Vol. 30 no. 2 2014, pages 295–297
doi:10.1093/bioinformatics/btt658

Advance Access publication November 13, 2013

SeqDepot: streamlined database of biological sequences and precomputed features

Luke E. Ulrich[1,2,*] and Igor B. Zhulin[2,3]

[1]Agile Genomics, LLC, Mount Pleasant, SC 29466, USA, [2]Department of Microbiology, University of Tennessee, Knoxville, TN 37996, USA and [3]Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

Associate Editor: John Hancock

ABSTRACT

Summary: Assembling and/or producing integrated knowledge of sequence features continues to be an onerous and redundant task despite a large number of existing resources. We have developed SeqDepot—a novel database that focuses solely on two primary goals: (i) assimilating known primary sequences with predicted feature data and (ii) providing the most simple and straightforward means to procure and readily use this information. Access to >28.5 million sequences and 300 million features is provided through a well-documented and flexible RESTful interface that supports fetching specific data subsets, bulk queries, visualization and searching by MD5 digests or external database identifiers. We have also developed an HTML5/JavaScript web application exemplifying how to interact with SeqDepot and Perl/Python scripts for use with local processing pipelines.

Availability: Freely available on the web at http://seqdepot.net/. REST access via http://seqdepot.net/api/v1. Database files and scripts may be downloaded from http://seqdepot.net/download.

Contact: ulrich.luke+sci@gmail.com

Received on September 11, 2013; revised on October 31, 2013; accepted on November 7, 2013

1 INTRODUCTION

Primary sequences and their associated features are vital to all bioinformatics research. Invariably, determining gene and protein function begins with sequence similarity searches (Altschul *et al.*, 1997) followed by identifying intrinsic features that strictly depend on the sequence itself. Many computational tools and libraries exist for elucidating intrinsic features such as signal peptides (Petersen *et al.*, 2011), transmembrane regions (Cserzo *et al.*, 2003; Sonnhammer *et al.*, 1998), protein domains (Haft *et al.*, 2013; Letunic *et al.*, 2012; Punta *et al.*, 2012), motifs (Attwood *et al.*, 2012; Jacobs *et al.*, 2009), subcellular location (Emanuelsson *et al.*, 2007) and structural elements (Lees *et al.*, 2010). When considered together and in context, these higher level details synergistically improve the understanding of macromolecular function.

Obtaining such functional details requires building a computational pipeline that analyzes sequences with the desired toolset and/or fetching results from external servers. Developing a custom pipeline is computationally expensive. Moreover, predicting intrinsic features produces the same results for identical sequences, and recomputing this information wastes valuable resources (Rattei *et al.*, 2010). Many public resources such as Pfam (Punta *et al.*, 2012), SMART (Letunic *et al.*, 2012) and Gene3D (Lees *et al.*, 2010) provide precomputed results; however, significant effort must be exercised to properly map sequence identifiers, handle coverage discrepancies and integrate results from independent resources.

SeqDepot addresses the aforementioned issues by consolidating a wide array of precomputed features across known protein space in a single database. SeqDepot enables users to rapidly retrieve and process with minimal effort both primary and intrinsic feature data for existing sequences, which makes it different from other tools (e.g. InterPro) that dynamically compute this information for novel sequences.

InterPro (Hunter *et al.*, 2012) and Similarity Matrix of Proteins (SIMAP; Rattei *et al.*, 2010) aggregate large masses of precomputed data for classifying all proteins in UniProt (Consortium, 2012) into families and calculating all-versus-all sequence similarity analyses, respectively; however, simply obtaining and consuming features for a given set of sequences is neither straightforward nor streamlined. InterPro's web interface permits querying a single sequence per request or up to 25 per request via web services. If precomputed matches do not exist, then the sequences are queued for analysis with InterProScan (Quevillon *et al.*, 2005), which may take 30 minutes per sequence to complete. SIMAP provides several avenues for data access including the web browser, database flat files and most notably via a powerful SOAP-based web service. Sequence searches with the web browser are limited to a single sequence with a maximum of 500 hits, and effectively using SOAP interfaces requires non-trivial programming. Exported results (e.g. flat files) from both InterPro and SIMAP for all predictive tools are merged into a fixed number of fields and necessitate additional parsing logic to disentangle and properly structure the data of interest.

As high-throughput sequencing projects exponentially reveal additional sequences, researchers are required to analyze larger amounts of sequence data. Even with the vast number of databases and web services available today, one must expend considerable time and effort fetching sequences and features, installing diverse software packages, generating results, restructuring data and even building local databases. There remains a compelling need for a single resource that both maps raw sequences to precomputed feature data at scale and is easily queried and

[*] To whom correspondence should be addressed.

processed by both humans and computers. SeqDepot fills this gap with the overarching goals of being both comprehensive and extremely simple to use.

2 IMPLEMENTATION

All data are stored in a MongoDB (http://mongodb.org) database. An 'Aseq'—a unique amino acid sequence regardless of its source or species—constitutes the core document entity. Each Aseq is uniquely identified by an 'Aseq ID' (amino acid sequence identifier)—a 22-character string derived by encoding the binary MD5 digest of the uppercase sequence characters with a slightly modified Base64 scheme. This scheme removes terminal padding characters and converts all forward slashes and plus signs to URL 'friendly' underscores and dashes, respectively. Each Aseq contains basic descriptors including its sequence, length and status. Aseqs are further decorated with predicted features and external database references.

Raw sequences and precomputed features hosted by the SIMAP project (Rattei *et al*., 2010) are downloaded, parsed and inserted into the database. Identifiers and sequences contained in the NCBI non-redundant, UniProt (Consortium, 2012) and Protein Data Bank (Rose *et al*., 2013) databases are subsequently merged to provide additional coverage and cross-referencing capabilities. Adding to the many predictive features precomputed by SIMAP (Rattei *et al*., 2010), we also predict transmembrane regions with the Dense Alignment Surface TM filter (Cserzo *et al*., 2003) and Extra Cytoplasmic Function domains (Staroń *et al*., 2009). The entire database is publicly accessible via a RESTful (representational state transfer) Application Programming Interface (API). Updates are performed on a quarterly basis to integrate novel sequences and associated features. This includes refreshing precomputed data with results from current versions of each feature prediction tool.

3 FEATURES

3.1 Comprehensive

SeqDepot contains >28.5 million unique amino acid sequences sourced from major sequence databases and >300 million intrinsic features.

3.2 Non-redundant

By definition, all Aseqs are non-redundant and source-agnostic. Thus, all derived data associated with a specific sequence is easily consolidated in a single document.

3.3 Intrinsic identifiers

Aseq IDs solely depend on the full-length ungapped sequence characters (i.e. intrinsic) and may be easily generated independent of any external database. Such universally unique and size-efficient identifiers are permanently stable and ideal for mapping sequences between databases.

3.4 Flexible well-documented RESTful API

The REST API encapsulates the gateway to all data within SeqDepot and has been designed to make accessing, visualizing and consuming relevant data extremely easy. The data for any given Aseq may be retrieved by requesting a URI with the general form http://seqdepot.net/api/v1/aseqs/{id}.{format}?{parameters}, where {id} is a supported identifier type, {format} is the desired output format (json, png or svg for JSON, PNG image or SVG image, respectively) and {parameters} optionally modifies the response (e.g. limit the fields returned). To visualize the domain architecture for a sequence, simply replace {format} with either png or svg to produce a PNG or SVG image file of the predicted biological elements. Textual results and error messages are encoded in JSON (http://json.org). Virtually every programming language contains libraries for transforming JSON encoded data into native data structures with a few lines of code. This completely circumvents the need to write custom parsers and requires minimal effort to process results in downstream applications. Complete documentation covering all aspects of the REST API including several examples is located at http://seqdepot.net/api.

3.5 Visualization

Generate raster or vector images visualizing the full domain architecture for any Aseq.

3.6 Bulk querying

Supports up to 1000 queries per request.

3.7 Cross-referencing

Aseqs may also be referenced using the hexadecimal MD5 digest of the sequence or a variety of external database identifiers including UniProt (Consortium, 2012), GenBank (GI) or Protein Data Bank (Rose *et al*., 2013) identifiers.

3.8 Utility scripts

Perl and Python modules and a companion script, sdQuery.pl, facilitate interacting with SeqDepot and performing many common tasks including the following: parse FASTA files, generate Aseq IDs, fetch sequences and/or precomputed data, cross-reference sequences, download visualizations and much more.

ACKNOWLEDGEMENTS

The authors thank Davi Ortega for useful feedback, system administration and helping with various implementation details. They are grateful for access to the Newton HPC facilities and thank Gerald Ragghianti for his assistance with using the high-performance cluster.

Funding: This work was supported in part by the National Institutes of Health grant [GM072285].

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Attwood,T.K. *et al*. (2012) The PRINTS database: a fine-grained protein sequence annotation and analysis resource—its status in 2012. *Database*, **2012**, bas019.

Consortium,T.U. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res*., **40**, D71–D75.

Cserzo,M. *et al*. (2003) TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. *Bioinformatics*, **20**, 136–137.

Emanuelsson,O. *et al*. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat*. *Protoc*., **2**, 953–971.

Haft,D.H. *et al*. (2013) TIGRFAMs and genome properties in 2013. *Nucleic Acids Res*., **41**, D387–D395.

Hunter,S. *et al*. (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res*., **40**, D306–D312.

Jacobs,G.H. *et al*. (2009) Transterm: a database to aid the analysis of regulatory sequences in mRNAs. *Nucleic Acids Res*., **37**, D72–D76.

Lees,J. *et al*. (2010) Gene3D: merging structure and function for a thousand genomes. *Nucleic Acids Res*., **38**, D296–D300.

Letunic,I. *et al*. (2012) SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res*., **40**, D302–D305.

Petersen,T.N. *et al*. (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat*. *Methods*, **8**, 785–786.

Punta,M. *et al*. (2012) The Pfam protein families database. *Nucleic Acids Res*., **40**, D290–D301.

Quevillon,E. *et al*. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res*., **33**, W116–W120.

Rattei,T. *et al*. (2010) SIMAP—a comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters. *Nucleic Acids Res*., **38**, D223–D226.

Rose,P.W. *et al*. (2013) The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res*., **41**, D475–D482.

Sonnhammer,E.L. *et al*. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc*. *Int*. *Conf*. *Intell*. *Syst*. *Mol*. *Biol*., **6**, 175–182.

Staroń,A. *et al*. (2009) The third pillar of bacterial signal transduction: classification of the extracytoplasmic function (ECF) sigma factor protein family. *Mol*. *Microbiol*., **74**, 557–581.