Βιοπληροφορική

Διάλεξη 9η :

Μοτίβα αλληλουχιών και χρήση τους :
Regular expressions, profiles, motifs, fingerprints, blocks
και Hidden Markov Models (HMMs).
Εύρεση γονιδίων : HMMs.

Εισαγωγή

Regexps, profiles, motifs, fingerprints, blocks, HMMs ...

Στόχος όλων αυτών των προσεγγίσεων είναι η εύρεση και ταξινόμηση μοτίβων σε συγγενείς αλληλουχίες με στόχο τη χρήση τους για τη διάγνωση ομολογιών με μη χαρακτηρισμένες αλληλουχίες. Οι προσεγγίσεις αυτές κυρίως στηρίζονται στην κωδικοποίηση της πληροφορίας που περιέχεται σε μια στοίχιση πολλών (ομόλογων) αλληλουχιών, και τη χρήση αυτής της πληροφορίας για το χαρακτηρισμό νέων αλληλουχιών. Εξαίρεση είναι τα Hidden Markov Models για τα οποία είναι εφικτή (άλλα τεχνικά δύσκολη) η ταυτόχρονη βελτιστοποίηση της στοίχισης και του μοτίβου.

Εισαγωγή

Ανάλογα με τον τρόπο κωδικοποίησης της πληροφορίας που περιέχεται σε ένα multiple sequence alignment, οι μέθοδοι διακρίνονται σε

- Regular expressions (regexp) και fuzzy regular expressions.
- Fingerprints
- Blocks
- Profiles
- Hidden Markov Models (profiles και motifs)

Regular expressions

Αυτή είναι η πλέον απλή προσέγγιση η οποία προ-σπαθεί να χαρακτηρίσει μια οικογένεια αλληλουχιών με βάση ένα συντηρημένο μοτίβο. Το μοτίβο αυτό έχει τη μορφή μίας consensus αλληλουχίας η οποία περιγράφει ποιά κατάλοιπα μπορούν να είναι (ή να μην είναι) στις διάφορες θέσεις του μοτίβου. Τα regular expressions λόγω του τρόπου δημιουργίας τους χρησιμοποιούν τη στοίχιση των αλληλουχιών μόνο για δημιουργία του regexp (η πληροφορία ολόκληρης της στοίχισης δε χρησιμοποιείται για την ανάλυση νέων αλληλουχιών). Τα regexps έχουν τη ρίζα τους στα regular expressions του unix (awk, grep, sed, ..., αλλά τώρα πια και στην perl).

Regular expressions

ADLGAVFALCDRYFQ
SDVGPRSCFCERFYQ
ADLGRTQNRCDRYYQ
ADIGQPHSLCERYFQ

[AS]-D-[IVL]-G-x4-{PG}-C-[DE]-R-[FY]2-Q

Regular expressions

Προβλήματα

Για να μειωθεί το ποσοστό των false positives τα regexps αγνοούν την ομοιότητα μεταξύ καταλοίπων και απαιτούν την ύπαρξη ταυτότητας μεταξύ των αλληλουχιών και του regexp. Για παράδειγμα, οι κάτωθι δύο αλληλουχίες δεν θα ανιχνεύονταν από το regexp του προηγούμενου παραδείγματος.

AELGRTQNRCDRYYQ
ADLGAAVFALCDRYFQ

[AS]-D-[IVL]-G-x4-{PG}-C-[DE]-R-[FY]2-Q

Regular expressions

Προβλήματα

Το δυαδικό αποτέλεσμα από τη χρήση των regexps (θετικό ή αρνητικό, χωρίς κάποιας μορφής βαθμολόγηση), κάνει δύσκολη τη δημιουργία εκφράσεων που να αποδίδουν ικανοποιητικά στις έρευνες των βάσεων δεδομένων. Το αποτέλεσμα είναι η μειωμένη διαγνωστική τους αξία για το χαρακτηρισμό νέων αλληλουχιών. Επιπλέον, η συντήρηση των regexps είναι δύσκολη (με την προσθήκη νέων αλληλουχιών η αρχική στοίχιση αλλάζει και πρέπει να επαναπροσδιοριστούν).

Regular expressions

Προβλήματα

Η χρήση ενός μόνο μοτίβου για το χαρακτηρισμό μιας ολόκληρης οικογένειας είναι εφικτή μόνο για οικογένειες οι οποίες έχουν μια υψηλά συντηρημένη περιοχή με έκταση 10-20 κατάλοιπα. Ο λόγος είναι ότι καθώς το μήκος του regexp μειώνεται, αυξάνεται η πιθανότητα να βρεθεί το μοτίβο (τυχαία) σε μη συσχετιζόμενες αλληλουχίες.

Παράδειγμα βάσης

PROSITE patterns

Είναι μια δευτερογενής flat-file βάση, με την ιδιαιτερότητα ότι σε κάθε καταχώρηση αντιστοιχούν δύο αρχεία. Το πρώτο αρχείο περιέχει μια αναλυτική περιγραφή του μοτίβου, μαζί με λεπτομέρειες για τη βιολογική του σημασία και σχετικές βιβλιογραφικές αναφορές. Το δεύτερο αρχείο περιέχει το καθ'αυτό μοτίβο.

Παραδείγματα : Prosite (doc)

```
{PDOC00211}
{PS00238; OPSIN}
{BEGIN}
**************************************************
* Visual pigments (opsins) retinal binding site *
**************************************************
```

Visual pigments [1,2] are the light-absorbing molecules that mediate vision. They consist of an apoprotein, opsin, covalently linked to the chromophore cis-retinal. Vision is effected through the absorption of a photon by cis-retinal which is isomerized to trans-retinal. This isomerization leads to a change of conformation of the protein. Opsins are integral membrane proteins with seven transmembrane regions that belong to family 1 of G-protein coupled receptors (see <PDOC00210>).

In vertebrates four different pigments are generally found. Rod cells, which mediate vision in dim light, contain the pigment rhodopsin. Cone cells, which function in bright light, are responsible for color vision and contain three or more color pigments (for example, in mammals: red, blue and green).

In Drosophila, the eye is composed of 800 facets or ommatidia. Each ommatidium contains eight photoreceptor cells (R1-R8): the R1 to R6 cells are outer cells, R7 and R8 inner cells. Each of the three types of cells (R1-R6, R7 and R8) expresses a specific opsin.

Παραδείγματα : Prosite (doc)

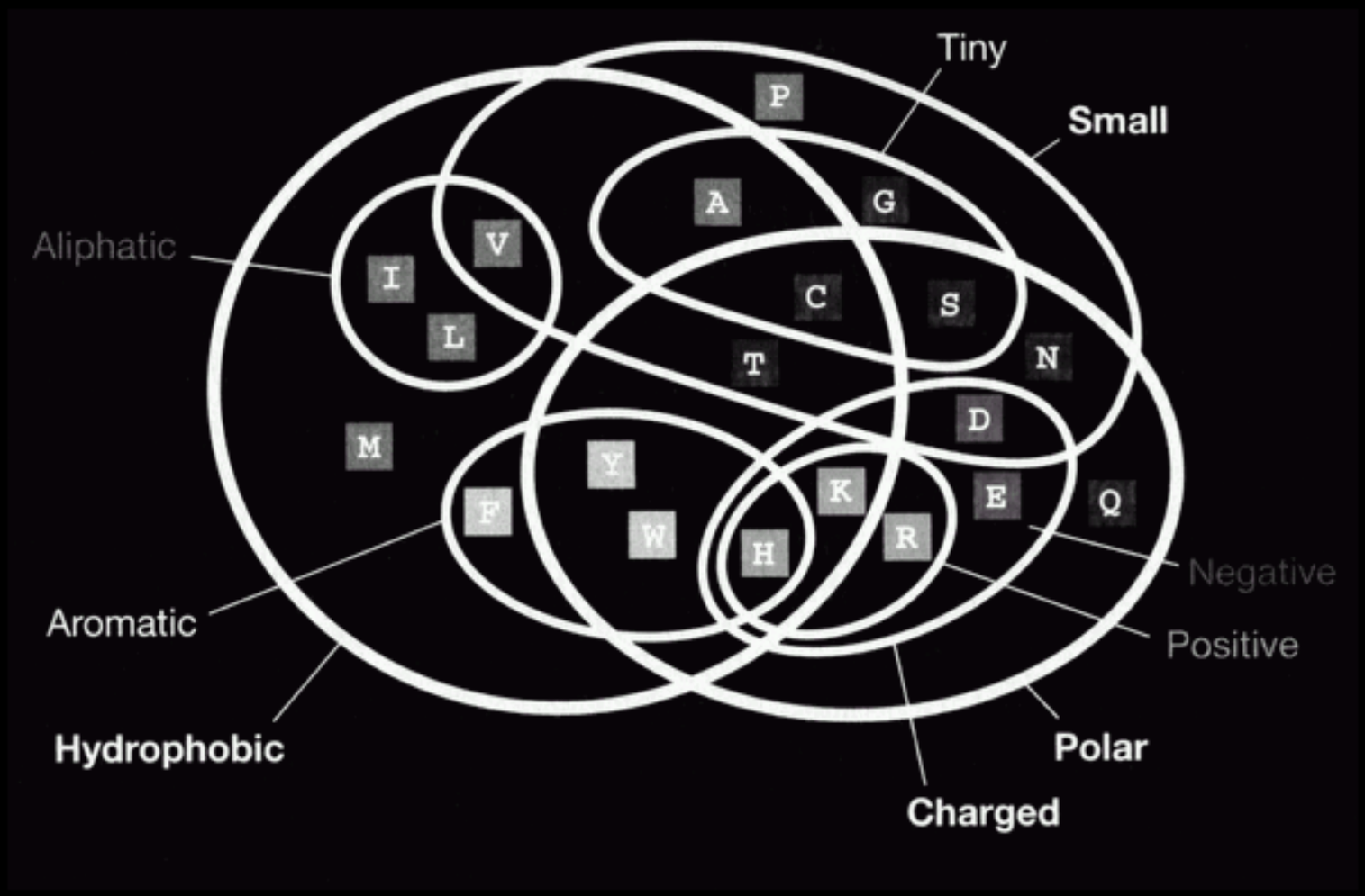
```
Proteins evolutionary related to opsins include:

 - Squid retinochrome, also known as retinal photoisomerase, which converts
   various isomers of retinal into 11-cis retinal.
 - Mammalian opsin 3 (Encephalopsin) that may play a role in encephalic
   photoreception.
 - Mammalian opsin 4 (Melanopsin) that may mediate regulation of circadian
   rhythms and acute suppression of pineal melatonin.
 - Mammalian retinal pigment epithelium (RPE) RGR [3], a protein that may also
   act in retinal isomerization.

The attachment site for retinal in the above proteins is a conserved lysine
residue in the middle of the seventh transmembrane helix. The pattern we
developed includes this residue.

-Consensus pattern: [LIVMFWAC]-[PSGAC]-x(3)-[SAC]-K-[STALIMR]-[GSACPNV]-
                    [STACP]-x(2)-[DENF]-[AP]-x(2)-[IY]
                    [K is the retinal binding site]
-Sequences known to belong to this class detected by the pattern: ALL.
-Other sequence(s) detected in Swiss-Prot: 2.
-Last update: December 2001 / Pattern and text revised.

[ 1] Applebury M.L., Hargrave P.A.
     Vision Res. 26:1881-1895(1986).
[ 2] Fryxell K.J., Meyerowitz E.M.
     J. Mol. Evol. 33:367-378(1991).
[ 3] Shen D., Jiang M., Hao W., Tao L., Salazar M., Fong H.K.W.
     Biochemistry 33:13117-13125(1994).

{END}
```

Παραδείγματα : Prosite (regexp)

```
ID   OPSIN; PATTERN.
AC   PS00238;
DT   APR-1990 (CREATED); DEC-2001 (DATA UPDATE); DEC-2001 (INFO UPDATE).
DE   Visual pigments (opsins) retinal binding site.
PA   [LIVMFWAC]-[PSGAC]-x(3)-[SAC]-K-[STALIMR]-[GSACPNV]-[STACP]-x(2)-[DENF]-
PA   [AP]-x(2)-[IY].
NR   /RELEASE=41.25,134803;
NR   /TOTAL=193(192); /POSITIVE=189(188); /UNKNOWN=0(0); /FALSE_POS=4(4);
NR   /FALSE_NEG=1; /PARTIAL=4;
CC   /TAXO-RANGE=??E??; /MAX-REPEAT=1;
CC   /SITE=5,retinal;
DR   Q9H1Y3, OPN3_HUMAN, T; Q9WUK7, OPN3_MOUSE, T; Q9UHM6, OPN4_HUMAN, T;
DR   Q9QXZ9, OPN4_MOUSE, T; P22269, OPS1_CALVI, T; P06002, OPS1_DROME, T;
DR   P28678, OPS1_DROPS, T; Q25157, OPS1_HEMSA, T; P35360, OPS1_LIMPO, T;
DR   O15973, OPS1_PATYE, T; Q94741, OPS1_SCHGR, T; P08099, OPS2_DROME, T;
.....................................................................
DR   O14718, OPSX_HUMAN, T; O35214, OPSX_MOUSE, T; P23820, REIS_TODPA, T;
DR   P47803, RGR_BOVIN , T; P47804, RGR_HUMAN , T;
DR   P17645, OPS3_DROVI, P; O18911, OPSG_ODOVI, P; O18914, OPSR_CANFA, P;
DR   O18912, OPSR_HORSE, P;
DR   Q9Z2B3, RGR_MOUSE , N;
DR   Q9CL24, OADB_PASMU, F; P22056, POLS_ONNVG, F; Q99NF8, RP17_MOUSE, F;
DR   P09009, TERM_BPPRD, F;
3D   1BOJ; 1BOK; 1F88; 1HZX; 1JFP; 1KPN; 1KPW; 1KPX; 1LN6;
DO   PDOC00211;
//
```

Fuzzy regular expressions

Ένας τρόπος μείωσης των προβλημάτων με την αυστηρότητα των regexps είναι μέσω της χρήσης διευρυμένων εκφράσεων οι οποίες να λαμβάνουν υπόψη τους π.χ. τις φυσικοχημικές ιδιότητες των καταλοίπων.



Fuzzy regular expressions

ADLGAVFALCDRYFQ
SDVGPRSCFCERFYQ
ADLGRTQNRCDRYYQ
ADIGQPHSLCERYFQ

[AS]-D-[IVL]-G-x4-{PG}-C-[DE]-R-[FY]2-Q

[ASGT]-D-[IVLM]-G-x5-C-[DENQ]-R-[FYW]2-Q

Fuzzy regular expressions

Το πρόβλημα με αυτού του τύπου τις εκφράσεις έχει ήδη αναφερθεί : καθώς η γενικότητα της έκφρασης αυξάνει, αυξάνουν και τα ψευδώς θετικά αποτελέσματα. Παράδειγμα :

Έκφραση	Hits
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]-[AG]-[VLI]-I-[DENQ]	51506

Fingerprints

Συχνά περισσότερα από ένα μοτίβα είναι συντηρημένα σε μια οικογένεια. Η χρήση όλων των συντηρημένων μοτίβων μιας οικογένειας οδηγεί στη δημιουργία ενός αποτυπώματος χαρακτηριστικού για την οικογένεια. Η ευαισθησία της μεθόδου μπορεί να αυξηθεί με τη χρήση πινάκων συχνότητας εμφάνισης καταλοίπων (αντί για χρήση regular expressions). Οι πίνακες αυτοί βελτιστοποιούνται μέσω διαδοχικών ερευνών των βάσεων και ενσωμάτωση καινούργιων αλληλουχιών. Η γνωστότερη βάση δεδομένων για fingerprints είναι η PRINTS.

Fingerprints

	C	A	G	N	S	P	F	L	Y	H	Q	V	K	D	E	I	W	R	M
YVTVQH**K**KLRT**P**L	0	2	0	0	0	0	0	0	7	0	0	1	0	0	0	0	2	0	0
YVTVQH**K**KLRT**P**L	0	3	0	0	0	0	0	2	0	0	0	4	0	0	0	3	0	0	0
YVTVQH**K**KLRT**P**L	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0
AATMKF**K**K**L**RH**P**L	0	1	1	0	3	0	0	1	0	0	0	3	0	0	0	0	0	0	2
AATMKF**K**K**L**RH**P**L	0	1	1	0	0	0	0	0	0	0	3	0	4	0	0	0	0	1	0
YIFATT**K**S**L**RT**P**A	0	1	0	0	1	0	2	0	1	3	0	0	0	0	0	0	0	0	0
VATLRY**K**K**L**RQ**P**L	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0
YIFGGT**K**S**L**RT**P**A	0	0	0	0	6	0	0	0	0	0	0	0	6	0	0	0	0	0	0
WVFSAA**K**S**L**RT**P**S	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0
WIFSTS**K**S**L**RT**P**S	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	10	0
YLFSKT**K**S**L**QT**P**A	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0
YLFTKT**K**S**L**QT**P**A	0	0	0	0	0	12	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

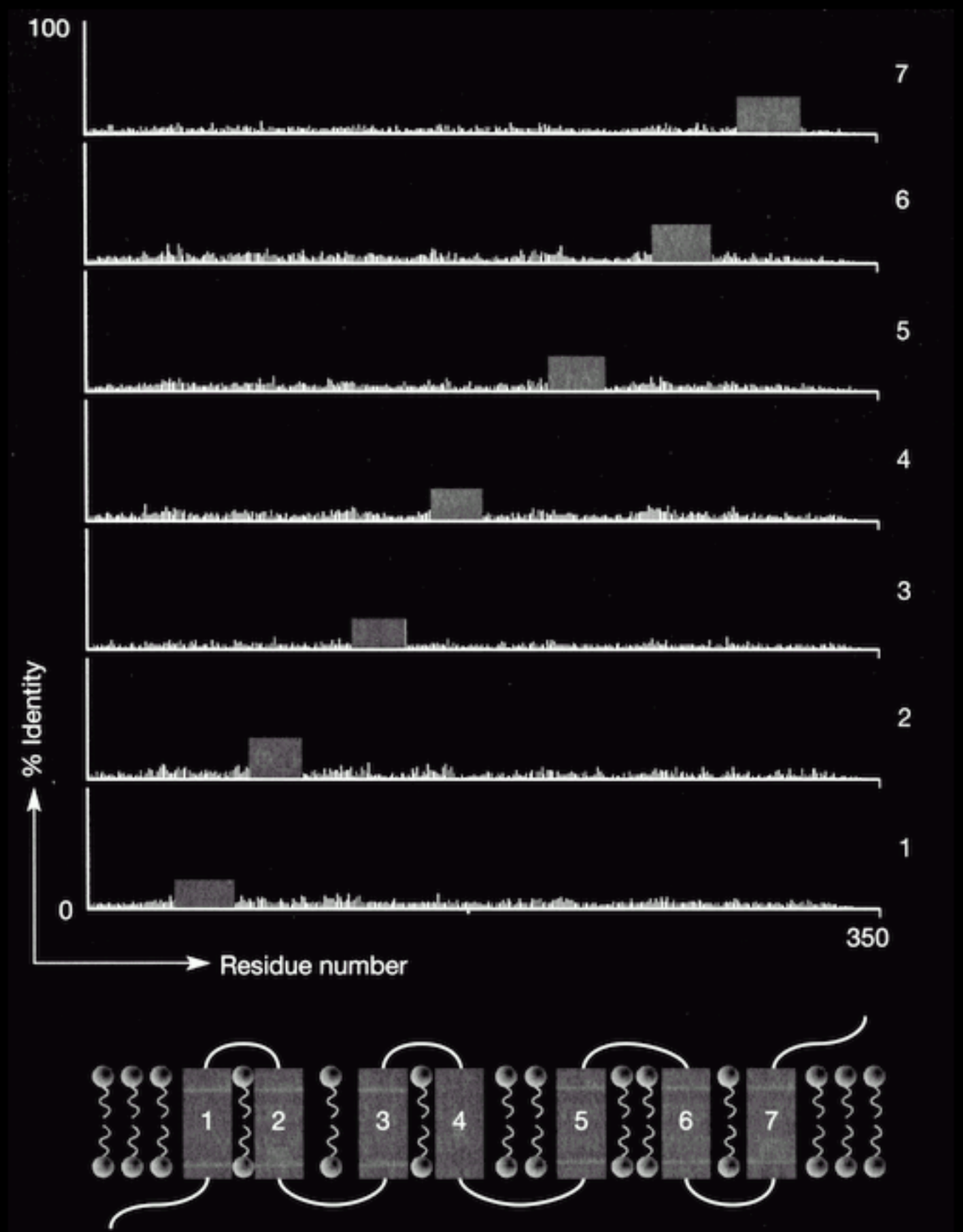
Fingerprints

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

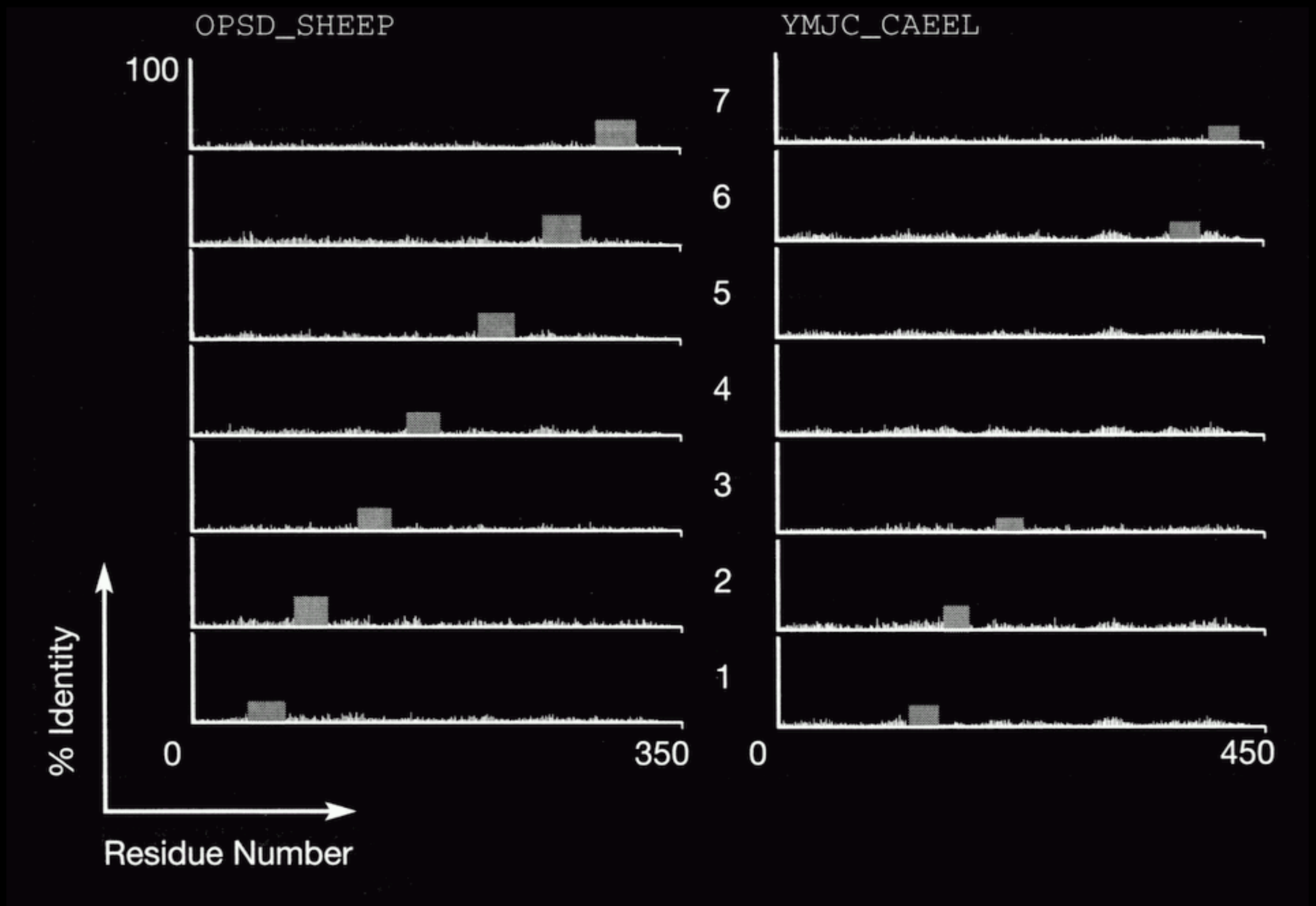
T	C	A	G	N	S	P	F	L	Y	H	Q	V	K	D	E	I	W	R	M
0	0	4	0	0	0	0	8	4	34	0	0	15	0	0	0	1	7	0	0
0	4	15	0	0	0	0	0	7	0	0	0	37	0	0	0	10	0	0	0
50	0	0	0	0	3	0	18	0	0	0	0	0	0	0	0	0	0	0	2
3	0	12	2	1	8	0	3	6	0	0	0	14	0	0	0	15	2	0	7
9	2	2	2	1	1	0	0	0	0	1	25	0	20	0	6	0	0	4	0
14	0	2	0	0	4	0	14	0	8	31	0	0	0	0	0	0	0	0	0
0	0	1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

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

Fingerprints



Fingerprints



Blocks

Η βασική ιδέα της βάσης BLOCKS είναι παρόμοια με αυτή της PRINTS. Η βασική διαφορά έγκειται στο ότι στην περίπτωση της BLOCKS τα μοτίβα δεν έχουν τη μορφή πινάκων συχνοτήτων καταλοίπων, αλλά στοιχισμένων μοτίβων τα οποία έχουν βρεθεί με τη χρήση πινάκων υποκατάστασης (BLOSUM62). Η ταυτοποίηση νέων αλληλουχιών γίνεται (όπως και για την PRINTS) μέσω αναζήτησης διαδοχικών μοτίβων κατά μήκος της νέας αλληλουχίας τα οποία να προέρχονται από την ίδια οικογένεια (και με την ίδια σειρά).

Blocks

```
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
```

Hidden Markov Models

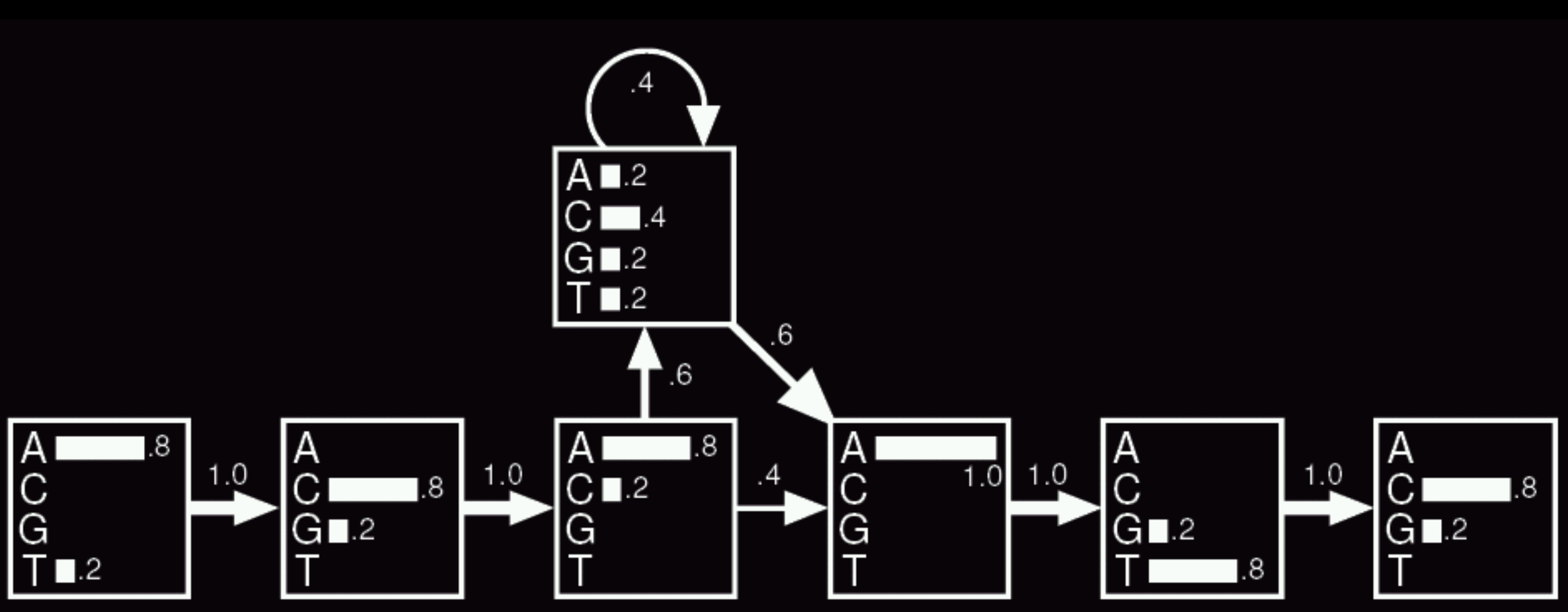
Παράδειγμα

```
A C A - - - A T G
T C A A C T A T C
A C A C - - A G C
A G A - - - A T C
A C C G - - A T C
```

Hidden Markov Models

Παράδειγμα

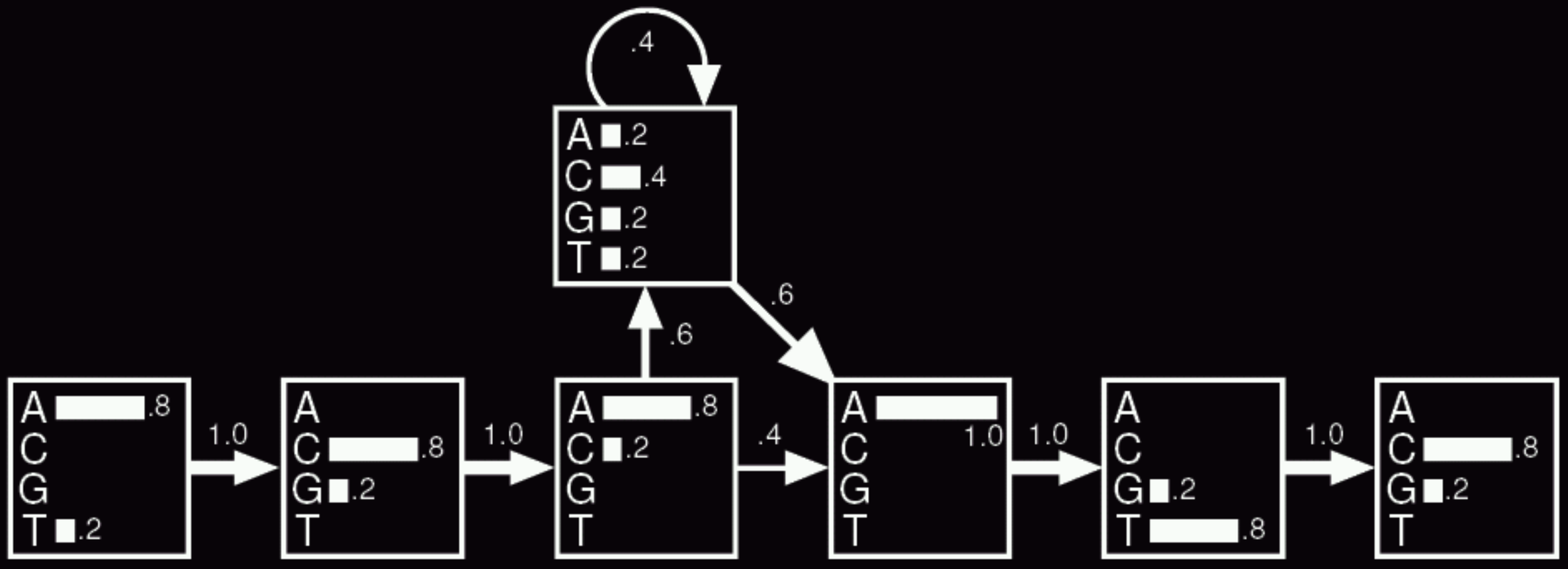
```
A C A - - - A T G
T C A A C T A T C
A C A C - - A G C
A G A - - - A T C
A C C G - - A T C
```



Hidden Markov Models

Παράδειγμα

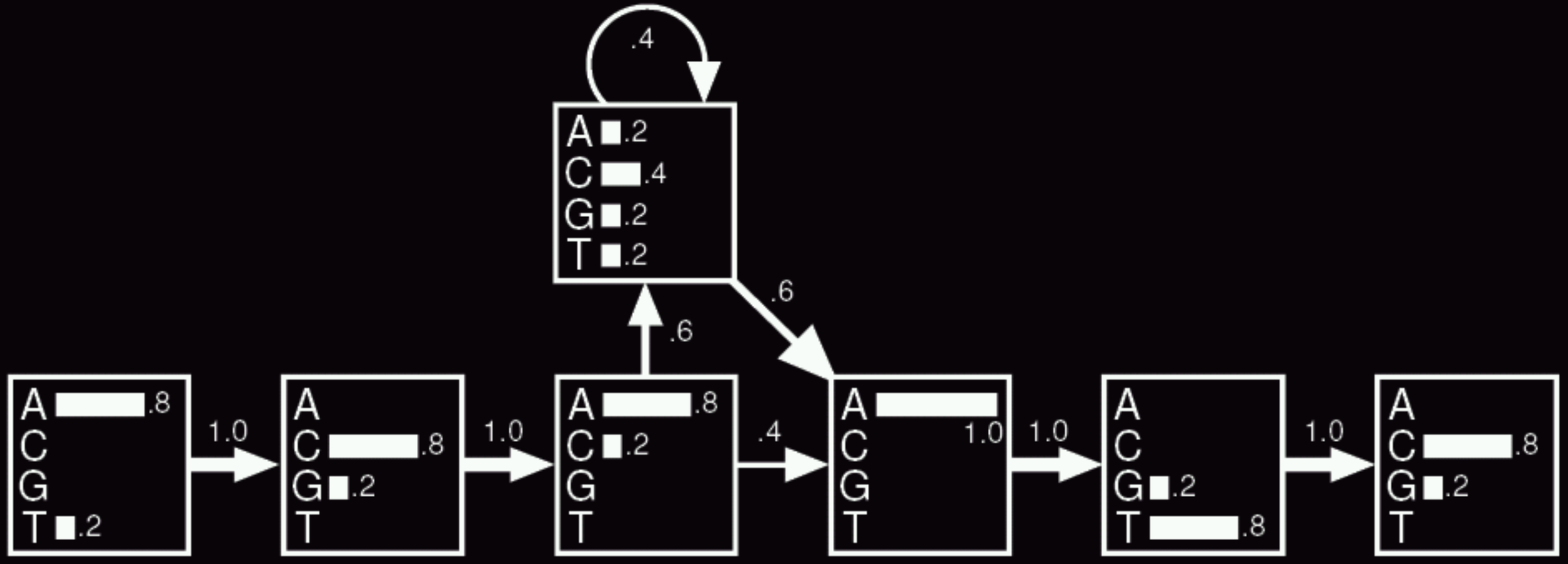
P(ACACATC) = 0.8 x 1.0 x 0.8 x 1.0
x 0.8 x 0.6 x 0.4 x 0.6
x 1.0 x 1.0 x 0.8 x 1.0
x 0.8 = 0.047



Hidden Markov Models

Παράδειγμα

P(TGCTAGG) = 0.2 x 1.0 x 0.2 x 1.0
x 0.2 x 0.6 x 0.2 x 0.6
x 1.0 x 1.0 x 0.2 x 1.0
x 0.2 = 0.000023



Hidden Markov Models

Log-odds

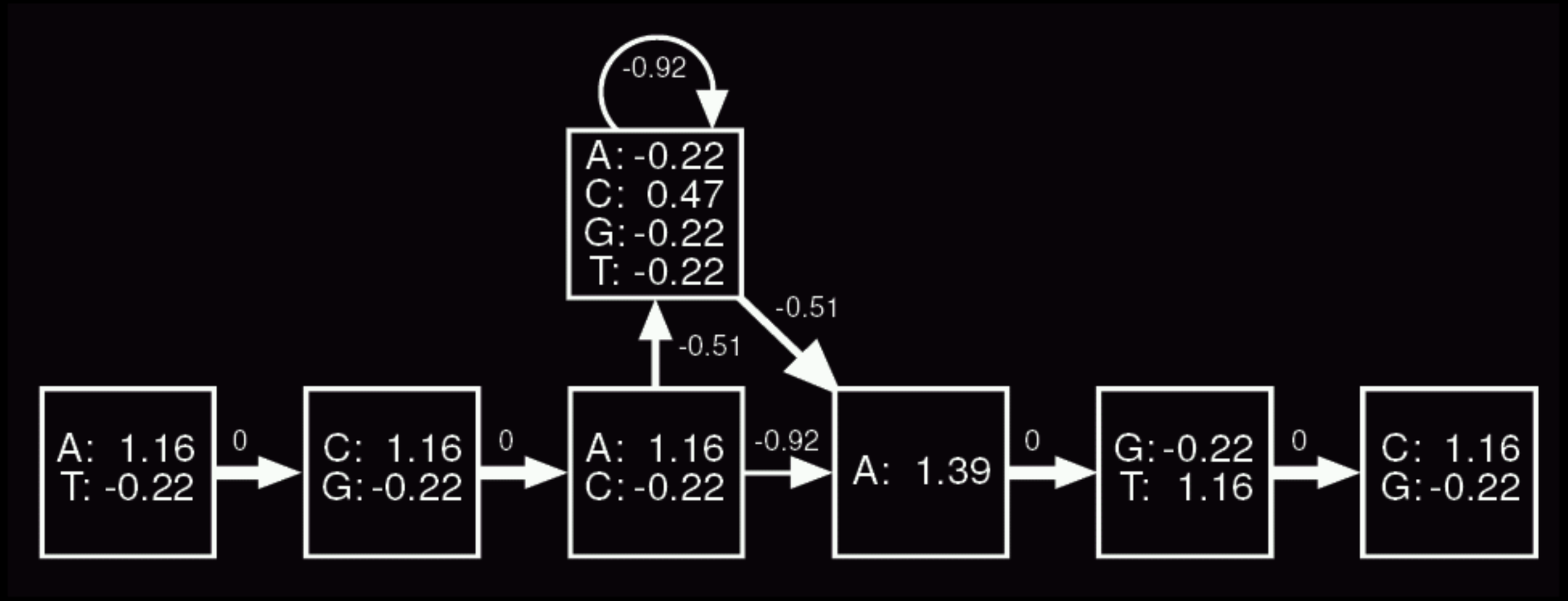
Η χρήση της πιθανότητας ως μέσο βαθμολόγησης έχει το πρόβλημα ότι εξαρτάται από το μήκος των αλληλουχιών. Το πρόβλημα λύνεται με τη χρήση των log-odds : Είναι ο λογάριθμος (φυσικός ή με βάση το 2 ή ...) του πηλίκου της πιθανότητας που υπολογίζουμε από το HMM δια της πιθανότητας του μηδενικού μοντέλου. Το μηδενικό μοντέλο είναι μία τυχαία αλληλουχία καταλοίπων με το ίδιο μήκος. Για παράδειγμα, για μια DNA αλληλουχία S με μήκος L βάσεων :

log-odds(S) = log [P(S) / 0.25^L]
= log P(S) - L log 0.25

Hidden Markov Models

Log-odds

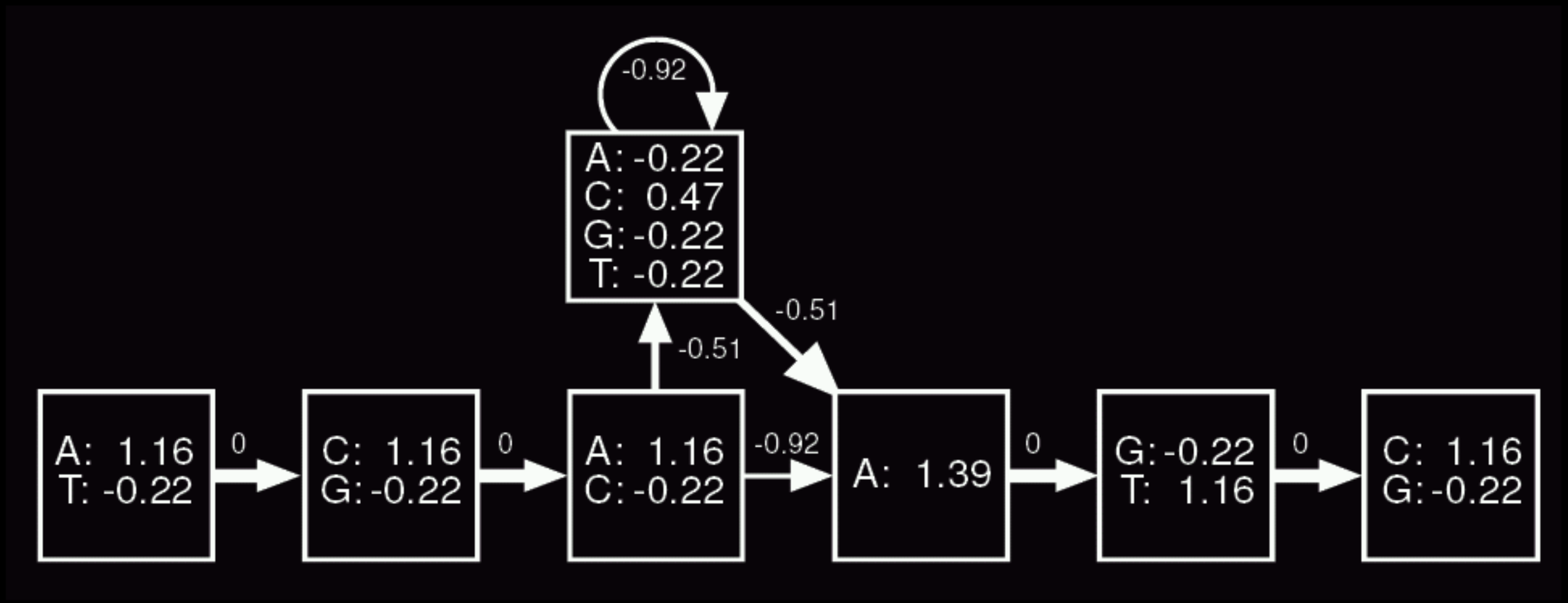
Η χρήση των log-odds μπορεί να επεκταθεί και σε καθ'αυτό το HMM ώστε αντί να πολλαπλασιάζουμε πιθανότητες, να προσθέτουμε log-odds :



Hidden Markov Models

Παράδειγμα

P(TGCTAGG) = -0.22 - 0.22 -0.22 -0.51
-0.22 -0.51 + 1.39 -0.22
-0.22 = -0.95



Hidden Markov Models

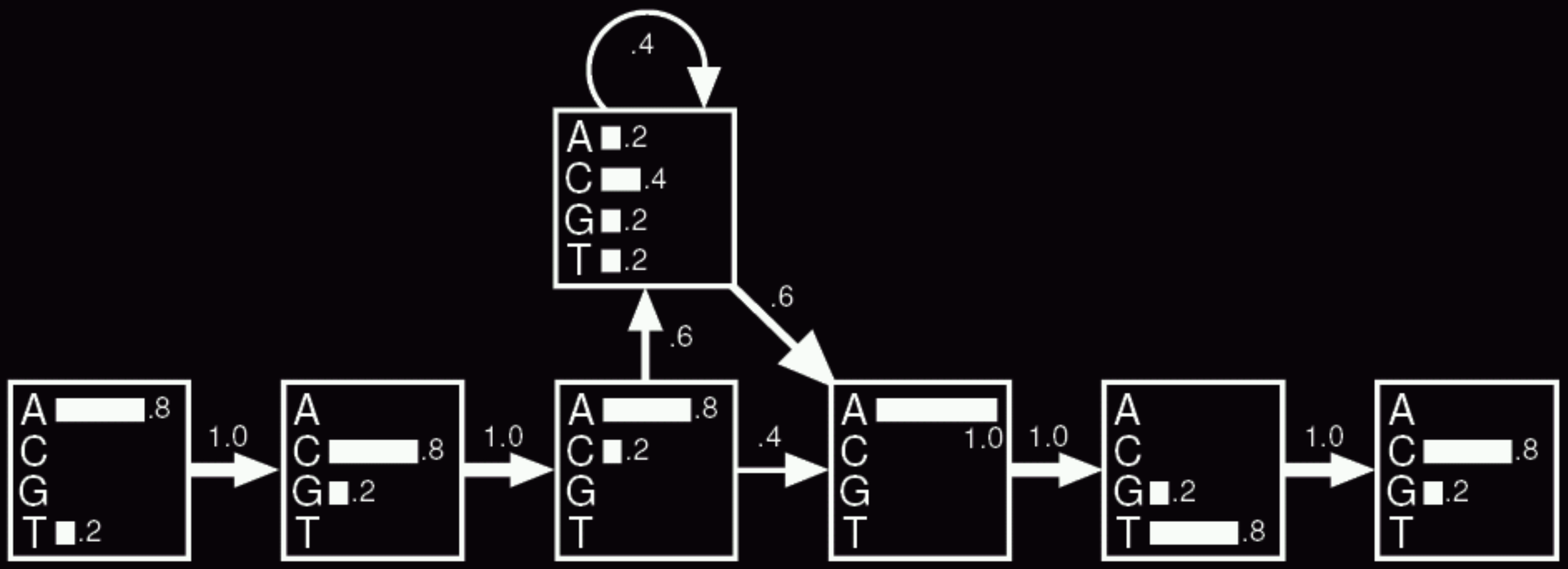
Παράδειγμα

Αλληλουχία	100·P	Log
A C A - - - A T G	3.3	4.9
T C A A C T A T C	0.0075	3.0
A C A C - - A G C	1.2	5.3
A G A - - - A T C	3.3	4.9
A C C G - - A T C	0.59	4.6
A C A C - - A T C	4.7	6.7
T G C T - - A G G	0.0023	-0.97

[AT]-[CG]-[AC]-[ACGT]*-A-[TG]-[GC]

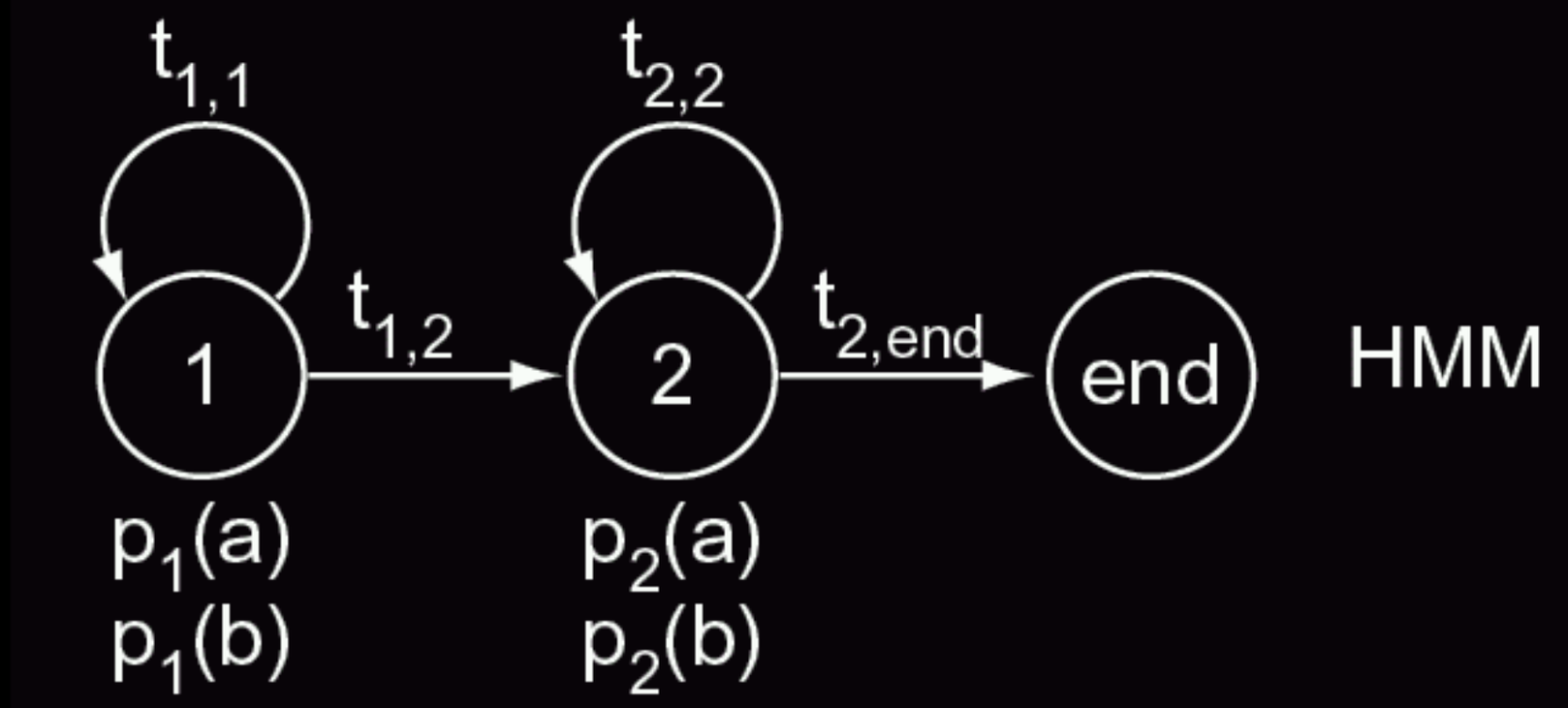
Hidden Markov Models

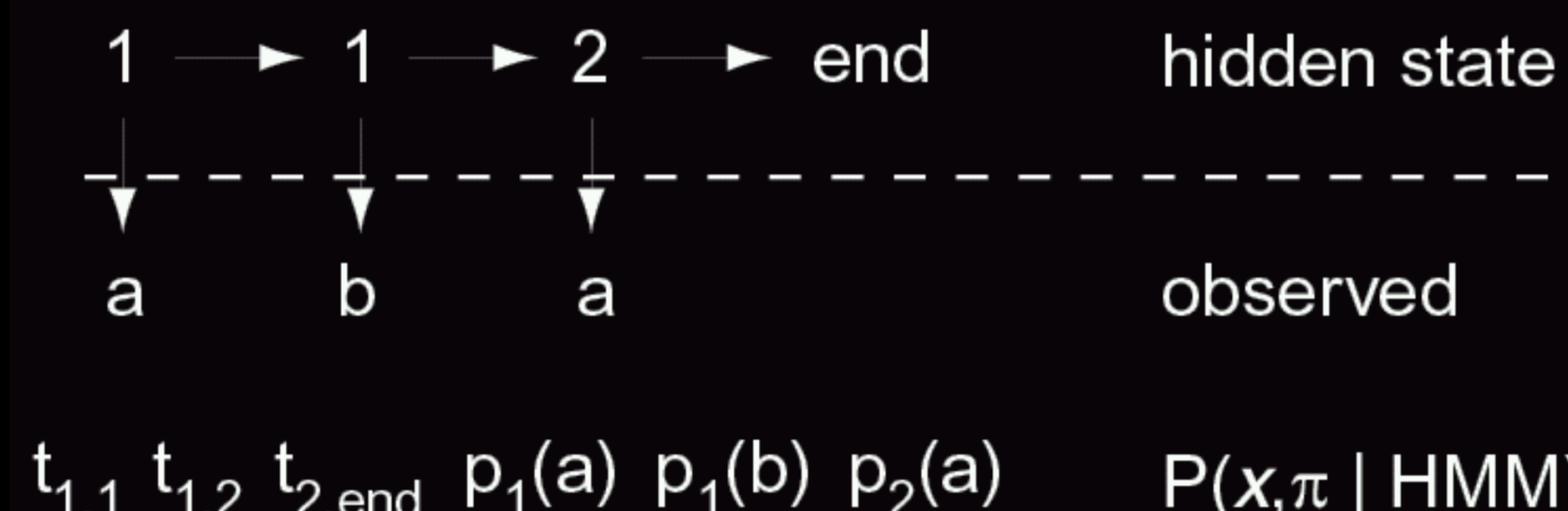
Μια 'στοχαστική' μηχανή.



Hidden Markov Models

Μια 'στοχαστική' μηχανή.

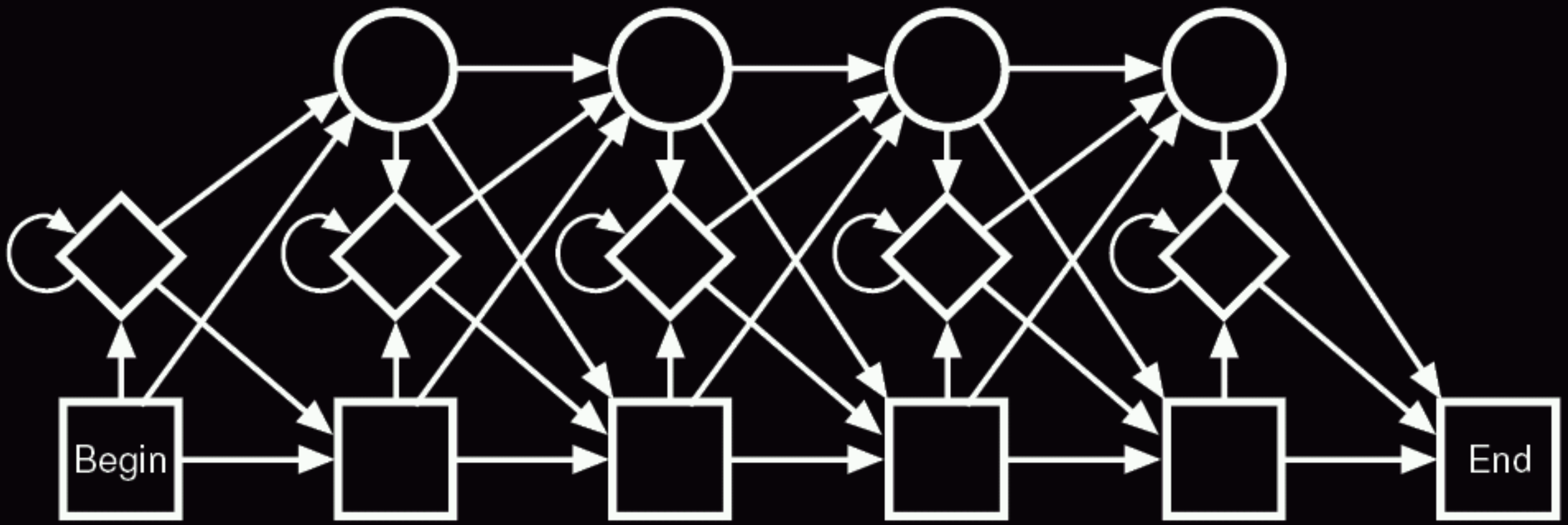




$t_{1,1}$ $t_{1,2}$ $t_{2,end}$ $p_1(a)$ $p_1(b)$ $p_2(a)$ $P(x,\pi \mid HMM)$

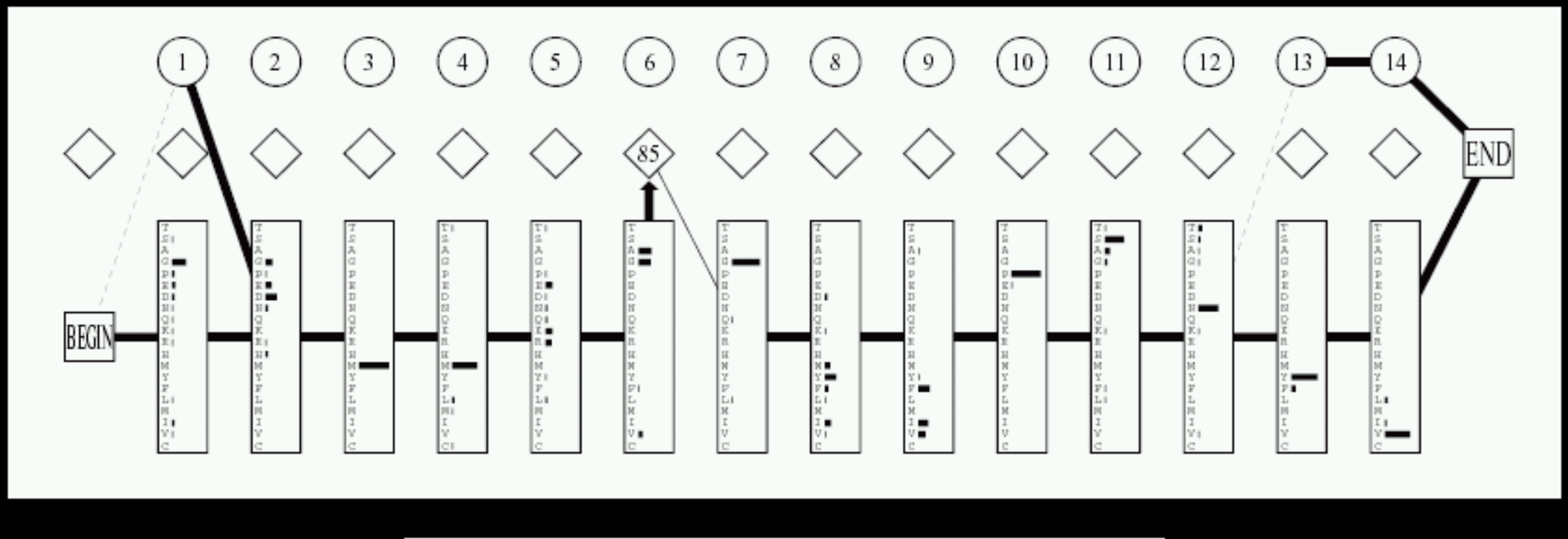
Profile HMMs

Είναι ένα HMM με δομή τέτοια ώστε να επιτρέπει (και να προβλέπει) γεγονότα εισαγωγών/διαγραφών. Χαρακτηρίζεται από τρεις πιθανές καταστάσεις (states) : τις κυρίως, αυτές των εισαγωγών, και αυτές των διαγραφών :

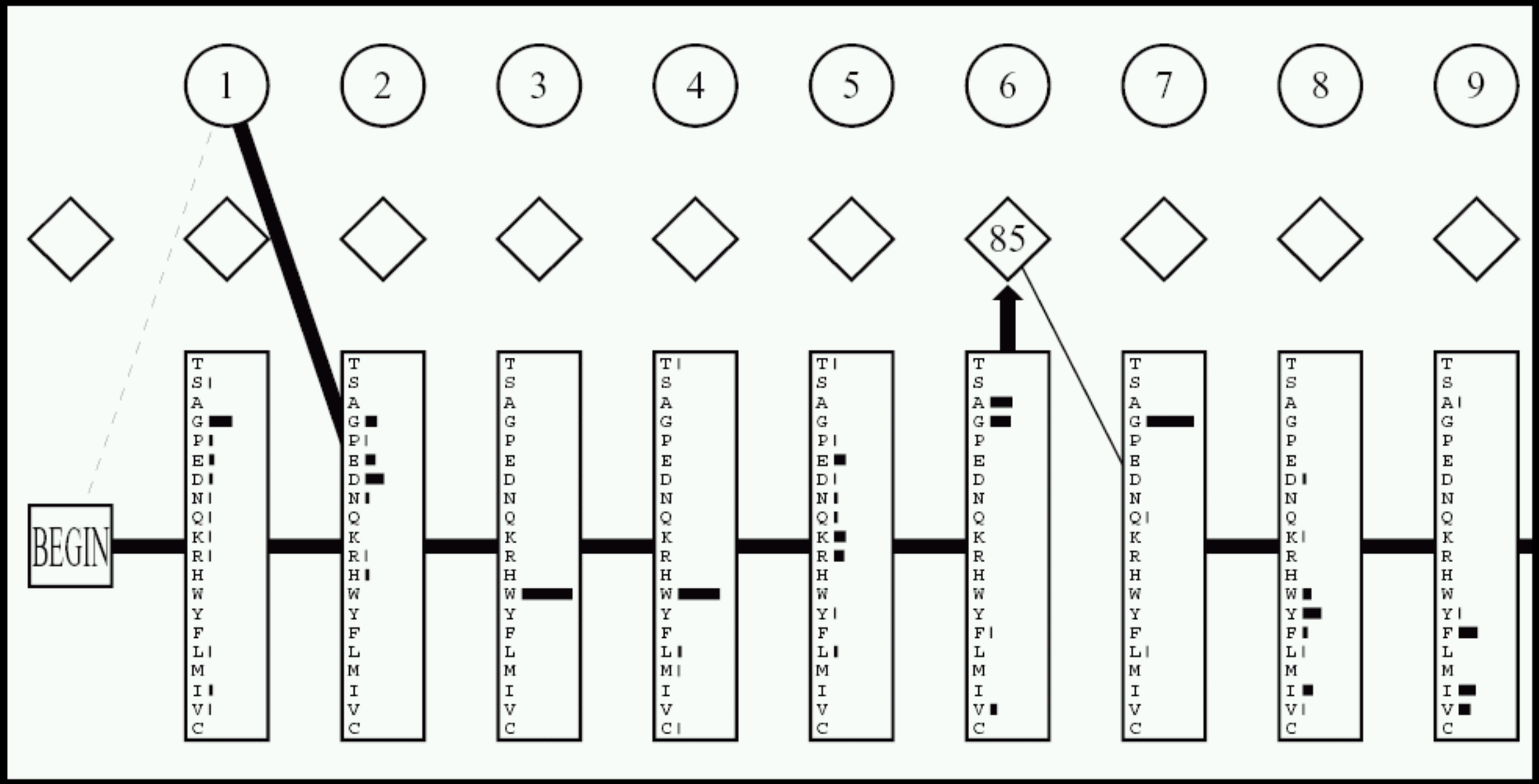


Profile HMMs

Profile HMMs

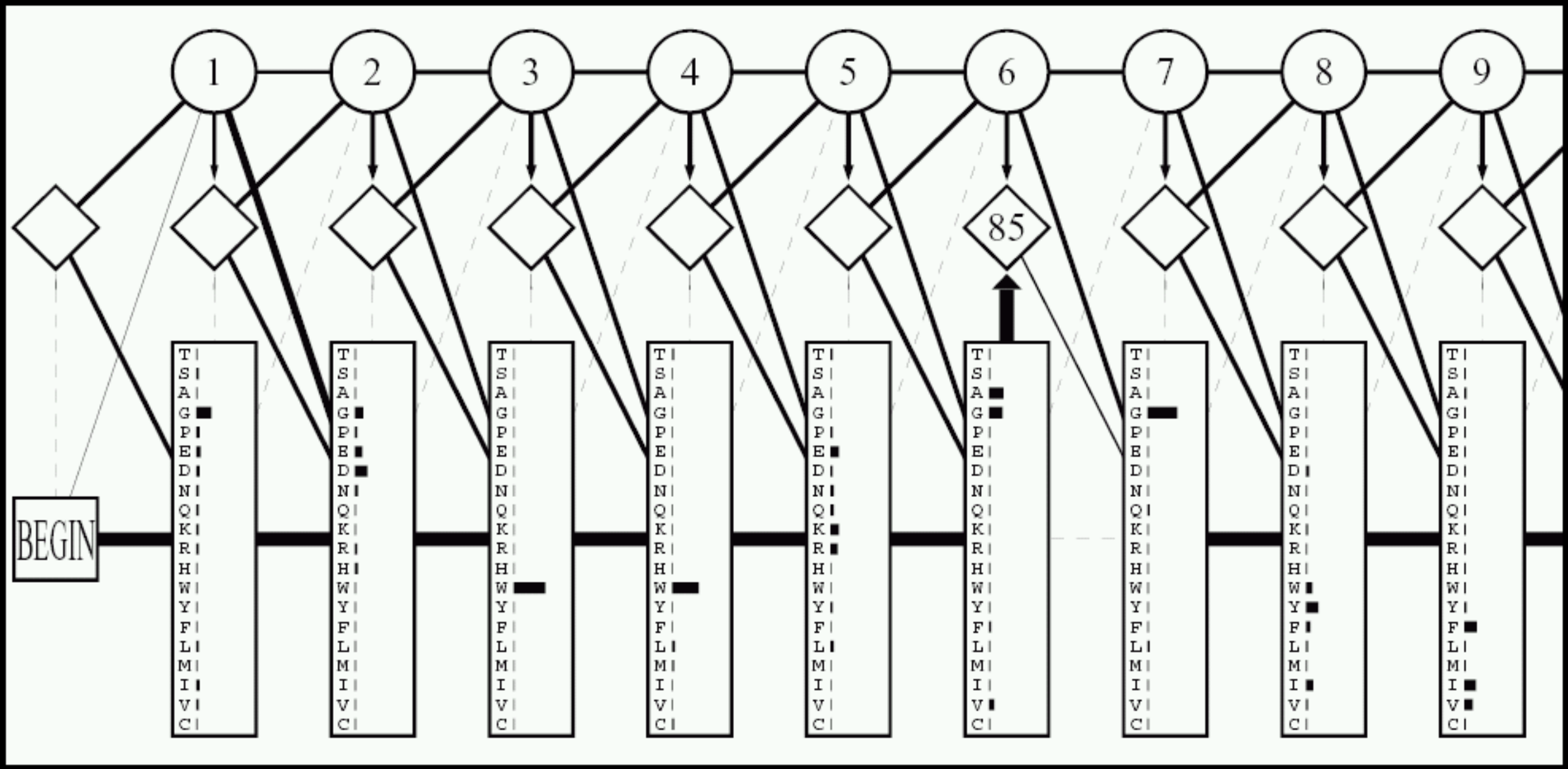


Profile HMMs



Profile HMMs

Pseudocounts



Profile HMMs

Χρήση

Η χρήση των HMMs για ταυτοποίηση ομολογιών νέων αλληλουχιών είναι διαισθητικά προφανής : με δεδομένο ένα HMM και μια νέα αλληλουχία, επιδιώκουμε να τα στοιχίσουμε έτσι ώστε η πιθανότητα (ή το log-odds) που προκύπτει από την εφαρμογή του HMM στην αλληλουχία να μεγιστοποιείται. Η εκφώνηση του προβλήματος θυμίζει τον αλγόριθμο των N&W, και όντως η επίλυση του γίνεται με τον αλγόριθμο του Viterbi, έναν αλγόριθμο δυναμικού προγραμματισμού. Ένας παρόμοιος αλγόριθμος (γνωστός με το όνομα 'Forward algorithm') βρίσκει το άθροισμα των πιθανοτήτων των στοιχίσεων του HMM με την αλληλουχία.

Profile HMMs

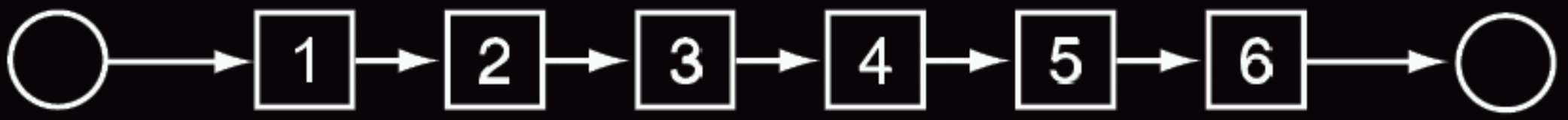
Βάσεις δεδομένων

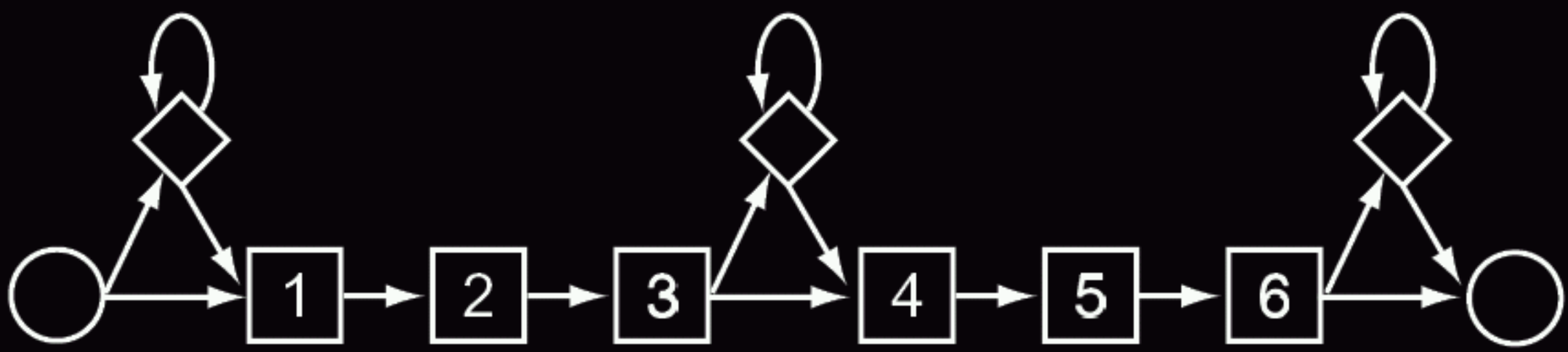
Οι δύο πλέον γνωστές βάσεις HMMs (με τις αντίστοιχες τους στοιχίσεις) είναι η PROSITE Profiles και η Pfam.
Η PROSITE έχει ήδη αναφερθεί για τα μοτίβα (regexps) που περιέχει. Για οικογένειες που τα regexps δεν αρκούν, η βάση συντηρεί και ένα τμήμα με HMM profiles (ενώ υπάρχουν και περιπτώσεις που τα προφίλ και τα regexp συνυπάρχουν για την ίδια οικογένεια).
Ένα παράδειγμα καταχώρησης από την PROSITE Profiles είναι :

PROSITE profiles

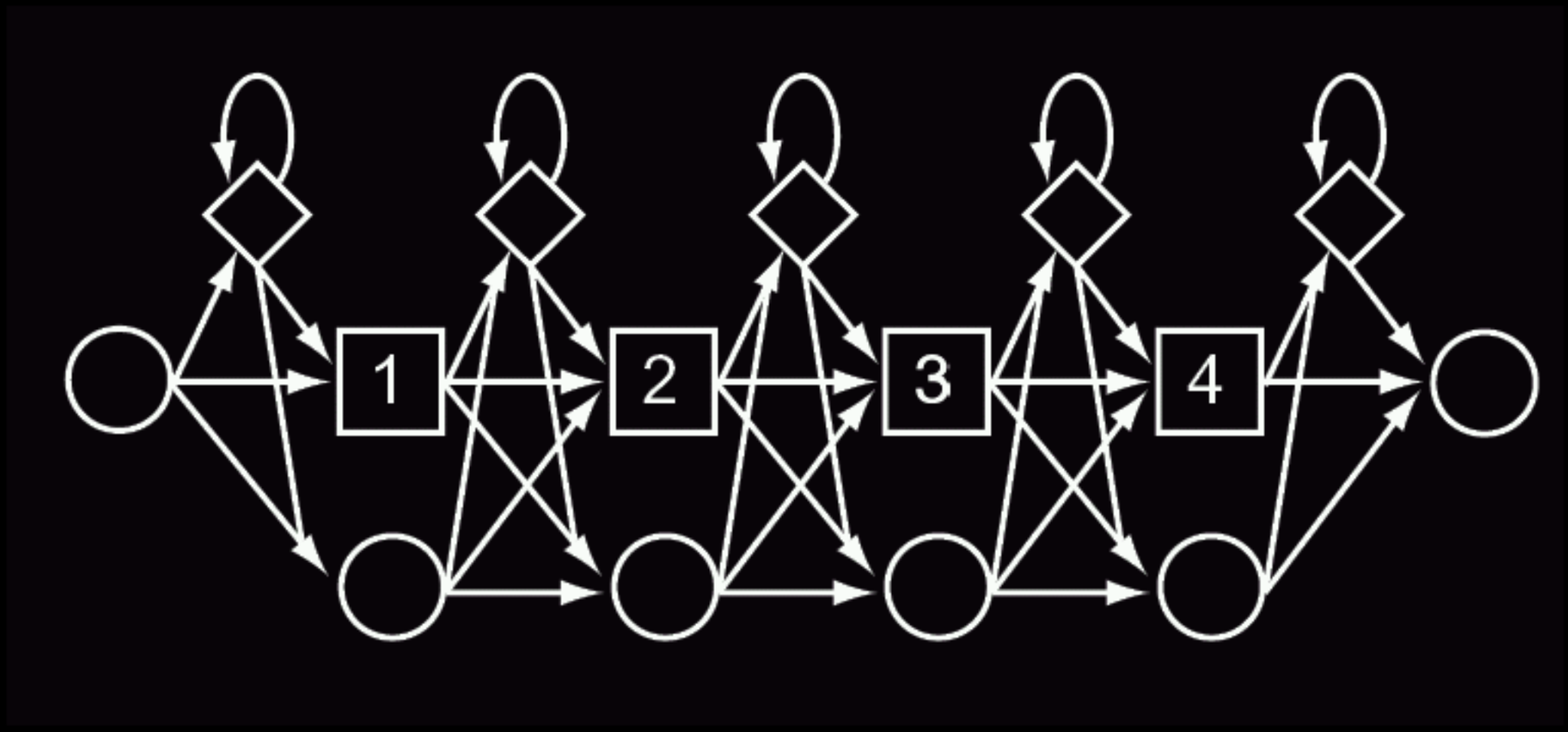
```
/DEFAULT: MI=-26; I=-3; IM=0; MD=-26; D=-3; DM=0;
/M: SY='F';M=-2,-3,-3,-4,2,-3,-2,1,-2,0,-1,-2,-3,-3,-4,-2,-1,0,-5,2;
/M: SY='I';M=-1,-5,-2,-3,-2,-3,0,1,1,-1,1,-1,-2,-1,1,-1,0,1,-4,-4;
/M: SY='A';M=2,-3,1,0,-5,2,-2,-1,-1,-3,-2,1,1,0,-2,2,2,0,-8,-5;
/M: SY='L';M=-3,-8,-5,-4,2,-6,-2,2,-4,6,4,-3,-3,-2,-3,-3,-2,1,-3,0;
/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;
```

Blocks, fingerprints & profiles

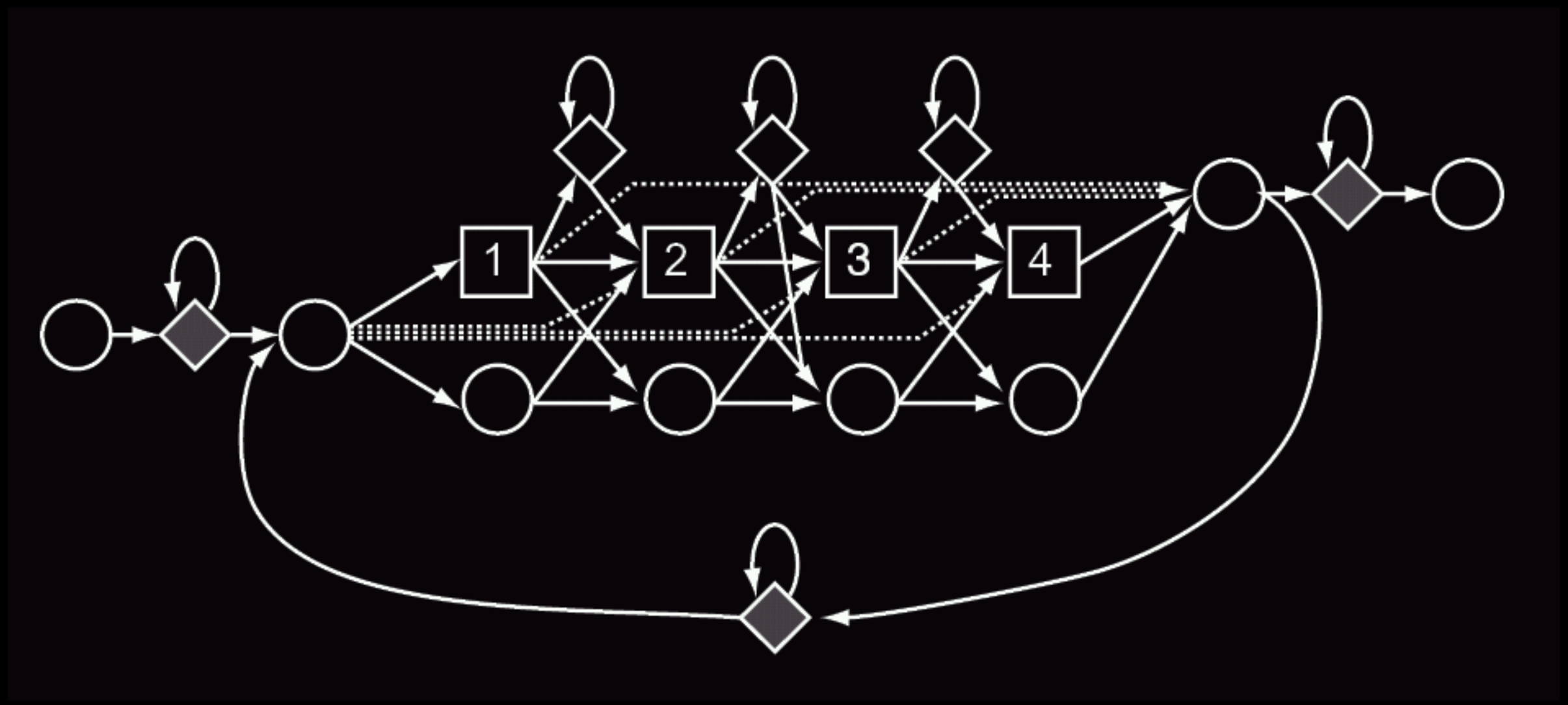




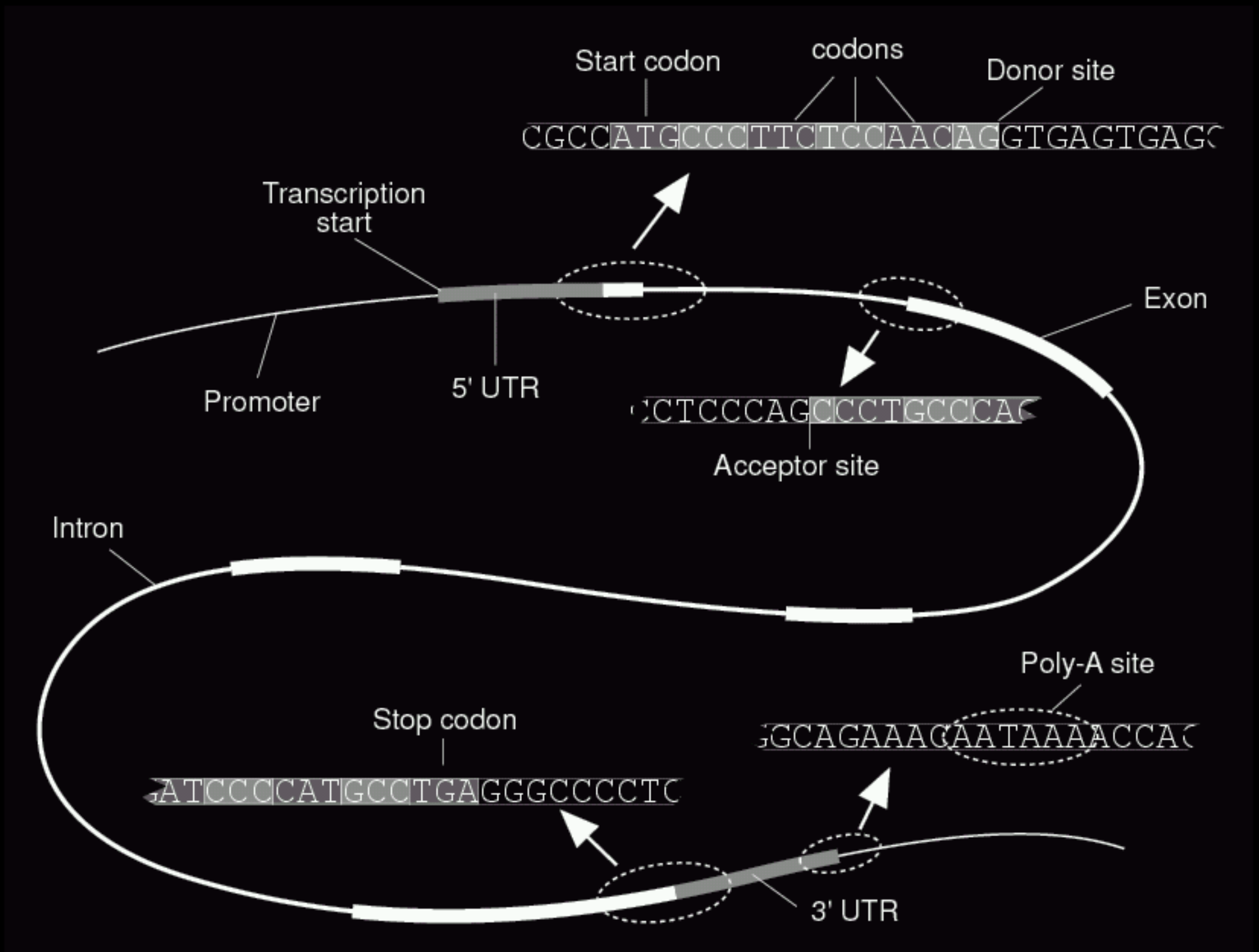
Blocks, fingerprints & profiles



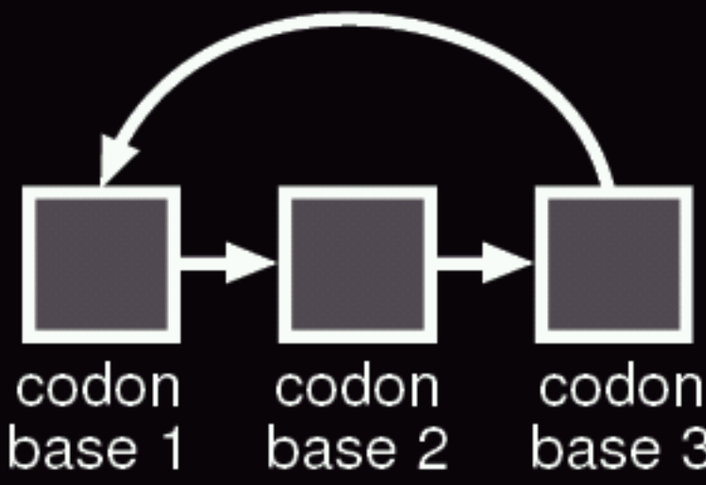
Blocks, fingerprints & profiles

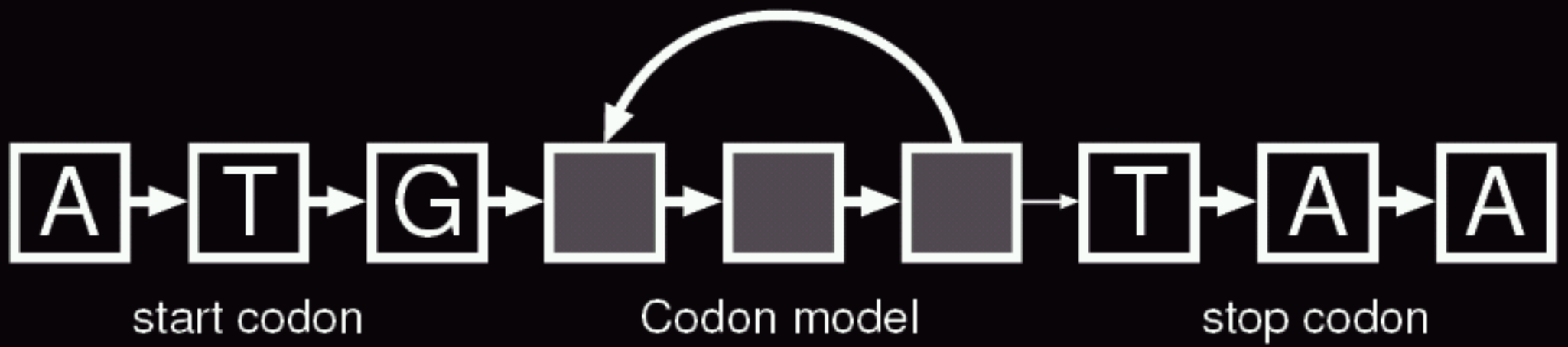


HMMs και εύρεση γονιδίων

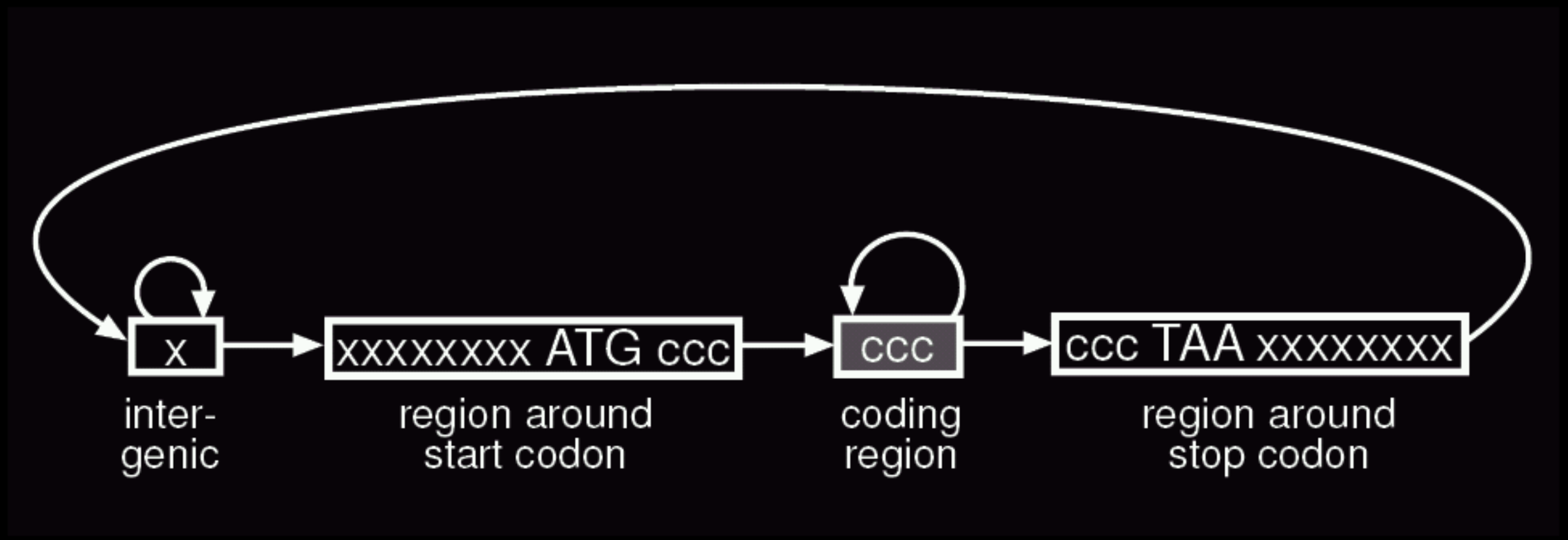


HMMs και εύρεση γονιδίων





HMMs και εύρεση γονιδίων



HMMs και εύρεση γονιδίων