chem-bla-ics

A new CDK default fingerprinter?

Egon Willighagen

Published July 16, 2010

Citation
Willighagen, E. (n.d.). In *chem-bla-ics*. chem-bla-ics. https://doi.org/10.59350/485a6-g8f87

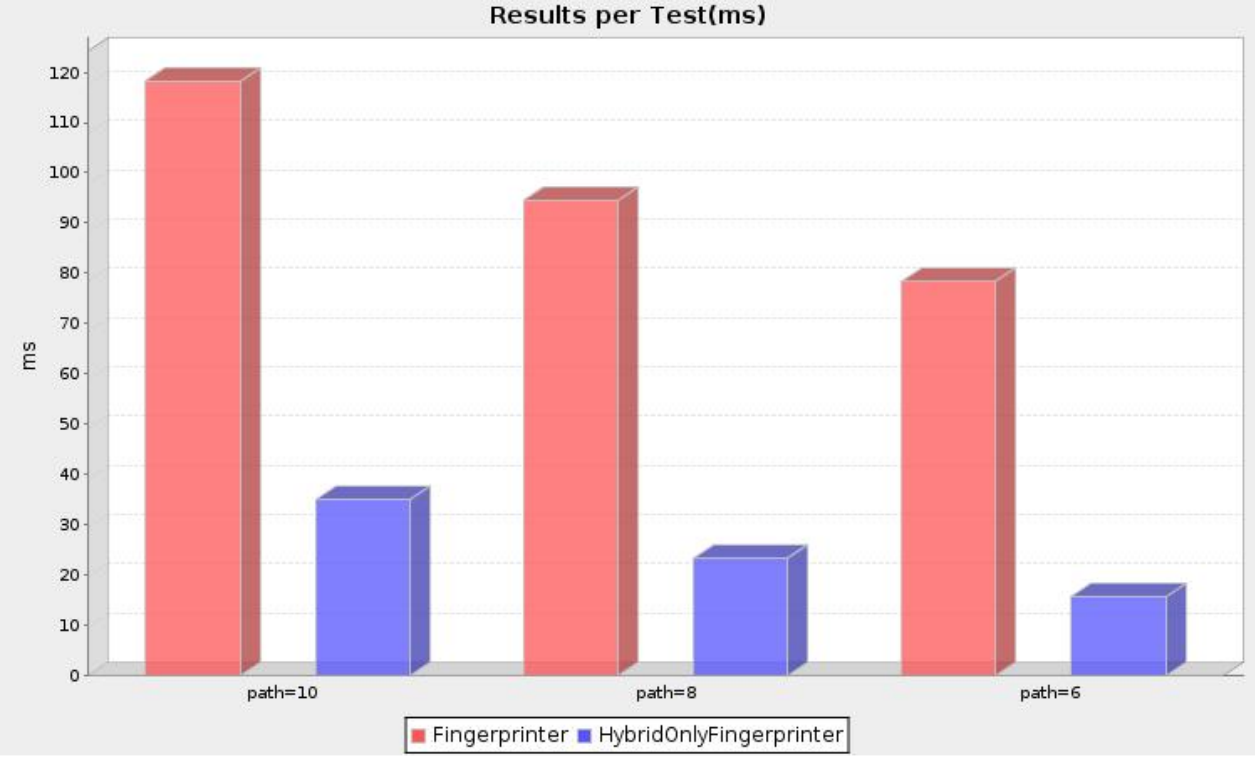
Keywords
Cdk, Fingerprint

Copyright
Copyright © Egon Willighagen 2010. Distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

chem-bla-ics

The current default fingerprinter in the CDK depends on aromaticity, but that concept is algorithmically difficult to define, and even experimentally there are multiple dimensions to this concept. Moreover, calculating aromaticity is not cheap, as it requires detecting of ring systems. The purpose why aromaticity is actually included is this: people expect a ethenol moiety to match phenol.

Now, an alternative is to not use aromaticity, but hybridization information instead: an aromatic bond is basically just a bond between two sp2-hybridized atoms. Removes some algorithmic complexity and speeds up the calculation:



The definition of the fingerprint has changed, and a bond between two sp2-hybridized atoms may not be aromatic. We can therefore expect that the fingerprint will give more false positives with substructure search. I'm hoping that Rajarshi can find some time to compare this new fingerprint in his excellent analysis he did some time ago.

The source code can be found in my GitHub repository, with the new class HybridOnlyFingerprinter.

Update: The fingerprinter has been renamed to HybridizationFingerprinter, and is part of the CDK 1.3.9 release.