DOI 10.1002/pmic.200701194

Proteomics 2008, *8,* 1756–1757

REPORT Proteomics FASTA Archive and Reference Resource

Jayson A. Falkner, James A. Hill and Philip C. Andrews*

Department of Biological Chemistry, University of Michigan, Ann Arbor, MI, USA

A FASTA file archive and reference resource has been added to ProteomeCommons.org. Motivation for this new functionality derives from two primary sources. The first is the recent FASTA standardization work done by the Human Proteome Organization's Proteomics Standards Initiative (HUPO-PSI). Second is the general lack of a uniform mechanism to properly cite FASTA files used in a study, and to publicly access such FASTA files post-publication. An extension to the Tranche data sharing network has been developed that includes web-pages, documentation, and tools for facilitating the use of FASTA files. These include conversion to the new HUPO-PSI format, and provisions for both citing and publicly archiving FASTA files. This new resource is available immediately, free of charge, and can be accessed at http://www.proteomecommons.org/data/fasta/. Source-code for related tools is also freely available under the BSD license.

Received: December 28, 2007
Revised: January 8, 2008
Accepted: January 9, 2008

Keywords:
FASTA / HUPO-PSI / Proteome commons / ProteomeCommons.org / Tranche

ProteomeCommons.org is an active community resource for proteomics that includes aggregation of proteomics-related news, tools, and data. Recent guidelines by several journals and general demand from researchers inspired creation of the Tranche data sharing network for scientific data (http://tranche.proteomecommons.org), which is focused on archiving proteomics data. ProteomeCommons.org continues to host and support development of the Tranche project, and now the data section of ProteomeCommons.org is primarily a search enabled index of resources hosted on Tranche. Described here is new functionality built upon Tranche with the explicit purpose of providing easy access to FASTA files as well as simplifying publication of FASTA files and creating proper references for published FASTA files. Tranche provides virtually unlimited storage space and an elegant digital hash scheme for citing stored data and verifying that data has not changed since publication. This new resource provides a complete set of documentation and an intuitive interface for aiding publication of FASTA files *via* Tranche. Additionally, modules for Tranche have been created that will do common FASTA file manipulations, including concatenation of files, generation of decoy sequences, and conversion into the HUPO-PSI FASTA format. Finally, ProteomeCommons.org also now actively archives and versions the more commonly used FASTA files from various protein sequence databases.

The primary results of this work are the web-pages added to ProteomeCommons.org found at http://www.proteomecommons.org/data/fasta/. These pages include the following:

(i) A brief history of the FASTA format and PSI standardization
(ii) Guide for uploads and downloads of FASTA files
(iii) Guide for proper citation of FASTA files in manuscripts
(iv) Guide for using the Tranche FASTA module including:

Correspondence: Dr. Philip C. Andrews, 300 N. Ingalls, 11th Fl. Rm. 1198, Ann Arbor, MI 48104, USA
E-mail: andrewsp@umich.edu
Fax: +1-734-615-4941

Abbreviation: HUPO-PSI, HUPO Proteomics Standards Initiative

* Additional corresponding author: Jayson A. Falkner, E-mail: jfalkner@umich.edu

(a) Merging multiple FASTA files
(b) Creating decoy sequences for FDR estimation
(c) Converting to the HUPO-PSI FASTA format

(v) Links to related FASTA resources

In addition to the new tools and documentation we have also archived existing, commonly used FASTA files. The archival work is not exclusive and suggestions are welcome for additional resources to archive. Currently the list of archived FASTA protein databases includes:

(i) Swiss-Prot/Tremble (http://expasy.org/sprot/)
(ii) National Center for Biological Informatics (NCBI) nr (http://www.ncbi.nlm.nih.gov/)
(iii) International Protein Index (IPI) (http://www.ebi.ac.uk/IPI/)
(iv) TheGPM.org's cRAP (http://www.thegpm.org/crap/)

Finally, updates have been made to Wikipedia's FASTA format page (http://en.wikipedia.org/wiki/FASTA_format). These updates are intended to aid in the "Googleability" of the HUPO-PSI FASTA work, and the Wikipedia page also directs visitors to ProteomeCommons.org if they wish to use the tools mentioned above.

ProteomeCommons.org, Tranche, and the FASTA resource described above are primarily supported by the National Resource for Proteomics and Pathways (NRPP) sponsored in part by NCRR grant # P41-RR018627 and NCI subcontract #27XS115. Feedback relating to any of these resources or participation interest in maintaining the aforementioned resources is welcomed. Please send related communications to the public ProteomeCommons.org e-mail list at http://groups.google.com/group/ProteomeCommons.

The authors have declared no conflict of interest.

