Bio::DB::NextProt: A Perl Module for neXtProt Database Information Retrieval

The neXtProt database is a comprehensive knowledge platform recently adopted by the Chromosome-centric Human Proteome Project as the main reference database. The primary goal of the project is to identify and catalog every human protein encoded in the human genome. For such, computational approaches have an important role as data analysis and dedicated software are indispensable. Here we describe Bio::DB::NextProt, a Perl module that provides an object-oriented access to the neXtProt REST Web services, enabling the programatically retrieval of structured information. The Bio::DB::NextProt module presents a new way to interact and download information from the neXtProt database. Every parameter available through REST API is covered by the module allowing a fast, dynamic and ready-to-use alternative for those who need to access neXtProt data. Bio::DB::NextProt is an easy-to-use module that provides automatically retrieval of data, ready to be integrated into third-party software or to be used by other programmers on the fly. The module is freely available from from CPAN (metacpan.org/release/Bio-DB-NextProt) and GitHub (github.com/Leprevost/Bio-DB-NextProt) and is released under the perl_5 license.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.372v1 | CC-BY 4.0 Open Access | received: 24 Apr 2014, published: 24 Apr 2014

Bio::DB::NextProt: A Perl Module for neXtProt Database Information Retrieval

Felipe da Veiga Leprevost[1]

[1]**Laboratory for Proteomics and Protein Engineering, Fiocruz, Curitiba, Brazil.**

ABSTRACT

The neXtProt database is a comprehensive knowledge platform recently adopted by the Chromosome-centric Human Proteome Project as the main reference database. The primary goal of the project is to identify and catalog every human protein encoded in the human genome. For such, computational approaches have an important role as data analysis and dedicated software are indispensable. Here we describe Bio::DB::NextProt, a Perl module that provides an object-oriented access to the neXtProt REST Web services, enabling the programatically retrieval of structured information. The Bio::DB::NextProt module presents a new way to interact and download information from the neXtProt database. Every parameter available through REST API is covered by the module allowing a fast, dynamic and ready-to-use alternative for those who need to access neXtProt data. The most important aspect of having a programatically way to access the database is to provide third-party software the capability to dynamically interact with neXtProt database. Bio::DB::NextProt is an easy-to-use module that provides automatically retrieval of data, ready to be integrated into third-party software or to be used by other programmers on the fly. The module is freely available from from CPAN (metacpan.org/release/Bio-DB-NextProt) and GitHub (github.com/Leprevost/Bio-DB-NextProt) and is released under the perl_5 license.

Keywords: Perl, neXtProt, database

INTRODUCTION

The Human Proteome Organization (HUPO), settled in 2001, established the goal to promote international collaboration and foster the development of new technologies related to proteomics and human diseases (States et al., 2006). The Human Proteome Project (HPP) is a successful HUPO initiative to coordinate laboratories around the world with the common goal of detecting all protein products predicted by the Human Genome Project (Marko-Varga et al., 2013). The initiative is organized in two distinct fronts: the Chromosome-based Human Proteome Project (C-HPP) and the Biology/Disease Human Proteome Project (B/D-HPP). The primary goal of the C-HPP is to identify all human proteins and catalog these findings by a chromosomal perspective (Hancock et al., 2011), with the help of the international community where each participating country is responsible for at least one chromosome. According to the literature, there are several protein coding genes annotated as having uncharacterized products, without taking into account SNPs and splicing variants (Li et al., 2011). The confirmation on the proteins existence should be accomplished by means of three distinct strategies; Mass Spectrometry, Antibodies, and Knowledgebase.

Standardization and effective computational approaches become fundamental on how the cataloging is done and in the availability of biological databases. The neXtProt database, developed at the Swiss Institute of Bioinformatics (SIB), has become the main reference to the C-HP Project (Lane et al., 2012); This database contains most of the known information about the predicted protein coding genes on each chromosome and its status (i.e. predicted by homologous sequences, computational methods, laboratory evidence, etc). Here we describe Bio::DB::NextProt, a Perl module designed to interact on the fly with the neXtProt REST API and FTP services that are made available on-line. The module offers an extensible list of object-oriented methods designed to access and retrieve information from the neXtProt database.

METHODS

Implementation

The Perl programming language is a powerful tool to deal with Internet protocols. Having being especially designed to process text and being backed up by a code repository such as CPAN, Perl presents itself as

an ideal programming language to constitute software working in the interface of communication between different applications and the Web.

REST API

neXtProt provides a REST API (www.nextprot.org/rest/) for providing information on different aspects of the identified human proteins and chromosomes. The HTTP protocol allows requesting information based on proteins features such as functionality, localizations, variants, isoforms, expression, among others. The request can be done manually, by using a standard web browser or command line and passing the desired parameters on the URI. The Bio::DB::NextProt module organizes in an object-oriented structure the methods necessary to interact with the REST interface and its parameters, greatly simplifying the interaction with the API. The module supports all available parameters from the API and retrieves the information in HTML, JSON, and XML formats.

Among the many functionalities made available by the Bio::DB::NextProt we highlight the possibility of retrieving chromosome information by requesting the access of the chr_report tables from the neXtProt FTP server. The module allows the automatic access to C-HPP reports on each chromosome, an important resource with information about the evidence on each predicted protein, automatically restructuring the information in data structures ready to be used by the user's application. Below, in Figure 1, is an example on how the information is organized internally.

FTP Interface

Another key aspect of the Bio::DB::NextProt is the capability of access and download different reports and lists from neXtProt FTP site. The objects posses a list of method configured to download and output different and important reports like the customized report files for the Human Proteome Project. It is also possible to obtain the mapping lists between the neXtProt accession numbers and the identifiers from different databases like Emsembl and Refseq.

```
NX_A7E2V4                    {
    antibody          "yes",
    description       "Zinc finger SWIM domain-containing
    disease           "no",
    existence         "protein level",
    has_3d            "no",
    isoforms          5,
    gene_name         "ZSWIM8",
    position          "10q22.2",
    proteomics        "yes",
    ptms              6,
    start_position    75545340,
    stop_position     75561551,
    variants          67
}
```

Figure 1. Example of the data structure representing one protein from chromosome 10 with its features.

CONCLUSION

Bioinformatics is one of the bottlenecks in the C-HPP project. The Bio::DB::NextProt poses as a standard building block for the development of new tools devoted for this project. Its broader applications allows a fast and practical retrieval of information that is automatically structured to be used by other softwares or by a bioinformatician that needs to process the data. Approaches like the one described here will help researchers working in projects like the C-HPP, facilitating the access to structured information.

AVAILABILITY AND REQUIREMENTS

Bio::DB::NextProt is implemented as an Perl package, is freely available from the Comprehensive Perl Archive Network (CPAN / metaCPAN) and can be installed though cpan shell. The source code is also available at GitHub (github.com/Leprevost/Bio-DB-NextProt).

ACKNOWLEDGMENTS

I would like to thank Dr. Paulo Costa Carvalho for revising this software application

REFERENCES

Hancock, W., Omenn, G., LeGrain, P., and Paik, Y.-K. (2011). Proteomics, human proteome project, and chromosomes. *Journal of Proteome Research*, 10(1):210–210.

Lane, L., Argoud-Puy, G., Britan, A., Cusin, I., Duek, P. D., Evalet, O., Gateau, A., Gaudet, P., Gleizes, A., Masselot, A., and et al. (2012). nextprot: a knowledge platform for human proteins. *Nucleic Acids Research*, 40(D1):D76–D83.

Li, M., Wang, I. X., Li, Y., Bruzel, A., Richards, A. L., Toung, J. M., and Cheung, V. G. (2011). Widespread rna and DNA sequence differences in the human transcriptome. *Science*, 333(6038):53–58.

Marko-Varga, G., Omenn, G. S., Paik, Y.-K., and Hancock, W. S. (2013). A first step toward completion of a genome-wide characterization of the human proteome. *Journal of Proteome Research*, 12(1):1–5.

States, D. J., Omenn, G. S., Blackwell, T. W., Fermin, D., Eng, J., Speicher, D. W., and Hanash, S. M. (2006). Challenges in deriving high-confidence protein identifications from data gathered by a hupo plasma proteome collaborative study. *Nature Biotechnology*, 24(3):333–338.