COMPUTATIONAL BIOPHYSICS COURSE (CS61060) TERM PROJECT 4

Project title: Fast multiple similar genetic sequence alignment based on the centre star strategy

Summary: Multiple sequence alignment (MSA) is important work, but bottlenecks arise in the massive MSA of homologous DNA or genome sequences. Try to implement trie trees to accelerate the centre star MSA strategy. The expected time complexity will be decreased to linear time from square time. The algorithm for centre star strategy is also discussed below.

Algorithm 1. Improved Centre Star Algorithm Based on Trie Trees

Input: n DNA Sequences, $S_1, S_2, \ldots S_n$
Output: n aligned DNA Sequences $S'_1, S'_2, \ldots, S'_n$
1. **For** each DNA Sequence, S_i,
2. Partition S_i into k segments $\{S_{i1}, S_{i2}, \ldots S_{ik}\}$ with equal lengths;
3. Construct trie tree $\mathcal{T}_i$ for the segments set $S_i = \{S_{i1}, S_{i2}, \ldots S_{ik}\}$
4. **for** j from 1 to n, $j \neq i$
5. search $\mathcal{T}_i$ in S_j, and set m_{ij} as the segment appearance times; record all of the appearances in $\mathcal{A}_{ij}$
6. end **for**
7. calculate $m_i = \sum_{j=1, j\neq i}^{n} m_{ij}$
8. end **For**
9. $m^* = argmax_{i=1,2,\ldots,n} m_i$, set S_{m^*} as the centre star sequence
10. **For** each i from 1 to n, $i \neq m^*$
11. Partition S_i, and S_{m^*} according $\mathcal{A}_{im^*}$, align the mismatched regions and obtain the pairwise alignment; record all of the positions of inserted spaces in $\mathcal{P}_{im^*}$ and $\mathcal{P}_{m^*i}$.
12. end **For**
13. **For** i from 1 to n, $i \neq m^*$
14. sum $\mathcal{P}_{m^*i}$ to $\mathcal{P}_{m^*}$
15. end **For**
16. obtain the final result, S'_{m^*}, according to $\mathcal{P}_{m^*}$
17. **For** i from 1 to n, $i \neq m^*$
18. compare $\mathcal{P}_{m^*i}$ with $\mathcal{P}_m$, and update $\mathcal{P}_{im^*}$, then obtain the final result, S'_i
19. end **For**

Datasets: You can download any of the online available datasets of genetic sequences. You can also download the RNA datasets from the google drive link given below.

https://drive.google.com/drive/folders/12Ld7wVSdsWAIvKFduRTaJd_D92C_L7Ul?usp=sharing