Nucleic Acids Research Advance Access published December 8, 2009

Nucleic Acids Research, 2009, 1–4
doi:10.1093/nar/gkp958

The Protein Naming Utility: a rules database for protein nomenclature

Johannes Goll*, Robert Montgomery, Lauren M. Brinkac, Seth Schobel, Derek M. Harkins, Yinong Sebastian, Susmita Shrivastava, Scott Durkin and Granger Sutton

The J. Craig Venter Institute, Rockville, MD 20850, USA

Received August 15, 2009; Revised October 6, 2009; Accepted October 13, 2009

ABSTRACT

Generation of syntactically correct and unambiguous names for proteins is a challenging, yet vital task for functional annotation processes. Proteins are often named based on homology to known proteins, many of which have problematic names. To address the need to generate high-quality protein names, and capture our significant experience correcting protein names manually, we have developed the Protein Naming Utility (PNU, http://www.jcvi.org/pn-utility). The PNU is a web-based database for storing and applying naming rules to identify and correct syntactically incorrect protein names, or to replace synonyms with their preferred name. The PNU allows users to generate and manage collections of naming rules, optionally building upon the growing body of rules generated at the J. Craig Venter Institute (JCVI). Since communities often enforce disparate conventions for naming proteins, the PNU supports grouping rules into user-managed collections. Users can check their protein names against a selected PNU rule collection, generating both statistics and corrected names. The PNU can also be used to correct GenBank table files prior to submission to GenBank. Currently, the database features 3080 manual rules that have been entered by JCVI Bioinformatics Analysts as well as 7458 automatically imported names.

INTRODUCTION

During the annotation phase of a typical modern genomics project, functional names are assigned to identified genes and proteins in an automated or semi-automated fashion. Ideally, before such names are submitted to public sequence databases, they should be manually reviewed by experts to ensure that they are consistent, syntactically correct and unambiguous. However, with the scale of genomic data produced by next-generation sequencing technology and with increasingly automated functional annotation processes, the manual correction of names is no longer feasible. This issue is further complicated by the prevalence of ambiguous names resulting from the lack of interspecies naming conventions (1). New proteins are often named based on homology to existing proteins and many existing proteins have syntactically incorrect or ambiguous names, producing transitive annotation errors. Consequently, poor-quality names have proliferated in both public databases and the scientific literature.

The need for consistent and unambiguous names has led to the development of a number of conventions for naming genes and proteins [UniProt protein nomenclature (2), HUGO human gene name nomenclature (3) and various other model organism databases (4–7)]. In addition, the biological text mining community has created dictionaries to resolve gene/protein synonyms to improve the identification of genes and proteins in scientific articles (1,8).

The Broad Institute has developed BioNames, a tool to resolve these difficulties using collections of hard-coded regular expressions (https://sourceforge.net/projects/microbiomeutil). Here, we present our solution to this problem in the form of the Protein Naming Utility (PNU), a web-based database to store and apply customizable sets of naming rules to correct and standardize gene and protein names within an annotated genome or metagenome. The database provides an intuitive web interface that allows users to create and maintain their own naming rules and organize these rules in projects that can be shared with the community.

NAMING RULES AND DATA

The PNU does not distinguish between protein or gene names: the term 'name' is used as a synonym for either. The PNU features two distinct types of naming rules: 'full matches' and 'partial matches'. A 'full match' replaces full

*To whom correspondence should be addressed. Tel: +1 301 795 7763; Fax: +1 301 294 3142; Email: jgoll@jcvi.org

Table 1. List of partial match actions

Action	Match Value	Replace Value	Example Input	Example Output
full replace	DUF	conserved hypothetical protein	hypothetical protein (DUF 1092)	conserved hypothetical protein
partial replace	7-DHC	7-dehydrocholesterol	7-DHC reductase	7-dehydrocholesterol reductase
remove	homolog	N/A	putative repressor homolog	putative repressor
merge duplicates	outative	N/A	putative kinase putative	putative kinase
move to beginning	putative	N/A	acyltransferase, putative	putative acyltransferase
move to end	putative	N/A	putative calicivirin	calicivirin, putative
regular expr. warning	/Salmonella/i	N/A	Salmonella invasin chaperone	WARNING
regular expr. local	/acyl-[cC]o[aA]/acyl-CoA/		acyl-coa dehydrogenase	acyl-CoA dehydrogenase
regular expr. global	/[Gg]nat family/GNAT family/g		acetyltransferase, Gnat family	acetyltransferase, GNAT family

For full and partial replace actions, users need to enter two input fields (match and replace value), while the other actions need only one input field. Perl-styled regular expressions can be used for the three regular expression actions. The example input and output columns demonstrate the respective action. All may match multiple names.

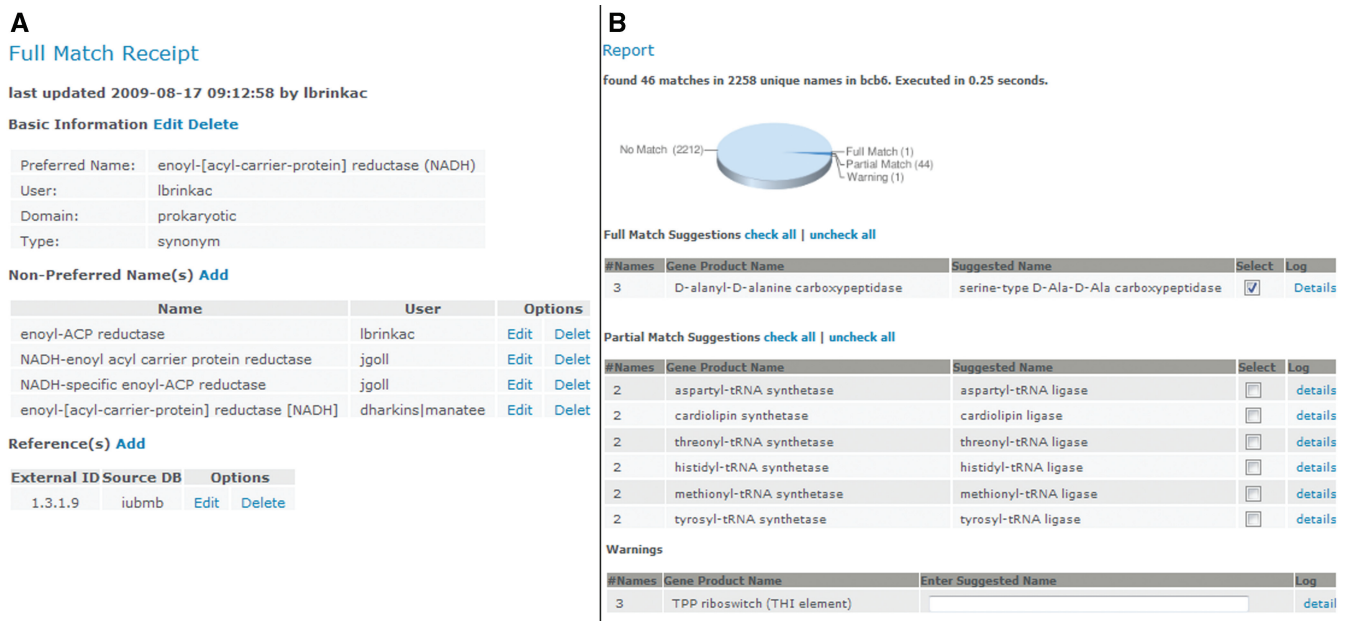

Figure 1. Screenshots of the user interface (**A**) Full match entry: this 'full match' entry links four nonpreferred names to one preferred name, here 'enoyl-[acyl-carrier-protein] reductase (NADH)'. The preferred name may be linked to an external reference, here EC 1.3.1.9 of the IUBMB (9). (**B**) PNU report: the report provides basic statistics in the heading. The table contains five columns: the number of entries for the respective input name, the input name, the PNU naming suggestion, a user confirmation check box and a link to further details. The bottom row in the figure represents a warning. If the user chooses to change the name associated with the warning, they can input the new name in the blank field under 'Enter Suggested Name'. Checked and entered names will then be used to correct and update the imported file.

name A (nonpreferred name, e.g. a synonym or misspelling) with full name B (preferred name), while a 'partial match' matches only a component of the name. Partial matches either trigger a partial name change or a 'warning'. A 'warning' allows the user to flag a matched name as suspicious and enter an alternative name when checking names (Figure 1B). A summary of all 'partial match' actions is given in Table 1.

The deployed public version of the PNU comes preloaded with 11 115 rules (577 'partial matches' and 10 538 'full matches'). Of these, 3080 have been manually curated by expert annotators and 7458 'full matches' are synonym pairs from the IUBMB database (9). New JCVI rules are continuously added, improved and made available through the PNU by JCVI analysts. Users can enter and modify rules by setting up their own PNU account via the web interface, detailed below.

USER INTERFACE

Entering rules

Users can create their own PNU account which will allow them to customize their work environment. During the account-creation process, users have the option either to enter their rules from scratch or to build upon the most current JCVI rules in the PNU database. These will then be imported to the user's profile. After this initial set up,

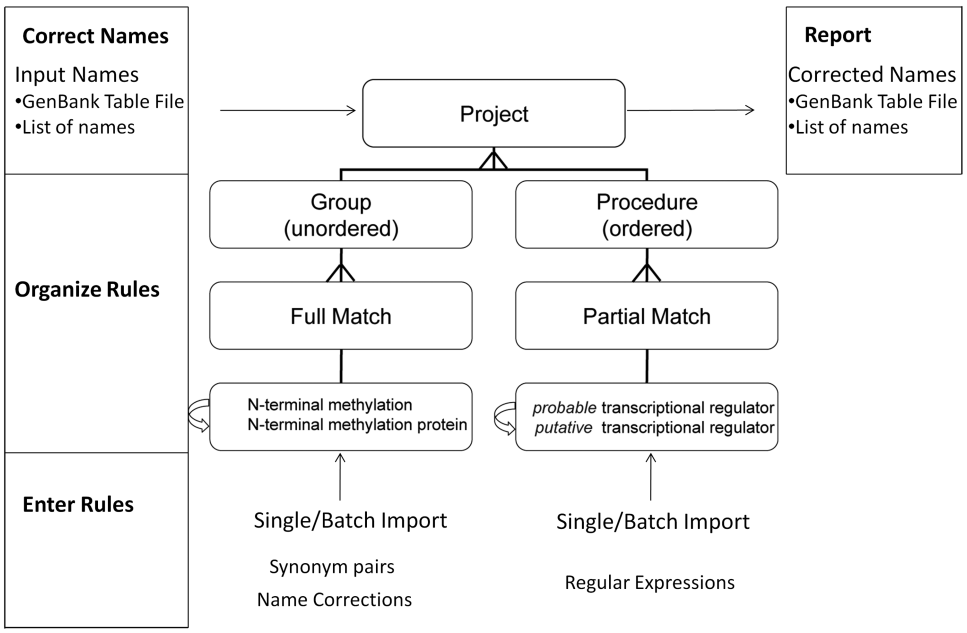

Figure 2. Overview of PNU use cases and project customization. Rules are entered via the web interface (either one by one or in a batch) and may be organized into groups, procedures and projects. Projects specify the set of rules that are used to correct names in input files. The PNU report allows users to verify name changes before correcting names (Figure 2B).

users can create their own rules (entered one by one or uploaded in batch) or modify existing ones (Figure 1A).

Organizing rules

Rules are organized into PNU 'projects', with the goal of helping the user to organize, share and apply rules. A project may contain several 'groups' and 'procedures' (Figure 1). Each 'group' contains several 'full matches' while each 'procedure' contains several 'partial matches'. The order of 'partial matches' in a 'procedure' matters as 'partial matches' are executed sequentially, i.e. the output of the first 'partial match' becomes the input of the second 'partial match' and so forth. The interface allows users to adjust the order of 'procedures' and 'partial matches'. The following constraints apply for 'full matches': a nonpreferred name cannot match an existing preferred name and vice versa. For all other types ('partial matches', 'groups', 'procedures' and 'projects'), the name must be unique. Users can share projects with the community by checking its 'public project' attribute.

Correcting names

The web interface provides an easy to use reporting tool to check names against a set of naming rules stored in a PNU project. By default, the JCVI project is selected. Users can apply their own custom PNU project or select from other shared projects. The PNU report lists the overall number of matches, 'full matches', 'partial matches' and 'warnings' that have been found among the set of unique input names (Figure 1B). Each row represents a suggested naming operation including the number of input entries with the respective name and the PNU suggested name. For each 'warning', the user can enter an alternative name in a text box. After the user has accepted relevant replacements and entered alternative names for 'warnings', a file can be downloaded with the original names corrected and replaced.

DISCUSSION

In this article, we have presented the PNU, a new web-based database for storing and applying protein naming rules. The PNU allows users to correct names in an automated fashion, leveraging curated JCVI names and incorporating their own. This will help relieve researchers from extensive manual curation of their genomes. The option to correct names in GenBank table files will aid researchers in submitting GenBank-acceptable names on the first attempt. We are reviewing past and current genome submissions for common issues flagged by GenBank to constantly improve the JCVI rule base. However, the JCVI project is only one take on naming and others are entitled to create and share their own projects. To allow users to apply rules programmatically, we plan to implement a PNU web services interface. Finally, users are requested to suggest additional features of interest.

AVAILABILITY OF THE DATABASE

A database schema (Supplementary Figure S1), a MySQL dump file and a tab delimited list of JCVI 'full matches' are available for download at: http://www.jcvi.org/pn-utility/download.php.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Funding for open access charge: National Institute of Allergy and Infectious Disease (contract HHSN266200400038C).

Conflict of interest statement. None declared.

REFERENCES

1. Fundel,K. and Zimmer,R. (2006) Gene and protein nomenclature in public databases. *BMC Bioinformatics*, **7**, 372.
2. The UniProt Consortium. (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, **37**, D169–D174.
3. Eyre,T.A., Ducluzeau,F., Sneddon,T.P., Povey,S., Bruford,E.A. and Lush,M.J. (2006) The HUGO Gene Nomenclature Database, 2006 updates. *Nucleic Acids Res.*, **34**, D319–D321.
4. Dwight,S.S., Balakrishnan,R., Christie,K.R., Costanzo,M.C., Dolinski,K., Engel,S.R., Feierbach,B., Fisk,D.G., Hirschman,J., Hong,E.L. *et al.* (2004) Saccharomyces genome database: underlying principles and organisation. *Brief Bioinform.*, **5**, 9–22.
5. Drysdale,R.A. and Crosby,M.A. (2005) FlyBase: genes and gene models. *Nucleic Acids Res.*, **33**, D390–D395.
6. Bult,C.J., Eppig,J.T., Kadin,J.A., Richardson,J.E. and Blake,J.A. (2008) The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Res.*, **36**, D724–D728.
7. Dwinell,M.R., Worthey,E.A., Shimoyama,M., Bakir-Gungor,B., DePons,J., Laulederkind,S., Lowry,T., Nigram,R., Petri,V., Smith,J. *et al.* (2009) The Rat Genome Database 2009: variation, ontologies and pathways. *Nucleic Acids Res.*, **37**, D744–D749.
8. Liu,H., Hu,Z.Z., Zhang,J. and Wu,C. (2006) BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics*, **22**, 103–105.
9. McDonald,A.G., Boyce,S. and Tipton,K.F. (2009) ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic Acids Res.*, **37**, D593–D597.