Subject: Metrics to use for UMAP analysis with SkelSpheres descriptor
Posted by Christophe on Thu, 08 Sep 2022 09:45:03 GMT
View Forum Message <> Reply to Message

Hello everyone,

I'd like to use the SkelSpheres descriptor generated by DW to run a UMAP analysis.

As far as I understand, the SkelSpheres descriptor is a 1024 bins matrix. My question : Are these 1024 bins filled with binary (0/1) or counted (integers) or continues (real) data ?

I need to be sure of this answer to select the correct metric available and required to perform ordination of compounds with the new UMAP functionality provided in the last version of DW.

Thanks

Subject: Re: Metrics to use for UMAP analysis with SkelSpheres descriptor
Posted by nbehrnd on Thu, 08 Sep 2022 20:53:56 GMT
View Forum Message <> Reply to Message

Dear Christophe,

curious about the description «Technically, it [the SkelSpheres Descriptor] is a byte vector with a resolution of 1024 bins.» (help menu in DW, chapter Similarity & Descriptors, section Molecule or Reaction Similarity and Descriptors), I found an open access publication[1] going a bit more in detail. In section 4.2, the authors describe it with

«This descriptor was developed by Actelion. It is a vector of integers which represents the occurrence of different substructures in a molecule. Five circular layers with increasing bond distance are located for each atom in the molecule. Hydrogen atoms are not considered. This results in five fragments starting with the naked central atom, adding one layer at a time. Every fragment is encoded as a canonical string (id-code), similar to the generation of canonical SMILES. The canonical id-code includes the stereochemistry of the encoded fragment, which is a feature missing in other molecular descriptors. The string is then assigned to one of 1024 fields n in a vector. Therefore, the hash value of the id-code is calculated and the corresponding value in the vector is increased by one. The Hashlittle algorithm from Jenkins is used as a binning function which takes a text string as input and returns an integer value between 0 (inclusive) and 1024 (exclusive). [...] To consider the molecular scaffold without the influence of the hetero atoms, the whole calculation is repeated while replacing the hetero atoms with carbon. The resulting hash values are used to increment the corresponding fields in the vector. By adding this skeleton information to the descriptor vector the similarity calculation between two descriptor vectors becomes a bit insensitive to the exact position of the hetero atoms in two molecules. This directs the similarity value toward the perception of similarity by medicinal chemists. For them the exact position of a hetero atom is not as discriminating as it would be for the spheres descriptor without the skeleton coding part. The additional consideration of the scaffold information and the use of a histogram instead of a binary vector distinguishes the SkeletonSpheres descriptor from circular fingerprints.»

So far however, I don't understand the concept of "byte" in byte vector when entering the integers as elements of the vector either, which could be crucial.

Norwid

[1] The Screening Compound Collection: A Key Asset for Drug Discovery. C. Boss, J. Hazemann, T. Kimmerlin, M. von Korff, U. Lüthi, O. Peter, T. Sander, R. Siegrist, Chimia 2017, 71, 667-677, DOI: 10.2533/chimia.2017.667 (https://chimia.ch/chimia/article/view/2017_667 open access).

Subject: Re: Metrics to use for UMAP analysis with SkelSpheres descriptor
Posted by Christophe on Fri, 09 Sep 2022 08:23:58 GMT
View Forum Message <> Reply to Message

Hi Norwid,

Thank you for the explanation. I know how to manage matrices containing binary data or occurrences (integers) as well as reals or even combinations of all these. But it seems to me from your explanations, that in addition to columns that would contain the occurrence of different substructures, the SkelSpheres Descriptor algorithm would also, in addition, converts strings into numbers. In this case, I confess that I am not sure which similarity/distance index to use.

Thank you for your time.

Christophe
