Virgil: A Databank of Links between GDB and GenBank

Frédéric Achard Frederic.Achard@infobiogen.fr
Emmanuel Barillot Emmanuel.Barillot@infobiogen.fr
Gis Infobiogen
7 rue Guy Môquet, BP 8, 94801 Villejuif Cedex, France

Abstract

This paper focuses on a specific type of information frequently used by researchers in Genetics: links between genome objects. It emphasizes the fact that, at present, links are not sufficiently characterized and describes our work to address this problem: the design of a prototype databank to store links between genome databases. Because this global repository is of concern for many people, we welcome and encourage feedback from the community.

1 Introduction

Links between genome databases are becoming a major concern for the genome community. Researchers in biology are unsatisfied with the current situation regarding this type of data. Jean Weissenbach, head of the French national sequencing center (CNS) says:

> "Despite the ever-increasing avalanche of data, we are still frustrated because not all of the information is accessible. For instance, links between different sets of data are difficult to retrieve or are missing from databases."[14]

At the 1997 Pacific Symposium on Biocomputing, Thure Etzold views the databanks of genome cross–references as a mean for facilitating data access:

> "Databanks of links are presently scarce but they will become more abundant once their importance and usefulness is better understood."[8]

In the Ecocyc database, Peter Karp and colleagues incorporated "heavyweight links". In addition to the source and the target, a link contains the link type, the author, the date and method of creation. This work is described in [12], along with a discussion on links and how to use them in the context of molecular biology databases.

Clearly, the genome community needs links of better quality: meaning links that are consistent, available, documented and maintained. We think that a specialized database will serve better for this purpose than leaving this task to database managers. A simple and strong argument is the consistency: only one instance of a link has to be managed, not two or more.

2 Interoperating genome databases

2.1 Present situation

Interoperation allows a collection of distributed programs and distributed data to cooperate. The genome databases are not yet interconnected in a satisfactory way either for programmers or for users. At most they offer reliable hypertext links to other databases but rarely permit programmed data gathering. In this case, links provide a convenient way to overcome the dissemination of the

information. If links are properly defined, a user may go from the localization of a gene in GDB [10] to either data in the GDB "mini-federation" such as the literature citation, or data that exist in remote databases such as nucleic acid sequences in GenBank [5] or the associated clinical syndromes in OMIM [13].

2.2 Enhancing the documentation of links

Following the explosion of the Web, numerous links have been created which are, for a large part, poorly documented and insufficiently maintained (one could argue that the creation of a link into the Web is too easy). More attention is given to the data that exist in the major databases. Maintainers of these databases are making a great effort to cross-reference genomic data and provide links of good quality. In spite of this, the links suffer from lack of characterization, at least for those publicly available. Information on the links themselves is, for the most part, either nonexistent or non accessible.

From a user's perspective, the quality of a link is essential (remarks to moderate this statement are given below). Too often, especially on the Web, a link yields unwanted information. The relevant data needs to be sorted out. For example, the following situations frequently occur when traversing a link:

- The link is incorrect. It frequently happens when links are automatically generated by softwares (false–positive links).
- The link is out of date. Sometimes the link has been out for years and knowledge has advanced since its creation. This happens frequently for genome information because of the progress in biology. Either changes in one of the two objects or in their relationship can make the link obsolete. In this case, the lexical interconnection which defined the link is no longer valid.
- One of the linked objects has been moved to another location or the method of access has been modified. Every Net-surfer has experienced http errors, such as "`404 Not Found. The requested URL /xxx was not found on this server`". In this case, the physical interconnection which defined the link is broken, at least temporarily.
- The link points to data which are clearly related, though unrelated to the user's viewpoint. At the meta–data level, terms are understood by users with regard to their experience, culture or motives. To assess the quality of a link, one should be aware of the semantics that has been used to create the link, if any. For example, cross-references from a gene would point to different data whether the link is defined by a physician, a geneticist or a molecular biologist. Clearly, a lexical interconnection (made out of terms), if solely used to define a link, would be insufficient.

The quality of a link is a highly subjective notion. It depends both on the information sought and the expectation of the seeker, as shown in the previous section. This is why it is best to focus on the means to dispense links that are documented. The main advantage is that one can estimate the effect of following a link. If a link is provided along with some material to 'subjectively', that is, depending on an individual, pre-assess the quality of a link, it will dramatically reduce the number of times a wrong link is hit.

2.3 Enhancing the distribution and management of links

The links need to be managed. Indeed, better link characterization means that one has to manage more complicated data for every link. Maintaining up to date data for each linked objects addresses non trivial issues [11]. Moreover, keeping pace with rapid evolution in biological knowledge, one has to change the links concurrently. Obviously the more data the genome research generates, the more data there will be to link.

It is a fact that the links are getting more complicated, more unstable and more numerous. Likewise, these data need to be populated. To face this situation, links data need to be more prominent. This will induce a two-fold advance for genome research. First the navigation in the genomic information hyperspace will be more rewarding. Second, the links are the basis for an advanced interoperation between genome databases.

The links need to be available for the entire scientific community. Any user or program or database should be able to retrieve a link.

3 Documented links

3.1 Object identifiers

For each link, one has to deal with three different biological objects: the link itself and the two biological objects which are linked. In the context of a world wide effort for genome research, it is crucial that objects can be addressed in an unequivocal way. The task to assign one unique identifier to each biological object can be delegated to a central server that delivers them on request. The benefits are that every object would be registered, with, possibly, a creator name on any object. On the down side, it would be difficult to set up such a system. A function that delivers an identifier guaranteed to be unique is rather straightforward to implement. The delivery is done with no prior knowledge of any already existing identifier but it does require that all databases use the same function. Such a function could be implemented with the concatenation of (i) a network address and (ii) the results of a function never returning twice the same value —any strictly increasing function will do. The Distributed Computing Environment standard also defines specifications to generate such identifiers (`http://www.osf.org/dce`).

Since a consensus should be obtained before starting either of the two previous schema, a third alternative was preferred. Note that this scheme was proposed by Ken Fasman [9] and is already in use at the GDB. Each object is identified by a "local" unique identifier, prefixed with a mnemonic for the database. It leaves the attribution of unique identifiers to the responsibility of each database manager. For example, `gbk:M13902` will refer to the GenBank entry which has the accession number `M13902`. The same mechanism is used to unequivocally identify each link.

3.2 Origin of the links

All the major database maintainers are now convinced of the pressing need to provide links that point to external databases. Although there is an obvious lack of coordination between those efforts, the good news is that the quantity and the quality of the available links are increasing. As part of the proposal for creating a federation of public biological databases, the GDB provides dependable link objects that comes with an unique identifier.

At the same time, information retrieval systems such as genXref [2] or ENTREZ [15] focus on the means to automatically generate cross-references.

Individuals doing research in their field of expertise are constantly relating different pieces of information. The possibility to store and document these data within a database of links allows the sharing of such individual expertise.

3.3 Characterization of a link

In the scope of biological studies, a link is simply an interconnection between two biological objects. Without any further indication, it is sometimes difficult to predict what is hidden behind a link. The goal of a proper characterization is to define the nature of the link and to provide some idea of its quality.

type: describes the kind of data that are linked, i.e. the meaning of the link. This piece of information is specially important when one of the databases contains several types of object. This is the case of the GDB: one has to know before following a link if it points to a gene, a probe or any other object that might exist in GDB.

author: is the name of the individual, the organization or the program that can be credited for the link. Note that there can be any number of authors for one particular link. It is likely that quality will be enhanced if someone has to sign for the creation of a link. In the long term, some authors may be recognized for their expertise in link creation.

belief value: is a normalized value (between 0 and 1) to rate the quality of the link. Obviously, the rating is author dependent. For example, an author can assign a belief value of `1` to guarantee a link, or `undef` if the rate is unknown. Likewise, the rating can be automatically generated with systems such as genXref.

dates: a number of dates are useful to characterize the pieces of information that constitute a link (see Section 4.2)

data related by the link: give some information on the two objects that are linked. The point of providing such data is to allow a user to pre-assess the quality of a link, with the subjectivity of his/her judgment. Moreover, it can also prevent the user from actually fetching the linked remote objects in case the information provided is sufficient.

An example is given in Section 4.2 to illustrate the characterization of a link.

4 Virgil, a test case

4.1 A databank of links

Not only is there an urgent need for documented links, but there is also an urgent need for those links to be managed and easily available available to the scientific community. Virgil is a prototype bank of links which aim is to address those issues. The main goal of Virgil will be to manage cross-references between objects from genome databases, i.e., to receive submissions and to fetch, store, update, annotate and distribute links. For the prototype stage of Virgil, we focus on providing links between GDB and GenBank. Indeed, these databases are of primary importance for genome research. In their day-to-day work, researchers in genetics often need to gather information from both databases.

At present, Virgil stores links between GDB gene objects and GenBank human sequences. It is arranged as a flat file library of 23,769 entries.

A set of 9,458 links was extracted from the GDB itself. The data were originally generated by heuristic matching on locus name. Data also come from direct author submission or third party annotations.

Another set of 17,395 links was automatically created with genXref [2]: a system to infer links between independent genome objects by term signature comparison.

We estimated the quality of the data contained in Virgil, the results are displayed in figure 1. A link is judged non relevant (false–positive) if there are evidences to prove that the gene (a GDB object) does not physically map to the sequence (a GenBank entry). Note that it is a severe restriction. For example, a link between a gene and its pseudogene is considered as a false–positive link.

4.2 A Virgil entry

With references to the broad ideas developed in Section 3, we show as an example a link entry between the human gene Involucrin as found in the GDB and its exon 1 sequence as found in GenBank. This Virgil entry is displayed in Figure 2.

Figure 1: Precision estimation (proportion of relevant links)

Origins of the links	Precision	Standard error
`Genome DataBase`	95 %	±4 %
`genXref`	83 %	±6 %

Figure 2: Virgil entry which describes a link between the human gene Involucrin as found in the GDB and its exon 1 sequence as found in GenBank.

```
LID    vgl:19465; gdb:119355 * gbk:M13903; 01-AUG-1996
TYP    vgl:19465; GDB[gene] * Genbank[sequence]
OR1    genXref; v1.0; 0.71; 01-AUG-1996
OR2    GDB; v6.1; undef; 25-MAY-1995
xxx
IDa    gdb:119355; IVL; 25-MAY-1995
DFa    involucrin
xxx
IDb    gbk:M13903; HUMINV2; 06-JAN-1995
DFb    Human involucrin gene, exon 2.
//
```

The Link IDentifier (`LID` field) is `vgl:19465`, where `vgl` stands for Virgil. The entry describes the link between a GDB object and a Genbank object and was created on the `01-AUG-1996`. The type of the link is given in the `TYP` field. This entry has two known origins. The first is described in the `OR1` field: the author is `genXref, v1.0`; the belief value is `0.71`; and the date is `01-AUG-1996`. The second is described in the `OR2` field: the author is `GDB, v6.1` and the belief value is unknown, although it may be subsequently filled in by the author.

Next, relevant data on the objects `a` and `b` are given. They come, as is, from the original databases. Each object is described within two fields. The field `IDa` (resp. `IDb`) contains the database identifier, name and date for the object a (resp. b). The field `DFa` (resp. `DFb`) contains the definition for the object a (resp. b).

4.3 Populating Virgil data

In the first stage, we deliberately format Virgil on a simple data model: a structured flat file library. The goal is to test whether the approaches we choose are reasonable. There are currently three means to access Virgil (see Virgil's Web page for details: `http://www.infobiogen.fr/services/virgil/Home.html`)

- The library can be ftp'ed as a whole. It allows the possibility for any site to integrate these data into a customized environment.
- Virgil is incorporated in SRS [7]; see Section 5 for a short SRS description.
- A dedicated Web interface to access Virgil is being developed.

Virgil was easy to build and populate because of its format which allows a good readability. On the other hand, the basic operations of data management need to be developed from scratch. At the moment, the addition and the deletion of entries are supported. The short-term goal is to port Virgil

to database management system that provides facilities for the following operations: merging two link objects, forking a link object into two objects, modifying a link object (e.g., for annotation, update) and checking data integrity for each of these operations. We are currently in the process of porting Virgil data to an object oriented database: EYEDB. This database management system is being developed by Sysra Informatique in close collaboration with Infobiogen. Its interface is compliant with the ODMG–93 standards [6].

5 Related work, discussion

There are a few works that make extensive use of cross-references to build complex information systems dedicated to genome data. The getDB system [3] achieves integration of information via linkDB, a bank of the links explicitly specified within any of the sixteen molecular databases that compose getDB. Similarly, SRS creates a virtual federation of genome databases. A language allows one to describe the structure of a flat file library and to define means to extract links between libraries. The program processes indices to allow navigation through all the libraries. A limitation of the SRS system is that it applies only to flat file libraries, not relational or object oriented systems. ENTREZ is an information retrieval system that integrates a subset of MEDLINE along with publicly available nucleotide and protein databases. The system pre-computes similarity between entries, using different methods such as BLAST [4] for sequence similarity or document neighboring [15] for text information.

The links that exist within these systems are, in the present state, difficult to export because they are dedicated to unique systems. Moreover, they do not have the minimum documentation as discussed in Section 3. Virgil, however, was created to be an open bank. Eventually, we mean Virgil to become a central repository of cross-references between genome databases. Individual users or organizations will be able to both submit links to the database and to query any information on cross-references. Systems such as getDB could hand over the burden needed to manage the links. On one hand, they will pass along their cross-reference data to Virgil. On the other hand, they could retrieve from Virgil any of the submitted cross-references plus other cross-references of interest that have been submitted by third parties.

Our middle-term projects is to facilitate link distribution via an object request broker (ORB). CORBA is a new standard that holds great promise for facilitating distributed heterogeneous application development and system integration [1]. Virgil addresses the problem of semantic interoperation, relying on the data provided by the community. With CORBA, our aim is to address the other aspects of interoperation: namely, the physical interconnection and the syntaxic interconnection. Indeed, it offers a standardized IDL (Interface Definition Language) to describe the means of access for distributed data. An ORB server transparently locates and delivers remote objects. The progress of genome research will no doubt be enhanced if the users have the possibility to leave aside problems due to disparate query and access mechanisms and can concentrate on the conceptual data integration.

Acknowledgments

We are thankful to Philippe Dessen and Guy Vaysseix for numerous discussions; Claude Scarpelli and Philippe Gesnouin for informatics support; and Jennifer Fitzpatrick and Miroslav Hill for critical reading of the manuscript.

References

[1] Frédéric Achard and Emmanuel Barillot. "Ubiquitous distributed objects with CORBA." In Russ Altman, Keith Dunker, Lawrence Hunter, and Teri Klein, editors, *Pacific Symposium on Biocomputing '97*, pages 39–50. World Scientific, 1997.

[2] Frédéric Achard and Philippe Dessen. "Automatic generation of links between heterogeneous genomic databases." In *International IEEE Symposium on Intelligence in Neural and Biological Systems*, pages 78–83, 1995.

[3] Yutaka Akiyama, Susumu Goto, Ikuo Uchiyama, and Minoru Kanehisa. "Linkdb: A database of cross links between molecular biology databases." In *Second Meeting on the Interconnection of Molecular Biology Databases*, 1995.

`http://www.ai.sri.com/people/pkarp/mimbd/95/abstracts.html`.

[4] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. "Basic local alignment search tool." *Journal of Molecular Biology*, 215:403–410, 1990.

[5] Dennis A. Benson, Mark Boguski, David J. Lipman, and James Ostell. "Genbank." *Nucleic Acids Research*, 24(1):1–5, 1996.

[6] R. G. G. Cattell. *The Object Database Standard : ODMG–93.* Morgan Kaufman, 1996.

[7] Thure Etzold and Patrick Argos. "SRS - an indexing and retrieval tool for flat file data libraries." *CABIOS*, 9(1):49–57, 1993.

[8] Thure Etzold and G. Verde. "Using view for retrieving data from extremely heterogeneous databanks." In Russ Altman, Keith Dunker, Lawrence Hunter, and Teri Klein, editors, *Pacific Symposium on Biocomputing '97*, 134–141, 1997.

[9] Kenneth H. Fasman. "Restructuring the genome database: A model for a federation of biological databases." *Journal of Computational Biology*, 1(2):165–171, 1994.

[10] Kenneth H. Fasman, Stanley I. Letovsky, Robert W. Cottingham, and David T. Kingsbury. "Improvements to the GDB Human Genome Data Base." *Nucleic Acids Research*, 24(1):57–63, 1996.

[11] Peter D. Karp. "Models of identifiers." In *Second Meeting on the Interconnection of Molecular Biology Databases*, 1995. `http://www.ai.sri.com/people/pkarp/mimbd/95/abstracts.html`.

[12] Peter D. Karp. "Database links are a foundation for interoperability." *Trends in Biotechnology*, 14:273–279, 1996.

[13] Victor A. McKusick. *Mendelian Inheritance in Man.* Baltimore: Johns Hopkins University Press, 1994. `http://www3.ncbi.nlm.nih.gov/omim/`.

[14] Jean Weissenbach. "Landing on the genome" (in editorial). *Science*, 479, October 1996.

[15] W. John Wilbur and Leona Coffee. "The effectiveness of document neighboring in search enhancement." *Journal of the American Society of Information Science*, 43(5):358–370, 1992.