The Web as a Distributed Biochemical Reactor: Semantically Enabled Metabolic Fate Prediction

Leonid L. Chepelev[1,*] and Michel Dumontier[1,2,3]
[1] Department of Biology, Carleton University, 1125 Colonel by Drive, Ottawa, Canada
[2] Institute of Biochemistry, Carleton University, 1125 Colonel by Drive, Ottawa, Canada
[3] School of Computer Science, Carleton University, 1125 Colonel by Drive, Ottawa, Canada

ABSTRACT

Over the past decade, we have witnessed Semantic Web Technologies in a tremendous range of applications, from representation and integration of knowledge to enable rich, multi-domain queries to automation of computational workflows using a range of Semantic Web Service platforms. By the virtue of its design, the Semantic Automated Discovery and Integration (SADI) framework provides excellent flexibility and facile discoverability and integration of disparate semantic web services. In this manuscript, we shall demonstrate the application of SADI in the design and implementation of a semantically enabled biochemical reactor framework. In this framework, enzymes and reactions are represented as distinct web services with embedded computational cheminformatics and computational chemistry tools for prediction of reaction feasibility and kinetics. We use a standalone client to reason over molecular descriptions and invoke appropriate SADI services in order to create a biochemical model of metabolism. We provide a simple example of metabolic fate prediction for benzene to demonstrate that this unique modelling framework can provide customizable dynamic and static models of metabolism.

1 INTRODUCTION AND OVERVIEW

Metabolic fate prediction (MFP) is a central problem for the pharmaceutical industries and numerous academic researchers alike. Despite prior extensive work in the development of a vast array of algorithms and approaches to enable the various components of MFP such as molecular similarity matching, reaction prediction, reaction energy calculation, protein binding, and toxicological prediction to name a few, their integration to address arbitrary MFP problems has remained an arduous task accessible *only through manual efforts* of human experts. There has been *no effort* to address MFP through a single, customizable system. Though commercial MFP packages exist, they offer only a closed collection of enzymes and predictive algorithms with no ability to interface with the existing and ever-growing collections of chemical data and services, let alone knowledgebases.

Before we could begin addressing MFP with SADI web services, we established Semantic Web-based solutions for overcoming barriers to the unification of chemical information and resources on the representational, conceptual, and computational levels. With Chemical Entity Semantic Specification (CHESS), we were able to represent chemical entities, from atoms to reactions, in RDF (Chepelev *et al.*, 2011a). Development of CHEMINF, gave us a standardized way of annotating chemical entities with information relevant to their subsequent classification and use in further investigations such as QSAR studies and reaction feasibility calculations relevant to MFP (Hastings *et al.*, 2011). Conceptual barriers to integration of chemical information were addressed by generating machine-understandable, unambiguous definitions of classes of small molecules, be they structural or functional, within chemical OWL ontologies and the use of said ontologies to classify arbitrary chemical compounds (Chepelev *et al.*, 2012). Finally, an example of overcoming computational interoperability barriers in cheminformatics has been achieved with the SADI framework coupled with machine reasoning and web service orchestration by the SHARE client (Chepelev *et al.*, 2011b).

With these components in place, it has become possible to address MFP for arbitrary small molecules. Recall that many known enzymes operate on a well-defined class of substrates and catalyze a well-defined set of biological transformations. It is thus possible to define enzymatic inputs semantically, using the aforementioned chemical ontology construction methodology: an arbitrary molecular collection can be screened for similarities which are then used to generate an OWL description for the chemical class to which all molecules in said collection belong (Chepelev et al., 2012). Any arbitrary molecule defined in CHESS format can then be accurately classified as a potential enzyme substrate using a reasoner. Formally defined enzyme substrate classes can be used in SADI service input descriptions, paving the way for an iterative application of this approach to known enzymes and the possibility of turning the World Wide Web into a distributed *in silico* biochemical reactor.

Furthermore, since all enzymes obey the same fundamental laws of physics, it is also possible to employ uniform and consistent approximations to compute the relevant kinetic constants for each enzymatic reaction. In theory, given a sufficiently advanced computer and enough starting information, it is possible to accurately compute the reaction energy profile and obtain reaction kinetics for each reaction in an arbitrary pathway. Practically, such calculations may be impossible due to missing enzyme structures and a lack of ability to obtain a reasonable homology model, or infeasible due to prohibitive computational cost for all but the very simplest of enzymatic systems. In such cases, approximations or database lookups are currently necessary.

Our approach to MFP on cellular scale calls for the construction of thousands of enzyme-mimicking SADI services:

* Corresponding author: leonid.chepelev@gmail.com

each with a chemical ontology-defined input class, and each producing a fully characterized CHESS-encoded reaction and product description. Each service therefore contains the reactor, the thermodynamic approximation module, and the kinetics approximation module embedded. These services are called upon in an appropriate, sequence based on the input molecule using a standalone client, with the orchestration made possible by SADI-conforming service input description and molecular classification through reasoning over an OWL-encoded chemical ontology.

In this work, we demonstrate the semantic distribution and subsequent concerted action of these services for MFP of benzene, with a very limited set of both enzymatic and non-catalyzed transformations. Resulting from this is a dynamic model of benzene metabolism which allows us to supplement the information resulting from the assessment of the toxicity of all of the chemical intermediates in this metabolic pathway with their bioaccumulation information. We argue that this helps pinpoint the metabolites with the greatest impact on small molecule toxicity, and therefore to identify potential molecular sites that could be modified to avoid a particular metabolic route. Finally, we discuss how the integrated MFP framework presented here can be easily extended for more accurate and comprehensive predictions.

2 METHODS

Our MFP framework distributes the biochemical reactor functionality with enzyme and reaction mimicking services that are identified within a central registry and integrates these services using a standalone Semantic Service Matching Client that iterates over the service registry until no further unique reactions can be identified (Fig. 1).

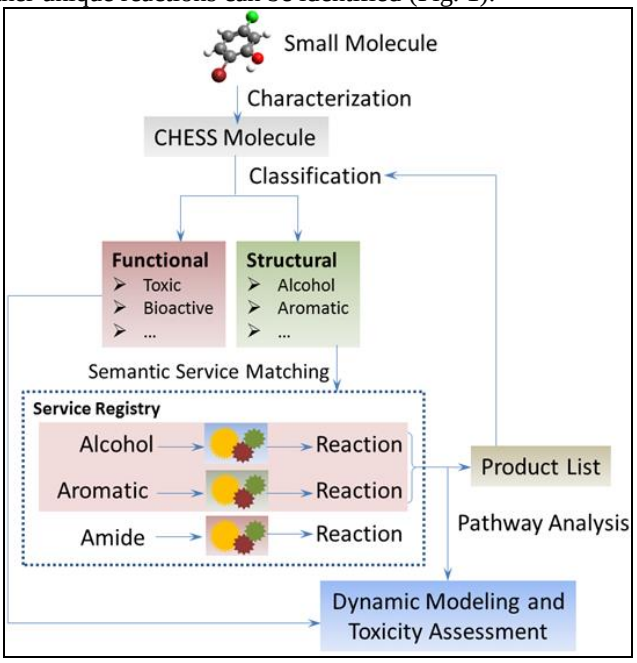

Fig. 1. Overview of the proposed SADI-enabled MFP framework.

2.1 Chemical Classification and Service Selection

For semantically enabled molecular classification, we used our previously developed work (Chepelev *et al.*, 2012). Briefly, a CHESS molecular representation that includes an enumeration of molecular fingerprint features of a given molecule is used as an input to the Pellet reasoner along with a chemical ontology containing definitions of input chemical class for every enzyme-mimicking web service.

The resultant classification of a given molecule then permits us to construct a list of web services to potentially query, given that they have not been queried with the same request in the prior iterations within a given MFP task. Iterative matching of reaction products to enzymes continues until no further new reactions are identified. While these operations can be partially achieved with the involvement of the SHARE client, we have used our own standalone client to expedite this workflow.

2.2 Reactive Transformations

For this demonstration, transformation patterns have been obtained from manually curated literature to cover a limited set of biochemical enzymes. These patterns were encoded as reactive SMARTS for RDKit 2012.03.1 to generate reaction products based on the input molecule. Each resultant reaction was then assessed for feasibility and kinetics.

2.3 Reaction Feasibility and Kinetics

Reaction feasibility was approximated using a multi-tiered procedure. First, sufficient substrate similarity was assured, and best similarity score as per the Tanimoto similarity coefficient for the queried substrate against a collection of known substrates was computed with the Chemistry Development Kit (CDK) 1.2.5. If the arbitrarily set cutoff value of 0.6 was met, reaction feasibility was then assessed using reaction free energy calculations in gas-phase at the AM1 level in MOPAC 7.1 software. Binding kinetics (K_M) were approximated by scaling the known K_M for the best matching substrate by the previously calculated similarity score. While these operations do not provide the most rigorously correct kinetics, we shall point out that this was not the purpose of this study.

2.4 Dynamical Model Simulation

The resultant reactions were collected from the SADI web service outputs in RDF and converted into an SBML file, which was then submitted to COPASI software for time course analysis to generate a dynamic model of benzene metabolism.

3 RESULTS

The target organism studied here was *Homo sapiens*. The reactions chosen to model in this study were the hydroxylation of aromatic moieties including benzene by CYP2E1, the auto-oxidation of hydroquinones and catechols (actually

a rolled-up and simplified auto-oxidative pathway), the two-electron reduction of quinones by enzymes (we are aware of non-enzymatic two-electron reduction of o-quinones, but assume Quinone Reductase, QR, does it much more efficiently), and the constant elimination of all the hydroxylated species lacking a ring carbonyl through Phase II metabolism. The predicted transformations and corresponding kinetics are shown (**Table 1**, **Fig. 2**, **Fig. 3**).

Substrate	Product	Reaction	k_{cat} (s^{-1})	K_M (mM)
Benzene	Phenol	CYP2E1	10.4	227.4
Phenol	Catechol	CYP2E1	$1.2x10^4$	40.4
Phenol	Hydroquinone	CYP2E1	$1.2x10^4$	40.4
Catechol	Hydroxyquinol	CYP2E1	13.2	47.6
Hydroquinone	Hydroxyquinol	CYP2E1	32.8	31.6
Catechol	o-Quinone	Auto-oxidation	0.5	-
Hydroxyquinol	2-HO p-Quinone	Auto-oxidation	30.3	-
Phenol	Eliminated	Phase II	0.1	-
Pyrogallol	3-HO o-Quinone	Auto-oxidation	2.2	-
3-HO o-Quinone	Pyrogallol	QR	$2.0x10^3$	346.2
2-HO p-Quinone	Hydroxyquinol	QR	$2.4x10^2$	346.2

Table 1. Examples of the transformations predicted for benzene.

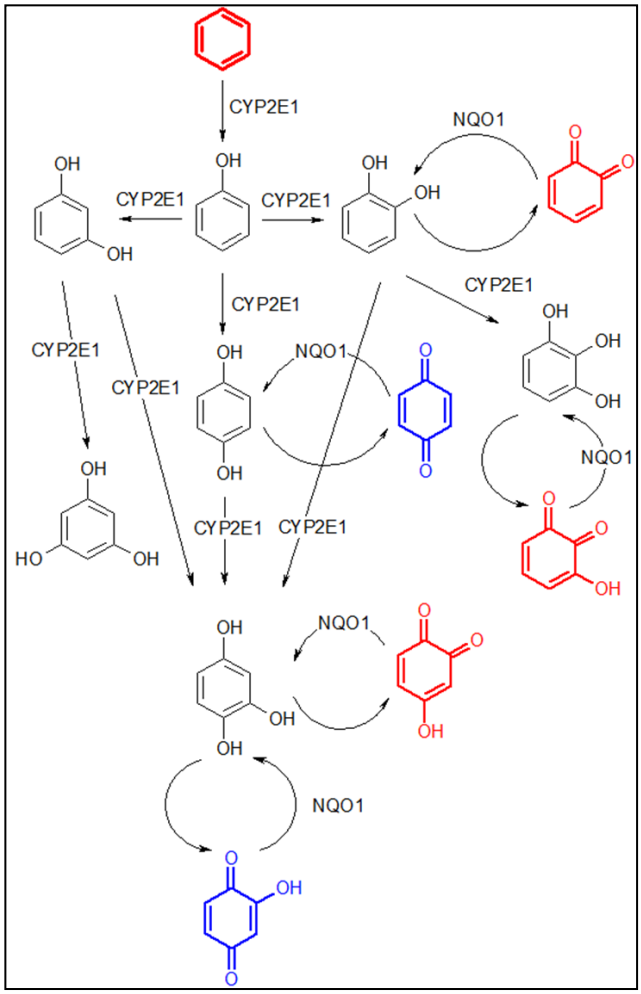

Fig. 2. Visualization of predicted benzene metabolism. Highly toxic compounds are shown in red, moderately toxic compounds are shown in blue.

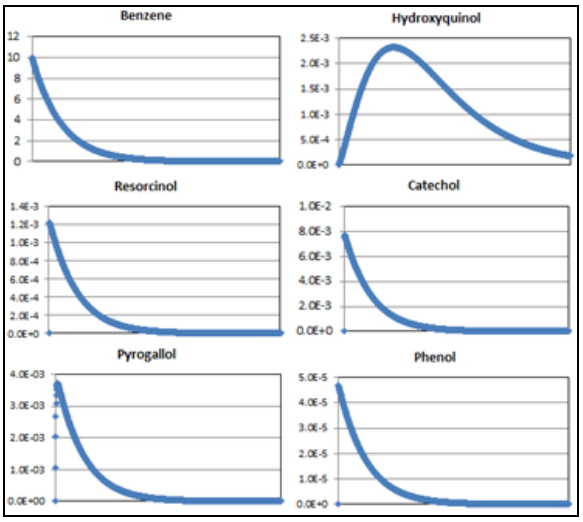

Fig. 3. Dynamics of concentration changes over time for selected benzene metabolism intermediates.

4 DISCUSSION AND CONCLUSIONS

With just four types of enzymatic and non-enzymatic transformations (**Table 1, Fig. 2**) we have been able to predict the majority of the transformations described for benzene in humans in the literature (Snyder and Hedli, 1996). In addition to this, we have characterized biologically plausible transformations that have not yet been described. As can be seen in the examples provided in Table 1, each reaction and enzyme mimicking service correctly identified an exhaustive list of products for each incoming substrate, which allowed us to construct not only the static (**Fig. 2**) but also the dynamic (**Fig. 3**) predictive models of benzene metabolism, whereby enzyme CYP2E1 activates benzene, which subsequently undergoes redox cycling through enzyme NQO1 and constant elimination in Phase II metabolism. Each enzyme mimicking service had a relatively computationally inexpensive approximation of kinetics built in, but unfortunately, quantitatively accurate MFP may be difficult to attain without much more time-consuming calculations. Fortunately, however, much more accurate and more computationally expensive enzyme web services can be constructed and selected at will in one's MFP task with minor modifications of an initial query. Thus, *derivation of accurate kinetics was not the focus of this study*. The correct matching of substrates to reactions and automated construction of an overall qualitative picture of benzene metabolism, however, is a testament to the viability of SADI MFP.

The process of constructing an integrated model for MFP of an arbitrary small molecule is quite arduous and requires a human expert to integrate a number of computational packages and databases. In this work, we achieved seamless integration and streamlining of this procedure using our standalone client, which examined the relevant structural features of the molecule at hand, logically inferred its chemical classification, and identified the enzymes that operated

on molecules of a given class, accurately and efficiently. For example, we have not observed products of benzene metabolism with more than three alcohol functional groups - the definition of the input class for the CYP2E1 service allowed compounds with a maximum of two hydroxyl groups.

Our framework is easily extensible, provided that new reaction simulation services are registered within a central service registry, no matter which physical machine they reside upon. In this sense, the World Wide Web has just become a distributed biochemical reactor. Although this reactor contains few services currently, the highly generalizable methods we developed for service generation allow for rapid expansion of our library. Unlike commercial packages for MFP, the framework described here harnesses the power of any computational package, method, and database that has been distributed through a SADI web service – an ever growing list. Thus, the framework we have proposed can potentially surpass the predictive ability of any commercial MFP software currently available. Furthermore, the individual components of our framework can also be reused for workflows not envisioned or pursued here.

On the positive side, a number of barriers preventing integrative research and comprehensive model construction have been struck down with the introduction of reactive semantic web services. Formalization of the specification and annotation transformation rules, identification and characterization of reactions, kinetics, algorithms, and biochemical entities paves the way for a more clearly defined, more reproducible, and more machine-computable science.

On the other hand, one may be wary of the erosion of too many barriers – a web as a biochemical reactor may be an interesting idea, but what happens when the reactor needs virtual walls and partitions to model e.g. human metabolism, as opposed to the entire possible metabolic space of all life? Thankfully, the annotation of web services and their parametric execution can help in such partitioning (Wilkinson *et al.*, 2011). For instance, it would be possible to select a Density Functional Theory method over a semi-empirical method with the modification of a single service parameter. In fact, service parameters could be used to control for multiple factors, including species of origin of a given enzyme, enzyme's subcellular location within a given organelle, pH of the medium, or the precise enzyme sequence and structure to be used. This information would also appear in the result set and influence the simulation (e.g. realization of a biochemical transformation in a given compartment). Using a SADI client such as SHARE, one may also specify the requisite web service parameters and properties by further characterizing the output as well (e.g. if only reactions that take place in cytosol are requested, only cytosolic web service-based enzymes will be used to complete a calculation). The importance of parameters is felt especially acutely when considering the fact that many rule-based systems and statistically-based models for predicting the nature and feasibility of biochemical transformations are geared towards the statistical majority of cases, while it is the minority that may have severe adverse effects to a given compound. Parametric service control enables the examination of the impact of such variations.

In the future, we are planning to expand our collection of web services both, quantitatively and quantitatively. While time-consuming, this task is quite simple: all enzyme services follow the same basic approach to input class definition – that is, automated learning from a collection of small molecules known as substrates for a given enzyme; all rely on a simple reaction definition; all employ standard, modular and reusable routines to estimate reaction kinetics.

With the prototype framework presented here, we have demonstrated that metabolic fate prediction can be undertaken using a first principles-based semantic biochemical reactor. This framework employs CHESS representations in order to automatically classify chemical compounds into semantically defined structural and functional classes, and draws on a collection of self-organizing semantic web services to carry out an array of reactions on the appropriate compounds and to predict their dynamic distribution. With this framework, it has become possible to mechanistically trace the origins of toxicity of small molecules, and to suggest possible routes for small molecule modification to lower this toxicity. Thanks to the readily extensible nature of the framework, it can be further developed in the future to produce increasingly accurate toxicological predictions, and to provide increasingly rich information on the interactions and effects of the produced compounds.

ACKNOWLEDGEMENTS

The authors thank Canada's Advanced Research and Innovation Network (CANARIE) and the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

Chepelev L.L., and Dumontier M. (2011 a). Chemical Entity Semantic Specification: Knowledge representation for efficient semantic cheminformatics and facile data integration. *J Cheminf.*, **3**, 20.

Chepelev L.L., and Dumontier M. (2011 b). Semantic Web integration of Cheminformatics resources with SADI framework. *J Cheminf.*, **3**, 16.

Chepelev L.L., Hastings J., Ennis M., Steinbeck C., Dumontier M. (2012) Self-organizing ontology of biochemically relevant small molecules. *BMC Bioinformatics*, **13**, 3.

Hastings J., Chepelev L.L., Willighagen E., Adams N., Steinbeck C., and Dumontier, M. (2011). The Chemical Information Ontology: Provenance and Disambiguation for Chemical Data on the Biological Semantic Web. *PLoS ONE* **6(10)**: e25513.

Snyder R., and Hedli C.C. (1996). An overview of benzene metabolism. *Environ Health Perspect.*, **104**, 1165-1171.

Wilkinson, M.D., Vandervalk, B., and McCarthy L. (2011) The Semantic Automated Discovery and Integration (SADI) Web service Design-Pattern, API and Reference Implementation. *J Biomed. Semantics*, **2**, 8.