Vol.2010-BIO-20 No.13
2010/3/5

マルチスケールシミュレーションのための Web サービスとデータ探索に関する研究

木戸善之[†1] 福本貴紀[†2] 野村泰伸[†1,†3]
倉智嘉久[†1,†4] 松田秀雄[†1,†2]

生体機能を解明するためには，ゲノム，化合物，タンパク質，細胞，器官を統合したマルチスケールシミュレーションがフィジオーム・システムバイオロジーとして必要不可欠となる．そのため生理機能を統合的に理解するためにマルチスケールシミュレーションを可能にする基盤が必要となる．一方，マルチスケールシミュレーションを行うためのシミュレーション，データ検索／提供サービスは各スケールで出そろい始めている．しかしこれらは研究者らが研究に必要なデータやサービスを自身で探さなければならない．そこで本研究では，ネームスペースサービスを利用した Web サービスおよびデータの探索システムを設計し関連するディレクトリサービスなどとの比較を行った．

A Study on Discovery of Web Service and Modeling Data for Multi-scale Simulation

Yoshiyuki Kido,[†2] Takanori Fukumoto,[†2]
Taishin Nomura,[†1,†3] Yoshihisa Kurachi[†1,†4]
and Hideo Matsuda [†1,†2]

Multi-scale simulation is mainly aimed at integrating scientific knowledge through computational methods and across different physiological levels; ranging from molecular level, genomic level, cellular level, organ level up to human body part level. Such a framework of modeling has a fundamental role in the process of understanding and clarifying physiological mechanisms. Although several Internet services for multi-scale simulation (such as single level simulations, similarity searchs and data analysis services) has been released, researchers and users have expressed diffuculties in picking up the appropriate and the necessary service and model data for their own researches. Therefore, an urgent need for an infrastructure to enable multiscale simulations is in order. This article suggests an architecture of discovery of services and modeling data for multi-scale simulation using a namespace service.

1. はじめに

次世代シーケンサや高速な計算機，医療技術の発達などにより，生理現象に関する研究が加速しており，計算機クラスタを用いた並列化によって大量データの解析やシミュレーションなどが可能となりつつある．生理現象における研究ではこうした科学技術の進歩によって生体機能の解明が期待されているが，医学，工学，理学，薬学など生理現象に関わる研究ではそれぞれの分野の従来研究が進められており，横断的な学問としての生理学 (フィジオーム・システムバイオロジー)[1)] は未だ黎明期であると言える．生理現象を解明するためには，各分野における個々のシミュレーションでは不十分であり，生理現象をマルチスケールシミュレーションによる全体的なシミュレーションが必要不可欠である．

例えばカルシウム依存性カリウムイオンチャネルは，カルシウムをシグナルとして開閉し細胞内でのカリウムイオンの濃度を調節する．このカリウムチャネルが正常に機能しない場合，心筋における活動電位の調整ができなくなり．心不全などの心臓疾患の由来となりうる．つまりカリウムイオンチャネルを構成するたんぱく質がどのように構造変化するのかはシミュレーションを用い解明することが可能であるが，心筋に与える影響は分子，たんぱく質スケールのシミュレーションでは不十分であり，スケールの推移によって変化する影響を心筋，つまり器官スケールまで統合的にシミュレーションが必要となる．

生理現象のマルチスケールシミュレーションを行なうには研究者らが各スケールの数理モデルデータを組み合わせて構築する．生理モデルデータベースはイオン・分子レベルから器官に至るまで，様々なスケールに応じたモデルデータを集約している．モデルデータは生理現象の数理モデルをノードとしてグラフ構造で表したデータである．図 1 は Hodgkin Huxley

†1 大阪大学臨床医工学融合研究教育センター
Center for Advanced Medical Engineering and Informatics, Osaka University
†2 大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University
†3 大阪大学大学院基礎工学研究科
Graduate School of Engineering Science, Osaka University
†4 大阪大学大学院医学系研究科
Graduate School of Medicine, Osaka University

方程式[2] を記述した細胞の電位活動の数理モデルを表している．例の Hodgkin Huxley 方程式は神経興奮による生体膜電位の電位差を定式化した数理モデルであり，細胞膜にあるイオン・チャネルの開閉が膜電位に与える影響を定量的に示すことができる．このようなモデルデータはモジュールと呼ばれるパーツに分解され，図 1 の様なグラフ構造として可視化する事ができる．図 1 の左側はモデルデータ全体のグラフ構造であり，右側のダイアログはノードの一部分の詳細であり，入力をナトリウムイオン，カリウムイオン，刺激による電流などとし，電圧を出力する関数方程式が記述されている．

一方，近年の Web サービスをベースとしたクラウドコンピューティング，グリッドコンピューティング技術[3] などの発達により広域分散処理が可能になりつつあり，そのため既に作成されたシミュレーションプログラム[4] やデータ加工や類似性検索の Web サービスが公開されている．また NCBI[5] や KEGG[6]，統合データベースプロジェクト[7] など様々な公共機関が提供するライフサイエンスのツールが Web サービスで公開されている．この様なツールは遺伝子配列情報やタンパク質立体構造の類似性検索などがあり，統合的なマルチスケールシミュレーションに有用である．

グラフ構造のモデルデータを複数用い，かつインターネット上に公開されているバイオインフォマティクス・ツールを利用して，マルチスケールシミュレーションのための数理モデルを構築する必要があるが，他スケールの部位がどこにどのような影響を及ぼすのかを研究者らが手作業で網羅的に調べるのは困難である．

そこで本研究では生理現象マルチスケールシミュレーションの基盤構築を目的とし，網羅的なモデルデータおよび Web サービスの検索基盤システムを提案する．具体的にはネームスペースサービスを利用し，Web サービスやモデルデータの情報を木構造データベースによって分類することで，ユーザの Web サービスおよびモデルデータ探索を支援する．

2. 技術要素と関連技術

本節ではネームスペースサービスについての詳細と関連する技術について述べる．

2.1 Resource Namespace Service

Resource Namespace Service (RNS)[8] とは，Open Grid Forum (OGF)[9] で提唱されているグリッドコンピューティング技術の標準仕様である．Web サービスやファイルなどをリソースとして定義し，広域に分散したリソースを一意の名前で解決するためのサービスの仕様である．Domain Name Service (DNS) と類似しており，階層化したデータベースによって，データを階層構造で分類する事が可能である．また RNS は分散データベースの構造を持っており，別の RNS サーバへの参照を持つことで，ある階層以下は別の RNS へ問い合せるといったデータベースの分散化も行うことができる．

グリッドコンピューティングの認証技術である Grid Security Infrastracture (GSI)[10] を利用し，ユーザのデータを他のユーザに公開することなく，分散した別のホストへ認証の委譲を行うことが可能となる．GSI 認証を取り入れる事によってシングル・サインオンが可能となる．またユーザは公開するデータと非公開のデータ，および委譲先の別サービスについても同様のアクセス制御を受けることとなり，より可用性の高い運用が可能となる．また認証はオプションであり，インターネット上に公開する公共のネームスペースサービスとしても運用できる．

RNS に問い合わせることでクライアントは Endpoint Reference (EPR) を得ることができる．EPR は WS-Addressing[11] によって定義された標準化が進められている SOAP メッセージの 1 つであり，回答するサービスとは別のサービスを示すホスト名を記述することが可能となる．また XML (eXtensible Markup Language) で記述されているため，アドレス以外のメタデータも記述することが可能であり，サービスへのパラメータなどをメタデータとして扱うことができる．EPR には Web Service Descrption Launguage (WSDL) で記述した Web サービスの Application Programming Interface (API) 情報を直接記述することができる．クライアントは EPR を受け取ることで，Web サービスの論理名，IP アドレス，プロトコル，パラメータ情報など，Web サービスを利用するために必要な情報を受け取ることとなる．

また RNS の仕様は分散ファイルシステムのネームスペースを定義するために策定されたことから，既存のファイルシステムとの親和性を考慮し，階層を'/' のデリミタで区切り，ファイルパス名として階層を扱うことを定義している (図 2 参照)．RNS の参照実装は，サービス側はグリッドサービスとして実装されており，rns-ls，rns-add，rns-rm といったコマンドによって検索，登録，削除などの基本操作が可能となっている．また図 3 に，RNS への問い合わせ例を示す．

2.2 Lightweight Directory Access Protocol

Lightweight Directory Access Protocol (LDAP)[12] は汎用性の高いディレクトリサービスである．階層構造を持った識別名でデータを管理しており，RNS と同様にデータを階層構造に分類することが可能である．データモデルは各エントリに対して複数の属性をもち，属性は 1 つ以上の値を持つ．そのため 1 つのエントリに対しメタデータを記述することができるため，データに対する外部参照 URL などが記述できる．LDAP は分散データベー

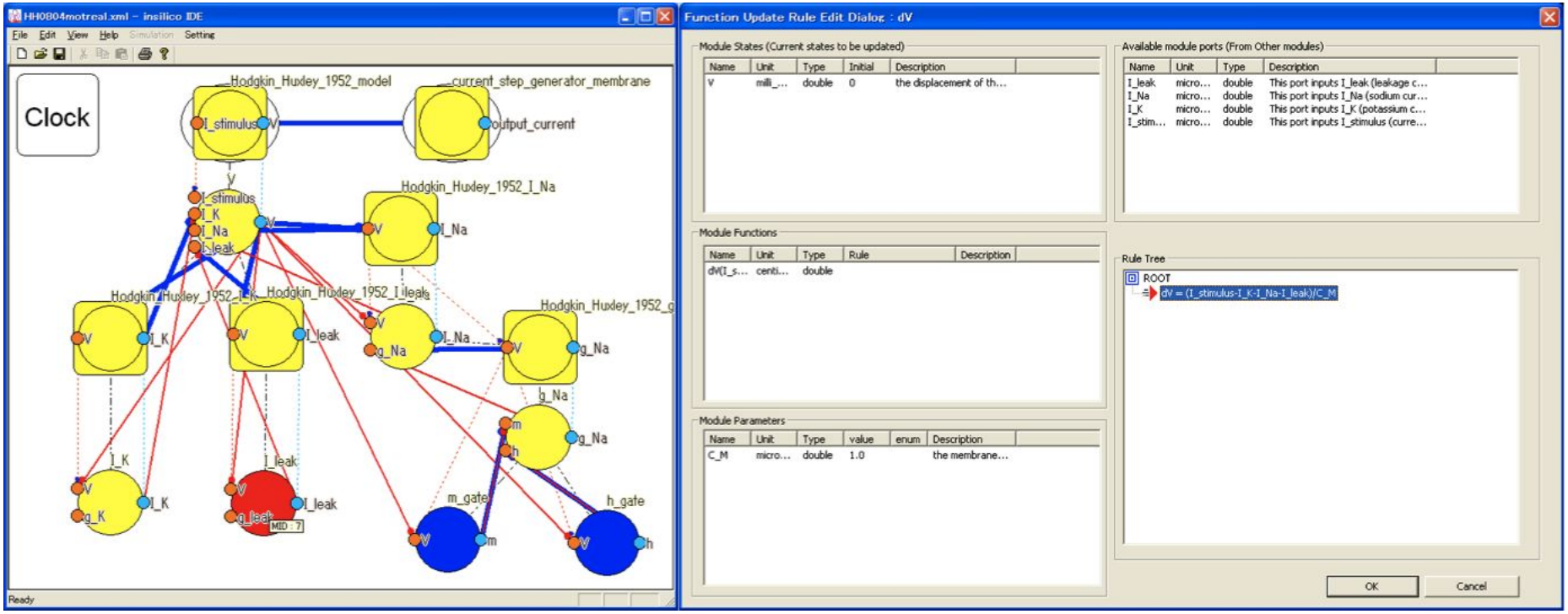

図 1 Hodgkin Huxley 方程式モデルの例
Fig. 1 Model Data of Hodgkin Huxley Equation

スでもありデータベースの分散化を行うことができる．その際の認証は Kerberos を用いることで，分散された LDAP サーバに対してシングル・サインオンが可能となる．LDAP では階層的にデータを管理することができるが，階層についても識別子と言われる属性名と，それに対する値を持つことが義務付けられている (図 4 参照)．LDAP の実装としては OpenLDAP が広く普及しており，クライアント・サーバモデルとして実装されている．基本的な操作は，ldapsearch，ldapadd，ldapdelete などのコマンドにより，エントリの検索，登録，削除が可能となっている．また図 5 に LDAP への問い合わせ例を示す．

2.3 LCG File Catalog

EGEE (Enabling Grid for e-Science)[13] が開発しているグリッドミドルウェア，gLite のコンポーネントであり，これらのミドルウェアは欧州原子核研究機構の大型ハドロン衝突型加速装置から得られる大量の物理実験データを，研究所，大学など様々な機関で共有するためのものである．その中で LCG File Catalog (LFC)[14] は，分散したストレージにまたがってファイルを管理するためのサービスである．ファイルを論理名と物理名で管理し，ユーザは分散管理されたデータファイルを操作する際，ホスト名や，ハッシュコードなどで生成された物理名を意識することなく論理名でファイル操作することが可能となる．gLite では認証基盤として，GSI から拡張した VOMS と言われるミドルウェアを採用しており，仮想組織 (Virtual Organization:VO) 毎にユーザ，ホスト，組織を管理している．LFC ではデータベースの分散化を行う仕様がないため LFC の運用ルールとしては，VO 毎に 1 つ

```
$ ldapsearch -LLL -x -w ******* -D ”cn=admin,dc=example,dc=com” ”cn=*,dc=gfs,dc=ogf,dc=grid”
dn: cn=file1,dc=gfs,dc=ogf,dc=grid
cn: file1
ou: EPR1 <?xml ...
$ ...
```

図 5 LDAP への問い合わせ例
Fig. 5 Example of Query to LDAP

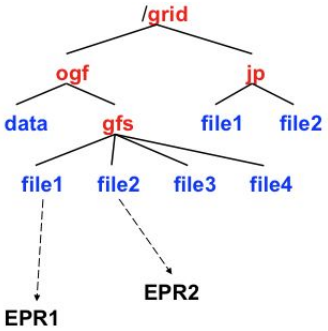

図 2 RNS の階層データ
Fig. 2 Hierarchical namespace on RNS

```
$ rns-ls /grid/ogf/gfs
file1
file2
file3
file4
$
```

図 3 RNS への問い合わせ例
Fig. 3 Example of Query to RNS

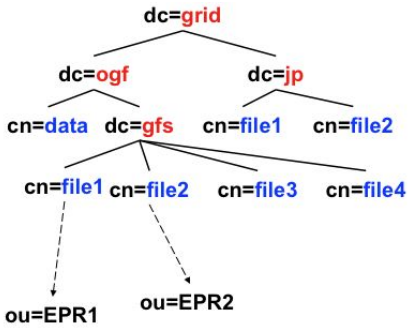

図 4 LDAP の階層データ
Fig. 4 Hierarchical namespace on LDAP

の LFC サーバを設けており，VO をまたいだファイルの共有では LFC のデータベースをミラーリングすることで運用を行っている．図 6 に LFC への問い合わせ例を示す．

3. システム概要と設計

本節では，提案するシステムについて説明する．概要図を図 7 に示す．シミュレーションに必要な Web サービスやデータが探索しやすい様に予め RNS サーバに登録し，ユーザは RNS を検索することで必要な Web サービスやデータを取得，利用する．その際，計算機からもユーザからも可読性が高いオントロジなどを利用する．ライフサイエンス分野では Open Bio Ontology[15] によって様々なオントロジが公開されているため，それらを利用しオントロジに登録されている用語をエントリとして RNS に登録する．オントロジとは抽象的な用語を上位に，具象的な用語を下位にした木構造をとる用語の集合の事を指す．またオントロジにはメタデータを用い意味を付加することが出来るため，RNS にオントロジを取り込む事でユーザは意味を追いながらデータを探索することが可能となる．さらにオントロジによって計算機は演繹的推論が可能となり，ユーザが注目しているモデルデータを元に，結合する新たなモデルデータを探索することが可能となる．

RNS はデータベース分散することが可能なことから，自身の所属機関ネットワーク内に RNS サービスを構築し，公共の RNS をインタネットを経由して参照することが可能である．そのため公共の RNS は共有し，かつ独自のデータを独自の RNS サービスに登録することが可能になり，非公開の Web サービス，データと公開されている Web サービス，データを同様に扱うことができる．この際，GSI 認証を利用することでアクセス制御が可能となり，ユーザ毎に公開，非公開のデータを作ることができる．

Vol.2010-BIO-20 No.13
2010/3/5

```
$ lfc-ls -l /grid/ogf/gfs
drwxr-xr-x 0 root root 0 Nov 12 17:36 file1
-rwxr-xr-x 0 root root 0 Oct 06 02:24 file2
-rwxr-xr-x 0 root root 0 Nov 12 18:22 file3
-rwxr-xr-x 0 root root 0 Nov 06 01:05 file4
$
```

図 6 LFC への問い合わせ例
Fig. 6 Example of Query to LFC

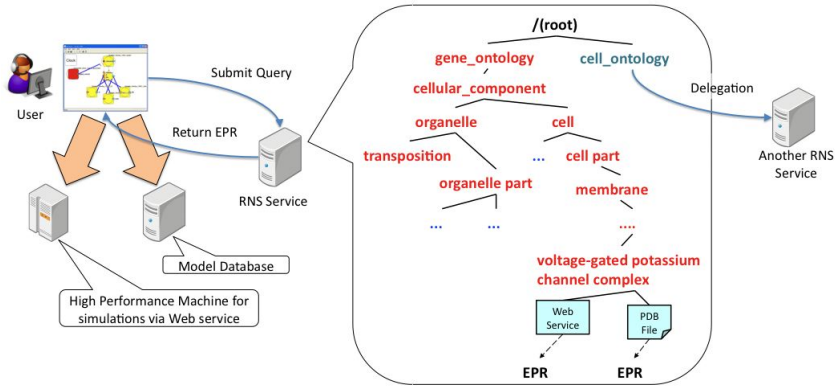

図 7 システム概要図
Fig. 7 System Overview

4. 評価と比較

この節ではシステム設計するにあたって，RNS との関連技術との比較を行った．LDAP の参照実装は OpenLDAP を利用した．また RNS と LDAP のそれぞれの参照実装について性能評価計測を行った．表 1 にその結果を示す．登録については rns-add コマンド，ldapadd コマンドをそれぞれ実行し 1 件の登録時間の平均を取った．検索については rns-ls コマンドと，ldapsearch コマンドを実行し 1000 件の結果が返ってくるまでの平均時間を出した．計測に利用した計算機は同一であり，CPU は Core2Duo E4600 2.40GHz，メインメモリは 2GByte のものを使用した．いずれもクライアントとサーバを同一計算機上で動作させている．

表 1 RNS と LDAP の評価
Table 1 Evaluation of RNS and LDAP

	RNS	OpenLDAP
検索	4.14 秒	1.47 秒
登録	1.59 秒	0.02 秒

表 2 関連技術の比較
Table 2 Comparing Implementations

	RNS	OpenLDAP	LFC
セキュリティ・GSI 認証	○	×	○
セキュリティなし問い合わせ	○	○	×
分散化	○	○	×
EPR の登録	○	△	×
クエリー形式	ファイルパス	識別子	ファイルパス

RNS の参照実装は，サーバ側をグリッドサービス，クライアント側を Java で実装している．またデータベースとしては Apache Derby を利用している．一方，OpenLDAP はクライアントサーバモデルとして実装しており，検索，登録ともに OpenLDAP が速い結果となっている．しかし一方で OpenLDAP では，データ件数が多くなると，検索時間，登録時間ともに増加することが Wang らの報告に示されている[16]．本研究では，Web サービスやモデルデータの探索のためにネームスペースサービスを利用するため，特にモデルデータのエントリは増加の一途を辿る．そのためエントリの増加と比例して検索時間も増加する OpenLDAP は，使用目的とそぐわない．

また表 2 に技術要素についての比較をまとめた．以下の箇条書きに技術要素の各項目について述べる．

- セキュリティ・GSI 認証：GSI 認証はサービスを分散，委譲するためには欠かせないセキュリティ技術である．
- セキュリティなし問い合わせ：公共のネームスペースサービスを考慮すると，認証なしでアクセスが可能であることも必要である．
- 分散化：本研究でのシステムではユーザが複数であることを想定しているため，分散化による負荷集中を避けることも重要である．
- EPR の登録：EPR の登録は Web サービスをエントリとして登録するためには必要である．OpenLDAP は自由に属性を追加することができるため，EPR に記載されてい

る転送先 URL やサービスのパラメータなどが登録可能である．しかし EPR は XML 形式であることから，XML 形式の妥当性を確認する必要がある．EPR 属性の妥当性を検証するすべは OpenLDAP には実装されていない．

- クエリー形式：コマンドラインで操作する場合，RNS，LFC はファイルパス形式であるため，Shell などによるファイル操作と同様の操作が可能となる．一方，OpenLDAP では識別子形式でエントリを検索しなければならない．

以上の項目から，本研究のシステムでは RNS が最適であると言える．

5. おわりに

本研究では生理現象マルチスケールシミュレーションの基盤構築を目的とし，網羅的なモデルデータおよび Web サービスの検索基盤システムを提案し，ネームスペースサービスと関連技術について比較検討を行った．性能評価としては OpenLDAP が RNS を上回る結果を示したが，OpenLDAP のスケーラビリティや機能面から考慮すると，本研究でのシステムでは RNS が優位である．

今後の課題としては，RNS の性能向上，分散化における負荷分散におけるスケジューリング，またエントリ検索におけるオントロジを利用した演繹的推論の実装などがあげられる．

6. 謝　　辞

本研究は文部科学省グローバルＣＯＥプログラム (医・工・情報学融合による予測医学基盤創成) の支援を受けた．また本研究の一部は文部科学省科学技術研究委託事業「研究コミュニティ形成のための資源連携技術に関する研究」，および日本学術振興会科学研究費 (スタートアップ 20800025) により実施したものである．

参 考 文 献

1) Nomura, T.: Challenges of Physiome Projects, *IEEJ Transactions on Electronics, Information and Systems* , Vol.127, Issue 10, pp.1491–1497, (2007).
2) Hodgkin, A. and Huxley, A.: A Quantitative Description of Membrane Current and Its Application to Conduction and Excitation in Nerve, *The Journal of Physiology*, Vol.117, Issue 4, pp.500–544, (1952).
3) Foster, I., Kesselman, C., and Tuecke, S.: The Anatomy of the Grid, *International Journal of Supercomputer Applications* , (2001).
4) Ichikawa, K., Date, S., Krishnan, S., Li, W., Nakata, K., Yonezawa, H., Nakamura, H. and Shimojo, S.: Opal OP: An Extensible Grid-enabling Wrapping Tool for Legacy Applications, *GCA 2007: Proceedings of the 3rd International Workshop on Grid Computing* , pp.117–127, (2007).
5) National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov
6) Kanehisa, M. and Goto, S.: KEGG:Kyoto Encyclopedia of Genes and Genomes, *Nucleic Acids Research* , Vol.28, No.1, pp.27–30, (2000).
7) 統合データベースプロジェクト, http://lifesciencedb.jp
8) Pereira, M., Tatebe, O., Luan, L. and Anderson, T.: Resource Namespace Service Specification, *Global Grid Forum 17, Grid File System Workshop Document* , (2006).
9) The Open Grid Forum, http://www.ogf.org
10) Foster, I. and Kesselman, C.: Globus: A Metacomputing Infrastructure Toolkit, *International Jounal of Supercomputer Applications* , Vol.11, No.2, pp.115–128, (1997).
11) Gudgin, M., Hadley, M. and Rogers, T.: Web Services Addressing 1.0 – Core, *W3C Recommendation* , (2006).
12) Wahl, M., Howes, T. and Kille, S.: Lightweight Directory Access Protocol (v3), *RFC 2251* , (1997).
13) The Enabling Grids for E-Science, http://www.eu-egee.org
14) Baud, J.-P., Casey, J., Lemaitre, S. and Nicholson, C.: Performance analysis of a file catalog for the LHC computing grid, *HPDC 2005: Proceedings of the 14th IEEE International Symposium on High Performance Distributed Computing* , pp.91–99, (2005).
15) The Open Bio Ontology, http://www.bioontology.org
16) Wang, X., Schulzrinne, H., Kandlur, D., Verma, D.: Measurement and Analysis of LDAP Performance. *Conference on Measurement and Modeling of Computer Systems*, ACM, pp.156–165, (2000).