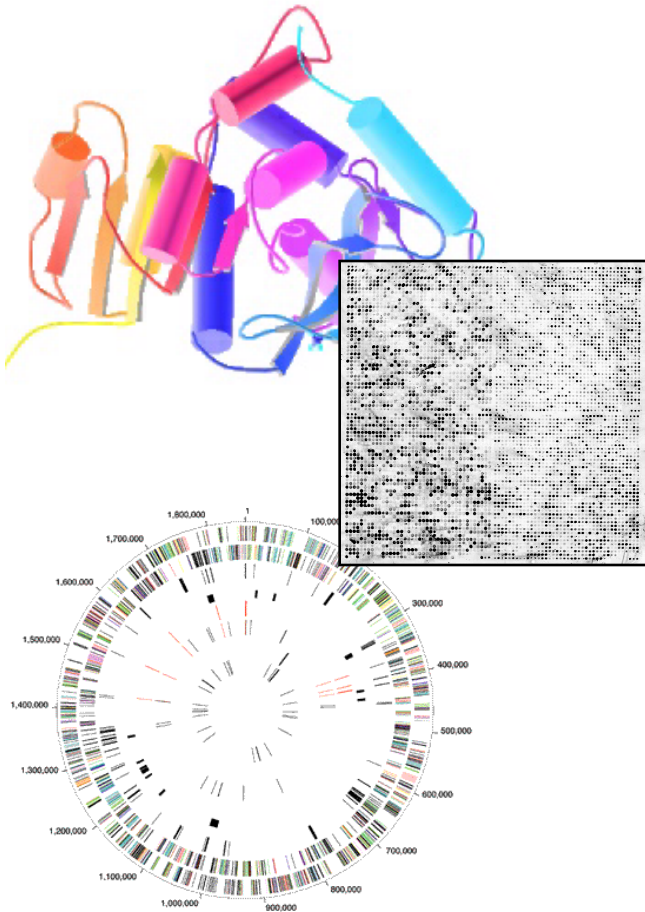


# BIOINFORMATICS

## Multiple Sequences



Mark Gerstein  
Yale University  
[GersteinLab.org/courses/452](http://GersteinLab.org/courses/452)  
(MG lect. #3, last edit in spring '16)

# Multiple Sequence Alignment Topics

- Multiple Sequence Alignment
- Motifs
  - Fast identification methods
- Profile Patterns
  - Refinement via EM
  - Gibbs Sampling
- HMMs
- Applications
  - Protein Domain databases
  - Regression vs expression

- One of the most essential tools in molecular biology

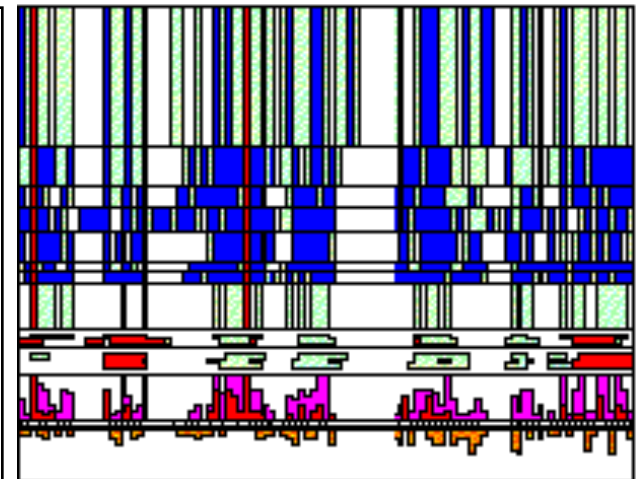
It is widely used in:

- Phylogenetic analysis
- Prediction of protein secondary/tertiary structure
- Finding diagnostic patterns to characterize protein families
- Detecting new homologies between new genes and established sequence families

# Multiple Sequence Alignments

- Practically useful methods only since 1987
- Before 1987 they were constructed by hand
- The basic problem: no dynamic programming approach can be used
- First useful approach by D. Sankoff (1987) based on phylogenetics

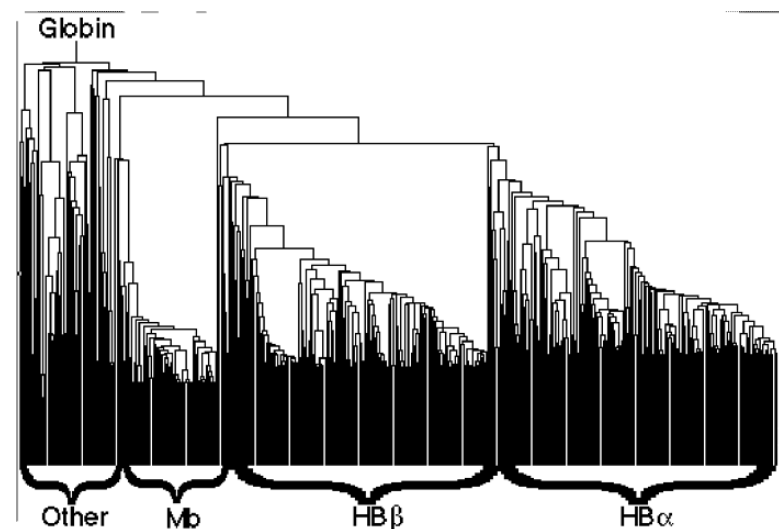
AGRI_CHICK	154	CVCPAS	...GS...	GVa	ESI	VCGS	DGK	YRSE	DLINKHAC	...DK	...	QEN	VFKK	FDGAC	201					
AGRI_RAT	165	GLCPTT	...GF...	GAp	DGT	VCGS	DGVD	YFSE	QQLLSHAC	...AS	...	QEH	IFKK	ENFGC	212					
FSA_HUMAN	116	CVCPAD	...CS...	NItw	KGP	VCGD	DGK	TYRNE	CALLKARC	...KE	...	QPE	LEVQ	YQGRG	164					
FSA_PIG	116	CVCPAD	...CS...	NItw	KGP	VCGD	DGK	TYRNE	CALLKARC	...KE	...	QPE	LEVQ	YQGRG	164					
FSA_RAT	116	CVCPAD	...CS...	NItw	KGP	VCGD	DGK	TYRNE	CALLKARC	...KE	...	QPE	LEVQ	YQGRG	164					
FSA_SHEEP	109	CVCPAD	...CS...	NItw	KGP	VCGD	DGK	TYRNE	CALLKARC	...KE	...	QPE	LEVQ	YQGRG	157					
IAC1_BOVIN	14	CKVYTEA	...CT...	RE	...	YNP	ICDS	AAKTY	SNECTF	...ONEKM	..NN	...	DAD	IHFNF	FGEC	61				
IAC2_BOVIN	7	CAEPKDP	...KVYCT	RE	...	SNP	HCCS	NGET	YGNKCAF	...CKAVM	..KS	...	GGK	INLKH	RGGC	57				
IACA_PIG	7	QNVYRSH	...LFFCT	...	RQ	...	MDP	ICG	NGKSY	AMPGIF	...CSEKG	..LR	...	NQK	PDFG	HWGHC	57			
IACS_PIG	12	GDVYRSH	...LFFCT	...	RE	...	MDP	ICG	NGKSY	AMPGIF	...CSEKL	..GR	...	NEK	PDFG	HWGHC	62			
IAC_MACFA	33	QARYQLPG	...CH	RD	...	FNP	VCGD	DMIT	YFNECTL	...QMKIR	..ES	...	GQN	IKLRR	GGC	81				
IOV7_CHICK	94	GSPYLQVVRD	...GNTMVA	CH	RI	...	LKP	VCGS	DSFT	YDNECGI	...QAYNA	..BH	...	HTN	ISKLH	DGEC	150			
IOVO_ABUPI	8	GSDHPKP	...ACL	...	QE	...	QKPL	CGS	DNKTY	DNKGSF	...QNAVV	..DS	...	NGT	ITLSH	FGKC	56			
IOVO_ALECH	6	GSEYPKP	...ACT	...	LE	...	YRPL	CGS	PSKTY	GNKGNF	...QNAVV	..BS	...	NGT	ITLSH	FGKC	54			
IPSG_VULVU	68	GTEYSDM	...CT	MD	...	YRPL	CGS	DGK	NYSNK	GIF	...QNAVV	..RS	...	RGT	IFLAK	HGEC	115			
IPST_ANGAN	12	CGEMSAMHA	...CH	MN	...	FAP	VCGD	DGNT	YFNEGSL	...CFQRQ	..NT	...	KTD	ELITK	DDRC	61				
IPST_BOVIN	9	GTNEVNG	...CH	RI	...	YNP	VCGD	DGVT	YSNEGLL	...QMENK	..ER	...	QTP	VLIQK	SGGC	56				
IPST_PIG	9	GTSEVSG	...CH	KI	...	YNP	VCGD	DGVT	YSNEGVL	...QSENK	..KR	...	QTP	VLIQK	SGGC	56				
IPST_SHEEP	9	GTNEVNG	...CH	RI	...	YNP	VCGD	DGVT	YSNEGLL	...QMENK	..ER	...	QTP	VLIQK	SGGC	56				
OATP_HUMAN	439	GNVDCN	...CHS	...	KI	...	WDP	VCGD	NGLS	YLSA	CLA	...GC	..ET	..SI	...	GTG	INMV	FQNC	485	
OATP_RAT	439	GNTRCS	...CHS	...	TNt	...	WDP	VCGD	NGV	YMSA	CLA	...GC	..KKF	..VT	...	GTN	M	VFQDC	SC	486
PE60_PIG	37	GEMTESPD	...CHS	...	RI	...	YDP	VCGD	DGVT	YSECKL	...CLARI	..EN	...	KQD	IQIVK	DGEC	86			
PGT_RAT	444	GRRDCS	...CHS	...	DSf	...	FHP	VCGD	NGV	YVSP	GHA	...GC	...	SS	...	TNT	SSEAS	KEPI	488	
PSG1_MOUSE	33	GHDVAVG	...CH	RI	...	YDP	VCGD	DGVT	YFNEGVL	...QFENR	..KR	...	IEP	VLRK	GGGC	80				
QR1_COTJA	466	GICQDPA	...ACHS	...	tKD	...	YKR	VCGD	DNKTY	DGTC	QLFGTK	QLEGTKM	...	GROL	HLDY	MGAC	521			
SCI1_RAT	424	GVCQDPET	...CHp	...	aKI	...	LDQ	ACG	DNKTY	YASS	CHLFATK	QLEGTKK	...	GHK	QLDY	FGAC	479			
SPRC_BOVIN	93	GVCQDP.TS	...CHap	...	iGE	...	FER	VCS	DNKTY	YDSS	CHFFATK	QLEGTKK	...	GHK	LHLDY	IGFC	149			
SPRC_CAEEL	74	GECISK	...CHp	...	ldgDP	...	MDR	VCA	NNOT	YFISL	CDLYRER	OLCKR	..KS	kecska	fNAK	VHLEY	LGEC	135		
SPRC_MOUSE	92	GVCQDP.TS	...CHap	...	iGE	...	FER	VCS	DNKTY	YDSS	CHFFATK	QLEGTKK	...	GHK	LHLDY	IGFC	148			
SPRC_XENLA	90	GVCQDPST	...CHts	...	vGE	...	FEL	ICG	DNKTY	YDSS	CHFFATK	QLEGTKK	...	GHK	LHLDY	IGFC	146			



(LEFT, adapted from Sonhammer et al. (1997). "Pfam," Proteins 28:405-20. ABOVE, G Barton AMAS web page)

# Progressive Multiple Alignments

- Most multiple alignments based on this approach
- Initial guess for a phylogenetic tree based on pairwise alignments
- Built progressively starting with most closely related sequences
- Follows branching order in phylogenetic tree
- Sufficiently fast
- Sensitive
- Algorithmically heuristic, no mathematical property associated with the alignment
- Biologically sound, it is common to derive alignments which are impossible to improve by eye



(adapted from Sonhammer et al. (1997). "Pfam," Proteins 28:405-20)

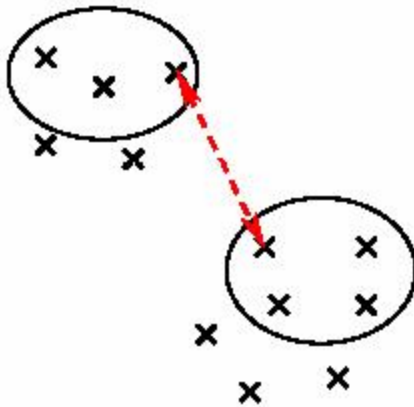
# Clustering approaches for multiple sequence alignment

- Clustal uses average linkage clustering

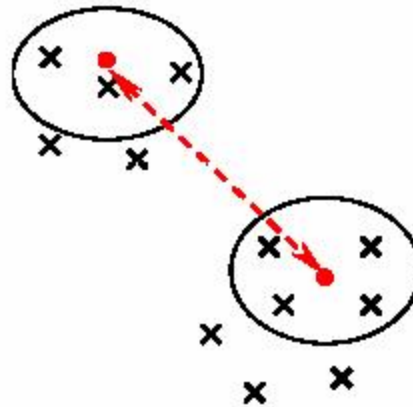
◇ also called UPGMA

Unweighted Pair Group Method with Arithmetic mean

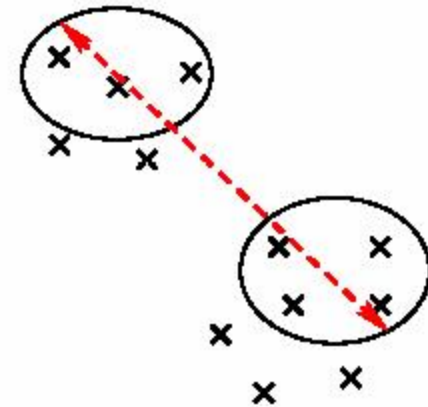
- Simple linkage



- Average linkage



- Complete linkage



<http://compbio.pbworks.com/f/linkages.JPG>

Ca28\_Human

ELSAHATPAFTAVLTSPLPASGMPVKFDRTLYNGHSGYNPATGI FTCPVGGVYYFAYHVV  
VKGTNVWVALYKNNVPATYTYDEYKKG YLDQASGGAVLQLRPNDQVWVQIPSDQANGLYS  
TEYIHSSFSGFLLCPT

C1qb\_Human

DYKATQKIAFSATRTINVPLRRDQTI RFDHVI TNMNNNYEPRSGKFTCKVPGLYYFTYHA  
SSRGNLCVNLMRGRERAQKVVTFC DYAYNTFQVTTGGMVLKLEQGENVFLQATDKNSLLG  
MEGANSIFSGFLLFPD

Cerb\_Human

VRSGSAKVAFSAIRSTNHEPSEMSNR TMI IYFDQVLVNI GNNFDSERSTFIAPRKG IYSF  
NFHVVKVYNRQTIQVSLMLNGWPVISA FAGEDQDVTREAA SNGVLIQMEKGDRA YLKLERG  
NLMGGWKYSTFSGFLLVFPD

COLE\_LEPMA.264

RGPKGPPGESVEQIRSAFSVGLFPSRSFP PPSLPVKFDFYNGEGHWDPTLNKFNVTYP  
GVYLF SYHITVRNRPVRAALV VNGVRKLRTRDSL YGQDIDQASNLALLHLTDGDQVWLET  
LRDWNGXYSSSEDDSTFSGFLLYPDTKKPTAM

HP27\_TAMAS.72

GPPGPPGMTV NCHSKGTSFAFAVKANELPPAPSQPVI FKEALHDAQGHFDLATGVFTCPVP  
GLYQFGFHIEAVQRAVKVSLMRNGTQVMEREAEAQDGYEHISG TAILQLGMEDRVWLENK  
LSQTDLERGTVQAVFSGFLIHEN

HSUPST2\_1.95

GIQGRKGEPEGAYVYRSAFSVGL ETYVTIPNMP IRF'TKIFYNQONHYDGSTGKFHCNIP  
GLYYFAYHITVYMKDVKVS LFKKDKAMLFTYDQYQENNV DQASGSVLLHLEVGDQVWLQV  
YGEGERNGLYADNDNDSTFTGFLLYHDTN

2.HS27109\_1

ENALAPDFSKGSYRYAPMVAFFASHTYGMTIPGP ILENNLDVNYGASYTPRTGKFRI PYL  
GVYVFKYTIESFSAHISGFLVVDGIDKLAF ESENINSEIHCDRVL TGDALLELN YGQEVW  
LRLAKGTIPAKFPPVTTFSGYLLYRT

4.YQCC\_BACSU

VVHGWPWQKISGFAHANIGTTGVQYLK KIDHTKIAFN RVIKDSHNAFDTKNNRFIAPND  
GMYLIGAS IYTLNYTSYINFHLKVYLNGKAYK TLHHVRGDFQEKDNGMNLGLNGNATVPM  
NKGDYVEIWCYCN YGGDETLKRAVDDKNGVFNFFD

5.BSPBSXSE\_25

ADSGWTAWQKISGFAHANIGTTGRQAL IKGENNKIKYNRI IKDSHKLFDTKNNRFVASHA  
GMHLVSASLYIENTERYSNFELYVYVNGTKYKLMNQFRMPTPSNNSDNEFNATV TGSVTV  
PLDAGDYVEIYVYVGYS GDVTRYVTD SNGALNYFD

# C1Q - Example

MMCOL10A1\_1.483  
Calx\_Chick  
S15435  
CA18\_MOUSE.597  
Ca28\_Human  
MM37222\_1.98  
COLE\_LEPMA.264  
HP27\_TAMAS.72  
S19018  
Clqb\_Mouse  
Clqb\_Human  
Cerb\_Human  
2.HS27109\_1

SGMPLVSAHGVGTG-----MPVSAFTVILS--KAYPA---VGCPHPIYEILYNRQQHY  
-----ALTG-----MPVSAFTVILS--KAYPG---ATVPIKFDKILYNRQQHY  
-----GGPA-----YEMPAFTAELT--APFPP---VGGPVKFNKLLYNGRQNY  
HAYAGKKGKHGGPA-----YEMPAFTAELT--VPFPP---VGAPVKFDKLLYNGRQNY  
-----ELSA-----HATPAFTAFLT--SPLPA---SGMPVKFDRTLYNGHSGY  
----GTPGRKGEPEGE---AAYMYRSAFSVGLETRVTVP----NVPIRFTKIFYNQONHY  
-----RGPKGPPGE---SVEQIRSAFSVGLFSPRSFPP---PSLPVKFDKVFYNGEGHW  
-----GPPGPPGMTVNCHSKGTSFAVKAN--ELPPA---PSQPVIKFEALHDAQGHF  
-----NIRD-----QPRPAFSAIRQ---NPMT---LGNVVIKFDKVLTNQESPY  
-----D---YRATQKVAFSALRTINSPLR----PNQVIRFEKVITNANENY  
-----D---YKATQKIAFSATRTINVPLR----RDQTIREFDHVITNMNNNY  
-----V---RSGSAKVAFSAIRSTNHEPSEMSNRTMIYFDQVLVNIIGNNF  
---ENALAPDFSKGS---YRYAPMVAFASHTYGMTIP-----GPILFNNLDVNYGASY

. \* . : :

MMCOL10A1\_1.483  
Calx\_Chick  
S15435  
CA18\_MOUSE.597  
Ca28\_Human  
MM37222\_1.98  
COLE\_LEPMA.264  
HP27\_TAMAS.72  
S19018  
Clqb\_Mouse  
Clqb\_Human  
Cerb\_Human  
2.HS27109\_1

DPRSGIFTCKIPGIYYFSYHVHVKGT--HVWVGLYKNGTP-TMYTY---DEYSKGYLDTA  
DPRTGIFTCRIPGLYYFSYHVHAKGT--NVWVALYKNGSP-VMYTY---DEYQKGYLDQA  
NPQTGIFTCEVPGVYFYFAYHVHCKGG--NVWVALFKNNEP-VMYTY---DEYKKGFLDQA  
NPQTGIFTCEVPGVYFYFAYHVHCKGG--NVWVALFKNNEP-MMYTY---DEYKKGFLDQA  
NPATGIFTCPVGGVYFYFAYHVHVKGT--NVWVALYKNNVP-ATYTY---DEYKKGYLDTA  
DGSTGKFYCNIPGLYYFSYHITVYMK--DVKVSFLFKDKA-VLFTY---DQYQEKVNDQA  
DPTLNKFNVTYPGVYLFYHITVNRN--PVRAALVVNGVR-KLRTR---DSLQYQDIDQA  
DLATGVFTCPVPGLYQFGFHIEAVQR--AVKVSIMRNGTQ-VMERE---AEAQDG-YEHI  
QNHTGRFICAVPGFYFYFNQVISKWD--LCLFIKSSSGGQ-PRDSLSFSNTNKGFLQVL  
EPRNGKFTCKVPGLYYFTYHASSRGN---LCVNLVGRDRDRSMQKVVTFCDYAQNFTFQVT  
EPRSGKFTCKVPGLYYFTYHASSRGN---LCVNLMRGRER--AQKVVTFCDYAYNTFQVT  
DSERSTFIAPRKGIIYSFNHVVKVYNRQTIQVSLMLNGWP----VISAFAGDQDVTREAA  
TPRTGKFRIPYLVGVYVFKYTIESFSA--HISGFLVVDGIDKLAFESN-INSEIHCDRVL

. \* \* \* \* :

MMCOL10A1\_1.483  
Calx\_Chick  
S15435  
CA18\_MOUSE.597  
Ca28\_Human  
MM37222\_1.98  
COLE\_LEPMA.264  
HP27\_TAMAS.72  
S19018  
Clqb\_Mouse  
Clqb\_Human  
Cerb\_Human  
2.HS27109\_1

SGSAIMELTENDQVWLQLPNA-ESNGLYSSEYVHSSFSGFLVAPM-----  
SGSAVIDLMENDQVWLQLPNS-ESNGLYSSEYVHSSFSGFLFAQI-----  
SGSAVLLLRPGDRVFLQMPSE-QAAGLYAGQYVHSSFSGYLLYPM-----  
SGSAVLLLRPGDQVFLQNPFE-QAAGLYAGQYVHSSFSGYLLYPM-----  
SGGAVLQLRPNDQVWVQIPSD-QANGLYSTEYIHSSFSGFLLCPT-----  
SGSVLLHLEVGDQVWLQVYGDGDHNGLYADNVNDSTFTGFLLYHDTN-----  
SNLALLHLTDGDQVWLETLR--DWNGXYSSSEDDSTFSGFLLYPDTKKPTAM  
SGTAILQLGMEDRVWLENKL--SQTDLERG-TVQAVFSGFLIHEN-----  
AGGTVLQLRRGDEVWIEKDP--AKGRIYQGTEADSIKSGFLIFPS-----  
TGGVVLKLEQEEVHLQATD---KNSLLGIEGANSIFTGFLLFDP-----  
TGGMVLKLEQGENVFLQATD---KNSLLGMEGANSIFSGFLLFDP-----  
SNGVLIQMEKGDRAYLKLER---GN-LMGG-WKYSTFSGFLVFP-----  
TGDALLELNYQEVWLRRLAK---GTIPAKFPPVTTTFSGYLLYRT-----

. :: : : . : \* \* : \*

# Clustal Alignment

# Problems with Progressive Alignments

- Local Minimum Problem
  - Parameter Choice Problem

## 1. Local Minimum Problem

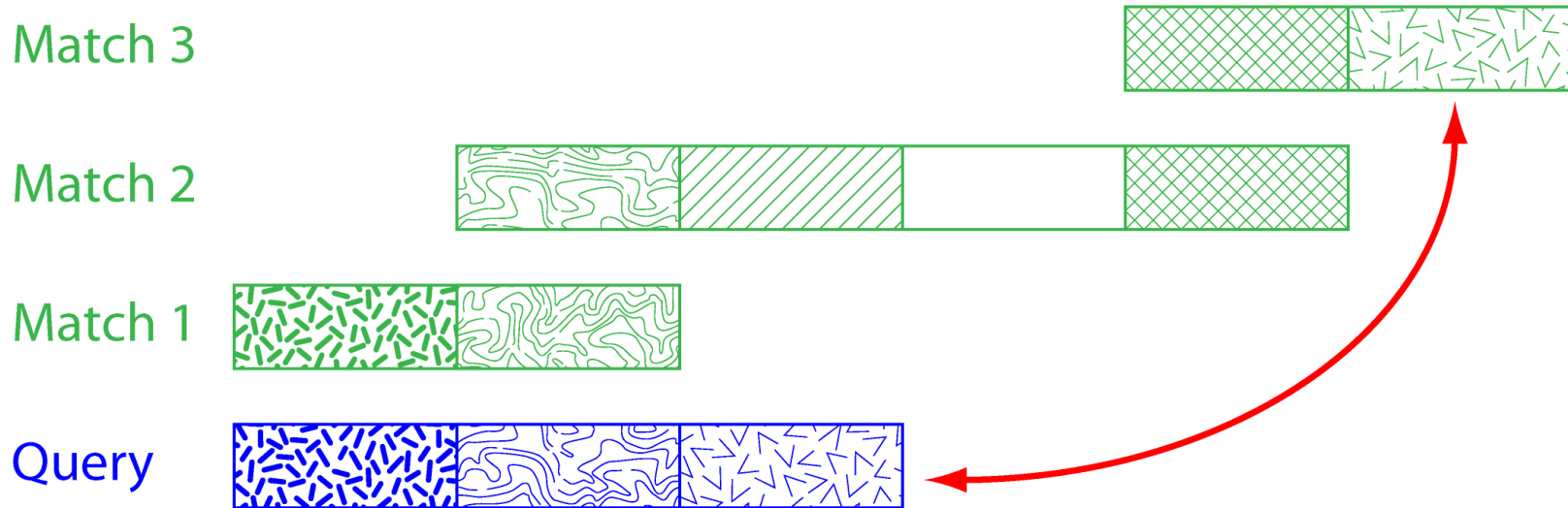
- It stems from greedy nature of alignment (mistakes made early in alignment cannot be corrected later)
- A better tree gives a better alignment (UPGMA neighbour-joining tree method)

## 2. Parameter Choice Problem

- - It stems from using just one set of parameters (and hoping that they will do for all)



# Domain Problem in Multiple Alignment



Fuse multiple alignment into:

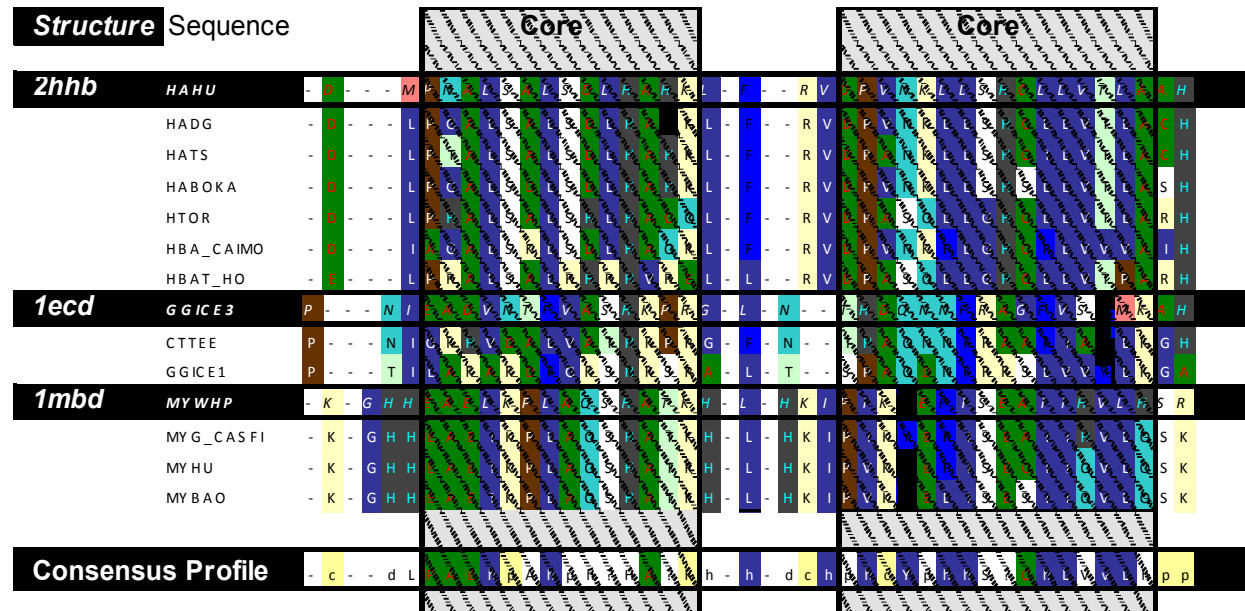
- **Motif**: a short signature pattern identified in the conserved region of the multiple alignment
- **Profile**: frequency of each amino acid at each position is estimated
- **HMM**: Hidden Markov Model, a generalized profile in rigorous mathematical terms

# Profiles

## Motifs

## HMMs

Can get more sensitive searches with these multiple alignment representations (Run the profile against the DB.)



# Multiple Alignment

## MOTIFS

## 2 different applications for motif analysis

- Given a collection of binding sites (or protein sequences with binding motifs), develop a representation of those sites that can be used to search new sites and reliably predict where additional binding sites occur.
- Given a set of sequences known to contain binding sites for a common factor, but not knowing where the sites are, discover the location of the sites in each sequence and a representation of the protein.

# Motifs

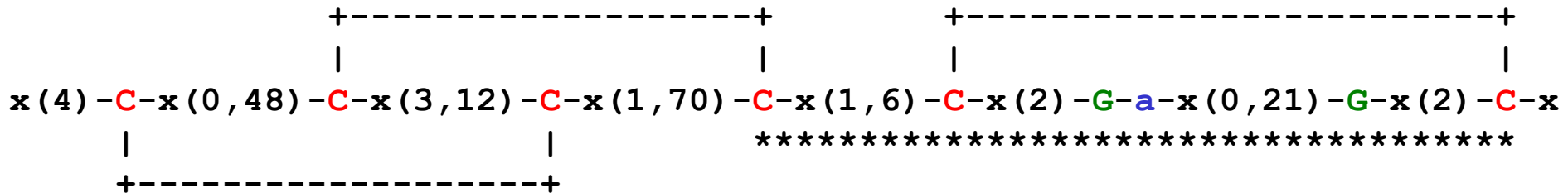
- several proteins are grouped together by similarity searches
- they share a conserved motif
- motif is stringent enough to retrieve the family members from the complete protein database
- PROSITE: a collection of motifs (1135 different motifs)

MMCOL10A1_1.483	SGSA	IME	L	TEND	QVWL	QLPNA	-ESNG	LYSSEYVHSSFS	SGFL	VAPM-----
Ca1x_Chick	SGSA	VID	L	MEND	QVWL	QLPNS	-ESNG	LYSSEYVHSSFS	SGFL	FAQI-----
S15435	SGSA	VLL	L	RPGD	RVFL	QMPSE	-QAAG	LYAGQYVHSSFS	SGYL	LYPM-----
CA18_MOUSE.597	SGSA	VLL	L	RPGD	QVFL	QNPFE	-QAAG	LYAGQYVHSSFS	SGYL	LYPM-----
Ca28_Human	SGGA	VLQ	L	RPND	QVWV	QIPSD	-QANG	LYSTEYIHSSFS	SGFL	LCPT-----
MM37222_1.98	SGSV	LLH	L	LEVGD	QVWL	QVYGD	GDHNG	LYADNVNDSTFT	TGFL	LYHDTN-----
COLE_LEPMA.264	SNLA	LLH	L	TDGD	QVWL	LETLR	--DWN	GXYSSSEDDSTFS	SGFL	LYPDTKKPTAM
HP27_TAMAS.72	SGTA	ILQ	L	GMED	RVWL	ENKL	--SQ	TLERG-TVQAVFS	SGFL	LIHEN-----
S19018	AGGT	VLQ	L	RRGDE	VWIE	EKDP	--AK	GRIYQGTEADSI	IFS	SGFLIFPS-----
C1qb_Mouse	TGGV	VLK	L	EQEE	VVHL	QATD	---K	NSLLGIEGANSIFT	TGFL	LFPD-----
C1qb_Human	TGGM	VLK	L	EQGEN	VFLQ	ATD	---K	NSLLGMEGANSIFS	SGFL	LFPD-----
Cerb_Human	SNGV	LIQ	M	EKGD	RAYL	KLER	---GN	-LMGG-WKYSTFS	SGFL	VFPL-----
2.HS27109_1	TGDAL	LE	L	NYGQ	EVWL	RRLAK	----G	TIPAKFPPVTTF	S	GYLLYRT-----
		::	:	:	:			*	*	:*

# Prosites Pattern -- EGF like pattern

A sequence of about thirty to forty amino-acid residues long found in the sequence of epidermal growth factor (EGF) has been shown [1 to 6] to be present, in a more or less conserved form, in a large number of other, mostly animal proteins. The proteins currently known to contain one or more copies of an EGF-like pattern are listed below.

- Bone morphogenic protein 1 (BMP-1), a protein which induces cartilage and bone formation.
- Caenorhabditis elegans developmental proteins lin-12 (13 copies) and glp-1 (10 copies).
- Calcium-dependent serine proteinase (CASP) which degrades the extracellular matrix proteins type ...
- Cell surface antigen 114/A10 (3 copies).
- Cell surface glycoprotein complex transmembrane subunit .
- Coagulation associated proteins C, Z (2 copies) and S (4 copies).
- Coagulation factors VII, IX, X and XII (2 copies).
- Complement C1r/C1s components (1 copy).
- Complement-activating component of Ra-reactive factor (RARF) (1 copy).
- Complement components C6, C7, C8 alpha and beta chains, and C9 (1 copy).
- Epidermal growth factor precursor (7-9 copies).



- 'C': conserved cysteine involved in a disulfide bond.
- 'G': often conserved glycine
- 'a': often conserved aromatic amino acid
- '\*': position of both patterns.
- 'x': any residue

-Consensus pattern: C-x-C-x(5)-G-x(2)-C  
 [The 3 C's are involved in disulfide bonds]

# Multiple Alignment

## PROFILES

# Profiles

<b>2hhb Human Alpha Hemoglobin</b>	R	V	D	C	V	A	Y	K	
HAHU	R	V	D	C	V	A	Y	K	100
HADG	R	V	D	C	V	A	Y	K	89
HTOR	R	V	D	C	A	A	Y	Q	76
HBA_CAIMO	R	V	D	P	V	A	Y	K	73
HBAT_HORSE	R	V	D	P	A	A	Y	Q	62
<b>1mbd Whale Myoglobin</b>	A	I	C	A	P	A	Y	E	
MYWHP	A	I	C	A	P	A	Y	E	100
MYG_CASFI	R	I	C	A	P	A	Y	E	85
MYHU	R	I	C	V	C	A	Y	D	75
MYBAO	R	I	C	V	C	A	Y	D	71
Eisenberg Profile Freq. A	1	0	0	2	2	9	0	0	↑ Identity
Eisenberg Profile Freq. C	0	0	4	3	2	0	0	0	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
Eisenberg Profile Freq. V	0	5	0	2	3	0	0	0	
Eisenberg Profile Freq. Y	0	0	0	0	0	0	9	0	
Consensus = Most Typical A.A.	R	V	D	C	V	A	Y	E	
Better Consensus = Freq. Pattern (PCA)	R	iv	cd	š	š	A	Y	μ	
	š = (A,2V,C,P); μ=(4K,2Q,3E,2D)								
Entropy => Sequence Variability	3	7	7	14	14	0	0	14	

Profile : a position-specific scoring matrix composed of 21 columns and N rows (N=length of sequences in multiple alignment)

**What happens with gaps?**



# EGF Profile Generated for SEARCHWISE

Cons	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Gap
V	-1	-2	-9	-5	-13	-18	-2	-5	-2	-7	-4	-3	-5	-1	-3	0	0	-1	-24	-10	100
D	0	-14	-1	-1	-16	-10	0	-12	0	-13	-8	1	-3	0	-2	0	0	-8	-26	-9	100
V	0	-13	-9	-7	-15	-10	-6	-5	-5	-7	-5	-6	-4	-4	-6	-1	0	-1	-27	-14	100
D	0	-20	18	11	-34	0	4	-26	7	-27	-20	15	0	7	4	6	2	-19	-38	-21	100
P	3	-18	1	3	-26	-9	-5	-14	-1	-14	-12	-1	12	1	-4	2	0	-9	-37	-22	100
C	5	115	-32	-30	-8	-20	-13	-11	-28	-15	-9	-18	-31	-24	-22	1	-5	0	-10	-5	100
A	2	-7	-2	-2	-21	-5	-4	-12	-2	-13	-9	0	-1	0	-3	2	1	-7	-30	-17	100
s	2	-12	3	2	-25	0	0	-18	0	-18	-13	4	3	1	-1	7	4	-12	-30	-16	25
n	-1	-15	4	4	-19	-7	3	-16	2	-16	-10	7	-6	3	0	2	0	-11	-23	-10	25
p	0	-18	-7	-6	-17	-11	0	-17	-5	-15	-14	-5	28	-2	-5	0	-1	-13	-26	-9	25
c	5	115	-32	-30	-8	-20	-13	-11	-28	-15	-9	-18	-31	-24	-22	1	-5	0	-10	-5	25
L	-5	-14	-17	-9	0	-25	-5	4	-5	8	8	-12	-14	-1	-5	-7	-5	2	-15	-5	100
N	-4	-16	12	5	-20	0	24	-24	5	-25	-18	25	-10	6	2	4	1	-19	-26	-2	100
g	1	-16	7	1	-35	29	0	-31	-1	-31	-23	12	-10	0	-1	4	-3	-23	-32	-23	50
G	6	-17	0	-7	-49	59	-13	-41	-10	-41	-32	3	-14	-9	-9	5	-9	-29	-39	-38	100
T	3	-10	0	2	-21	-12	-3	-5	1	-11	-5	1	-4	1	-1	6	11	0	-33	-18	100
C	5	115	-32	-30	-8	-20	-13	-11	-28	-15	-9	-18	-31	-24	-22	1	-5	0	-10	-5	100
I	-6	-13	-19	-11	0	-28	-5	8	-4	6	8	-12	-17	-4	-5	-9	-4	6	-12	-1	100
d	-4	-19	8	6	-15	-13	5	-17	0	-16	-12	5	-9	2	-2	-1	-1	-13	-24	-5	31
i	0	-6	-8	-6	-4	-11	-5	3	-5	1	2	-5	-8	-4	-6	-2	0	4	-14	-6	31
g	1	-13	0	0	-20	-3	-3	-12	-3	-13	-8	0	-7	0	-5	2	0	-7	-29	-16	31
L	-5	-11	-20	-14	0	-23	-9	9	-11	8	7	-14	-17	-9	-14	-8	-4	7	-17	-5	100
E	0	-20	14	10	-33	5	0	-25	2	-26	-19	11	-9	4	0	3	0	-19	-34	-22	100
S	3	-13	4	3	-28	3	0	-18	2	-20	-13	6	-6	3	1	6	3	-12	-32	-20	100
Y	-14	-9	-25	-22	31	-34	10	-5	-17	0	-1	-14	-13	-13	-15	-14	-13	-7	17	44	100
T	0	-10	-6	-1	-11	-16	-2	-7	-1	-9	-5	-3	-9	0	-1	1	3	-4	-16	-8	100
C	5	115	-32	-30	-8	-20	-13	-11	-28	-15	-9	-18	-31	-24	-22	1	-5	0	-10	-5	100
R	0	-13	0	2	-19	-11	1	-12	4	-13	-8	3	-8	4	5	1	1	-8	-23	-13	100
C	5	115	-32	-30	-8	-20	-13	-11	-28	-15	-9	-18	-31	-24	-22	1	-5	0	-10	-5	100
P	0	-14	-8	-4	-15	-17	0	-7	-1	-7	-5	-4	6	0	-2	0	1	-3	-26	-10	100
P	1	-18	-3	0	-24	-13	-3	-12	1	-13	-10	-2	15	2	0	2	1	-8	-33	-19	100
G	4	-19	3	-4	-48	53	-11	-40	-7	-40	-31	5	-13	-7	-7	4	-7	-29	-39	-36	100
Y	-22	-6	-35	-31	55	-43	11	-1	-25	6	4	-21	-34	-20	-21	-22	-20	-7	43	63	50
S	1	-9	-3	-1	-14	-7	0	-10	-2	-12	-7	0	-7	0	-4	4	4	-5	-24	-9	100
G	5	-20	1	-8	-52	66	-14	-45	-11	-44	-35	4	-16	-10	-10	4	-11	-33	-40	-40	100
E	2	-20	10	12	-31	-7	0	-19	6	-20	-15	5	4	7	2	4	2	-13	-38	-22	100
R	-5	-17	0	1	-16	-13	8	-16	9	-16	-11	5	-11	7	15	-1	-1	-13	-18	-6	100
C	5	115	-32	-30	-8	-20	-13	-11	-28	-15	-9	-18	-31	-24	-22	1	-5	0	-10	-5	100
E	0	-26	20	25	-34	-5	6	-25	10	-25	-17	9	-4	16	5	3	0	-18	-38	-23	100
T	-4	-11	-13	-8	-1	-21	2	0	-4	-1	0	-6	-14	-3	-5	-4	0	0	-15	0	100
D	0	-18	5	4	-24	-11	-1	-11	2	-14	-9	1	-6	2	0	0	0	-6	-34	-18	100
I	0	-10	-2	-1	-17	-14	-3	-4	-1	-9	-4	0	-11	0	-4	0	2	-1	-29	-14	100
D	-4	-15	-1	-2	-13	-16	-3	-8	-5	-6	-4	-1	-7	-2	-7	-3	-2	-6	-27	-12	100

Cons.  
Cys

2hhb	Human Alpha Hemoglobin	R	V	D	C	V	A	Y	K	
	HAHU	R	V	D	C	V	A	Y	K	100
	HADG	R	V	D	C	V	A	Y	K	89
	HTOR	R	V	D	C	A	A	Y	Q	76
	HBA_CAIMO	R	V	D	P	V	A	Y	K	73
	HBAT_HORSE	R	V	D	P	A	A	Y	Q	62

1mbd	Whale Myoglobin	A	I	C	A	P	A	Y	E	
	MYWHP	A	I	C	A	P	A	Y	E	100
	MYG_CASFI	R	I	C	A	P	A	Y	E	85
	MYHU	R	I	C	V	C	A	Y	D	75
	MYBAO	R	I	C	V	C	A	Y	D	71

Eisenberg Profile Freq. A	1	0	0	2	2	9	0	0	↑ Identity
Eisenberg Profile Freq. C	0	0	4	3	2	0	0	0	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
Eisenberg Profile Freq. V	0	5	0	2	3	0	0	0	
Eisenberg Profile Freq. Y	0	0	0	0	0	0	9	0	

Consensus = Most Typical A.A. 

R	V	D	C	V	A	Y	E
---	---	---	---	---	---	---	---

Better Consensus = Freq. Pattern (PCA) 

R	iv	cd	š	š	A	Y	μ
---	----	----	---	---	---	---	---

š = (A,2V,C,P); μ=(4K,2Q,3E,2D)

Entropy => Sequence Variability 

3	7	7	14	14	0	0	14
---	---	---	----	----	---	---	----

# Profiles formula for position $M(p,a)$

**$M(p,a)$  = chance of finding amino acid  $a$  at position  $p$**

$M_{\text{simp}}(p,a)$  = number of times  $a$  occurs at  $p$  divided by number of sequences

However, what if don't have many sequences in alignment?  $M_{\text{simp}}(p,a)$  might be biased. Zeros for rare amino acids. Thus:

$$M_{\text{cplx}}(p,a) = \sum_{b=1 \text{ to } 20} M_{\text{simp}}(p,b) \times Y(b,a)$$

$Y(b,a)$ : Dayhoff matrix for  $a$  and  $b$  amino acids

$$S(p,a) \sim \sum_{a=1 \text{ to } 20} M_{\text{simp}}(p,a) \ln M_{\text{simp}}(p,a)$$

2hhb	Human Alpha Hemoglobin	R	V	D	C	V	A	Y	K	
	HAHU	R	V	D	C	V	A	Y	K	100
	HADG	R	V	D	C	V	A	Y	K	89
	HTOR	R	V	D	C	A	A	Y	Q	76
	HBA_CAIMO	R	V	D	P	V	A	Y	K	73
	HBA_T_HORSE	R	V	D	P	A	A	Y	Q	62
1mbd	Whale Myoglobin	A	I	C	A	P	A	Y	E	
	MYWHP	A	I	C	A	P	A	Y	E	100
	MYG_CASFI	R	I	C	A	P	A	Y	E	85
	MYH_U	R	I	C	V	C	A	Y	D	75
	MYBAO	R	I	C	V	C	A	Y	D	71

Eisenberg Profile Freq. A  
Eisenberg Profile Freq. C  
⋮  
Eisenberg Profile Freq. V  
Eisenberg Profile Freq. Y

1	0	0	2	2	9	0	0
0	0	4	3	2	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0	5	0	2	3	0	0	0
0	0	0	0	0	0	9	0

↑  
Identity

Consensus = Most Typical A.A.  
Better Consensus = Freq. Pattern (PCA)  
ξ = (A,2V,C,P); μ=(4K,2Q,3E,2D)

R	V	D	C	V	A	Y	E
R	iv	cd	š	š	A	Y	μ

Entropy => Sequence Variability

3	7	7	14	14	0	0	14
---	---	---	----	----	---	---	----

# Profiles formula for entropy H(p,a)

$$H(p,a) = - \sum_{a=1 \text{ to } 20} f(p,a) \log_2 f(p,a),$$

where  $f(p,a)$  = frequency of amino acid  $a$  occurs at position  $p$  ( $M_{\text{simp}}(p,a)$ )

Say column only has one aa (AAAAA):

$$H(p,a) = 1 \log_2 1 + 0 \log_2 0 + 0 \log_2 0 + \dots = 0 + 0 + 0 + \dots = 0$$

Say column is random with all aa equiprobable (ACD..ACD..ACD..):

$$H_{\text{rand}}(p,a) = .05 \log_2 .05 + .05 \log_2 .05 + \dots = -.22 + -.22 + \dots = -4.3$$

Say column is random with aa occurring according to probability found in the sequence databases (ACAAAADAADDDDDAAAA....):

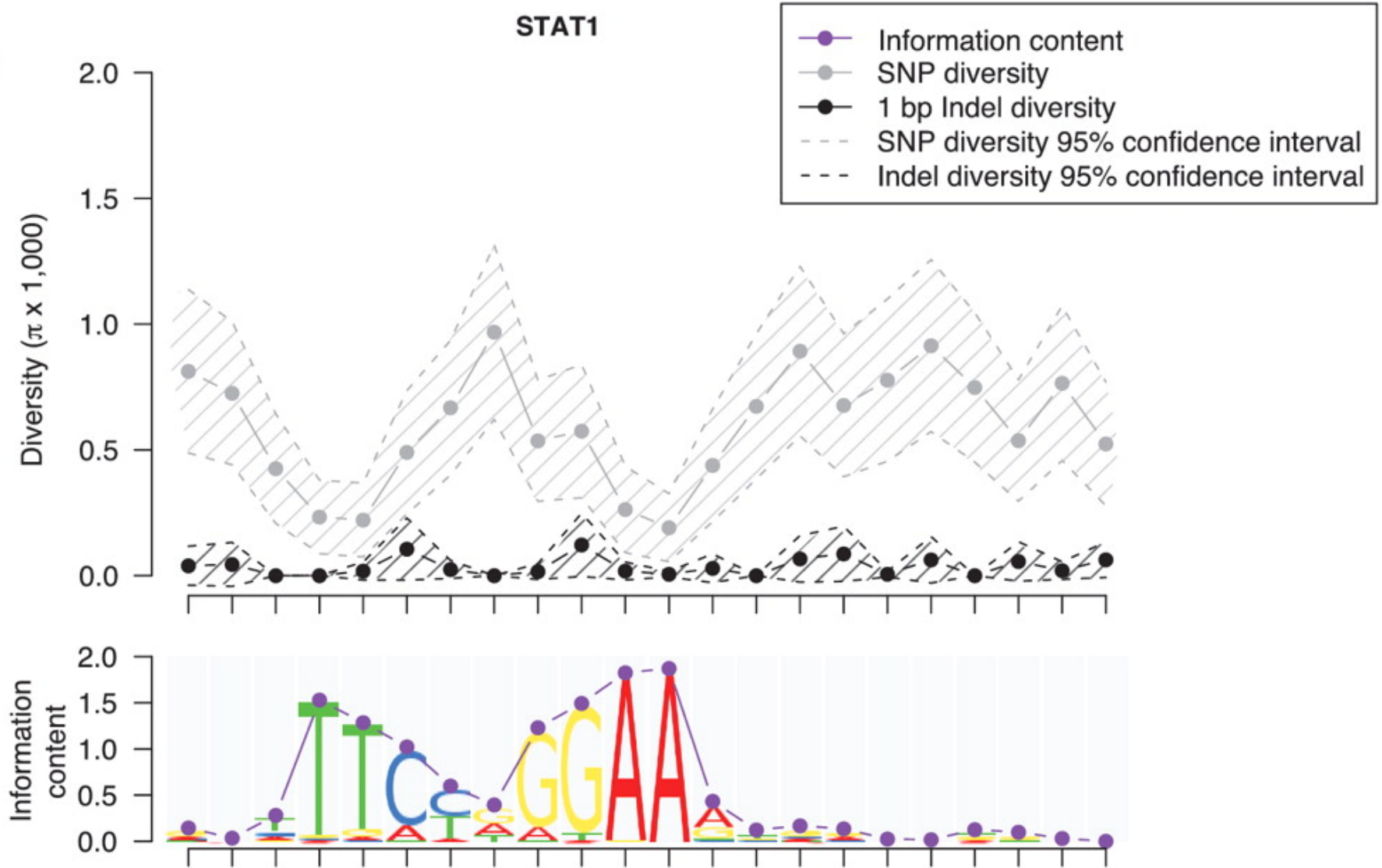
$$H_{\text{db}}(a) = - \sum_{a=1 \text{ to } 20} F(a) \log_2 F(a),$$

where  $F(a)$  is freq. of occurrence of  $a$  in DB

$$H_{\text{corrected}}(p,a) = H(p,a) - H_{\text{db}}(a)$$

# (A) Aggregation of nucleotide diversity across STAT1 motifs.

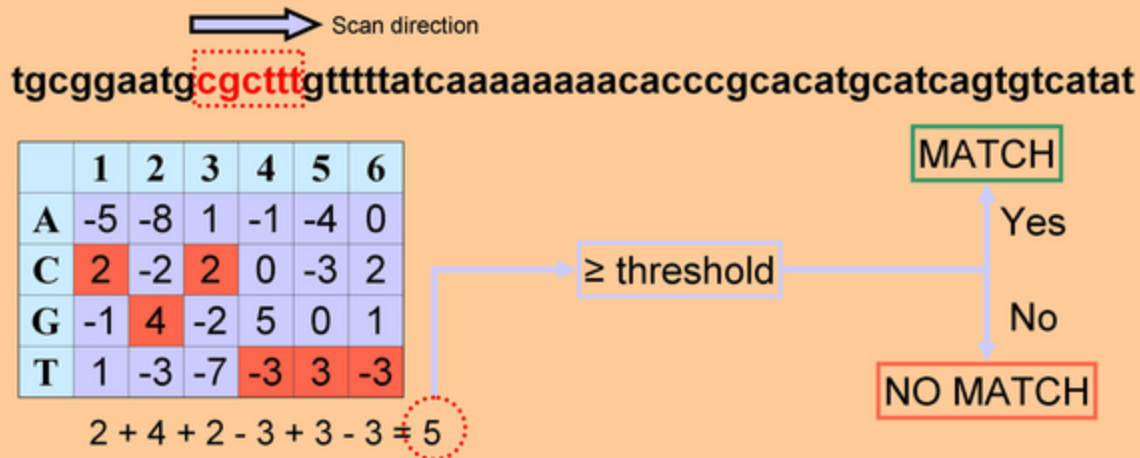
A



Mu X J et al. Nucl. Acids Res. 2011;39:7058-7076

# Scanning for Motifs with PWMs

Position Weight Matrices define an additive scheme for scoring sequence. Often, the weights are simply log likelihood ratios of observing a nucleotide in a binding site relative to genomic background. Sequences are scanned by scoring every site, on both the forward and reverse complement strands, and identifying matches as shown in the schematic below:



A particular site is evaluated by adding up the entries from the scoring matrix at each position, and comparing the sum to a match threshold. For log ratio PWMs, an empirically chosen threshold of 60% of the maximum positive score has been used by Harbison et al. and is approximately equal to cutoffs determined by the principled cross-validated method presented in Maclsaac et al. More sophisticated algorithms developed specifically for motif scanning are described briefly in Figure 3.

# Ψ-Blast

Parameters: overall threshold, inclusion threshold, iterations

- Automatically builds profile and then searches with this
- Also PHI-blast

© 1997 Oxford University Press

Nucleic Acids Research, 1997, Vol. 25, No. 17 3389-3402

## Gapped BLAST and PSI-BLAST: a new generation of protein database search programs








Stephen F. Altschul\*, Thomas L. Madden, Alejandro A. Schaffer<sup>1</sup>, Jinghui Zhang, Zheng Zhang<sup>2</sup>, Webb Miller<sup>2</sup> and David J. Lipman

National Center for Biotechnology Information, Bethesda, MD 20894, USA, <sup>1</sup>Laboratory of Molecular Biology, National Institutes of Health, Bethesda, MD 20894, USA, <sup>2</sup>Department of Engineering, Pennsylvania State University, University Park, PA 16802, USA

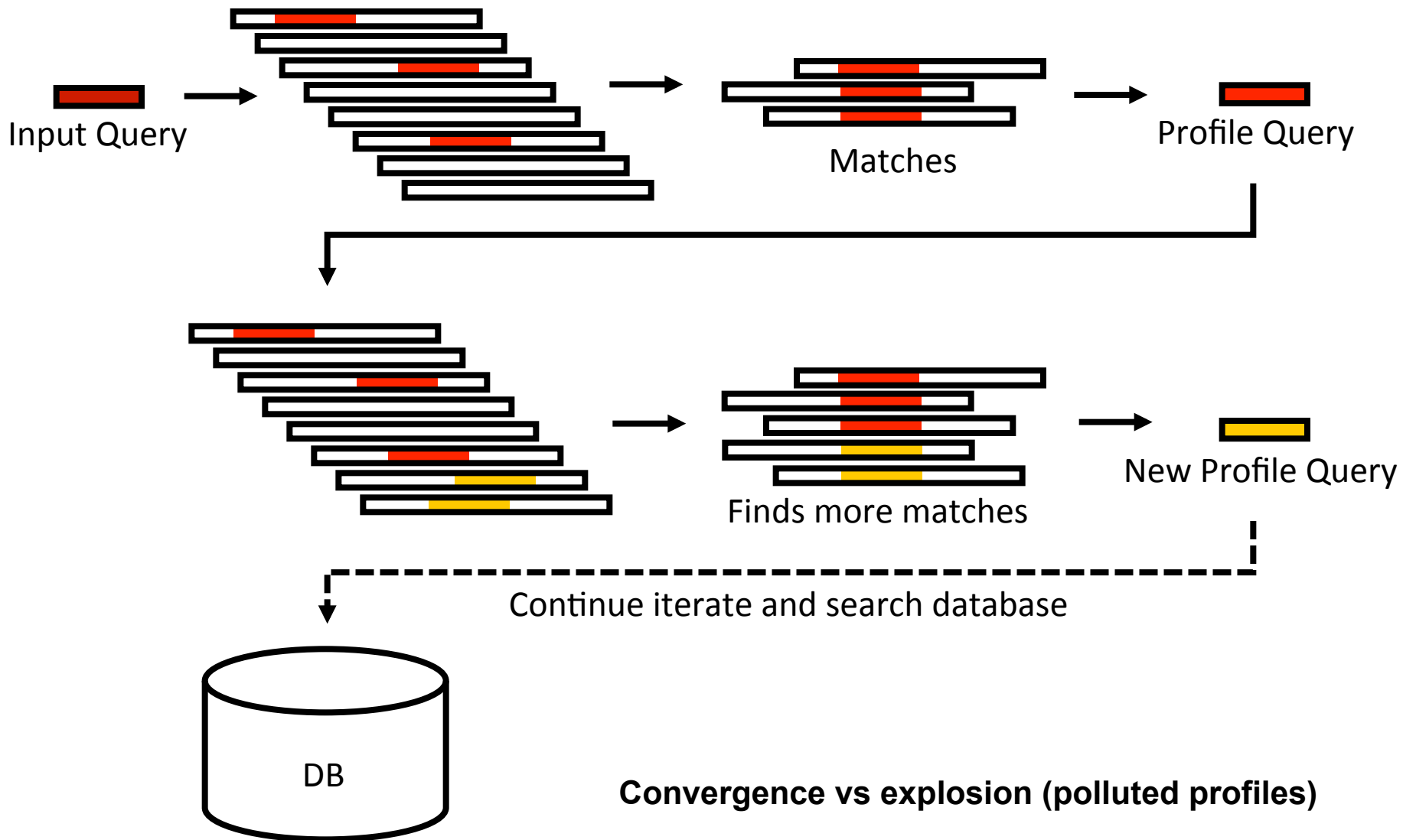
Received June 20, 1997; Revised and Accepted August 1, 1997

### ABSTRACT

The BLAST programs are widely used to search protein and DNA databases for sequence similarities. For protein comparisons, we have developed a new algorithm, Gapped BLAST, which uses a heuristic search of a database for a high-scoring pair of sequences. This method is computationally efficient and statistically rigorous. In addition, we have developed a new algorithm, PSI-BLAST, which uses an iterative search of a database for a high-scoring pair of sequences. This method is computationally efficient and statistically rigorous. In addition, we have developed a new algorithm, PSI-BLAST, which uses an iterative search of a database for a high-scoring pair of sequences. This method is computationally efficient and statistically rigorous.

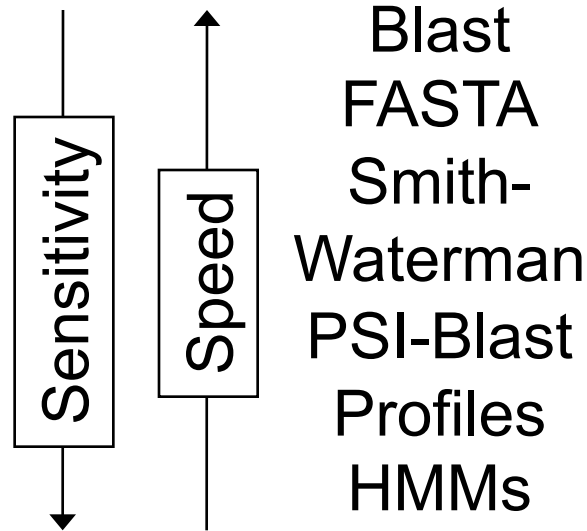
<u>Accession</u>	<u>Alignment</u>	<u>E-value</u>
P49789		
P49779		8e-27
P49775		6e-18
Q11066		3e-07
Q09344		4e-05
P49378		0.001
P32084		0.002

# PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool)



# Speed v Sensitivity Tradeoff

PSI Blast as a form of Semi-supervised learning





# Multiple Alignment: Probabilistic Approaches for Determining PWMs

- Expectation Maximization: Search the PWM space randomly
- Gibbs sampling: Search sequence space randomly.

# Expectation-Maximization (EM) algorithm

- Used in statistics for finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved latent variables.
- EM alternates between performing
  - an expectation (E) step, which computes an expectation of the likelihood by including the latent variables as if they were observed, and
  - a maximization (M) step, which computes the maximum likelihood estimates of the parameters by maximizing the expected likelihood found on the E step.
- The parameters found on the M step are then used to begin another E step, and the process is repeated.

# Alternating approach

1. Guess an initial weight matrix
2. Use weight matrix to predict instances in the input sequences
3. Use instances to predict a weight matrix
4. Repeat 2 [E-step] & 3 [M-step] until satisfied.

Examples: Gibbs sampler (Lawrence et al.)

MEME (expectation max. / Bailey, Elkan)

ANN-Spec (neural net / Workman, Stormo)

Another good source is Wes Craven's 776 course: <https://www.biostat.wisc.edu/~craven/776/lecture9.pdf>

[Adapted from B Noble, GS 541 at UW, <http://noble.gs.washington.edu/~wnoble/genome541/>]

# EM (again!)

```
EM  foreach subsequence of width W
    convert subsequence to a matrix
    do {
        re-estimate motif occurrences from matrix
        re-estimate matrix model from motif occurrences
    } until (matrix model stops changing)
    end
    select matrix with highest score
```

# Sample DNA sequences

>ce1cg

```
TAATGTTTGTGCTGGTTTTTGTGGCATCGGGCGAGAATA  
GCGCGTGGTGTGAAAGACTGTTTTTTTGATCGTTTTTCAC  
AAAAATGGAAGTCCACAGTCTTGACAG
```

>ara

```
GACAAAACGCGTAACAAAAGTGTCTATAATCACGGCAG  
AAAAGTCCACATTGATTATTTGCACGGCGTCACACTTTG  
CTATGCCATAGCATTTTTTATCCATAAG
```

>bglr1

```
ACAAATCCCAATAACTTAATTATTGGGATTTGTTATATA  
TAACTTTATAAATTCCTAAAATTACACAAAGTTAATAAC  
TGTGAGCATGGTCATATTTTTTATCAAT
```

>crp

```
CACAAAGCGAAAGCTATGCTAAAACAGTCAGGATGCTAC  
AGTAATACATTGATGTACTGCATGTATGCAAAGGACGTC  
ACATTACCGTGCAGTACAGTTGATAGC
```

# Motif occurrences

>celcg

taatgtttgtgctgggtttttgtggcatcgggcgagaata  
gcgcggtgggtgtgaaagactgtttt**TTTGATCGTTTTCAC**  
aaaatggaagtccacagtcttgacag

>ara

gacaaaaacgcgtaacaaaagtgtctataatcacggcag  
aaaagtccacattgatta**TTTGCACGGCGTCAC**actttg  
ctatgccatagcatttttatccataag

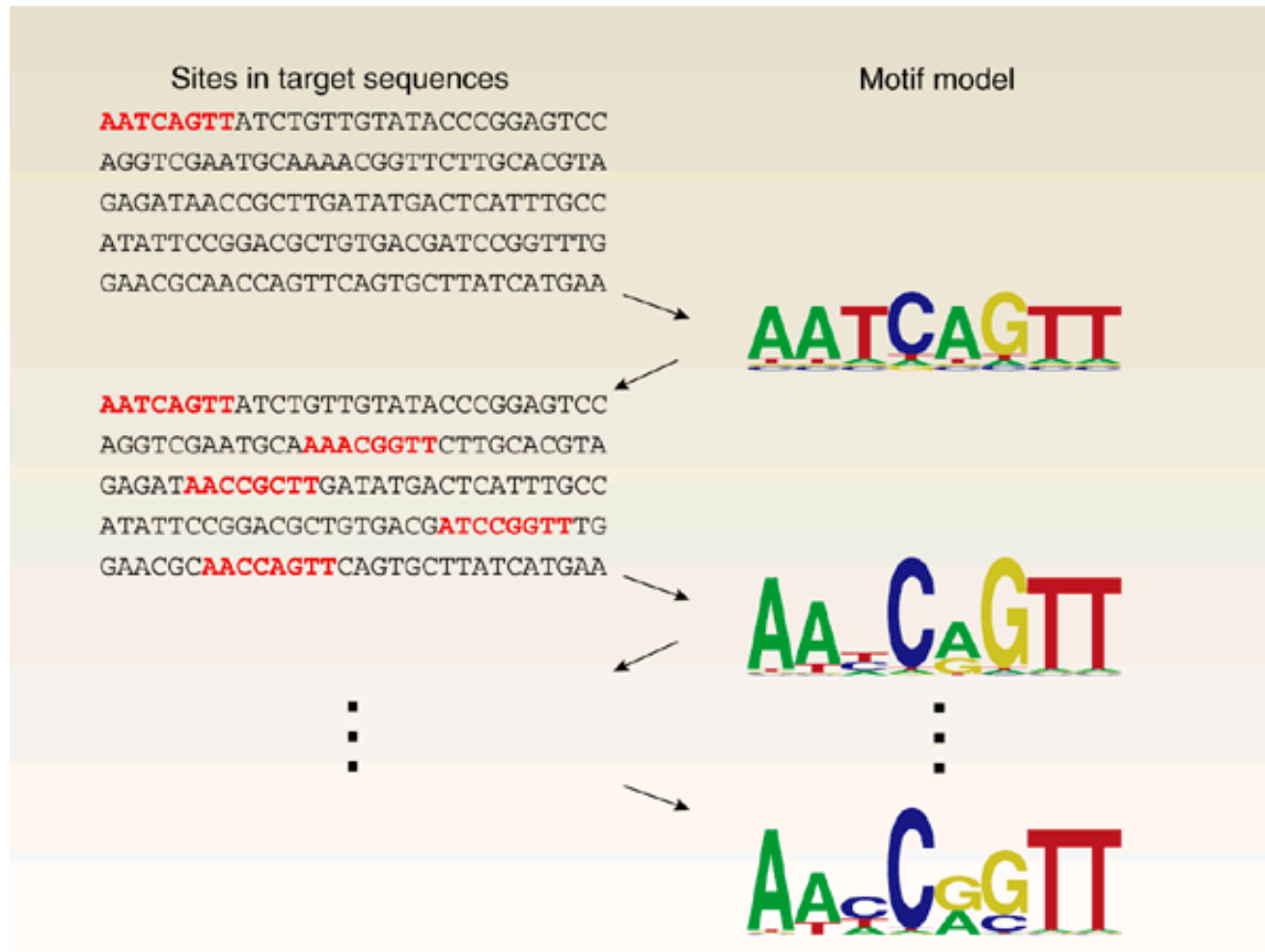
>bglr1

acaaatcccaataacttaattattgggatttgttatata  
taactttataaattcctaaaattacacaaagttaataac  
**TGTGAGCATGGTCAT**atTTTTatcaat

>crp

cacaaagcgaaagctatgctaaaacagtcaggatgctac  
agtaatacattgatgtactgcatgta**TGCAAAGGACGTC**  
**AC**attaccgtgcagtacagttgatagc

# How does EM algorithm work?



Starting from a single site, expectation maximization algorithms such as MEME alternate between assigning sites to a motif (left) and updating the motif model (right).

Note that only the best hit per sequence is shown here, although lesser hits in the same sequence can have an effect as well.

Specifically, in E step, estimate location of motif match. In M step, find most likely parameters of motif model given the locations.

# MEME - a practical program using EM

- Subsequences which occur in the input DNA sequence are used as the starting points from which EM converges iteratively to locally optimal motifs. This increases the likelihood of finding globally optimal motifs.
- Multiple occurrences of a motif are allowed. Algorithm is allowed to ignore sequences with no appearance of a shared motif. So, more resistance to noisy data.
- Motifs are probabilistically erased after they are found, so more than one motif can be found.

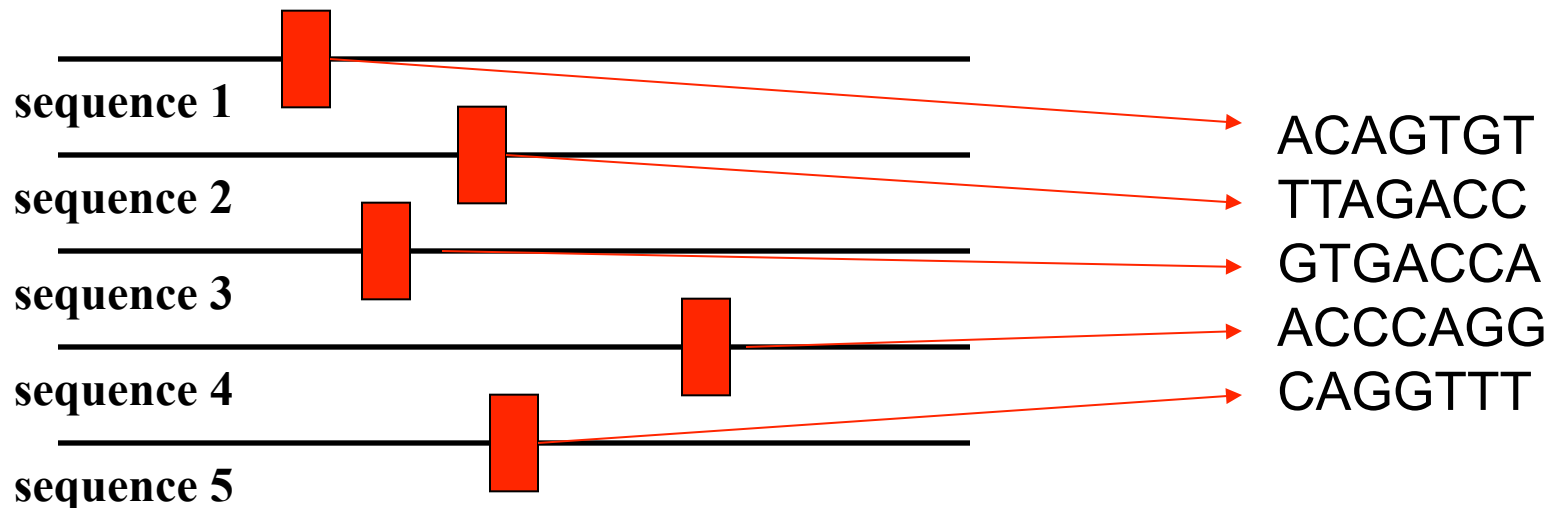


# Multiple Alignment

## Gibbs Sampling

# Initialization

- Step 1: Randomly guess an instance  $s_i$  from each of  $t$  input sequences  $\{S_1, \dots, S_t\}$ .



# Gibbs sampler

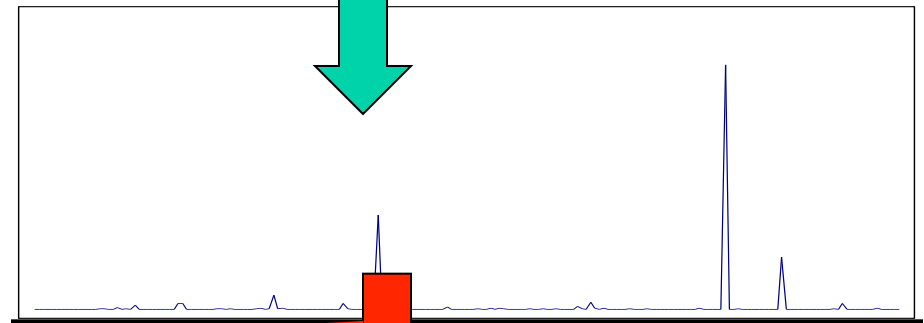
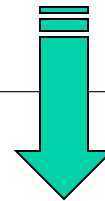
- Steps 2 & 3 (search):
  - Throw away an instance  $s_i$ : remaining  $(t - 1)$  instances define weight matrix.
  - Weight matrix defines instance probability at each position of input string  $S_i$
  - Pick new  $s_i$  according to probability distribution (not necessarily always the  $s_i$  giving the highest prob.)
- Return highest-scoring motif seen

# Sampler step illustration:

ACAGTGT  
TAGGCGT  
ACACCGT  
??????  
CAGGTTT



A	.45	.45	.45	.05	.05	.05	.05
C	.25	.45	.05	.25	.45	.05	.05
G	.05	.05	.45	.65	.05	.65	.05
T	.25	.05	.05	.05	.45	.25	.85



sequence 4

11%

ACAGTGT  
TAGGCGT  
ACACCGT  
**ACGCCGT**  
CAGGTTT



ACGCCGT:20%

ACGGCGT:52%

# Comparison

- Both EM and Gibbs sampling involve iterating over two steps
- Convergence:
  - EM converges when the PSSM stops changing.
  - Gibbs sampling runs until you ask it to stop.
- Solution:
  - EM may not find the motif with the highest score.
  - Gibbs sampling will provably find the motif with the highest score, if you let it run long enough.

# Multiple Alignment

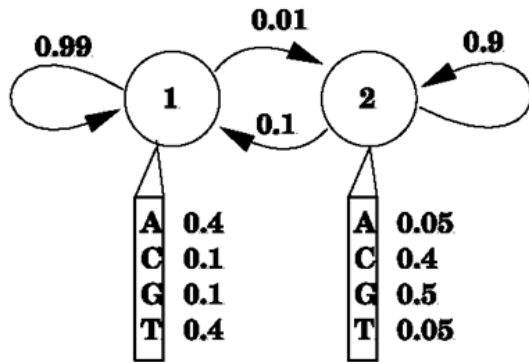
## HMMs

# Hidden Markov Model:

- a composition of finite number of states,
- each corresponding to a column in a multiple alignment
- each state emits symbols, according to symbol-emission probabilities

# HMMs

Starting from an initial state, a sequence of symbols is generated by moving from state to state until an end state is reached.



**state sequence (hidden):**

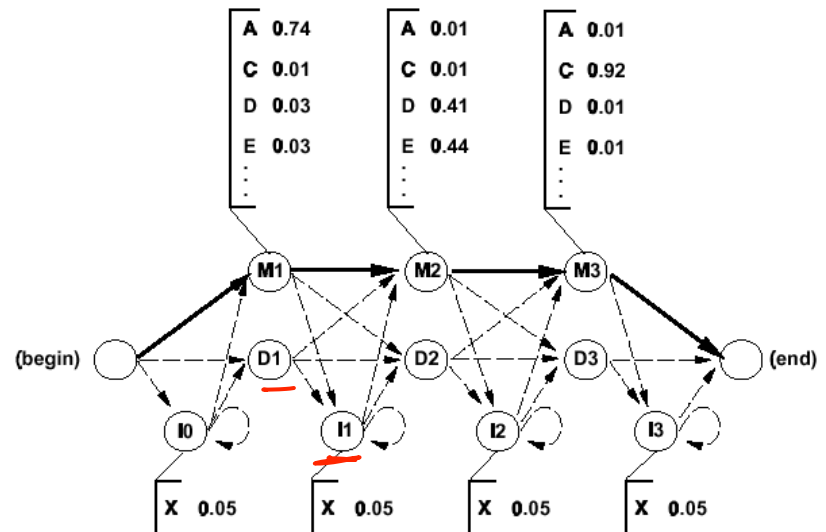
... (1) (1) (1) (1) (1) (2) (2) (2) (2) (1) (1) ...

transitions: ? 0.99 0.99 0.99 0.99 0.01 0.9 0.9 0.9 0.1 0.99

**symbol sequence (observable):**

... A T C A A G G C G A T ...

emissions: 0.4 0.4 0.1 0.4 0.4 0.5 0.5 0.4 0.5 0.4 0.4



(Figures from Eddy, Curr. Opin. Struct. Biol.)

# Profile HMMs

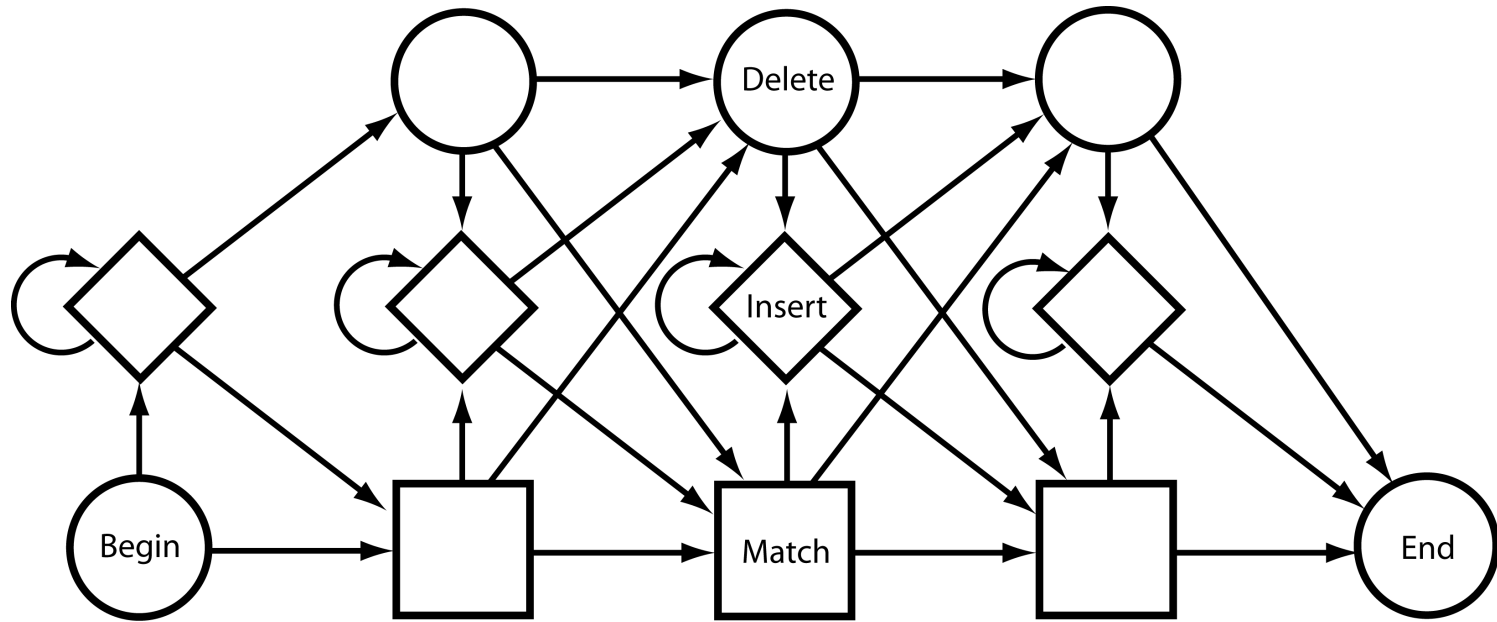


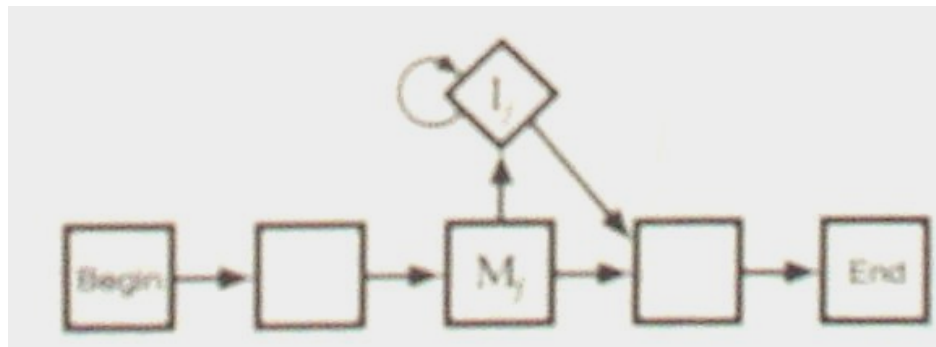
Figure 5.2, Durbin, Eddy, Krogh, and Mitchison, Biological Sequence Analysis (1998)



# Sequence profile elements

- Insertions:

C	A	-	T	G
C	A	T	T	G



# Algorithms

**Probability** of a path through the model

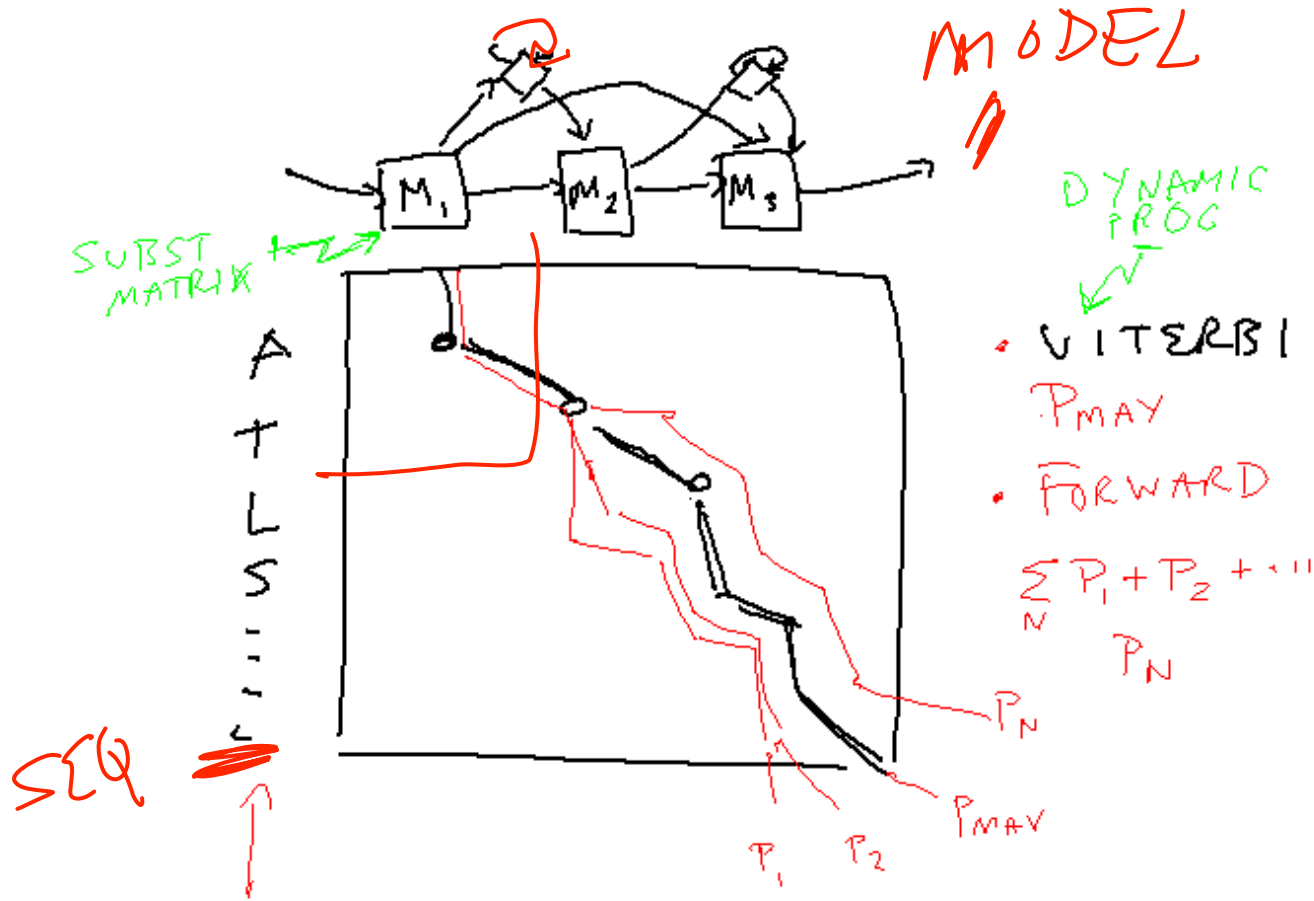
**Viterbi maximizes for seq**

**Forward sums of all possible paths**

**Forward Algorithm** – finds probability  $P$  that a model  $\lambda$  emits a given sequence  $O$  by summing over all paths that emit the sequence the probability of that path

**Viterbi Algorithm** – finds the most probable path through the model for a given sequence  
(both usually just boil down to simple applications of dynamic programming)

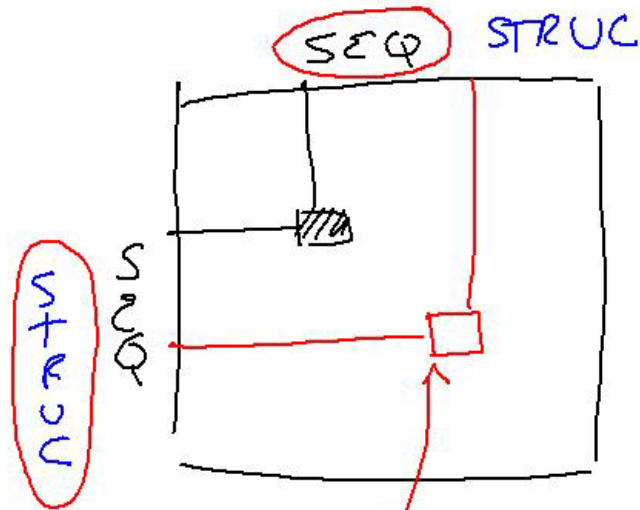
# HMM algorithms are similar to those in sequence alignment



Forward

$$P = \sum_{i=1}^N P_i + P_2 + \dots + P_N$$

# Sequence Alignment, Structure Alignment, Threading



▨ = SEQ IDENTITY  
FOR  
SEQ ALIGNMENT

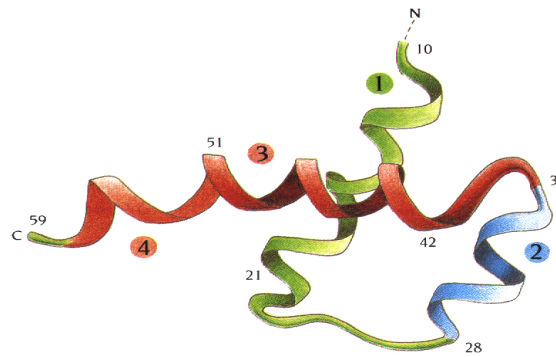
□ = STRUC COORD  
SIM.  
FOR  
STRUC ALIGNMENT

□ = MATCH OF  
SEQ TO  
3D STRUC  
FOR  
THREADING

DEGREE THAT  
RES  $i$  IN TOP  
SEQ MATCHES  
STRUC ENVIRON. OF  
 $j$  IN LEFT STRUC

# Domains

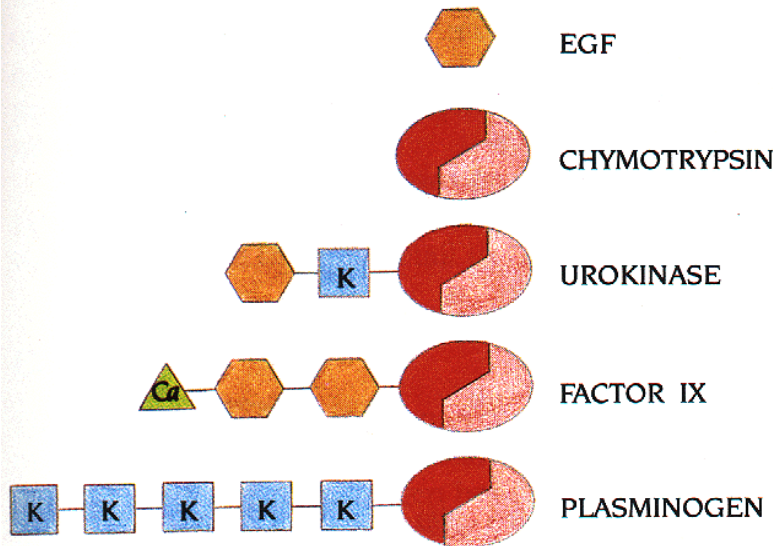
HMMs, Profiles,  
Motifs, and Multiple  
Alignments used to  
define domains







•Another example of the helix-loop-helix motif is seen within several DNA binding domains including the homeobox proteins which are the master regulators of development

(Figures from Branden & Tooze)

**Figure 2.19** Organization of polypeptide chains into domains. Small protein molecules like the epidermal growth factor, EGF, comprise only one domain. Others like the serine proteinase chymotrypsin are arranged in two domains that are both required to form a functional unit (Chapter 15). Many of the proteins that are involved in blood coagulation and fibrinolysis, such as urokinase, factor IX, and plasminogen have long polypeptide chains that comprise different combinations of domains homologous to EGF and serine proteinases and, in addition, calcium-binding domains and Kringle domains.



-  Domains that are homologous to the epidermal growth factor, EGF, which is a small polypeptide chain of 53 amino acids;
-  Serine proteinase domains that are homologous to chymotrypsin, which has about 245 amino acids arranged in two domains;
-  Kringle domains that have a characteristic pattern of three internal disulphide bridges within a region of about 85 amino acid residues;
-  Calcium-binding domain (see Figure 2.13).

- Several motifs ( $\beta$ -sheet, beta-alpha-beta, helix-loop-helix) combine to form a compact globular structure termed a domain or tertiary structure
- A domain is defined as a polypeptide chain or part of a chain that can independently fold into a stable tertiary structure
- Domains are also units of function (DNA binding domain, antigen binding domain, ATPase domain, etc.)

# Applications of HMMs: Protein Domain Databases

Pfam <http://pfam.sanger.ac.uk/>

SMART <http://smart.embl-heidelberg.de/>

CDD

Interpro

etc.