Nucleic Acids Research, 2005, Vol. 33, Database issue ***D501–D504***
doi:10.1093/nar/gki025

NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins

Kim D. Pruitt*, Tatiana Tatusova and Donna R. Maglott

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Rm 6An.12J, 45 Center Drive, Bethesda, MD 20892-6510, USA

Received September 15, 2004; Revised and Accepted September 21, 2004

ABSTRACT

The National Center for Biotechnology Information (NCBI) Reference Sequence (RefSeq) database (http://www.ncbi.nlm.nih.gov/RefSeq/) provides a non-redundant collection of sequences representing genomic data, transcripts and proteins. Although the goal is to provide a comprehensive dataset representing the complete sequence information for any given species, the database pragmatically includes sequence data that are currently publicly available in the archival databases. The database incorporates data from over 2400 organisms and includes over one million proteins representing significant taxonomic diversity spanning prokaryotes, eukaryotes and viruses. Nucleotide and protein sequences are explicitly linked, and the sequences are linked to other resources including the NCBI Map Viewer and Gene. Sequences are annotated to include coding regions, conserved domains, variation, references, names, database cross-references, and other features using a combined approach of collaboration and other input from the scientific community, automated annotation, propagation from GenBank and curation by NCBI staff.

INTRODUCTION

RefSeq is a public database of nucleotide and protein sequences with corresponding feature and bibliographic annotation. The RefSeq database is built and distributed by the NCBI, a division of the National Library of Medicine located at the US National Institutes of Health. NCBI makes RefSeq publicly available, at no cost, over the internet via FTP, Entrez query (1), Basic Local Alignment Search Tool (BLAST) (2,3) programs, and incorporation in a wide range of NCBI resources.

NCBI builds RefSeq from the sequence data available in the archival database GenBank (4), which is a comprehensive public repository of sequences submitted to, and exchanged among, GenBank in the US, the EMBL Data Library in the UK and the DNA Data Bank of Japan. In addition, the annotated RefSeq record and/or supplementary information may be provided by multiple collaborations established with nomenclature groups, model organism databases and other facets of the scientific community. RefSeq records indicate the source GenBank data, include references and annotations relevant to the gene, transcript and protein, and indicate curation with attribution to the curation group.

The RefSeq collection is unique in providing a curated, non-redundant, explicitly linked nucleotide and protein database representing significant taxonomic diversity. Genomic and protein sequence datasets are provided for the majority of organisms included; transcript records are currently provided for a subset of the eukaryotic collection. The RefSeq database provides a critical foundation for integrating sequence, genetic and functional information, and is used internationally as a standard for genome annotation. The collection is curated on an ongoing basis by collaborating groups and by NCBI staff. Sequence records are presented in a standard format and are subject to computational validation.

DISTINCTION FROM GENBANK

The RefSeq collection is derived from the primary submissions available in GenBank. GenBank is a redundant archival database that represents sequence information generated at different times, and may represent several alternate views of the protein, names or other information. In contrast, RefSeq represents a nearly non-redundant collection that is a synthesis and summary of available information, and represents the 'current' view of the sequence information, names and other annotations.

RefSeq records can be distinguished from GenBank records by the format of the accession series. RefSeq accession numbers are formatted as two alphabetic characters, followed by an

*To whom correspondence should be addressed. Tel: +1 301 435 5950; Fax: +1 301 480 2918; Email: pruitt@ncbi.nlm.nih.gov

Table 1. Annual growth of the RefSeq collection

Date	FTP release	Species	Number of records		
			Genomic	Transcript	Protein
6/30/2003	1	2005	64 729	211 803	785 143
7/5/2004	6	2467	68 592	247 639	1 050 975

underscore ('_'), optionally followed by four alphabetic characters (specific to the NZ_ prefix), followed by six, eight or nine numerals. GenBank accessions never include an underscore. Different alphabetic prefixes have implied meaning in terms of both the process of generation and the type of molecule represented. A full definition of the RefSeq accession numbers is available on the RefSeq Web site (http://www.ncbi.nlm.nih.gov/RefSeq/key.html#accessions).

GROWTH

The RefSeq database continues to grow in pace with the large-scale genome and cDNA sequencing projects (see Table 1). As new complete genome assemblies become available, they are incorporated into the RefSeq collection. Most organisms are represented in the collection only after some genomic sequence data (nuclear, plastid, mitochondrial or other genomic molecules) becomes available; however, transcript and protein records may be provided for a subset of eukaryotic model organisms prior to the availability of genomic sequence data.

ANNOTATION

Annotation of RefSeq records originates from several sources including the original GenBank submission, collaborating groups, NCBI computational analysis, user feedback and manual curation at NCBI. For example, collaboration supports the RefSeq representation of *Saccharomyces cerevisiae*, *Drosophila melanogaster* and *Arabidopsis thaliana*, which are directly contributed by the Saccharomyces Genome Database (SGD)(5), FlyBase (6) and The Institute for Genomic Research (TIGR), respectively. Similarly, the entire viral RefSeq collection is reviewed and curated by the NCBI Viral Genome Advisors group. See the RefSeq Collaborators page for more information about contributions from collaborators (http://www.ncbi.nlm.nih.gov/RefSeq/collaborators.html). All RefSeq records include explicit cross-links between the nucleotide and protein cognates and to Entrez Gene (7), which provides gene-oriented access to the RefSeq collection. Additional links, annotated as 'db_xref' notations, are provided on some records to organism-specific genome resources such as Mouse Genome Informatics (MGI) (8) or FlyBase.

For other species, including *Apis mellifera* (honey bee), *Gallus gallus* (chicken), *Homo sapiens* (human), *Mus musculus* (mouse) and *Rattus norvegicus* (rat), genome annotation is provided by a NCBI computational process that utilizes transcript alignments, protein support and a hidden Markov model (HMM) *ab initio* prediction algorithm (see the NCBI Handbook; http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books). Genomic RefSeq records that are annotated by this process represent genes, transcripts and proteins, and include additional feature annotation to represent STS markers. The available RefSeq transcript dataset, with the 'NM_' accession prefix, is an important reagent in this annotation pipeline.

Comprehensive representation of the proteins, explicitly linked to a RefSeq nucleotide record, is a major focus of the RefSeq project. The goal is to represent the full-length protein product; however, partial protein products are represented for some genomes when partial protein annotation is contributed by a collaborator or when proteins are predicted from incomplete genome sequence data. Proteins are annotated by computation and curation. Conserved domains are calculated by an automatic process using data maintained in the NCBI Conserved Domain Database (CDD) (9); this annotation provides hints about possible function. Likewise, variation features that are located in the coding region are automatically calculated from data available in the NCBI dbSNP database (10). Additional features including Enzyme Commission (EC) numbers, other landmark regions of the protein sequence and references may be added by curation either by an external collaborator or by NCBI staff.

Transcript records are provided for a subset of eukaryotic species, including those in the Chordata taxonomic lineage, to represent protein-coding sequences, transcribed pseudogenes, ribosomal RNAs and other small RNAs. Annotation results from a mixture of automated and curatorial analysis. Variation features are calculated automatically from data in the dbSNP database, and the nucleotide region corresponding to the annotated protein conserved domains are also provided automatically (as a miscellaneous feature, or 'misc_feat'). Other features, such as polyadenylation signals and sites, alternate transcription start sites and RNA editing sites, are provided by curation.

CURATION AND QUALITY CONTROL

RefSeq sequences are validated to confirm the following: (i) accurate nucleotide-to-protein sequence correspondence; (ii) valid ASN.1 format and (iii) for species supported by collaboration with official nomenclature groups, current preferred name and symbol designations. Validation of map location is available for species that are annotated via the NCBI annotation pipeline.

NCBI staff review and manually modify a subset of the RefSeq collection including those provided for viruses, some bacteria, mammals and some additional species. The goal of this manual curation is to provide accurate and full-length sequence data, to ensure accurate sequence-to-gene associations, to expand the collection by adding previously unrepresented genes and/or alternate splice products, and to provide additional feature annotation to represent mature peptide products, regions of interest and/or to highlight less frequent biological events such as non-AUG initiation sites (11) or selenoproteins (12). The curation status is annotated on RefSeq records, as a COMMENT feature; the status terms used include model, predicted, provisional, inferred, validated and reviewed, with the latter two indicating that sequence-level curation has taken place. Curation status terms are documented on the RefSeq Web site (http://www.ncbi.nlm.nih.gov/RefSeq/key.html#status).

Several processes are used to identify records that will benefit most from staff review. For instance, records targeted for review include those that differ relative to available genomic

Table 2. RefSeq information, access and feedback

Resource	URL
RefSeq home page	http://www.ncbi.nlm.nih.gov/RefSeq/
FTP—RefSeq release	ftp://ftp.ncbi.nih.gov/refseq/release/
BLAST home page	http://www.ncbi.nlm.nih.gov/BLAST/
Entrez home page	http://www.ncbi.nlm.nih.gov/Entrez/
RefSeq feedback form	http://www.ncbi.nlm.nih.gov/RefSeq/update.cgi
Contact NCBI Help Desk	info@ncbi.nlm.nih.gov
Subscribe to RefSeq announce	http://www.ncbi.nlm.nih.gov/mailman/listinfo/refseq-announce

sequence, those with significant protein length variation compared to homologous groups calculated by the NCBI HomoloGene resource (13), and those for which there are no related proteins other than the GenBank record used to construct the RefSeq. Several additional tests for transcript and protein quality are in place but are not enumerated here. In addition, review is based on user feedback that identifies additional data or errors. We welcome user feedback to help maintain and improve the RefSeq collection. A feedback form is provided online, or users can contact the main NCBI Help Desk (see Table 2).

RETRIEVING DATA

The RefSeq collection can be accessed multiple ways at NCBI, including by Entrez query, BLAST, FTP, and links provided from NCBI databases and resources (see Table 2).

Entrez query

RefSeq results are included in the results returned when performing a global query of the Entrez databases from the NCBI or Entrez homepage. Returned results can be restricted to include only RefSeq records by going to the homepage of the nucleotide or protein database and either using the Entrez Limits page to select 'Only from RefSeq' or adding one of the RefSeq-specific property restrictions directly to the entered text query. For example, a query to retrieve all RefSeq nucleotide records that include the name 'BRCA1' somewhere in the record is formatted as BRCA1 AND srcdb_refseq[prop]. The RefSeq Web site provides definitions of the available property restrictions (http://www.ncbi.nlm.nih.gov/RefSeq/key.html#query).

Entrez queries from the Entrez home page, where it is possible to query against all of the Entrez databases at once, will also return results to the Entrez Gene and Genomes (14) databases, which are both components of the RefSeq project. Entrez Gene integrates gene-specific annotation from RefSeq records with other sources of information, and thus provides a gene-oriented view of data about genes (7). When there is sequence for a complete genome or chromosome, the data are also included in the Entrez Genome database, which provides multiple tools to display and analyze the information.

BLAST and BLink

RefSeq records are included in the main BLAST nr databases and are also made available in genome-specific BLAST database collections (listed at http://www.ncbi.nlm.nih.gov/BLAST/). Hits to RefSeq records can be immediately identified by the distinct format of the accession numbers. BLAST nr results can be configured to show only those hits to the RefSeq collection by entering the Entrez property query on the format page (e.g. srcdb_refseq[prop]).

RefSeq records are also included in the pre-computed BLAST analysis that is done to provide Entrez links to related sequences (nucleotide or protein) and to BLink, a visualization tool for the related protein sequences dataset. The BLink interface includes an option to show only RefSeq proteins.

FTP

The complete RefSeq collection is made available for anonymous FTP as bi-monthly releases in conjunction with daily and cumulative updates between the release cycles. The RefSeq release is structured to provide access to the full RefSeq collection or to a portion of the collection organized by main taxonomic categories (e.g. plant, viral, vertebrate_mammalian) or molecules of interest (e.g. organelle, plasmid). Documentation includes an indication of files and sequences provided, sequences that have been removed since the previous release, and a full description of the release structure and content. Announcements about large changes, problems and the availability of a RefSeq release are emailed to the refseq-announce email list (see Table 2). Additional FTP data is provided for some organisms of interest, including the transcript and protein dataset for human, mouse and rat. Users may be interested in subscribing to refseq-announce@ncbi.nlm.nih.gov to receive information about the RefSeq releases and planned modifications as they occur over time.

Links

Multiple NCBI databases and resources include links to RefSeq records. Links to RefSeq records can be found in many Entrez databases and resources including Gene, UniGene, HomoloGene, Map Viewer, UniSTS.

REFERENCES

1. Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol.*, **266**, 141–162.
2. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
3. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
4. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2005) GenBank. *Nucleic Acids Res.*, **33**, D34–D38.
5. Christie,K.R., Weng,S., Balakrishnan,R., Costanzo,M.C., Dolinski,K., Dwight,S.S., Engel,S.R., Feierbach,B., Fisk,D.G., Hirschman,J.E. *et al.* (2004) Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res.*, **32**, 311–314.
6. FlyBase Consortium (2003) The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res.*, **31**, 172–175.
7. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2005) Entrez Gene: Gene-centered information at NCBI. *Nucleic Acids Res.*, **33**, D54–D58.
8. Bult,C.J., Blake,J.A., Richardson,J.E., Kadin,J.A., Eppig,J.T., Baldarelli,R.M., Barsanti,K., Baya,M., Beal,J.S., Boddy,W.J. *et al.* (2004) The Mouse Genome Database (MGD): integrating biology with the genome. *Nucleic Acids Res.*, **32**, 476–481.

9. Marchler-Bauer,A., Anderson,J.B., DeWeese-Scott,C., Fedorova,N.D., Geer,L.Y., He,S., Hurwitz,D.I., Jackson,J.D., Jacobs,A.R., Lanczycki,C.J. *et al*. (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res*., **31**, 383–387.
10. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*., **29**, 308–311.
11. Touriol,C., Bornes,S., Bonnal,S., Audigier,S., Prats,H., Prats,A.C. and Vagner,S. (2003) Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons. *Biol. Cell*., **95**, 169–178.
12. Copeland,P.R. (2003) Regulation of gene expression by stop codon recoding: selenocysteine. *Gene*, **312**, 17–25.
13. Wheeler,D.L., Church,D.M., Edgar,R., Federhen,S., Helmberg,W., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., *et al*. (2005) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res*., **33**, D39–D45.
14. Tatusova,T.A., Karsch-Mizrachi,I. and Ostell,J.A. (1999) Complete genomes in WWW Entrez: data representation and analysis. *Bioinformatics*, **15**, 536–543.