Pacific Symposium on Biocomputing 15:465-476(2010)

SYNTHESIS OF PHARMACOKINETIC PATHWAYS THROUGH KNOWLEDGE ACQUISITION AND AUTOMATED REASONING

LUIS TARI, SAADAT ANWAR, SHANSHAN LIANG, JÖRG HAKENBERG, CHITTA BARAL

Department of Computer Science and Engineering,
Arizona State University, Tempe, AZ 85287, USA

Biological pathways are seen as highly critical in our understanding of the mechanism of biological functions. To collect information about pathways, manual curation has been the most popular method. However, pathway annotation is regarded as heavily time-consuming, as it requires expert curators to identify and collect information from different sources. Even with the pieces of biological facts and interactions collected from various sources, curators have to apply their biological knowledge to arrange the acquired interactions in such a way that together they perform a common biological function as a pathway. In this paper, we propose a novel approach for automated pathway synthesis that acquires facts from hand-curated knowledge bases. To comprehend the incompleteness of the knowledge bases, our approach also obtains facts through automated extraction from Medline abstracts. An essential component of our approach is to apply logical reasoning to the acquired facts based on the biological knowledge about pathways. By representing such biological knowledge, the reasoning component is capable of assigning ordering to the acquired facts and interactions that is necessary for pathway synthesis. We demonstrate the feasibility of our approach with the development of a system that synthesizes pharmacokinetic pathways. We evaluate our approach by reconstructing the existing pharmacokinetic pathways available in PharmGKB. Our results show that not only that our approach is capable of synthesizing these pathways but also uncovering information that is not available in the manually annotated pathways.

1. Introduction

Developing systems and algorithms to assist and guide biological researchers in reverse engineering and synthesis of biomolecular systems has been a long term goal for the fields of computational biology and bioinformatics. An important task of modeling biomolecular systems is the building of pathways for various biological processes. In building pathways for processes such as pharmacokinetics, the ability to collect and integrate information contained in existing databases and the literature is important so that partial pathways can be built. With the partial pathways, a researcher can identify what gap in knowledge needs to be filled.

Several knowledge bases have been created for the need of pathway information, focusing on different aspects of networks and pathways. Reactome [1], KEGG [2] and HumanCyc [3] are examples of knowledge bases for metabolic pathways, while Biocarta[1] and Panther [4] consist of knowledge bases for signaling pathways. PharmGKB [5] is a knowledge base for drug-oriented genomic information that includes pharmacokinetic and pharmacodynamic pathways. These knowledge bases rely on manual curation by experts and the pathway information is very precise but far from being complete due to the intensive process required in the annotation of pathways. It is therefore necessary to investigate another paradigm for the curation of pathways to speed up the annotation process.

In this paper, we propose a novel approach for the automated synthesis of pharmacokinetic pathways by first acquiring the necessary pharmacokinetic facts and interactions of the target drug from existing curated knowledge bases. As the curation effort of these knowledge bases is yet to be completed, our approach includes an automated text extraction component that extracts facts and interactions from Medline abstracts. The inclusion of knowledge extracted from Medline abstracts can lead to the synthesis of more comprehensive pathways, as compared to using only the curated knowledge bases for building pathways. In the synthesis of pathways, it is essential to indicate the ordering of the interactions in a pathway, as a pathway is a series of interactions that are triggered by one another, in which the consequences of the interactions include the activation of certain biological functions or generation of products such as metabolites as a result of drug metabolism. To assign the ordering of the interactions, our approach includes the logical representation of the general properties and behavior of pharmacokinetic pathways. Automated reasoning can then be applied to the acquired knowledge so that ordering of the interactions can be assigned to synthesize pathways. The inclusion of such reasoning capabilities on top of mining relevant knowledge distinguishes

[1] Biocarta – http://www.biocarta.com

our approach from typical data-driven and knowledge-driven approaches in pathway curation. Data can be collected by means of large-scale protein-protein interaction networks from experimental sources such as two-hybrid screening (Y2H) [6]. On the other hand, knowledge can be extracted by automated text extraction techniques to produce large-scale interaction networks and pathways [7-10]. While these networks of interactions provide insights to biological discovery, it is not always straightforward to identify which interactions are indeed biologically related within a large network of interactions. To determine which parts of the interaction networks indeed correspond to pathways, pathway curators rely on visualization and pathway building tools to synthesize pathways [11]. Pathway building tools such as CellDesigner [12] or proprietary tools such as Ingenuity IPA[2] and ActiveMotif[3] based on manually curated databases allow curators to visualize and assemble pathways from an interaction network. Such methodology still heavily depends on the biological knowledge of the expert curators for pathways synthesis. An alternative is the use of graph-theoretic methods on interaction networks (see [13] for a survey) or network alignment methods over interaction networks of multiple species to uncover conserved modules as pathways [14, 15].

It becomes apparent that fully-automated systems for pathway synthesis are required to have the capabilities of acquiring various kinds of information from multiple sources, as well as assigning appropriate order to the acquired interactions. To automatically arrange the interactions for pathway synthesis, our work includes a reasoning component for this purpose. With proper representation of pathways for a particular biological process such as pharmacokinetics, reasoning can be applied to the acquired knowledge and automatically assigned the appropriate ordering of the interactions to synthesize pharmacokinetic pathways. The approach of utilizing biological domain knowledge has been applied to various applications, such as the generation of metabolic networks based on stoichiometric constraints [16] and hypothesis generation in signaling pathways [17]. The need of formulating biological domain knowledge and applying reasoning to infer pathways from text is highlighted as new challenges in [18]. In pharmacokinetic pathways, an example of a biological property is that drug metabolites are generated as a result of drug metabolism. In other words, a metabolized drug is a precondition for the generation of drug metabolites to occur, and the effect of the interaction is the production of drug metabolites. By encoding the logic representation in the form of pre- and post-conditions of pharmacokinetic properties that describe the course of drug disposition in the body, which includes drug absorption, distribution, metabolism and excretion, reasoning can be applied to the interactions in order to find a sequence that satisfies the pharmacokinetic properties. Finding a sequence of actions is known as *planning* in the field of artificial intelligence, and planning is considered as one kind of *reasoning*. More specifically, planning can be described as given the initial states and the goals of the problem, find a sequence of actions such that the goals can be achieved from the initial states. In the case of the pharmacokinetics effects of drugs, the initial state is when a drug is administered and the goal state is when the drug is eliminated after being metabolized. The expected plan is the actions required for the administered drug to be delivered to the systemic circulation for drug consumption. With the acquired facts and interactions and the biological properties of pharmacokinetics in logical representation, the reasoning component of our proposed system arranges the interactions so that a series of interactions is generated as a pathway model.

The rest of the paper is outlined as follows. We describe the basic properties of pharmacokinetics that we encode in our system in Section 2. In Section 3, the processes of acquiring the necessary facts and interactions from existing knowledge bases and Medline abstracts are described. In addition, the reasoning component is illustrated in how the pharmacokinetic properties and behavior are encoded in order to synthesize pathways. In Section 4, we demonstrate the feasibility of our approach by illustrating the precision and recall of generated pathways. We concluded in Section 5.

2. Pharmacokinetics

Pharmacokinetics is concerned with the relationships between various processes during the course of the drug consumption in the body. The study of pharmacokinetics is important to biologists and drug designers, as the bioavailability of a drug, i.e. the effectiveness of a drug when it is absorbed into the systemic circulation, is heavily

[2] Ingenuity Pathway Analysis Tool: http://www.ingenuity.com
[3] Active Motif: http://www.activemotif.com

dependent on the processes involved in pharmacokinetics. When a drug is taken orally, the drug is absorbed in the intestine, and the corresponding drug transporters distribute the drug to the appropriate cellular locations of the intestinal cells. The drug is then delivered to the liver through the bloodstream, and the relevant drug transporters in the liver cells distribute the drug for metabolism by the enzymes. Drugs that are taken intravenously would bypass the drug absorption phase. The pharmacokinetics of a drug includes several processes such as the *distribution* of a drug through different tissues, the *metabolism* of a drug, the *excretion* of a drug, and the *absorption* of a drug into the systemic circulation [19]. Several essential elements are involved in different processes of pharmacokinetics, namely *drug transporters*, *enzymes* and *metabolites*. The typical processes involved in pharmacokinetic pathways are shown in Figure 1. Drug transporters are responsible for drug *distribution* for absorption (as in Region B_1 in Figure 1 (left)), metabolism (Region B_2 in Figure 1 (left)) and excretion (Region B_3 in Figure 1 (left)), and they can be expressed in many tissues such as intestine and liver [20]. Once the target drug is distributed into an appropriate cellular location, the enzymes play the role of metabolizing the drug (as in Region A in Figure 1 (left)), which take place mainly in the liver. Metabolites are produced as a result of the metabolism of the drug, shown in Region C.

Identifying the pharmacokinetic mechanism of a drug is essential in avoiding potential side effects of drug-drug interactions, even though in most cases the processes of drug disposition for co-administered drugs typically do not affect one another. However, drugs that are strong inducers or inhibitors of certain enzymes can influence the bioavailability of drugs that are metabolized by these enzymes [21]. The drug ketoconazole is an example of a powerful inhibitor that is known to inhibit CYP3A enzymes, which are responsible for the metabolism of a wide variety of drugs, such as midazolam. Such inhibition of CYP3A enzymes can affect the drug-metabolizing activity and lead to the increase of the bioavailability CYP3A substrates. On the other hand, drugs that are potent inducers of CYP3A enzymes, such as carbamazepine, can cause a reduction of the effect of CYP3A substrates. With the increasing availability of clinical drugs that are inducers or inhibitors of enzymes, the study of pharmacokinetic drug interactions becomes more critical [21]. While a drug can be involved in various parts of the body, our focus in the synthesis of pharmacokinetic pathways is on the processes involving drug absorption, distribution, metabolism and elimination in the intestine and liver.

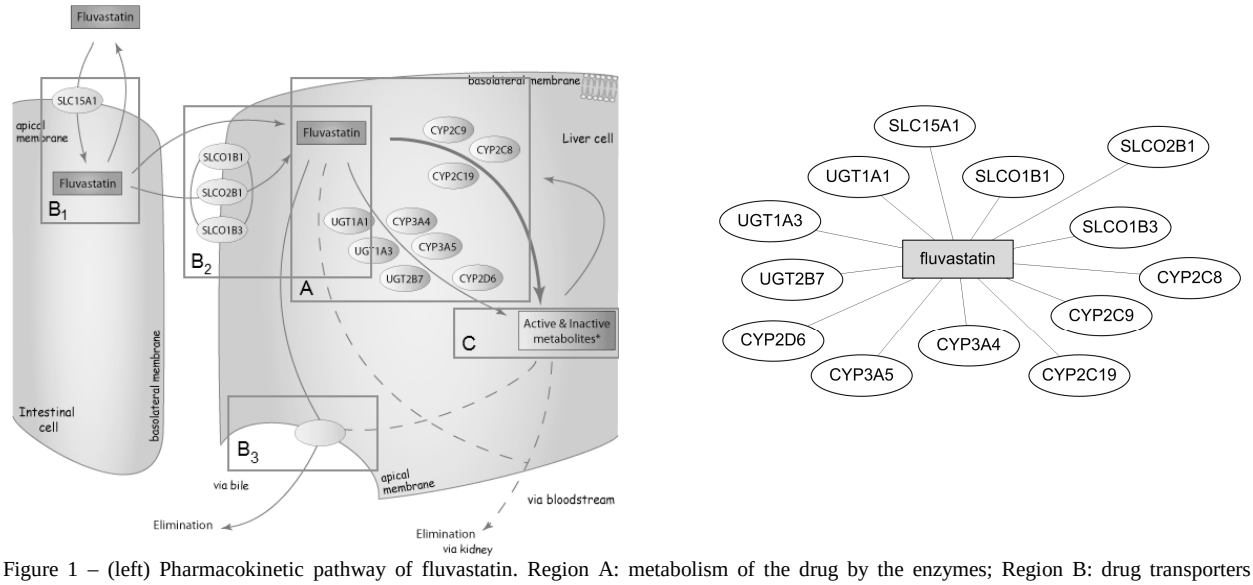

Figure 1 – (left) Pharmacokinetic pathway of fluvastatin. Region A: metabolism of the drug by the enzymes; Region B: drug transporters distribute the drug for absorption in intestine in B_1, for metabolism in B_2 and for elimination in B_3; Region C: the drug is metabolized to metabolites by the enzymes. (Diagram source: PharmGKB); (right) A network representation of the drug-protein interactions for fluvastatin.

3. Methods

The goal of our system is to synthesize the pharmacokinetic pathway of a given drug. Our approach in constructing pathways can be described as a two-stage approach: (i) *fact and interaction extraction* from knowledge bases and text; (ii) inferences of pathways through *reasoning* with the extracted facts and interactions based on the biological

knowledge of pharmacokinetic pathways as described in Section 2. Figure 2 illustrates how the facts and interactions extracted in step (i) are utilized together with the biological background knowledge to construct pathways.

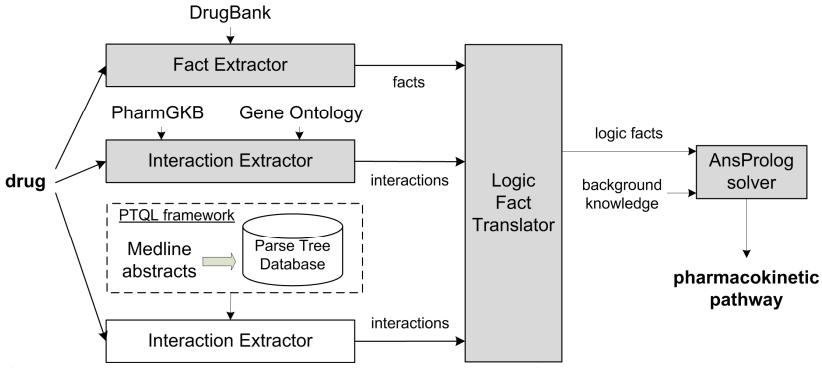

Figure 2 - An overview of the system architecture. Facts and interactions are acquired from knowledge bases such as Drugbank, PharmGKB and Gene Ontology annotations, as well as our PTQL framework for text extraction from Medline abstracts. The shaded components correspond to the novel features of this paper. The facts and interactions are translated into logic facts so that together with the background knowledge, the logic program solver (AnsProlog solver) assigns ordering to the interactions for the synthesis of pharmacokinetic pathways.

The first stage involves various kinds of fact extraction, such as identifying which proteins are drug transporters, as well as interaction extraction, such as finding which enzymes play a role in metabolizing the drug of interest. Fetching the facts and interactions alone leads to the formation of a network of interactions, but the resulting network lacks the information that describes which of the interactions appear ahead of the others. Using the gene-drug interaction network in Figure 1 (right) as an example, the outcome of the extraction process is the interactions between the drug fluvastatin and the proteins, such as SLCO1B1 and CYP3A4. However, with the extracted interactions alone, it is unclear whether the interaction between SLCO1B1 and fluvastatin should precede or follow the interaction between CYP3A4 and fluvastatin. In the synthesis of pathways, ordering of the interactions is essential, as a pathway is a series of interactions, in which the consequence of an interaction affects the subsequent interactions. For instance, the CYP and UGT enzymes in Figure 1 (left) are able to metabolize fluvastatin in the liver cells only if the drug is distributed by the drug transporters SLCO1B1, SLCO2B1 and SLCO1B3. In other words, distribution of the drug in the liver cells is a prerequisite for the drug metabolizing interaction to take place. In our approach, such kind of preconditions and postconditions of interactions are encoded as logic rules so that the reasoning component in step 2 assigns an ordering to the interactions extracted in step 1. We describe the details of each of the steps in the rest of the section.

3.1. *Fact and interaction extraction from knowledge bases*

The first stage of pathway synthesis is to identify and recognize the facts involved in the pharmacokinetic pathways of the drug of interest. As the pharmacokinetics of drug behaves differently depending on its dosage form, it is essential to obtain such information in order to synthesize pathways correctly. Drug metabolism can take place in different organs, and a drug that is known to be metabolized in the liver cells would have a different pathway from a drug that is metabolized in other cells such as the intestinal cells. DrugBank [22] is a rich resource to obtain such kind of facts about drugs. The field “dosage form” in DrugBank provides information such as whether a drug is taken orally or intravenously, and the metabolism of a drug can be obtained from the field “biotransformation”. With DrugBank, logic facts in the form of `is_taken(Drug, Method)` and `metabolism(Drug, Organ)` are generated, in which `Drug` is the name of drug of interest, `Method` is either `orally` or `intravenously` and `Organ` can be `liver` or `intestine`. Using the drug fluvastatin as an example, the facts are written as logic facts `is_taken(fluvastatin, orally)` and `metabolism(fluvastatin, liver)`.

To construct pharmacokinetic pathways, it is important to identify the interactions between drug and enzymes. Several resources are used to obtain interactions between drugs and enzymes. The DrugBank knowledge base provides the metabolizing enzymes as well, but the list of enzymes for each drug is not comprehensive as only the main enzymes are included. Other than DrugBank, PharmGKB [5] is another rich resource that provides information

about genes, drugs and diseases. An extensive list of interactions between drugs and proteins can be found in PharmGKB relationships. The interactions are categorized into several types namely "pharmacokinetics" (PK), "pharmacodynamics" (PD), "molecular and cellular functional assays" (FA) and "clinical outcome" (CO). For the purpose of the pharmacokinetic pathway synthesis, the interactions labeled as pharmacokinetics ("PK") are utilized. However, obtaining the pharmacokinetic interactions is not sufficient for pathway synthesis, as it is important to realize whether the proteins involved in the interactions are enzymes or transporters. Such kind of information can be inferred from the Gene Ontology (GO) annotations [23].

GO is a hierarchy of controlled vocabulary that includes three independent ontologies for biological process, molecular function and cellular component. Standardized terms in GO describe roles of genes and gene products in any organism. Curators annotate the functions of proteins by assigning GO terms, and such annotation is known as GO annotations. The terms "metabolic process" (GO:0008152) and "transporter activity" (GO:0005215) from the GO "biological process" and "molecular function" sub-ontologies can be utilized to identify enzymes and proteins. Given a drug-protein interaction, a protein is considered as an enzyme if the protein is annotated under the subcategories of the term "metabolic process" according to the GO annotation. Similarly, a protein annotated in one of the subcategories of the term "transporter activity" is regarded as a transporter. Since the protein is obtained from a drug-protein interaction, the transporter is regarded as a drug transporter. With PharmGKB and GO, logic facts in the form of `enzyme(Protein)`, `metabolizes(Enzyme, Drug)`, `transporter(Protein)` and `distributes(Transporter, Drug)`are generated. Using the drug fluvastatin as an example, the facts and interactions are written into these logic facts: `enzyme(cyp3a4)`, `metabolizes(cyp3a4, fluvastatin)`, `transporter(slco1b1)` and `distributes(slco1b1, fluvastatin)`.

3.2. *Automated text extraction of facts and interactions*

While the sources provide an extensive amount of information for the synthesis of pharmacokinetic pathways, much of the information is resided in the biomedical literature. In particular, drug transporters, drug-enzyme metabolic relations and the metabolites produced as a result of drug-enzyme metabolic relations can be found in Medline abstracts. Extraction of such facts and interactions requires the utilization of syntactic and lexical clues from sentences. One way of extracting metabolic relations between the target drug and enzymes from text can be performed based on cooccurrences of drug names, gene/protein names and the word "metabolized". However, using cooccurrences is not sufficient for the precision needed in pathway synthesis. Suppose the target drug for extraction is fluvastatin, the relations `metabolizes(CYP2C9, fluvastatin)` and `metabolizes(CYP3A4, fluvastatin)` are extracted from the sentence "*Fluvastatin is metabolized by CYP2C9, while simvastatin, lovastatin and atorvastatin are metabolized by cytochrome P450 3A4 (CYP3A4).*" (PMID:16714062) based on cooccurrences, in which the relation `metabolizes(CYP3A4, fluvastatin)` is an incorrect relation according to the sentence. By utilizing the syntactic pattern that a drug-enzyme metabolic relation is feasible if the word "metabolized" and the gene/protein mention appear in the same verb phrase, then only the correct relation `metabolizes(CYP2C9, fluvastatin)` can be extracted from the sentence. This example shows that it is important to extract drug-enzyme metabolic relations with the use of syntactic patterns. With the diverse extraction needs in the synthesis of pharmacokinetic pathways, it is not feasible to develop individual extraction systems for each specific extraction goal. This implies the need of a flexible system that is designed for generic extraction.

Our PTQL framework [24, 25] provides the flexibility to perform such kind of diverse extraction. A sample list of logic facts that are generated through PTQL extraction is described in Table 1. The central piece of our extraction framework is the parse tree database, which is composed of the syntactic structures for each of the sentences in the entire collection of Medline abstracts. Each of the Medline abstracts is represented as a hierarchical representation with the syntactic and semantic information, which includes the recognition of biological entities. BANNER [26] and MetaMap [27] were used for gene/protein mentions and drug names, while gene/protein mentions are normalized by GNAT [28] for their standardized form in order to avoid name ambiguity. By storing the syntactic and semantic information in the database, performing extraction becomes a matter of writing extraction queries. Since standard relational database queries such as SQL are not ideal for expressing queries that involve linguistic patterns, we

developed a query language called *parse tree query language* (*PTQL*) that are used to express linguistic patterns for extraction. While PTQL queries are expressive, writing the queries manually for extraction is time-consuming and potentially limits the recall of the system. As an alternative, our PTQL framework provides a component that is capable of generating PTQL queries from keyword-based queries. For instance, in the extraction of drug-enzyme metabolic relations, we simply issue the keyword-based query

```
<type:DRUG> and <class:metabolism> and <type:PROTEIN>
```

where `<type:DRUG>` and `<type:PROTEIN>` correspond to matching any mentions of drug name and gene/protein names, and `<class:metabolism>` corresponds to lexical variants of the word "metabolism", which includes "metabolize", "metabolized" and "metabolizes". Using a small corpus of Medline abstracts, the component first retrieves sentences that are relevant to the keyword-based query as in a typical search engine. By utilizing the parse tree database, grammatically similar sentences are retrieved and their common grammatical patterns are utilized in forming PTQL queries automatically. The resulting PTQL queries include the necessary syntactic patterns and they are applied to the full parse tree database that includes all Medline abstracts for extraction. This component is used for the extraction of (i) drug-enzyme metabolic relations; (ii) proteins responsible for drug elimination; (iii) protein expression in liver and intestinal cells. Extraction of these kinds of relations is essential in the synthesis of pharmacokinetic pathways, as existing sources such as DrugBank and PharmGKB lack such information or in a format that is not easily readable by computers.

• *Protein expression in liver and intestinal cells.* In the synthesis of pharmacokinetic pathways, it is essential to find out whether a protein is expressed in the liver or intestinal cells. The following keyword-based query is used for the extraction of protein expression in liver cells:

```
<type:PROTEIN> and <class:liver>
```

The keywords "hepatic" and "liver" are used to represent `<class:liver>`. Similarly, `<class:intestine>` is used for the extraction of protein expression in intestinal cells, in which "intestinal", "gastrointestinal" are used for `<class:intestine>`. With the keyword-based queries, the generated PTQL queries retrieve the sentences and relations as shown in Table 1.

• *Proteins responsible for drug elimination.* Among the interactions between the target drug and its drug transporters, it is necessary to find out the roles of each of the drug transporters, as drug transporters are known to be involved in various roles such as drug distribution, absorption and elimination. Finding the exact roles of transporter is essential for the assignment of the ordering of the interactions. For instance, if a drug transporter is responsible for drug elimination, we know that the drug transporter cannot distribute the drug until the drug is metabolized. Using the keyword-based query

```
<type:PROTEIN> and <class:elimination> and <type:DRUG>
```

where `<class:elimination>` is represented by the keywords "elimination" and "excretion", an example of sentences that indicate the role of drug elimination for a drug transporter is shown in Table 1.

• *Drug-metabolites relations.* For the extraction of relations, typically named entity recognizers are applied to recognize the entities involved in the target relations. However, in the case of metabolites, the lack of dictionaries or training data means that we need to rely on lexical hints for the identification of metabolites. For example, the sentence "The antioxidative effects of the metabolites of fluvastatin (M2, M3, M4 and M7) …" indicates that M2, M3, M4 and M7 are the metabolites of fluvastatin based on the lexical hints "*metabolites of fluvastatin*".

Table 1 – Sample logic facts and evidence sentences for each type of relations that are extracted by our PTQL framework. Note that the protein names have been replaced with their official gene symbols in the logic facts.

Logic facts	Evidence sentences
`metabolizes(CYP3A4, fluvastatin)`	Fluvastatin is metabolized by CYP2C9 (PMID:16714062)
`is_expressed(ABCC2, liver)`	These decreases in hepatic Mrp2 may contribute to cholestasis (PMID:17959626)
`is_expressed(SLC15A1, intestine)`	PPARalpha plays critical roles in intestinal PEPT1 expression. (PMID:16751172)
`eliminates(ABCB1)`	Colchicine is also a substrate of P-glycoprotein, a transporter involved in cellular efflux and elimination of numerous drugs. (PMID:15494379)
`metabolite(desmethyldesipramine, desipramine)`	desipramine in rats may be attributed not only to the inhibition of the norepinephrine transporter by desipramine but also to the inhibition of serotonin transporter by the active metabolite desmethyldesipramine. (PMID: 17850785)

3.3. *Ordering of interactions through reasoning*

Once the relevant facts and interactions are acquired from knowledge bases and Medline abstracts, the last step is to utilize the facts and interactions to generate pharmacokinetic pathways. As the interactions themselves do not reveal any kind of ordering, the goal is to represent the fundamental behavior and properties of pharmacokinetics so that the representation can be utilized to assign ordering of the interactions through reasoning. Implementation of the reasoning component requires a language that is ideal in specifying what kind of reasoning to be performed rather than how the reasoning is performed. This is analogous to declarative programming language such as SQL, in which the users specify what is intended to be found rather than how the search mechanism of the database system should be performed to answer the queries. AnsProlog [29, 30] is a declarative language that is useful for reasoning, as well as capable for reasoning with incomplete information. We first describe how AnsProlog is applied to the representation of pharmacokinetic properties.

The core idea of the representation of pharmacokinetics is to encode the *pre-* and *post-conditions* of interactions, also known as the executability and direct effects of actions. *Timepoints* are used to define the logical ordering of the interactions. Interaction I_1 occurs before interaction I_2 if I_1 is assigned with a timepoint that is smaller than the timepoint for I_2. Using the interaction that involves the generation of metabolites as an example, the pre-condition of such generation is that the target drug has to be metabolized. The post-condition of the interaction is the production of metabolites. Such mechanism is represented by the following AnsProlog logic rules:

```
o(converts(D, M),Loc, T) :- h(metabolized(D),T),
metabolites(D, M), metabolism(D, Loc), not
h(converted(D),T).

h(converted(D),T+1) :- o(converts(D, M), Loc, T),
metabolites(D,M).
```

The first logic rule states that the pre-conditions for the action *converts(D, M)* occur at timepoint *T* in location *Loc* (which can be either the liver or intestinal cell), denoted as `o(converts(D,M),Loc,T)`. For instance, `o(converts(fluvastatin,m2),liver,3)` indicates that the drug fluvastatin is converted into the metabolite M2 in the liver cells at timepoint 3. The following are the pre-conditions, which are specified to the right of the "if" symbol `:-` in the rule, for the action *converts(D, M)*:

- the drug *D* has been metabolized at timepoint *T*, denoted as `h(metabolized(D),T)`;
- metabolite *M* is known to be a metabolite of *D*, denoted as `metabolites(D, M)`;
- metabolism of *D* is known to take place in *Loc*, denoted as `metabolism(D, Loc)`;
- it is not known that *D* has been converted into metabolites in the previous timepoints, denoted as `not h(converted(D),T).`

Notice that `metabolism(D, Loc)` and `metabolites(D, M)` are logic facts that are obtained from the extraction of knowledge bases and text. The second logic rule states the post-condition of the action *converts(D, M)*, which is to indicate the drug *D* is converted in the next timepoint *T*+1 when the action occurs at timepoint *T*. With the timepoints, we can observe that fluvstatin is converted into metabolites with `h(converted(fluvastatin),4)` and this conversion occurs as a result of the action `o(converts(fluvastatin, m2), liver, 3).`

Another example of pharmacokinetic behavior is that a drug can only be metabolized by some enzymes if the drug is distributed to the appropriate location in the liver by a drug transporter. In addition, elimination of the target drug can only take place when the drug is metabolized and metabolites are produced. Our logical representation also includes the fact that an orally-taken drug is transported to the intestines, and drugs that are taken intravenously are transported to the liver. To mimic the behavior of typical pharmacokinetic pathways, we include logic rules to ensure that all interactions involved in the intestinal cells have to occur ahead of the interactions in the liver cells. By encoding rules that represent the pharmacokinetic behavior of drugs, interactions are assigned with timepoints to indicate the ordering of the interactions in the pharmacokinetic pathway. With the logic rules and facts, the model, known as *answer sets* in AnsProlog, is computed by an answer set solver called clasp [31]. The resulting answer sets correspond to the pharmacokinetic pathway of the target drug.

Our approach for pathway synthesis can be summarized in the following steps:

1. Given the drug of interest, knowledge bases that include DrugBank, PharmGKB and the Gene Ontology annotations are utilized to acquire information such as interactions between drug and proteins, as well as the enzymes and drug transporters involved in the interactions.
2. To complement the information fetched from manually curated knowledge bases, information such as protein expression in liver/intestinal cells, the roles of drug transporters, drug metabolites are extracted from Medline abstracts using our PTQL extraction framework.
3. With the generic logical representation of pharmacokinetic pathways, facts and interactions acquired in steps (i) and (ii) are utilized so that the pharmacokinetic interactions are assigned with timepoints to reveal their ordering in the resulting pharmacokinetic pathway.

4. Synthesis of pharmacokinetic pathways

In this section, we illustrate our approach with the synthesis of pharmacokinetic pathways for two drugs: repaglinide and parvastatin. The pathways of another 18 drugs are synthesized by our system and presented as supplementary material in our website: http://www.kbpathway.org/. The pharmacokinetic pathways of these drugs have been manually annotated and made available in PharmGKB. Our system provides the output in the form of logic facts, as well as GPML files that can be visualized in Cytoscape [32] with the Cerebral plug-in [33], which takes advantages of the protein cellular locations in generating the layout of the pathways. We evaluate the performance of our system based on the performance of the extraction.

4.1. *Repaglinide pharmacokinetic pathway*

The logical representation of the pathway generated by our system is shown in Table 2. The model indicates that repaglinide is administered orally in the initial step, represented as `h(is_taken(repaglinide,orally),0)`. The drug consumption leads to the presence of the drug in the intestinal cells (`h(is_present(repaglinide, intestine),1)`), and the drug is transported to the liver cells through the bloodstream (`o(transports(repaglinide,intestine),liver,1)`). The drug repaglinide becomes present in the liver cells (`h(is_present(repaglinide,liver),2)`) and it is distributed by the hepatic drug transporter SLCO1B1 (`o(distributes(slco1b1,repaglinide),liver,2)`). Metabolism of repaglinide by the enzymes CYP3A4 and CYP2C8 (`o(metabolizes(cyp3a4,repaglinide),liver,2)`, `o(metabolizes(cyp2c8, repaglinide),liver,2)`) occurs after the distribution, and repaglinide becomes metabolized (`h(metabolized(repaglinide,liver),3)`). As a result of the drug metabolism, metabolites M1 and M4 are generated (`o(converts(repaglinide,m1),liver,3)`, `o(converts(repaglinide,m4),liver,3)`). The last timepoint indicates that repaglinide is no longer present in the liver, represented by `-h(is_present( repaglinide,liver),4)`, in which the symbol "`-`" corresponds to negation.

Table 2 – The output of the logical representation of the pharmacokinetic pathway of repaglinide generated by our system.

Timepoint	Events
0	h(is_taken(repaglinide,orally),0).
1	h(is_present(repaglinide,intestine),1). o(transports(repaglinide,intestine),liver,1)
2	-h(is_present(repaglinide,intestine),2). h(is_present(repaglinide,liver),2). o(distributes(slco1b1,repaglinide),liver,2). o(metabolizes(cyp3a4,repaglinide),liver,2). o(metabolizes(cyp2c8,repaglinide),liver,2).
3	h(metabolized(repaglinide,liver),3). o(converts(repaglinide,m1),liver,3). o(converts(repaglinide,m4),liver,3).
4	-h(is_present(repaglinide,liver),4).

The manually annotated pathway and the version synthesized by our system are shown in Figure 3. One distinctive difference is that our current approach is not capable of finding which of the enzymes are responsible for which metabolites. Here we assume that the enzymes CYP3A4 and CYP2C8 are responsible for the metabolism of repaglinide, and metabolites M1 and M4 are generated in the process. In the next illustration, we show that our automated pathway synthesis approach is capable of uncovering components that are not described in manually annotated pathways.

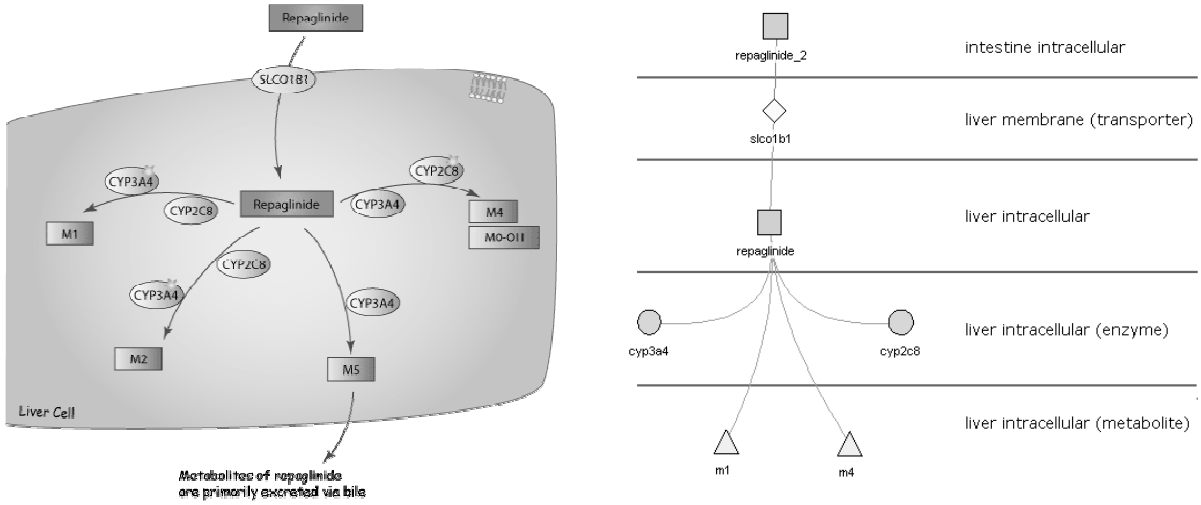

Figure 3 – (left) The manually curated pharmacokinetic pathway of the drug repaglinide from PharmGKB; (right) the pharmacokinetic pathway of repaglinide synthesized by our system.

4.2. *Pravastatin pharmacokinetic pathway*

Here we demonstrate that our method is capable of producing extra elements in the pathways compared to the manually annotated pathways. We use the synthesis of pravastatin pharmacokinetic pathway as an example. The manually annotated pathway in Figure 4 (left) lacks the information that states which enzymes are responsible for metabolism and what metabolites are generated as a result of the metabolism. Such information is included in our synthesized version of the pathway. As shown in Figure 4 (right), enzymes CYP3A4, CYP2C8 and CYP2C9 metabolize the drug pravastatin, and it is metabolized into the metabolites SN-38, 3alpha-hydroxy-iso-pravastatin and 3'alpha-iso-pravastatin. The drug-enzyme metabolic relations are originated from PMID: 17178259 according to the relationships in PharmGKB corresponding evidences. The resulting metabolites are extracted by our PTQL extraction framework, and the evidence sentences in Table 3 indicate the correctness of these facts and interactions.

Table 3 – Evidence sentences for the metabolites (underlined) of pravastatin extracted by our PTQL extraction framework.

PMID	Evidence	Correctness
16027406	Plasma concentrations of *pravastatin* and its active *metabolite*, 3alpha-hydroxy-iso-pravastatin, were measured, and pharmacokinetics was assessed.	Correct
10490896	In addition, as in the liver, *pravastatin* was metabolized in the small intestine by sulfation and subsequent degradation to its main *metabolite* 3'alpha-iso-pravastatin.	Correct
16515396	These genetic variants have been shown to lead to altered pharmacokinetics of OATP1B1 substrates, mainly *pravastatin*, but also the irinotecan metabolite SN-38	Incorrect

4.3. *Evaluation and analysis*

We evaluate the performance of our pathway synthesis method by finding how many of the interactions can be recovered (i.e. the coverage) with respect to 20 pharmacokinetic pathways available in PharmGBK. Table 4 shows the coverage when each of the sources DrugBank, PharmGBK relations and PTQL extraction is utilized for pathway synthesis. While the use of PharmGKB relations achieves the best coverage of the three sources with 47.27%, the coverage for the three sources combined results in 56.97%. We further manually evaluated the extraction performance by our PTQL framework. As DrugBank and PharmGKB do not provide information about drug metabolites, it is essential for the PTQL framework to be capable of extracting metabolites. Another important evaluation criterion is to determine if such approach in pathway synthesis can uncover more information than manually annotated pathways. Table 5 indicates that our extraction framework achieves a high precision of 84.0% and 82.72% for the extraction of enzymes and transporters, and metabolites. In particular, among the enzymes, transporters and metabolites uncovered by our method, our system produces 24 extra enzymes and transporters as well as 48 metabolites that are correct but not included in the 20 manually annotated pathways. The high quality of

extraction suggests that using PTQL framework for extraction is a valuable source to complement the existing knowledge bases for pathway synthesis.

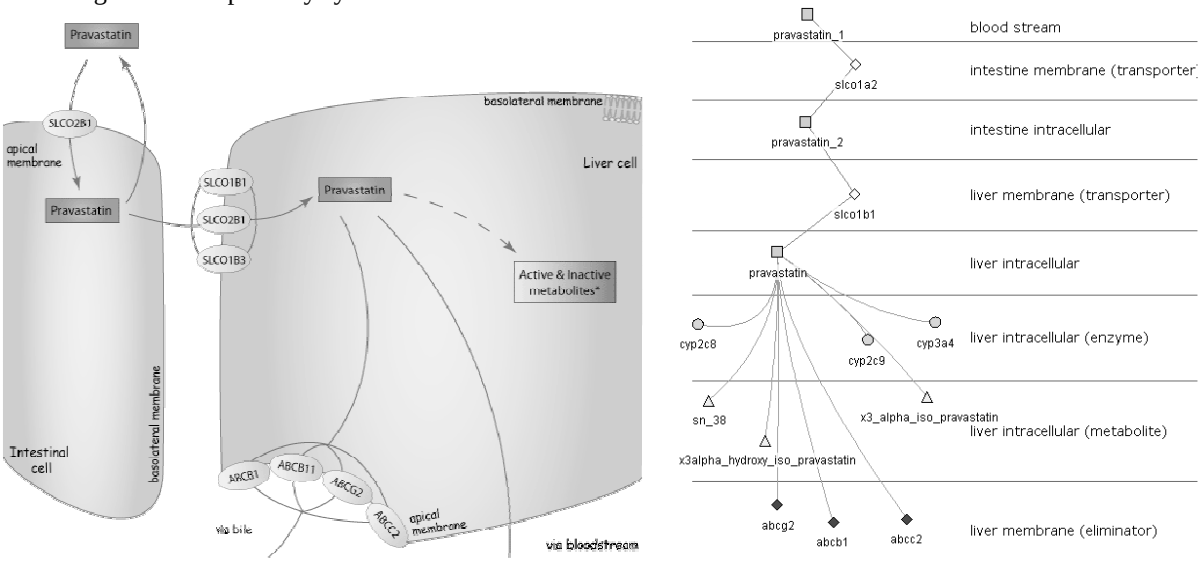

Figure 4 – (left) The manually curated pharmacokinetic pathway of the drug pravastatin from PharmGKB; (right) the pharmacokinetic pathway of pravastatin synthesized by our system.

Table 4 – Coverage of each of the sources in the pharmacokinetic pathways for the 20 manually annotated pharmacokinetic pathways in PharmGKB.

Sources	Coverage
DrugBank	20.61%
PTQL extraction	34.23%
PharmGKB	47.27%
All	56.97%

Table 5 – Precision and recall for PTQL extraction of enzymes and transporters, as well as metabolites. The ones that are not in the PharmGKB annotated pathways are considered as "Extra", and their correctness is evaluated.

	Precision	Recall	Extra (Precision)
Enzymes and transporters	84.00%	34.23%	19/31 = 61.29%
Metabolites	82.72%	65.05%	48/62 = 77.42%

We further analyzed the correctness of the placement of the interactions in the synthesized pathways compared to 20 pharmacokinetic pathways available in PharmGKB. Our synthesized pathways are generally consistent with the annotated pathways. However, our synthesized pathways do not include information that specifies which enzymes are responsible for the production of a particular drug metabolite. Such limitation is due to the fact that metabolites are not found in DrugBank and PharmGKB relationships, and our current text extraction is limited to the extraction from individual sentences. Drug-enzyme-metabolite relations can rarely be found within individual sentences.

Among the 20 pharmacokinetic pathways we evaluate, 2 of them are taken intravenously according to DrugBank, namely imipramine and irinotecan. In our logical representation of pharmacokinetics, we assume that no interactions occur in the intestine for drugs that are taken intravenously, unlike the orally-taken drugs that require absorption in the intestine. The imipramine pathway in PharmGKB only shows the interactions in liver cells, but the irinotecan pathway includes interactions in both the intestinal and liver cells, with the interactions in the liver cells appear ahead of the ones in the intestinal cells. Our current modeling of pharmacokinetic properties does not capture this behavior. In terms of metabolism, 17 of the drugs we evaluate are known to be metabolized in the liver cells according to PharmGKB, so by default we assume the metabolism of the other 3 drugs, namely atorvastatin, repaglinide, rosuvastatin, takes place in the liver cells. This assumption is valid except for the atorvastatin, lovastatin and simvastatin pathways that indicate the drugs are metabolized in both the intestinal and liver cells. In our current modeling, we assume that drug metabolism and drug distribution for elimination take place in the same cell. This is

not the case for the drugs clopidogrel, fluvastatin and pravastatin, in which the manually annotated pathways suggest that drug transporters take the drug from intracellular to extracellular in the intestinal cells, even metabolism of these drugs occur in the liver cells. Our current model also does not capture the transformation of a metabolite to another through enzymes, as suggested by the pathways for phenytoin and tamoxifen.

5. Conclusion

The study of pharmacokinetics is essential in identifying the effectiveness of drugs in the systemic circulation. In particular, variability in drug response is largely influenced by genetics. In this paper, we extend our previous work in synthesizing biological networks [25] by including a reasoning component for the synthesis of pharmacokinetic pathways. The use of reasoning distinguishes our approach from existing methods in generating networks of interactions so that ordering of interactions can be assigned through reasoning. Such ordering is critical for the representation of pathways, in which the effects of interactions trigger the subsequent interactions. Our results show that our approach is capable of synthesizing pharmacokinetic pathways in high quality and identifying components that are not in the manually annotated pathways. With the partial pathways generated by our approach, curators can utilize the synthesized pathways as a first step of curation and add their findings to expand the pathway annotation. Through the synthesized pharmacokinetic pathways, drug designers can examine the impact of drug response due to the genetic variations of the gene products involved in the pathways. Identifying relations between drug response and genetic variations is a critical step in realizing personalized medicines.

For future work, we will implement a web-based version of our approach so that pharmacokinetic pathways can be created based on the drugs specified by the users. Announcements will be made on our website at http://www.kbpathway.org when the web-based version of the implementation becomes publicly available. We also plan to expand our work to handle close-loop interactions, which cannot be captured in our current approach. Information such as drug-drug interactions will be included to identify drugs that inhibit or induce enzymes responsible for the metabolism of other drugs. Such information can be useful to drug designers as well as physicians to learn the potential side-effects of drugs due to drug-drug interactions. We also plan to apply our approach to other kinds of pathways, such as pharmacodynamics and signaling pathways.

Acknowledgements

We would like to thank the comments made by the anonymous reviewers. We would also like to acknowledge the support of these research grants: NSF 0412000, SFAZ CAA 0131-07 and SFAZ CAA 0289-08.

References

1. G. Joshi-Tope, M. Gillespie, I. Vastrik, et al. Reactome: a knowledgebase of biological pathways. *Nucl. Acids Res.*, **33**, suppl 1, D428-432 (2005).
2. M. Kanehisa, S. Goto, S. Kawashima, et al. The KEGG resource for deciphering the genome. *Nucleic Acids Res. Database Issue*, **32**, D277-80 (2004).
3. P. Romero, J. Wagg, M. L. Green, et al. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.*, **6**, 1, R2 (2005).
4. H. Mi, N. Guo, A. Kejariwal, et al. PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic Acids Res.*, **35**, D247 (2007).
5. T. E. Klein, J. T. Chang, M. K. Cho, et al. Integrating genotype and phenotype information: an overview of the PharmGKB project. *Pharmacogenomics Journal*, **1**, 3, 167 (2001).
6. P. Uetz, L. Giot, G. Cagney, et al. A comprehensive analysis of protein--protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 6770, 623-627 (2000).
7. D. Rajagopalan and P. Agarwal. Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics*, **21**, 6, 788-793 (2005).

8. C. Friedman, P. Kra, H. Yu, et al. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, **17 Suppl 1**, S74-82 (2001).
9. J. C. Park, H. S. Kim and J. J. Kim. Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. *Pac. Symp. Biocomputing*, 396-407 (2001).
10. M. Theobald, N. Shah and J. Shrager. Extraction of Conditional Probabilities of the Relationships Between Drugs, Diseases, and Genes from PubMed Guided by Relationships in PharmGKB. *AMIA* (2009).
11. G. A. Viswanathan, J. Seto, S. Patil, et al. Getting started in biological pathway construction and analysis. *PLoS Computational Biology*, **4**, 2 (2008).
12. A. Funahashi, M. Morohashi, H. Kitano, et al. CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *Biosilico*, **1**, 5, 159-162 (2003).
13. T. Aittokallio and B. Schwikowski. Graph-based methods for analysing networks in cell biology. *Brief Bioinform*, **7**, 3, 243-255 (2006).
14. X. Guo and A. J. Hartemink. Domain-oriented edge-based alignment of protein interaction networks. *Bioinformatics*, **25**, 12, 240-1246 (2009).
15. R. Sharan, S. Suthram, R. M. Kelley, et al. Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 6, 1974-1979 (2005).
16. M. L. Mavrovouniotis, G. Stephanopoulos and G. Stephanopoulos. Computer-aided synthesis of biochemical pathways. *Biotechnol. Bioeng.*, **36**, 1119-1132 (1990).
17. N. Tran, C. Baral, V. J. Nagaraj, et al. Knowledge-based framework for hypothesis formation in biochemical networks. *Bioinformatics*, **21**, suppl 2, ii213-219 (2005).
18. K. Oda, J. D. Kim, T. Ohta, et al. New challenges for text mining: mapping between text and manually curated pathways. *BMC Bioinformatics*, **9 Suppl 3**, S5 (2008).
19. Z. A. Sharif. Pharmacokinetics, metabolism, and drug-drug interactions of atypical antipsychotics in special populations. *J. Clin. Psychiatry*, **5**, 22-25 (2003).
20. N. Mizuno, T. Niwa, Y. Yotsumoto, et al. Impact of Drug Transporter Studies on Drug Discovery and Development. *Pharmacol. Rev.*, **55**, 3, 425-461 (2003).
21. D. Greenblatt, L. von Moltke, J. Harmatz, et al. Pharmacokinetics, pharmacodynamics, and drug disposition. *Neuropsychopharmacology: the Fifth Generation of Progress*, 507–24 (2002).
22. D. S. Wishart, C. Knox, A. C. Guo, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucl. Acids Res.*, **34**, suppl 1, D668-672 (2006).
23. E. Camon, D. Barrell, V. Lee, et al. The Gene Ontology Annotation (GOA) Database - An integrated resource of GO annotations to the UniProt Knowledgebase. *Silico Biology*, **4**, 1, 5-6 (2004).
24. P. H. Tu, C. Baral, Y. Chen, et al. Generalized text extraction from molecular biology text using parse tree database querying. *Arizona State University*, **TR-08-004** (2008).
25. L. Tari, J. Hakenberg, G. Gonzalez, et al. Querying parse tree database of Medline text to synthesize user-specific biomolecular networks. *Pacific Symposium on Biocomputing (PSB'09)*. (2009).
26. R. Leaman and G. Gonzalez. BANNER: An executable survery of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing (PSB'08)*, 652-663 (2008).
27. Aronson, A. R. *MetaMap: Mapping Text to the UMLS Metathesaurus.* (2006).
28. J. Hakenberg, C. Plake, R. Leaman, et al. Inter-species normalization of gene mentions with GNAT. (2008).
29. M. Gelfond and V. Lifschitz. The Stable Model Semantics for Logic Programs. *International Symposium on Logic Programming*, 1070-1080 (1988).
30. M. Gelfond and V. Lifschitz. Classical Negation in logic programs and disjunctive databases. *New Generation Computing*, 365-387 (1991).
31. M. Gebser, B. Kaufmann, A. Neumann, et al. clasp: A Conflict-Driven Answer Set Solver. *Proceedings of the Ninth International Conf. on Logic Programming and Nonmonotonic Reasoning (LPNMR'07)*, 260-265 (2007).
32. P. Shannon, A. Markiel, O. Ozier, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 11, 2498-2504 (2003).
33. A. Barsky, J. L. Gardy, R. E. W. Hancock, et al. Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. *Bioinformatics*, **23**, 8, 1040-1042 (2007).