Βιοπληροφορική Ι

Παντελής Μπάγκος
Αναπληρωτής Καθηγητής

Πανεπιστήμιο Θεσσαλίας
Λαμία, 2015

Διάλεξη 4

Hidden Markov Models (HMMs)

Μαρκοβιανά μοντέλα εξάρτησης

- Η πιθανότητα εμφάνισης ενός νουκλεοτιδίου είναι δεσμευμένη στο αμέσως προηγούμενό του
- Κατάλληλα μοντέλα για βιολογικές ακολουθίες, καθώς αντικατοπτρίζουν την έννοια της πληροφορίας που αυτές περιέχουν
- 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.

Το μοντέλο Markov (Markov Model)

Σε μια ακολουθία DNA $\mathbf{x} = x_1, x_2, ..., x_n \quad x_i \in \{A, T, G, C\}$

θεωρούμε ότι η εμφάνιση ενός νουκλεοτιδίου εξαρτάται απο 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:

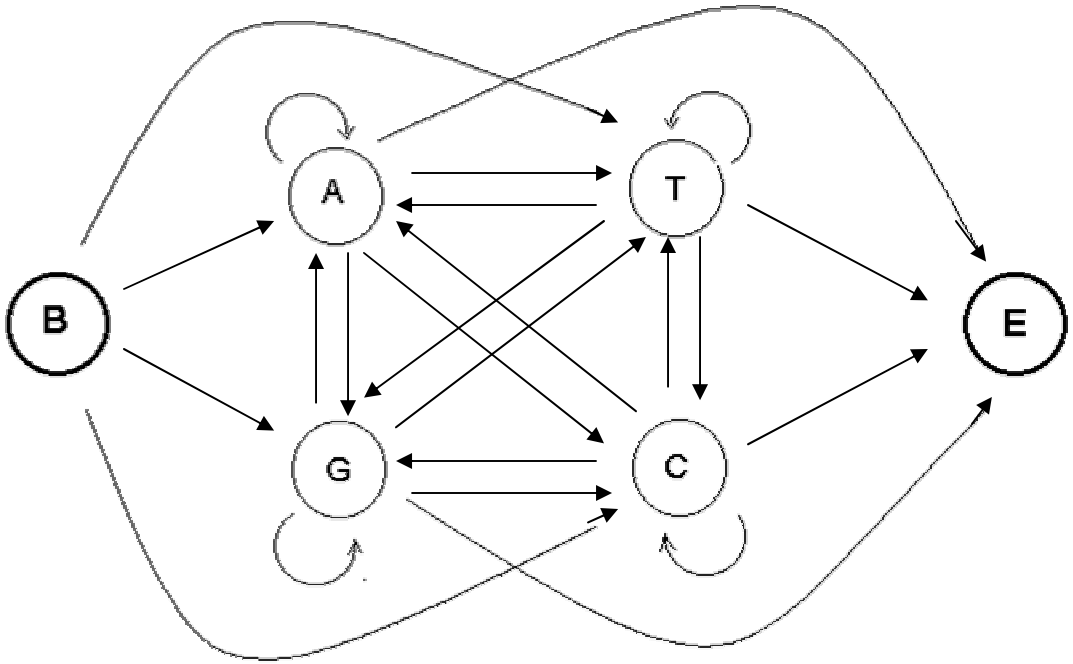
$$P(\mathbf{x}) = P(x_n, x_{n-1}, ..., x_1) = P(x_x \mid x_{n-1}, ..., x_1) P(x_{n-1} \mid x_{n-2}, ..., x_1) ... P(x_1)$$

και επειδή:

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

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

Το Markov Model διαγραμματικά



Ο πίνακας μεταβάσεων

		Θέση i			
		A	**C**	**G**	**T**
Θέση i-1	**A**	P_{AA}	P_{AC}	P_{AG}	P_{AT}
	C	P_{CA}	P_{CC}	P_{CG}	P_{CT}
	G	P_{GA}	P_{GC}	P_{GG}	P_{GT}
	T	P_{TA}	P_{TC}	P_{TG}	P_{TT}

Maximum Likelihood Estimates

$$P(\mathbf{x}) = P(x_k, x_{k-1}, ..., x_1) \prod_{i=k+1}^{L} P(x_i \mid x_{i-1}, x_{i-2}, ...x_{i-k})$$

$$\log P(\mathbf{x}) = \log P(x_k, x_{k-1}, ..., x_1) + \log \prod_{i=k+1}^{L} P(x_i \mid x_{i-1}, x_{i-2}, ...x_{i-k}) =$$

$$= \log P(x_k, x_{k-1}, ..., x_1) + \sum_{i=k+1}^{L} \log P(x_i \mid x_{i-1}, x_{i-2}, ...x_{i-k})$$

Θέτουμε τη μερική παράγωγο ως προς τις παραμέτρους ιση με μηδέν ικανοποιώντας τους περιορισμούς:

$$\hat{a}_{s_k ... s_1 s_0} = \frac{n_{s_k, ..., s_1, s_0}}{\sum_{s_k, ..., s_1, \forall s_0 \in Q}^{m} n_{s_k, ..., s_1, s_0}}$$

Εφαρμογή- LR test

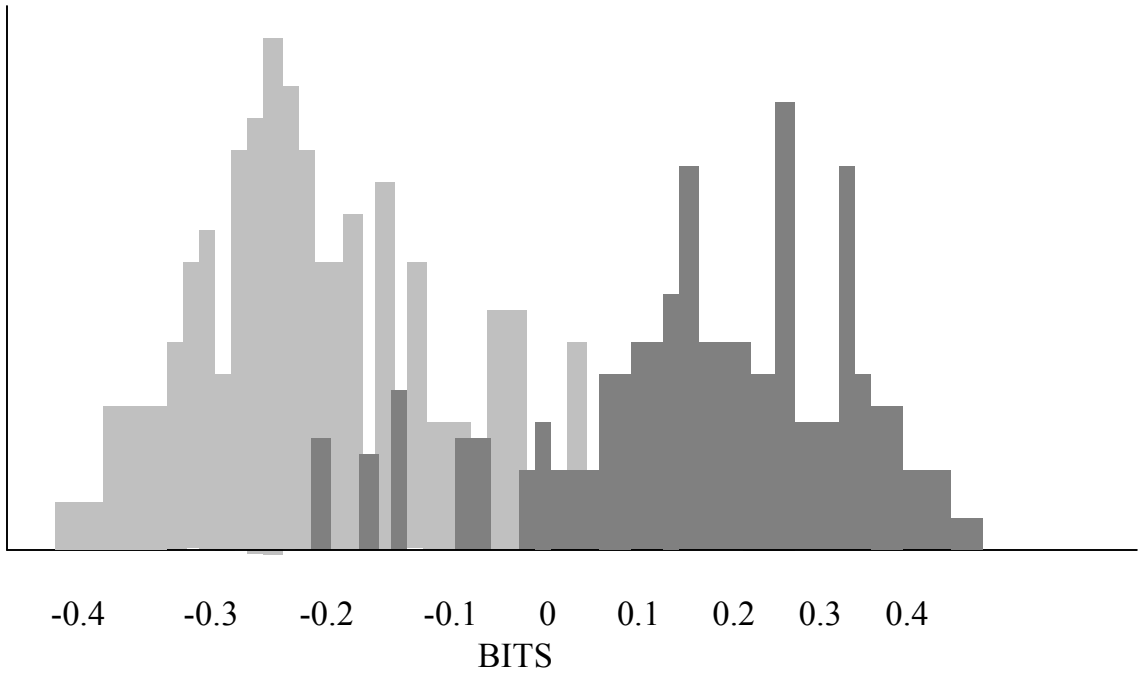
Έστω ότι έχουμε δύο μοντέλα (+) και (-), στα οποία υποθέτουμε ότι ισχύουν διαφορετικοί πίνακες μεταβάσεως. Υπολογίζουμε τους Ε.Μ.Π. για τις μεταβάσεις, ως εξής:

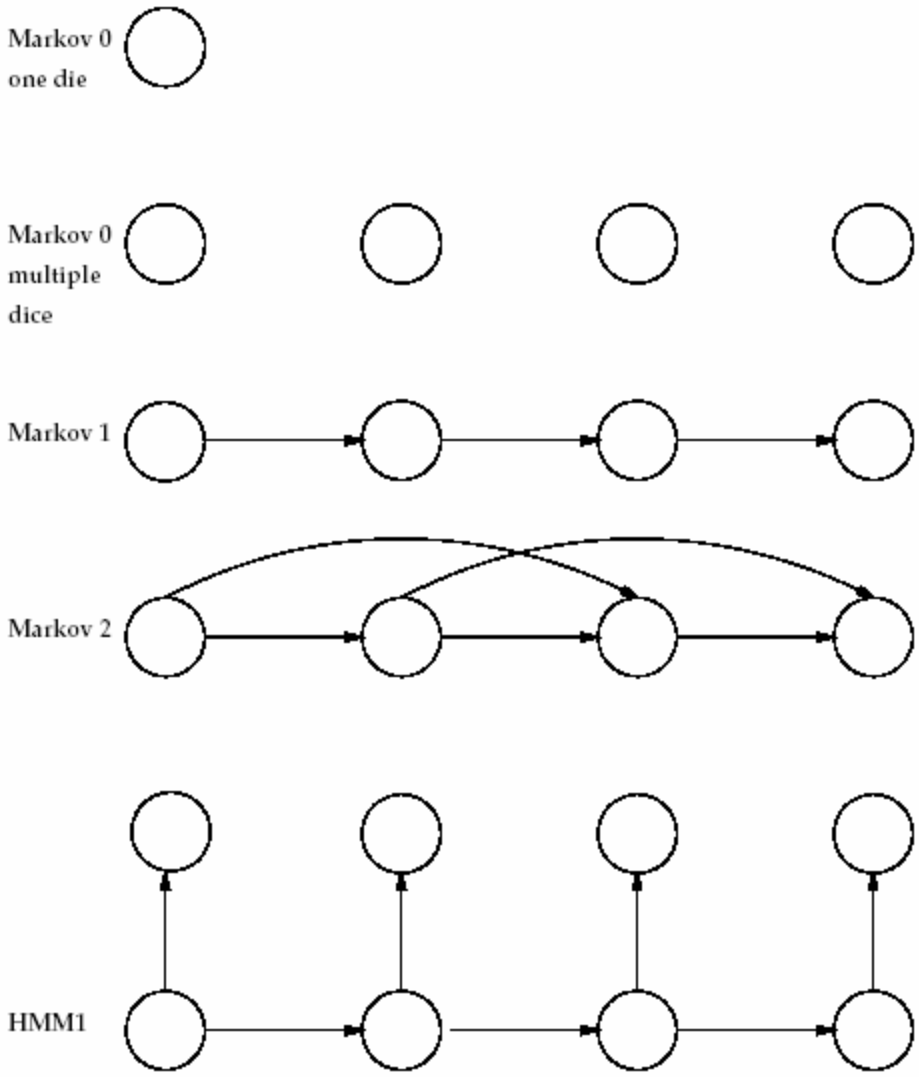
$$p^{+}_{ab} = \frac{c^{+}_{ab}}{\sum c^{+}_{ax}} \qquad p^{-}_{ab} = \frac{c^{-}_{ab}}{\sum c^{-}_{ax}}$$

Κατόπιν υπολογίζουμε τα log-odds, για κάθε νουκλεοτίδιο, και τα αθροίζουμε για ένα «παράθυρο», ή και για ολόκληρη την ακολουθία.

$$S(x) = \log \frac{P(x \mid +)}{P(x \mid -)} = \sum_i \log \frac{p^{+}_{x_{i-1}x_i}}{p^{-}_{x_{i-1}x_i}} = \sum_i \beta_{x_{i-1}x_i}$$

Εφαρμογή στην εύρεση γονιδίων- CpG islands




Markov 0
one die
Markov 0
multiple
dice
Markov 1
Markov 2
HMM1

Ανώτερης τάξης αλυσίδες Markov

- Μια k τάξης αλυσίδα ορίζεται πολύ απλά:

$$P(x_i \mid x_{i-1}, \ldots, x_1) = P(x_i \mid x_{i-1}, x_{i-2}, \ldots, x_{i-n}) = \alpha_{x_n \ldots x_{i-1} x_i}$$

- Μεγάλο πρόβλημα καθώς απαιτούνται εκθετικά αυξανόμενα αριθμός δεδομένων εκπαίδευσης [(20-1)*20^{k-1} παράμετροι]
- Ιδιαίτερα σε περιπτώσεις πρωτεϊνών όπου έχουμε μεγάλο αλφάβητο
- Ανάγκη άλλων προσεγγίσεων

Προσεγγίσεις Ανώτερης τάξης αλυσίδες Markov

- Variable Length Markov chains (VLMC)
- Mixture Transition Distribution (MTD)
- Parsimonious Markov chains (PMC)

Το Hidden Markov Model

Στην περίπτωση που οι δυο περιοχές διαδέχονται η μια την άλλη, μέσα στην ίδια ακολουθία, πιο «κομψό» μαθηματικά είναι το Hidden Markov Model.

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

2145264366365616662321 45

−−−−−−−−−++++++++++++−−−−−−

\+ = 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 αντιστοίχηση του αποτελέσματος, με την πραγματική κατάσταση.

Ο ορισμός του μοντέλου

Ένα Hidden Markov Model, είναι ένα μοντέλο ***M*** που περιέχει 3 στοιχεία ***Σ, Q, θ***.

$$\mathbf{M} = (\Sigma, Q, \theta)$$

- ***Σ***, το αλφάβητο των δυνατών ενδεχομένων (π.χ. 1,2,..6 για το ζάρι, A,T,G,C για το DNA, κλπ)
- ***Q***, το σύνολο των δυνατών καταστάσεων του μοντέλου (μεροληπτικό αμερόληπτο, για το ζάρι, γονίδιο – όχι γονίδιο για το DNA, κλπ)
- ***θ***, το σύνολο πιθανοτήτων που διέπουν το μοντέλο και μπορεί να είναι:
  1. Πιθανότητες μεταβάσεως (transitions) από κατάσταση σε κατάσταση, και
  2. Πιθανότητες γεννήσεως (emissions), με τις οποίες παράγονται τα σύμβολα σε κάθε κατάσταση.

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

Ορισμοί

Ακολουθία συμβόλων:

$$\mathbf{x} = x_1, x_2, \ldots, x_{L-1}, x_L$$

transition probabilities:

$$a_{kl} = P(\pi_i = l \mid \pi_{i-1} = k)$$

emission probabilities:

$$e_k(b) = P(x_i = b \mid \pi_i = k)$$

begin probabilities:

$$a_{Bk} = P(\pi_1 = k \mid B)$$

end probabilities:

$$a_{kE} = P(E \mid \pi_i = k)$$

Πιθανοφάνεια

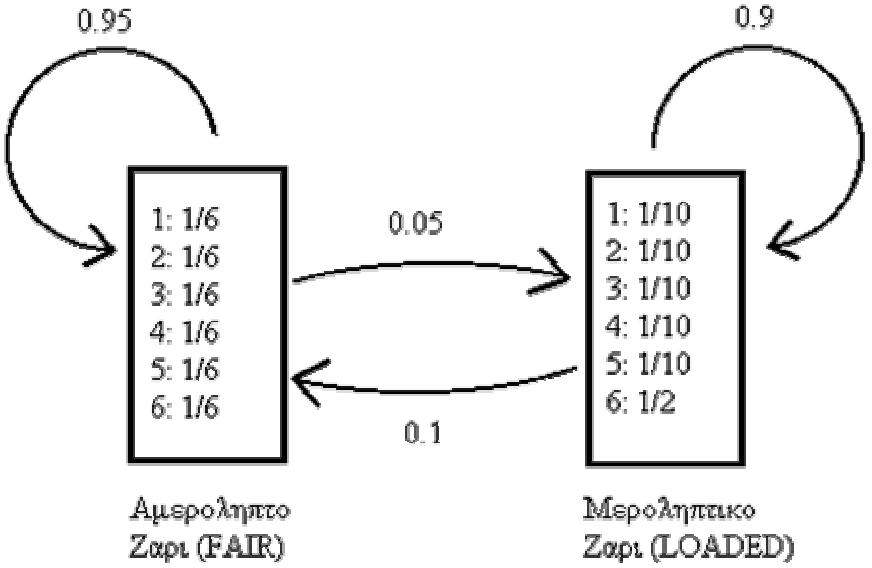
Η από κοινού πιθανότητα μιας ακολουθίας **x** και του μονοπατιού *π*

$$P(\mathbf{x},\pi) = P(x_L, x_{L-1},\ldots,x_1,\pi) = a_{B\pi_1}\prod_{i=1}^{L} e_{\pi i}(x_i)a_{\pi_i\pi_{i+1}}$$

Για να υπολογίσουμε την συνολική πιθανότητα, μιας ακολουθίας **x**, δεδομένου του μοντέλου, θα πρέπει να αθροίσουμε για όλες τις πιθανές αλληλουχίες καταστάσεων, δηλαδή να αθροίσουμε την συνεισφορά στη συνολική πιθανότητα όλων των πιθανών μονοπατιών *π*.

$$P(\mathbf{x}\,|\,\theta) = \sum_{\pi} P(\mathbf{x},\pi\,|\,\theta) = \sum_{\pi} a_{B\pi_1}\prod_{i=1}^{L} e_{\pi i}(x_i)a_{\pi_i\pi_{i+1}}$$

Το Hidden Markov Model σχηματικά



Το μεροληπτικό καζίνο

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 συγκεκριμμένη 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

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

… 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).

Αλγόριθμος Forward

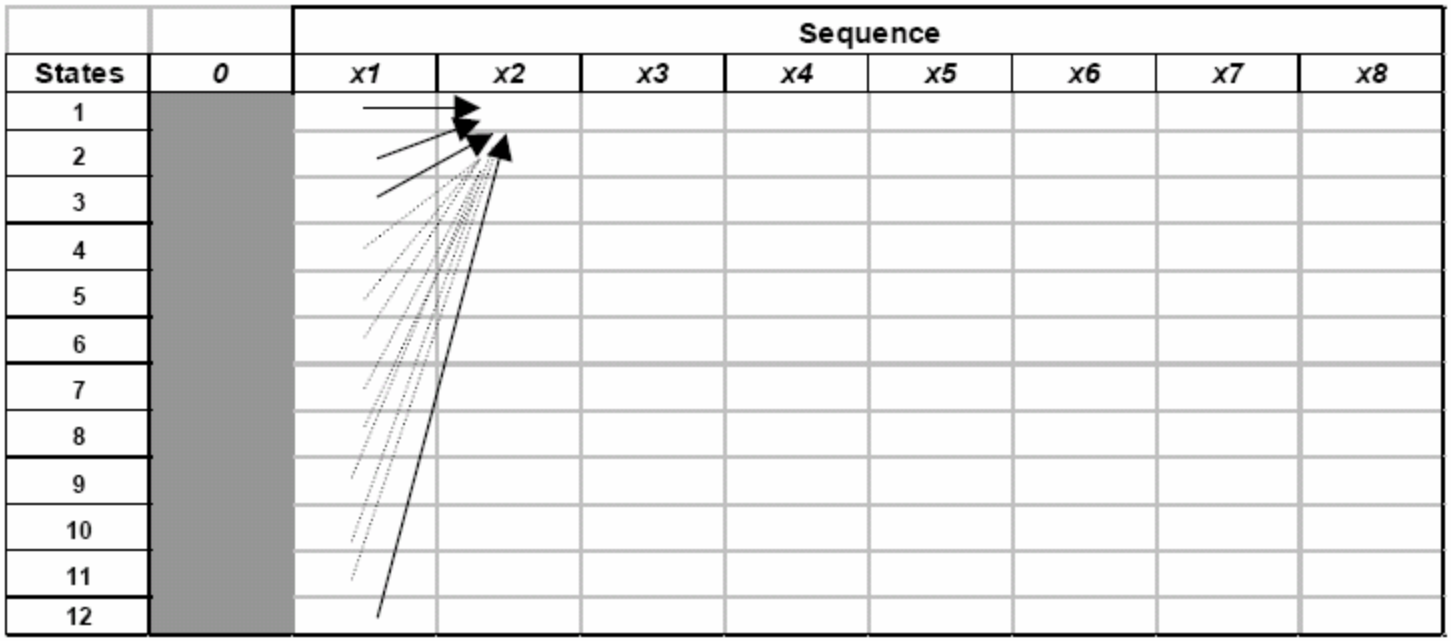
$$\forall k \neq B,\ i = 0 : f_B(0) = 1,\ f_k(0) = 0$$

$$\forall 1 \leq i \leq L : f_l(i) = e_l(x_i) \sum_k f_k(i-1) a_{kl}$$

$$P(\mathbf{x}|\theta) = \sum_k f_k(L) a_{kE}$$

Ο αλγόριθμος αυτός, κατασκευάζει έναν πίνακα με διαστάσεις *N(L+1)*, όπου *N* ο αριθμός των καταστάσεων και *L* το μήκος της ακολουθίας, και θεωρεί μια ενδιάμεση μεταβλητή $f_k(i)$ για κάθε θέση *i* και κατάσταση *k* της ακολουθίας. Η ποσότητα αυτή, πρακτικά, είναι ίση με την από κοινού πιθανότητα της αλληλουχίας έως το κατάλοιπο *i*, και του μονοπατιού που αντιστοιχεί στην κατάσταση *k*. Δηλαδή:

$$f_k(i) = P(x_1, x_2, ..., x_i, \pi_i = k)$$



Εικόνα 2.6 Διαγραμματική απεικόνιση του πίνακα Forward, για ένα υποθετικό μοντέλο με 12 καταστάσεις (states), και μια ακολουθία από 8 κατάλοιπα. Για τον υπολογισμό της τιμής ενός κελιού (π.χ. του $f_1(2)$), υπολογίζονται οι συνεισφορές όλων των προηγούμενων κελιών στη θέση 1 της ακολουθίας (βέλη).

Εντελώς ανάλογος είναι ο αλγόριθμος Backward (Durbin et al., 1998; Rabiner, 1989), ο οποίος διαφέρει μόνο ως προς την κατεύθυνση προς την οποία διατρέχει την αλληλουχία. Η ενδιάμεση μεταβλητή που χρησιμοποιείται, ονομάζεται πλέον $b_k(i)$, και ορίζεται για κάθε i ως η πιθανότητα της ακολουθίας από την θέση i+1 έως το τέλος, δεδομένου ότι στη την θέση i συναντάμε την κατάσταση k. Δηλαδή:

$$b_k(i) = P(x_{i+1},...,x_L \mid \pi_i = k)$$

Άρα ο αλγόριθμος, διατυπώνεται ως εξής:

Αλγόριθμος Backward

$$\forall k,\ i = L:\ b_k(L) = a_{kE}$$

$$\forall\ 1 \le i < L:\ b_k(i) = \sum_l a_{kl} e_l(x_{i+1}) b_l(i+1)$$

$$P(\mathbf{x}|\theta) = \sum_l a_{Bl} e_l(x_1) b_l(1)$$

Όμοια, αν δεν υπάρχουν καταστάσεις λήξεως, στην αρχικοποίηση, οι αντίστοιχες πιθανότητες τίθενται ίσες με 1. Το τελικό αποτέλεσμα του αλγορίθμου, είναι ακριβώς όμοιο με αυτό του Forward.

Αλγόριθμος Viterbi

Αλγόριθμος Viterbi

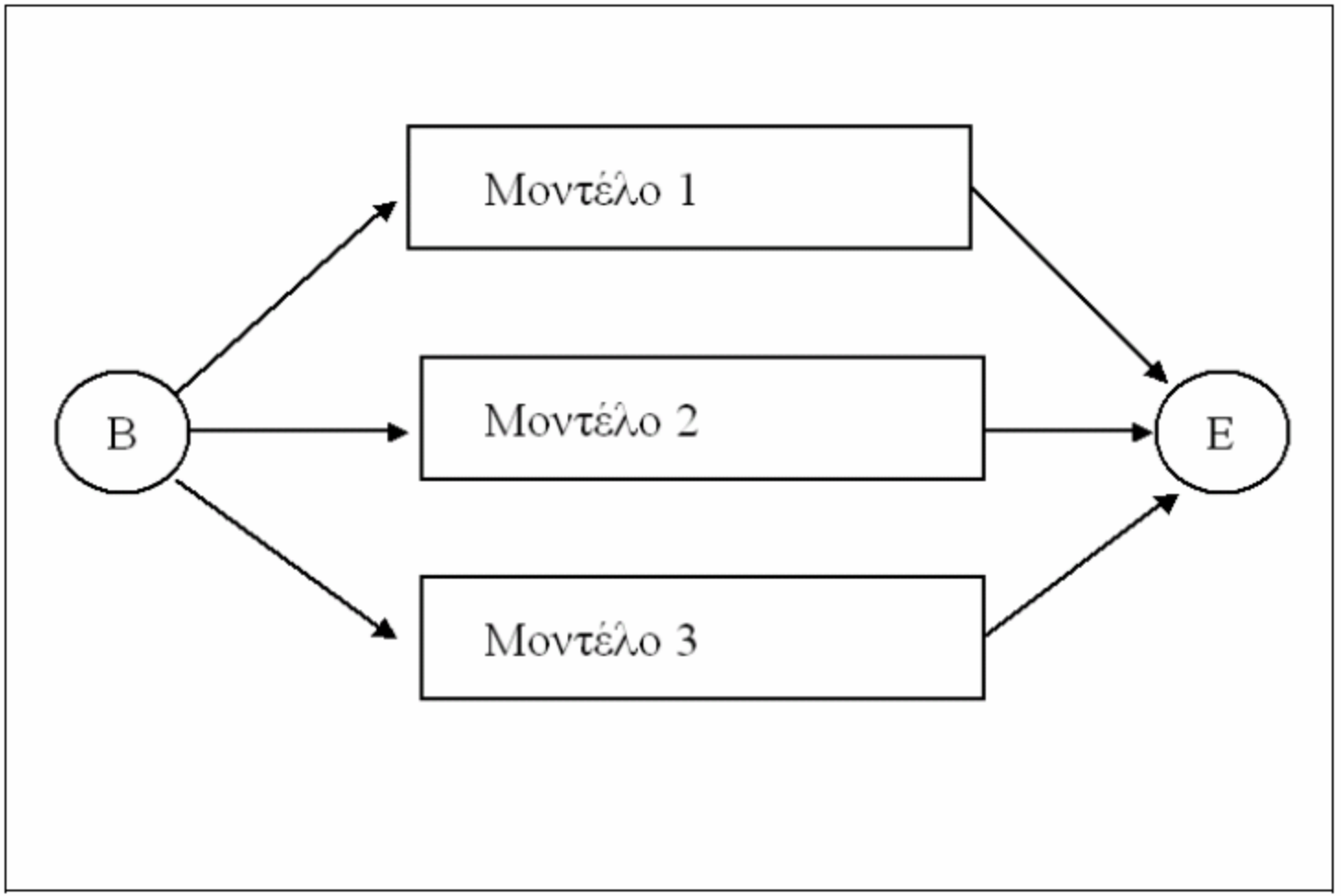
$$\forall k \neq B,\ i = 0:\ u_B(0) = 1,\ u_k(0) = 0$$

$$\forall\ 1 \leq i \leq L:\ u_l(i) = e_l(x_i) \max_k \{u_k(i-1) a_{kl}\}$$

$$P(\mathbf{x}, \pi^{\max} \mid \theta) = \max_k \{u_k(L) a_{kE}\}$$

Ο αλγόριθμος του Viterbi, είναι στην ουσία όμοιος με τον Forward, με τη μόνη διαφορά να βρίσκεται στο ότι τα διαδοχικά αθροίσματα αντικαθίστανται από μεγιστοποιήσεις. Σε αυτή την περίπτωση με π^{max}, συμβολίζουμε το μονοπάτι με τη μεγαλύτερη πιθανότητα και η πιθανότητα αυτή συμβολίζεται με $P(\mathbf{x}, \pi^{max}|\theta)$. Προφανώς, ισχύει ότι $P\left(\mathbf{x}, \pi^{\max} \mid \theta\right) \leq P\left(\mathbf{x} \mid \theta\right)$. Ένα επιπλέον χαρακτηριστικό του αλγόριθμου αυτού, είναι το ότι απαιτεί την ύπαρξη ενός ξεχωριστού πίνακα στον οποίο θα κρατούνται δείκτες (pointers), για την καλύτερη (πιθανότερη) κατάσταση σε κάθε θέση της ακολουθίας. Με αναδρομή (back-tracking), σε αυτόν τον πίνακα, ανακτά κανείς στο τέλος, το ίδιο το πιθανότερο μονοπάτι.

Αποκωδικοποίηση forward



“Εκ των υστέρων” αποκωδικοποίηση

Εναλλακτικά μπορεί να υπολογισθεί η πιθανότητα: $P(\pi_i = k \mid \mathbf{x})$
δηλαδή, η εκ των υστέρων πιθανότητα το συγκεκριμμένο νουκλεοτίδιο να προήλθε απο μια κατάσταση

$$P(\mathbf{x}, \pi_i = k) = P(x_1, x_2, ..., x_i, \pi_i = k) P(x_{i+1}, ..., x_n \mid x_1, ..., x_i, \pi_i = k)$$
$$= P(x_1, x_2, ..., x_i, \pi_i = k) P(x_{i+1}, ..., x_n \mid \pi_i = k)$$

Κάνοντας χρήση των Forward και Backward:

$$P(\mathbf{x}, \pi_i = k) = f_k(i) b_k(i)$$

Τέλος, σύμφωνα με το θεώρημα Bayes θα έχουμε:

$$P(\pi_i = k \mid \mathbf{x}) = \frac{f_k(i) b_k(i)}{P(\mathbf{x})}$$

Με τον τύπο αυτό, μπορούμε να υπολογίσουμε την πιθανότητα μια παρατήρηση να προέρχεται από μια συγκεκριμένη κατάσταση. Μπορούμε επίσης να ορίσουμε μια άλλη αλληλουχία καταστάσεων, για την οποία ισχύει:

$$\hat{\pi}_i = \arg\max_k P(\pi_i = k \mid \mathbf{x})$$

Πλεονεκτήματα:

• στις περιπτώσεις που τα εναλλακτικά μονοπάτια έχουν πολύ μικρές διαφορές στις προβλεπόμενες πιθανότητες.

• όταν μια κατάσταση έχει πολύ μικρή πιθανότητα και το μονοπάτι με την μέγιστη πιθανότητα, δεν την «επισκέπτεται» ποτέ.

Μειονεκτήματα:

• Μπορεί να προβλεφθεί μια πιθανότητα η οποία δεν είναι έγκυρη για το μοντέλο (μια μη επιτρεπτή μετάβαση).

Αλγόριθμος Baum-Welch

Στην ιδανική (όσο και ανέφικτη) περίπτωση, κατά την οποία γνωρίζουμε τα ακριβή μονοπάτια για τις ακολουθίες εκπαίδευσης, ο υπολογισμός των ΕΜΠ είναι αρκετά απλός. Συγκεκριμένα, δεν έχουμε παρά να καταμετρήσουμε πόσες φορές παρατηρήθηκε μια συγκεκριμένη μετάβαση από κάθε κατάσταση, και πόσες φορές ένα αμινοξύ εμφανίστηκε σε κάθε κατάσταση. Άρα οι ΕΜΠ, για τις πιθανότητες μετάβασης θα είναι:

$$\hat{a}_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}}$$

και για τις πιθανότητες γεννήσεως,

$$\hat{e}_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b)}$$

όπου τα αθροίσματα στους παρονομαστές, εκτείνονται σε όλο το εύρος των παραμέτρων.

$$l(\mathbf{x},\theta) = \log P(\mathbf{x},\theta) = \sum_{\pi} \log P(\mathbf{x},\pi \mid \theta)$$

Επειδή από το θεώρημα του Bayes είναι γνωστό ότι:

$$P(\pi \mid \mathbf{x},\theta) = \frac{P(\mathbf{x},\pi \mid \theta)}{P(\mathbf{x} \mid \theta)}$$

θα έχουμε:

$$\log P(\mathbf{x} \mid \theta) = \log P(\mathbf{x},\pi \mid \theta) - \log P(\pi \mid \mathbf{x},\theta)$$

Τότε αν πολλαπλασιάσουμε με $P(\pi \mid \mathbf{x},\theta^{\mathbf{t}})$, και αθροίσουμε για όλα τα πιθανά μονοπάτια π, θα έχουμε:

$$\log P(\mathbf{x} \mid \theta) = \sum_{\pi} P(\pi \mid \mathbf{x},\theta^{\mathbf{t}}) \log P(\mathbf{x},\pi \mid \theta) - \sum_{\pi} P(\pi \mid \mathbf{x},\theta^{t}) \log P(\pi \mid \mathbf{x},\theta)$$

Τον πρώτο όρο του παραπάνω αθροίσματος τον ονομάζουμε:

$$Q(\theta \mid \theta^{t}) = \sum_{\pi} P(\pi \mid \mathbf{x},\theta^{\mathbf{t}}) \log P(\mathbf{x},\pi \mid \theta)$$

Για να μεγιστοποιηθεί η πιθανοφάνεια, θέλουμε :

$$\log P(\mathbf{x} \mid \theta) \geq \log P(\mathbf{x} \mid \theta^{t})$$

για κάθε σετ παραμέτρων θ, άρα:

$$\log P(\mathbf{x} \mid \theta) - \log P(\mathbf{x} \mid \theta^{t}) = Q(\theta \mid \theta^{t}) - Q(\theta^{t} \mid \theta^{t}) + \sum_{\pi} P(\pi \mid \mathbf{x},\theta^{t}) \log \frac{P(\pi \mid \mathbf{x},\theta^{t})}{P(\pi \mid \mathbf{x},\theta)}$$

και επειδή ο τελευταίος όρος είναι η σχετική εντροπία και είναι πάντα θετικός εκτός αν $\theta=\theta^t$ θα έχουμε:

$$\log P(\mathbf{x} \mid \theta) - \log P(\mathbf{x} \mid \theta^t) \geq Q(\theta \mid \theta^t) - Q(\theta^t \mid \theta^t)$$

Τότε αν διαλέξουμε το σύνολο των παραμέτρων που μεγιστοποιεί τη συνάρτηση Q, δηλαδή:

$$\theta^{t+1} = \arg\max_{\theta} Q(\theta \mid \theta^t)$$

$$P(\mathbf{x}, \pi \mid \theta) = \prod_{k=1} \prod_{b} \left[e_k(b)\right]^{E_k(b,\pi)} \prod_{k=0} \prod_{l=1} a_{kl}^{A_{kl}(\pi)}$$

όπου, $E_k(b,\pi)$ και $A_{kl}(\pi)$, είναι οι συνολικές εμφανίσεις του συμβόλου b, και των μεταβάσεων στην κατάσταση l αντίστοιχα, από την κατάσταση k, σε ένα μονοπάτι π.

Αντικαθιστώντας

$$Q(\theta \mid \theta^t) = \sum_{\pi} P(\pi \mid \mathbf{x}, \theta^{\mathrm{t}}) \left[\sum_{k=1} \sum_{b} E_k(b,\pi) \log e_k(b) + \sum_{k=0} \sum_{l=1} A_{kl}(\pi) \log \alpha_{\mathrm{kl}} \right]$$

Θα δείξουμε παρακάτω, ότι οι αναμενόμενες τιμές $E_k(b)$ και A_{kl} των παραμέτρων, αθροιζόμενες για όλα τα μονοπάτια, μπορούν να εκφραστούν σαν συνάρτηση των μεταβλητών $f_k(i), b_k(i)$, ποσότητες που υπολογίζονται από τους αλγορίθμους forward και backward που είδαμε παραπάνω.

Υπολογισμός των Expected counts

$$P(\mathbf{x}, \pi_i = k, \pi_{i+1} = l \mid \theta) = P(x_1, x_2, ..., x_L, \pi_i = k, \pi_{i+1} = l \mid \theta) =$$

$$= P(x_1, x_2, ..., x_i, \pi_i = k \mid \theta) P(x_{i+1}, x_{i+2}, ..., x_L, \pi_{i+1} = l \mid x_1, x_2, ..., x_i, \pi_i = k, \theta)$$

και επειδή δεν υπάρχει εξάρτηση ούτε των παρατηρήσεων ούτε των καταστάσεων από προηγούμενες παρατηρήσεις, καταλήγουμε:

$$P(\mathbf{x}, \pi_i = k, \pi_{i+1} = l \mid \theta) =$$

$$P(x_1, x_2, ..., x_i, \pi_i = k \mid \theta) P(x_{i+1}, x_{i+2}, ..., x_L, \pi_{i+1} = l \mid \pi_i = k, \theta)$$

βλέπουμε ότι:

$$f_k(i) = P(x_1, x_2, ..., x_i, \pi_i = k)$$

Επιπλέον,

$$P(x_{i+1}, x_{i+2}, ..., x_L, \pi_{i+1} = l \mid \pi_i = k, \theta) =$$

$$P(x_{i+1}, \pi_{i+1} = l \mid \pi_i = k, \theta) P(x_{i+2}, ..., x_L \mid x_{i+1}, \pi_{i+1} = l, \pi_i = k, \theta)$$

Ο πρώτος όρος του γινομένου, γίνεται:

$$P(x_{i+1}, \pi_{i+1} = l \mid \pi_i = k, \theta) =$$
$$= P(\pi_{i+1} = l \mid \pi_i = k) P(x_{i+1} \mid \pi_{i+1} = l) =$$
$$= a_{kl} e_l(x_{i+1})$$

ενώ ο δεύτερος,

$$P(x_{i+2}, \ldots, x_L \mid x_{i+1}, \pi_{i+1} = l, \pi_i = k, \theta) =$$
$$= P(x_{i+2}, \ldots, x_L \mid \pi_{i+1} = l) = b_l(i+1)$$

Αντικαθιστώντας

$$P(x_{i+1}, x_{i+2}, \ldots, x_L, \pi_{i+1} = l \mid \pi_i = k, \theta) = a_{kl} e_l(x_{i+1}) b_l(i+1)$$

και τελικά:

$$P(\mathbf{x}, \pi_i = k, \pi_{i+1} = l \mid \theta) = f_k(i) a_{kl} e_l(x_{i+1}) b_l(i+1)$$

απ' όπου με χρήση του θεωρήματος του Bayes:

$$P(\pi_i = k, \pi_{i+1} = l \mid \mathbf{x}, \theta) = \frac{f_k(i) a_{kl} e_l(x_{i+1}) b_l(i+1)}{P(\mathbf{x})}$$

όπου $f_k^j(i), b_l^j(i)$ είναι οι ποσότητες που υπολογίζονται από τους αλγορίθμους forward και backward. Με όμοιο τρόπο,

Είδαμε επίσης ότι ισχύει:

$$P(\pi_i = k \mid \mathbf{x}) = \frac{f_k(i) b_k(i)}{P(\mathbf{x})}$$

Τότε, από τον ορισμό της αναμενόμενης τιμής, έχουμε για τις μεταβάσεις:

$$A_{kl} = \sum_{\pi} P(\pi|\mathbf{x}, \theta^t) A_{kl}(\pi) = \frac{1}{P(\mathbf{x})} \sum_i f_k(i) a_{kl} e_l(x_{i+1}) b_l(i+1)$$

και αντίστοιχα για τις πιθανότητες γεννήσεως:

$$E_k(b) = \sum_{\pi} P(\pi|\mathbf{x}, \theta^t) E_k(b, \pi) = \frac{1}{P(\mathbf{x})} \sum_{\{i|x_i=b\}} f_k^j(i) b_k^j(i)$$

και αντικαθιστώντας

$$Q(\theta \mid \theta^t) = \sum_{k=1} \sum_b E_k(b) \log e_k(b) + \sum_{k=0} \sum_{l=1} A_{kl} \log \alpha_{kl}$$

Η συνάρτηση Q μεγιστοποιείται από τους ΕΜΠ (έχει τα ίδια ακρότατα με τη log-likelihood)

Συνοπτικά ο αλγόριθμος

- Υπολογισμός των Α και Ε
- Υπολογισμός των ΕΜΠ
- Επανάληψη μέχρι να συγκλίνει

Χαρακτηριστικά του αλγορίθμου Baum-Welch

- Είναι υπολογιστικά απλός
- Είναι σίγουρο ότι συγκλίνει
- Δεν χρειάζεται επιλογή παραμέτρων

Αλλά επίσης:

- Δεν συγκλίνει πάντα στο ολικό μέγιστο
- Δεν είναι smooth (αν μια παράμετρος μηδενιστεί δεν αλλάζει ξανά τιμή)
- Δεν λειτουργεί on-line (για κάθε κατάλοιπο ξεχωριστά)

Gradient Descent Training

Η μέθοδος Gradient-Descent, είναι μια γενική ευριστική μέθοδος ελαχιστοποίησης ενέργειας. Αν θεωρήσουμε μια συνάρτηση, f με n μεταβλητές:

$$f(\mathbf{x}) = f(x_1, x_2, ..., x_n)$$

η οποία είναι παραγωγίσιμη, τότε ένα τοπικό της ελάχιστο, στο πολυδιάστατο σημείο

$$(\mathbf{x}^0) = (x_1{}^0, x_2{}^0, ..., x_n{}^0)$$

μπορεί να προσδιοριστεί, προσεγγίζοντας διαδοχικά το σημείο μέσω της σχέσης:

$$f(\mathbf{x}^{t+1}) = f(\mathbf{x}^t) - \eta \Delta f(\mathbf{x})$$

όπου, Δ είναι το διάνυσμα των μερικών παραγώγων της συνάρτησης και η ένας αρκετά μικρός ρυθμός μάθησης (learning rate). Στην περίπτωση μας, ως «ενέργεια» μπορεί να οριστεί το αντίθετο του λογαρίθμου της πιθανοφάνειας (negative log-likelihood), ενώ οι παράμετροι είναι φυσικά το σύνολο των πιθανοτήτων μεταβάσεως και γεννήσεως. Για κάθε παράμετρο ω του μοντέλου, η ανανέωση επιτυγχάνεται θέτοντας:

$$\omega^{t+1} = \omega^t - \eta \frac{\partial \ell(\mathbf{x} \mid \theta)}{\partial \omega}$$

Υπολογισμός μερικών παραγώγων

$$\frac{\partial \log P(\mathbf{x}|\theta)}{\partial \omega} = \frac{1}{P(\mathbf{x}|\theta)} \frac{\partial P(\mathbf{x}|\theta)}{\partial \omega}$$

$$= \frac{1}{P(\mathbf{x}|\theta)} \frac{\partial P(\mathbf{x},\pi|\theta)}{\partial \omega}$$

$$= \frac{1}{P(\mathbf{x}|\theta)} \sum_{\pi} P(\mathbf{x},\pi|\theta) \frac{\partial \log P(\mathbf{x},\pi|\theta)}{\partial \omega}$$

$$= \sum_{\pi} P(\pi|\mathbf{x},\theta) \frac{\partial \log P(\mathbf{x},\pi|\theta)}{\partial \omega}$$

$$\frac{\partial \log P(\mathbf{x},\pi|\theta)}{\partial a_{kl}} = \frac{\partial \left(\sum_{k=0} \sum_{l=1} A_{kl}(\pi) \log a_{kl} \right)}{\partial a_{kl}}$$

$$= A_{kl}(\pi) \frac{\partial \left(\sum_{k=0} \sum_{l=1} \log a_{kl} \right)}{\partial a_{kl}}$$

$$= \frac{A_{kl}(\pi)}{a_{kl}}$$

Άρα για τα transitions, έχουμε:

$$\frac{\partial \log P(\mathbf{x}\,|\,\theta)}{\partial a_{kl}} = \sum_{\pi} P(\pi\,|\,\mathbf{x},\theta)\frac{A_{kl}(\pi)}{a_{kl}} = \frac{A_{kl}}{a_{kl}}$$

και για τα emissions:

$$\frac{\partial \log P(\mathbf{x}\,|\,\theta)}{\partial e_k(b)} = \sum_{\pi} P(\pi\,|\,\mathbf{x},\theta)\frac{E_k(\pi,b)}{e_k(b)} = \frac{E_k(b)}{e_k(b)}$$

Παρατηρούμε ότι οι μερικές παράγωγοι της likelihood ως προς τις παραμέτρους του μοντέλου είναι ίδιες με αυτές της συνάρτησης Q

Normalization

- Αναγκαίο βήμα έτσι ώστε οι πιθανότητες να είναι μεταξύ 0-1
- Softmax μετασχηματισμός

$$a_{kl} = \frac{\exp(z_{kl})}{\sum_{l'} \exp(z_{kl'})}$$

Πραγματοποιώντας τώρα, την ελαχιστοποίηση με τη μέθοδο Gradient Decent, όχι στα a_{kl}, αλλά στα z_{kl}:

$$z_{kl}^{t+1} = z_{kl}^{t} - \eta \frac{\partial \ell^{t}}{\partial z_{kl}}$$

παίρνουμε τις ανανεωμένες παραμέτρους για τις πιθανότητες μετάβασης:

$$a_{kl}^{(t+1)} = \frac{a_{kl}^{(t)} \exp\left(-\eta \frac{\partial \ell^{(t)}}{\partial z_{kl}}\right)}{\sum_{l'} a_{kl'}^{(t)} \exp\left(-\eta \frac{\partial \ell^{(t)}}{\partial z_{kl'}}\right)}$$

$$a_{kl}^{(t+1)} = \frac{a_{kl}^{(t)} \exp\left(-\eta \frac{\partial \ell^{(t)}}{\partial z_{kl}}\right)}{\sum_{l'} a_{kl'}^{(t)} \exp\left(-\eta \frac{\partial \ell^{(t)}}{\partial z_{kl'}}\right)}$$

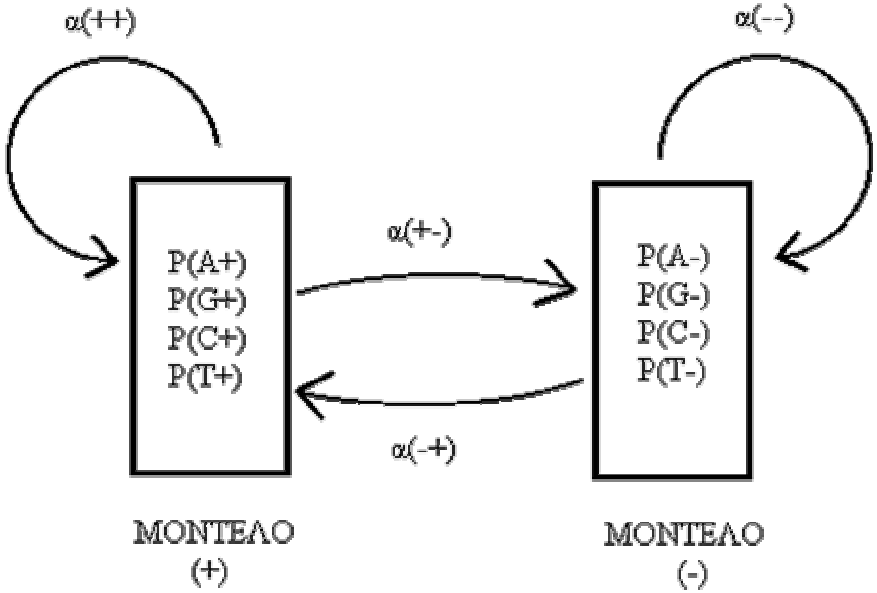
Με αλλαγή μεταβλητής μπορούμε να υπολογίσουμε επίσης τις μερικές παραγώγους του αντίθετου του λογαρίθμου της πιθανοφάνειας ως προς τις βοηθητικές παραμέτρους z_{kl}:

$$\frac{\partial \ell}{\partial z_{kl}} = -\left[A_{kl} - a_{kl} \sum_{l'} A_{kl'} \right]$$

Αντικαθιστώντας, παίρνουμε μια έκφραση η οποία εξαρτάται πλέον μόνο από τις τιμές των παραμέτρων στην προηγούμενη επανάληψη και από τις αναμενόμενες τιμές τους:

$$\alpha_{kl}^{(t+1)} = \frac{\alpha_{kl}^{(t)} \exp\left(-\eta \left[A_{kl} - a_{kl} \sum_{l'} A_{kl'} \right]\right)}{\sum_{l'} \alpha_{kl'}^{(t)} \exp\left(-\eta \left[A_{kl} - a_{kl} \sum_{l'} A_{kl'} \right]\right)}$$

Ένα παράδειγμα...



συνέχεια...

Πιθανότητες μεταβάσεως:

	1	0
1	0.90	0.10
0	0.10	0.90

Πιθανότητες γεννήσεως :

	A	T	G	C
1	0.70	0.10	0.10	0.10
0	0.25	0.25	0.25	0.25

συνέχεια...

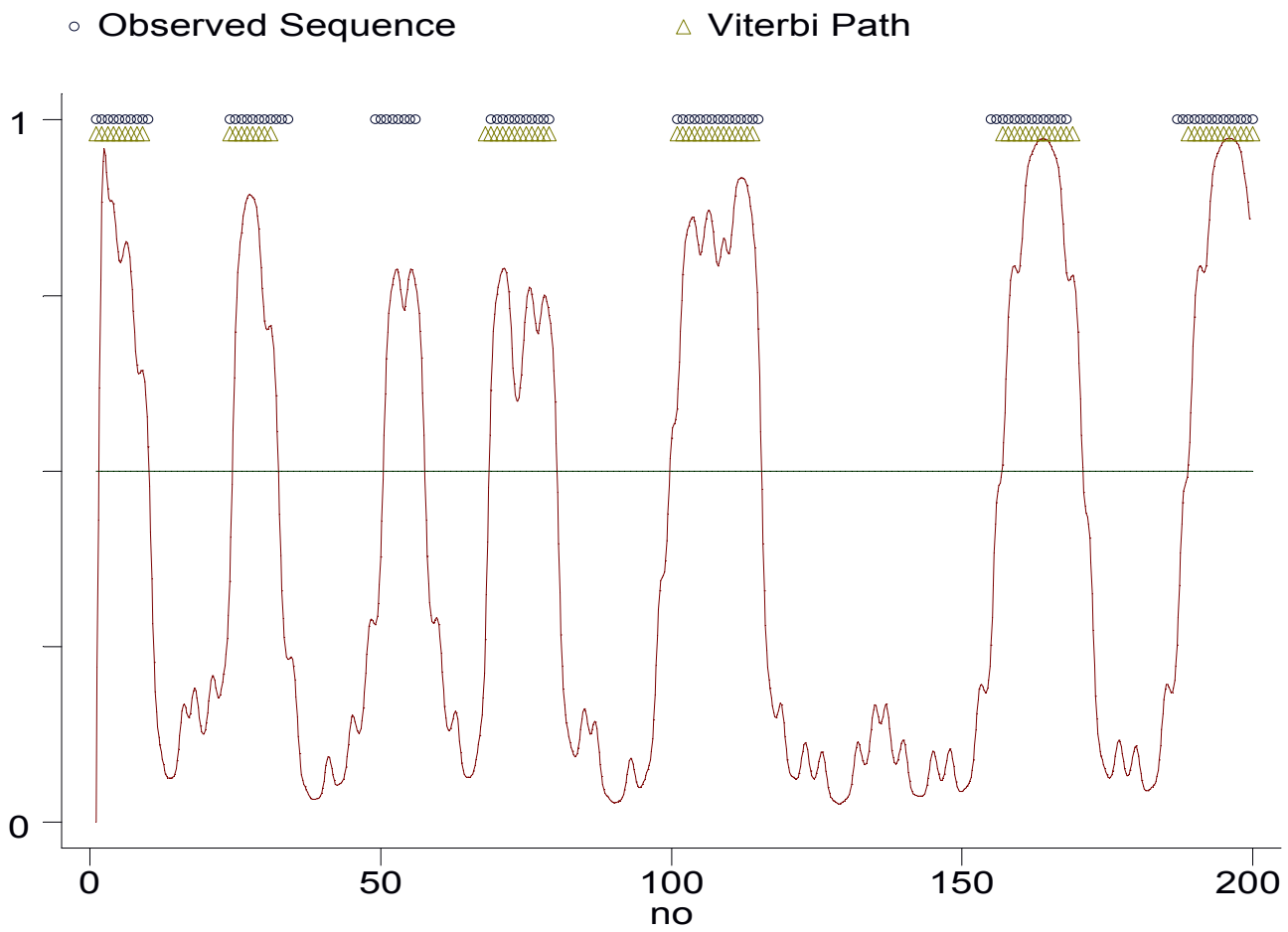
Έστω μια ακολουθία DNA, η οποία προέρχεται από το παραπάνω μοντέλο:

```
AAACAAGAATGCGCACACTACGCAAAAACAATTAGTCGCACTCACGATGAAACAAATTACCACGGTGAA
111111111100000000000011111111110000000000000111111100000000000001

AACGAATAAACCTCAGAGGCCCAGCGTATATAAACAAGATAAAAACCTAGTCAGCACTCTGACCAGACG
1111111110000000000000000000011111111111111000000000000000000000000

AGCTCACGACTTGAGGATAAGAAAAAAACAACAGCTCACGACTTGAGGATAAGAAAAAAACA
000000000000000011111111111111000000000000000001111111111111
```

συνέχεια...



συνέχεια…

Αν όμως οι πιθανότητες μεταβάσεως άλλαζαν:

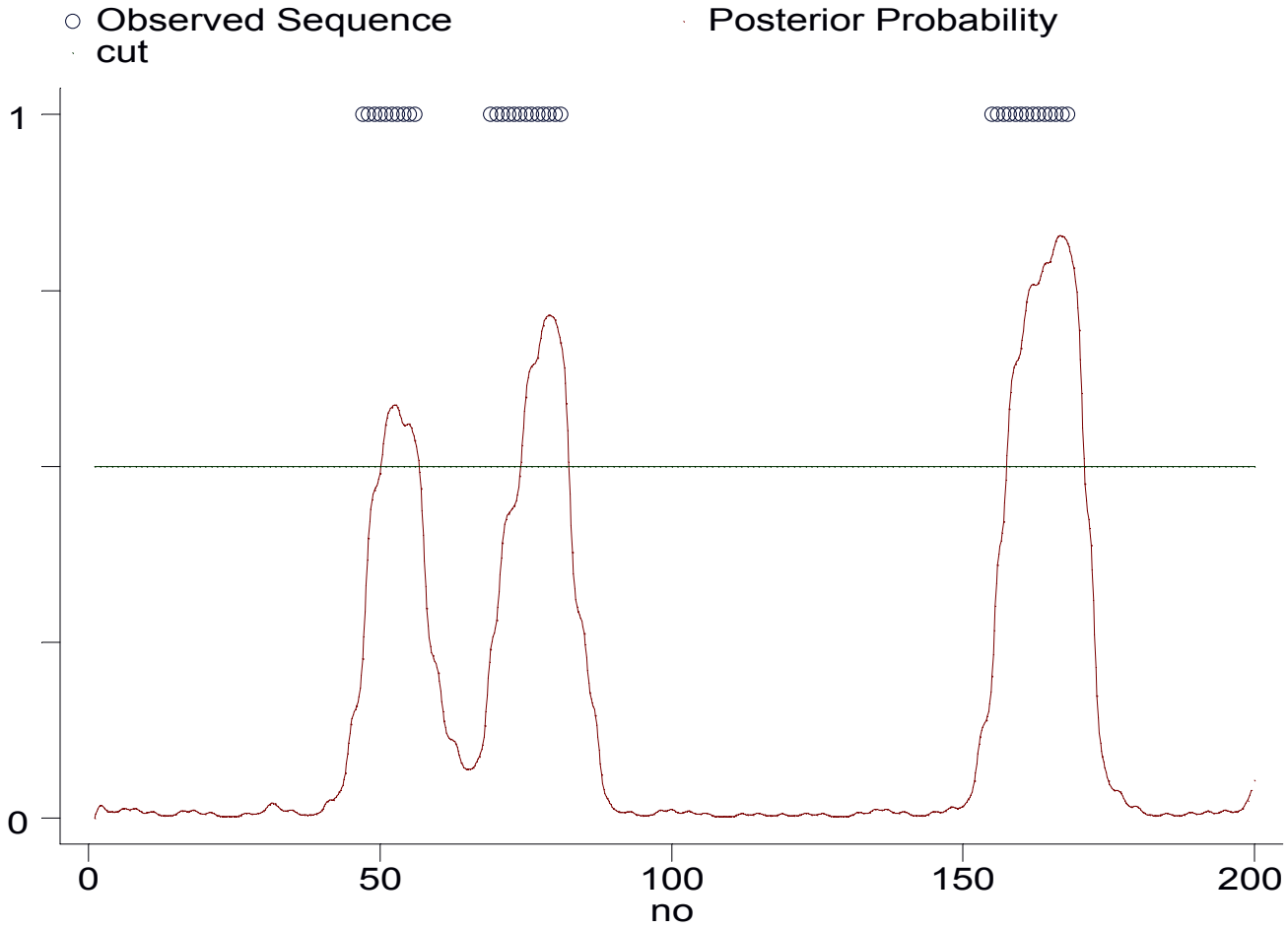
Πιθανότητες μεταβάσεως:

	1	0
1	0.98	0.02
0	0.03	0.97

Πιθανότητες γεννήσεως :

	A	T	G	C
1	0.60	0.10	0.10	0.10
0	0.25	0.25	0.25	0.25

συνέχεια...


Observed Sequence
cut
Posterior Probability
1
0
0
50
100
150
200
no

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

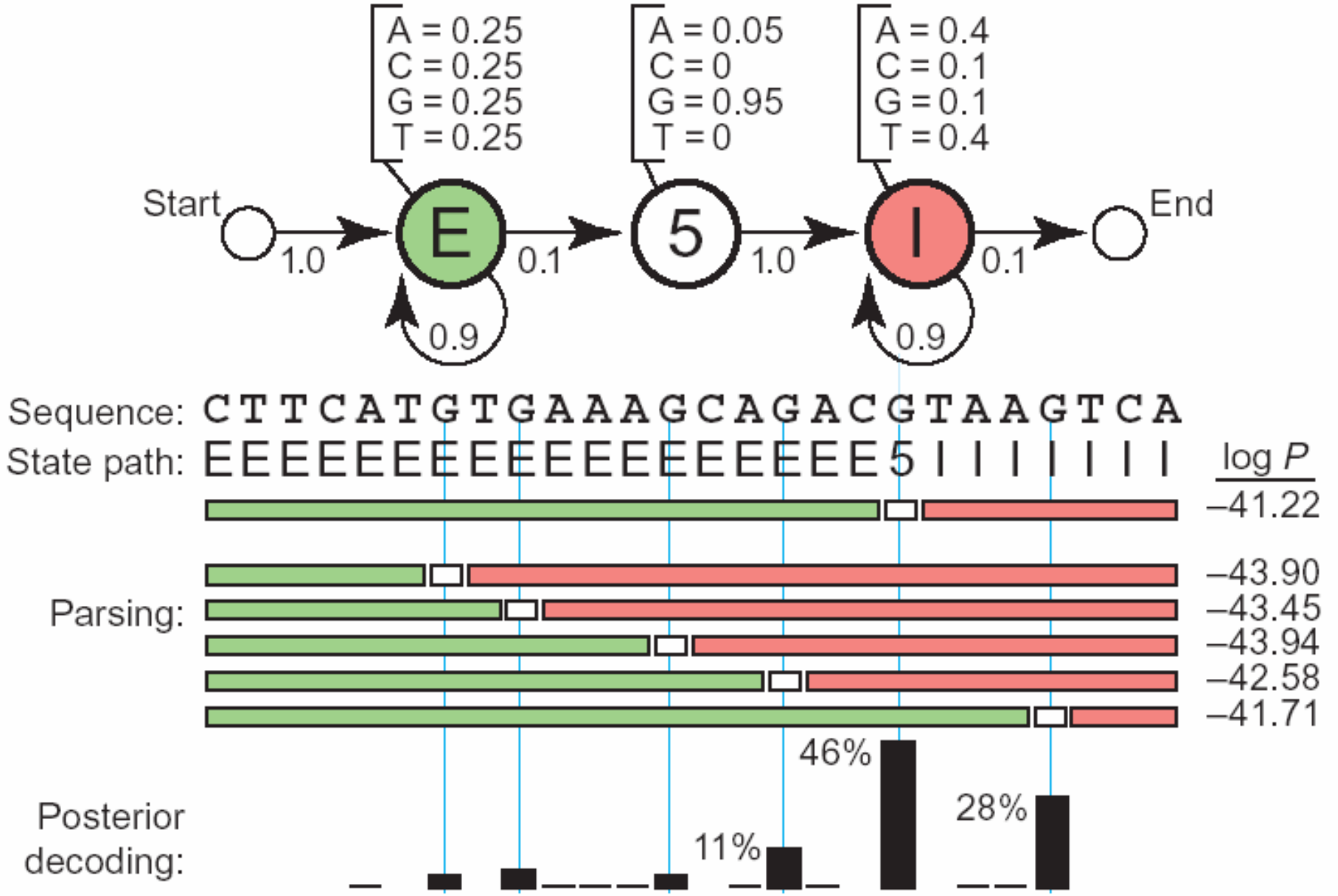


Figure 1 A toy HMM for 5′ splice site recognition. See text for explanation.

Labeled Sequences

- Αναγκαίο βήμα αν πρόκειται να κατασκευάσουμε μεγάλο μοντέλο με καταστάσεις οι οποίες ομαδοποιούνται
- Απαιτούνται αλλαγές στους αλγορίθμους

Labeled sequences

Συγκεκριμένα, με την τεχνική αυτή, κάθε ακολουθία συμβόλων

$$\mathbf{x} = x_1, x_2, ..., x_{L-1}, x_L ,$$

συνοδεύεται, και από μια ακολουθία σημάνσεων (labels)

$$\mathbf{y} = y_1, y_2, ..., y_{L-1}, y_L$$

Στη συγκεκριμένη περίπτωση της πρόγνωσης των διαμεμβρανικών τμημάτων, οι σημάνσεις είναι 3: μια για τα διαμεμβρανικά τμήματα (M), μια για την εσωτερική περιοχή (I) και μια για την εξωτερική (O). Επιπλέον, είναι αναγκαίο πλέον να ορίσουμε μια κατανομή για την πιθανότητα σύμπτωσης μιας κατάστασης με μια δεδομένη σήμανση. Στην πράξη, ομαδοποιούμε τις καταστάσεις σε ομάδες οι οποίες έχουν μια βιολογική σημασία, δηλαδή ομαδοποιούμε τις καταστάσεις που αντιστοιχούν σε διαμεμβρανικά τμήματα κ.ο.κ. Χρειαζόμαστε έτσι, μια μεταβλητή $\delta_k(c)$ που δηλώνει την πιθανότητα. η κατάσταση k να έχει σήμανση c. Η κατανομή που ακολουθεί αυτή η μεταβλητή, είναι προφανώς διωνυμική, αλλά σε όλες τις εφαρμογές που θα χρησιμοποιήσουμε, είναι απλώς μια δίτιμη συνάρτηση (delta function) που παίρνει απλώς την τιμή 1 αν η κατάσταση συμφωνεί με τη σήμανση και 0 σε αντίθετη περίπτωση. Δηλαδή, δεν επιτρέπουμε σε μια κατάσταση να συμπίπτει με περισσότερες από μια σημάνσεις.

Όπως γίνεται πλέον φανερό, με την εισαγωγή των σημάνσεων, ένας τρόπος να επιτύχουμε «μάθηση μετά διδασκάλου», είναι να θεωρήσουμε ως αντικειμενική συνάρτηση την από κοινού πιθανότητα $P(\mathbf{x},\mathbf{y}|\theta)$ των ακολουθιών $\mathbf{x}$ με τις σημάνσεις $\mathbf{y}$, δεδομένου του μοντέλου:

$$P(\mathbf{x},\mathbf{y}\,|\,\theta)=\sum_{\pi}P(\mathbf{x},\mathbf{y},\pi\,|\,\theta)=\sum_{\pi\in\Pi_y}P(\mathbf{x},\pi\,|\,\theta)=\sum_{\pi\in\Pi_y}a_{B\pi_1}\prod_{i=1}^{L}e_{\pi i}(x_i)a_{\pi_i\pi_{i+1}}$$

Τροποποιημένος αλγόριθμος Forward

$$\forall k\neq B,\ i=0:\ f_B(0)=1,\ f_k(0)=0$$

$$\forall 1\le i\le L:\ f_l(i)=e_l(x_i)\delta_l\left(y_i\right)\sum_k f_k(i-1)a_{kl}$$

$$P(\mathbf{x},\mathbf{y}|\theta)=\sum_k f_k(L)a_{kE}$$

Τροποποιημένος αλγόριθμος Backward

$$\forall k,\ i=L:\ b_k(L)=a_{kE}$$

$$\forall\ 1\le i<L:\ b_k(i)=\sum_l a_{kl}e_l(x_{i+1})\delta_l\left(y_{i+1}\right)b_l(i+1)$$

$$P(\mathbf{x},\mathbf{y}|\theta)=\sum_l a_{Bl}e_l(x_1)b_l(1)$$

			Sequence							
			I	I	I	M	M	M	O	O
States	Labels	0	x1	x2	x3	x4	x5	x6	x7	x8
1	I					f=0			f=0	
2	I									
3	I									
4	I									
5	M		f=0			f calculated as usual				
6	M									
7	M									
8	M									
9	O						f=0			
10	O									
11	O									
12	O									

Διαγραμματική απεικόνιση του πίνακα Forward για σημασμένες ακολουθίες. Έχουμε ένα μοντέλο με 12 υποθετικές καταστάσεις, και μια ακολουθία με 8 κατάλοιπα για τα οποία είναι γνωστές οι σημάνσεις (labels). Είναι φανερό ότι για τα κατάλοιπα και τις καταστάσεις που δεν συμφωνούν με την σήμανση, οι τιμές του πίνακα απλά μηδενίζονται.

Maximum Likelihood για Labeled sequences

Με την εισαγωγή των αλγορίθμων για σημασμένες ακολουθίες, είναι εφικτό πλέον να πραγματοποιήσουμε εκτίμηση μέγιστης πιθανοφάνειας:

$$\theta^{ML} = \arg\max_{\theta} P(\mathbf{x},\mathbf{y} \mid \theta)$$

Όλοι οι αλγόριθμοι που είδαμε ότι ισχύουν για τις μη σημασμένες ακολουθίες, ισχύουν με μικρές παραλλαγές και εδώ. Λόγω του ότι, οι ακολουθίες και οι σημάνσεις είναι ανεξάρτητες,

$$A_{kl} = \frac{1}{P(\mathbf{x},\mathbf{y} \mid \theta)} \sum_i f_k(i) a_{kl} e_l(x_{i+1}) \delta_l\left(y_{i+1}\right) b_l(i+1)$$

$$E_k(b) = \frac{1}{P(\mathbf{x},\mathbf{y} \mid \theta)} \sum_{\{i \mid x_i^j = b\}} f_k(i) b_k(i)$$

όπου οι ποσότητες $P(\mathbf{x},\mathbf{y}|\theta)$, $f_i(i)$ και $b_k(i)$, υπολογίζονται πλέον από τους τροποποιημένους αλγόριθμους που είδαμε παραπάνω. Με όλα τα παραπάνω, μπορούμε άνετα να πραγματοποιήσουμε εκπαίδευση μέγιστης πιθανοφάνειας, είτε με τη μέθοδο των Baum-Welch είτε με τη μέθοδο Gradient Descent.

Conditional Maximum Likelihood

- Με την εισαγωγή των labels, μπορούμε να πραγματοποιήσουμε εκπαίδευση που μεγιστοποιεί την πιθανότητα:

$$\theta^{CML} = \arg\max_{\theta} P(\mathbf{y} \mid \mathbf{x}, \theta) = \arg\max_{\theta} \frac{P(\mathbf{x}, \mathbf{y} \mid \theta)}{P(\mathbf{x} \mid \theta)}$$

- Η Πιθανοφάνεια γίνεται:

$$\ell = -\log P(\mathbf{y} \mid \mathbf{x}, \theta) = \ell_c - \ell_f$$

όπου:

$$\ell_c = -\log P(\mathbf{x},\mathbf{y} \mid \theta)$$

$$\ell_f = -\log P(\mathbf{x} \mid \theta)$$

Με τους δείκτες *c* και *f*, ονομάζουμε αντίστοιχα την πιθανοφάνεια που υπολογίζεται στη φάση όπου οι σημάνσεις λαμβάνονται υπόψη (clamped phase), και αυτή στην οποία οι σημάνσεις δεν υπολογίζονται (free-running phase).

Conditional Maximum Likelihood

- Απαιτεί διπλάσιο υπολογιστικό χρόνο από ML
- Αποδίδει καλύτερα όταν τα labels είναι καλής ποιότητας
- Εκπαίδευση μόνο με Gradient descent
- Περισσότερο ευαίσθητος στις αρχικές τιμές των παραμέτρων

1-best decoding

Ο αλγόριθμος 1-best (Krogh, 1997), είναι μια τροποποίηση του αλγορίθμου N-best, ο οποίος είχε προταθεί παλαιότερα για αναγνώριση ομιλίας (Schwartz and Chow, 1990). Στην ουσία, πρόκειται για έναν ευριστικό αλγόριθμο δυναμικού προγραμματισμού, ο οποίος αναζητά την εύρεση της πιο πιθανής αλληλουχίας σημάνσεων $\mathbf{y}^{\max}$ αντί αυτή της πιο πιθανής αλληλουχίας καταστάσεων. Ο αλγόριθμος, για κάθε θέση i της ακολουθίας, αποθηκεύει όλες τις πιθανές «ενεργές υποθέσεις» h_{i-1} για τη σήμανση, οι οποίες αποτελούνται από όλες τις πιθανές αλληλουχίες σημάνσεων μέχρι εκείνο το σημείο. Κατόπιν, για κάθε κατάσταση l «προωθεί» τις υποθέσεις προσθέτοντας στο τέλος κάθε μια από τις πιθανές σημάνσεις y_i και διαλέγει την καλύτερη. Η όλη διαδικασία επαναλαμβάνεται ως το τέλος της ακολουθίας. Σε αντίθεση με τον αλγόριθμο του Viterbi, ο αλγόριθμος 1-best δεν χρειάζεται αναδρομή αλλά έχει και μεγαλύτερες υπολογιστικές απαιτήσεις τόσο σε μνήμη όσο και σε πραγματοποιούμενες πράξεις.

Αλγόριθμος 1-best

$$i = 1: \ \gamma_l(h_1) = a_{Bl} e_l(x_1)$$

$$\forall 1 < i \le L: \ \gamma_l(h_i \mathrm{y}_i) = e_l(x_i) \sum_k \gamma_k(h_{i-1}) a_{kl}$$

$$P\left(\mathbf{x}, \mathbf{y}^{\max} \mid \theta\right) = \sum_k \gamma_k(h_L) a_{kE}$$

Posterior-Viterbi decoding

Ορίζονται οι επιτρεπτές μεταβάσεις:

$$\delta(k,l)=\begin{cases}1, \text{ if } a_{kl}>0\\ 0, \text{ otherwise}\end{cases}$$

Τελικά, το βέλτιστο επιτρεπτό εκ των υστέρων μονοπάτι π^{PV}, δίνεται από τη σχέση:

$$\pi^{PV}=\arg\max_{\pi}\prod_{i=1}^{L}\delta(\pi_i,\pi_{i+1})P(\pi_i\,|\,\mathbf{x})$$

Ο συνολικός αλγόριθμος, ο οποίος παρουσιάζεται παρακάτω, είναι στην ουσία μια παραλλαγή του αλγορίθμου Viterbi, στην οποία οι πιθανότητες γεννήσεως αντικαθίστανται από τις εκ των υστέρων πιθανότητες και οι πιθανότητες μετάβασης από την δίτιμη συνάρτηση που είδαμε παραπάνω.

Αλγόριθμος Posterior-Viterbi

$$\forall k\neq B,\ i=0:\ u_B(0)=1,\ u_k(0)=0$$

$$\forall\ 1\le i\le L:\ u_l(i)=P(\pi_i=l\,|\,\mathrm{x})\max_k\{u_k(i-1)\delta(k,l)\}$$

$$P(\mathbf{x},\pi^{PV}\,|\,\theta)=\max_k\{u_k(L)\delta(k,E)\}$$

Optimal Accuracy Posterior Decoding

Παραλλαγή του Posterior-Viterbi, η οποία υπολογίζει το μονοπάτι:

$$\pi^{OAPD} = \arg\max_{\pi} \sum_{i=1}^{L} \left\{ \delta(\pi_i, \pi_{i+1}) \left(\sum_k P(\pi_i \mid \mathbf{x}) \lambda_k(c) \right) \right\}$$

Συνολικά:

Optimal Accuracy Posterior Decoder algorithm

$$\forall k \neq B,\ i = 0:\ A_B(0) = 0,\ A_k(0) = -\infty$$

$$\forall\ 1 \leq i \leq L:\ A_l(i) = P\left(y_i = c^l \mid \mathbf{x}, \theta\right) + \max_k \left\{ A_k(i-1)\delta(k,l) \right\}$$

$$P(\mathbf{x}, \pi^{OAPD} \mid \theta) = \max_k \left\{ A_k(L)\delta(k,E) \right\}$$

Σύγκριση HMM-Regular Expressions

- Κάθε κανονική έκφραση μπορεί να αναπαρασταθεί με ένα HMM
- Το αντίστροφο δεν ισχύει
- Παραδείγματα

Regular Expressions

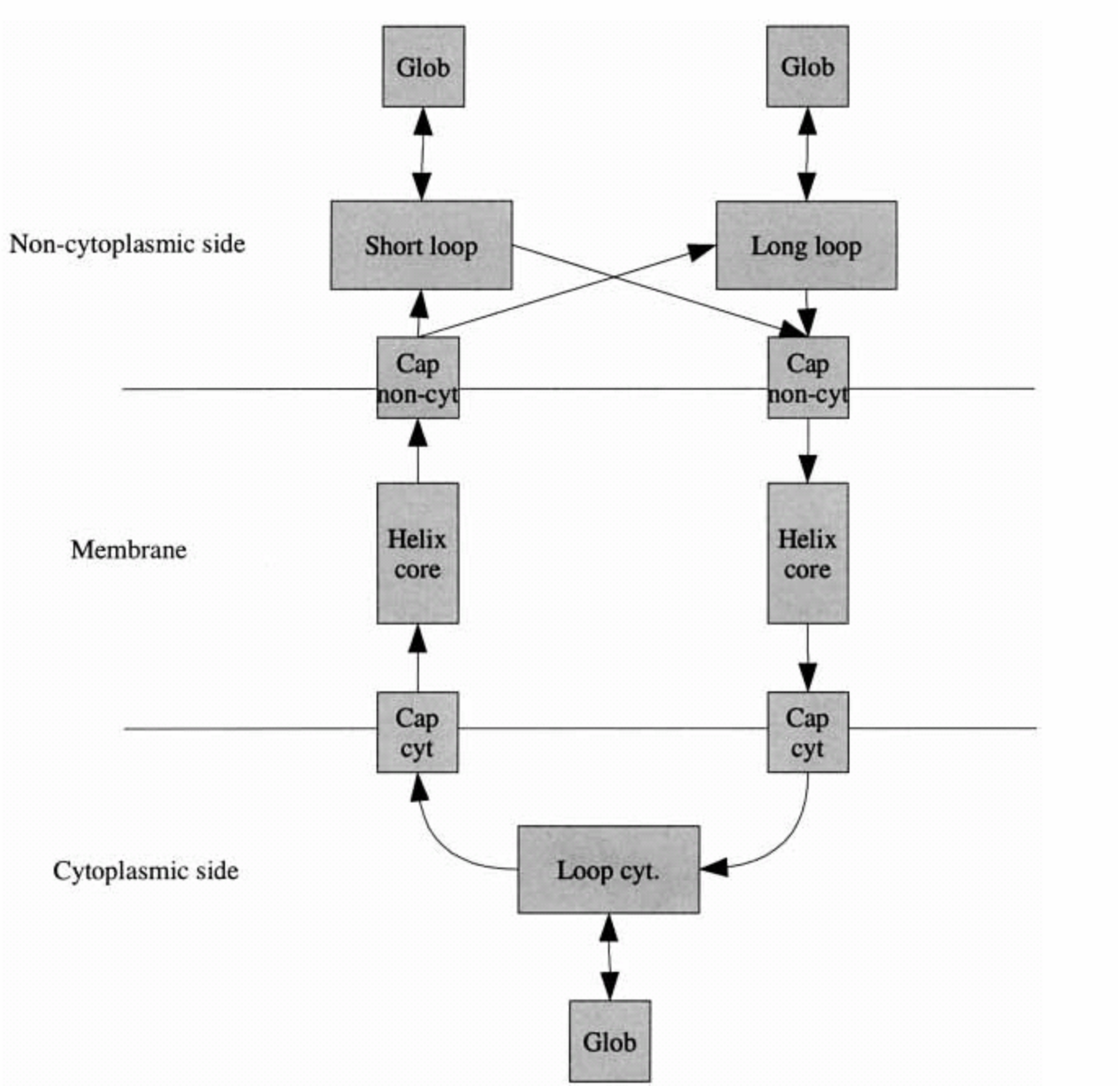
- `[AC]AG-GT[AG]AGT`
- `{DERK}(6)-[LIVMFWSTAG](2)`
  `[LIVMFYSTAGCQ]- [AGS]- C`

Πλεονεκτήματα του Hidden Markov Model

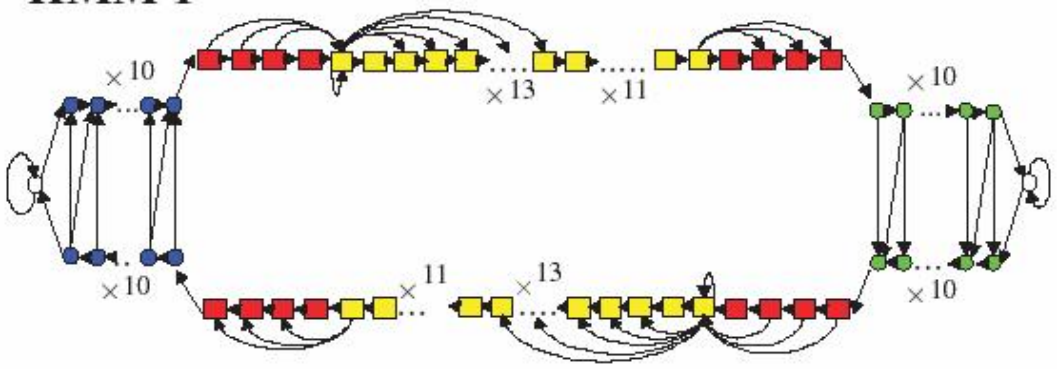
- Άριστη μαθηματική θεμελίωση, και πιθανοθεωρητική ερμηνεία των αποτελεσμάτων.
- Ύπαρξη κατάλληλων αλγορίθμων για την υλοποίηση του.
- Μπορεί να ενσωματώσει, σχεδόν κάθε είδους πληροφορία για το βιολογικό πρόβλημα, χωρίς να αναγκαστούμε να καταφύγουμε σε ευριστικές (heuristic) μεθόδους.
- Αν δομηθεί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

Εφαρμογές του Hidden Markov Model στη Βιολογία

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

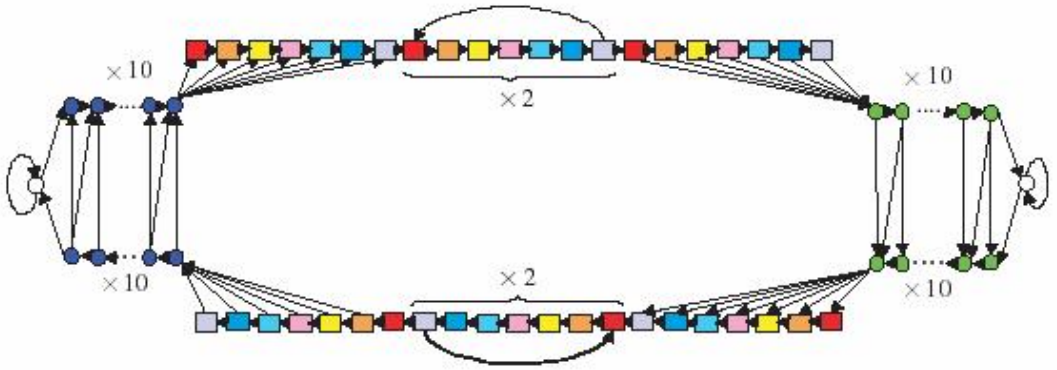

Glob
Glob
Non-cytoplasmic side
Short loop
Long loop
Cap non-cyt
Cap non-cyt
Membrane
Helix core
Helix core
Cap cyt
Cap cyt
Cytoplasmic side
Loop cyt.
Glob

HMM 1

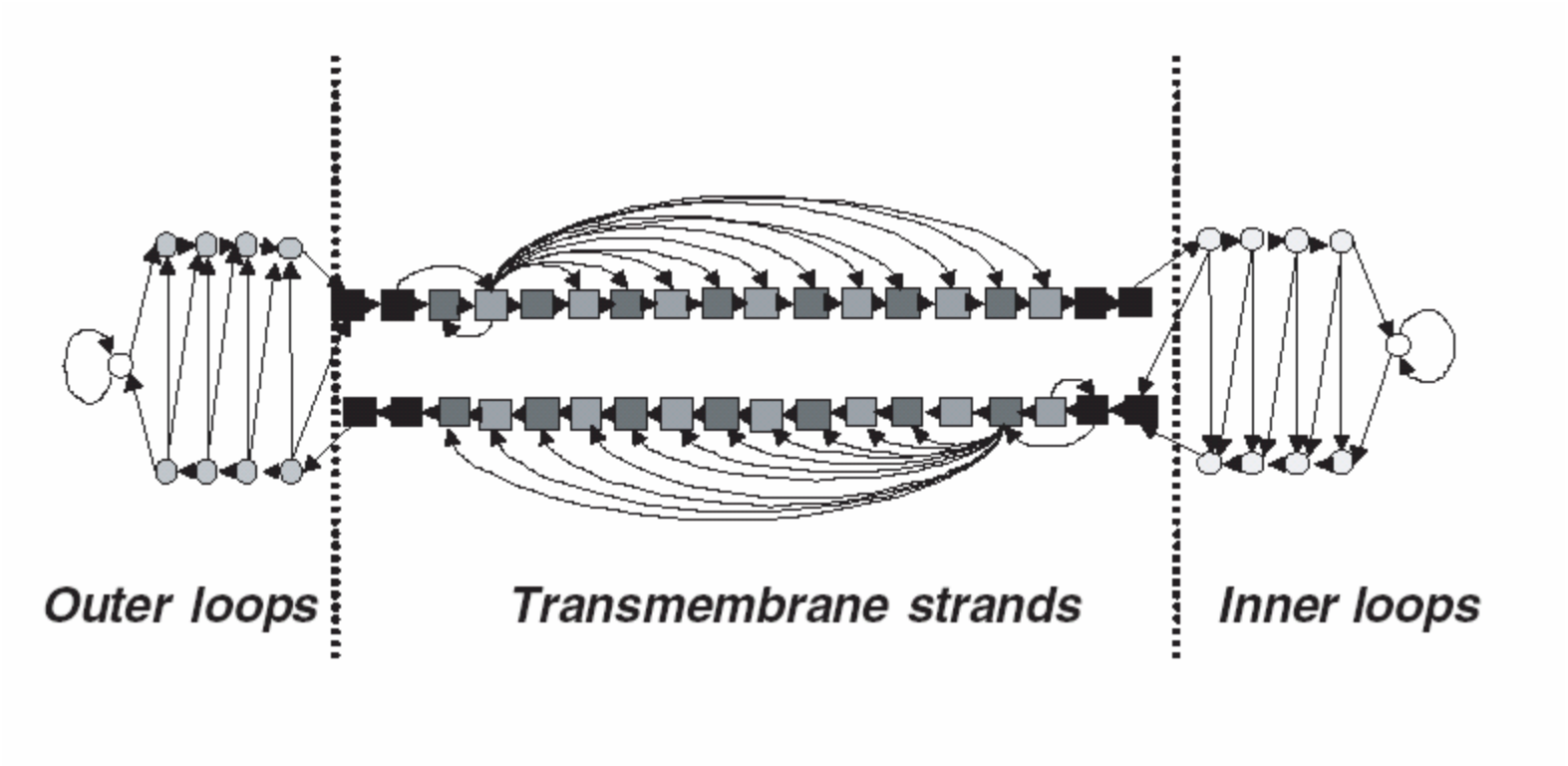


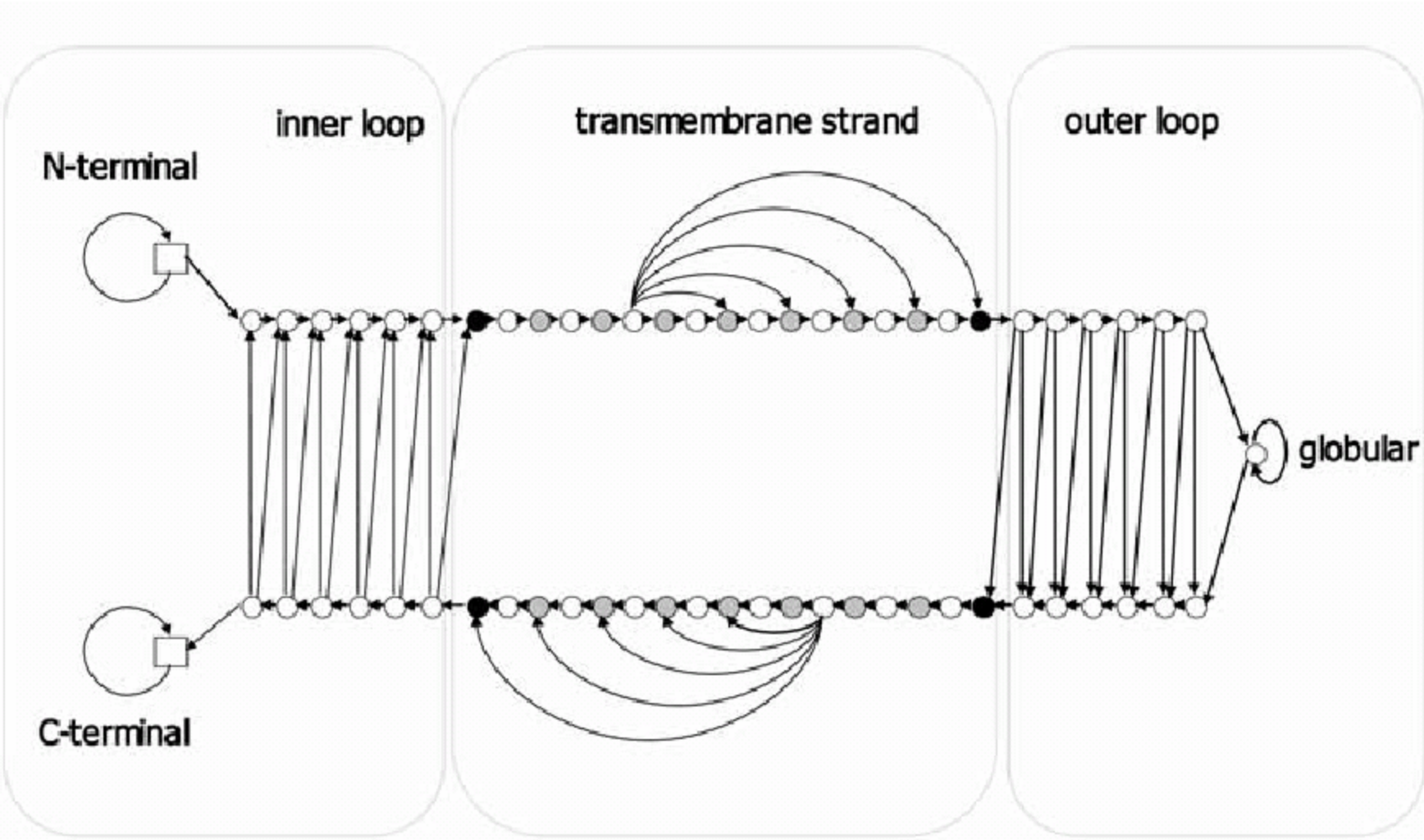
Outer Side Transmembrane Inner Side

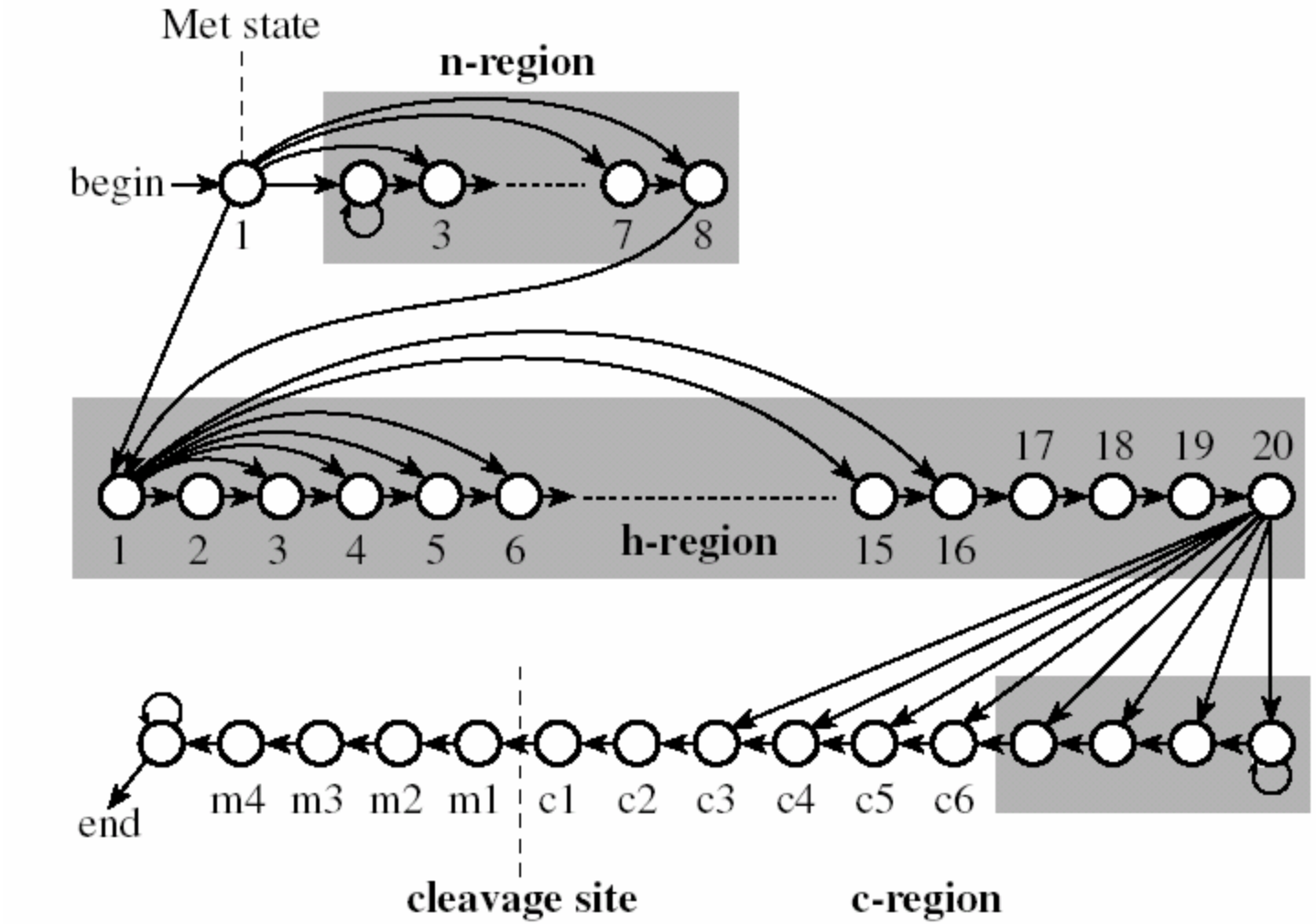
HMM 2

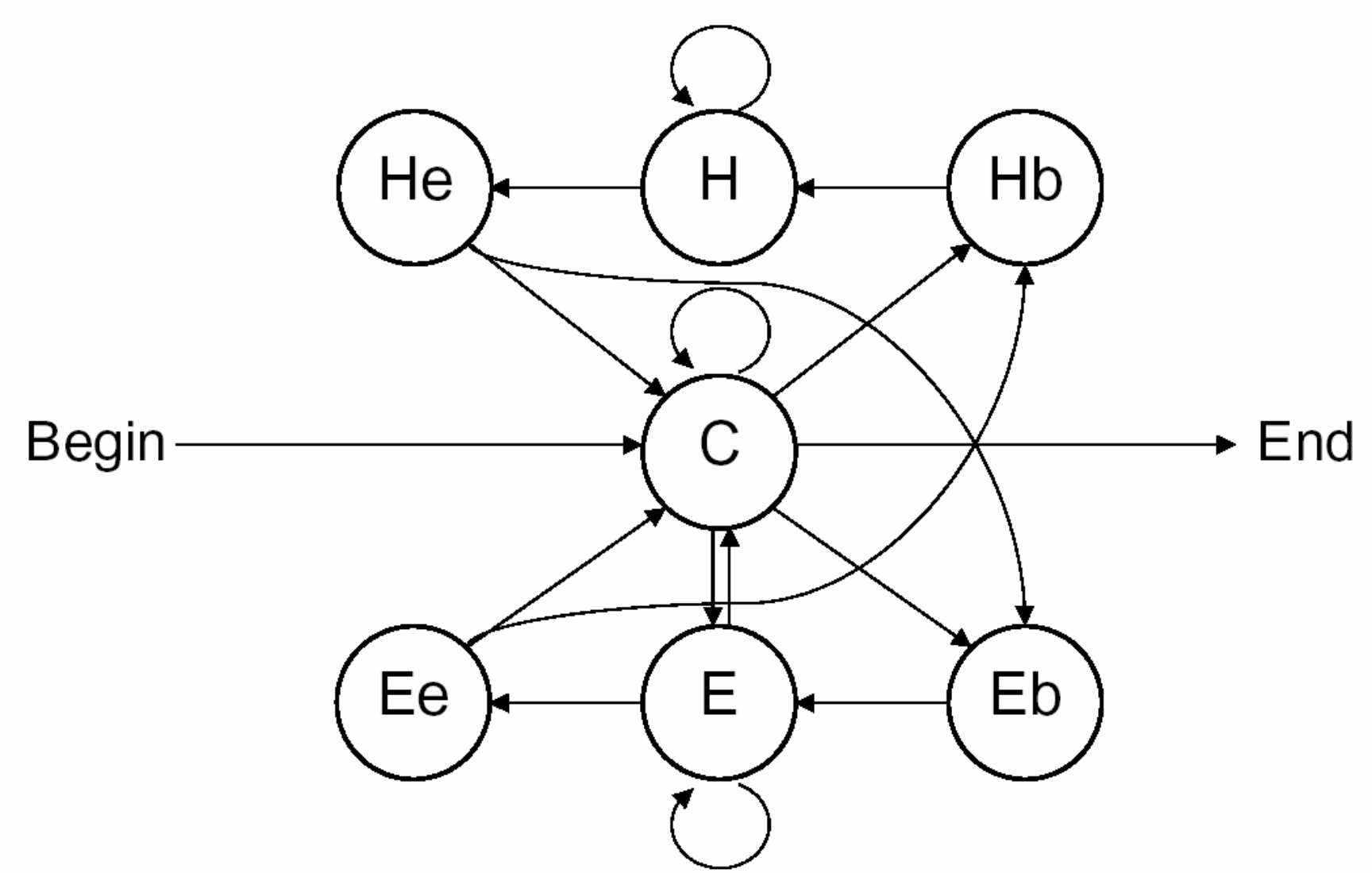


Outer Side Transmembrane Inner Side


Outer loops
Transmembrane strands
Inner loops

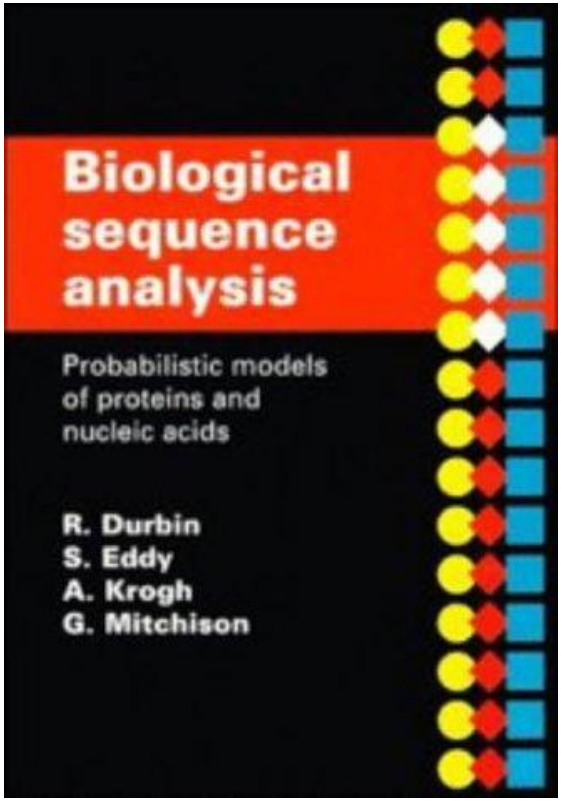

inner loop
transmembrane strand
outer loop
N-terminal
globular
C-terminal


Met state
n-region
begin
1
3
7
8
17
18
19
20
1
2
3
4
5
6
h-region
15
16
end
m4
m3
m2
m1
c1
c2
c3
c4
c5
c6
cleavage site
c-region


He
H
Hb
Begin
C
End
Ee
E
Eb

Επεκτάσεις του HMM

- HMMSDO
- Input-Output HMM (IOHMM)
- Factorial HMM
- Hierarchical HMM
- Partially Hidden Markov Model (PHMM)
- Hidden Neural Networks (HNN)
- και πολλά άλλα …


Biological sequence analysis
Probabilistic models of proteins and nucleic acids
R. Durbin
S. Eddy
A. Krogh
G. Mitchison


BIOINFORMATICS
THE MACHINE LEARNING APPROACH
SECOND EDITION
PIERRE BALDI AND SØREN BRUNAK

Software

- http://www.dina.dk/~sestoft/bsa/Match3.java
- http://www.cfar.umd.edu/~kanungo/software/software.html
- http://hmmer.wustl.edu/