AliasServer: a web server to handle multiple aliases used to refer to proteins

Florian Iragne, Aurélien Barré, Nicolas Goffard and Antoine de Daruvar∗

Centre de Bioinformatique Bordeaux, Université V. Segalen Bordeaux 2, 146 rue Léo Saignat, 33076 Bordeaux Cedex, France

Received on November 10, 2003; revised on February 2, 2004; accepted on February 5, 2004
Advance Access publication April 1, 2004

ABSTRACT

Summary: AliasServer provides services that facilitate the assembly of data or datasets that make use of different identifiers for referring to the same protein. This resource relies on a database which contains, for a given organism, a non-redundant list of protein sequences associated with a set of aliases.

Availability: AliasServer is available as an interactive Web server at http://cbi.labri.fr/outil/alias/ and as a web service using a SOAP interface. The complete tool, including sources and data, is available for local installations upon request.

Contact: antoine.daruvar@pmtg.u-bordeaux2.fr

Supplementary information: Technical documentation is available at http://cbi.labri.fr/outil/alias/asdoc.pdf

INTRODUCTION

The variety of identifiers used to refer to the same biological object, such as a protein, is a recurrent source of problems. When different aliases are used for the same protein, bringing together data or datasets requires a substantial amount of time and effort. Aliases have various origins: systematic names defined by sequencing projects, accession numbers used by public database, protein names used by the biologists, etc. While the designation of a universally accepted unique identifier would theoretically solve the problem, this has absolutely no chance of happening given the multiplication and growth of molecular biology databases. Therefore, solutions must be proposed to help in the handling of the diversity of aliases.

In this respect, species oriented databases like SGD (Cherry et al., 1998) for Saccharomyces cerevisiae, which collect and centralize aliases, are very helpful. However, the help provided by such resources is limited for different reasons: they are specialized to a given species; they are not designed to easily incorporate new alias definitions; and they do not offer, to users and programs, dedicated ‘alias management services’ such as the translation of identifiers from one type to another, like systematic names into accession numbers.

∗To whom correspondence should be addressed.

We have developed AliasServer in an attempt to overcome these limitations. Our system is designed to load and store aliases attached to the proteomes of an extendable list of organisms and to provide services through both an interactive web interface and a SOAP-based application programming interface (API) (Box et al., 2000, http://www.w3.org/TR/SOAP/).

PRINCIPLES

For each protein of a given organism, AliasServer uses a unique reference identifier: a 64-bit key, which is computed from the amino acid sequence using the CRC algorithm (Press et al., 1992). For large proteomes (i.e. human) which contain up to 100 000 proteins, the probability of having the same key for different sequences is <10−9 and thus can be neglected. All aliases found for a protein are then attached to this reference.

Most aliases are values of a particular field (i.e. accession number, name, etc.) of a given database [e.g. SWISSPROT (Boeckmann et al., 2003), PIR (Wu et al., 2003), etc.]. We consider that each combination of field and database defines a type of alias. These types are independent name spaces with possible collisions, i.e. P25114 is an accession number in both SWISSPROT and PIR databases but corresponds to different proteins. Consequently, each alias in AliasServer is a unique combination of a species identifier (i.e. Rattus norvegicus), a type (i.e. SWISSPROT accession number) and a character string (i.e. P25114).

AliasServer uses a relational database, hosted on a PostgreSQL server, to store reference identifiers, aliases and related data. Input data must be submitted in a defined XML format which is described in the technical documentation.

Loading aliases in AliasServer

There are two steps when loading proteome data for a given organism in AliasServer.

The first step initializes the database of reference CRC identifiers, and requires as input an exhaustive and non-redundant set of protein sequences. This set can be obtained from any
reference database [like SGD for *S.cerevisiae* or any GenBank complete genome entry (Benson et al., 2003)] and contains a first list of aliases for each sequence.

The second step enriches the database with new aliases. These aliases may have any origin: public databases, literature, privately used identifiers, etc. For each sequence, a list of aliases is submitted. One of the aliases from this list must already exist in AliasServer in order to be able to map those aliases to a reference identifier. When this mapping fails, the corresponding aliases are kept in the log section of the AliasServer database. These unmapped aliases will be automatically submitted again after any further addition of new aliases in an iterative process.

During the loading of aliases, AliasServer detects ambiguities: this is typically the case when a list of aliases are mapped to different reference identifiers. Such ambiguities cannot be resolved by the system automatically and are stored in a log file.

**SERVICES**

AliasServer provides a configurable query interface, accessible either through the web interface or its API, using SOAP requests.

Typically, an AliasServer query aims at retrieving aliases corresponding to a list of one or several identifiers. Various options can then be specified to refine the query: restrict the search to a particular organism, restrict the search to a particular alias type, only retrieve aliases corresponding to a given type, etc.

**An example of use of AliasServer**

A biologist wants to study the domain composition of yeast proteins that are involved in protein–protein interactions. For this purpose, the biologist decides to combine interaction data retrieved from MIPS (Mewes et al., 2002), where proteins are identified by their systematic name, with data from BIND (Bader et al., 2003), where NCBI ‘GenInfo Identifiers’ (GI) are used as identifiers. For the domain composition of the proteins, the biologist choose the PRODOM database (Servant et al., 2002), which uses SWISSPROT and TREMBL references.

In order to combine these data, the biologist needs to convert all protein identifiers in to a unique and unambiguous name space. The biologist can easily do this by writing a simple script which submits SOAP queries to AliasServer.

**CONCLUSIONS**

Many biologists and bioinformaticians have faced or will face problems related to the multiplicity of aliases used to identify biological sequences. Most probably, tens of thousands of data files and scripts dealing with lists of aliases have been created worldwide for various organisms. Those individual efforts are largely redundant and do not result in a shared resource.

AliasServer aims at providing a rational means for building a collective resource by providing online services for the management of aliases and a standard interface to load new data. At the time of writing, 14 proteomes were already available in AliasServer, including *S.cerevisiae* with 6273 reference sequences and 70 945 aliases, and *Homo sapiens* with 36 771 reference sequences and 188 436 aliases.

**REFERENCES**


