Bioinformatics

Secondary database searching

Bioinformatics - lectures

- Introduction
- Information networks
- Protein information resources
- Genome information resources
- DNA sequence analysis
- Pairwise sequence alignment
- Multiple sequence alignment
- Secondary database searching
- Analysis packages
- Protein structure modelling

Secondary database searching

- why search secondary databases?
- secondary databases
- regular expressions
- fingerprints
- blocks
- profiles
- Hidden Markov Models

Why search secondary databases?

- Interpretation of the results from primary database searches is sometimes difficult:
  - X.000.000 sequences from XX.000 organisms
  - complex and redundant search outputs
  - irrelevant matches of low-complexity sequences, repetitive sequences, modular sequences
  - local regions of similarity in multi-domain proteins
  - truncated description lines

- Secondary database searches enable to identify both homology and more exacting orthology.

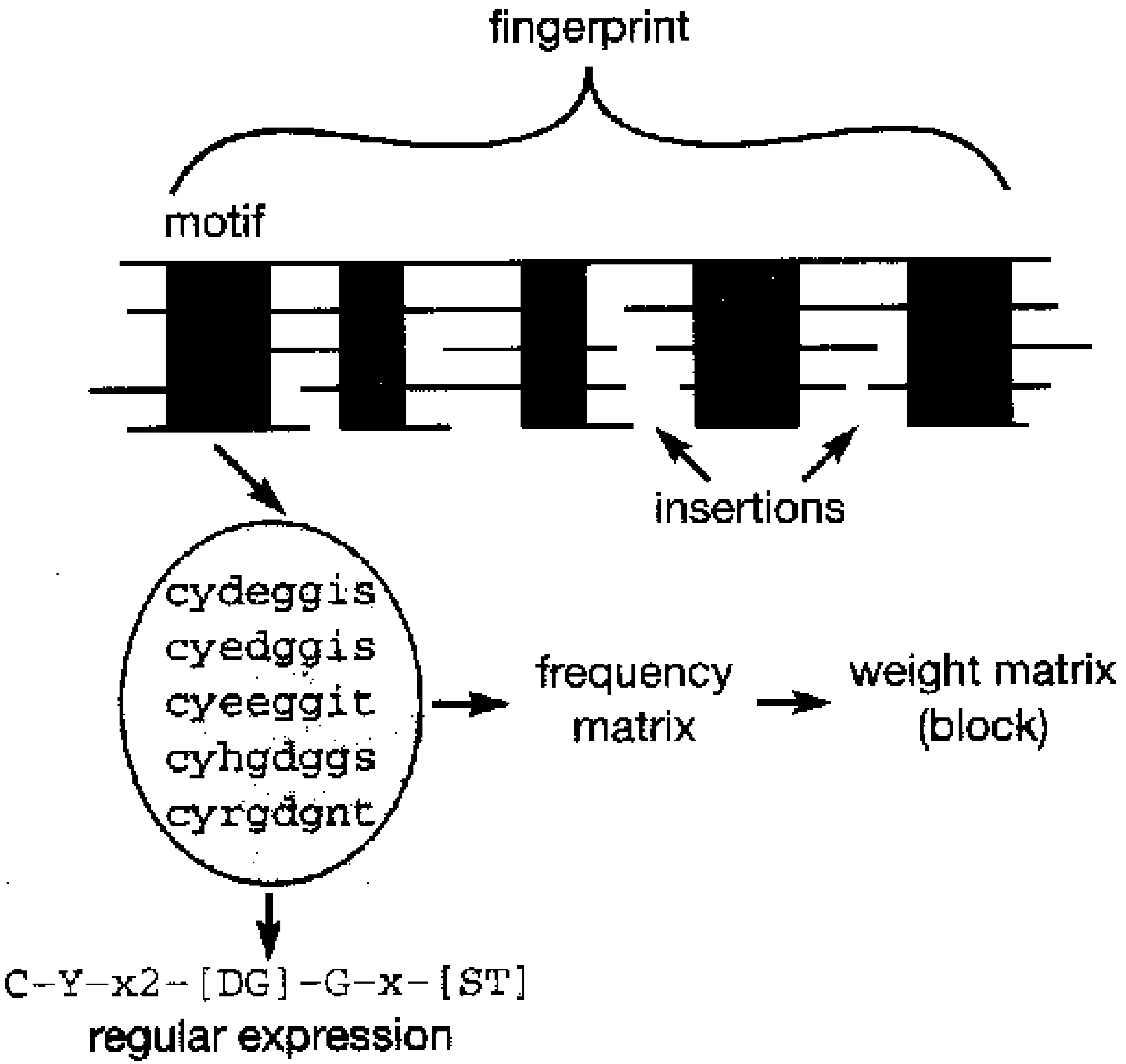
Secondary databases

- Contains information derived from primary sequence data, typically in the form of abstractions: regular expressions, fingerprints, blocks, profiles or Hidden Markov Models.
- These abstractions represent destillations of the most conserved features of multiple alignments.
- The abstractions are useful for discrimination of family membership for newly determined sequences.

Secondary databases

- PROSITE - regular expressions
- PRINTS - fingerprints
- BLOCKS - blocks
- PROFILES - profiles
- PFAM - Hidden Markov Models
- IDENTIFY - fuzzy regular expressions

Terms used in sequence analysis methods



fingerprint

motif

insertions

cydeggis
cyedggis
cyeeggit
cyhgdggs
cyrgdgnt

frequency matrix

weight matrix (block)

C-Y-x2-[DG]-G-x-[ST]

regular expression

Regular expressions

- Regular expression reduces the sequence data to the most conserved residue information.

```
Multiple alignment          Regular expression
ADLGAVFALCDRYFQ             [AS]-D-[IVL]-G-X5-C-[DE]-R-[FY]2-Q
SDVGPRSCFCERFYQ
ADLGRTQNRCDRYYQ
ADIGQPHSLCERYFQ
```

- Limitations:
  - stringent pattern - retrieves only identical matches and can miss remote relatives
  - fuzzier pattern - better chance to detect remote relatives, but results in more noisy output
  - single motif may not be sufficient to infer the function

Regular expressions

- Regular expressions works most effectively when a particular protein family can be characterised by a highly conserved motif (10-20 residues).

- Limitation: short patterns (3-4 residues) are not sufficiently discriminative.

Asp-Ala-Val-Ile-Asp (DAVID)	71 exact matches in OWL29.6
Asp-Ala-Val-Glu (DAVE)	1088 exact matches in OWL29.6

Regular expressions

- Rules - short patterns that can be used to provide a guide to possible existence of functional sites:

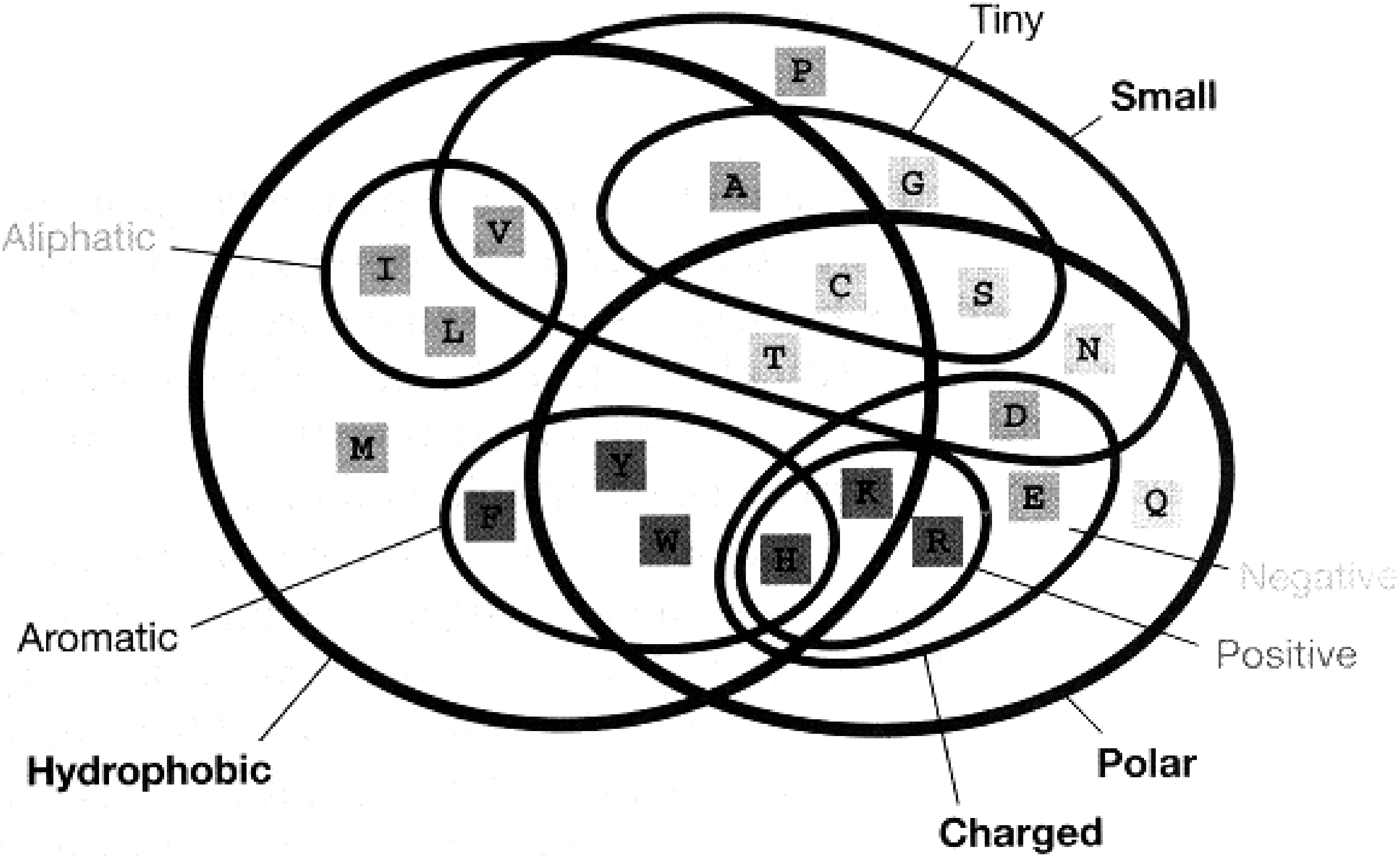
Functional site	Regular expression
N-glycosilation site	N-{P}-[ST]-{P}
Protein kinase C phosphorylation site	[ST]-X-[RK]
Casein kinase II phosphorylation site	[ST]-X(2)-[DE]
Asp adn Asn hydroxylation site	C-X-[DN]-X(4)-[FY]-X-C

Regular expressions

- Fuzzy regular expressions - regular expressions with introduced fuzziness into patterns using groups of amino acids with similar biochemical properties (FYW - aromatic, HKR - basic, etc.).

Multiple alignment	Fuzzy regular expression
ADLGAVFALCDRYFQ	[ASGPT]-D-[IVLM]-G-X5-C-[DENQ]-R-[FYW]2-Q
SDVGPRSCFCERFYQ	
ADLGRTQNRCDRYYQ	
ADIGQPHSLCERYFQ	

Residue	*Property*	*Colour*
Asp, Glu	Acidic	red
His, Arg, Lys	Basic	blue
Ser, Thr, Asn, Gln	Polar neutral	green
Ala, Val, Leu, Ile, Met	Hydrophobic aliphatic	white
Phe, Try, Trp	Hydrophobic aromatic	purple
Pro, Gly	Special structural properties	brown
Cys	Disulphide bond former	yellow


Tiny
Small
Aliphatic
Aromatic
Hydrophobic
Charged
Polar
Positive
Negative
P
G
A
V
I
L
C
S
N
T
D
M
Y
K
E
Q
F
W
H
R

Regular expressions

- Introduction fuzziness into regular expressions increases the number of matches retrieved from the sequence database:

Regular expression	No. of exact matches (OWL29.6)
D-A-V-I-D	71
D-A-V-I-[DENQ]	252
[DENQ]-A-V-I-[DENQ]	925
[DENQ]-A-[VLI]-I-[DENQ]	2739
[DENQ]-[AQ]-[VLI]2-[DENQ]	51506

Fingerprints

- Motivation: there are often more than one conserved region present in multiple alignment.
- Groups of motifs excised from the sequence and converted into matrices populated by the residue frequencies observed at each position.
- Unweighted scoring system - no additional mutation or substitution matrices are employed.
- Weighted scoring system - additional matrices are employed resulting in less sparse matrix, but poor signal-to-noise performance.

(a)

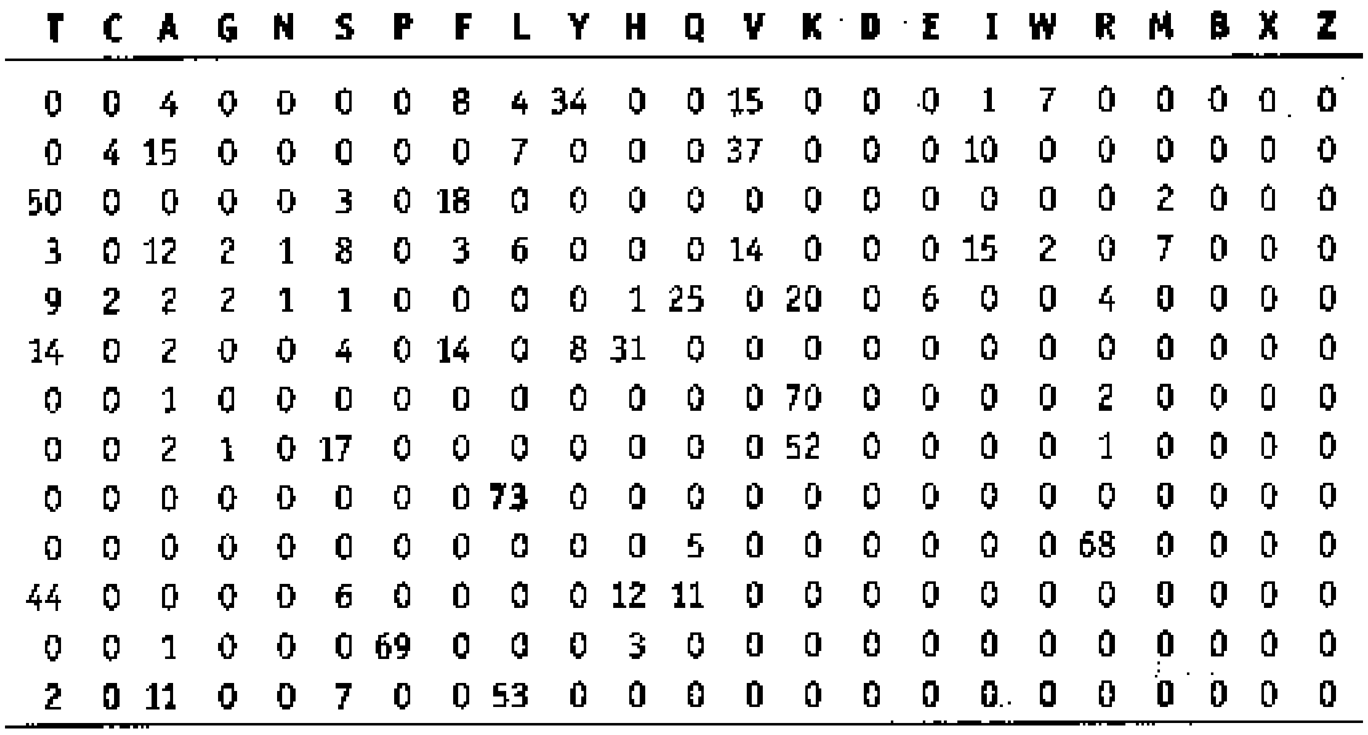
YVTVQH**K**KLRT**P**L
YVTVQH**K**KLRT**P**L
YVTVQH**K**KLRT**P**L
AATMKF**K**K**L**RH**P**L
AATMKF**K**K**L**RH**P**L
YIFATT**K**S**L**RT**P**A
VATLRY**K**K**L**RQ**P**L
YIFGGT**K**S**L**RT**P**A
WVFSAA**K**S**L**RT**P**S
WIFSTS**K**S**L**RT**P**S
YLFSKT**K**S**L**QT**P**A
YLFTKT**K**S**L**QT**P**A

(b)

T	C	A	G	N	S	P	F	L	Y	H	Q	V	K	D	E	I	W	R	M	B	X	Z
0	0	2	0	0	0	0	0	0	7	0	0	1	0	0	0	0	2	0	0	0	0	0
0	0	3	0	0	0	0	0	2	0	0	0	4	0	0	0	3	0	0	0	0	0	0
0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	0	3	0	0	1	0	0	0	3	0	0	0	0	0	0	2	0	0	0
0	0	1	1	0	0	0	0	0	0	0	3	0	4	0	0	0	0	1	0	0	0	0
0	0	1	0	0	1	0	2	0	1	3	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0
0	0	0	0	0	6	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	10	0	0	0	0
0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	4	0	0	2	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0

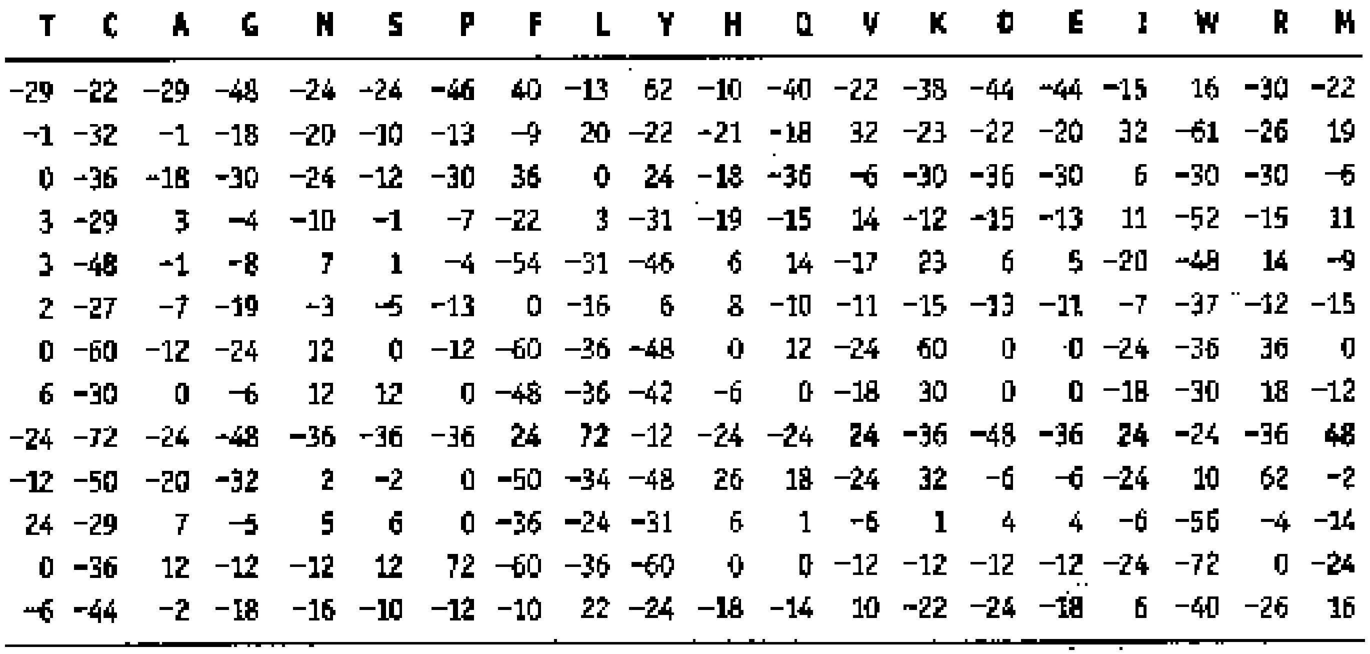
Example of frequency matrix derived from initial unweighted motif (a) and PAM-weighted matrix (b)

(a)



T	C	A	G	N	S	P	F	L	Y	H	Q	V	K	D	E	I	W	R	M	B	X	Z
0	0	4	0	0	0	0	8	4	34	0	0	15	0	0	0	1	7	0	0	0	0	0
0	4	15	0	0	0	0	0	7	0	0	0	37	0	0	0	10	0	0	0	0	0	0
50	0	0	0	0	3	0	18	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
3	0	12	2	1	8	0	3	6	0	0	0	14	0	0	0	15	2	0	7	0	0	0
9	2	2	2	1	1	0	0	0	0	1	25	0	20	0	6	0	0	4	0	0	0	0
14	0	2	0	0	4	0	14	0	8	31	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	70	0	0	0	0	2	0	0	0	0
0	0	2	1	0	17	0	0	0	0	0	0	0	52	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	73	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	68	0	0	0	0
44	0	0	0	0	6	0	0	0	0	12	11	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	69	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
2	0	11	0	0	7	0	0	53	0	0	0	0	0	0	0	0	0	0	0	0	0	0

(b)



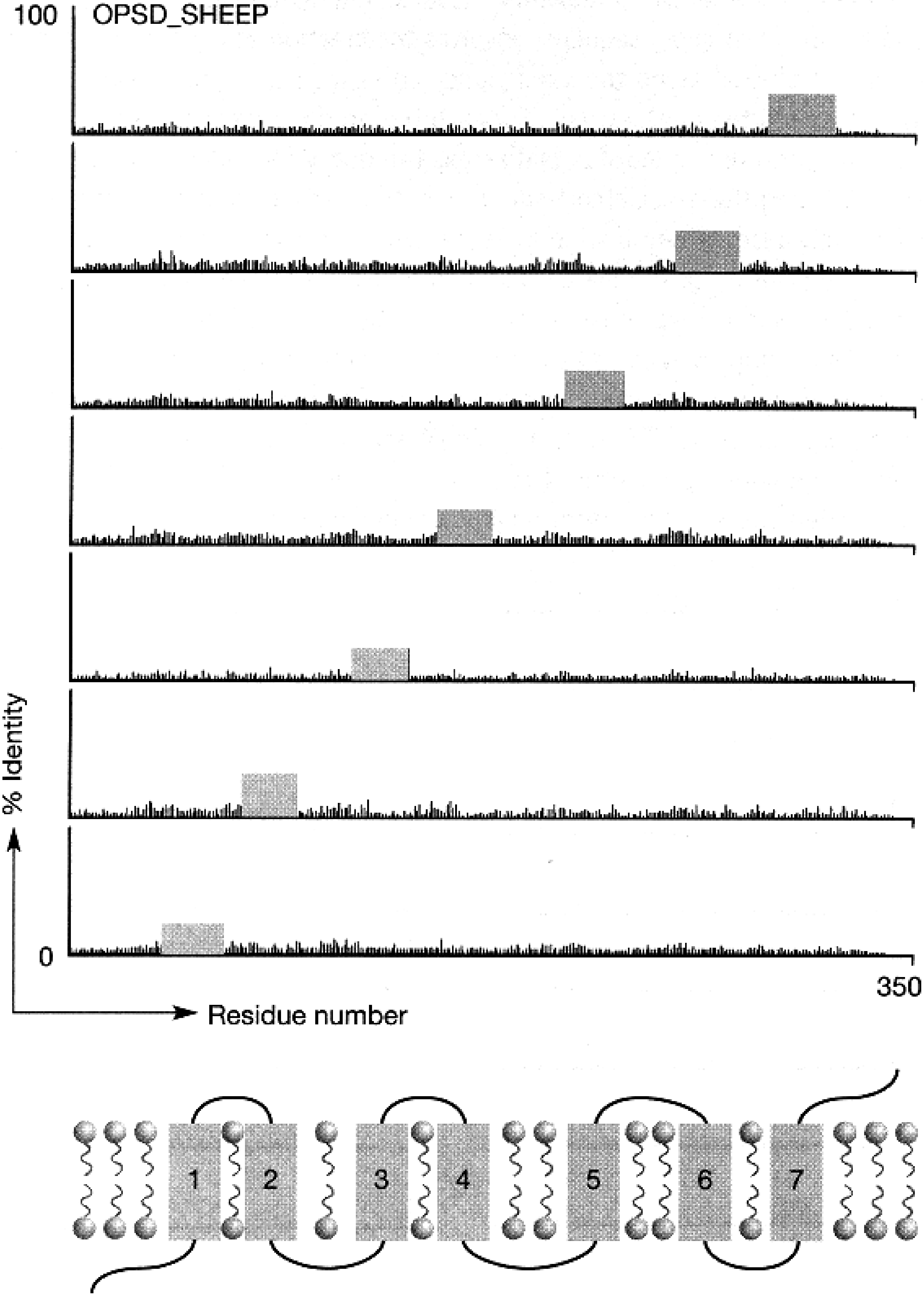
T	C	A	G	N	S	P	F	L	Y	H	Q	V	K	D	E	I	W	R	M
−29	−22	−29	−48	−24	−24	−46	40	−13	62	−10	−40	−22	−38	−44	−44	−15	16	−30	−22
−1	−32	−1	−18	−20	−10	−13	−9	20	−22	−21	−18	32	−23	−22	−20	32	−61	−26	19
0	−36	−18	−30	−24	−12	−30	36	0	24	−18	−36	−6	−30	−36	−30	6	−30	−30	−6
3	−29	3	−4	−10	−1	−7	−22	3	−31	−19	−15	14	−12	−15	−13	11	−52	−15	11
3	−48	−1	−8	7	1	−4	−54	−31	−46	6	14	−17	23	6	5	−20	−48	14	−9
2	−27	−7	−19	−3	−5	−13	0	−16	6	8	−10	−11	−15	−13	−11	−7	−37	−12	−15
0	−60	−12	−24	12	0	−12	−60	−36	−48	0	12	−24	60	0	0	−24	−36	36	0
6	−30	0	−6	12	12	0	−48	−36	−42	−6	0	−18	30	0	0	−18	−30	18	−12
−24	−72	−24	−48	−36	−36	−36	24	72	−12	−24	−24	24	−36	−48	−36	24	−24	−36	48
−12	−50	−20	−32	2	−2	0	−50	−34	−48	26	18	−24	32	−6	−6	−24	10	62	−2
24	−29	7	−5	5	6	0	−36	−24	−31	6	1	−6	1	4	4	−6	−56	−4	−14
0	−36	12	−12	−12	12	72	−60	−36	−60	0	0	−12	−12	−12	−12	−24	−72	0	−24
−6	−44	−2	−18	−16	−10	−12	−10	22	−24	−18	−14	10	−22	−24	−18	6	−40	−26	16

Blocks

- Conserved motifs are located by a first motif-finding algorithm: search for the spaced residue triplets (e.g., Ala-X-X-Val-X-Trp); a block score is weighted using BLOSUM 62 substitution matrix.

- Validation of blocks by a second motif-finding algorithm: search for the highest-scoring set of blocks in the correct order without overlapping.

- Sequences are clustered to avoid a bias due to identical sequences.


100
OPSD_SHEEP
7
6
5
4
3
2
1
% Identity
0
350
Residue number
1
2
3
4
5
6
7

```
 CCKR_HUMAN ( 362) SSCVNPIIYCFMNKRFR 3
   CCKR_RAT ( 378) SSCVNPIIYCFMNKRFR 3

 FML2_HUMAN ( 294) NSCLNPMLYVFVGQDFR 4
 FMLR_HUMAN ( 293) NSCLNPMLYVFMGQDFR 4
 FMLR_MOUSE ( 304) NSCLNPMLYVFMGQDFR 4
 FMLR_RABIT ( 295) NSCLNPMLYVFMGQDFR 4

 GASR_CANFA ( 388) SACVNPLVYCFMHRRFR 5
 GASR_HUMAN ( 382) SACVNPLVYCFMHRRFR 5
 GASR_PRANA ( 385) SACVNPLVYCFMHRRFR 5
 GASR_RABIT ( 387) SACVNPLVYCFMHRRFR 5
   GASR_RAT ( 387) SACVNPLVYCFMHRRFR 5

 ET1R_BOVIN ( 361) NSCINPIALYFVSKKFK 9
   ET1R_RAT ( 361) NSCINPIALYFVSKKFK 9
 ETBR_BOVIN ( 377) NSCINPIALYLVSKRFK 9
 ETBR_HUMAN ( 378) NSCINPIALYLVSKRFK 9
   ETBR_PIG ( 379) NSCINPIALYLVSKRFK 9
   ETBR_RAT ( 378) NSCINPIALYLVSKRFK 9

 OPSD_LOLFO ( 307) SAIHNPMIYSVSHPKFR 12
 OPSD_OCTDO ( 308) SAIHNPIVYSVSHPKFR 12
 OPSD_TODPA ( 306) SAIHNPMIYSVSHPKFR 12

 P2UR_HUMAN ( 296) NSCLDPVLYFLAGQRLV 13
 P2UR_MOUSE ( 298) NSCLDPVLYFLAGQRLV 13
   P2UR_RAT ( 297) NSCLDPVLYFLAGQRLV 13

    5H6_RAT ( 312) NSTMNPIIYPLFMRDFK 16

 EDG1_HUMAN ( 302) NSGTNPIIYTLTNKEMR 21

 EBI2_HUMAN ( 300) NCCMDPFIYFFACKGYK 23

 OXYR_HUMAN ( 321) NSCCNPWIYMLFTGHLF 24
   OXYR_PIG ( 323) NSCCNPWIYMLFTGHLF 24
 V1AR_HUMAN ( 340) NSCCNPWIYMFFSGHLL 18
   V1AR_RAT ( 346) NSCCNPWIYMFFSGHLL 18

 PER3_BOVIN ( 337) NQILDPWVYLLLRKILL 35
 PER3_HUMAN ( 338) NQILDPWVYLLLRKILL 35

 YN84_CAEEL ( 331) SCVAYPLIFTLLNRGIR 100
```

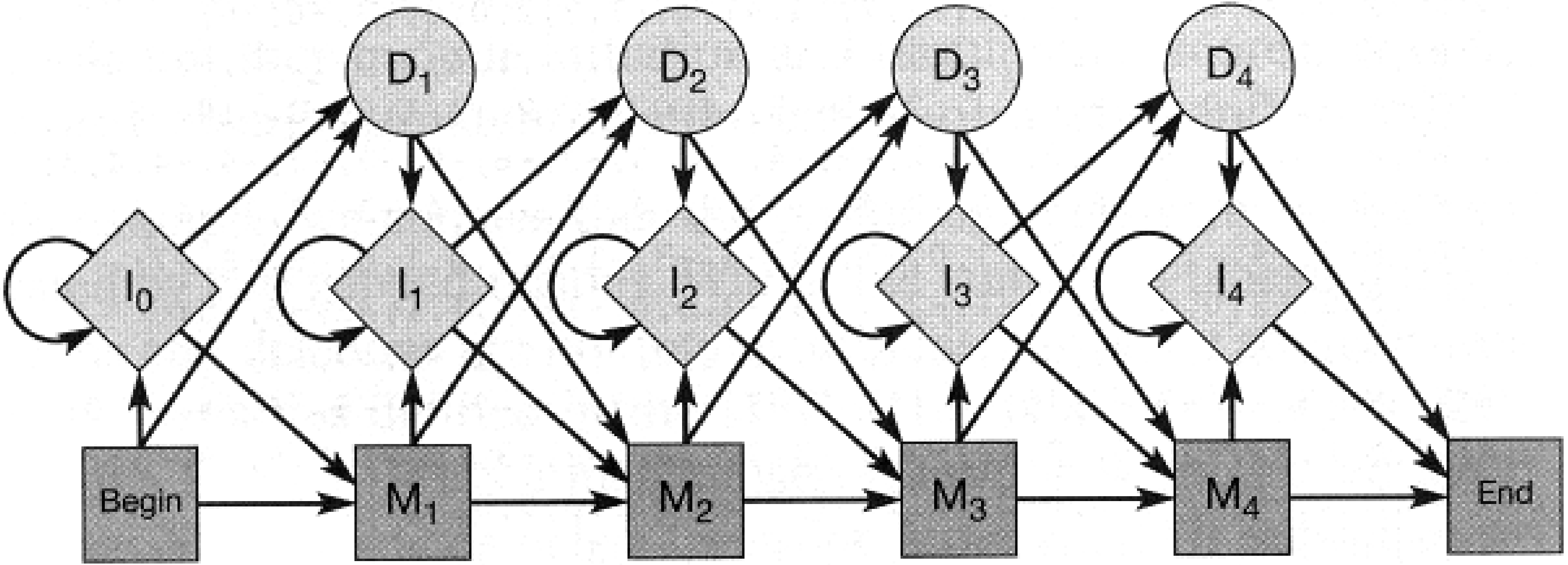
Profiles

- Based on entire sequences.
- Profiles define which residues are allowed at given positions, which positions are highly conserved and which degenerate, which positions can tolerate insertions.
- The scoring system may include evolutionary weights and results from structural analysis.

```
/DEFAULT: MI=-26; I=-3; IM=0; MD=-26; D=-3; DM=0;
/M: SY='F';M=-2,-3,-3,-4,2,-3,-2,1,-2,0,-1,-2,-3,-3,-4,-2,-1,0,-5,2;
/M: SY='I';M=-1,-5,-2,-3,-2,-3,0,1,1,-1,1,-1,-2,-1,1,-1,0,1,-4,-4;
/M: SY='A';M=2,-3,1,0,-5,2,-2,-1,-1,-3,-2,1,1,0,-2,2,2,0,-8,-5;
/M: SY='L';M=-3,-8,-5,-4,2,-6,-2,2,-4,6,4,-3,-3,-2,-3,-3,-2,1,-3,0;
/M: SY='Y';M=-4,-2,-6,-6,9,-7,0,-1,-5,-1,-3,-3,-6,-5,-6,-4,-4,-4,-1,11;
/M: SY='D';M=1,-6,3,3,-7,0,0,-2,-1,-4,-3,2,0,1,-2,0,0,-2,-9,-6;
/M: SY='Y';M=-5,-3,-6,-6,10,-7,-1,-1,-2,-1,-2,-3,-6,-5,-5,-4,-4,-4,-1,11;
/M: SY='K';M=-1,-6,1,1,-4,-2,0,-2,2,-3,-1,1,-1,1,1,0,0,-3,-7,-6;
/M: SY='A';M=1,-4,1,0,-5,1,-1,-1,0,-3,-1,1,0,0,0,1,1,-1,-7,-6;
/M: SY='R';M=0,-5,0,0,-5,-1,0,-1,1,-3,-1,1,0,1,1,0,0,-2,-5,-5;
/M: SY='R';M=0,-5,1,1,-6,0,1,-2,1,-4,-2,1,0,1,2,1,0,-2,-5,-5;
/M: SY='E';M=1,-6,2,2,-6,0,0,-2,-1,-4,-2,1,1,1,-1,0,0,-3,-8,-6;
/M: SY='D';M=0,-6,2,2,-6,0,1,-3,0,-5,-3,2,-1,2,-1,0,0,-4,-7,-4;
/M: SY='D';M=0,-8,4,3,-6,0,0,-2,-1,-3,-2,2,-2,2,-2,0,-1,-3,-9,-6;
/M: SY='L';M=-2,-8,-5,-5,2,-5,-3,3,-4,7,5,-4,-3,-3,-4,-3,-2,3,-4,-2;
/M: SY='S';M=1,-4,1,1,-5,1,0,-2,1,-4,-2,1,0,0,0,1,1,-2,-6,-5;
/M: SY='F';M=-3,-7,-6,-6,6,-5,-3,3,-2,5,3,-4,-5,-4,-5,-4,-3,1,-3,3;
/M: SY='Q';M=-1,-6,0,0,-3,-2,1,-1,1,-2,0,0,-1,1,1,-1,0,-1,-6,-4;
/M: SY='K';M=-1,-8,0,1,-3,-2,0,-2,3,-3,0,1,0,2,2,0,0,-3,-6,-6;
/M: SY='G';M=2,-5,1,0,-7,7,-3,-4,-2,-6,-4,1,-1,-2,-4,2,0,-2,-10,-8;
/M: SY='D';M=1,-7,5,4,-8,1,1,-3,0,-5,-3,2,-1,2,-2,0,0,-4,-10,-6;
/M: SY='I';M=0,-5,-1,-2,-2,-2,-1,2,0,0,1,-1,-2,0,0,-1,0,1,-6,-5;
/M: SY='L';M=-2,-6,-5,-5,3,-5,-3,4,-3,6,4,-4,-4,-3,-4,-3,-2,3,-5,0;
/M: SY='Q';M=-1,-5,-1,-1,-3,-2,0,0,0,-2,-1,0,-1,0,0,-1,0,-1,-6,-3;
/M: SY='V';M=0,-4,-3,-4,-1,-3,-3,5,-3,3,3,-2,-2,-2,-3,-2,0,5,-8,-4;
/M: SY='L';M=-1,-6,-3,-3,-1,-3,-2,2,-3,3,2,-2,-2,-2,-3,-2,-1,2,-5,-3;
/M: SY='D';M=0,-6,3,3,-6,0,1,-3,2,-5,-2,2,-1,2,1,0,0,-4,-7,-5;
/M: SY='K';M=-1,-6,0,0,-2,-1,0,-3,3,-4,-1,1,-1,0,1,0,0,-3,-6,-4;
/M: SY='N';M=1,-4,1,1,-5,0,0,-2,0,-3,-2,1,1,0,-1,1,1,-1,-7,-5;
   /I: MI=0; I=-1; MD=0; /M: SY='X'; M=0; D=-1;
/M: SY='G';M=1,-5,0,0,-5,1,-2,-1,-2,-3,-2,0,0,-1,-2,0,0,-1,-8,-6;
/M: SY='G';M=1,-6,3,3,-7,3,0,-4,-1,-5,-4,2,-1,1,-2,1,0,-3,-10,-6;
/M: SY='W';M=-9,-12,-9,-11,1,-11,-4,-8,-5,-3,-6,-6,-8,-7,3,-4,-8,-9,26,0;
/M: SY='W';M=-7,-9,-9,-9,0,-9,-4,-5,-5,-1,-4,-6,-7,-6,2,-3,-6,-6,18,-1;
/M: SY='K';M=-1,-7,0,0,-3,-2,0,-2,2,-3,-1,1,-1,1,2,0,-1,-3,-5,-5;
/M: SY='G';M=2,-3,0,-1,-6,3,-3,-2,-3,-4,-3,0,0,-2,-3,1,0,0,-10,-6;
/M: SY='Q';M=-2,-6,0,0,-3,-3,1,-2,0,-2,-1,0,-2,1,1,-1,-1,-3,-5,-3;
   /I: MI=0; I=-2; MD=0; /M: SY='X'; M=0; D=-2;
/M: SY='T';M=0,-4,-1,-1,-4,0,-2,0,-1,-2,0,0,-1,-1,-1,0,1,0,-7,-5;
/M: SY='T';M=0,-5,0,0,-3,-1,-1,-1,1,-3,-1,1,-1,0,0,1,1,-1,-6,-4;
/M: SY='G';M=0,-5,0,-1,-5,3,-2,-3,-1,-5,-3,0,-1,-1,-1,1,0,-2,-7,-6;
/M: SY='K';M=0,-6,1,1,-5,-1,1,-2,2,-4,-1,1,-1,2,2,0,0,-3,-6,-6;
/M: SY='R';M=-1,-6,-1,-1,-5,-3,1,-1,1,-3,-1,0,-1,1,3,-1,-1,-2,-2,-6;
/M: SY='G';M=1,-5,0,0,-6,6,-3,-3,-3,-5,-4,0,-1,-2,-4,1,0,-2,-10,-6;
/M: SY='W';M=-5,-5,-5,-5,2,-6,-2,-2,-4,-1,-3,-3,-6,-5,-3,-3,-4,-4,4,3;
/M: SY='F';M=-3,-5,-6,-6,6,-5,-3,4,-1,3,2,-4,-4,-5,-4,-3,-2,2,-4,3;
/M: SY='P';M=2,-4,-1,-1,-7,-1,0,-3,-2,-4,-3,-1,8,0,0,1,0,-2,-8,-7;
/M: SY='G';M=1,-3,0,0,-4,2,-1,-2,0,-3,-2,0,0,-1,-1,1,1,-1,-6,-5;
/M: SY='N';M=1,-5,2,1,-5,0,1,-2,1,-4,-2,2,0,0,0,1,1,-2,-7,-4;
/M: SY='Y';M=-5,-1,-7,-7,10,-8,-1,-1,-5,-1,-3,-3,-7,-6,-6,-4,-4,-5,0,13;
/M: SY='V';M=0,-3,-3,-5,-2,-2,-3,5,-3,2,2,-2,-2,-3,-4,-1,0,5,-8,-5;
/M: SY='E';M=1,-6,2,3,-6,0,0,-2,1,-4,-2,1,0,2,0,0,0,-3,-8,-6;
/M: SY='P';M=0,-5,-1,-1,-2,-2,-1,-2,-1,-3,-2,0,1,-1,-2,0,-1,-2,-6,-3;
```

Hidden Markov Models

- Based on entire sequences.
- HMMs are probabilistic models consisting of a number of interconnecting states - linear chains of match, delete or insert states.
- Each position in the multiple alignment is assigned to either match, insert or delete state.
- Construction: seed alignment, iterative sequence gathering, final alignment (all automatic).


D1
D2
D3
D4
I0
I1
I2
I3
I4
Begin
M1
M2
M3
M4
End