Multiple Sequence Alignment Based on Profile Alignment of Intermediate Sequences

Yue Lu and Sing-Hoi Sze

RECOMB 2007

Presented by: Wanxing Xu

March 6, 2008

Content

- Biology Motivation
- Computation Problem
- Algorithm
- Performance

Content

- Biology Motivation
- Computation Problem
- Algorithm
- Performance

Biology Motivation

- Multiple Sequence Alignment:
  - Assess sequence conservation of protein domains, tertiary and secondary structures and even individual amino acids or nucleotides.
  - Evolutionary relationships or sequence conservation among homologous.
  - Simultaneously compare several sequences.

Content

- Biology Motivation
- Computation Problem
- Algorithm
- Performance

Content

- Biology Motivation
- Computation Problem
- Algorithm
- Performance

Computation Problem

- Methods:
  - Pairwise alignments
  - Prograssive alignment construction
  - Iterative methods
  - Hidden Markov models
- Problems:
  - Accuracy
  - Computational complexity

Content

- Biology Motivation
- Computation Problem
- Algorithm
- Performance

Content

- Biology Motivation
- Computation Problem
- Algorithm
- Performance

Algorithm-Introduction

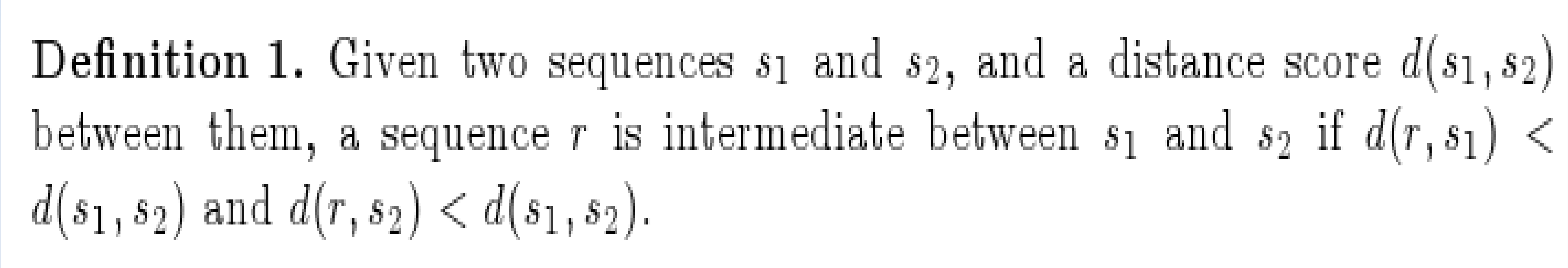
- Incorporate additional hits into the input sequences
  - Hits that are not intermediate will introduce noise
  - Use carefully defined intermediate sequences
- Align profiles instead of the sequences
  - Construct a profile for each sequence
  - Align the profiles by modifying the pair-HMM
  - Obtain a secondary structure prediction

Algorithm

- Finding intermediate sequences
- Choosing intermediate sequences
- Constructing sequence profiles
- Alignment via modified pair-HMM

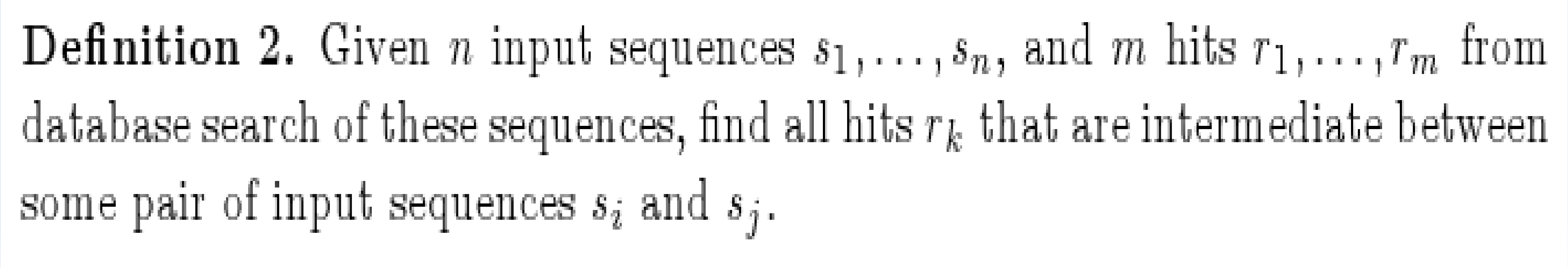
Finding Intermediate Sequence

- Definitions of Intermediate Sequence
  - Between two input sequences:



Definition 1. Given two sequences s_1 and s_2, and a distance score $d(s_1, s_2)$ between them, a sequence r is intermediate between s_1 and s_2 if $d(r, s_1) < d(s_1, s_2)$ and $d(r, s_2) < d(s_1, s_2)$.

  - Between multiple sequences:



Definition 2. Given n input sequences $s_1, \ldots, s_n$, and m hits $r_1, \ldots, r_m$ from database search of these sequences, find all hits r_k that are intermediate between some pair of input sequences s_i and s_j.

Finding Intermediate Sequence

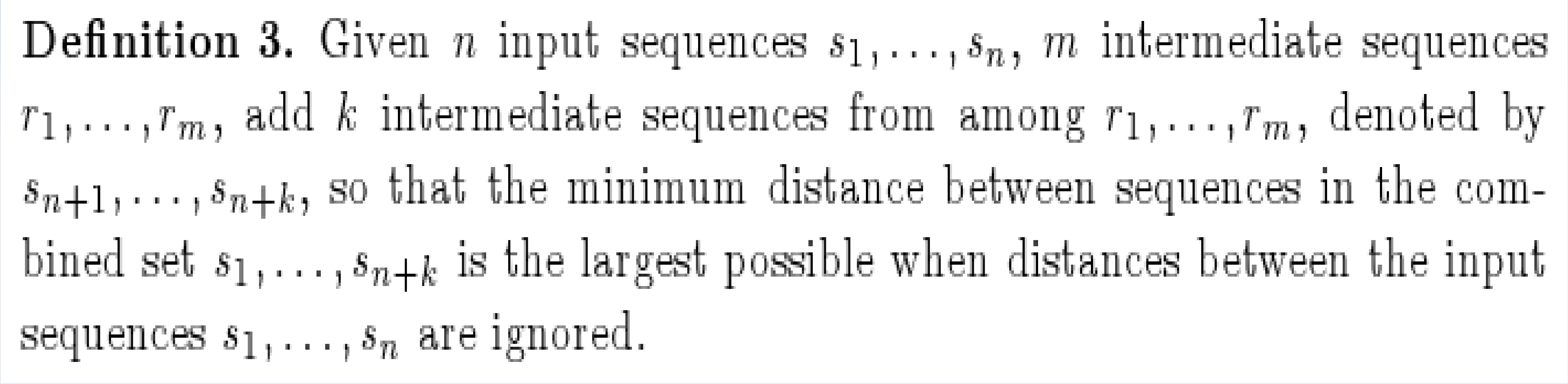
- No need to compute pairwise distances between the potentially very large number of hits.
- The number of pairwise distance score computations: $O(mn+n^2)$
- The number of score comparisons is $O(mn^2)$.

Choosing Intermediate Sequences

- The number of intermediate sequences can be very large
- Use a subset of intermediate sequences
- Similar sequences are likely to contain redundant information
- Choose a small subset of intermediate sequences using a greedy strategy
- Goal: identify a combined set of sequences as divergent as possible

Choosing Intermediate Sequences

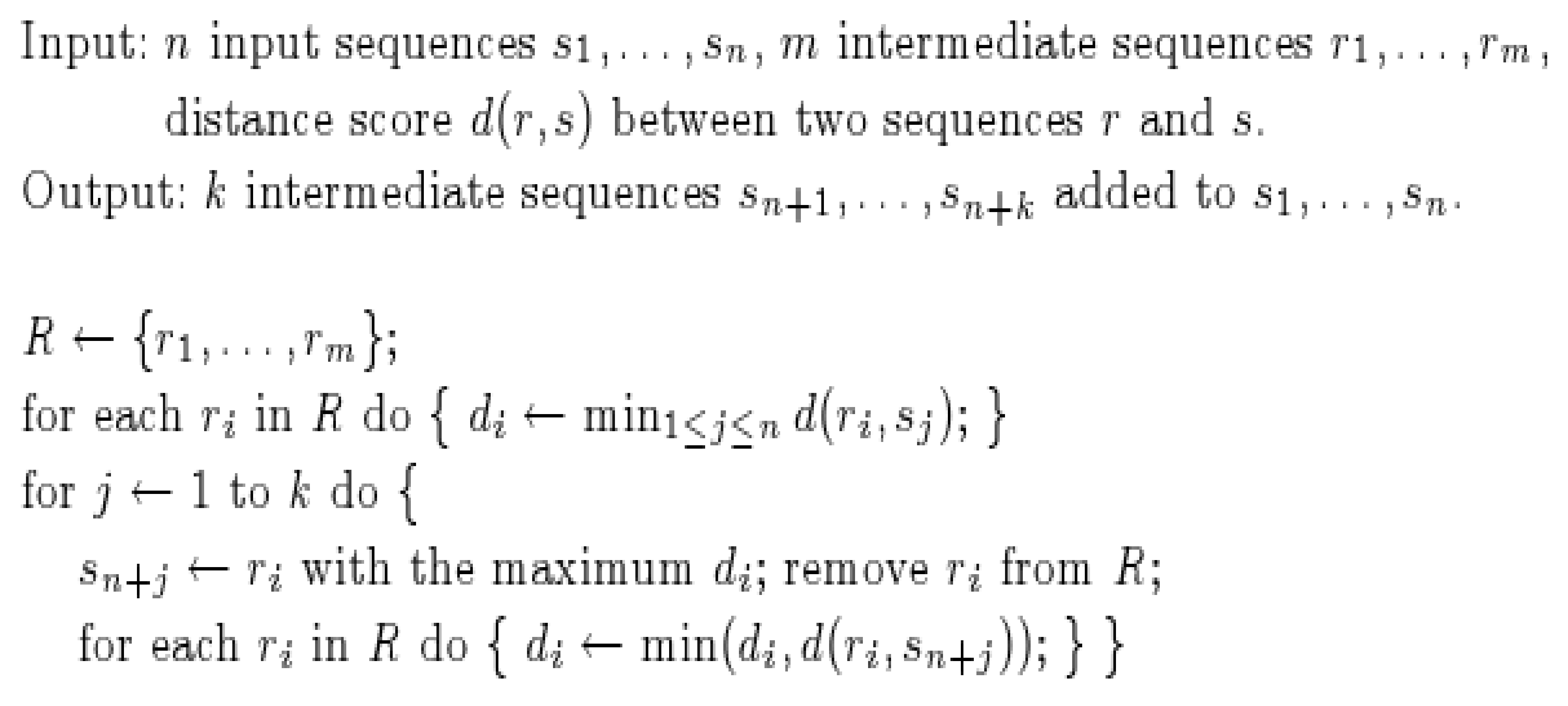
- Definition



Definition 3. Given n input sequences $s_1, \ldots, s_n$, m intermediate sequences $r_1, \ldots, r_m$, add k intermediate sequences from among $r_1, \ldots, r_m$, denoted by $s_{n+1}, \ldots, s_{n+k}$, so that the minimum distance between sequences in the combined set $s_1, \ldots, s_{n+k}$ is the largest possible when distances between the input sequences $s_1, \ldots, s_n$ are ignored.

Choosing Intermediate Sequences

- Greedy algorithm



Input: n input sequences $s_1, \ldots, s_n$, m intermediate sequences $r_1, \ldots, r_m$,
distance score $d(r, s)$ between two sequences r and s.

Output: k intermediate sequences $s_{n+1}, \ldots, s_{n+k}$ added to $s_1, \ldots, s_n$.

```
$R \leftarrow \{r_1, \ldots, r_m\}$;
for each $r_i$ in $R$ do { $d_i \leftarrow \min_{1 \le j \le n} d(r_i, s_j)$; }
for $j \leftarrow 1$ to $k$ do {
    $s_{n+j} \leftarrow r_i$ with the maximum $d_i$; remove $r_i$ from $R$;
    for each $r_i$ in $R$ do { $d_i \leftarrow \min(d_i, d(r_i, s_{n+j}))$; } }
```

Choosing Intermediate Sequences

- Iteratively add the farthest intermediate sequence.
- Does not guarantee optimum divergence, but still reasonable.
- The number of pairwise score computations is $O(m(n+k))$.

Constructing Sequence Profiles

- Assign each intermediate sequence r_i (i=1..m) to the most similar sequence s_j (j=1..$n+k$).
- Use star alignment for each sequence s_j and the intermediate sequence assigned to it.
- The relative frequency of each residue of s_j is used to construct a profile as a probability distribution.

Constructing Sequence Profiles

- If the number of very closely related sequences assigned to s_j is very large, It will have over-contribution.
- Solution: before choosing intermediate sequences, remove sequences from the original set so that none of the remaining sequences are very similar to each other.

Modified Pair-HMM

- Original model:

X	x	x	-
Y	y	-	y
Probability	e(x,y)	e(x)	e(y)

- δ: the gap opening probability
- ε: the gap extension probability

Modified Pair-HMM

- Add the probability distribution of residues at each position:
  - $p_1(x,i)$: residue x at position i in X.
  - $p_2(y,j)$: residue y at position j in Y.
- New emission probability of state M:

$$e'(i,j)=\sum_x \sum_y p_1(x,i)p_2(y,j)e(x,y)$$

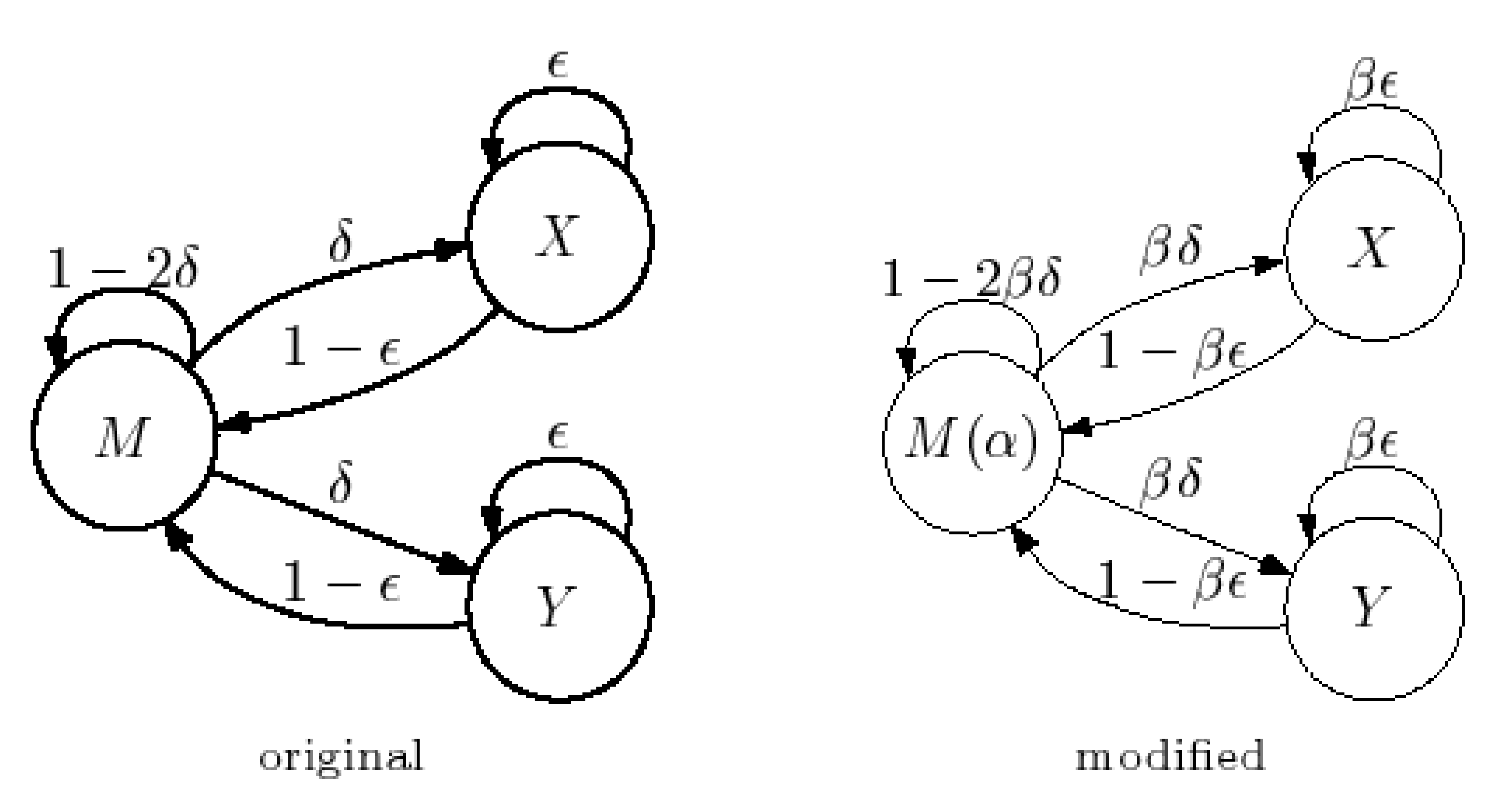
$$e'(i)=\sum_x p_1(x,i)e(x) \quad e'(j)=\sum_y p_2(x,i)e(y)$$

Modified Pair-HMM

- Secondary structure predictions:
  - In state *M*, introduce an additional parameter *α*
  - Subdivide the emission probability *e'*(*i*,*j*) into two cases to obtain the state *M*(*α*) with emission probability *αe'*(*i*,*j*) if (x,y) at position *i* in *X* and *j* in *Y* have the same secondary structure type.
  - (1-*α)e'*(*i*,*j*) otherwise.
- Decrease in emission will allow more gaps:
  - Use *β* to compensate for the change

Modified Pair-HMM

- Secondary structure prediction



Content

- Biology Motivation
- Computation Problem
- Algorithm
- Performance

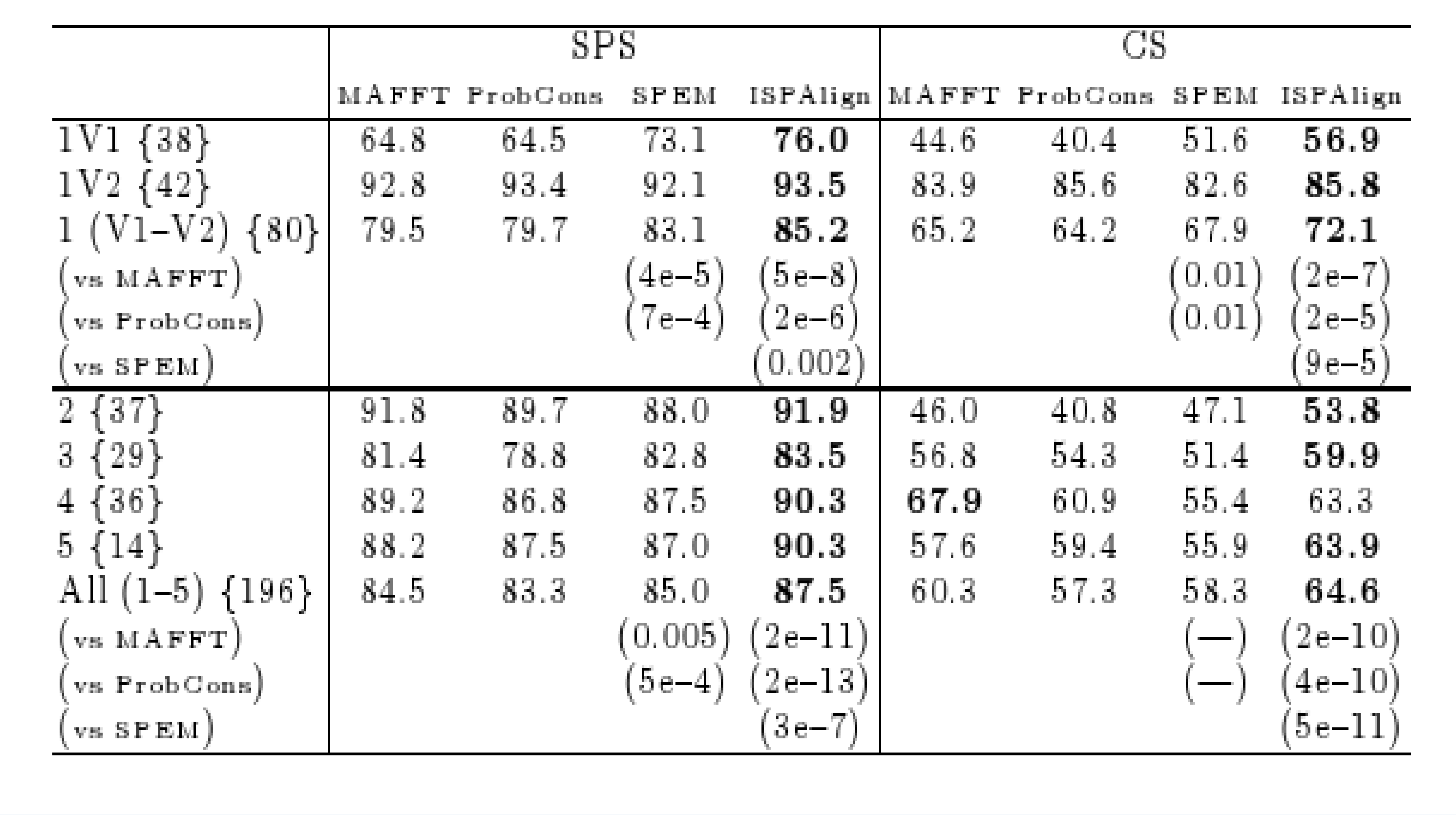
Content

- Biology Motivation
- Computation Problem
- Algorithm
- Performance

Performance

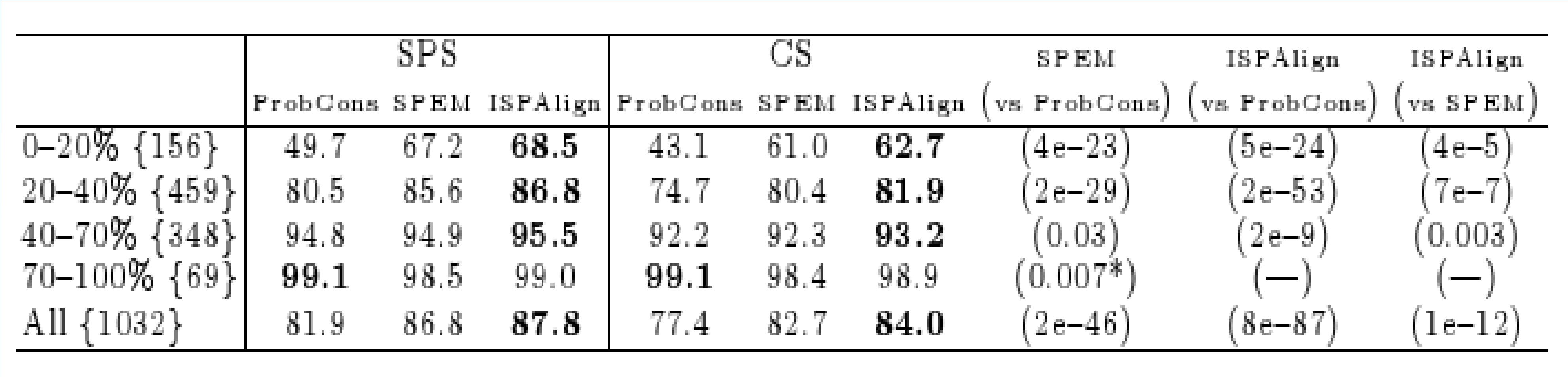
- Benchmark Sets:
  - BAliBASE 3.0
  - HOMSTRAD
  - PREFAB
  - SABmark
- Compare with:
  - MAFFT 5.8
  - ProbCons 1.10
  - SPEM

Performance

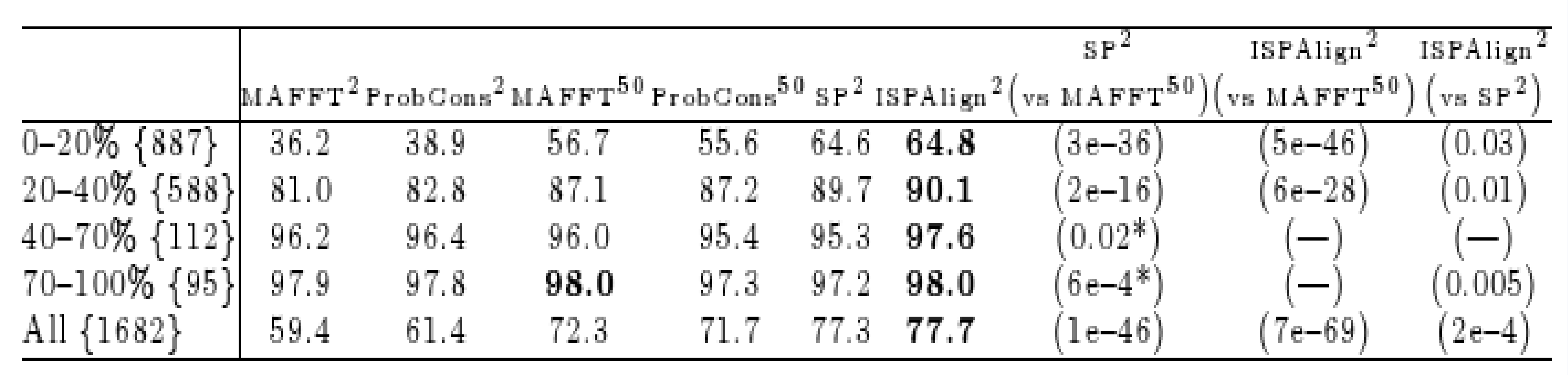


	SPS				CS			
	MAFFT	ProbCons	SPEM	ISPAlign	MAFFT	ProbCons	SPEM	ISPAlign
1V1 {38}	64.8	64.5	73.1	**76.0**	44.6	40.4	51.6	**56.9**
1V2 {42}	92.8	93.4	92.1	**93.5**	83.9	85.6	82.6	**85.8**
1 (V1–V2) {80}	79.5	79.7	83.1	**85.2**	65.2	64.2	67.9	**72.1**
(vs MAFFT)			(4e–5)	(5e–8)			(0.01)	(2e–7)
(vs ProbCons)			(7e–4)	(2e–6)			(0.01)	(2e–5)
(vs SPEM)				(0.002)				(9e–5)
2 {37}	91.8	89.7	88.0	**91.9**	46.0	40.8	47.1	**53.8**
3 {29}	81.4	78.8	82.8	**83.5**	56.8	54.3	51.4	**59.9**
4 {36}	89.2	86.8	87.5	**90.3**	**67.9**	60.9	55.4	63.3
5 {14}	88.2	87.5	87.0	**90.3**	57.6	59.4	55.9	**63.9**
All (1–5) {196}	84.5	83.3	85.0	**87.5**	60.3	57.3	58.3	**64.6**
(vs MAFFT)			(0.005)	(2e–11)			(—)	(2e–10)
(vs ProbCons)			(5e–4)	(2e–13)			(—)	(4e–10)
(vs SPEM)				(3e–7)				(5e–11)

Performance

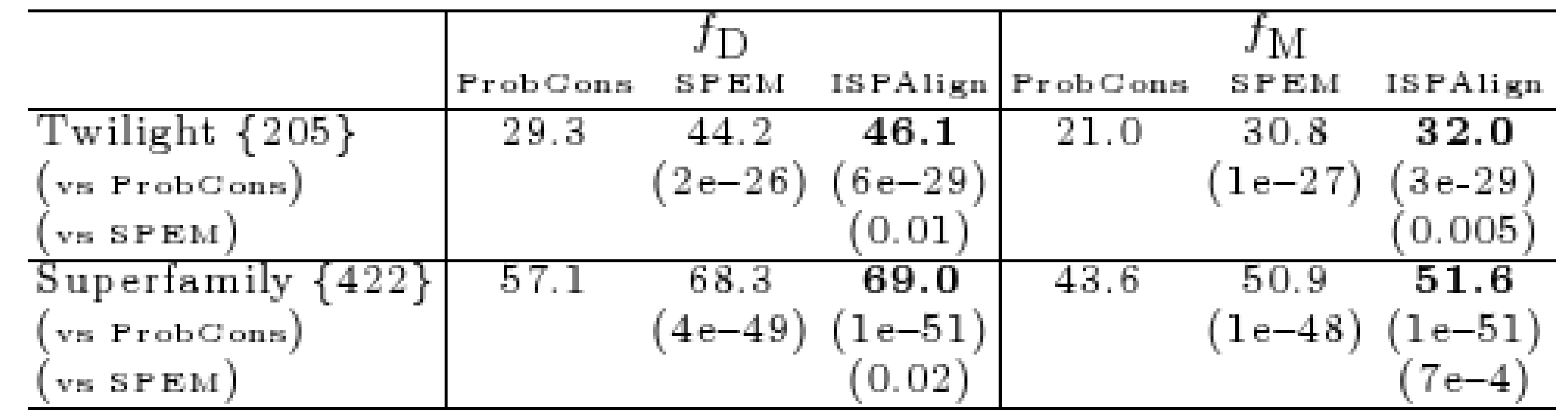


	SPS			CS			SPEM	ISPAlign	ISPAlign
	ProbCons	SPEM	ISPAlign	ProbCons	SPEM	ISPAlign	(vs ProbCons)	(vs ProbCons)	(vs SPEM)
0–20% {156}	49.7	67.2	**68.5**	43.1	61.0	**62.7**	(4e–23)	(5e–24)	(4e–5)
20–40% {459}	80.5	85.6	**86.8**	74.7	80.4	**81.9**	(2e–29)	(2e–53)	(7e–7)
40–70% {348}	94.8	94.9	**95.5**	92.2	92.3	**93.2**	(0.03)	(2e–9)	(0.003)
70–100% {69}	**99.1**	98.5	99.0	**99.1**	98.4	98.9	(0.007*)	(—)	(—)
All {1032}	81.9	86.8	**87.8**	77.4	82.7	**84.0**	(2e–46)	(8e–87)	(1e–12)

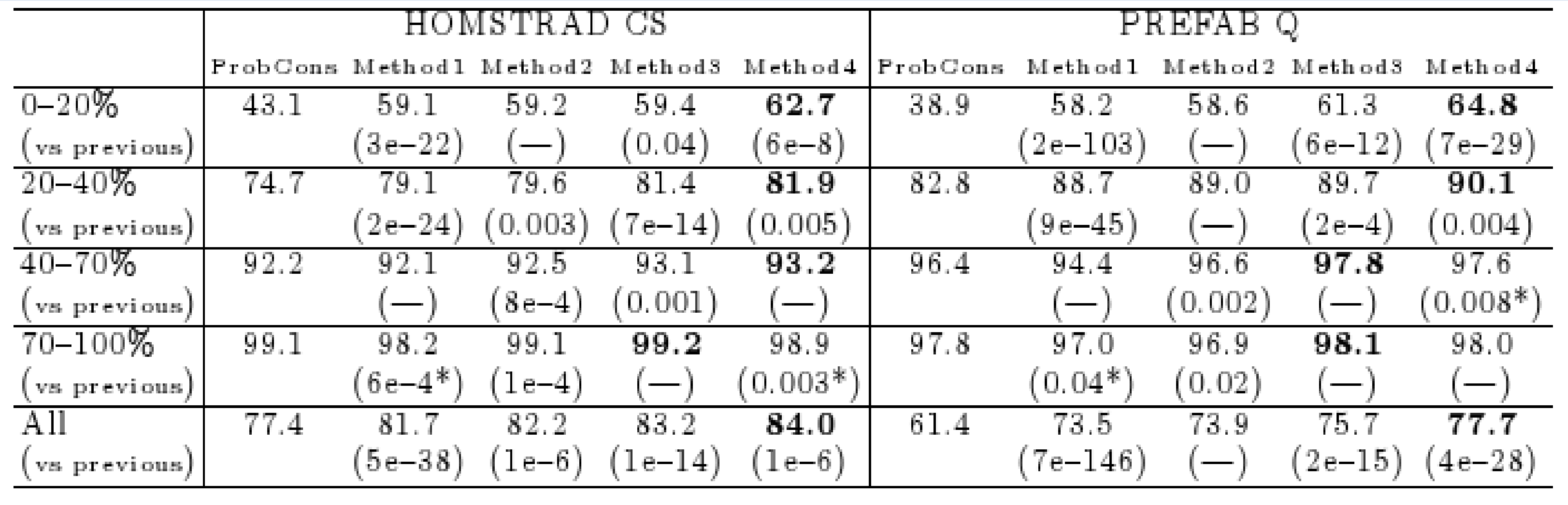


							SP^2	ISPAlign^2	ISPAlign^2
	MAFFT^2	ProbCons^2	MAFFT^{50}	ProbCons^{50}	SP^2	ISPAlign^2	(vs MAFFT^{50})	(vs MAFFT^{50})	(vs SP^2)
0–20% {887}	36.2	38.9	56.7	55.6	64.6	**64.8**	(3e–36)	(5e–46)	(0.03)
20–40% {588}	81.0	82.8	87.1	87.2	89.7	**90.1**	(2e–16)	(6e–28)	(0.01)
40–70% {112}	96.2	96.4	96.0	95.4	95.3	**97.6**	(0.02*)	(—)	(—)
70–100% {95}	97.9	97.8	**98.0**	97.3	97.2	**98.0**	(6e–4*)	(—)	(0.005)
All {1682}	59.4	61.4	72.3	71.7	77.3	**77.7**	(1e–46)	(7e–69)	(2e–4)

Performance



	f_D			f_M		
	ProbCons	SPEM	ISPAlign	ProbCons	SPEM	ISPAlign
Twilight {205}	29.3	44.2	**46.1**	21.0	30.8	**32.0**
(vs ProbCons)		(2e−26)	(6e−29)		(1e−27)	(3e−29)
(vs SPEM)			(0.01)			(0.005)
Superfamily {422}	57.1	68.3	**69.0**	43.6	50.9	**51.6**
(vs ProbCons)		(4e−49)	(1e−51)		(1e−48)	(1e−51)
(vs SPEM)			(0.02)			(7e−4)



	HOMSTRAD CS					PREFAB Q				
	ProbCons	Method1	Method2	Method3	Method4	ProbCons	Method1	Method2	Method3	Method4
0–20%	43.1	59.1	59.2	59.4	**62.7**	38.9	58.2	58.6	61.3	**64.8**
(vs previous)		(3e−22)	(—)	(0.04)	(6e−8)		(2e−103)	(—)	(6e−12)	(7e−29)
20–40%	74.7	79.1	79.6	81.4	**81.9**	82.8	88.7	89.0	89.7	**90.1**
(vs previous)		(2e−24)	(0.003)	(7e−14)	(0.005)		(9e−45)	(—)	(2e−4)	(0.004)
40–70%	92.2	92.1	92.5	93.1	**93.2**	96.4	94.4	96.6	**97.8**	97.6
(vs previous)		(—)	(8e−4)	(0.001)	(—)		(—)	(0.002)	(—)	(0.008*)
70–100%	99.1	98.2	99.1	**99.2**	98.9	97.8	97.0	96.9	**98.1**	98.0
(vs previous)		(6e−4*)	(1e−4)	(—)	(0.003*)		(0.04*)	(0.02)	(—)	(—)
All	77.4	81.7	82.2	83.2	**84.0**	61.4	73.5	73.9	75.7	**77.7**
(vs previous)		(5e−38)	(1e−6)	(1e−14)	(1e−6)		(7e−146)	(—)	(2e−15)	(4e−28)

Future Work

- Adding intermediate sequence
  - Rather than a fixed number, the number to add depends on the number of the input.
  - Or until the minimum distances fall below a threshold.
- Retain the pair-HMM using a set of confirmed secondary structures.
- Use other profile method
- Use 3D structures if possible

References

- Do, C.B., Mahabhashyam, M.S., Brudno, M., Batzoglou, S.: ProbCons: probabilistic consistency-based multiple sequence alignment. Genome Res. 15 (2005) 330–340
- Durbin, R., Eddy, S., Krogh, A., Mitchison, G.: Biological sequence analysis. Cambridge University Press (1998)
- Gusfield, D.: Efficient methods for multiple sequence alignment with guaranteed error bounds. Bull. Math. Biol. 55 (1993) 141–154
- Salamov, A.A., Suwa, M., Orengo, C.A., Swindells, M.B.: Combining sensitive database searches with multiple intermediates to detect distant homologues. Protein Eng. 12 (1999) 95–100
- Zhou, H., Zhou, Y.: SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structures. Bioinformatics 21 (2005) 3615–3621
- http://en.wikipedia.org/wiki/Multiple_sequence_alignment
- http://en.wikipedia.org/wiki/Hidden_Markov_Model
- http://lectures.molgen.mpg.de/MSA/Intro/index.html

Thank you!

Questions or Comment?