A Framework for Molecular Biology Database Integration Using Context Graph

Nawaz Khan
School of Computing Science
Middlesex University, UK.
n.x.khan@mdx.ac.uk

Shahedur Rahman
School of Computing Science
Middlesex University, UK.
s.rahman@mdx.ac.uk

A. G. Stockman
Queen Mary College
University of London, UK
tonys@dcs.qmul.ac.uk

Abstract

This paper proposes a framework based on the context of the web data for interactions with different biological data resources so that a consolidated view of data can be achieved. A formal description of context of resources and their relationships have been described here using context graph. The context increases the interoperability by providing the description of the resources and the navigation plan for accessing the web based databases. A higher level construct is developed to implement the context in RDF for web interactions. The interaction among the resources is achieved by using a context based integration domain. The integration domain allows to navigate and execute the query plan within the resource databases.

1. Introduction

Most of the database integration research are based on query optimisation or query script writing [3], [4], [9]. Other researchers have suggested metadata based integration of molecular biology databases [1], [3], [8]. They have used metadata to store database structural description for schema comparison. This is a widely used approach for schema transformation to create datawarehouse, e.g., OPM model [1]. The drawbacks of datawarehouses are highlighted in papers [5], [6] and [7].

Martin [10] in a review article on 'Trends in Biotechnology' emphasised the need for web semantics. He proposed to use Resource Document Framework (RDF) [12] for resource description. However, the RDF model needs to be further extended to higher-level constructs to deal with biological web complexity. For example the same data may exist at different web under different context or it may not follow the traditional approach for relational algebraic operation. In this context if the higher level constructs can include the context of the web resources then it will increase the semantics of the webs. This approach will have the following unique features:

(i) resource context descriptions - implementation of these contexts into RDF will be able to increase the interoperability by providing the description of the resources and the navigation plans for accessing the web based databases.

(ii) design flexibility for navigational plans – this will enable to choose the participating resources and its query need.

The proposed approach for using context in integration domain along with navigational plan for heterogeneous multiple biological resource integration is novel. The web based approach provides an alternative to the use of generic schema for database integration as proposed in [6], [9] and [11].

Proceedings of the 17th IEEE Symposium on Computer-Based Medical Systems (CBMS'04)
1063-7125/04 $ 20.00 © 2004 IEEE

2. Resource Description and Navigation

The following sections look into the structure of resource description and its representation. The sections also highlight the novel concept of integration and search initiation.

2.1 Context Graph for Resource Mapping

A graph *G* is defined with three interrelated subsets as: *G=(S, E, L),* where *S* denotes an object in resource, i.e., page or any particular content, *E* defines the edges and *L* defines the element of an image object. *S* can be expressed as $\{u_1(v_1)\ldots u_n(v_n)\}$ where *u* denotes the resource name or ID and *v* denotes the operator. *u(v)* denotes an object *v* of a particular resource *u*. If any node is linked with another node, then it can be expressed with edge *E* as $E \subseteq$ u x u; where each $e \in$ E is represented as *e= u1.u2* if *e* is an edge linking resource u_1 and u_2. *L* denotes a labelled image object element with a list of values. The context graph also describes the entry point to nodes. In order to access an entry point node directly, an operator needs to have a constant value. Figure 1 shows the proposed context graph model for PDB, GDB and OMIM. This context graph model describes how the PDB, GDB and OMIM objects are related to each other and how they provide access to other target pages either by means of node name (direct linking) or by using search forms. The diagram in Figure 1 shows only those portions of the schema which are related to the integration part or which are related to share a common view for integration.

Context graph interpretation for resource mapping: Each interpretation in context graph *I* defines a mapping *M.* A set of values for particular mapping *M* is assumed to have a set of values called *V*. The map requires to contain triple *has h, provide p* and *access a*. An interpretation of *I* for mapping *M* is defined by a nonempty set *R* of resources, called the domain of *I* and superset of value *V*. Resource access *a* points to the set of resources, *R1, R2..Rn*, if any value *x* are in *R1, R2..Rn* where *I(x)* identifies arguments for which the resources are true. A resource *R* is composed of a set of elements *h* where $h=\{e_1, e_2, \ldots e_n\}$. A resource provides a set of values *p* where *I(p)= true*. The mapping is illustrated in Figure 2. The mapping is for the *resource {r1:h, r1:p, r1:a}* where *<{r1:h, r1:p, r1:a}>= true* if any value *x* in *r1* for which interpretation *I(x)* is in resource *{r2:h, r2:p, r2:a}* and *<{r2:h, r2:p, r2:a}>=true* if and only if any value *y* in *r2* for which interpretation *I(y)* is in resource *{rn:h, rn:p, rn:a}* and *<{rn:h, rn:p, rn:a }*>=true. In such a case the map denotes all the objects in the resources.

Context representation in RDF: A context represented in RDF as a statement set with additional structural and logical properties is explained as follows:

rdfc:context is a subclass of *rdfc:StatementSet*, and it represents a context. By inheritance this consists of a set of statements. Context implements a set of statements which provide values, properties and resources. A context can be a set of contexts for any integration domain.

rdftype:has indicates values that is a member of a context, and which is also asserted to be true for a particular resource. This corresponds to the values that the resource contains.

rdftype:provide indicates properties to show the particular reason for accessing the resource. This is true for one particular resource in one context, but this can also be true for any other resources in another context.

rdfctype:access are the remote or local URI targets which have access to any particular element.

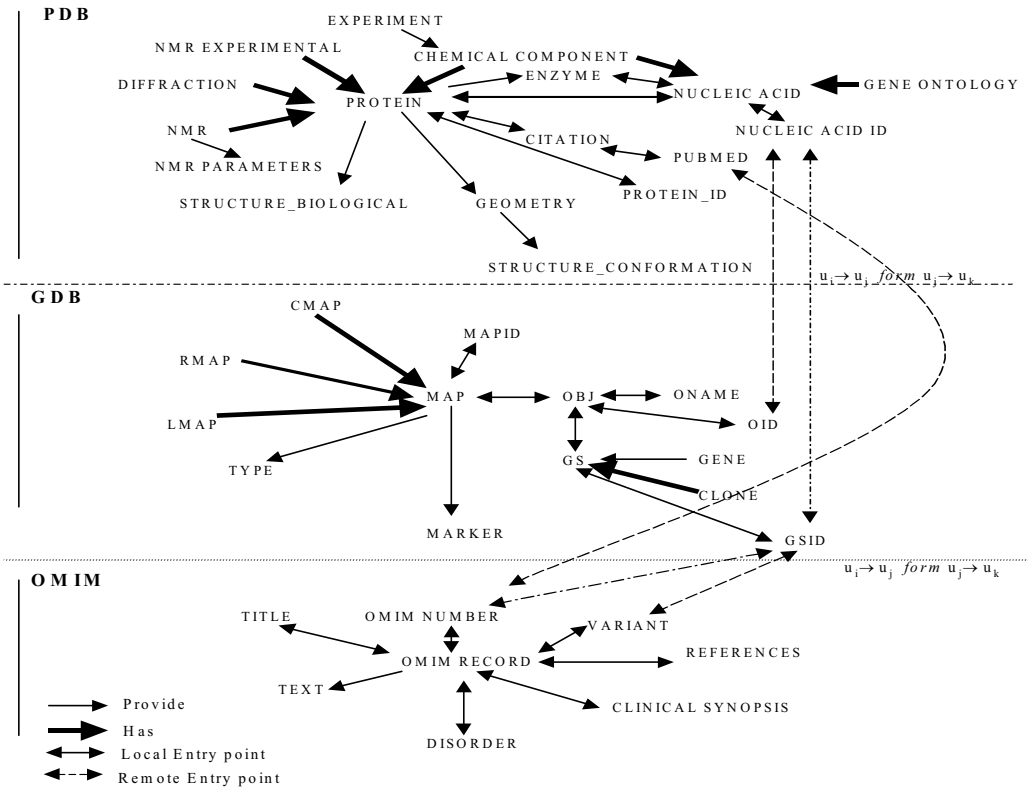

Figure 1. Context graph for web contents and links

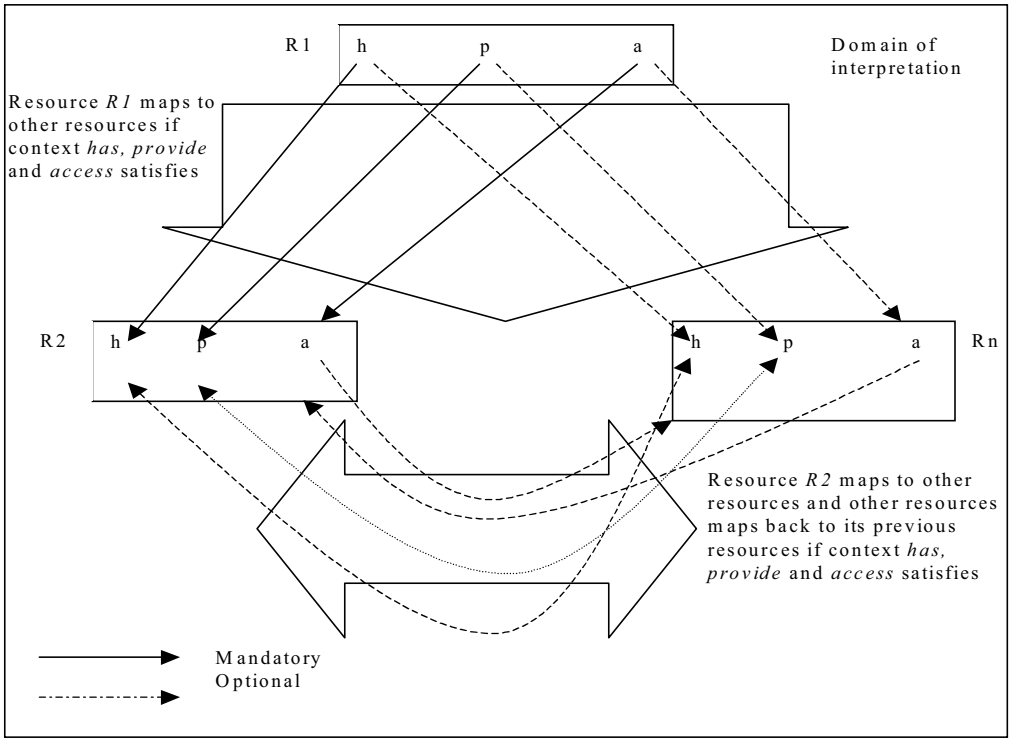

Figure 2. Resource mapping

2.2 Integration Domain Using Context

A context is a collection of attributes of any resource where the resource is described in terms of its values, objects that it is providing and connection to other resources. The set of expressed contexts for each resource is integrated by creating a unifying context for all the contexts. This unifying context is used as the domain or range of integration. It leads to all the resources which are associated to each other and which represent a collection of statements describing the objects present within it. This also allows any context to hold another context without knowing the detail physical structure. This provides a modular approach for describing any high-level relationships among the components. The following example shows how to integrate PDB, GDB and OMIM.

```
[IntegrationDomain]----rdf:type → [rdfc:type]
{
[ProteinDataBank] --- provide → [ProteinID]
[ProteinDataBank] --- provide → [NucleicAcidID]
            {
            [ProteinID] ----------has → "value"
            [NucleicAcidID] ----------has → "value"
            }
      [GenomeDataBank] ---provide → [GenomeID]
            {
            [GenomeID] ---------has → "value"
            }
      [OmimRecord] ----provide→ [OmimNumber]
            {
            [OmimNumber] ---- has → "value"
            }
[ProteinDataBank]-- access → [GenomeDataBank]
[GenomeDataBank] --- access → [OmimRecord]
[ProteinDataBank] --- access → [OmimRecord]
}
```

2.3. Dispatcher to Initiate Search

A search initiation mechanism is described in Figure 3. It is a bottom up approach to receive elements from each node and to pass through to the next node to receive more elements from the nodes. Finally, when all the elements are collected from the target nodes then these elements are combined into one object.

A *Dispatcher* submits operators to the individual resource database to establish the link. For the search mechanism hyperlink is carried out by the Dispatch operator [4], i.e, GenomeID, NucleicAcidID or ProteinID. Basic functions of this search initiator, Dispatcher, are as follows: *i.* split the search operators; *ii*. allocate them to multiple bioinformatics sources; *iii.* determine each search operator as a sub-plan of the total output; and *iv*. create dynamic memory to hold the subset of the output for further integration. To link the relevant bioinformatics resources, it is necessary to link to the requested page *p* directly in order to optimise the hyperlink data searching. For example, if data set *d* is distributed over a number of biological resources, D_p, D_g *and* D_o, then to retrieve *d*, the following steps are performed: *i*. access to the D_p, D_g and D_o resources*; ii.* retrieve the required pages p_p, p_g, p_o from D_p, D_g *and* D_o; *iii*. access the required set of elements e_p, e_g and e_o from the pages p_p, p_g, p_o respectively, and *iv*. then pass the elements of the pages to the result integrator to embed the elements into a single page *p.* The steps are shown in Figure 4. The overall objective of the dispatcher is to apply a set of search operators *O* to the respective data resource *R* as described in mapping linker and let $O_i(R_i)$ ($1=<i=<n$, where *n* is a finite value) be a set of derived facts related to the overall search result.

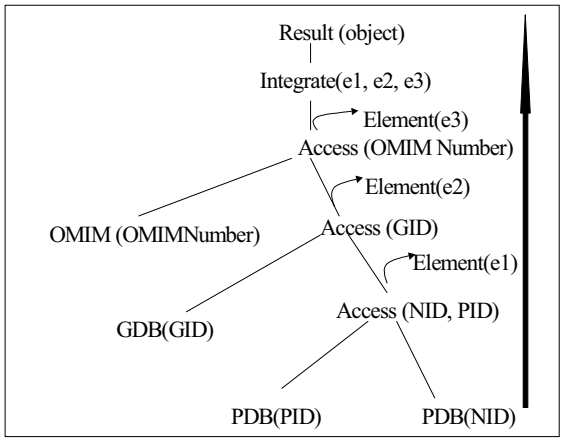

Figure 3. Search initiation mechanism

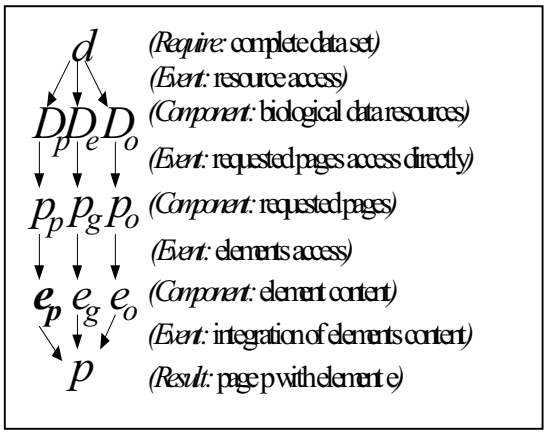

Figure 4. Dispatching and collecting elements

3. An Example of Integrating Biological Data Sources

An example to find information on Alzheimer disease is shown here to illustrate the process. The *Dispatcher* aims to dispatch the searching operators individually to the respective data resources for unifying them into a single page. Individual resource along with their operators (collected from RDF) are sent to the respective databases to find the information on Alzheimer's disease. DOM [13] interfaces are used for the resources and operators to traverse along the contents.

Figure 5 shows the final combined result in HTML, e.g., mutation rate, likelihood in male and female, phenotypic details, diagnosis environment, pathological lesion details and coding region. The combined result has collected these elements information from different biological resources. Table 1 shows the sources of these elements. All the elements which are collected from different resources are embedded in a single page, called 'Trait Analysis' (Figure 5).

Trait Analysis
Mutation Rate: Peak lod score with Gm 1.37
Likelihood in male and female: Theta =0.05

Phynotypic Details: Protease Inhibitor Domain Of Alzheimer'S Amyloid beta-Protein Precursor (APPI) - Chain A 1AAP:A : VREVCSEQAE TGPCRAMISR WYFDVTEGKC APFFYGGCGG NRNNFDTEEY GGGGS SSS EEE EEEETTTTEE EEEEE SSS S BSSHHH CMAVCGSA HHHHH Protease Inhibitor Domain Of Alzheimer'S Amyloid beta-Protein Precursor (APPI) - Chain B 1AAP:B: VREVCSEQAE TGPCRAMISR WYFDVTEGKC APFFYGGCGG NRNNFDTEEY GGGGS SS EEE EEEETTTTEE EEEEE SSS S BSSHHH CMAVCGSA HHHHT
Diagnosis Environment:
CSF analysis of hyperphosphorylated tau protein. (Phosphorylation at serine 199 tau-199) for the antemortem diagnosis of AD
Genetic Event Type: Mutation in APP (A-beta amyloid precursor protein gene) gene. Autosomal dominat trait in families
Abnormality Frequency:7 of 21 families. 3 generation transmission, Familial incident 43%
Sex: Male
Ethnicity: Afro-British
Region: South Africa
No. of case Studies:001

Pathological lesion
Details: Alzheimer's disease/cerebral hamorrhage, Dimential, Trisomy 21, Parkinsonism, long tract signs. Overall Mean onset Age 45.7 and overall mean age of death 53.3. 20% cases
Coding Region: Point mutation; From GCA to GGA. Observed in g.275267>g and p.A692G. Region is Ex17 and N-term

Figure 5. Element collection and integration for Alzheimer disease

Table 1. Elements Collected from Different Sources

Elements	Resources
Mutation rate	Online Mendalian Inheritance in Men
Likelihood in male and female	Online Mendalian Inheritance in Men
Phenotytypic details	Protein Data Bank
Diagnosis Environment	Online Mendalian Inheritance in Men
Pathological lesion	Online Mendalian Inheritance in Men
Coding Region	Genome Data Bank
Genetic Event Type	Genome Data Bank
Sex	Local database
Ethnicity	Local database
No. of cases	Local database

4. Conclusion

Genetic variance analysis can not be accomplished as a standalone process, instead, it requires to combine information from other data resources so that a meaningful insight of gene mechanism can be derived or hypothesised. This entails developing a framework and creating a cooperative environment so that readily available data, i.e. web data, can be combined and correlated. This paper has highlighted some of these important issues, namely importance of web semantics and resource mapping. The paper has also suggested a novel approach to describe the participating resources using a context graph. Higher level constructs of RDF are used to describe the context of the web and to formulate an integration domain for navigation within the databases.

5. References

[1]Chen, I.A., Markowitz, V.M. An Overview of the Object-Protocol Model (OPM) and OPM Data Management Tools. Information Systems. Information Systems, 20, 1995. pp393--418.

[2]Freidman, M., Levy, A and Millstein, T. Navigational Plans for Data Integration. In proceedings of the Sixteenth National Conference on Artificial Intelligence and Eleventh Conference on Innovative Application of Artificial Intelligence, AAAI-Press , 1999. pp 67-73

[3]Kemp, G.J.L., Angelopulos, N. and Gray P.M.D. A Schema-Based Approach to Building a Bioinformatics Database Federation. Proceedings IEEE International Symposium on Bioinformatics and Biomedical Engineering. IEEE Computer Society Press, 2000, pp 13-20.

[4]Kemper, A and Wiesner, C. Hyper Queries: Dynamic Distributed Query Processing on the internet, Proc. of VLDB Conference. 2001, Pp 551-560.

[5]Khan, N., Rahman, S and Clarkson, G. T.; An approach to develop human gene disorder database for intelligent variance analysis of genes and its products. 12th International Conference on Database and Expert System (DEXA, 2001). Proc. IEEE Computer society press. 2001. Pp 301-305

[6] Khan, N and Rahman, S; A conceptual object modelling of gene mutation data. German Conference of Bioinformatics, GCB01. Germany, Proc. 2001 Pp. 187-190

[7]Khan, N., Stockman, A.G. and Rahman, S. A Cooperative Environment for Genetic Variance Analysis Using Component Database for Database Integration. 15th IEEE Conference on Medical Based Systems (CMBS, 2002). Proc. IEEE Computer Society Press. 2002. Pp 365-368

[8]Kohler, J, Lange, M, Hofestadt, R and Schulze-Kremer, S, Logical and Semantic Database Integration. Proc. 2nd IEEE Symposium on Bioinformatics and Bioengineering. IEEE Press, 2000. pp 77-80.

[9]Markowitz, V.M. Heterogeneous Molecular Biology Database, Vol 2, no: 4, 1995, Journal of Computational Biology, 1995. pp: 537-538.

[10] Martin, C.R. A Can we integrate bioinformatics data on the internet? Meeting report, Trends in Biotechnology, Elsevier Science Press, Vol. 19, No. 9, 2001. pp. 327-328.

[11] Schonbach, C., Kowalski-Saunders. P. and Brusic, V.; "Data warehousing in molecular biology", Briefings in Bioinformatics, vol: 1, no: 2, 2000. pp: 190-198.

[12] http://www.w3.org Resource Description Framework (RDF) Model and Syntax Specification.

[13] http://www.w3.org/TR/REC-DOM-Level-1, DOM Specification Level 1