Multiple Sequence Alignment

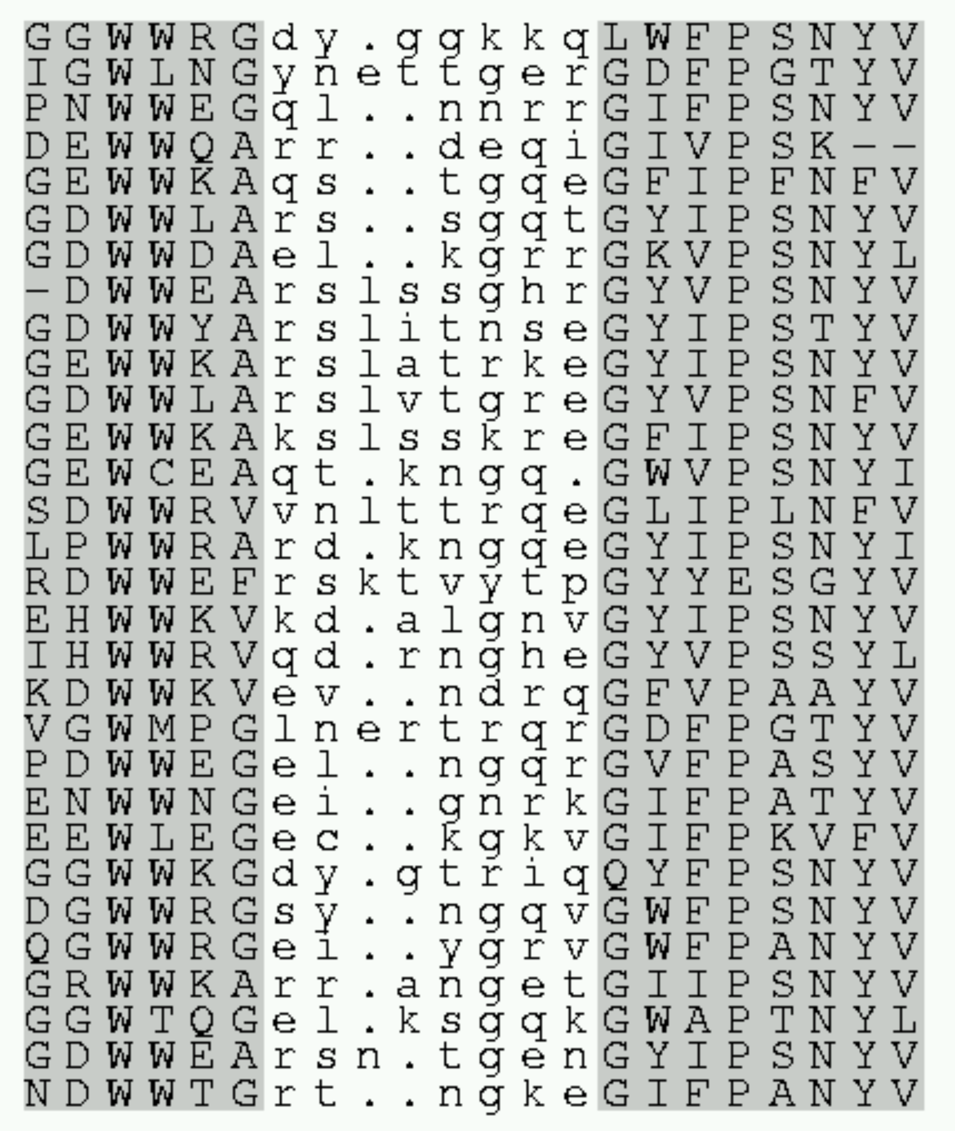
BMI/CS 576
www.biostat.wisc.edu/bmi576.html
Colin Dewey
cdewey@biostat.wisc.edu

Fall 2011

Multiple Sequence Alignment: Task Definition

- Given
  - a set of more than 2 sequences
  - a method for scoring an alignment
- Do:
  - determine the correspondences between the sequences such that the alignment score is maximized

Multiple Alignment of SH3 Domain



```
G G W W R G d y . g g k k q L W F P S N Y V
I G W L N G y n e t t g e r G D F P G T Y V
P N W W E G q l . . n n r r G I F P S N Y V
D E W W Q A r r . . d e q i G I V P S K - -
G E W W K A q s . . t g q e G F I P F N F V
G D W W L A r s . . s g q t G Y I P S N Y V
G D W W D A e l . . k g r r G K V P S N Y L
- D W W E A r s l s s g h r G Y V P S N Y V
G D W W Y A r s l i t n s e G Y I P S T Y V
G E W W K A r s l a t r k e G Y I P S N Y V
G D W W L A r s l v t g r e G Y V P S N F V
G E W W K A k s l s s k r e G F I P S N Y V
G E W C E A q t . k n g q . G W V P S N Y I
S D W W R V v n l t t r q e G L I P L N F V
L P W W R A r d . k n g q e G Y I P S N Y I
R D W W E F r s k t v y t p G Y Y E S G Y V
E H W W K V k d . a l g n v G Y I P S N Y V
I H W W R V q d . r n g h e G Y V P S S Y L
K D W W K V e v . . n d r q G F V P A A Y V
V G W M P G l n e r t r q r G D F P G T Y V
P D W W E G e l . . n g q r G V F P A S Y V
E N W W N G e i . . g n r k G I F P A T Y V
E E W L E G e c . . k g k v G I F P K V F V
G G W W K G d y . g t r i q Q Y F P S N Y V
D G W W R G s y . . n g q v G W F P S N Y V
Q G W W R G e i . . y g r v G W F P A N Y V
G R W W K A r r . a n g e t G I I P S N Y V
G G W T Q G e l . k s g q k G W A P T N Y L
G D W W E A r s n . t g e n G Y I P S N Y V
N D W W T G r t . . n g k e G I F P A N Y V
```

Figure from A. Krogh, An Introduction to Hidden Markov Models for Biological Sequences

Motivation for MSA

- establish input data for phylogenetic analyses
- determine evolutionary history of a set of sequences
  - At what point in history did certain mutations occur?
- discovering a common motif in a set of sequences (e.g. DNA sequences that bind the same protein)
- characterizing a set of sequences (e.g. a protein family)
- building *profiles* for sequence-database searching
  - PSI-BLAST generalizes a query sequence into a profile to search for remote relatives

Scoring a Multiple Alignment

- key issue: how do we assess the quality of a multiple sequence alignment?
- usually, the assumption is made that the individual *columns* of an alignment are independent

$$Score(m) = G + \sum_i S(m_i)$$

gap function

score of i^{th} column

- we’ ll discuss two methods
  - sum of pairs (SP)
  - minimum entropy

Scoring an Alignment: Sum of Pairs

- compute the sum of the pairwise scores

$$S(m_i) = \sum_{k<l} s(m_i^k, m_i^l)$$

m_i^k = character of the *k*th sequence in the *i* th column

s = substitution matrix

Scoring an Alignment: Minimum Entropy

- basic idea: try to minimize the *entropy* of each column
- another way of thinking about it: columns that can be communicated using few bits are good
- information theory tells us that an optimal code uses $-\log_2 p$ bits to encode a message of probability *p*
  - Frequently sent messages require few bits
  - Rarely sent messages require many bits

Scoring an Alignment: Minimum Entropy

- the messages in this case are the characters in a given column
- the entropy of a column is given by:

$$S(m_i) = -\sum_a c_{ia} \log_2 p_{ia}$$

$m_i =$ the *i* th column of an alignment *m*

$c_{ia} =$ count of character *a* in column *i*

$p_{ia} =$ probability of character *a* in column *i*

Dynamic Programming Approach

- can find optimal alignments using dynamic programming
- generalization of methods for pairwise alignment
  - consider *k*-dimension matrix for *k* sequences (instead of 2-dimensional matrix)
  - each matrix element represents alignment score for *k* subsequences (instead of 2 subsequences)
- given *k* sequences of length *n*
  - space complexity is

$$O(n^k)$$

Dynamic Programming Approach

$$\alpha_{i_1,i_2,\ldots,i_k} = \max \begin{cases} \alpha_{i_1-1,i_2-1,\ldots,i_k-1} & + \; S(x^1_{i_1}, x^2_{i_2}, \ldots, x^k_{i_k}) \\ \alpha_{i_1,i_2-1,\ldots,i_k-1} & + \; S(-, x^2_{i_2}, \ldots, x^k_{i_k}) \\ \alpha_{i_1-1,i_2,\ldots,i_k-1} & + \; S(x^1_{i_1}, -, \ldots, x^k_{i_k}) \\ & \vdots \\ \alpha_{i_1,i_2,\ldots,i_k-1} & + \; S(-, -, \ldots, x^k_{i_k}) \\ & \vdots \end{cases}$$

max score of alignment
for subsequences
$x^1_{i_1}, x^2_{i_2}, \ldots, x^k_{i_k}$

Dynamic Programming Approach

- given k sequences of length n
  - time complexity is

$$O(k^2 2^k n^k)$$

if we use sum of pairs

$$O(k 2^k n^k)$$

if column scores can be computed in $O(k)$, as with entropy

Heuristic Alignment Methods

- since time complexity of DP approach is exponential in the number of sequences, heuristic methods are usually used
- *progressive alignment*: construct a succession of pairwise alignments
  - star approach
  - tree approaches, like CLUSTALW
  - etc.

- iterative refinement
  - given a multiple alignment (say from a progressive method)
    - remove a sequence, realign it to profile of other sequences
    - repeat until convergence

Star Alignment Approach

- given: *k* sequences to be aligned

  $x_1, \ldots, x_k$

  - pick one sequence x_c as the "center"
  - for each $x_i \neq x_c$ determine an optimal alignment between x_i and x_c
  - merge pairwise alignments
- return: multiple alignment resulting from aggregate

Star Alignment Example

Given:

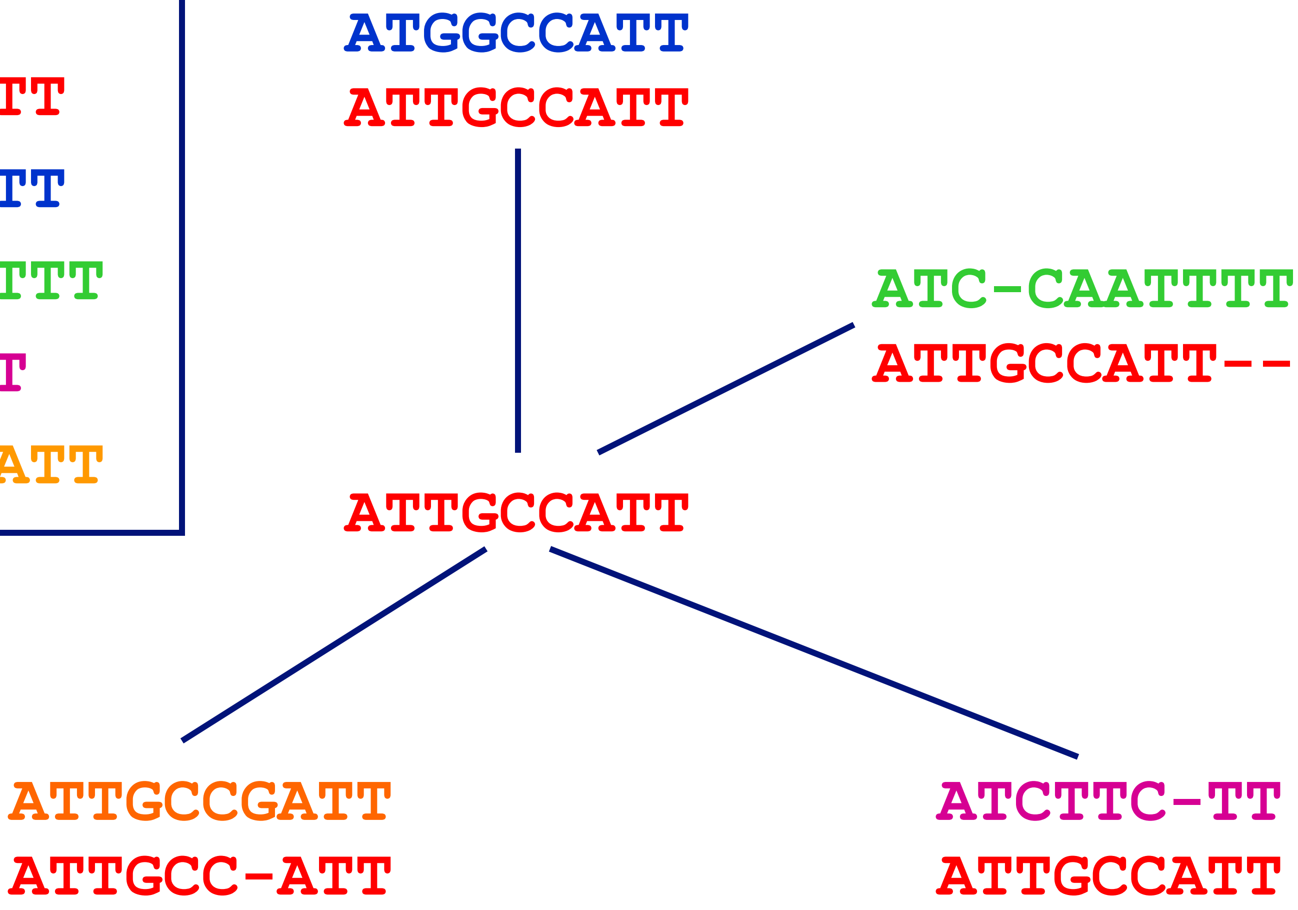
ATTGCCATT

ATGGCCATT

ATCCAATTTT

ATCTTCTT

ATTGCCGATT



```
ATGGCCATT
ATTGCCATT
```

```
ATC-CAATTTT
ATTGCCATT--
```

ATTGCCATT

```
ATTGCCGATT
ATTGCC-ATT
```

```
ATCTTC-TT
ATTGCCATT
```

Star Alignment Example

- merging pairwise alignments

	present pair	alignment
1.	ATGGCCATT ATTGCCATT	ATTGCCATT ATGGCCATT
2.	ATC-CAATTTT ATTGCCATT--	ATTGCCATT-- ATGGCCATT-- ATC-CAATTTT

Star Alignment Example

present pair

alignment

3.

```
ATCTTC-TT
ATTGCCATT
```

```
ATTGCCATT--
ATGGCCATT--
ATC-CAATTTT
ATCTTC-TT--
```

4.

```
ATTGCCGATT
ATTGCC-ATT
```

```
ATTGCC-ATT--
ATGGCC-ATT--
ATC-CA-ATTTT
ATCTTC--TT--
ATTGCCGATT--
```

shift entire columns
when incorporating a gap

Star Alignments: Aggregating Pairwise Alignments

- "once a gap, always a gap"
- shift entire columns when incorporating gaps

Star Alignments:
Approaches to Picking the Center

Two possible approaches:

1. try each sequence as the center, return the best multiple alignment
2. compute all pairwise alignments and select the string x_c that maximizes:

$$\sum_{i \neq c} \mathrm{sim}(x_i, x_c)$$

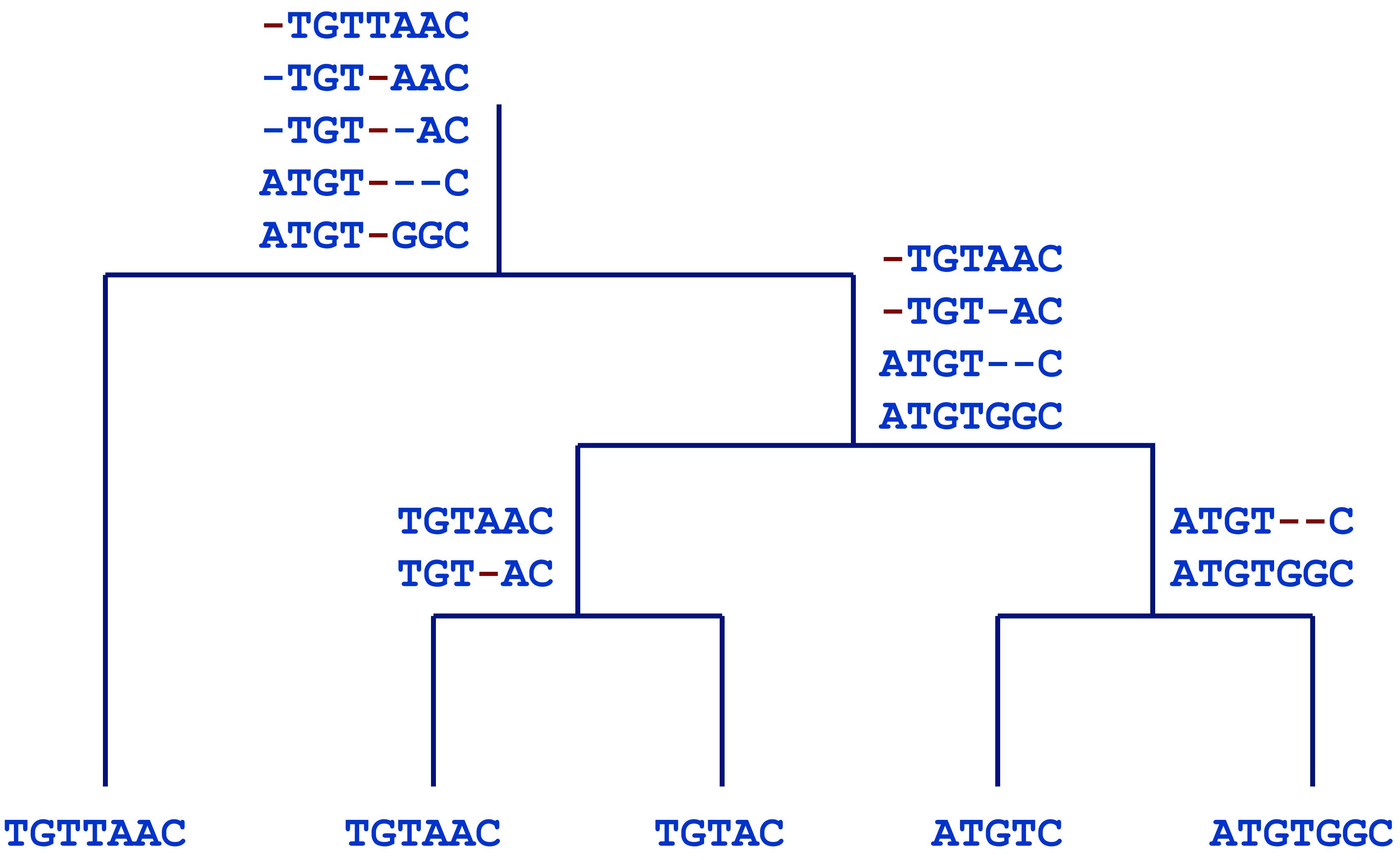
Tree Alignments

- basic idea: organize multiple sequence alignment using a *guide tree*
  - leaves represent sequences
  - internal nodes represent alignments
- determine alignments from bottom of tree upward
  - return multiple alignment represented at the root of the tree
- one common variant: the CLUSTALW algorithm [Thompson et al. 1994]

Doing the Progressive Alignment in CLUSTALW

- depending on the internal node in the tree, we may have to align a
  - a sequence with a sequence
  - a sequence with a *profile* (partial alignment)
  - a *profile* with a *profile*
- in all cases we can use dynamic programming
  - for the profile cases, use SP scoring

Tree Alignment Example



```
-TGTTAAC
-TGT-AAC
-TGT--AC
ATGT---C
ATGT-GGC
```

```
-TGTAAC
-TGT-AC
ATGT--C
ATGTGGC
```

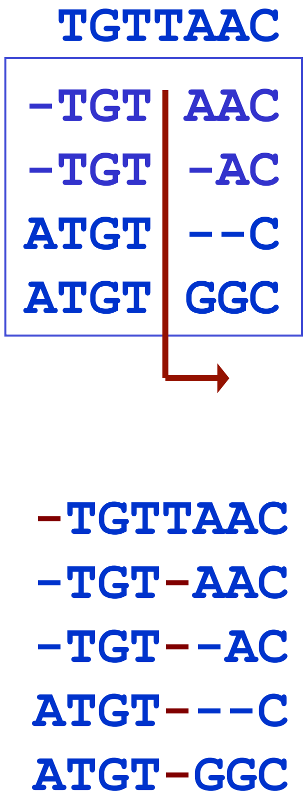
```
TGTAAC
TGT-AC
```

```
ATGT--C
ATGTGGC
```

TGTTAAC TGTAAC TGTAC ATGTC ATGTGGC

Aligning Profiles

- aligning sequences/profiles to profiles is essentially pairwise alignment
  - shift entire columns when incorporating gaps



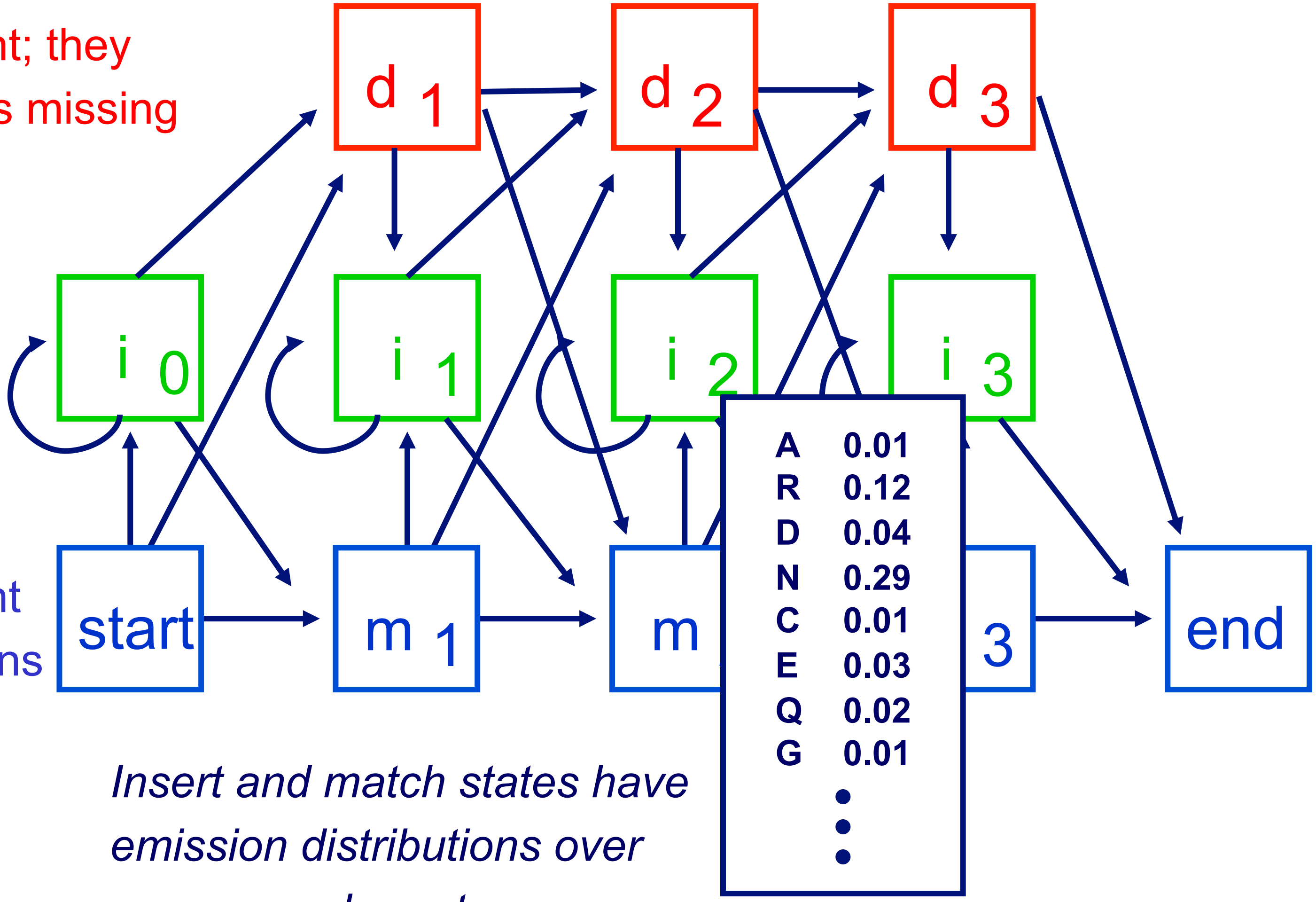
```
 TGTTAAC

-TGT AAC
-TGT -AC
ATGT --C
ATGT GGC


-TGTTAAC
-TGT-AAC
-TGT--AC
ATGT---C
ATGT-GGC
```

Profile HMMs

- profile HMMs are commonly used to model families of sequences



Delete states are silent; they Account for characters missing in some sequences

Insert states account for extra characters in some sequences

Match states represent key conserved positions

A	0.01
R	0.12
D	0.04
N	0.29
C	0.01
E	0.03
Q	0.02
G	0.01

•
•
•

Insert and match states have emission distributions over sequence characters

Multiple Alignment with Profile HMMs

insert states

delete states (silent)

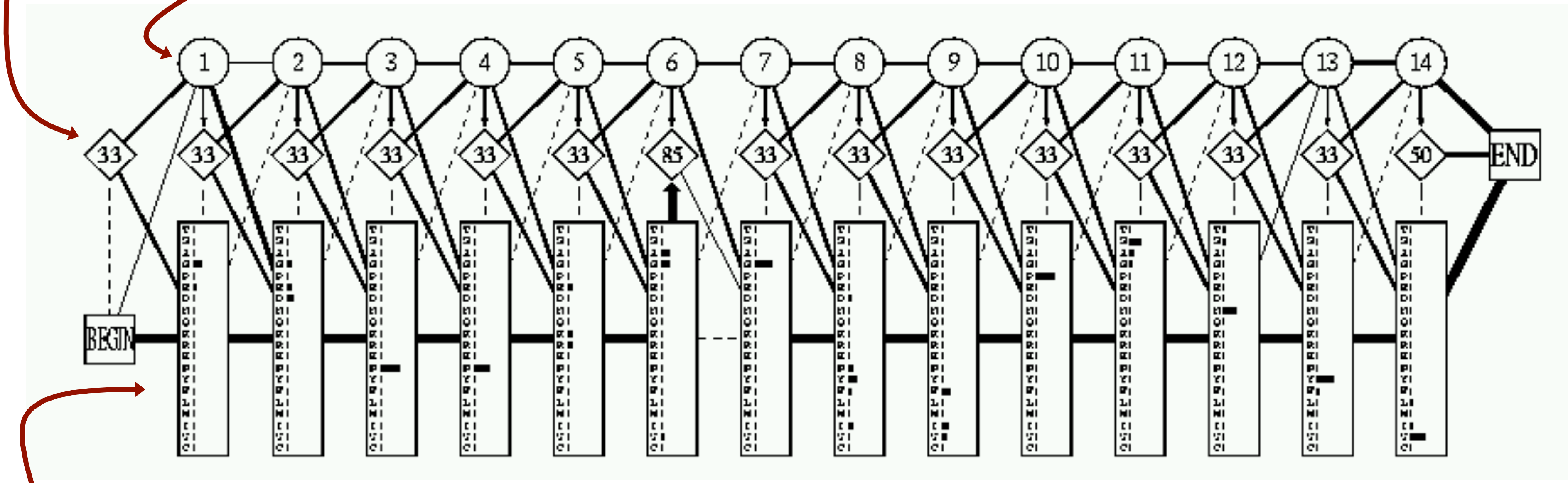


Figure from A. Krogh, An Introduction to Hidden Markov Models for Biological Sequences

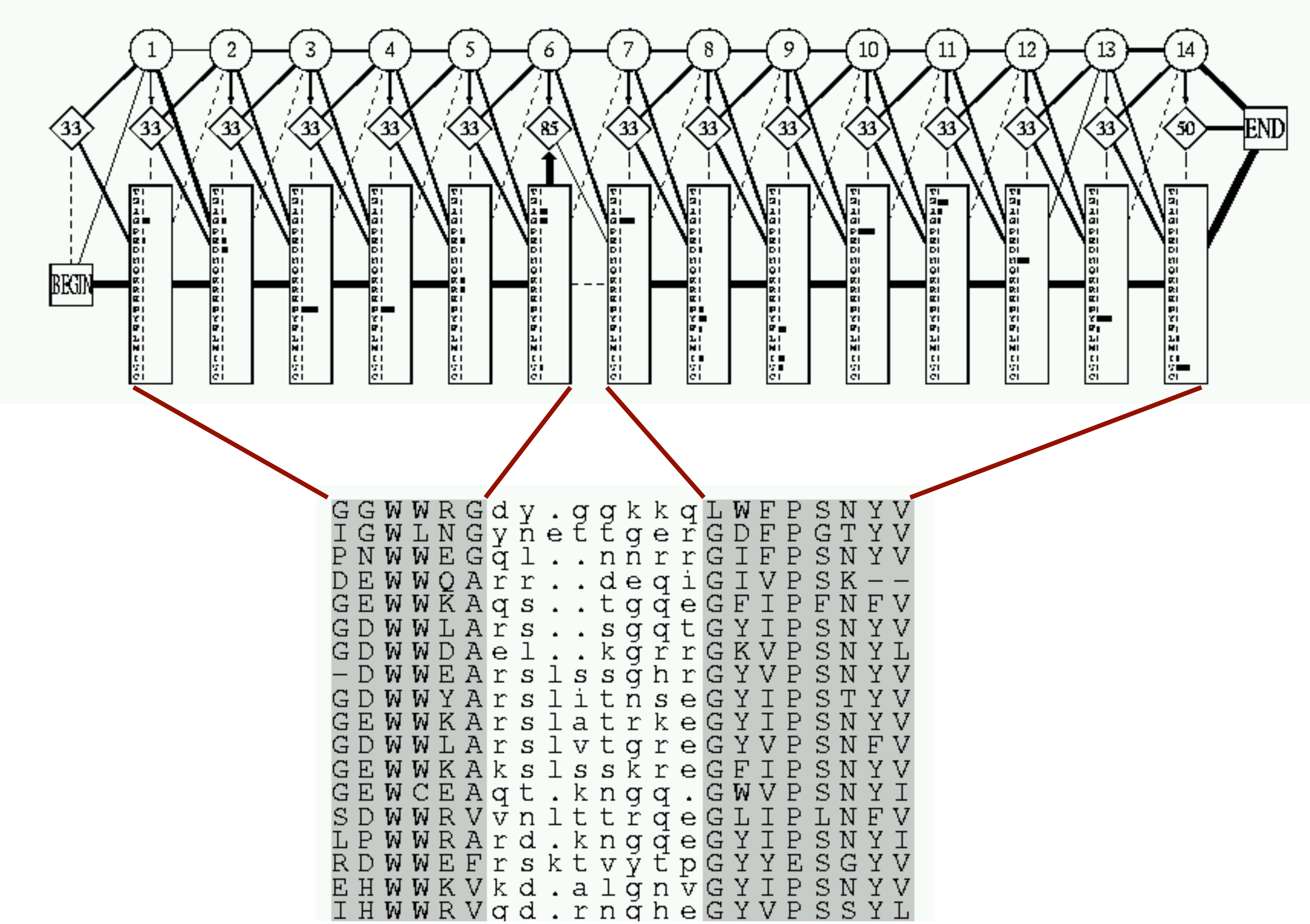
match states

Multiple Alignment with Profile HMMs

- given a set of sequences to be aligned
  - use Baum-Welch to learn parameters of model
  - may also adjust length of profile HMM during training

- to compute a multiple alignment given the profile HMM
  - run the Viterbi algorithm on each sequence
  - Viterbi paths indicate correspondences among sequences

Multiple Alignment with Profile HMMs



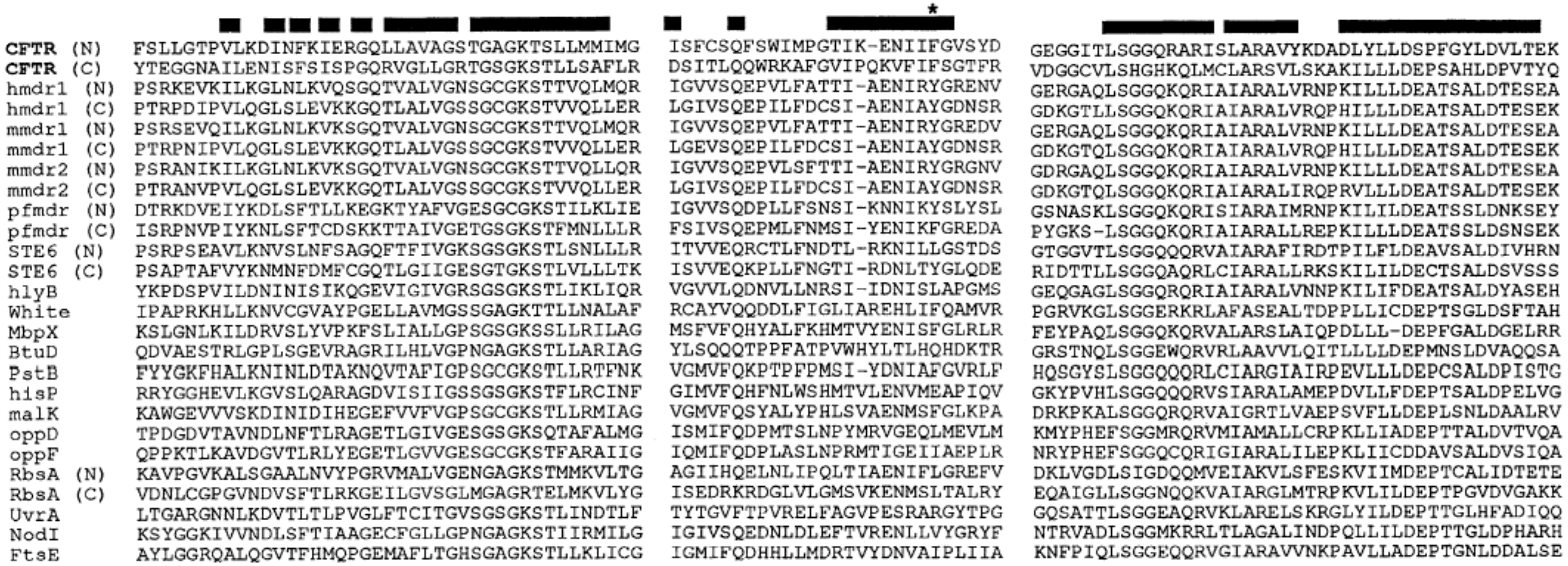
```
G G W W R G d y . g g k k q L W F P S N Y V
I G W L N G y n e t t g e r G D F P G T Y V
P N W W E G q l . . n n r r G I F P S N Y V
D E W W Q A r r . . d e q i G I V P S K - -
G E W W K A q s . . t g q e G F I P F N F V
G D W W L A r s . . s g q t G Y I P S N Y V
G D W W D A e l . . k g r r G K V P S N Y L
- D W W E A r s l s s g h r G Y V P S N Y V
G D W W Y A r s l i t n s e G Y I P S T Y V
G E W W K A r s l a t r k e G Y I P S N Y V
G D W W L A r s l v t g r e G Y V P S N F V
G E W W K A k s l s s k r e G F I P S N Y V
G E W C E A q t . k n g q . G W V P S N Y I
S D W W R V v n l t t r q e G L I P L N F V
L P W W R A r d . k n g q e G Y I P S N Y I
R D W W E F r s k t v y t p G Y Y E S G Y V
E H W W K V k d . a l g n v G Y I P S N Y V
I H W W R V q d . r n g h e G Y V P S S Y L
```

Multiple Alignment Case Study: The Cystic Fibrosis Gene

- cystic fibrosis (CF)
  - recessive genetic disease caused by a defect in a single-gene
  - causes the body to produce abnormally thick mucus that clogs the lungs and the pancreas

- the cystic fibrosis conductance regulator (CFTR) gene
  - gene and its role in CF identified in 1989 [Riordan et al., *Science*]
  - most common mutation is called ΔF508; a deletion of a phenylalanine (F) at position 508 in the CFTR protein
  - the CFTR protein controls the movement of salt and water into and out of cells; mutations in CFTR block this movement, causing mucus problem

So What Does CFTR Do?
A CFTR Multiple Alignment



```
                                                                       *
CFTR (N)    FSLLGTPVLKDINFKIERGQLLAVAGSTGAGKTSLLMMIMG  ISFCSQFSWIMPGTIK-ENIIFGVSYD  GEGGITLSGGQRARISLARAVYKDADLYLLDSPFGYLDVLTEK
CFTR (C)    YTEGGNAILENISFSISPGQRVGLLGRTGSGKSTLLSAFLR  DSITLQQWRKAFGVIPQKVFIFSGTFR  VDGGCVLSHGHKQLMCLARSVLSKAKILLLDEPSAHLDPVTYQ
hmdr1 (N)   PSRKEVKILKGLNLKVQSGQTVALVGNSGCGKSTTVQLMQR  IGVVSQEPVLFATTI-AENIRYGRENV  GERGAQLSGGQKQRIAIARALVRNPKILLLDEATSALDTESEA
hmdr1 (C)   PTRPDIPVLQGLSLEVKKGQTLALVGSSGCGKSTVVQLLER  LGIVSQEPILFDCSI-AENIAYGDNSR  GDKGTLLSGGQKQRIAIARALVRQPHILLLDEATSALDTESEK
mmdr1 (N)   PSRSEVQILKGLNLKVKSGQTVALVGNSGCGKSTTVQLMQR  IGVVSQEPVLFATTI-AENIRYGREDV  GERGAQLSGGQKQRIAIARALVRNPKILLLDEATSALDTESEA
mmdr1 (C)   PTRPNIPVLQGLSLEVKKGQTLALVGSSGCGKSTVVQLLER  LGEVSQEPILFDCSI-AENIAYGDNSR  GDKGTQLSGGQKQRIAIARALVRQPHILLLDEATSALDTESEK
mmdr2 (N)   PSRANIKILKGLNLKVKSGQTVALVGNSGCGKSTTVQLLQR  IGVVSQEPVLSFTTI-AENIRYGRGNV  GDRGAQLSGGQKQRIAIARALVRNPKILLLDEATSALDTESEA
mmdr2 (C)   PTRANVPVLQGLSLEVKKGQTLALVGSSGCGKSTVVQLLER  LGIVSQEPILFDCSI-AENIAYGDNSR  GDKGTQLSGGQKQRIAIARALIRQPRVLLLDEATSALDTESEK
pfmdr (N)   DTRKDVEIYKDLSFTLLKEGKTYAFVGESGCGKSTILKLIE  IGVVSQDPLLFSNSI-KNNIKYSLYSL  GSNASKLSGGQKQRISIARAIMRNPKILILDEATSSLDNKSEY
pfmdr (C)   ISRPNVPIYKNLSFTCDSKKTTAIVGETGSGKSTFMNLLLR  FSIVSQEPMLFNMSI-YENIKFGREDA  PYGKS-LSGGQKQRIAIARALLREPKILLLDEATSSLDSNSEK
STE6 (N)    PSRPSEAVLKNVSLNFSAGQFTFIVGKSGSGKSTLSNLLLR  ITVVEQRCTLFNDTL-RKNILLGSTDS  GTGGVTLSGGQQQRVAIARAFIRDTPILFLDEAVSALDIVHRN
STE6 (C)    PSAPTAFVYKNMNFDMFCGQTLGIIGESGTGKSTLVLLLTK  ISVVEQKPLLFNGTI-RDNLTYGLQDE  RIDTTLLSGGQAQRLCIARALLRKSKILILDECTSALDSVSSS
hlyB        YKPDSPVILDNINISIKQGEVIGIVGRSGSGKSTLIKLIQR  VGVVLQDNVLLNRSI-IDNISLAPGMS  GEQGAGLSGGQRQRIAIARALVNNPKILIFDEATSALDYASEH
White       IPAPRKHLLKNVCGVAYPGELLAVMGSSGAGKTTLLNALAF  RCAYVQQDDLFIGLIAREHLIFQAMVR  PGRVKGLSGGERKRLAFASEALTDPPLLICDEPTSGLDSFTAH
MbpX        KSLGNLKILDRVSLYVPKFSLIALLGPSGSGKSSLLRILAG  MSFVFQHYALFKHMTVYENISFGLRLR  FEYPAQLSGGQKQRVALARSLAIQPDLLL-DEPFGALDGELRR
BtuD        QDVAESTRLGPLSGEVRAGRILHLVGPNGAGKSTLLARIAG  YLSQQQTPPFATPVWHYLTLHQHDKTR  GRSTNQLSGGEWQRVRLAAVVLQITLLLLDEPMNSLDVAQQSA
PstB        FYYGKFHALKNINLDTAKNQVTAFIGPSGCGKSTLLRTFNK  VGMVFQKPTPFPMSI-YDNIAFGVRLF  HQSGYSLSGGQQQRLCIARGIAIRPEVLLLDEPCSALDPISTG
hisP        RRYGGHEVLKGVSLQARAGDVISIIGSSGSGKSTFLRCINF  GIMVFQHFNLWSHMTVLENVMEAPIQV  GKYPVHLSGGQQQRVSIARALAMEPDVLLFDEPTSALDPELVG
malK        KAWGEVVVSKDINIDIHEGEFVVFVGPSGCGKSTLLRMIAG  VGMVFQSYALYPHLSVAENMSFGLKPA  DRKPKALSGGQRQRVAIGRTLVAEPSVFLLDEPLSNLDAALRV
oppD        TPDGDVTAVNDLNFTLRAGETLGIVGESGSGKSQTAFALMG  ISMIFQDPMTSLNPYMRVGEQLMEVLM  KMYPHEFSGGMRQRVMIAMALLCRPKLLIADEPTTALDVTVQA
oppF        QPPKTLKAVDGVTLRLYEGETLGVVGESGCGKSTFARAIIG  IQMIFQDPLASLNPRMTIGEIIAEPLR  NRYPHEFSGGQCQRIGIARALILEPKLIICDDAVSALDVSIQA
RbsA (N)    KAVPGVKALSGAALNVYPGRVMALVGENGAGKSTMMKVLTG  AGIIHQELNLIPQLTIAENIFLGREFV  DKLVGDLSIGDQQMVEIAKVLSFESKVIIMDEPTCALIDTETE
RbsA (C)    VDNLCGPGVNDVSFTLRKGEILGVSGLMGAGRTELMKVLYG  ISEDRKRDGLVLGMSVKENMSLTALRY  EQAIGLLSGGNQQKVAIARGLMTRPKVLILDEPTPGVDVGAKK
UvrA        LTGARGNNLKDVTLTLPVGLFTCITGVSGSGKSTLINDTLF  TYTGVFTPVRELFAGVPESRARGYTPG  GQSATTLSGGEAQRVKLARELSKRGLYILDEPTTGLHFADIQQ
NodI        KSYGGKIVVNDLSFTIAAGECFGLLGPNGAGKSTIIRMILG  IGIVSQEDNLDLEFTVRENLLVYGRYF  NTRVADLSGGMKRRLTLAGALINDPQLLILDEPTTGLDPHARH
FtsE        AYLGGRQALQGVTFHMQPGEMAFLTGHSGAGKSTLLKLICG  IGMIFQDHHLLMDRTVYDNVAIPLIIA  KNFPIQLSGGEQQRVGIARAVVNKPAVLLADEPTGNLDDALSE
```

Figure from Riordan et al, *Science* 245:1066-1073, 1989.

Multiple Alignment Case Study: the Cystic Fibrosis Gene

- two key features of the protein made apparent in multiple sequence alignment (and other analyses)
  - membrane-spanning domains
  - ATP-binding motifs
- these features indicated that CFTR is likely to be involved in transporting ions across the cell membrane

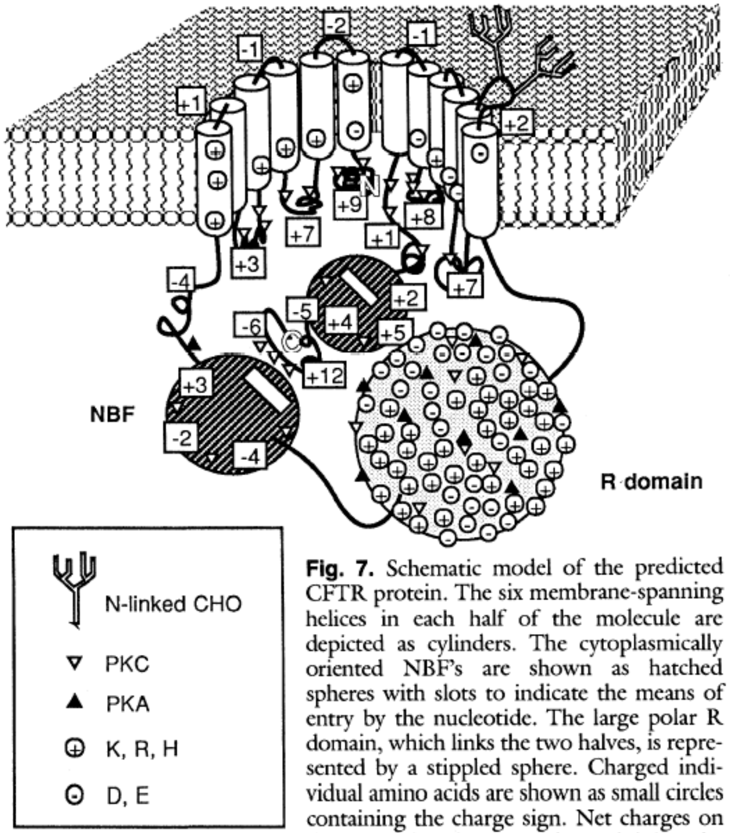


Fig. 7. Schematic model of the predicted CFTR protein. The six membrane-spanning helices in each half of the molecule are depicted as cylinders. The cytoplasmically oriented NBF's are shown as hatched spheres with slots to indicate the means of entry by the nucleotide. The large polar R domain, which links the two halves, is represented by a stippled sphere. Charged individual amino acids are shown as small circles containing the charge sign. Net charges on the internal and external loops joining the membrane cylinders and on regions of the NBF's are contained in open squares. Potential sites for phosphorylation by protein kinases A or C (PKA or PKC) and N-glycosylation (N-linked CHO) are as indicated. K, Lys; R, Arg; H, His; D, Asp; and E, Glu.

Figure from Riordan et al, *Science* 245:1066-1073, 1989.

Notes on Multiple Alignment

- as with pairwise alignment, can compute *local* and *global* multiple alignments
- dynamic programming is not feasible for most cases -- heuristic methods usually used instead

Summary: Some Methods for Multiple Sequence Alignment

method	alignment types	search
multi-dimensional dynamic programming	global/local	dynamic programming
Star	global	greedy via pairwise alignments
CLUSTALW (tree)	global	greedy via pairwise alignment
profile HMMs	global/local	Baum-Welch (EM) to learn model, Viterbi to reocover alignments
EM/MEME Gibbs sampling Random projections etc.	local	EM Gibbs sampling random projections