RetAlign

An efficient solution for MSA using alignment networks

Adrienn Szabó
Phd student of
Eötvös University, Budapest (ELTE)
and
DMS Group
Institute for Computer Science and Control,
Hungarian Acedemy of Sciences

June 30, 2014

Table Of Contents

About me

Education

- MSc: Software engineer, Budapest University of Technology and Economics (2008)
- PhD: Data mining techniques on biological data (supervisors: *András Benczúr, István Miklós*), Eötvös University, Budapest (ongoing)

About me

Research interests

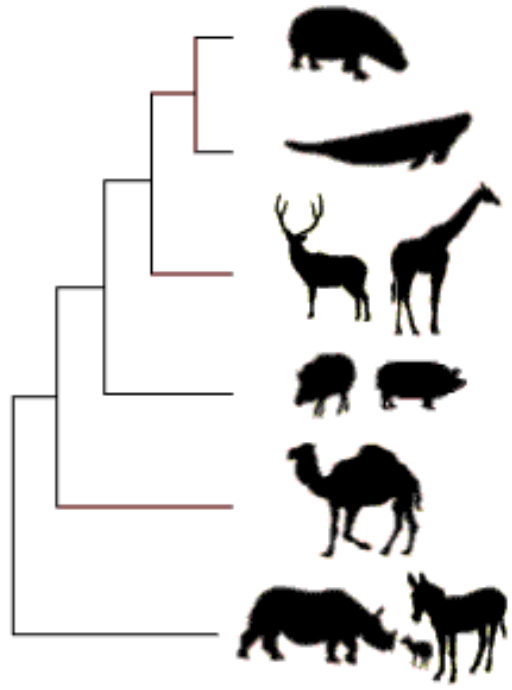
- Bioinformatics, especially multiple sequence alignment, and problems with a lot of data
- Data mining, machine learning, text mining, especially on biological datasets

Work

- Developer and research assistant at Data Mining and Search Group (head: András Benczúr), MTA SZTAKI (2007 -)
- Software engineer intern at Google Zürich (2009)

MSA – Introduction

- Multiple sequence alignment (MSA): alignment of three or more biological sequences
- Needed for phylogenetic analysis, function prediction of proteins, etc.



```
                 :   :      **  *                  : .   :    :: *. *:*   :
1mnmC      TKPYRGHR-FTKENVRILESWFAKNIENPYLDTKGLENLMKNTSLSRIQIKNWVSNRR---RKEKTITIAPEL
1au7A      RKRKR-RTTISIAAKDALERHFG---EHSKPSSQEIMRMAEELNLEKEVVRVWFCNRRQREKRVKT-SLNQSL
1akhA      KSPKG-KSSISPQARAFLEEVFR---RKQSLNSKEKEEVAKKCGITPLQVRVWFINKRMRSK----------
1fjlA      KQRRS-RTTFSASQLDELERAFE---RTQYPDIYTREELAQRTNLTEARIQVWFQNRRARLRKQ----HTSVS
1ftt       MRRKR-RVLFSQAQVYELERRFK---QQKYLSAPEREHLASMIHLTPTQVKIWFQNHRYKMKRQAK-DKAAQQ
1ftz       MDSKRTRQTYTRYQTLELEKEFH---FNRYITRRRRIDIANALSLSERQIKIWFQNRRMKSKKDRTLDSSPEH
```

Basics – pairwise sequence alignment

- The standard edit distance based formulation of sequence alignment leads to $\mathcal{O}(L^2)$
- Dynamic programming: *Smith-Waterman* and *Needleman-Wunsch* algorithms

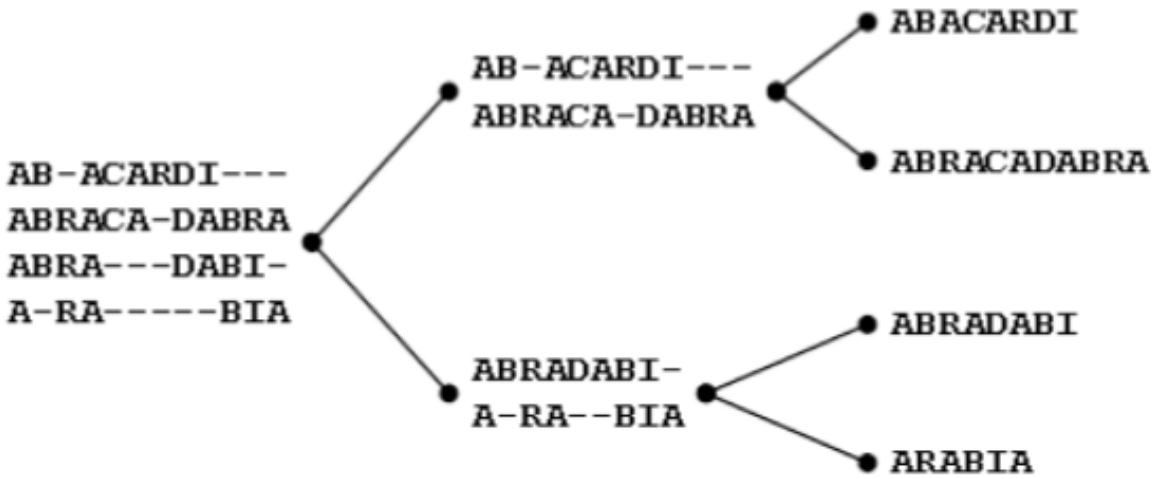
```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                    ****: .***:  * *:** * :*****.:* ********..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:***********:***:**.: .****************    :  *.: :
```

Problems with multiple sequence alignment

- For straightforward dynamic programming solutions, each additional sequence multiplies the time and memory required
- Finding the optimal alignment is NP-complete
- Corner-cutting methods shrink the search space, but are still exponential in memory and running time
- Heuristics applied: *progressive alignment*

Progressive alignment

- A guide tree is used, and pairwise alignments at each inner node
- Polynommial running time
- Once a gap has been inserted it can not be removed



RetAlign - main idea

- Store a set of optimal and suboptimal alignments at each step of the progressive alignment procedure
- Propagate the partial networks at each inner node of a guide tree upwards
- Essentially we are extending the *Waterman-Byers* algorithm to align a network of alignments to another network of alignments

RetAlign - data structure

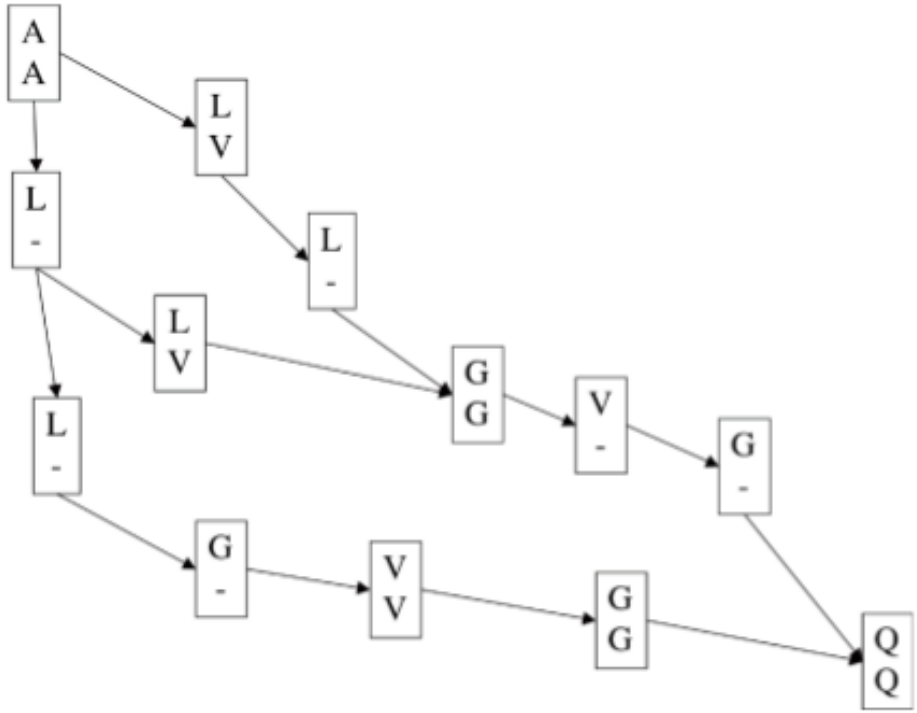
We used a special data structure:

***x*-network:** a set of alignment paths that contain the optimal pairwise alignment and all suboptimal paths that have an aligment score above the optimal score minus x

Note: this is a DAG

RetAlign - data structure

This network shows three different alignments of the sequences ALLGVGQ and AVGQ:



Outline of the RetAlign algorithm

1. Build or load a guide tree for the sequences
2. Bottom-up, for each node v of the tree:
   - calculate the x_v-network of its children's sub-networks using the generalized Waterman-Byers algorithm
3. Return the best scored alignment from the x-network calculated at the root of the guide tree

Measuring performance

- Tested and evaluated on the *BAliBASE* datasets, that contain more than 6000 sequences
- Compared with the most widely used MSA packages: ClustalW, MAFFT and FSA

Accuracy comparison

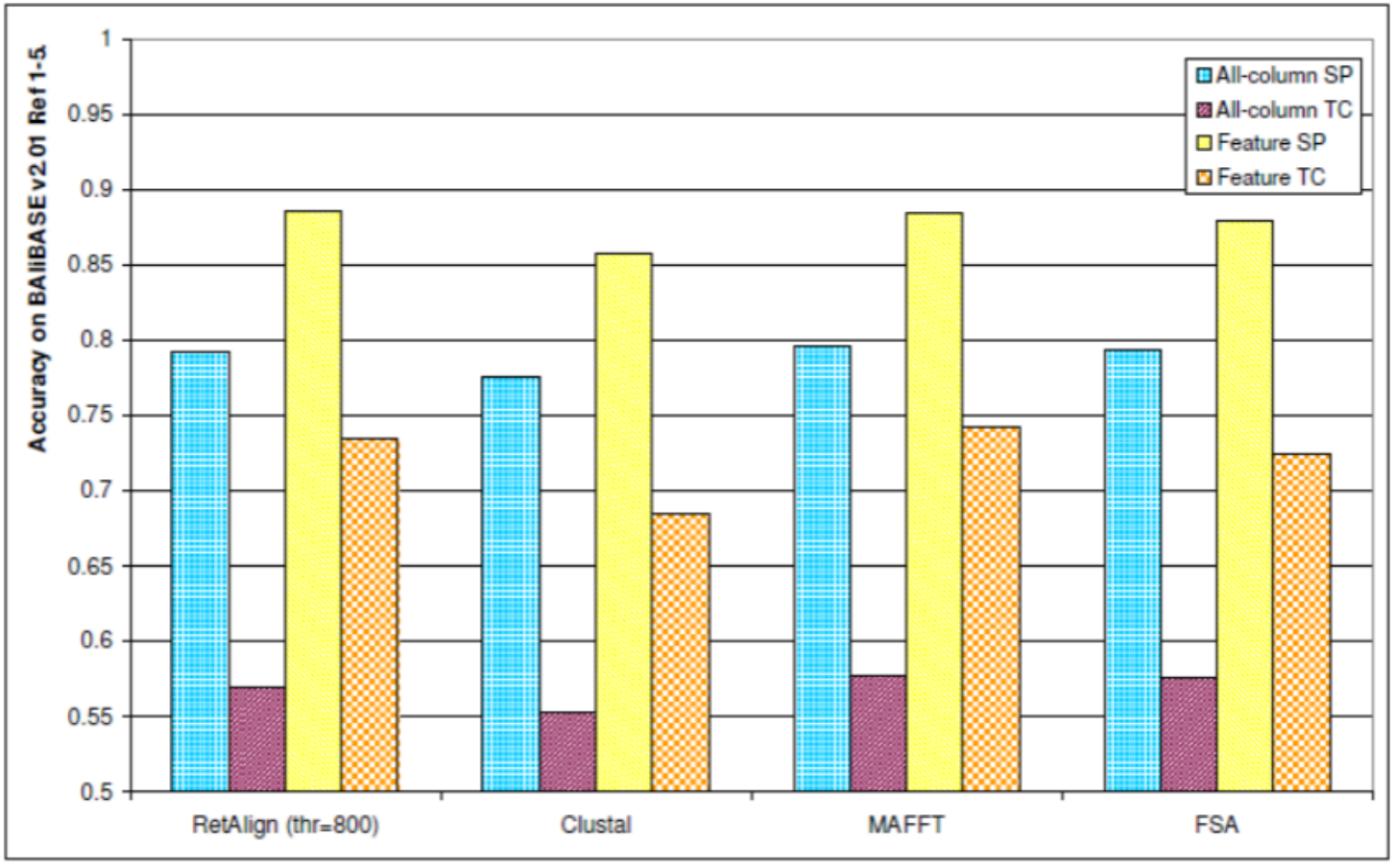


Figure : Comparison of alignment software on BAliBASE v2.01 Refs [1-5]. Alignment accuracy of multiple alignment programs compared to that of RetAlign as measured on BAliBASE v2.01 Reference sets [1-5] using the provided bali_score tool (SP and TC scores calculated on all of the columns versus on columns containing features are all shown). RetAlign was run with sequence weighting on, a single guide tree iteration and with a reticular threshold of 800. FSA was run in maximum sensitivity mode. MAFFT was run with the -auto switch and ClustalW with the default settings.

Current and future work

Working on a sequel paper: how to build up an alignment network from multiple separete MSA alignments?

- different input parameters for the underlying MSA algorithm
- sampling
- measure performance

References and sources

- Publication:
  Adrienn Szabó, Ádám Novák, István Miklós, Jotun Hein: *Reticular alignment: A progressive corner-cutting method for multiple sequence alignment*, BMC Bioinformatics, 2010
- References:
  - `http://en.wikipedia.org/wiki/Sequence_alignment`
  - `http://en.wikipedia.org/wiki/Multiple_sequence_alignment`
- Sources of pictures:
  - `http://upload.wikimedia.org/wikipedia/commons/8/86/Zinc-finger-seq-alignment2.png`
  - `http://cnx.org/content/m15807/latest/`

Questions?