Supplementary material - Succinct Inerval-Splitting Tree for Scalable Similarity Search of Compound-Protein Pairs with Property Constaints

Yasuo Tabei, Akihiro Kishimoto, Masaaki Kotera and Yoshihiro Yamanishi

1 Experiments on large-scale compound sets

1.1 Setups

The metabolic network is an important biochemical network involving enzymatic reactions among chemical compounds, but a large number of metabolic pathways remain unknown. Therefore, the prediction of substrate-product pairs (compound-compound pairs converted to each other by enzymatic reactions) is a challenging issue in recent bioinformatics toward the de novo reconstruction of metabolic pathways.

1.2 Results

We tested SITA, MT and BIN on their abilities to search for substrate-product pairs in metabolic pathways. We calculated the average search time of $2,000$ queries on a large-scale database consisting of $243,438,006$ compound-compound pairs, where each pair is represented by a fingerprint with the dimension of 1,758 based on differential chemical substructures (Kotera et al, In Proceedings of ISMB/ECCB2013). In addition, we used the absolute difference of the molecular weights of each compound-compound pair as a property.

Figure 1 shows the average search time of each algorithm for $\epsilon = 0.6$, $\epsilon = 0.8$, $\delta = 0.5$ and $\delta = 5$. SITA outperformed MT and BIN by 2 or 3 orders of magnitude. The search time was significantly reduced by applying SITA. The performance difference tends to be smaller for lager δ of 5 and smaller ϵ of 0.6, Table 1 details the average search time of each algorithm on $243,438,006$ fingerprints, where $|P_1|$ is the number of candidate fingerprints chosen by database partitioning, $|P_1 \cap [i_l, i_r]|$ is the number of candidate fingerprints chosen by database partitioning plus sorting and binary search, #Rank is the number of rank operations and $|P_{NC}|$ is the number of solutions.

Figure 2 depicts the average search time of various ϵ with $\delta = 0.5$ and $\delta = 5$. SITA significantly outperformed the other algorithms. The performance difference tends to be smaller for larger ϵ.

Figure 3 shows the memory for the number of compound-compound pairs. MT consumed a large amount of memory of 114 GB for $243,438,006$ compound-compound pairs. SITA consumed 89 GB, which was smaller than MT's memory and was almost the same as BIN's memory of 86 GB.

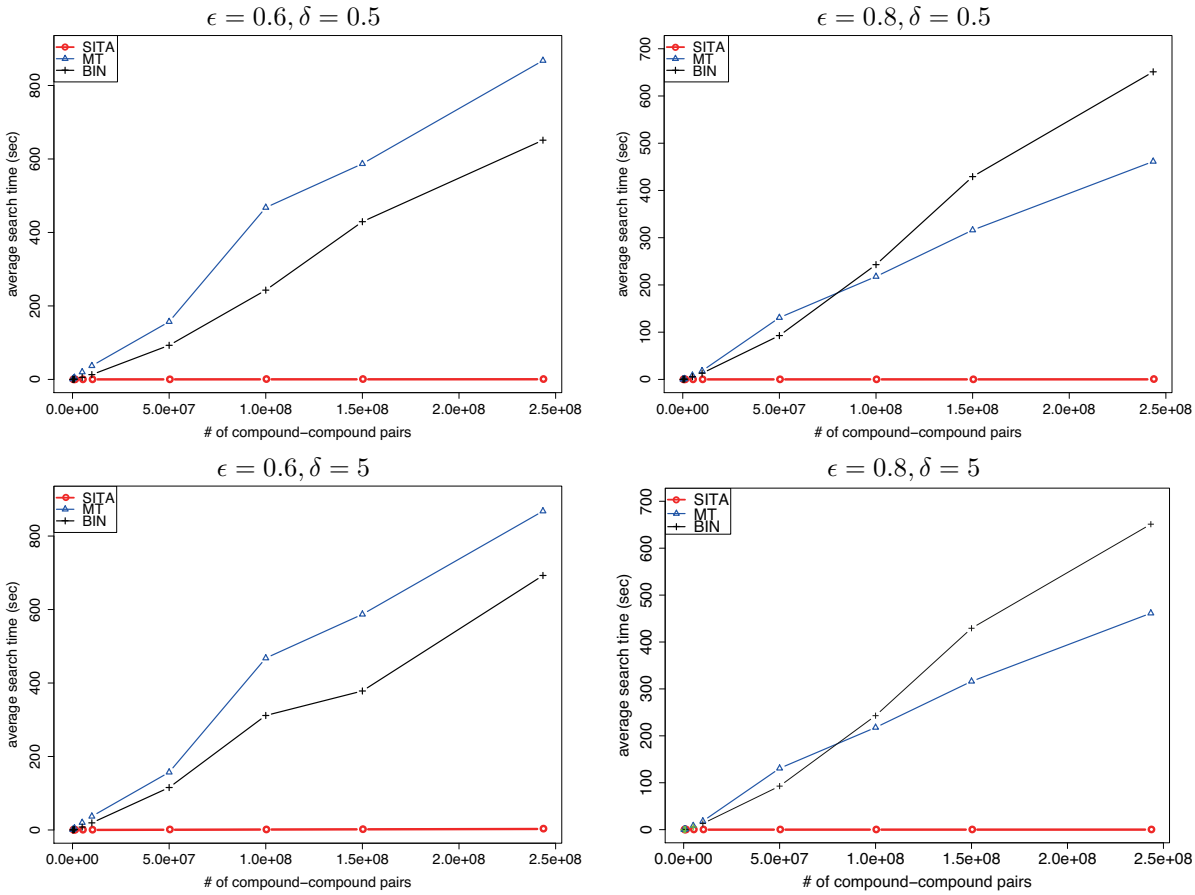

Figure 1: Average search time for the number of compound-compound pairs in seconds

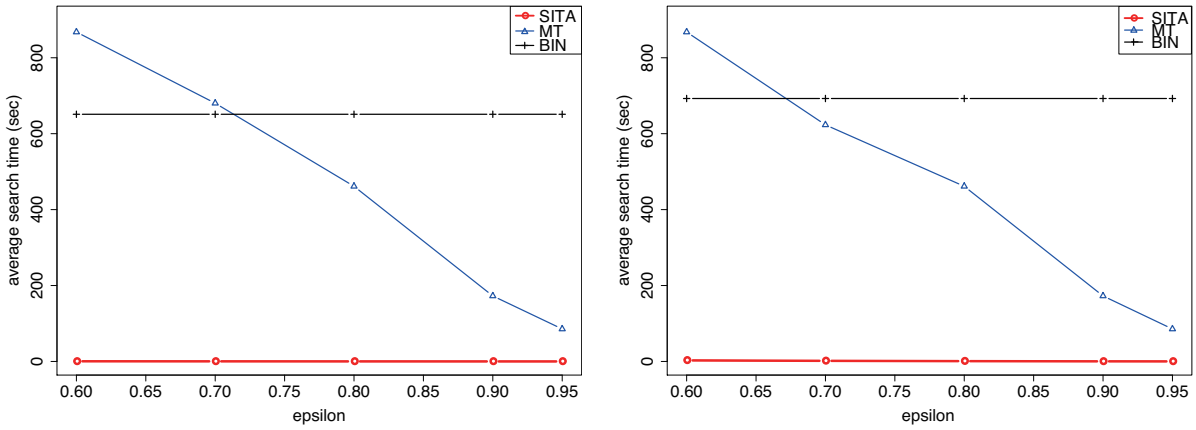

Figure 2: Search time for various epsilons by fixing $\delta = 0.5$ (left) and $\delta = 5$ (right)

Table 1: Performance summary on 243,438,006 compound-compound pairs

	SITA ($\epsilon = 0.6, \delta = 0.5$)	SITA ($\epsilon = 0.8, \delta = 0.5$)	MT ($\epsilon = 0.6$)	MT ($\epsilon = 0.8$)	BIN ($\delta = 0.5$)
ave. search time (sec)	0.41 ± 0.26	0.12 ± 0.08	867.80 ± 482.02	461.38 ± 284.20	651.15 ± 196.20
$\lvert P_1 \rvert$	$142,387,920$	$70,133,488$			
$P_1 \cap [i_l, i_r]$	$325,922$	$161,772$			
#Rank	$9,996,452$	$2,866,385$			
$\lvert P_{NC} \rvert$	760	34			

	SITA ($\epsilon = 0.6, \delta = 5$)	SITA ($\epsilon = 0.8, \delta = 5$)	MT ($\epsilon = 0.6$)	MT ($\epsilon = 0.8$)	BIN ($\delta = 5$)
ave. search time (sec)	2.75 ± 2.29	0.80 ± 0.71	867.80 ± 482.02	461.38 ± 284.20	692.66 ± 200.297
$\lvert P_1 \rvert$	$142,387,920$	$70,133,488$			
$P_1 \cap [i_l, i_r]$	$384,670$	$1,580,351$			
#Rank	$3,184,670$	$1,580,351$			
$\lvert P_{NC} \rvert$	$6,397$	206			

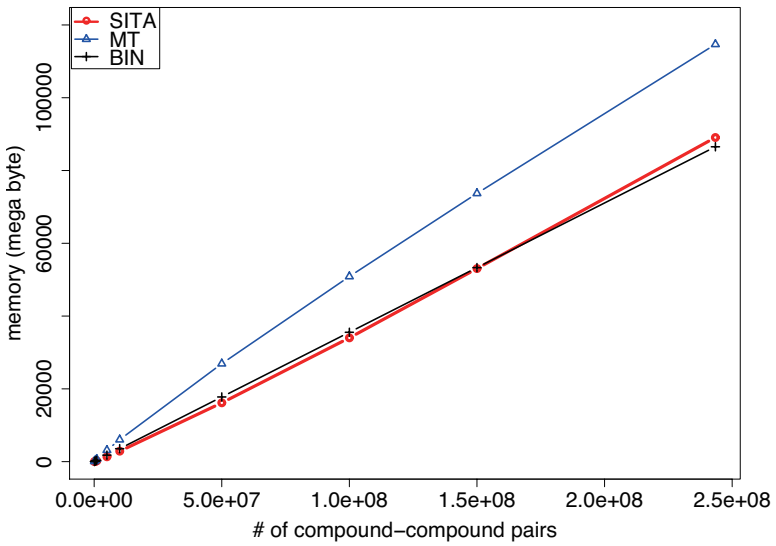

Figure 3: Memory usage for each method