Chapter 8
Multiple sequence alignment

Chaochun Wei

Spring 2019

Contents

1. Reading materials
2. Multiple sequence alignment
   - basic algorithms and tools
   - how to improve multiple alignment

Reading materials

Book

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). Biological Sequence Analysis. Cambridge University Press. Chapter 5, 6

(Errata page: http://selab.janelia.org/cupbook_errata.html)

Multiple Alignment

- What can one learn from a multiple alignment?
- How can a multiple alignment be used?
- How is a good multiple alignment obtained?

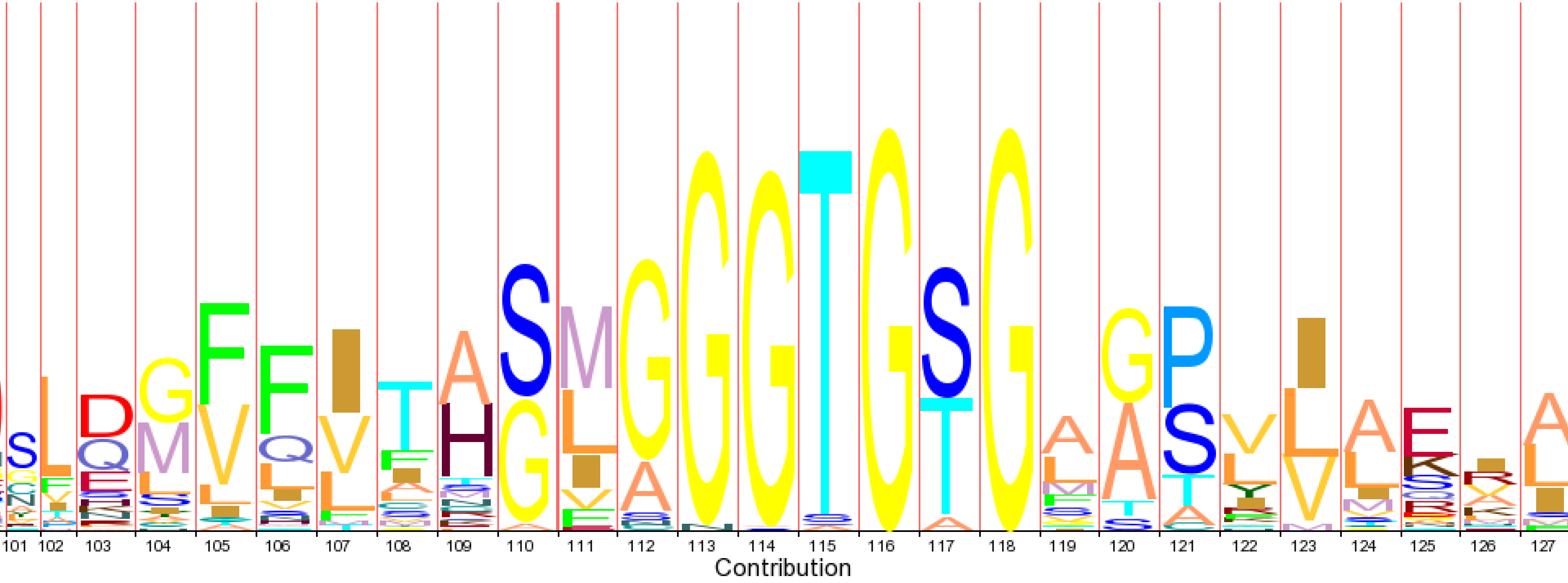
```
Q9GPZ8_DICDI/51-243  ..PEVGKKAT EESIEELMNQ IGDT...QML FVTAGMGGGT GTGGAAVIAS
  FTSZ_ARATH/74-267  ..PLLGEQAA EESKDAIANA LKGS...DLV FITAGMGGGT GSGAAPVVAQ
Q9XJ33_CYACA/92-292  ..PEAGRVAA EESKEDIAKA LQGG...DLV FVTAGMGGGT GTGAAPIVAD
   FTSZ_MYCKA/9-202  ..PEVGRXAA EDAKDDIEEL LRGA...DMV FVTAGEGGGT GTGGAPVVAS
   FTSZ_CORGL/9-202  ..PEVGRASA EDHKNEIEET IKGA...DMV FVTAGEGGGT GTGAAPVVAG
 Q9RWN5_DEIRA/4-197  ..PKVGEEAA VEDRDRIKEY LDDT...DML FITAGMGGGT GTGSAPVVAE
  FTSZ_MYCPU/11-202  ..PEVGKKAA EESIVEIKEK LKGA...DMV IITSGMGGGT GTGASPIIAK
  FTSZ_PORGI/17-211  ..PEVARRAA EASEADIRKI LDDG.HTRMV FVTAGMGGGT GTGAAPVIGR
Q9S344_9BACT/15-205  ..PARARQAA EETLDDIKGM LNDG..TKMA FITAGMGGGT GTGAAPVIAR
   FTSZ_AQUAE/8-201  ..PEVGEEAA LEDIDKIKEI LRDT...DMV FISAGLGGGT GTGAAPVIAK
Q19490_CAEEL/49-246  ..YTIGKELI DVVMDRVRRL TERCQSLQGF LIFHSFGGGT GSGFTSLVME
  TBA1_SCHPO/53-250  ..YTVGKEMI DSVLERIRRM ADNCSGLQGF LVFHSFGGGT GSGLGALLLE
O36040_SPIVO/27-224  ..NTIGKEVI DLVLDRIRKL ADDCSGLQGF IMFHSFGGGT GSGLGALLLE
Q9UVR1_9ZYGO/30-229  ..YTEGAELL DQVLDTIRQD VERCDLLSGF QLCHSIAGGT GSGMGSLMLQ
Q20823_CAEEL/45-245  ..YEQGAEIV DKVLSVIRRE AEAADSLEGF QLIHSLGGGT GSGLGSLLIS
  TBBP_DROME/46-243  ..HTDGAAIL DQVLENTRRE VESVDSLQGF QLLHSIGGGT GSGLTSLIME
   TBG_EUPAE/46-244  ..YTDAEKVQ DEILEMIDRE ADGSDSLEGF VLTHSIAGGT GSGFGSYLLE
   TBG_CHLRE/46-247  ..YTQGEAVQ ETLLDMIDRE AEYCDSLEGF NMCHSIAGGT GSGMGSYMLE
  TBG1_DROME/46-247  ..YSQGEKLQ EEVFDIIDRE ADGSDSLEGF ILCHSIAGGT GSGMGSFIME
Q94771_9TRYP/46-249  ..YEMGDTVQ ETLFDMIERE AENSDSLEGF VLTHSIAGGT GSGMGSYLLE
   TBG_USTVI/46-246  ..YAAGERVY EEVMEMIDRE AEGSDSLEGF MLLHSIAGGT GSGLGSYLLE
   TBG_SCHJP/46-247  ..YAHAEKIF EDIVDMIDRE AEGSDSLEGF SLLHSIAGGT GSGLGSYLLE
O15812_DICDI/46-244  ..YKQGESFY DDIFDMIDRE ADGSESLEGF LLTHSISGGT GSGMGSYILE
O00849_TETTH/46-246  ..YQEANKIQ DDLLDMIDRE ADTSDSFEAF LLIHSIAGGT GSGVGSYLLE
   TBG_CAEEL/47-249  ..YCQGQEVQ EKIMDIIIRE AENTNNLDGI LFTHSVSGGT GSGTGSLLLE
   TBG_ENTHI/45-242  ..YYTTEKMS .EIEEIIDRE VEHCDSLEGF FFCHSICGGT GSGLGSKIME
   TBG_YEAST/48-246  ..YDIGTRNQ DDILNKIDKE IDSTDNFEGF QLLHSVAGGT GSGLGSNLLE
   TBG_CANAL/75-282  ..YKYGTEEE ETLLNLIDRE VDKCDNLSNF QLFHSVAGGT GSGVGSKMLE
Q9NI44_9TRYP/49-280  ..MEYGDKYI DSITETVREQ VERCDSIQSF LIMHSLSGGT GAGLGTRVLG
   TBD_HUMAN/46-242  ..SVHGPRHE ESIMNIIRKE VEKCDSFSGF FIIMSMAGGT GSGLGAFVTQ
```

```
Q9GPZ8_DICDI/51-243   ..PEVGKKAT EESIEELMNQ IGDT...QML FVTAGMGGGT GTGGAAVIAS
   FTSZ_ARATH/74-267  ..PLLGEQAA EESKDAIANA LKGS...DLV FITAGMGGGT GSGAAPVVAQ
Q9XJ33_CYACA/92-292   ..PEAGRVAA EESKEDIAKA LQGG...DLV FVTAGMGGGT GTGAAPIVAD
    FTSZ_MYCKA/9-202  ..PEVGRXAA EDAKDDIEEL LRGA...DMV FVTAGEGGGT GTGGAPVVAS
    FTSZ_CORGL/9-202  ..PEVGRASA EDHKNEIEET IKGA...DMV FVTAGEGGGT GTGAAPVVAG
 Q9RWN5_DEIRA/4-197   ..PKVGEEAA VEDRDRIKEY LDDT...DML FITAGMGGGT GTGSAPVVAE
   FTSZ_MYCPU/11-202  ..PEVGKKAA EESIVEIKEK LKGA...DMV IITSGMGGGT GTGASPIIAK
   FTSZ_PORGI/17-211  ..PEVARRAA EASEADIRKI LDDG.HTRMV FVTAGMGGGT GTGAAPVIGR
Q9S344_9BACT/15-205   ..PARARQAA EETLDDIKGM LNDG..TKMA FITAGMGGGT GTGAAPVIAR
    FTSZ_AQUAE/8-201  ..PEVGEEAA LEDIDKIKEI LRDT...DMV FISAGLGGGT GTGAAPVIAK
Q19490_CAEEL/49-246   ..YTIGKELI DVVMDRVRRL TERCQSLQGF LIFHSFGGGT GSGFTSLVME
  TBA1_SCHPO/53-250   ..YTVGKEMI DSVLERIRRM ADNCSGLQGF LVFHSFGGGT GSGLGALLLE
O36040_SPIVO/27-224   ..NTIGKEVI DLVLDRIRKL ADDCSGLQGF IMFHSFGGGT GSGLGALLLE
Q9UVR1_9ZYGO/30-229   ..YTEGAELL DQVLDTIRQD VERCDLLSGF QLCHSIAGGT GSGMGSLMLQ
Q20823_CAEEL/45-245   ..YEQGAEIV DKVLSVIRRE AEAADSLEGF QLIHSLGGGT GSGLGSLLIS
  TBBP_DROME/46-243   ..HTDGAAIL DQVLENTRRE VESVDSLQGF QLLHSIGGGT GSGLTSLIME
   TBG_EUPAE/46-244   ..YTDAEKVQ DEILEMIDRE ADGSDSLEGF VLTHSIAGGT GSGFGSYLLE
   TBG_CHLRE/46-247   ..YTQGEAVQ ETLLDMIDRE AEYCDSLEGF NMCHSIAGGT GSGMGSYMLE
  TBG1_DROME/46-247   ..YSQGEKLQ EEVFDIIDRE ADGSDSLEGF ILCHSIAGGT GSGMGSFIME
Q94771_9TRYP/46-249   ..YEMGDTVQ ETLFDMIERE AENSDSLEGF VLTHSIAGGT GSGMGSYLLE
   TBG_USTVI/46-246   ..YAAGERVY EEVMEMIDRE AEGSDSLEGF MLLHSIAGGT GSGLGSYLLE
   TBG_SCHJP/46-247   ..YAHAEKIF EDIVDMIDRE AEGSDSLEGF SLLHSIAGGT GSGLGSYLLE
O15812_DICDI/46-244   ..YKQGESFY DDIFDMIDRE ADGSESLEGF LLTHSISGGT GSGMGSYILE
O00849_TETTH/46-246   ..YQEANKIQ DDLLDMIDRE ADTSDSFEAF LLIHSIAGGT GSGVGSYLLE
   TBG_CAEEL/47-249   ..YCQGQEVQ EKIMDIIIRE AENTNNLDGI LFTHSVSGGT GSGTGSLLLE
   TBG_ENTHI/45-242   ..YYTTEKMS .EIEEIIDRE VEHCDSLEGF FFCHSICGGT GSGLGSKIME
   TBG_YEAST/48-246   ..YDIGTRNQ DDILNKIDKE IDSTDNFEGF QLLHSVAGGT GSGLGSNLLE
   TBG_CANAL/75-282   ..YKYGTEEE ETLLNLIDRE VDKCDNLSNF QLFHSVAGGT GSGVGSKMLE
Q9NI44_9TRYP/49-280   ..MEYGDKYI DSITETVREQ VERCDSIQSF LIMHSLSGGT GAGLGTRVLG
   TBD_HUMAN/46-242   ..SVHGPRHE ESIMNIIRKE VEKCDSFSGF FIIMSMAGGT GSGLGAFVTQ
```

What can one learn from a multiple alignment?

- Some regions tend to be more highly conserved than others
- Gaps are often clustered
- May be conservation of types of residues (eg. hydrophylic/hydrophobic) even if the residues them selves are variable
- Can plot conservation to get an overview of how it varies

Logo of a section of the tubulin protein family



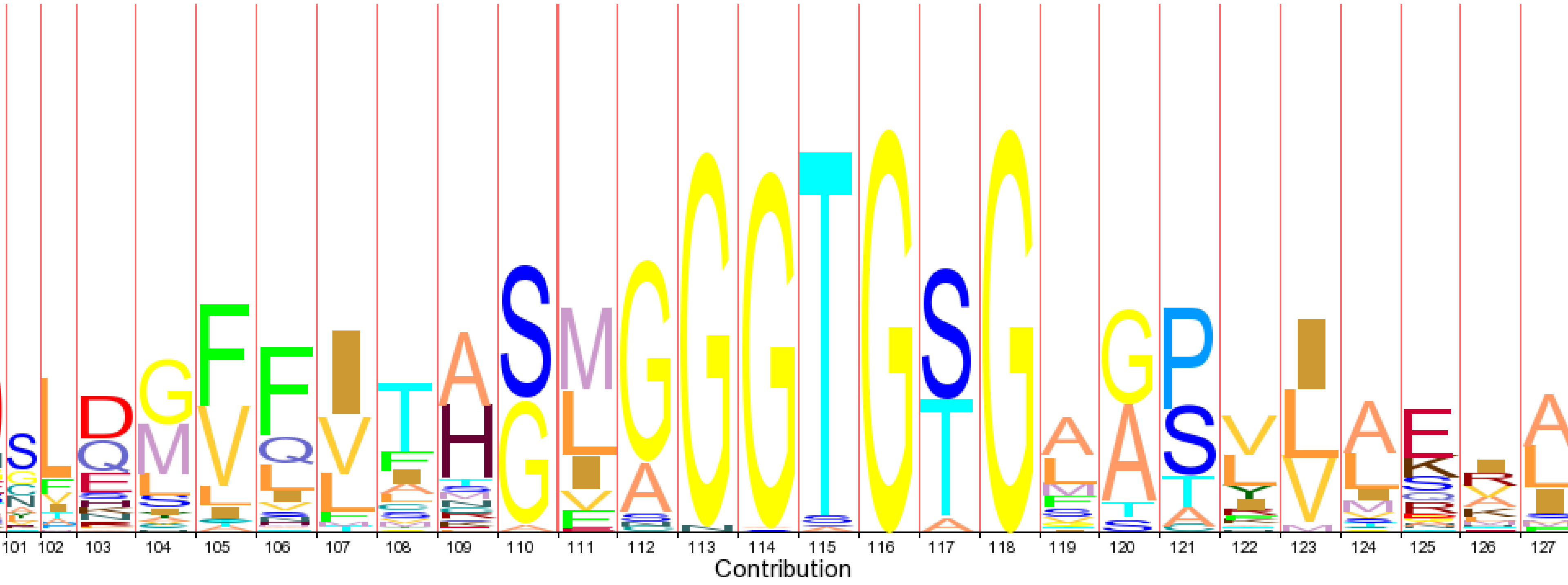
How can a multiple alignment be used?

- Insights into protein structure/function
  - Highly conserved positions/regions mostly likely required for function
  - Indels and hydrophilic regions usually on surface
- Better, more sensitive searches
  - Uses more information about protein’s features to identify homologs
  - Position-specific scoring function

Table 2 - The log odds matrix for BLOSUM 62

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	4	0	−2	−1	−2	0	−2	−1	−1	−1	−1	−2	−1	−1	−1	1	0	0	−3	−2
C		9	−3	−4	−2	−3	−3	−1	−3	−1	−1	−3	−3	−3	−3	−1	−1	−1	−2	−2
D			6	2	−3	−1	−1	−3	−1	−4	−3	1	−1	0	−2	0	−1	−3	−4	−3
E				5	−3	−2	0	−3	1	−3	−2	0	−1	2	0	0	−1	−2	−3	−2
F					6	−3	−1	0	−3	0	0	−3	−4	−3	−3	−2	−2	−1	1	3
G						6	−2	−4	−2	−4	−3	0	−2	−2	−2	0	−2	−3	−2	−3
H							8	−3	−1	−3	−2	1	−2	0	0	−1	−2	−3	−2	2
I								4	−3	2	1	−3	−3	−3	−3	−2	−1	3	−3	−1
K									5	−2	−1	0	−1	1	2	0	−1	−2	−3	−2
L										4	2	−3	−3	−2	−2	−2	−1	1	−2	−1
M											5	−2	−2	0	−1	−1	−1	1	−1	−1
N												6	−2	0	0	1	0	−3	−4	−2
P													7	−1	−2	−1	−1	−2	−4	−3
Q														5	1	0	−1	−2	−2	−1
R															5	−1	−1	−3	−3	−2
S																4	1	−2	−3	−2
T																	5	0	−2	−2
V																		4	−3	−1
W																			11	2
Y																				7

```
FTSZ_AQUAE/8-201     ..PEVGEEAA LEDIDKIKEI LRDT...DMV FISAGLGGGT GTGAAPVIAK
Q19490_CAEEL/49-246  ..YTIGKELI DVVMDRVRRL TERCQSLQGF LIFHSFGGGT GSGFTSLVME
                          * *       *                  *    **** * *
```



```
FTSZ_AQUAE/8-201     ..PEVGEEAA LEDIDKIKEI LRDT...DMV FISAGLGGGT GTGAAPVIAK
Q19490_CAEEL/49-246  ..YTIGKELI DVVMDRVRRL TERCQSLQGF LIFHSFGGGT GSGFTSLVME
                          * *       *                  *    **** * *
```

Scoring multiple alignments

- Common to use “sum of pairs” using the standard pairwise scoring
- An alignment of residue X in the query with the position Y of the alignment that contains the set Y_i of residues gets:

$$\begin{aligned}\text{Score}(X,Y) &= \sum_i s(X,Y_i) \\ &= \sum_i \ln[P(X,Y_i)/P(X)P(Y_i)] \\ &= \sum_i \ln[P(X|Y_i)/P(X)]\end{aligned}$$

Sum-of-Pairs scoring (cont)

- $\mathrm{Score}(X,Y) = \sum_i \ln[P(X|Y_i)/P(X)]$

  we can pre-compute the score for any X

- ➡ “Profile” for a multiple alignment

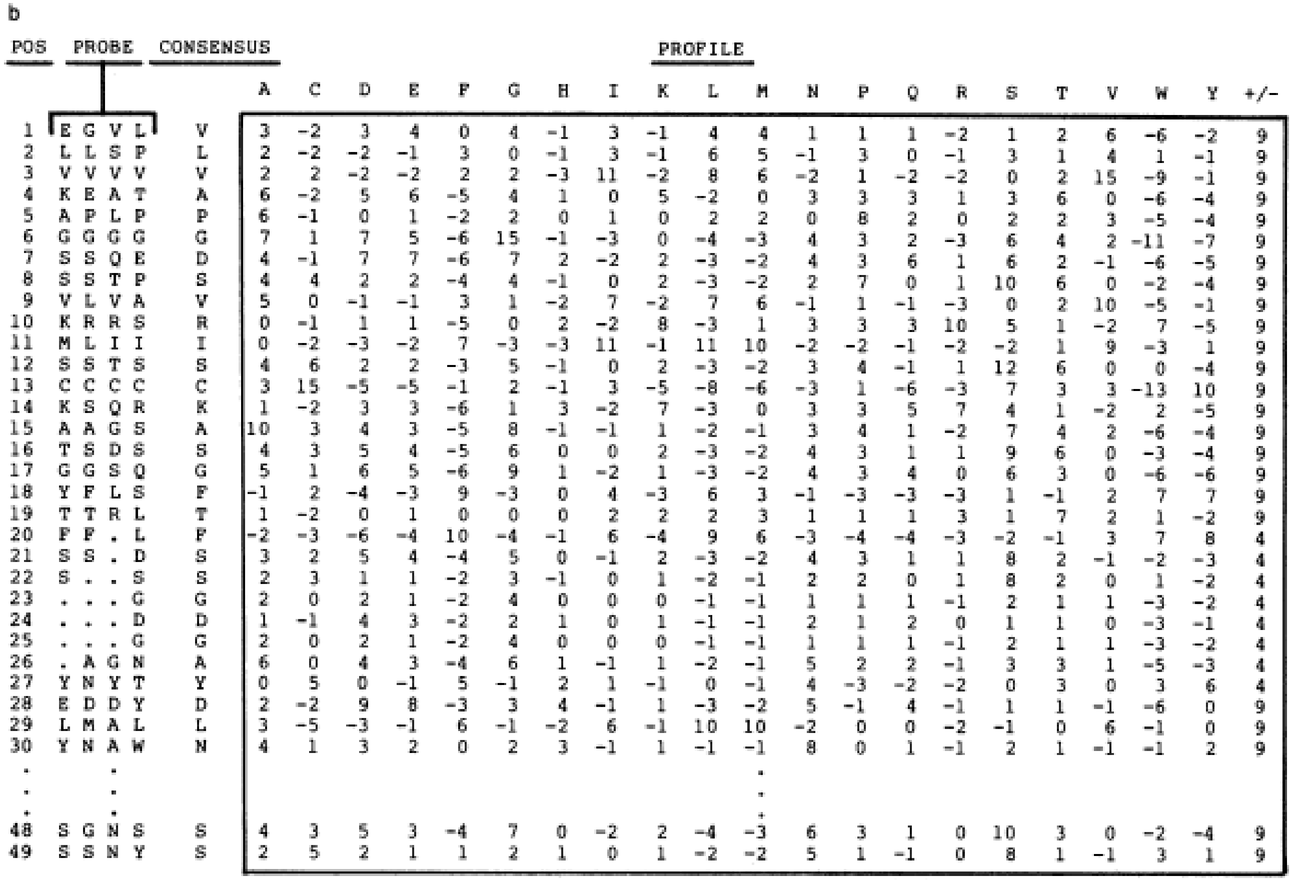
- Important Point: highly variable position tend toward 0 for all scores, while highly conserved positions maintain the s(X,Y) scores, increasing their contribution to the Score

Proc. Natl. Acad. Sci. USA
Vol. 84, pp. 4355–4358, July 1987
Biochemistry

Profile analysis: Detection of distantly related proteins

(amino acid/sequence comparison/protein structure/globin structure/immunoglobulin structure)

MICHAEL GRIBSKOV*, ANDREW D. MCLACHLAN†, AND DAVID EISENBERG*

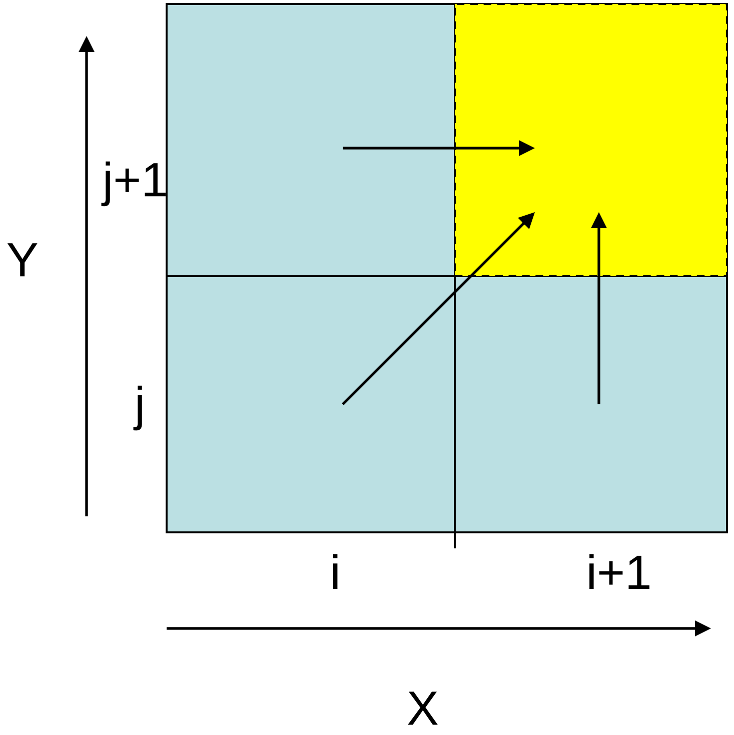


b

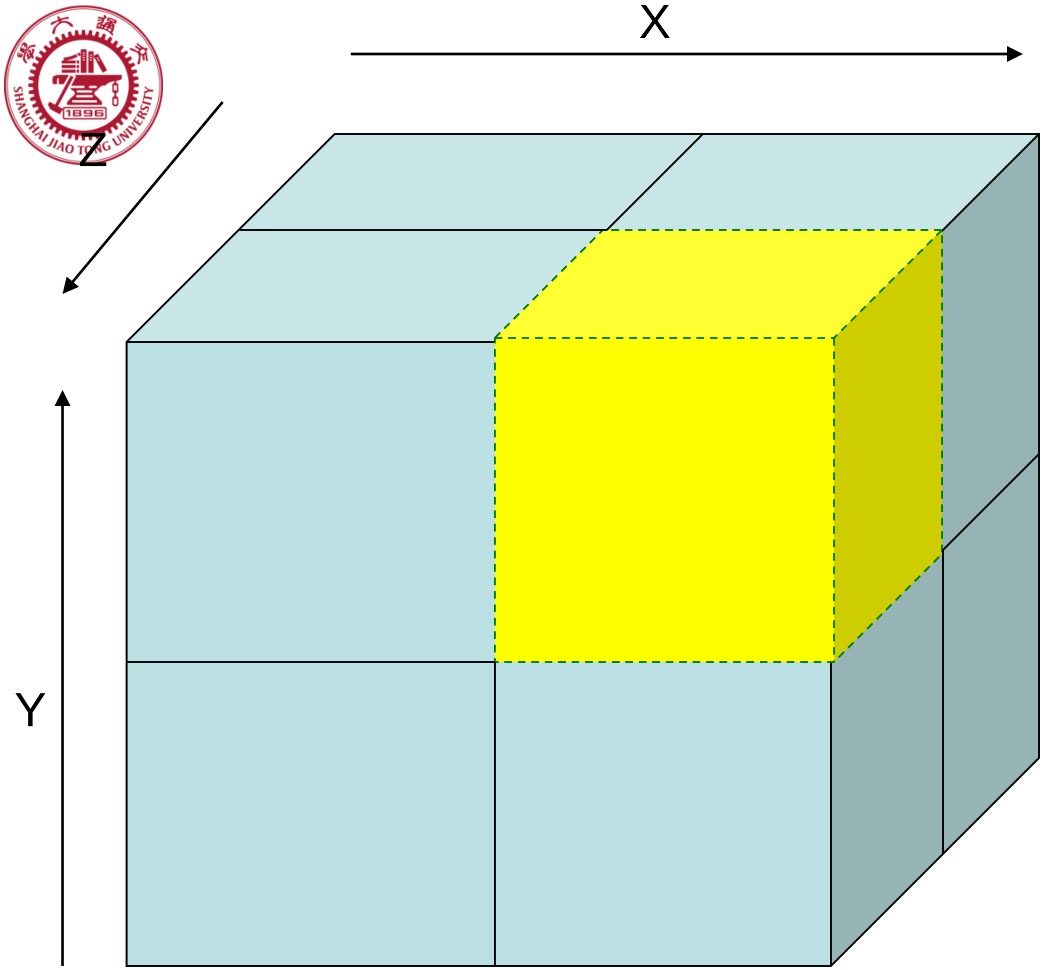
POS	PROBE	CONSENSUS	PROFILE																				
			A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	+/-
1	E G V L	V	3	−2	3	4	0	4	−1	3	−1	4	4	1	1	1	−2	1	2	6	−6	−2	9
2	L L S P	L	2	−2	−2	−1	3	0	−1	3	−1	6	5	−1	3	0	−1	3	1	4	1	−1	9
3	V V V V	V	2	2	−2	−2	2	2	−3	11	−2	8	6	−2	1	−2	−2	0	2	15	−9	−1	9
4	K E A T	A	6	−2	5	6	−5	4	1	0	5	−2	0	3	3	3	1	3	6	0	−6	−4	9
5	A P L P	P	6	−1	0	1	−2	2	0	1	0	2	2	0	8	2	0	2	2	3	−5	−4	9
6	G G G G	G	7	1	7	5	−6	15	−1	−3	0	−4	−3	4	3	2	−3	6	4	2	−11	−7	9
7	S S Q E	D	4	−1	7	7	−6	7	2	−2	2	−3	−2	4	3	6	1	6	2	−1	−6	−5	9
8	S S T P	S	4	4	2	2	−4	4	−1	0	2	−3	−2	2	7	0	1	10	6	0	−2	−4	9
9	V L V A	V	5	0	−1	−1	3	1	−2	7	−2	7	6	−1	1	−1	−3	0	2	10	−5	−1	9
10	K R R S	R	0	−1	1	1	−5	0	2	−2	8	−3	1	3	3	3	10	5	1	−2	7	−5	9
11	M L I I	I	0	−2	−3	−2	7	−3	−3	11	−1	11	10	−2	−2	−1	−2	−2	1	9	−3	1	9
12	S S T S	S	4	6	2	2	−3	5	−1	0	2	−3	−2	3	4	−1	1	12	6	0	0	−4	9
13	C C C C	C	3	15	−5	−5	−1	2	−1	3	−5	−8	−6	−3	1	−6	−3	7	3	3	−13	10	9
14	K S Q R	K	1	−2	3	3	−6	1	3	−2	7	−3	0	3	3	5	7	4	1	−2	2	−5	9
15	A A G S	A	10	3	4	3	−5	8	−1	−1	1	−2	−1	3	4	1	−2	7	4	2	−6	−4	9
16	T S D S	S	4	3	5	4	−5	6	0	0	2	−3	−2	4	3	1	1	9	6	0	−3	−4	9
17	G G S Q	G	5	1	6	5	−6	9	1	−2	1	−3	−2	4	3	4	0	6	3	0	−6	−6	9
18	Y F L S	F	−1	2	−4	−3	9	−3	0	4	−3	6	3	−1	−3	−3	−3	1	−1	2	7	7	9
19	T T R L	T	1	−2	0	1	0	0	0	2	2	2	3	1	1	1	3	1	7	2	1	−2	9
20	F F . L	F	−2	−3	−6	−4	10	−4	−1	6	−4	9	6	−3	−4	−4	−3	−2	−1	3	7	8	4
21	S S . D	S	3	2	5	4	−4	5	0	−1	2	−3	−2	4	3	1	1	8	2	−1	−2	−3	4
22	S . . S	S	2	3	1	1	−2	3	−1	0	1	−2	−1	2	2	0	1	8	2	0	1	−2	4
23	. . . G	G	2	0	2	1	−2	4	0	0	0	−1	−1	1	1	1	−1	2	1	1	−3	−2	4
24	. . . D	D	1	−1	4	3	−2	2	1	0	1	−1	−1	2	1	2	0	1	1	0	−3	−1	4
25	. . . G	G	2	0	2	1	−2	4	0	0	0	−1	−1	1	1	1	−1	2	1	1	−3	−2	4
26	. A G N	A	6	0	4	3	−4	6	1	−1	1	−2	−1	5	2	2	−1	3	3	1	−5	−3	4
27	Y N Y T	Y	0	5	0	−1	5	−1	2	1	−1	0	−1	4	−3	−2	−2	0	3	0	3	6	4
28	E D D Y	D	2	−2	9	8	−3	3	4	−1	1	−3	−2	5	−1	4	−1	1	1	−1	−6	0	9
29	L M A L	L	3	−5	−3	−1	6	−1	−2	6	−1	10	10	−2	0	0	−2	−1	0	6	−1	0	9
30	Y N A W	N	4	1	3	2	0	2	3	−1	1	−1	−1	8	0	1	−1	2	1	−1	−1	2	9
.	.												.										
.	.												.										
.	.												.										
48	S G N S	S	4	3	5	3	−4	7	0	−2	2	−4	−3	6	3	1	0	10	3	0	−2	−4	9
49	S S N Y	S	2	5	2	1	1	2	1	0	1	−2	−2	5	1	−1	0	8	1	−1	3	1	9

How is a good multiple alignment obtained?

- Can extend dynamic programming (DP) method (Smith-Waterman or Needleman-Wunch) to N>2 sequences



$$\max \left\{ \begin{array}{c:c:c} X & X & -- \\ Y & -- & Y \end{array} \right.$$



The seven neighboring cells are the seven possible paths for the optimal alignment

$$\max \left\{ \begin{array}{c|c|c|c|c|c|c} X & X & X & X & -- & -- & -- \\ Y & Y & -- & -- & Y & Y & – \\ Z & – & Z & – & Z & -- & Z \end{array} \right.$$

DP on multiple sequences

- Can extend standard DP to N sequences by using N-dimensional matrix, filling in optimal scores for each element using a defined scoring system, such as sum-of-pairs
- Problem: complexity is $O(L^N)$ for N sequences of length L

Impact of Computational Complexity

- Suppose your algorithm can run on N=10 sequences of length L=1000.
- You then get 1000 times as much of the limiting resource.
- How many sequences can you now run on, as a function of the complexity O(..) of that limiting resource?

Initially N=10 and L=1000

Then increase limiting resource by 1000-fold

Assuming overhead costs and all other terms are neglible.

Algorithm Complexity	New problem size
O(N)	10,000
O(N log N)	~2,000
O(N^2)	316
O(N^3)	100
O(L^N)	?

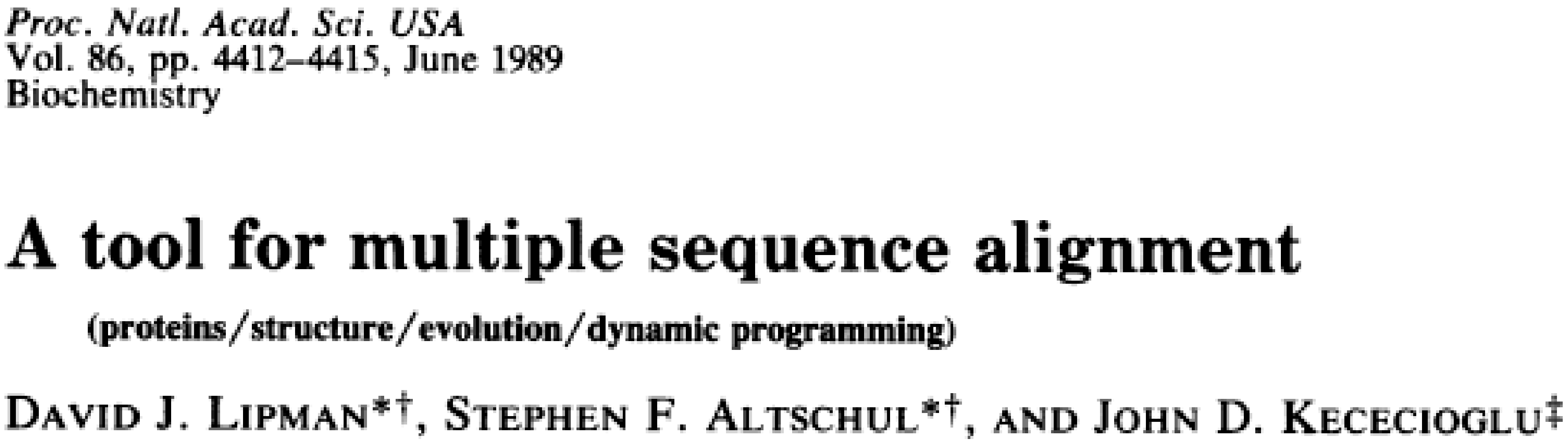
Initially N=10 and L=1000

Then increase limiting resource by 1000-fold

Assuming overhead costs and all other terms are neglible.

Algorithm Complexity	New problem size
O(N)	10,000
O(N log N)	~2,000
O(N^2)	316
O(N^3)	100
O(L^N)	11

Making multiple sequence alignment more efficient. MSA program uses pair-wise alignments to define “search space” in which to apply DP to find optimal alignment. Doesn’t have to fill in entire N-dim matrix, only those sections that can contribute to the optimal alignment. Uses branch-and-bound to determine the alignment space to be considered.



Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 4412–4415, June 1989
Biochemistry

A tool for multiple sequence alignment

(proteins/structure/evolution/dynamic programming)

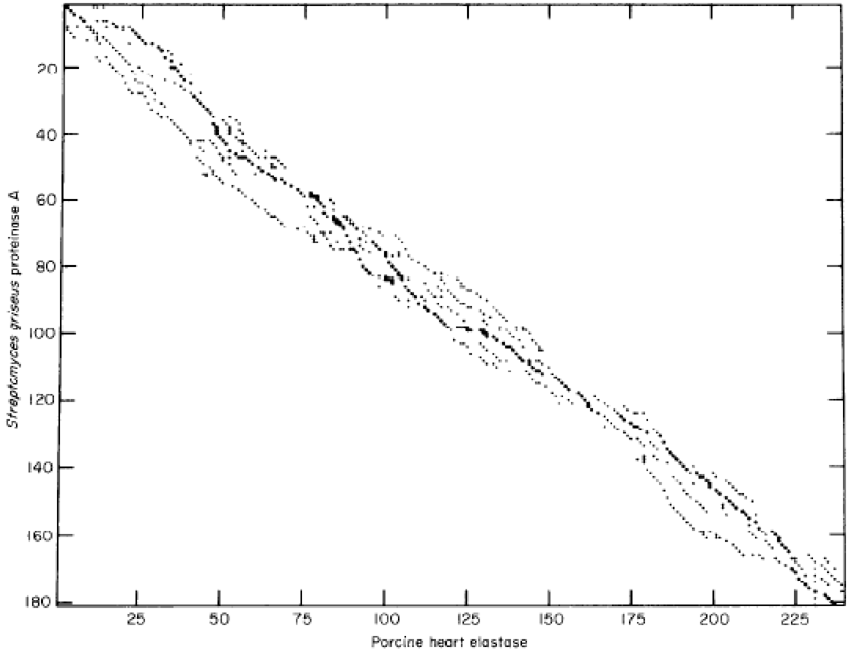
DAVID J. LIPMAN[*†], STEPHEN F. ALTSCHUL[*†], AND JOHN D. KECECIOGLU[‡]

Determining and displaying sub-optimal alignments. Can be used to set boundaries for MSA

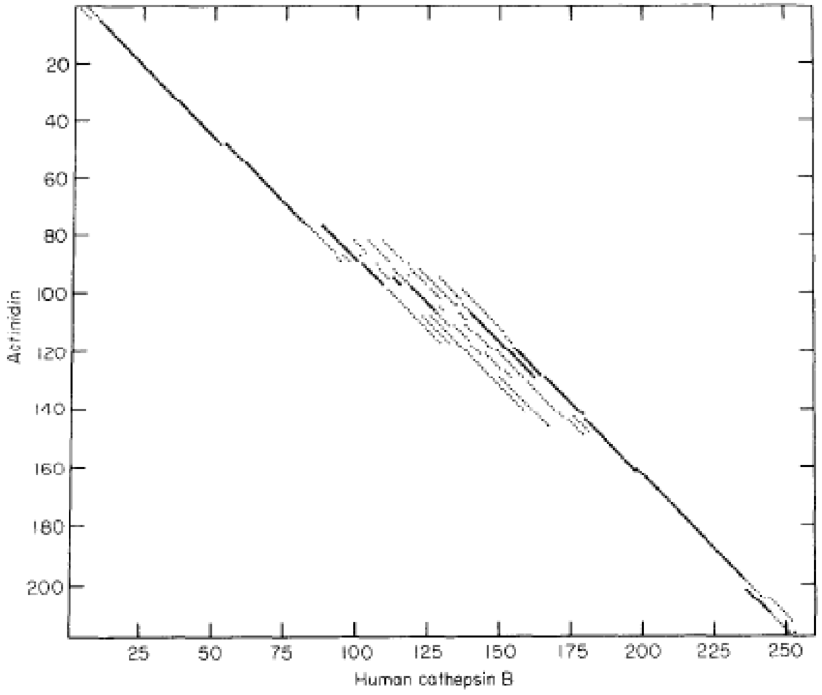
M(x,y) = Forward(x,y) +
Backward(x,y)

Can show all cells within some % of optimum score. Can be used to define boundaries for multi-sequence optimization.

Zuker, M (1991) JMB 221:403-420



(a)



(a)

How is a good multiple alignment obtained?

- Can standard dynamic programming (DP) method (Smith-Waterman or Needleman-Wunch) to N>2 sequences
  - $O(L^N)$ limits applicability

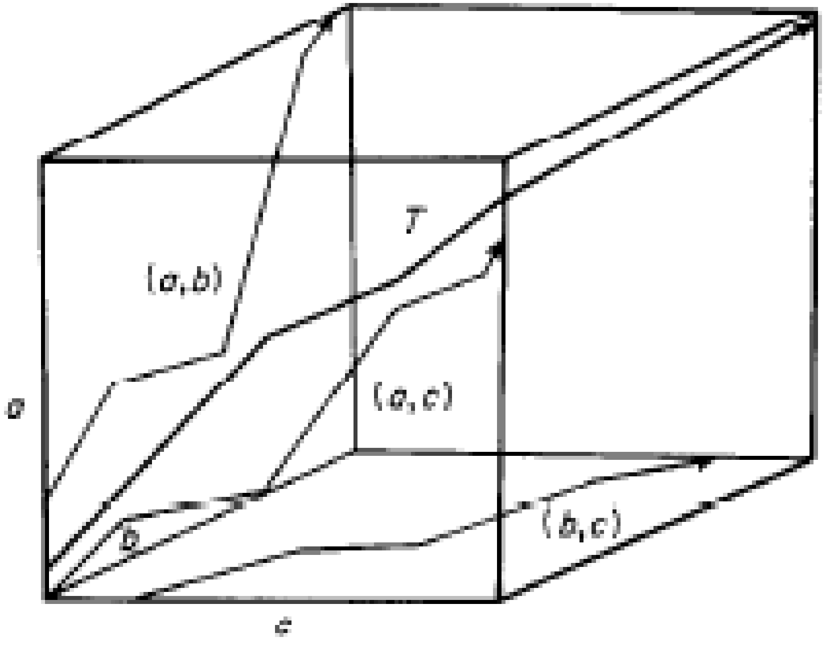
- Need good heuristic that returns near-optimal alignments in reasonable time/space

“Progressive Alignment”

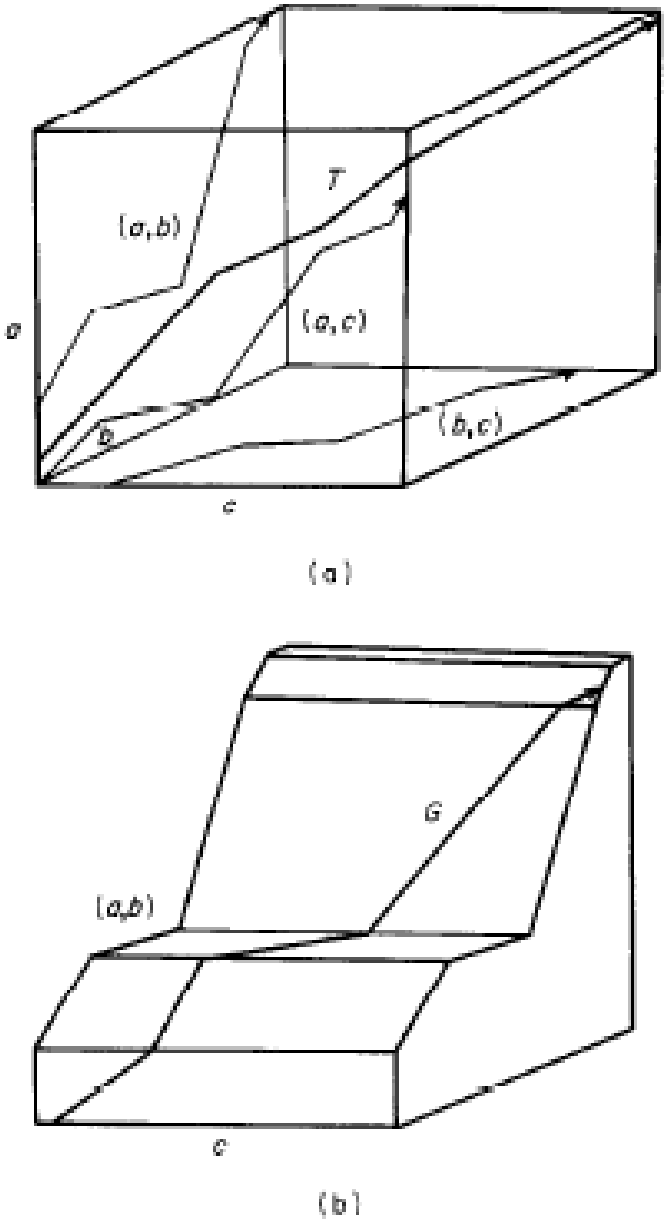
- Always do pairwise alignments
- Use DP to get optimal alignment of pairs
- Once a pair is aligned, that alignment is fixed in subsequent steps

- Some programs allow for the revising of previous steps, optimization of total score

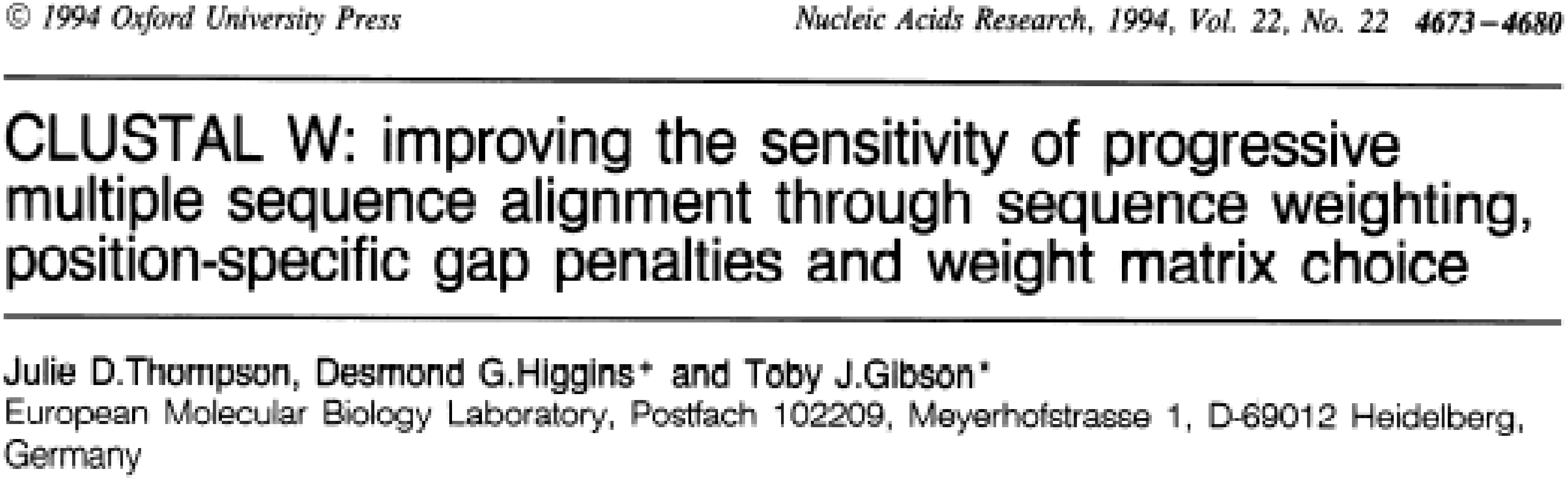
Three sequences (a,b,c) with optimal alignment T and pairwise alignments (a,b), (a,c), (b,c)



(a)



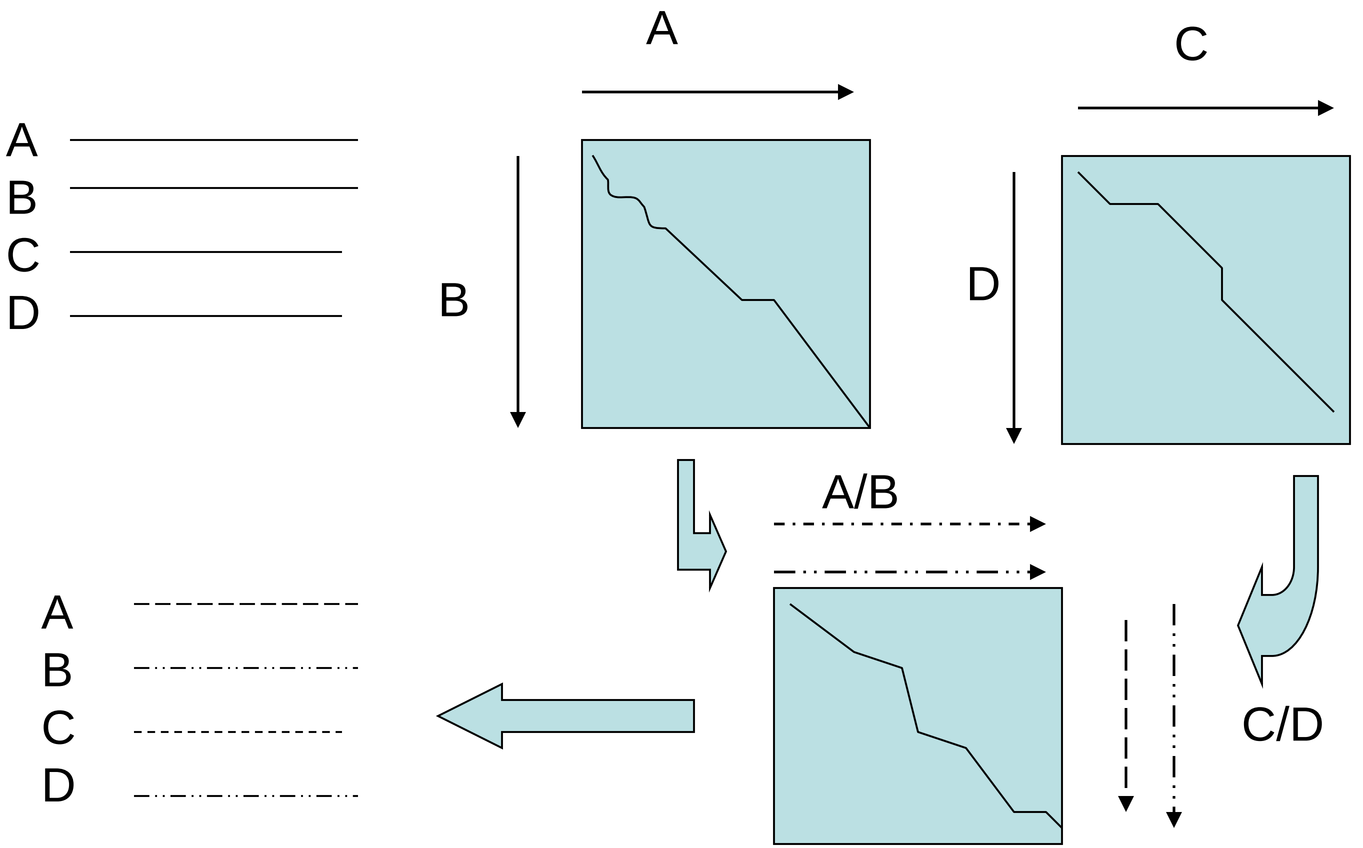
Subbiah and Harrison, (1989) J Mol Biol. 209:539-48.



Nucleic Acids Research, 1994, Vol. 22, No. 22 4673–4680

CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice

Julie D.Thompson, Desmond G.Higgins* and Toby J.Gibson*
European Molecular Biology Laboratory, Postfach 102209, Meyerhofstrasse 1, D-69012 Heidelberg, Germany

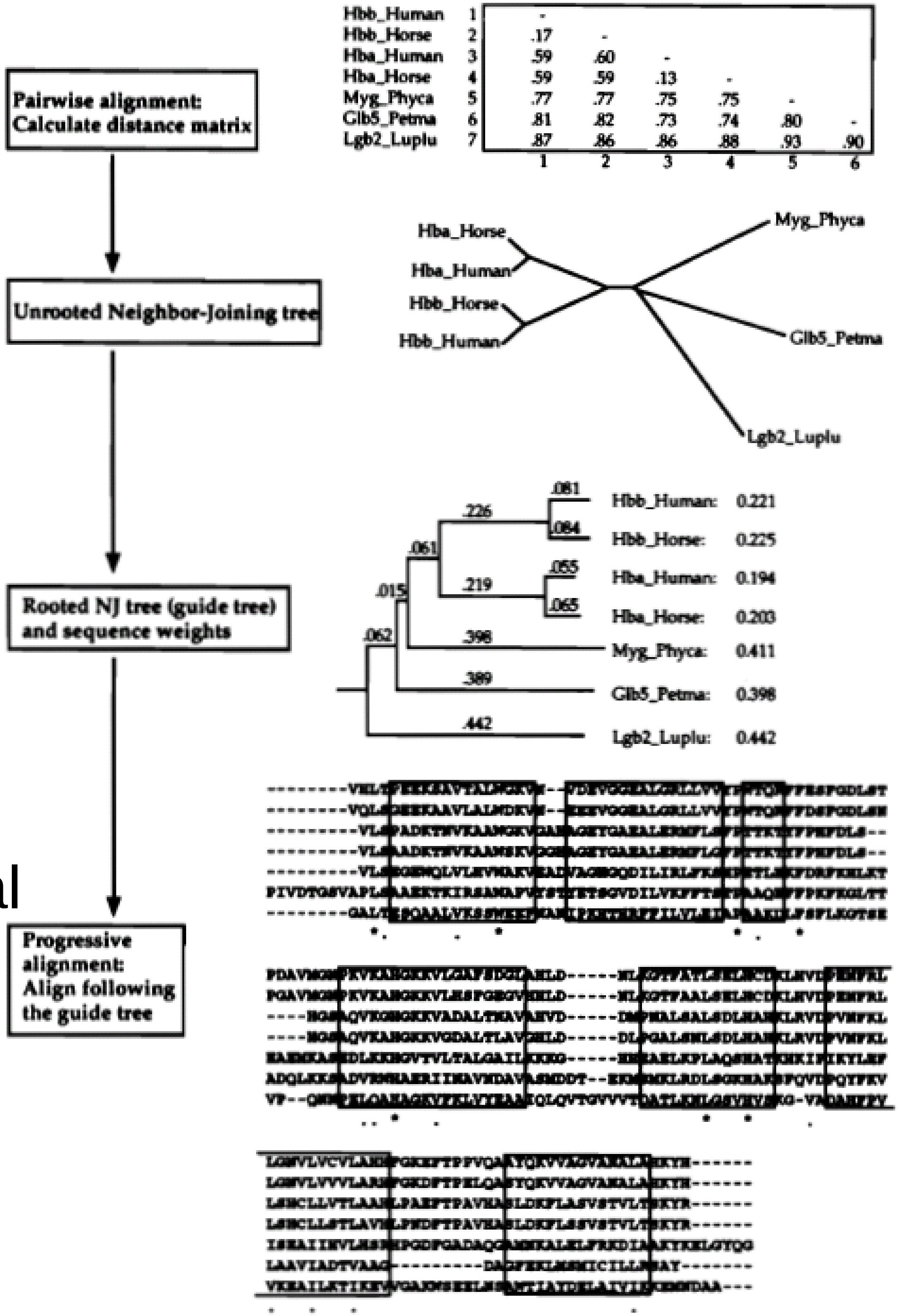


Overview of ClustalW:

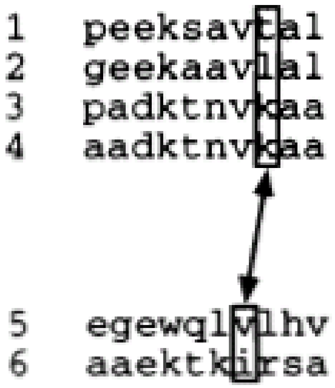
1.Get pairwise “distances”
2.Determine tree
3.Follow order of tree to do
pairwise alignments

After each step the alignment is fixed. This generates a complete multiple alignment of the sequences using optimal pairwise alignments (with DP) at each step.

Scoring is SoP with heuristic Modifications (next slide).



Hbb_Human	1	-					
Hbb_Horse	2	.17	-				
Hba_Human	3	.59	.60	-			
Hba_Horse	4	.59	.59	.13	-		
Myg_Phyca	5	.77	.77	.75	.75	-	
Glb5_Petma	6	.81	.82	.73	.74	.80	-
Lgb2_Luplu	7	.87	.86	.86	.88	.93	.90
		1	2	3	4	5	6



Without sequence Weights:

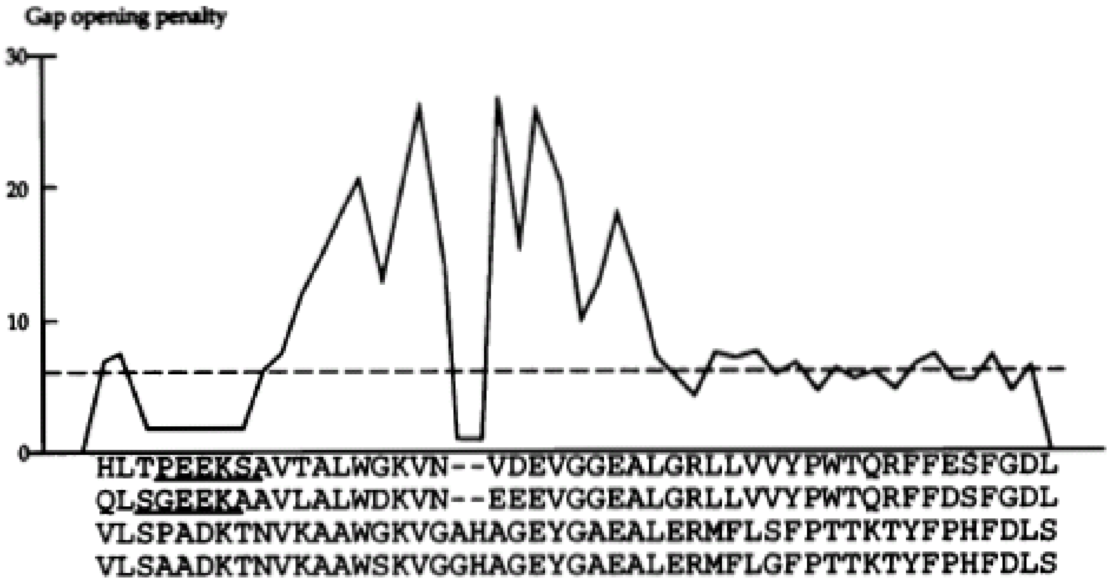
Score = M(t,v)
+ M(t,i)
+ M(l,v)
+ M(l,i)
+ M(k,v)
+ M(k,i)
+ M(k,v)
+ M(k,i)/8

With sequence Weights W_i:

Score = $M(t,v)*W_1*W_5$
+ $M(t,i)*W_1*W_6$
+ $M(l,v)*W_2*W_5$
+ $M(l,i)*W_2*W_6$
+ $M(k,v)*W_3*W_5$
+ $M(k,i)*W_3*W_6$
+ $M(k,v)*W_4*W_5$
+ $M(k,i)*W_4*W_6/8$

Sequence weighting: Based on shared tree lengths, avoids problems from overly biased samples

Gap penalty adjustment: Increases/reduces gap opening penalty depending on local alignment features; New gaps cluster with previous ones, and in hydrophylic regions



Thought Exercise

- Consider two sets of proteins: {A,B,C}, {X,Y,Z}
- Within each set, any pair of proteins is ~15% identical which puts them in the “twilight zone” where is it difficult to determine if they are homologs or not (or equivalently their E-values are ~1)
- Set {A,B,C} proteins are unrelated
- Set {X,Y,Z} proteins are homologous
    - What differences do you expect in their alignments?
    - How would the multiple alignments differ between the two?

Multiple Alignment Lecture 2

Improved Progressive Alignments

- Faster
- More accurate
- Consistency objective

Alternative scoring systems

Position-specific scoring (Profiles)

Probabilistic modeling: Profile-HMMs

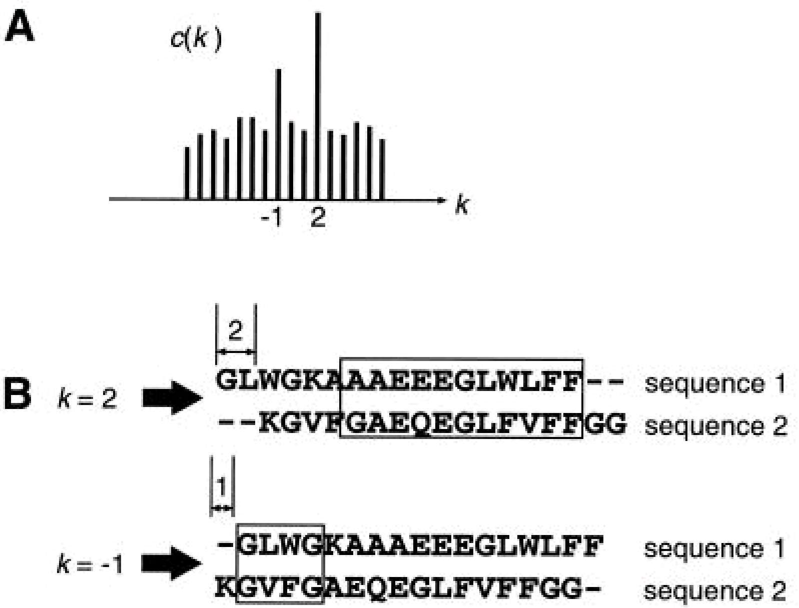
More recent improved methods Faster and/or more accurate

- *See recent reviews by:*
  - *Edgar and Batzoglou, Current Opin. Struct. Biol. (2006) 16:368-373*
  - *Notredame, PLoS Comp Biol. (2007) 3:e123*

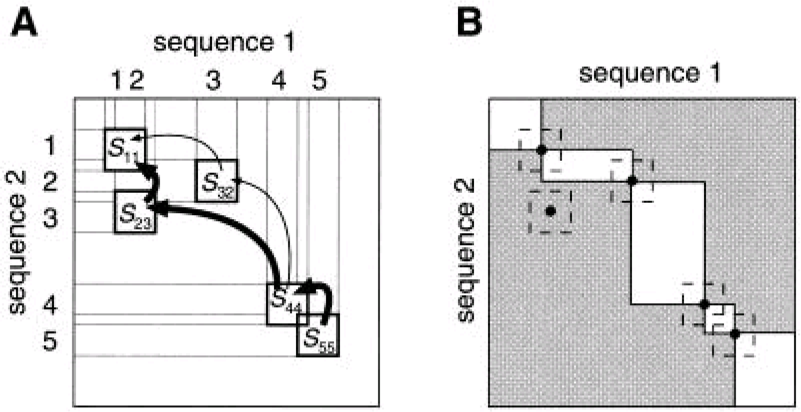
- FFT for speed; combine local and global alignments; iterative refinements; use additional types of information (such as structure) if available; maximize consistency with pairwise alignments

MAFFT – multiple alignment using Fast Fourier Transform, Katoh et al., Nucleic Acids Res. 30:3059-3066 (2002)

- Recode aa sequence into lists of properties (eg. volume, polarity)
- Considering all possible shifts of ungapped sequences, identify the shifts with high similarity
- Can be computed in O(L*ln*L) time instead of O(L^2)

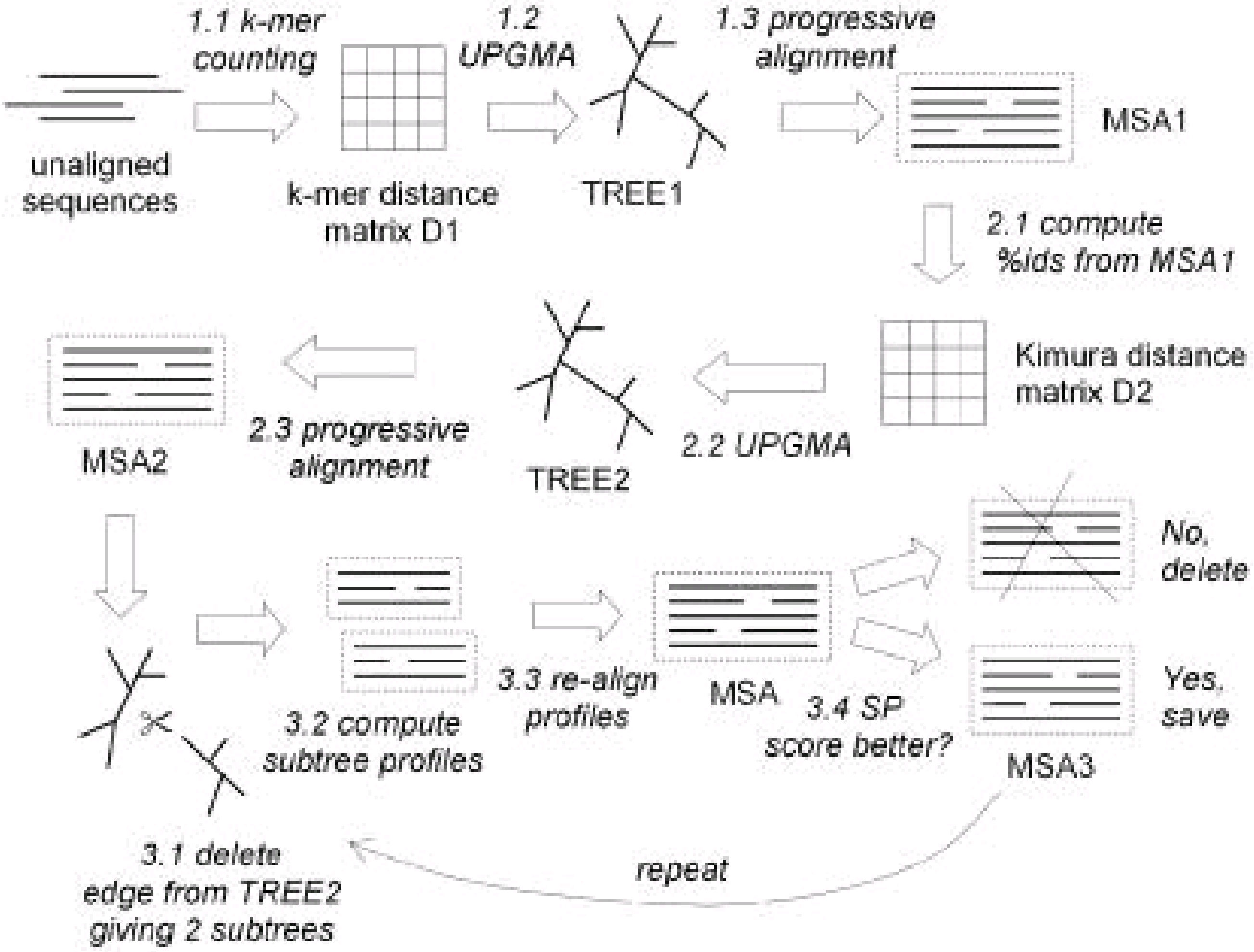


- Gives locally aligned, ungapped segments
- Can be “stitched” together with DP to give global alignment



- The order of pairwise alignments is still based on a guide tree
- The whole process can be iterated to refine the alignment
  -At each iteration the alignment from the previous iteration is used for the guide tree, and the overall alignment can be broken into pieces that are optimized separately

MUSCLE: a multiple sequence alignment method with reduced time and space complexity, RC Edgar, *BMC Bioinformatics,2004, 5:113*


1.1 k-mer counting
1.2 UPGMA
1.3 progressive alignment
unaligned sequences
k-mer distance matrix D1
TREE1
MSA1
2.1 compute %ids from MSA1
Kimura distance matrix D2
2.2 UPGMA
TREE2
2.3 progressive alignment
MSA2
3.1 delete edge from TREE2 giving 2 subtrees
3.2 compute subtree profiles
3.3 re-align profiles
MSA
3.4 SP score better?
No, delete
Yes, save
MSA3
repeat

If only first 2 steps: O($N^2L + NL^2$)

If third refinement step is included: O(N^3L)

Avoids first step, all-by-all alignment from ClustalW, which is O(N^2L^2)

An alternative scoring system (objective function)

- **Maximize consistency in multiple alignment with each of the optimal pairwise alignments**

- **Basic idea: given three sequences A, B, C**

  Pairwise alignments of A:B and B:C

  infers an alignment of A:C

  How well does that match the pairwise

  alignment of A:C ?

Goal: Find most consistent multiple alignment.

Outline of T-Coffee

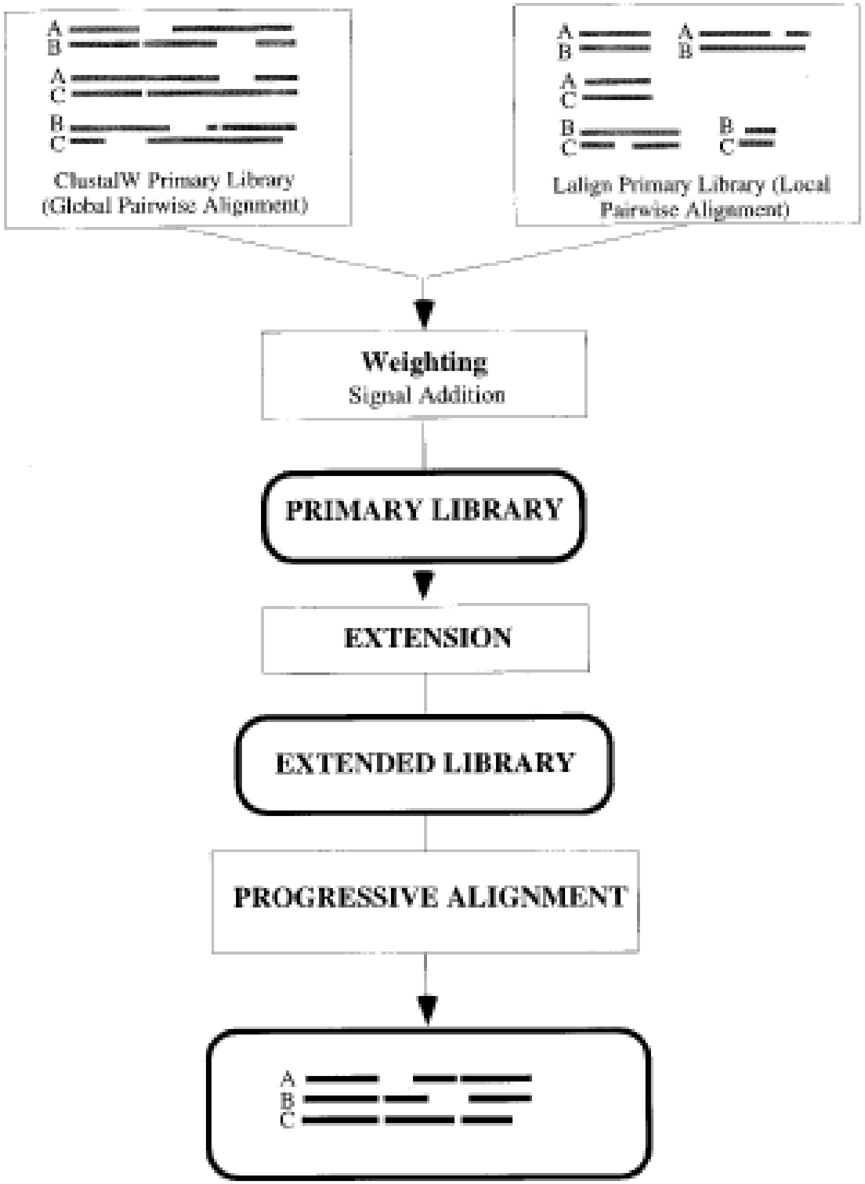


Figure 1. Layout of the T-Coffee strategy; the main steps required to compute a multiple sequence alignment using the T-Coffee method. Square blocks designate procedures while rounded blocks indicate data structures.

ProbCons: Probabilistic consistency-based multiple sequence alignment

Chuong B. Do, Mahathi S.P. Mahabhashyam, Michael Brudno and Serafim Batzoglou

Genome Res. 2005 15: 330-340

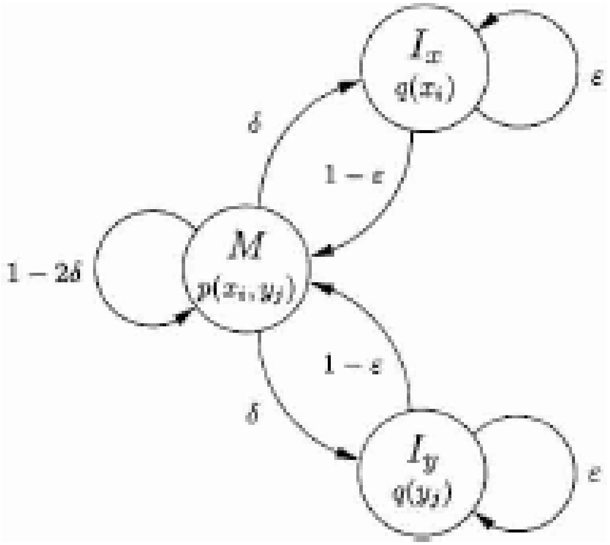


Figure 1. Basic pair-HMM for sequence alignment between two sequences, *x* and *y*. State *M* emits two letters, one from each sequence, and corresponds to the two letters being aligned together. State I_x emits a letter in sequence *x* that is aligned to a gap, and similarly state I_y emits a letter in sequence *y* that is aligned to a gap. Finding the most likely alignment according to this model by using the Viterbi algorithm corresponds to applying Needleman–Wunsch with appropriate parameters. The logarithm of the emission probability function $p(.,.)$ at *M* corresponds to a substitution scoring matrix, while affine gap penalty parameters can be derived from the transition probabilities δ and ε (Durbin et al. 1998).

ProbsCon details:

1. Pairwise alignment probabilities for all pairs of sequences; forward-backward using a similarity matrix (BLOSSUM62)
2. Find maximum *expected accuracy* alignment; i.e. alignment with maximum number of expected correct aligned pairs
3. Probabilistic consistency transform; find highest accuracy alignment of X:Y by $\sum_Z \sum_k P(x_i:z_k)P(y_j:z_k)$
4. Guide tree determination based on expected accuracy
5. Progressive alignment based on expected accuracy

Refinement can be done at the end if desired

Revisit the scoring system issue

- Sum-of-Pairs (SoP) assumes a single similarity matrix is appropriate for all positions – the same as for pair-wise alignments
- Want to have a position specific scoring matrix (PSSM) – Profiles implement this using SoP

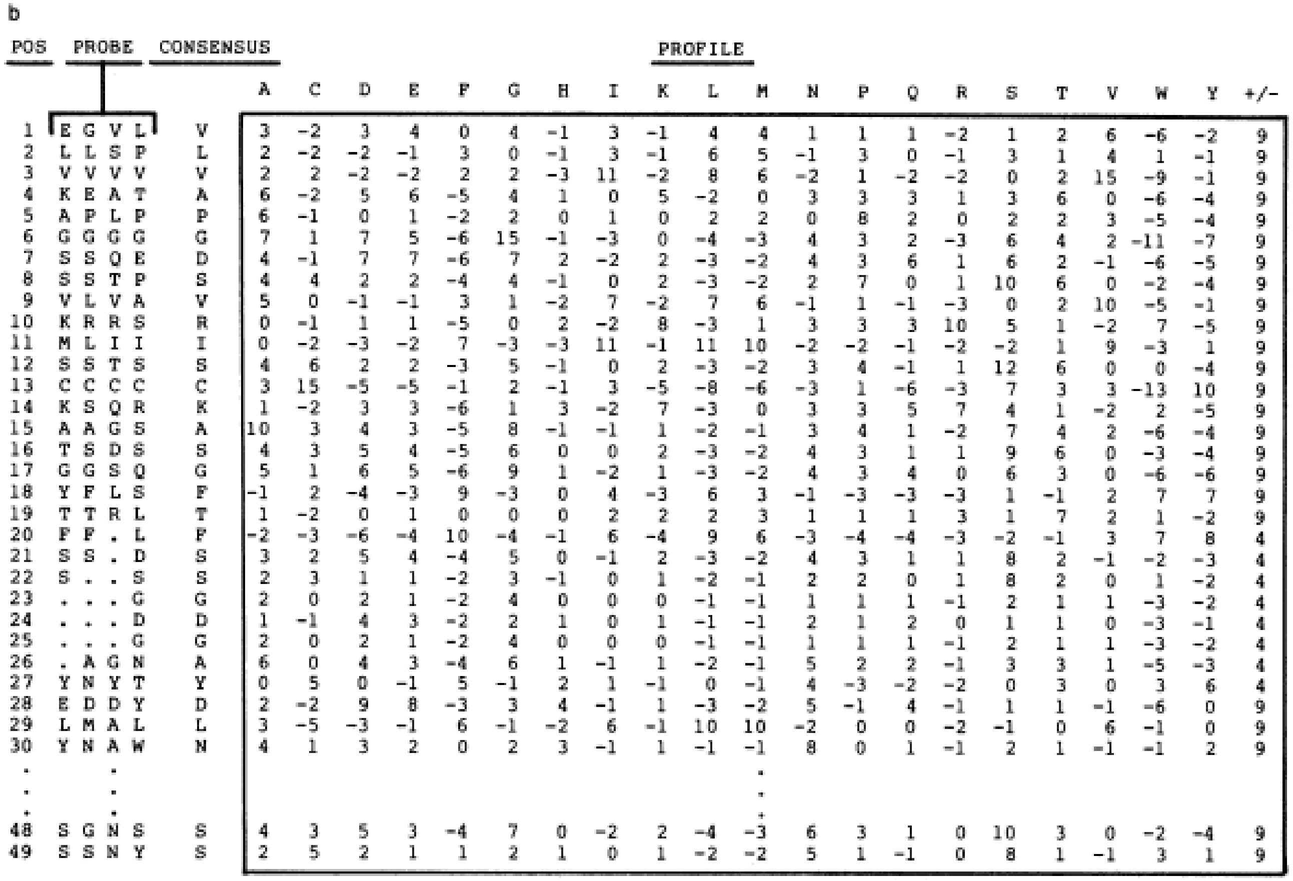
- HMM-profiles provide probabilistic scoring that is position specific

Proc. Natl. Acad. Sci. USA
Vol. 84, pp. 4355–4358, July 1987
Biochemistry

Profile analysis: Detection of distantly related proteins

(amino acid/sequence comparison/protein structure/globin structure/immunoglobulin structure)

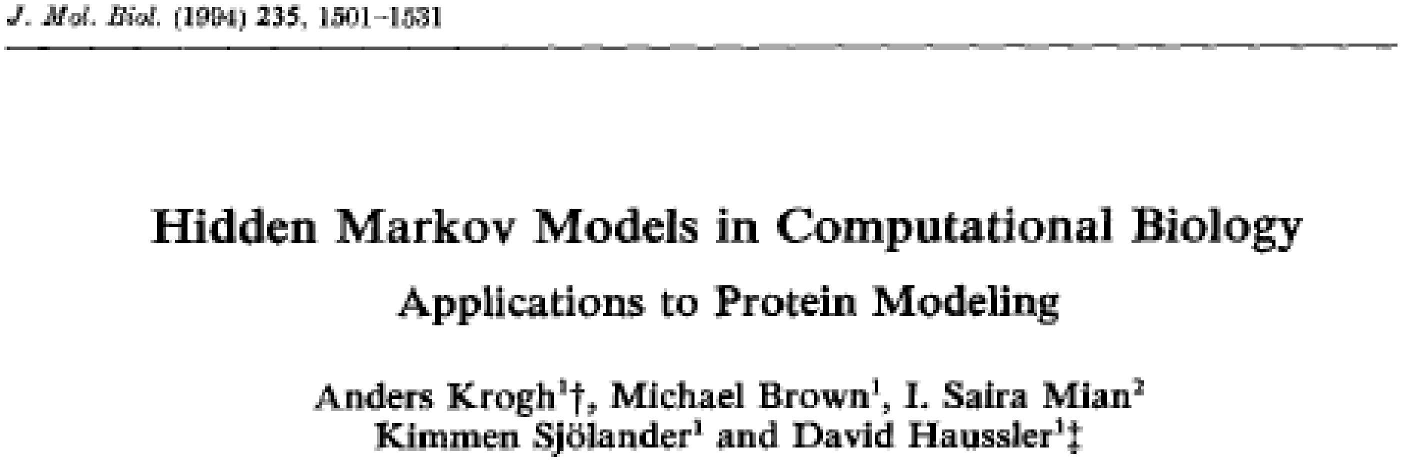
MICHAEL GRIBSKOV*, ANDREW D. MCLACHLAN†, AND DAVID EISENBERG*



b

POS	PROBE	CONSENSUS									PROFILE											
			A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	+/-
1	E G V L	V	3	-2	3	4	0	4	-1	3	-1	4	4	1	1	1	-2	1	2	6	-6	-2	9
2	L L S P	L	2	-2	-2	-1	3	0	-1	3	-1	6	5	-1	3	0	-1	3	1	4	1	-1	9
3	V V V V	V	2	2	-2	-2	2	2	-3	11	-2	8	6	-2	1	-2	-2	0	2	15	-9	-1	9
4	K E A T	A	6	-2	5	6	-5	4	1	0	5	-2	0	3	3	3	1	3	6	0	-6	-4	9
5	A P L P	P	6	-1	0	1	-2	2	0	1	0	2	2	0	8	2	0	2	2	3	-5	-4	9
6	G G G G	G	7	1	7	5	-6	15	-1	-3	0	-4	-3	4	3	2	-3	6	4	2	-11	-7	9
7	S S Q E	D	4	-1	7	7	-6	7	2	-2	2	-3	-2	4	3	6	1	6	2	-1	-6	-5	9
8	S S T P	S	4	4	2	2	-4	4	-1	0	2	-3	-2	2	7	0	1	10	6	0	-2	-4	9
9	V L V A	V	5	0	-1	-1	3	1	-2	7	-2	7	6	-1	1	-1	-3	0	2	10	-5	-1	9
10	K R R S	R	0	-1	1	1	-5	0	2	-2	8	-3	1	3	3	3	10	5	1	-2	7	-5	9
11	M L I I	I	0	-2	-3	-2	7	-3	-3	11	-1	11	10	-2	-2	-1	-2	-2	1	9	-3	1	9
12	S S T S	S	4	6	2	2	-3	5	-1	0	2	-3	-2	3	4	-1	1	12	6	0	0	-4	9
13	C C C C	C	3	15	-5	-5	-1	2	-1	3	-5	-8	-6	-3	1	-6	-3	7	3	3	-13	10	9
14	K S Q R	K	1	-2	3	3	-6	1	3	-2	7	-3	0	3	3	5	7	4	1	-2	2	-5	9
15	A A G S	A	10	3	4	3	-5	8	-1	-1	1	-2	-1	3	4	1	-2	7	4	2	-6	-4	9
16	T S D S	S	4	3	5	4	-5	6	0	0	2	-3	-2	4	3	1	1	9	6	0	-3	-4	9
17	G G S Q	G	5	1	6	5	-6	9	1	-2	1	-3	-2	4	3	4	0	6	3	0	-6	-6	9
18	Y F L S	F	-1	2	-4	-3	9	-3	0	4	-3	6	3	-1	-3	-3	-3	1	-1	2	7	7	9
19	T T R L	T	1	-2	0	1	0	0	0	2	2	2	3	1	1	1	3	1	7	2	1	-2	9
20	F F . L	F	-2	-3	-6	-4	10	-4	-1	6	-4	9	6	-3	-4	-4	-3	-2	-1	3	7	8	4
21	S S . D	S	3	2	5	4	-4	5	0	-1	2	-3	-2	4	3	1	1	8	2	-1	-2	-3	4
22	S . . S	S	2	3	1	1	-2	3	-1	0	1	-2	-1	2	2	0	1	8	2	0	1	-2	4
23	. . . G	G	2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
24	. . . D	D	1	-1	4	3	-2	2	1	0	1	-1	-1	2	1	2	0	1	1	0	-3	-1	4
25	. . . G	G	2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
26	. A G N	A	6	0	4	3	-4	6	1	-1	1	-2	-1	5	2	2	-1	3	3	1	-5	-3	4
27	Y N Y T	Y	0	5	0	-1	5	-1	2	1	-1	0	-1	4	-3	-2	-2	0	3	0	3	6	4
28	E D D Y	D	2	-2	9	8	-3	3	4	-1	1	-3	-2	5	-1	4	-1	1	1	-1	-6	0	9
29	L M A L	L	3	-5	-3	-1	6	-1	-2	6	-1	10	10	-2	0	0	-2	-1	0	6	-1	0	9
30	Y N A W	N	4	1	3	2	0	2	3	-1	1	-1	-1	8	0	1	-1	2	1	-1	-1	2	9
⋮	⋮												⋮										
48	S G N S	S	4	3	5	3	-4	7	0	-2	2	-4	-3	6	3	1	0	10	3	0	-2	-4	9
49	S S N Y	S	2	5	2	1	1	2	1	0	1	-2	-2	5	1	-1	0	8	1	-1	3	1	9

Profile HMMs



J. Mol. Biol. (1994) 235, 1501–1531

Hidden Markov Models in Computational Biology

Applications to Protein Modeling

Anders Krogh[1]†, Michael Brown[1], I. Saira Mian[2]
Kimmen Sjölander[1] and David Haussler[1]‡

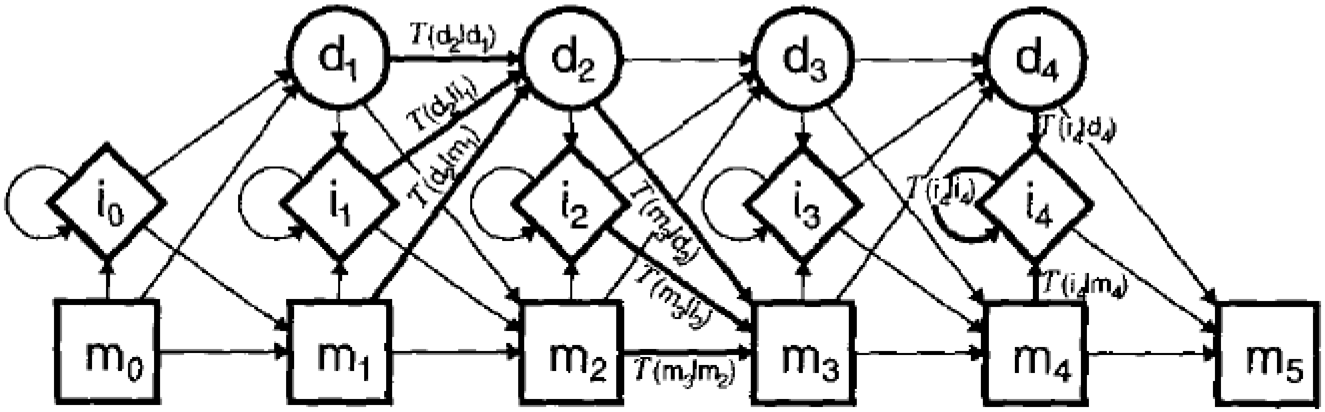
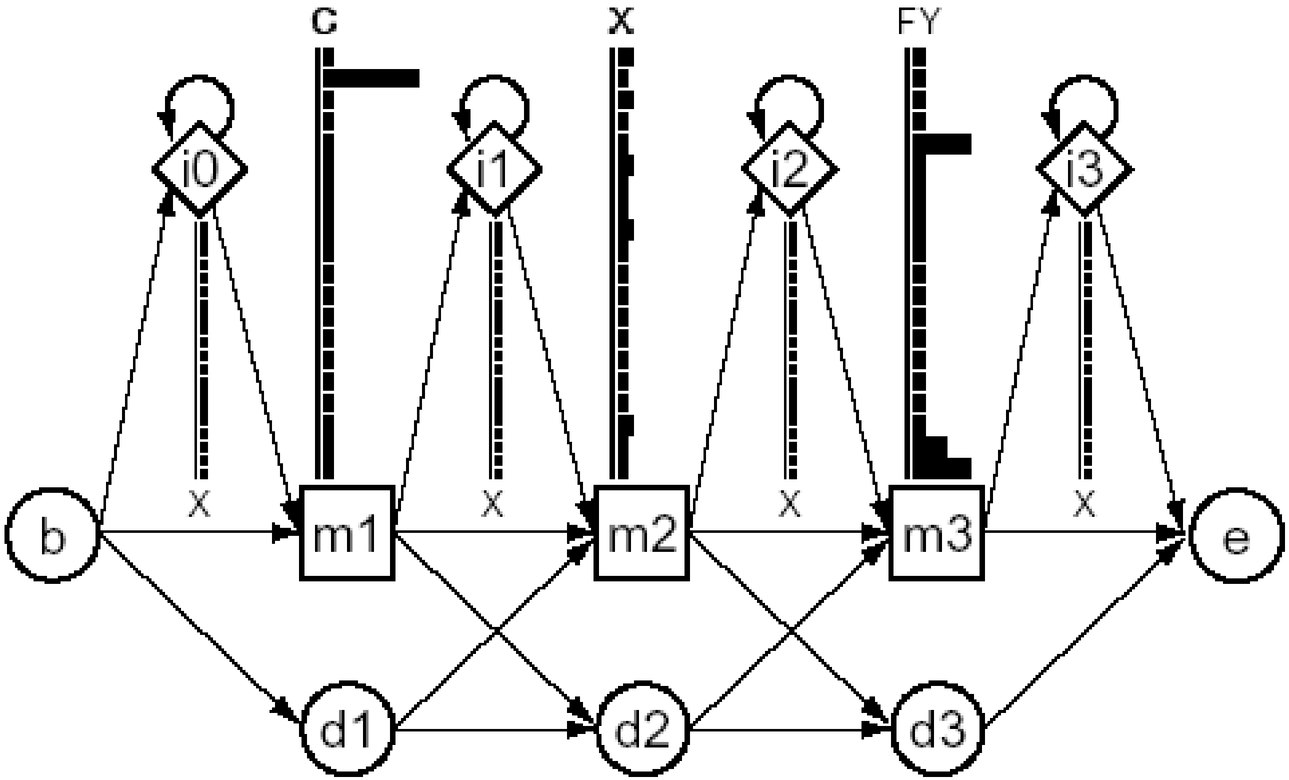


Figure 1. The model.

1	**2**	3
C	A	F
C	G	W
C	D	Y
C	V	F
C	K	**Y**



Review: “Profile hidden Markov models”
by Eddy SR. *Bioinformatics.* 1998;14(9):755-63.

HMM-Profiles:

- Given an alignment, can estimate parameters
  - Emission Probabilities
  - Transition probabilities
    - Pfam database of HMM-profiles

      www.sanger.ac.uk/Software/**Pfam**/
- Given an HMM and another sequence, can find best alignment by Viterbi (i.e. DP)

- Can iterate between those steps (EM):

  start with unaligned sequences and end up with an alignment and a model that represents the family

Limitations: over-fitting from small sample sizes

use of priors can help
choice of model architecture, refinement
weighting of sequence contributions

Parameters obtained from an alignment

- All of the transition and emission probabilities can be obtained from the alignment just by “counting” how often each occurs
- Need a large sample size to estimate all of the parameters accurately
- Can add pseudocounts to avoid 0’s
  - Laplace “add 1” rule is common
- Can use more complex priors (Dirichlet) that differ for different residues and even mixtures of Dirichlet priors

Find best alignment of a sequence to an HMM

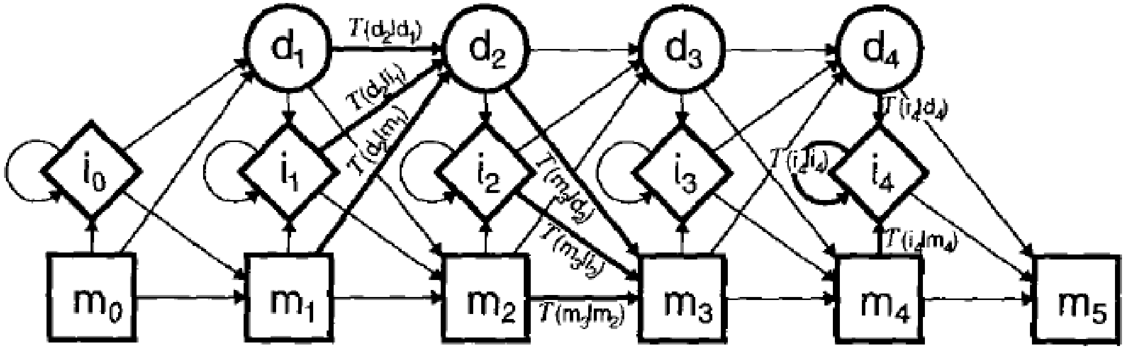


Figure 1. The model.

Viterbi algorithm

$$V_j^M(i) = \log \frac{e_{M_j}(x_i)}{q_{x_i}} + \max \begin{cases} V_{j-1}^M(i-1) + \log a_{M_{j-1}M_j} \\ V_{j-1}^I(i-1) + \log a_{I_{j-1}M_j} \\ V_{j-1}^D(i-1) + \log a_{D_{j-1}M_j} \end{cases}$$

$$V_j{}^I(i) = \log \frac{e_{I_j}(x_i)}{q_{x_i}} + \max \begin{cases} V_{j-1}^M(i-1) + \log a_{M_{j-1}I_j} \\ V_{j-1}^I(i-1) + \log a_{I_{j-1}I_j} \\ V_{j-1}^D(i-1) + \log a_{D_{j-1}I_j} \end{cases}$$

$$V_j^D(i) = \max \begin{cases} V_{j-1}^M(i-1) + \log a_{M_{j-1}D_j} \\ V_{j-1}^I(i-1) + \log a_{I_{j-1}D_j} \\ V_{j-1}^D(i-1) + \log a_{D_{j-1}D_j} \end{cases}$$

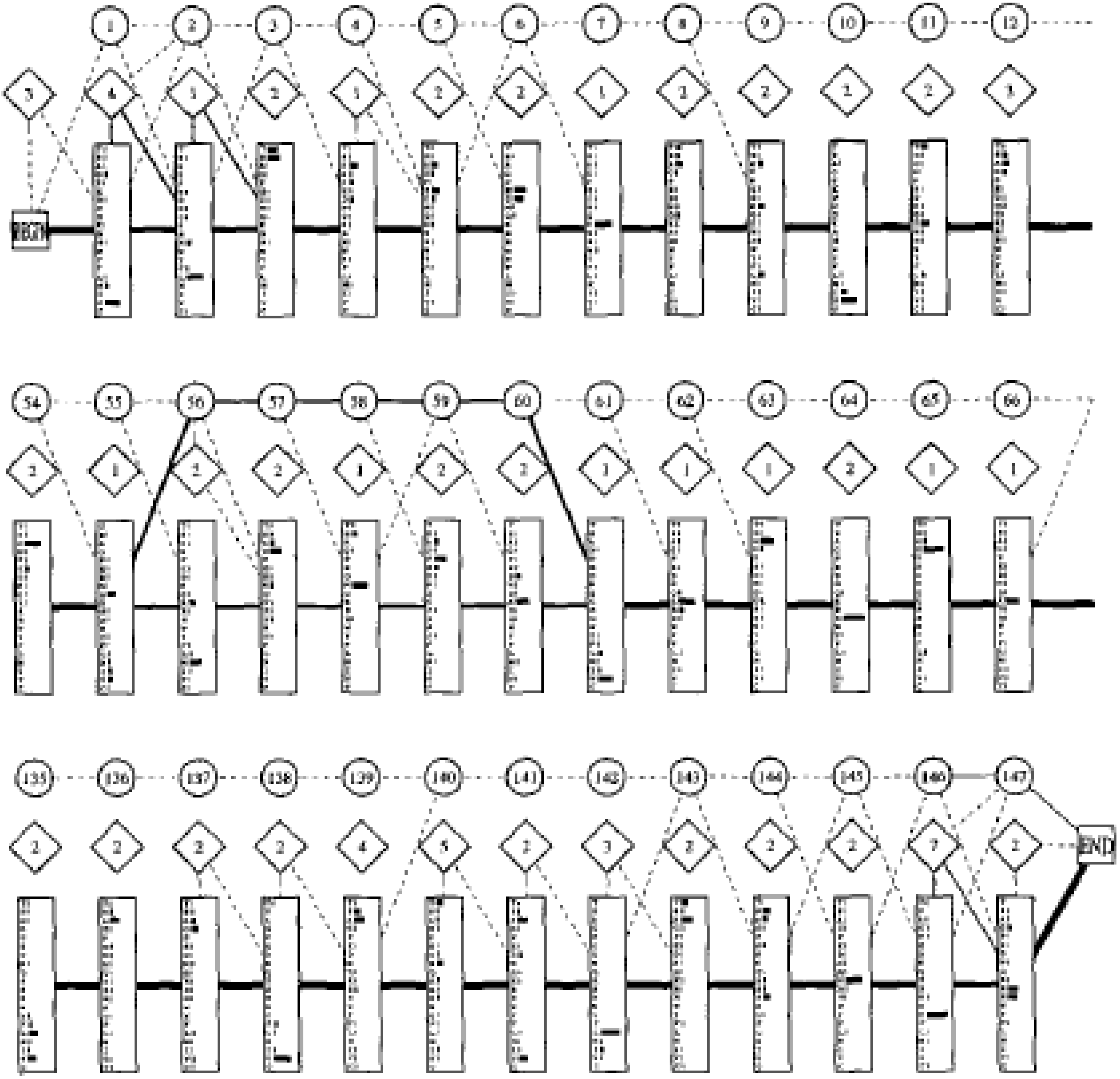


Figure 8. Parts of the final globin model. The position numbers are shown in the delete states.

Find probability that a sequence is “generated” by an HMM

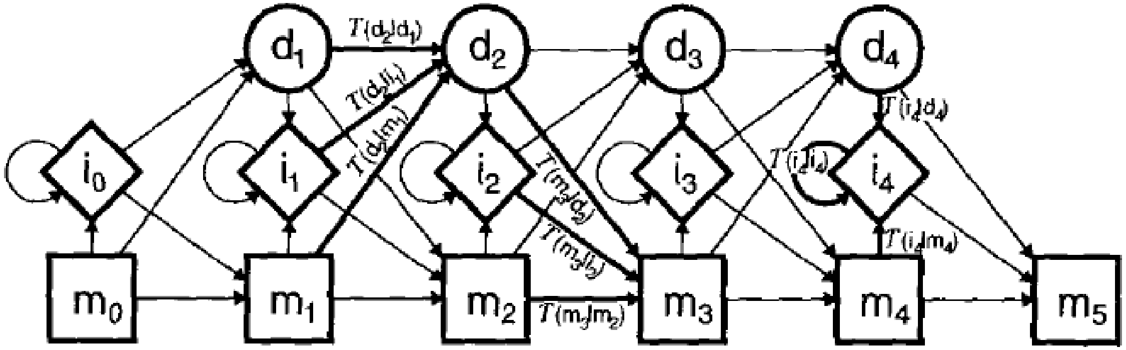


Figure 1. The model.

Forward algorithm

$$F_j^M(i) = \log \frac{e_{M_j}(x_i)}{q_{x_i}} + \log \begin{cases} a_{M_{j-1}M_j} \exp\left(F_{j-1}^M(i-1)\right) \\ +a_{I_{j-1}M_j} \exp\left(F_{j-1}^I(i-1)\right) \\ +a_{D_{j-1}M_j} \exp\left(F_{j-1}^D(i-1)\right) \end{cases}$$

$$F_j^I(i) = \log \frac{e_{I_j}(x_i)}{q_{x_i}} + \log \begin{cases} a_{M_{j-1}I_j} \exp\left(F_{j-1}^M(i-1)\right) \\ +a_{I_{j-1}I_j} \exp\left(F_{j-1}^I(i-1)\right) \\ +a_{D_{j-1}I_j} \exp\left(F_{j-1}^D(i-1)\right) \end{cases}$$

$$F_j^D(i) = \log \begin{cases} a_{M_{j-1}D_j} \exp\left(F_{j-1}^M(i-1)\right) \\ +a_{I_{j-1}D_j} \exp\left(F_{j-1}^I(i-1)\right) \\ +a_{D_{j-1}D_j} \exp\left(F_{j-1}^D(i-1)\right) \end{cases}$$

Acknowledgement

Most of the slides in this chapter were provided by Prof. Gary Stormo.