BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 5 2010, pages 700–702
doi:10.1093/bioinformatics/btp705

Databases and ontologies Advance Access publication January 6, 2010

dbTEU: a protein database of trace element utilization

Yan Zhang and Vadim N. Gladyshev*

Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA

Associate Editor: Burkhard Rost

ABSTRACT

Summary: Biological trace elements are required for numerous biological processes and by all organisms. We describe a database, dbTEU (DataBase of Trace Element Utilization), that features known transporters and user proteins for five trace elements (copper, molybdenum, nickel, cobalt and selenium) and represents sequenced organisms from the three domains of life. The manually curated dbTEU currently includes ~16 500 proteins from >700 organisms, and offers interactive trace element, protein, organism and sequence search and browse tools.

Availability and Implementation: dbTEU is freely available at http://gladyshevlab.bwh.harvard.edu/trace_element/

Contact: vgladyshev@rics.bwh.harvard.edu

Received on October 17, 2009; revised on December 3, 2009; accepted on December 16, 2009

1 BACKGROUND

Biological trace elements are used in small amounts but are utilized by all organisms. These micronutrients, such as iron (Fe), zinc (Zn), copper (Cu), molybdenum (Mo), manganese (Mn), nickel (Ni), cobalt (Co), selenium (Se) and iodine (I), play important roles in a variety of biological processes by providing proteins with unique catalytic, coordination, redox functions and also have other functions. Metals are used either as coordinated ions or as part of more complex metal-containing cofactors (such as molybdopterin for Mo and vitamin B_{12} for Co) (Mendel *et al*., 2007). Se is mainly used in the form of selenocysteine (Sec), known as the 21st amino acid in proteins inserted co-translationally into nascent polypeptide chains (Böck *et al*., 1991). Investigation of proteins that utilize trace elements (metalloproteins and selenoproteins) as well as other proteins involved in trace element homeostasis (such as transporters, regulators and chaperones) and cofactor biosynthesis pathways may provide important information for understanding utilization of trace elements and their common and unique features.

Several databases have previously been developed that provide sequence and other information for selected trace elements, such as Metalloprotein Database & Browser (MDB) (Castagnetto *et al.*, 2002) and SelenoDB (Castellano *et al.*, 2008). In addition, efforts on evolutionary processes of metal utilization produced sets of metalloproteins in several organisms (Andreini *et al.*, 2006; Dupont *et al.*, 2006). These resources contain useful, but partial information about trace element utilization and are restricted to a limited number of organisms.

Here, we describe dbTEU (database of trace element utilization) which includes all known transporters and user proteins (excluding metal-binding chaperones and regulators) for five trace elements: Cu, Mo, Ni, Co and Se, in >700 sequenced organisms across the three domains of life. Approximately 16 500 proteins are stored in the database that represent known proteins and processes dependent on these five trace elements. In addition, the dbTEU offers multiple search and browse options, such as BLAST homology searches, and allows compilation of sets of trace element utilization proteins in individual organisms.

2 DATABASE CONTENT

We previously carried out comparative genomic analyses of utilization of Cu, Mo, Ni, Co and Se (Kryukov and Gladyshev, 2004; Lobanov *et al*., 2009; Ridge *et al*., 2008; Zhang and Gladyshev, 2008a; Zhang and Gladyshev, 2008b; Zhang *et al*., 2009). Members of known metal transporters and metalloproteins were analyzed by several approaches, including conserved domain, genomic context, phylogenetic and metal-binding ligand analyses (Zhang and Gladyshev, 2009). The identification of selenoproteins is also reliable because several independent algorithms with excellent performance have been developed to identify these proteins in sequence databases (Kryukov *et al.*, 1999, 2003; Kryukov and Gladyshev, 2004; Zhang and Gladyshev, 2005).

In dbTEU v. 1.0, we compiled the largest protein database specific for the utilization of five trace elements, which features all sequenced genomes from the three domains of life (as of December 2008; NCBI). Based on trace element and protein category, proteins were divided into eight groups: (i) Cu transporter, (ii) Cu user, (iii) Mo transporter, (iv) Mo user, (v) Ni/Co transporter, (vi) Ni user, (vii) Co user (Co/B_{12}-dependent proteins) and (viii) Se user (selenoproteins). For each entry in dbTEU, we provide the following information: (i) organism name, (ii) kingdom, (iii) dependence on a particular trace element, (iv) protein category, (v) protein family description, (vi) GenBank accession number, (vii) protein sequence, (viii) structure information (BLAST search against PDB sequence database) and (ix) functional association (search against STRING, a database of known and predicted functional associations).

As a novel resource for the analyses of trace elements in sequenced organisms, dbTEU provides large protein datasets for understanding trace element utilization and evolution in the three domains of life. Compared with published databases, dbTEU greatly extends the information about utilization of trace elements and should be a valuable resource for researchers working in this field. Our database

*To whom correspondence should be addressed.

is also the first attempt to define metalloproteomes, i.e. sets of proteins that represent all or almost all metal-dependent proteins in sequenced organisms.

3 DATABASE INTERFACE AND TOOLS

The database system is implemented in openSUSE 11.1 (Linux). It employs Apache as web server and Perl for interactive web pages. The dbTEU supports various search options in order to facilitate access to and utilization of the information stored in the database.

The 'Browse organism' page shows the list of all examined prokaryotes and the majority of examined eukaryotes (Fig. 1A). Organisms are classified into phyla or clades. Since some eukaryotic genomes are incompletely sequenced, we selected a subset of the examined eukaryotes (including all major model organisms) for some eukaryotic phyla. By clicking on the organism name, a user can retrieve all trace element-related proteins in the selected organism. The results are shown as a tab-delimited text format. Each protein entry contains several items, including 'Trace element', 'Protein category', 'Protein family', 'NCBI accession number', 'Sequence', 'Structural information (PDB)' and 'Functional association (STRING)'. Proteins are sorted by 'Trace element', 'Protein category' and 'Protein family'. The NCBI accession number is linked to GenBank. Although GenBank entries contain protein sequence information, errors were found for some proteins, including the majority of selenoproteins (i.e. the Sec. codon in selenoprotein genes is often annotated as stop signal). Thus, in dbTEU, we provide sequence information for each protein upon clicking on 'show sequence' button (for selenoproteins, Sec. residues are represented by U and highlighted in red). In addition, we provide links for BLAST searches against PDB sequence database and functional association networks predicted by STRING (http://string.embl.de/). Figure 1B shows one example: clicking on the actinobacterium *Arthrobacter chlorophenolicus* shows all trace element utilization proteins in this organism. Figure 1C–E show further information: the accession number link, 'show sequence' feature and STRING functional prediction, respectively.

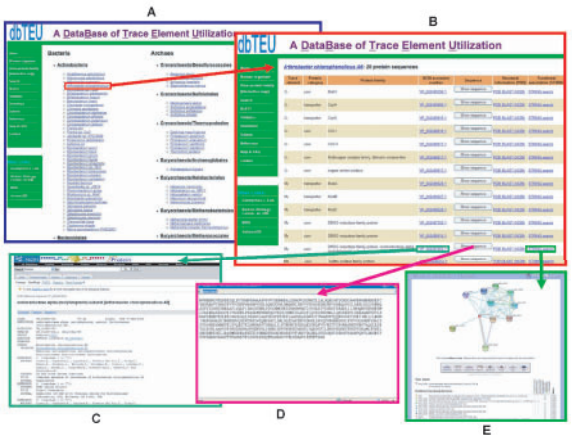

Fig. 1. Examples of the use of 'Browse organism' option to analyze trace element utilization in individual organisms. (**A**) The 'Browse organism' window with all examined organisms from the three domains of life; (**B**) All trace element utilization proteins in the actinobacterium *Arthrobacter chlorophenolicus*; (**C**) The corresponding NCBI annotation; (**D**) The sequence window; (**E**) The STRING functional prediction window.

The 'View protein family' page provides user with an interactive map of all protein families in dbTEU. User may retrieve members of a protein family by simply clicking on its name.

With the 'Search' page, user can search for predefined protein information via three options: protein category, protein family and keyword search. In addition, if a protein family contains a large number of subfamilies, user may use specific subfamily name (e.g. laccase or ceruloplasmin instead of multicopper oxidase). The output format of all three options is the same. It is also similar to that of the browse tool (see above) except that two additional columns, 'Organism' and 'Kingdom', are included for database search results.

We also provide online BLAST service based on NCBI wwwBLAST v. 2.2.18 to support sequence similarity searches against the entire dbTEU sequence database. To use BLAST, the user needs to input a query protein sequence and select a BLAST program (currently blastp or blastx). The user can modify other default settings, such as *E*-value cutoff, low complexity filter and the number of alignments shown in the BLAST output.

dbTEU offers users a real-time monitor program that shows the distribution of protein categories in the current database. When new proteins are added into dbTEU, the 'Statistics' page automatically updates the number of each category. In addition, by clicking on the number that follows each category, users can retrieve the corresponding protein information with the same format as the database search output.

All proteins or subsets of protein sequences (e.g. Cu transporters or selenoproteins) can be downloaded in FASTA format with the annotation of protein family, GenBank accession number (if available), trace element, protein category, source organism and phylum. These files are stored in text format and are available on the 'Download' page.

4 PERSPECTIVES

We focused on developing the tools for the analysis of utilization of five trace elements, Cu, Mo, Ni, Co and Se, in sequenced organisms across the three domains of life. We plan to update our database periodically to include newly identified proteins involved in trace element utilization, especially new transporters and user proteins. We also plan to include information on additional proteins (such as proteins involved in molybdopterin and Sec biosynthesis and metal chaperones) and additional trace elements.

Funding: National Institutes of Health (GM061603).

Conflict of Interest: none declared.

REFERENCES

Andreini,C. *et al*. (2006) Zinc through the three domains of life. *J. Proteome Res.*, **5**, 3173–3178.

Böck,A. *et al*. (1991) Selenocysteine: the 21st amino acid. *Mol. Microbiol.*, **5**, 515–520.

Castagnetto,J.M. *et al*. (2002) MDB: the Metalloprotein Database and Browser at The Scripps Research Institute. *Nucleic Acids Res.*, **30**, 379–382.

Castellano,S. *et al*. (2008) SelenoDB 1.0: a database of selenoprotein genes, proteins and SECIS elements. *Nucleic Acids Res.*, **36**, D332–D338.

Dupont,C.L. *et al*. (2006) Modern proteomes contain putative imprints of ancient shifts in trace metal geochemistry. *Proc. Natl Acad. Sci. USA.*, **103**, 17822–17827.

Kryukov,G.V. *et al*. (1999) New mammalian selenocysteine-containing proteins identified with an algorithm that searches for selenocysteine insertion sequence elements. *J. Biol. Chem.*, **274**, 33888–33897.

Kryukov,G.V. *et al*. (2003) Characterization of mammalian selenoproteomes. *Science*, **300**, 1439–1443.

Kryukov,G.V. and Gladyshev,V.N. (2004) The prokaryotic selenoproteome. *EMBO Rep.*, **5**, 538–543.

Lobanov,A.V. *et al*. (2009) Eukaryotic selenoproteins and selenoproteomes. *Biochim. Biophys. Acta*, **1790**, 1424–1428.

Mendel,R.R. *et al*. (2007) Metal and cofactor insertion. *Nat. Prod. Rep.*, **24**, 963–971.

Ridge,P.G. *et al*. (2008) Comparative genomic analyses of copper transporters and cuproproteomes reveal evolutionary dynamics of copper utilization and its link to oxygen. *PLoS One*, **3**, e1378.

Zhang,Y. and Gladyshev,V.N. (2005) An algorithm for identification of bacterial selenocysteine insertion sequence elements and selenoprotein genes. *Bioinformatics*, **21**, 2580–2589.

Zhang,Y. and Gladyshev,V.N. (2008a) Molybdoproteomes and evolution of molybdenum utilization. *J. Mol. Biol.*, **379**, 881–899.

Zhang,Y. and Gladyshev,V.N. (2008b) Trends in selenium utilization in marine microbial world revealed through the analysis of the global ocean sampling (GOS) project. *PLoS Genet.*, **4**, e1000095.

Zhang,Y. *et al*. (2009) Comparative genomic analyses of nickel, cobalt and vitamin B_{12} utilization. *BMC Genomics*, **10**, 78.

Zhang,Y. and Gladyshev,V.N. (2009) Comparative genomics of trace elements: emerging dynamic view of trace element utilization and function. *Chem. Rev.*, **109**, 4828–4861.