Subject: Why is clustering limited up to 10000 molecules?
Posted by Nastasia on Wed, 22 Jul 2015 14:37:15 GMT
View Forum Message <> Reply to Message

I am trying to cluster my molecules, but the number of my molecules is far beyond 10000, actually it's about 50000. However, there are some duplicate molecules, which may be joined, but when I do that I loose my different descriptors important for each case, as they are just merged into some multiple group. Is it possible to increase the size for clusters?

Subject: Re: Why is clustering limited up to 10000 molecules?
Posted by thomas on Thu, 23 Jul 2015 19:40:59 GMT
View Forum Message <> Reply to Message

Clustering is a very old functionality and not used very often. The used algorithm is reproducible and is not based on random starting point as other faster algorithms. The flipside, however, is that is not a particularly fast one and its memory consumption and processor usage grow quadratic with the size of the data file. Without a limit people would experience error messages or extreme processing times causing lots of frustration. I am aware that DataWarrior needs an alternative clustering algorithm and it is on the feature list for the future, but till now I considered clustering low priority, because often there are better methods to reach a certain goal.

Kind regards, Thomas

Subject: Re: Why is clustering limited up to 10000 molecules?
Posted by avkitex on Sun, 28 Feb 2016 18:03:50 GMT
View Forum Message <> Reply to Message

Dear autors. What if I would like to cluster ~100000 compounds?
I understand that it could consume huge amount of time and ram.

Best,
Nikita

Subject: Re: Why is clustering limited up to 10000 molecules?
Posted by thomas on Thu, 03 Mar 2016 20:49:57 GMT
View Forum Message <> Reply to Message

Dear Nikita, the current DataWarrior version is limited to 10.000 compounds. For 100.000 compounds the current algorithm would need 20GB of memory and many hours to finish. The only way to do it would be to patch the datawarrior.jar file (OSX or Linux only) or to change the source code and to recompile.

Regards, Thomas

Subject: Re: Why is clustering limited up to 10000 molecules?
Posted by avkitex on Thu, 03 Mar 2016 20:54:28 GMT
View Forum Message <> Reply to Message

Dear, Thomas

I have servers with about 30Gb ram and a lot of time. And I woult like to try doing it.
Can you tell me how to patch and recompile jar file? And what is the best way to run it in console mode? (I have molecules in sdf or mol2 format. I would like to have a clusterisation tree in newick or compatable format.)

Looking forward to hear from you.

Best,
Nikita

Subject: Re: Why is clustering limited up to 10000 molecules?
Posted by thomas on Fri, 04 Mar 2016 21:24:48 GMT
View Forum Message <> Reply to Message

Dear Nikita,

if you use Linux or OSX, I will send you a customized datawarrior.jar file without limitation. But don't complain, if the clustering process seems to take forever...

If your server runs Linux, you can install the datawarrior directory on the server and login from remote via 'ssh -X' and then launch datawarrior the usual way. Then you all DataWarrior windows open on the client machine while it executes on the server and uses server RAM and server cores.

Best wishes, Thomas

Subject: Re: Why is clustering limited up to 10000 molecules?
Posted by dataviz on Wed, 22 Mar 2017 10:34:15 GMT
View Forum Message <> Reply to Message

Hello
it is indeed true that to be able to run the cluster function on a dataset larger than 10,000 would be really great considering all the new data available today... This can be done externally but Warrior is so convenient that in a future release, it really would be nice...
Sincerely

Subject: Re: Why is clustering limited up to 10000 molecules?
Posted by thomas on Thu, 30 Mar 2017 18:35:00 GMT

View Forum Message <> Reply to Message

Hi Dataviz,

I have lifted the limit from 10.000 to 100.000 structures, but display a warning above 20.000 compounds. Nevertheless, clustering of significantly more structures than 20.000 will require a Linux or Macintosh computer, where it is easy to increase the memory maximum that DataWarrior is allowed to use. The change will be available with the next update.

Regards, Thomas

Subject: Re: Why is clustering limited up to 10000 molecules?
Posted by E3ubiquitinligase on Thu, 10 Jan 2019 20:46:46 GMT
View Forum Message <> Reply to Message

How do we manually increase the memory and compound limit for Datawarrior on a Mac? The new tSNE feature is working unbelievably well and I'd like to be able to apply it to >100k molecules. It hovers around 2.46 GB memory while crunching the tSNE so I'd love to be able to reove the existing limits.

Many thanks!
Mike

Subject: Re: Why is clustering limited up to 10000 molecules?
Posted by thomas on Thu, 10 Jan 2019 23:00:48 GMT
View Forum Message <> Reply to Message

on the Mac this is easy. You need to increase the -Xmx setting in the Info.plist file. The details are explained in the manual: on the bottom of this page
http://www.openmolecules.org/help/installation.html#Installa tion

Thomas

Subject: Re: Why is clustering limited up to 10000 molecules?
Posted by E3ubiquitinligase on Fri, 11 Jan 2019 06:21:38 GMT
View Forum Message <> Reply to Message

Works wonderfully - thank you!!

Mike