Bioinformatics Advance Access published January 23, 2016

BISQUE: locus- and variant-specific conversion of genomic, transcriptomic, and proteomic database identifiers

Michael J. Meyer[1,2,3,†], Philip Geske[1,2,†] and Haiyuan Yu[1,2,*]

[1] Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York, 14853, USA

[2] Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, New York, 14853, USA

[3] Tri-Institutional Training Program in Computational Biology and Medicine, New York, New York, 10065, USA

Associate Editor: Dr. Janet Kelso

ABSTRACT

Summary: Biological sequence databases are integral to efforts to characterize and understand biological molecules and share biological data. However, when analyzing these data, scientists are often left holding disparate biological currency—molecular identifiers from different databases. For downstream applications that require converting the identifiers themselves, there are many resources available, but analyzing associated loci and variants can be cumbersome if data is not given in a form amenable to particular analyses. Here we present BISQUE, a web server and customizable command-line tool for converting molecular identifiers and their contained loci and variants between different database conventions. BISQUE uses a graph traversal algorithm to generalize the conversion process for residues in the human genome, genes, transcripts, and proteins, allowing for conversion across classes of molecules and in all directions through an intuitive web interface and a URL-based web service.

Availability: BISQUE is freely available via the web using any major web browser (http://bisque.yulab.org/). Source code is available in a public GitHub repository (https://github.com/hyulab/BISQUE).

Contact: haiyuan.yu@cornell.edu

Supplementary Information: Data are available online.

1 INTRODUCTION

The proliferation of genomic and proteomic databases has helped us organize and understand biological molecules and phenomena, but has left the scientific community using many different naming conventions for the same, or intrinsically related biological entities: genes, proteins, and transcripts (Cunningham, et al., 2015; Gray, et al., 2015; Pruitt, et al., 2014; UniProt-Consortium, 2015). While there are many tools to convert the identifiers themselves, there are insufficient resources to convert loci and variants annotated in reference to these identifiers (Supplementary Table 1). Due to pervasive sequencing and variant annotation, this has led to a routine burden on biologists to convert from one naming convention to another, and may lead to errors when building upon other labs' research (McCarthy, et al., 2014).

Here we present BISQUE (The **Bi**ological **Sequ**ence **E**xchange), a multi-interface utility for converting human genomic, transcriptomic, and proteomic loci and variants from their reported form into forms useful for downstream research. BISQUE is an extensible conversion framework deployed as a web server (http://bisque.yulab.org) for user-friendly conversion among the most popular human database identifiers. It is also available as a programmatic web service, downloadable as a customizable standalone application (http://github.com/hyulab/bisque), and importable as a Python module.

2 METHODS

BISQUE is designed to track the manifestation of variants in coding regions of human biomolecules. It can trace genomic variants to their effects in higher order molecules (transcripts and proteins) and, conversely, discover the source of proteome aberrations in lower order genomic sequences (genome and genes). As such, BISQUE catalogues positions in molecules that are functional in their transcription or translation across all catalogued molecular classes (genome/genes, transcripts, and proteins). To further functional discovery, BISQUE incorporates two peripheral databases for investigation, dbSNP (Sherry, et al., 2001) and the PDB (Berman, 2000), allowing users to quickly determine the relationships between known genomic variants and their potential effects on protein structures.

The nodes in the BISQUE core conversion graph (Fig. 1a) include the latest human genome builds (GRCh37 and GRCh38) (Benson, et al., 2015), Ensembl gene, transcript, and protein (Cunningham, et al., 2015), RefSeq transcript and protein (Pruitt, et al., 2014), UniProt protein (UniProt-Consortium, 2015), dbSNP (Sherry, et al., 2001), and the PDB (Berman, 2000). The connections between these nodes in the graph represent all potential traversals to produce conversions, beginning at an origin node determined by user input and a destination node based on a selection by the user. BISQUE then identifies the optimal path through the conversion graph from the origin to the destination based on the available edges (Supplementary Note 1.5). Conversions are computed stepwise along this path as single conversions between one input node and one output node. At each step, BISQUE first checks if a mapping exists for the input identifier in the selected output database. If a mapping exists, any associated loci and mutations will be converted according to the rules that describe the edge and its direction between the nodes (Supplementary Note 1.6). For instance, when converting from a genomic locus to a transcript locus, introns must be removed, strand-sense (whether or not the transcript is on the sense (+) or antisense (-) genomic DNA strand) must be taken into account, and variant bases complemented in a transcript annotated in reference to the antisense strand. The output of each stepwise conversion is fed as input to the next stepwise conversion. Figure 1b describes an example conversion, showing the input and output of each step required to convert from genomic variant to protein substitution.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

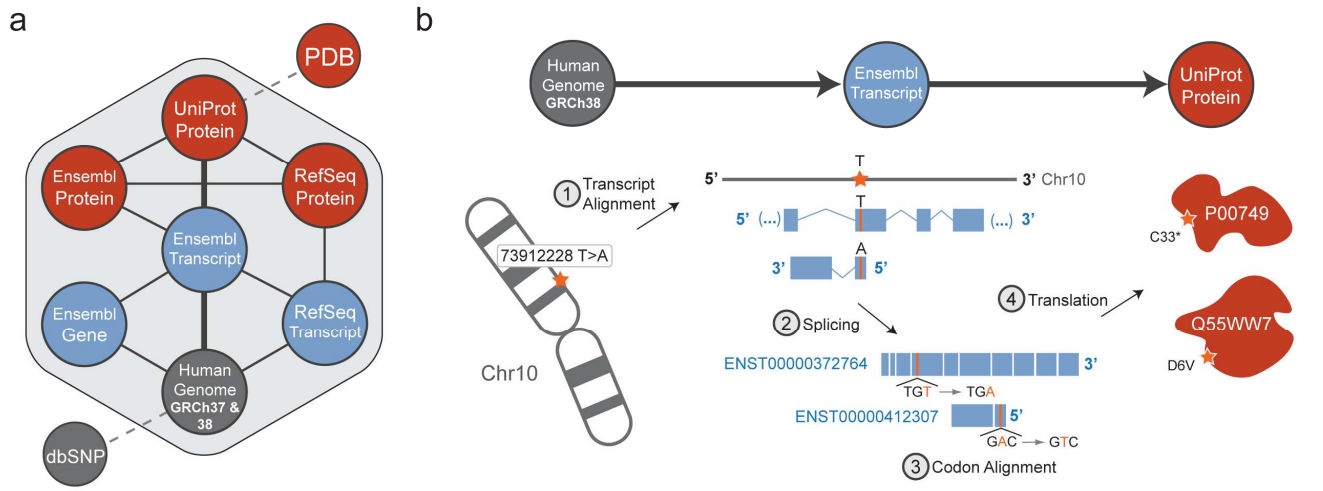

Fig. 1. (a) The core BISQUE conversion graph, including all possible starting and ending points for a conversion with BISQUE. (b) The steps for conversion of a genomic variant to UniProt amino acid substitution(s) by traversing a conversion path through Ensembl transcript. The shown genomic variant maps to two Ensembl transcripts, one on the forward strand and one on the reverse. BISQUE uses genomic alignments of transcripts from Ensembl's database, removes introns and non-coding exons (UTRs and alternatively spliced regions), and codon-aligns the result to match amino acids in UniProt proteins.

3 USAGE

BISQUE is available in several forms to increase its usability for a variety of applications. For small queries (up to 1000 conversions at a time), the web server is recommended. For larger queries and systematic integration, the web service or command-line application may be more appropriate.

3.1 Web server

The BISQUE web server is both a frontend to the conversion engine and a portal to learning more about available web functions and accessing BISQUE in its other forms (discussed below). Queries can be submitted either one at a time or in batch (by pressing the '+' button) through the interactive form on the home page. Batch queries may be manually entered into a form or uploaded as a text file in a variety of formats, including the Variant Call Format (VCF). Conversion results are presented in a table which is exportable in several popular formats including CSV, JSON, and XML.

3.2 Web service

BISQUE also provides simple programmatic access to its conversion engine through URL-based queries that produce tab-delimited, plain text output. By passing query parameters through CGI fields, users may rapidly access BISQUE's conversion capabilities in their own scripts (examples using both Python and Perl are provided on the About page), or incorporate BISQUE conversions into publicly available web servers and databases.

3.3 Command-line application

The BISQUE conversion engine is encapsulated in a separate, downloadable command-line application. Two versions of this tool are available for download through the BISQUE web server—a lite and a full version. The full version download includes all MySQL tables required for mapping so that a user may install BISQUE without any external dependencies. The lite version contains all of the same functionality as the full version, but relies on an internet connection to access data from the public BISQUE MySQL database.

The command-line application is open source, written in Python, and importable as a Python module, providing users with another avenue to incorporate BISQUE into their scripts. This also gives more advanced users the ability to modify BISQUE to meet their own needs, for instance to expand the coverage of database identifiers available for conversion by adding and removing edges and nodes from the conversion graph. For most new databases this is achievable through included utilities and documentation, and may not require modifying the code.

4 DISCUSSION

While there are some tools available to perform limited conversion of positions and variants associated with database identifiers (Supplementary Table 1), these often exist as coupled conversion-analysis tools, which lack the all-to-all conversion capability of BISQUE. Furthermore, variant conversion is typically coupled with much more computationally expensive functions, such as variant annotation. What is lacking in these tools is modularity—a core design principle of complex systems that allows functional components to be repurposed for other tasks. BISQUE embraces modularity in its own internal structure through a generalized conversion framework, which enables the expansion of its conversion capability. BISQUE is also a module itself, with the ability to perform multiple conversions for a scripting task or to act as the conversion engine for a database or analysis tool. In this way, BISQUE will be very valuable to the scientific community as it publicizes vital biological infrastructure.

Funding: This work was supported by the National Institute of General Medical Sciences [R01 GM104424 to H.Y.]

Conflict of Interest: None declared.

REFERENCES

Benson, D.A., *et al.* (2015) GenBank, *Nucleic Acids Res*, **43**, D30-35.
Berman, H.M. (2000) The Protein Data Bank, *Nucleic Acids Research*, **28**.
Cunningham, F., *et al.* (2015) Ensembl 2015, *Nucleic Acids Res*, **43**, D662-669.
Gray, K.A., *et al.* (2015) Genenames.org: the HGNC resources in 2015, *Nucleic Acids Res*, **43**, D1079-1085.
McCarthy, D.J., *et al.* (2014) Choice of transcripts and software has a large effect on variant annotation, *Genome Med*, **6**, 26.
Pruitt, K.D., *et al.* (2014) RefSeq: an update on mammalian reference sequences, *Nucleic Acids Res*, **42**, D756-763.
Sherry, S., *et al.* (2001) dbSNP: the NCBI database of genetic variation, *Nucleic acids research*, **29**, 308-311.
UniProt-Consortium (2015) UniProt: a hub for protein information, *Nucleic Acids Res*, **43**, D204-212.