



AliasServer

AliasServer : a service to find your way in the maze of biological sequence aliases

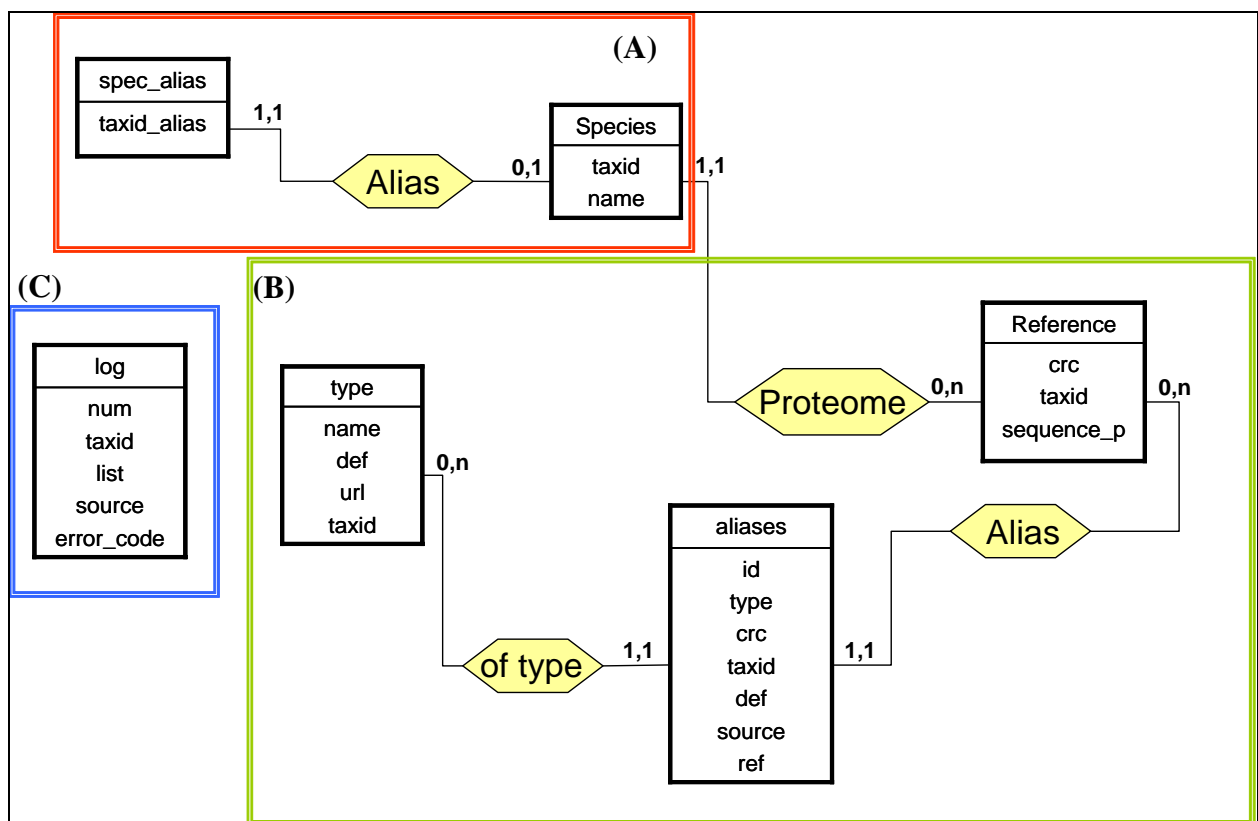
The variety of identification used to refer to the same biological object, such as gene or protein, is a recurrent and annoying problem. When different aliases are used, bringing together data or data sets often requires substantial amount of time and efforts. Those aliases have various origins: systematic naming defined in the frame of sequencing projects, public databases identifiers or accession numbers, gene or protein names used by the biologists, etc. If the designation of a unique identifier universally accepted and used 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 for the handling of aliases diversity. That's the aim of AliasServer.

Reference and aliases

For each sequence of a given organism proteome, AliasServer uses the CRC keys, computed from the amino acid sequence, as the unique reference identifier.

Aliases are derived from different sources which are independent name spaces with possible collisions: i.e. P25114 is an accession number in both SWISSPROT and PIR databases but correspond to different proteins. Consequently, each alias in AliasServer is a unique combination of identifier (i.e. P25114), type (i.e. SWISSPROT accession number) associated to a reference in a given species.

AliasServer Database



The database is hosted on a PostgreSQL server. We can distinguish three parts in this database schema :

- A. Species are managed in the red part. One table stored information about species; the other allows declaring species aliases (useful when databank don't use the same taxonomic number for same species).
- B. References and aliases are managed in the green part. This component is species oriented, allowing to store the same reference for different species.
- C. Ambiguities are stored in a log table (blue part) for further manually or automatic curation.

Data insertion

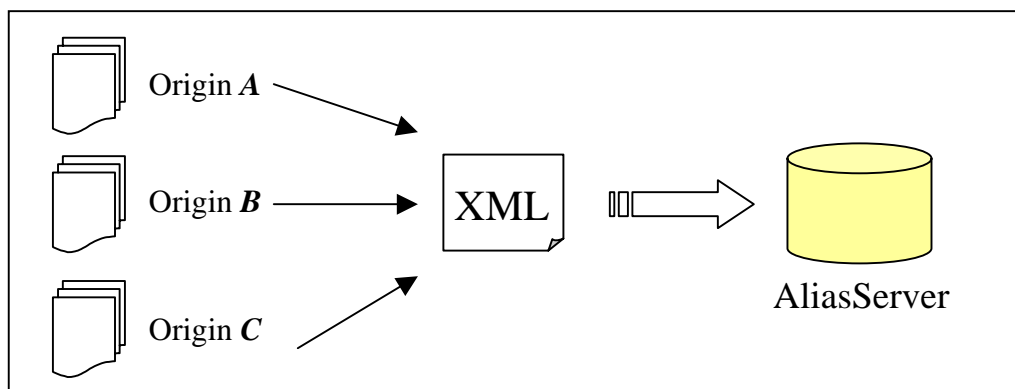
There's two step when loading proteome data for a given organism in AliasServer :

1. Creation of references, and first level of alias, for this organism using an exhaustive set of proteins. These references (their exhaustivity and quality) will be very important for the next step when aliases have to be mapped on them.
2. Enrichment of database with new aliases corresponding to references. A list of aliases corresponding to a protein is submitted and have to map to a reference (one of the aliases submitted should be already present in the database).

Ambiguities appear when a list of submitted aliases can not be mapped to a reference, or when this list is mapped to more than one reference. The first case is stored to be resubmitted automatically in a recursive process when new aliases will be added (adding new aliases can resolve this type of ambiguity). The second case can not be automatically resolved and need a manual curation.

Input format

Input data containing aliases can have various origins (general databank like GenBank, or species oriented databank like SGD for *S. cerevisiae*). These data must be converted into a defined XML format to be inserted in the database.



DTD :

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<! ELEMENT list(key*)>
<! ATTLIST list taxid>
<! ATTLIST list source>
<! ELEMENT key(alias*)>
<! ATTLIST key seq?>
<! ATTLIST key id?>
<! ELEMENT alias(#PCDATA)>
<! ATTLIST alias type>
<! ATTLIST alias source?>
<! ATTLIST alias def?>
```

The value of the attribute *source* in *list* tag (here GenBank) and in *alias* tag (SwissProt for the alias P59641 of the third key) can be different. In this case, his value in the *alias* tag has the priority.

In *key* tag, *seq* attribute contain the peptide sequence of the protein while *id* attribute contain the CRC key computed from the sequence.

Attributes of *key* tag are optional. In consequence, we can define two XML format, corresponding to the two step of data insertion, from this DTD :

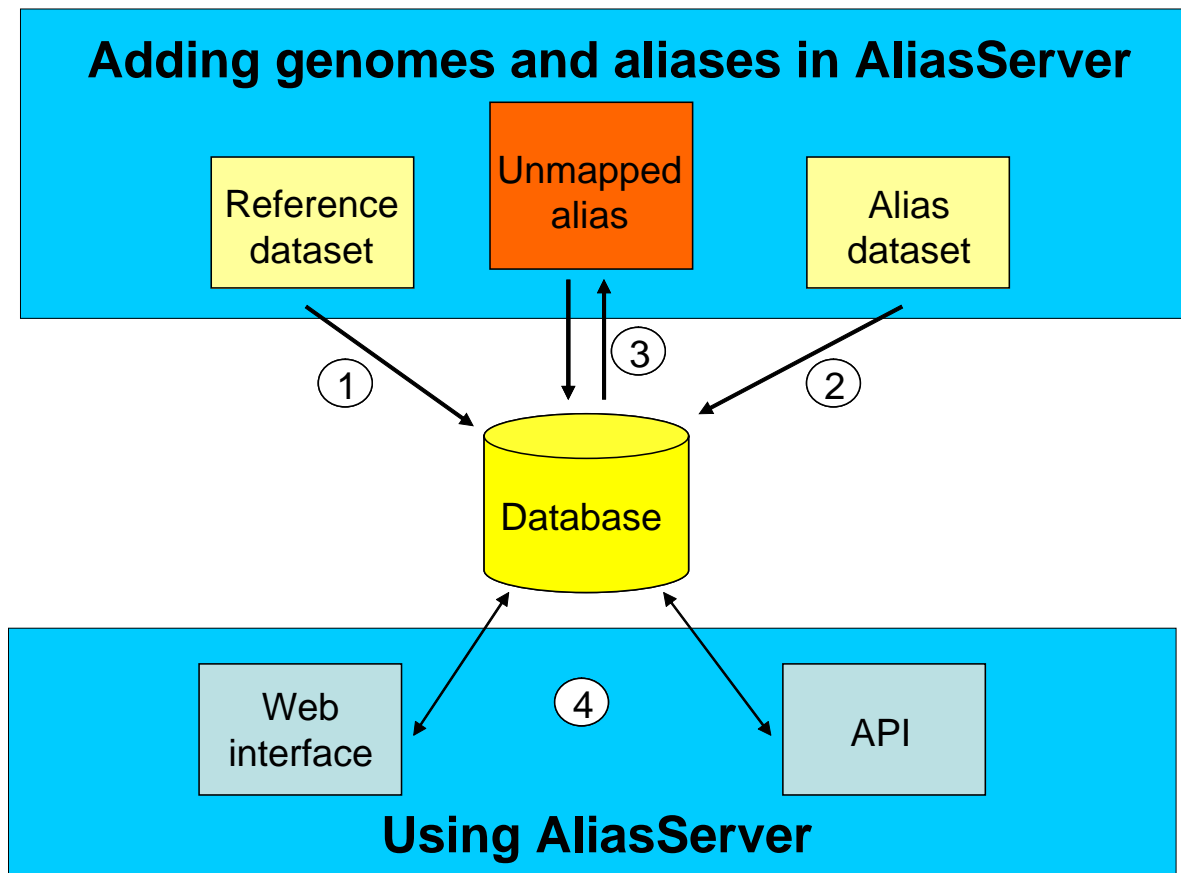
- If these attributes are not empty, the *key* tag is used to create references and *alias* tag corresponds to the first level of aliases. This XML of type *reference* is used in the first insertion step.
- But if these attributes are empty, the XML file is only used to add aliases to existing references of the organism. This XML of type *alias* is used in the second insertion step.

TYPE REFERENCE	<pre><?xml version="1.0" encoding="ISO-8859-1"?> <list taxid="4932" source="GenBank"> <key id="CRC" seq="XXXXXXXXXX"> <alias type="GENE_NAME" source="GenBank" def="def"> TPD3 </alias> <alias type="SWISSPROT_AC" source="SwissProt" def="def"> 1P31383 </alias> <alias type="SYSTEMATIC" source="GenBank" def="def"> YAL016W </alias> </key> </list></pre>
TYPE ALIAS	<pre><?xml version="1.0" encoding="ISO-8859-1"?> <list taxid="4932" source="GenBank"> <key id="" seq=""> <alias type="GENE_NAME" source="GenBank" def="def"> TPD3 </alias> <alias type="SWISSPROT_AC" source="SwissProt" def="def"> 1P31383 </alias> <alias type="SYSTEMATIC" source="GenBank" def="def"> YAL016W </alias> </key> </list></pre>

The species TaxID must be declared in the species part of the database.

All *type* value in XML *alias* tag must be declared in the *type* table of the database to avoid orthographic errors.

AliasServer framework



- 1) **Initialization** : a dataset is submitted which contains the complete list of references identifiers and a first set of aliases
- 2) **Enrichment** : new aliases are submitted to be mapped on reference identifiers
- 3) **Unmapped aliases** are automatically submitted again iteratively
- 4) **Querying AliasServer** *via* web or API interface

AliasServer web form

The image shows the AliasServer Query Form interface. At the top left is the CBiB logo (Centre de Bioinformatique de Bordeaux). The main title is 'AliasServer Query Form' in a yellow box. The form contains several input fields and options:

- (1) A text input field labeled 'Enter your list of identifier' with a 'CLEAR' button.
- (2) A text input field labeled 'Or give a path to a file' with a 'Browse' button.
- (3) A dropdown menu labeled 'Search on:' with 'All species' selected.
- (4) Two dropdown menus labeled 'Which identifier type submit?' and 'Which identifier type in result?'. Both show a list of identifier types: CROCH, GENBANK_LAC, GENE_NAME, GI, HGNC_ID, PIR_ID, REFSEQ, SGD_ID, and SGD_ID2.
- (5) A checkbox labeled 'Only retrieve one synonym'.
- A 'SUBMIT' button at the bottom.
- At the bottom left, there is a link 'Or submit a proteic sequence' with a 'SEQUENCE' button below it.

Fig 1 : AliasServer query form for identifier submission

- 1) List of submitted identifiers (separated by '@') keyboarded in the form, or imported as a flat file (in same format)
- 2) Restrict search to one or all species
- 3) Type of identifier submitted
- 4) Type of identifier asked in response
- 5) Option to retrieve only one alias per type asked in response. The chosen alias is the first found for each type.

Submit a sequence

(1) Paste one proteic sequence here :

(2) Which identifier type in result ?

- ALL
- CROCH
- GENBANK_AC
- GENE_NAME
- GI
- HGNC_ID
- PIR_ID
- PEPSEQ
- SGO_ID
- SGO_ID2

(3) Which species ?

All species

SUBMIT

Fig 2 : AliasServer query form for sequence submission

If no identifier is available to query AliasServer, a peptide sequence can be used.

- 1) Peptide sequence from which a CRC will be computed, and submitted to AliasServer
- 2) Type of identifier asked in response
- 3) Restrict search to one or all species

The screenshot shows the AliasServer response interface. At the top, there is a blue box labeled "Result(s)". Below it, a search bar contains the identifiers "dnaa" and "gyrB" (marked with a red (1)). To the right, there is a "Save selected results" button (marked with a red (2)). Below the search bar, the results for "dnaa" are displayed, including the organism name "Mycoplasma genitalium : dnaa" (marked with a red (3)) and a CRC key "CRC64 = D4DF45A0F4768ED6" (marked with a red (4)). A red arrow points from the "Save selected results" button to a small icon in the results section. Below this, a table lists various aliases for "dnaa" (marked with a red (5)), including "dnaA", "DHAA", "12045709", "NP_073143.1", "P35888", "Q49343", and "DHAA_MYCGE". Each row in the table includes a "GeneBank" link (marked with a red (6)) and a "Web link" icon.

Alias	Gene Name	Description	GeneBank	Web Link
dnaA	GENE_NAME	chromosomal replication initiator protein dnaA	GeneBank	
DHAA	GENE_NAME	Chromosomal replication initiator protein dnaA	GeneBank	(6)
12045709	GI	chromosomal replication initiator protein dnaA	GeneBank	
NP_073143.1	REFSEQ	chromosomal replication initiator protein dnaA Mycoplasma genitalium	GeneBank	
P35888	SWISSPROT_AC	Chromosomal replication initiator protein dnaA	GeneBank	
Q49343	SWISSPROT_AC	Chromosomal replication initiator protein dnaA	GeneBank	
DHAA_MYCGE	SWISSPROT_ID	Chromosomal replication initiator protein dnaA	GeneBank	

Fig 3 : AliasServer response interface

- (1) Submitted identifiers (dnaA and gyrB in this exemple)
- (2) Save chosen results card (using checkbox) in FASTA format
- (3) Result card for identifier 'dnaa' in M. genitalium
- (4) Save a result card in FASTA or tabular format
- (4) Reference found for this result card (CRC key)
- (5) Aliases list ordered by type
- (6) Web link

AliasServer SOAP

getAliases options

List are separated by ','. A list is composed of 1 or more elements

MANDATORY

-id List of identifier to search

OPTIONAL

-species Taxid list for organisms targeted

-out Output file

-ti Identifier type list corresponding to identifier list

-to List of identifier type to return

-one Return only one value for each identifier type asked [T/F]

-format Output format (txt, tab, list or fasta / default=txt)

-all Results only if each identifier match for each species and each asked type (-out). Disable if no output type is given. [T/F]

SPECIES LIST :

TAXID NAME

-> 4932 Saccharomyces cerevisiae
-> 2097 Mycoplasma genitalium
-> 2104 Mycoplasma pneumoniae
-> 2107 Mycoplasma pulmonis
-> 28227 Mycoplasma penetrans
-> 2130 Ureaplasma urealyticum
-> 9606 Homo sapiens
-> 224308 Bacillus subtilis subsp. subtilis str. 168
-> 71421 Haemophilus influenzae Rd
-> 6239 Caenorhabditis elegans
-> 233150 Mycoplasma gallisepticum str. R
-> 83333 Escherichia coli
-> 7227 Drosophila melanogaster
-> 4896 Schizosaccharomyces pombe

TYPE LIST :

-> CRC64
-> GENE_NAME
-> GI
-> HGNC_ID
-> PIR_ID
-> REFSEQ
-> SGD_ID
-> SGD_ID2
-> SWISSPROT_AC
-> SWISSPROT_ID
-> SYSTEMATIC
-> TREMBL_AC
-> TREMBL_ID

**Dynamically
asked to database**