D344–D350 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 11 October 2007*
doi:10.1093/nar/gkm791

ChEBI: a database and ontology for chemical entities of biological interest

Kirill Degtyarenko[1,*], Paula de Matos[1], Marcus Ennis[1], Janna Hastings[1], Martin Zbinden[1], Alan McNaught[1], Rafael Alcántara[1], Michael Darsow[1], Mickaël Guedj[1] and Michael Ashburner[2]

[1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD and [2]Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK

Received August 13, 2007; Revised September 14, 2007; Accepted September 17, 2007

ABSTRACT

Chemical Entities of Biological Interest (ChEBI) is a freely available dictionary of molecular entities focused on 'small' chemical compounds. The molecular entities in question are either natural products or synthetic products used to intervene in the processes of living organisms. Genome-encoded macromolecules (nucleic acids, proteins and peptides derived from proteins by cleavage) are not as a rule included in ChEBI. In addition to molecular entities, ChEBI contains groups (parts of molecular entities) and classes of entities. ChEBI includes an ontological classification, whereby the relationships between molecular entities or classes of entities and their parents and/or children are specified. ChEBI is available online at http://www.ebi.ac.uk/chebi/

INTRODUCTION

Any chemical compound naturally occurring in living organisms can be called a 'biochemical compound'. Biochemical compounds can be classified according to their structure, physico-chemical properties or biological function. Most biologists conveniently divide all biochemical compounds into 'biopolymers', which consist of macromolecules, and the rest, which consist of 'small molecules' (see, for instance, MetaCyc Taxonomy of Compounds (1) (http://biocyc.org/META/class-tree?object=Compounds). This dichotomy is faithfully mirrored by 'traditional' bioinformatics in the sense that information-rich macromolecules live in their databases such as the EMBL Nucleotide Sequence Database (2) and UniProt (3) independently from all other molecules, whether small or large.

Although 'small molecules' appear to be less complex entities than macromolecules, their naming, citation and representation in databases is not a trivial task. Most genetically encoded biomacromolecules are easily represented as one-dimensional (1D) strings, while a two-dimensional (2D) sketch remains the most adequate portrait of a 'small molecule'. Several algorithms of linear notation have been developed, e.g. SMILES (4). However, linear notation, as for any other structural core data, cannot be used in speech (and should not be used in free text). Good annotation practice for biological databases is to use either consistent and widely recognized terminology or unique identifiers (to look up the molecule of interest from a dedicated database) (5).

It is an unfortunate fact that chemical data has for a long time been neglected by the computational biology/bioinformatics community. In publications, it is almost never featured as something worthy of attention on its own, but either in conjunction with one or another 'omics' or as part of a 'data integration' project. We consider this approach to be fundamentally flawed and that an open-access, good quality resource for chemical entities or chemical reactions has an absolute value, not just in the context of metabolic pathways or protein ligands. In order to address this issue, in 2002 a project was initiated at the European Bioinformatics Institute (EBI) to create a definitive, freely available dictionary of Chemical Entities of Biological Interest (ChEBI; pronounced /'keb1/). The primary motivation was to provide a high quality, thoroughly annotated controlled vocabulary to promote the correct and consistent use of unambiguous biochemical terminology throughout the molecular biology databases at the EBI. However, it became clear that this aim could not be achieved outside of a wider context, namely that of general chemistry and chemical nomenclature. Since its first public release (21 July 2004), ChEBI has grown to represent more than 12 000 molecular entities, groups and classes. The term 'molecular entity' refers to any constitutionally or isotopically distinct atom, molecule, ion, ion pair, radical, radical ion, complex, conformer, etc., identifiable as a separately distinguishable

*To whom correspondence should be addressed. Tel: +44 12 23 49 46 59; Fax: +44 12 23 49 44 68; Email: kirill@ebi.ac.uk

entity (6). A group is a defined linked collection of atoms or a single atom within a molecular entity (7). ChEBI includes classes of molecular entities (e.g. 'alkanes') as well as classes of groups (e.g. 'alkyl groups'). The scope of ChEBI encompasses not only 'biochemical compounds' but also pharmaceuticals, agrochemicals, laboratory reagents, isotopes and subatomic particles.

DATABASE DESCRIPTION

Main principles

- The terminology used in ChEBI is 'definitive' in the sense that it is explicitly endorsed, where applicable, by international bodies such as IUPAC (http://www.iupac.org/) (general chemical nomenclature) and NC-IUBMB (http://www.iubmb.org/) (biochemical nomenclature).
- Nothing held in the database is proprietary or derived from a proprietary source that would limit its free distribution/availability to anyone.
- Every data item in the database is fully traceable and explicitly referenced to the original source.
- The entirety of the data is available to all without constraint as, for example, MySQL table dumps and Open Biomedical Ontologies (OBO) format flat files (http://oboedit.org/).

Although the initial objective of ChEBI was to standardize biochemical terminology, the need to store and represent the 2D chemical structures has been recognized from the start. In accordance with the principles outlined above, ChEBI has adopted open standards for chemical structure representation, such as the IUPAC International Chemical Identifier (InChI) (8) and will shortly also incorporate Chemical Markup Language (CML) (9). The connectivity and stereochemistry (2D structure) for the majority of the small organic molecules in ChEBI (including isotope-labelled ones) can be unambiguously represented as InChI strings.

Data sources

In order to create ChEBI, data from a number of different sources were incorporated and then merged. Data for the initial release were drawn from three main sources:

- *IntEnz*—the Integrated relational Enzyme database of the EBI (10). IntEnz contains the Enzyme Nomenclature, the recommendations of the NC-IUBMB on the Nomenclature and Classification of Enzyme-Catalysed Reactions.
- *KEGG COMPOUND*—One part of the LIGAND composite database (http://www.genome.jp/kegg/ligand.html) of the Kyoto Encyclopedia of Genes and Genomes (KEGG) (11).
- *Chemical Ontology*—Originally developed as 'Chemical Ontology' by Michael Ashburner and Pankaj Jaiswal, the initial alpha release was subsumed into ChEBI and is currently in the process of being refined and extended.

These and a number of further sources of terminology and other data used and cross-referenced in ChEBI are summarized in Table 1. Names created by ChEBI annotators and based on the recommendations of international authorities are assigned the relevant source as shown in Table 2. When no data source can be indicated, the source of the term is shown as 'ChEBI'.

Table 1. Online databases cross-referenced in ChEBI

Source	URL	Reference
ChemIDplus	http://chem.sis.nlm.nih.gov/chemidplus/	
COMe	http://www.ebi.ac.uk/come/	(14)
EMBL Nucleotide Sequence Database	http://www.ebi.ac.uk/embl/	(2)
IntEnz	http://www.ebi.ac.uk/intenz/	(10)
KEGG LIGAND, including KEGG COMPOUND KEGG DRUG KEGG GLYCAN	http://www.genome.ad.jp/kegg/ligand.html	(11)
LIPID MAPS	http://www.lipidmaps.org/	(15)
MolBase	http://winter.group.shef.ac.uk/molbase/	
MSDchem	http://www.ebi.ac.uk/msd-srv/msdchem/	
NIST Chemistry WebBook	http://webbook.nist.gov/chemistry/	
PDB	http://www.rcsb.org/	(16)
RESID	http://www.ebi.ac.uk/RESID/	(17)
UM-BBD	http://umbbd.msi.umn.edu/	(18)
UniProt	http://www.uniprot.org/	(3)
WebElements	http://www.webelements.com/	

Table 2. Recommendations used to derive terminology in ChEBI

Source	Comment
CBN	Name based on the recommendations of the IUPAC-IUB Commission on Biochemical Nomenclature, the forerunner of JCBN, which was discontinued in 1977.
IUBMB	Name based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB; http://www.chem.qmul.ac.uk/iubmb/). Of particular relevance is Glossary of Chemical Names (http://www.chem.qmul.ac.uk/iubmb/enzyme/glossary.html) used in the Enzyme Nomenclature.
IUPAC	Name based on the recommendations of the International Union of Pure and Applied Chemistry (http://www.chem.qmul.ac.uk/iupac/).
JCBN	Name based on the recommendations of the IUPAC-IUBMB Joint Commission on Biochemical Nomenclature (http://www.chem.qmw.ac.uk/iupac/jcbn/), a body jointly responsible to both IUBMB and IUPAC, which deals with matters of biochemical nomenclature that have importance in both biochemistry and chemistry.

Database design

ChEBI is designed as a relational database, which is implemented in an Oracle database server. A number of utility applications, implemented mainly in Java and Unix scripts, provide the additional functionality around the

database, such as the loading of data from external sources. Specialized web-based interfaces provide for both public access to the data and restricted access to the annotation tool.

Database contents

ChEBI ID. *ChEBI ID* is a unique and stable identifier for the entity, for example, CHEBI:16236. It is semantically free and may be cited by external users.

ChEBI Names. *ChEBI Name* is the name for an entity recommended for use by the biological community. In general, traditional names have been retained by ChEBI, but these may have been modified to enhance clarity, avoid ambiguity and follow more closely current IUPAC recommendations on chemical nomenclature. For instance, CHEBI:18258 has the ChEBI Name '3,3′,5-triiodo-L-thyronine' rather than the more ambiguous 'triiodothyronine' or '3,5,3′-triiodothyronine'.

ChEBI ASCII Name is the ChEBI Name provided in ASCII format if the original includes special characters that require a Unicode presentation. For instance, β-D-galactopyranosyl-(1→4)-α-D-galactopyranose (CHEBI:36227) has the ChEBI ASCII Name 'beta-D-galactopyranosyl-(1→4)-alpha-D-galactopyranose'.

Definition. Where appropriate, the meaning of class names is explained by means of a short verbal 'definition'. For instance, the definition of organosulfonic acids (CHEBI:33551) is 'Organic derivatives of sulfonic acid in which the sulfo group is linked directly to carbon'.

Structural diagrams. ChEBI stores 2D or 3D structural diagrams as connection tables in MDL molfile format (http://www.mdl.com/solutions/). One entity can have one or more connection tables. Accordingly, one or more structures may be displayed for an entity. Where there is more than one structure available, one of these is designated the 'default structure' and is displayed on the Result screen for the entity; the additional ones may be viewed by clicking on the 'more structures' link beside this main displayed structure. By default, the diagrams are shown as static PNG images generated by MarvinBeans from ChemAxon (http://www.chemaxon.com/), while clicking on 'Applet' will open an interactive MarvinView applet which allows the structure to be manipulated. Clicking on 'Image' restores the static image view. A link is provided beneath a structure to the corresponding MDL molfile.

IUPAC InChI. The IUPAC InChI™ (8) is a non-proprietary identifier for chemical substances that can be used in printed and electronic data sources, thus enabling easier linking of diverse data compilations. It expresses chemical structures in terms of atomic connectivity, tautomeric state, isotopes, stereochemistry and electronic charge in order to produce a sequence of machine-readable characters unique to the respective molecule. In ChEBI, the InChI string is derived from the default structure using the free InChI software (http://www.iupac.org/inchi/).

SMILES. SMILES (Simplified Molecular Input Line Entry System) is a simple but comprehensive chemical line notation, created in 1986 by David Weininger (4) and further extended by Daylight Chemical Information Systems, Inc. (http://www.daylight.com/smiles/). SMILES specifically represents a valence model of a molecule and is widely used as a data exchange format. In ChEBI, the SMILES string is automatically generated from the default structure using ChemAxon MarvinBeans.

Formula. Where possible, formulae are assigned for entities and groups. For compounds consisting of discrete molecules, this is generally the molecular formula, a formula in accordance with the relative molecular mass (or the structure). To facilitate searching and downloading of data from external sources, the use of subscripts to indicate multipliers is avoided.

Ontology. Every ChEBI entry contains a list of parent and children entries and the names of the relationships between them (see the section 'ChEBI Ontology' below).

IUPAC name(s). The IUPAC name is a name provided for an entity based on current recommendations of IUPAC. It need not be fully systematic as it makes use of 'retained names'. For instance, the IUPAC name for abietic acid (CHEBI:28987) is abieta-7,13-dien-18-oic acid, based on the retained name 'abietane', rather than the fully systematic name (1*R*,4a*R*,10a*R*)-1,4a-dimethyl-7-(propan-2-yl)-1,2,3,4,4a,5,6,10,10a-decahydrophenanthrene-1-carboxylic acid (which is cited in ChEBI within the list of synonyms for this compound). In most cases, a single IUPAC name is provided for a molecular entity or a group. For organic compounds this name will, if necessary, be amended when the IUPAC rules for providing a 'Preferred IUPAC Name' for any organic compound are published (http://www.iupac.org/projects/2001/2001-043-1-800.html).

Synonyms. Synonyms are alternative names for an entity which either have been used in EBI or external sources or have been devised by the annotators based on recommendations of IUPAC, NC-IUBMB or their associated bodies. The source of each synonym is clearly identified (see 'Data sources' section). Systematic names may also be included in this section.

Database cross-references. A field 'Database Links' contains one or more manually entered accession numbers for entries in public databases (Table 1) relevant to the given ChEBI entry. In addition, a separate page called 'Automatic Xrefs' contains automatically generated cross-references to the IntAct (12) and UniProt (3) databases. Note that the cross-references to UniProt are based on text matching and can change from release to release.

Registry Number(s). The Chemical Abstracts Service (CAS) Registry Number is a unique numeric identifier assigned to a substance when it enters the CAS REGISTRY database (http://www.cas.org/EO/regsys.html).

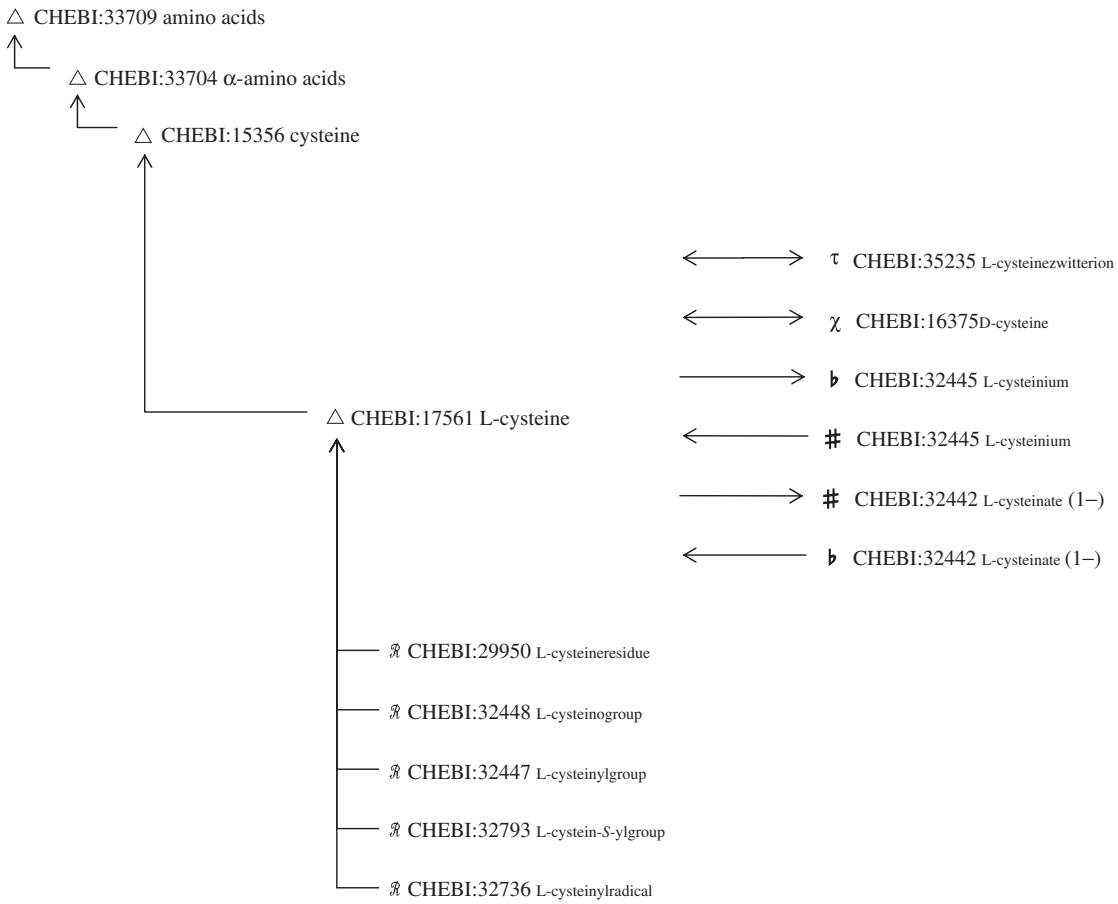

Figure 1. Fragment of ChEBI Ontology.

Registry Numbers have no chemical significance and are assigned in sequential order to unique, new substances identified by CAS scientists for inclusion in the database. Other registry numbers which may be displayed are Beilstein and Gmelin Registry Numbers (http://www.mdl.com/).

Comment(s). A free-text comment may be added to some terms especially in cases where confusing terminology has been historically used. A comment may relate to a single term or to the entry as a whole. For instance, the entry for riboflavin (CHEBI:17015) contains a comment relating to the synonym 6,7-dimethyl-9-D-ribitylisoalloxazine: 'Uses obsolete isoalloxazine skeletal numbering system'.

ChEBI Ontology

Ontologies are structured controlled vocabularies; generally they are graph-theoretic structures consisting of 'terms', which form the nodes of the graphs, linked by 'relations', which form the edges between the nodes (13). Ontologies are generally organized as directed acyclic graphs, i.e. any child term may have one or more parent terms. However this is not the case with ChEBI since a number of cyclic relationships are included (see below). A growing number of ontologies relevant to biological sciences are available from the OBO Foundry (http://obofoundry.org/).

ChEBI Ontology consists of four sub-ontologies (Table 3):

- Molecular Structure, in which molecular entities or parts thereof are classified according to structure;
- Biological Role, which classifies entities on the basis of their role within a biological context (e.g. antibiotic, coenzyme, hormone);
- Application, which classifies entities, where appropriate, on the basis of their intended use by humans (e.g. pesticide, drug, fuel);
- Subatomic Particle, which classifies particles smaller than atoms.

The Molecular Structure sub-ontology is largely constructed from terms approved by IUPAC and IUBMB. However the relationships between the terms are unique to the ChEBI Ontology (Figure 1). The Molecular Structure sub-ontology employs both plural and singular terms, where the plural terms refer to classes of chemical compounds. This follows the widely accepted practice in chemical nomenclature where classes are often named after a specific representative. For instance,

Table 3. ChEBI sub-ontologies

Sub-ontology	Definition	Example
Molecular structure	A description of the molecular entity or part thereof based on its composition and/or the connectivity between its constituent atoms.	CHEBI:23091 ChEBI ontology └◇CHEBI:24431 molecular structure └ΔCHEBI:23367 molecular entities └ΔCHEBI:33259 homoatomic molecular entities └ΔCHEBI:33262 elemental oxygen └ΔCHEBI:33263 diatomic oxygen └ΔCHEBI:15379 dioxygen └ΔCHEBI:26689 singlet dioxygen └ΔCHEBI:27140 triplet dioxygen
Subatomic particle	A particle smaller than an atom.	CHEBI:23091 ChEBI ontology └◇CHEBI:36342 subatomic particle └ΔCHEBI:33233 fundamental particle └ΔCHEBI:36338 lepton └ΔCHEBI:10545 electron
Biological role	A role played by the molecular entity or part thereof within a biological context.	CHEBI:23091 ChEBI ontology └◇CHEBI:24432 biological role └ΔCHEBI:33280 molecular messenger └ΔCHEBI:24621 hormone └ΔCHEBI:28918 (*R*)-adrenaline
Application	Intended use of the molecular entity or part thereof by humans.	CHEBI:23091 ChEBI ontology └◇CHEBI:33232 application └ΔCHEBI:25944 pesticide └ΔCHEBI:22153 acaricide └ΔCHEBI:38593 fenazaquin

phenols (CHEBI:33853) is a class that includes a specific compound, phenol (CHEBI:15882).

Relationships. There is, in the OBO community, an effort to standardize the relationships used in biomedical ontologies (13). Two of the relationships used in ChEBI are defined by the Relations Ontology (http://www.obofoundry.org/cgi-bin/detail.cgi?id=relationship), that is *is a* and *is part of*, but others are new and are specifically required by ChEBI (Table 4). Another significant difference from a 'classic' OBO such as Gene Ontology is that some of the ChEBI relationships are necessarily cyclic. The relationship 'A *is conjugate acid of* B' means that the relationship 'B *is conjugate base of* A' is always true, while the relationships 'E *is tautomer of* K' and 'R *is enantiomer of* S' also mean that 'K *is tautomer of* E' and 'S *is enantiomer of* R' are always true. The members of these cyclic relationships are placed at the same hierarchical level of the ontology. The relationships were introduced out of a need to formalize the differences between terms that are often (incorrectly) interchangeably used, especially in the biochemical literature. For example, 'lactate' is frequently used as a synonym of 'lactic acid'. In ChEBI, lactate (CHEBI:24996) *is conjugate base of* lactic acid (CHEBI:28358).

Web access

ChEBI can be accessed via the web at http://www.ebi.ac.uk/chebi/. The user can browse the database via the ChEBI Periodic Table or via the ontology as well as by use of a simple or advanced textual search. For instance, an InChI (or a fragment of an InChI) can be used as a query. A true chemical (sub)structure search facility is planned for the near future.

Web Services

The main aim of ChEBI Web Services is to provide programmatic access to the ChEBI database in order to aid our users in integrating ChEBI into their applications. Web Services (http://www.w3.org/2002/ws/) provide a standard means of interoperating between different software applications. To ensure that software from various sources work well together, this technology is built on open standards such as Simple Object Access Protocol (SOAP), a messaging protocol for transporting information, and Web Services Description Language (http://www.w3.org/TR/wsdl), a standard method of describing Web Services and their capabilities. For the transport layer itself, Web Services utilize most of the commonly available network protocols, especially Hypertext Transfer Protocol (HTTP). ChEBI provides SOAP access to its database. ChEBI Web Services are made available at http://www.ebi.ac.uk/chebi/webServices.do. To obtain 'lightweight' ontology objects, the Ontology Lookup Service (19) (http://www.ebi.ac.uk/ontology-lookup/) can be used as alternative Web Services.

Feedback

A SourceForge Forum (http://sourceforge.net/projects/chebi/) has been established and is used to report bugs, discuss annotation issues and request new ChEBI terms or entries. Alternatively, one can send an email to the ChEBI help desk via a form (http://www.ebi.ac.uk/chebi/emailChebiForward.do).

PubChem deposition

PubChem (http://pubchem.ncbi.nlm.nih.gov/) is an open repository of chemical structure established at the National Center for Biotechnology Information.

Table 4. Relationships in ChEBI ontology.

Relationships in ChEBI ontology			
Relationship	Symbol	Description	Example
is a	Δ	Relationship between more specific and more general concepts.	(*R*)-lactate (CHEBI:16004) *is a* lactate (CHEBI:24996)
is part of	◇	Relationship between part and whole.	tetracyanonickelate(2−) (CHEBI:30025) *is part of* potassium tetracyanonickelate(2−) (CHEBI:30071)
is conjugate acid of	#	A pair of relationships used to connect acids with their conjugate bases.	lactic acid (CHEBI:28358) *is conjugate acid of* lactate (CHEBI:24996)
is conjugate base of	b		lactate (CHEBI:24996) *is conjugate base of* lactic acid (CHEBI:28358)
is tautomer of	t	A cyclic relationship used to show the interrelationship between two tautomers.	9*H*-purine (CHEBI:35589) *is tautomer of* 1*H*-purine (CHEBI:35586)
is enantiomer of	X	A cyclic relationship used in instances when two entities are mirror images of and non-superposable upon each other.	D-cysteine (CHEBI:16375) *is enantiomer of* L-cysteine (CHEBI:17561)
has functional parent	f	The relationship between two molecular entities (or classes of entities), one of which possesses one or more chacteristic groups from which the other can be derived by functional modification.	codeine (CHEBI:16714) *has functional parent* morphine (CHEBI:17303)
has parent hydride	*	The relationship between an entity and its parent hydride.	perfluorodecane (CHEBI:38851) *has parent hydride* decane (CHEBI:32894)
is substituent group from	/	The relationship between a substituent group (or atom) and its parent molecular entity, from which it is formed by loss of one or more protons or simple groups such as hydroxy groups.	ethyl group (CHEBI:37807) *is substituent group from* ethane (CHEBI:23975) L-cysteino group (CHEBI:32448) *is substituent group from* L-cysteine (CHEBI:17561)

The default structures of molecular entities from ChEBI are deposited into PubChem monthly.

Internationalization

In addition to English, the ChEBI home page is available in French, German, Russian and Spanish; the User Manual and Frequently Asked Questions are available in English and German.

AVAILABILITY

All data in the database and on the FTP server is non-proprietary or is derived from a non-proprietary source. It is thus freely accessible and available to anyone. In addition, each data item is fully traceable and explicitly referenced to the original source. Apart from web access, the entire ChEBI data is provided in four different formats and can be downloaded from the FTP server (ftp://ftp.ebi.ac.uk/pub/databases/chebi/).

Flat-file table dumps—ChEBI is stored in a relational database and we currently provide the ChEBI tables in a flat-file tab delimited format. There are various spreadsheet tools available to import this into a relational database. The files are stored in the same structure as the relational database.

Oracle binary table dumps—ChEBI provides an Oracle binary table dump that can be imported into an Oracle relational database using the 'imp' command. The parameter file import.par should reside in the same directory when the import is done. The correct command to execute is:

imp database_name/database_password@Instance_name PARFILE = import.par

Generic Structured Query Language (SQL) *table dumps*—ChEBI provides a generic SQL dump which consists of SQL insert statements. The archive file called generic_dump.zip consists of 12 files which contain SQL table insert statements of the entire database. The file called compounds.sql should always be inserted first in order to avoid any constraint errors. Included in the folder are MySQL and PostgreSQL scripts for creating the tables in the user's database. These insert statements should be usable in any database which accepts SQL as its query language.

OBO ontology format—ChEBI provides the ChEBI ontology in OBO format version 1.2 (http://www.geneontology.org/GO.format.obo-1_2.shtml). The open-source ontology editor OBO-Edit (20) (http://oboedit.org/) can be used to view the OBO file.

CONCLUSIONS

Compared with established commercial chemistry resources, ChEBI is a small database. However the

strength of ChEBI lies in its quality. The ultimate goal of ChEBI is to provide and promote a 'gold standard' of annotation for molecular entities, which comprises a controlled vocabulary (standardized and unambiguous terminology), graphical representations of molecular structure (clear and unambiguous 2D diagrams), and defined logical relationships between concepts (ontology).

ACKNOWLEDGEMENTS

The ChEBI project is supported by the European Commission under FELICS, contract number 021902 (RII3). We are grateful to Rolf Apweiler, Volker Ast, David Binns, Bernard de Bono, Dimitris Dimitropoulos, Henning Hermjakob, Pankaj Jaiswal, Michael Kleen, Ernst Kretschmann, Nicolas Le Novère, Mark Rijnbeek, Karine Robbe, Philippe Rocca-Serra, Janet Thornton, Daniela Wieser and all users of ChEBI. Funding to pay the Open Access publication charges for this article was provided by the European Bioinformatics Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Caspi,R., Foerster,H., Fulcher,C.A., Hopkinson,R., Ingraham,J., Kaipa,P., Krummenacker,M., Paley,S., Pick,J. *et al.* (2006) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.*, **34**, D511–D516.
2. Kulikova,T., Akhtar,R., Aldebert,P., Althorpe,N., Andersson,M., Baldwin,A., Bates,K., Bhattacharyya,S., Bower,L. *et al.* (2007) EMBL Nucleotide Sequence Database in 2006. *Nucleic Acids Res.*, **35**, D16–D20.
3. The UniProt Consortium. (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **35**, D193–D197.
4. Weininger,D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, **28**, 31–36.
5. Degtyarenko,K., Ennis,M. and Garavelli,J. (2007) "Good annotation practice" for chemical data in biology. *In Silico Biol.*, **7**, S1, 06.
6. McNaught,A.D. and Wilkinson,A. (eds.) (1997) *Compendium of Chemical Terminology ("The Gold Book")* 2nd edn. Blackwell Scientific Publications, Oxford, p. 262
7. McNaught,A.D. and Wilkinson,A. (eds.) (1997) *Compendium of Chemical Terminology ("The Gold Book")* 2nd edn. Blackwell Scientific Publications, Oxford, p. 177
8. Heller,S.R. and McNaught,A.D. (2006) The IUPAC International Chemical Identifier, InChI. In Coghill,A.M. and Garson,L.R. (eds), *The ACS Style Guide,* 3rd edn. Oxford University Press, New York, pp. 101–102.
9. Murray-Rust,P. and Rzepa,H.S. (2003) Chemical markup, XML, and the World Wide Web. 4. CML schema. *J. Chem. Inf. Comput. Sci.*, **43**, 757–772.
10. Fleischmann,A., Darsow,M., Degtyarenko,K., Fleischmann,W., Boyce,S., Axelsen,K.B., Bairoch,A., Schomburg,D., Tipton,K.F. *et al.* (2004) IntEnz, the integrated relational enzyme database. *Nucleic Acids Res.*, **32**, D434–D437.
11. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
12. Kerrien,S., Alam-Faruque,Y., Aranda,B., Bancarz,I., Bridge,A., Derow,C., Dimmer,E., Feuermann,M., Friedrichsen,A. *et al.* (2007) IntAct—open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**, D561–D565.
13. Smith,B., Ceusters,W., Klagges,B., Köhler,J., Kumar,A., Lomax,J., Mungall,C., Neuhaus,F., Rector,A.L. *et al.* (2005) Relations in biomedical ontologies. *Genome Biol.*, **6**, R46.
14. Degtyarenko,K. and Contrino,S. (2004) COME: the ontology of bioinorganic proteins. *BMC Struct. Biol.*, **4**, 3.
15. Sud,M., Fahy,E., Cotter,D., Brown,A., Dennis,E.A., Glass,C.K., Merrill,A.H.Jr, Murphy,R.C., Raetz,C.R.H. *et al.* (2007) LMSD: LIPID MAPS structure database. *Nucleic Acids Res.*, **35**, D527–D532.
16. Berman,H., Henrick,K., Nakamura,H. and Markley,J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.
17. Garavelli,J.S. (2004) The RESID database of protein modifications as a resource and annotation tool. *Proteomics*, **4**, 1527–1533.
18. Ellis,L.B.M., Roe,D. and Wackett,L.P. (2006) The University of Minnesota Biocatalysis/Biodegradation Database: the first decade. *Nucleic Acids Res.*, **34**, D517–D521.
19. Côté,R.G., Jones,P., Apweiler,R. and Hermjakob,H. (2006) The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, **7**, 97.
20. Day-Richter,J., Harris,M.A., The Gene Ontology OBO-Edit Working Group, Haendel,M. and Lewis,S. (2007) OBO-Edit—an ontology editor for biologists. *Bioinformatics*, **23**, 2198–2200.