Combining a bisector tree with the Tanimoto distance for similarity searches and beyond

(Kyushu Institute of Technology) Francois Berenger*, Yoshihiro Yamanishi

A bisector tree [3] is a data structure from computational geometry to do static spatial indexing of points. It allows to do fast and exact nearest neighbor searches (and other queries) in an N-dimensional space, provided a metric to measure the distance between any two points in that space exists.

We have made an open source implementation of a bisector tree (https://github.com/UnixJunkie/bisect-tree). It is bucketized, such that several nearby molecules can be put into the same bucket. The (maximum) bucket size is a user-chosen parameter. Our implementation proposes two heuristics, in order to find good vantage points [4,5,6] during tree construction, to accelerate subsequent queries.

In this presentation, we report on the indexing and querying of millions of molecules and the associated challenges.

Figure 1: the Stanford bunny (~35000 3D points), as guillotined by the first layer of a bisector tree.

Figure 2: Brute force nearest neighbor query time (median of five experiments), as a function of the number of molecules.

*beren314@bio.kyutech.ac.jp
Figure 3: time to find all molecules within Tanimoto distance $\leq 0.01$ from a query molecule (median of five experiments), as a function of the bucket size $K$ and the number of molecules $N$.


*beren314@bio.kyutech.ac.jp*