Multiple Alignments: CLUSTALW

* identical

: conserved substitutions

. semi-conserved substitutions

```
gi|2213819   CDN-ELKSEAIIEHLCASEFALR------------MKIKEVKKENGDKK 223
gi|12656123  ----ELKSEAIIEHLCASEFALR------------MKIKEVKKENGD-- 31
gi|7512442   CKNKNDDDNDIMETLCKNDFALK------------IKVKEITYINRDTK 211
gi|1344282   QDECKFDYVEVYETSSSGAFSLLGRFCGAEPPPHLVSSHHELAVLFRTDH 400
                 : .   : *  . . *:*               . :*:
```

Red: AVFPMLW (Small & hydrophobic)

Blue: DE (Acidic)

Magenta: RHK (Basic)

Green: STYHCNGQ (Hydroxyl, Amine, Basic)

Gray: Others

Multiple Alignments

- **Family alignment for the ITAM domain (Immunoreceptor tyrosine-based activation motif)**

```
CD3D_MOUSE/1-2   EQLYQPLRDR EDTQ-YSRLG GN
Q90768/1-21      DQLYQPLGER NDGQ-YSQLA TA
CD3G_SHEEP/1-2   DQLYQPLKER EDDQ-YSHLR KK
P79951/1-21      NDLYQPLGQR SEDT-YSHLN SR
FCEG_CAVPO/1-2   DGIYTGLSTR NQET-YETLK HE
CD3Z_HUMAN/3-0   DGLYQGLSTA TKDT-YDALH MQ
C79A_BOVIN/1-2   ENLYEGLNLD DCSM-YEDIS RG
C79B_MOUSE/1-2   DHTYEGLNID QTAT-YEDIV TL
CD3H_MOUSE/1-2   NQLYNELNLG RREE-YDVLE KK
CD3Z_SHEEP/1-2   NPVYNELNVG RREE-YAVLD RR
CD3E_HUMAN/1-2   NPDYEPIRKG QRDL-YSGLN QR
CD3H_MOUSE/2-0   EGVYNALQKD KMAEAYSEIG TK
Consensus/60%    -.lYpsLspc pcsp.YspLs pp
```

Simple Modular Architecture Research Tool

Multiple Alignment

How to Score Multiple Alignments?

- Sum of Pairs Score (SP)
  - Optimal alignment: $O(d^N)$ [Dynamic Prog]
  - Approximate Algorithm: Approx Ratio 2
    - Locate Center: $O(d^2N^2)$
    - Locate Consensus: $O(d^2N^2)$

Consensus char: char with min distance sum

Consensus string: string of consensus char

Center: input string with min distance sum

Multiple Alignment Methods

- Phylogenetic Tree Alignment (NP-Complete)
  - Given tree, task is to label leaves with strings
- Iterative Method(s)
  - Build a MST using the distance function
- Clustering Methods
  - Hierarchical Clustering
  - K-Means Clustering

Multiple Alignment Methods (Cont'd)

- Gibbs Sampling Method
  - Lawrence, Altschul, Boguski, Liu, Neuwald, Winton, *Science,* 1993
- Hidden Markov Model
  - Krogh, Brown, Mian, Sjolander, Haussler, *JMB,* 1994

Multiple Sequence Alignments (MSA)

- Choice of Scoring Function
  - Global vs local
  - Gap penalties
  - Substitution matrices
  - Incorporating other information
  - Statistical Significance
- Computational Issues
  - Exact/heuristic/approximate algorithms for optimal MSA
  - Progressive/Iterative/DP
  - Iterative: Stochastic/Non-stochastic/Consistency-based
- Evaluating MSAs
  - Choice of good test sets or benchmarks (BAliBASE)
  - How to decide thresholds for good/bad alignments

Progressive MSA: CLUSTALW

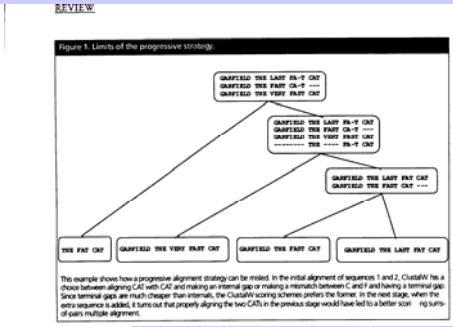

REVIEW

Figure 1. Limits of the progressive strategy.

This example shows how a progressive alignment strategy can be misled. In the initial alignment of sequences 1 and 2, ClustalW has a choice between aligning CAT with CAT and making an internal gap or making a mismatch between C and F and having a terminal gap. Since terminal gaps are much cheaper than internals, the ClustalW scoring schemes prefers the former. In the next stage, when the extra sequence is added, it turns out that properly aligning the two CATs in the previous stage would have led to a better scori ng sums-of-pairs multiple alignment.

C. Notredame, *Pharmacogenomics*, **3**(1), 2002.

Software for MSA

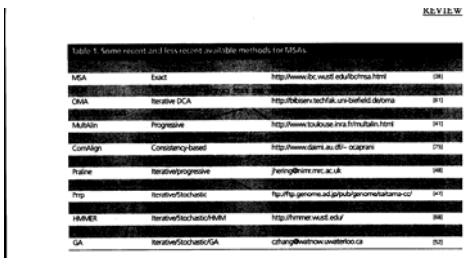

REVIEW

Table 1. Some recent and less recent available methods for MSAs

MSA	Exact	http://www.ibc.wustl.edu/tcf/msa.html	[39]
OMA	Iterative DCA	http://bibserv.techfak.uni-bielefeld.de/oma	[41]
MultAlin	Progressive	http://www.toulouse.inra.fr/multalin.html	[41]
ConAlign	Consistency-based	http://www.daimi.au.dk/~ocaprani	[76]
Praline	Iterative/progressive	jhering@mrc.mrc.ac.uk	[48]
Pmp	Iterative/Stochastic	ftp://ftp.genome.ad.jp/pub/genome/saitama-cc/	[41]
HMMER	Iterative/Stochastic/HMM	http://hmmer.wustl.edu/	[66]
GA	Iterative/Stochastic/GA	czhang@watnow.uwaterloo.ca	[52]

C. Notredame, *Pharmacogenomics*, **3**(1), 2002.

MSA: Conclusions

- Very important
  - Phylogenetic analyses
  - Identify members of a family
  - Protein structure prediction
- No perfect methods
- Popular
  - Progressive methods: CLUSTALW
  - Recent interesting ones: Prrp, SAGA, DiAlign, T-Coffee
- Review of Methods [C. Notredame, *Pharmacogenomics,* **3**(1), 2002]
  - CLUSTALW works reasonably well, in general
  - DiAlign is better for sequences with long insertions & deletions (indels)
  - T-Coffee is best available method

Profile Method

PROFILE METHOD, [M. Gribskov et al., '90]

Location in Seq.	Sequence							Protein Name
	1	2	3	4	5	6	7	
14	G	V	S	A	S	A	V	Ka RbtR
32	G	V	S	E	M	T	I	Ec DeoR
33	G	V	S	P	G	T	I	Ec RpoD
76	G	A	G	I	A	T	I	Ec TrpR
178	G	C	S	R	E	T	V	Ec CAP
205	C	L	S	P	S	R	L	Ec AraC
210	C	L	S	P	S	R	L	St AraC
13	G	V	N	K	E	T	I	Br MerR

FREQUENCY TABLE

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

Profile Method

FREQUENCY TABLE

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

WEIGHT MATRIX

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1	0	106	0	101	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	21	78	0	0	0	0	0	0	0	44	0	0	0	0	0	0	0	0	0	
3	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	102	
4	21	0	32	0	0	0	0	38	33	0	0	0	0	0	0	86	39	0	0	
5	21	0	0	62	0	0	0	0	0	0	0	0	74	0	0	0	0	0	72	
6	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
7	0	0	0	0	0	0	0	99	0	0	0	0	0	0	0	0	0	0	0	

$$W_{\text{weight}}[i, A.A] = \log\left(\frac{\text{Freq}[i, A.A]}{\text{Background}[A.A]}\right) \cdot 100$$

Profile Method

WEIGHT MATRIX

	A	C	E	G	I	K	L	M	P	S	D
1	0	108	0	101	0	0	0	0	0	0	0
2	21	78	0	0	0	0	44	0	0	0	0
3	0	0	0	23	0	0	0	0	46	0	0	102
4	21	0	32	0	38	32	0	0	0	86	39	0
5	21	0	62	23	0	0	0	74	0	0	0	72
6	21	0	0	0	0	0	0	0	0	0	69	0
7	0	0	0	0	98	0	44	0	0	0	0	0

Given the following protein sequence:

M T E D L F G D L Q D D T I L A H L D N
P A E D T S R F P A L L A E L N D L L R
G E L S R L G V D P A H S L E I V V A I
C K H L G G G Q V Y I P R G Q A L D S L
I R D L R I W N D F N G R N V S E L T T
R Y G V T F N T V Y K A I R R M R R L K

9

CpG Islands

- Regions in DNA sequences with increased occurrences of substring "*CG*"
- Rare: typically *C* gets methylated and then mutated into a T.
- Often around promoter or "start" regions of genes
- Few hundred to a few thousand bases long

Problem 1:

- Input: Small sequence S
- Output: Is S from a CpG island?
  - Build Markov models: M+ and M —
  - Then compare

Markov Models

+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

−	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

How to distinguish?

- Compute

$$S(x) = \log\left(\frac{P(x \mid M+)}{P(x \mid M-)}\right) = \sum_{i=1}^{L} \log\left(\frac{p_{x(i-1)xi}}{m_{x(i-1)xi}}\right) = \sum_{i=1}^{L} r_{x(i-1)xi}$$

r=p/m	A	C	G	T
A	-0.740	0.419	0.580	-0.803
C	-0.913	0.302	1.812	-0.685
G	-0.624	0.461	0.331	-0.730
T	-1.169	0.573	0.393	-0.679

Score(GCAC)
= .461-.913+.419
< 0.
GCAC not from CpG island.

Score(GCTC)
= .461-.685+.573
> 0.
GCTC from CpG island.

Problem 1:

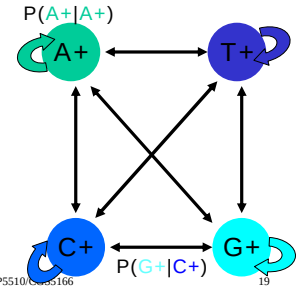
- Input: Small sequence S
- Output: Is S from a CpG island?
  - Build Markov Models: M+ & M-
  - Then compare

Problem 2:

- Input: Long sequence S
- Output: Identify the CpG islands in S.
  - Markov models are inadequate.
  - Need Hidden Markov Models.

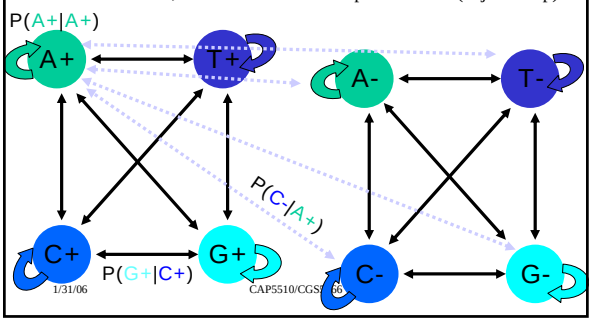
Markov Models

+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182



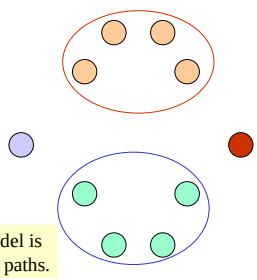
CpG Island + in an ocean of – First order Hidden Markov Model

MM=16, HMM= 64 transition probabilities (adjacent bp)



Hidden Markov Model (HMM)

- States
- Transitions
- Transition Probabilities
- Emissions
- Emission Probabilities



- What is hidden about HMMs?

Answer: The path through the model is hidden since there are many valid paths.

How to Solve Problem 2?

- Solve the following problem:

Input: Hidden Markov Model M,
parameters Θ, emitted sequence S

Output: Most Probable Path Π

How: Viterbi's Algorithm (Dynamic Programming)

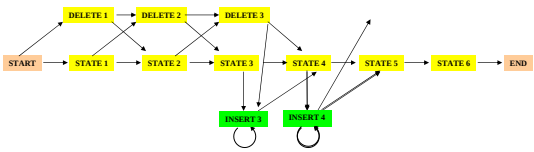
Define Π[i,j] = MPP for first j characters of S ending in state i

Define P[i,j] = Probability of Π[i,j]

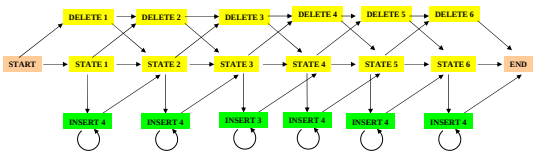
- **Compute state i with largest P[i,j].**

Profile HMMs with InDels

- Insertions
- Deletions
- Insertions & Deletions



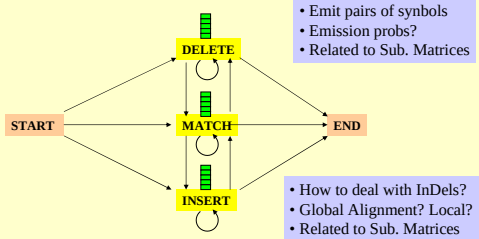
Profile HMMs with InDels



Missing transitions from DELETE j to INSERT j and from INSERT j to DELETE j+1.

How to model Pairwise Sequence Alignment

LEAPVE
LAPVIE



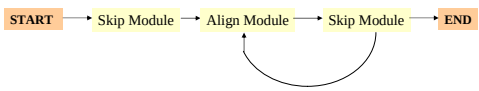
Pair HMMs
- Emit pairs of synbols
- Emission probs?
- Related to Sub. Matrices

- How to deal with InDels?
- Global Alignment? Local?
- Related to Sub. Matrices

How to model Pairwise Local Alignments?

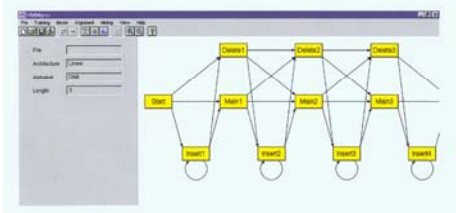
START → Skip Module → Align Module → Skip Module → END

How to model Pairwise Local Alignments with gaps?

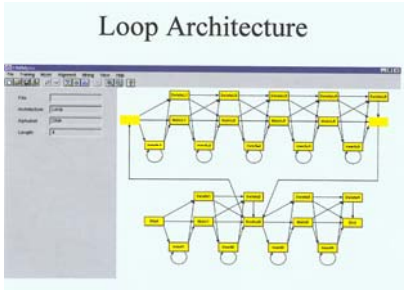


Standard HMM architectures

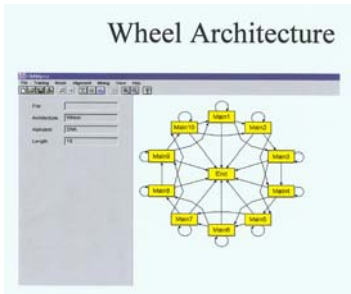
Linear Architecture



Standard HMM architectures



Standard HMM architectures



Profile HMMs from Multiple Alignments

```
HBA_HUMAN    VGA--HAGEY
HBB_HUMAN    V----NVDEV
MYG_PHYCA    VEA--DVAGH
GLB3_CHITP   VKG------D
GLB5_PETMA   VYS--TYETS
LGB2_LUPLU   FNA--NIPKH
GLB1_GLYDI   IAGADNGAGV
```

Construct Profile HMM from above multiple alignment.

HMM for Sequence Alignment

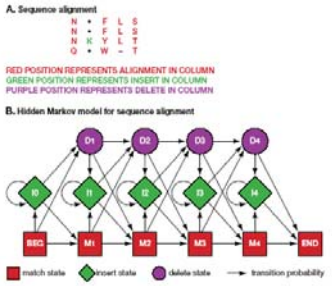


FIGURE 5.16. Relationship between the sequence alignment and the hidden Markov model of the alignment (Krogh et al. 1994). This particular form for the HMM was chosen to represent the sequence, structural, and functional variation expected in proteins. The model accommodates the identities, mismatches, insertions, and deletions expected in a group of related proteins. (*A*) A section of an msa. The illustration shows the columns generated in an msa. Each column may include matches and mismatches (*red* positions), insertions (*green* positions), and deletions (*purple* positions). (*B*) The HMM. Each column in the model represents the possibility of a match, insert, or delete in each column of the alignment in *A*. The HMM is a probabilistic representation of a section of the msa. Sequences can be generated from the HMM by starting at the beginning state labeled BEG, and then by following

Problem 3: LIKELIHOOD QUESTION

- Input: Sequence S, model M, state i
- Output: Compute the probability of reaching state i with sequence S using model M
  - Backward Algorithm (DP)

Problem 4: LIKELIHOOD QUESTION

- Input: Sequence S, model M
- Output: Compute the probability that S was emitted by model M
  - Forward Algorithm (DP)

Problem 5: LEARNING QUESTION

- Input: model structure M, Training Sequence *S*
- Output: Compute the parameters Θ
- Criteria: ML criterion
  - maximize P(S | M, Θ) HOW???

Problem 6: DESIGN QUESTION

- Input: Training Sequence *S*
- Output: Choose model structure M, and compute the parameters Θ
  - No reasonable solution
  - Standard models to pick from

Iterative Solution to the LEARNING QUESTION (Problem 5)

- Pick initial values for parameters Θ_0
- Repeat
  - Run training set S on model M
  - Count # of times transition $i \Rightarrow j$ is made
  - Count # of times letter x is emitted from state i
  - Update parameters Θ
- Until (some stopping condition)

Entropy

- Entropy measures the variability observed in given data.

$$E = -\sum_c p_c \log p_c$$

- Entropy is useful in multiple alignments & profiles.
- Entropy is max when uncertainty is max.

G-Protein Couple Receptors

- Transmembrane proteins with 7 α-helices and 6 loops; many subfamilies
- Highly variable: 200-1200 aa in length, some have only 20% identity.
- [Baldi & Chauvin, '94] HMM for GPCRs
- HMM constructed with 430 match states (avg length of sequences) ; Training: with 142 sequences, 12 iterations

GPCR - Analysis

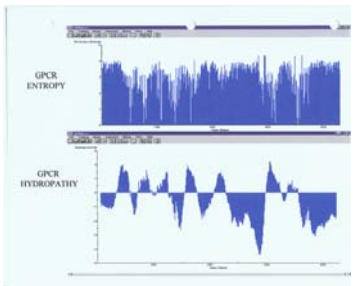
- Compute main state entropy values

$$H_i = -\sum_a e_{ia} \log e_{ia}$$

- For every sequence from test set (142) & random set (1600) & all SWISS-PROT proteins
  - Compute the negative log of probability of the most probable path π

$$Score(S) = -\log(P(\pi \mid S, M))$$

GPCR Analysis



Entropy

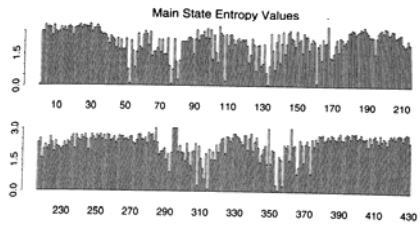


Figure 8.1: Entropy Profile of the Emission Probability Distributions Associated with the Main States of the HMM After 12 Cycles of Training.

GPCR Analysis (Cont'd)

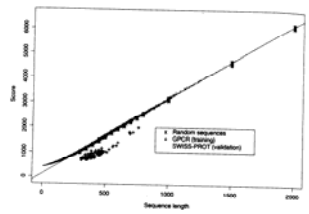


Figure 8.2: Scores (Negative Log-likelihoods of Optimal Viterbi Paths). Represented sequences consist of 147 GPCR training sequences, all sequences from the SWISS-PROT database of length less than or equal to 2000, and 220 randomly generated sequences with same average composition as the GPCRs of length 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800 (20 at each length). The regression line was obtained from the 220 random sequences. The horizontal distances in the histogram correspond to finalized scores (6).

Applications of HMM for GPCR

- Bacteriorhodopsin
  - Transmembrane protein with 7 domains
  - But it is not a GPCR
  - Compute score and discover that it is close to the regression line. Hence not a GPCR.
- Thyrotropin receptor precursors
  - All have long initial loop on INSERT STATE 20.
  - Also clustering possible based on distance to regression line.

HMMs – Advantages

- Sound statistical foundations
- Efficient learning algorithms
- Consistent treatment for insert/delete penalties for alignments in the form of locally learnable probabilities
- Capable of handling inputs of variable length
- Can be built in a modular & hierarchical fashion; can be combined into libraries.
- Wide variety of applications: Multiple Alignment, Data mining & classification, Structural Analysis, Pattern discovery, Gene prediction.

HMMs – Disadvantages

- Large # of parameters.
- Cannot express dependencies & correlations between hidden states.