https://doi.org/10.1093/nar/gkab1112

Database resources of the national center for biotechnology information

Eric W. Sayers[*], Evan E. Bolton, J. Rodney Brister, Kathi Canese, Jessica Chan, Donald C. Comeau, Ryan Connor, Kathryn Funk, Chris Kelly, Sunghwan Kim, Tom Madej, Aron Marchler-Bauer, Christopher Lanczycki, Stacy Lathrop, Zhiyong Lu, Francoise Thibaud-Nissen, Terence Murphy, Lon Phan, Yuri Skripchenko, Tony Tse, Jiyao Wang, Rebecca Williams, Barton W. Trawick, Kim D. Pruitt and Stephen T. Sherry

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received September 18, 2021; Revised October 20, 2021; Editorial Decision October 21, 2021; Accepted November 18, 2021

ABSTRACT

The National Center for Biotechnology Information (NCBI) produces a variety of online information resources for biology, including the GenBank® nucleic acid sequence database and the PubMed® database of citations and abstracts published in life science journals. NCBI provides search and retrieval operations for most of these data from 35 distinct databases. The E-utilities serve as the programming interface for the most of these databases. Resources receiving significant updates in the past year include PubMed, PMC, Bookshelf, RefSeq, SRA, Virus, dbSNP, dbVar, ClinicalTrials.gov, MMDB, iCn3D and PubChem. These resources can be accessed through the NCBI home page at https://www.ncbi.nlm.nih.gov.

INTRODUCTION

NCBI overview

The National Center for Biotechnology Information (NCBI), a center within the National Library of Medicine at the National Institutes of Health, was created in 1988 to develop information systems for molecular biology (1). In this article, we provide a brief overview of the NCBI collection of databases, followed by a summary of resources that we significantly updated in the past year. We provide more complete discussions of NCBI resources on the home pages of individual databases, on the NCBI Learn page (https://www.ncbi.nlm.nih.gov/learn/), and in the NCBI Handbook (https://www.ncbi.nlm.nih.gov/books/NBK143764/).

NCBI databases

NCBI maintains a diverse set of 35 databases that together contain 3.6 billion records (Table 1 and Figure 1), most of which are available through the Entrez retrieval system (2) at https://www.ncbi.nlm.nih.gov/search/. Each database supports text searching using simple Boolean queries, downloading of data in various formats, and linking records between databases based on asserted relationships. Records retrieved in Entrez can be displayed in many formats and downloaded singly or in batches. An Application Programming Interface for Entrez functions (the E-utilities) is available, and detailed documentation is provided at https://eutils.ncbi.nlm.nih.gov/.

Data sources and collaborations

NCBI receives data from three sources: direct submissions from researchers, national and international collaborations or agreements with data providers and research consortia, and internal curation efforts. For example, NCBI manages the GenBank database (3) and participates with the EMBL-EBI European Nucleotide Archive (ENA) (4) and the DNA Data Bank of Japan (DDBJ) (5) as a partner in the International Nucleotide Sequence Database Collaboration (INSDC) (6). Details about direct submission processes are available from the NCBI Submit page (https://www.ncbi.nlm.nih.gov/home/submit.shtml) and from the resource home pages (e.g. the GenBank page, https://www.ncbi.nlm.nih.gov/genbank/). More information about the various collaborations, agreements, and curation efforts are also available through the home pages of the individual resources.

[*] To whom correspondence should be addressed. Tel: +1 301 496 2475; Fax: +1 301 480 9241; Email: sayers@ncbi.nlm.nih.gov

Published by Oxford University Press on behalf of Nucleic Acids Research 2021.

Table 1. NCBI databases (as of 4 September 2021)

Database	Records	Description
Literature		
PubMed	33 027 761	Scientific and medical abstracts/citations
PubMed Central	7 325 415	Full-text journal articles
NLM Catalog	1 629 799	Index of NLM collections
Bookshelf	892 126	Books and reports
MeSH	348 370	Ontology used for PubMed indexing
Genomes		
Nucleotide	476 054 019	DNA and RNA sequences from GenBank and RefSeq
BioSample	19 473 659	Descriptions of biological source materials
SRA	15 919 320	High-throughput DNA/RNA sequence read archive
Taxonomy	2 492 889	Taxonomic classification and nomenclature catalog
Assembly	1 083 900	Genome assembly information
BioProject	536 242	Biological projects providing data to NCBI
Genome	64 815	Genome sequencing projects by organism
BioCollections	8 468	Museum, herbaria, and biorepository collections
Genes		
GEO Profiles	128 414 055	Gene expression and molecular abundance profiles
Gene	33 664 932	Collected information about gene loci
GEO DataSets	4 784 603	Functional genomics studies
PopSet	366 935	Sequence sets from phylogenetic/population studies
HomoloGene	141 268	Homologous gene sets for selected organisms
Clinical		
dbSNP	1 076 992 604	Short genetic variations
dbVar	7 117 914	Genome structural variation studies
ClinVar	1 071 071	Human variations of clinical significance
ClinicalTrials.gov	388 717	Registry of clinical studies and results database
MedGen	335 277	Medical genetics literature and links
GTR	77 498	Genetic testing registry
dbGaP	1 405	Genotype/phenotype interaction studies
Proteins		
Protein	968 236 913	Protein sequences from GenBank and RefSeq
Identical Protein Groups	448 096 579	Protein sequences grouped by identity
Protein Clusters	1 137 329	Sequence similarity-based protein clusters
Structure	181 772	Experimentally-determined biomolecular structures
Protein Family Models	179 133	Conserved domain architectures, HMMs, and BlastRules
Conserved Domains	62 852	Conserved protein domains
Chemicals		
PubChem Substance	284 180 803	Deposited substance and chemical information
PubChem Compound	110 628 849	Chemical information with structures, information and links
PubChem BioAssay	1 391 308	Bioactivity screening studies
BioSystems	983 968	Molecular pathways with links to genes, proteins and chemicals

RECENT DEVELOPMENTS

Literature updates

PubMed. NCBI has continued to update the new PubMed with features and improvements following its launch in May 2020 and the retirement of the legacy site in October 2020. We have added many of the most requested updates, such as a custom date range filter, additional display formats and sort options, MyNCBI highlighting preferences, 'Related Information' links on Abstract pages, and more convenient positioning of the 'Save', 'Email' and 'Send to' buttons based on usability testing where we observe volunteers interacting with various design options. In addition, users can now email up to 1000 citations at a time. We also updated the Single Citation Matcher and Batch Citation Matcher to the new PubMed cloud environment and deployed an updated relevance-based algorithm supporting Best Match retrieval that includes feedback from clicked articles (7). This year we reviewed and refined the protocols for updating the model that places articles with the greatest number of clicks for a given query type near the top of the result list. The result increases the likelihood that a searcher will easily find articles of interest.

The rapid surge in biomedical research on the COVID-19 pandemic in 2021 has made it increasingly challenging to navigate and analyze the biomedical literature related to COVID-19. In response, we updated the PubMed Clinical Queries interface following a usability study (similar to that employed with PubMed) to support user-friendly, efficient searching on clinical and disease-related topics. Clinical Queries currently includes filters for Clinical Studies and COVID-19 articles (https://pubmed.ncbi.nlm.nih.gov/clinical/?term=nursing+home&filter_category=covid-19&clinical_study_category=therapy&covid_category=prevention&clinical_study_scope=broad) based on the eight topic categories in LitCovid, a spe-

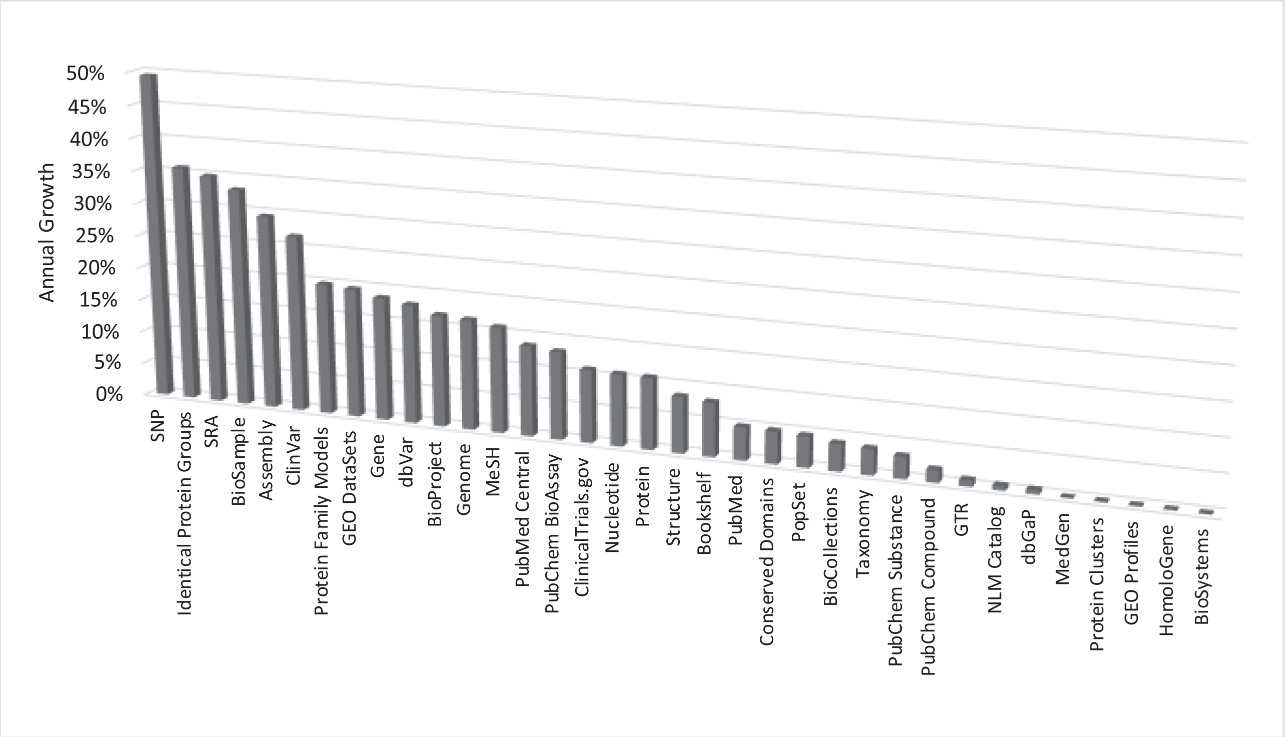

Figure 1. Annual growth rates of the number of records in each NCBI database as of 4 September 2021.

cialized literature hub for tracking SARS-CoV-2 (8,9).

PubMed Central (PMC). In January 2021, PMC reached a significant milestone in surpassing 7 million journal articles and author manuscripts publicly available in the archive. This milestone is the result of contributions from publishers, journals, authors, and data providers in addition to partnerships with numerous research funding organizations. Two additional projects that launched in 2021 also continued to add content to the archive for public discovery and use. First, the Public Health Emergency COVID-19 Initiative (https://www.ncbi.nlm.nih.gov/pmc/about/covid-19/), a collaboration with more than 50 publishers and societies launched in March 2020, added 185 000 articles in human and machine-readable formats to PMC through September 2021 and continues to enable human and machine-readable access on a global scale to the most current research on COVID-19 and SARS-CoV-2. Second, the NIH Preprint Pilot, which started in June 2020 as a pilot project to include preprints resulting from NIH-funded research in PMC (https://www.ncbi.nlm.nih.gov/pmc/about/nihpreprints/), remained in Phase 1, limiting the scope of the effort to NIH-supported preprints reporting on COVID-19 research. Through September 2021, 2700 preprint records had been identified as in scope for the pilot and made available in PMC and PubMed; just under half of these preprints (48%) had been linked to a peer-reviewed journal article version.

To further support machine-readable access to journal articles, manuscripts, and preprints made available under license terms that allow for text mining and other types of reuses (e.g. Creative Commons licenses), NCBI updated the structure of the PMC FTP directories and added a cloud distribution option. Beginning in July 2021, articles from the PMC Open Access Subset and Author Manuscript Dataset were discoverable in the Registry of Open Data on Amazon Web Services (AWS) in full-text XML and plain text formats. Additionally, in June 2021 NCBI announced the release of PMC Labs (https://www.ncbi.nlm.nih.gov/labs/pmc/), the first phase of a project aimed at creating a more modern and accessible user experience in PMC. Ongoing direct feedback, usability testing, and user research on PMC Labs will guide future improvements to the PMC website.

Bookshelf. The NCBI Bookshelf provides free online access to over 9600 books and documents in life science and healthcare from over 150 content providers. This past year, in collaboration with participating U.S. health agencies, international health organizations, and other sponsors and publishers of evidence-based clinical or public health information, Bookshelf made available a collection of COVID-19 guidelines and reviews (https://www.ncbi.nlm.nih.gov/books/about/covid/) that meet Bookshelf's content selection criteria (https://www.ncbi.nlm.nih.gov/books/about/scientificeval/). The targeted search query (https://www.ncbi.nlm.nih.gov/books?term=%22covid+19%22%5BResource+ Type%5D&cmd=DetailsSearch) is regularly updated with new and updated guidelines and reviews.

Genome updates

RefSeq. The RefSeq prokaryotic collection has grown from 198 640 genomes in August 2020 to 223 542 as of 6 August 2021 (10). Because this collection is large and unevenly distributed across the taxonomic tree, we also provide a representative set, containing exactly one genome per species, available as both nucleotide and protein BLAST databases. The number of species in that set has grown from 11 735 in August 2020 to 14 606, reflecting the increase in the diversity of the RefSeq collection.

We recently updated the Prokaryotic Genome Annotation Pipeline (PGAP) that is used to annotate all RefSeq genomes (except for 15 reference genomes). The update addresses limitations on comparative analyses caused by a lack of common gene identifiers across genomes. This new module in PGAP is similar to one used in the Eukaryotic Genome Annotation Pipeline (11); it assigns gene symbols (e.g. recA) to annotated genes using an orthology calculation between the PGAP-annotated and the reference genome for five common species. After reannotation of all RefSeq genomes for these five species in scope, the percentage increase in the number of gene symbols ranged from 25% to 110%. Now, 73% of PGAP-annotated *Escherichia coli* genes and 79% of *Bacillus subtilis* genes have symbols (35% for *Mycobacterium tuberculosis*, 40% for *Acinetobacter pittii* and 46% for *Campylobacter jejuni*). PGAP is available as a stand-alone package, and as a service available to GenBank submitters (https://www.ncbi.nlm.nih.gov/genome/annotation_prok/).

The RefSeq eukaryotic collection has now grown to 1290 genomes from 1272 species as of 6 August 2021, including 71 species updated with new genomes in the last 12 months. Genomes from 780 species are now annotated with NCBI's Eukaryotic Genome Annotation Pipeline, including all vertebrates and most other multicellular eukaryotes, with most models completely based on RNA-seq and/or protein alignment evidence. We have added reporting of BUSCO statistics (12) computed on the longest protein from each gene. BUSCO provides an assessment of the completeness of the annotated gene set based on single-copy orthologs. These statistics are available in the annotation reports for recent annotation releases and for over 800 RefSeq genomes in data returns from E-utility calls to NCBI Assembly. Over half of the RefSeq eukaryote annotations produced by NCBI score over 98% complete by BUSCO using the most specific BUSCO lineage available for each species. One-to-one orthologs for fish species are now computed using the zebrafish genome as a reference, and orthologs for insects compared to the *Drosophila melanogaster* genome are available on the NCBI Gene FTP site. The collection of curated, well-supported, protein-coding transcripts (NM_ accessions) for human has grown to over 62 000 transcripts, of which 99.5% exactly match the GRCh38 genome and 87% have been revised based on CAGE and polyA-seq data. In addition to quarterly updates to the RefSeq GRCh38.p13 genome annotation, we provide yearly updates for the prior GRCh37.p13 genome to support clinical users that have not yet updated to the current reference assembly. To help properly annotate SNPs on RefSeq transcripts that have sequence differences relative to the genome, we have added transcript-to-genome alignments in BAM format with each annotation release on FTP. Finally, the data for a collection of human genes involved in coronavirus infection and disease have been revised and are available in NCBI gene with a query for 'coronavirus related'[filter].

Sequence read archive. NCBI maintains the NIH Sequence Read Archive (SRA), an archival database designed to support storage, retrieval, and analysis of next-generation nucleotide sequence data. A single copy of the archive now includes 11.5 Petabytes of publicly available data and another 4.9PB of controlled-access dbGaP data. This is one of the largest publicly available, biological data repositories, and while the archive holds tremendous promise for biomedical research, the sheer size of this dataset makes it difficult to store, retrieve, and analyze. NCBI is currently implementing more efficient approaches to data storage and improving data accessibility. As part of the NIH Science and Technology Research Infrastructure for Discovery, Experimentation and Sustainability (STRIDES) Initiative (https://datascience.nih.gov/strides/), NCBI is now maintaining the entire SRA on two commercial cloud platforms, AWS and Google Cloud Platform (GCP) (https://www.ncbi.nlm.nih.gov/sra/docs/sra-cloud/). Moving SRA to the cloud provides several benefits: long-term sustainability and scalability, increased download speeds, unlimited concurrent downloads, and access to the originally submitted data files. To support SRA sequence data selection and analysis, SRA run metadata, BioProject data, and BioSample data are available through GCP BigQuery and AWS Athena to support selection and analysis of SRA sequence data (https://www.ncbi.nlm.nih.gov/sra/docs/sra-cloud-based-examples/). The metadata available in BigQuery and Athena includes results from the k-mer-based Sequence Taxonomic Analysis Tool that describe the organismal content present in samples (13).

Both originally submitted and normalized SRA files are available through a combination of 'hot' and 'cold' storage tiers on both AWS and GCP cloud platforms. The SRA toolkit can be used to retrieve data from hot cloud storage to local and cloud locations, and for cloud-based users, will pull data from the appropriate commercial cloud platform to user generated buckets (https://github.com/ncbi/sra-tools/wiki/02.-Installing-SRA-Toolkit). A web-based interface can also be used to retrieve hot and/or cold stored data and move it to user-created cloud buckets (https://www.ncbi.nlm.nih.gov/Traces/cloud-delivery/). More information about selecting and retrieving SRA data from cloud platforms can be found at https://www.ncbi.nlm.nih.gov/sra/docs/data-delivery/.

NCBI remains committed to continuing to expand SRA and improving access to the archive based on FAIR data principles. As part of this effort, normalized SRA files are also available through the AWS Open Data Program (ODP) (https://registry.opendata.aws/ncbi-sra/) where they can be egressed or downloaded for free by both cloud and non-cloud users. This program supports enhanced access to a variety of cloud-based and non-cloud data users, and several tools can be used to download and/or egress based on specific use cases, including the SRA toolkit and tools maintained by AWS.

Virus. In response to the COVID-19 pandemic, NCBI has provided enhanced access to >700 000 SARS-CoV-2 SRA samples through the AWS ODP and the Google Public Dataset Program. These samples are identified using STAT organismal content analysis (13) and user-provided sample descriptions. We provide several data files for each sample on the cloud platforms, including originally formatted sequence files submitted to SRA, normalized format SRA sequence files, and SRA aligned read files where submitted reads are aligned to contigs assembled from runs using the SARS-CoV-2 RefSeq guided assembly. We annotate these contigs with VIGOR3 (14), then translate and include them in BLASTn and BLASTp databases available in the AWS ODP. Nucleotide variations relative to the RefSeq reference (NC_045512.2) are also calculated for each run and made available in variant call format (VCF) files (14) and SPDI (Sequence, Position, Deletion, Insertion) format (15). Finally, SRA, BioSample and BioProject metadata and analysis data calculated by NCBI are available through BigQuery and Athena tables (https://www.ncbi.nlm.nih.gov/sra/docs/aligned-metadata-tables/).

Clinical updates

dbSNP. The Database of Single Nucleotide Polymorphisms (dbSNP) is a repository of human genomic variations and frequency data that includes both common and rare single-nucleotide variations and other small-scale variations. In 2021, dbSNP reached a milestone of over 1 billion RS numbers with build release 155 (June 2021). This dbSNP release included the NCBI Allele Frequency Aggregator (ALFA) data (https://www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/) with over 900 million variants and allele frequencies for 12 populations from 192 000 dbGaP subjects. We improved the Entrez SNP search to support queries using HGVS, protein position, SPDI, LitVar, and ALFA attributes. We also redesigned the RefSNP page to include more intuitive horizontal tabs for variant information including ClinVar clinical significance, allele frequency, and publications (e.g. https://www.ncbi.nlm.nih.gov/snp/rs732609?horizontal_tab=true). In addition, we added a function to retrieve RefSNP flank sequences with a user-specified length.

dbVar. In the past year dbVar added 26 new human studies for structural variation (SV) that include 1 221 062 SV regions and 1 241 943 SV calls, bringing the total in the June 2021 release to 207 studies, 6 964 796 SV regions, and 37 081 907 calls. Included in dbVar are structural variants from ClinVar (nstd102), gnomAD from 15 000 genomes (nstd106), Genome in a Bottle (nstd175), 1000 Genomes 30X (nstd206), NCBI Curated Common Structural Variants with frequency (nstd186), and other population studies (nstd200 and nstd207). dbVar now had improved displays of selected structural variation datasets, including somatic and presumed normal SV, available on TrackHub, the NCBI Genome Data Viewer, the UCSC genome browser, and other genomic browsers. Additionally, we improved mouse-over tips, track colors, and links to ClinVar.

ClinicalTrials.gov. Launched in 2000, ClinicalTrials.gov serves as an essential, transparent, and trusted part of the research ecosystem by providing patients, healthcare professionals, and researchers access to information on ongoing and completed clinical research (16). Study information is provided, voluntarily or to meet legal and policy requirements (17), by the sponsor or investigator responsible for the research. As of August 2021, the ClinicalTrials.gov registry listed over 384 000 clinical studies from all 50 U.S. states and 220 countries, and the ClinicalTrials.gov results database reached a milestone of listing summary results information for over 50 000 clinical studies, about half of which have not been published in the biomedical literature.

ClinicalTrials.gov is designed to support several use cases. These include helping potential participants discover recruiting studies, allowing researchers to identify gaps and redundancies across the clinical research enterprise, aiding consumers of evidence-based medicine by reducing publication bias and preventing selective publication of outcomes, and enabling journal editors to evaluate results reported in manuscripts against prespecified research plans. To advance these several uses further, NLM launched a multi-year effort in October 2019 to modernize ClinicalTrials.gov by delivering an improved user experience on a cloud-based platform that will accommodate growth and enhance efficiency (https://clinicaltrials.gov/ct2/about-site/modernization/). Initial work has focused on technical infrastructure enhancements and obtaining broad and focused user input to develop prototypes using an agile development process with iterative design based on user testing.

In response to the COVID-19 public health emergency, ClinicalTrials.gov has expedited the processing of COVID-19 related study submissions to facilitate discovery and access. As of August 2021, over 6200 studies related to COVID-19 were listed (https://clinicaltrials.gov/ct2/results?cond=COVID-19) and another nearly 4700 COVID-19 studies from the WHO registry search portal (https://clinicaltrials.gov/ct2/who_table) were available from ClinicalTrials.gov. In November 2020, the NIH Director issued a statement calling on researchers to swiftly share COVID-19 clinical research information by registering and submitting results to ClinicalTrials.gov as quickly as possible, ahead of legal and policy deadlines (https://www.nih.gov/about-nih/who-we-are/nih-director/statements/nih-calls-clinical-researchers-swiftly-share-covid-19-results).

Protein updates

MMDB. NCBI has upgraded both the MMDB (Molecular Modeling Database) and GenBank sequence record tools so that they can process the complete set of Protein Data Bank (PDB) molecular structure data. With advances in structure determination and the ability to process much bigger molecular assemblies, the PDB has been releasing larger structures only in the more flexible mmCIF-format (https://mmcif.wwpdb.org/). The latter allows for longer 'chain identifiers', as opposed to single alphanumeric characters, which are necessary to accommodate very large structural assemblies with many different polypeptide chains. Whereas previously all protein and nucleotide sequences derived from PDB structures could be retrieved using accessions like '1ABC_X', instances such as '1ABC_XX' or '1ABC_WXYZ', or '1ABC_abc' can now be found. As of this writing there are about 1200 such structures containing long chain identifiers, and corresponding entries have been added to the sequence databases; the MMDB structure database and sequence databases reflect the complete set of all the structures in the Protein Data Bank and we continue to update them weekly.

iCn3D. With the release of iCn3D 3.0, users can perform three-dimensional structure analyses on lists of structures using Node.js scripts, based on the npm icn3d package (https://www.npmjs.com/package/icn3d). Some example scripts are provided at https://github.com/ncbi/icn3d/tree/master/icn3dnode and demonstrate how to analyze protein-ligand interactions, protein-protein interactions, and change of interactions due to residue mutations (18). Additional examples demonstrate how to compute electrostatic potentials using DelPhi (19), solvent accessible surface areas, and more. Users can easily generate their own Node.js scripts based on the functions in iCn3D: https://www.ncbi.nlm.nih.gov/Structure/icn3d/icn3d.html#classstructure. iCn3D also allows users to align multiple chains from different structures to a reference chain and visualize the corresponding super-impositions, to calculate the intramolecular symmetry dynamically (20), and visualize structures as 2D cartoons at the level of chains, protein domains, and protein secondary structure.

Chemical updates

In the past year, data from >50 new sources were integrated into PubChem (21–24), which now provides chemical information for over 110 million compounds collected from more than 800 data sources. The newly added data include annotations from authoritative sources, such as the Acute Exposure Guideline Levels (AEGLs) from the U.S. Environmental Protection Agency (EPA), the Drug-Induced Liver Injury Rank (DILIrank) dataset from the U.S. Food and Drug Administration (FDA), the carcinogen classification from the International Agency for Research on Cancer (IARC), and the Chemical Abstracts Service (CAS) registry numbers from CAS Common Chemistry. Notably, patent information from Google Patents (https://patents.google.com/) was integrated with PubChem along with corresponding patent metadata used to generate PubChem Patent Summary pages (https://pubchemdocs.ncbi.nlm.nih.gov/patents).

PubChem updated its data model used for storing bioassay information (https://pubchemblog.ncbi.nlm.nih.gov/2021/02/25/updates-to-the-pubchem-assay-data-model/). The new bioassay data model makes it easier to handle and display panel assay data. It also supports UTF-8 characters, which often appear in assay data (for example, Greek letters in target names such as β-galactosidase) or units found in experimental protocols (e.g. °C and °F). A full data specification is available at the PubChem FTP site (https://ftp.ncbi.nlm.nih.gov/pubchem/). PubChem also updated PubChemRDF (25), machine-readable PubChem data formatted using the Resource Description Framework (RDF; https://www.w3.org/RDF/) (https://pubchemblog.ncbi.nlm.nih.gov/2020/10/19/pubchemrdf-1--7%ce%b2-has-been-released/). A major change in this update is the addition of a new subdomain, called Pathway, which encodes information on biological pathways and their relationship with genes, proteins, and chemicals.

FOR FURTHER INFORMATION

The resources described here include documentation, other explanatory materials, and references to collaborators and data sources on their respective web sites. An alphabetical list of NCBI resources is available from a link above the category list on the left side of the NCBI home page. The NCBI Help Manual and the NCBI Handbook (www.ncbi.nlm.nih.gov/books/NBK143764/), both available as links in the common page footer, describe the principal NCBI resources in detail. The NCBI Learn page (www.ncbi.nlm.nih.gov/learn/) provides links to documentation, tutorials, webinars, courses, and upcoming conference exhibits. A variety of video tutorials are available on the NLM YouTube channel that can be accessed through links in the standard NCBI page footer. User-support staff are available to answer questions at info@ncbi.nlm.nih.gov, and users can view support articles at https://support.nlm.nih.gov. Updates on NCBI resources and database enhancements are described on the NCBI Insights blog (https://ncbiinsights.ncbi.nlm.nih.gov/), NCBI social media sites (FaceBook, Twitter and LinkedIn), and the several mailing lists and RSS feeds that provide updates on services and databases. Links to these resources are in the NCBI page footer and on NCBI Insights.

ACKNOWLEDGEMENTS

The authors would like to thank all the NCBI staff who through their dedicated efforts continue to allow NCBI to provide our full collection of services to the community.

FUNDING

Intramural Research Program of the National Library of Medicine, National Institutes of Health. Funding for open access charge: Intramural Research Program of the National Library of Medicine, National Institutes of Health.
Conflict of interest statement. None declared.

This paper is linked to: doi:10.1093/nar/gkab1053.

REFERENCES

1. Sayers,E.W., Beck,J., Bolton,E.E., Bourexis,D., Brister,J.R., Canese,K., Comeau,D.C., Funk,K., Kim,S., Klimke,W. *et al.* (2021) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **49**, D10–D17.
2. 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.
3. Sayers,E.W., Cavanaugh,M., Clark,K., Pruitt,K.D., Schoch,C.L., Sherry,S.T. and Karsch-Mizrachi,I. (2021) GenBank. *Nucleic Acids Res.*, **49**, D92–D96.

4. Harrison,P.W., Ahamed,A., Aslam,R., Alako,B.T.F., Burgin,J., Buso,N., Courtot,M., Fan,J., Gupta,D., Haseeb,M. *et al.* (2021) The european nucleotide archive in 2020. *Nucleic Acids Res.*, **49**, D82–D85.
5. Fukuda,A., Kodama,Y., Mashima,J., Fujisawa,T. and Ogasawara,O. (2021) DDBJ update: streamlining submission and access of human data. *Nucleic Acids Res.*, **49**, D71–D75.
6. Arita,M., Karsch-Mizrachi,I. and Cochrane,G. (2021) The international nucleotide sequence database collaboration. *Nucleic Acids Res.*, **49**, D121–D124.
7. Fiorini,N., Canese,K., Starchenko,G., Kireev,E., Kim,W., Miller,V., Osipov,M., Kholodov,M., Ismagilov,R., Mohan,S. *et al.* (2018) Best match: new relevance search for PubMed. *PLoS Biol.*, **16**, e2005343.
8. Chen,Q., Allot,A. and Lu,Z. (2021) LitCovid: an open database of COVID-19 literature. *Nucleic Acids Res.*, **49**, D1534–D1540.
9. Chen,Q., Allot,A. and Lu,Z. (2020) Keep up with the latest coronavirus research. *Nature*, **579**, 193.
10. Li,W., O'Neill,K.R., Haft,D.H., DiCuccio,M., Chetvernin,V., Badretdin,A., Coulouris,G., Chitsaz,F., Derbyshire,M.K., Durkin,A.S. *et al.* (2021) RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic. Acids. Res.*, **49**, D1020–D1028.
11. Pruitt,K.D., Brown,G.R., Hiatt,S.M., Thibaud-Nissen,F., Astashyn,A., Ermolaeva,O., Farrell,C.M., Hart,J., Landrum,M.J., McGarvey,K.M. *et al.* (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, **42**, D756–D763.
12. Manni,M., Berkeley,M.R., Seppey,M., Simao,F.A. and Zdobnov,E.M. (2021) BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.*, **38**, 4647–4654.
13. Katz,K.S., Shutov,O., Lapoint,R., Kimelman,M., Brister,J.R. and O'Sullivan,C. (2021) STAT: a fast, scalable, MinHash-based k-mer tool to assess Sequence Read Archive next generation sequence submissions. *Genome Biol.*, **22**, 270.
14. Wang,S., Sundaram,J.P. and Stockwell,T.B. (2012) VIGOR extended to annotate genomes for additional 12 different viruses. *Nucleic Acids Res.*, **40**, W186–W192.
15. Holmes,J.B., Moyer,E., Phan,L., Maglott,D. and Kattman,B. (2020) SPDI: data model for variants and applications at NCBI. *Bioinformatics*, **36**, 1902–1907.
16. Zarin,D.A., Tse,T., Williams,R.J. and Rajakannan,T. (2017) Update on trial registration 11 years after the ICMJE policy was established. *N. Engl. J. Med.*, **376**, 383–391.
17. Zarin,D.A., Fain,K.M., Dobbins,H.D., Tse,T. and Williams,R.J. (2019) 10-Year update on study results submitted to clinicaltrials.gov. *N. Engl. J. Med.*, **381**, 1966–1974.
18. Xiang,Z. and Honig,B. (2001) Extending the accuracy limits of prediction for side-chain conformations. *J. Mol. Biol.*, **311**, 421–430.
19. Nicholls,A. and Honig,B. (1991) A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson-Boltzmann equation. *J. Comp. Chem.*, **12**, 435–445.
20. Tai,C.H., Paul,R., Dukka,K.C., Shilling,J.D. and Lee,B. (2014) SymD webserver: a platform for detecting internally symmetric protein structures. *Nucleic Acids Res.*, **42**, W296–300.
21. Kim,S., Chen,J., Cheng,T., Gindulyte,A., He,J., He,S., Li,Q., Shoemaker,B.A., Thiessen,P.A., Yu,B. *et al.* (2021) PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.*, **49**, D1388–D1395.
22. Kim,S., Chen,J., Cheng,T., Gindulyte,A., He,J., He,S., Li,Q., Shoemaker,B.A., Thiessen,P.A., Yu,B. *et al.* (2019) PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.*, **47**, D1102–D1109.
23. Kim,S., Thiessen,P.A., Bolton,E.E., Chen,J., Fu,G., Gindulyte,A., Han,L., He,J., He,S., Shoemaker,B.A. *et al.* (2016) PubChem Substance and Compound databases. *Nucleic Acids Res.*, **44**, D1202–D1213.
24. Kim,S. (2016) Getting the most out of PubChem for virtual screening. *Expert Opin. Drug Discov.*, **11**, 843–855.
25. Fu,G., Batchelor,C., Dumontier,M., Hastings,J., Willighagen,E. and Bolton,E. (2015) PubChemRDF: towards the semantic annotation of PubChem compound and substance databases. *J. Cheminform.*, **7**, 34.