Assignment 4—Grundlagen der Bioinformatik

Anastasia Grekova 4149666 and Huajie Chen 4199962

May 13, 2018

1 Profile alignment

match - score : 2

mismatch - score : -2

gap penalty : 2

Vereinfachte Version

	-	V	C	L	W	C
	-	V	C	M	F	C
- - -	0	-12	-24	-36	-48	-60
I I I	-12	-12	-24	-36	-48	-60
E E E	-24	-24	-24	-36	-48	-60
C C C	-36	-36	-12	-24	-36	-36
I M I	-48	-48	-24	-20	-32	-44
E Q E	-60	-60	-36	-32	-32	-44
C C C	-72	-72	-48	-44	-44	-20

2 Three different multiple sequence approaches

2.1 Clustal Omega

It is accurate but also allows alignments of almost any size to be produced. In benchmark tests, it is distinctly more accurate than most widely used, fast methods and comparable in accuracy to some of the intensive slow methods. With Clustal Omega, we use a modified version of mBed (Blackshields et al, 2010), which has complexity of O(N log N), and which produces guide trees that are just as accurate as those from conventional methods.

2.2 Kalign

The Kalign algorithm follows a strategy analogous to the standard progressive method for sequence alignment. Pairwise distances are calculated, a guide tree is constructed and sequences/profiles are aligned in the order given by the tree. In contrast to existing methods, the Wu-Manber approximate string-matching algorithm is used in the distance calculation and optionally in the dynamic programming used to align the profiles.

2.3 MAFFT

A multiple sequence alignment program, MAFFT, has been developed. The CPU time is drastically reduced as compared with existing methods. MAFFT includes two novel techniques. Homologous regions are rapidly identified by the fast Fourier transform (FFT), in which an amino acid sequence is converted to a sequence composed of volume and polarity values of each amino acid residue. We

propose a simplified scoring system that performs well for reducing CPU time and increasing the accuracy of alignments even for sequences having large insertions or extensions as well as distantly related sequences of similar length. Two different heuristics, the progressive method (FFT-NS-2) and the iterative refinement method (FFT-NS-i), are implemented in MAFFT.

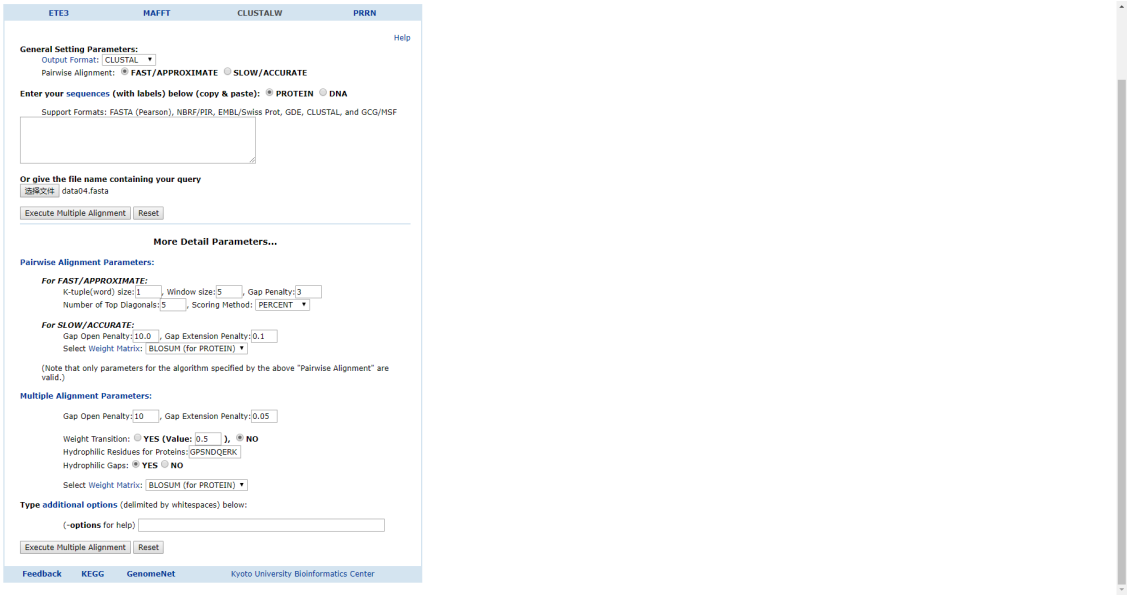
3 Application of a multiple alignment tool

The program "ClusterW" on http://www.genome.jp/tools-bin/clustalw was used and hereunder is the generated results.

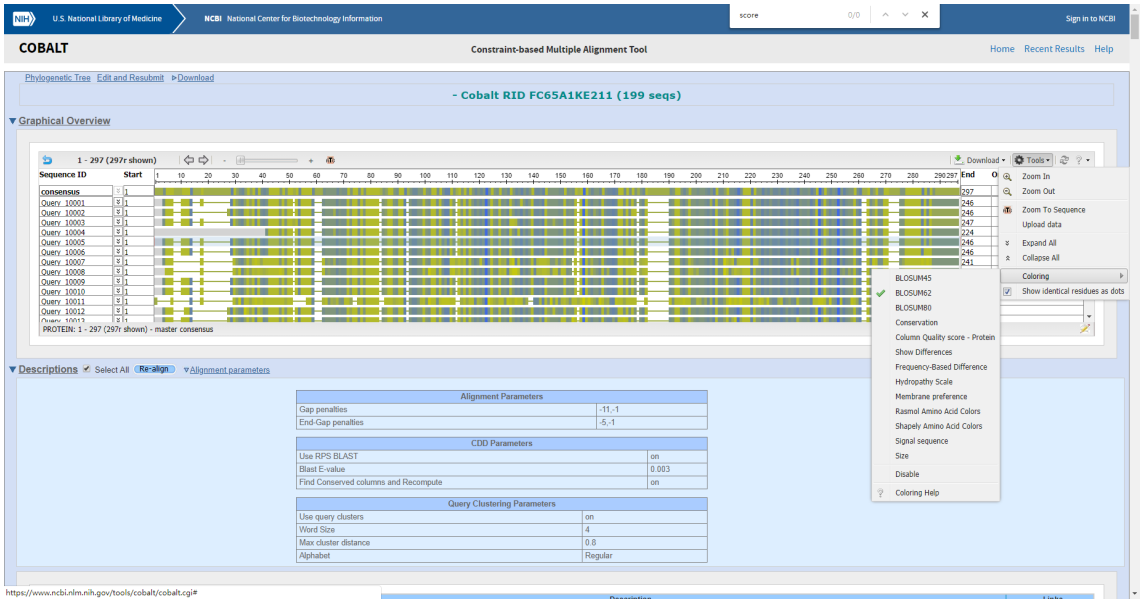
```
Group 159: Sequences:   4      Score:3685
Group 160: Sequences:   9      Score:3531
Group 161: Sequences:  10      Score:3700
Group 162: Sequences:  11      Score:3618
Group 163: Sequences:  16      Score:3557
Group 164: Sequences:   2      Score:3944
Group 165: Sequences:  18      Score:3644
Group 166: Sequences: 100      Score:3357
Group 167: Sequences: 145      Score:3436
Group 168: Sequences:   2      Score:4061
Group 169: Sequences:   3      Score:4049
Group 170: Sequences:   4      Score:4020
Group 171: Sequences:   2      Score:4082
Group 172: Sequences:   6      Score:3991
Group 173: Sequences: 151      Score:3727
Group 174: Sequences:   2      Score:3982
Group 175: Sequences:   3      Score:3882
Group 176: Sequences:   4      Score:3816
Group 177: Sequences: 155      Score:3473
Group 178: Sequences:   2      Score:3981
Group 179: Sequences:   3      Score:3780
Group 180: Sequences:   2      Score:3959
Group 181: Sequences:   3      Score:3926
Group 182: Sequences:   6      Score:3728
Group 183: Sequences:   7      Score:3602
Group 184: Sequences: 162      Score:3308
Group 185: Sequences:   2      Score:4057
Group 186: Sequences:   3      Score:4041
Group 187: Sequences:   4      Score:3930
Group 188: Sequences: 166      Score:3552
Group 189: Sequences: 167      Score:3456
Group 190: Sequences: 185      Score:1747
Group 191: Sequences:   2      Score:3698
Group 192: Sequences:   3      Score:1866
Group 193: Sequences:   2      Score:1865
Group 194: Sequences:   5      Score:1655
Group 195: Sequences: 190      Score:1577
Group 196: Sequences: 197      Score:1288
Group 197:                     Delayed
Group 198:                     Delayed
Alignment Score 19361687

CLUSTAL-Alignment file created  [clustalw.aln]
```

The alignment score is 19361687.
The parameters lie in this image hereunder.



Since this software did not provide us with the color scheme, another software from NCBI called "COBALT" was used. The specific parameters are in this image and the applied color scheme is Blosum62.


U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
Sign in to NCBI
COBALT
Constraint-based Multiple Alignment Tool
Home Recent Results Help
Phylogenetic Tree Edit and Resubmit Download
- Cobalt RID FC65A1KE211 (199 seqs)
Graphical Overview
1 - 297 (297r shown)
Download
Tools
Sequence ID
Start
End
consensus
Query_10001
Query_10002
Query_10003
Query_10004
Query_10005
Query_10006
Query_10007
Query_10008
Query_10009
Query_10010
Query_10011
Query_10012
PROTEIN: 1 - 297 (297r shown) - master consensus
Zoom In
Zoom Out
Zoom To Sequence
Upload data
Expand All
Collapse All
Coloring
Show identical residues as dots
BLOSUM45
BLOSUM62
BLOSUM80
Conservation
Column Quality score - Protein
Show Differences
Frequency-Based Difference
Hydropathy Scale
Membrane preference
Rasmol Amino Acid Colors
Shapely Amino Acid Colors
Signal sequence
Size
Disable
Coloring Help
Descriptions
Select All
Re-align
Alignment parameters
Alignment Parameters
Gap penalties
-11,-1
End-Gap penalties
-5,-1
CDD Parameters
Use RPS BLAST
on
Blast E-value
0.003
Find Conserved columns and Recompute
on
Query Clustering Parameters
Use query clusters
on
Word Size
4
Max cluster distance
0.8
Alphabet
Regular
https://www.ncbi.nlm.nih.gov/tools/cobalt/cobalt.cgi#