BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 19 2008, pages 2272–2273
doi:10.1093/bioinformatics/btn424

Databases and ontologies

The Synergizer service for translating gene, protein and other biological identifiers

Gabriel F. Berriz[1] and Frederick P. Roth[1,2,*]

[1]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 250 Longwood Avenue and [2]Center for Cancer Systems Biology, Dana-Faber Cancer Institute, Boston, MA 02115, USA

Received on March 16, 2007; revised on June 12, 2008; accepted on August 8, 2008

Advance Access publication August 12, 2008

Associate Editor: John Quackenbush

ABSTRACT

Summary: The Synergizer is a database and web service that provides translations of biological database identifiers. It is accessible both programmatically and interactively.

Availability: The Synergizer is freely available to all users interactively via a web application (http://llama.med.harvard.edu/synergizer/translate) and programmatically via a web service. Clients implementing the Synergizer application programming interface (API) are also freely available. Please visit http://llama.med.harvard.edu/synergizer/doc for details.

Contact: fritz_roth@hms.harvard.edu

With the wealth of information available in biological databases has come a proliferation of 'namespaces', i.e. schemes for naming biological entities (genes, proteins, metabolites, etc.). For example, a single gene might be identified as 'IL1RL1' in the HGNC symbol namespace, 'ENSG00000115602' in the Ensembl gene id namespace, and 'Hs.66' in the Unigene namespace, while a protein product of that gene might be identified as 'NP_003847' in the RefSeq peptide namespace and 'IPI00218676' within the International Protein Index (IPI) namespace.

The lack of standardized gene and protein identifiers remains a fundamental hindrance to biological research, and is particularly obstructive to strategies based on integrating high-throughput data from disparate sources (e.g. combining mRNA expression data with protein interaction and functional annotations).

A very common task is the translation of an ordered set of identifiers from one namespace to another. The Synergizer is a database, associated with both programmatic and interactive web interfaces, with the sole purpose of helping researchers (both bench scientists and bioinformaticians) perform this deceptively simple task.

The simplest way to describe the use of the programmatic interface is via a short example Perl script (see Fig. 1) using the Perl module `Synergizer::DemoClient` (available for download; see Availability above).

The key functionality of the Synergizer application programming interface (API) is represented here by the function `translate` (line 11). When executed, this function generates a remote procedure call in the form of a JSON-encoded object, and sends it via HTTP to a remote Synergizer server. This server translates the identifiers in the 'ids' argument (lines 7–9) from one namespace (here designated as the 'domain') to another (designated the 'range'), using mappings provided by the specified 'authority' [in this case Ensembl (Hubbard *et al.*, 2002)]. These results are returned via HTTP to the script, where they are assigned to the variable `$translated` as a reference to an array of arrays, one array per original input identifier (since an input identifier may return zero, one or several translations). Some identifiers (e.g. 'pxn' in the example) that belong to the domain namespace, but for which no equivalent in the range namespace was found, will return no translations. To highlight inputs that were not found in the domain namespace (e.g. '?test?'), these identifiers are translated to the `undefined` value. For further details, please consult the Synergizer API (see Availability above).

It is important to note that although the example above is written in Perl, the Synergizer service is language independent (as well as platform independent). The API for the service is publicly available and it is a simple matter to write API-conforming clients in Perl, Python, PHP, Ruby, Java, JavaScript or any other modern programming language.

A second illustration of the service is its web front end, (see Availability above), which is itself a Synergizer client application (illustrating the language independence of the Synergizer API, this client is written in JavaScript as opposed to the earlier example in Perl).

Although several tools are available to translate biological identifiers (for example, see references Bussey *et al.*, 2003; Côté *et al.*, 2007; Draghici *et al.*, 2006; Huang *et al.*, 2007; Iragne *et al.*, 2004; Kasprzyk *et al.*, 2004; Reimand *et al.*, 2007), the Synergizer has some features that will make it particularly useful to bioinformaticians.

Perhaps the Synergizer's greatest asset is its simplicity. It is designed to perform a single task, bulk translation of biological database identifiers from one naming scheme (or *namespace*) to another, as quickly and simply as possible. The service obtains its information from authorities, such as Ensembl (Hubbard *et al.*, 2002) and NCBI (Wheeler *et al.*, 2008), that publish detailed correspondences between their identifiers and those used by external databases. In general, we say that two identifiers are 'synonymous', according to a specific authority, when the authority assigns them to its same internal identifier. (For brevity, we refer to the authority's internal identifier as the 'peg'.) For example, we would say that, according to authority Ensembl, the identifiers IL1RL1 and Q01638

*To whom correspondence should be addressed.

© 2008 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

```
 1  use Synergizer::DemoClient ':all';
 2  my $args =
 3    {authority => 'ensembl',
 4     species   => 'Homo sapiens',
 5     domain    => 'hgnc_symbol',
 6     range     => 'entrezgene',
 7     ids       => [qw(c1ql4 scn5A ?test?
 8                      pxn RORC rOrC MYC
 9                      lnx1)]};

11  my $translated = translate($args);

13  my @unrecognized = cull($translated);
14  print $_->format for @$translated;
15  print "\nnot found:\n";
16  print "$_\n" for @unrecognized;
17  __END__
```

Output:

```
c1ql4 338761
scn5A 6331 | 650400 | 731231
pxn
RORC 6097
rOrC  6097
MYC 4609 | 731404
lnx1 84708

not found:
?test?
```

Fig. 1. Use of a typical Synergizer client.

are synonymous because it assigns both to its internal identifier (peg) ENSG00000141510. For the same reason, 6 01 203 and 5998 are synonymous, but in this second example the synonyms are simple numbers that give no indication of the database of origin. For this reason, every identifier listed by the Synergizer service belongs to a specific 'namespace'. In this example, 6 01 203 belongs to the namespace mim_gene_accession, and 5998 belongs to the namespace hgnc_id. More formally, 'namespace', as we use the term, is a collection of identifiers, all generated by the same organization, in a controlled way.

Regarding namespaces, it is worth noting that some providers of biological information follow the practice of prepending a prefix to the identifier to indicate the database of origin (e.g. HGNC:IL1RL1), which has the same effect of segregating identifiers according to namespace. The Synergizer system generally does not follow this practice.

Also, the identifiers discussed until now, which belong to well-defined namespaces, must be distinguished from those that are proposed *ad hoc*, one-at-a-time, by the researchers who first describe them in the literature. These *ad hoc* identifiers are only within the scope of the Synergizer service where they correspond to a tightly controlled namespace for which some authority offers correspondences to other namespaces.

For each authority, Synergizer uses a peg that is specific to that authority. For example, for Ensembl currently it is the Ensembl gene id, and for NCBI it is the Entrez gene id. But the choice of peg for a given authority is an implementation detail that may change in the future.

By this procedure, we generate a repository of synonym relationships between database identifiers. When we do this we often find discrepancies among various authorities. The reasons for these discrepancies are varied. They range from simple time lags between databases, to policy differences among the authorities on the assignment of external identifiers to their respective internal identifiers, and even to more substantive disagreements, at the scientific level, on gene assignments. Rather than attempting to resolve these discrepancies, the synonym relationships served by each authority are kept separate within the Synergizer system.

The Synergizer's schema has been designed to preserve the provenance of all synonym relationships, and to accommodate new sources of synonym information over time.

To access the Synergizer's interactive web interface visit the link listed under Availability above. To use the interface, simply paste the identifiers to be translated in the input field (or, alternatively, enter the name of a local file from which to upload the identifiers). Then, choose the domain and range namespaces. It is also possible to specify the special catchall domain namespace '__ANY__' (although we note that specifying the domain namespace recommended where possible since it is less prone to ambiguous translation). By default the output is in the form of an HTML table, but there is also the option to obtain the output in the form of a spreadsheet.

Currently the Synergizer supports synonyms from two different authorities Ensembl and NCBI, and holds a total of just over 20 million synonym relations covering over 70 species and over 150 namespaces.

ACKNOWLEDGEMENTS

We thank J. Beaver, E. Birney, C. Bult, R. Gerszten, A. Kasprzyk, D. Maglott, J. Mellor, T. Shtatland and M. Tasan for helpful discussions, and technical and editorial advice.

Funding: National Institutes of Health. (grants HG003224, HG0017115, and HL081341), Keck Foundation.

Conflict of Interest: none declared.

REFERENCES

Bussey,K.J. *et al.* (2003) MatchMiner: a tool for batch navigation among gene and gene product identifiers. *Genome Res.*, **4**, 1–7.

Côté,R. *et al.* (2007) The Protein Identifier Cross-Reference (PICR) service: reconciling protein identifiers across multiple source databases. *BMC Bioinformatics*, **8**, 401.

Draghici,S. *et al.* (2006) Babel's tower revisited: a universal resource for cross-referencing across annotation databases. *Bioinformatics*, **22**, 2934–2939.

Huang,H. *et al.* (2007) Integration of bioinformatics resources for functional analysis of gene expression and proteomic data. *Frontiers in Bioscience*, **12**, 5071–5088.

Hubbard,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.

Iragne,F. *et al.* (2004) Aliasserver: a web server to handle multiple aliases used to refer to proteins. *Bioinformatics*, **20**, 2331–2332.

Kasprzyk,A. *et al.* (2004) Ensmart: A generic system for fast and flexible access to biological data. *Genome Res.*, **14**, 160–169.

Reimand,J. *et al.* (2007) g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.*, **35(suppl. 2)**, W193–W200.

Wheeler,D.L. *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36(suppl. 1)**, D13–D21.