Situational Modeling: Defining Molecular Roles in Biochemical Pathways and Reactions

Michel Dumontier[1,2,3]
[1] Department of Biology, [2]School of Computer Science, [3]Institute of Biochemistry,
Carleton University, 1125 Colonel By Drive,
K1S 5B6, Ottawa, Canada
michel_dumontier@carleton.ca

Abstract. Central to a coherent understanding of cellular biology is a faithful representation of biochemical processes as it pertains to its molecular participants. Current representations underspecify our knowledge because they fail to indicate the roles of the molecular components during relevant processes. Here, we describe a knowledge representation using OWL2 that overcomes previous limitations in modeling biochemical events and has clear implications for the accurate functional/role based annotation of molecular components.

Keywords: semantic web, knowledge representation, ontology, life sciences, OWL-DL, biochemistry.

1 Introduction

Crucial to the success of *in silico* biology is the development of a comprehensive biochemical knowledge base (BKB) capable of answering complex questions about biochemical-related phenomena. To do so, a BKB should exhibit detailed and accurate knowledge representation (KR) of biochemical events such as energy generation or signal transduction while identifying the roles contributions from involved components (from photons to organelles). While numerous biochemical representations have been put forward over the past two decades to deal with the exponentially increasing biological knowledge, these neither share a common conceptualization (ontology) nor adopt a formal representation (syntax and semantics). Importantly, the functions or roles of molecular components are generally underspecified because they are either asserted without reference to the relevant process and hence erroneously appear to occur under any condition, or do not allow the semantic annotation of the parts of a molecule that are critically involved in the process. Hence, lack of granularity and incompatible representational diversity hinders knowledge discovery by increasing the time and effort of data integration, semantic annotation and subsequent data mining.

To address these issues we present an outline for an expressive biochemical knowledge representation in the context of recent additions to the Web Ontology Language (OWL2). This KR is sufficiently developed to capture various aspects of biochemical reactions by focusing on the roles/functions of molecular participants, at various levels of processual detail.

Our example system examines the first reaction in glycolysis in liver cells: the phosphorylation of glucose by the glucokinase enzyme. The reaction involves glucokinase as the catalyst, glucose and magnesium complexed ATP (Mg^{2+}ATP) as the reactants and results in the formation of glucose-6-phosphate (G6P) and Mg^{2+}ADP as the products. During this reaction, the γ phosphate is transferred from ATP to glucose. The reaction can be written as follows:

$$\text{Glc} + Mg^{2+}\text{ATP} \xrightarrow{\text{GK}} \text{G6P} + Mg^{2+}\text{ADP} \qquad \textbf{GKR}$$

Our goal is to represent this reaction with sufficient knowledge to answer several questions:

Q1: In which processes does glucokinase play the role of catalyst?
Q2: Glucose is a substrate in which biochemical reactions?
Q3: In which reactions is a phosphate transferred?
Q4: During which process does glucose form part of an enzyme complex?
Q5: What are the products of GKR?
Q6: What is the role of Mg^{2+}?
Q7: From what molecules is G6P derived from?

2 Notation

Ontological entities are denoted using camel case. Class names start with a capital letter (e.g. Molecule) with boldfaced natural language labels (e.g. **molecule** or **molecules**), properties are italicized and the first letter is lowercase (e.g. *hasPart*). Fully defined classes are underlined (e.g. Enzyme). ***All modeling is at the class level.*** Queries are specified using the Manchester OWL syntax.

3 Biochemical Situational Modeling

Situational models represent a situation (an event, a sequence of events or a collection of events). Situational models consider entities, their qualities, roles and functions, in the context of temporal and spatial locations. Our representation is inspired by the Basic Formal Ontology (BFO) [1, 2], although other upper level ontologies GFO[3], DOLCE[4] have similar philosophies. Common to each is that there exists **continuants**, a class of entities that persist in time (e.g. objects, qualities, spatial regions), and **occurrents**, a class of entities that extend in time (processes, process aggregates, temporal intervals). In turn, continuants may be divided into **independent continuants** (e.g. objects) and **dependent continuants** (e.g. qualities, roles). Real world objects *x* can be associated with numerous **qualities** *y* (e.g. hair color, weight). Although we would like to qualify the values of qualities with time (e.g. partial charge *y* of atom *x* during process *z*), OWL currently only allows the expression of binary relations. Thus, two choices present themselves to describe changing values: 1) a single instance of the quality could be associated with multiple instances of observed/measured values, and the latter are associated with an occurrent; 2) each

value is represented by a different quality instance that is associated with the occurrent. We currently favor the latter approach in our representation, although we note other efforts to develop a common representation (e.g. Ontology of Biomedical Investigation).

Function vs Role: An important aspect of situational modeling involves the contextual realization of functions or roles. The difference between functions and roles is not particularly obvious in molecular systems, and may in fact be redundant. For instance, the function of an enzyme is to catalyze a reaction, or more specifically, to increase the rate of reaction by reducing the activation energy. Every time a protein executes such functionality, it necessarily realizes the enzyme role. Functionality appears intrinsic, while roles are extrinsic and context dependent [5]. Functionality is therefore a kind of *default* description (e.g. that every enzyme has the function of catalyzing a reaction), whether they actually do execute this function or not. Most of our current biochemical knowledge, embodied as functional annotation based on the Gene Ontology, captures this context independent aspect of functionality. This is significantly problematic because molecules may exhibit conflicting functionality that is only executed in different situations. In contrast to views expressed by Arp & Smith [5], we do not believe that roles should be specified or instantiated unless they are coupled with the situations in which they are realized. Roles can encompass context-specific functionality as well as other descriptions in which no functionality is executed (e.g. a molecule can act as a spectator – by simply being in close proximity to the reaction). In this paper, we describe molecular situations using roles (see Figure 1 for examples).

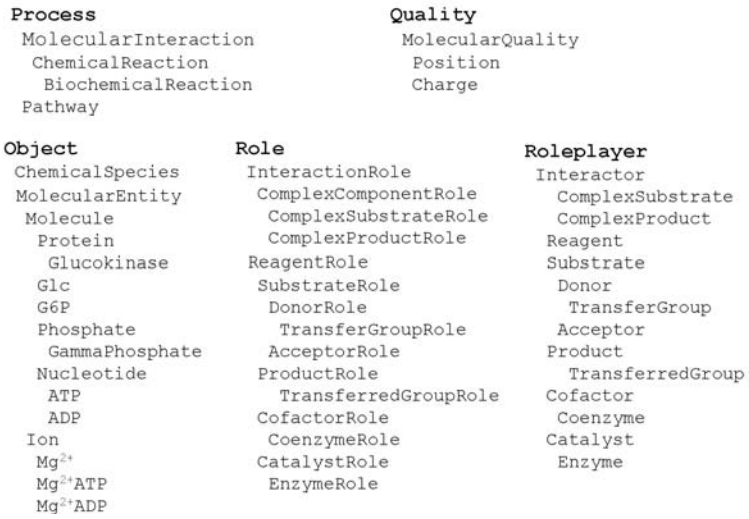

Figure 1 Example ontology containing biomolecules and their qualities, roles and processes. Roleplayers are defined classes for automatic classification based on existential restrictions to roles.

Roles are realizable dependent entities, that is, they are borne by independent continuants and are realized by occurrents. Two basic relations connect entities of these types: *realizes(x,y)*, relating an occurrent *x* to a realizable entity *y*, and *hasBearer(x,y)*, connecting an independent continuant *x* with a realizable entity *y*. Figure 2 shows the relationship between occurrent, realizable entity and independent continuant, and how it applies to modeling the role of GK in GKR. We can now query the OWL KB to ask for reactions that have certain participants, and that these reactions are realizing specific roles, such as the enzyme role (Question 1).

```
Process that realizes some (EnzymeRole that isRoleOf some Glucokinase)
```

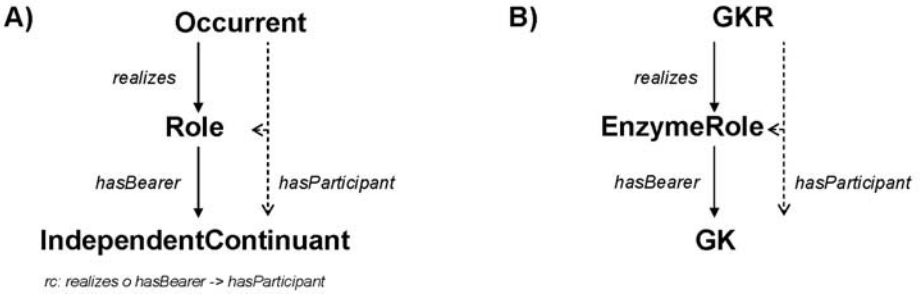

Figure 2 A) Entities such as functions and roles are realized during processes. Realizable entities are participants because *realizes* is a sub-property of *hasParticipant*. Objects are inferred to be participants of the process via a *realizes* ∘ *hasBearer* role chain. B) The enzyme role is realized during glucokinase-catalyzed glucose phosphorylation.

Roleplayers: As part of our natural language description of events, we often talk of a protein *being* an enzyme, rather than playing the role of an enzyme. OWL provides

the means by which one can fully define the necessary and sufficient conditions for class membership. A **roleplayer** is a defined class of entities that must have a relation to a role as part of the necessary and sufficient conditions. The *hasRole(x,y)* predicate defines a relation between an independent continuant and a role, and is a sub-property of *hasBearer*. For instance, we define an **enzyme** as any object that holds at least one instance of the enzyme role. Having a defined class automatically infers membership using an OWL reasoner, and makes possible querying the knowledge base for role holding objects. Thus, we can determine in which reactions is glucose a substrate (Question 2) by asking:

```
BiochemicalReaction that hasParticipant some (Glc and Substrate)
```

3.1 Don't forget roles for parts!

In biochemistry, functionality is often executed by parts of a molecule. Well characterized parts are known as functional groups [6], and these also can have important roles in biochemical events. In our representation, parts may realize roles or functions by participating in biochemical events (Figure 3). For instance, we might like to capture the fact that the gamma phosphate of the ATP molecule is transferred to glucose. Thus, the transfer group role is realized during this process by the phosphate. We can now ask Q3:

```
BiochemicalReaction that realizes some (TransferGroupRole that isRoleOf some
Phosphate)
```

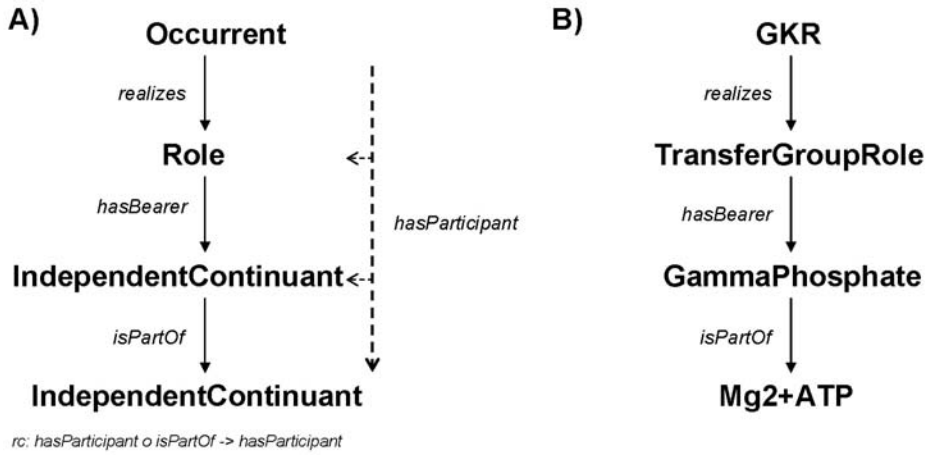

Figure 3 A) Parts of objects can also realize roles during a process. The object whole is inferred to be a participant of the process via the *hasParticipant* ∘ *isPartOf* role chain. B) The gamma phosphate group of Mg^{2+}ATP bears the role of the transfer group during glucokinase-catalyzed glucose phosphorylation.

Usefully, any part of a molecule can be semantically annotated as having a function or role during some biochemical event. Taken together, we represent GKR as a richer

description containing the roles of molecular components in a biochemical reaction (Figure 4).

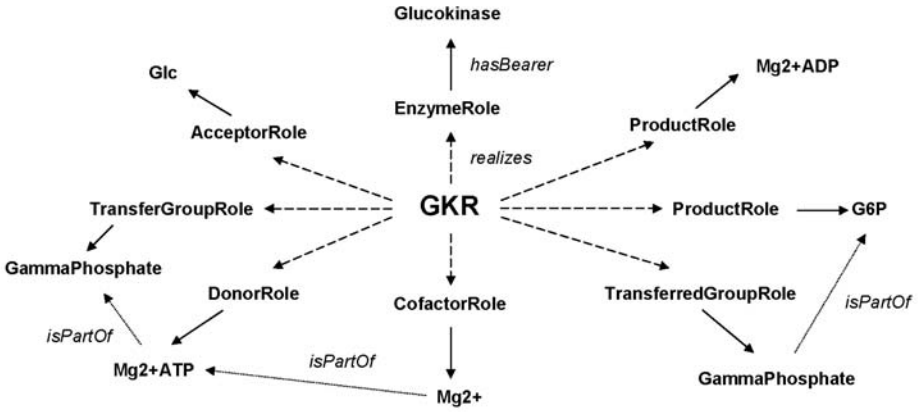

Figure 4 Role-based knowledge representation for the glucokinase-mediated phosphorylation of glucose (GLC) to glucose-6-phosphate (G6P). Various roles are realized (dashed arrow: *realizes*) by reaction participants (solid arrow: *hasBearer*) as the biochemical reaction unfolds. Glucokinase plays the role of enzyme by lowering the activation energy of the reaction, in presence of the double charged magnesium ion (Mg^{2+}) co-factor. As a donor, Mg^{2+}ATP transfers its gamma phosphate to the GLC acceptor which results in the formation of products Mg^{2+}ADP and G6P (contains the transferred phosphate).

At this point, we identify a critical weakness of OWL in that it cannot easily represent cyclic class expressions. For instance, we would like to represent that the cofactor role played by Mg^{2+} is also part of the Mg^{2+}ATP complex that plays the donor role, where both roles are realized in GKR (Figure 4). The resulting (partial) class expression fails to capture this dependency:

```
GKR ::= Reaction
 and realizes exactly 1 (DonorRole that isRoleOf (Mg2+ATP that hasProperPart
       exactly 1 Mg2+ that hasRole some (CofactorRole that isRealizedIn some
        GKR)))
 and realizes exactly 1 (CofactorRole that isRoleOf (Mg2+ that isProperPartOf
       exactly 1 (Mg2+ATP that hasRole some (DonorRole that isRealizedIn some
        GKR)))
```

3.2 Event Decomposition

The breakdown of a complex process into simpler events is important in biochemistry. For instance, the progress of a biochemical reaction can be described by changes in substrate structure through one or more transition states to finally form the products. In our knowledge representation (Figure 5), we ensure that knowledge captured at these finer granular processual parts still relate to the process whole. This is accomplished to a large part by invoking a *hasPart* ∘ *hasParticipant* -> *hasParticipant* role chain.

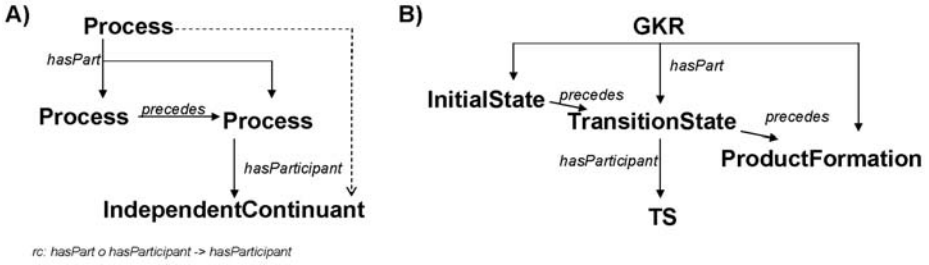

Figure 5 A) Participants of sub-processes are also participants of process wholes using the *hasPart* ◦ *hasParticipant* role chain. B) GKR can be broken down into a number of steps (intial state, transition state, product formation) to indicate the progression of the chemical reaction.

The decomposition of the reaction mechanism is equally important and is enabled by this representation. Thus, we can transform XML-based approaches [7] with a more expressive OWL representation. However, as described above in section 3.1, the representation requires a structured object rather than tree-like class expression.

3.3 Chemical Persistence and Transformation

Dependent continuants such as qualities, functions and roles act as pivots between objects and processes, and our knowledge representation ensures that objects persist with a single identity throughout their lifetime. That is to say, there is no need to create another distinct instance of the same object in so as to place it in a particular spatial-temporal context with certain attributes. Much debate in online forums questions whether the slightest chemical modification leads to creation of an entirely distinct entity, or whether it is the same entity with some attribute. However, a fundamental aspect of chemistry is that identity is intrinsically linked to chemical structure. As such, changes to structure lead to changes in identity.

A biochemical reaction results in the conversion of at least one object into at least one other different object, represented using the *derivesFrom* predicate (Figure 6). In OWL2, we can specify that the same instance cannot derive from itself with the irreflexive characteristic. We can also specify which molecules can be derived from by applying a universal restriction on *derivesFrom*.

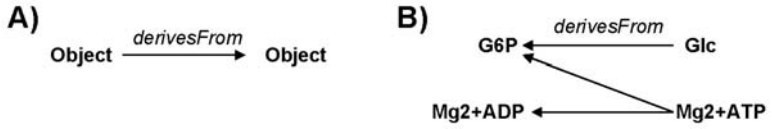

Figure 6 A) Derivation is the transformation of objects into new fundamental entities. B) The fate of chemically modified biochemical species can be captured as a result of chemical transformations.

The formation of G6P from Glc occurs by preferential binding of glucose followed by Mg^{2+}ATP [8]. Two representations for this information are shown in Figure 7. The

first representation uses *derivesFrom* to indicate that the complex is formed from components. The second representation associates roles of the molecules before the formation of the complex, and after.

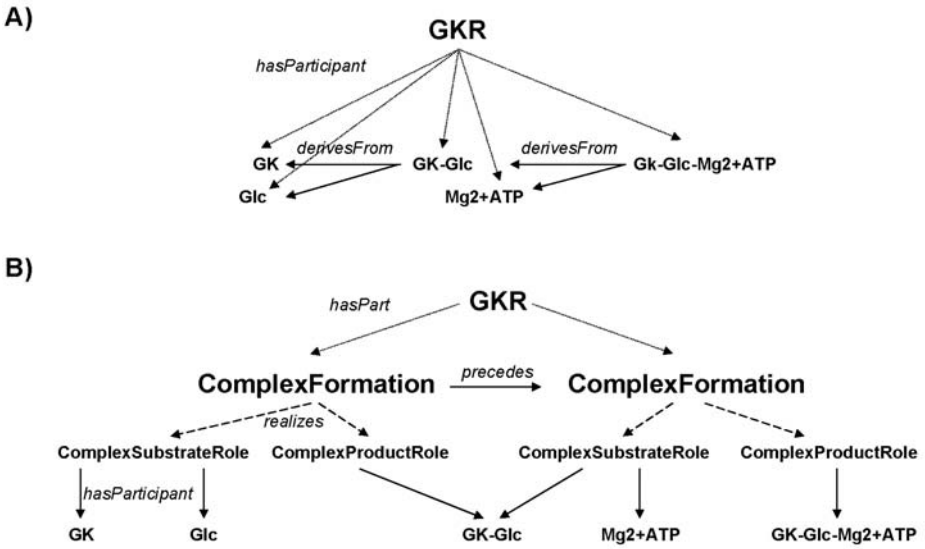

Figure 7 Two representations for the formation of molecular complexes involved in the kinetic mechanism of Glucokinase-catalyzed glucose phosphorylation. A) The use of a predicate to relate entities and B) the use of processual classes and roles to indicate complex substrates and resulting complex products.

While the first provides a temporal progression of species via a predicate, the second explicitly details the roles of each component at every part of the complex formation. Therefore, it becomes possible to query the knowledge base with respect to the role of the participant, such that it becomes possible to find reactions where glucose is a component in complex formation. We can now during which reaction does glucose form part of an enzyme complex (Q4):

```
BiochemicalReaction that hasPart some (ComplexFormation that hasParticipant some
Glucose)
```

4 Discussion

4.1 What about whether some do?

This class-based representation aims to capture the molecular behavior by assigning roles during biochemical reactions. While we can ask the knowledge base about any kind of biochemical reaction, we cannot ask about the roles or participants directly

(Q5-7). That is to say, we would like to learn what we know about a particular concept – how it is used. In a sense, we would like to ask "are there *some* glucose that are substrates?", rather than "are *all* glucose substrates?". We have noticed, however, that a knowledge base could determine how objects related to the processes, and be able to answer questions about "some" objects or roles. For instance, from Figure 4, we know that some Mg^{2+} bear a co-factor role that is realized in the reaction. Protégé 4 does something akin to this with its "class usage" tab. Thus, this approach could serve as a portal to querying circumstantial knowledge.

4.2 Comparison with Existing Approaches

Most conceptualization and representation of biochemical knowledge has been as the result of representing knowledge in relational databases. Enzyme [9], and later IntEnz [10], describe enzymatic reactions (as a string) primarily based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) for which an Enzyme Commission (EC) number has been assigned. BRENDA [11] is a comprehensive resource on biochemical reactions and enzyme kinetics for which publication references are given. English descriptions of mechanistic details along with substrates and corresponding products, regulation, co-factors, activators, inhibitors, kinetic parameters (km, kcat) under varying conditions (pH, organism), effective temperature range, tissue/cell distribution, subcellular localization, complex, and roles in disease. The data is available under a highly restrictive license. MACIE [12] stepwise describes enzyme mechanisms in natural language for a wide variety of reactions, and also identifies over 15 molecular roles. BioCyc use a frame-based representation [13] which links reactions to enzyme-catalyzed reactions in a relational manner, rather than that of subsumption. Further, the two sides of the reaction are conceptualized as "left" and "right" so as to avoid the directionality implied by using "reactants" and "products", as many biochemical reaction are reversible. While this is true for mass action kinetics (as opposed to micro-scale particle dynamics), the thermodynamic feasibility (directionality) of a reaction is captured by the change in Gibbs free energy of the system (defined by equilibrium between substrates and products) under standard conditions. Thus, these are not representations of chemical reactions *per se*, but rather the end concentrations of substrates and products from collections of billions of chemical reactions occurring in both the forward and reverse directions.

BioPAX is an OWL-based knowledge representation for biochemical reactions and pathways developed by a consortium of pathway and interaction databases as well as interested parties. The development of BioPAX was largely influenced by BioCyc which is reflected in the data model and property names (e.g. LEFT is the name of the object property that links an object at the beginning of the biochemical event). Since enzymes modulate processes, and roles are indicated by predicates "CONTROLLER" and "CONTROLLED", respectively, this representation is generally incompatible with upper level ontologies. BioPAX also fails to capitalize on consistent URI naming as a means to integrate data, and does not associate related knowledge in a way that can be reasoned about (imports of controlled vocabularies are only that). Recent

demonstration of the utility of BioPAX data [14] was largely limited by an initial syntactic matching of contents.

A a simple representation of a biochemical reaction in OWL was put forward as an n-ary design pattern [15]. Role-based representation is achieved by use of special predicates (e.g. has_substrate or has_product). This approach leads to a proliferation of predicates, one for each role, and whose expressivity is limited to available OWL property characteristics (e.g. transitive, reflexive, irreflexive, functional, inverse functional, anti-symmetric, disjunction). Clearly, this approach cannot be combined in such a way to take advantage of OWL's class constructors (e.g. union, intersection, negation, cardinality, existential and universal restrictions). Hence, the creation of sophisticated expressions (e.g. a substrate role, but not an acceptor role) cannot be realized using predicate expressivity alone.

5 Additional OWL Requirements

5.1 Need for Structured Objects/Description Graphs

Our knowledge representation could benefit from the incorporation of structured objects (aka description graphs) [16] into OWL. For instance, class-based representation of the biochemical reaction in Figure 4 and the sequential complex formation in Figure 7B makes references to objects that are distantly linked in the tree structure, and are better represented as a structured object. We have also previously made the case for description graphs in the representation of cyclic molecules [17, 18], which cannot currently be done at the class level with OWL.

5.2 Nonstructural restrictions

We find that the benefits of role chains are challenged by the drawbacks of nonstructural restrictions on properties, as they can no longer be used to define cardinality restrictions. While we've managed to overcome such problems by restructuring our representation, it would be infinitely more useful to have a better explanation of the inconsistencies by OWL reasoners (FaCT++, Pellet).

6 Conclusion

We have presented a rich knowledge representation for biochemical events compatible with upper level ontology. We use recent additions to the OWL language to infer relations and facilitate knowledge discovery. We anticipate that the instantiation of this representation with existing biochemical databases will create new opportunities for data integration and knowledge discovery.

Acknowledgments: We would like to thank members of BioPAX-OBO, particularly Alan Ruttenberg, Oliver Ruebenacker, Andrea Splendiani for valuable discussions during our working group sessions. We thank our anonymous reviewers for raising issues that have certainly improved the quality of this manuscript. This work was supported in part by an NSERC Discovery Grant.

7 References

1. Grenon, P., Smith, B., Goldberg, L.: Biodynamic ontology: applying BFO in the biomedical domain. Stud Health Technol Inform **102** (2004) 20-38
2. Grenon, P.: Temporal Qualification and Change with First–Order Binary Predicates. In International Conference on Formal Ontology in Information Systems (FOIS 2006), Baltimore, Maryland (USA) (2006)
3. Herre, H., Heller, B., Burek, P., Hoehndorf, R., Loebe, F., Michalek, H.: General Formal Ontology (GFO) – A Foundational Ontology Integrating Objects and Processes. Onto-Med Report 8, Research Group Ontologies in Medicine, Institute of Medical Informatics, Statistics and Epidemiology. University of Leipzig, Leipzig, Germany (2006)
4. Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A., Schneider, L.: WonderWeb Deliverable D17. The WonderWeb Library of Foundational Ontologies and the DOLCE ontology.: (2002) Preliminary Report (ver. 2.0, 15-08-2002)
5. Arp, R., Smith, B.: Function, Role, and Disposition in Basic Formal Ontology. Nature Precedings (2008)
6. Villanueva-Rosales, N., Dumontier, M.: Describing chemical functional groups in OWL-DL for the classification of chemical compounds. OWL Experiences and Design, Innsbruck, Austria. (2007)
7. Sankar, P., Aghila, G.: Ontology aided modeling of organic reaction mechanisms with flexible and fragment based XML markup procedures. J Chem Inf Model **47** (2007) 1747-1762
8. Ning, J., Purich, D.L., Fromm, H.J.: Studies on the kinetic mechanism and allosteric nature of bovine brain hexokinase. J Biol Chem **244** (1969) 3840-3846
9. Bairoch, A.: The ENZYME data bank in 1999. Nucleic Acids Res **27** (1999) 310-311
10. Fleischmann, A., Darsow, M., Degtyarenko, K., Fleischmann, W., Boyce, S., Axelsen, K.B., Bairoch, A., Schomburg, D., Tipton, K.F., Apweiler, R.: IntEnz, the integrated relational enzyme database. Nucleic Acids Res **32** (2004) D434-437
11. Barthelmes, J., Ebeling, C., Chang, A., Schomburg, I., Schomburg, D.: BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. Nucleic Acids Res **35** (2007) D511-514
12. Holliday, G.L., Almonacid, D.E., Bartlett, G.J., O'Boyle, N.M., Torrance, J.W., Murray-Rust, P., Mitchell, J.B., Thornton, J.M.: MACiE (Mechanism, Annotation and Classification in Enzymes): novel tools for searching catalytic mechanisms. Nucleic Acids Res **35** (2007) D515-520
13. Karp, P.D., Riley, M.: Representations of metabolic knowledge. Proc Int Conf Intell Syst Mol Biol **1** (1993) 207-215
14. Luciano, J.S., Stevens, R.D.: e-Science and biological pathway semantics. BMC Bioinformatics **8 Suppl 3** (2007) S3
15. Stevens, R., Aranguren, M.E.n., Wolstencroft, K., Sattler, U., Drummond, N., Horridge, M., Rector, A.: Using OWL to Model Biological Knowledge. (2006)

16. Motik, B., Grau, B.C., Sattler, U.: Structured Objects in OWL: Representation and Reasoning. 17th Int. World Wide Web Conference (WWW 2008). ACM Press, Beijing, China (2008) 169-182
17. Dumontier, M., Villanueva-Rosales, N.: Modeling Life Science Knowledge with OWL 1.1. OWL Experiences and Design, Washington D.C. (2008)
18. Konyk, M., De Leon, A., Dumontier, M.: Chemical Knowledge for the Semantic Web. In: Bairoch, A., Boulakia, S.C. and Froidevaux, C. (ed.): DILS, Vol. 5109. Springer, Evry, France (2008) 169-176