

MSD-mine Practical Session

EMBL-EBI, Welcome Trust Genome Campus

This is a demonstration of a standalone replica version of the MSD-mine web application accessing a local standalone version of the MSD Search Database (MSDSD) in MySQL. All the software together with the demo database is self-contained and wrapped up in a single CR-ROM disc that may be installed and used directly in any Windows or Linux machine. There are no other special pre-requisites apart from 4.5 GB of free disk space.

You may get a copy (subject to licensing restrictions) of this MSD-demo CD-ROM disc and follow these steps in order to repeat the MSD tutorial later. The purpose of this CD-ROM is for demonstration and educational purposes only. Please read carefully the licence.txt file before start using the CD-ROM disc.

This CR-ROM includes

- a) Java 1.4 preinstalled for windows and linux
- b) MySQL 4.0 preinstalled for windows and linux
- c) The QueryForm Database Tool
- d) Rhino 1.5 from Mozilla
- e) MySQL ODBC drivers 02.50.38, setup for windows
- f) Jakarta Tomcat 4.1 servlet container
- g) The MSDSD (MSD search database) a cut-down demo version preloaded in MySQL
- h) The MSD-mine servlet web application setup to run locally and access the local MySQL database.
- i) A demonstration setup of the MSDSD incremental update mechanism
- j) Example files and scripts
- k) The standalone static MSDSD data warehouse documentation

Ignore steps 1 to 3 if you have already worked out the MSDSD practical session and your MySQL server is already up and running

1) Install the MSD-demo CD-ROM disc

This may have been done already for you in advance, in order to save time so ask your tutor before you start.

Follow the steps in the install.txt file of the CD-ROM disc and run the install.bat script. This will unzip the msddemo.zip file on drive C:\ and create the C:\msddemo directory.

Test that everything is OK. Open "My Computer" or the Windows "Explorer" and check that C:\msddemo exists and is populated with files and directories

2) Configure MSD-demo

Run the script C:\msddemo\bin\config.bat. This will create a virtual drive O: and will copy the MySQL ini files in the appropriate locations. You will have to rerun this script after each reboot of the host in order to re-create the virtual drive O:

Test that everything is OK. Open "My Computer" or the Windows "Explorer" and check that the virtual drive O: exists. Also check that the file C:\my.cnf (MySQL configuration file) exists

3) Start the MySQL database

Run the script O:\bin\startdb.bat. Leave the MySQL server window open

```
040114 10:37:02 InnoDB: Started  
mysqld: ready for connections
```

4) Start the MSD-mine web server

This is the middle-tier of the MSD-mine web application. It operates on your local MSDSD on your local mySQL server and provides a web service that supports the MSD-mine web application. The MSD-mine is a Java servlet that runs inside the Jakarta Tomcat container.

Run `O:\bin\startimine.bat`. Leave the MSD-mine server window open

```
15-Jan-2004 10:40:10 org.apache.commons.modeler.Registry loadRegistry
INFO: Loading registry information
. . . .
```

```
INFO: Initializing Coyote HTTP/1.1 on port 8080
Starting service Tomcat-Standalone
Apache Tomcat/4.1.24
```

Now you may access the MSD-mine web-application through a browser like Internet Explorer or Mozilla. Double click on the "O:\bin\MSD-mine.url" URL link file or access the URL <http://localhost:8080/imine/servlet/imine.hinterface.FilterBuild> from your browser

5) First MSD-mine example scenario (1.1): Querying

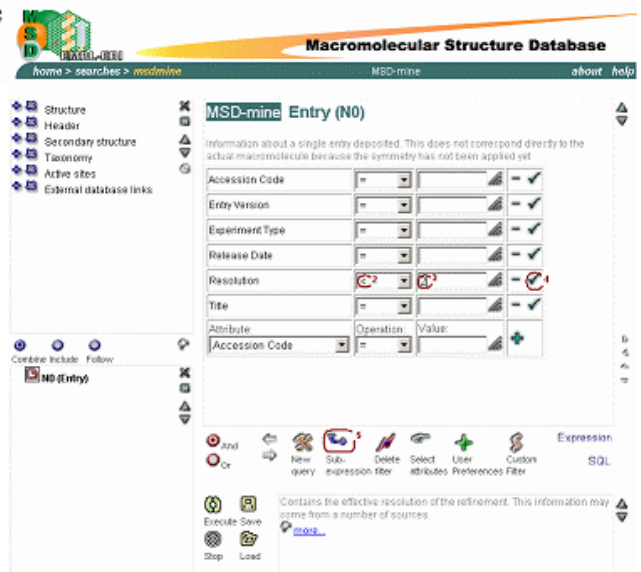
This example demonstrates the MSD-mine querying and filtering and its capabilities of guiding the user in building flexible ad-hoc queries with handy metadata and documentation.

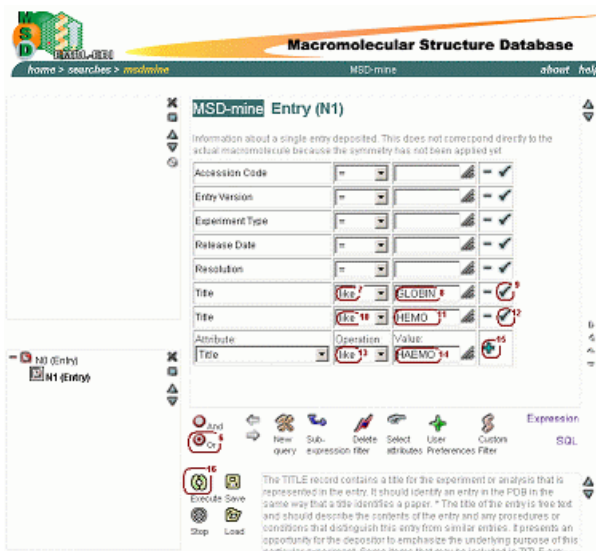
We will get all entries with resolution < 2.0 and having HAEMOGLOBIN or something similar in their title.



We first choose to use the "Structure" mart (section of the database) And then we choose to use the entity "Entry".

Now we are ready to define our first constraint for "Entry". We want the entries that have a Resolution that is better (less) than 2.0.

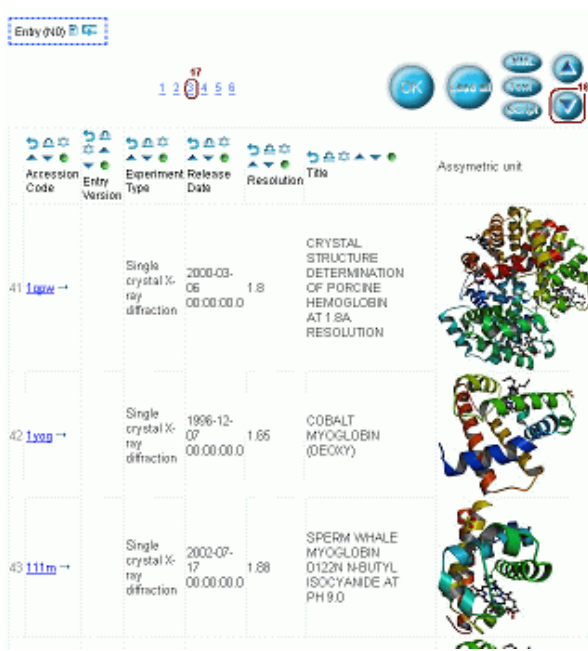




We also want to put constraints in the Title of the entries, in order to find haematological interesting proteins. So we want the title has to contain the work "HAEMO" or "HEMO" or "GLOBIN". Since this is an logical "Or" expression we need to define a sub-expression in order to say that the resolution is < 2.0 and (the title contains the word HAEMO or HEMO or GLOBIN). So we click on the Sub-expression icon

Now we give that we want to logically "Or" the constraints and that Title contains (is like) GLOBIN or HEMO or HAEMO.

We then click the execute button



The "Result page" appears that contains the results of our query. We may click the next page button a couple of times in order to load more records or directly a link to a specific result page. We click "OK" when we our done to stop the query and close the results window.

When done we press the "New Query" icon on the "Filter build" page to continue with our next example.

6) Second MSD-mine example scenario (2.1): Charts – Statistics

This scenario demonstrates the data analysis and charts generation capabilities of MSD-mine. What we will examine is the distribution of assembly types, which means the we will get a chart of the number of monomeric, dimeric, etc assemblies in the MSD database.

It is important to understand that MSD-mine does is not only accessible from your own workstations but it is in fact a web server that provides its service on the network. Ask the hostname or IP address of the workstation next to you (ask the person next to you to run "hostname" or "ipconfig" on a cmd prompt) and change the "localhost" on your URLs with this. Now you are using his MSD-mine server and MSDSD on mySQL. Now proceed with the example:

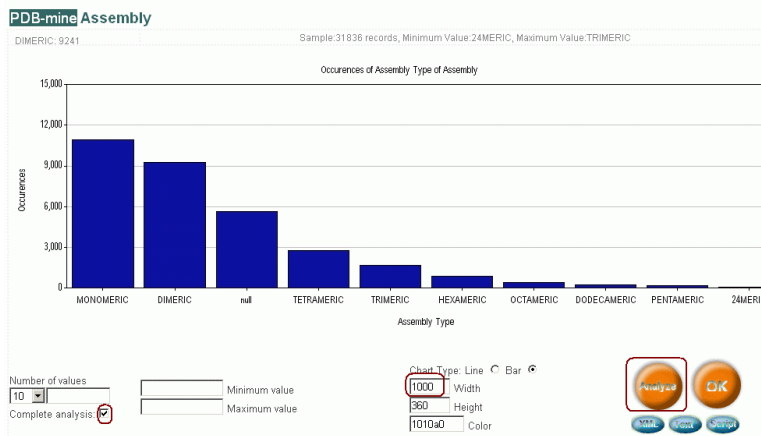


Firstly we will choose the "Assembly" entity of the "Structure" mart as the starting entity. Add the constraint that "Assembly Serial != 0" to use real assemblies. We simply then press the "Execute" button to get the list of assemblies in the MSD database

Accession Code	Assembly Serial	Assembly Type	Number of chains	Assembly image	Select (all)
3 2mr →	1	OCTAMERIC	8		<input type="checkbox"/>
4 2pr →	1	TRIMERIC	3		<input type="checkbox"/>
6 2sk →	1	MONOMERIC	1		<input type="checkbox"/>

In the results page we will see the first 40 assemblies and their types. By clicking on the analyse icon for the "Assembly Type" attribute we will be able to get a chart with its distribution.

We choose to do a complete analysis of the whole MSD database (not just of the 40 records we have loaded), so will check the "complete analysis" checkbox. We also want to make our chart a bit wider (1000 pixels) so we change the corresponding text box and then we press the analyse button.



By moving the mouse cursor on top of say the dimeric bar we can see that 1248 of these assemblies are dimeric.

In the results we see that the analysis was done on the 3697 assemblies that exist on the MSD database and the chart with the 10 most popular types. By moving the mouse cursor on top of say the dimeric bar we can see that 1248 of these assemblies are dimeric.

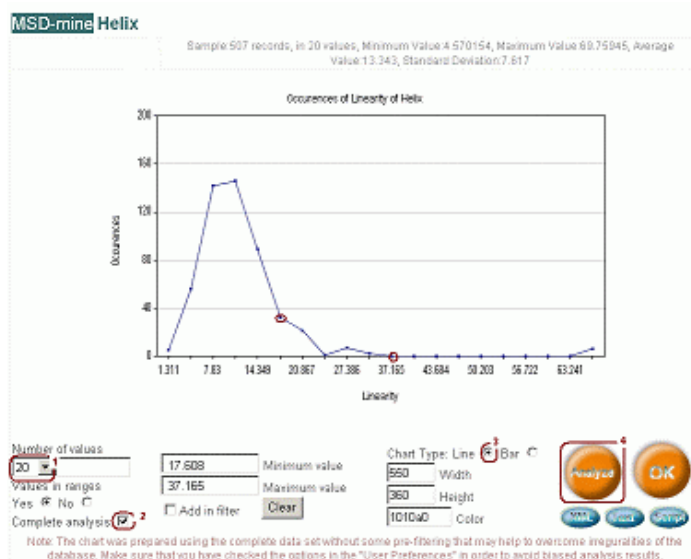
7) Third MSD-mine example scenario (2.5): Combining entities - Advanced analysis

In this example we will demonstrate a more advanced scenario of using MSD-mine, where we will combine (join) entities and perform advanced analysis operations like drill-down and analysing an attribute on the basis of another. We will examine the active site contacts of helices that are part of beta-alpha-beta motifs

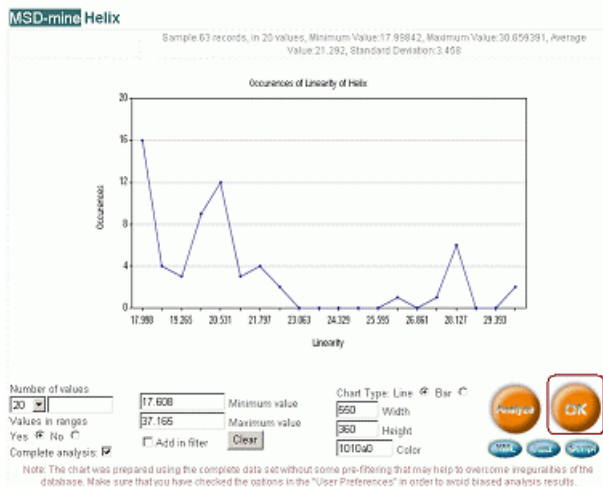
We start from the "Motif" entity (of secondary structure mart) and combine it with the "Helix" entity. We further use the "contacts" relationship to find the Residue contacts of the helix. Using the "Select attributes" icon we choose to "List all" the attributes of the contacts. We then press the "Execute" icon.

Accession Code	Assembly Serial	Chain Code	Motif Serial	Helix Serial	Model Serial	Motif Type	Strand 1Serial	Strand 2Serial	Linearity	Number of Residues	Pitch	Unit Rise	Ligand Code 3Letter	Bond Strength	Cc Cc
31 1e2k	1	A	1	8	1	BETA_ALPHA_BETA 5	5	6	16.548	12	5.261	1.416	SO4	10	4
32 1dtko	1	A	1	3	1	BETA_ALPHA_BETA 2	3	3	8.571	16	5.353	1.437	W04	4	5
33 1dli	1	A	1	1	1	BETA_ALPHA_BETA 1	2	2	10.558	12	5.453	1.419	NAD	3	7
34 1dli	1	A	1	1	1	BETA_ALPHA_BETA 1	2	2	10.558	12	5.453	1.419	NAD	10	4

After navigating through the results we choose to analyse the linearity of these helices and we press the analyse icon for the linearity attribute.

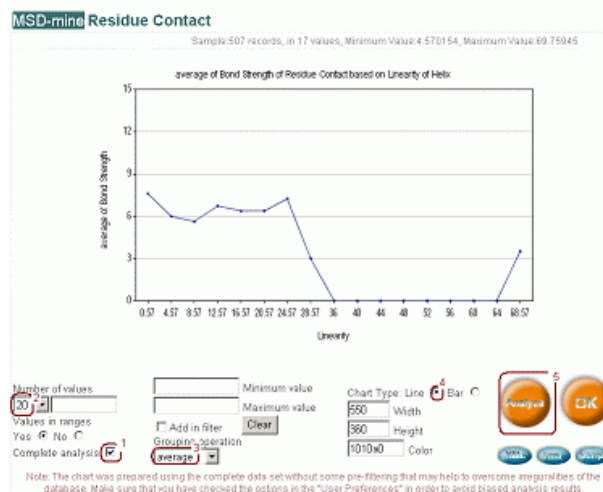


We want to do a "complete analysis" of the linearity of the helices of these contacts splitting the whole range of linearity in 20 sub-ranges and get back a line chart. After selecting the appropriate options we press the "Analyse" icon.



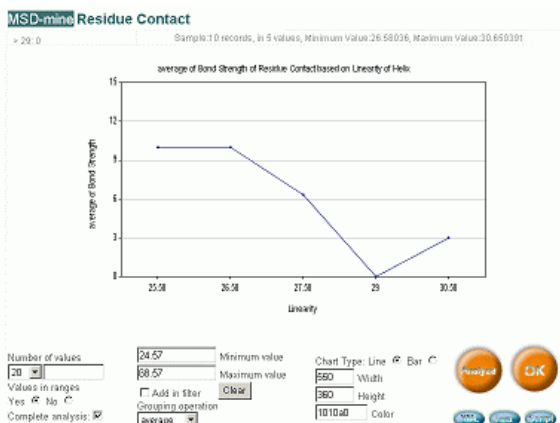
We decide to focus (drill-down) on an area of interest for the range of linearity between ~ 18-37, for that we click on the corresponding points on the line chart and then we press the "analyse" button once more.

We want now to examine if there is a relationship between the linearity of the helices of these



residue contacts and the bond strength of the contacts, so we decide to examine the average bond strength for different values of linearity. We close the analysis page by clicking the "OK" icon and back on the results page we choose to use linearity as the basis attribute by click the "analysis base" icon for the linearity attribute. We then click the "analyse" icon of the bond strength icon. We choose once more to split the range of linearity in 20 sub-ranges, do a complete analysis and get a line chart. We also choose to use the "average" as the grouping operation. Our chart now has determined the whole range of linearity in our result set, has split the range in 20 sub-ranges, has sorted the

result records in 20 subsets based on the range that their linearity falls in and has calculated the average bond strength of these 20 subsets.



We see that while the bond strength seems to remain constant and unaffected by the linearity of the helices, there is an area of irregularity for linearity between 24-62. So we click on the corresponding icons in the chart and we decide to focus in this area by pressing once more the "Analyse" icon. We then may see that while there were 507 records in the initial result set, there are just 10 in the area or the irregularity. It is easy to understand then that this irregularity is rather a result of statistical error than any indication of a relationship between these 2 factors.

8) Stop the MSD-mine web server and the mySQL database

Run the script O:\bin\stopimine.bat to stop MSD-mine and O:\bin\stopdb.bat to shutdown mySQL

