An Analysis of Differences In Biological Pathway Resources

Lucy L. Wang, John H. Gennari, and Neil F. Abernethy
Department of Biomedical Informatics and Medical Education,
University of Washington
Seattle, Washington 98195
Email: lucylw@uw.edu, gennari@uw.edu, neila@uw.edu

***Abstract*—Integrating content from multiple biological pathway resources is necessary to fully exploit pathway knowledge for the benefit of biology and medicine. Differences in content, representation, coverage, and more occur between databases, and are challenges to resource merging. We introduce a typology of representational differences between pathway resources and give examples using several databases: BioCyc, KEGG, PANTHER pathways, and Reactome. We also detect and quantify annotation mismatches between HumanCyc and Reactome. The typology of mismatches can be used to guide entity and relationship alignment between these databases, helping us identify and understand deficiencies in our knowledge, and allowing the research community to derive greater benefit from the existing pathway data.**

***Keywords*—pathway database, knowledge representation, resource comparison**

I. Introduction

Describing and studying biological pathways is necessary for understanding biological and disease processes. Biological functions and processes follow from complex networks of interactions among gene products and molecules. Through the study of pathways of known biochemical reactions, we can gain deeper insights into these interactions. Many of these relationships and reactions have been catalogued in pathway resources such as Reactome, BioCyc, KEGG, and others [1–5].

As of April 2016, PathGuide, a pathway resource aggregator, lists 547 pathway resources [6], each providing specialized knowledge in niche areas of biology. Efforts have been made to integrate some of these databases. PathwayCommons catalogs human pathway resources under a unified biological pathway exchange umbrella (BioPAX), allowing easier querying of pathways across 22 different resources [7, 8]. Tools such as Consensus Pathway DB [9] and hiPATHDB [10] offer querying and visualization of pathways from multiple databases. Statistical frameworks like R Spider seek to probabilistically combine protein interactions from various pathway databases into merged networks [11]. These tools improve querying of multiple resource, and pave the way towards more comprehensive network models of human biological processes.

Some work has also been done in inter-resource comparison, quantifying the overlap between different databases [12–15]. These comparison studies emphasize differences in entity membership in pathways and differing counts of unique entities and pathways, but do not focus on cross-resource entity alignment. Existing tools for entity normalization of proteins [16] and metabolites [17] may provide a starting point for alignment. Other studies emphasize aligning metabolic pathways of different species in order to find analogous but missing relationships [18, 19], merging resources for combined network analysis [11, 20], or defining conserved pathway elements across existing pathway resources [21]. However, although they represent progress, the tools and studies mentioned above accomplish goals that do not include aligning representations across resources.

Given the number and uniqueness of pathway resources, inter-resource merging is a challenge. In order to successfully align and integrate the content of multiple knowledge bases, we must contend with variability in content correctness, standards usage, knowledge representation choices, and coverage. Pathway data sharing standards such as BioPAX, SBML, and PSI MI [8, 22, 23] assist in the interchange of pathway resources, but even resources available in the same standard still retain differences in content and representation. Nonetheless, our goal is to align knowledge, so that users can benefit from a semantic union across multiple resources.

To align resources, we must comprehensively understand the types of differences one may encounter. Stobbe et al. have made an excellent start in this direction, providing numerous examples and descriptions of the sorts of differences among metabolic pathway resources [13, 24]. Here, we extend this work, aiming at a typology

This work was supported by the National Library of Medicine (NLM) Training Grant T15LM007442.

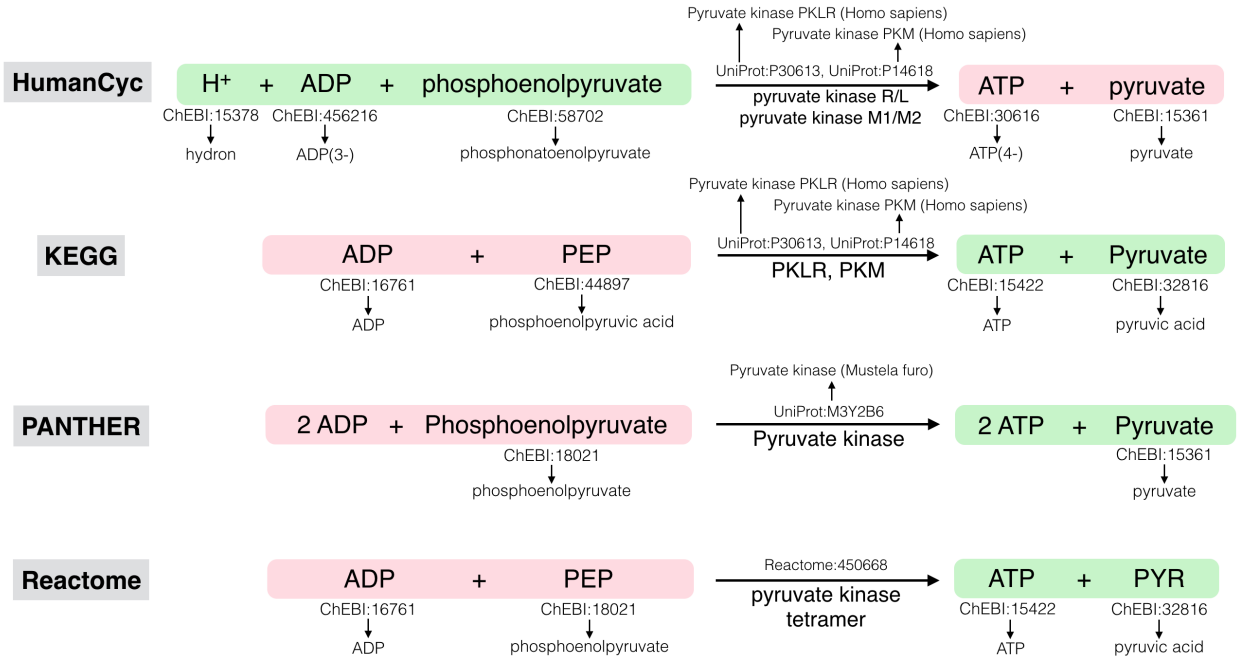

Fig. 1: The conversion of phosphoenolpyruvate and ADP into pyruvate and ATP assisted by the enzyme pyruvate kinase, as represented by HumanCyc, KEGG, PANTHER, and Reactome. The display name for each entity is given, along with ChEBI or UniProt identifiers where available. Entities related to the reaction by the BioPAX left property are red, and entities related by the BioPAX right property are green.

of mismatches among pathway resources. In particular, we describe and give examples of mismatches in (a) annotation, (b) existence, (c) reaction semantics, and (d) granularity. By classifying mismatches, we enable the better understanding and discussion of resource differences, and allow for improved consensus formation in multiple pathway resource applications.

We also present some results in quantifying annotation mismatches between two popular human pathway resources: HumanCyc and Reactome. Results demonstrate the pervasiveness of representational differences and suggest further work towards consensus pathway representations. Understanding the types of mismatches that exist between resources is a first step towards expanding and deriving the full benefit of our pathway knowledge.

II. Mismatches in Pathway Resources: A Typology

To provide examples of mismatches, we retrieved reaction representations from HumanCyc, KEGG, PANTHER, and Reactome. Fig. 1 shows several different representations of a step of glycolysis in *Homo sapiens*: the conversion of phosphoenolpyruvate and ADP to pyruvate and ATP modulated by the enzyme pyruvate kinase. In this single, well-studied biochemical reaction, we see a variety of important mismatches, of which a subset are described below.*

A. Annotation

We first consider annotation mismatches on the participating physical entities. Inconsistencies arise when two pathway resources refer to the same entity with different identifiers or different names. Pyruvate is represented by all four resources (Fig. 1), but is annotated with two identifiers, ChEBI:15361 (HumanCyc and PANTHER) and ChEBI:32816 (KEGG and Reactome). The ChEBI:15361 entity "pyruvate" and ChEBI:32816 entity "pyruvic acid" are conjugate acids and bases of one another in ChEBI. The display name for pyruvate also differs between resources, and is given as "pyruvate" (HumanCyc), "Pyruvate" (KEGG, PANTHER), or "PYR" (Reactome). Differences in identifiers and names are also seen for all other participants in this reaction.

*Pathways were retrieved from Reactome v55 (http://reactome.org) and HumanCyc v19.5 (http://humancyc.org) BioPAX3 exports and through PathwayCommons v7 [7]. Glycolysis pathways for KEGG and PANTHER are located at http://purl.org/pc2/7/#Pathway_307add3cea6530288cc1016267ec055b and http://identifiers.org/panther.pathway/P00024 respectively and are supplemented by the pathway diagrams at http://kegg.jp and http://pantherdb.org.

In order to resolve these mismatches, we must either enforce consistent labeling of entities across resources, or somehow infer alignment of similar but differently annotated entities across resources. The former strategy is usually impractical; in this case, we can infer similarity by treating ChEBI identifiers that refer to conjugate acid/base pairs as synonyms.

A second type of annotation mismatch occurs when entities lack cross-referenced identifiers, e.g., no identifiers are given for ADP or ATP in PANTHER pathways. Other features such as string name, entity relationships, and local network topology can be used to align entities between resources when identifiers are insufficient.

B. Existence

Existence refers to missing or extraneous physical entities, reactions, relationships, or information, e.g., entities that participate in a reaction or reactions that are members of a pathway in one resource but not another, or a connection between two reactions that occurs in one resource but not another. In Reactome, for example, the conversion of fructose 6-phosphate to fructose 2,6-biphosphate is a reaction in the glycolysis pathway. This reaction is not included in the glycolysis pathway of the other three resources. Although the reaction involves entities that participate in glycolysis, there is uncertainty in whether it is important to the overall process.

Another example of an existence mismatch is the inclusion of H+ in the conversion of phosphoenolpyruvate to pyruvate in HumanCyc (Fig. 1). The ion is included in order to balance reaction charge, but according to BioPAX3 documentation, reaction participants should be neutral and ions such as H+ and Mg2+ are not recommended for inclusion [25]. Other potential existence mismatches could occur if one resource lacks or is missing relevant information about a relationship between two entities, or one resource specifically negates the existence of a relationship asserted in another resource.

Existence mismatches can be resolved by taking the most common representation between many resources (democratic) or by integrating all possible representations (exhaustive). Although an exhaustive consensus method is unlikely to leave out information, it may, however, produce a large and unwieldy alignment.

C. Reaction semantics

Many differences in reaction representation have been described in Stobbe et al, such as using the terms left and right, product and substrate, and input and output to describe participants in reactions [24]. In BioPAX, the properties conversionDirection, stepDirection, left, and right are used to indicate reaction direction, as well as the identities of reactants and products [25]. In KEGG, PANTHER, and Reactome, phosphoenolpyruvate is labeled left and pyruvate right, with a reaction direction of left-to-right. However, in HumanCyc, phosphoenolpyruvate is labeled right and pyruvate left and the reaction direction is right-to-left, a choice dictated by the Enzyme Commission system [2]. Even though HumanCyc is in the minority, its choice follows recommendations from the BioPAX3 specifications [25].

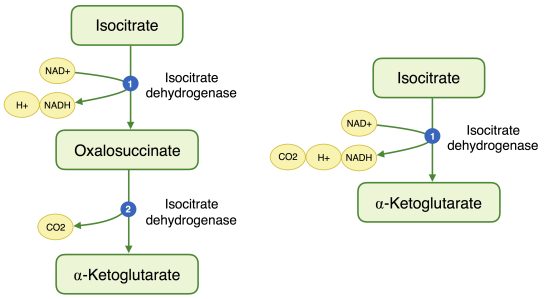

Fig. 2: The oxidative decarboxylation of isocitrate can be represented as a two-step process with an oxalosuccinate intermediary *(left)* and as a one-step process *(right)*.

Resolving this type of semantic mismatch between resources requires knowledge about the ordering of reactions, which can be derived from pathway design, or when reactions are taken out of context, may depend on chemical kinetics and the reacting environment. For well-studied pathways, a consensus ordering usually exists. When participant left and right labels differ between resources and ordering is unclear, the BioPAX pathwayOrder object (designed to relay reaction topology) can sometimes be used along with reaction direction to infer the correct sequence.

D. Granularity

Mismatches of granularity occur when resources represent the same entity or process in different degrees of detail. One example is complex naming. Many reaction enzymes are complexes made up of multiple protein subunits. A reaction may be annotated with a protein modifier, when in actuality, it is catalyzed by a complex: a dimer, trimer etc. In Fig. 1, Reactome makes this distinction by annotating to the "pyruvate kinase tetramer," a complex composed of the pyruvate kinase protein referenced from the other three resources. Due to the lack of standardized complex naming, however, we often cannot easily align complexes and proteins between resources.

Another type of granularity mismatch occurs at the reaction level. For example, one resource may choose to

represent the elementary steps of a reaction, including intermediate chemical species. A single reaction in one resource may be represented as several in another, with the same ultimate inputs and outputs. For example, the oxidative decarboxylation of isocitrate is a two step process, modified by the enzyme isocitrate dehydrogenase, producing α-ketoglutarate from isocitrate via an oxalosuccinate intermediate. The reaction can be represented both with and without the intermediate species, as in Fig. 2. In these cases, we can study the ultimate inputs and outputs of ordered reaction sequences to determine the appropriate reaction-level alignment.

III. Annotation differences between two resources

We identify and enumerate mismatches in entity annotation between two exemplar resources: HumanCyc and Reactome. Compared to other mismatches, a disagreement in the annotation of entities could be viewed as primary: if two resources disagree on physical entities, then they are also likely to disagree on the reactions and pathways in which these entities participate.

The most confident match between entities in two resources arises when both identifiers and names match. For example, the molecule ATP matches both on name and ChEBI identifier for KEGG and Reactome (Fig. 1). Less confident are identifier matches without string name matches (e.g. HumanCyc and KEGG use different names for the entity cross-referenced to UniProt:P30613), and string name matches without identifier matches (e.g. HumanCyc and PANTHER cross-reference to different ChEBI identifiers for the entity named "phosphoenolpyruvate").

From HumanCyc and Reactome, we extract all proteins and small molecules with cross-referenced identifiers (UniProt for proteins, ChEBI for small molecules) and names. String names are taken as all objects of the BioPAX properties name, displayName, and standardName on the entity of interest. Using only string names and UniProt/ChEBI identifiers, there are four possible ways that entities can match between these two resources. Entities in HumanCyc can match to entities in Reactome on ID and name (+I/+N), ID but not name (+I/-N), name but not ID (-I/+N), and on neither ID nor name (-I/-N). For this initial analysis, we define string name matches as case-insensitive equivalence, so small differences in spelling do not produce a match.

For each entity in HumanCyc, SPARQL queries are used to determine whether a matching entity exists in Reactome, and similarly, Reactome entities are matched to HumanCyc entities. Resulting matches for proteins are given in Table I. Out of 2737 unique HumanCyc proteins, 2078 (75.9%) match to Reactome entities using identifiers and/or string names. Out of 16949 unique Reactome proteins, 2973 (17.5%) match to a HumanCyc protein on identifiers and/or name. Reactome references many protein isoforms, causing the large imbalance in unique protein counts between the two resources. These match ratios are illustrated in Fig. 3.

TABLE I: Proteins matches between HumanCyc and Reactome on UniProt identifiers and names

HumanCyc protein matches to Reactome

	+N	-N	Total
+I	1264	759	2023
-I	55	659	714
Total	1319	1418	2737

Reactome protein matches to HumanCyc

	+N	-N	Total
+I	1495	1390	2885
-I	88	13976	14064
Total	1583	15366	16949

TABLE II: Small molecule matches between HumanCyc and Reactome on ChEBI identifiers and names

HumanCyc small molecule matches to Reactome

	+N	-N	Total
+I	247	140	387
-I	479	744	1223
Total	726	884	1610

Reactome small molecule matches to HumanCyc

	+N	-N	Total
+I	425	276	701
-I	890	1300	2190
Total	1315	1576	2891

Table II shows matches for small molecules. In HumanCyc, 866 (53.8%) out of 1610 small molecules match on annotation to an entity in Reactome. In Reactome, 1591 (55.0%) out of 2891 small molecules match to entities in HumanCyc, with a large proportion (890 out of 1591) matching on string names only.

Cross-referenced identifiers are the gold standard of matching between two resources. Therefore, groups +I/+N and +I/-N likely consist of true matches. Group -I/+N can be used to learn about representational differences. Some of the cross-references for entities in this group point to secondary accession identifiers, which

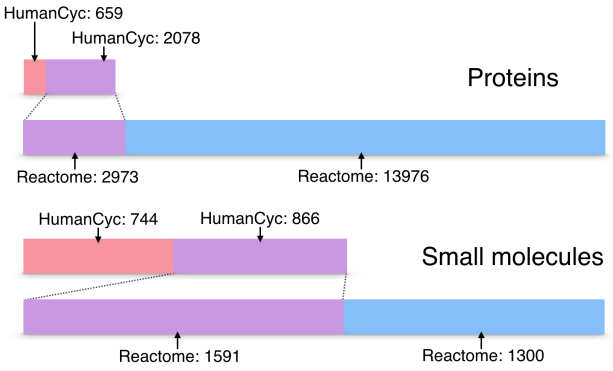

Fig. 3: Protein *(top)* and small molecule *(bottom)* matches between HumanCyc and Reactome based on annotation, consisting of unmatched HumanCyc entities (red), unmatched Reactome entities (blue), and matched entities between the two resources (purple).

redirect to other identifiers in the same database. For example, UniProt:A0AVP9 redirects to UniProt:Q8IWU4, the entity Zinc transporter 8. For small molecules only, we also find annotation to ChEBI conjugate acids or bases (e.g., HumanCyc annotates with ChEBI:456216 (ATP(3-)), a conjugate base of ChEBI:16761 (ATP), which is used in Reactome), or annotation to tautomers (e.g., ChEBI: 16828 and ChEBI:57912 for L-tryptophan and the L-tryptophan zwitterion respectively). Annotation mismatches of the above subtypes are detected by querying the UniProt or ChEBI APIs using the BioServices 1.4.8 Python package [26].

Within -I/+N matches, the 55 HumanCyc and 88 Reactome proteins had 208 pairwise string name matches. Of these, 28 pairs had cross-referenced identifiers that are UniProt secondary accession IDs, indicating that they likely refer to the same entity. We could not confirm the identities of the other 180 pairs through UniProt accession identifiers. For small molecules in the -I/+N group, the 479 HumanCyc and 890 Reactome molecules had 1869 pairwise string name matches. Of these, at least 1506 pairs referred to similar entities. Annotation to ChEBI conjugate acids or bases accounted for the majority of these (1122), followed by annotation with ChEBI tautomer IDs (301), and ChEBI secondary accession numbers (83).

IV. Discussion

In order to reduce redundancy and errors when merging information from different knowledge bases, we must correctly align entities and other assertions between resources. Entity alignment is a necessary first step before we can clarify higher-order concepts such as complexes, reactions, and pathways. As demonstrated using HumanCyc and Reactome, many proteins and small molecules can be matched between two resources using annotation features such as cross-referenced identifiers and string names. Among entities that share only string names, many have related identifiers that can be matched computationally. Related identifiers can be used to help improve the accuracy of annotations.

Moving beyond annotation, other issues of semantics and granularity come into play. For future work, we intend to incorporate other features, such as entity relationships and graph properties like degree and bipartite connectivity to assist in entity alignment.

Several limitations exist in this work. First, we only compared entities between two pathway resources, HumanCyc and Reactome. We expect to expand our analysis to include other resources as well. Although some of our current methods rely on BioPAX, our general ideas about physical entities and their annotations can be applied to data represented using other biological pathway knowledge standards.

Another limitation arises in the way we identify annotation mismatches. We only assessed proteins and small molecules with UniProt or ChEBI identifiers, excluding those entities without cross-references or with cross-references to other databases. This was partially for simplicity and partially to limit the size of the comparison problem. For example, an agreement on one set of identifiers and a disagreement on another yields yet another class of mismatches.

Lastly, we were limited by our use of 100% string name matching to identify potential matched entities. By doing so, we limit our ability to detect positive matches and yield more conservative results, e.g., “fructose 1,6 bisphosphate” does not match to “D-fructose 1,6-bisphosphate”; the second is a stereoisomer of the first (generic) molecule, and they may play similar roles in reactions. Fuzzy string matches may perform better. However, we want to minimize the false positive rate, e.g., “fructose 1,6-bisphosphate” and “fructose 2,6-bisphosphate” only differ by one character but refer to different molecules. With these caveats, the typology we present affords an opportunity to test different algorithms for the systematic alignment of pathway resources.

V. Conclusion

The complexity of pathway content is a barrier to resource integration, but as described above, we are also challenged by representational and content differences. Standards like BioPAX help clarify some differences between resources, but they do not solve all problems

of interoperability. In order to draw from the spectrum of knowledge we have built as a community, the content of these resources must be aligned and integrated into something greater than the parts. Doing so involves identifying the differences between resources, and resolving those differences to understand shared meaning. Our results show that a sizable portion of physical entities can be aligned between pathway resources using existing cross-referenced identifiers and string names. However, annotation features alone are likely insufficient for matching a majority of entities between resources. Knowledge of entity relationships, reaction semantics, granularity, and more about these resources is necessary to create and evaluate potential alignments. Much of the work can be done computationally, and the typology above should guide the engineering of future matching algorithms.

To align and integrate knowledge across resources, the research community must have strategies for resolving these different sorts of mismatches. Some mismatches, such as those of annotation, can largely be resolved using the existing data. Other issues of semantics, such as differences in how standard languages are used to express the same knowledge, pose a bigger challenge. Resource developers should be allowed to make different choices in knowledge representation. However, this flexibility should not come at the cost of increased error or decreased interoperability. A better understanding of how specific mismatches occur will provide an incentive for resources to work toward interoperable data and representations.

ACKNOWLEDGEMENTS

The authors thank Peter Karp for helpful comments on an early draft of this paper.

REFERENCES

[1] D. Croft, A. Mundo, and R. Haw et al. The reactome pathway knowledgebase. *Nucleic Acids Res*, 42(Database issue):D472–477, 2013.

[2] P. Romero, J. Wagg, M. Green, D. Kaiser, M. Krummenacker, and P. Karp. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biology*, 6(R2):1–17, 2004.

[3] M. Kanehisa and S. Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28:27–30, 2000.

[4] P. Thomas, M. Campbell, and A. Kejariwal et al. Panther: a library of protein families and subfamilies indexed by function. *Genome Res*, 13:2129–2141, 2003.

[5] M. Kutmon, A. Riutta, and N. Nunes et al. Wikipathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Res*, 44(D1):D488–D494, 2016.

[6] G. Bader, M. Cary, and C. Sander. Pathguide: a pathway resource list. *Nucleic Acids Res*, 34(Database issue):D504–506, 2005.

[7] E. Cerami, B. Gross, and E. Demir et al. Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res*, 39(Database issue):D685–690, 2011.

[8] E. Demir, M. Cary, and S. Paley et al. The biopax community standard for pathway data sharing. *Nature Biotechnology*, 28(9):935–42, 2010.

[9] A. Kamburov, C. Wierling, H. Lehrach, and R. Herwig. Consensuspathdb – a database for integrating human functional interaction networks. *Nucleic Acids Res*, 37(Database issue):D623–628, 2009.

[10] N. Yu, J. Seo, and K. Rho et al. hipathdb: a human-integrated pathway database with facile visualization. *Nucleic Acids Res*, 40(Database issue):D797–802, 2012.

[11] A. Antonov, E. Schmidt, S. Dietmann, M. Krestyaninova, and H. Hermjakob. R spider: a network-based analysis of gene lists by combining signaling and metabolic pathways from reactome and kegg databases. *Nucleic Acids Res*, 38(Web Server issue):W78–83, 2010.

[12] D. Soh, D. Dong, Y. Guo, and L. Wong. Consistency, comprehensiveness, and compatibility of pathway databases. *BMC Bioinformatics*, 11:449–64, 2010.

[13] M. Stobbe, S. Houten, G. Jansen, A. van Kampen, and P. Moerland. Critical assessment of human metabolic pathway databases: a stepping stone for future integration. *BMC Systems Biology*, 5:165–183, 2011.

[14] T. Altman, M. Travers, A. Kothari, R. Caspi, and P. Karp. A systematic comparison of the metacyc and kegg pathway databases. *BMC Bioinformatics*, 14:112, 2013.

[15] S. Chowdhury and R. Sarkar. Comparison of human cell signaling pathway databases – evolution, drawbacks and challenges. *Database*, page bau126, 2015.

[16] Y. Hu, Y. Li, H. Lin, Z. Yang, and L. Cheng. Integrating various resources for gene name normalization. *PLoS One*, 7(9):e43558, 2012.

[17] G. Wholgemuth, P. Haldiya, E. Willighagen, T. Kind, and O. Fiehn. The chemical translation service – a web-based tool to improve standardization of metabolomic reports. *Bioinformatics*, 26(20):2647–8, 2010.

[18] F. Ay and T. Kahveci. Submap: aligning metabolic pathways with subnetwork mappings. *Journal of Computational Biology*, 18(3):219–35, 2011.

[19] R. Alberich, M. Llabrés, D. Sánchez, M. Simeoni, and M. Tuduri. Mp-align: alignment of metabolic pathways. *BMC Systems Biology*, 8:58, 2014.

[20] D. Petrochilos, A. Shojaie, J. Gennari, and N. Abernethy. Using random walks to identify cancer-associated modules in expression data. *BioData Mining*, 6(1):17, 2013.

[21] A. Morgat, E. Coissac, E. Coudert, K. Axelsen, G. Keller, and A. Bairoch et al. Unipathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Research*, 40(Database issue):D761–769, 2012.

[22] M. Hucka, A. Finney, and H. Sauro et al. The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–31, 2003.

[23] H. Hermjakob, L. Montecchi-Palazzi, and G. Bader et al. The hupo psi's molecular interaction format – a community standard for the representation of protein interaction data. *Nature Biotechnology*, 22:177–83, 2004.

[24] M. Stobbe, G. Jansen, P. Moerland, and A. van Kampen. Knowledge representation in metabolic pathway databases. *Brief Bioinform*, 15(3):455–470, 2014 May.

[25] BioPAX Workgroup. Biopax - biological pathways exchange language, level 3, release version 1 documentation. July 2010.

[26] T. Cokelaer, D. Pultz, L. Harder, J. Serra-Musach, and J. Saez-Rodriguez. Bioservices: a common python package to access biological web services programmatically. *Bioinformatics*, 29(24):3241–2, 2013.