ChEBI: The online chemical dictionary for small molecules

Namrata Kale [1]
Chemical biology
Beginner
1 hour

Chemical Entities of Biological Interest (ChEBI) [2] is a user-friendly online chemical dictionary that focuses on the nomenclature, structure, and biological properties of 'small' molecules that may be encountered by anyone working in (or merely with an interest in) molecular biology. This course will show you what kind of information is available in ChEBI and how to access and query the database.

Learning objectives:

- Evaluate what ChEBI is and how it can be useful in your day-to-day research
- Learn what data are stored and how ChEBI classifies molecules (understanding ChEBI ontology).
- Be able to carry out different types of search.

What is ChEBI?

ChEBI [3] is a freely available, open source bioinformatics and cheminformatics resource that provides manually annotated information on 'small' chemical entities – constitutionally or isotopically distinct atoms, molecules, ion pairs, radicals, radical ions, complexes, conformers, or anything else that is a distinguishable entity (Figure 1).

In biology, metabolites are a typical example of small molecules [4]. Molecules directly encoded by the genome (e.g. nucleic acids, proteins and peptides derived from proteins by cleavage) are not as a rule included in ChEBI.
Figure 1 The ChEBI homepage is easily accessible by clicking on the ChEBI Link in the Chemical biology panel on the EBI services page [5].

ChEBI focuses on nomenclature endorsed by the International Union of Pure and Applied Chemistry (IUPAC [6]) and the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB [7]). A major feature of ChEBI is that it includes a Chemical Ontology [8] – a form of classification to specify relationships between molecular entities or classes of entities and their parents and/or children in a structured way.

Where does the data come from?

Chemical entities curated into ChEBI originate from a variety of sources, including:

- Automatic loading of data from selected resources, such as the Ligand [9] section of the Kyoto Encyclopedia of Genes and Genomes (KEGG Compound), the Human Metabolome Database (HMDB), etc;
- Data annotated elsewhere (e.g. user submissions), which are subsequently checked and updated by ChEBI curators;
- Manual annotation [10] by the ChEBI team, covering a variety of requests from users via the ChEBI SourceForge channel, user emails, etc.
The ChEBI star rating system

- Preliminary data gets 1 star and is not normally displayed on the ChEBI website.
- Data from externally curated sources (e.g. imported from KEGG) but not yet checked by ChEBI gets 2 stars and appears "as is" on the ChEBI website.
- After an entry has been updated with additional links, cross-references, etc, and checked by a member of ChEBI curation team, it is upgraded to 3 stars.

Why do we need ChEBI?

Molecular biologists and biochemists encounter small molecules on a regular basis – after all, small molecules participate in all the processes of life: for example, as neurotransmitters, as intermediates in cell metabolism, in biological pathways and in drug design. However, although the field of ‘small molecules’ is important, it is not the prime interest of most biologists. As a consequence, small molecules frequently appear in databases simply as free text, which leads to problems. We will discuss some examples of these in the next section.

Did you know?

Where ChEBI records have been merged, the ChEBI IDs of the merged records are never lost. One is retained as the primary ID whilst the others are listed as secondary IDs. Searching for either primary or secondary IDs will take you to the same record.

Examples of why we need ChEBI

Chemical Nomenclature - part 1

Problem:

Often a single molecule can have a large number of valid chemical names (Figure 2), whilst a single or a chemically ambiguous common name can be used for more than one molecule. This causes major problems when trying to find all references to a compound in the scientific literature.
Paracetamol, acetaminophen, 4-acetamidophenol, N-(4-hydroxyphenyl)acetamide, panadol, tylenol...?

**Figure 2** This drug is generally known as paracetamol in the UK but as acetaminophen in US. It also has numerous names based on its structure (including those conforming to either IUPAC [6] or Chemical Abstracts naming rules), brand names, etc.

**ChEBI Solution:**
ChEBI reduces ambiguity by providing a unique and recommended name along with a stable ChEBI identifier. It also includes a recommended IUPAC name and a collection of synonyms, including brand names and International Nonproprietary Names (INNs) for drugs, which can be used by text mining applications.

**Try it yourself**
1. Go to [ChEBI homepage](#) [2] and type “tylenol” into the search box. What is the ChEBI name for this compound?

2. One of the secondary IDs (to the right of the structure) is CHEBI:46191. Return to ChEBI homepage and search for "46191". What is the ChEBI ID of the record you find?
Chemical Nomenclature - part 2

Problem:
Another problem with chemical nomenclature is the use of same chemical name for more than one substance (Figure 3).

Figure 3 Adrenaline can have several possible structures.

ChEBI Solution:
ChEBI names are more specific and are followed by an appropriate definition.

Try it yourself
Go to ChEBI homepage [2] and type "adrenaline" into the search box. Find the biologically active isomer.

Stereochemistry
Problem:
Missing stereochemical representation of biologically important molecules in the research literature.

ChEBI Solution:
ChEBI provides stereospecific illustrations in the form of a 2D and 3D figure, where possible. Additional structural information is also available in the form of IUPAC [6] International Chemical Identifier [13] (InChI) and Simplified Molecular Input Line Entry System (SMILES [14]).

Try it yourself
Go to ChEBI homepage [15] and type “cholesterol” into the search box. Find the stereochemistry.

Annotation of Bioinformatics data

Problem:
Chemical annotations are commonly captured as free text. The terminologies used vary from one annotator to another, so it can be difficult to interpret the intended meaning.

ChEBI Solution:
ChEBI uses standard and consistent vocabularies or ontologies with clearly defined terms (see ChEBI Ontology [16]) for annotations that enable semantic compatibility for machine searching of biological information.

Updating existing information

Problem:
It can take many years to identify natural products, involving a long series of publications that cover the initial identification of a molecular skeleton (and its possible revision), followed by identification of some stereocentres (and their possible subsequent revisions), and eventually the full identification of the complete structure.

ChEBI Solution:
ChEBI is a manually curated database where the information stored is frequently updated with useful literature resources and database cross-references.
How ChEBI can help with your research

When to use ChEBI

You can use ChEBI to:

- find the stereochemistry and structure of an entity by searching for any relevant name;
- find the preferred chemical terminology for an entity by searching for any name, formula or registry numbers associated with it, including Chemical Abstract Service Registry Number (CAS-RN) and Reaxys numbers;
- view the relationship between chemicals and their classes using the ChEBI ontology.

When not to use ChEBI

ChEBI cannot help you:

- to find biological macromolecules, such as nucleic acids, proteins and large peptides;
- if you need a large set of molecules for molecular modelling, binding studies, etc; the ChEMBL database [17] would be more appropriate.

If the entity you are interested in is not listed in ChEBI

- then you can submit it to ChEBI yourself using the ChEBI Submission Tool;
- or you can ask the ChEBI team to add the entity to the database using ChEBI's SourceForge Curation [18] Requests channel.

How to search ChEBI

You can search ChEBI [2] (Figure 4) using either a free text search or a targeted search via the 'Advanced Search' option. The wildcard character: * is available for both the simple and the advanced search. The following sections will take you through both of these options.
Searching the ChEBI database using the simple text search

Use the search box on the ChEBI home page to enter simple search terms:

1. Name of molecule e.g. paracetamol;
2. ChEBI identifier e.g. CHEBI:16526;
3. IUPAC [6] International Chemical Identifier (InChI) e.g. InChI=1S/C02/2-1-3;
4. Molecular formula e.g. $C_6H_{12}O_6$;
5. Chemical Abstract Service registry number (CAS-RN) e.g. 124-38-9.

Guided example: searching for a simple molecule:

Let's start with a simple molecule search e.g. paracetamol as shown in Figure 5.
Figure 5 ChEBI search by typing the name of a simple molecule.

Once the search is complete, a results page will be returned showing the ChEBI matches to your query (Figure 6). Search results may be downloaded as TSV [19], XML [20] or SDF [21] files for import into other applications.

Figure 6 Search results page for paracetamol showing paracetamol and its structural derivatives.

Try it yourself?

Type the molecular formula C₆H₁₂O₆ into the search box. How many entries can you find?
Advanced search

Searches in ChEBI can be customised and filtered using advanced search options (Figure 7):

- drawing a structure and searching by "exact", "substructure" or "similar structure" searches;
- restricting text searches to individual fields in the database (e.g. searching in the 'name' and 'synonym' fields only), and combining the results of several such searches using AND, OR, and BUT NOT logic;
- filtering the results by stars to search for only 3-star entries;
- controlling the number of results displayed by using drop-down boxes associated with "results per page" and "total results";
- combining the results of a structure search with one or more text searches;
- the Advanced Search page is also used for searching the ChEBI Ontology (but more about this later!)

![Advanced Search page in ChEBI showing text search and structure search.]

Structure search

- The sketch pad (powered by Ketcher [22]) is an open source tool that can be used for sketching the structure (or substructure) of a molecule of interest (Figure 8).
- Bonds, rings and electronic charges can be selected from the icons on the left-hand-side of
the sketch pad, whilst icons for the commonly encountered elements are on the right. Rarer elements are available via a periodic table icon (bottom right).

- "Radio buttons" to the right of the sketchpad can be used to select an exact search ("Find this entity"), a substructure search ("Find compounds which contain this structure") or a similarity search ("Find compounds which resemble this structure") in ChEBI [12].

**Figure 8** Showing sketchpad with drawing tools and template for periodic table.

**Guided example for a structure search in ChEBI**

**Try it yourself?**

- Draw a structure in the sketch pad using the drawing tools.
- Select "find compounds which contain this structure" for a substructure search.
- The search results are shown in a grid (Figure 9).
- Results are paginated if more than 15 results are returned.
- Click the relevant ChEBI accession [23] number (hyperlinked under the search result) to go to the entry page of that entity.
- Structures can be zoomed by hovering over the relevant compound.
**Advanced text search in ChEBI has several options**

Type your query as free text in the search box and filter results by using or combining categories - an example is shown in Figure 10:
Figure 10 Advanced text search options in ChEBI.

- All – this allows you to search all the categories;
- ChEBI ID – allows searching for specific ChEBI identifiers;
- ChEBI name – will search only for ChEBI names matching your search term;
- Definition – the ChEBI definitions;
- All Names – will search all the synonyms, ChEBI Names and IUPAC [6] names available for this compound;
- IUPAC name – will search for IUPAC names;
- Database Links – allows searching for accession [23] numbers from other sources;
- Formula – will search for formula;
- Mass – the molecular weight;
- Charge – allows searching for the charge;
- CAS Registry Number [24] - will search for CAS Registry Numbers matching your search criteria;
- InChI/InChIKey [25] - will search for InChIs matching your search criteria;
- SMILES [14] - will search for SMILES matching your search criteria.

The plus and minus tabs can be used to add additional categories.
ChEBI ontology

Introduction to ChEBI ontology - part 1

ChEBI entities are annotated using standard vocabularies or ontologies. These consist of three sub-ontologies: (see Figure 11)

1. **Molecular structure**, in which molecular entities or parts of entities are classified according to their structure;
2. **Role**, in which entities are classified on the basis of their role within a biological context, e.g. as antibiotics, antiviral agents, coenzymes, enzyme inhibitors, or on the basis of their intended use by humans, e.g. as pesticides, detergents, healthcare products, fuel;
3. **Subatomic particle**, in which entities are classified as particles that are smaller than atoms.

![ChEBI ontology for (R)-adrenaline.](image)

Introduction to ChEBI ontology - part 2

The ChEBI [12] ontology [8] uses two generic plus several chemistry-specific relationships (Figure 12).
**Figure 12** Various ontology relationships in ChEBI.

- **Is a**: Entity A is an instance of Entity B. For example, ethanol *is a* primary alcohol;
- **Has part**: Indicates a relationship between a part and the whole;
- **Is conjugate base of and is conjugate acid of**: Relationships used to connect acids with their conjugate bases. For example, L-cysteine *is conjugate base of* L-cysteinium (Figure 13A);
- **Is tautomer of**: Cyclic relationship used to show the relationship between two tautomers (Figure 13B);
- **Is enantiomer [26] of**: Cyclic relationship used in instances when two entities are non-superposable mirror images of each other. For example, D-cysteine *is enantiomer* of L-cysteine (Figure 13C);
- **Has functional parent**: Denotes the relationship between two molecular entities or classes, one of which possesses one or more characteristic groups from which the other can be derived by functional modification. For example, S-(4-bromophenyl)-L-cysteine *has functional parent* L-cysteine (Figure 13D).
• **Has parent hydride**: Denotes the relationship between an entity and its parent hydride (Figure 14E);
• **Is substituent group from**: Indicates the relationship between a substituent group/atom and its parent molecular entity (Figure 14F);
• **Has role**: Denotes the relationship between a molecular entity and the particular behaviour which the entity may exhibit either by nature or by human application, for example, morphine *has role* as an opioid analgesic.
Browsing the ChEBI ontology

To view the ontology classification for any entity in ChEBI, click on the "ChEBI Ontology" tab at the top part of any ChEBI entry page (Figure 15)

Scrolling down shows the various relationships in a "graph" or "tree view" of the entity. The graph view shows the structural classification (is a and has part relationships). It is fully interactive (Figure 16):

- use the zoom and navigation tools to the left of the graph to move and resize the graph;
- click and drag on any name to to move individual class names, to clarify the graph if necessary;
- click on any name to see the definition for that entry;
• hover over any line to view the relationship being depicted;
• thick blue lines indicate hidden relationships - click on them to reveal the hidden paths;
• the "Expanded view" radio button can be used to display all of the hidden paths.

![ChEBI ontology for paracetamol.](image)

**Figure 16** ChEBI ontology for paracetamol.

### Searching the ChEBI ontology

The ChEBI ontology can be searched using "Filter by Ontology Term" on the ChEBI Advanced Search Page (Figure 17).

- Select a desired relationship (e.g. *is a, has role etc.*) from the drop-down menu of "Select relationship";
- Enter the ontology term (e.g. fatty acid, alcohol etc.) into the ontology search box. When entering the ontology term into the ontology search box, it is important to type in a few characters, allow time for possible matches to be suggested, then to select the desired term from the suggested list (type a few more characters one at a time if the desired term isn't on the initially-suggested list).
- Extra relationships can be added/removed by using the plus/minus tabs adjacent to query fields.
Summary

- ChEBI is an encyclopedia focused on small chemical entities.
- ChEBI focuses on high quality manual annotation, non-redundancy, and provision of a chemical ontology rather than full coverage of the vast chemical space.
- ChEBI ontology provides standardised descriptions of molecular entities that enable other databases at the EMBL-EBI and worldwide to annotate their entries in a consistent fashion.

Exercises

The exercises allow you to apply your knowledge gained on this course by providing examples of how ChEBI can be used and asking you to solve the task given. You can start by clicking on one of the exercise titles provided.

If you need help to complete this section you can look in the 'Need some help?' and 'Want to know how we did it?' sections.
ChEBI Advanced Search

Scenario

"Lipids" is a loosely defined term for naturally occurring molecules, including glycerides (fats and oils), phospholipids and sterols. As a researcher, you want to explore the different intermediates or metabolites (e.g fatty acids) involved in the biosynthesis of lipids.

Exercise

Use ChEBI to find relevant fatty acids.

Need some help?

1. Add your search criteria in the text box.
2. In the output view for your search, open the molecule of interest.
3. Find the relevant information in the entry page.

Want to know how we did it?

1. Use the ChEBI Advanced search page.
2. Use the 'Filter by Ontology Term' query to find all the entities which have a relationship is_a fatty acid.
3. To filter results with chemical structures, click on 'Yes' in the option to 'Filter by Chemical Structure: Results only with chemical structures?' (at the bottom of the text query section).
4. Click on the desired molecule to gain the relevant information you are looking for.

Quiz: ChEBI quiz

<table>
<thead>
<tr>
<th>Questions:</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attempts allowed:</td>
<td>Unlimited</td>
</tr>
<tr>
<td>Available:</td>
<td>Always</td>
</tr>
<tr>
<td>Pass rate:</td>
<td>75 %</td>
</tr>
<tr>
<td>Backwards navigation:</td>
<td>Allowed</td>
</tr>
</tbody>
</table>
Your feedback

Please tell us what you thought about this course. Your feedback is invaluable and helps us to improve our courses and thus enhance your learning experience.

Learn more

Publications:


Keep up to date:

Follow ChEBI on Twitter [30].

Get help and support on ChEBI

Support

- For more detailed information about how to use ChEBI, see the ChEBI user manual [31].
- For information about the ChEBI mailing lists and forums, see our SourceForge page [32].
- For other support-related enquiries, please contact the ChEBI support team [33].

Funding

ChEBI is funded by BBSRC, grant agreement number BB/K019783/1 within the "Bioinformatics and biological resources" fund.

Contributors
Namrata Kale [34]

EMBL-EBI
Scientific Database Curator - Steinbeck team: Cheminformatics and metabolism

Namrata S. Kale obtained her PhD in natural product chemistry from Indian Institute of Technology, New Delhi, India. She is currently a member of the Cheminformatics and Metabolism group at EMBL-EBI, where she works as a curator for ChEBI and MetaboLights databases.

Gareth Owen [35]

EMBL-EBI
Scientific Database Curator

Gareth Owen is a member of the Cheminformatics and Metabolism group at the EBI, where he works as curator and project manager for the ChEBI database. Gareth obtained his PhD in synthetic organic chemistry from Leeds University. He continued practising bench chemistry in a collaborative project with the Biotechnology unit at Sheffield University, synthesising radioactive intermediates that were used as part of an effort to produce morphine from microorganisms. He subsequently moved into the area of cheminformatics, designing and building both reaction and molecule databases for ORAC Ltd and later for Synopsys and Accelrys, before joining EMBL-EBI in 2010.
molecules

Links
[1] http://www.ebi.ac.uk/training/online/trainers/Namrata%20Kale
[2] https://www.ebi.ac.uk/chebi/
[4] http://www.ebi.ac.uk/training/online/glossary/small-molecules
[8] http://www.ebi.ac.uk/training/online/glossary/ontology
[9] http://www.ebi.ac.uk/training/online/glossary/ligand
[12] http://www.ebi.ac.uk/training/online/glossary/chebi
[14] http://www.ebi.ac.uk/training/online/glossary/smiles
[15] https://www.ebi.ac.uk/chebi
[16] http://www.ebi.ac.uk/training/online/course/chebi-online-chemical-dictionary-small-molecules/chebi-ontology
[17] https://www.ebi.ac.uk/chembl/
[18] http://www.ebi.ac.uk/training/online/glossary/curation
[19] http://www.ebi.ac.uk/training/online/glossary/tsv
[20] http://www.ebi.ac.uk/training/online/glossary/xml
[21] http://www.ebi.ac.uk/training/online/glossary/sdf
[22] http://scitouch.net/opensource/ketcher
[23] http://www.ebi.ac.uk/training/online/glossary/accession
[24] http://www.ebi.ac.uk/training/online/glossary/cas-registry-number
[25] http://www.ebi.ac.uk/training/online/glossary/inchikey
[26] http://www.ebi.ac.uk/training/online/glossary/enantiomer
[30] https://twitter.com/chebit
[31] https://www.ebi.ac.uk/chebi/userManualForward.do
[33] https://www.ebi.ac.uk/chebi/emailChebiForward.do
[34] http://www.ebi.ac.uk/training/online/trainers/Namrata%20Kale
[35] http://www.ebi.ac.uk/training/online/trainers/gowen