

pHMMs - HMMER

Μαργαρίτα Θεοδωροπούλου
Πανεπιστήμιο Θεσσαλίας, Λαμία 2023

Μηχανική μάθηση (Machine Learning)



Μηχανική Μάθηση

“The field of study that gives computers the ability to learn without being explicitly programmed”

Arthur Samuel (1959)

*"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, **if** its performance at tasks in T, as measured by P, improves with experience E"*

Tom M. Mitchell (1997)

Αλγόριθμοι Μηχανικής Μάθησης

- **Artificial Neural Networks (ANNs)**
- Dynamic Bayesian Networks (DBNs)
- Support Vector Machines (SVMs)
- Decision Trees , Random Forests
- Conditional Random Fields
- **Hidden Markov Models (HMMs)**
- Hidden Neural Networks (HNNs)

Μηχανική Μάθηση

Επιβλεπόμενη Μάθηση (*Supervised Learning*)

Στη μάθηση με επίβλεψη το σύστημα καλείται να "μάθει" μια έννοια ή συνάρτηση από ένα σύνολο δεδομένων (*labeled data*), η οποία αποτελεί περιγραφή ενός μοντέλου.

Μη επιβλεπόμενη Μάθηση (*Unsupervised Learning*)

Στη μάθηση χωρίς επίβλεψη το σύστημα πρέπει μόνο του να ανακαλύψει συσχετίσεις ή ομάδες σε ένα σύνολο δεδομένων (*unlabeled data*), δημιουργώντας πρότυπα, χωρίς να είναι γνωστό αν υπάρχουν, πόσα και ποια είναι.

Μερικώς επιβλεπόμενη Μάθηση (*Semi-supervised Learning*)

Το σύστημα σε αυτή την περίπτωση καλείται να μάθει τόσο από γνωστά δεδομένα (*labeled data*) όσο και από άγνωστα (*unlabeled data*).



Μάθηση εννοιών στον Άνθρωπο!

Μηχανική Μάθηση

*"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , **if** its performance at tasks in T , as measured by P , improves with experience E "*

Tom M. Mitchell (1997)

Μοντέλο

Παρατήρηση, ανάλυση του προβλήματος

Σύνολο εκπαίδευσης

Αντιπροσωπευτικό, αξιόπιστο

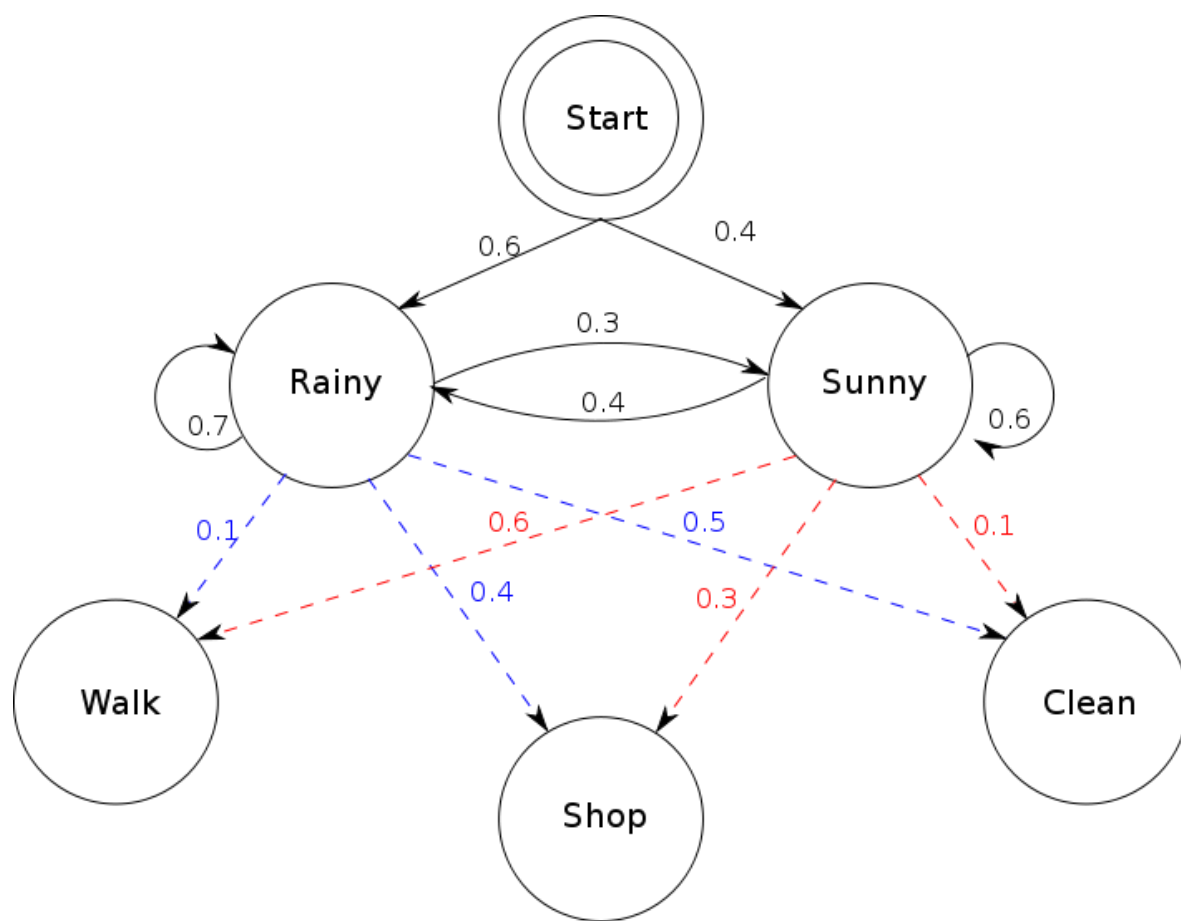
Σύνολο ελέγχου

Ανεξάρτητο από σύνολο εκπαίδευσης

Δείκτες επιτυχίας

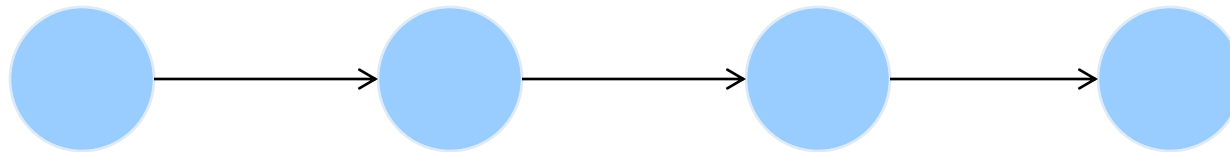
Κατάλληλοι για την φύση του προβλήματος

Hidden Markov Models (HMMs)



Μαρκοβιανά Μοντέλα Εξάρτησης (Markov Models)

- **Πιθανοθεωρητικά** (στοχαστικά) μοντέλα
- Η θεώρηση μιας ακολουθίας ενδεχομένων ως **αλυσίδα Markov** πολύ απλά στηρίζεται στην ιδέα ότι κάθε ένα από τα ενδεχόμενα **εξαρτάται** μόνο από το αμέσως προηγούμενό του, ή αλλιώς το κάθε ενδεχόμενο καθορίζει με κάποια **πιθανότητα** το αμέσως επόμενο του.



- Κατάλληλα μοντέλα για **βιολογικές ακολουθίες**, καθώς αντικατοπτρίζουν την έννοια της πληροφορίας που αυτές περιέχουν.

Μαρκοβιανά Μοντέλα Εξάρτησης (Markov Models)



- Παραδείγματα από τις φυσικές γλώσσες (στα αγγλικά το **Q** ακολουθείται με μεγαλύτερη πιθανότητα από **U**, παρά από κάποιο άλλο).
- Ο **Markov** εμπνεύστηκε τις ομώνυμες αλυσίδες απο ένα ποίημα του Pushkin (*Eugene Onegin*).

>20.000 γράμματα

57% σύμφωνα

43% φωνήεντα

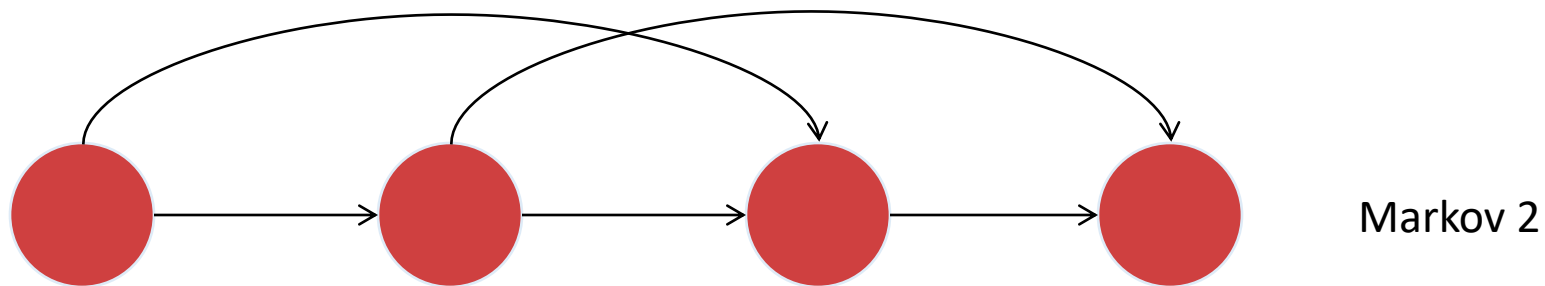
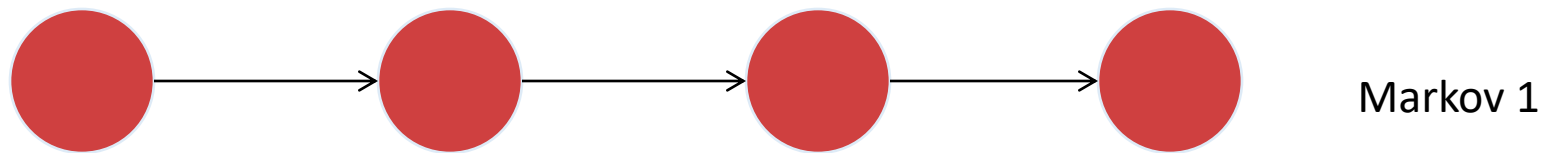
αλλά

Φ --> **Σ** **87%**

Σ --> **Φ** **66%**

*“My **uncle** -- high **ideals** inspire him;
but when past **joking** he **fell** sick,
he **really** forced **one** to **admire** him –
and **never** played a shrewder **trick**.
Let **others** **learn** from his **example**!
But God, how **deadly** dull to **sample**
sickroom **attendance** **night** **and** **day**
and **never** stir a **foot** **away**! ...”*

Μαρκοβιανά Μοντέλα Εξάρτησης (Markov Models)



Μαρκοβιανά Μοντέλα Εξάρτησης (Markov Models)

ATTGTAATCTCACGGTGTACGCGCATGCACAGTCAGT

Σε μια ακολουθία DNA $\mathbf{X} = x_1, x_2, \dots, x_n$ $x_i \in \{A, T, G, C\}$
θεωρούμε ότι η εμφάνιση ενός νουκλεοτιδίου εξαρτάται από το
αμέσως προηγούμενο του

$$P_{ab} = P(x_i = b \mid x_{i-1} = a)$$

και η πιθανότητα αυτή ονομάζεται πιθανότητα μεταβάσεως
η ολική πιθανότητα της ακολουθίας θα είναι:

$$P(\mathbf{X}) = P(x_n, x_{n-1}, \dots, x_1) = P(x_n \mid x_{n-1}, \dots, x_1) P(x_{n-1} \mid x_{n-2}, \dots, x_1) \dots P(x_1)$$

και επειδή:

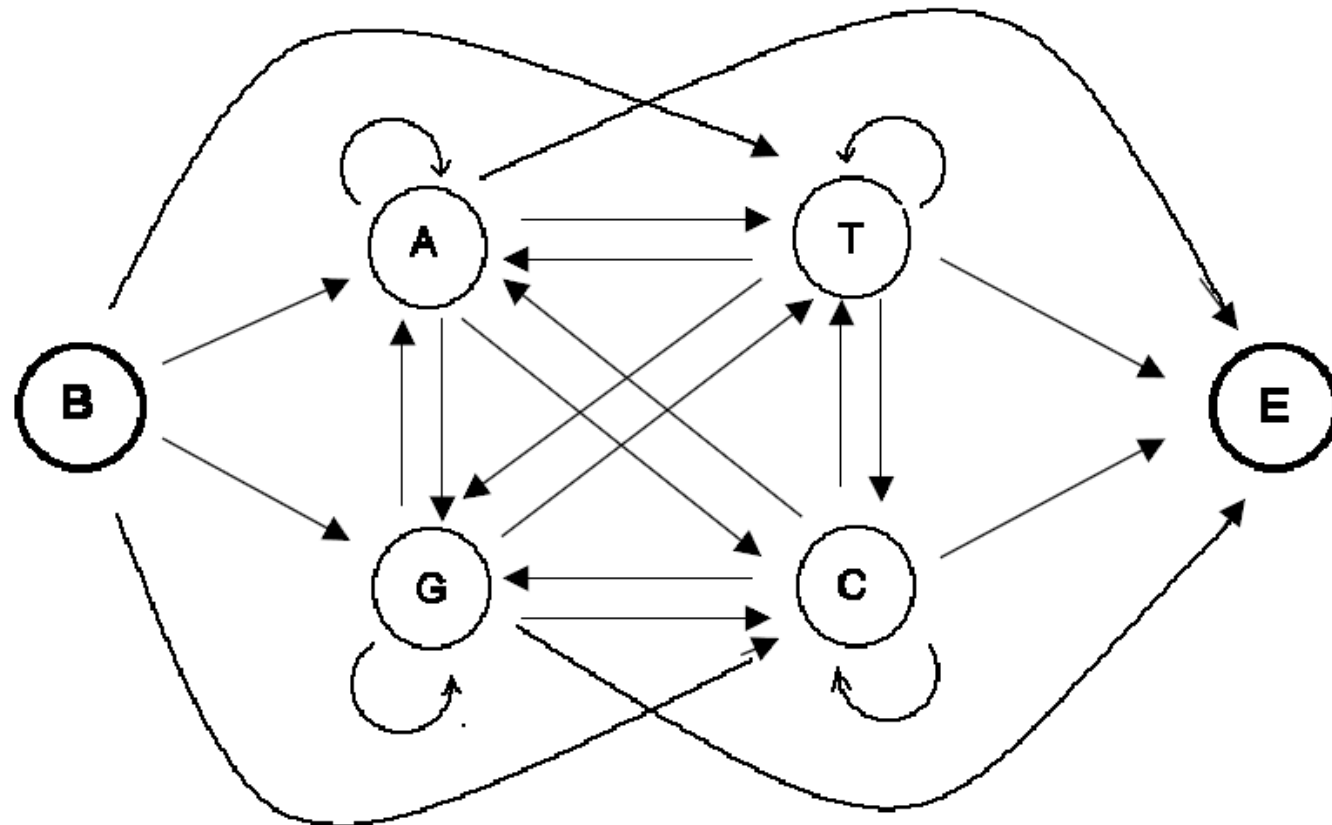
$$P(x_i \mid x_{i-1}, \dots, x_1) = P(x_i \mid x_{i-1}) = p_{x_i x_{i-1}}$$

$$P(x_n \mid x_{n-1}) P(x_{n-1} \mid x_{n-2}) \dots P(x_1) = P(x_1) \prod_{i=2}^n P(x_i \mid x_{i-1})$$

Μαρκοβιανά Μοντέλα Εξάρτησης (Markov Models)

States (καταστάσεις)

Transitions (μεταβάσεις)



Hidden Markov Models

Αν φανταστούμε ένα καζίνο, το οποίο χρησιμοποιεί κάθε φορά με συγκεκριμένη πιθανότητα, ένα **αμερόληπτο** (fair) ζάρι και ένα **μεροληπτικό** (loaded) ζάρι.

21452643**6636561666**232145
-----**+++++**-----

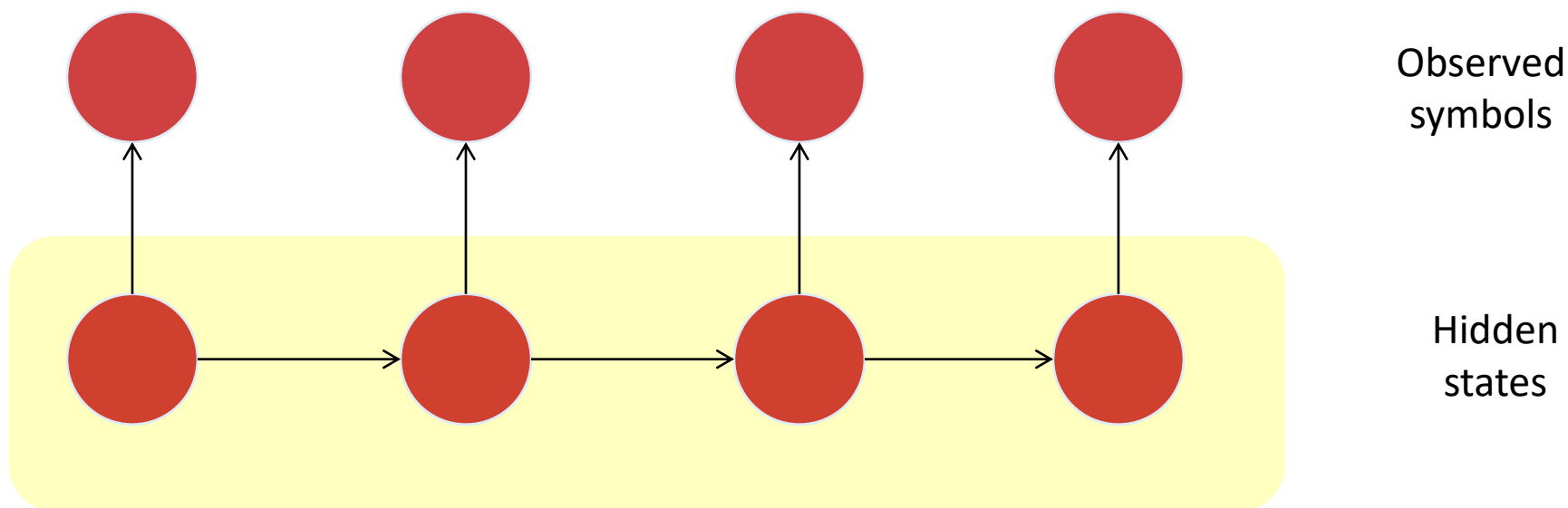
+ = μεροληπτικό

- = αμερόληπτο

Ο παίχτης κάθε φορά γνωρίζει μόνο το αποτέλεσμα, και όχι την φύση του ζαριού. Το μοντέλο που μπορεί να περιγράψει αυτή την κατάσταση είναι το **Hidden Markov Model**, και ονομάζεται έτσι (κρυμμένο) γιατί πλέον δεν υπάρχει 1-1 αντιστοίχιση του αποτελέσματος, με την πραγματική κατάσταση.

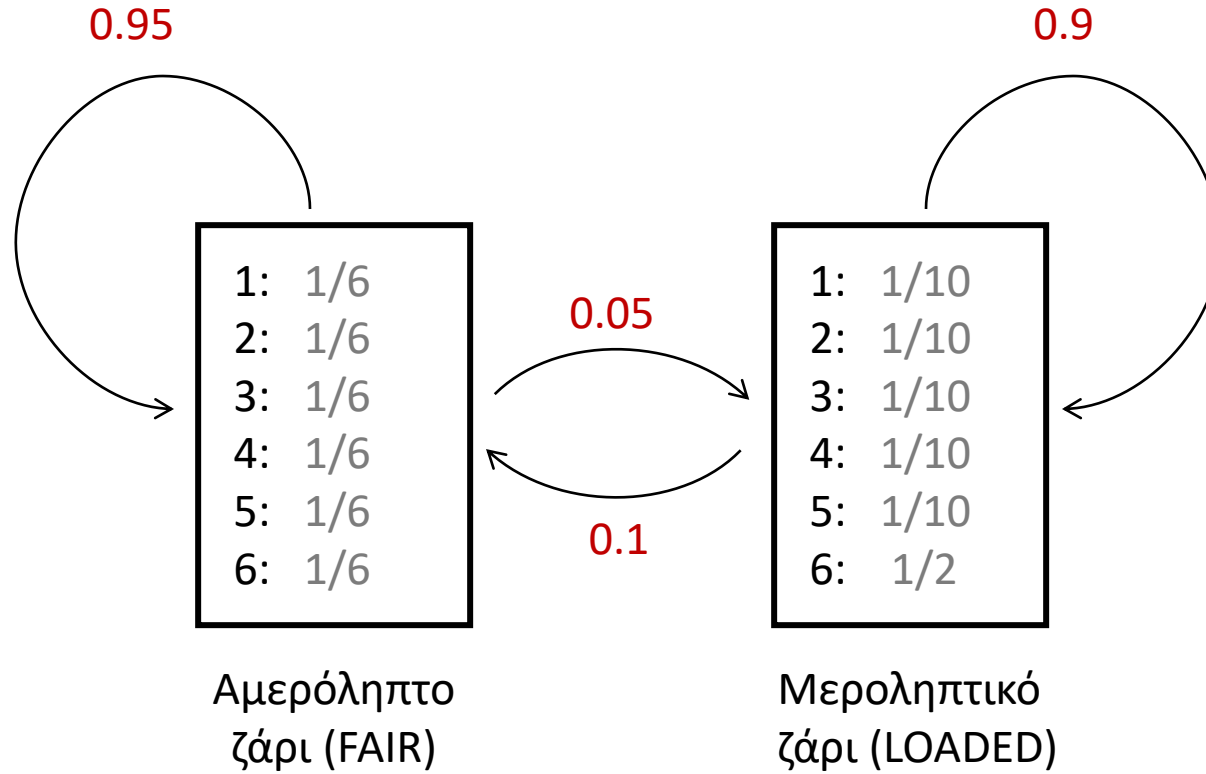
Hidden Markov Models

Ένα Hidden Markov Model (HMM), αποτελείται από ένα σύνολο **κρυφών καταστάσεων**, ένα σύνολο **παρατηρούμενων συμβόλων**, και δυο σύνολα πιθανοτήτων, τις **πιθανότητες μετάβασης** και τις **πιθανότητες εκπομπής** (ή γεννήσεως).

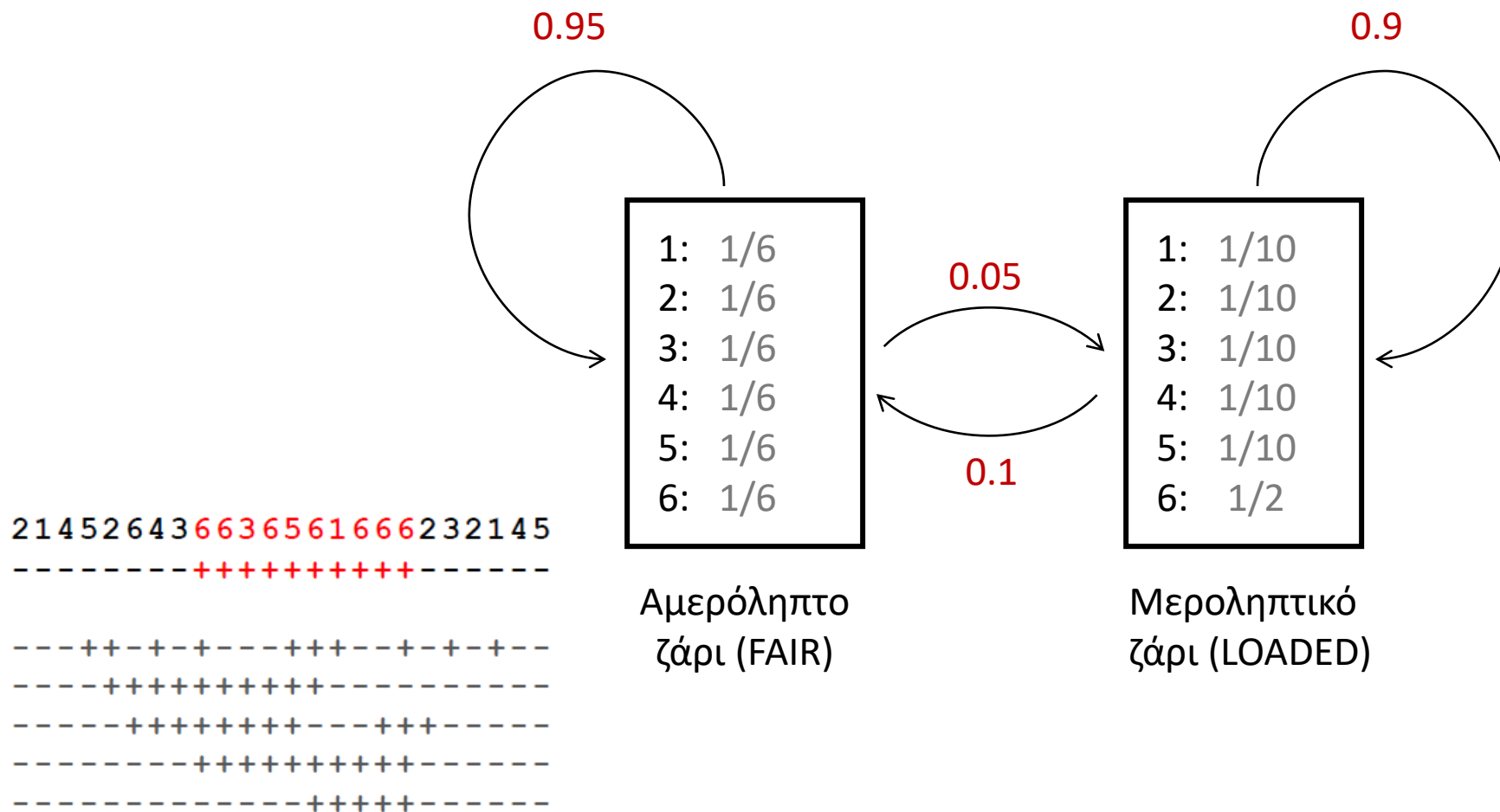


! Πρέπει να τονιστεί, ότι σε ένα HMM η **μαρκοβιανή ιδιότητα** ισχύει για τις καταστάσεις του μοντέλου (states), και όχι για τα σύμβολα.

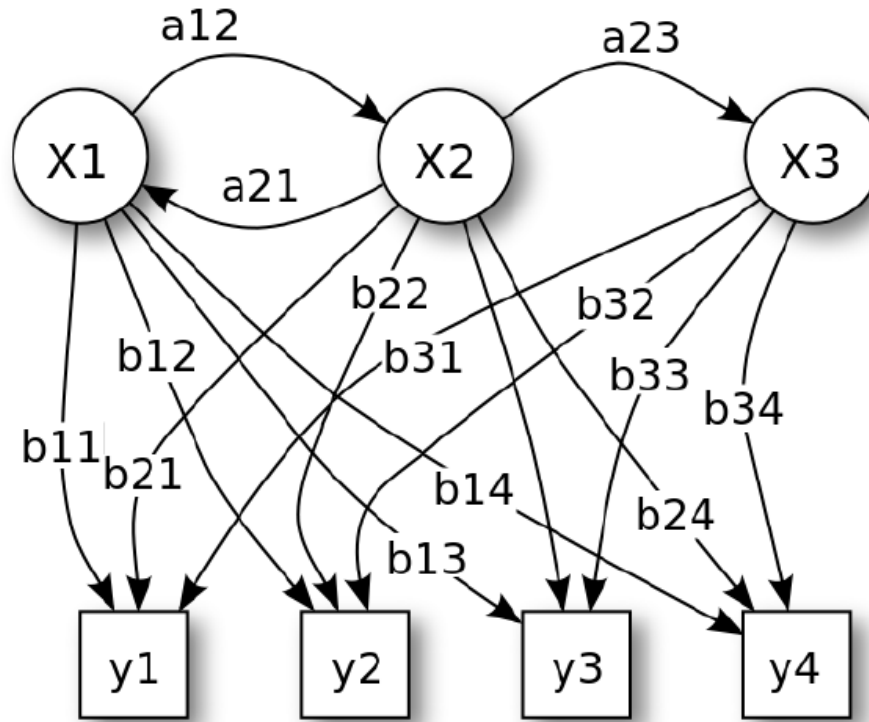
Hidden Markov Models



Hidden Markov Models



Hidden Markov Models



x — states

y — possible observations

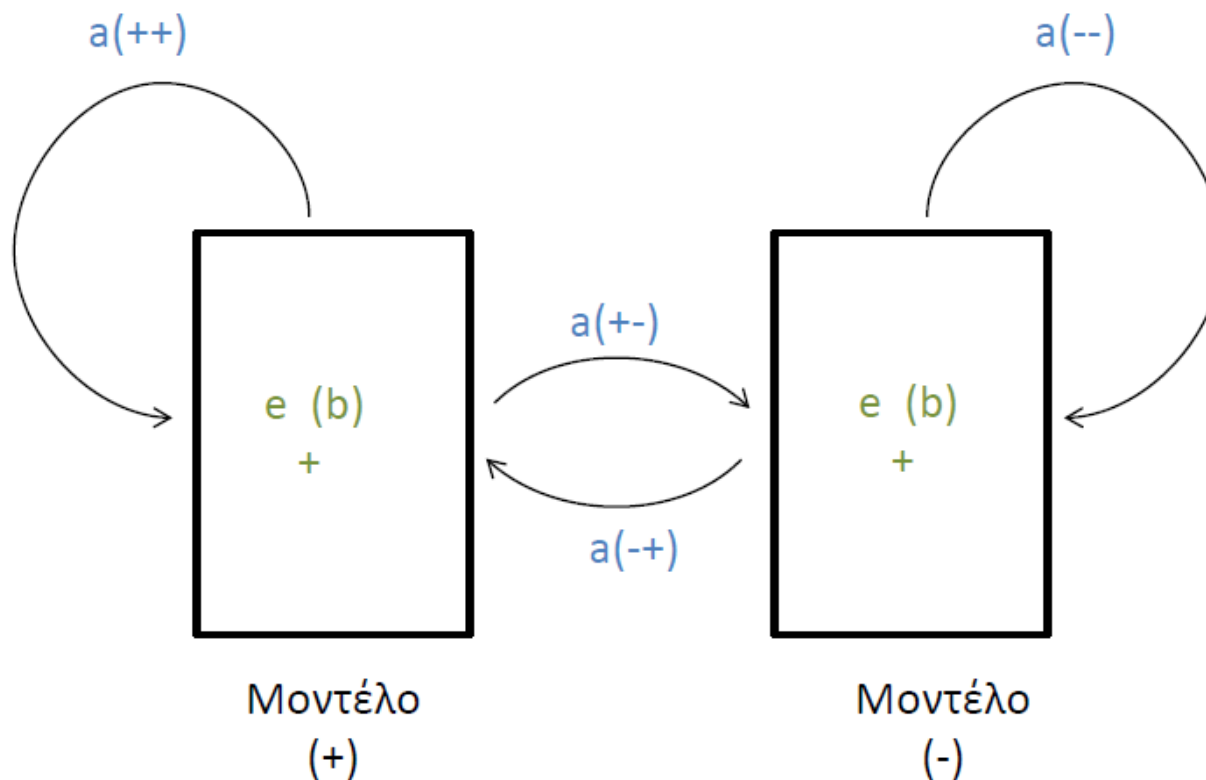
a — state transition probabilities

b — output probabilities (emmission)

Hidden Markov Models

. . AEDGPRGSDADK**LIAVILIGFVLIFVALVC**VTYTRED . .
. . -----**+++++**----- . .

- : μη διαμεμβρανικό τμήμα
- + : διαμεμβρανικό τμήμα



Τα 3 βασικά ερωτήματα σε ένα HMM . . .

Εκτίμηση

- Δεδομένου του μοντέλου, πως θα υπολογίσουμε την ολική πιθανότητα μιας ακολουθίας συμβόλων.
 $P(\mathbf{x}|\theta)$

Αποκωδικοποίηση

- Πως θα βρούμε την πιο πιθανή αλληλουχία καταστάσεων (path) από την οποία έχει διέλθει το μοντέλο, για να δώσει την συγκεκριμένη ακολουθία συμβόλων.

$$\pi^* = \arg \max_{\pi} P(x, \pi)$$

Εκπαίδευση

- Πως θα τροποποιήσουμε τις παραμέτρους του μοντέλου, έτσι ώστε να μεγιστοποιηθεί η συνολική πιθανοφάνεια των ακολουθιών

$$\theta_{\text{ML}} = \arg \max P(\mathbf{x}|\theta)$$

. . . και οι απαντήσεις τους

Εκτίμηση

- Αλγόριθμος FORWARD, αλγόριθμος δυναμικού προγραμματισμού, που υπολογίζει την συνολική πιθανότητα της ακολουθίας, χωρίς να διέλθει από όλα τα δυνατά μονοπάτια (αλληλουχίες καταστάσεων).

Αποκωδικοποίηση

- Αλγόριθμος του VITERBI, αλγόριθμος δυναμικού προγραμματισμού, που μέσω αναδρομής (recursion) υπολογίζει την πιο πιθανή αλληλουχία καταστάσεων για τη δεδομένη ακολουθία και το δεδομένο μοντέλο. (Εναλλακτικά NBEST).

Εκπαίδευση

- Αλγόριθμος των BAUM-WELCH (η αλλιώς FORWARD-BACKWARD), ειδική περίπτωση του αλγόριθμου EM (Expectation-Maximization), ο οποίος χειρίζεται τα δεδομένα σαν δεδομένα με ελλειπής τιμές (missing values) και υπολογίζει Ε.Μ.Π. για τις παραμέτρους του μοντέλου (Εναλλακτικά Gradient Descent).

Hidden Markov Models

Ενα HMM, αφού υπολογιστούν οι παραμέτροι του (**πιθανότητες μεταβάσεως** κλπ) απο ένα σύνολο πρωτεϊνών γνωστής δομής (training set), χρησιμοποιείται για την **πρόγνωση-αποκωδικοποίηση**, σε ένα σύνολο πρωτεϊνών άγνωστης δομής (test set).

Οι μέθοδοι αποκωδικοποίησης, δηλαδή εύρεσης της αλληλουχίας των καταστάσεων εάν είναι γνωστή η αλληλουχία των συμβόλων, είναι βασικά 2, η αποκωδικοποίηση Viterbi, και η εκ των υστέρων αποκωδικοποίηση (posterior decoding). Συνήθως σε περιπτώσεις πολύπλοκων μοντέλων , είναι πιο χρήσιμη η εκ των υστέρων αποκωδικοποίηση.

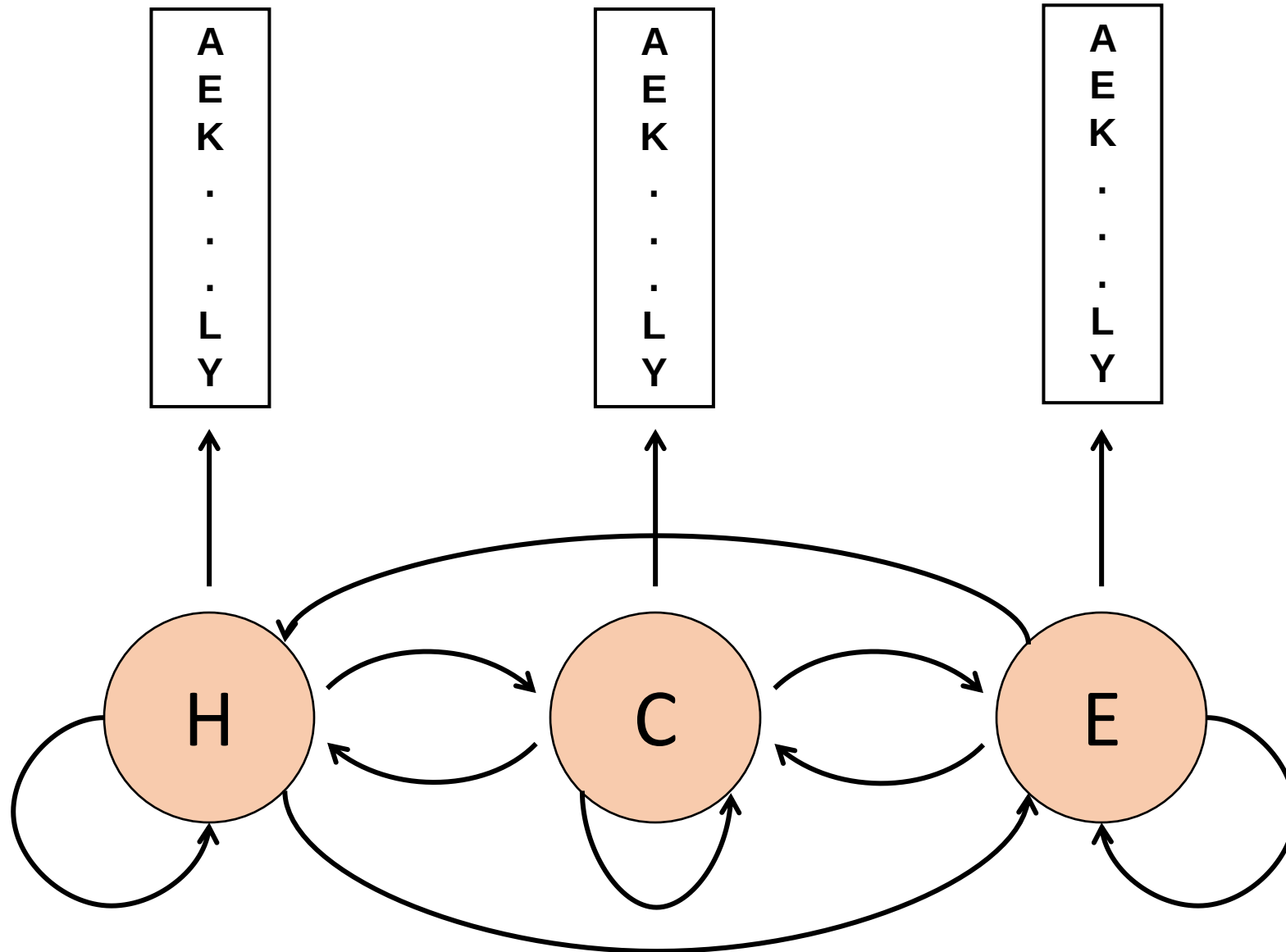
Hidden Markov Models (πλεονεκτήματα)

- Άριστη μαθηματική θεμελίωση, και πιθανοθεωρητική ερμηνεία των αποτελεσμάτων.
- Ύπαρξη κατάλληλων αλγορίθμων για την υλοποίηση του.
- Μπορεί να ενσωματώσει, σχεδόν κάθε είδους πληροφορία για το βιολογικό πρόβλημα, χωρίς να αναγκαστούμε να καταφύγουμε σε ευριστικές (heuristic) μεθόδους.
- Αν δομηθεί το μοντέλο, η εκπαίδευση του και η εκτίμηση πάνω σε αυτό γίνεται με μια διαδικασία από την αρχή μέχρι το τέλος.

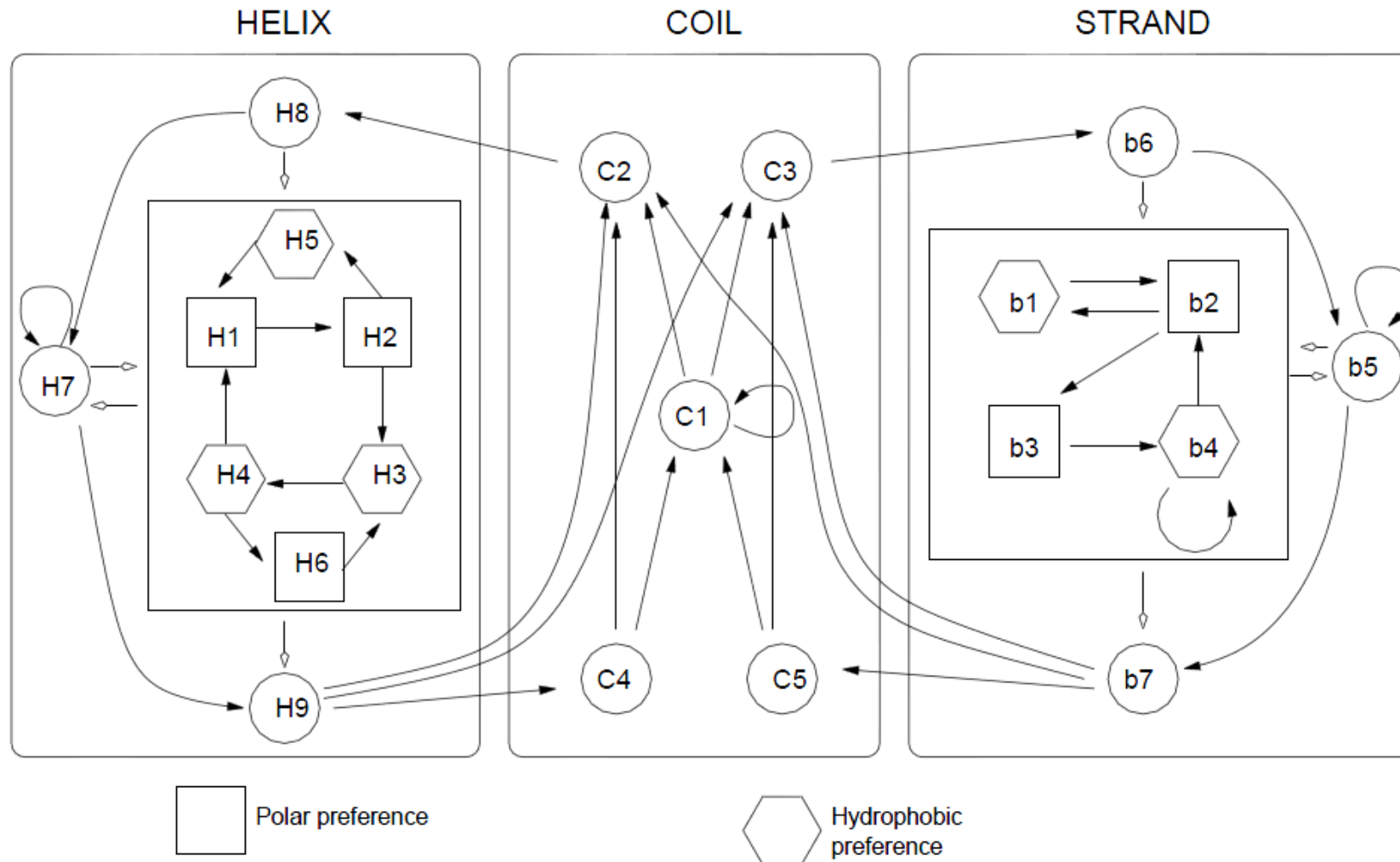
Hidden Markov Models (εφαρμογές)

- Εύρεση πιθανών γονιδίων (εσώνια – εξώνια).
- Πρόγνωση δευτεροταγούς δομής πρωτεϊνών.
- Πρόγνωση διαμεμβρανικών τμημάτων πρωτεϊνών.
- Πρόγνωση των πεπτιδίων οδηγητών.
- Εύρεση ομόλογων οικογενειών ακολουθιών.
- και πολλά άλλα

Hidden Markov Models



Hidden Markov Models



Hidden Markov Models

Σύνολο εκπαίδευσης
(αντιπροσωπευτικό, μη ομόλογο, labeled data)

>**protein1**

KRAIYPGTFDPITNGHIDIVTRATQMFDHVILAIASPSKKPMFTLEERVALAQQAT
CEEEEEEECCCCCHHHHHHHHHHHHCCEEEEEECCCCCCCCCHHHHHHHHHHC

>**protein2**

NSTVVSNSSELILNLTPIALAYTVQSLPLIATQPAWLGTIADNYSKWRWVSLRI
CEEEEEEEEEEECEECCCCEEEEEECCCCCHHHHHHCEEEEEEEEEEE

.
. .

>**protein200**

MVSKLSQLQTELLAALLESGLSKEALIQALG
CCCCCHHHHHHHHHHHHCCCHHHHHHHCC

Hidden Markov Models

Σύνολο ελέγχου
(αντιπροσωπευτικό, μη ομόλογο, ανεξάρτητο από το
σύνολο εκπαίδευσης !)

>**P08172**

MNNSTNSSNNSLALTSPYKTFEVVFIVLVAGSLSLVTIIGNILVMVSIKVNRLQTV
NNYFLFSLTNSSNNSLALTSPYKTFEVVFIVLVAGSLSLVTIIGNILVMVSIKVNRLH
LQTVNNYFLFSLACADLIIGVF

>**P08173**

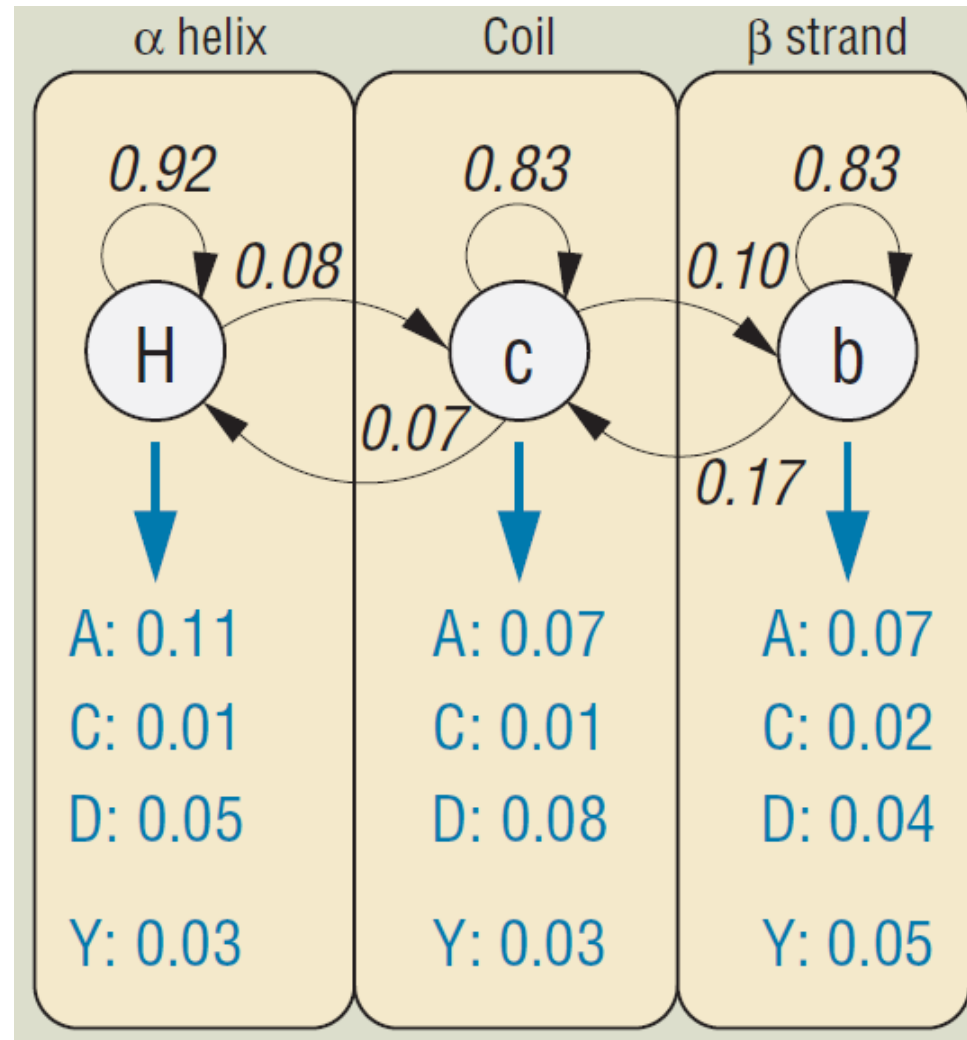
MANFTPVNGSSGNQSVRLVTSSSHNRYETVEMVFIATVTGSLSLVTTVGNILVMLSI
KVNRLQTVNNYFLFSLA

.
. .

>**P50052**

MKGNSTLATTSKNITSKNITSKNITSKNITSKNITSGLHFGLVNISGNNESTLNCSQ
KPSDKHLDAIPILYYIIFVIGFLVNIVVVTLFCCQIVVVTLFCCQKGP

Hidden Markov Models



Hidden Markov Models

Name	Web site
PHD/PHDpsi	http://www.predictprotein.org/
PSIPRED	http://bioinf.cs.ucl.ac.uk/psipred/
PROF (king)	http://www.aber.ac.uk/~phiwww/prof/
SSpro	http://scratch.proteomics.ics.uci.edu/
Porter	http://distill.ucd.ie/porter/
APSSP2	http://www.imtech.res.in/raghava/apssp2/
SAM-T06	http://www.soe.ucsc.edu/research/compbio/SAM_T06/T06-query.html/
YASPIN	http://www.ibi.vu.nl/programs/yaspinwww/
Jpred (v3)	http://www.compbio.dundee.ac.uk/jpred/

Hidden Markov Models (εφαρμογές)

Splice site recognition

- Εύρεση της 5'-περιοχής ματίσματος
- Προϋποθέσεις:
 - α) Τα εξώνια έχουν ίδιες πιθανότητες εμφάνισης βάσεων
 - β) Τα εσώνια έχουν πολύ A/T
 - γ) Στο σημείο ματίσματος υπάρχει σχεδόν πάντα G

Hidden Markov Models (εφαρμογές)

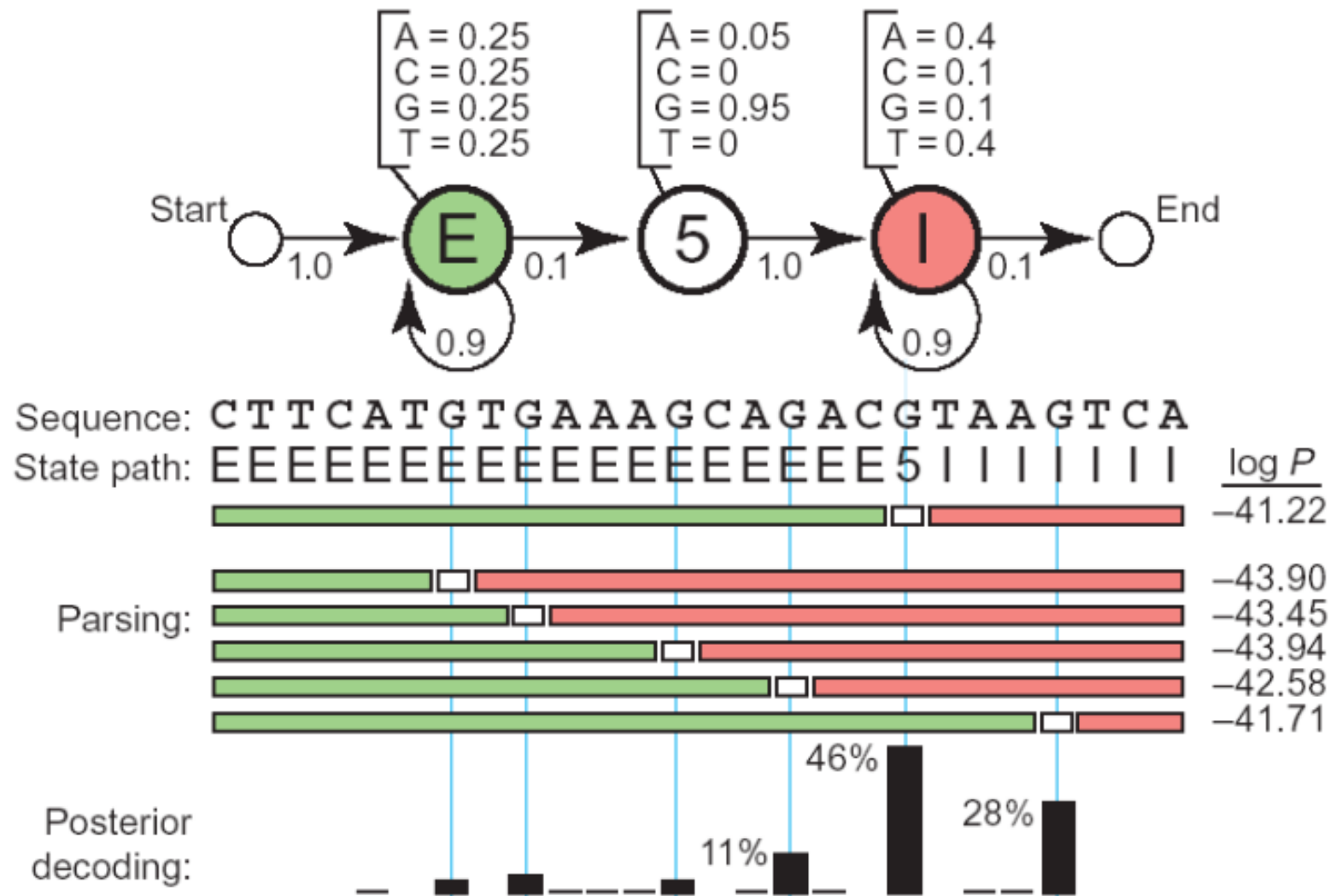
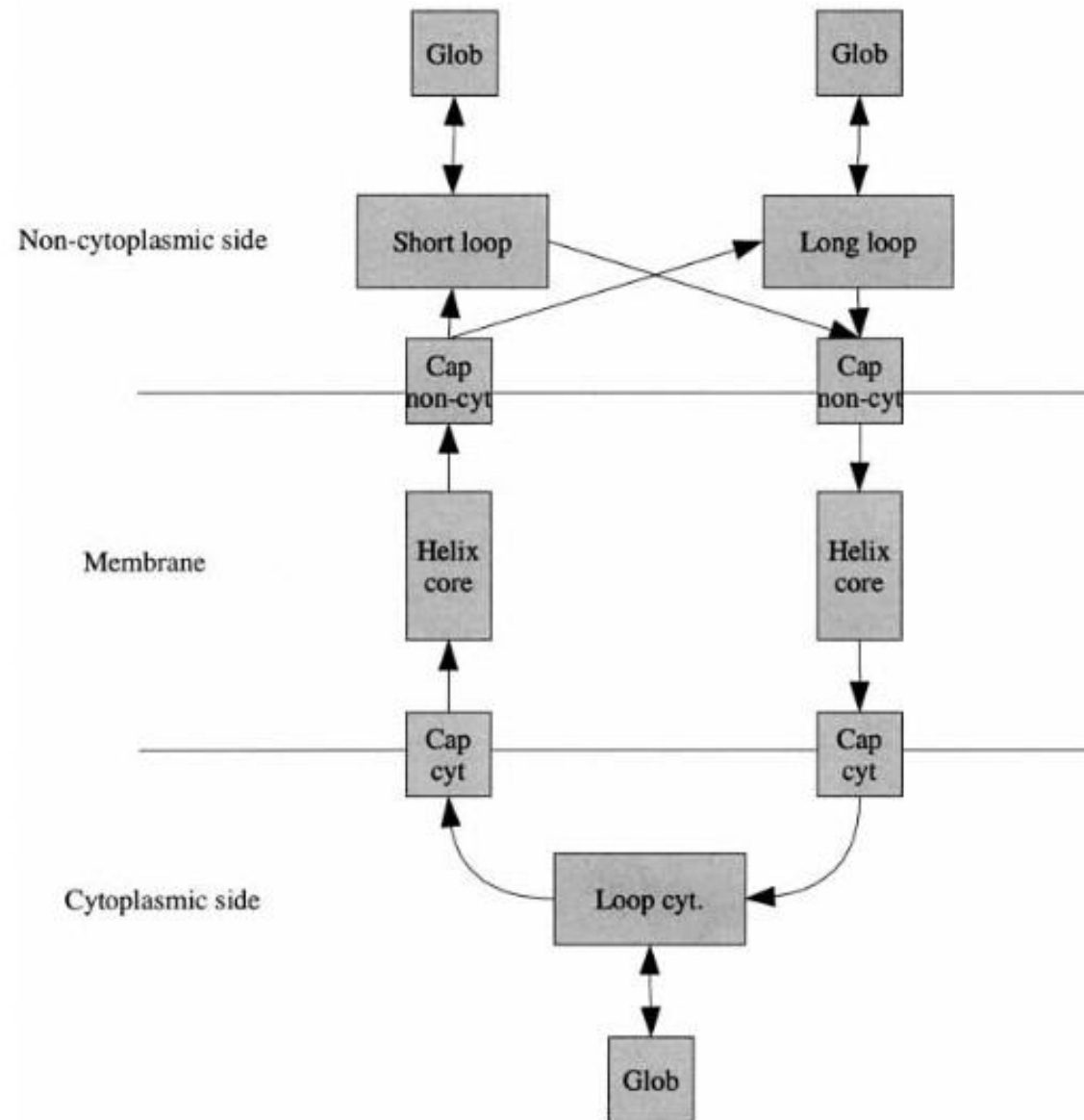


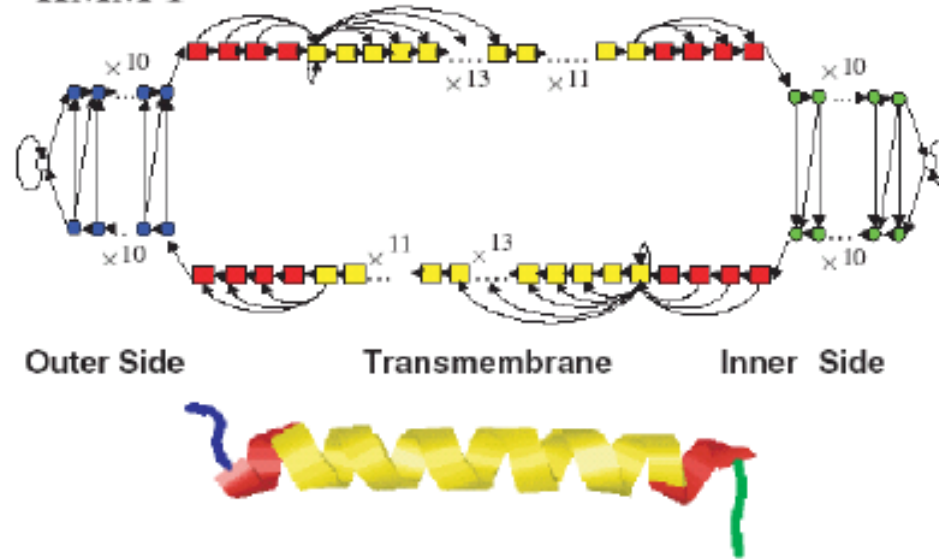
Figure 1 A toy HMM for 5' splice site recognition. See text for explanation.

Hidden Markov Models (εφαρμογές)

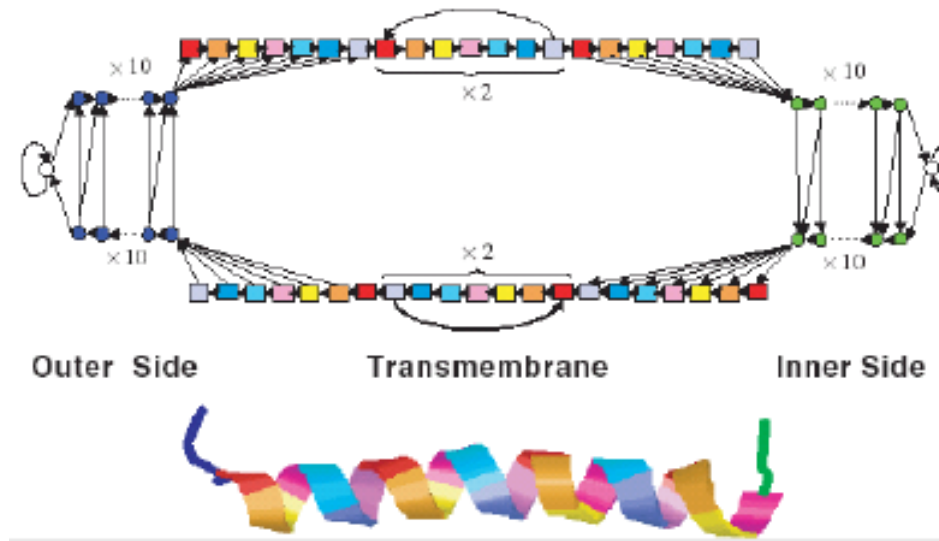


Hidden Markov Models (εφαρμογές)

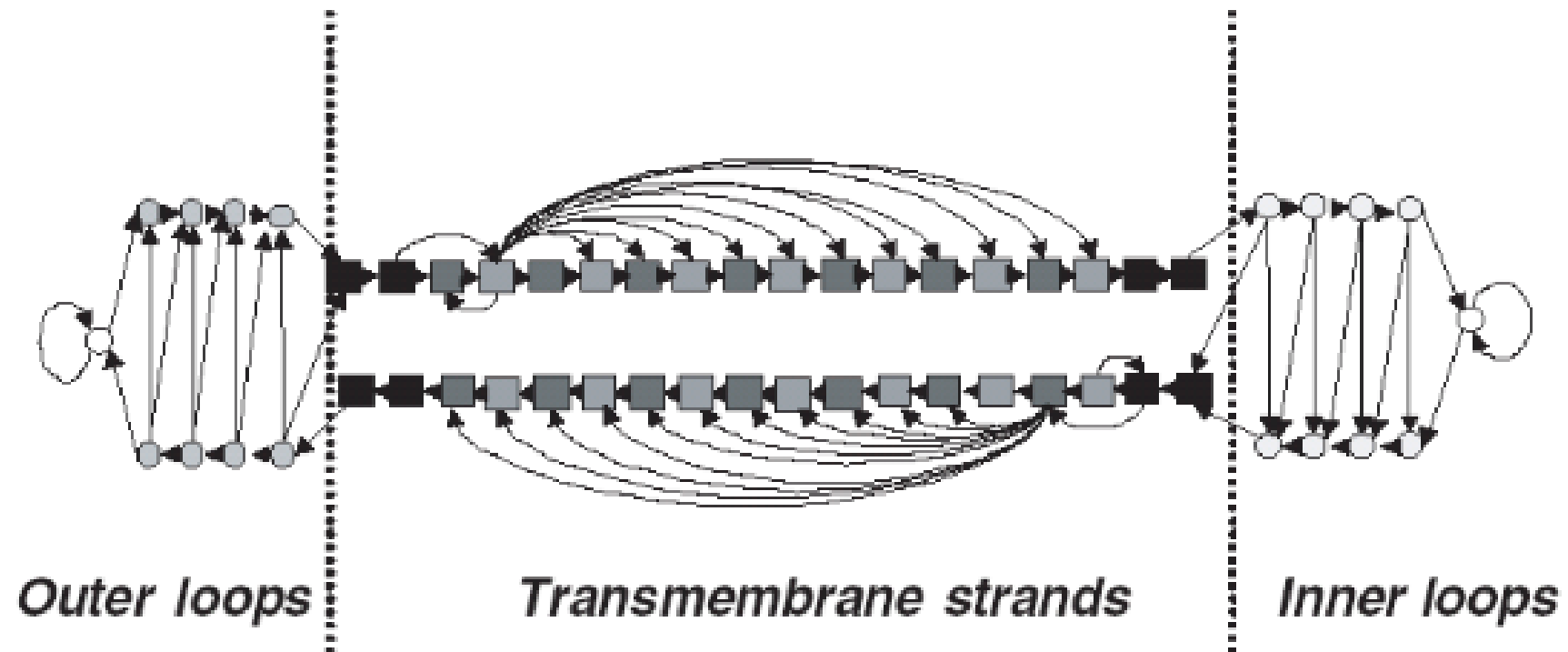
HMM 1



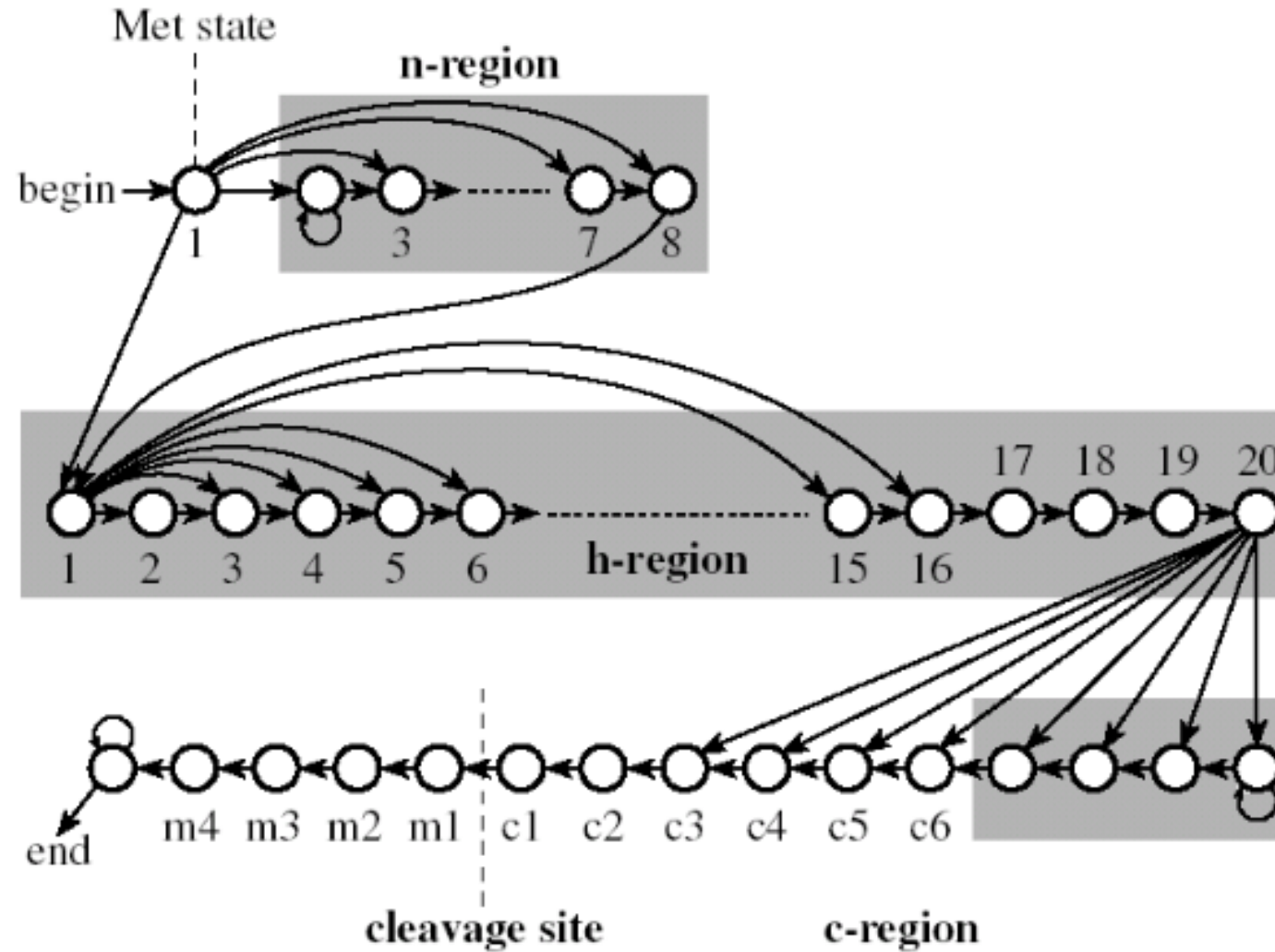
HMM 2



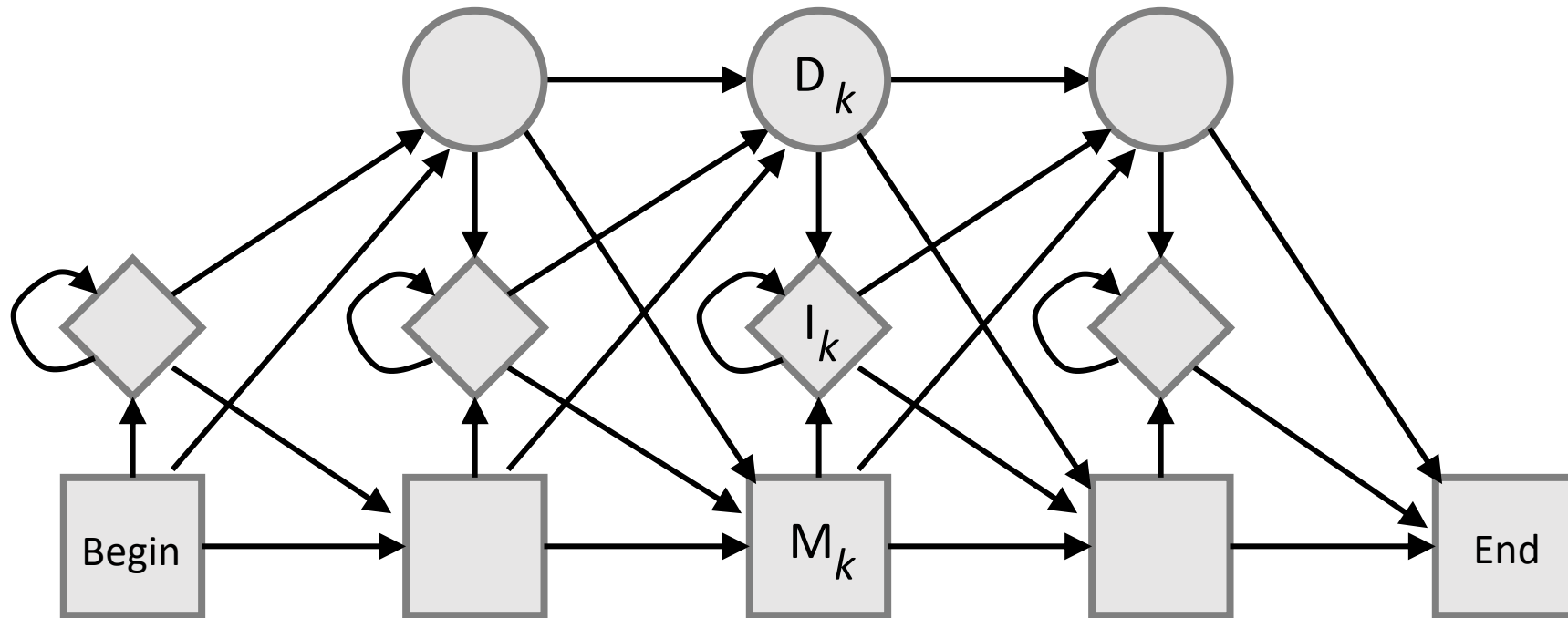
Hidden Markov Models (εφαρμογές)



Hidden Markov Models (εφαρμογές)



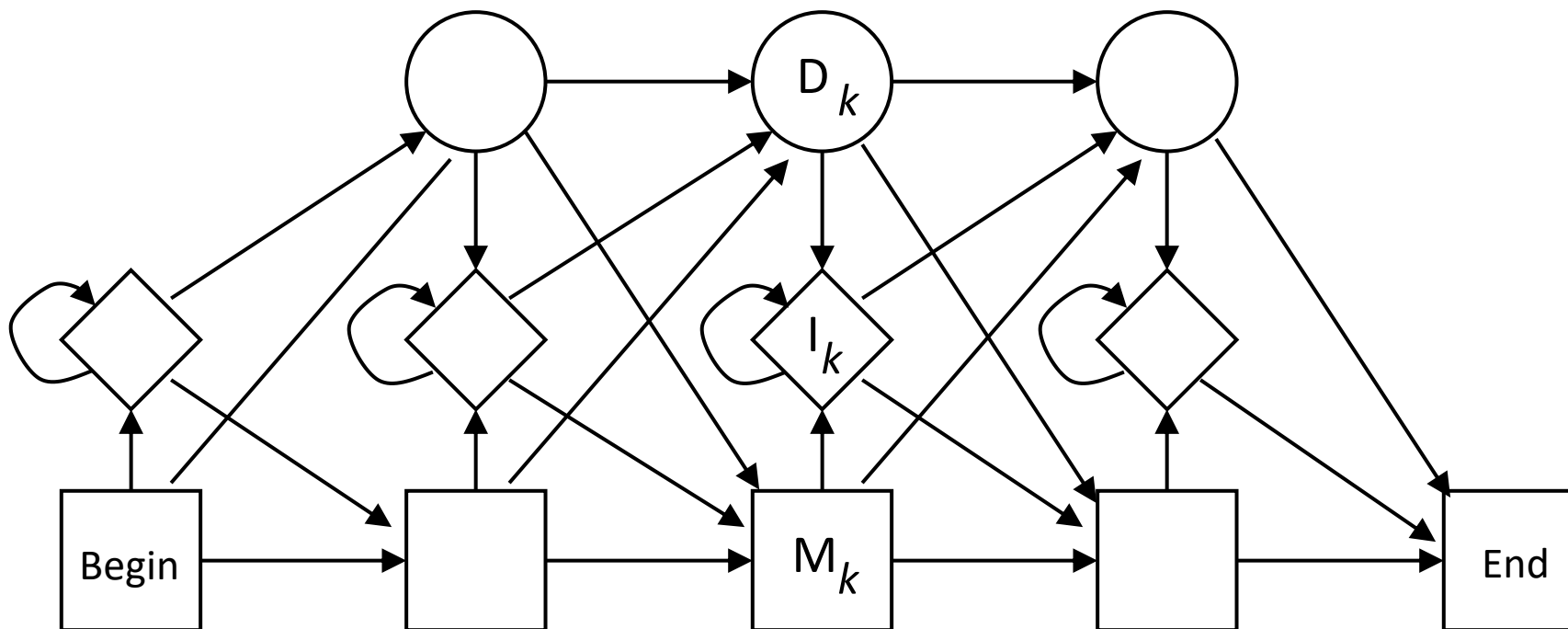
profile Hidden Markov Models (pHMMs)



profile Hidden Markov Models (pHMMs)

- Ένα HMM με “left-to-right” αρχιτεκτονική (μονόδρομη κατεύθυνση των μεταβάσεων).
- Διαθέτει ειδικά states τα οποία δεν έχουν emissions (non-emitting states)
- Περιγράφουν στατιστικά μια πολλαπλή στοίχιση.
- Προτάθηκαν απο τους Hughey και Krogh (1996).

profile Hidden Markov Models (pHMMs)



Οι καταστάσεις που παρατηρούνται σε ένα τέτοιο μοντέλο (εκτός αυτών της εκκίνησης και του τερματισμού) χωρίζονται σε 3 κατηγορίες :

Καταστάσεις Σύμπτωσης (Match states) M_k τετράγωνα

Καταστάσεις Εισαγωγής (Insertion states) I_k ρόμβοι

Καταστάσεις Απαλοιφής (Deletion states) D_k κύκλοι

profile Hidden Markov Models (pHMMs)

Ιδιαιτερότητες

- Αντίστοιχα ορίζονται οι πιθανότητες γέννησης οι οποίες γεννούν τα σύμβολα σε κάθε κατάσταση.
- Έτσι υπάρχει και εδώ μια αλληλουχία καταστάσεων η οποία είναι κρυφή και μια αλληλουχία συμβόλων που είναι φανερή, και θεωρούμε ότι παράγεται από την αλληλουχία των καταστάσεων.
- Οι καταστάσεις σύμπτωσης και εισαγωγής, είναι κανονικές καταστάσεις οι οποίες συνδέονται μέσω των πιθανοτήτων γέννησης με την εμφάνιση συμβόλων.
- Οι μεν καταστάσεις σύμπτωσης αντιστοιχούν σε στήλες της πολλαπλής στοίχισης οι οποίες στοιχίζονται καλά και άρα αντιστοιχούν σε περιοχή με ομοιότητα, ενώ οι καταστάσεις εισαγωγής, αντιστοιχούν σε περιοχές στις οποίες έχουμε εισαγωγή χαρακτήρων που δεν στοιχίζονται καλά.
- Οι περιοχές αυτές, οι οποίες δεν υπάρχουν στις υπόλοιπες ακολουθίες, εμφανίζονται ως κενά τα οποία μοντελοποιούνται μέσω των σιωπηρών καταστάσεων απαλοιφής.

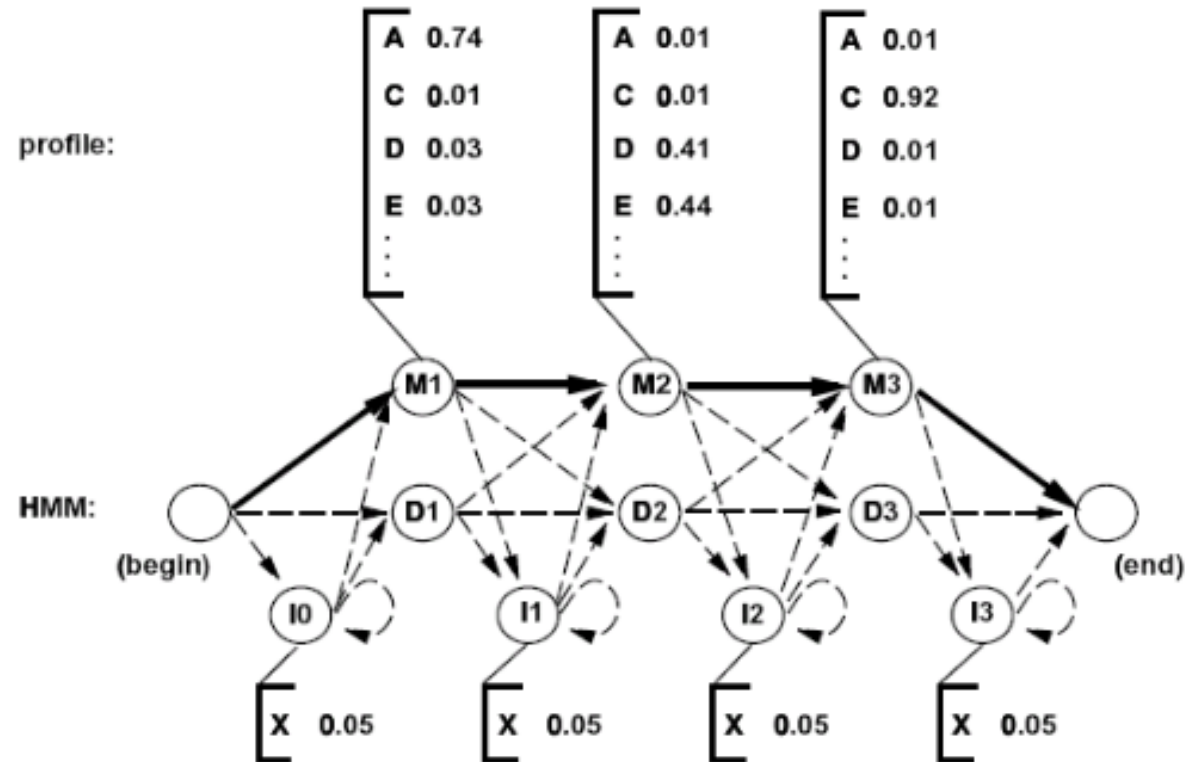
profile Hidden Markov Models (pHMMs)

multiple alignment:

-	A	D	T	C
W	A	E	-	C
-	V	E	-	C
-	A	D	-	C
-	A	E	-	C

consensus:

A	D/E	C
---	-----	---



profile Hidden Markov Models (pHMMs)

Πλεονεκτήματα

- Με την εισαγωγή των διαφορετικών καταστάσεων σύμπτωσης και εισαγωγής, γίνεται μια σημαντική τομή σε σχέση με τις κλασσικές μεθόδους στοίχισης, οι οποίες καθώς δεν προϋποθέτουν ένα μοντέλο δεν διαχωρίζουν τις πληροφοριακές θέσεις στην στοίχιση από τις απλές τυχαίες εισαγωγές.
- Επιπλέον, για πρώτη φορά οι ποινές για την εισαγωγή κενών (gap penalties), δεν τίθενται εκ των προτέρων αλλά εκτιμώνται από τα δεδομένα και αναπαρίστανται με καθαρά πιθανοθεωρητικό τρόπο, αποκλείοντας την υποκειμενική παρέμβαση.
- Έτσι, με τέτοια μοντέλα, είμαστε σε θέση να πραγματοποιήσουμε ιδιαίτερα ευαίσθητες αναζητήσεις και να εντοπίσουμε απομακρυσμένες ομολογίες (remote homologies), τις οποίες οι παραδοσιακοί αλγόριθμοι στοίχισης δεν θα μπορούσαν να εντοπίσουν.

profile Hidden Markov Models (pHMMs)

EGF domain structural alignment:

```
1ixa      VDGDQCE  SNPCLNGGSCCKDD  INSYECWCPFGFEGKNCEL
1apo      KDGDQCE  GHPCLNQGHCKDG  IGDYTCTCAEGFEGKNCEFSTR
1epi      NSYPGCPSSYDGYCLNGGVCMHIESLDSYTCNCVIGYSGDRCQTRDLRWWELR
4tgf      VVSHFNDCPDSHTQFCFH  GTCRFLVQEDKPACVCHSGYVGARCEHADLLA
```

ClustalW alignment:

```
1ixa      VDGDQCESN  P  CLNGGSCCKDD  INSYECWCPFGFEGKNCE  L
1apo      KDGDQCEGH  P  CLNQGHCKDG  IGDYTCTCAEGFEGKNCE  FSTR
1epi      NSYPGCPSSYDGYCLNGGVCMHIES  LDSYTCNCVIGYSGDRCQTRDLRWWELR
4tgf      VVSHFND  CPDSHTQF  CFHG  TCRFLVQEDKP  ACVCHSGYVGARCE  HADLLA
```

HMM simulated annealing alignment:

```
1ixa      V  DGDQCE  SNPCLNGGSCCKDD  INSYECWCPFGFEGKNCE  L
1apo      K  DGDQCE  GHPCLNQGHCKDG  IGDYTCTCAEGFEGKNCEF STR
1epi      N  SYPGCPSSYDGYCLNGGVCMHIESLDSYTCNCVIGYSGDRCQTRDLRWWELR
4tgf      VVSHFNDCPDSHTQFCFH  GTCRFLVQEDKPACVCHSGYVGARCE  HADLL  A
```

profile Hidden Markov Models (pHMMs)

Χρήσεις

- Πολλαπλή στοίχιση
- Κατασκευή χαρακτηριστικών profiles
- Εντοπισμός remote homologues
- Αναζητήσεις σε βάσεις δεδομένων

Software

<http://www.cfar.umd.edu/~kanungo/software/software.html>

- HMMER
 - (<http://hmmer.org/>)
- SAM
 - (<http://www.cse.ucsc.edu/research/compbio/sam.html>)
- pftools package
 - (<http://web.expasy.org/pftools/>)

Software

- HMM Editor
http://sysbio.rnet.missouri.edu/multicom_toolbox/HMMEditor%201.0.html
- GHMM Library – HMMEd
<http://ghmm.sourceforge.net/index.html>
<http://ghmm.sourceforge.net/hmmed.html>
- Skylign
<http://skylign.org/>

HMM-ModE

Το 2007 προτάθηκε από τους Srivastava και συνεργάτες η χρήση τόσο θετικού όσο και αρνητικού συνόλου εκπαίδευσης κατά την κατασκευή pHMM, με στόχο την ακόμα μεγαλύτερη βελτίωση της διακριτικής ικανότητας μεταξύ οικογενειών και υπο-οικογενειών (κυρίως αυτών με πολύ μεγάλη ομολογία).

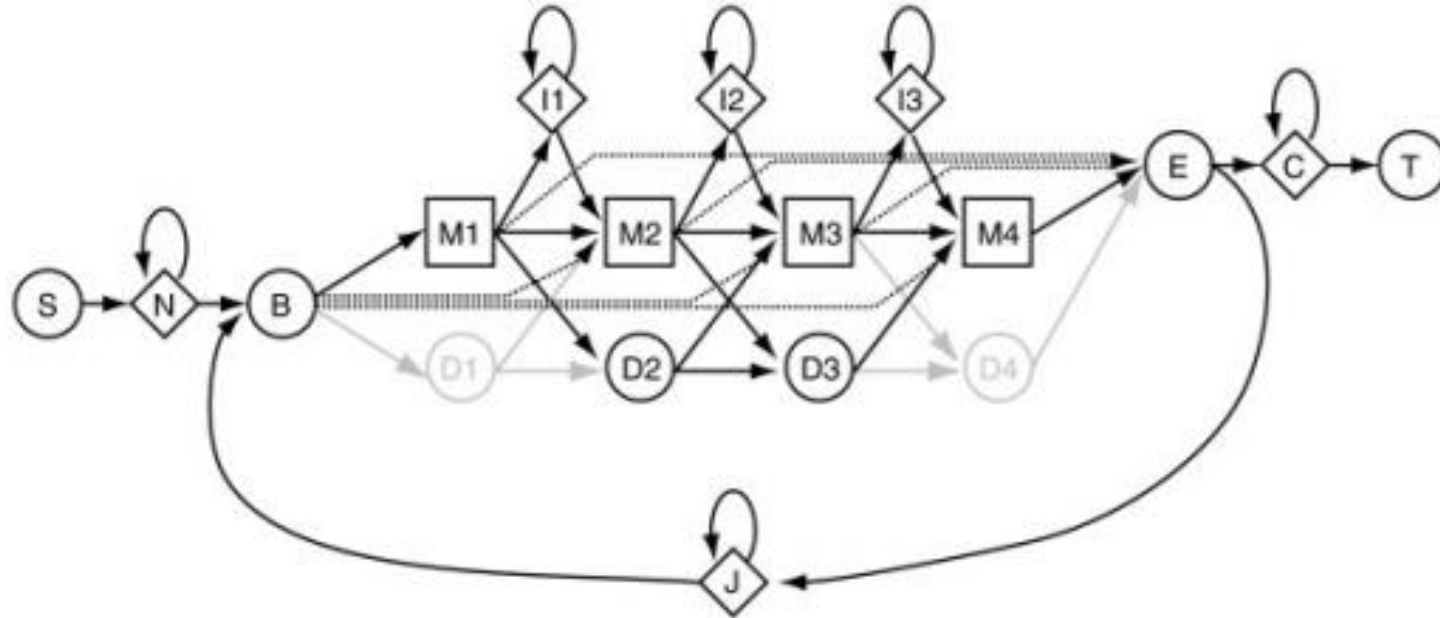
Το HMM-ModE είναι ένα πρωτόκολλο που χρησιμοποιεί επιπλέον ένα αρνητικό σύνολο εκπαίδευσης (πρωτεΐνες που δεν ανήκουν στην οικογένεια αλλά μοιάζουν με τα μέλη) για να τροποποιήσει τις πιθανότητες εκπομπής και μετάβασης του τελικού pHMM κατάλληλα ώστε να διαχωρίζονται οι εν λόγω πρωτεΐνες.

HMM-ModE – Improved classification using profile hidden Markov models by optimising the discrimination threshold and modifying emission probabilities with negative training sequences

<https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-104>

<https://bmccresnotes.biomedcentral.com/articles/10.1186/1756-0500-7-483>

HMMER



- GNU license
- Ανανεώνεται συνεχώς
- Τρέχει σε όλες τις πλατφόρμες
- Ποικιλία εργαλείων
- Ευέλικτη αρχιτεκτονική

HMMER

- **hmmbuild:** Πρόγραμμα με χρήση του οποίου, ξεκινώντας από μια αρχική πολλαπλή στοίχιση, κατασκευάζεται ένα μοντέλο HMM το οποίο να την περιγράφει.
- **hmmalign:** Πρόγραμμα με το οποίο μια σειρά ακολουθιών οι οποίες προέρχονται από ένα HMM, στοιχίζονται σε μια πολλαπλή στοίχιση. Η πολλαπλή στοίχιση, επιτυγχάνεται μέσω διαδοχικών στοιχίσεων των ακολουθιών με το μοντέλο.
- **hmmsearch:** Πρόγραμμα το οποίο, πραγματοποιεί αναζητήσεις ενός μοντέλου HMM έναντι μιας βάσης ακολουθιών πρωτεϊνών.
- **phmmer:** Πρόγραμμα το οποίο πραγματοποιεί αναζήτηση μια πρωτεϊνικής αλληλουχίας έναντι μιας βάσης δεδομένων πρωτεϊνών (ανάλογο με το BLASTP)
- **jackhmmmer:** Πρόγραμμα το οποίο πραγματοποιεί επαναληπτικές αναζητήσεις μια πρωτεϊνικής αλληλουχίας έναντι μιας βάσης δεδομένων πρωτεϊνών (ανάλογο με το PSI-BLAST)
- **hmmsearch:** Πρόγραμμα με το οποίο πραγματοποιούνται αναζητήσεις μιας ή περισσότερων ακολουθιών έναντι μιας βάσης δεδομένων από μοντέλα HMM. Πρέπει να τονιστεί εδώ, ότι αν έχουμε μια ακολουθία και ένα HMM, τα δυο παραπάνω προγράμματα επιστρέφουν ακριβώς το ίδιο αποτέλεσμα. Αν διαφέρουν, είτε οι ακολουθίες είτε τα μοντέλα, τότε δίνουν άλλο αποτέλεσμα, λόγω του διαφορετικού τρόπου υπολογισμού της στατιστικής σημαντικότητας.

HMMER

- **nhmmer:** Πρόγραμμα που πραγματοποιεί αναζήτηση μιας ακολουθίας DNA, μιας στοίχισης ή ενός pHMM, έναντι μιας βάσης ακολουθιών DNA. (ανάλογο με το BLASTN)
- **nhmmscan:** Πρόγραμμα που πραγματοποιεί αναζήτηση μιας ακολουθίας DNA έναντι μιας βάσης δεδομένων από DNA profile HMM.
- **hmmconvert:** Πρόγραμμα που μετατρέπει μοντέλα HMM από και προς τη μορφή του HMMER3.
- **hmmemit:** Πρόγραμμα, με το οποίο 'εκπέμπεται' η καλύτερη (ανάλογα με τον ορισμό) ακολουθία η οποία θα μπορούσε να παραχθεί από το μοντέλο.
- **hmmcompress:** Μετατρέπει μια βάση δεδομένων HMM σε δυαδικό κώδικα για το hmmscan.
- **hmmstat:** δείχνει συνοπτικά στατιστικά για μια βάση δεδομένων HMM.

Μέτρα εκτίμησης της αξιοπιστίας των μεθόδων

Για να μπορέσουμε να μετρήσουμε την επιτυχία και την αξιοπιστία των προγνώσεων που προέρχονται από μια μέθοδο, έχουν προταθεί διάφορα μέτρα. Τα περισσότερα από αυτά, ισχύουν τόσο για την περίπτωση της **τοπικής πρόγνωσης (per-residue prediction)** όσο και για την **κατάταξη αλληλουχιών σε κατηγορίες (per-protein classification)**.

Ως παράδειγμα για την αξιολόγηση των προγνώσεων σε επίπεδο αμινοξικών καταλοίπων, θα υπολογίσουμε διάφορα μέτρα αξιοπιστίας για την πρόγνωση α-ελικοειδών καταλοίπων, με την πρόγνωση να έχει αναχθεί σε δυο κατηγορίες (ελικοειδή κατάλοιπα/μη-ελικοειδή κατάλοιπα).

N = αριθμός πρωτεϊνών

TP = πρωτεΐνες που προβλέφθηκαν α τύπου και παρατηρήθηκαν α τύπου

TN = πρωτεΐνες που προβλέφθηκαν όχι α τύπου και παρατηρήθηκαν όχι α τύπου

FN = κατάλοιπα που προβλέφθηκαν όχι α τύπου και παρατηρήθηκαν α τύπου

FP = κατάλοιπα που προβλέφθηκαν α τύπου και παρατηρήθηκαν όχι α τύπου

Μέτρα εκτίμησης της αξιοπιστίας των μεθόδων

Πρόγνωση			Παρατήρηση		
Πρωτεΐνη 1	α τύπου	α τύπου	Πρωτεΐνη 5	α τύπου	α τύπου
Πρωτεΐνη 2	α τύπου	α τύπου	Πρωτεΐνη 6	α τύπου	α τύπου
Πρωτεΐνη 3	α τύπου	όχι α τύπου	Πρωτεΐνη 7	α τύπου	όχι α τύπου
Πρωτεΐνη 4	όχι α τύπου	όχι α τύπου	Πρωτεΐνη 8	όχι α τύπου	α τύπου

Παρατήρηση

Πρόγνωση		α τύπου	όχι α τύπου
	α τύπου	TP	FP
	όχι α τύπου	FN	TN

Παρατήρηση

Πρόγνωση		α τύπου	όχι α τύπου
	α τύπου	4	2
	όχι α τύπου	1	1

Μέτρα εκτίμησης της αξιοπιστίας των μεθόδων

$$Q = \frac{TP + TN}{TP + TN + FP + FN} \cdot 100\%$$

Ακρίβεια (Accuracy): το συνολικό ποσοστό των καταλοίπων που έχουν προβλεφθεί σωστά (Q)

$$Sn = \frac{TP}{TP + FN} \quad \text{Ευαισθησία (Sensitivity, Sn)} \quad Sp = \frac{TN}{TN + FP} \quad \text{Ειδικότητα (Specificity, Sp)}$$

$$C = \frac{TP \times TN - FP \times FN}{\sqrt{(TP \times FN)(TP \times FP)(TN \times FP)(TN \times FN)}}$$

Συντελεστής συσχέτισης του Matthews (C)

Ο συντελεστής αυτός, είναι ισοδύναμος του γνωστού συντελεστή συσχέτισης του Pearson όταν εφαρμοστεί σε δίτιμα δεδομένα και παίρνει τιμές από το -1 (τελείως αντίθετη πρόγνωση), έως το +1 (τέλεια πρόγνωση), με το 0 να αντιστοιχεί στην τελείως τυχαία πρόγνωση. Το μεγάλο πλεονέκτημα του συντελεστή συσχέτισης είναι ότι συνδυάζει όλες τις τιμές του πίνακα σε μία αριθμητική τιμή.

