A Benchmark Study of Multiple Sequence Alignment Methods

Ellen Nie
Hang Yu
CS 466

Introduction

- Multiple alignment of protein sequences has become a fundamental tool in molecular biology
  - evolutionary studies
  - protein structure/function
  - inter-molecular interactions
- Computing the optimal multiple sequence alignment is a **NP-complete** problem.
- A number of approximation algorithms were developed as an alternative.

Review of MSA techniques

- **Progressive alignment:** compute an alignment from the "bottom-up" on the basis of a guide tree
- **Iterative alignment:** refines an initial alignment by iteratively dividing the alignment into two profiles and realigning them.
- **Divide and Conquer:** divides the sequence dataset into subsets, which are aligned and then merged together.
- **Consistency:** uses a set of alignments in order to inform the alignment.
- Estimation of the alignment under a **statistical model**: uses sequence profiles, profile HMMs to represent multiple sequence alignments.

Representative MSA tools

- Progressive Method
  - Consistency-based: **T-Coffee (2000)**
- Iterative Method:
  - Divide-and-conquer: **PASTA (2015)**
  - Matrix-based: **MUSCLE (2004)**
- Hidden Markov Model: **CLUSTA-OMEGA (2011)**

Summary of MSA tools

MSA Technique	Algorithm Name	Version	Download Link
Progressive - Consistency	T-Coffee	11.00.8cbe486	http://www.tcoffee.org/
Iterative - Matrix	MUSCLE	v3.8.31	http://www.drive5.com/muscle
Iterative - Divide and Conquer	PASTA	--	https://github.com/smirarab/pasta.git
Hidden Markov Model	Clustal-Omega	1.2.4	http://www.clustal.org/omega/

Protein Database

- BAliBASE4: a database of simulated protein sequences specifically developed for MSA methods assessments

- Pick 6 datasets out of 10.

  - RV1: cases with small numbers of equidistant sequences
  - RV2: families with one or more “orphan” sequences;
  - RV3: a pair of divergent subfamilies, with less than 25% identity between the two groups;
  - RV5: sequences with large internal insertions and deletions.
  - RV9: protein families with linear motifs often found in disordered regions
  - RV10: large, complex protein families, designed to reproduce today's sequence exploration requirement

Assessment Procedure

- Accuracy
  - Sum-of-Pairs (SP) score: the sum of all pair-wise induced alignment scores.
  - Total Column (TC) score: the number of columns that are identical (including gaps) in the two alignments
- Error
  - SPFN rate: the fraction of true pairs that are not recognized in the alignment
  - SPFP rate: the fraction of recognized pairs that are not true pairs
- Efficiency
  - Average time to finish each dataset

FastSP: Alignment Comparison

- an open-source Java program that can be used to compute these metrics.

"java -jar FastSP.jar -r reference_alignment_file -e estimated_alignment_file"

- Available for download from: https://github.com/smirarab/FastSP.git.

References

1. L. Wang and T. Jiang. “On the complexity of multiple sequence alignment”. *Journal of Computational Biology*, 1:337–348, 1994.

2. Notredame, C., Higgins, D.G., and Heringa, J. (2000). *“*T-Coffee: a novel method for fast and accurate multiple sequence alignment*”*. *Journal of Molecular Biology*, 302:205–217.

3. Berger, M.P., and Munson, P.J. (1991). “A novel randomized iterative strategy for aligning multiple protein sequences*”*. *CABIOS*, 7:479-484.

4. K. Liu, S. Raghavan, S. Nelesen, C. R. Linder, T. Warnow (2009). "Rapid and Accurate Large-Scale Coestimation of Sequence Alignments and Phylogenetic Trees". *Science*, 324:1561-1564.

5. K. Liu, T. J. Warnow, M.T. Holder, et al (2012). “SATé-II: Very Fast and Accurate Simultaneous Estimation of Multiple Sequence Alignments and Phylogenetic Trees”. *Syst Biol* : 61 (1): 90-106. doi: 10.1093/sysbio/syr095

References

6. Bahr, A., Thompson, J. D., Thierry, J.-C., & Poch, O. (2001). BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. Nucleic Acids Research, 29(1), 323–326.

6. Robert C. Edgar (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res,* 32 (5): 1792-1797. doi: 10.1093/nar/gkh340

7. Robert C. Edgar (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004;5:113. doi: 10.1186/1471-2105-5-113.

8. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karpuls, K., Li, W., … Lopez, R. (2011, November 10). Fast, scalable generation of high‐quality protein multiple sequence alignments using Clustal Omega | Molecular Systems Biology. Retrieved from http://msb.embopress.org/content/7/1/539