UniChem: Quick tour

Chemical biology
Beginner

This quick tour provides a brief introduction to UniChem [2], a compound identifier mapping service. It covers the scientific principles behind the tool, and outlines the use of UniChem and UniChem 'Connectivity Search' to directly locate identical and related molecules in various sources.

Learning objectives:

- Gain a basic understanding of UniChem and the benefits it brings to navigating small molecule resources
- Know how to conduct simple searches with UniChem and UniChem Connectivity Search

What is UniChem?

UniChem is a simple, large-scale, non-redundant database of pointers between chemical structure identifiers. Its purpose is to optimise the efficiency with which structure-based hyperlinks are built and maintained between chemistry databases. This is particularly suitable for creating links 'on the fly' by the use of REST web services. Primarily, this service has been designed to maintain cross references between EMBL-EBI chemistry data. These include primary chemistry databases (ChEMBL [3] and ChEBI [4]); and secondary resources which may contain small molecule information (Expression Atlas [5], PDBe [6]).

Why do we need UniChem?

When exploring small molecule information held in a particular database, a common question for many people might be "is there additional data on the same or similar molecules in another resource?". Sometimes hyperlinks on the page can direct you to other websites and repositories where you can find this data. However, these links can quickly become out of date, or only link to a small number of other resources; often the links will not be present at all.

This course focuses on how UniChem can be used to look for such additional data. If implemented within a database, UniChem can work in the background to create and maintain these links for the benefit of all resource users. The advantage is that UniChem is free and adds no ongoing maintenance cost to the resource that employs it. If UniChem is not integrated into a site, users can still find related molecules by using the UniChem web tool directly.

What can I do with UniChem?
You can use UniChem [7] to:

- Search for identical small molecule data across a wide variety of chemically aware web resources, using either src_compound_ids [8] (the identifiers for the small molecules [9] in different resources), InChI [10] or InChIKeys.

You can use UniChem 'Connectivity Search' to:

- Search for molecules with the same atom connectivity but different stereochemistry or isotopic composition or existing in a different salt form, using either src_compound_ids or InChIKeys.

You can use UniChem and 'Connectivity Search' to:

- Automatically generate hyperlinks from within a resource to other resources. An example can be seen at the bottom of this ChEMBL [11] compound report page [12]. Ask your developers to include similar automatically generated UniChem hyperlinks in your resource.

Searching using UniChem

Our molecule of interest

Suppose you are researching the small molecule yohimbine - an indole alkaloid that acts as an antagonist of alpha₂ adrenoceptors. You have located the yohimbine entry [13] in ChEBI, but wish to find other resources that might have additional information. This might include entries in other EMBL-EBI databases, other data resources, online compound libraries, patent databases or journal articles.

Look through the ChEBI entry and you will find an ID number - 10093 - the identifier [14] that denotes yohimbine in the ChEBI database. We can use this as our query term for UniChem [7]. Copy this ID and then navigate to UniChem [2].
Figure 1 ChEBI webpage showing yohimbine metadata

UniChem interface

UniChem [7] has a simple user interface as shown in the figure below.

The home page provides extra background on UniChem and more advanced usage of the resource. There is also a series of example queries at the bottom of the page that can help you to gain familiarity with using UniChem. Both of these are hidden as expandable boxes on the webpage. You can find additional information in the menu on the left-hand side, including a FAQ section, contact information and guidance for developers wanting to add UniChem to their own resource.
**src_compound_id search**

Once you have pasted the [ChEBI](https://www.ebi.ac.uk/chembl) [15] ID, you will need to click on the 'src_compound_id' button. This tells [UniChem](https://www.ebi.ac.uk/training/online) [7] that you are searching by compound ID, instead of using an InChI/InChIKey [16] identifier [14].

Clicking this button will activate a drop-down menu of chemical resource identifiers. Select the ChEBI option (number 7) and then submit your query.
UniChem search results

The results page provides a list of links to other resources that also contain entries on yohimbine. These include ChEMBL [17], KEGG [18], PubChem [19] and numerous patent records [20]. Each result provides a src_compound_id for the relevant resource (you can repeat the search using one of these new codes), a UniChem Identifier [14] (UCI) and the standard InChIKey [16] for the molecule (more on these later).
InChls and InChIKey search

InChls are International Chemical Identifiers, developed by the International Union of Pure and Applied Chemistry (IUPAC [21]). They are a text representation of molecule information that is in a human-readable form. They are formed from a number of layers that encode various properties of the molecule. These include the chemical formula, atom connections and hydrogen atoms in the first layer; charge, stereochemical and isotopic information are encoded in the 2nd, 3rd and 4th layers, respectively. The 5th and 6th layers are not normally used in standard InChls. InChIKeys are a hashed version of a full InChI used to facilitate speedier searching.

UniChem [7] provides the option to input either InChI or InChIKey [16] identifiers instead of src_compound_ids. Just paste in your query string, select the appropriate search criteria and click to submit the search. Continuing on with the previous example of yohimbine, you can try a UniChem search using the identifiers below:

- InChI=1S/C21H26N2O3/c1-26-21(25)19-15-10-17-20-14(13-4-2-3-5-16(13)22-20)8-9-23(17)11-12(15)6-7-18(19)24/h2-5,12,15,17-19,22,24H,6-11H2,1H3/t12-,15-,17-,18-,19+/m0/s1
- BLGXFZZNTVWALAY-SCYLSFHTSA-N
This query will return the same result data as on the previous tutorial - linking to resources with information on the same molecule.

You will notice that the molecule results all have the same InChIKey [16]. This is acceptable if you are only interested in data around this particular form of yohimbine. However, you might want to include other molecular forms of a molecule that are, for most purposes, identical to your original query. To undertake analyses of this nature, you can make use of UniChem Connectivity Search.

### Searching using UniChem Connectivity Search

#### Connectivity search

Our previous searches with UniChem have shown that yohimbine has an identical InChIKey for each entry in the accessed databases. These results do not include other forms of the yohimbine molecule: those with common atomic connectivity but differing stereochemistry or isotopic forms. These alternate forms might be of interest to an analyst's work, but would not be identified using the normal UniChem search tool.

Connectivity Search is an expanded version of the standard UniChem tool that allows us to find these related molecules. For example, you might be interested in data on yohimbine hydrochloride, tritiated forms of yohimbine, or even stereochemical variants of yohimbine, such as rauwolscine. You can see in the ChEMBL entries below how different molecules have very subtle differences in structure.
Figure 6 Chemical structures for: CHEMBL15245, CHEMBL1327758 and CHEMBL1472740

**Searching for alternative stereochemical and isotopic forms I**

Using the ChEBI ID for yohimbine (10093) you can conduct a wide search for data relating to the other molecular forms of the molecule. Instead of using the standard UniChem tool, access the 'Connectivity Search' [23] tool. As before, paste the ChEBI ID into the text box, select the 'src_compound_id' button, then select 'chebi' from the drop down list, and finally click submit. The results we returned are shown below.
Searching for alternative stereochemical and isotopic forms II

The search we conducted on the previous slide has returned a large number of hits for our query ID - 551 in total. With the basic UniChem search we only had a single hit in the ChEMBL database (CHEMBL15245). Now, with the Connectivity Search, we have 14 hits from ChEMBL. These results include our original exact match as well as stereochemical and isotopic variants of yohimbine.

The relationship between these hits and the query molecule can be discerned by exploring the links or by studying the columns in the result set, which indicate how different layers of the InChI vary from the equivalent layer in the query molecule. For example, sorting on the 'i' column shows hits that differ in isotopic composition to the query molecule. A table legend above the results provides details on the content displayed in each column.

Finding all alternative forms of a molecule

We have so far shown ways to access information and data on a particular small molecule, and similar variants of the query. It is also possible to use Connectivity Search to retrieve links to compounds which exist in different salt forms, such as Yohimbine Hydrochloride. The next slide will cover how to undertake this type of search.
To locate salt forms of a small molecule of interest we will conduct another Connectivity Search, as described on the previous slide. Once again we can use the ChEBI ID for yohimbine (10093). Instead of hitting 'Submit Query', click on the drop-down menu ‘...or click here and refine your query...’.

This opens an expanded query interface with additional options for optimising your search. There are a total of 8 different options for modifying the Connectivity Search (A-H, with 'H' only available for web-service calls). The use of most of these is beyond the scope of this introductory training, but full details can be found on the documentation page [24] for connectivity search. For now, simply select radio button number '4' (ie: 'Run all 0-3') under option 'C', and then run the query.

Note that the results page now contains a much larger number of hits. This is because now, not only salt forms of yohimbine, and mixtures containing yohimbine have been retrieved, but also salt forms and mixtures of stereochemical and isotopic variants. This can amount to rather a lot of information, which can be difficult to interpret. For this reason, sortable columns (annotated with information on how the query molecule differs from the retrieved molecule) allow you to intelligently browse the result set. These columns include flags to indicate common salt partners, and how the individual layers of the InChI differ between query and retrieved molecule. A table legend within an expandable drop-down box gives full details on the meanings of these columns.
UniChem 'Connectivity Search' results page for all alternative yohimbine structures

UniChem developer resources

UniChem [7] is not only available via a webpage, but can also be accessed as a web service via a REST [25] API [26].

If you are a developer and want to retrieve entries programmatically from the UniChem database, further information on the RESTful API [27] and the Connectivity Search [24] are available.

Get help and support on UniChem

Additional information on UniChem can be found at:

- The UniChem General Info [28] pages
- The UniChem FAQ [29]

If you cannot find the information you require on these pages, you can also email the unichem [at] ebi.ac.uk (UniChem team) for help.

You can also view a webinar introducing UniChem recorded on the 13 May 2015 here [30].
References


Contributors

**Jon Chambers** [1]
EMBL-EBI
Data Integration & Development Officer

Jon is a Data Integration and Development Officer for the ChEMBL team at EMBL-EBI.

He obtained a degree in biochemistry and then a PhD in molecular pharmacology. He then worked for many years in the pharmaceutical industry, working on cloned GPCRs and then various screening and data integration projects. He joined the ChEMBL Group at EMBL-EBI in 2010.

**Tom Hancocks** [34]
Tom Hancocks works as a Scientific Training Officer for the Training Team at EMBL-EBI.

He studied Human Genetics at the University of Leeds and McMaster University in Hamilton, Ontario; before completing an MSc in Analytical Genomics at the University of Birmingham.

Tom has worked for the NHS in diagnostic genetics and as a bioinformatics trainer for healthcare scientists and clinicians.

Funding

UniChem was developed with funding from EU-OPENSCREEN [35]. UniChem Connectivity Search and this e-learning module were developed as part of the BioMedBridges [36] project, funded by FP7 Capacities Specific Programme, grant agreement no. 284209.