BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 10 2007, pages 1282–1288
doi:10.1093/bioinformatics/btm098

Databases and ontologies

UniRef: comprehensive and non-redundant UniProt reference clusters

Baris E. Suzek*, Hongzhan Huang, Peter McGarvey, Raja Mazumder and Cathy H. Wu

Protein Information Resource, Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, Washington, DC 20007, USA

Received on January 25, 2007; revised on March 2, 2007; accepted on March 7, 2007

Advance Access publication March 22, 2007

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Redundant protein sequences in biological databases hinder sequence similarity searches and make interpretation of search results difficult. Clustering of protein sequence space based on sequence similarity helps organize all sequences into manageable datasets and reduces sampling bias and overrepresentation of sequences.

Results: The UniRef (UniProt Reference Clusters) provide clustered sets of sequences from the UniProt Knowledgebase (UniProtKB) and selected UniProt Archive records to obtain complete coverage of sequence space at several resolutions while hiding redundant sequences. Currently covering >4 million source sequences, the UniRef100 database combines identical sequences and subfragments from any source organism into a single UniRef entry. UniRef90 and UniRef50 are built by clustering UniRef100 sequences at the 90 or 50% sequence identity levels. UniRef100, UniRef90 and UniRef50 yield a database size reduction of ~10, 40 and 70%, respectively, from the source sequence set. The reduced redundancy increases the speed of similarity searches and improves detection of distant relationships. UniRef entries contain summary cluster and membership information, including the sequence of a representative protein, member count and common taxonomy of the cluster, the accession numbers of all the merged entries and links to rich functional annotation in UniProtKB to facilitate biological discovery. UniRef has already been applied to broad research areas ranging from genome annotation to proteomics data analysis.

Availability: UniRef is updated biweekly and is available for online search and retrieval at http://www.uniprot.org, as well as for download at ftp://ftp.uniprot.org/pub/databases/uniprot/uniref

Contact: bes23@georgetown.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Clustering of similar sequences aids in the identification of homologs, family classification, phylogenetic analysis, and related genomic and proteomic analysis tasks. A number of algorithms, tools and databases have been developed for clustering protein sequences (Enright and Ouzounis, 2000; Hobohm *et al*., 1992; Li *et al*., 2001; Mika and Rost, 2003; Paccanaro *et al*., 2006; Petryszak *et al*., 2005; Pipenbacher *et al*., 2002). We have developed and disseminated the UniProt Reference Clusters (UniRef) as one of the key components of the Universal Protein Resource (UniProt), to provide complete coverage of protein sequence space at resolutions of 100, 90 and 50% identity (Wu *et al*., 2006).

Many commonly used protein sequence databases contain redundant sequences, including identical sequences of the same length, identical subfragments and small sequence variations. Thus, the information content of these sequence databases is not linearly proportional to their size. Representative sequence databases can be used to provide the same amount of biological information with a smaller set of sequences (Park *et al*., 2000). UniRef is designed to remove sequence redundancy and reduce the number of sequences representing the sequence space. Furthermore, to improve biological interpretation of sequence data, UniRef has tight linkage of the clusters and members to functional annotation in UniProtKB. One major challenge in UniRef database development is to keep pace with the rapid expansion of protein sequence space while maintaining a biweekly update schedule. We have developed an automatic procedure to generate the UniRef databases in a timely fashion. In this article, we describe the automatic procedure, the design and content, as well as the applications of UniRef databases.

2 SYSTEM AND METHODS

2.1 Source data

To achieve comprehensive coverage of protein sequence space, the UniProtKB and UniParc (Leinonen *et al*., 2004) databases maintained by the UniProt Consortium (2007) were used to generate the sequence set needed for the UniRef databases. The datasets are: (i) all UniProtKB entries, (ii) all splice variants extracted from UniProtKB/Swiss-Prot entries and (iii) all UniParc entries that contain at least one active Ensembl (Hubbard *et al*., 2007) from human, mouse, rat, fly, dog, chicken, Fugu, Tetraodon or Xenopus species, RefSeq (Pruitt *et al*., 2007) or Protein Data Bank (PDB) (Kouranov *et al*., 2006) cross-reference, but no cross-reference to UniProtKB or UniProtKB splice variant.

*To whom correspondence should be addressed.

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/bioinformatics/article/23/10/1282/197795 by U.S. Department of Justice user on 16 August 2022

2.2 Generation of UniRef100, UniRef90 and UniRef50 clusters

The UniRef databases are generated in a hierarchical fashion (Fig. 1): UniRef100 clusters are generated first using sequences from UniProtKB and UniParc, UniRef90 clusters are then generated using UniRef100 clusters and UniRef50 clusters are generated using UniRef90 clusters.

UniRef100 clusters are created in three steps. First, the CD-HIT algorithm (Li and Godzik, 2006; Li *et al.*, 2001) is used to cluster all sequences with a 100% sequence identity threshold. Then the overlapping regions are checked to remove cluster members having gapped alignments with the longest sequence designated as the UniRef100 seed sequence. The cluster members removed in this step are treated as single-member clusters. Finally, as the CD-HIT algorithm does not process sequences <11 amino acids in length, these small fragments are tested for sequence identity and clustered when possible to achieve comprehensive coverage of the underlying sequence space. Small fragments not clustered are treated as single-member clusters.

Primary UniRef90 clusters are generated from the UniRef100 seed sequences using CD-HIT with a 90% sequence identity threshold. Likewise, primary UniRef50 clusters are generated using the UniRef90 seed sequences at 50% sequence identity. Since the primary UniRef90 and UniRef50 clusters are created with seed sequences only, additional members are added to the primary clusters to attain complete membership in the final clusters. All members of a UniRef100 cluster whose seed sequence is present in a primary UniRef90 cluster are added to the UniRef90 cluster. All members of this final UniRef90 cluster are similarly added to the appropriate primary UniRef50 cluster.

This procedure creates non-overlapping clusters, with each source sequence assigned to exactly one UniRef100, one UniRef90 and one UniRef50 cluster. Furthermore, the UniRef clusters have parent–child relationships at different identity levels, where each UniRef100 cluster and UniRef90 cluster are the child clusters of the UniRef90 and UniRef50 clusters that contain their corresponding seed sequences.

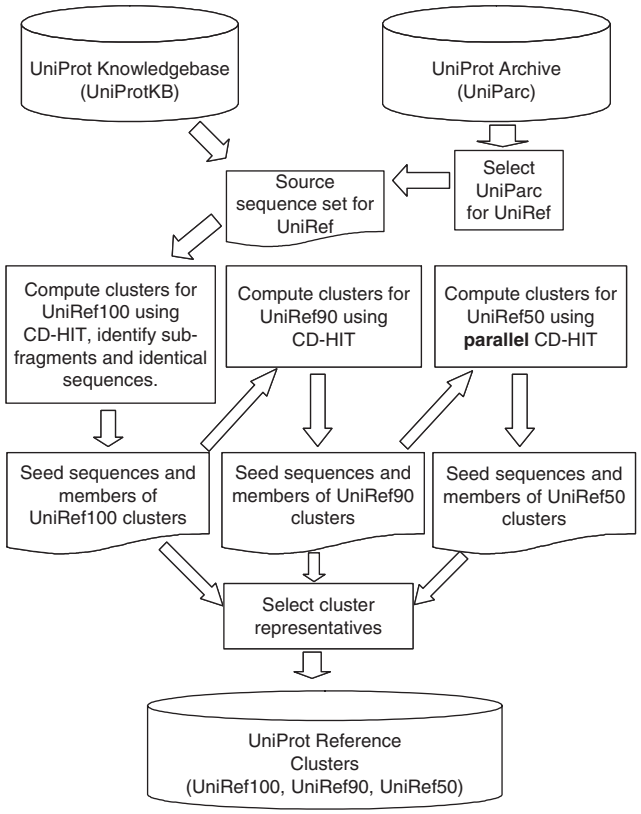

Fig. 1. Hierarchical generation of UniRef databases from UniProtKB and selected UniParc source sequences.

2.3 Optimization of UniRef creation and update procedures

Creating UniRef50 clusters by running CD-HIT at the 50% identity level is highly computationally intensive. To keep UniRef in sync with the biweekly update cycle of UniProt databases for the ever increasing number of source sequences, we developed a parallel clustering procedure using CD-HIT, as well as an incremental update procedure.

2.3.1 Parallel CD-HIT procedure The parallel procedure was developed for the creation of UniRef50 clusters using the CD-HIT algorithm. Nevertheless, the procedure can be used to cluster any large sequence set at low identity levels that requires heavy computation using CD-HIT. The parallelization method is also generally applicable to other clustering algorithms. The algorithm is described as follows (Fig. 2):

Given k nodes in a computer cluster

(1) *Sequence sorting*: sort the source sequences (all seed sequences from UniRef90) by length in descending order to generate sequence set S.

(2) *Sequence grouping*: split the S into $k+1$ parts S_i, $i=0, 1, 2, \ldots, k$, each containing roughly the same total number of residues, and none of the sequence in S_i is shorter than those in S_{i+1}.

(3) *Parallel clustering*: This step will take up to k rounds.

For the first round:

(a) Run CD-HIT at 50% level on the first group S_0 to compute the sequence cluster set C.

(b) Place sequences in each of the remaining k groups (i.e. $S_1, S_2, \ldots, S_k$) into a separate computation node and, on each node, run CD-HIT at 50% level to recruit additional member sequences to C by clustering sequences under the seed sequences of S_0.

(c) Combine the newly clustered sequences on each computation node (step b) with members under the seed sequences from S_0 (step a). This will complete the generation of the cluster set C for the seed sequences from the first group S_0.

(d) For the remaining sequences not clustered in step b, retain the sequences of the next group S_1 as the first group and combine all the remaining sequences into a new sequence set for the next round.

As in the first round, the sequences in the first group will be clustered first and all sequences in the new sequence set will be split into k parts for parallel clustering.

(4) Concatenate the clusters generated in each round to complete the generation of UniRef50.

The parallel CD-HIT procedure speeds up the creation of UniRef50 clusters by at least an order of magnitude. For example, the UniRef50 creation for 850 000 source sequences took 4 weeks without the parallel procedure and reduced to 40 h when parallelized on 30 Intel® Xeon® 2.4 GHz processors.

2.3.2 Incremental update procedure To further reduce computation time, we have recently developed a procedure for the incremental update of UniRef clusters during minor UniProt releases.

Downloaded from https://academic.oup.com/bioinformatics/article/23/10/1282/197795 by U.S. Department of Justice user on 16 August 2022

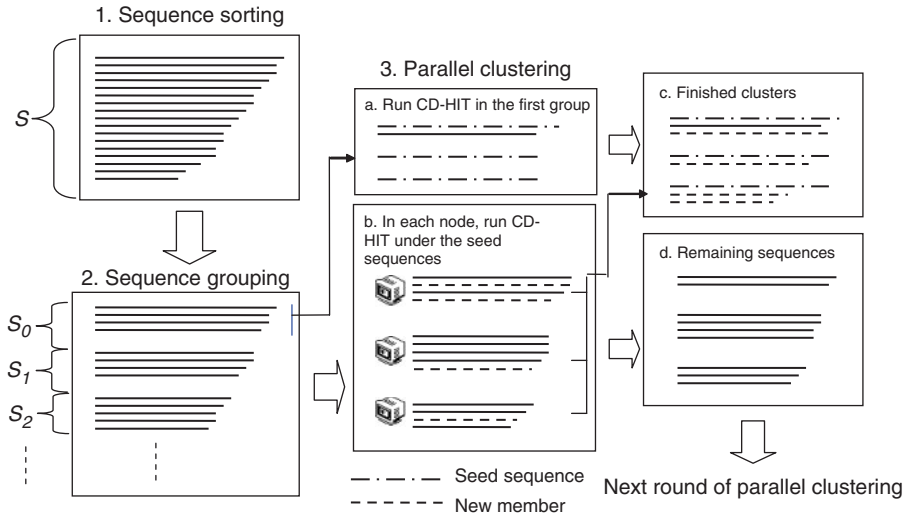

Fig. 2. Parallel clustering and creation of UniRef50 clusters using the CD-HIT algorithm.

The UniProt minor releases (major releases are numbered x.0, minor releases are x.1–x.n) usually have <10% new sequences.

We designate the source sequence set (UniProtKB and selected UniParc sequences) from the previous release S_0, the source sequence set of the new release S_1, and the UniRef100 cluster set (seed sequences and non-seed members) from the previous release C_0. The steps for incremental update are as follows:

(1) Create the new cluster set C from C_0 by removing UniRef100 non-seed members in C_0 that are no longer in S_1.

(2) Identify all entries in S_1 but not in S_0 to generate the new sequence set S.

(3) Identify seed sequences that are no longer in S_1, remove the corresponding UniRef100 clusters from the cluster set C and incorporate the non-seed members in S_1 to the new sequence set S.

(4) Cluster sequences in S under the remaining UniRef100 seed sequences in the cluster set C for recruitment of non-seed members using the incremental mode of the CD-HIT program.

(5) Cluster all remaining entries in S that cannot be clustered in step 4 using the regular (default) mode of the CD-HIT concatenate clusters computed in steps 4 and 5 to generate the final UniRef100 cluster set for the new release.

The UniRef90 and UniRef50 are then updated similarly using the new versions of UniRef100 and UniRef90 as source datasets, respectively.

The difference in the final clustering results between incremental and full update is minimal. Our data indicate that >95% of UniRef50 clusters are identical between seven consecutive incremental updates (releases 9.0–9.7) and one full update (for release 9.7), given the 15% sequence difference during these releases. This translates into <1% difference on average for each minor release.

2.4 Selection of representative member

The seed sequence produced by CD-HIT is the longest member of the cluster, since it contains the most complete sequence information for computational purposes. However, for biological discovery the longest sequence may not always be the most richly annotated or informative. There is often more biologically relevant information (e.g. name, function, cross-references) available on other cluster members. To computationally select the representative member containing the most biological information, the following rules are sequentially applied.

(1) Quality of entry annotation: order of preference is a member from UniProtKB/Swiss-Prot, UniProtKB/TrEMBL, then UniParc.

(2) Meaningful name: members with protein names that do not contain words such as 'hypothetical' or 'probable' are preferred.

(3) Organism: members from model organisms are preferred.

(4) Sequence length: longest sequence is preferred.

3 RESULTS AND DISCUSSION

3.1 Database coverage, size reduction and cluster distribution

The UniRef databases are generated and released every 2 weeks in conjunction with updates of the UniProtKB and UniParc databases. The most recent major release of the UniRef databases (release 9.0, October 31, 2006) consists of 3 867 872 (UniRef100), 2 475 002 (UniRef90) and 1 292 284 (UniRef50) clusters, respectively. The source dataset consists of 4 362 881 sequences, including 3 571 161 UniProtKB sequences and 791 720 selected UniParc sequences.

As a comprehensive, non-redundant reference database, UniRef100 is similar in scope to the nr protein sequence database produced by NCBI (Wheeler *et al.*, 2007). Most sequences in the two databases are the same since both are derived from a similar set of sequence repositories that include EMBL/GenBank/DDBJ CDS translations, UniProtKB, PIR-PSD (now merged with UniProtKB) and PDB. When we compare the sequence sets represented in compatible releases of UniRef100 (Release 9.0, October 31, 2006) and nr (November 1, 2006), we identify 185 326 sequences unique to UniRef100, of which 140 651 are from Ensembl and the rest from RefSeq and UniProtKB. The 9440 sequences unique to nr are from Protein Resource Foundation (PRF; http://www.prf.or.jp) and

Downloaded from https://academic.oup.com/bioinformatics/article/23/10/1282/197795 by U.S. Department of Justice user on 16 August 2022

RefSeq. The unique PRF entries previously not represented in UniProtKB have been reviewed and incorporated into UniProtKB. The difference in RefSeq and UniProtKB coverage is due to different update cycles at NCBI and UniProt, but nr also contains many subfragments that are merged in UniRef clusters. Indeed, we generally achieved ~4% size reduction from nr when identical subfragments were removed from nr using the UniRef100 generation algorithm.

UniRef100, UniRef90 and UniRef50 yield a database size reduction of 11, 43 and 70%, respectively, with respect to the source sequence set. The size reduction of UniRef50 with respect to UniRef100 is similar to that reported in earlier studies done on representative sequence databases (Park *et al.*, 2000). The size reduction has increased over time as seen in Figure 3, which compares growth of the source and UniRef datasets since UniRef inception. The large increase in number of UniRef clusters in release 2.4 (August 31, 2004) coincided with the incorporation of over 1 million environmental sequences into UniParc, which were used as part of the UniRef source sequence set. These environmental sequences have been removed from the UniRef source sequences since major release 4.0 (February 1, 2005) because of their minimal level of annotation. The slower smoother rate of growth seen in UniRef90 and UniRef50 during releases 2.4 and 4.0 illustrates how they are minimizing sequence redundancy and bias as the majority of new sequences cluster with existing sequences.

Figure 4 shows how UniRef databases cluster the sequence space in terms of the distribution of cluster size. We see a logarithmic distribution with an R^2 value approaching 1.0, consistent with the power law distributions that have been noted for a variety of biological phenomena including the distribution of protein domains and families (Luscombe *et al.*, 2002). Notable are the large numbers of single-member clusters at these high levels of sequence identity, ranging from 64% of all clusters in UniRef50 to 78 and 93% in UniRef90 and UniRef100. On the other end of the cluster distribution are a number of very large clusters, illustrating the high level of redundancy and overrepresentation of certain groups of sequences. Currently the largest cluster in UniRef100 consists of >1100 highly conserved Histone H3 proteins from a wide variety of arthropod species. The largest cluster in UniRef90 contains >3600 matrix 1 proteins of influenza A virus, while the largest UniRef50 cluster has ~75 000 Pol polyproteins from human, feline, and simian immunodeficiency viruses (HIV, FIV and SIV).

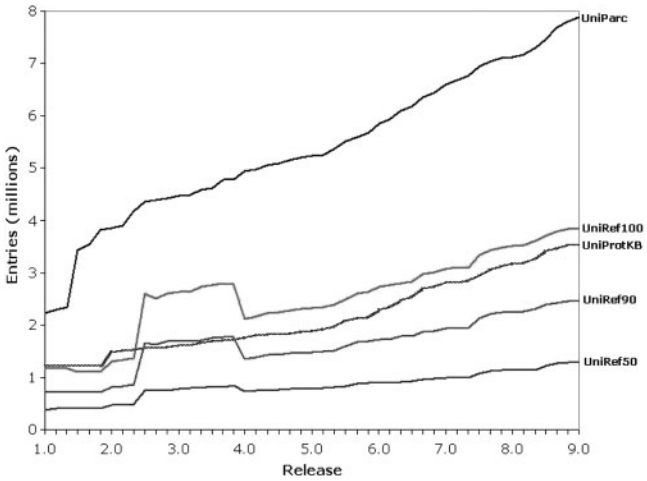

Fig. 3. Growth of UniRef databases in relationship to the UniProtKB and UniParc source databases. Environmental sequences were introduced to UniRef in release 2.4 (August 31, 2004) and later removed in release 4.0 (February 1, 2005). A colour version of this figure is available as supplementary material online.

3.2 Database content and availability

3.2.1 Content of UniRef records UniRef records contain the following information:

(1) General cluster information:

- UniRef ID (cluster ID derived from the accession number of the representative member, e.g. UniRef90_P69905 used the UniProtKB primary accession P69905)
- Cluster name (derived from the protein name of the representative member)
- Member count (number of source sequences in the cluster)
- Common taxonomy (the lowest common taxonomic node shared by all the members with the corresponding NCBI taxonomy identifier)
- Representative member (accession number of the representative sequence)
- Seed member (accession number of the seed sequence)
- Parent clusters (UniRef ID of the parent UniRef90/UniRef 50 clusters)
- Protein sequence of the representative member with sequence length and CRC64 checksum

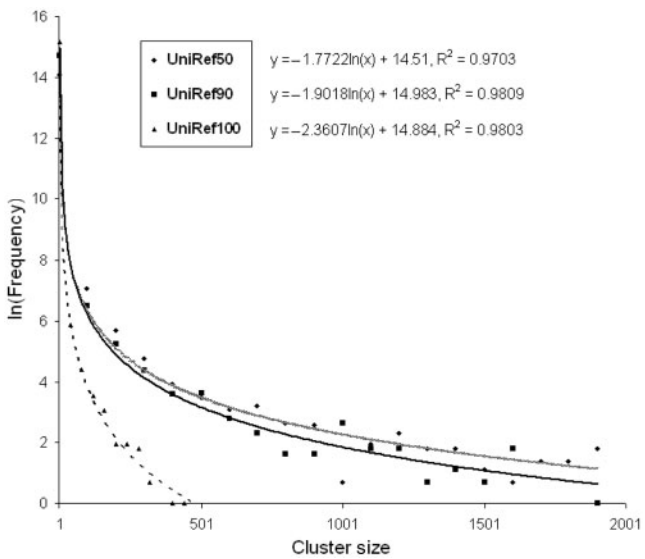

Fig. 4. The power law distribution of UniRef clusters in terms of cluster size and frequency.

Downloaded from https://academic.oup.com/bioinformatics/article/23/10/1282/197795 by U.S. Department of Justice user on 16 August 2022

(2) Cluster membership (information for source sequences):

- Source database (UniProtKB or UniParc) and the corresponding ID and accession number of each sequence
- UniRef identifiers (ID of the child UniRef100/UniRef90 clusters where the source sequence belongs)
- Protein name (extracted from UniProtKB or, for UniParc entries, extracted from the underlying RefSeq, PDB or Ensembl databases. Some UniParc entries are missing protein name.)
- Organism name and NCBI taxonomy ID (In cases where a UniParc entry contains multiple Ensembl entries, the source information is prioritized by model organism in the following order: human, mouse, rat, fly, dog, chicken, Fugu, Tetraodon and Xenopus. Some UniParc entries are missing organism information.)
- Sequence length

3.2.2 Availability The biweekly releases of the UniRef databases are formatted in XML and downloadable by FTP from ftp.uniprot.org. A FASTA format version containing only the name and sequence of representative members is also available for download. The UniProt web site allows direct retrieval of the UniRef records in multiple formats using the following search strings (URLs):

- HTML: http://www.uniprot.org/entry/UniRef90_P69905
- XML: http://www.uniprot.org/entry/UniRef90_P69905?format=xml
- FASTA: http://www.uniprot.org/entry/UniRef90_P69905?format=fasta

The web site also supports searches on various attributes of the UniRef clusters, including UniRef cluster ID, protein names, organism names and database identifiers. Note that since all UniRef database clusters are recreated every 2 weeks based on an updated sequence set, clusters and their representative sequences may change from release to release, resulting in different cluster IDs. Therefore, the best way to search for a cluster of interest is to use a member's accession number. The web site provides an easy way to view the information in a UniRef cluster and links all members to source UniProtKB/UniParc entries with additional information and annotation. The UniRef datasets are also available for BLAST searches on the UniProt web site. Batch retrieval and ID mapping for UniRef IDs are available at http://www.pir.uniprot.org/search/batch.shtml and http://www.pir.uniprot.org/search/idmapping.shtml, respectively.

3.3 Applications of UniRef databases

UniRef is currently being used worldwide by researchers for a broad range of applications in the areas of functional annotation, family classification, systems biology, structural genomics, phylogenetic analysis and mass spectrometry. The most common use for the UniRef100 and UniRef90 datasets has been to exploit the non-redundancy and informative description lines and links to UniProtKB to assist annotation of genome sequencing projects (Cannon *et al*., 2005; Childs *et al*., 2007; Joron *et al*., 2006; Mudge *et al*., 2005; Pavy *et al*., 2005, 2006; Ramirez *et al*., 2005; Sato *et al*., 2005; Stover *et al*., 2005; Yan *et al*., 2006) and in software development for functional annotation (Koski *et al*., 2005). For example, UniRef is used as a reference database for the BLAST search to assign functional annotation for the TIGR plant transcript assemblies (Childs *et al*., 2007).

Many groups take advantage of the reduced sampling bias in UniRef to develop and improve profile-based models (PSI-BLAST, HMMER) for specific gene and structural motifs (Casbon and Saqi, 2006; Kinjo and Nishikawa, 2006; McGuffin *et al*., 2006; Peng *et al*., 2005; Rojas *et al*., 2005; Wang and Samudrala, 2006). Several studies have utilized the UniRef clusters to aid in development of their own gene-family- or domain-specific clusters (Flaus *et al*., 2006; Novatchkova *et al*., 2006; Silverstein *et al*., 2005). Evolutionary and phylogenetic studies (Barnosa *et al*., 2006) have also found uses for UniRef, as have a number of studies in the areas of structural genomics (Fernandez-Fuentes and Fiser, 2006; Jakobsson *et al*., 2005; Maurer-Stroh and Eisenhaber, 2005; Overton and Barton, 2006), systems biology (Ng *et al*., 2006) and global proteome analysis (Frith, 2006). In particular, many above applications use UniRef as the representative of the universe of protein sequences based on sequence conservation captured in UniRef50 clusters. For instance, UniRef50 is used to select non-redundant PDB structures and to develop scoring for structural genomics target ranking (Overton and Barton, 2006). In another example, UniRef50 is used to derive average protein physical properties for the prediction of protein prenylation motifs (Maurer-Stroh and Eisenhaber, 2005).

UniRef is also useful for mass spectrometry (MS)-based proteomics data analysis. As UniRef100 is the most comprehensive set of unique, non-redundant protein sequences, including sequences of splice variants, and contains links to UniProt annotation on post-translational modifications, a number of groups have used UniRef100 as the underlying database for protein identification from tandem mass spectra using different search engines (Gagne *et al*., 2005; Perkins *et al*., 2006; Vgenopoulou *et al*., 2006). UniRef can also be utilized for functional analysis and profiling of proteomic data, including reanalysis of published data, where proteins have been identified based on other protein sequence databases. Both UniRef100 and UniRef90 have been used in the iProXpress proteomic expression analysis system for the biological analysis of several proteomes (Chi, 2006; Hu *et al*., 2007; Huang *et al*., 2007). As peptides of the same protein from the MS experiments may be assigned to more than one protein entry, sequences mapped to the same UniRef100 or UniRef90 clusters are examined to select the most richly annotated UniProtKB entry from the same source organism.

4 CONCLUSIONS

The UniRef databases have been created to provide a complete coverage of the protein sequence space, while removing sequence redundancy and reducing the number of sequences, thereby increasing the speed of similarity searches and simultaneously

Downloaded from https://academic.oup.com/bioinformatics/article/23/10/1282/197795 by U.S. Department of Justice user on 16 August 2022

improving detection of distant relationships with a more even sampling of sequence space. It further facilitates biological discovery by providing information such as protein name and taxonomy in the UniRef clusters, as well as the direct reference to UniProtKB with rich functional annotation. The biweekly update cycle ensures that UniRef provides up-to-date clusters, keeping pace with the rapid growth of protein sequences. These major features—the comprehensive sequence coverage, the reduction of sequence redundancy and the tight linkage to functional annotation in UniProtKB—have allowed UniRef to be utilized for a variety of applications ranging from genome annotation and family classification to phylogenic analysis and proteomics data analysis.

ACKNOWLEDGEMENTS

This project is supported by the UniProt grant U01 HG02712 from the National Institutes of Health. The authors are grateful to Dr Weizhong Li for his support on CD-HIT, and to our UniProt Consortium collaborators at the Swiss Institute of Bioinformatics and European Bioinformatics Institute for their fruitful discussions. The funding for the Open Access publication charges was provided by UniProt grant U01 HG02712 from National Institutes of Health, USA.

Conflict of Interest: none declared.

REFERENCES

Barnosa,D. *et al.* (2006) Divergent paralogous in Uniref50 enriched-COG clusters depicted by Phylip neighbor trees rooted with Taxbrowser tables. *Abstract ISMB2006*, Retrieved September 30, 2006 from http://ismb2006.cbi.cnptia.embrapa.br/poster_abstract_lb.php?id=LB-56.

Cannon,S.B. *et al.* (2005) Databases and information integration for the Medicago truncatula genome and transcriptome. *Plant Physiol.*, **138**, 38–46.

Casbon,J.A. and Saqi,M.A. (2006) On single and multiple models of protein families for the detection of remote sequence relationships. *BMC Bioinformatics*, **7**, 48.

Chi,A. *et al.* (2006) Proteomic and bioinformatic characterization of the biogenesis and function of melanosomes. *J. Proteome Res.*, **5**, 3135–3144.

Childs,K.L. *et al.* (2007) The TIGR Plant Transcript Assemblies database. *Nucleic Acids Res.*, **35**, D846–D851.

Enright,A.J. and Ouzounis,C.A. (2000) GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics*, **16**, 451–457.

Fernandez-Fuentes,N. and Fiser,A. (2006) Saturating representation of loop conformational fragments in structure databanks. *BMC Struct. Biol.*, **6**, 15.

Flaus,A. *et al.* (2006) Identification of multiple distinct Snf2 subfamilies with conserved structural motifs. *Nucleic Acids Res.*, **34**, 2887–2905.

Frith,M.C. *et al.* (2006) The abundance of short proteins in the mammalian proteome. *PLoS Genet.*, **2**, e52.

Gagne,J.P. *et al.* (2005) Proteome profiling of human epithelial ovarian cancer cell line TOV-112D. *Mol. Cell. Biochem.*, **275**, 25–55.

Hobohm,U. *et al.* (1992) Selection of representative protein data sets. *Protein Sci.*, **1**, 409–417.

Hu,Z.Z. *et al.* (2007) Comparative bioinformatics analyses and profiling of lysosome-related organelle proteomes. *Int. J. Mass Spectrom.*, **259**, 147–160.

Huang,H. *et al.* (2007) Challenges and solutions in proteomics. *Curr. Genomics*, **8**, 21–28.

Hubbard,T.J. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.

Jakobsson,E. *et al.* (2005) Structure of human semicarbazide-sensitive amine oxidase/vascular adhesion protein-1. *Acta Crystallogr. D. Biol. Crystallogr.*, **61**, 1550–1562.

Joron,M. *et al.* (2006) A conserved supergene locus controls colour pattern diversity in heliconius butterflies. *PLoS Biol.*, 4.

Kinjo,A.R. and Nishikawa,K. (2006) CRNPRED: highly accurate prediction of one-dimensional protein structures by large-scale critical random networks. *BMC Bioinformatics*, **7**, 401.

Koski,L.B. *et al.* (2005) AutoFACT: an automatic functional annotation and classification tool. *BMC Bioinformatics*, **6**, 151.

Kouranov,A. *et al.* (2006) The RCSB PDB information portal for structural genomics. *Nucleic Acids Res.*, **34**, D302–D305.

Leinonen,R. *et al.* (2004) UniProt archive. *Bioinformatics*, **20**, 3236–3237.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Li,W. *et al.* (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.

Luscombe,N.M. *et al.* (2002) The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. *Genome Biol.*, **3**, RESEARCH0040.

Maurer-Stroh,S. and Eisenhaber,F. (2005) Refinement and prediction of protein prenylation motifs. *Genome Biol.*, **6**, R55.

McGuffin,L.J. *et al.* (2006) High throughput profile-profile based fold recognition for the entire human proteome. *BMC Bioinformatics*, **7**, 288.

Mika,S. and Rost,B. (2003) UniqueProt: creating representative protein sequence sets. *Nucleic Acids Res.*, **31**, 3789–3791.

Mudge,J. *et al.* (2005) Highly syntenic regions in the genomes of soybean, Medicago truncatula, and Arabidopsis thaliana. *BMC Plant Biol.*, **5**, 15.

Ng,A. *et al.* (2006) pSTIING: a 'systems' approach towards integrating signalling pathways, interaction and transcriptional regulatory networks in inflammation and cancer. *Nucleic Acids Res.*, **34**, D527–D534.

Novatchkova,M. *et al.* (2006) DOUTfinder – identification of distant domain outliers using subsignificant sequence similarity. *Nucleic Acids Res.*, **34**, W214–W218.

Overton,I.M. and Barton,G.J. (2006) A normalised scale for structural genomics target ranking: the OB-Score. *FEBS Lett.*, **580**, 4005–4009.

Paccanaro,A. *et al.* (2006) Spectral clustering of protein sequences. *Nucleic Acids Res.*, **34**, 1571–1580.

Park,J. *et al.* (2000) RSDB: representative protein sequence databases have high information content. *Bioinformatics*, **16**, 458–464.

Pavy,N. *et al.* (2005) Generation, annotation, analysis and database integration of 16 500 white spruce EST clusters. *BMC Genomics*, **6**, 144.

Pavy,N. *et al.* (2006) Automated SNP detection from a large collection of white spruce expressed sequences: contributing factors and approaches for the categorization of SNPs. *BMC Genomics*, **7**, 174.

Peng,K. *et al.* (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, **7**, 208.

Perkins,D.N. *et al.* (2006) Mascot online help manual, Retrieved November 28, 2006, from http://www.matrixscience.com/help/seq_db_setup_uniref.html

Petryszak,R. *et al.* (2005) The predictive power of the CluSTr database. *Bioinformatics*, **21**, 3604–3609.

Pipenbacher,P. *et al.* (2002) ProClust: improved clustering of protein sequences with an extended graph-based approach. *Bioinformatics*, **18** (Suppl. 2), S182–S191.

Pruitt,K.D. *et al.* (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.

Ramirez,M. *et al.* (2005) Sequencing and analysis of common bean ESTs. Building a foundation for functional genomics. *Plant Physiol.*, **137**, 1211–1227.

Rojas,A.M. *et al.* (2005) Death inducer obliterator protein 1 in the context of DNA regulation. Sequence analyses of distant homologues point to a novel functional role. *FEBS J.*, **272**, 3505–3511.

Sato,S. *et al.* (2005) Comprehensive structural analysis of the genome of red clover (Trifolium pratense L.). *DNA Res.*, **12**, 301–364.

Silverstein,K.A. *et al.* (2005) Genome organization of more than 300 defensin-like genes in Arabidopsis. *Plant Physiol.*, **138**, 600–610.

Stover,N.A. *et al.* (2006) Tetrahymena Genome Database (TGD): a new genomic resource for Tetrahymena thermophila research. *Nucleic Acids Res.*, **34**, D500–D503.

The UniProt Consortium (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **35**, D193–D197.

Vgenopoulou,I. *et al.* (2006) Specific modification of a Na^+ binding site in NADH:quinone oxidoreductase from Klebsiella pneumoniae with dicyclohexylcarbodiimide. *J. Bacteriol.*, **188**, 3264–3272.

Downloaded from https://academic.oup.com/bioinformatics/article/23/10/1282/197795 by U.S. Department of Justice user on 16 August 2022

Wang,K. and Samudrala,R. (2006) Incorporating background frequency improves entropy-based residue conservation measures. *BMC Bioinformatics*, **7**, 385.

Wheeler,D.L. *et al*. (2007) Database resources of the national center for biotechnology information. *Nucleic Acids Res*., **35**, D5–D12.

Wu,C.H. *et al*. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*., **34**, D187–D191.

Yan,H. *et al*. (2006) Genomic and genetic characterization of rice Cen3 reveals extensive transcription and evolutionary implications of a complex centromere. *Plant Cell*, **18**, 2123–2133.

Downloaded from https://academic.oup.com/bioinformatics/article/23/10/1282/197795 by U.S. Department of Justice user on 16 August 2022