Prospects for an Ontology-Grounded Language Service Infrastructure

Yoshihiko Hayashi
Graduate School of Language and Culture, Osaka University
1-8 Machikaneyama, Toyonaka, 5600043 Japan
hayashi@lang.osaka-u.ac.jp

Abstract

Servicization of language resources (LR) and technologies (LT) on an appropriately designed and adequately operated infrastructure is a promising solution for sharing them effectively and efficiently. Given this rationale, this position paper reviews relevant attempts around the Language Grid, and presents prospects for an ontology-grounded language service infrastructure. As the associated issues may have substantial depth and stretch, collaborations among international and inter-cultural experts are finally called for.

1 Introduction

Servicization of language resources (LR) and technologies (LT) on an appropriately designed and adequately operated infrastructure is a promising solution for effectively and efficiently sharing them. Such an infrastructure would enable: (a) More non-expert users to have accesses to LR/LT without being too much bothered by cumbersome IPR issues; (b) virtual/dynamic language resources to be realized as language services through useful combination of the existing language services. To enjoy the benefit particularly described in (b), however, we need to address the issue of *interoperability* (Calzolari, 2008).

In the rest of this position paper: The notion of an ontology-grounded language service infrastructure is first introduced; An ontological construct for describing language services and the associated linguistic elements, referred to as *language service ontology*, is then sketched out; By reviewing the attempts around the Language Grid (Ishida, 2006; Ishida, 2011), including the language service ontology, issues and the prospects for an ontology-grounded language service infrastructure is then discussed. As the associated issues may have substantial depth and stretch, collaborations among international and inter-cultural experts are finally called for.

2 Language Service Infrastructure

A language service infrastructure is a software platform on which effective and efficient dissemination and utilization of servicized language resources will be possible. As nicely demonstrated by the Language Grid, such an infrastructure can provide a solid foundation for supporting activities of certain types. For example, the primary goal of the Language Grid was to support a range of activities associated with intercultural collaboration. However, such an infrastructure can attract more audiences as originally intended, if it could provide easier access to a reasonable set of language resources; the Language Grid, for instance, has been utilized by researches in the field of information and communication sciences.

Therefore a language service infrastructure should be designed, built, and operated while considering a wide variety of potential users, which include not only activists/end-users (service consumers) but also LR/LT experts (service providers). In addition, further cooperations among language service infrastructures should be considered as probably discussed in this workshop.

Given the potential benefits of language resource servicization, as discussed in the previous section, one of the most important features of a language service infrastructure is to provide a sufficient set of actual services, each classified into a reasonable service type. This is particularly important, as a service interface (or application program interface: API) should be specified according to the type of a service. To enable this, we primarily have to have a reasonable list or taxonomy of language service types.

As of February 2011, the Language Grid accommodates more than 100 Web services, which

Proceedings of Workshop on Language Resources, Technology and Services in the Sharing Paradigm, pages 1–7,
Chiang Mai, Thailand, November 12, 2011.

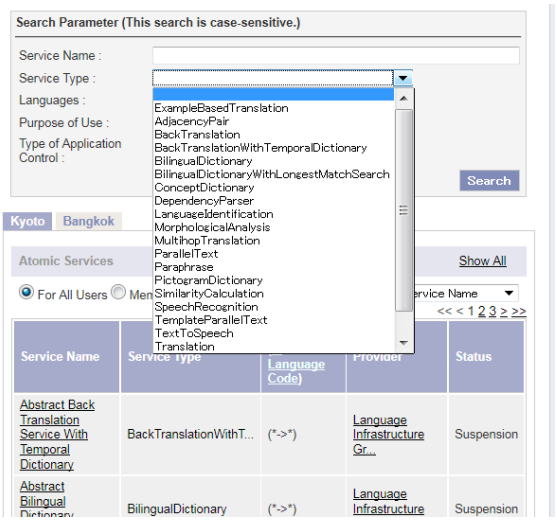

Figure 1: Language services in the Language Grid

are classified into one of the around 20 service types[1]. A user can utilize the provided language services through accordingly defined APIs. Figure 1 shows a screenshot from the Language Grid Web site, where a user can search for a language service based on the service type and/or supported languages.

To identify possible language service types and to further organize them structurally, we should, at least, consider two aspects: (1) functionality of the service, and (2) the input/output data types. The issue of interoperability arises here: as the underlying language resources are independently developed, they essentially exhibit idiosyncrasies in many aspects. A promising approach to partly address this issue would be to have a comprehensive vocabulary, or an ontological construct, so as to we can define and describe a language service type and the accordingly defined interface.

3 Language Service Ontology

Among the relevant attempts (Klein and Potter, 2004; Villegas et al., 2010), one came out from around the Language Grid is an ontological construct referred to as *language service ontology* (Hayashi et al., 2011). The language service ontology is intended to cover not only language services but their necessary elements including types of linguistic data object.

Figure 2 illustrates the top-level of the proposed language service ontology. The upper half of the diagram depicts our notion of the fundamental structure of a language service: (1) a language service is provided by a language process; (2) a language process operates upon linguistic objects by using language data resources; (3) a language data resource consists of linguistic objects; (4) a language data resource is created by organizing a set of linguistic objects each processed by language processes.

It should be noted here that the linguistic object class includes a range of linguistic annotations as well as linguistic expressions, which are the targets of annotations. These types of abstract objects comprise the data to/from NLP tools/systems, as well as the content of language data resources.

The lower half of the diagram, on the other hand, additionally introduces some important classes. Each box in the diagram denotes a top-level class in the whole ontology; some of these classes further induce corresponding sub-ontologies (Hayashi et al., 2011).

Among these top-level classes, `LanguageService` is functionally the top-most one: a language service is provided by an instance of `LanguageProcessingResource` class. Note that a language data resource does not provide a language service by itself; as it is a static resource, it is always activated through an access mechanism, which is an instance of a language processing resource subclass.

A language processing resource takes `LinguisticObject` as the input/output, and may use `LanguageDataResource`. `LanguageDataResource` consists of `LinguisticObject`, which might have been brought about by the results of `LanguageProcessingResource`. The language processing resources should be further classified according to their functionalities; the functionality is largely characterized by the types of associated objects. More specifically, the types of used language resources and/or the types of input/output language objects induce the taxonomy of language processing resources as displayed in Fig. 3

`LinguisticObject`, according to Saussure tradition, can have linguistic forms (`LinguisticExpression`) and meanings (`LinguisticMeaning`), where the former denotes the latter. Additionally, a linguistic meaning can be described by `TextualDescription`.

[1] `http://langrid.org/service_manager/language-services`

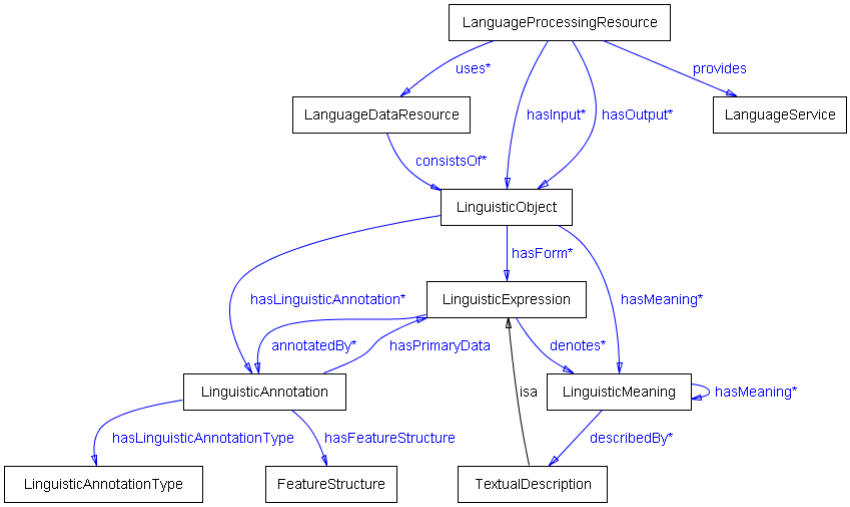

Figure 2: Top-level of the Language Service Ontology

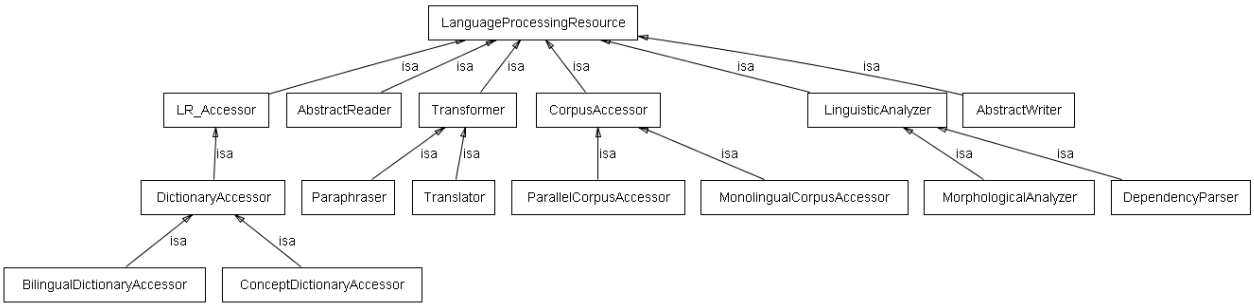

Figure 3: Taxonomy of the Language Processing Resources

Note here that an instance of the linguistic meaning class functions as a place holder for representing a semantic equivalent relation among linguistic objects. On the other hand, a `LinguisticObject` instance can be annotated by instances of `LinguisticAnnotation`, which should have actual annotation content represented with `FeatureStructure`.

4 Prospects for an Ontology-Grounded Language Service Infrastructure

4.1 Two issues uncovered

Each language service in the Language Grid is classified as one of the around twenty service types, including: CONCEPT DICTIONARY, MORPHOLOGICAL ANALYSIS, DEPENDENCY PARSER, and TRANSLATION. Each service type specifies its API, which includes data type specification for the input/output. The input/output data types, as also discussed previously, contributes to forming the taxonomy of language processing resources. Table 1 demonstrates this by listing major Language Grid service types and relating them to classes in the language service ontology. Note here that the ontology classes shown in the table are placed relatively upper in the taxonomy.

Through this review, the following two issues are uncovered.

- Although the language service ontology has been formalized so as to be comprehensive and linguistically-sound, the consensus among the related experts has not yet been reached. Also the current coverage may not be sufficient, insisting that the language service ontology has to be further expanded and revised.

- Although the set of Language Grid service types has been developed so as to be compatible with the language service ontology, there are no direct connections between them, insisting that actual utility of the language

Table 1: Major Language Grid Service Types and the Associated Ontology Classes

Service type	Ontology class	Input type	Output type
TRANSLATION	`Translator`	sentence string	sentence string
PARAPHRASE	`Paraphraser`	sentence string	sentence string
CONCEPT DICTIONARY	`DictionaryAccessor`	query string	lexical entry
BILINGUAL DICTIONARY	`DictionaryAccessor`	query string	lexical entry
PARALLEL CORPUS	`CorpusAccessor`	query string	annotation
MORPHOLOGICAL ANALYSIS	`LinguisticAnalyzer`	sentence string	morphological annotation
DEPENDENCY PARSER	`LinguisticAnalyzer`	sentence string	dependency annotation

service ontology is still not obvious, hence should be attested and demonstrated.

We will look at these issues in more detail.

4.2 Refining the language service ontology

The language service ontology should be considerably expanded and detailed in order for it to be used as an effective vocabulary for describing a wide variety of language services and the elements.

To accomplish this, we first need to identify the current and potential language service types and the elements. An actual language service infrastructure such as the Language Grid provides us with a concrete list of such elements, we however have to go beyond to further enrich the list; this, at least, requires collaborations among LR/LT experts. We however may further need to incorporate user requirements, particularly in a collaborative environment, for example the one offered by the Language Grid. Figure 4 generally illustrates necessary steps toward the goal, where we have to:

- Identify possible language service types. To this end, bottom-up activities, such as "LREC2010 Map of Language Resources, Technologies and Evaluation"[2], are crucially important. In parallel, we need to establish more connections with potential user communities of various kinds to discover novel service functionalities.

- Classify and describe the service types. We first have to clarify the dimensions of classification. Obviously, input/output linguistic data type and language processing functionality are two important things. We then need to organize ontological knowledge that includes a taxonomy of application-oriented use intentions as well as LR/LT domain ontologies: these domain ontologies can partly be organized by basing on the relevant international standards for linguistic data modeling, as further noted below.

- Facilitate the Web-servicization. We will be able to facilitate this by giving a wrapper template for each service type. Ontological knowledge would be further beneficial, as they could be utilized in (semi-)automatic service composition as discussed later.

A note on another role of LR standards: In further detailing some of the important sub-ontologies, on the other hand, we believe it is crucial to incorporate relevant international standards to deal with the issue of interoperability. In this sense, we have been looking at Linguistic Annotation Framework (LAF) (Ide and Romary, 2004) and Lexical Markup Framework (LMF) (Francopoulo et al., 2009) and the associated standards discussed in ISO[3]. LAF has been incorporated into our ontology not only for specifying the input/output data type of NLP tools, but also for defining the content type of corpora; while LMF has been introduced to develop a taxonomy of lexicon classes, which obviously forms a part of the language data resource taxonomy.

Figure 5 depicts how a particular class for syntactic annotation can be defined in the language service ontology by incorporating the Syntactic Annotation Framework (SynAF) (Declerck, 2008) standard, which is a subtype of general LAF in the sense that it focuses on syntactic annotations. Similarly Fig 6 shows that subtypes of lexicon class can be defined in terms of types of lexical entry, and the types of lexical entry should be speci-

[2] `http://www.lrec-conf.org/lrec2010/?LREC2010-Map-of-Language-Resources`

[3] `http://www.tc37sc4.org/`

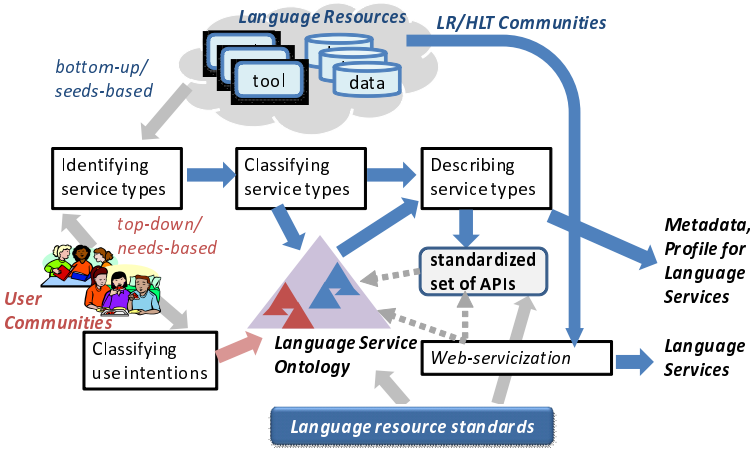

Figure 4: Steps toward standardized service APIs

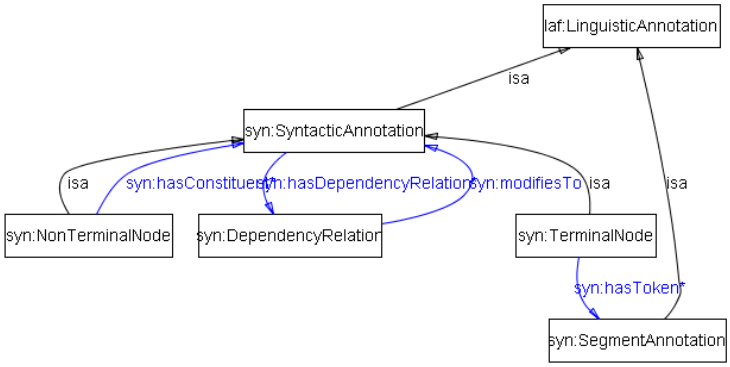

Figure 5: Ontologization of LAF and SynAF

fied by incorporating the *ontologized* LMF specification.

4.3 Linking service specifications with service ontology

The current standard for giving the concrete technical specification to a Web service (type) is to assign a Web Service Description Language (WSDL)[4] document to the Web service. Although a WSDL document defines the service name, functions, and input/output data types, it does not provide any semantic annotation to the elements. For example, the input/output data types defined in a WSDL document do not give us any ideas about which abstract linguistic object type is associated with which concrete data type. Therefore, to ensure the interoperability of a service and its service description, the WSDL document should be associated with the background service ontology in some way.

Among several possible solutions to this issue[5], we see adoption of the W3C recommendation Semantic Annotations for WSDL and XML Schema (SAWSDL)[6] could be a reasonable first step. The most prominent reason for this is its simplicity: as semantic annotations are just added to a WSDL document, the current Web service practices around WSDL can be maintained; SAWSDL does not require any special language for representing semantic models for the annotations, meaning that we could interrelate a WSDL document with the language service ontology. In fact, with the `sawsdl:modelReference` construct provided by SAWSDL, we can semantically annotate a WSDL document by making references to the classes in the language service ontology.

Although this solution could be a reasonable first step toward the full-fledged semantic Web services as discussed in (Yu, 2007), we will

[4] `http://www.w3.org/TR/wsdl`

[5] (Villegas et al., 2010) also discuss this topic and adopt a MyGrid approach (Wolstencroft et al., 2007), where descriptions about service invocation are also separated from the service ontology.

[6] `http://www.w3.org/TR/sawsdl/`

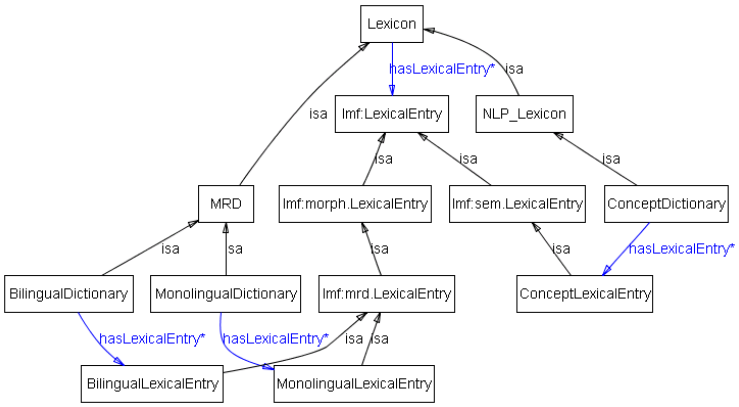

Figure 6: Lexicon Taxonomy as based-on LMF

have to develop an external mechanism for service discovery and compositions on top of the language service ontology and semantically augmented descriptions of the set of language Web service types. Furthermore, if we are stepping forward to the direction of planning-based automatic service composition, we have to device a system for representing goals and statuses. This is an area where almost nothing has been worked out, particularly with respect to the language service ontology.

5 Discussion

In this section, two distinct topics are discussed as below.

The first topic is about the activities for achieving an effective linguistic service infrastructure or software platform. A number of activities can be mentioned; among them, UIMA (Hahn et al., 2008) has gained a prominent position, particularly in text mining applications. U-Compare (Kano et al., 2009) is one of the representative software platforms that utilizes UIMA as the foundation. U-Compare, in particular, has stressed on task-dependent comparison and evaluation of the linguistic processing elements, and provides utilities to accomplish these tasks. A type system for a range of linguistic annotations with the UIMA framework is proposed in (Hahn et al., 2007), sharing common objectives with a part of the language service ontology. Heart of Gold (Schäfer, 2008) is another example of software platform, in which XML together with XSLT play a crucial role. In Heart of Gold, the integration of shallow and deep NLP components is particularly focused on. It should be noted that these platforms, in general, center on the effective creation of a so-called NLP pipeline, and pay little attention to access to lexical resources.

The second topic is just associated with the access to lexical resources. Maybe needless to say, there exist types of resource and/or types of resource access that do not suit well with the query-based access usually provided by language Web services. For example, an access requesting transferring large amount of data would be impossible or prohibited. Moreover, types of access requiring long computational time, for example one that demands complex corpus statistics figures, would be inadequate in a language service infrastructure. Nevertheless, as pointed out at the beginning of this paper, easier access to lexical resources might allow the users to realize a virtual/dynamic resource, that actually does not exist as a whole. One might expect classes of hybrid dictionary, as exemplified in (Hayashi, 2011), to be virtually realized in a language service infrastructure on a query-driven and an on-demand basis.

6 Concluding Remarks

This position paper argued that realizing and maintaining a standardized set of Web APIs is crucially important, and the APIs should be formally classified and described by grounding on a shared ontological foundation. However it is obvious that we have to address a number of issues to achieve the goal. Therefore this paper broke down some of the important issues by reviewing the attempts made around the Language Grid project, and showed general steps and presented some detailed proposals, in hope of making some contribution toward the goal. As the issues however may

have substantial depth and stretch, collaborations among international experts, as discussed in (Calzolari and Soria, 2010), are called for. We also argued that user involvements, particularly in a collaborative environment, would be necessary to identify possible language services and resources that are definitely required but remained unaware to the LR/LT experts.

Acknowledgments

The presented work was largely supported by the Strategic Information and Communications R&D Promotion Programme (SCOPE) of the Ministry of Internal Affairs and Communications of Japan. The author would like to thank Toru Ishida, Yohei Murakami, Chiharu Narawa, and other Language Grid members, as well as the international experts in collaboration: Thierry Declerck (DFKI, Germany), Nicoletta Calzolari, Monica Monachini, Claudia Soria (ILC-CNR, Italy), and Paul Buitelaar (DERI, Ireland).

References

Nicoletta Calzolari. 2008. Approaches towards a ' Lexical Web ' : the Role of Interoperability. *Proc. ICGL2008*, pp.34–42.

Nicoletta Calzolari, and Claudia Soria. 2010. Preparing the Field for an Open and Distributed Resource Infrastructure: the Role of the FlaReNet Network. *Proc. LREC2010*.

Thierry Declerck. 2008. A Framework for Standardized Syntactic Annotation. *Proc. LREC2008*.

Gil Francopoulo, Núria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. 2009. Multilingual Resources for NLP in the Lexical Markup Framework (LMF). *Language Resources and Evaluation*, Vol.43, No.1, pp.57–70.

Udo Hahn, Ekaterina Buyko, Rico Landefeld, Matthias Muhlhausen, Michael Poprat, Katrin Tomanek , and Joachim Wermter. 2008. An overview of JCoRe, the JULIE lab UIMA component repository, *Proc. LREC'08 Workshop on Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP*, pp.1–7.

Udo Hahn, Ekaterina Buyko, Katrin Tomanek , Scott Piao, John McNaught, Yoshimasa Tsuruoka, and Sophia Ananiadou. 2007. An annotation type system for a data-driven NLP pipeline. *Proc. the Linguistic Annotation Workshop*, pp.33–40.

Yoshihiko Hayashi, Thierry Declerck, Nicoletta Calzolari, Monica Monachini, Claudia Soria, and Paul Buitelaar. 2011. Language Service Ontology. Toru Ishida (editor). *The Language Grid: Service-Oriented Collective Intelligence for Language Resource Interoperability*, Springer.

Yoshihiko Hayashi. 2011. A Representation Framework for Cross-lingual/Interlingual Lexical Semantic Correspondences. *Proc. IWCS2011*, pp.155–164.

Nancy Ide, and Laurent Romary. 2004. International Standard for a Linguistic Annotation Framework. *Journal of Natural Language Engineering*, Vol.10, No.3–4, pp.211–225.

Toru Ishida. 2006. Language Grid: An Infrastructure for Intercultural Collaboration. *Proc. SAINT-06*, Keynote address, pp.96–100.

Toru Ishida (editor). 2011. *The Language Grid: Service-Oriented Collective Intelligence for Language Resource Interoperability*, Springer.

Kano, Yoshinobu, William A. Baumgartner Jr., Luke McCrohon, Sophia Ananiadou, K. Bretonnel Cohen, Lawrence Hunter and Jun'ichi Tsujii. 2009. U-Compare: share and compare text mining tools with UIMA, *Bioinformatics*, 25(15), pp.1997–1998.

Ewan Klein, and Stephen Potter. 2004. An Ontology for NLP Services. *Proc. LREC2004 Workshop on REgistry of Linguistic Data Categories.*

Urlich Schäfer. 2008. *Integrating Language Processing Components with XML*. VDM Verlag.

Liayng Yu. 2007. Introduction to the Semantic Web and Semantic Web Services. Chapman & Hall/CRC.

Martha Villegas, Núria Bel, Santiago Bel, and Victor Rodríguez. 2010. A Case Study on Interoperability for Language Resources and Applications. *Proc. LREC2010.*

Katy Wolstencroft, Pinar Alper, Duncan Hull, Christopher Wroe, Phillip Lord, Robert Stevens, and Carole Goble. 2007. The myGrid Ontology: Bioinformatics Service Discovery. *International Journal of Bioinformatics Resesearch and Applications*, Vol. 3, No. 3, pp.303–325.