International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:9, No:7, 2015

Secure Bio Semantic Computing Scheme

Hiroshi Yamaguchi, Phillip C.-Y. Sheu, Ryo Fujita, Shigeo Tsujii

***Abstract*—**In this paper, the secure BioSemantic Scheme is presented to bridge biological/biomedical research problems and computational solutions via semantic computing. Due to the diversity of problems in various research fields, the semantic capability description language (SCDL) plays and important role as a common language and generic form for problem formalization. SCDL is expected the essential for future semantic and logical computing in Biosemantic field. We show several example to Biomedical problems in this paper. Moreover, in the coming age of cloud computing, the security problem is considered to be crucial issue and we presented a practical scheme to cope with this problem.

***Keywords*—**Biomedical applications, private information retrieval (PIR), semantic capability description language (SCDL), semantic computing.

I. Introduction

SINCE the development of chain termination of a DNA sequencing method by Sanger and his colleague in 1977 [10] and the subsequent development of computational methods for data retrieval and analysis [11]-[15] bioinformatics has become a new area of research. Many new experimental technologies have been rapidly developed that include: systematic analysis of gene expression profiles at the transcriptional level as well as the translational level using DNA microarrays, 2D protein gel electrophoresis and mass spectroscopy [16], [17]; yeast two –hybrid system for detection of protein-protein interactions; and NMR or X-ray crystallography for the resolution of protein 3D structures. These advances and new technologies has resulted in the rapid accumulation of immense amounts and types of data. These data can be found and data-mined in primary database containing large-scale experimental data such as GenBank and secondary databases providing biology knowledge. As most biomedical database and analysis tools are scattered across different web sites users have to partition their jobs manually into several tasks and do them separately. It is desired that different databases and analysis tools be normalized, integrated and encompassed with a semantic interface such that users of biological data and tools could communicate with the system in natural language and a workflow could be allowed to concentrate on their research and not the job of interfacing disparate systems and data sets. Usability is of importance to the future of bioinformatics tools. In this paper, we describe the BioSemantic System which is a framework that allows heterogeneous tools and data to be integrated via a service-oriented architecture (SOA) for declarative access.

This paper is organized as follows. In Section II, Secure BioSematic System is described. Section III gives an introduction the needs of needs for BioSemantic system. Section IV presents "Using Semantic Capability Description Language (SCDL) for integration tools and data for biomedical applications. In Section V presents conclusion.

Hiroshi Yamaguchi, Ryo Fujita, and Shigeo Tsujii are with the Research and Development Initiative, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo, 112-8551 Japan (e-mail: yamaguchivc@cap.ocn.ne.jp).

Phillip C.-Y. Sheu is with the Department of Electrical Engineering and Computer Science, University of California Irvine.

II. BioSemantic System

The ultimate objective of the BioSemantic System is to provide an integrated framework for prospective users to facilitate their work, such as biological and biomedical knowledge retrieval, management, discovery, capture, sharing delivery and presentation [6]. As illustrated in Fig. 1, the system is able to provide a number of different web services, which can be incrementally plugged into the system. Each service has its own database as well as functions to perform the tasks mentioned above. Accordingly, a common language for these supported service bases to communicate with the system is necessary to formalize and formulate a variety of problems. SCDL is thus proposed to meet this requirement, and will be introduced in more detail in the next section.

The current system relies on [9] as the core technology, which is a development environment that provides an object relational layer on top of relational data sources that could assist designers generate a global schema to capture the semantics of compound objects. Objects are defined within a global schema and wrapped by Java classes. Data are stored in different data sources and manipulated by *semanticObject*™ transparently without depending on further data sources. The global schema is mapped to local data sources by a mapping module. Using the Object Designer, the user can declare object classes as well as define their operations and behaviors. The data associated with actual objects are stored in the data sources. An SNL (Structured Natural Language) Parser is also provided to allow the user to compose their queries in SCDL, using Web Tools. Hence, a solution developed in *semanticObject*™ is extensible and user programmable based SCDL. We envision that the system is being used as follows. User will define the problem by composing an SCDL query or an SCDL program. Users will define the problem by composing and SCDL query or an SCDL program. The SCDL request is parsed into a set of queries in SemanticObjects after service search and service syntheses are done.

A. Semantic Capability Description Language

Semantic Capability Description Language (SCDL) is an SQL-like description language that may be utilized to describe the functionality and capability of a database driven web service, with an objective to support automatic service

composition. The syntax of SCDL for a web service WS is similar to that of SQL, as expressed in the following generic form;

$$SELECT\ outputs(O_1,\ldots,O_m),$$
$$FROM\ inputs(I_1,\ldots,I_m),$$
$$variables(R_1,\ldots,R_n),$$
$$other variables(S_1,\ldots,S_k)$$
$$WHERE\ p(inputs, outputs, other variables)$$
$$GROUPBY\ (H_1,\ldots,H_j)$$

where $O_1,\ldots,O_m$ are output objects; $I_1,\ldots,I_m$ are input objects; $R_1,\ldots,R_n$ are some range variables; $S_1,\ldots,S_k$ are sets that may be derived from the inputs and the range variables; $H_1,\ldots,H_j$ are the variables based on which to group the output objects; and $p(inputs, outputs, other\ variables)$ is a formula that describes the relationships among the inputs, the outputs and the variables. Like SQK-99, SCDL allows variables to be typed, and it allows a function to be included as a condition in the WHERE clause. A major difference between SCDL and SQL is that SCDL allows "exponential variables", where the domain of an exponential variable could be the set of all subsets of an existing set, and variations of exponential variables.

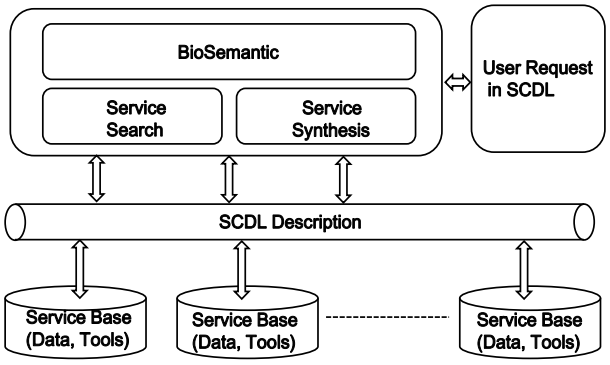

Fig. 1 Architecture of BioSemantic System

III. Needs of Security for BioSemantic System

In the age of Internet accessing remote database is common and information is the most sought after and costliest commodity. In such a situation it is very important not only to protect information but also to protect the identity of the information that a user is interested. Cloud computing environment is constructed by various computing and networking technologies such as virtual machine and distributed computing. Virtual machine can support individual processes depending on the abstraction level, while an important challenge of distributed computing is location transparency. These technologies of virtualization and transparency offer benefits on cost-effectiveness and scalability, while the extensive use of virtualization in implementing cloud infrastructure brings unique security concerns for customers and service providers. This new problem on security issues are concerned not only to software related service, but also to various cloud service functions. Typical examples are Private Information Retrieval (PIR), Encrypted Computing such as Privacy Preserving Data Mining (PPDM). PIR protocol allows a user to retrieve an item from a server in possession of a database without revealing. Which item is retrieved that was introduced in 1995 by Chor, Goldreich, Kushilevitz and Sudan in the information-theoretic setting. Encrypted Computing addresses the challenge to safely outsource data processing onto remote computing resources by protecting programs and data even during processing. This allows users to confidently outsource computation over confidential information independently from the trustworthiness or the security level of the remote delegate. PPDM is a scenario in which two parties owning confidential databases wish to run a data mining algorithm on the union of their databases, without revealing any unnecessary information. This is caused the existence of curious service system administrators who can follow the user's privacy due to the lack of virtual machine and distributing computing technologies. We depict the new security problems in cloud computing environment Fig. 1 which is considered to be applied not only for software related issue but also for logical document, natural language based information processing etc. As depicted in Fig. 1, the user submit a problem (software related issues, logical document, natural language based query) to Cloud Service Provider (CSP) without revealing the problem. And then, CSP executes the required process without recognizing the contents of problem protecting his confidentiality.

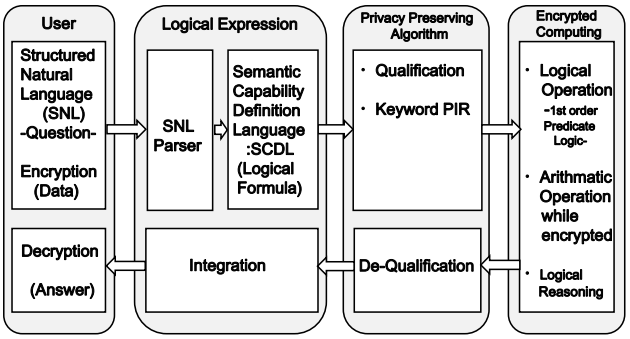

Fig. 2 Workflow of Secure BioSemantic System

A. User

1) Structured Natural Language

Question is described in structured natural language and parsed into a set of SCDL.

2) Encryption

Sensitive data are encrypted and stored in the database prior to send a question by user. In an encrypted computing step, these data are computed while encrypted.

3) Decryption

Result of encrypted computing is decrypted by user and obtained a solution.

B. Logical Expression

1) Parsing

SNL paser is also provided to allow user to compose their queries in SCDL, using Web Tools.

2) Logical Formula

SCDL compose a logic formula.

C. Privacy Preserving Algorithm

1) Qualification

Symbols including Sensitive information are qualified to another ID for the sake of preserving the privacy.

2) Private Information Retrieval (PIR) by Keyword

In the age of Internet accessing remote database is common and information is the most sought after and costliest commodity. In such a situation it is very important not only to protect information but also to protect the identity of the information that a user is interested. Consider the case, when an investor wants to know value of certain stock, but is reductant to reveal the identity of that stock, because it may expose his future intentions with regard to that stock. The scheme or protocol which facilitates a user to access database and receive desired information without exposing the identitiy of information was first introduced by [1] in which the bit-based information of identity. Later the block-based information of identity was presented by [2] and more efficient scheme by [4], [3].

3) De-Qualification

Qualified symbols are de-qualified in the stage of encrypted computing.

D. Encrypted Computing

1) Logical Operation

Logical operation is performed along with logical formula (SCDL).

2) Arithmetic Operation While Encrypted

Arithmetic operation is performed while encrypted. Algorithms for this purpose are presented such as homomorphic encryption scheme [7], secret sharing scheme [8], and three-party secure function evaluation protocol [5].

3) Logical Reasoning

Informally, two kinds of **logical reasoning** can be distinguished in addition to formal deduction: induction and abduction. Given a precondition or *premise*, a conclusion or *logical consequence* and a rule or *material conditional* that implies the conclusion, where the precondition is given.

IV. Using SCDL for Integration Tools and Data for Biomedical Applications

Current bioinformatics tools or databases are very heterogeneous in terms of data formats, database schema and terminologies. Semantic Capability Description Language (SCDL) plays an important role as a common language and generic form for problem formalization. Several queries as well as their corresponding SCDL descriptions are provided as biomedical examples [6].

A. Notations and Definitions

The notations used in the rest of section are listed below:

- A DNA sequences is a string of nucleotide bases q_i, where

$$q_i \in NA = \{A, T, C, G\}, i = 1, \ldots, n; n \in Z^+; \quad (1)$$

- An RNA sequence is a string of nucleotide bases q_i, where

$$q_i \in NB = \{A, U, C, G\}, i = 1, \ldots, n; n \in Z^+; \quad (2)$$

And each element has an attribute called charge and an attribute called molecular_weight;

- A protein sequence is a string sequence of amino acid a_i, where

$$a_i \in AA = \begin{Bmatrix} F, I, C, W, L, P, H, Q, I, M, T, \\ N, K, S, R, V, A, D, E, G \end{Bmatrix}, \quad (3)$$
$$i = 1, \ldots, n; n \in Z^+;$$

- The predicate *blast* A,B) is true if nucleotide_sequence A *blasts* nucleotide_sequence B;
- λ^{NA} designates the set of all possible DNA sequences;
- λ^{NB} designates the set of all possible DNA sequences;
- ξ^{AA} designates the set of all possible protein structures;
- The function (s_u, s_v). Similar ( ) calculates the structural and/or sequence similarity to compare it with a predefined threshold *t*.

An SNL (Structured Natural Language) parser is also provided to allow the user to compose their queries in SCDL, using Web Tools. Hence a solution developed is extensible and user programmable based on SCDL. We envision that the system is being used as follows. Users will define the problem by composing and SCDL query. The SCDL request is parsed into a set of queries after service search and service syntheses are done.

B. Notations and Definitions

Primary sequence analyses of genes and proteins represent a fundamental class of applications that are routinely performed. These analyses depend solely on the underlying nucleic acid sequences for genes, and the amino acid sequence for proteins. These analyses often cover the BLAST, alignment, and prediction of protein demonstrates the wide applicability of SCDL and BioSemantic System to primary sequence, structural analyses and alignment problems.

1. BLAST Problem

Perhaps one of the most common tasks in biological research today is that of identifying genes and proteins related or similar to a particular sequence. We show the example applying our scheme for bio-information data. The user makes a query to our scheme as "Find nucleotide sequences from a database that are similar to a given sequence". This query is described in the form of The SCDL language as follows;

Example 1.

・**SNL**; "Find nucleotide sequences from a database that are similar to a given sequence"

・**SCDL**:

$$\begin{aligned} &SELECT\ N \\ &FROM\ \lambda^{NB}(input)\,s,\ \lambda^{NBA}(input)\,s' \\ &WHERE\ blast(s, s'), \\ &Similar(\ ) \geq t \end{aligned}$$

where s and s' are input protein that is a number of the protein structured database. (s, s').

2. Sequence Alignment Problem

Another common problem is that of aligning multiple sequences of nucleic acids and/or proteins. The object is to identify which regions are conserved and which are different. This problem becomes complicated by the fact that there can be intervening sequences of varying lengths that play little or no functional/structural role. The SCDL describing a representative query for finding subsequence pairs that match with statistical significance is presented below:

Example 2

・**SNL**; Given two nucleotide or amino acid sequences, align based on matching residues and minimizing mis-matches and gaps.

・**SCDL** :

$$\begin{aligned} &SELECT\ (s_u, s_v) \\ &FROM\ \lambda^{NA}(input)\,u \in Q,\ \lambda^{NA}(input)\,v \in Q, \\ &\qquad \omega^u s^u, \omega^v s^v,\ float(input)\,t \\ &WHERE\ u \neq v\ AND\ (s_u, s_v), Match(\ ) \leq t \end{aligned}$$

where u and v belong to the set Q of sequence, s_v is any subsequence that may be derived from v.

Fig. 3 depicts the workflow of evaluating similarity of encrypted nucleotide sequence with "S000387" and "S004082" and evaluation result while encrypted. Nucleotide sequence "S000387", "S004082" and evaluation result are kept secret because each data is encrypted.

A. Predict Protein Families, Domains and Functions

In biological systems the structure of a macromolecule such as a protein determines its function. Much effort has gone into analyses of primary sequences to predict the structure and function of expressed proteins. This includes the prediction of protein family and domain. While it is often the case that similar primary sequences result in similar 3D protein structure, in many cases different primary sequences also can result in similar 3D structures. Many tools and algorithms have been generated to perform these types of analyses and predictions. Following is a representative example.

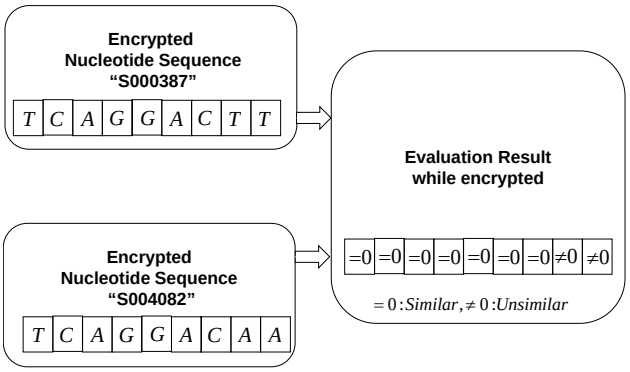

Fig. 3 Private Sequence Evaluation Scheme

Example 3

・**SNL**; Find all proteins from a database that share a structural similarity to a give protein.

・**SCDL** :

$$\begin{aligned} &SELECT\ y \\ &FROM\ \xi^{AA}(input)\,u,\ \xi^{AA}(input) \\ &\qquad v \in db, \omega^{(v)} s_v, \omega^{(u \times db)}(u, v) \\ &\qquad \omega^{(u \times v)} p(s_u, s_v),\ float(input)\,t \end{aligned}$$

where u is an input protein that is a number of the protein structure database db, v is an input protein that is a member of the protein structure database db, s_u is any substructure that may be derived from u *and* s_v is any substructure that may be derived from v. The function (s_u, s_v). Similar calculates the structural similarity to compare it with a predefined threshold t.

V. Conclusion

In this paper, we presented the outline of Biosemantic system and Semantic Capability Definition Language (SCDL) realizing logical formulation. SCDL is expected the essential for future semantic and logical computing in Biosemantic field. We showed several examples to Biomedical problems in this paper. Moreover, in the coming age of cloud computing, the security problem is considered to be crucial issue and we presented a practical scheme to cope with this problem.

Acknowledgment

This study is supported by the Project entitled "Development of Public-key Cryptosystem for confidential communication among Organizations (17201)" of the National Institute of Information and Communications Technology (NICT).

REFERENCES

[1] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, "Private information retrieval," Proc. of 36th FOCS, 1995, pp. 41-50, 1995.
[2] B. Chor and N. Gilboa, "Computationally private information retrieval," Proc. of 29th STOC, 1997, pp. 304-313, 1997.
[3] S. Tsujii, H. Yamaguchi, and T. Morizumi, "Proposal on concept of encryption theory based on logic -Toward realization of confidentiality preserving retrieval and creation of answer by natural language-," SCIS 2013,
[4] S. Tsujii, H. Yamaguchi and T. Morizumi, "Proposal on concept of encrypted state processing at semantic layer -Toward realization of confidentiality preserving retrieval and creation of answer by natural language-", ISEC 2012 Technical Report, 2012.
[5] K. Chida, K. Hamada, D. Ikarashi, and K. Takahashi, "A three-party secure function evaluation with lightweight verifiability revisited," CSS2010, 2010.
[6] S. Wang, R.-M. Hu, H. C. W. Hsiao, D. A. Hecht, K.-L. Ng, R.-M. Chen, P. C.-Y. Sheu, and J. J. P. Tsai, "Using SCDL for integrating tools and data for complex biomedical applications," International Journal of Semantic Computing, 2(2), pp, 291-308, June 2008.
[7] R. Crammer, R. Gennaro, and B. Shoenmakers, "A secure and optimally efficient multi-authority election scheme," Advances in Cryptology -EUROCRYPT'97, LNCS1233. pp. 103-118, 1997.
[8] A. Shamir, "How to share a secret," Communication of the ACM, 22(11), pp. 612-613, 1979.
[9] P. C.-Y. Sheu and A. Kitazawa, "From Semantic Objects to Semantic Software Engineering," International Journal of Semantic Computing, 1(1), pp. 11-28, 2007.
[10] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," Proc. Natl. Acd. Sci. USA, 74(12), pp. 5463-5467, 1977.
[11] P. H. Sellers, "Pattern recognition in genetic sequences," Proc. Natl. Acd. Sci. USA, 76(7), p. 3041, 1979.
[12] A. Sepulveda, M. Pieber, M. A. Soto, and J. C. Toha, "Storage and retrieval of biomolecule sequences," K/ Theor. Biol. 103(2) pp. 331-332, 1983.
[13] H. Peltola, H. Söderlund, and E. Ukkonen, "Algorithms for the search of amino acid patterns in nucleic acid sequences," Nucleic Acids Research, 14(1), pp.99-107, 1986.
[14] P. Gilna, L. J. Tomlinson, and C. Burks, "Submission of nucleotide sequence data to GenBank," Journal of General Microbiology, 135(7), pp. 1779-1786, 1989.
[15] W. R. Pearson and W. Miller, "Dynamic programming algorithms for biological sequence comparison," Methods in Enzymology, 210, pp. 575-601, 1992.
[16] J. Gollub, C. A. Ball, and G. Sherlock, "The Stanford Microarray Database: a user's guide," Methods in Molecular Biology, 338, pp. 191-208, 2006.
[17] O. Langella, M. Zivy, and J. Joets, "The PROTICdb database for 2-DE proteomics," Methods in Molecular Biology, 355, pp. 279-303, 2007.