Fast alignment of fragmentation trees
Supplementary Material

Franziska Hufsky[1,2], Kai Dührkop[1], Florian Rasche[1], Markus Chimani[3], and Sebastian Böcker[1]

[1] Chair for Bioinformatics, Friedrich-Schiller-University, Jena, Germany,
{franziska.hufsky,kai.duehrkop,florian.rasche,sebastian.boecker}@uni-jena.de
[2] Max Planck Institute for Chemical Ecology, Beutenberg Campus, Jena, Germany
[3] Algorithm Engineering, Friedrich-Schiller-University, Jena, Germany,
markus.chimani@uni-jena.de

1 Proof of Lemma 1 and Theorem 1

The running time for computing $S_{u,v}[A,B]$ is dominated by the computation of $deleteL_{u,v}[A,B]$ and $deleteR_{u,v}[A,B]$, as well as $joinL_{u,v}[A,B]$ and $joinR_{u,v}[A,B]$. To compute $deleteR_{u,v}[A,B]$ for all $A \subseteq C(u)$ and $B \subseteq C(v)$ we have to iterate over all $A' \subseteq A$ and all $b \in B$. Iterating over all subsets $A' \subseteq A \subseteq C(u)$ needs 3^{d_u} time. Iterating over all $b \in B \subseteq C(v)$ needs $2^{d_v} \cdot d_v$ time. This leads to an overall running time of $O(3^{d_u} \cdot 2^{d_v} \cdot d_v)$ for the computation of $deleteR_{u,v}[A,B]$ and $O(2^{d_u} \cdot 3^{d_v} \cdot d_u)$ for $deleteL_{u,v}[A,B]$. For $joinL_{u,v}[A,B]$ and $joinR_{u,v}[A,B]$ the proof is similar.

To prove the correctness of recurrence (2) we have to distinguish three cases: nodes u,v are matched, u is deleted or v is deleted. We know $S_{u,v}[A,\emptyset] = S_{u,v}[\emptyset,B] = 0$.

Matching u,v is covered by $match_{u,v}[A,B]$. To match two nodes u,v we have to match at least one outgoing edge from u with one outgoing edge from v and therefore one child $a \in A$ with one child $b \in B$. For each $a \in A$ and $b \in B$, $S(a,b)$ is the maximal score of the local alignment of the two subtrees rooted in a and b, $\sigma(ua,vb)$ is the score of matching ua to vb. Furthermore $S_{u,v}[A-\{a\},B-\{b\}]$ is the score of the optimal local alignment of the remaining children. The sum is the optimal score for $S_{u,v}[A,B]$ when matching ua with vb. Since at least one outgoing edge from u has to be matched to one outgoing edge from v, this is in fact the maximum over all children $a \in A$ and $b \in B$.

The deletion of an outgoing edge from u is covered by $deleteL_{u,v}[A,B]$. When deleting edge ua from the left tree, we have to bipartition the children of v depending on whether they match with children of u or of a. Therefore we have to iterate over all subsets $B' \subseteq B$. The maximum score of a local alignment with subtree rooted in a and v, using all children of a and the children in $B' \subseteq B$ from v, is already given by $S_{a,v}[C(a),B']$. The remaining children $B-B'$ from v are matched with the children from u, where the maximum score is also already given by $S_{u,v}[A-\{a\},B-B']$. Furthermore, we have to add the costs for deleting an edge given by $\sigma(ua,\lambda)$. The sum of these three values is maximized over all bipartitions of B. Deleting a node u means deleting one of its outgoing edges, so we further have to maximize over all children $a \in A$.

The reasoning when deleting v is analogous.

Computing the maximum over all three cases and 0 results in the maximum score for $S_{u,v}[A,B]$, that is the score of an optimal local alignment with subtree rooted in u and v, respectively, such that at most the children A of u and B of v are used in the alignment. □

2 Correlation with chemical similarity

Correlation of fragmentation tree fingerprint similarity with chemical similarity of the molecules was suggested by Rasche *et al.* [1], and carried out for the *MassBank* and *Orbitrap* dataset. Here, we evaluate correlation of alignment-based similarity with chemical similarity for the *Hill* dataset. We tightly follow the protocol of [1].

We compute fragmentation tree alignments for all pairs of trees in the *Hill* dataset. We normalize similarities by perfect scores: For each fragmentation tree we compute the alignment score of the tree against itself, then use the minimum of the two scores, taken to the power of 0.5. We use columns of the resulting similarity matrix as fingerprints (or feature vectors), corresponding to similarities of the respective tree to all other trees in the entire dataset. The *fingerprint similarity* of two compounds is the Pearson coefficient of the corresponding two columns in the original similarity matrix. To measure chemical similarity of two molecules, we apply the widely-used PubChem/Tanimoto score [2,3]. We then correlate these two measures for all compound pairs, excluding self-comparisons of a compound against itself.

Using all trees of the *Hill* dataset, we obtain a Pearson correlation $r = +0.40$ ($r^2 = 0.16$) and a Spearman correlation $\rho = +0.31$ ($\rho^2 = 0.10$). The correlation coefficients increase when we limit our analysis to fragmentation trees with a lower bound on the number of neutral losses: Pearson correlation is $r = +0.41$, $+0.42$, and $+0.43$ for trees with at least 5, 7, or 10 losses, respectively. We observe a similar effect for the Spearman correlation.

Note that the correlation is weaker than those reported in [1] for *MassBank*, *Orbitrap*, and a third dataset, containing 44 compounds measured on an API QSTAR QTOF spectrometer by Applied Biosystems. (This last dataset was too small to be used for running time evaluations.) There, Pearson correlations of up to $r = +0.65$ ($r^2 = 0.42$) were observed. We attribute this to two facts: Fragmentation trees in the *Hill* dataset are larger than those in the other datasets and their sizes vary stronger. This may interfere with the alignment method to some extend. Furthermore, the *Hill* dataset is very heterogeneous, with most compound classes containing only a single compound, and few compound pairs have chemical similarity above 0.6. Hence, the chemical similarity measured by PubChem/Tanimoto score may contain less “signal” and more “noise”. In view of the non-informative compound classes, we refrained from clustering the compounds based on their chemical similarity, as suggested in [1].

References Supplementary Material

1. F. Rasche, K. Scheubert, F. Hufsky, T. Zichner, M. Kai, A. Svatoš, and S. Böcker. Identifying the unknowns by aligning fragmentation trees. *Anal. Chem.*, Mar. 2012. In press, doi:10.1021/ac300304u.
2. D. J. Rogers and T. T. Tanimoto. A computer program for classifying plants. *Science*, 132(3434):1115–1118, 1960.
3. Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, and S. H. Bryant. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, 37(Web Server issue):W623–W633, 2009.