

91學年度上學期「生物資訊學」課程

---

# Pattern Search

張傳雄

國立陽明大學 遺傳學研究所

10-21-2002



# Searching the databases

---

## ▶ (Text) String Searching

- e.g., ATGC, “DEAD”
- Entrez (<http://www.ncbi.nlm.nih.gov/Entrez/>)
- OMIM (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM>)
- LocusLink (<http://www.ncbi.nlm.nih.gov/LocusLink/>)

## ▶ Sequence Similarity Searching

- FastA (Fast alignment)
- BLAST (Basic Local alignment Search Tool)

## ▶ Structure Similarity Searching

- VAST (protein structures)  
(<http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml>)

## ☛ ▶ Pattern Searching

- Aval restriction enzyme cutting site: “C(T,C)CG(A,G)”
- leucine zipper: “LX{6}LX{6}LX{6}L”

---

# Application example of pattern search (I)

Location of the transcription factor binding  
site in the upstream sequence

DHFR-undefined-site-1

CTF/CBP-hs |

hsp70.5 |

**CCAAT\_site\_4** |

NFI.2 |

GR-intron-site-3      **CAAT\_site(1)** |

GR-intron-site-4      |      IE1.2      |      |

AGCTGCAAGGGACACAGGAAAGGGCTGATTG**CCAAT**CCTTTTCAGACATCGCAAAACTTCC

301 -----+-----+-----+-----+-----+-----+-----+ 360

TCGACGTTCCCTGTGTCCTTTCCCGACT**TAAC**GGTTAGGAAAGTCTGTAGCGTTTTGAAGG

GR-intron-site-2

TATA-box.2 |

his3-Tr-TATA |

Ad2MLP\_US.3 |

(TFIID/TBF)-RS |

GAL1-TATA |

**TATA-box-CS** |

GR-MT-IIA      | |

D3      c-mos\_DS1      |      TATA      |

|      |      |      |      |

CGATGCATGTGCGATAATGGTTTGTCCCTAGAGCT**TATA**TAAACAGGCACACATGGCGGCTA

361 -----+-----+-----+-----+-----+-----+-----+ 420

GCTACGTACACGCTATTACCAAACAGGATCTCGATATATTTGTCCGTGTGTACCGCCGAT

PEA3\_RS

Ets-1\_CS |

TCF-2-alpha\_CS |

CP2-gamma-FBG      |      |      PTF1-beta-consensus      |

CAGTGGCTTCTACAAGTTTCAGAGGAAGCCGAGGGCAGCTTAGTTACTGAAGGAGAGATGG

421 -----\*-----+-----+-----+-----+-----+-----+-----+ 480

GTCACCGAAGATGTTCAAGTCTCCTTCGGCTCCCGTCGAATCAATGACTTCCTCTCTACC

---

# Application example of pattern search (II)

The discovery of zinc finger

# Repeats in TFIIIA

---

“..., contains an unusually large number of **Cys** and **His** residues. At first sight these residues appeared to us to form roughly periodic groupings. We therefore made a systematic search for repeats in both amino acid sequence and the cDNA, using the diagonal comparison matrix method and the damped Needleman and Wunsch method (see Materials and methods).”

Miller, J., McLachlan, A.D., and Klug, A. (1985) Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. EMBO J. 4, 1609-1614.

# Local sequence alignment

---

1	YICSFADCGAAYNKNWKLQ*AHLC*KH	37
2	TGEK*PFPCKEEGCEKGFITSLHHLT*RHSL*TH	67
3	TGEK*NFTICDSDGCDLRFITKANMK*KHFNRFH	98
4	NIKICVYVCHFENCGKAFKKHNQLK*VHQF*SH	129
5	TQQL*PYECPHEGCDKRFSLPSRLK*RHEK*VH	159
6	AG---*--YPCKKDDSCSFVGKTWTLYLKHVAECH	188
7	QD---*LAVC--DVCNRKFRHKDYLR*DHQK*TH	214
8	EKERTVYLCPRDGCDRSYTTAFNLR*SHIQSFH	246
9	EEQR*PFVCEHAGCGKCFAMKKSLE*RHSV*VH	276
	TGEK*PYVVC..DGCDKRFITKK..LK*RH..* .H	Consensus

# Proposed structure of a zinc finger by MRC group

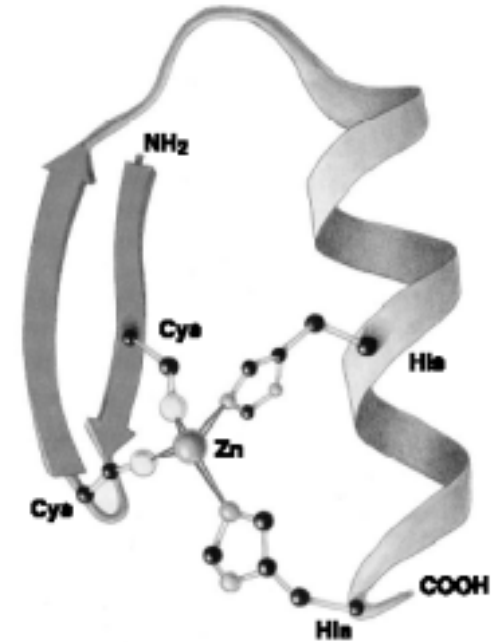
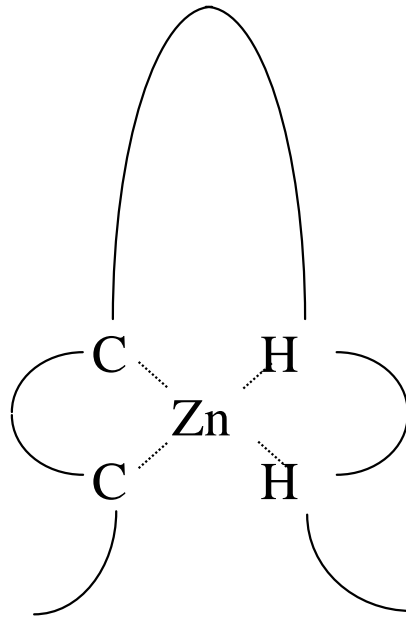
---

TFIIIA

complexed with Zinc

S, N

able to form a coordination  
bond with Zinc



RIKEN Review No. 35 (May, 2001):102

# Any known sequences with this pattern

(C<sub>2</sub>H<sub>2</sub>)?

1	YICSFADCGAAYNKNWKLQ*AHLC*KH	37
2	TGEK*PFPCKEEGCEKGF <sup>T</sup> SLHHL <sup>T</sup> *RHSL* <sup>T</sup> H	67
3	TGEK*NFTCDSDGCDLRF <sup>T</sup> TKANMK*KHFNRFH	98
4	NIKICVYVCHFENC <sup>G</sup> KAFKKHNQLK*VHQF*SH	129
5	TQQL*PYECPHEGCDKRFSLPSRLK*RHEK*VH	159
6	AG--* <sup>-</sup> YPCKKDDSCSFVGKTWTL <sup>Y</sup> LKHVAECH	188
7	QD--*LAVC--DVCNRKFRHKDYLR*DHQK*TH	214
8	EKERTVYLCPRDGC <sup>D</sup> RSYTTAFNLR*SHIQSFH	246
9	EEQR*PFVCEHAGCGKCFAMKKSLE*RHSV*VH	276
	TGEK*PYV <sup>C</sup> ..DGCDKRETKK..LK*RH..* <sup>.</sup> H	Consensus

---

# Berg's zinc-finger pattern

Patterns of metal-binding domains described in Berg, J.M. (1986)  
Potential metal-binding domains in nucleic acid binding proteins.  
Science 232, 485-487

Name	Offset	Pattern ..
TFIIIA	1	CX{2,5}CX{12,12}HX{2,3}H

# Consensus sequence to pattern

---

## MSA of 9 sequences

1	YI	CS	FAD	CGA	AY	NKN	WKL	Q*	AH	LC*	KH	37																					
2	TGEK*	PF	CKE	EG	CE	KG	FT	SL	HHL	T*	RH	SL*	TH	67																			
3	TGEK*	NF	TC	DSD	GCD	LR	FT	TK	AN	MK*	KH	FNR	FH	98																			
4	NIK	IC	VY	V	CH	FEN	CG	KAF	KK	HN	QL	K*	VH	QF*	SH	129																	
5	TQQL*	PY	EC	PHE	GCD	KRF	SL	PS	RL	K*	RHE	K*	VH	159																			
6	AG--*	-Y	P	CK	KDD	SC	SF	V	GKT	WT	LY	L	KH	V	A	E	C	H	188														
7	QD--*	L	A	M	C--	D	V	C	N	R	K	F	R	H	K	D	Y	L	R*	D	H	Q	K*	T	H	214							
8	E	K	E	R	T	V	Y	L	C	P	R	D	G	C	D	R	S	Y	T	T	A	F	N	L	R*	S	H	I	Q	S	F	H	246
9	EE	Q	R*	P	F	V	C	E	H	A	G	C	G	K	C	F	A	M	K	K	S	L	E*	R	H	S	V*	V	H	276			

Consensus  
sequence

TGEK\*PYMC..DGCDKRFTTKK..LK\*RH..\* .H



- Pattern of consensus sequence:  $CX\{2,5\}CX\{12,12\}HX\{2,3\}H$

# Convention to express patterns in GCG environment

---

- Use 1-letter code to express amino acids
- (F,V) means F or V at this location
- X{2,4} means 2 to 4 X (any amino acids)
- {,4} = {0,4}
- {4,} = {4, 350,000}

# Protein pattern search programs

---

- Look for motifs in a sequence
  - GCG: Motifs
  - EMBOSS:
  - others: ScanProsite
- Look for sequences containing a motif
  - GCG: FindPatterns
  - EMBOSS:
  - others: ScanProsite

# Nucleic acid pattern search programs

---

- Look for motifs in a sequence
  - GCG: Map
  - EMBOSS:
  - others: SignalScan
- Look for sequences containing a motif
  - GCG: FindPatterns
  - EMBOSS:
  - others:

# Two subpatterns in zinc-finger motif

---

C-X<sub>2</sub>-C

[(F,Y)-X-C-X-X-{X-X}-C-X-X-X-F] -[X<sub>4</sub>]-  
Cys-Cys loop Tip

H-X<sub>3</sub>-H

[X-L-X-X-H-X-X-X-H]-[X<sub>5</sub>]  
His-His loop Linker

Berg, J.M. (1988) Proposed structure for the zinc-binding domains from transcription factor IIIA and related proteins. Proc. Natl. Acad. Sci. 85, 99-102.

# Structure of C-X<sub>2</sub>-C

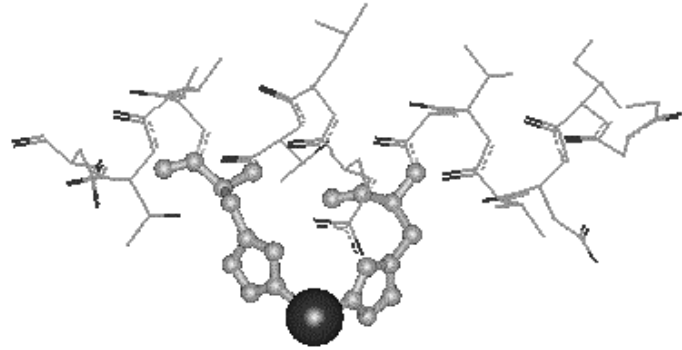
---

- Rubredoxin has two "C-X<sub>2</sub>-C" sequences
- Regulatory subunit of trans-carbamoylase has one "C-X<sub>2</sub>-C" and one "C-X<sub>4</sub>-C" sequences
- These two structures have similar conformations - lie in a loop at the base of an antiparallel  $\beta$ -sheet.



---

## Structure of H-X<sub>3</sub>-H



- Thermolysin (Zn), hemerythrin (Fe), and hemocyanin (Cu) all have a "H-X<sub>3</sub>-H" sequence
- These three structures all lie in an  $\alpha$ -helix.

---

# Structure Prediction of Zinc Finger Motifs by Homology

Rationale:

Sequence implies structure implies function.

- Murray-Rust, 1994

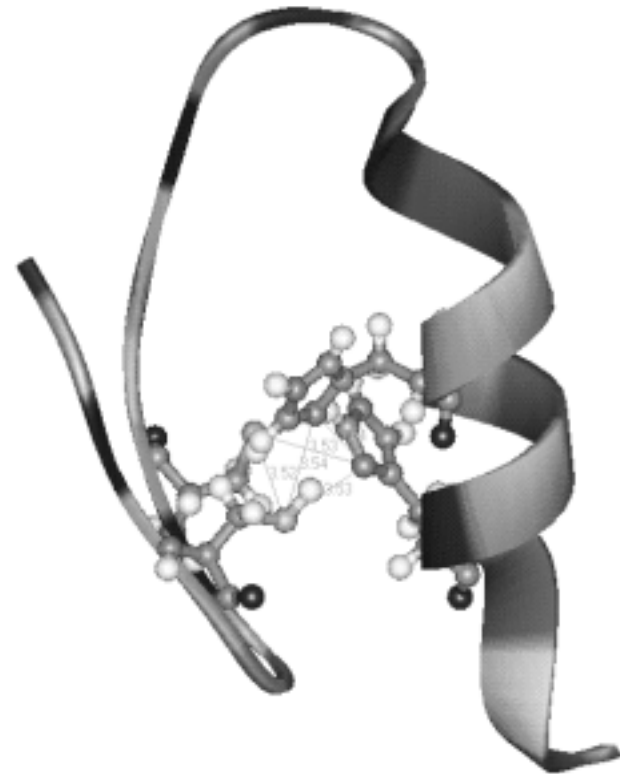
# Model building

C-X<sub>2</sub>-C

antiparallel  $\beta$ -sheet

H-X<sub>3</sub>-H

$\alpha$ -helix

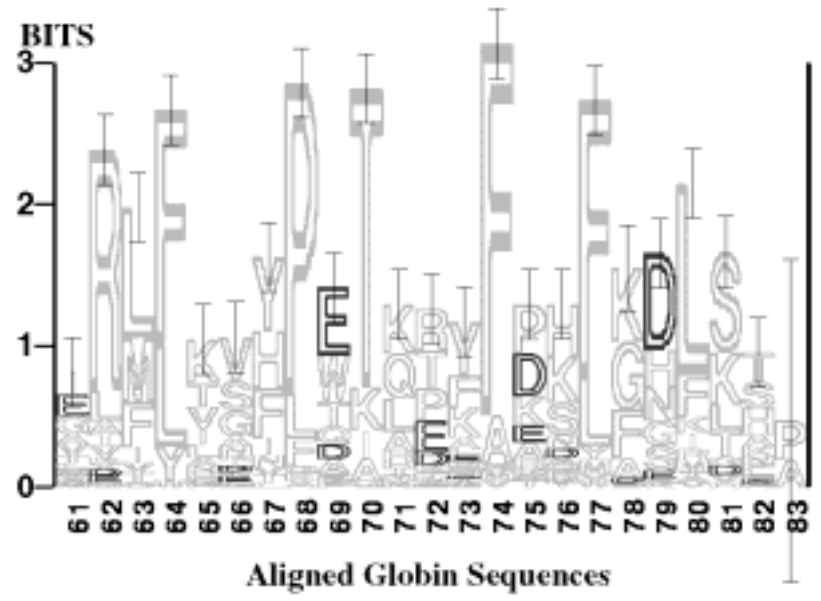
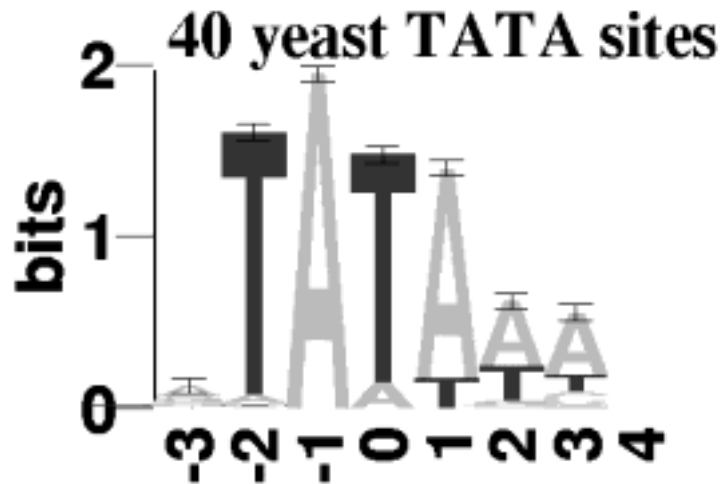


# Summary of the zinc finger studies

---

- Locate repeats: Dot Matrix Analysis
- Identify pattern: Multiple Sequence Alignment
- Find known sequences with similar patterns:  
search for a pattern against a sequence databank
- Predict structure by homology: search pdb
- A successful work based on knowledge of  
protein properties: model building

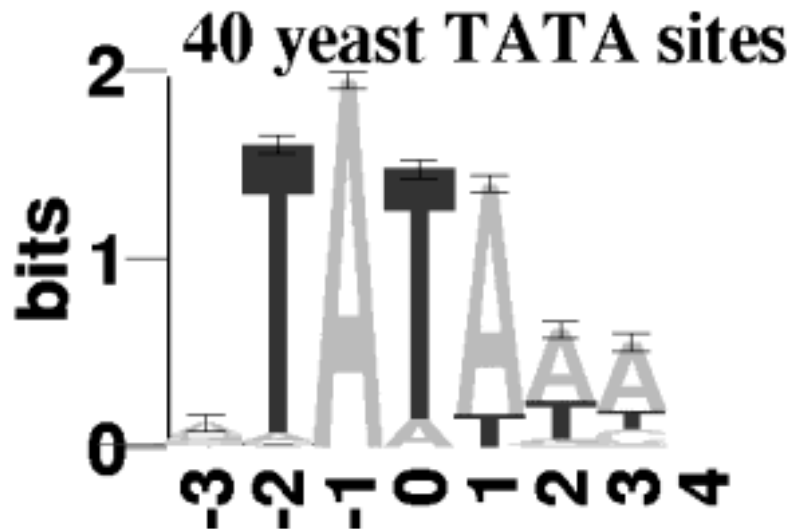
# Sequence logos



<http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html>

# The problem of using a pattern (consensus sequence)

---



TATAAA is too strict;

(T,A)A(T,A)(A,T)(A,T)(A,T)  
is too loose, the following seq  
will show up:

TATTTT

AATTTT

TAAAAA

AAAAAA

# Sequence diversity

---

- Examples
    - antibody-antigen interaction
    - transcription factor binding
    - ... *etc.*
- ⇒ Pattern may be too loose or too strict.
- The disadvantage of using discrete patterns
    - limited descriptive power  
(no weight can be attributed to alternatives)

# Position-specific scoring matrices

---

- An alternative method of sequence comparisons than alignments (NW, SW, BLAST, FastA).
- Position-specific algorithms place importance on regions that are conserved and de-emphasize regions that are not conserved.

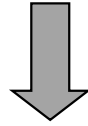
## Tools

- Profile analysis: a kind of PSSM (Position Specific Scoring Matrix)
- Iterative PSSM searches using Blast (PSI-Blast)
- HMMER

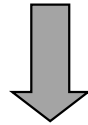
# Concept of profile

---

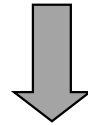
Common ancestor (Homology)



Structure conservation



Position dependent sequence  
conservation



Position specific scoring matrix (PSSM)

- A profile conserves all of the information in the alignment, whereas a consensus sequence removes this information.

# PSSM

---

- PSSM
  - Fixed length position specific scoring matrices.
  - Each position of the motif has different scoring matrix! depending on the type of base aligned at that position.

e.g. PSI-Blast creates PSSM starting from BLOSUM62. And modifying it to reflect the base composition of the alignments.

ProDom uses PSI-Blast to generate motifs and PSSM.

# Profiles

---

- Profiles are generated from multiple sequence alignments. The information in the alignment is represented quantitatively as a table of position-specific values and gap penalties. This table is called a profile.
- A profile is a table where we find for each amino acid position the frequency of each of the 20 amino acids (Profile = position-specific scoring table)
  - i.e., a position-dependent scoring matrix that has N rows and 20+ columns. N is the length of the profile.
- The first 20 columns of each row specify the probability for finding, at that position in the target sequence, each of the 20 amino acid residues.
- The >20 column(s) contain(s) a penalty (penalties) for insertions/deletions at that position.



# Profile features

---

- Use a weight matrix to capture position specific information about conserved/variable residues
- Position specific gap penalties
- Use dynamic programming for matching to allow insertions and deletions
- Computationally less expensive than multiple alignment, yet sensitive
- Use a local similarity approach to deal with modular nature of proteins
- Can use sub-optimal or non-intersecting alignment to find repeated features

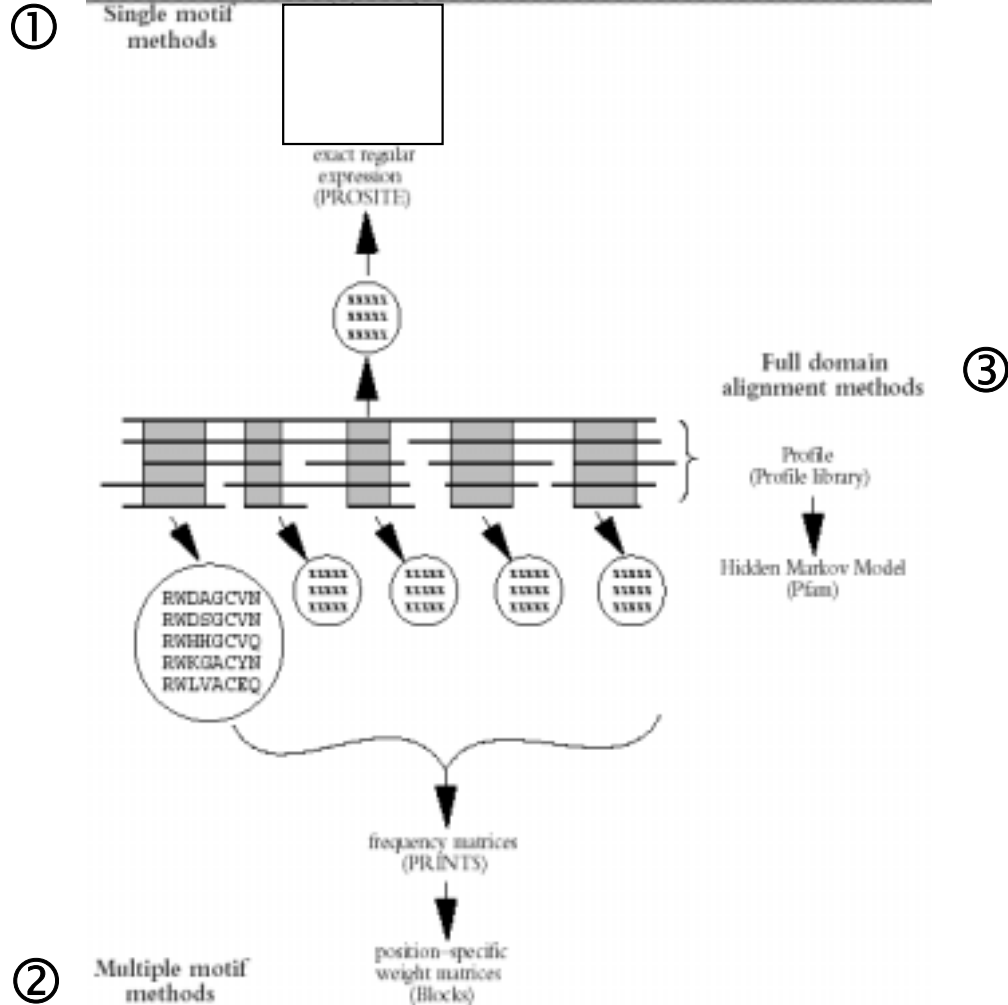
# Various types of profiles

---

- Alignment profile
  - scores for matches, substitutions, insertions
  - used for PROSITE profiles
- PSSM (Position-specific scoring matrix)
  - used in the BLOCKS database
- Hidden Markov model
  - Bayesian statistical approach
  - used for Pfam

# Methods for building pattern databases

(BRIEFINGS IN BIOINFORMATICS. VOL 1. NO 1. 45-59. FEBRUARY 2000)



# In GCG

## How to construct a profile?

---

Pre-align the domain or motif  
(*e.g.* Blast, Fasta in GCG)

Multiple sequence alignment  
(*e.g.* pileup in GCG)

Create a profile (*e.g.*  
profilemake in GCG)

# Compute frequency from counts

---

Position \ Bases	-2	-1	0	1	2	3
A	9	214	63	142	118	8
C	22	7	26	31	52	13
G	18	2	29	38	29	5
T	193	19	124	31	43	216
Subtotal	242	242	242	242	242	242



Position \ Bases	-2	-1	0	1	2	3
A	0.04	0.88	0.26	0.59	0.49	0.03
C	0.09	0.03	0.11	0.13	0.21	0.05
G	0.07	0.01	0.12	0.16	0.12	0.02
T	0.80	0.08	0.51	0.13	0.18	0.89
Subtotal	1.00	1.00	1.00	1.00	1.00	1.00

The counts were obtained from the alignment result of 242 sequences. Simply divide the counts of each nucleotide at a given position by total sequence number. You will get the frequency of each nucleotide at each position.

# Compute odds from frequency

---

Position \ Bases	-2	-1	0	1	2	3
A	0.04	0.88	0.26	0.59	0.49	0.03
C	0.09	0.03	0.11	0.13	0.21	0.05
G	0.07	0.01	0.12	0.16	0.12	0.02
T	0.80	0.08	0.51	0.13	0.18	0.89
Subtotal	1.00	1.00	1.00	1.00	1.00	1.00



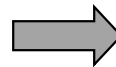
Position \ Bases	-2	-1	0	1	2	3
A	0.15	3.54	1.04	2.35	1.95	0.13
C	0.36	0.12	0.43	0.51	0.86	0.21
G	0.30	0.03	0.48	0.63	0.48	0.08
T	3.19	0.31	2.05	0.51	0.71	3.57

Assume the composition of each nucleotide is 25% here.  
Divide frequency by the composition will give you the odds of finding each nucleotide at a given position.

# Convert to log(odds), so you can use addition instead of multiplication

---

Position \ Bases	-2	-1	0	1	2	3
A	0.15	3.54	1.04	2.35	1.95	0.13
C	0.36	0.12	0.43	0.51	0.86	0.21
G	0.30	0.03	0.48	0.63	0.48	0.08
T	3.19	0.31	2.05	0.51	0.71	3.57



Position \ Bases	-2	-1	0	1	2	3
A	- 2.75	1.82	0.06	1.23	0.96	- 2.92
C	- 1.46	- 3.11	- 1.22	- 0.96	- 0.22	- 2.22
G	- 1.75	- 4.92	- 1.06	- 0.67	- 1.06	- 3.60
T	1.67	- 1.67	1.04	- 0.96	- 0.49	1.84

Think about how do we use PSSM in previous slides  
- you add the score up rather than multiplying them up.

# Profile (weight matrix): make model

242 species sequences

MSA

Consensus sequence  
(length=6bp)

statistics

Position \ Bases	-2	-1	0	1	2	3
A	0.04	0.88	0.26	0.59	0.49	0.03
C	0.09	0.03	0.11	0.13	0.21	0.05
G	0.07	0.01	0.12	0.16	0.12	0.02
T	0.80	0.08	0.51	0.13	0.18	0.89
Subtotal	1.00	1.00	1.00	1.00	1.00	1.00

normalization

Position \ Bases	-2	-1	0	1	2	3
A	9	214	63	142	118	8
C	22	7	26	31	52	13
G	18	2	29	38	29	5
T	193	19	124	31	43	216
Subtotal	242	242	242	242	242	242

Make odds

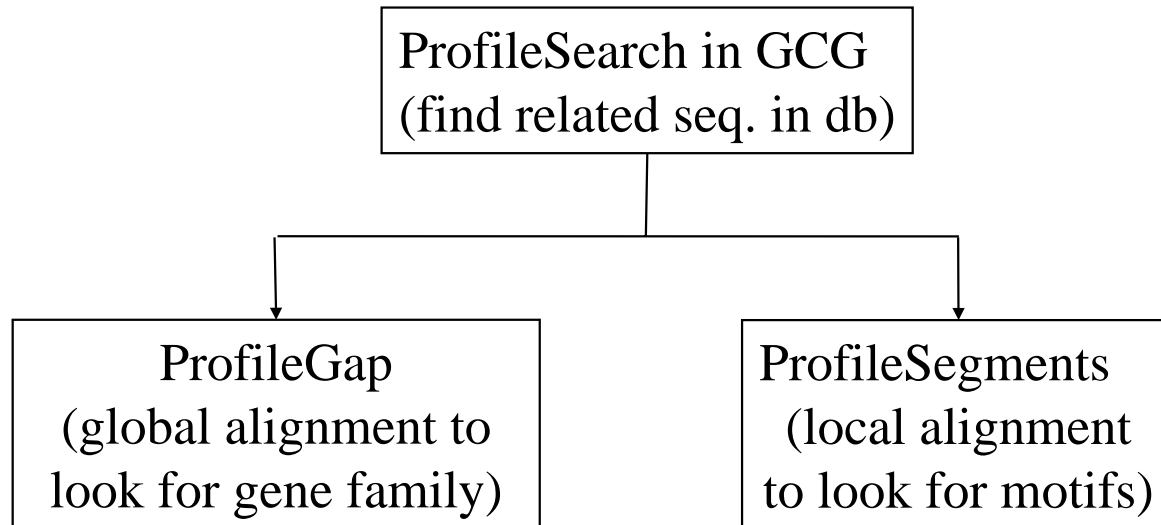
Position \ Bases	-2	-1	0	1	2	3
A	0.15	3.54	1.04	2.35	1.95	0.13
C	0.36	0.12	0.43	0.51	0.86	0.21
G	0.30	0.03	0.48	0.63	0.48	0.08
T	3.19	0.31	2.05	0.51	0.71	3.57

Log(odds)

Position \ Bases	-2	-1	0	1	2	3
A	-2.75	1.82	0.06	1.23	0.96	-2.92
C	1.46	-3.11	1.22	0.96	0.22	2.22
G	1.75	-4.92	1.06	0.67	1.06	3.60
T	1.67	-1.67	1.04	-0.96	-0.49	1.84

# How to use a profile?

---



# Align a sequence with PSSM

---

	C	T	A	T	A	A	T	C	Total Score
A	-38	19	1	12	10	-48			
C	-15	-38	-8	-10	-3	-32			
G	-13	-48	-6	-7	-10	-48			
T	17	-32	8	-9	-6	19			
Score	-15	-32	1	-9	10	-48			-93



	C	T	A	T	A	A	T	C	Total Score
A		-38	19	1	12	10	-48		
C		-15	-38	-8	-10	-3	-32		
G		-13	-48	-6	-7	-10	-48		
T		17	-32	8	-9	-6	19		
Score		17	19	8	12	10	19		85

Shift PSSM along the seq and sum up the score

The best location will have the  
highest score

	C	T	A	T	A	A	T	C	Total Score
A		-38	19	1	12	10	-48		
C		-15	-38	-8	-10	-3	-32		
G		-13	-48	-6	-7	-10	-48		
T		17	-32	8	-9	-6	19		
Score		17	19	8	12	10	19		85



	C	T	A	T	A	A	T	C	Total Score
A		-38	19	1	12	10	-48		
C		-15	-38	-8	-10	-3	-32		
G		-13	-48	-6	-7	-10	-48		
T		17	-38	8	-9	-6	19		
Score		-38	-38	1	12	-6	-32		-101

# Profile (weight matrix): use model

New sequence

	C	T	A	T	A	A	T	C	Total Score
A		-38	19	1	12	10	-48		
C		-15	-38	-8	-10	-3	-32		
G		-13	-48	-6	-7	-10	-48		
T		17	-32	8	-9	-6	19		
Score		17	19	8	12	10	19		85

	C	T	A	T	A	A	T	C	Total Score
A		-38	19	1	12	10	-48		
C		-15	-38	-8	-10	-3	-32		
G		-13	-48	-6	-7	-10	-48		
T		17	-32	8	-9	-6	19		
Score		-15	-32	1	-9	10	-48		-93

	C	T	A	T	A	A	T	C	Total Score
A			-38	19	1	12	10	-48	
C			-15	-38	-8	-10	-3	-32	
G			-13	-48	-6	-7	-10	-48	
T			17	-38	8	-9	-6	19	
Score			-38	-38	1	12	-6	-32	-101

# Protein Blast

---

- BLASTP
- PSI-BLAST (position specific iterated)  
& PHI-BLAST (pattern hit initiated)  
for searching short nearly exact matches
- RPS (reverse position specific)-BLAST  
for conserved domain search

---

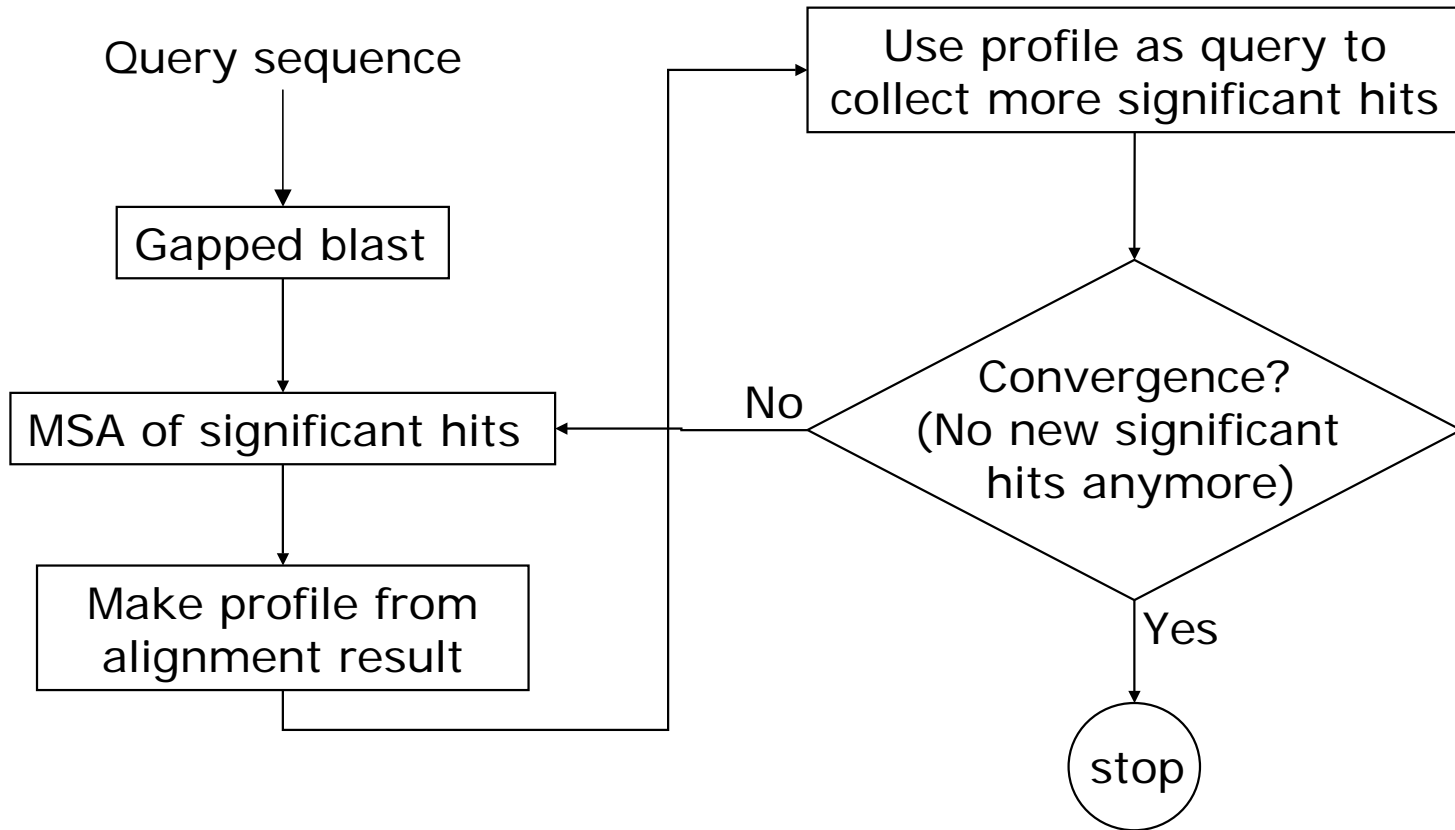
# Iterative database search using PSSM

PSI- BLAST (position specific iterated)

& PHI-BLAST (pattern hit initiated)

# Steps in PSI-Blast

---



# PSI-Blast

---

Scoring matrix: almost always BLOSUM 62

gap scores: Gap opening 11, gap extension 2. They are position independent. This is a major limitation of PSI-BLAST and it is unlikely we can do much about it since the  $E$  value calculation (and  $K$ ) is dependent on the gap scores.

E-value cutoff: a variable, commonly .0005. One must realize that the E-value calculation does not alter the ranking of hits to a particular query. Instead, it forces hits of different queries to be judged by the same cutoff.

Can one improve the E-value calculation procedure?

Can one use different cutoffs for different query sequences?

# Performance of PSI-Blast compared to other pairwise alignment algorithms

---

The number of SWISS-PROT sequences yielding alignments with  $E$ -value  $\leq 0.01$

Protein family	Query	Smith–Waterman	Original BLAST	Gapped BLAST	PSI-BLAST
Serine protease	P00762	275	273	275	286
Serine protease inhibitor	P01008	108	105	108	111
Ras	P01111	255	249	252	375
Globin	P02232	28	26	28	623
Hemagglutinin	P03435	128	114	128	130
Interferon $\alpha$	P05013	53	53	53	53
Alcohol dehydrogenase	P07327	138	128	137	160
Histocompatibility antigen	P10318	262	241	261	338
Cytochrome P450	P10635	211	197	211	224
Glutathione transferase	P14942	83	79	81	142
H <sup>+</sup> -transporting ATP synthase	P20705	198	191	197	207
Normalized running time		36	1.0	0.34	0.87

Normalized running times are the mean ratio of program running time to that for the original BLAST. The time for PSI-BLAST includes the time for the initial BLAST search.

*Nucleic Acids Research*, 1997, 25(17): 3389–3402

# Applications of PSI-Blast

---

A more sensitive search method  
Discover novel motifs

What to do when your search returns homologs with  
weak similarities ?

- Run PSI-BLAST
- Look for sequence motifs (may help identify distant members of protein families)

---

# PHI-Blast

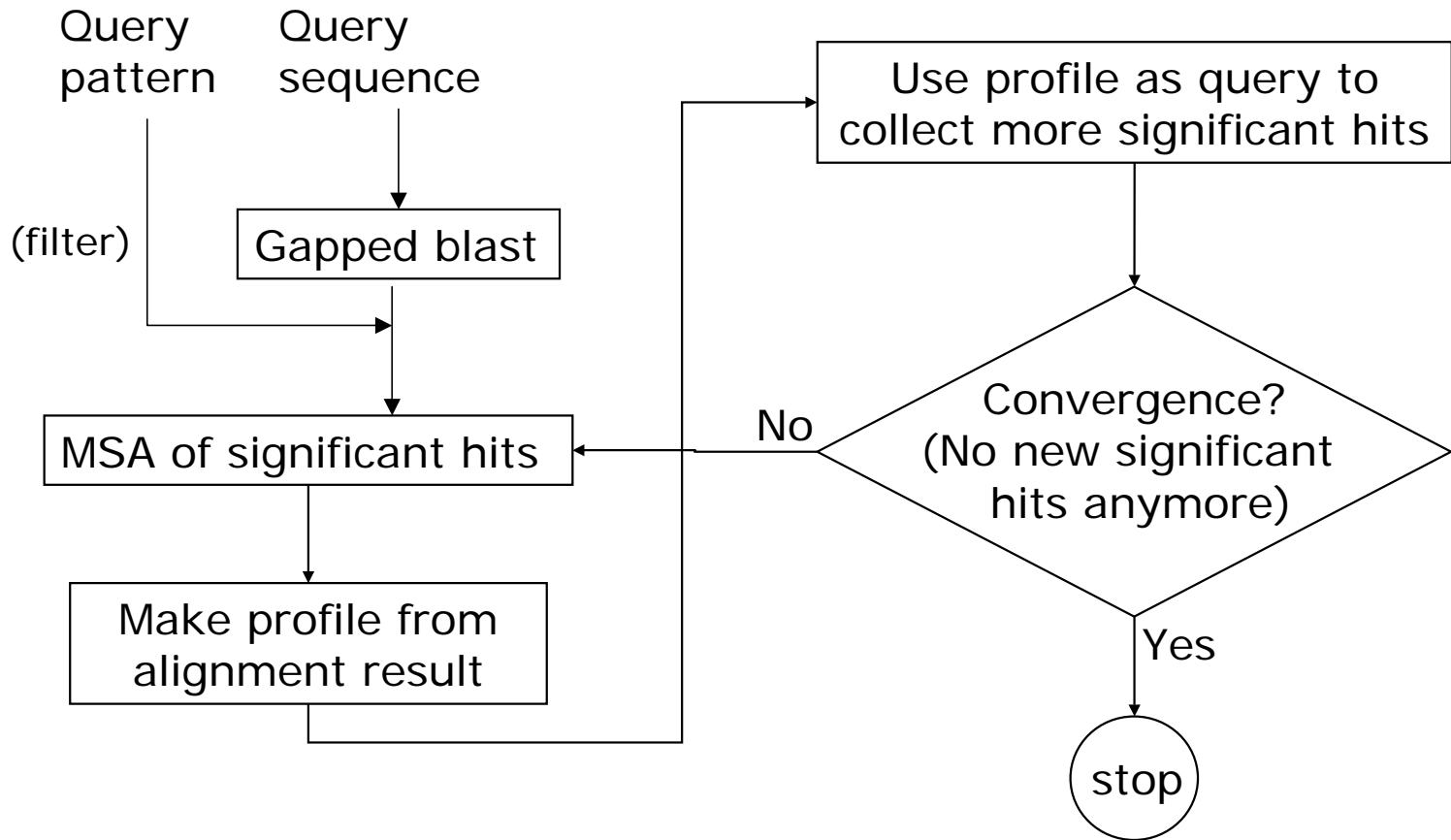
Similar to PSI-Blast, but start with a pattern rather than a sequence

# Syntax of a pattern

---

- [LFYT] = L or F or Y or T
- x(5) = xxxxx (x = any residues)
- x(2,4) = xx or xxx or xxxx
- Example:
  - ID ER\_TARGET; PATTERN. PA [KRHQSA]-[DENQ]-E-L>. HI (19 22) HI (201 204)
- \* <http://www.ncbi.nlm.nih.gov/blast/html/PHIsyntax.html>

# Steps in PHI-Blast



# Find all known domains in a sequence

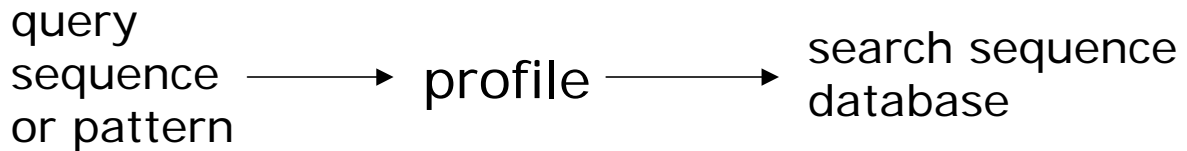
---

- Database: CDD (conserve domain database)
  - Pfam: <http://pfam.wustl.edu/>
  - smart (simple modular architecture research tool): <http://smart.embl-heidelberg.de/>
- Tools
  - CD search: use RPS-blast (reverse position specific blast) to search for domains in a given sequence

# RPS-Blast

---

- PSI- or PHI- blast : find novel motifs



- RPS-blast : find all known motifs in a query sequence
  - Using CDD database to search query sequence



- CDD : Conserve Domain Database
  - SMART
  - Pfam
  - NCBI

# Search for proteins with similar domain organization

- DART: Domain Architecture Retrieval Tool
  - Use data from precomputed RPS-blast analysis
- SMART: Simple modular architecture research tool
  - ?

# Summary of PSSM-related Blast analysis

---

- Discover novel motifs
  - PSI-blast
- Find all known motifs in a sequence
  - RPS-blast against CDD
- Search for proteins with similar domain organization
  - DART (Domain Architecture Retrieval Tool)

# The problem of using a profile

- when there are gaps, ...

---

MSA :

rat	A	C	G	G	A
ecoli	A	C	-	-	A
cow	A	-	-	-	A
corn	A	U	-	-	A

---

A	1	0	0	0	1
C	0	0.5	0	0	0
G	0	0	0.25	0.25	0
U	0	0.25	0	0	0
-	0	0.25	0.75	0.75	0

---

(\* This example was taken from Richard Hughey's handouts)



- Use HMM (Hidden Markov Model)

# Solution - use “state” expression

---

## Hidden Markov Model (HMM)

Patterns , profiles, ... *etc.* can be viewed as special cases of HMM

- Hidden-Markov Models alignments (HMMER)
  - Alignments based on profiles with variable size gaps lengths between consecutive positions.

# What is an HMM?

---

- A statistical model consisting of a number of interconnecting states – they are essentially linear chains of match, delete or insert states that attempt to encode the sequence conservation within aligned families.
  - Probabilities or costs (negative log-probabilities) are associated with each omission and each transition between states.
  - To align a sequence is to find the highest-probability (lowest-cost) path (sequence of states) through the HMM.
- Similar to a “weight matrix” that can recognise gaps and treat them in a systematic way.
- If it is used to represent a gene, then it will have different “states” that represent introns, exons, and intergenic regions.

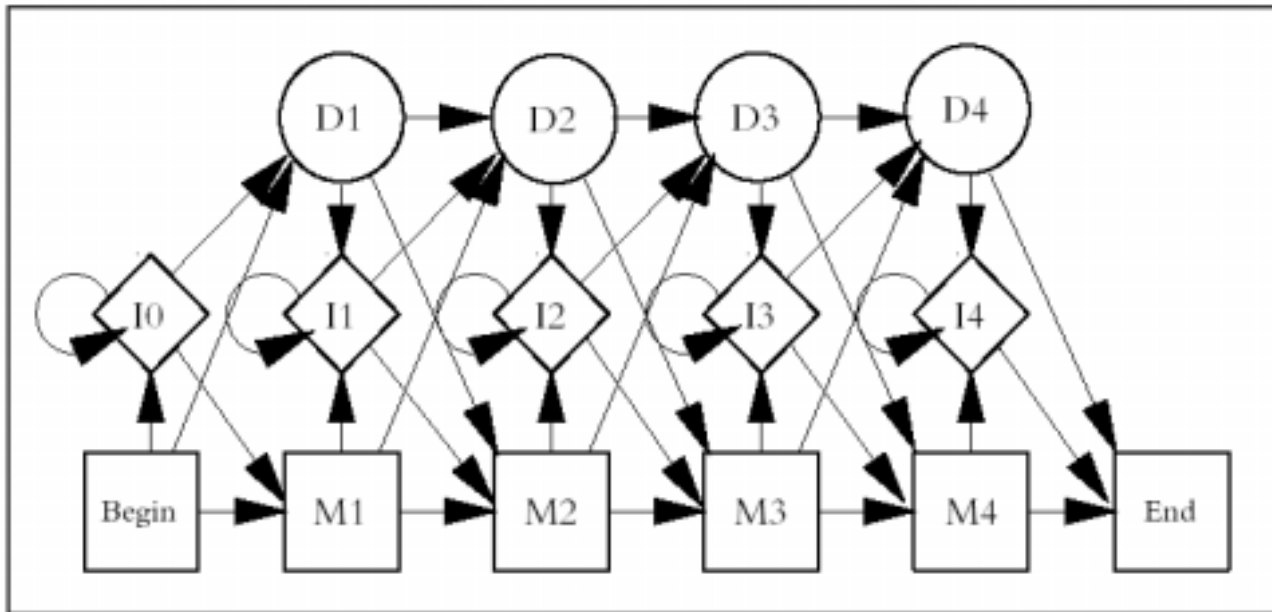
# HMM

---

- No alignment necessary. HMM searches are similar to Profile searches – used to perform position-dependent scoring.
- HMM scoring is a statistical model represented as a series of transitions between discrete states. In this case, the states are amino acids in a protein.
- In principle, HMMs can be developed from unaligned sequences by successive rounds of optimization, but in practice, protein profile HMMs are simply built from multiple sequence alignments.
- Work best with training set.
  - The most valuable data : Well-trained model

# HMM

- Each position of an alignment is represented as a match (M), an insert (I), or a delete (D) state in the HMM. This allows a query sequence to be aligned by the most probable state transition to each of its residues.

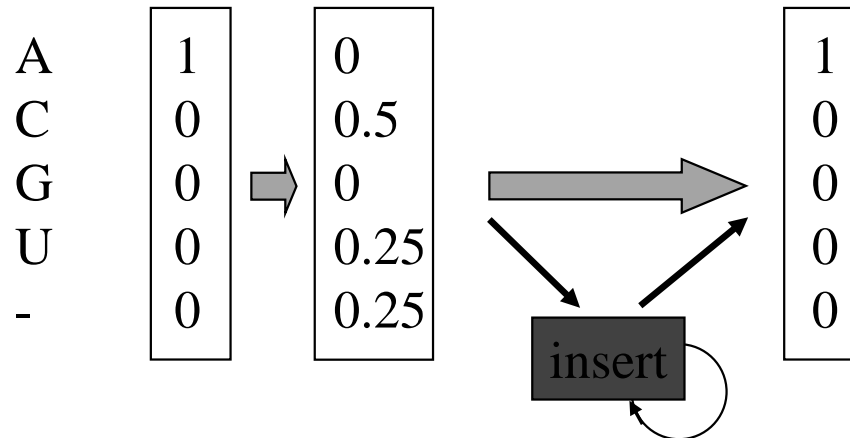


(BRIEFINGS IN BIOINFORMATICS. VOL 1. NO 1. 45-59. FEBRUARY 2000)

# Solution - use “state” expression

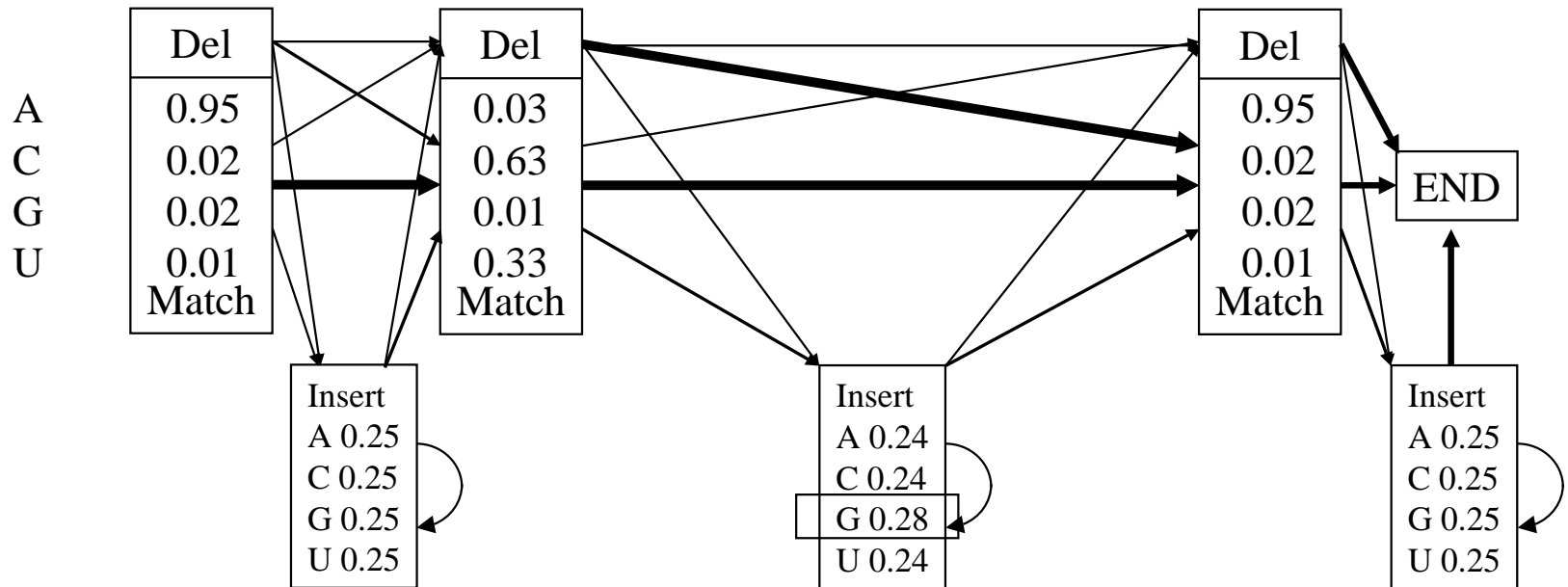
---

rat	A	C	G	G	A
ecoli	A	C	-	-	A
cow	A	-	-	-	A
corn	A	U	-	-	A



# There are two processes assoc. with each state: emission and transition

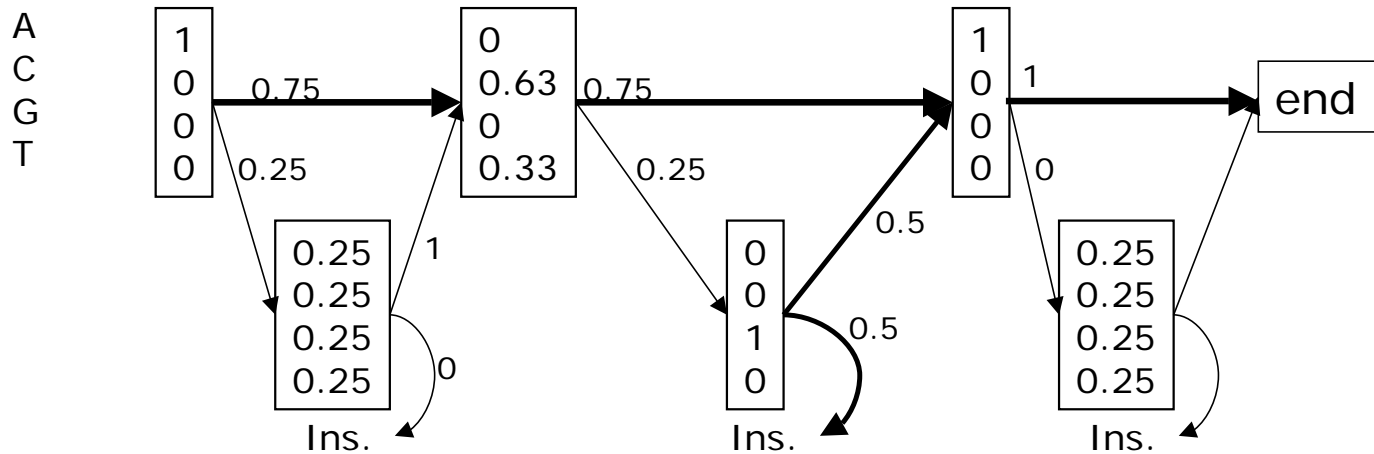
rat	A	C	G	G	A
ecoli	A	C	-	-	A
cow	A	-	-	-	A
corn	A	U	-	-	A



# HMM

- consider insertion/deletion path  $\longrightarrow$  allow gaps
- profiles can be viewed as special cases of HMM

1	A	C	G	G	A
2	A	C	-	-	A
3	A	-	-	-	A
4	A	T	-	-	A



# Major pattern databases

---

- Nucleic acids
  - transfac: transcription factor database
  - epd: eukaryotic promoter database
- Protein
  - primary: prosite, blocks, prints, pfam,,  
prodom, and smart
  - integrated: interpro, iproclass, ...*etc.*

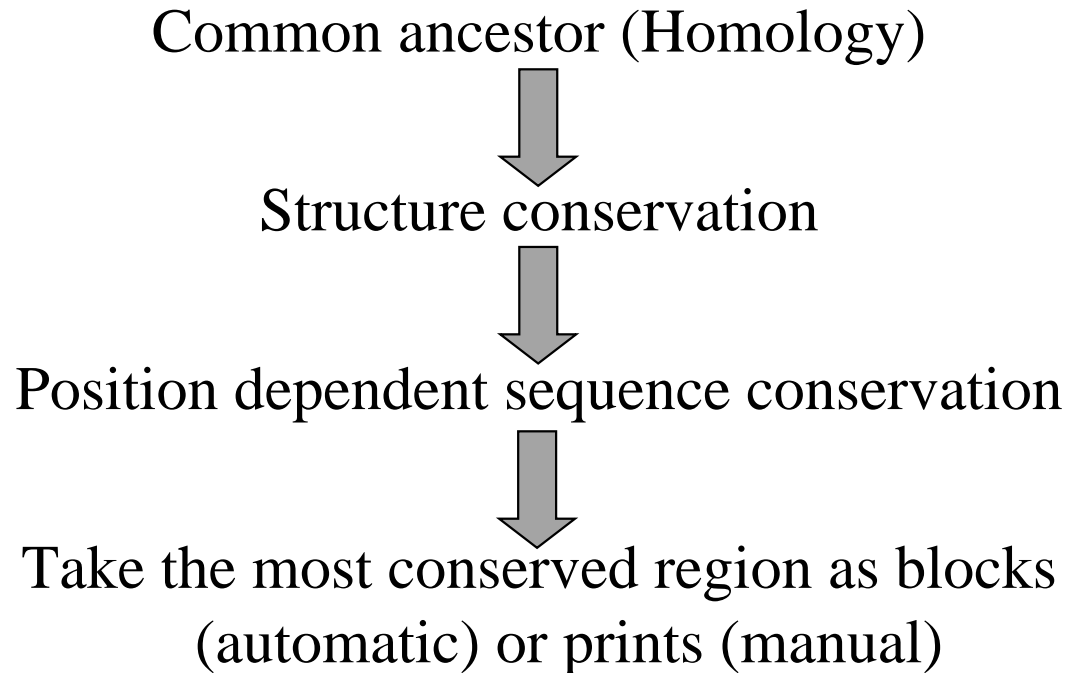
# Protein family databases

---

- Protein Clustering
  - PIR (PSD, ALN, ASDB): Superfamilies and Families; FASTA Clustering
  - COG (Clusters of Orthologous Groups) of Complete Genomes
  - ProtoMap: Automated Hierarchical Classification of Proteins
- Protein Domains
  - PIR (PSD, ALN): Homology Domains
  - Pfam: Alignments and HMM Models of Protein Domains
  - ProDom: Protein Domain Families
- Protein Motifs
  - PROSITE: Protein Patterns and Profiles
  - BLOCKS: Protein Sequence Motifs and Alignments
  - PRINTS: Protein Sequence Motifs and Signatures
- Integrated Family Databases
  - *i*ProClass: Superfamilies/Families, Domains, Motifs, Rich Links
  - InterPro: Integration of Pfam, P RINTS, PROSITES, ProDom
  - MetaFam: Supersets of > 10 Family Databases

# Blocks and Prints

---



---

Identify protein family by using blocks or prints

All the blocks or prints that characterize a domain should all exist

# ProDom

---

- Created by exhaustive PSI-blast analysis for those protein families defined in Prosite
- Cluster sequences and present them in a tree form
- Connect to structure-prediction servers, ... *etc.*

# Pfam

---

- Pfam = Protein families
- Based on prosite
- HMM models of aligned domains (full domain)
- database of multiple alignments of protein domains or conserved protein regions.
- Pfam-A are accurate human crafted multiple alignments
- Pfam-B is an automatic clustering of the rest of SWISS- PROT and TrEMBL proteins

# Protein domains

---

- Structural domains: conserved autonomous folding unit
- Functional domains: autonomous functional unit.  
The active site may be between two structural domains.  
Sometimes the function is sequence specific rather than structural specific (e.g., the nucleus localization signal): motifs
- Evolutional domains: observed combination of structural domains

# Types of domains

---

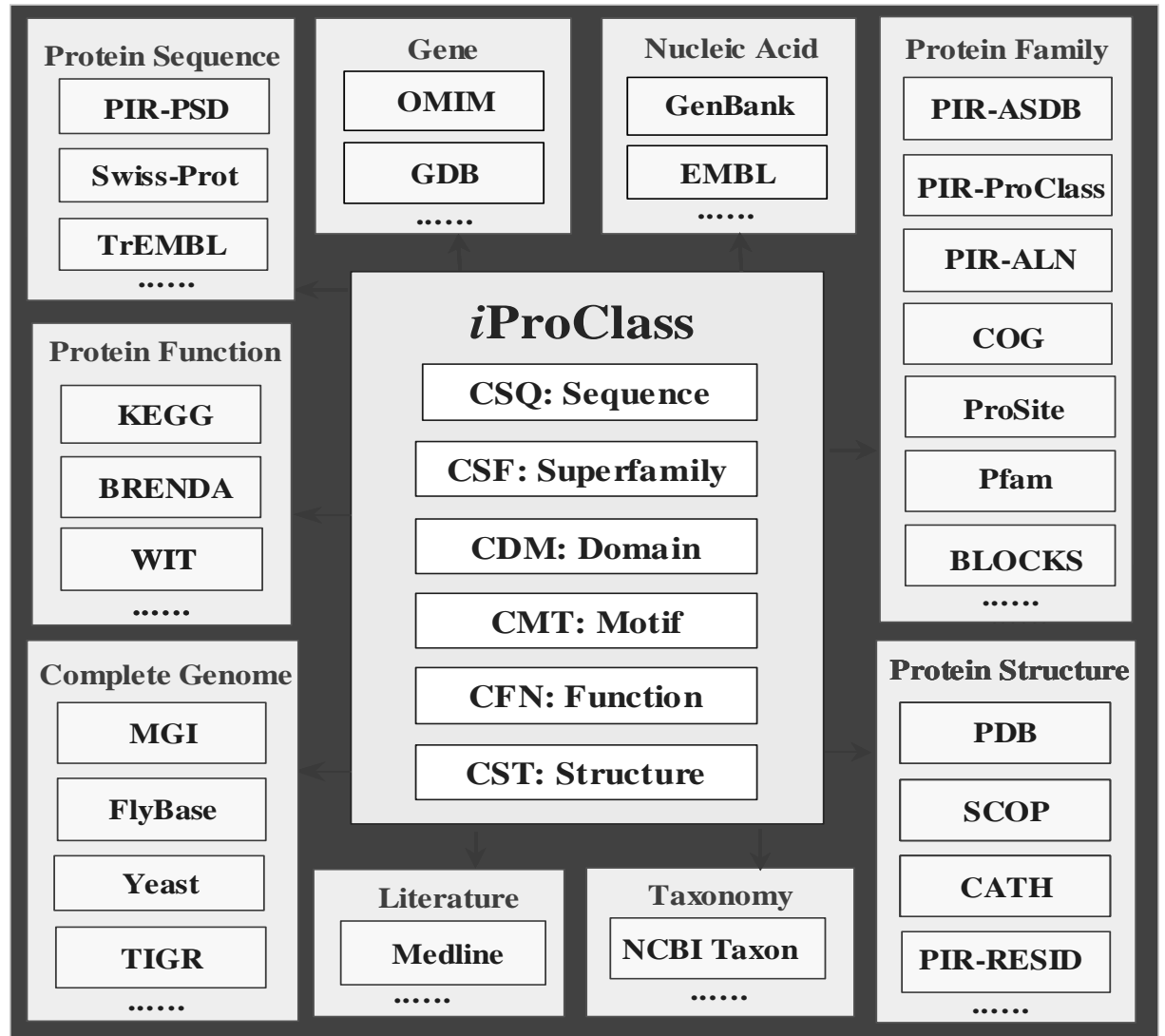
- Enzymatic domains
  - Relatively more conserved as they are directly related to function. They can be easily found by automatic procedures
- Regulatory domains
  - There are lots of sequence variations, because these domains may not have a strong selection pressure. As a result, it is difficult to find by automatic method.

# SMART

---

- Goal
  - to compile regulatory domains, such as those in the signal pathway, ... *etc.*
- Method
  - PSI-blast analysis and manual curation
  - Refer to tertiary structure in defining domains

# iProClass Overview



# InterPro

---

- Integrated resource of protein families, domains and sites
- Unified interface to Prosite, PRINTS, ProDom and Pfam
- Collaborators: EBI, SIB, University of Manchester, Sanger Centre, GENE-IT, CNRS/INRA, LION Bioscience AG and University of Bergen