

Contents

Part I Background

1 Multiple Sequence Assignments: An Introduction	3
1.1 Introduction	3
1.2 Areas of Application of MSAs	4
1.2.1 Preserved Sequence Sections: Motifs and Domains	4
1.2.2 Prediction of Function and Structure	5
1.2.3 Phylogeny	5
1.2.4 MSAs as Everyday Tools	5
1.3 Representation Formats of MSAs	6
1.3.1 FASTA	8
1.3.2 Clustal	10
1.3.3 MSF	10
1.3.4 PHYLIP	11
1.3.5 NEXUS	12
1.3.6 Tree Formats	12
1.3.7 Graphical Visualizations	13
2 How Do MSA Programs Work?	17
2.1 Introduction	17
2.2 Pairwise Sequence Alignments	17
2.2.1 A Naive Method	17
2.2.2 Dynamic Programming	18
2.2.3 Gaps and the Similarity Matrix	20
2.2.4 Global and Local Alignments	23
2.3 Multiple Sequence Alignments	25
2.3.1 The Central Problem	25
2.3.2 Solution 1: The Progressive Method	25
2.3.3 Solution 2: The Iterative Method	26
2.3.4 Solution 3: The Consistency-Based Method	26
2.3.5 Solution 4: The Probabilistic Method	27
2.3.6 Solution 5: The Meta or Ensemble Method	28
2.3.7 Methods of the Future	29
2.3.8 Special Cases Require Special Methods	30

2.4	Further Topics	31
2.4.1	Structure-Based MSAs	31
2.4.2	BLAST and Co.	32
2.4.3	Alignment-Free Methods	33
2.4.4	Genome and Chromosome Alignments	34
3	Overview of Current MSA Programs	35
3.1	Introduction	35
3.2	List of Common MSA Programs	36
3.2.1	DFalign	36
3.2.2	Clustal	36
3.2.3	ClustalW	36
3.2.4	SAGA	37
3.2.5	PRRP	38
3.2.6	DIALIGN	38
3.2.7	DIALIGN2	39
3.2.8	T-Coffee	39
3.2.9	MAFFT	40
3.2.10	POA	41
3.2.11	PRALINE	41
3.2.12	Align-M	42
3.2.13	MUSCLE	42
3.2.14	3D-Coffee	43
3.2.15	Kalign	43
3.2.16	DIALIGN-T	44
3.2.17	ProbCons	45
3.2.18	M-Coffee	45
3.2.19	R-Coffee and RM-Coffee	45
3.2.20	DIALIGN-TX	46
3.2.21	PRALINE™	46
3.2.22	PRANK	47
3.2.23	PSI-Coffee and TM-Coffee	47
3.2.24	MSAProbs	48
3.2.25	PicXAA	48
3.2.26	AlignMe	49
3.2.27	Clustal Omega	49
3.2.28	ReformAlign	50
3.2.29	DECIPHER	50

Part II Which Program Fits My Data?

4	Details of the Analysis	55
4.1	Benchmarking	55
4.1.1	A Marathon for MSA Programs	55
4.1.2	The Crux of MSA Benchmarking	56

4.2	Benchmark Datasets Used (Table 4.1)	57
4.2.1	BAlIbASE	57
4.2.2	BRAlIbASE	60
4.2.3	Bench	61
4.2.4	Artificial Benchmarks: Rose, IRMBASE and DIRMBASE	63
4.2.5	HOMSTRAD	64
4.3	Scores	65
4.3.1	Sum-of-Pairs	65
4.3.2	The Column Score	66
4.3.3	Modeler Score	66
4.3.4	Cline’s Shift Score	67
4.3.5	Transitive Consistency Score	67
5	Decision Aid	71
5.1	Introduction	71
5.2	Implementation	72
5.3	Results	72
5.3.1	RNA	72
5.3.2	DNA	76
5.3.3	Proteins	78
	Glossary	89
	Literature	95
	Index	101