Subject: name vs score 2D plot
Posted by sublimeuser on Mon, 06 Mar 2023 07:36:46 GMT
View Forum Message <> Reply to Message

Hi everyone,

Just wanted to know how to represent a 2D plot - I have 2 columns, one containing scores and the other compound names.

I just wanted to know how to represent a simple 2D plot with names on the x axis and scores on y axis. In 2D view, I only see the score column assigned.

I'm sure this is a very simple thing to do, but for some reason it escapes me.

Many thanks!

Subject: Re: name vs score 2D plot
Posted by nbehrnd on Mon, 06 Mar 2023 08:57:06 GMT
View Forum Message <> Reply to Message

Hello sumblimeuser,

with the 2D plot already created, there will be a button «xy» (next to the wrench symbol). This allows to select the columns from the array / spread sheet, and if the corresponding columns possess a header line, this will be used / updated to annotate the 2D plot, too.

Regards,

Norwid

File Attachments
1) column_selection.png, downloaded 338 times

Subject: Re: name vs score 2D plot
Posted by sublimeuser on Mon, 06 Mar 2023 14:16:30 GMT
View Forum Message <> Reply to Message

Hi Norwid and thank you for the reply!

That's just the problem: while one of the columns is indeed assigned, the other one is unassigned (compound names). How do I assign the name column as well?

Many thanks!

Subject: Re: name vs score 2D plot
Posted by nbehrnd on Mon, 06 Mar 2023 19:10:53 GMT
View Forum Message <> Reply to Message

Hi Sublimeuser,

access the representation of the array / the spread sheet. Mark the (empty) column header in question (left mouse click), then open the pull-down menu available by a right mouse click; select «Set Column Alias».

Regards,

Nprwid

Subject: Re: name vs score 2D plot
Posted by sublimeuser on Mon, 06 Mar 2023 21:14:41 GMT
View Forum Message <> Reply to Message

Defined new aliases for both columns. Unfortunately it still does not work. :(

Subject: Re: name vs score 2D plot
Posted by nbehrnd on Tue, 07 Mar 2023 14:52:35 GMT
View Forum Message <> Reply to Message

Hi sublimeuser,

is it possible for you to share a typical, yet not confidential .sdf presenting the problems to you? I speculate perhaps the problem correlates with the syntax in the input file.

In an instance of Linux Debian 12/bookworm with openbabel (vesion 3.1.1) I created two .sdf files where the «compound name» is the standard InChI key

``` shell
$ obabel -:"O" -:"CCOCC" -:"c1ccncc1" -h --gen3d --append "inchikey" -O testV2000.sdf
$ obabel -:"O" -:"CCOCC" -:"c1ccncc1" -h --gen3d --append "inchikey" -x3000 -O testV3000.sdf
```

This is to cover both the elder V2000 format, as well as the contemporary V3000 one DW prefers on file I/O as .sdf. The reading from the CLI was in pattern of

``` shell
$ datawarrior ./testV3000.sdf &
```

run without noticeable difficulty, nor the adjust the header "compound name" by "InChI key" - there was a prompt (automatic) update in the 2D and 3D plot on abscissa and ordinate.

The four tests run DataWarrior 5.5.0 in said Debian, either the one provided by a pristine installation with the installer fetched by 2022-05-18 (i.e., the state by 2021-04-03), or by subsequent application of the update fetched today (2023-03-07) about the state reached by 2024-02-24.

With regards,

Norwid

File Attachments
1) testV2000.sdf, downloaded 422 times
2) testV3000.sdf, downloaded 401 times
3) testV2000.dwar, downloaded 400 times

Subject: Re: name vs score 2D plot
Posted by sublimeuser on Wed, 08 Mar 2023 10:47:28 GMT
View Forum Message <> Reply to Message

Hi nbehrnd!

As an input, I'm simply using a .csv file with 2 columns (with headers) -> Ligand name and Docking Score. Could this be the source of the problem?

Thanks!

Subject: Re: name vs score 2D plot
Posted by nbehrnd on Wed, 08 Mar 2023 13:36:08 GMT
View Forum Message <> Reply to Message

Hello,

reading a tabulator separated, two column ASCII file by DW (by Ctrl + O) did not yield a problem; the 2D plot generated on the fly was there, change of the column's alias provided the instantaneous update of the 2D plot, interchange of the selection for abscissa and ordinate of the 2D plot: so far, no replication of the problem you report.

Regards,

Norwid

File Attachments
1) test_tsv.txt, downloaded 289 times

2) test_tsv.dwar, downloaded 382 times

Subject: Re: name vs score 2D plot
Posted by sublimeuser on Wed, 08 Mar 2023 14:55:48 GMT
View Forum Message <> Reply to Message

Hi nbehrnd,

It works perfectly if I produce a test file with 5 ligand entries. If fails however,with the large number of rows in my original file (>300.000) so I suspect that DataWarrior is simply not able to plot such a large number of entries.

Thank you for your help once again!

Subject: Re: name vs score 2D plot
Posted by nbehrnd on Wed, 08 Mar 2023 20:32:22 GMT
View Forum Message <> Reply to Message

Hi sublimeuser,

a set of 300k indeed might a bit large (for DW). In such a case I recommend to have a look into AWK. Though initially written for text processing, this «Swiss pocket knive» understands some mathematics you can use to filter by thresholds. Based on the assumption your raw data file is organized as two-column ASCII like `test_tsv.txt` with docking score in the second column, it can be used e.g.

+ to report only the data with an entry in the second column higher than 0.2:

```` shell
$ awk '{if ($2 > 0.2) print}' test_tsv.txt
````

+ to report only the data where the second column's entries are in the in the interval between 0.2 and 0.8:

```` shell
$ awk '{if ($2 > 0.2 && $2 < 0.8) print}' test_tsv.txt
````

which you can redirect into a permanent record either by overwriting the old content (`>`), or by append (`>>`). In case of access to an installation of Linux, you can combine this with a line count (`wc -l`) you either can run on the newly written record

``` shell
```

```
$ awk '{if ($2 > 0.2 && $2 < 0.8) print}' test_tsv.txt > records.txt && wc -l records.txt
2 records.txt
```

or pipe

``` shell
$ awk '{if ($2 > 0.2 && $2 < 0.8) print}' test_tsv.txt | wc -l
2
```

Though this equally removes the headers, the newly written (filtered) record files should be less resource hungry and hence accessible to DW (Edit -> Paste Special -> Paste Without a Header Row).

With regards,

Norwid

Subject: Re: name vs score 2D plot
Posted by thomas on Sat, 18 Mar 2023 16:55:26 GMT
View Forum Message <> Reply to Message

Hi Norwid and Sublimeuser,

there are a couple of defined maxima applied to graphical views in DataWarrior to prevent an explosion of in-memory-data. Usually, these limits are meant to be high enough to not restrict any reasonable graph drawing. For instance category columns can be assigned to an axis if they have not more than 256 distinct categories (your compound name column should be classified as category column). The idea was that more than 256 distinct objects on an axis cannot be displayed well anymore. Admittedly, 256 seems very low, especially considering that the limitation stays in place when zooming in. I will increase the limit to 10000.

Thomas
