Subject: AllFragFp
Posted by Christophe on Fri, 12 Apr 2024 08:48:34 GMT
View Forum Message <> Reply to Message

Hello everyone,

While updating my old version of DW, I've just seen that a new Descriptor, i.e. AllFragFP was available. Does anyone could tell me about any difference(s) with the FragFP ?
I can't find its description on the online User Manual.
Thanks a lot
Christophe

Subject: Re: AllFragFp
Posted by nbehrnd on Tue, 16 Apr 2024 20:48:21 GMT
View Forum Message <> Reply to Message

Hello Christophe,

AllFragFP could be something still very early in implementation into DW and thus not (not yet) documented. If one queries the source code of DataWarrior,[1] and its assisting openchemlib[2] on GitHub, only the later contains this very string only once in file DescriptorConstants.java, lines 88 to 96 as a `DESCRIPTOR_TYPE_MOLECULE`.

Norwid

[1] https://github.com/thsa/datawarrior
[2] https://github.com/Actelion/openchemlib

Subject: Re: AllFragFp
Posted by Christophe on Fri, 19 Apr 2024 14:53:41 GMT
View Forum Message <> Reply to Message

Thank you Norwid

Subject: Re: AllFragFp
Posted by thomas on Wed, 24 Apr 2024 12:45:18 GMT
View Forum Message <> Reply to Message

AllFragFp is substantially different from FragFp: it is hashed and uses 2048 bits. AllFragFp internally generates all substructures of a given molecule with up to 6 connected bonds including stereo chemistry. These substructures are converted into a canonical representation from which a hash code between 0 to 2047 is generated, for which the corresponding bit is set. The original idea was to accelerate the substructure search by a more discriminating descriptor than the FragFp. If the AllFragFp descriptor is available in a DataWarrior file, then DataWarrior uses that

for substructure pre-screening. Since the sub-structure search is usually fast for mot more than some hundred thousand molecules, one shouldn't bother to use the AllFragFp. For many millions, however, it makes a significant difference.

Regarding the value of similarities calculated by this descriptor, I didn't really investigate it applicability domain. It certainly will produce very fine grained similarity values, but the SkeletonSpheres descriptor is will probably generate more intuitive ones, because by design single atom replacements cause less large losses of similarity compared to other substructure based descriptors.

Subject: Re: AllFragFp
Posted by Christophe on Fri, 26 Apr 2024 07:45:51 GMT
View Forum Message <> Reply to Message

Hello Thomas,

Thank you for this more detailed reply.
If I've understood correctly, the AllFragFP is more akin to what is more commonly known in chemometrics as a path fingerprints, whereas the SkelSpheres descriptor would be of the Extended Connectivity FP type. Is this correct?

Christophe

Subject: Re: AllFragFp
Posted by thomas on Thu, 23 May 2024 20:48:01 GMT
View Forum Message <> Reply to Message

Hi Christophe,

somewhat: SkelSpheres is like Extended Connectivity FP, but it includes full stereo features and it inlcudes a set of fragments, which contain the sckeleton only (no atom types). This makes it more tolerant to single atom replacements, which don't cause a drastic drop of similarity anymore.

AllFragFp is not just linear paths, it covers all fragments with up to 6 bonds (circular, chains, combinations) that are a substructure of the given molecule.

Thomas
