

# Using MSDchem to Search the PDB Ligand Dictionary

The Protein Data Bank (PDB; *UNIT 1.9*) is an extremely valuable resource for understanding the three-dimensional (3-D) structure of proteins and interacting ligands. The PDB datafiles, however, do not provide clear and unambiguous information about chemical properties (e.g., bond orders, atom elements, and charges) for biological molecules. The possible ways that atoms in the molecules entered in the PDB are connected to form ligands and polymer residues are calculated from atom distances in the 3-D space. This is exactly what many protein visualization packages do—successfully in most, but not all, cases. Experimental errors and inaccuracies complicate things further; as a result, information about important chemical characteristics such as aromatic rings and chiral atoms is not directly accessible to scientists who want to understand the chemical structure of ligands they encounter in a PDB file.

The Macromolecular Structure Database (MSD), one of three that maintains the Worldwide Protein Data Bank (wwPDB), provides MSDchem, the definitive database of chemical records of PDB ligands (Bernstein et al., 1977). MSDchem contains data supplementary to the PDB archive that is exchanged among members of wwPDB (Berman et al., 2005; Golovin et al., 2004). These data provide explicit chemical definitions for standard and modified amino acids, nucleic acids, drugs, inhibitors, cofactors, and other chemical species included in PDB entries.

MSDchem is of use to structural biologists who want to resolve the chemical identity of a small molecule's 3-D structure and to chemists who are interested in a ligand's biological structure and function. MSDchem utilizes chemical software packages and resources including CACTVS (Ihlenfeldt et al., 1992; <http://www2.chemie.uni-erlangen.de/software/cactvs>) and CORINA (Gasteiger et al., 1990; <http://www.molnet.de/software/corina>). The CACTVS toolkit implements several checks on chemical consistency and functions to introduce additional molecular properties such as explicit stereodescriptors, aromatic flags, chemical drawings with PDB atom names, and unique SMILES strings (Weininger, 1988). CORINA is used to produce coordinates of an ideal 3-D conformation of each PDB ligand. MSDchem is an integral part of the Macromolecular Structure Search Database (MSDSD; Boutselakis et al., 2004) and is updated on a weekly basis with new and revised ligand definitions, resulting from significant curation and clean-up efforts by wwPDB. Many MSD and wwPDB tools reference this data (e.g., the Ligand Depot service described in *UNIT 1.9*).

The MSDchem search service offers various options for searching the ligand dictionary based on name, chemical formula, subgraph matching, or fingerprint similarity, as well as any combination of the above. While searching for a ligand using part of its code, name, synonym, or formula is useful in following literature or PDB file references, looking for molecules that contain a given chemical structure (subgraph searching) can be valuable when only an outline of the chemical diagram is known or when identifying variants of molecules that are expected to have similar chemical behavior in their common parts. On the other hand, chemical fingerprint similarity can be used to find ligands composed of a similar set of smaller subgroups, which may be connected differently but which have similar localized chemistry. Based on the results Web pages users may investigate, visualize, and export ligand structures or refer back to the relevant PDB entries. The MSDchem database is available for export in various formats: a ready-to-use relational database, collections of commonly used chemical data files, or SMILE string listings.

Four protocols are included in this unit, the first of which covers the simplest search option where the three-letter code or a part of the molecular name is known (Basic Protocol 1). This is the most popular option because it provides an overview of the ligand details from a literature reference. The next two protocols cover searching using a molecular formula or chemical fragment (Basic Protocol 2) and subgraph matching (Basic Protocol 3). These types of searching provide increasingly more powerful and accurate options for interactive use of MSDchem. Basic Protocol 4, which involves exporting the ligand dictionary, is for users who need to apply their own tools and methods to a local copy of the data collection.

## **BASIC PROTOCOL 1**

### **SEARCHING FOR LIGANDS USING THE THREE-LETTER PDB CODE OR MOLECULAR NAME**

The most common reason for using MSDchem is to have a look at the chemical diagram and properties of a ligand mentioned in a PDB file or to search the literature by either its common three-letter PDB code or a chemical name. This protocol demonstrates how to use MSDchem to perform this fundamental task and familiarizes the user with the MSDchem Web pages.

#### ***Necessary Resources***

##### *Hardware*

Computer with Internet access

##### *Software*

An up-to-date Internet browser, such as Internet Explorer 3.0 or later (<http://www.microsoft.com/ie>); Netscape 4.75 or later (<http://browser.netscape.com>); Firefox 1.0 or later (<http://www.mozilla.org/firefox>); or Safari (<http://www.apple.com/safari>)

#### ***Search for ligands using the three-letter PDB code***

1. Open the MSDchem search home page (<http://www.ebi.ac.uk/msd-srv/msdchem>; Fig. 14.3.1).

*This page is the starting point for simple and advanced searches of the ligand dictionary with combinations of individual search constraints and access to export functionality and relevant documentation.*

*The main area of the page provides version and summary information about the status of the database and the various text fields, controls, and buttons for selecting the search operators and invoking the constraint editors in order to build the value of constraints. Documentation about search fields can be found by following links from the data item labels (like "Molecule Name") or by using the adjacent question marks. There are various search operators for each search field that can be selected from the drop-down menu next to each search item name, and the most frequently used one is preselected.*

*The top header area of the page provides Web links to the MSD group page at EBI, the MSD Web services toolbox, introductory MSDchem documentation, and e-mail address contact for feedback and questions. There is also a link for accessing the "Energy types" section of the MSDchem data that is used as a source for refinement dictionaries of crystallographic software packages (Krissinel et al., 2004) and a direct shortcut back to the MSDchem search home page.*

*The left-hand menu area contains references to MSDchem guide, relevant literature and citations, and acknowledgments to software and resource contributors of MSDchem. This area has also links to alternative search pages and access to the ligand index and export pages.*

**Figure 14.3.1** The MSDchem search home page. The figure illustrates how to find the ligand with a three-letter code of ATP.

2. Add the three-letter code or the name of the ligand of interest. For example, type ATP in the “3 letter code” text field.

*The alternative Code text field is used when entering the MSD extended code, which in cases of topological variants can be different from the three-letter code. In the Molecule name text field, the user may input a part or a pattern of a molecular name. Both \* and % are accepted as wildcard expressions (that match any number of characters). When no wildcards are used, they are automatically assumed at both ends, and searches are case insensitive. For example, the ligand MIT, with the common name ARGATROBAN, systematic name (2R,4R)-4-methyl-1-(N2-[(3S)-3-methyl 1,2,3,4-tetrahydroquinolin-8-yl]sulfonyl)-L-arginylpiperidine-2-carboxylic acid, and synonyms MD-805 and MIT-SUBISHI INHIBITOR will match all following molecule name expressions: inhibitor, \*tetrahydroquinolin\* acid, and ARGA%.*

3. Click on the Search button and view the list of ligand results. There is a row for each one of the ligands (in this example, only one) that match the search criteria in the result page, with summary details that include the three-letter code, the common name, and the formula, as well as a small overview image of its chemical drawing.

On the top of the page there are links to the list of PDB entries and binding site details (Golovin et al., 2005) for the set of these ligands.

*Documentation can be found by following links from the data item names in the column headers just below the line reporting the number of results.*

#### **View a ligand details page**

4. Click on the three-letter code (the PDB reference) to navigate to an individual ligand details page. The resulting page is shown in Figure 14.3.2. In this page there is extensive data about the molecule, e.g., common and systematic names, stereo and nonstereo SMILE strings, formula and molecular charge, and number of total and heavy (non-hydrogen) atoms. There is also a larger chemical diagram of the molecule with atoms identified by their common PDB atom names. This diagram and other information in this page provide an understanding of the chemical context of the atoms observed in the PDB experiment. Atoms are colored based on their element type, bond orders, and stereo configurations, while aromatic bonds are displayed using gray instead of the black color used for other bond types.
5. Click on the Atoms link on the left hand side of the page to obtain more detailed data. The view shown in Figure 14.3.3 appears, providing a list of the atoms of the ligand with explicit stereodescriptors, aromatic flags, atomic charges, idealized 3-D coordinates, and other data at the atomic level.

*Additionally, one may get a summary page with all the entities that are part of or associated with the ligand using the Contents or the Complete contents links.*

*As usual, the documentation for all data items in all pages is available from links accessed by clicking on the data item names or the close-by question marks.*

#### **Visualize the ligand in three dimensions**

6. Select a coordinate set from the Library drop-down menu on the left side of the ligand details page (Fig. 14.3.2) by choosing one of the menu items described below.

**Ideal:** idealized 3-D coordinates that are generated automatically by the CORINA (Gasteiger et al., 1990) software package. CORINA does not use experimental data but only the molecule connectivity, bond orders, and chirality to produce a conformation of the molecule that is energetically favorable in isolation and visually elegant in the 3-D space.

**PDB:** the set of representative coordinates that wwPDB curators have manually chosen from all the occurrences of the ligand in PDB files. The “PDB representative conformation” of a ligand is chosen from the PDB file of an experiment with the best possible resolution after the curators make sure that there are no errors or conflicts with this coordinate set and the chemical structure of the ligand as given by its chemical diagram. This conformation is the result of the interaction of the ligand with a protein and is useful in understanding its biological function.

*The other menu option, PDB+H (representative non-hydrogen atom coordinates with idealized hydrogen coordinates), is not used in this step since hydrogen atoms are not visible for reasons of clarity.*

7. Select the viewer of preference from the Viewers drop down menu by choosing one of the following:

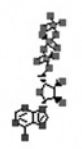
Jmol applet viewer, which will work with any browser without any other prerequisites, but is missing some functionality of other popular viewers, or Rasmol/rastop variant viewer, which must be installed by the user.


*This viewer must be configured as the “chemical/x-pdb” mime type handler of the computer, associated with .pdb files.*

home > searches > MSDChem MSD: Ligand Chemistry ? Energy types ? about help

**MSD Ligand Chemistry** [Get PDB entries](#)  
[Get PDB sites](#)

Molecule  
1 results

RecordCode	3 letter code	Extended Code	Molecule name	Stereo smile	Formula
1	ATP	ATP	ADENOSINE-5'-TRIPHOSPHATE		C10 H16 N5 O13 P3

 [EMBL-EBI](#) **Macromolecular Structure Data**

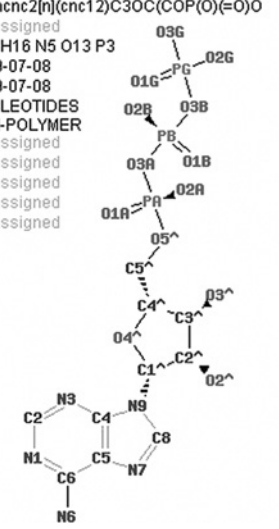
home > searches > MSDChem MSD: Ligand Chemistry ? Energy types ?

**MSD Ligand Chemistry**

Molecule  
ATP  
->

Distinct chemical molecule that is composed by atoms and bon

Code	ATP
3 letter code	ATP
Extended Code	not assigned
1 letter code	not assigned
Molecule name	ADENOSINE-5'-TRIPHOSPHATE
All atoms	47
All atoms except h	31
Formal charge	0
Stereo smile	Nc1ncnc2[n](cnc12)[C@@H]3O[C@H](C
Non stereo smile	Nc1ncnc2[n](cnc12)C3OC(COP(O)=O)O
Systematic name	not assigned
Formula	C10 H16 N5 O13 P3
Defined at	1999-07-08
Last modified at	1999-07-08
Classification	NUCLEOTIDES
Hetgroup type	NON-POLYMER
Polymer topology	not assigned
Polymer code	not assigned
Polymer sub type	not assigned
Obsoleted	not assigned
Parent	not assigned



Output:  Format:  Library:  Viewers:

hydrogens

[PDB entries](#)  
[Site Interactions](#)  
[Binding statistics](#)  
[As a ligand](#)  
[As ligand environment](#)

Retrieve: [XML](#) - [Perl](#) - [JavaScript](#)

**Figure 14.3.2** The MSDchem result page (top), listing the ligand with the three-letter code of ATP that matches the search criteria, and the ligand details page (bottom) with information about the ligand properties. Links to ligand content and related data, visualization and export functionality, and the PDB nomenclature chemical diagram are provided.

- Click on the View button to obtain one of the views (idealized or representative) shown in Figure 14.3.4.

#### Export ligand data in different file formats

- Select sdf from the Format drop down menu on the left side of the ligand details page (Fig. 14.3.2).

MSD Ligand Chemistry

Atom of molecule  
ATP.ATP-> Atoms

Atom of a chemical element, that composes a molecule

Record	Atom name in the molecule	EBI Ordering	PDB name	PDB Ordering	Atom stereochemistry	Element symbol	Is leaving atom	Is ring atom	Charge	Incomplete valence	X coordinate	Y coordinate	Z coordinate
1	PG	10	PG	10		P	N	N	0		-6.8507	1.2003	-2.263
2	O1G	20	O1G	20		O	N	N	0		-6.872	1.7402	1.1401
3	O2G	30	O2G	30		O	N	N	0		-7.9919	2.1234	-1.0361
4	O3G	40	O3G	40		O	N	N	0		-7.4212	-3.027	-1.1391
5	PB	50	PB	50	R	P	N	N	0		-4.4462	.2559	-.13
6	O1B	60	O1B	60		O	N	N	0		-4.3045	.8104	1.2349
7	O2B	70	O2B	70		O	N	N	0		-5.057	-1.2312	-.0444
8	O3B	80	O3B	80		O	N	N	0		-5.4334	1.1929	-.9901
9	PA	90	PA	90	R	P	N	N	0		-2.0713	-.7457	.068
10	O1A	100	O1A	100		O	N	N	0		-2.6694	-2.0974	.1437
11	O2A	110	O2A	110		O	N	N	0		-1.957	-.1258	1.5495
12	O3A	120	O3A	120		O	N	N	0		-3.0024	.203	-.8404
13	O5*	130	O5*	130		O	N	N	0		-.604	-.844	-.5873
14	C5*	140	C5*	140		C	N	N	0		.1706	-1.6948	.2601
15	C4*	150	C4*	150	R	C	N	Y	0		1.5843	-1.8311	-.309
16	O4*	160	O4*	160		O	N	Y	0		2.2342	-.5422	-.3552
17	C3*	170	C3*	170	S	C	N	Y	0		2.4651	-2.6838	.6309
18	O3*	180	O3*	180		O	N	N	0		2.5342	-4.033	.1651
19	C2*	190	C2*	190	R	C	N	Y	0		3.8562	-2.0116	.5556
20	O2*	200	O2*	200		O	N	N	0		4.8272	-2.9264	.0433
21	C1*	210	C1*	210	R	C	N	Y	0		3.6478	-.8309	-.4185
22	N9	220	N9	220		N	N	Y	0		4.4258	.332	.0158
23	C8	230	C8	230		C	N	Y	0		4.0122	1.3023	.8793
24	N7	240	N7	240		N	N	Y	0		4.9554	2.1841	1.0422
25	C5	250	C5	250		C	N	Y	0		6.0333	1.8336	.3002
26	C6	260	C6	260		C	N	Y	0		7.3035	2.3917	.0776
27	N6	270	N6	270		N	N	N	0		7.6816	3.5648	.7069
28	N1	280	N1	280		N	N	Y	0		8.1354	1.7638	-.7472
29	C2	290	C2	290		C	N	Y	0		7.7832	.6441	-1.352
30	N3	300	N3	300		N	N	Y	0		6.6026	.0883	-1.1783
31	C4	310	C4	310		C	N	Y	0		5.7043	.6441	-.3719
32	HOG 2	320	2HOG	320		H	N	N	0		-8.7254	2.1005	-.5463
33	HOG 3	330	3HOG	330		H	N	N	0		-7.5229	-.6163	-1.0482
34	HOB 2	340	2HOB	340		H	N	N	0		-5.1329	-1.5549	-.9525
35	HOA 2	350	2HOA	350		H	N	N	0		-1.5636	.7525	1.4553
36	H5* 1	360	1H5*	360		H	N	N	0		-.2964	-2.6783	.3124
37	H5* 2	370	2H5*	370		H	N	N	0		.2215	-1.2631	1.2597
38	H4*	380	H4*	380		H	N	N	0		1.5508	-2.275	-1.304
39	H3*	390	H3*	390		H	N	N	0		2.0787	-2.6519	1.6496
40	HO3	400	*HO3	400		H	N	N	0		3.0945	-4.5155	.7882
41	H2*	410	H2*	410		H	N	N	0		4.1576	-1.6464	1.5373
42	HO2	420	*HO2	420		H	N	N	0		4.8672	-3.6676	.6629
43	H1*	430	H1*	430		H	N	N	0		3.9312	-1.1195	-1.4307
44	HR	440	HR	440		H	N	N	0		3.0443	1.3348	1.3574

Figure 14.3.3 MSDchem ligand data at the atomic level that can be accessed from a ligand details page.

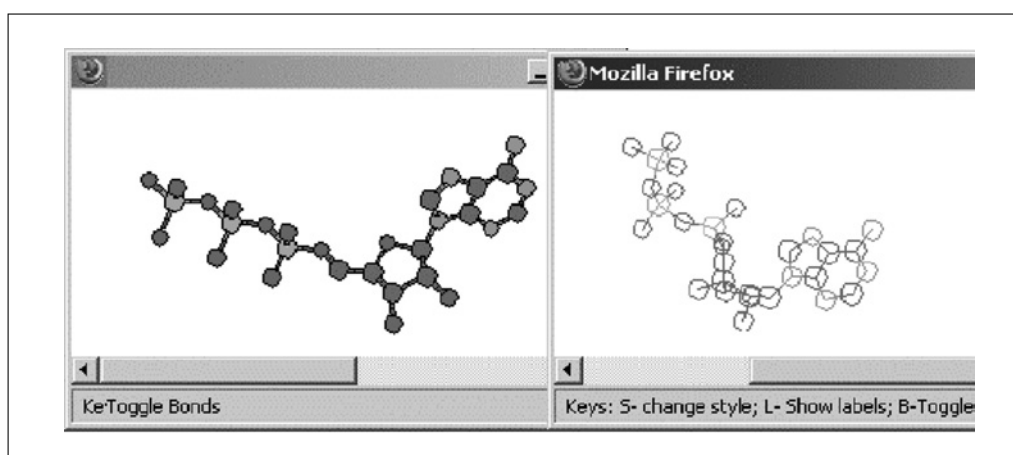


Figure 14.3.4 Three-dimensional visualizations of a ligand using the Jmol applet for idealized versus representative coordinates from MSDchem.

Using MSDchem to Search the PDB Ligand Dictionary

14.3.6

Other choices include PDB, crystallographic mmCif, CML (Chemical Mark-up Language), and XYZ file formats. The reason that MSDChem offers all these alternative export formats is that the most common ones are not designed to store every important piece of information. PDB ligand export files can be easily incorporated and used in parallel with files from the actual PDB archive but are missing placeholders for important chemical properties like bond orders. SDF/MDL format, on the other hand, is one of the most popular formats in chemoinformatics in that it is able to store the definitive chemical properties. However, it has no place for PDB atom labels, which are used to provide direct literature references for ligands on the atom level. Crystallographic mmCif is the format of the wwPDB exchange designed to solve the problems of incomplete ligand representation but is not been widely used by chemical and visualization software. CML is an XML-based format ideal for programmatic use and slowly gaining in popularity, while XYZ is a very primitive format supported by various general purpose 3-D visualization packages.

10. Select the HTML option from the Output drop down menu.

Other choices allow a user the options to either download the file on the hard disk to later open using a text editor or another program or have a look at the contents of the file directly on a separate browser window.

11. From the Library drop down menu choose either ideal (to use idealized) or PDB (to use representative coordinates). The Hydrogens checkbox specifies whether to include hydrogen atoms in the exported files; uncheck this option in order to exclude hydrogens.

Files that include hydrogen atoms provide a more complete data set, but excluding the hydrogen atoms may often simplify visualization and processing of the really significant chemical structure of the heavier atoms. If hydrogen atoms are required, use the pdb+H option from the Library menu to get representative PDB heavy atom coordinates together with CACTVS (Ihlenfeldt et al., 1992) idealized hydrogen coordinates. This is usually a good idea since hydrogen coordinates are often missing from the PDB, and the hydrogen atoms will have null (zero) coordinates in the exported file. Most export formats do not distinguish between an atom in the three-dimensional point (0,0,0) from an unobserved atom, with unpredictable results.

12. Click on the Save button. The pop-up window shown in Figure 14.3.5 will appear.

13. Access the PDB entries that include the ligand, and their binding site information.

On the bottom of the left menu area of the ligand details page, there are links that redirect the user to the list of PDB entries containing the ligand from MSDlite (Golovin et al.,

The screenshot shows a web browser window displaying the export options for a ligand (ATP) and the resulting SDF/MDL file content. The browser window is titled "Mozilla Firefox" and shows the file "ATP.sdf" with the following content:

```

ATP.sdf
-ISIS-          3D
47 49 0 0 0 0 0 0 0 0 0 1 V2000
46.1070 45.1820 56.9500 P 0 0 0 0 0
45.7790 46.3300 56.0520 O 0 0 0 0 0
47.3820 44.4970 56.6260 O 0 0 0 0 0
45.9720 45.5300 58.3750 O 0 0 0 0 0
43.9110 43.7400 55.6550 P 0 0 0 0 0
42.9750 42.7220 55.9860 O 0 0 0 0 0
43.6030 44.7670 54.6780 O 0 0 0 0 0
45.0410 44.0150 56.7380 O 0 0 0 0 0
  
```

The left sidebar shows the export options:

- Contents: Complete
- Output: html
- Format: sdf
- Library: pdb+H
- Viewers: jmol
- Hydrogens:
- Buttons: Save, View

**Figure 14.3.5** Exporting ligand data with representative heavy-atom and idealized hydrogen coordinates in the SDF/MDL chemical file format using MSDchem.

2004), as well as to details about its binding sites from MSDsite (Golovin et al., 2005). For example, following the link for binding statistics labeled “As a Ligand” produces a chart with the relative frequencies of interactions with various amino acids. The next link below, labeled “As ligand environment,” is useful for standard and modified amino acids that can be part of a bound molecule’s environment.

## **SEARCHING FOR LIGANDS USING A FORMULA OR FRAGMENT EXPRESSION**

Remembering ligand three-letter codes and avoiding mistakes in spelling molecular names are not easy. Additionally, it may be desirable to perform searches that do not target a single ligand, but rather a class of ligands that share some common chemical characteristics (e.g., atoms of particular elements or common chemical groups). A simple way of performing this type of search is by using a formula range expression or a pharmacophore fragment expression. Following the steps of this protocol will facilitate converting these elements into a formula or fragment-based search option that may significantly reduce the number of candidate ligands that have to be inspected.

### ***Necessary Resources***

#### *Hardware*

Computer with Internet access

#### *Software*

An up-to-date Internet browser, such as Internet Explorer 3.0 or later (<http://www.microsoft.com/ie>); Netscape 4.75 or later (<http://browser.netscape.com>); Firefox 1.0 or later (<http://www.mozilla.org/firefox>)

### ***Use the formula expression editor***

1. Open the MSDchem search home page (<http://www.ebi.ac.uk/msd-srv/msdchem>; Fig. 14.3.1)
2. Enter a formula range expression (a space-separated list of chemical elements followed by a value or a range for the number of times this element is allowed in the ligand formula).

*The syntax for a formula range expression is as follows.*

*[<Element> | <Element><Value> | <Element> <Minimum>-<Maximum>]*

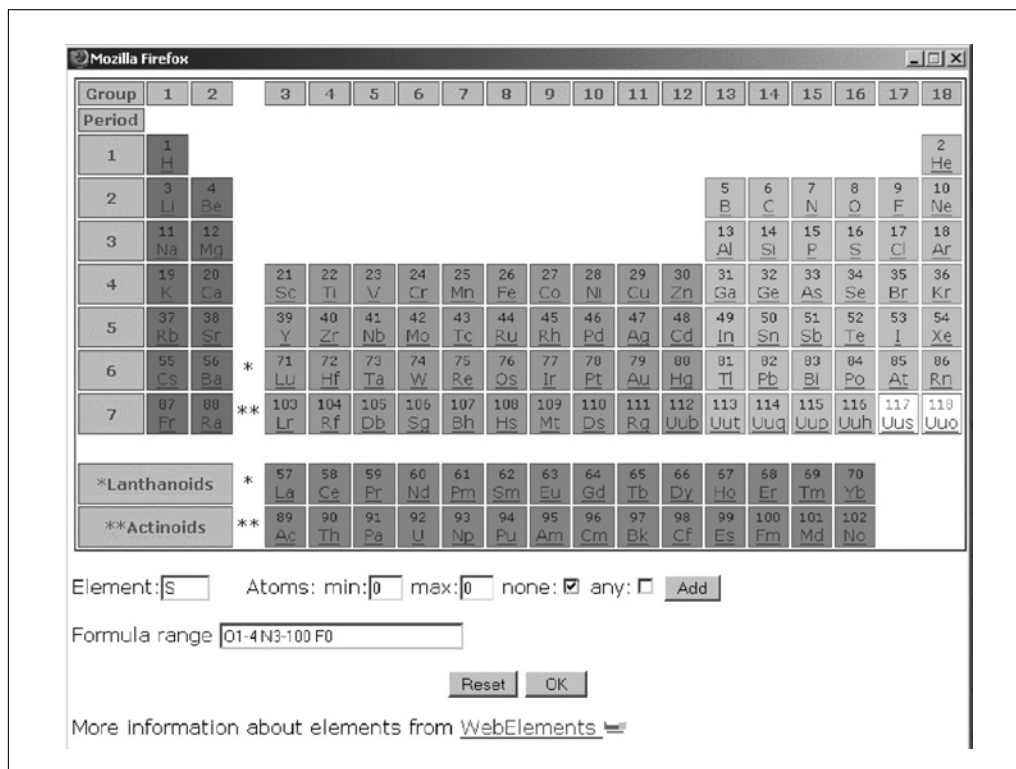
*A range has to be given in the form of “minimum value” - “maximum value” (separated with the “-” character). Elements given without a specified value or a range are equivalent to <Element>1 while <Element>0 means that the particular element is not allowed at all. Elements not given in the expression at all may or may not be part of the ligand formula.*

*For example: C3-6 N2 F (three to six carbon atoms, exactly two nitrogen atoms, a fluorine, and anything else); O1-4 N3-100 Cl1-100 F0 S0 (no more than four oxygen atoms, at least three nitrogen atoms, at least one chlorine, no fluorine or sulfur, and anything else).*

3. Alternatively, click on the “edit” button on the same line as the Formula text field to bring up the formula expression editor window shown in Figure 14.3.6.

*Using the formula expression editor is optional, but it is a fast and easy way to build an expression in an interactive way, without having to worry that the formula range expression being queried is incorrect.*

4. If using the formula expression editor window, leave the default “formula range” search operator selected and specify, for example, ligands with one to four oxygen



**Figure 14.3.6** The formula expression editor screen used in an example to obtain MSDchem ligands with one to four oxygen atoms, at least three nitrogen atoms, no fluorine, and no sulfur.

atoms, more than three nitrogen atoms, and no fluorine or sulfur, by performing the following steps:

- Click on O (oxygen) and type in 1 and 4 as the min and max values. Click on Add.
- Click on N (nitrogen) and specify 3 as the min value. Click on Add.
- Click on F (fluorine) and tick the “none” checkbox. Click on Add.
- Repeat the same step for S (sulfur).
- When finished, click on OK to transfer the expression in the search page.

*In order to add a constraint for a new element on the formula, first click on the corresponding element, choose either the “any” or “none” check box, or alternatively provide input values into the min and max fields and click on the Add button to append the new constraint in the formula expression. Just clicking on an element name and immediately on the OK button will generate the expression equivalent to “any number of atoms of this element” (the value of 100 is realistically a number for no actual upper limit).*

#### **Use the fragment expression editor**

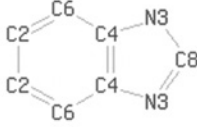
- Go to the MSDchem search home page (<http://www.ebi.ac.uk/msd-srv/msdchem>). Click on the “edit” button that is on the same line as the Fragments text field to bring up the fragment expression editor window shown in Figure 14.3.7. Enter a fragment name, followed by a value or a range, for the number of times this fragment is allowed in the ligand formula, as in the following formal description:

[<Fragment> | <Fragment> <Value> | <Fragment> <Min> - <Max>]

Mozilla Firefox

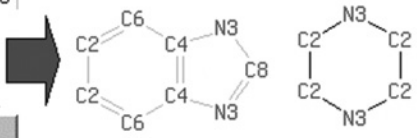
## Select chemical fragment pattern

acetylurea	acridine	acridone			
actinophenoxazine	adenine	alkaloid			
barbit	barbiturates	barbiturgroup			
<u>benzimidazole</u>	benzodiazepine	benzofuran			
benzoisoquinoline	benzothiadiazide	benzothiazole			
benzothiophen	benzoxazole	bilirubin			
biotin	carbazole	cephalosporin			
chromen	cinnoline	coumarine			
cyclobutane	cyclohexane	cyclopentane			
cyclopropane	cytosine	deoxyribose			
dibenzofuran	dibenzothiophen	dithiolane			
flavin	furan	furanose			
glycerophos	guanine	imidazole			
indole	inosine	isoquinoline			
isoxadiazole	isoxazole	naphtyridine			
napthalene	oxadiazole	oxazole			
oxazolinedione	oxepin	peptide			
penicillin	phenanthrene	phenanthridine			
phenanthroline	phenazine	phenothiazine	phenyl	phthalazine	piperazine
porphin	prostto	pteridine	pteroyl	purine	pyran
pyranose	pyrazine	pyrazole	pyridazine	pyridine	pyrimidine
pyrole	quinazoline	quinoline	quinoxaline	rauwolfia	ribose
steroid	succinimide	thiadiazole	thiazole	thiepin	thiophen
tolcol	vitaminAcore	xanthen			



Fragment:  min:  max:  none:  any:

Fragment expression



**Figure 14.3.7** The fragment expression editor screen used in an example to obtain MSDchem ligands with two or more benzimidazole and without any piperazine groups.

The benzimidazole and piperazine groups used are shown Figure 14.3.7. There are fragments for about 90 common functional groups that are chosen to be large and characteristic enough to locate real pharmacophores. The selection, which is based on published literature, is expected to be revised in the future.

Using the fragment editor is convenient because it has a context sensitive list of various predefined fragments. Whenever the mouse cursor is moved above a group name, the corresponding fragment is displayed on the top right area of the editor.

Fragments and their display images often include wildcard "green colored - X" types for atom elements and wildcard "green colored - any" orders for bonds. Aromatic bonds are displayed with gray color. In order to add a constraint for a new fragment click on the corresponding group name, choose either the "any" or "none" check box. Alternatively, provide input values into the "min" - "max" field (for the number of times the fragment should appear in the molecule).

6. For example, use the fragment expression editor in building an expression for ligands that have at least two benzimidazole and no piperazine groups by performing the following steps:
  - a. Click on the benzimidazole group and specify 2 in the min text field.
  - b. Click on the Add button to specify at least two benzimidazole groups.

- c. Click on the piperazine group and tick the "none" check box.
  - d. Click on the Add button and then (when finished with all fragments) click on the OK button. The search home page shown in Figure 14.3.1 will appear, with the search fields filled out according to the selection made in the previous step.
7. Click on the Search button on the search page to get the list of PDB ligands with one to four oxygen atoms, at least three nitrogen atoms, no fluorine or sulfur, at least two benzimidazole groups, and no piperazine groups to obtain the view shown in Figure 14.3.8.

MSD Ligand Chemistry		<a href="#">Get PDB entries</a>	<a href="#">Get PDB sites</a>
8 results			
RecordCode	Molecule name	Stereo smile	Formula
1	<a href="#">BAH</a> BIS(5-AMIDINO-2-BENZIMIDAZOLYL)METHANE KETONE HYDRATE		C17 H18 N8 O2
2	<a href="#">BAK</a> BIS(5-AMIDINO-2-BENZIMIDAZOLYL)METHANE KETONE		C17 H16 N8 O1
3	<a href="#">BAO</a> BIS(5-AMIDINO-2-BENZIMIDAZOLYL)METHANONE		C17 H14 N8 O1
4	<a href="#">BOZ</a> BIS(5-AMIDINO-2-BENZIMIDAZOLYL)METHANONE ZINC		C17 H18 N8 O1 ZN1
5	<a href="#">E96</a> 4-[(4-HYDROXY-PHENYL)-1H-BENZIMIDAZOLE-5-YL]-BENZIMIDAZOLE-2-YL-[4-HYDROXY-BENZENE]		C26 H18 N4 O2
6	<a href="#">E97</a> [3-(4-(2-(4-(3-DIMETHYLAMINO-PROPOXY)-PHENYL)-3H,3H-[5,5']BIBENZOIMIDAZOLYL-2-YL)-PHENOXY)-PROPYL)-DIMETHYL-AMINE]		C36 H40 N6 O2
7	<a href="#">IBB</a> 5-(2-IMIDAZOLINYL)-2-[2-(4-HYDROXYPHENYL)-5-BENZIMIDAZOLYL]BENZIMIDAZOLE		C23 H20 N6 O1
8	<a href="#">TBZ</a> 2-(4-METHOXYPHENYL)-5-(3-AMINO-1-PYRROLIDINYL)-2,5',2',5"-TRIS-BENZIMIDAZOLE		C32 H29 N6 O1

**Figure 14.3.8** MSDchem ligands that satisfy particular formula range and fragment expression constraints.

Found 16 hits (1 pages). Showing hits 1 to 16.		
<a href="#">109d</a> <a href="#">View</a>	STRUCTURE OF A BIS-BENZIMIDAZOLE DRUG BOUND TO THE DNA DUPLEX C-G-C-G-A-A-T-T-C-G-C-G	2.00Å
<a href="#">1c1w</a> <a href="#">View</a>	RECRUITING ZINC TO MEDIATE POTENT, SPECIFIC INHIBITION OF SERINE PROTEASES	1.90Å
<a href="#">1c2d</a> <a href="#">View</a>	RECRUITING ZINC TO MEDIATE POTENT, SPECIFIC INHIBITION OF SERINE PROTEASES	1.65Å
<a href="#">1c2e</a> <a href="#">View</a>	RECRUITING ZINC TO MEDIATE POTENT, SPECIFIC INHIBITION OF SERINE PROTEASES	1.65Å
<a href="#">1c2f</a> <a href="#">View</a>	RECRUITING ZINC TO MEDIATE POTENT, SPECIFIC INHIBITION OF SERINE PROTEASES	1.70Å
<a href="#">1c2g</a> <a href="#">View</a>	RECRUITING ZINC TO MEDIATE POTENT, SPECIFIC INHIBITION OF SERINE PROTEASES	1.65Å
<a href="#">1c2h</a> <a href="#">View</a>	RECRUITING ZINC TO MEDIATE POTENT, SPECIFIC INHIBITION OF SERINE PROTEASES	1.40Å
<a href="#">1c2i</a> <a href="#">View</a>	RECRUITING ZINC TO MEDIATE POTENT, SPECIFIC INHIBITION OF SERINE PROTEASES	1.47Å
<a href="#">1c2j</a> <a href="#">View</a>	RECRUITING ZINC TO MEDIATE POTENT, SPECIFIC INHIBITION OF SERINE PROTEASES	1.40Å
<a href="#">1ftd</a> <a href="#">View</a>	5'-D(CP*GP*CP*GP*AP*AP*TP*TP*CP*GP*CP*G)-3'-SYMMETRIC BIS-BENZIMIDAZOLE COMPLEX	2.00Å
<a href="#">1xuh</a> <a href="#">View</a>	TRYPsin-KETO-BABIM-CO+2, PH 8.2	2.37Å
<a href="#">1xuj</a> <a href="#">View</a>	TRYPsin-KETO-BABIM, ZN+2-FREE, PH 8.2	1.50Å
<a href="#">1xuj</a> <a href="#">View</a>	TRYPsin-KETO-BABIM-ZN+2, PH 8.2	1.92Å
<a href="#">263d</a> <a href="#">View</a>	ISCHELICITY AND PHASING IN DRUG-DNA SEQUENCE RECOGNITION: CRYSTAL STRUCTURE OF A TRIS(BENZIMIDAZOLE)-OLIGONUCLEOTIDE COMPLEX	2.20Å

**Figure 14.3.9** List of PDB entries referring to MSD atlas pages that include ligands that satisfy particular formula range and fragment expression constraints.

8. Follow links in the results list for details about each individual ligand using the same steps explained in Basic Protocol 1.
9. Obtain a list of PDB entries from the links on the top of the page for information on where these ligands can be found or about their binding site details (e.g., see Fig. 14.3.9).
10. Follow the four-character PDB file name links in the list of PDB entries to the atlas pages that provide summary information about the PDB entries, or follow the “view” links to activate a protein 3-D visualization applet for the whole PDB entry.

### **PERFORMING A CHEMICAL SUBGRAPH SEARCH**

Formula and chemical fragment searching is appropriate in cases where little is known about the ligand chemical structure. Often one may not remember the three-letter PDB code or chemical name of a ligand but may still easily draw up the diagram of a significant part of its chemical structure. In cases where the connectivity diagram for a reasonable fraction of the molecule is known or there is a ligand that is quite similar in terms of chemical structure, the steps in this protocol may be used to search for a chemical subgraph, i.e., a subset of the atoms and bond of the target ligand drawn up in a chemical diagram editor. This procedure will return a restricted list of more accurate candidate molecules that include the input chemical structure, and it can be also used to look for ligand variants.

A convenient and popular way to encode a chemical structure into a text string is by using SMILE strings that are equivalent to chemical formulas but also incorporate the atom connectivity and chemical properties. A nonstereo SMILE will encode all the fundamental information found on a chemical diagram but with unspecified atom chirality. This information is about atom elements, formal charges and connectivity, as well as bond orders. Nonstereo SMILES will not be able to distinguish between two different stereo-isomers, while stereo SMILES, which also encode stereo descriptors of atoms and bonds, will. It is usually a lot more convenient to use the nonstereo SMILES as the search criteria and then to visually inspect all the stereoisomers. Reading and writing SMILE strings is a rather difficult exercise for larger molecules and this where a molecular editor will definitely help.

#### ***Necessary Resources***

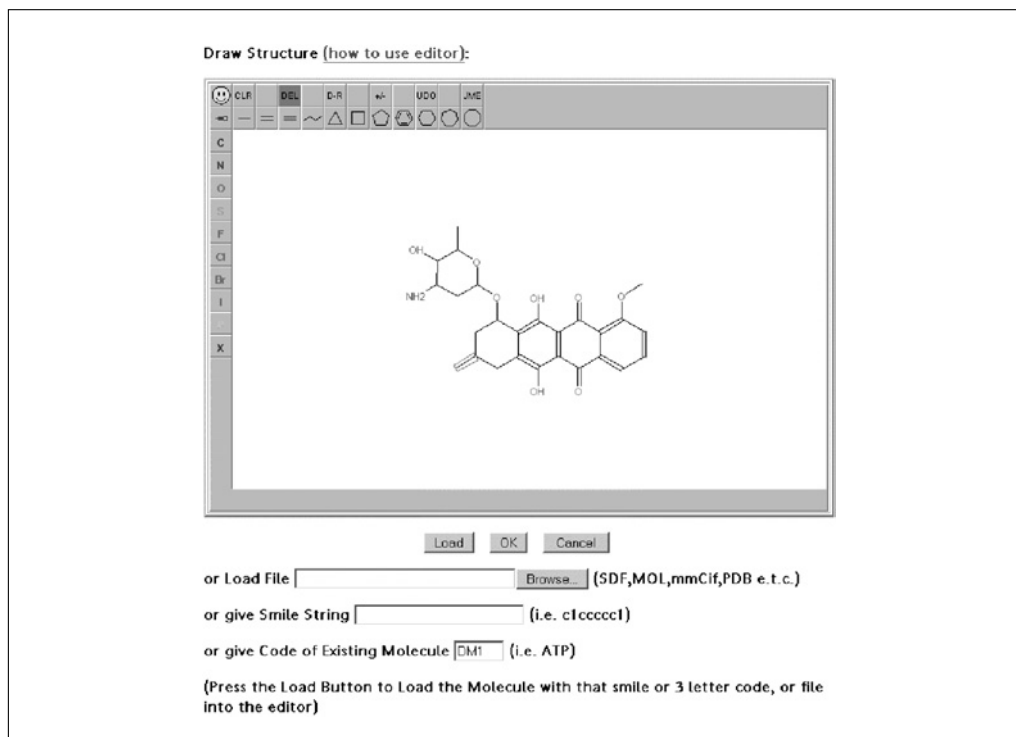
##### *Hardware*

Computer with Internet access

##### *Software*

An up-to-date Internet browser, such as Internet Explorer 3.0 or later (<http://www.microsoft.com/ie>); Netscape 4.75 or later (<http://browser.netscape.com>); Firefox 1.0 or later (<http://www.mozilla.org/firefox>)

1. Open the MSDchem search home page (<http://www.ebi.ac.uk/msd-srv/msdchem>; Fig. 14.3.1).
2. Open the JME molecular editor by clicking on the “edit” button on the same line as the “Non stereo smile” text field. Sketch the ligand structure or modify the ligand structure of a similar known ligand (Fig. 14.3.10).



**Figure 14.3.10** Screen used for loading the molecular structure diagram of DM1 on JME editor and modifying it by removing its noncharacteristic atoms or groups in order to prepare a subgraph search criteria that will match molecules with the same main structure as DM1.

- For example, to load the chemical diagram of daunomycin, type the three-letter code (DM1 in this case), a chemical file name (SDF/MOL, mmCif, PDB), or a SMILE string into the appropriate field on the molecular editor page and click Load. The molecular diagram of DM1 will appear on the JME editor.

*The JME molecular editor is a Java applet embedded in a Web page that offers the functionality of drawing a chemical diagram. It has controls to add bonds and small groups (common rings) on the molecule, modify existing atom elements and bond orders, and remove bonds together with their atoms. In order to use it, first sketch the connectivity diagram of the desired substructure and then finalize it by modifying noncarbon heavy elements and bond orders. Use of hydrogen atoms is not recommended, and the editor will not allow input of a disconnected chemical graph.*

- Click on the delete (DEL) button of JME, and then click on all bonds of the OH and C-C=O groups linked to the C12 ring atom on the left bottom of the structure to remove the hydroxyl and methylCO nonring groups linked to atom C12.
- Click on the OK button to transfer the SMILE string of this chemical subgraph on the search page.

*Alternatively, skip using the molecular editor altogether by directly inputting the SMILE string or a three-letter code that will be used as a subgraph, assuming that these groups are not characteristic for the substructure being searched in this example. The result will be a more generalized chemical subgraph of DM1.*

- Leave the default “has substructure” search operator selected.
- Click on Search to get the list of molecules that include the substructure. The partial result is shown in Figure 14.3.11. There are ten ligands that match the specified subgraph search criteria. The images in the results list indicate that they are very similar

## Molecule

10 results

RecordCode	3 letter code	Extended Code	Molecule name	Stereo smile	Formula
1	<a href="#">BDA</a>	<a href="#">BDA</a>	4- METHYLBENZYL- N- BIS[DAUNOMYCIN]		C62 H66 N2 O20
2	<a href="#">CMD</a>	<a href="#">CMD</a>	3'- DESAMINO- 3'- (3- CYANO- 4- MORPHOLINYL)- DOXORUBICIN		C32 H34 N2 O12
3	<a href="#">DM1</a>	<a href="#">DM1</a>	DAUNOMYCIN		C27 H29 N1 O10
4	<a href="#">DM2</a>	<a href="#">DM2</a>	DOXORUBICIN		C27 H29 N1 O11
5	<a href="#">DM6</a>	<a href="#">DM6</a>	4'- EPIDOXORUBICIN		C27 H30 N1 O11
6	<a href="#">DM8</a>	<a href="#">DM8</a>	2'- BROMO- 4'- EPIDAUNORUBICIN		C27 H28 N1 O10 BR1
7	<a href="#">DMM</a>	<a href="#">DMM</a>	3'- DESAMINO- 3'- (2- METHOXY- 4- MORPHOLINYL)- DOXORUBICIN		C32 H37 N1 O13
8	<a href="#">ERT</a>	<a href="#">ERT</a>	METHYL (4R)- 2- ETHYL- 2,5,12- TRIHYDROXY- 7- METHOXY- 6,11- DIOXO- 4- ([2,3,6- TRIDEOXY- 3- (DIMETHYLAMINO)- BETA- D- RIBO- HEXOPYRANOSYL]OXY)- 1H,2H,3H,4H,6H,11H- TETRACENE- 1- CARBOXYLATE		C31 H37 N1 O11

**Figure 14.3.11** Eight of the ten daunomycin-like ligands that contain the reduced chemical graph of DM1 as a subgraph, retrieved using the MSDchem “has substructure” search functionality.

molecules, and even the molecular names for most of them suggest that they fall in the class of daunomycin/doxorubicin variants.

*In the case of ERT, the molecule name does not resemble anything, and this is an obvious example where only a subgraph search will identify the similarity of this ligand with daunomycins and doxorubicins while a name based one would not.*

*Subgraph searching is nontrivial problem, and this means that often the results are not instantaneous. The user is warned by a pop-up window that the search may require a couple of minutes to finish, but in practice, in the majority of cases, the results are available a lot faster.*

home > searches > MSDsite > search result Hit list: [this page](#) [about help](#)

Items: 1 - 20 of 40 Next > | Last >> Hetero: BDA|CMD|DM1|DM2|DM6|DM8|DMM|ERT|MAR|NOD

amend search | get: [sxml](#), [pdb](#), [of](#)

Press a button to sort by the column, press it again to change the order

Type	ID	Classification	Released	Resolution	R-Factor	Hetero list	<a href="#">AstexViewer™@MSD-EBI</a> <a href="#">- RasMol script viewer</a> <a href="#">- Original PDB</a>
X-ray	427d	deoxyribonucleic acid	2000-01-24	1.1		DM1	<a href="#">Atlas</a> <a href="#">PDBsum</a> <a href="#">SF</a>
X-ray	1x0k	deoxyribonucleic acid	2003-06-10	1.17		DM1 NA	<a href="#">Atlas</a> <a href="#">PDBsum</a> <a href="#">SF</a>
X-ray	1e11	deoxyribonucleic acid	1993-01-15	1.2		DM1 NA	<a href="#">Atlas</a> <a href="#">PDBsum</a> <a href="#">SF</a>
X-ray	1e68	deoxyribonucleic acid	2005-02-22	1.2		MAR	<a href="#">Atlas</a> <a href="#">PDBsum</a> <a href="#">SF</a>
X-ray	1d35	deoxyribonucleic acid	1995-01-15	1.3		CH2 MAR MG NH2	<a href="#">Atlas</a> <a href="#">PDBsum</a>
X-ray	1e20	deoxyribonucleic acid	2003-05-13	1.34	0.193	DM2 TL	<a href="#">Atlas</a> <a href="#">PDBsum</a>
X-ray	152d	deoxyribonucleic acid	1994-05-31	1.4		DM1	<a href="#">Atlas</a> <a href="#">PDBsum</a>
X-ray	1d54	deoxyribonucleic acid	1995-01-15	1.4		DM6	<a href="#">Atlas</a> <a href="#">PDBsum</a> <a href="#">SF</a>
X-ray	2d34	deoxyribonucleic acid	1992-04-15	1.4		CH2 DM1 MG	<a href="#">Atlas</a> <a href="#">PDBsum</a>
X-ray	151d	deoxyribonucleic acid	1994-05-31	1.4		DM2	<a href="#">Atlas</a> <a href="#">PDBsum</a>
X-ray	308d	deoxyribonucleic acid	1997-02-12	1.5	0.175	DM1 GLC M05 M06	<a href="#">Atlas</a> <a href="#">PDBsum</a> <a href="#">SF</a>
X-ray	1ie2	dna, rna	2003-07-29	1.5	0.198	DM1	<a href="#">Atlas</a> <a href="#">PDBsum</a> <a href="#">SF</a>
X-ray	2des	deoxyribonucleic acid	1995-05-15	1.5		DMM MG NA	<a href="#">Atlas</a> <a href="#">PDBsum</a> <a href="#">SF</a>
X-ray	1d10	deoxyribonucleic acid	1992-10-15	1.5		DM1 NA SPM	<a href="#">Atlas</a> <a href="#">PDBsum</a>
X-ray	1d15	deoxyribonucleic acid	1992-07-15	1.5		DM6 SPM	<a href="#">Atlas</a> <a href="#">PDBsum</a>
X-ray	1d33	deoxyribonucleic acid	1992-04-15	1.5		CH2 DM1 MG	<a href="#">Atlas</a> <a href="#">PDBsum</a>
X-ray	1d36	deoxyribonucleic acid	1995-05-15	1.5		MAR MG	<a href="#">Atlas</a> <a href="#">PDBsum</a>
X-ray	1da0	deoxyribonucleic acid	1993-07-15	1.5		DM1	<a href="#">Atlas</a> <a href="#">PDBsum</a>
X-ray	1da4	deoxyribonucleic acid	1999-09-15	1.6	0.205	DM2	<a href="#">Atlas</a> <a href="#">PDBsum</a> <a href="#">SF</a>
X-ray	395d	deoxyribonucleic acid	1999-01-20	1.6		NOD	<a href="#">Atlas</a> <a href="#">PDBsum</a> <a href="#">SF</a>

**Figure 14.3.12** The 40 PDB entries that include the 10 daunomycin-like ligands and access to their binding site details from MSDsite.

- Click on the Get PDB entries URL; there are 40 PDB entries that include these ligands.

*By browsing through, one can also identify similarities in the biological function.*

- At the top right of the results screen, click on the Get PDB sites link to view details about the binding sites of these ten ligands in PDB entries. The view shown in Figure 14.3.12 (MSDsite search result page) appears. Follow a link for each PDB entry to visualize interactions of the ligands with the macromolecule, as well as further details and statistics regarding strength and distance of the interactions.

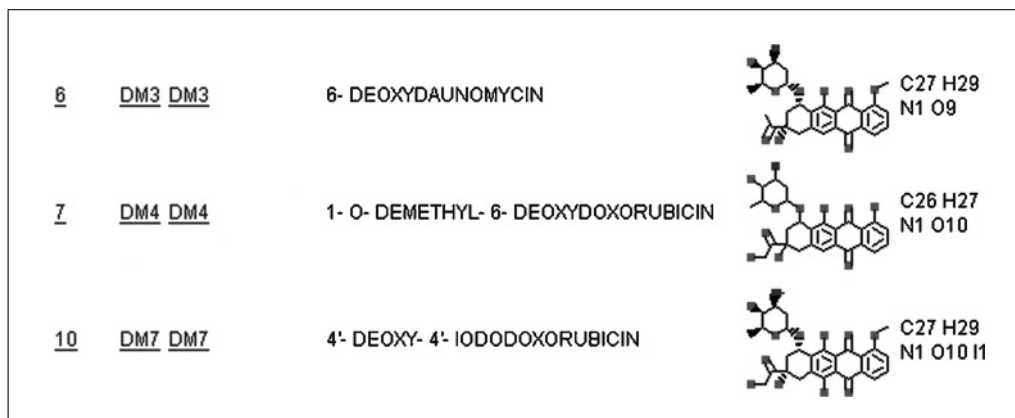
*There are more search options based on molecule graphs. The drop down menu for search operators next to the “Non stereo smile” label allow options for “exact structure” and “is substructure of” a particular structure. The exact structure search is instantaneous and will find all ligands that are stereoisomers of the input, using the “Non stereo smile” data field. If using the “Stereo smile” data field, which is just below, and the same “exact structure” search operator, one can search for particular stereoisomers but will need to input the correct stereo configuration in JME. The “Non stereo smile” “substructures of” search operator will return ligands that are included as subgraphs in the chemical graph provided as input.*

### Search using fingerprint similarity

- Open the MSDchem search home page (<http://www.ebi.ac.uk/msd-srv/msdchem>; Fig. 14.3.1).
- Type DM1 directly into the “fingerprint” text field.

*For known ligands there is no need to use the JME editor in order to draw up their chemical structure. Giving a three-letter PDB code directly instead of a SMILE string is enough for MSDchem to automatically use its chemical structure as the search criteria.*

*The Fingerprint search field uses fingerprints prepared using the existence of each one of 500 segments in the predefined library of the CACTVS system (Ihlenfeldt et al., 1992). This search option will give useful results mainly for big input molecules where the resulting hits will have almost the same segment groups (at least 99% common groups).*



**Figure 14.3.13** Three more hits for DM1 daunomycin-like ligands revealed using the MSDchem fingerprint similarity searching.

12. Click on the Search button.

*Giving the reduced DM1 as input will also return molecules that do not include the complete input structure but are still quite close in structure. For example, three more daunomycin/doxorubicin variants are found in this search (shown in Fig. 14.3.13) that were missed using the subgraph search. In this case the fingerprint similarity search proved quite useful, although one must take into account that this option is in general more unpredictable than subgraph searching.*

## BASIC PROTOCOL 4

### EXPORTING THE LIGAND DICTIONARY

Searching and downloading data for individual ligands is sufficient in most cases, but there are still times when it may be useful to have the complete ligand dictionary as a local resource for convenience or systematic use. The volume of data in the database is manageable and the MSDchem service offers several options for downloading it in various formats.

#### *Necessary Resources*

##### *Hardware*

Computer with Internet access

##### *Software*

An up-to-date Internet browser, such as Internet Explorer 3.0 or later (<http://www.microsoft.com/ie>); Netscape 4.75 or later (<http://browser.netscape.com>); Firefox 1.0 or later (<http://www.mozilla.org/firefox>)

A compression/decompression utility that can handle gzip-compressed tar files (WinZip for Windows, <http://www.winzip.com>; gzip, <http://www.gnu.org/software/gzip/gzip.html>, and tar, <http://www.gnu.org/software/tar>, for Linux and other Unix systems)

#### *View the ligand index pages*

1. Open the MSDchem search home page (<http://www.ebi.ac.uk/msd-srv/msdchem>). Click on link to “ligand index and download” at lower left hand side to open an MSDchem ligand index page.

*Ligand index pages provide direct links and data about ligands organized on a very simple layout. Ligands are listed numerically (0 to 9) and alphabetically (A to Z) according to the first character of their three-letter code. There are links to all of the index pages on the top of each index page for easy navigation. Each ligand is presented visually using a*

home > searches MSDchem MSD: Ligand Chemistry? about help

0 1 2 3 4 5 6 7 8 9 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Download zip file with the complete ligand collection (idealised coordinates & explicit hydrogen atoms) in:  
[SDF CML PDB mmCif](#)  
[MSDchem XML FTP area](#)

The screenshot displays a grid of 15 ligand entries, each consisting of a chemical structure, a three-letter code, and a list of file formats (sdf, cml, pdb, cif). The entries are:

- A**: Adenosine 5'-monophosphate
- A12**: Phosphomethylphosphonic acid adenosine ester
- A15**: 3'-3'-dichlorophenol-1,8-3H-benzo [d,e] isochromen-1-one
- A1A**: 6-amino hexanoic acid
- A1E**: Tetrahydroquinolin-2(1H)-one
- A1P**: 9-(2-deoxy-β-D-ribofuranosyl)-β-D-erythro-pentofuranosyl-β-D-purine-2-amine
- A23**: Adenosine 5'-phosphate 2'-3'-cyclic phosphate
- A24**: 6-nitro-5-(2-morpholin-4-yl-ethylamino)carbonyl-phenyl-β-D-galactopyranoside
- A26**: 2-cyano-3-hydroxy-N-(4-trifluoromethyl-phenyl)-butyramide
- A2E**: Tetrahydroquinolin-2(1H)-one
- A2G**: N-acetyl-2-deoxy-2-amino-β-D-galactose
- A2L**: 3'-O-methoxyethyl-adenosine 5'-monophosphate
- A2M**: 2'-methyl-adenosine 5'-monophosphate
- A2P**: Adenosine 2',5'-diphosphate
- A32**: Morpholin-4-yl-propylaminocarbonyl-phenyl-β-D-galactopyranoside

**Figure 14.3.14** The MSDchem ligand index page for the letter A with a list of the 525 ligands that have a three-letter code starting with the character A. There are links to access and download data for each one of them as well as for the whole ligand collection.

*small image of its chemical diagram, together with its common name and three-letter code for ligand identification. There are also links for the ligand details page and for chemical file format downloads. These pages can be useful for browsing ligands with interesting structure and for access to general purpose Web search engines that use components of ligand names.*

2. Click on the letter A to obtain the view shown in Figure 14.3.14.

### Download ligand files

3. Click on one of the four links at the top of each ligand page (Fig. 14.3.14) to download a gzip-compressed tar file for the complete ligand collection in a single compressed archive file. The file contains separate entries (~7000 as of this writing) for each ligand in one of the SDF/MDL, CML, PDB, or mmCif formats.
4. Alternatively, export an SDF/MDL, CML, PDB, or mmCif file for an individual ligand (e.g., DM1) by using the corresponding link in the ligand index page for letter D and clicking on the format link associated with DM1 (e.g., CML). The CML file for this entry will appear as shown in Figure 14.3.15.

*Alternatively, the file may be saved in a temporary area and opened later using the appropriate program. Run the utility and export the files on a local directory.*

### Download a single file with summary data for all ligands

5. Click on the Export button on the MSDchem search home page (<http://www.ebi.ac.uk/msd-srv/msdchem>) to get a list of all ligands, together with their common names and SMILE strings, in a single XML file. Choose the output format from the Retrieve drop-down menu to get the list as a Perl or JavaScript data structure for easier programmatic processing.

```

- <msd:entry dictRef="http://www.ebi.ac.uk/msd-srv/msdchem/ligand/reference.htm">
- <molecule id="msd1334" formalCharge="0">
  <identifier convention="msd:code3Letter">DM1</identifier>
  <identifier convention="msd:extendedCode">DM1</identifier>
  <identifier convention="msd:code1Letter"/>
  <name>DAUNOMYCIN</name>
- <formula convention="msd:stereoSmiles">
  COc1cccc2C(=O)c3c(O)c4C[C@](O)(C[C@H](O)[C@H]5C[C@H](N)[C@H](O)[C@H](C)O5)c4c(O)c
  </formula>
- <formula convention="msd:NonStereoSmiles">
  COc1cccc2C(=O)c3c(O)c4CC(O)(CC(OC5CC(N)C(O)C(C)O5)c4c(O)c3C(=O)c12)C(C)=O
  </formula>
  <formula convention="ElementCountWhitespace">C27 H29 N1 O10</formula>
  <scalar type="xsd:date" dictRef="msd:DefinedAt">1999-07-08</scalar>
  <scalar type="xsd:date" dictRef="msd:LastModifiedAt">1999-07-08</scalar>
- <name convention="msd:SystematicName">
  (1S,3S)-3-acetyl-3,5,12-trihydroxy-10-methoxy-6,11-dioxo-1,2,3,4,6,11-hexahydrotetracen-1-yl
  3-amino-2,3,6-trideoxy-a-L-lyxo-hexopyranoside
  </name>
  <scalar dictRef="msd:RCSBClassification">3 AND MORE RING SYSTEMS</scalar>
  <scalar dictRef="msd:PolymerTopology"/>
  <scalar dictRef="msd:PolymerCode"/>
  <scalar dictRef="msd:SupercededBy"/>
- <atomArray>
- <atom id="msd48374" elementType="C" hydrogenCount="1" formalCharge="0" x3="6.
  z3="-_3491">
  <label value="C1"/>
  </atom>
- <atom id="msd48375" elementType="C" hydrogenCount="1" formalCharge="0" x3="6.
  z3=".085">
  <label value="C2"/>
  </atom>

```

**Figure 14.3.15** The CML file for ligand DM1 as exported through the ligand index page for the letter D.

## COMMENTARY

### Background Information

#### *The PDB ligand dictionary*

The information available from the ligand dictionary is not part of the historical PDB archive. The PDB data bank files do not provide a clear chemical definition for ligands and amino/nucleic acids. Nevertheless, the PDB nomenclature based on three-letter codes and atom names clearly suggests that ligands with the same three-letter code should be the same chemical species and that atoms should superimpose within a stereochemical diagram. Of course, an explicit process to validate this rule has not always been in place, and the PDB archive has many errors propagated over the years. Furthermore, deriving the chemical identity of a ligand using a set of 3-D coordinates from a PDB entry is not a reliable operation, especially since there are inaccurate or unavailable experimental data in many cases.

The international body that manages the PDB is the Worldwide Protein Data Bank (wwPDB; Berman et al., 2005), with the mis-

sion of maintaining a single archive, freely and publicly available to the global community. wwPDB was founded by the Research Collaboratory for Structural Bioinformatics (RCSB PDB USA), the Macromolecular Structure Database group (MSD-EBI Europe) and the PDBj (Japan). All three organizations serve as deposition, data processing, and distribution sites for the PDB archive. Each site additionally provides its own view of the primary data with a variety of tools and resources for the global community.

There is an ongoing effort in wwPDB to address the problem of missing chemical definitions and to provide references and cleaned-up PDB data. At the time of curation of PDB entries, extensive work is done in carefully examining ligand chemistry issues and resolving them in cooperation with the depositors. The ligand dictionary (Westbrook et al., 2005), exchanged in the wwPDB (Chemical Component Information dictionary), that forms the basis of MSDchem, is a first step toward

achieving this goal. The wwPDB partners are making a systematic effort to finalize and resolve any remaining issues. This effort will ultimately be consolidated into a common dictionary. MSDchem incorporates the results of this effort, and at times it may provide data and corrections that are ahead of this common dictionary.

### ***Ligand stereochemistry in MSDchem ligand dictionary***

The MSDchem ligand dictionary includes explicit stereodescriptors using the absolute Chan-Ingold-Prelog notation as part of the definition of atoms and bonds in order to cope with the PDB implicit convention, i.e., different stereoisomers should be identified by different three-letter codes. Additionally, the MSDchem back-end system utilizes cheminformatics programs and libraries like the CACTVS software package (Ihlenfeldt et al., 1992) and the CORINA Web service (Gasteiger et al., 1990) in order to enrich the ligand dictionary with important chemical information and to validate and clean-up the data collection.

The MSDchem view is that PDB coordinates and atom names are not fundamental properties of a ligand, which is defined as a complete, distinct stereoisomer of a chemical compound. Representative coordinates are used as speculative chemistry, and they require manual curation since a set of coordinates may be compatible with different isomers. On the other hand, there may be conflicts with the depositor's view of ligand chemistry introduced as errors in experimental data or in the refinement process. Therefore, the unique chemical identity of a ligand in MSDchem is based on its stereo SMILE string in the CACTVS canonical unique form, including automatic detection of aromaticity and tautomerism.

UNIT 1.9 describes in detail the PDB archive data and the RCSB PDB Web tools as well as Ligand Depot, the RCSB's Web search system for small molecule information. Ligand Depot, among others, provides access to various small molecule sites and resources, one of which is MSDchem.

### ***Role of the MSDchem database***

The MSDchem database provides the framework for the contribution of the Macromolecular Structure Database group in the wwPDB ligand curation and clean-up effort and for the correct processing of new PDB ligands and entries. It is based on the wwPDB chemical component information dictionary

(Westbrook et al., 2005) and is an exchange resource; nevertheless, it also introduces several extensions like the use of explicit stereo-configuration descriptors as part of the ligand identity.

Additionally, MSDchem has identified several cases where the same chemical species has been defined more than once using different three-letter codes. In these cases, one unique three-letter identifier has been selected and the remaining codes have been marked as obsolete in the MSDchem database in order to stop their use in new PDB entries. These obsolete ligands will be highlighted in all MSDchem result pages, with direct links to the entries that supersede them.

### ***Topological variants***

The MSDchem database also includes topological variants of small molecules (typically standard and modified amino/nucleic acids) that form polymeric chains. Amino acids, for example, usually have four entries for the same small molecule in the dictionary, one each for the free, N-terminal, O-terminal, and linking variants. This is apparent from the value of the "Extended code" column that is included in MSDchem results and ligand details pages. There is no predefined formula for the extended code values, and the format depends on the type of polymerization. However, the following standard naming conventions are used: <three-letter code>\_LFOH for L-form of a free amino acid with OH group added; <three-letter code>\_LL for L-form of a linking amino acid residue; <three-letter code>\_LSN3 for L-form of a starting N-terminal (NH<sub>3</sub><sup>+</sup> group); <three-letter-code>\_LEO2 for L-form of an ending O-terminal (O<sub>2</sub><sup>-</sup> group).

The results from the protocols presented in the unit were retrieved from MSDchem release 28-2006.06.25. MSDchem is following the PDB weekly release cycle.

### **Critical Parameters and Troubleshooting**

Even though the production of MSDchem requires systematic checking and corrections, it still contains several known errors and inaccuracies. These errors may be due to missing or problematic experimental data (e.g., incomplete single set of fully observed coordinates for some ligands or inconsistencies between the experimental coordinates and the chemical description), bugs or inaccuracies in chemical software packages (core dumps or error messages during

stereo-identification and idealized coordinate generation), or incomplete manual curation (e.g., valence inconsistencies or missing hydrogen atoms).

It is important to understand that construction of the MSDchem database and back-end system is an ongoing effort that will gradually improve the quality of the data collection and merge it into wwPDB. In addition, MSDchem provides a reference definition for a ligand that is not required to be consistent with data in the PDB archive. The task of validating and correcting the PDB data with ligand definitions is a separate effort that is only loosely associated with the quality of the data in MSDchem.

Finally, the MSDchem Web service has an additional restriction to avoid server and user overload. There is a maximum limit of 300 hits in any search. On exceeding this limit, the first 300 results are displayed along with a clear warning on the result page that the search has to be further refined.

### Acknowledgements

The Macromolecular Structure Database (MSD) group is part of the European Bioinformatics Institute (EBI), which is one of the outstations of the European Molecular Biology Laboratories (EMBL) located in the Wellcome Trust Genome Campus at Hinxton Cambridge, UK.

Peter Keller, Sameer Velankar, Jawahar Swaminathan, John Ionides, Harry Boutselakis, Adel Golovin, and many other members of the MSD group have significantly contributed to MSDchem, together with all partners of wwPDB who are committed to the ligand dictionary exchange effort. MSD funding has been provided from the EU-Templor project, by the Wellcome trust and EMBL/EBI core support.

Finally, chemical software development projects like CACTVS and CORINA play a crucial role in providing supporting technology in the back-end of the MSDchem database.

### Literature Cited

Berman, H., Nakamura, H., and Henrick, K. 2005. The Protein Data Bank (PDB) and the World-Wide PDB <http://www.wwpdb.org>. In *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*. Section 4.6. (M. Dunn, L. Jorde, P. Little, and S. Subramaniam, eds.) <http://www.mrw.interscience.wiley.com/ggpb/articles/g406303/frame.html>. John Wiley & Sons, Hoboken, N.J.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F. Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. 1977. Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112:535-542.

Boutselakis, H., Dimitropoulos, D., Henrick, K., Ionides, J., John, M., Keller, P.A., McNeil, P., Pineda, J., and Suarez-Uruena, A. 2004. The European Bioinformatics Institute macromolecular structure relational database technology. In *Database Annotation in Molecular Biology*. pp. 223-240. John Wiley & Sons, Hoboken, N. J.

Gasteiger, J., Rudolph, C., and Sadowski, J. 1990. Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comp. Method.* 3:537-547.

Golovin, A., Oldfield, T.J., Tate, J.G., Velankar, S., Barton, G.J., Boutselakis, H., Dimitropoulos, D., Fillon, J., Hussain, A., Ionides, J.M.C., John, M., Keller, P.A., Krissinel, E., McNeil, P., Naim, A., Newman, R., Pajon, A., Pineda, J., Rachedi, A., Copeland, J., Sitnov, A., Sobhany, S., Suarez-Uruena, A., Swaminathan, J., Tagari, M., Tromm, S., Vranken, W., and Henrick, K. 2004. E-MSD: An integrated data resource for bioinformatics. *Nucl. Acids Res.* 32:D211-D216.

Golovin, A., Dimitropoulos, D., Oldfield, T., Rachedi, A., and Henrick, K. 2005. MSDsite: A database search and retrieval system for the analysis and viewing of bound ligands and active sites. *Proteins* 58:190-199.

Ihlenfeldt, W.D., Takahashi, Y., Abe, H., and Sasaki, S. 1992. CACTVS: A chemistry algorithm development environment. In *Daijuukagakutouronkai Dainijuukai Kouzoukaseisoukan Shinpojiumu Kouenyoushishuu* (K. Machida and T. Nishioka, eds.) pp. 102-105. Kyoto University Press, Kyoto, Japan.

Krissinel, E.B., Winn, M.D., Ballard, C.C., Ashton, A.W., Patel, P., Potterton, E.A., McNicholas, S.J., Cowtan, K.D., and Emsley, P. 2004. The new CCP4 Coordinate Library as a toolkit for the design of coordinate-related applications in protein crystallography. *Acta. Crystallogr. D Biol. Crystallogr.* 60:2250-2255.

Weininger, D. 1988. SMILES 1. Introduction and encoding rules. *J. Chem. Inf. Comput. Sci.* 28:31.

Westbrook, J.D., Henrick, K., Ulrich, E., and Berman, H.M. 2005. Classification and use of macromolecular data. Appendix 3.6.2. The Protein Databank exchange dictionary. In *International Tables for Crystallography, Vol. G: Definition and Exchange of Crystallographic Data* (S. Hall and B. McMahon, eds.) pp. 195-197. Springer, Dordrecht, The Netherlands.

### Key References

Berman et al., 2005. See above.

*A description of the wwPDB consortium, its organization, and goals.*

Dutta, S., Burkhardt, K., Bluhm, W.F., and Helen, B. 2006. Using the tools and resources of the RCSB Protein Data Bank. *In* Current Protocols in Bioinformatics (A.D. Baxevanis, R.D.M. Page, G.A. Petsko, L.D. Stein, and G.D. Stormo, eds.) pp. 1.9.1-1.9.40. John Wiley & Sons, Hoboken, N. J.

*Explains various concepts about the PDB, the wwPDB, and tools that are provided by the RCSB partner, as well as the corresponding Ligand Depot service databases and suite of Web tools.*

Golovin et al., 2004. See above.

*A consistent overview of the activities and policies of the MSD group at EBI and of the concepts of the MSD.*

Westbrook et al., 2005. See above.

*A description of the process of the wwPDB exchange, which is the basis of the MSDchem database.*

### Internet Resources

<http://www.ebi.ac.uk/msd-srv/msdchem>

*The MSDchem search home page.*

<http://www.ebi.ac.uk/msd/index.html>

*Contains information about the MSD group and the MSD suite of tools and services.*

<http://www.ebi.ac.uk/msd-srv/msdlite>

*The MSDlite search system provides overview atlas pages for PDB entries, using the MSD database.*

<http://www.ebi.ac.uk/msd-srv/msdsite>

*The MSDsite Web service that provides details about ligand occurrences and binding sites of small molecules in PDB entries.*

<http://www.ebi.ac.uk/msd-srv/docs/dbdoc>

*Contains information about the MSDSD public search relational database and how to download and use it.*

<http://www.ebi.ac.uk/msd-srv/docs/moldoc/help.html>

*The molecule subgraph containment package used by the MSDchem search system.*

<http://deposit.pdb.org/public-component-erf.cif>  
*The Chemical Component Information dictionary that is exchanged in wwPDB.*

<http://www2.chemie.uni-erlangen.de/software/cactvs>

*The CACTVS chemistry algorithm development environment, the main software package used by MSDchem database and Web service*

<http://www2.chemie.uni-erlangen.de/software/corina>

*The CORINA Web service for fast and efficient generation of high-quality 3-D molecular models used to generate idealized coordinates for ligands.*

<http://www.molinspiration.com/jme>

*The home page of the JME Molecular Editor Java applet used by MSDchem Web service.*

<http://jmol.sourceforge.net>

*The home page of the Jmol, free, open source 3-D molecule viewer used by MSDchem Web service.*

<http://www.mdli.com>

*Information about the definition of the popular MDL CTfile Formats.*

<http://www.acdlabs.com>

*The ACD-labs chemical software package used at the time of curation of new ligands.*

[http://users.unimi.it/~ddl/vega/index\\_noanim.htm](http://users.unimi.it/~ddl/vega/index_noanim.htm)

*The VEGA Molecular modeling software package used in the back-end of the MSDchem database.*

---

Contributed by Dimitris Dimitropoulos,

John Ionides, and Kim Henrick  
European Bioinformatics Institute  
Hinxton, Cambridgeshire  
United Kingdom