Chemical and Metabolic Pathway Semantic Similarity

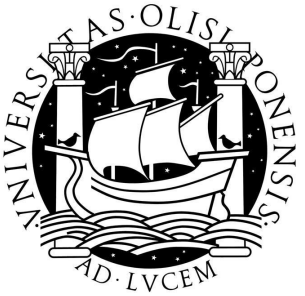
Tiago Grego, João D Ferreira, Catia Pesquita, Hugo Bastos, Diogo V Viçosa, João Freire, and Francisco M Couto

DI–FCUL–TR–2010–01

DOI:10455/3335

(http://hdl.handle.net/10455/3335)

March 2010



Published at Docs.DI (http://docs.di.fc.ul.pt/), the repository of the Department of Informatics of the University of Lisbon, Faculty of Sciences.

Chemical and Metabolic Pathway Semantic Similarity

Tiago Grego
tgrego@fc.ul.pt

João D Ferreira
ferreira.jds@gmail.com

Catia Pesquita
cpesquita@xldb.di.fc.ul.pt

Hugo Bastos
hbastos@xldb.di.fc.ul.pt

Diogo Vila Viçosa
diogo.vicosa@fc.ul.pt

João Freire
joaofreire@fm.ul.pt

Francisco M Couto
fcouto@di.fc.ul.pt

March 5, 2010

Abstract

Background: Similarity measures for the comparison of metabolic pathways can provide a better understanding of evolutionary relationships among species or strains and have an important biotechnological value for the pharmaceutical industry. Semantic similarity applied to enzymes has been used; here we show an approach using metabolites. While there is a number of methods to compare and align metabolic pathways based on metabolites, they are usually based only on their structural information neglecting their biological information.

Results: In this work we present an alternative approach for measuring the semantic similarity between metabolic pathways by comparing their metabolites. This comparison is based on the Chemical Entities of Biological Interest ontology, and can be useful in toxicology and drug discovery for assessment of biological activity of chemical compounds. We implemented a software capable of measuring the similarity between metabolic pathways present in the Kyoto Encyclopedia of Genes and Genomes database and a preliminary analysis shows the effectiveness of the proposed approach.

Conclusions: We have shown that semantic similarity can be applied to pathways whose chemical compounds are annotated in the Chemical Entities of Biological Interest ontology. This work resulted in the creation of a software, CMPSim, accessible as a web-tool at `http://xldb.di.fc.ul.pt/biotools/cmpsim/`. It can be used to obtain similarity measures between chemical compounds and metabolic pathways.

1 Background

The emergence of biological databases dedicated to chemical compounds and metabolic pathways has enhanced the development of methods to compare and align them. Comparative analysis of metabolic pathways can reveal important

information on both evolution of organisms and potential pharmacological targets [1]. Furthermore, it can be helpful to assess similarity between metabolic networks [2].

Alignment and comparison of metabolic pathways usually involves representing the pathway as a graph whose elements correspond to participants in the pathway's reactions, namely enzymes (as the edges) or compounds (as nodes). To align and compute the similarity between metabolic pathways, several aspects may be considered, including the structure of the graph [3] or the similarity between the individual reactions [4], enzymes [3–7] and metabolites (chemical compounds) [8].

The most common approach to perform pathway alignment is based on the sequence similarity between their intervening enzymes [3]. This strategy assumes a correlation between pathway similarity and sequence similarity, but this may not always exist due to analogous enzymes [9]. Other strategies consider semantic similarity [5], where enzyme annotation similarity can be based on their Enzyme Commission (EC) numbers [10] or their Gene Ontology (GO) annotations [11]. Less common approaches include comparing the topology of the pathways using graph-matching algorithms [6], and comparing the metabolites that take part in the pathways by means of chemical compound similarity [8].

For the purpose of this work, we consider a metabolic pathway as the set of its metabolites. As such, the similarity score is based on the similarity between metabolites. Most methods for calculating chemical similarity are based on the compound's two- and three-dimensional structure [12]. In these methods the molecular structures are usually represented by fixed-size or variable-size molecular fingerprints [13, 14]. These fingerprints are then compared, for example, by using the Jaccard-Tanimoto coefficient [15]. However, while these methods may be effective in some areas, when it comes to biological interest, structure is not the most informative aspect of a molecule. For instance, while L-amino acids are used to synthesize proteins, their stereo-isomers, D-amino acids, are much less frequent in nature and their role is totally different [16]. Their structural similarity is almost absolute, but biologically they are very different.

We implemented our approach in a web-tool that measures the similarity between chemical compounds or metabolic pathways. The approach uses the ChEBI ontology as a common schema and the Kyoto Encyclopedia of Genes and Genomes database as source of chemical annotations for the metabolic pathways. This web-tool is available at `http://xldb.di.fc.ul.pt/biotools/cmpsim/`.

2 Data Sources

2.1 Kyoto Encyclopedia of Genes and Genomes

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a collection of databases categorized into systems information, genomic information and chemical information. The different KEGG databases are highly integrated in an effort to constitute a computer representation of the biological system [17].

One of the main components of KEGG is the PATHWAY database, which contains a collection of graphical representations of the known pathways and

lists of enzymes, reactions and metabolites within them. Pathway maps in KEGG are species-independent and exist for metabolism, genetic information processing, environmental information processing such as signal transduction, and various other cellular processes and human diseases. Pathways are also annotated with species, so that inter-species analysis can be performed.

KEGG pathway maps are organized in a two level hierarchy that groups together closely related metabolic pathways. In the top level we can find general terms such as Metabolism, Genetic Information Processing, Environmental Information Processing, Cellular Processes, Human Diseases and Drug Development. These top level terms are further detailed in the lower level with variable degrees of specificity (see **Figure 1**).

Each metabolic pathway entry integrates information from other databases in KEGG such as the intervening enzymes (KEGG ENZYME), chemical reactions (KEGG REACTION) and chemical compounds (KEGG COMPOUND).

KEGG COMPOUND is a chemical structure database for metabolic compounds and other chemical substances that are relevant to biological systems. We use the entries in KEGG COMPOUND database as chemical annotations of the metabolic pathways in the KEGG PATHWAY maps.

2.2 ChEBI

Chemical Entities of Biological Interest (ChEBI) is a freely available dictionary of small molecular entities such as any constitutionally or isotopically distinct atoms, molecules, ions etc., identifiable as a separately distinguishable entity that is either a product of nature or a synthetic product used to intervene in the processes of living organisms [18].

Classes of molecular entities and part-molecular entities are also included, enabling ChEBI to be organized as an ontology, structuring molecular entities into classes and defining the relations between them. Several relationship types exist in ChEBI, and a number of them are reciprocal in nature. The ontology is subdivided into three separate sub-ontologies:

- **Molecular structure**, in which the entities are classified according to composition and structure.
- **Role**, in which entities are classified on the context of their role within a biological context.
- **Subatomic particle**, that classifies particles which are smaller than atoms.

The graph of this ontology contains almost 550,000 nodes representing terms. Some terms are not chemical compounds but part of compounds, such as functional groups, that make the ontology structure possible. Also, for each individual chemical compound, there may be several identifiers, which come from different annotations that waere later identified as the same compound. To better picture the ontology, **Figure 2** shows a simplified view of the ChEBI ontology for the chemical compound *(R)-adrenaline* (CHEBI:28918).

For each compound entry in ChEBI there is an extensive list of synonyms and manually curated cross-references to other non-proprietary databases, including KEGG COMPOUND.

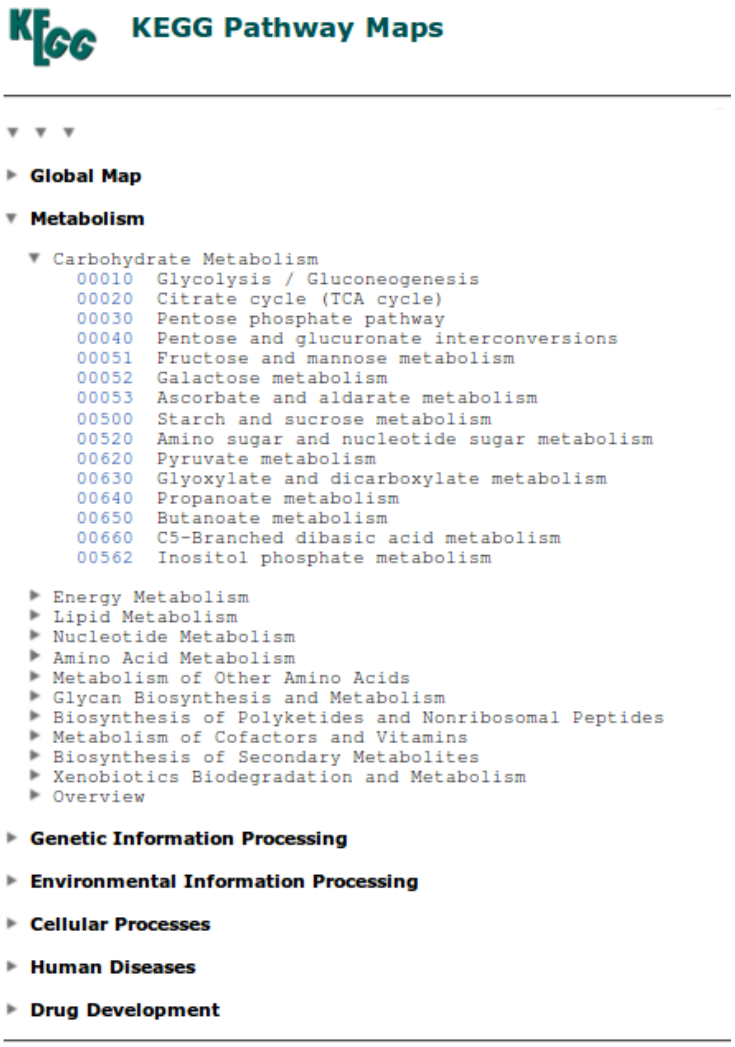


Figure 1: A two level hierarchical organization in KEGG groups together closely related metabolic pathways. In the top level we can find general terms. The second level further details the terms, and the available pathways are included into one of this second level classes. This organization can be browsed in `http://www.genome.jp/kegg-bin/get_htext?htext=br08901.keg`.

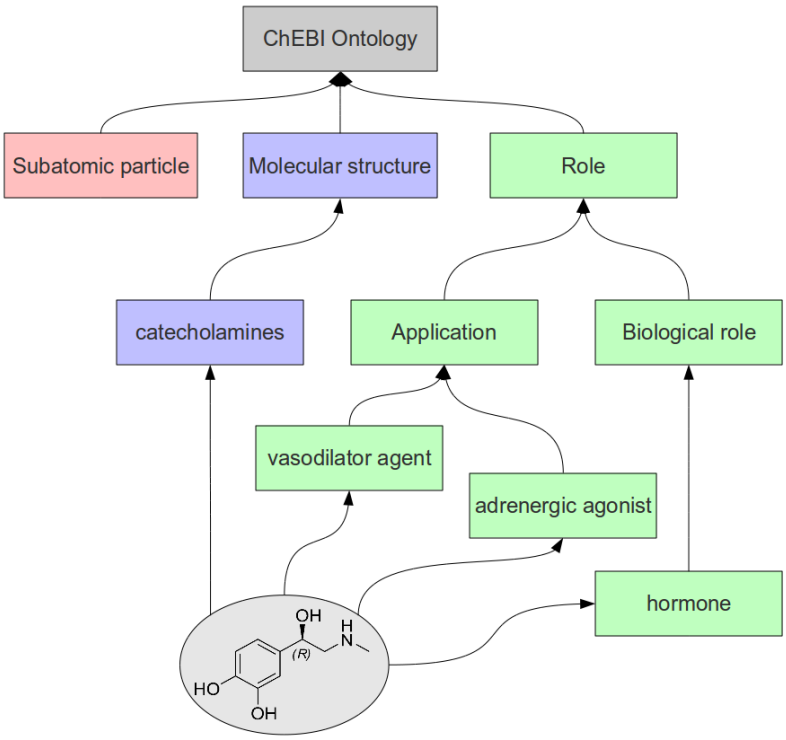


Figure 2: A simplified ontology for the compound *(R)-adrenaline* is presented, where we see that a compound can be described by several terms in different sub-graphs.

3 Methods

We developed an approach to calculate the similarity between metabolic pathways. Given a pair of metabolic pathways the method returns a value between 0 and 1 as a measure of the similarity between them (0 represents total dissimilarity and 1 equality).

The first step in the similarity calculation consists in the retrieval of the chemical compounds present in each pathway. Then, each compound is mapped to the ChEBI ontology. Once we have this information we can compare the sets of compounds present in each pathway using semantic similarity measures.

simUI is a graph-based measure, which means that it considers not only the terms themselves, but also all of their ancestors in the graph of the ontology. Given two chemical compounds c_1 and c_2, and the sets of all their ancestral terms up to the root, $\text{asc}(c_1)$ and $\text{asc}(c_2)$ respectively, simUI is defined as the number of terms in the intersection of $\text{asc}(c_1)$ with $\text{asc}(c_2)$ divided by the number of terms in their union:

$$\text{simUI}(c_1, c_2) = \frac{\#\{\text{asc}(c_1) \cap \text{asc}(c_2)\}}{\#\{\text{asc}(c_1) \cup \text{asc}(c_2)\}}$$

Note that the term itself is included as an ancestor for this calculation.

However it is known that for ontologies where term specificity is not well correlated with term depth, methods based on information content (IC) are preferable [19].

Let $p(c)$ be the frequency of usage of a given term c in the corpus. Then, the information content of a term can be given by the expression:

$$\text{IC}(c) = -\log p(c)$$

Thus, a very frequent term is considered to be less informative and vice-versa. To obtain uniform IC values we need however to uniformize the previous equation by dividing it by the scale maximum (so as to obtain a value in a scale between 0 and 1). The expression for the uniform IC is:

$$\text{IC}_\text{u}(c) = \frac{\text{IC}(c)}{\max_c \text{IC}(c)}$$

simGIC is a hybrid measure, since it combines graph and IC properties. It is defined as the sum of the IC of each term in the intersection of $\text{asc}(c_1)$ and $\text{asc}(c_2)$ divided by the sum of the IC of each term in their union:

$$\text{simGIC}(c_1, c_2) = \frac{\sum_{t \in \text{asc}(c_1) \cap \text{asc}(c_2)} \text{IC}_u(t)}{\sum_{t \in \text{asc}(c_1) \cup \text{asc}(c_2)} \text{IC}_u(t)}$$

3.1 Metabolic Pathway similarity

The measures simUI and simGIC are defined to calculate the similarity between two chemical compounds, however a metabolic pathway can be annotated with many compounds and can be seen as a set of chemical compounds. So, when comparing two metabolic pathways we are comparing two sets of compounds.

A best-match average approach can be used to handle this issue. Given two metabolic pathways m_1 and m_2, with sets of chemical compounds $\text{cpds}(m_1)$

and cpds(m_2) respectively, the best-match average is given by the average of the similarities between each compound in cpds(m_1) and its most similar compound in cpds(m_2). This result is averaged with its reciprocal to obtain a symmetric score (this ensures that each compound is only compared to its most similar counterpart, avoiding potential biases [20]):

$$\begin{aligned} \text{sim}'(m_1, m_2) &= \text{avg}_{c_1}(\max_{c_2} \text{simUI}(c_1, c_2)),\ c_1 \in \text{cpds}(m_1),\ c_2 \in \text{cpds}(m_2) \\ \text{sim}_{\text{BMA}}(m_1, m_2) &= \frac{\text{sim}'(m_1, m_2) + \text{sim}'(m_2, m_1)}{2} \end{aligned}$$

A quick glance will show that the formula above uses simUI to compare compounds. However, the same formula is applied to the hybrid simGIC approach.

4 Implementation

4.1 Architecture

We implemented the approach to compare two KEGG metabolic pathways maps. For a KEGG pathway identifier, we first retrieve its chemical annotations in terms of the KEGG COMPOUND identifiers. Then we map those compounds to the ChEBI ontology. Using the ChEBI cross-references to the KEGG COMPOUND database, we were able to map 5,357 KEGG compounds into the ChEBI ontology.

The pathways are thus represented as sets of compounds identified in the ChEBI ontology, and the similarity between the two groups of compounds can be measured using the described semantic similarity approach.

Some modifications however needed to be made to the structure of ChEBI in order to enable the application of the semantic similarity methods. Since a directed acyclic graph (DAG) structure is required for the use of semantic similarity measure defined above, we had to perform some modifications to the structure of ChEBI. The graph structure of ChEBI was simplified by removing all the cyclic relationships and merging all nodes that correspond to the same chemical compound, redirecting the original relationships. By doing this we loose the independent structure of the three sub-ontologies, but produce a more cohesive structure, in the form of a DAG, that is able to support semantic similarity calculations without loss of information. To calculate similarities using only one of the sub-ontologies, a similar simplified sub-ontology using the terms of the target subgraph can be built. With this modification we can directly apply simUI in ChEBI to measure the similarity of two given compounds.

To use simGIC we need to perform the additional calculation of the IC of each chemical compound, which gives a measure of how informative a compound is when annotating a metabolic pathway. Using the complete KEGG pathway database as corpus, we calculated the frequency of annotation of each compound to metabolic pathways. To obtain this frequency, we first count for a given chemical compound the number of distinct metabolic pathways annotated to it or to one of its descendants, and then divide that number by the total number of annotations.

After these procedures, the similarity between metabolic pathways can then be calculated by best match average combined with either simUI or simGIC for

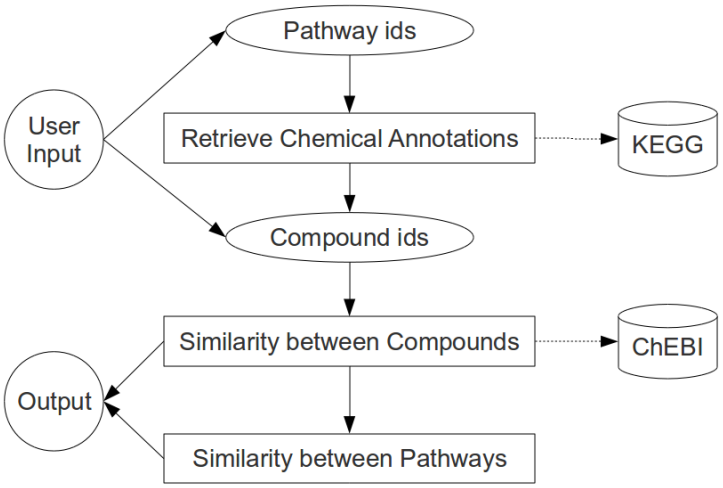


Figure 3: The user gives a pair of pathways or compounds as input, and after a series of steps where the KEGG and ChEBI databases are used, a result for the similarity is given.

compound similarity. The semantic similarity measure for both the compounds and metabolic pathways is given as a number between 0 and 1.

4.2 CMPSim web-tool

A web-based graphical interface for calculation of similarities between chemical compounds and metabolic pathways was developed, implementing the approach detailed above. A diagram of the architecture of the web-tool is given in **Figure 3** and **Figure 4** shows a screenshot of the CMPSim homepage.

The web-page has three main functionalities:

- **Search**, where the user can look for chemical compounds and metabolic pathways using keywords or identifiers.
- Calculation of the **semantic similarity between two compounds**, given by the user as ChEBI identifiers.
- Calculation of the **semantic similarity between two metabolic pathways**, given by the user as KEGG PATHWAY identifiers.

Search: The user can choose to make a generic search for a keyword, which will show the top 10 compounds and metabolic pathways that most resemble the search query. In the advanced search the user can specify if he is looking for compounds or pathways, which will show the top 20 results for the chosen class.

If an identifier is inserted as keyword, the correspondent compound or pathway is returned.

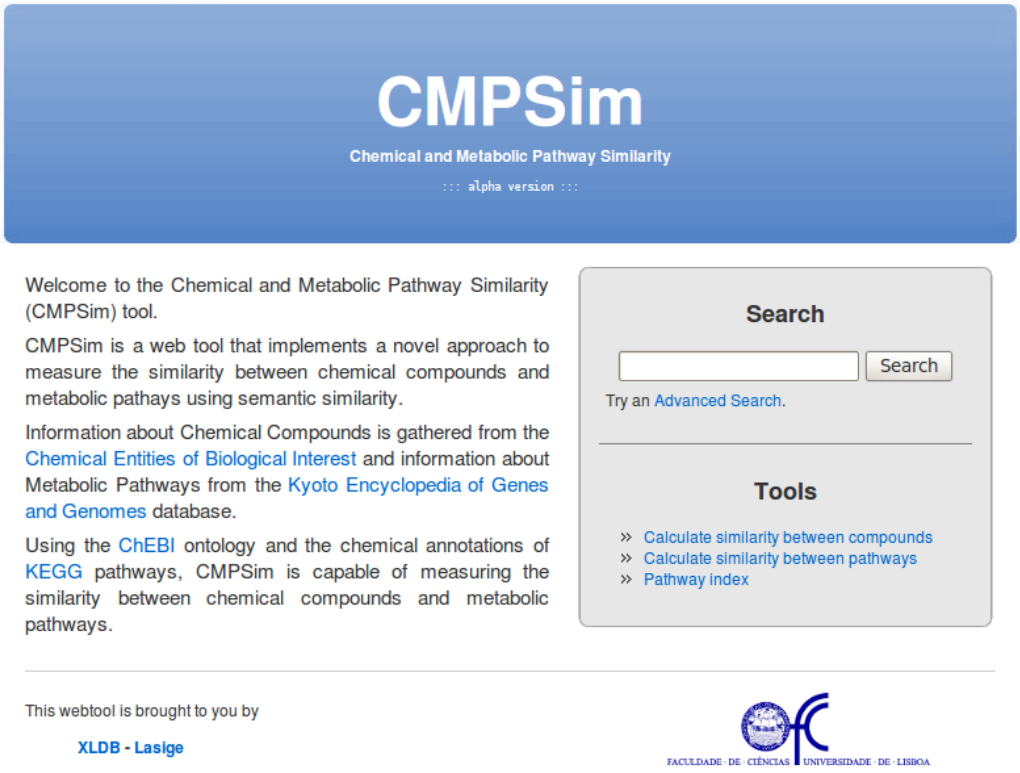


Figure 4: The homepage of CMPSim is shown. The user can perform search or similarity calculations using the right-side bar.

All returned ChEBI compounds and KEGG pathways are linked to their respective entry pages, which contain detailed information about them.

Compound Similarity: In this section the user is asked for the two ChEBI identifiers of the compounds to compare. The tool checks if the identifiers are valid, and if so, both simUI and simGIC values are presented. If the identifiers are not valid, the user is warned.

Pathway Similarity: In this section the user is asked for the two KEGG pathway identifiers of the metabolic pathways to compare. After the inserted identifiers are validated, the similarity between the two pathways is calculated using the best match average-simGIC measure. The user is also presented with the list of the compound annotations for each pathway.

5 Results

To evaluate the performance of our semantic similarity methodology, we calculated the similarity between all the pairs of available metabolic pathway maps in the KEGG PATHWAYS database using the simGIC between the intervening chemical compounds present in ChEBI and using a best match average approach.

There are 294 KEGG pathways that have compounds present in ChEBI. When the similarity between all the 43,071 distinct pairs of metabolic pathways is calculated we obtain the average distribution of values shown in **Figure 5**, where 50% of the similarity measures are lower than 0.089, and the average similarity is 0.120. The obtained distribution shows that most pathway pairs have low similarity and few have high similarity. This is in accordance with the idea that the number of pathways with high similarity should be small, since only a few are related.

We then analysed the results for the pathways "Photosynthesis" and "Oxidative Phosphorilation" (KEGG PATHWAY identifiers map00190 and map00195). These two pathways have a similar topology [21] (see **Figure 6**) and we expected to obtain a high similarity between them. All but one compound were successfully mapped into ChEBI (see **Table 1**). The score obtained was 0.62, a value that is on the top 0.8% of all similarity scores calculated, corresponding to a high similarity, as expected. This high measure was obtained because there are a lot of compounds in one pathway with a similar compound in the other, like the pairs NAD/NADP, NADH/NADPH. It is worth mentioning that most compounds in these pathways have a high information content, meaning they are present in a small number of pathways. Contrarily, water and oxygen (which is an ancestor of 13317 compounds) have a very low information content (~ 0.04 and ~ 0.01, respectively), which reflect their presence (or their descendants') in many pathways.

Table 2 shows the 20 most similar pairs of pathways. Only *significant* pathways were considered (under section Discussion, we explain the concenpt of *significance*). It is worth mentioning that the pair "Photosyntesis" vs "Oxidative Phosphorylation" is the 14^{th} entry.

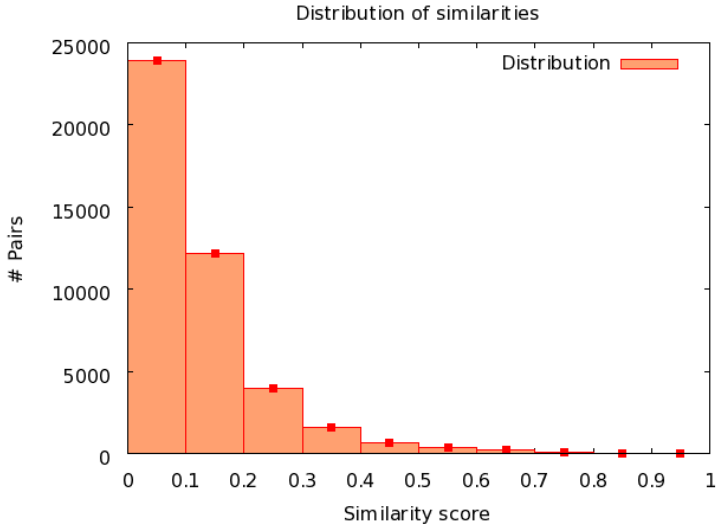


Figure 5: Histogram of similarity scored obtained for each pair of metabolic pathways, given in intervals of 0.1. We observe that most pairs have a low score and only a few are very similar.

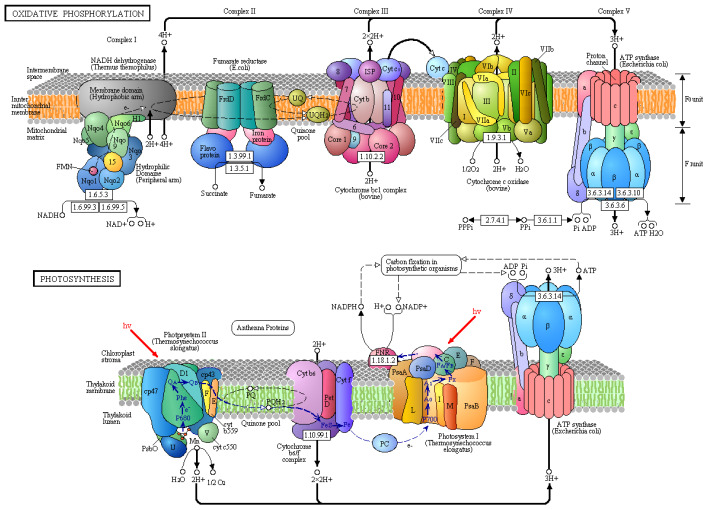


Figure 6: Photosynthesis and Oxidative Phosphorylation metabolic maps as can be found in KEGG

Table 1: Compounds belonging to pathways "Oxidative phosphorilation" (map190) and "Photosynthesis" (map195).

Oxidative Phosphorilation				Photosynthesis			
KEGG	**ChEBI**	**Name**	**IC**	**KEGG**	**ChEBI**	**Name**	**IC**
C00001	15377	water	0.042	C00001	15377	water	0.042
C00002	15422	ATP	0.535	C00002	15422	ATP	0.535
C00003	15846	NAD(+)	0.878	C00005	16474	NADPH	0.878
C00004	16908	NADH	0.878	C00006	18009	NADP(+)	0.806
C00007	25805	oxygen	0.011	C00007	25805	oxygen	0.011
C00008	16761	ADP	0.237	C00008	16761	ADP	0.237
C00009	18367	phosphate(3-)	0.657	C00009	18367	phosphate(3-)	0.657
C00013	18361	diphosphate(4-)	0.878	C00034	18291	manganese	0.806
C00042	15741	succinic acid	0.395	C00080	15378	hydron	0.633
C00061	17621	FMN	0.878	C02185	16323	plastoquinol-1	1.000
C00080	15378	hydron	0.633				
C00122	29806	fumarate(2-)	0.440	C02061	*No ChEBI correspondence*		
C00390	17976	ubiquinol	1.000				
C00399	16389	ubiquinone	0.878				
C00524	18070	cytochrome c	0.878				
C00536	18036	triphosphate(5-)	1.000				

Table 2: The 20 significant pathway pairs with higher similarity. Here, significant means that, for both pathways, more than 90% of the compounds, as retrieved from the KEGG PATHWAY database, must be mapped into ChEBI and also that the pathway must contain at least 10 compounds.

Path1		**Path2**		**Similarity**
map00062	Fatty acid elongation in mitochondria	map00071	Fatty acid metabolism	0.93341
map00562	Inositol phosphate metabolism	map04070	Phosphatidylinositol signaling system	0.76447
map01060	Biosynthesis of plant secondary metab...	map01070	Biosynthesis of plant hormones	0.67606
map00720	Reductive carboxylate cycle (CO2 fixa...	map01065	Biosynthesis of alkaloids derived fro...	0.67115
map00720	Reductive carboxylate cycle (CO2 fixa...	map01070	Biosynthesis of plant hormones	0.66587
map00640	Propanoate metabolism	map00720	Reductive carboxylate cycle (CO2 fixa...	0.65221
map00062	Fatty acid elongation in mitochondria	map00640	Propanoate metabolism	0.64906
map00062	Fatty acid elongation in mitochondria	map00720	Reductive carboxylate cycle (CO2 fixa...	0.64374
map00720	Reductive carboxylate cycle (CO2 fixa...	map01060	Biosynthesis of plant secondary metab...	0.63662
map00062	Fatty acid elongation in mitochondria	map00625	Tetrachloroethene degradation	0.63580
map00400	Phenylalanine, tyrosine and tryptopha...	map01070	Biosynthesis of plant hormones	0.63105
map00630	Glyoxylate and dicarboxylate metabolism	map00720	Reductive carboxylate cycle (CO2 fixa...	0.62892
map00071	Fatty acid metabolism	map00640	Propanoate metabolism	0.62478
map00190	Oxidative phosphorylation	map00195	Photosynthesis	0.62223
map00062	Fatty acid elongation in mitochondria	map00410	beta-Alanine metabolism	0.61715
map00410	beta-Alanine metabolism	map00640	Propanoate metabolism	0.61670
map00410	beta-Alanine metabolism	map00770	Pantothenate and CoA biosynthesis	0.61237
map01065	Biosynthesis of alkaloids derived fro...	map01070	Biosynthesis of plant hormones	0.61182
map00062	Fatty acid elongation in mitochondria	map00620	Pyruvate metabolism	0.61097
map04070	Phosphatidylinositol signaling system	map04664	Fc epsilon RI signaling pathway	0.60857

6 Discussion

There are still three issues that should be handled.

Firstly, the cross-references between ChEBI and KEGG are incomplete: out of the 3192 distinct compounds retrieved from KEGG pathways, only 2742, or 85.9%, are referenced in the ChEBI database. This means that we need to develop a proper significance metric for pathways. Consider the pathway "Biosynthesis of type II polyketide backbone", where only 5 out of the 19 compounds are mapped to ChEBI: when compared to "Fatty acid elongation in mitochondria", the similarity obtained is 0.85. However, this value is not very significant, since the 14 missing compounds may be very different from those in other pathway and, as such, change the result. This metric would reflect this bias and measure the significance of the similarity obtained. This metric should also address the problem of under annotation in the KEGG database. For instance, both pathways map04530 (Tight junction) and map05218 (Melanoma) have a single compound, C05981. As such, the similarity between them is computed as 1.0, even though they are distinct pathways. Table 2 considers as significant all pathways with at least 10 compounds and with more than 90% of coverage between ChEBI and KEGG.

Secondly, the best match approach does not weight the importance of each compound, but in the future, we intend to improve this approach by giving weight to the compounds based on their information content. This weighting is important in pathways with very frequent compounds, as their presence in two pathways should *not* be taken as a good evidence about their similarity.

Finally, to reduce the effect of the lack of coverage between ChEBI and KEGG, we are studying a new approach that combines both semantic and structural similarity which will be incorporated in a new version of CMPSim. This will guarantee that all compounds annotated to a pathway will be taken into consideration, even if no cross-reference to ChEBI could be found.

7 Conclusions

This paper presents a novel software that calculates semantic similarity between chemical compounds and metabolic pathways. This software is presented to the community as a web-tool with the intent of further development, and not as a final product.

The preliminary results inherent to such a new project show that the approach is meaningful and can potentially find unknown relationships between pathways. In the future, further analysis will be performed, including clustering techniques to group related pathways together and compare the obtained clusters with existing pathway classification schemes.

One of the main challenges we face is the incompleteness of KEGG annotation (some pathways contain as few as one metabolite) and ChEBI↔KEGG cross-references. To address this issue, we plan to develop a hybrid approach where structural similarity is combined with semantic similarity, thus reducing the effect of the incomplete mapping from KEGG to ChEBI.

Nevertheless, we have shown the feasibility of semantic similarity in the context of chemical compounds/metabolic pathways, and provided the community with a web-based tool that can compute such similarities.

8 Availability and requirements

Project name Chemical and Metabolic Pathway Similarity (CMPSim) Tool

Project home page `http://xldb.di.fc.ul.pt/biotools/cmpsim/`

Operating system Broswer Based - Platform independent

Programming languages HTML, Perl, PHP, Python

9 Authors contributions

This work was a collaboration with equal contribution from all authors, under supervision of FC. DVV, JFerreira and JFreire were responsible for the creation of the final sketch of the webtool. CP, HB and TG were responsible for the implementation of semantic similarity. TG and JFerreira wrote the manuscript. All authors reviewed the final form of the manuscript.

10 Acknowledgements

Work supported by FCT's Multiannual Funding Programme and PhD grants SFRH/BD/36015/2007, SFRH/BD/42481/2007 and SFRH/BD/48035/2008.

References

[1] T Dandekar, S Schuster, B Snel, M Huynen, and P Bork. Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochemical Journal*, 343:115–124, 1999.

[2] T. Hancock and H. Mamitsuka. A Markov Classification Model for Metabolic Pathways. *Algorithms for Molecular Biology*, 5(10), 2000. doi: {doi:10.1186/1748-7188-5-10}.

[3] L Zhenping, S Zhang, Y Wang, XS Zhang, and L Chen. Alignment of molecular networks by integer quadratic programming. *Bioinformatics*, 23 (13):1631, 2007.

[4] Y Li, D de Ridder, MJL de Groot, and MJT Reinders. Metabolic pathway alignment between species using a comprehensive and flexible similarity measure. *BMC Systems Biology*, 2(1):111, 2008.

[5] J.C. Clemente, K. Satou, and G. Valiente. Reconstruction of phylogenetic relationships from metabolic pathways based on the enzyme hierarchy and the gene ontology. *Genome Informatics*, 16(2):45, 2005.

[6] R.Y. Pinter, O. Rokhlenko, E. Yeger-Lotem, and M. Ziv-Ukelson. Alignment of metabolic pathways. *Bioinformatics*, 21(16):3401–3408, 2005.

[7] T. Shlomi, D. Segal, E. Ruppin, and R. Sharan. QPath: a method for querying pathways in a protein-protein interaction network. *BMC bioinformatics*, 7(1):199, 2006.

[8] Y Tohsato and Y Nishimura. Metabolic pathway alignment based on similarity between chemical structures. *Information and Media Technologies*, 3(1), 2008. ISSN 1881-0896.

[9] M.Y. Galperin, D.R. Walker, and E.V. Koonin. Analogous enzymes: independent inventions in enzyme evolution, 1998.

[10] Y. Tohsato, H. Matsuda, and A. Hashimoto. A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB'00)*, pages 376–383, 2000.

[11] J. Gamalielsson and B. Olsson. GOSAP: Gene Ontology-Based Semantic Alignment of Biological Pathways. *International Journal of Bioinformatics Research and Applications*, 4(3):274–294, 2008.

[12] P. Willett, J.M. Barnard, and G.M. Downs. Chemical similarity searching. *J. Chem. Inf. Comput. Sci*, 38(6):983–996, 1998.

[13] Gerald M. Maggiora and Veerabahu Shanmugasundaram. Molecular similarity measures. 275, May 2004.

[14] N. Nikolova and J. Jaworska. Approaches to measure chemical similarity-a review. *QSAR & Combinatorial Science*, 22, 2003.

[15] P. Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull. Soc. Vaudoise Sci. Nat*, 37:547–579, 1901.

[16] H. Wolosker, E. Dumin, L. Balan, and V.N. Foltyn. d-Amino acids in the brain: d-serine in neurotransmission and neurodegeneration. *FEBS Journal*, 275(14):3514–3526, 2008.

[17] M. Kanehisa, S. Goto, M. Hattori, K.F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic acids research*, 34(Database Issue):D354, 2006.

[18] K. Degtyarenko, P. Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcantara, M. Darsow, M. Guedj, and M. Ashburner. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 2007.

[19] C. Pesquita, D. Faria, A.O. Falcão, P. Lord, and F.M. Couto. Semantic Similarity in Biomedical Ontologies. 2009.

[20] Francisco Couto, Mário J. Silva, and Pedro Coutinho. Semantic similarity over the gene ontology: Family correlation and selecting disjunctive ancestors. In *ACM CIKM - Conference in Information and Knowledge Management*, October 2005.

[21] J.M. Berg, J.L. Tymoczko, and L. Stryer. *Biochemistry*, chapter 19–The Light Reactions of Photosynthesis. Freeman, 2007.