ANALYZING INCONSISTENCY TOWARD ENHANCING INTEGRATION OF BIOLOGICAL MOLECULAR DATABASES*

YI-PING PHOEBE CHEN[1,2] AND QINGFENG CHEN[1]

[1]Faculty of Science and Technology, Deakin University, Melbourne, VIC 3125, Australia,
[2]Australia Research Council Centre in Bioinformatics, Australia

Abstract: The rapid growth of biological databases not only provides biologists with abundant data but also presents a big challenge in relation to the analysis of data. Many data analysis approaches such as data mining, information retrieval and machine learning have been used to extract frequent patterns from diverse biological databases. However, the discrepancies, due to the differences in the structure of databases and their terminologies, result in a significant lack of interoperability. Although ontology-based approaches have been used to integrate biological databases, the inconsistent analysis of biological databases has been greatly disregarded. This paper presents a method by which to measure the degree of inconsistency between biological databases. It not only presents a guideline for correct and efficient database integration, but also exposes high quality data for data mining and knowledge discovery.

1 Introduction

In recent years, advanced experiment methods have resulted in the rapid growth of life science databases. Many biological databases have been developed for different purposes, such as GenBank and NCBI [1-2]. The enormous data in databases are meaningful for the exploration of their life origin and evolution, and to predict the function and structure of life systems. They have been commonly usedby biologists during data analysis.

Due to the increasingly complex and specific nature of biological databases, a complicated biological question has to be answered by consulting multiple biological databases. However, the knowledge of life systems is too detailed and complex to be completely comprehended. Such complexity presents a big challenge to merge knowledge from diverse databases. The heterogeneity of databases blocks the accessibility to them [3-4]. In other words, the inconsistent structures and terminologies of biological databases result in a significant lack of interoperability. Thus, it creates a demand for data preprocessing.

As an important cleaning action, the integration of biological databases is significant when dealing with the heterogeneity of biological databases. However, the twisted and deformed biological data often demand additional knowledge so that the values held in databases can be specified and constrained. This causes considerable difficulties for data integration.

* This work is partially supported by a Discovery Grant DP0559251 from the Australian Research Council.

Technical and semantic problems are two key issues which present themselves when integrating biological databases. The former can be solved because most current biological databases are implemented on relational database management systems (RDBMS) that provide standard interfaces like JDBC and ODBC for data and metadata exchange [5-6]. Nevertheless, the solution of semantic problems remains unsolved..

Modern bioinformatics demand knowledge extracted from databases for communication purposes. For example, a user's query of a protein kinase may refer to hundreds of databases. There are two options to integrate knowledge from databases: (1) standardising the nomenclature of diverse databases; and (2) creating bridges between databases even if they differ radically in structure and nomenclature. The former have encountered resistance from database maintainers and specialists who hesitate to change preferred terminology [4]. As a tradeoff scheme, the latter has been commonly applied in the integration phase of biological databases. Among them, ontology-based biological database integration is one of the representative methods designed to capture knowledge from databases.

There have been many attempts to develop standards that can be applied tobioinformatics ontologies and which subsequently exploit biological information. For example, EcoCyc ontology [7] covers E.coli gene, metabolism, regulation and signal transduction, and Gene ontology (GO) [8] describes drosophila, moused and yeast gene function, process and cellular location and structure. Recently, ontology-based semantic integration of biological databases was presented in [2, 9]. Philippi [6] proposed a method for the ontology-based semantic integration of life science databases using XML technology. To enhance semantic interoperability, there have been considerable efforts to solve nomenclature-mapping problems and standardise the naming of functional relations and processes and their arguments such as ontology-mapping in GO community [10]. It provides a comprehensive list of synonyms that can be used immediately to improve indexing and search over the literature. However, no effort has been made to analyse the inconsistency of biological databases which would effectively lead to the enhancement of database integration.

This paper presents a method by which to analyse inconsistencies between biological databases using ontology. The method is able to find out the databases that are inappropriate for integration or need to be further improved. This not only reduces the search space but also generates high quality data for accurate and efficient data mining and knowledge discovery. Algorithms and experiments are presented to further demonstrate our approaches.

The remainder of this paper is organised as follows. Section 2 presents basic concepts. The approaches by which to analyse the inconsistency between biological databases are presented in Section 3. In Section 4, experiments are presented. Section 5 concludes this paper.

2 Basic Concepts

2.1 Problem Description

The increasing biological databases relating to genome sequences and protein structures and functions are challenging the traditional approaches for knowledge acquisition. To answer a complex biological question, hundreds of biological databases can be consulted. It is critical to guarantee accessibility to the databases. However, the discrepant structures and nomenclature of databases have an effect on their communication capabilities.

Although some biological data publishing and collection use HTML (Hypertext Markup Language) format, this method cannot describe complex structure documents. Besides, the varied organisation, storage and publication of biological data leads to different information types. For example, the representative database NCBI (National Center for Biotechnology Information) adopts mostly the binary ASN.1 [1], while flatfiles are used in GenBank [2]. The differences in the information types result in heterogeneities between biological databases and prevent us from obtaining high quality data for data analysis. Additionally, the information derived only from a single database does not enable us to obtain a comprehensive understanding, and the knowledge acquisition is inconvincible.

There have been some efforts to establish a link between disparate databases, such as data warehousing and database federation. Nevertheless, the increasing new data and databases have led efforts to reach a terminological impasse whereby databases have to agree on nomenclature and compatible formats before a link is able to be built. Nevertheless, database maintainers and specialists in certain research fields find it difficult to accept such a link. Ontology has been recently used to create bridges between biological databases. However, there are still some problems in the ontology-based integration of biological information. Problems include:

- Ontologies with independent terminologies and structures are often incompatible. This causes difficulties when acquiring knowledge from databases.
- Heterogeneities such as synonym result in a significant lack of interoperability among biological databases. This blocks the generation of high quality data.
- Semantic inconsistency has been widely ignored. The integration of biological databases with high discrepancies cannot guarantee efficient data mining.

Such inconsistencies surrounding biological databases and when the databases are appropriate for further processing, present major ambiguities. The analysis of the inconsistent nature of biological databases assists us in sorting out the appropriate databases from which high quality data can be derived. Hence, it is imperative to develop approaches by which to measure the inconsistency of biological databases and ensure reliable integration of biological databases.

2.2 Symbols and Formal Semantics

Suppose A and L denote atom symbols and proposition formulae respectively. In particular, *A* can contain α and $\neg\alpha$ for some atoms α. Let $\wedge$, $\neg$ and $\rightarrow$ be logical connectives. Let C, $c \in A$ be concepts such as *Gene* and *Protein*, *CV* for control vocabulary, r for relationship, and α, β and γ for attributes in general. Let $\equiv$ be logical equivalence. A model of a formula ϕ is a possible set of atoms where ϕ is true in the usual sense.

Controlled vocabularies are a set of named concepts that may have an identifier. The concepts or their identifiers are often used as database entries. Its definition is as follows.

DEFINITION 2.1. Suppose *t*, *def*, *id* and *sn* present term, definition, identifier and synonyms respectively. Let *C* be the set of concepts of *databases. Hence, we have*

$$\textit{Controlled Vocabulary } CV := \{c \mid c = (t, def, id, sn) \in C\}$$

An example of *Gene Ontology* (*GO*) is as follows. Each concept (*biological process*) has a term (*recommended name*), an identifier (*id: GO: number*), definition (*explanation and references*) and synonyms (*other names*). The *definition* of each *biological process* is provided by brief description and references to relevant literature or web links.

Ontology includes relationships as well as concepts. The relationships consist of '*is-a*' (*Specification relationships*) and '*part-of*' (*Partitive relationships*), by which concepts can correlate with each other. Although '*part-of*' relationship can be defined, only the transitive '*is-a*' hierarchy is required for querying databases. For example, '*Enzyme* is one kind of *Protein*', '*Protein* is one kind of *Macromolecule*' and '*Membrane* is part of *Cells*'. Therefore, ontology can be viewed as a tree, where the nodes and directed edges present concepts and relationships respectively.

DEFINITION 2.2. *Let O be ontology, and r be relationships that link concepts. Ontology can be defined as a set of tuples.*

$$\textit{Ontology } O := \{(c_1, c_2, r) \mid c_1, c_2 \in CV, \text{ and } r : c_1 \rightarrow c_2\}$$

where $c_1 \rightarrow c_2$ presents a relationship r from c_1 to c_2, such as 'c_1 *is-a* c_2'.

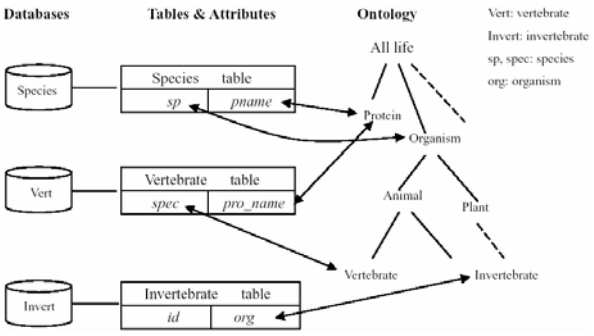

Figure 1. Biological database attributes are linked to ontology concepts. Attributes *pname* and *pro_name* from databases *Species* and *Vert* have different attribute names, but they are correlated by a common concept *protein* of the ontology.

EXAMPLE 2.1. In Figure 1, Vertebrate, Animal, Plant and Organism are connected by transitive ‘*is-a*’ relation. (*Animal, Organism*, *Animal* → *Organism*), (*Plant, Organism, Plant* → *Organism*), (*Vertebrate, Animal, Vertebrate* → *Animal*) *and* (*Invertebrate, Animal*, *Invertebrate* → *Animal*) represent tuples of ontology.

To analyse the inconsistency of biological databases, the above need to be defined semantically using ontology. One of the key processes is to link tables and attributes to a specified ontology. Subsequently, users can execute queries via hierarchies, such as ‘*is-a*’, to derive information from databases. Four operators to describe the interactions among attributes, tables and ontology are given below. Let $Att_1 \in DB_1$ and $Att_2 \in DB_2$ be database attributes. Let CV_1 and CV_2 be controlled vocabularies.

(1) **Mapping:** Let $Att \in DB$ be database attributes. Let *O* and *c* be ontology and concepts respectively. **‘***maps*(*O*, *Att*, *c*)’ states the attribute *Att* in *DB* can be mapped to a corresponding concept *c* via *ontology O*.

(2) **Cross-reference:** Let *CV* be controlled vocabulary. ‘*cross-reference*(*CV*, (Att_1, Att_2), *c*)’ states that if Att_1 and Att_2 can be linked to a common concept *c* by *cross-reference* of *CV*, they are semantically equivalent owing to *c*.

(3) **Translation:** ‘*translates*((CV_1, CV_2), (Att_1, Att_2), *c*)’ states that database attribute Att_1 and Att_2 can be translated to a common concept *c* using the controlled vocabularies CV_1 and CV_2. Thus, it is feasible to relate database entries that use different terms for the same thing, such as the *English species name* and *Systematic species name* in Figure 2.

(4) **Taxonomy:** Let c_i *and* c_j be concepts. ‘*is-a*(c_i, c_j)’ states that c_i is a sub-concept of c_j and c_j is a sup-concept of c_i. For simplicity, the operator ‘*is-a*’ below implies both ‘*is-a*’ and ‘*part-of*’ relationships mentioned above. Actually, the ‘*is-a*’ relationship holds transitivity. Hence, we have

$$\forall\, c_1, \ldots, c_n \in \mathrm{O},\ \ is\text{-}a(c_1, c_2) \wedge is\text{-}a(c_2, c_3) \wedge \ldots \wedge is\text{-}a(c_{n-1}, c_n) \rightarrow\ is\text{-}a(c_1, c_n)$$

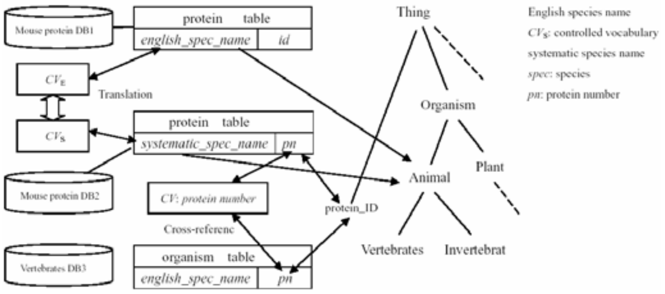

Figure 2. Translation by mapping synonymous concepts of controlled vocabularies is used to link databases with synonyms. Database attributes corresponding to the same concept and sharing the same controlled vocabulary can be viewed as cross-references of attributes.

The above axioms describe possible processes in response to a user's query on biological databases. Ontology plays a central role in mapping database attributes to common concepts or translating attributes between different controlled vocabularies, such as *English controlled vocabulary* and *Systematic controlled vocabulary* in Figure 2.

Additionally, *queries* operator usually intends to search in *attribute* for specified *terms* as mentioned above. Hence, a user's query can be classified into two categories in terms of entries regarding *attribute*:

- if the queried attribute is found in databases, it will be mapped to a corresponding concept of ontology, and will enable other database attributes to be linked together;
- if no database attribute is defined as the queried attribute, a corresponding concept of ontology is selected. Its sub-concepts and super-concepts will be searched to find the attribute.

Although the latter is complex, it can eventually get back to the former pathway via ontology. In either case, the queries bring about a collection of results, which can be used to measure the inconsistency found in biological databases. Usually, users specify a term *T* along with queries. *T* is able to reduce the searched concepts that are irrelevant to the queries. Suppose *Att* is the queried attribute by users, and its mapping concept of ontology *O* is *C*. Hence, we have

1. $sub(C, T) = \{c \mid \forall\, c,\ is\text{-}a(c, C),\ c \sqsupseteq T\}$
2. $sup(C, T) = \{\ c \mid \forall\, c,\ is\text{-}a(C, c),\ c \sqsupseteq T\ \}$

where $\sqsupseteq$ denotes a inclusion relationship in view of semantics.

EXAMPLE 2.2. Suppose a queried database attribute is *Animal* with a specified term *parrot*. Hence, in Figure 1, we have *sub*(*Animal*, *parrot*) = {*Vertebrate*}, *sup*(*Animal*, *parrot*) = {*Organism*}. Without the term *parrot*, *sub*(*Animal*, *parrot*) = {*Vertebrate*, *Invertebrate*}, *sup*(*Animal*, *parrot*) = {*Organism*}.

From the observation, the database attributes should be semantically defined as specific as possible, which can avoid searching unrelated databases.

DEFINITION 2.3. *Let* $ATT_{DB} = \{a_1, a_2, \ldots, a_n\}$ *be a set of attributes of biological database DB. The set of attributes derived from reference database and compared databases are denoted by* ATT_R and ATT_C *respectively.*

The reference database consists of multiple databases containing the queried attribute or the attribute that can be mapped to concepts of $sub(C, T) \cup sup(C, T)$. It is used to decide whether or not the attributes found in compared databases are consistent with the specified attribute. An example regarding ATT_R and ATT_C is given below.

EXAMPLE 2.3. In Figure 1, $ATT_{Species} = \{sp, pname\}$, $ATT_{Vert} = \{pr\_\ name, spec\}$ and $ATT_{Invert} = \{id, org\}$. If users query attribute *pname*, then $ATT_R = ATT_{Species} = \{sp, pname\}$ and $ATT_C = ATT_{Vert} \cup ATT_{Invert} = \{pro\ name, spec, id, org\}$..

DEFINITION 2.4. *Let* $\models$ *be a supporting relationship. For a set of database attributes* ATT_{DB}, $ATT_{DB} \models$ *is defined as follows.*

(1) if the queried database attribute α is found in current databases, we have

- $ATT_R \models \alpha$ iff ATT_R contains α
- $ATT_c \models \neg\alpha$ iff ATT_c contains β that is a database attribute of compared databases, which has a common concept with α.

(2) if the queried database attribute α is not found in databases but can be mapped to a concept *C* in ontology, we have

- $ATT_R \models \alpha_1$ iff ATT_R *contains* α_1 and $maps(O, \alpha_1, c)$

- $ATT_C \vDash \neg\alpha_1$ iff ATT_c *contains* α_2 that is the corresponding database attribute of *c* in compared databases

Here *O* and *c* present the ontology and concepts in *sub*(*C*, *T*) ∪ *sup*(*C*, *T*) respectively. α_1 denotes a mapping attribute in reference database from *c*.

EXAMPLE 2.4. Suppose *'pname : mouse'* and *'animal : mouse'* are two queries on Figure 1, in which *pname* and *animal* are queried attributes, and *mouse* is a term that locates databases. For the query *'pname : mouse'*, database *Species* can be viewed as the reference database. The term *mouse* reduces the search space to database *Species* and *Vert.* Hence, we have $ATT_{Species} \vDash pname$ and $ATT_{Vert} \vDash \neg pname$. For the query *'animal : mouse'*, no database attribute is defined as *animal.* Nevertheless, the sub-concepts and super-concepts of *animal* in ontology can be mapped to this attribute. The search will be limited to *Vertebrate* and *Species* due to the term *mouse* that is mapped to attribute *spec* of *Vert* and *sp* of *Species. Vert* is selected as the reference database so $ATT_{Vert} \vDash spec$. The attribute *sp* of *Species* is viewed as a negative attribute of *spec,* namely $ATT_{Species} \vDash \neg spec$.

3 Analyzing Inconsistency of Biological Databases

3.1 Models of Queried Biological Databases Attributes

DEFINITION 3.1. *Suppose* $ATT \in \wp(L)$, $X \in \wp(A)$. *Let* ATT_{DB} *be attributes derived from* $DB \in \{R, C\}$. *Let* $X \vDash ATT$ *denote that* $X \vDash \alpha$ *holds for every* α *in ATT.*

$$model(ATT) = \{X \in \wp(A) \mid X \vDash ATT\}$$

where *ATT* denotes a set of database attributes. The model of *ATT* actually presents a set of atoms that support *ATT.*

For measuring inconsistency, we use compatibility of biological databases. The *consistentset* of a model is the set of database attributes that have identical names with corresponding reference attributes. The *conflictset* of a model consists of (1) the set of database attributes that are semantically equivalent; and (2) the *null* attribute that presents no attribute is semantically equivalent to the reference attribute. Actually, some databases may not contain the queried attribute at all.

DEFINITION 3.2. *Let* δ *be a selected reference attribute from reference database R. Let* $Y \in \wp(A)$ *be a model of* δ.. *The consistentset and conflictset are defined below.*

- $Consistentset(\alpha) = \{\alpha \mid \alpha \in Y, \alpha \equiv \delta\}$
- $Conflictset(\alpha) = \{\alpha \mid \alpha \in Y, \alpha \equiv \neg\delta, \text{or } \alpha \equiv null\}$

Based on *consistentset* and *conflictset* from minimal models, a measurement can be used to compute the inconsistency of minimal models in respect to specified database attributes.

DEFINITION 3.3. *The compatibility function from A into* [*0, 1*], *is defined below when* α *is not empty, and Compatibility*(∅) = *0.*

$$Compatibility(\alpha) = \frac{|Consistenset(\alpha)|}{|Consistentset(\alpha)| + |Conflictset(\alpha)|} \times 100\%$$

where $|Consistentset(\alpha)|$ and $|Conflictset(\alpha)|$ are the cardinality of $Consistentset(\alpha)$ and $Conflictset(\alpha)$ respectively. If $Compatibility(\alpha) = 0$, then we can say that the model Y has no opinion upon α and vice versa; if $Compatibility(\alpha) = 1$, it indicates that there are no negative attributes $\neg\alpha$ in the model Y; if $0< Compatibility(\alpha) <1$, it presents the model Y as partially inconsistent/consistent with respect to α.

The compatibility function quantifies the inconsistency of biological databases in relation to the queried database attributes. A queried attribute in biological databases is regarded as compatible or consistent in the case that the compatibility regarding this attribute is equal to, or greater than, the threshold minimal compatibility (*mincomp*) given by users or experts. Here, let $mincomp = 0.5$. Then

- *consistent* if $Compatibility(\alpha) \geq mincomp$
- *inconsistent* if $Compatibility(\alpha) < mincomp$

Ideally, we would like to achieve $Compatibility(\alpha) = 1$. Nevertheless, the inconsistency is the objective existence of discrepant terminologies and ontology used in diverse biological databases. If biological databases are found to be inconsistent, it is possible that they contain too many incompatible database attributes or most of current databases do not contain the queried database attribute. Therefore, the method presented in this paper is a prerequisite for the integration of biological databases.

3.2 Experiments

Table 1 represents the definition of attribute fields for five biological databases and refers to Gene ontology and NCBI databases. Among them, DB_2 and DB_5 are databases which regard *Species*, DB_1 and DB_4 are databases with respect to *Vertebrate*, and DB_3 is about *Invertebrate*. All database attributes under *DNA sequence* are linked to *DNA sequence* concept of the ontology. For the attributes under *Organism*, *org* is related to *Vertebrate* concept but *spec* is linked to *Organism* concept. The *null* value in this table means no such attribute is defined in corresponding biological databases, such as attributes of DB_1 under *Description* column. In particular, the attribute $spec_s$ of DB_5 is a systematic species name. A translation is, therefore, needed to search for this attribute.

Table 1. Attributes of biological databases.

DB	Author	DNA Sequence	Description	Identifier	Organism	Enzyme
DB_1	*au*	*dns*	*null*	*id*	*org*	*null*
DB_2	*author*	*seq_dna*	*desc*	*id*	*spec*	*ename*
DB_3	*au*	*seq_dna*	*desc*	*mid*	*org*	*ename*
DB_4	*au*	*seq_dna*	*null*	*gid*	$spec_s$	*enzyme*
DB_5	*au*	*seq_dna*	*null*	*id*	*org*	*ec_nr*

Measuring the inconsistency of biological databases mainly comprises of three steps: (1) input queried database attributes; (2) compute the compatibility of databases in relation to queried attributes; and (3) determine the consistency of databases. Two experiments are presented below. One is to query attribute '*enzyme* : *mouse*' via cross-reference, and the other is to query attribute '*animal* : *mouse*' using translation.

In the former, DB_3 is ignored for it does not meet the constraint *mouse*. DB_5 is selected because the reference database for *enzyme* is found in DB_5, which is mapped to concept *protein* of the ontology in Figure 1. According to the ontology, the attributes under *Enzyme* of DB_2, DB_4 and DB_5 use different terminology to represent the same concepts. The common concept *Enzyme* can be used for cross-reference among them. According to Definition 3.2, we can obtain *Consistentset*(*enzyme*) = {*enzyme*} from DB_5, and *Conflictset*(*enzyme*) = {*null*, *enmae*, *ename*} from DB_1, DB_2 and DB_4. Both *null* and *ename* are regarded as ¬*enzyme* when computing the compatibility of biological databases. Finally, we obtain *Compatibility*(*enzyme*) = 1 / 4 = 0.25 < *mincomp*. Therefore, the biological databases are inconsistent in relation to the database attribute *enzyme*.

As for the latter case, DB_3 is ignored in the same way. The database attribute *org* of DB_1 and DB_4 is linked to *Vertebrate* concept in the ontology, and *spec* of DB_2 and DB_5 is linked to *Organism* concept in Figure 2. Among them, the *spec* attribute in DB_2 and DB_3 needs to be translated to the corresponding attribute *specs* in DB_5 for it is a systematic species name. Hence, the *model*(*ATT*) = {*org, spec, org, spec*}. There are two possibilities by which to select the reference attribute here: (1) *org*; and (2) *spec*. If we use *org* as the reference attribute, we have *Consistentset*(*org*) = {*org*, *org*} with respect to DB_1 and DB_4, and *Conflictset*(*org*) = {*spec*, $spec_s$} in relation to DB_2 and DB_5. Therefore *Compatibility*(*org*) = 2 / 4 = 0.5 ≥ *mincomp*. Therefore, the biological databases are consistent in respect to the queried database attribute *animal* : *mouse* using *org*. On the other hand, if we use *spec* as the reference attribute, we have *Consistentset*(*spec*) = {*spec*}, *Conflictset*(*spec*) = {*org*, *org*, $spec_s$} and Co*mpatibility*(*spec*) = 1 / 4 = 0.25 < *mincomp*. Thus, they are inconsistent in respect to the database attribute (*animal, mouse*) using *spec*.

4 Conclusions

Knowledge acquisition from biological databases plays a nontrivial role in biological studies. However, the heterogeneity of biological databases has resulted in a significant lack of interoperability between them. The integration of biological databases is critical when dealing with heterogeneity but suffers from the twisted and deformed nature of biological data. Ontology-based integration of biological databases is an efficient way to capture knowledge from multiple sources. Nevertheless, no effort has been made to analyse the inconsistency in biological databases. This paper proposes a method to measure the inconsistency of biological databases via ontology. It assists in obtaining high quality data for data mining and knowledge discovery. We demonstrate our method by conducting experiments.

References

1. http://www.ncbi.nlm.nih.gov/.
2. Benson DA., Karsch-Mizrachi I., Lipman DJ., Ostell J. and Wheeler DL., GenBank Update, *Nucleic Acids Research*, vol 32 (Database issue), pp 23-26, 2004.
3. Stevens R., Goble C., Horrocks I. and Bechhofer S., OILing the Way to Machine Understandable Bioinformatics Resources, *IEEE Trans Inf Technol Biomed*, 6(2), pp 129-134, 2002.

4. Williams N., Bioinformatics: how to get databases talking the same language, *Science*, 275(5298), pp 301-302, 1997.
5. Kohler J., Philippi S. and Lange M., SEMEDA: ontology based semantic integration of biological databases, *Bioinformatics*, 19(18), pp 2420-2427, 2003.
6. Philippi S. and Kohler J., Using XML Technology for the Ontology-based Semantic Integration of Life Science Databases, *IEEE Trans Inf Technol Biomed*, 8(2), pp 154-160, 2004.
7. Karp P.D., Riley M., Saier M., Paulsen I.T., Paley S.M. and Pellegrini-Toole A., The EcoCyc and MetaCyc Databases, *Nucleic Acids Research*, 30(1), pp 59-61, 2000.
8. Yeh I., Karp PD., Noy NF. and Altman RB., Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO), *Bioinformatics*, 19(2), pp 241-248, 2003.
9. Kohler J., Philippi S. and Lange M., SEMEDA: ontology based semantic integration of biological databases, *Bioinformatics*, 19(18), pp 2420-2427, 2003.
10. http://www.geneontology.org/.