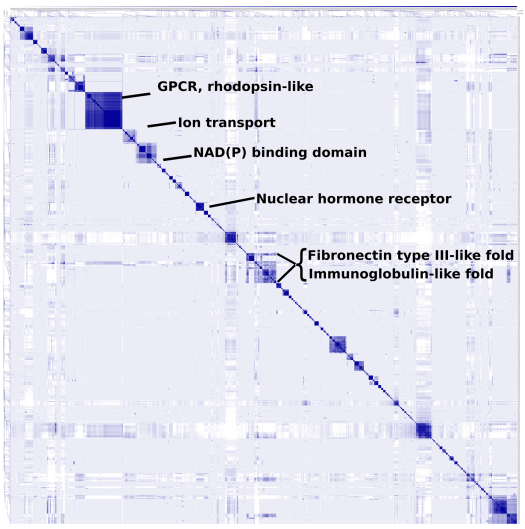

Figure 2. Heatmap with all 1541 drug targets reordered semantically by the hierarchical clustering algorithm. It was easy to identify various clusters as the GPCRs (the greatest group) and other cases showed in detail in other figures.

application that provides an extensive library of algorithms and visualization tools for integrative data analysis from a user-friendly interface.

III. Results

A similarity matrix was constructed from the values of the cosines computed as described on the previous section and used this matrix as input into the software named MeV. Then, it was applied the hierarchical clustering algorithm (HCL) implemented in MeV and it produced a heatmap with the targets semantically reordered (Fig. 2, 3, 5 and 6). Fig. 2 shows the heatmap for the whole ensemble. Fig. 3 shows in detail the region at the heatmap related to nuclear receptors (NR). The similarity measure was found to be efficient in discriminating the NR members in a second level grouping Peroxisome; Retinoid and Vitamin D receptors.

The Euclidean distance was also evaluated and let us to conclude that, referring to our application and data set, Euclidean distance and cosine angle measure provide similar clustering results (Fig. 4).

Similarly to the case of nuclear receptors, Fig. 5 shows a cluster (drug targets with NAD-P binding domain) larger than the NR cluster and with deeper hierarchy level. Fig. 6 illustrates the efficiency of the method to discover relationships hardly recognized by simple sequence similarity search – it shows an interesting relationship between Fibronectin type-III like folding and Immunoglobulin like folding – two domains that have low-similar sequence but high-similar structure and that co-occur in some protein-folding pathways [19].

The results were compared with the ones produced by sequence similarity (with BLASTALL) [20]. For the whole

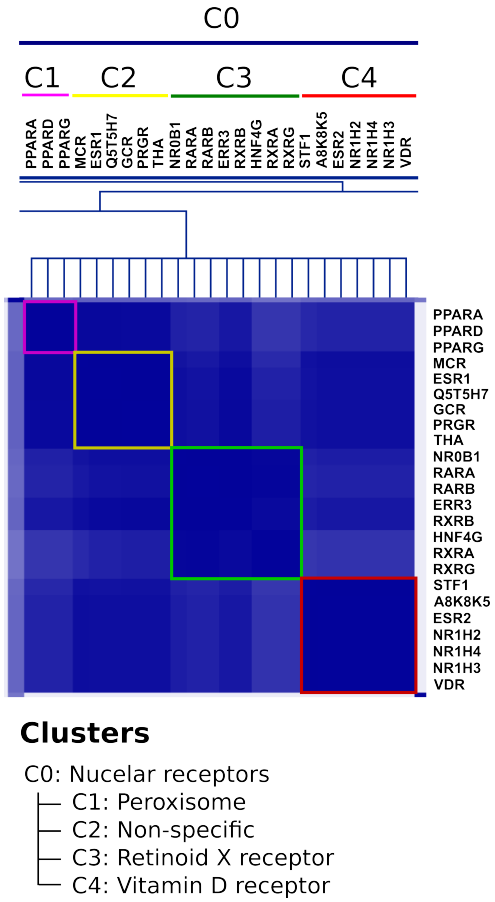

Figure 3. Zoom view of the region (of Fig. 2) related to hormone nuclear receptors. It is evident the in-depth consistency according to their additional annotations.

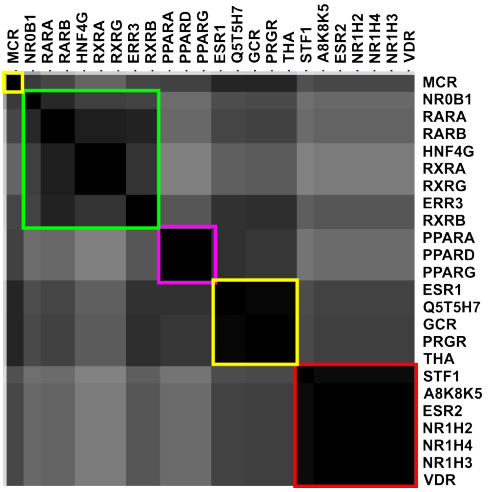

Figure 4. Zoom view of the region related to hormone nuclear receptors from the heatmap obtained from Euclidean distance. The clustering results provided by Euclidean distance for small samples were very similar to the ones provided by cosine measure.

data set, the BLAST bitscore did not provide good discrimination, but for a small sample the clustering results were very similar when using the SVD-based similarity score and the sequence similarity based score. It was performed two different clustering methods in this case: HCL and cluster affinity searching technique (CAST), both implemented in MeV. Fig. 7 illustrates the clustering results using semantic similarity and sequence alignment for 42 selected targets. Five non-unitary groups could be easily identified. One GPCR (PE2R3_HUMAN), left as orphan by the two clustering methods across the sequence similarity matrix. The same target was correctly grouped (in the context of biological annotations) with other GPCR by both, HCL and CAST, across the semantic-based similarity matrix.

To find potential "druggable" candidates, it was projected other proteins into the reduced space. As an example of interesting finding, the case of Kynurenine 3-monooxygenase (KMO) can be cited (Table I). The value of the distance-like coefficient is significantly low only for two known drug targets: ERG1 and SOX. ERG1 shares annotation with both KMO and SOX, but there are not shared annotation between SOX and KMO. So, the space transformation indicates a non *prima facie* relationship between KMO and SOX. That "secondary" relationship is not retrieved from the original data set or from the transformed space when it is added many factorized terms. The higher the number of terms of the factorization, the smaller the retrieval capability to discover hidden relationships (with many terms it is only possible to compute the coefficient for pairs whose members share some annotation, the remaining becomes equivalent to infinite). The kynurenine pathway is the main pathway for tryptophan metabolism and have been considered a pathway with a lot of potential sites for drug discovery in neuroscience [21].

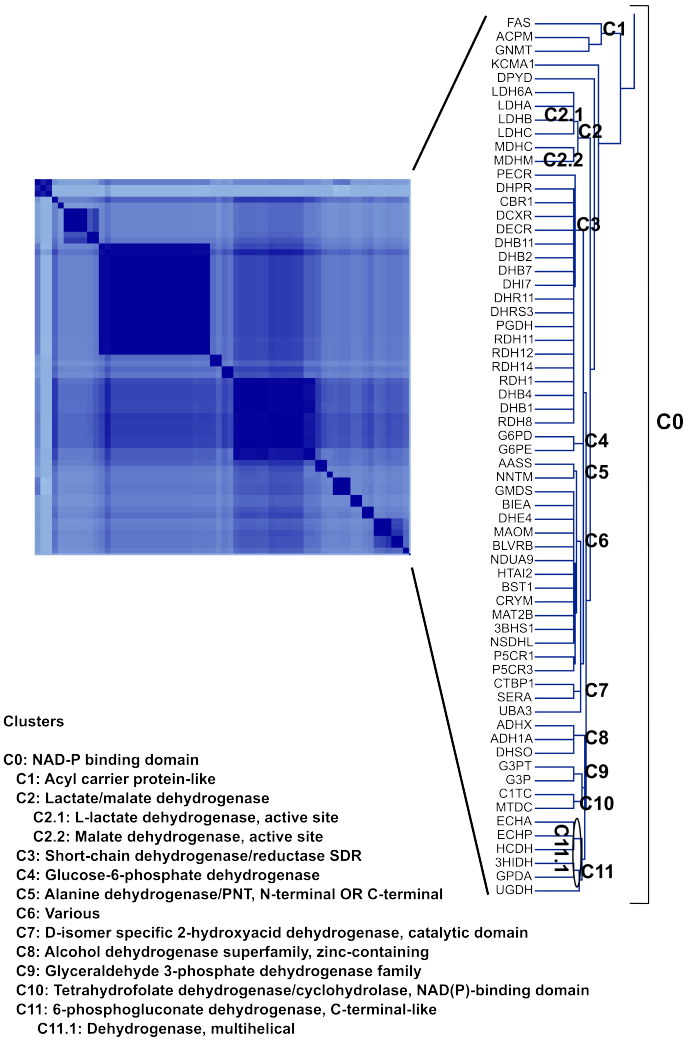

Figure 5. Zoom view of another region of the Fig. 2 related to drug targets with NAD-P binding domain. Again, it is evident the in-depth consistency – here, on a larger cluster than the one with nuclear receptors and showing deeper hierarchy level.

Particularly, KMO (a member of the kynurenine pathway) has the gene located in the chromosome region associated with schizophrenia [22]. On the other hand, it is known that glycine binds to SOX and it is used as an alternative therapy of schizophrenia [23], [24].

IV. Conclusion and Future Work

A semantic-based measure across the InterPro annotations of protein drug targets was developed. It was shown that this measure may be used for similar targets searching. Nowadays, public resources provide target similarity searching using a local BLAST algorithm. Our method has a fixed computational time consumption independently of the sequence size. New targets may be compared against the current set representing it by their biological annotations, projecting it on the U_k space and, then, computing the cosine

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-260-8

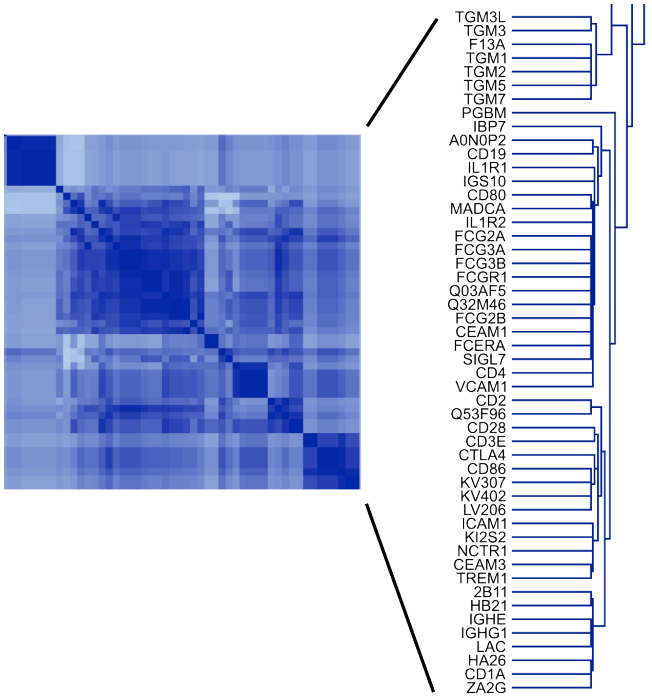

Figure 6. Zoom view of the region of the Fig. 2 showing targets with Fibronectin type III-like fold domain and/or Immunoglobulin-like fold domain. The correlation among these targets are shown as estimated [19].

Table I
RANKED LISTS FOR KMO_HUMAN OF SIMILAR TARGETS

rank	$k = 320$		$k = 800$		original	
	target	score	target	score	target	score
1	ERG1	0.0009	ERG1	0.0445	ERG1	0.2374
2	SOX	0.0022	-	∞	-	∞
3	CBPE	0.4656	-	∞	-	∞
4	SO1B1	0.5135	-	∞	-	∞
5	P85A	0.5512	-	∞	-	∞
6	DCK	0.5550	-	∞	-	∞

Each ranked list is given by the distance-like score computed from the k-dimensional space or from the original vector space (before apply SVD). The value considered infinite is 0.6931.

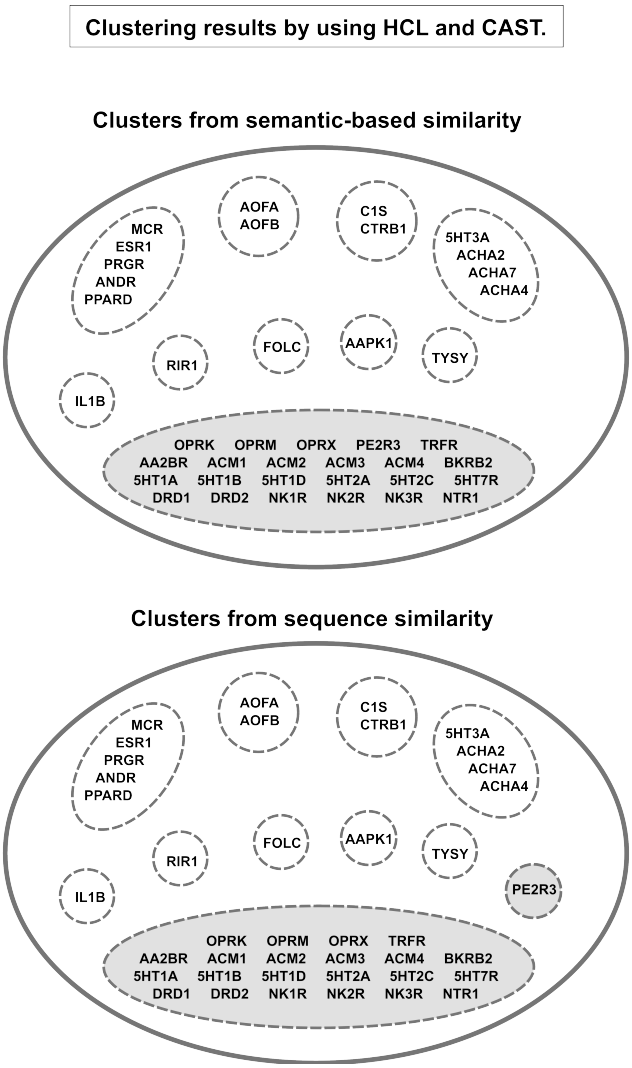

Figure 7. Clusters resulted by both HCL and CAST methods for a small sample with 42 targets. The algorithms were performed across the similarity matrices obtained from the semantic-based similarity measure and the BLAST bitscore based similarity. Clusters in grey denote some discrepancy between the two methods.

among the produced column vector with each row vectors of $V_k S_k$. The biological annotations of the new targets may be determined by InterProScan [25] over their sequence or may be inferred by the research by other experimental observations. Thus, it was shown that the effort exerted on annotation can be explored to order data semantically. The measure is consistent and complementary to BLAST-based sequence alignment approach allowing identification of similar and co-existent fold domains even for low-similar sequences. So, the measure can be potentially effective to discover hidden relationships that are hardly recognized by simple sequence similarity search. Furthermore, the methodology can be expanded to incorporate different kinds of descriptors (e.g., MeSH terms) to discover more specific drug target relationships.

We are going to expand our work on:

- Optimizing the applied algorithm and parameters (clustering algorithm, rank determination, etc.);
- Ensemble correlations and other cross-correlation analysis;
- Predicting new potential drug target candidates and new possible therapy applications;
- Applying the method for non-human targets;
- Incorporating other types of annotations to the descriptors set (e.g., MeSH, OMIM, and UMLS);
- Comparing the performance of the method with approaches using other types of decomposition: PCA and NMF.

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-260-8

REFERENCES

[1] M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran, R. Whaley, R. A. Glennon, J. Hert, K. L. H. Thomas, D. D. Edwards, B. K. Shoichet, and B. L. Roth, "Predicting new molecular targets for known drugs," Nature, vol. 462, pp. 175–181, 2009.

[2] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, and B. K. Shoichet, "Relating protein pharmacology by ligand chemistry," Nature Biotechnology, vol. 25, pp. 197–206, 2007.

[3] M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen, and P. Bork, "Drug target identification using side-effect similarity," Science, vol. 321, pp. 263–266, 2007.

[4] F. Zhu, B. Han, P. Kumar, X. Liu, X. Ma, X. Wei, L. Huang, Y. Guo, L. Han, C. Zheng, and Y. Chen, "Update of ttd: Therapeutic target database," Nucleic Acids Research – Database issue, vol. 38, pp. D787–D791, 2010.

[5] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali, "Drugbank: a knowledgebase for drugs, drug actions and drug targets," Nucleic Acids Research – Database issue, vol. 36, pp. D901–D906, 2008.

[6] A. L. Hopkins and C. R. Groom, "The druggable genome," Nature reviews. Drug discovery, vol. 1, no. 9, pp. 727–730, September 2002.

[7] A. P. Russ and S. Lampel, "The druggable genome: an update." Drug Discovery Today, vol. 10, no. 23-24, pp. 1607–1610, 2005.

[8] G. Jeh and J. Widom, "Simrank: A measure of structural-context similarity," in In KDD, 2002, pp. 538–543.

[9] P. Lord, R. Stevens, A. Brass, and C. Goble, "Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation," Bioinformatics, vol. 19, pp. 1275–1283, 2003.

[10] M. Chagoyen, P. Carmona-Saez, C. Gil, J. M. Carazo, and A. Pascual-Montano, "A literature-based similarity metric for biological processes," BMC Bioinformatics, vol. 7, pp. 363–375, 2006.

[11] L. Eldén, "Numerical linear algebra in data mining," Acta Numerica, vol. 15, pp. 327–384, 2006.

[12] L. Eldén, Matrix methods in data mining and pattern recognition. Fundamentals of Algorithms 4. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), 2007.

[13] M. W. Berry, S. T. Dumais, and G. W. O'Brien, "Using Linear Algebra for Intelligent Information Retrieval," University of Tennessee, Tech. Rep. UT-CS-94-270, 1995.

[14] D. del Castillo-Negrete, S. P. Hirshman, D. A. Spong, and E. F. D'Azevedo, "Compression of magnetohydrodynamic simulation data using singular value decomposition," J. Comput. Phys., vol. 222, pp. 265–286, March 2007.

[15] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," Journal of the American society for information science, vol. 41, no. 6, pp. 391–407, 1990.

[16] R. B. Cattell, "The Scree Test For The Number Of Factors," Multivariate Behavioral Research, vol. 1, no. 2, pp. 245–276, 1966.

[17] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa, "Kegg for representation and analysis of molecular networks involving diseases and drugs," Nucleic Acids Research, no. 38, pp. D355–D360, 2010.

[18] E. Howe, K. Holton, S. Nair, D. Schlauch, R. Sinha, and J. Quackenbush, "Mev: Multiexperiment viewer," Biomedical Informatics for Cancer Research, pp. 267–277, 2010.

[19] D. J. Leahy, "Implications of atomic-resolution structures for cell adhesion," Annual Review of Cell and Developmental Biology, vol. 13, pp. 363–393, November 1997.

[20] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic local alignment search tool," Journal of Molecular Biology, vol. 215, pp. 403–410, 1990.

[21] J. Rodgers, T. Stone, M. Barrett, B. Bradley, and P. Kennedy, "Kynurenine pathway inhibition reduces central nervous system inflammation in a model of human african trypanosomiasis," Brain, vol. 132, no. 5, pp. 1259–1267, May 2009.

[22] M. Holtze, P. Saetre, S. Erhardt, L. Schwieler, T. Werge, T. Hansen, J. Nielsen, S. Djurovic, I. Melle, O. A. Andreassen, H. Hall, L. Terenius, I. Agartz, G. Engberg, E. G. Jansson, and M. Schalling, "Kynurenine 3-monooxygenase (kmo) polymorphisms in schizophrenia: An association study," Schizophrenia Research, vol. 127, no. 1-3, pp. 270–272, 2011.

[23] J. Semba, "Glycine therapy of schizophrenia; its rationale and a review of clinical trials," Nihon Shinkei Seishin Yakurigaku Zasshi, vol. 18, no. 3, pp. 71–80, 1998.

[24] U. Heresco-Levy, M. Ermilov, P. Lichtenberg, G. Bar, and D. C. Javitt, "High-dose glycine added to olanzapine and risperidone for the treatment of schizophrenia." Biological psychiatry, vol. 55, no. 2, pp. 165–171, Jan. 2004.

[25] E. M. Zdobnov and R. Apweiler, "InterProScan – An integration platform for the signature-recognition methods in InterPro," Bioinformatics, vol. 17, no. 9, pp. 847–848, September 2001.

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-260-8