Subject: Filter out nasty functions
Posted by juliocoll on Sat, 13 May 2023 08:51:33 GMT
View Forum Message <> Reply to Message

dear Thomas
I am trying to design a filter for large tables of docked chemicals that, among other things, would automatically select for all toxic to eliminate them from the final table.

The partial macro looks like this:

<task name="calculateCompoundProperties">
propertyList=mutagenic tumorigenic reproEffective irritant nasty
structureColumn=Structure
</task>
<task name="changeCategoryFilter">
column=Mutagenic
settings=high low
duplicate=1
</task>
<task name="changeCategoryFilter">
column=Tumorigenic
settings=high low
duplicate=1
</task>
<task name="changeCategoryFilter">
column=Reproductive Effective
settings=high low
duplicate=1
</task>
<task name="changeCategoryFilter">
column=Irritant
settings=high low
duplicate=1
</task>
<task name="changeCategoryFilter">
column=Nasty Functions
settings=<multiple categories>
duplicate=1
</task>

The macro works well except for the nasty.
Despite using the <multiple categories>, the chemicals with "simultaneous nasties" per molecule were not removed.
For instance, the
Nc1c([C@@H](CC(C2=CNC(Nc3ccccc3)=CC2=O)=O)C(C(O[C@@H]2[C@@H]3CCCC2)=O)=C3O)cccc1 molecule predicted a double "nasty" like "polar activated DB; twice activated DB" but it was not removed among other 5434 molecules.

They may be not too many, but it would be better to remove any of those possible cases.
should those names be included at the settings?
That could be an enormous number of dual or perhaps nth number of possibilities!! 8o

Is there any other alternative code solution?

thanks for your attention
julio

Subject: Re: Filter out nasty functions
Posted by nbehrnd on Sun, 14 May 2023 20:24:37 GMT
View Forum Message <> Reply to Message

Dear Julio,

I would like to add two suggestions; how the task is presented/shared, and a revision of the SMILES string.

a) After recording a DW macro, it is possible to export this via Macro -> Export Macro as a file with file extension .dwam. This offers the advantage for an easier/faster import to replicate your observations (Macro -> Import Macro in a first step, Macro -> Run Macro to apply the instructions) by any subsequent reader of your post, because it typically is small enough to be attached to a message here (up to five files in total [e.g. incl. a small test data set] for a maximum over all files of 2MB). It equally prevents the omission of lines; in your most recent example, the opening how the macro is named (for the display within the DW session) and closing line </macro> was missing.

For the purpose of illustration, I enclose a small test set .dwar, and a macro to assign SMILES and compute Mw as .dwam.

b) Curious about the structure the SMILES strings describes, I relayed it to openbabel to write a .sdf, however without success.

```
$ obabel -:"Nc1c([C@@H](CC(C2=CNC(Nc3ccccc3)=CC2=O)=O)C(C(O[C@@H]2[C@@H]
3CCCC2)=O)=C3O)cccc1" -h --gen3d -O test.sdf -xv3000
============================
*** Open Babel Warning  in ParseSmiles
  Invalid SMILES string: 2 unmatched ring bonds.

0 molecules converted
```

Because equally ChemDraw test site[1] faces difficulties to process this one (Structure -> Load SMILES), as well as CDK Depict,[2] can you please check if the SMILES string shared contains

the complete information?

With regards,

Norwid

[1] https://chemdrawdirect.perkinelmer.cloud/js/sample/index.htm l#
[2] https://www.simolecule.com/cdkdepict/depict.html

File Attachments

1) 10Random_Molecules.dwar, downloaded 462 times
2) compute_SMILES_Mw.dwam, downloaded 466 times

Subject: Re: Filter out nasty functions
Posted by juliocoll on Wed, 17 May 2023 15:54:24 GMT
View Forum Message <> Reply to Message

dear nbehrnd,
Thank you for your attention and my apologies for the irregularities you mention of my previous comunication. I am trying to improve it here.

I am doing intense DataWarrior Build Evolutionary Library during the last 2-3 months with several different systems and cavities. I manage to make libraries of a few thousand children molecules per every experiment with 3 runs each. However, I also found that 8.8 to 66 % of the children generated were classified as toxic by DW chemical properties depending on the runs, parent, etc.

Therefore I designed macro 1 to save the results in dwar and sdf formats, before and after elimination of the toxic-children molecules automatically (I enclose the 1.dwam file)

I first found out the children with the smiles I sent you with the 2 nasties without being eliminated by macro 1. However, after sending the topic to the forum, I realized that other nasties were also not being eliminated !!!!!.

I reproduced those failures and select some of the toxic rows and other healthy for you from experiments B10 (~3500 rows) and B13 (~5400 raws) in the files selected-15B10... and selected-24B13... Evolutionary_Library.dwar Both of them came from DataWarrior Build Evolutionary Libraries after being "detoxicated" with macro 1.

The particular smiles that I sent you in the previous communication, corresponds to ID879 of the experiment B10. I visually confirmed that it was the same smiles that I sent you before (I hope).

Thank you for your help !
sincerely Julio

File Attachments

1) 1.dwam, downloaded 437 times

2) selected-15B10Evolutionary_Library.879sentbefore.dwar, downloaded 470 times
3) selected-24B13Evolutionary_Library.dwar, downloaded 469 times

---

Subject: Re: Filter out nasty functions
Posted by juliocoll on Sun, 21 May 2023 10:24:55 GMT
View Forum Message <> Reply to Message

I do not know how the following sentence appeared on the initial top title!

"to design a macro to filter all nasty functions their names should be included"

If that is the answer to my previous question on removing nasties, the next question would appear:

where in Datawarrior can I find out all the nasty names that DataWarrrior is checking when it filters them?

Any one knows it?

thank you in advance
julio

Subject: Re: Filter out nasty functions
Posted by nbehrnd on Tue, 23 May 2023 05:09:10 GMT
View Forum Message <> Reply to Message

Dear Julio,

so far, I understand your approach as following: departing on e.g., selected-15B10Evolutionary_Library.879sentbefore.dwar, you load the 1.dwam macro to filter out compounds which are not good enough for further consideration. In the attempt to replicate this, at level of exporting the results as .sdf files (the macro still running), there are multiple error messages. This either could be a) because of your macro, b) because of the version of DW I use (DW for Linux including the updates packaged by 2023-05-18), or c) a combination of the two.

I briefly tinkered a macro attached below which does the the filtering, however requires manual intervention to save the results as .dwar and .sdf. Conceptually, it builds on DW's assignment of toxicity properties -- like the one you built. This however is followed by calculating a column with an if clause; if you would apply this manually: Data -> Add Calculated Values. As for the formula used: DW's "if syntax" basically is

if(test condition, positive case, negative case)

as the general form, and

if (Mutagenic == "none" && Tumorigenic == "none" && ReproductiveEffective == "none" && Irritant == "none" && NastyFunctions == "", "retain entry", "skip entry")

as the one used here. You recognize e.g. Mutagenic as one of DW's assigned functions; here, DW is requested to check if the assignment was negative (as expressed as the string "none"), yet simultaneously (the &&) was fine about Tumorigenic, ReproductiveEffective, and Irritant. For NastyFunctions, I opted for an empty string as condition for a compound useful to retain. This is why, if an entry passes all these five tests well, its corresponding entry in the new column to build will be the string "retain entry", or else "skip entry".

The script then removes the filters of the individual properties (Tumorigenic, Irritant, etc) to leave only one if the compounds are good, or not for further work; for the ease of work with an additional green, or red background of the cell. For a much smaller set of molecules, this macro works well enough and does not yield the errors I observed with your macro. It however does not (yet) automatically save the results of «filtering» in a separate .sdf/.dwar file. Is this approach in line of what you like to accomplish?

As for «what defines a function nasty», one would have to check the source code,[1] as e.g. file /src/com/actelion/research/datawarrior/task/chem/DETaskCalcu lateChemicalProperties.java for example contains the string «nasty» 16 times and how this property is assigned either in this file's functions, or elsewhere in the source code. Maybe some criteria to mark compounds as not well suitable are similar to the ones in Lilly's criteria.[2]

Norwid

[1] https://github.com/thsa/datawarrior
[2] https://github.com/IanAWatson/Lilly-Medchem-Rules, https://github.com/IanAWatson/LillyMol_6_cmake

File Attachments

1) 20Random_Molecules.dwar, downloaded 450 times
2) suitable_compounds_color.dwam, downloaded 474 times
3) 20Random_Molecules_processed.dwar, downloaded 453 times
4) 2023-05-23T06.26.35 -- screenshots.png, downloaded 415 times

---

Subject: Re: Filter out nasty functions
Posted by juliocoll on Thu, 25 May 2023 20:07:57 GMT
View Forum Message <> Reply to Message

Thank you for your work Norwid !!! REALLY IMPRESSIVE !!.

It will take me sometime to digest all you sent me, test it with large EL dward files and look upon the urls you mention to understand what DW is doing. Hope I can reach the code. My last attempts were not very fruitful....

I will let you know of my advances on due time.

Thanks again, sincerely
julio

Subject: Re: Filter out nasty functions
Posted by juliocoll on Sat, 27 May 2023 19:11:40 GMT
View Forum Message <> Reply to Message

dear Norwid,
I tested an hybrid macro between my 1 and your suitable_compounds_color.
It was successful on a ~12000.dwar compound file from an Evolutionary Library (EL) !!!!!. :)

With your "skip entry" filter, it would be impossible to retain any of those nasty rows for further analysis. In the EL example tested only ~3800 compounds were retained !!! 8o

I am working under windows 10-64, which may be one of the reasons the macros did not exchange well. I relied mainly on recording macros to design mine. Nevertheless, I attached the final 11.dwam just in case you wanted to take a look or it may interest to others in the forum.

Thank you for the information on the nasties. My DW windows code says that there are 20 nasty compounds but I could not find their names.....

File Attachments
1) 11.dwam, downloaded 451 times

Subject: Re: Filter out nasty functions
Posted by nbehrnd on Mon, 29 May 2023 16:56:48 GMT
View Forum Message <> Reply to Message

Dear Julio,

I do not know the parameters to set up the evolutionary library you used -- motives may be similar to approved drugs, or natural products, or molecules derived from a different "seed library". Nor how the internally generated molecules eventually were filtered to be retained in the library; criteria may be sensitive to molecular patterns, scalar properties like molecular weight, or a

combination. Second, your script seems to be stuck in the rut once the first set of molecules was saved. Because there are only the two levels of "safe molecules", and "harmful molecules" by only one remaining filter, the manual intervention after DW's computation to save either one sub set may be considered "still acceptable" (cf. the silent attached). (The small test set of natural product-like compounds, more than half were not considered "safe" either.)

The DataWarrior macro/script should be portable and work equally well regardless of the operating system in which DataWarrior is working. It was not previously tested in Windows because (for a couple of years) DW no longer is satisfied with a 32bit, but requires a 64bit system.

Norwid

File Attachments

1) record.mp4, downloaded 409 times

Subject: Re: Filter out nasty functions
Posted by nbehrnd on Mon, 29 May 2023 17:07:39 GMT
View Forum Message <> Reply to Message

Dear Thomas,

the .dwam DataWarrior macros may grow over multiple "blocks" of tasks like in

<task name="saveFileAs">
fileName=#ask#
</task>

I would like to know if there are permitted special characters to add annotating comments. Percent sign, number sign, exclamation mark as in .tex, Python, or Fortran respectively do not seem suitable here. The addition of an empty line into a .dwam file breaks the macros' working.

Or, would the optional addition of comments add too much complexity to DW's working? This were plausible because the macro editor allows to change the sequence of the individual tasks by click-and-move of the individual tasks ahead.

Norwid

Subject: Re: Filter out nasty functions
Posted by juliocoll on Thu, 01 Jun 2023 07:40:00 GMT
View Forum Message <> Reply to Message

dear Norwick,
I am only using either drugs or natural products for the evolutionary libraries (EL). The criteria were only: docking scores (x4 relative weight), molecular weights between 400-500g/mol (x2) and

logP<3 (x1). I am evolving different parents and pdb complexes from several biological systems: coronavirus, monkey-vaccinia, rodenticides, new antibiotics, collagen hsp, and other. That´s all.

Thanks for the movie. GOOD IDEA!!!.
I will try to incorporate that method to the macro11 to avoid the need for the actual *.sdf manual elimination!

Just one more question:
Is it possible to save in a variable the number of rows at the bottom of the tables (Visible:... and Total:...)?
I am manually using those row numbers to easily differentiate *,dwar and *sdf files from different experiments. It will be great to automatically save those number in the file name.
I would like also to automatically incorporate into the variable a short label for each experiment.
I need the EL *.sdf files to convert them to *.pdbqt for AutoDockVina for consensus with docking-scores in nM affinities.....
When getting a large number of different files it is confusing for me to keep track of all those files to avoid mistakes even using different directories. I usually have a lot of those!!!

Thank you for your attention!
sincerely
julio
Any ideas?

Subject: Re: Filter out nasty functions
Posted by juliocoll on Fri, 02 Jun 2023 11:54:52 GMT
View Forum Message <> Reply to Message

dear Norwid,
I designed a new simplified version by taking into account your movie.
Is the 111.dwar.

I still failed to save the visible *.sdf though.
I also try unssuccesfully to automatically include anything else into the #ask# fileName.....

Thank you!
julio

File Attachments
1) 111.dwam, downloaded 452 times

Subject: Re: Filter out nasty functions

Posted by nbehrnd on Tue, 06 Jun 2023 21:10:42 GMT
View Forum Message <> Reply to Message

Dear Julio,

at present I understand the task ahead as you want to retain both sets of molecules as eventual separate files. With DataWarrior 5.5.0 including the updates packaged by 2023-05-18 (just downloaded again), I'm unable to write a .dwam macro which would save them directly as .sdf. Apparently, the filter -- here, to display either set «retain entry», or «skip» only acts to display. File -> Save Special -> SD-File, which would open the menu to .sdf file still exports molecules which fits either criterion. Coherent with this observation, the entry in Macro Editor annihilates the effort to filter the molecules, too.

This is the reason why the current bypass is to simulate a File -> Save Visible As with the .dwam macro to save the sets separately, however in DataWarrior's native .dwar format. Altogether with a test library, the edited .dwam macro, the two resulting sub sets; and a brief silent are attached. Obviously, this leaves the separate (not macro driven) action to open the .dwam files for a Save -> Save Special -> SD-File still on the table.

Norwid

File Attachments

1) 10Random_Molecules.dwar, downloaded 465 times
2) sucoco_edit.dwam, downloaded 441 times
3) 10Random_Molecules_good.dwar, downloaded 455 times
4) 10Random_Molecules_bad.dwar, downloaded 454 times
5) record.mp4, downloaded 441 times

Subject: Re: Filter out nasty functions
Posted by nbehrnd on Tue, 06 Jun 2023 22:01:25 GMT
View Forum Message <> Reply to Message

Dear Julio,

the report about the number of columns in DataWarrior's spreadsheet arguably is better kept separate as a (meta) analysis outside the program. It is possible to export the data (File -> Save Special -> Textfile) as a .txt with tabulator separated columns, which then is more accessible to other programs and can be queried e.g., by other programs, or home written scripts. AWK and Python can be handy for this; their report to the command line e.g. by the example below

python reporter.py example.txt

equally can be redirected into a permanent record (a choice at time of running the script).

Norwid

File Attachments
1) example.dwar, downloaded 454 times
2) example.txt, downloaded 316 times
3) reporter.py, downloaded 434 times

Subject: Re: Filter out nasty functions
Posted by juliocoll on Wed, 07 Jun 2023 10:51:00 GMT
View Forum Message <> Reply to Message

dear Norwid,
Thank you very much for your new method!

Nevertheless, it would be too complex to implement compared to typing the numbers when saving of the files just by looking those numbers at the bottom of the generated Tables....

I am really surprised by the Java's difficulties to automatically put any of those numbers into a variable that could then be copied into the names of the resulting files. DW EL does a similar labeling when proposing default names for the Evolutionary_library.dwar or sdf files in the corresponding save as dialogs!!!

It is hard to understand why DW is labeling those numbers accurately during and after the evolutionary library generation in EL windows, and supplies all the EL number details inside the resulting EL dward table file!!!. Yet those numbers cannot be extracted to name the files (????)....

Thank you for your efforts,I really appreciatted them
cheers,julio

Subject: Re: Filter out nasty functions
Posted by thomas on Thu, 08 Jun 2023 15:50:30 GMT
View Forum Message <> Reply to Message

Dear Julio,

sorry for the late answer. I didn't follow the conversation. I assume, though, that the solution is quite simple: If I run your macro (I had to add start and end macro tags and replace some spaces by TABs because the forum software did the opposite) on the CNS_NonCNS.dwar reference file, then it calculates all five properties, then creates default filters for all new columns, and then tries to apply 5 category filter settings. The problem is that the default filter for the 'Nasty Functions' column is not a category filter. It is a text filter because in this data file we have too many categories. I assume, you had the same issues. Your macro, however, tries to change a category filter 'Nasty Functions', which does not exist. I changed the macro to configure a text filter such that it hides all rows containing a ';', which is used to separate individual entries. Therefore, rows with multiple nasty functions have at least one ';' in the cell. The task now looks like this:

<task name="changeTextFilter">

column=Nasty Functions
settings=#inverse# #contains# ;
duplicate=1
</task>

(Note that the settings line's spaces should be TABs in reality)

I have attached the entire macro to this message.

File Attachments

1) tox_nasty.dwam, downloaded 442 times

Subject: Re: Filter out nasty functions
Posted by juliocoll on Fri, 09 Jun 2023 05:58:59 GMT
View Forum Message <> Reply to Message

dear Thomas,
THANK YOU fo your last proposal!
Unfortunately, it has failed in a couple of files with ~6000 rows.
Norwid proposal, filtering those rows previously labeled to be skip, works beautifully (see my last 111.darm version).

Still the problem of automatically label the numerous EL files remains.
I handled about 200 dwar and sdf files from EL with thousands of non-toxic children each. The actual file labeling as "Evolutionary-library.dwar" is unsufficient. The manual labeling is prone to errors.
If possible I suggested to automatically use the number of rows that EL shows in its Tables (visible and total). I am using that manually. It resulted almost unique given the variations from experiment to experiment. I failed to incorporate any variable to the *ask* command on the filtering macros.

you can check what I mean by looking into my last publication which makes extensive use of EL for generating diverse libraries:
https://chemrxiv.org/engage/chemrxiv/article-details/6479b8c fbe16ad5c57577cce

Can you help me?

sincerely
julio

Subject: Re: Filter out nasty functions
Posted by nbehrnd on Tue, 13 Jun 2023 19:33:02 GMT
View Forum Message <> Reply to Message

Dear Julio,

postscript to what DW considers a nasty (organic) function: DW builds on openchemlib,[1] a relevant part of it is `NastyFunctionDetector.java`.[2] With varying success one can recover a structure scheme (encoded in the idcode [A-Z, a-z, @, tilde, accent grave and circonflexe, backslash; no numbers] which is enclosed in double quotes) by copy-paste into DataWarrior's sketcher. Because sometimes the recovery failed (e.g., entry oxirane/aziridine), I enclose both the file in question, as well as the .dwar retaining the original description about this set of 50 motifs.

Sincerely,

Norwid

[1] https://github.com/Actelion/openchemlib
[2] https://github.com/Actelion/openchemlib/blob/master/src/main/java/com/actelion/research/chem/NastyFunctionDetector.java

File Attachments

1) nasty.dwar, downloaded 457 times
2) NastyFunctionDetector.java, downloaded 465 times

Subject: Re: Filter out nasty functions
Posted by juliocoll on Wed, 14 Jun 2023 14:34:58 GMT
View Forum Message <> Reply to Message

HEY! 8o
I REALLY ENJOY IT NORWID!!!

At least I have now a complete description of what nasty molecules are in the DW "nasties"!!!

I will try to decipher how they look like.....

THANK YOU VERY MUCH!!!

sincerely
julio

Subject: Re: Filter out nasty functions
Posted by juliocoll on Wed, 14 Jun 2023 15:47:23 GMT
View Forum Message <> Reply to Message

dear Norwid
I enclose the 50DWnasties.dwar translation of the DW codes from the NastyFunctionDetector.java file you kindly sent me.

Although not an organic chemistry expert, I could recognize most of the "nasties", except 3 of them which I could not decipher with the DW sketch and a few other which raise me some questions (?).

I really appreciate your effort!
THANKS AGAIN !

cheers
julio

File Attachments
1) 50DWnasties.dwar, downloaded 436 times

Subject: Re: Filter out nasty functions
Posted by thomas on Wed, 14 Jun 2023 19:30:03 GMT
View Forum Message <> Reply to Message

If Java strings contain a back slash '\' this needs to be encoded as double back slash '\\'. This explains, why three of the idcodes of the NastyFunctionDetector cannot be converted by DataWarrior. You need to replace '\\' again by '\' before pasting. You should also add a title line, e.g. "idcode<TAB>name". <TAB> should be a real TAB character and 'idcode' is a keyword for DataWarrior to tell it that this column contains idcode-encode chemical structures.

If you paste such a table with Edit->Paste, then DataWarrior creates a new document with the idcodes correctly decoded as substructure queries rather than complete structures. In this case the structure column is marked to contain substructures rather than structures. In this case it is important, because the nasty function substructures contain query features, which are lost if one copies the idcodes into an existing document's structure column. Then DataWarrior automatically converts substructures to structures, which removes atom/bond query features and considers open valences to be filled with hydrogen atoms.

I attach both, the text file with the idcodes and the dwar file containing the nasty function defining substructures.

Query feartures, which are not directly visible, can be recognized by their yellow atom/bond backgrounds. You can see all query features used by first opening the substructure in the structure editor and then opening the query feature dialog of a bond or atom. To open the editor double click the structure in the table. Then make sure the lasso tool is chosen in the editor. Then double click an atom or bond. The respective atom or bond query feature dialog opens and shows the current selection of query features for that atom or bond.

File Attachments
1) nasty_functions.txt, downloaded 338 times
2) nasty_functions.dwar, downloaded 462 times

Subject: Re: Filter out nasty functions
Posted by juliocoll on Thu, 15 Jun 2023 14:40:57 GMT
View Forum Message <> Reply to Message

Thanks for the files Thomas. MORE CLEAR NOW ! :)

The "nasties" became really complex....but I love to understand a bit more of the code behind.

I think already understood > 80 % of what you explain (i.e., both the use of // & / and the substructures/structures are clarified).

I could reconstruct successfuly your dwar table by using Edit/Paste (I was selecting & copying from the text file in notepad and then right click/paste into a dwar table).

I also deep-dive into the lazo-dialog: REALLY AMAZING!!!. It was surprising the amount of information it contains !!!. I need some time to digest all this new information to fully follow each of the nastic functions!

Thanks again and cheers,
julio