

Biological Sequence Analysis and Motif Discovery

Introductory Overview Lecture
Joint Statistical Meetings 2001, Atlanta

Jun Liu

Department of Statistics

Harvard University

<http://www.fas.harvard.edu/~junliu>

jliu@stat.harvard.edu

Topics to be covered

- **Basic Biology: DNA, RNA, Protein; genetic code.**
- **Biological Sequence Analysis**
 - **Pairwise alignment --- dynamic programming**
 - Needleman-Wunsch
 - Smith-Waterman
 - Blast ...
 - **Multiple sequence alignment**
 - Heuristic approaches
 - The hidden Markov model
 - **Motif finding in DNA and protein sequence**

Central Paradigm of Bioinformatics

Genetic
Information



Molecular
Structure

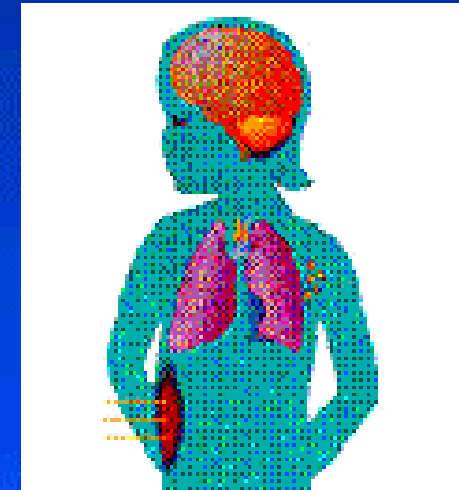
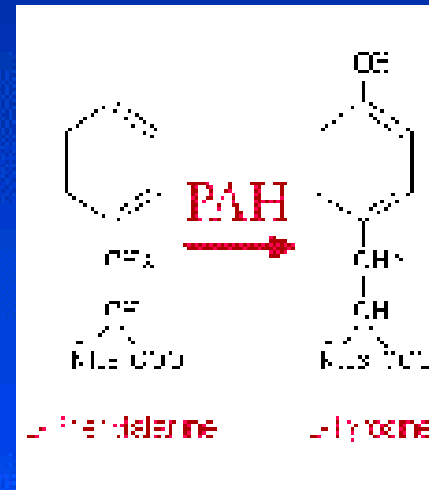
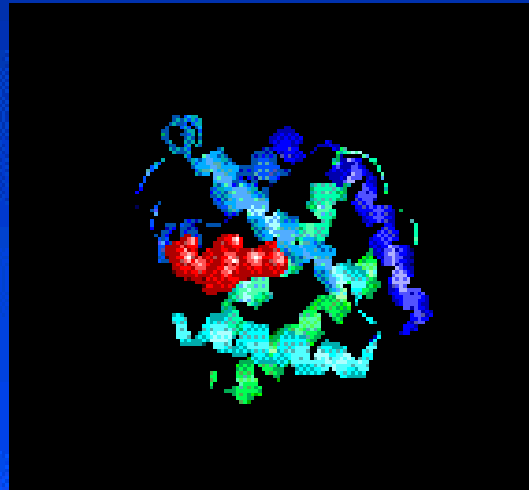


Biochemical
Function



Symptoms
(Phenotype)

SRAAINMHTVA
VSYQTVERVVN
VSTATVSRALA
CVTTTVSHVIN
SEVSAVSAILN
GVSEMTREDLN
TAYATINRVE
GSDPTVSRRLA
MSIATITRCSN
ISSETPERILE
FDISRLSHLFR
LRFSLAHLFR
MTVETISBLLE
TLEPHLHHLFR



Courtesy of Doug Brutlag



Buliding Blocks of Biological Systems:

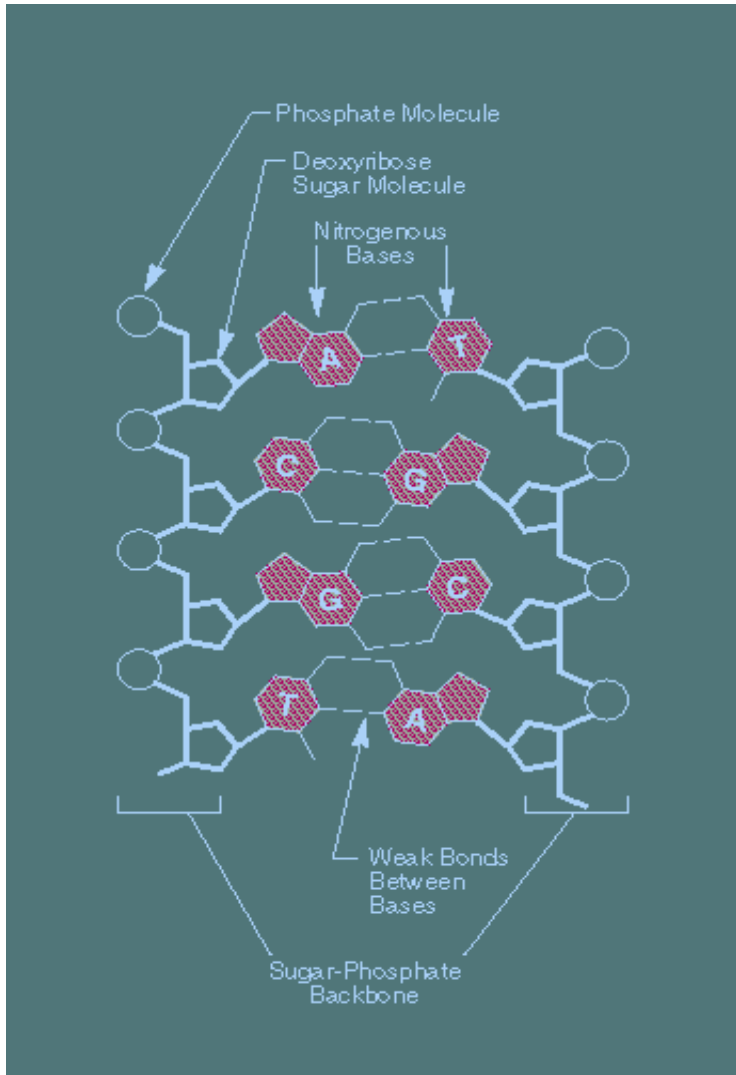
nucleotides and amino acids

- **DNA** (nucleotides, 4 types): information carrier/encoder.
- **RNA**: bridge from DNA to protein.
- **Protein** (amino acids, 20 types): action molecules.
- **Genetic code**: deciphering genetic information.

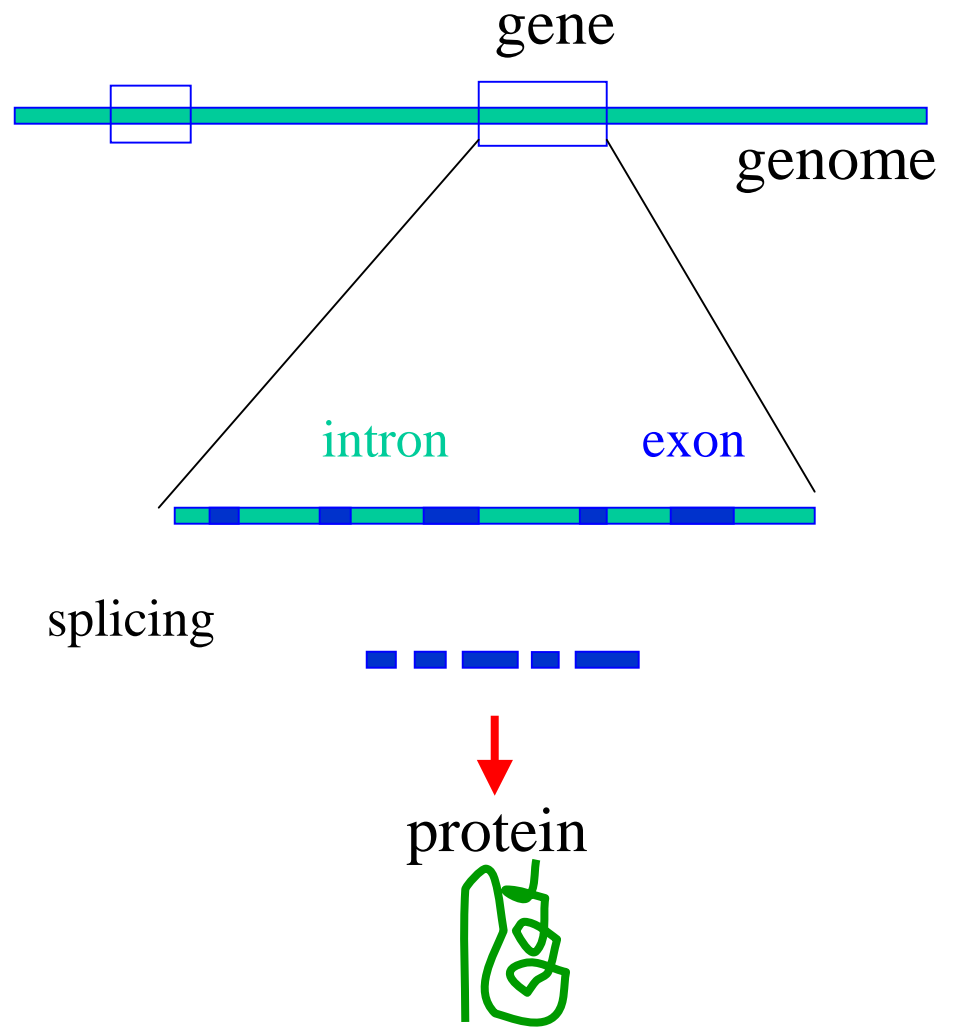
atgaatcgta ggggtttgaa cgctggcaat acgatgactt ctcaagcgaa
cattgacgac ggcagctgga aggcggtctc cgagggcgga



MNRRGLNAGNTMTSQANIDDGSWKAVSEGG

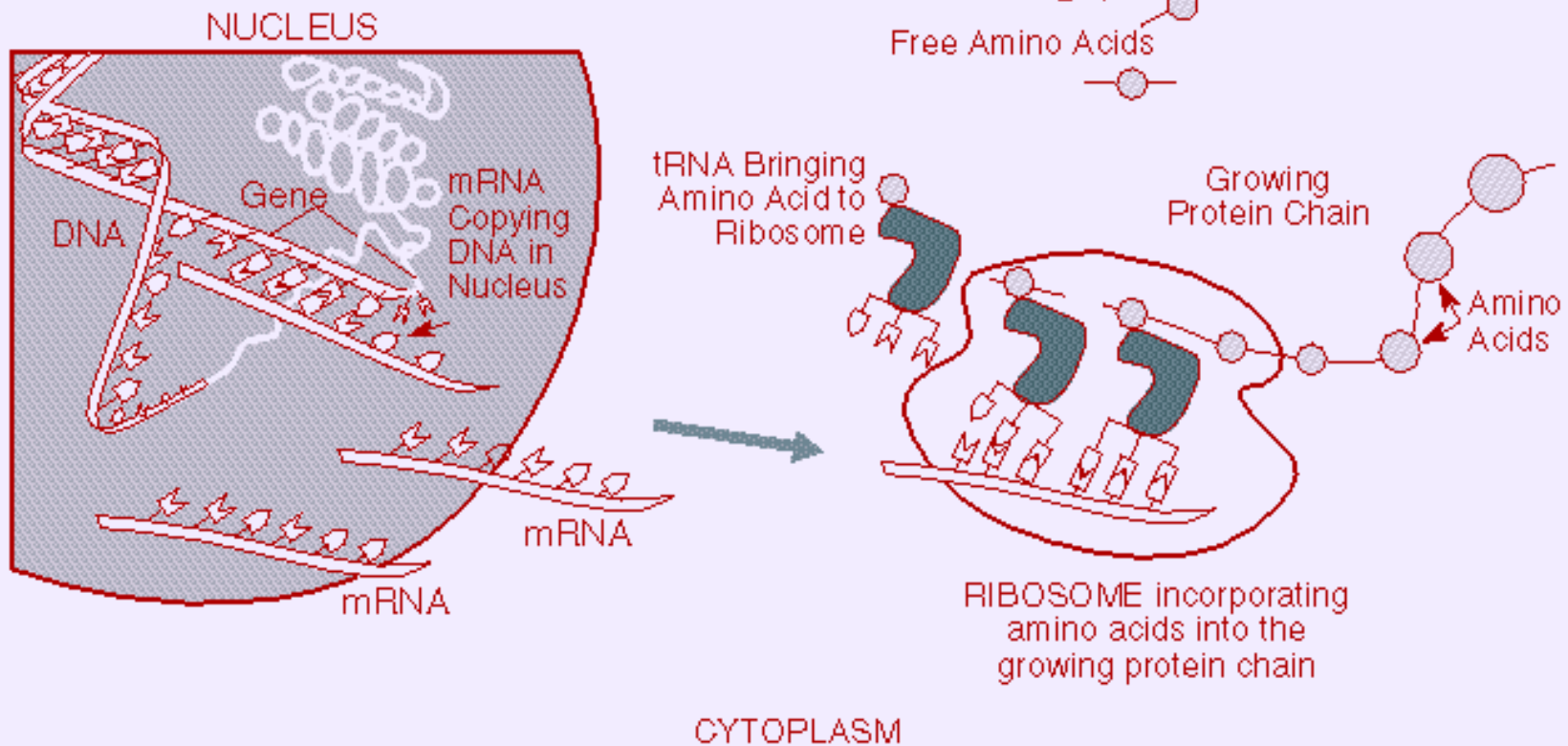


DNA sequences



From DNA to Protein

ORNL-DWG 94M-17360



Genetic Code

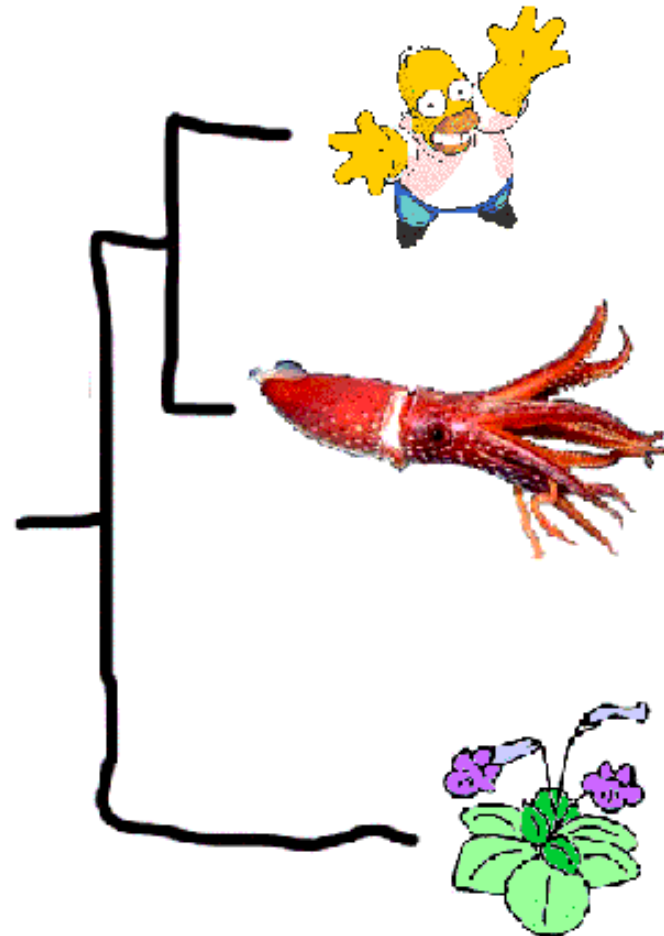
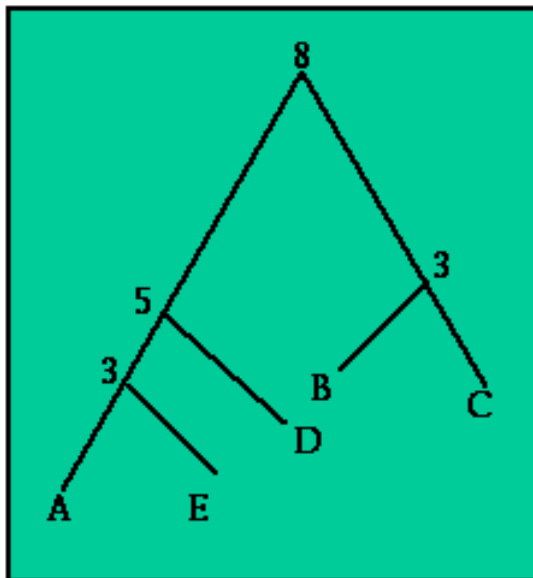
		Second Position of Codon					
		T	C	A	G		
First Position	T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T	
		TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C	
		TTA Leu [L]	TCA Ser [S]	TAA <i>Ter</i> [end]	TGA <i>Ter</i> [end]	A	
		TTG Leu [L]	TCG Ser [S]	TAG <i>Ter</i> [end]	TGG Trp [W]	G	
	C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T	
		CTC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C	
		CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A	
		CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G	
	A	ATT Ile [I]	ACT Thr [T]	AAT Asn [N]	AGT Ser [S]	T	
		ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C	
		ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A	
		ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G	
	G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T	
		GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C	
		GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A	
		GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G	

Main Resources for Data

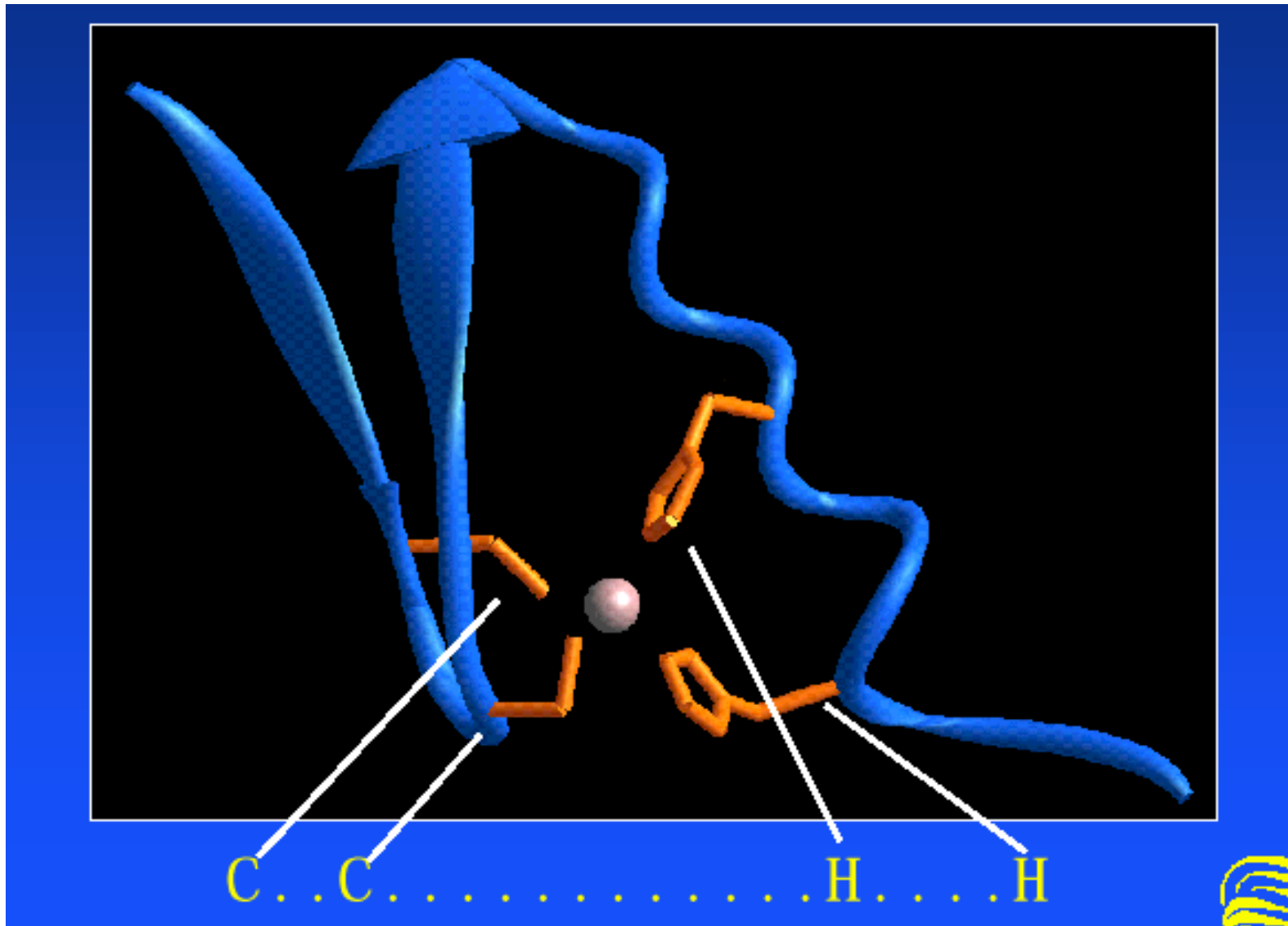
- **NCBI --- national center for biotechnology information.**
 - GenBank maintainance
 - BLAST searching and servers
 - Entrez database
 - Taxonomy database
 - Structure database
 - Bankit and SEQUIN submission software
 - **Web access** <http://www.ncbi.nlm.nih.gov>
- **EMBL --- European equivalent of NCBI.**
 - **Web access** <http://www.embl-heidelberg.de>
 - **EBI --- outstation of EMBL.** <http://www.ebi.ac.uk/>

Finding ‘Patterns’ in Biological Sequences

Everything is related to everything else in a logical way



A Motif



Consensus in Sequences

- A database for protein families and patterns:
 - **Prosite** <http://www.expasy.ch/prosite/>

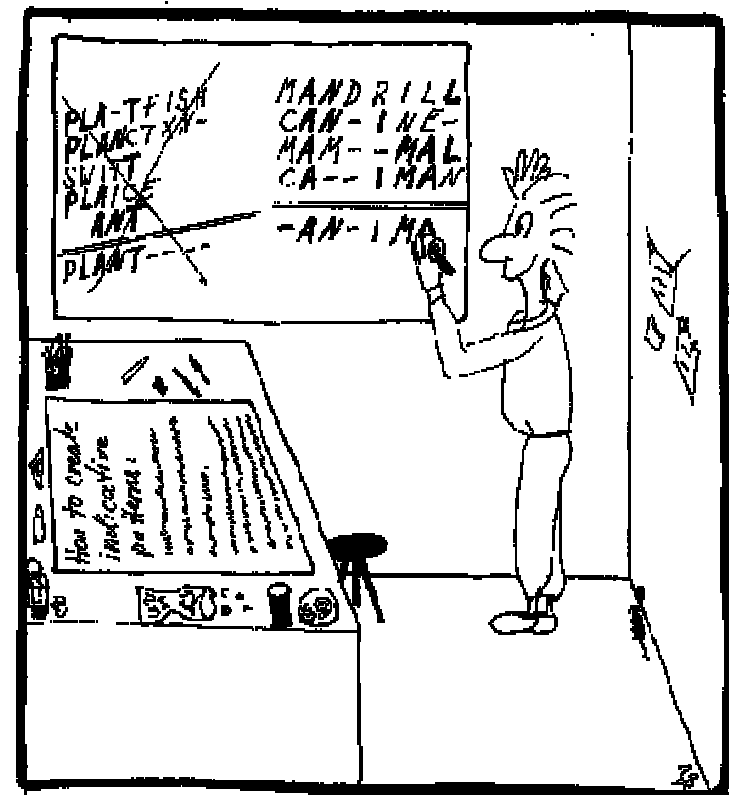
Pattern: [LIVM]-[ST]-A-[STAG]-H-C

Interpretation:

<A-x-[ST](2)-x(0,1)-V-*{LI}*

This pattern, which must be in the N-terminal of the sequence (^<'), means:
Ala-any-[Ser or Thr]-[Ser or Thr]-
(any or none)-Val-(any but Leu & Ile)

How we develop Prosite patterns!



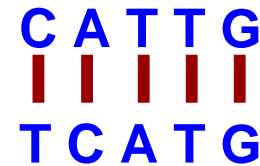
Pattern Finding Programs

- **Prosites web program** (<http://www.expasy.ch/prosite/>)
 - **ScanProsite** - Scan a sequence against PROSITE or a pattern against SWISS-PROT
 - **ProfileScan** - Scan a sequence against the profile entries in PROSITE
- **Motif web programs** (<http://motif.stanford.edu/identify/>)
 - IDENTIFY
 - SCAN
- **GCG SEQWEB programs (commercial)**
 - Stringsearch
 - Findpatterns
 - Motifs

Pairwise Sequence Alignment Methods

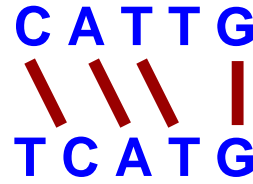
The Sequence Alignment problem

Given a pair of sequences,
how do we view them?

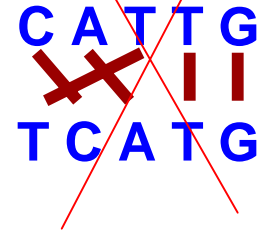


CATTG
TCATG

A better representation?

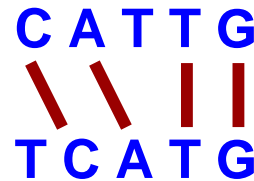


CATTG
TCATG

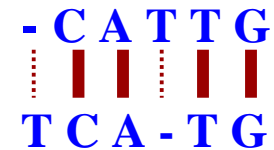


~~CATTG
TCATG~~

Or



CATTG
TCATG



-CATTG
TCA-TG

What determines the choice? Gap & mismatch penalties

Sequence Alignment and Typical Objective Function

X	220	230	240	250	X
F--SGGNTHIYMNHVEQCKEILRREPKELCVLISGLPYKFRYLSTKE-QLK-Y					
: :: : : : : : : ::::: ::					
GDFIHTLGDAHIYLNHIEPLKIQIQREPRPFPKLRILRKVEKIDDFKAEDFQIEGYN					
X	260	270	280	290	X

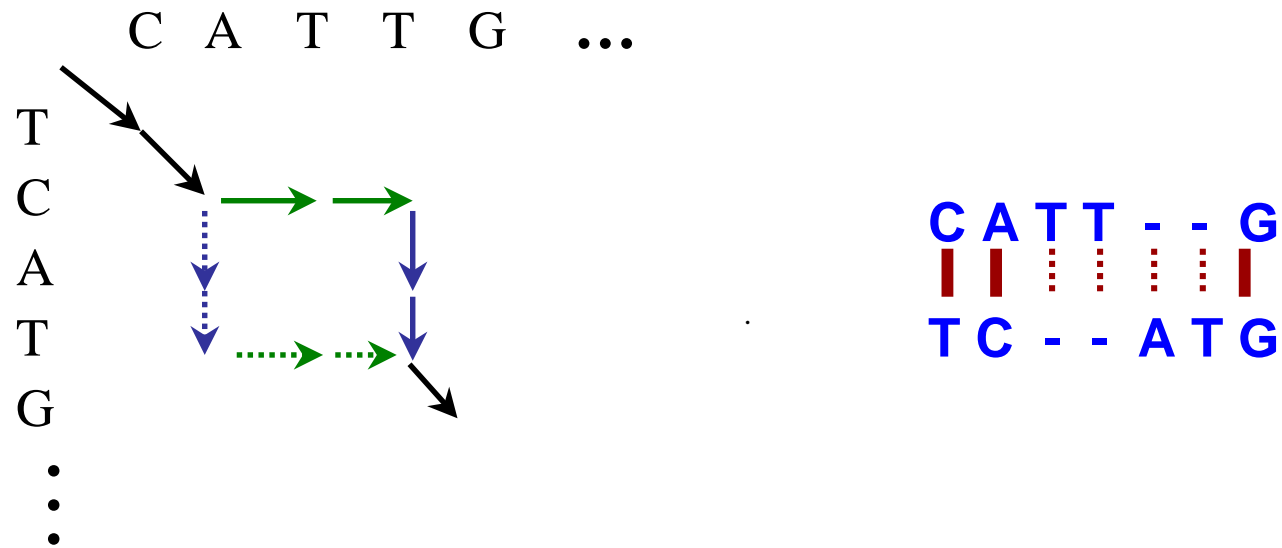
$$\text{Score} = \sum_{\text{Region Start}}^{\text{Region End}} \text{Similarity-weights} - \sum_{\text{Region Start}}^{\text{Region End}} \text{Penalties}$$

where:

$$\text{Penalty} = \text{Gap-penalty} + \text{Size-of-gap} \times \text{Gap-size-penalty}$$

Global: Needleman-Wunsch Algorithm

- Finding the optimal alignment via dynamic programming
- Example alignment

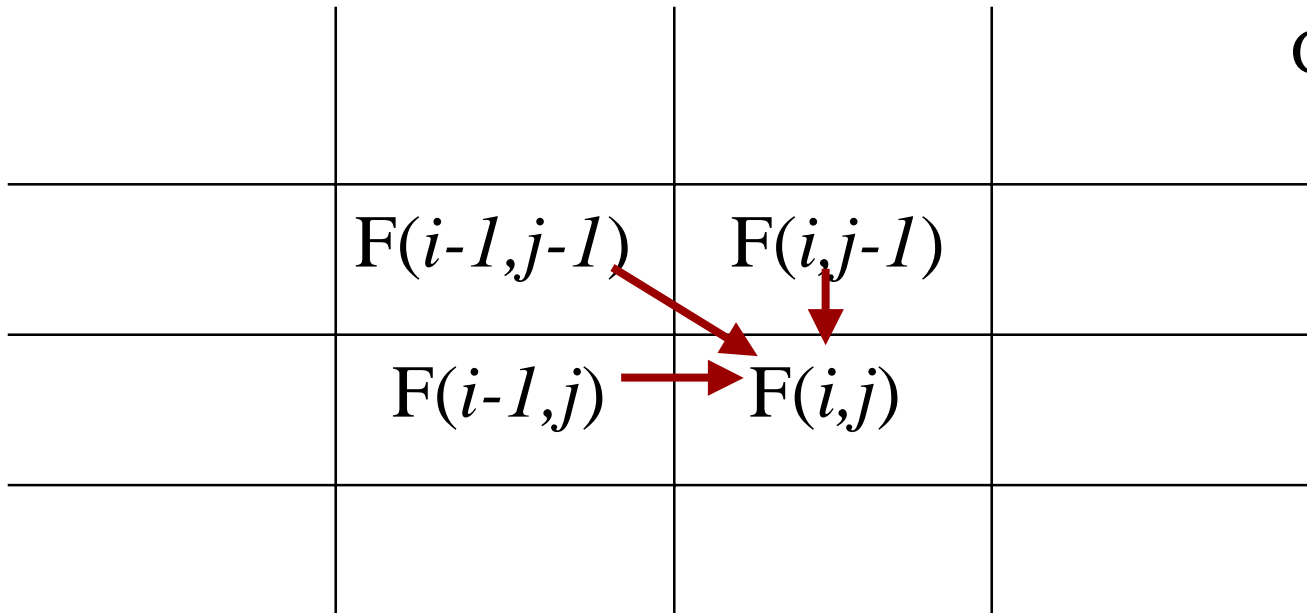


Alignment Recursion

$$F(i, j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - \gamma \\ F(i, j-1) - \gamma \end{cases}$$

E.g., $s(x, y) = 2I_{\{x=y\}}, \gamma = 1$

	C	A	T	T	G	
T	0	-1	-2	-3	-4	-5
C	-1	0	-1	0		
A	-2	1	0	-1		
T	-3	0	3	2	1	
G	-4	-1	3	5	4	
	-5	-2	2	4	5	6



Substitution/Scoring Matrices

- Pam matrices (*Dayhoff et al. 1978*) --- phylogeny-based.
PAM1: expected number of mutation = 1%

	C	S	T	F	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	12																				C
S	0	2																			S
T	-2	1	3																		T
F	-3	1	0	6																	F
A	-2	1	1	1	2																A
G	-3	1	0	-1	1	5															G
N	-4	1	0	-1	0	0	2														N
D	-5	0	0	-1	0	1	2	4													D
E	-5	0	0	-1	0	0	1	3	4												E
Q	-5	-1	-1	0	0	-1	1	2	2	4											Q
H	-3	-1	-1	0	-1	-2	2	1	1	3	6										H
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6									R
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5								K
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6							M
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5						I
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6					L
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4				V
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9			F
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10		Y
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17	W
	C	S	T	F	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

PAM250 matrix, log-odds representation

BLO(ck)SU(bstitution)M(atrix) (Henikoff & Henikoff 1992)

- Derived from a set (2000) of aligned and ungapped regions from protein families; emphasizing more on chemical similarities (versus how easy it is to mutate from one residue to another). BLOSUM x is derived from the set of segments of $x\%$ identity.

BLOSUM62 Matrix, log-odds representation

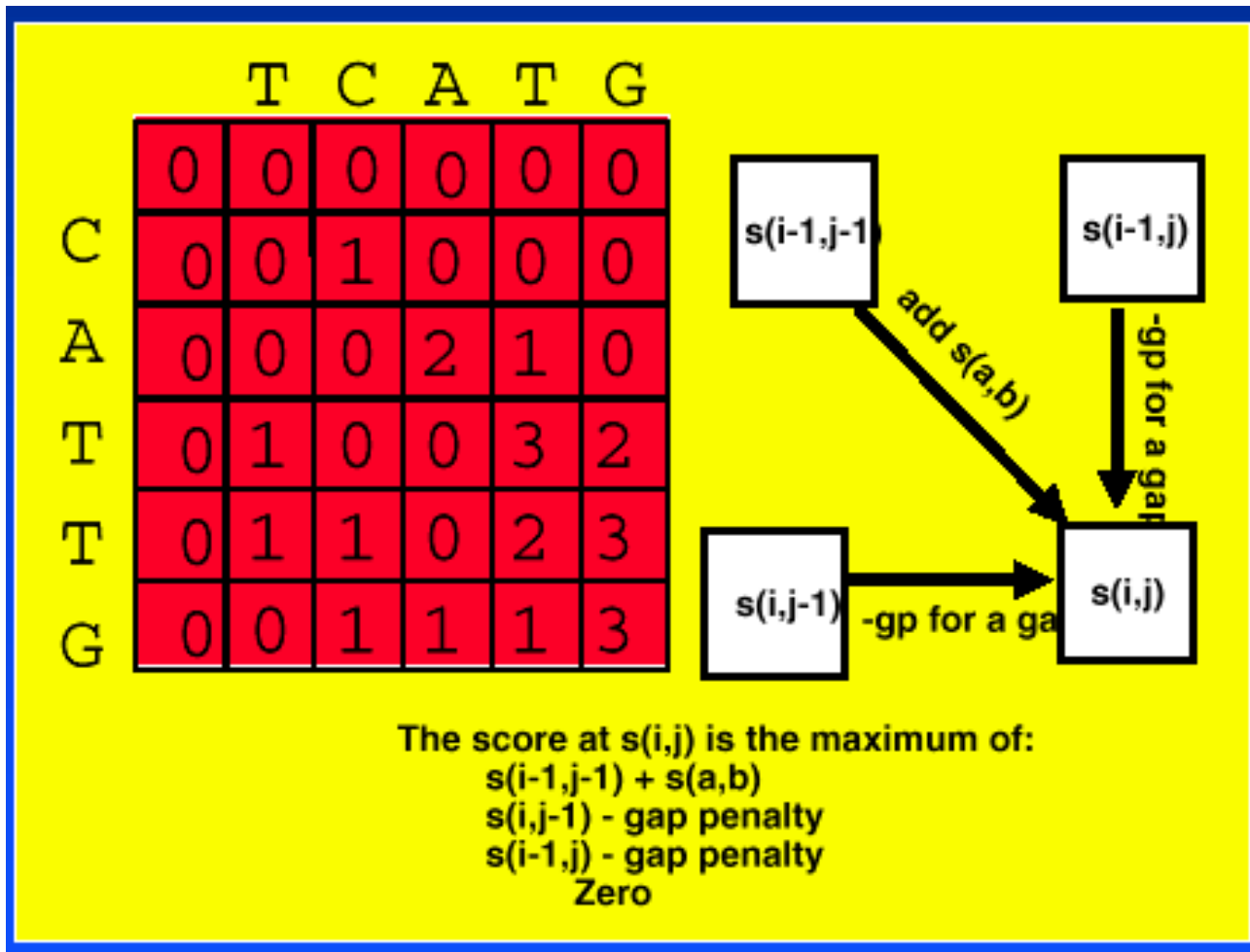
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
F	-1	1	5																		F
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1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

Gap Penalties

- Linear score: $\gamma(g) = -gd$
 - Typically: $d=8$, in unit of half-bits ($=4\log 2$)
- Affine score: $\gamma(g) = -d - (g-1)e$
 - d : gap opening penalty; e : gap extension penalty
 - Typical $d=12$, $e=2$ (in unit of half-bits)
- Gap penalty corresponds to log-probability of opening a gaps. For example, under the standard linear score, $P(g=k) = \exp(-dk) = 2^{-4k}$

Local Alignment: Smith-Waterman Algorithm

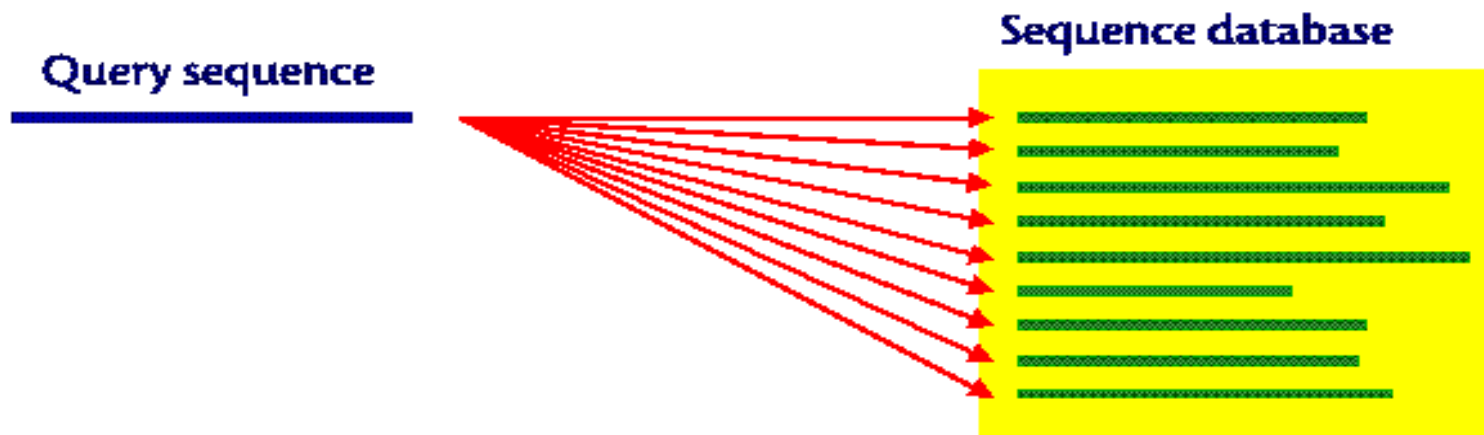
$$F(i, j) = \max \begin{cases} 0 \\ F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - d \\ F(i, j-1) - d \end{cases}$$



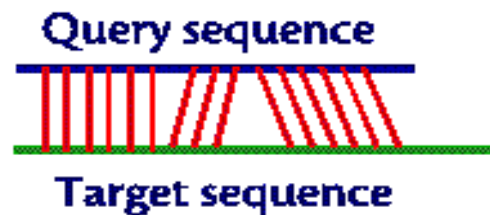
Sequence comparison and data base search

Overview

Similarity search (BLAST, FASTA)



Recognize by matching two sequences:



BLAST

(Altschul et al. 1990)

QTVGLMIVYDDA



- Create a word list from the query;
 - word length =3 for protein and 12 for DNA.
- For each listed word, find “neighboring words” (~ 50), $S(W, W') > T$
- For each sequence in the database, search exact matches to each word in the set.
- Extend the hits in both directions until score drops below X
- No gap allowed; use Karlin-Altschul statistics for significance
- New versions (>1.4) of BLAST gives gapped alignments.
- Compute Smith-Waterman for “significant” alignments
- BLASTP (protein), BLASTN (DNA), BLASTX (pr \rightarrow DNA).

BLAST 2.0

- Two word hits must be found within a window of A residues in order to trigger extension
- Gapped extension from the middle of ungapped HSPs
- Position-specific iterative (PSI-) BLAST.
 - Profile constructed on the fly and iteratively refined.
 - Begin with a single query, profile constructed from those significant hits; use the profile to do another search, and iterate the procedure till “convergence”

A Bayesian Model for Pairwise Alignment

Missing data --- Alignment matrix

$A_{i,j} = 1$ if residue i of sequence 1 aligns with residue j of sequence 2, 0 otherwise.

Observed data pair of sequence $R^{(1)}, R^{(2)}$

$$P(R_i^{(1)}, R_j^{(2)} \mid A, \Theta) = \Theta_{R_i^{(1)}} + \Theta_{R_j^{(2)}} + A_{i,j} \Psi_{R_i^{(1)}, R_j^{(2)}}$$

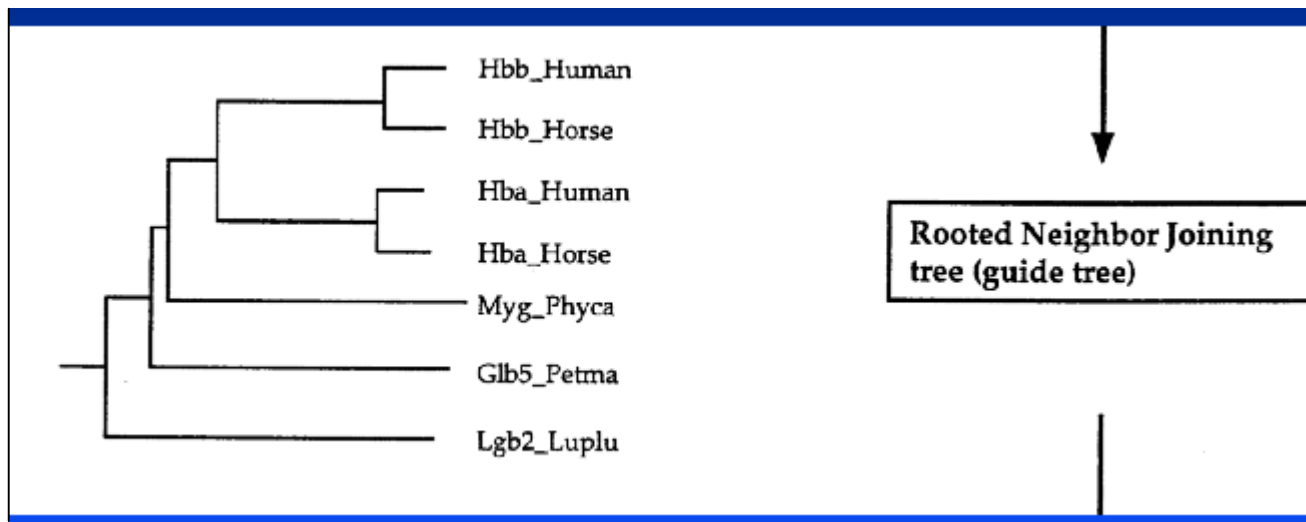
$$\Psi_{R^1, R^2} \quad \text{PAM or Blosom} \quad \sum_i A_{i,j} \leq 1 \quad \sum_j A_{i,j} \leq 1$$

ClustalW Step 1: Distance Matrix

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

Pairwise alignment:
Calculate distance matrix

ClustalW Step 2: bBuild the Tree



ClustalW Step 3: Progressive Alignment

```

-----VHLTPEEKSAVTALWGKV---VDEVGGEALGRLLVVYPWTQRFFESFGDLST
-----VQLSGEEEKAAVLALWDKVN---EEVVGGEALGRLLVVYPWTQRFFDSFGDLSN
-----VLSPADKTNVKAANGKVGAHAGEYGAELERMFLSFPTTKTYFPHFDLS--
-----VLSAADKTNVKAAWSKVGGHAGEYGAELERMFLGFPTTKTYFPHFDLS--
-----VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDKFKHLKT
PIVDTGSAVPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTFAAQEFFPKFKGLTT
-----GALTESQAALVKSSWEEFNANIPKHTHREFILVLEIAFAAKDLFSPLKGTSE
      *      *      *      *

```

```

PDAVMGNPKVKAHGKKVLGAFSDGLAHLD-----NLKGTFFATLSELHCDKLHVDPENFRL
PGAVMGNPKVKAHGKKVLHSPGEGVHHLA-----NLKGTFAALSELHCDKLHVDPENFRL
----HGSAQVKGHGKKVADALTNAVAHVDEL-----DMPNALSALSDLHAHKLKRVDPVNFKL
----HGSAQVKAHGKKVGDALTLAVGHLD-----DLPGALSNDLSDLHAHKLKRVDPVNFKL
EAEMKASEDLKKHGVTVLTAALGAILKKKG-----HHEAELKPLAQSHATKHKIPIKYLEF
ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLGSKHAKSFQVLPQYFKV
VP--QNNPELQAHAGKVFELVYEAAIQQLQVTVGVVVDATLKNLGSVHVSQKGVADAHFPV
      *      *      *

```

```

LGNVLVCVLAHHPGKEFTPPVQAAYQKVVAGVANALAHKYH-----
LGNVLVVVLAHHPGKDFTPPELQASYQKVVAGVANALAHKYH-----
LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTISKYR-----
LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTISKYR-----
ISEAIIHVLHSHRHPGDFGADAQGAMNKALELPRKDIAAKYKELGYQG
LAAVIADTVAAG-----DAGFEKLMISMICILLRSAY-----
VKEAIIKTIKEVWGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---

```

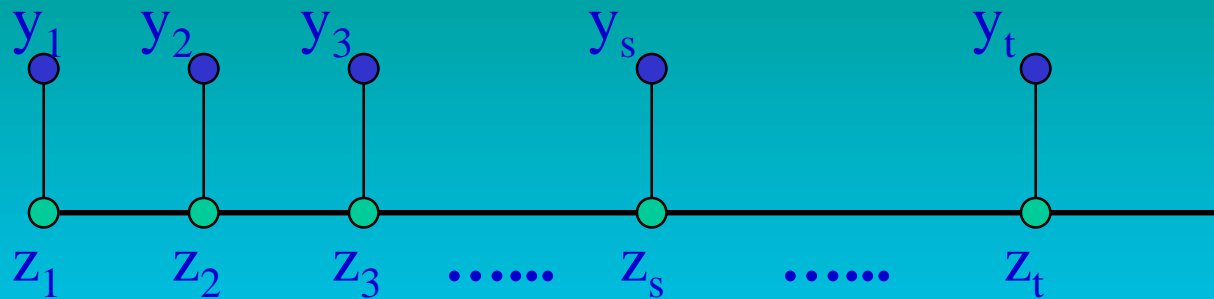
Progressive alignment:
Align following the guide tree

However

- No explicit model to guide for the alignment --- heuristic driven.
- Tree construction has problems.
- Overall, not sensitive enough for remotely related sequences.

The Hidden Markov Model

- For given z_s , $y_s \sim f(y_s / z_s, \theta)$, and the z_s follow a Markov process with transition $p_s(z_s / z_{s-1}, \phi)$.

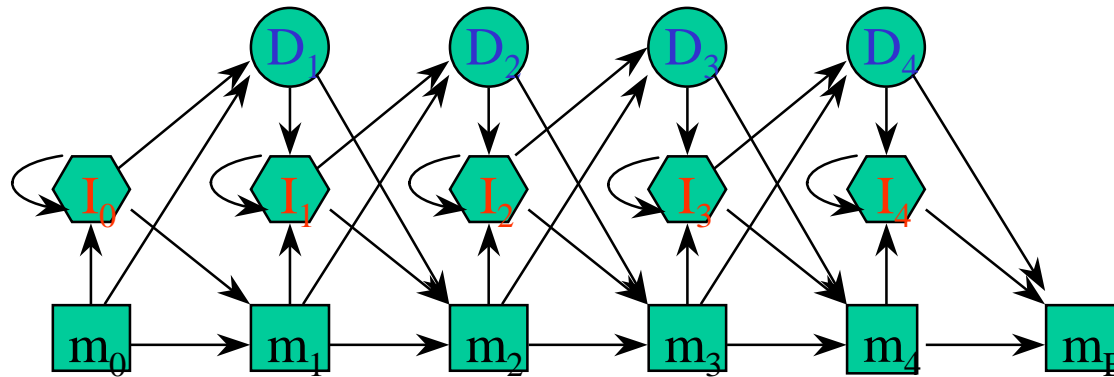


$$\pi_t(\underline{z}_t) = p(z_1, \dots, z_t | y_1, \dots, y_t; \phi, \theta)$$

“The State Space Model”

What Are Hidden in Sequence Alignment?

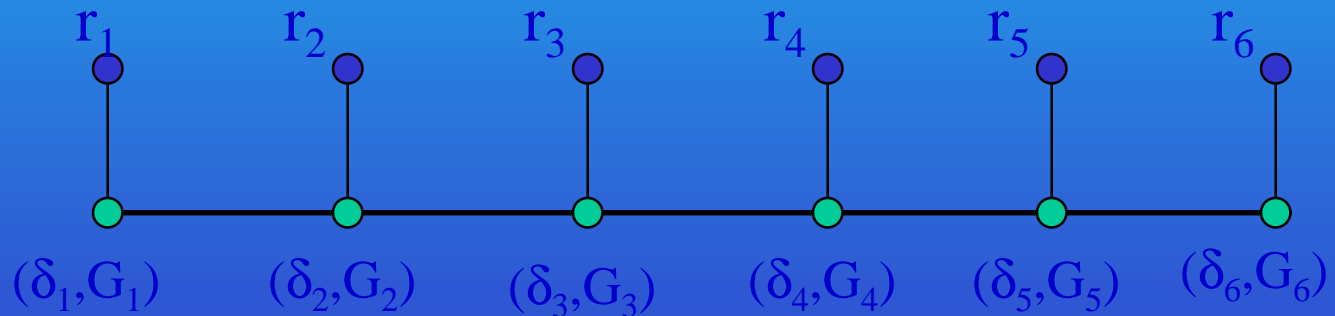
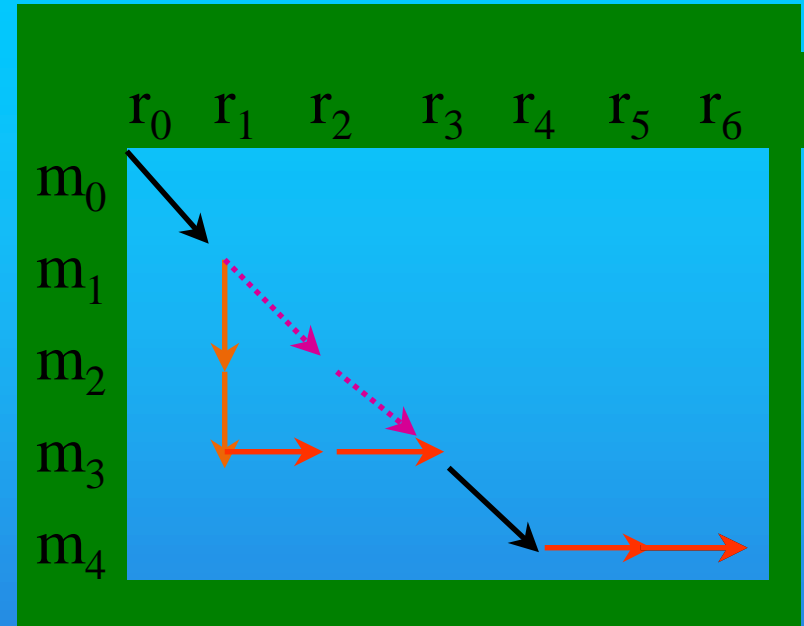
- *HMM Architecture*: transition diagram for the underlying Markov chain.



The *Path*

$\delta_i = \# \text{ of deletions}$

$G_i = \begin{cases} 0 & \text{if generated by insertion} \\ 1 & \text{if generated by a match} \end{cases}$



Solid path: $(0,1) \rightarrow (2,0) \rightarrow (2,0) \rightarrow (2,1) \rightarrow (2,0) \rightarrow (2,0)$
Dashed path: $(0,1) \rightarrow (0,1) \rightarrow (0,1) \rightarrow (0,1) \rightarrow (0,0) \rightarrow (0,0)$

Pfam alignment view (<http://pfam.wustl.edu/browse.shtml>)

Jalview alignment editor

File Edit Font View Colour Calculate Align Help

	10	20	30	40	50	60	70	80	90	100	110								
ENV HV1A2/33-509	PVWKEA	TTILFCAS	DARAYD	TEVHN	VWATHACV	PTDPNP	QEVV	LGNVTEN	FNMWK	---	NNMVE	QMOED	II	SLWD	QSLKPC	VKLTPL	CVTLN	CTD	-----
ENV HV1B1/34-511	PVWKEA	TTILFCAS	DAKAYD	TEVHN	VWATHACV	PTDPNP	QEVV	LGNVTEN	FNMWK	---	NNMVE	QMHED	II	SLWD	QSLKPC	VKLTPL	CVSLK	CTDL	KN-----
ENV HV1J3/33-523	PVWKEA	ATILFCAS	DAKAYD	TEVHN	VWATHACV	PTDPNP	QEVV	LGNVTEN	FNMWK	---	NNMVE	QMHED	II	SLWD	QSLKPC	VKLTPL	CVTLN	CTD	WGN-----
ENV HV1W1/33-510	PVWKEA	TTILFCAS	DAKAYS	TEAHK	VWATHACV	PTDPNP	QEVV	LGNVTEN	FNMWK	---	NNMVE	QMHED	II	SLWD	QSLKPC	VKLTPL	CVTLN	CTD	KNIT-----
ENV HV1BN/34-507	PVWKEA	NTILFCAS	DAKAYD	TEIHN	VWATHACV	PTDPNP	QELV	MGNVTEN	FNMWK	---	NNMVE	QMHED	II	SLWD	QSLKPC	VKLTPL	CVTLN	CTD	HDFNA-----
ENV HV1RH/33-519	PVWKEA	TTILFCAS	EAKAYK	TEVHN	VWAKHACV	PTDPNP	QEVV	LGNVTEN	FNMWK	---	NNMVE	QMHED	II	SLWD	QSLKPC	VKLTPL	CVTLN	CTD	A-----
ENV HV1OY/33-509	PVWKEA	TTILFCAS	DARAYA	TEVHN	VWATHACV	PTDPNP	QEVV	LGNVTEN	FNMWK	---	NNMVE	QMOED	II	SLWD	QSLKPC	VKLTPL	CVTLN	CTD	VNTTS-----
ENV HV1C4/35-522	PVWKEA	TTILFCAS	DAKAYD	TEAHN	VWATHACV	PTDPNP	QEVV	LGNVTEN	FNMWK	---	NNMVE	QMHED	II	SLWD	QSLKPC	VKLTPL	CVTLN	CTD	LNTN-----
ENV HV1ZH/33-511	PVWKEA	TTILFCAS	DAKAYD	TEKHN	VWATHACV	PTDPNP	QEL	SLGNVTEN	FNMWK	---	NNMVE	QMHED	VI	SLWD	QSLKPC	VKLTPL	CVTLN	CTD	SCHNITIK-----
ENV HV1EL/33-508	PVWKEA	TTILFCAS	DAKAYS	TEAHN	IWATHACV	PTDPNP	QEL	IALGNVTEN	FNMWK	---	NNMVE	QMHED	II	SLWD	QSLKPC	VKLTPL	CVTLN	CTD	SDELR-----
ENV HV1Z8/33-518	PVWKEA	TTILFCAS	DAKAYS	TEPEAHN	IWATHACV	PTDPNP	PRE	IEMGNVTEN	FNMWK	---	NNMVE	QMHED	II	SLWD	QSLKPC	VKLTPL	CVTLN	CTD	TNAGG-----
ENV HV1ND/33-501	PIWKEA	TTILFCAS	DAKAYK	TEAHN	IWATHACV	PTDPNP	QEL	IALGNVTEN	FNMWK	---	NNMVE	QMHED	II	SLWD	QSLKPC	VKLTPL	CVTLN	CTD	DELR-----
ENV HV1MA/33-513	PVWKEA	TTILFCAS	DAKAYS	TEVHN	IWATHACV	PTDPNP	QEL	IALGNVTEN	FNMWK	---	NNMVE	QMHED	II	SLWD	QSLKPC	VKLTPL	CVTLN	CTD	TVNG-----
ENV SIVCZ/33-496	PVWHDAD	PVLF	CASDAKAH	STEAHN	IWATHACV	PTDPS	QEVV	LGNVTEN	FNMWK	---	NNMVE	QMHED	II	SLWD	QSLKPC	VKLTPL	CVTLN	CTD	QSKANFS-----
ENV HV2BE/24-510	PAWKNA	SIP	LCATKNR	-----	DTWGT	IQCLP	DNDY	QEIILN	-VTEAF	DAWN	---	NTVTE	QAVED	VWHL	FETSIK	PCVKLTPL	CVAMNC	SRVQ	GNTPNP-----
ENV HV2CA/25-512	PAWKNA	SIP	LCATKNR	-----	DTWGT	IQCLP	DNDY	QEIILN	-VTEAF	DAWD	---	NTITE	QAIED	VWHL	FETSIK	PCVKLTPL	CVAMNC	-----	ISTSI-----
ENV HV2D1/24-501	PAWRNA	SIP	LCATKNR	-----	DTWGT	IQCLP	DNDY	QEIILN	-VTEAF	DAWD	---	NTVTE	QAIED	VWRL	FETSIK	PCVKLTPL	CVAMNC	NITSG	TATPSP-----
ENV HV2G1/23-502	PVWRNA	SIP	LCATKNR	-----	DTWGT	IQCKP	DNDY	QEIILN	-VTEAF	DAWD	---	NTVTE	QAVED	VWSL	FETSIK	PCVKLTPL	CVAMNC	-----	ST-----
ENV HV2NZ/24-502	PAWRNA	SIP	LCATKNR	-----	DTWGT	IQCLP	DNDY	QEIILN	-VTEAF	DAWN	---	NTVTE	QAVED	VWNL	FETSIK	PCVKLTPL	CVAMNC	TR	-----
ENV SIVM1/24-528	PAWRNA	SIP	LCATKNR	-----	DTWGT	TQCLP	DNDY	SELALN	-VTEAF	DAWE	---	NTVTE	QAIED	VWQL	FETSIK	PCVKLSPL	CIIMRC	NKSETD	KWGLTK-----
ENV HV2D2/24-513	PAWRNA	SIP	LCATKNR	-----	DTWGT	VQCLP	DNDY	TEIRLN	-ITEAF	DAWD	---	NTVTE	QAVDD	VWRL	FETSIK	PCVKLTPL	CVAMNC	SKTETN	---PGNA-----
ENV SIVA1/24-538	PVWKNSS	VQAF	CMPTT	-----	RLWA	TTCIP	DDHD	YTEVELN	-ITEAF	EAWD	---	NPLVK	QAES	NIHLL	FEQTL	KPCVKLSPL	CIKMNC	VELKGS	-----
ENV SIVAI/22-522	PVWKNSS	VQAF	CMPTT	-----	RLWA	TTCIP	DDHD	NTTEVELN	-ITEAF	EAWD	---	NPLVK	QAES	NIHLL	FEQTL	MRPCVKLSPL	CIKMSC	VELNGT	-----
ENV_SIVGB/47-569	PVWKEA	KTHLIC	ATDWS	-----	SLWV	TTCIP	SLPD	YDEVEIP	DIKEN	FTGLIRE	NQIV	YQAW	HAMG	SMLD	TILK	PCVKLIN	PYCVK	MCCO	ETENVS-----

More Restricted Models

(Liu et al. 99, Neuwald et al. 97)

- **Block Motif Propagation Model:** limit the total number of gaps but no deletions allowed.



Motif 1

Motif 2

Motif 3

Motif 4

Representative Alignment

C

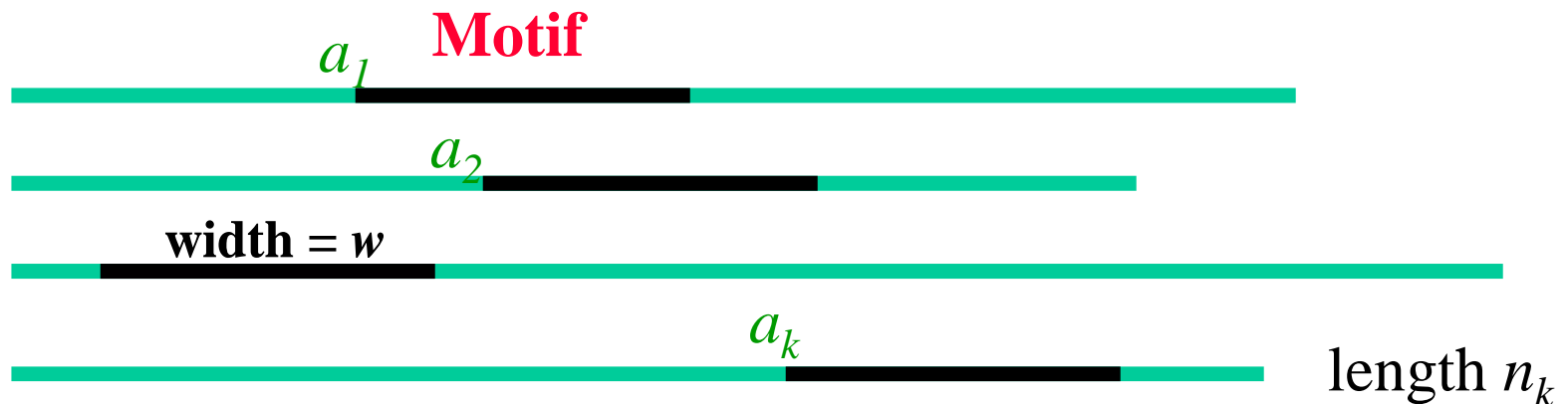
	helix:	*****	bbbbbbbbbb	ccccccccccc	eeeeeeeeeeeeeeeeeeee	
HBA_HUMAN	2	LSPADKTNVKAAWGKV	(6)	YGAEALERMFLSFPPTTKTYFPHF	(5)	SAQVKVGHGKKVADALTNVAHAVD
MYBO	3	LSDGEWQLVLNAWVKV	(6)	HGQEVLIIRLFTGHPETLEKFDKF	(11)	SEDLKKHGNTVLTALGGILKKKG
GLBT_CHITH	9	MTDAQVAAVKGDWEKI	(2)	SGVEILYFFLNKFPGNFPMFKKL	(9)	TAEFKDQADKIIAFLQGVIKLG
GLB3_LUMTE	24	CSEEDHRIVQKQWDIL	(12)	FGRLLLTKLAKDIPDVNDLFKRV	(6)	GPKFSAHALRILNGLDLAINLLD
GLB3_LUCPE	4	LTGPQKAALKSSWSRF	(6)	NGTNFYMDLFKAYPDTLTPFKSL	(11)	HPTMKAQALVFCDGMSSFVDNLD
1HLB	13	LTLAQKKIVRKTWHQL	(6)	FVTDVFIIRIFAYDPSAQNKFPQM	(10)	SRQMQAHAIRVSSIMSEYVEELD
VUU33205_1	4	FSDKQEGLVNGAYEAF	(6)	YSVVFYTTILEKAPAAKNLFSFL	(7)	NPKLTGHAEKLFGLVRDSAAQLR
HMPA_ALCEU	2	LTQKTQDIVKATAPVL	(6)	IIKCFYQRMFEAHPELKNVFNMA	(0)	HQEQQQQQALARAVYAYAENIE
GLB1_ARTSX	9	LSGLEKNAILDWTKV	(6)	VGKATFGKLFAAHPEYQGMFRFF	(10)	SPKFAAHTQRVVSALDQTLALN
GLB_ASCSU	162	INKHGRHAVRHQCMRS	(14)	NGIDLKMHFENYPSMREAFKDR	(10)	DPFFVKQGRILLACHLLCASVD
W01C9.5	24	LTPSQVSVVRRSRWHI	(7)	VLTRCFSRLESNCPIVSQCFQSA	(10)	VRTVADHAKYLLQLLDKIIEGDV
F46C8.7	61	LSKIQRRAIRFTWHRL	(11)	VFEVFDKLVKNLPNIRDMFSTR	(10)	TSTLRDHSKNCVKMIDSVIKNFD
R13A1.8	161	IDKESCEVVADSWRLV	(12)	FGLFVQRFVFSKIPMLRPLFGLS	(10)	NHPVRRHARLFTSILHISVKNVD
C52A11.2	32	LNKKDRTLLRETWQRL	(6)	VGLIFLDIVNDIEPDLKKVFGVD	(9)	MPKFGGHIIRFYEFMEQLTSMLG
F19H6.2	50	LTRRERILLEQSWRKT	(7)	IGSKIFFMVLTAQPDIAKIFGLE	(9)	DPRFRQHALVYTKTLDVIRNLD
F49E2.4	14	ITDEEVTAIRDVWRRRA	(4)	VGKKILQTLIEKRPKFAEYFGIQ	(11)	SKEFHLQAHRIQNFLDTAVGSLG
F52A8.4	70	PNVYEKELLRRTWSDE	(6)	LGSAIYCYIFDHNPNCKQLPFI	(10)	SKEFRSQALKFVQTLAQVVKNIY
C29F5.7	28	LNAKTKKLVIQEWPRV	(6)	LFTEIWHKSASTRSTSIKLAFGIA	(7)	NAAFGLSSTIQAFFYKLIITYE

	helix:	ffffffffffff	ffgggggggggggggggggg	hhhhhhhhhhhhhh				
HBA_HUMAN	(1)	MPNALSA LSDLHAHK	(1)	RVDPVNFKLLSHCLLVTL	(14)	SLDKFLASVSTVL	136	α-hemoglobin
MYBO	(1)	HEAEVKHLAESHANK	(1)	KIPVKYLEFISDAIIVHL	(14)	AMSKALELFRNDM	143	bovine myoglobin
GLBT_CHITH	(5)	AKALLNQLGTSHKAM	(0)	GITKDQFDQFRQALTELL	(14)	TVDLMFHVIFNAL	146	midge globin
GLB3_LUMTE	(4)	LDAALDHLAQHEVR	(1)	GVQKAHFKKFGEILATGL	(10)	AWKSCCLKGILTKEI	164	earthworm globin
GLB3_LUCPE	(4)	LVVLLQKMAKLHFNR	(0)	GIRIKELRDGYGVLLRYL	(12)	AWEDFIAYICRVQ	144	clam hemoglobin
1HLB	(3)	LPELLATLARTHDLN	(0)	KVGADHYNLFKVLMEAL	(14)	AWAKAFSVVQAVL	153	sea cucumber globin
VUU33205_1	(3)	GVVADAALGAVHSQK	(0)	AVNDAQFVVVKEALVKT	(14)	AVELAYDELAAAI	141	leghemoglobin
HMPA_ALCEU	(4)	LMAVLKNIANKHASL	(0)	GVKPEQYPIVGEHLLAAI	(14)	AWAQAYGNLADVL	133	flavo-hemoprotein
GLB1_ARTSX	(5)	FVYMIKELGLDHINR	(0)	GTDRSFVEYLKESLGDV	(3)	TVQSFGEIVNFL	140	brine shrimp globin
GLB_ASCSU	(4)	FHMYVHELMERHERL	(2)	QLPDQHWTFWKLFEFL	(12)	AWAVIGKEFAYEA	311	ascaris globin
W01C9.5	(0)	DSEFLREIGANHVCL	(4)	GFSTQEWDRFQEIIMVEVI	(14)	AWRL LICSFIELI	166	0.0003
F46C8.7	(10)	SENDPRVIGRAHSIL	(3)	GLAGNYWEKFGVEMIDVV	(14)	AWVIFTACLVDQM	216	0.000000004
R13A1.8	(5)	VAPT VFKYGERHYRP	(4)	HMTEENVRVFCAQIVCTV	(15)	SWIELMRYLGQKL	314	0.00000006
C52A11.2	(7)	AWQLVRKTGRSHVRQ	(8)	QMEKNYFEIVINVFIERL	(34)	VWKKFLNTVISQM	203	0.00000003
F19H6.2	(4)	LEVYFENLGKRHVAM	(3)	GFEPGYWETFAECMTQAA	(13)	AWRN LISCIISFM	193	< 0.0000000000001
F49E2.4	(6)	VFDMAHRIGQIHFYR	(2)	NFGADNWLVFKKVTVDQV	(49)	GWNK LMTVIVREM	193	0.000000004
F52A8.4	(4)	TESFLYMGQKHVKF	(3)	GFKHEYWDIFQDAMEFAL	(20)	VWRTLALYTTVHM	220	0.00000007
C29F5.7	(5)	VREACEQLGARHVDF	(3)	GFNSHFWDIFLVCMAEKI	(19)	AWQRVINSIVHQM	175	0.00005

Finding Repetitive Patterns

```
1 taatgtttgtgctgggttttgtggcatcgggcgagaatagecgcgtgggtgtgaaagactgtttttttga
2 gacaaaaacgcgtaacaaaagtgtctataatcacggcgagaaaagtccacattgattatttgcacggcg
3 acaaatcccaataacttaattattgggatttgttatataactttataaattcctaaaattacacaa
4 cacaaagcgaaagctatgctaaaacagtcaggatgctacagtaatacattgatgactgcatgtatgc
5 acggtgctacacttgtatgtagcgcacatctttctttacgggtcaatcagcatgggtgttaaatgacacg
6 agtgaattatttgaaccagatcgcattacagtgatgcaaacttqtaagtagatttccttaattgtgat
7 ggcataaaaaacggctaaattcttggtaaacgattccactaatttattccatgtcacacttttcgc
8 gctcggcggggttttttgttatctgcaattcagtaacaaaacggtgatcaaccctcaatttccctt
9 aacgcaattaatgtgagttagctcactcattagggcaccccaggctttacactttatgcttcgggetg
10 acattaccgccaattctgtaacagagatcacacaaagcgcacgggtggggcgtaggggcaaggaggatgg
11 ggaggaggcgggaggatgagaacacggcttctgtgaaactaaaccgagggtcatgtaaggaatttcgtga
12 gatcagcgtcgttttaggtgagttgttaataaagatttggaaattgtgacacagtgcaaattcagacac
13 gctgacaaaaaagattaaacataccttatacaagacttttttttcatatgcctgacggagttcacact
14 ttttttaaacattaaaattcttacgtaatttataatcttttaaaaaaagcatttaaatattgctccccga
15 cccatgagagtgaaattgttgtgatgtggttaacccaatttagaattcgggattgacatgtcttaccaa
16 ctggettaactatgcggcatcagagcagattgtactgagagtgcaccatattgcgggtgtgaaataccgc
17 ctgtgacggaagatcacttcgcagaaataaataaatcctgggtgtccctgttgataccgggaagccctgg
18 gatttttataactttaacttgttgatatttaaaaggatatttaattgtaataacgatactctggaaagtat
```

Motif Alignment Model



The missing data: Alignment variable: $A = \{a_1, a_2, \dots, a_k\}$

- Every **non-site positions** follows a common multinomial with $\mathbf{p}_0 = (p_{0,1}, \dots, p_{0,20})$
- Every position i in the motif element follows probability distribution $\mathbf{p}_i = (p_{i,1}, \dots, p_{i,20})$

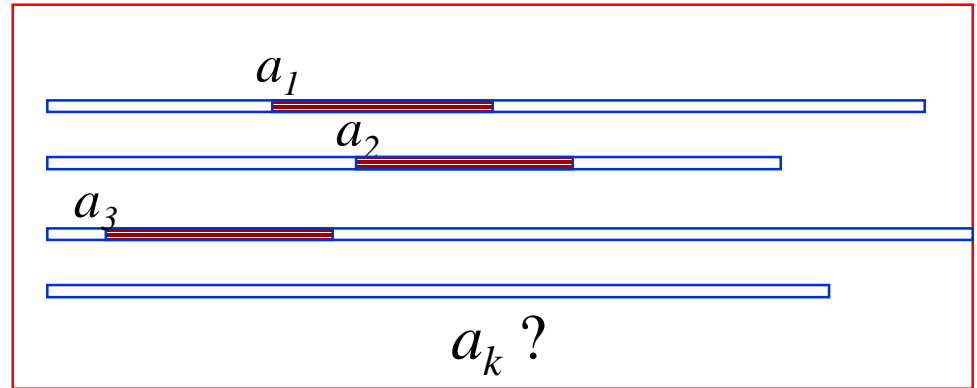
The Algorithm

- Initialized by choosing random starting positions

$$a_1^{(0)}, a_2^{(0)}, \dots, a_K^{(0)}$$

- Iterate the following steps many times:
 - Randomly or systematically choose a sequence, say, *sequence k*, to exclude.
 - Carry out the *predictive-updating* step to update a_k
- Stop when no more observable changes in likelihood
- Available Programs:
 - Gibbs motif sampler (<http://www.wadsworth.org/res&res/bioinfo>)
 - Bioprospector (<http://bioprospector.stanford.edu>)
 - MACAW (<http://www.ncbi.nlm.nih.gov>)

The PU-Step



1. Compute predictive frequencies of each position i in motif

C_{ij} = count of amino acid type j at position i .

C_{0j} = count of amino acid type j in all non-site positions.

$q_{ij} = (c_{ij} + b_j) / (K - 1 + B)$, $B = b_1 + \dots + b_K$ “pseudo-counts”

2. Sample from the predictive distribution of a_k .

$$P(a_k = l + 1) \propto \prod_{i=1}^w \frac{q_{i,R_k}(l+i)}{q_{0,R_k}(l+i)}$$

Using MACAW

The screenshot displays the MACAW software interface. The main window, titled "Untitled", shows a "Schematic" view of 18 sequences. The sequences are represented by horizontal lines, and the alignment is visualized as black bars connecting corresponding positions across the sequences. A scale at the top of the schematic ranges from 0 to 100. The bottom status bar indicates "18 sequences" and "IUPAC +5/-6".

Search Results window:

keep	m	len	Info.
18	19	0.260	
18	18	0.259	
18	20	0.250	
18	21	0.245	
18	22	0.243	
18	23	0.238	
18	24	0.235	

Buttons: View / Edit..., Link, Unlink, Keep, Help. Keep when linked

Idea 2: Mixture modeling

- View the dataset as a long sequence with k motif types:



- **Idea:** partition the input sequence into segments that correspond to different (unknown) motif models.
- It is a mixture model (unsupervised learning).
- Implement a predictive updating scheme.

Special Case: Bernoulli Sampler

- Sequence data: $\mathbf{R} = r_1 r_2 r_3 \dots r_N$

- Indicator variable: $\Delta = \delta_1 \delta_2 \delta_3 \dots \delta_N$

$$\delta_i = \begin{cases} 1, & \text{if it is the start of an element} \\ 0, & \text{if not.} \end{cases}$$

parameter for the motif model

- Likelihood: $\pi(\mathbf{R}, \Delta | \Theta, \varepsilon)$, ε is the prior prob for $\delta_i=1$

- Predictive Update:

$$\frac{\pi(\delta_k = 1 | \Delta_{[-k]}, R)}{\pi(\delta_k = 0 | \Delta_{[-k]}, R)} = \frac{\hat{\varepsilon}}{1 - \hat{\varepsilon}} \prod_{i=1}^w \left(\frac{\hat{p}_{i, r_{k+i-1}}}{\hat{p}_{0, r_{k+i-1}}} \right)$$

References (self)

- **Durbin R. et al. (1997).** *Biological Sequence Analysis*, Cambridge University Press, London.
- **Liu, J.S. (2001).** *Monte Carlo Strategies in Scientific Computing*, Springer-Verlag, New York.
- **Liu, J.S. and Lawrence, C.E. (1999).** Bayesian analysis on biopolymer sequences. *Bioinformatics*, 138-143.
- **Liu, J.S., Neuwald, A.F., and Lawrence, C.E. (1999)** . Markovian structures in biological sequence alignments. *J. Amer. Statist. Assoc.*, 1-15.
- **Neuwald, A.F., Liu, J.S., Lipman, D.J., and Lawrence, C.E. (1997)** . Extracting protein alignment models from the sequence database. *Nucleic Acids. Res.* **25**, 1665-1677.
- **Liu, J.S., Neuwald, A.F., and Lawrence, C.E. (1995)** . Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Amer. Statist. Assoc.* **90**, 1156-1170.
- **Lawrence, et al. (1993).** *Science* **262**, 208-214.