[P3] Combining a bisector tree with the Tanimoto distance for similarity searches and beyond

Francois Berenger[1], Yoshihiro Yamanishi[1]

[1]*System Cohort Division, Medical Institute of Bioregulation, Kyushu University, 3-1-1 Maidashi, Higashi-ku, Fukuoka 812-8582, Japan.*

This is not known to all chemoinformaticians, but the Tanimoto score (t) can be converted into a proper distance (1 - t) that satisfies the triangular inequality [1,2].

On the other hand, a bisector tree [3] allows to do fast but 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. Figure 1 shows an example point set embedded into a bisector tree.

We have implemented a bisector tree (cf. https://github.com/UnixJunkie/bisec-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 poster, we show the construction time of such a data structure. We also compare its performance versus brute force searches on a large virtual chemical library.

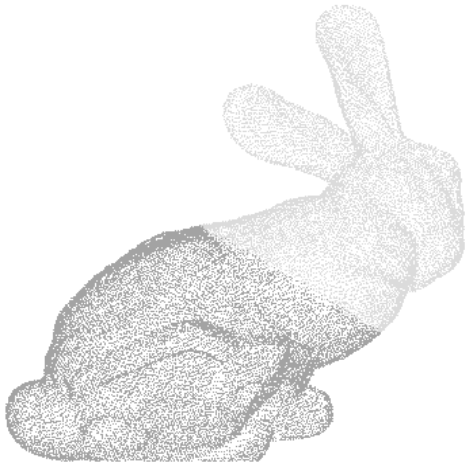

Figure 1: the Stanford bunny, made of 35947 3D points, as guillotined by the first layer of a bisector tree.

Bibliography:

[1] Lipkus, Alan H. "A proof of the triangle inequality for the Tanimoto distance." Journal of Mathematical Chemistry 26.1-3 (1999): 263-265.
[2] Kosub, Sven. "A note on the triangle inequality for the jaccard distance." arXiv preprint arXiv:1612.02696 (2016).
[3] Kalantari, Iraj, and Gerard McDonald. "A data structure and an algorithm for the nearest point problem." IEEE Transactions on Software Engineering 5 (1983): 631-634.
[4] Shapiro, Marvin. "The choice of reference points in best-match file searching." Communications of the ACM 20.5 (1977): 339-343.
[5] Peter N. "Data structures and algorithms for nearest neighbor search in general metric spaces." SODA. Vol. 93. No. 194. 1993.
[6] Brin, Sergey. "Near neighbor search in large metric spaces." (1995).