Κεφάλαιο 8: Μαρκοβιανά Μοντέλα

Σύνοψη
Στο κεφάλαιο αυτό, θα γίνει η απαραίτητη εισαγωγή στα μαρκοβιανά μοντέλα εξάρτησης και κατόπιν, παρουσίαση των κρυπτομαρκοβιανών μοντέλων (Hidden Markov Models) τα οποία αποτελούν ένα σημαντικό εργαλείο στη σύγχρονη βιοπληροφορική. Θα αναφερθούμε στα βασικά χαρακτηριστικά των μοντέλων αυτών και στη μαθηματική τους θεμελίωση, ενώ θα παρουσιαστούν σε βάθος οι διάφοροι αλγόριθμοι που χρησιμοποιούνται για τον υπολογισμό της πιθανοφάνειας, για την αποκωδικοποίηση και για την εκτίμηση παραμέτρων στα μοντέλα αυτά. Θα παρουσιαστούν επίσης, τα μοντέλα για σημασμένες αλληλουχίες, τα οποία αποτελούν μια επέκταση του βασικού HMM, η οποία βρίσκει πολλές εφαρμογές στην ανάλυση βιολογικών αλληλουχιών (πρόγνωση διαμεμβρανικών πρωτεϊνών, εύρεση γονιδίων κ.ο.κ.). Τέλος, θα γίνει ειδική αναφορά στο profile HMM το οποίο είναι άλλη μια παραλλαγή του βασικού μοντέλου, η οποία βρίσκει εφαρμογές στη μοντελοποίηση πρωτεϊνικών οικογενειών, στην εύρεση μακρινών ομολόγων και στην πολλαπλή στοίχιση.

Προαπαιτούμενη γνώση
Βασικές γνώσεις πιθανοτήτων. Κατανόηση των εννοιών της στοίχισης και πολλαπλής στοίχισης αλληλουχιών που μελετήθηκαν στα κεφάλαια 3 και 4.

8. Εισαγωγή

Στο κεφάλαιο αυτό θα μελετήσουμε μαθηματικά μοντέλα τα οποία ανήκουν σε μια μεγάλη οικογένεια στοχαστικών-πιθανοθεωρητικών μοντέλων, τα οποία ονομάζονται μοντέλα εξάρτησης του Markov ή αλλιώς Μαρκοβιανά μοντέλα. Θα εισαγάγουμε αρχικά την έννοια της αλυσίδας Markov (Markov Chain), η οποία βρίσκει σημαντικές εφαρμογές στη δημιουργία μοντέλων που περιγράφουν αλληλουχίες DNA ή και πρωτεϊνών. Η θεώρηση μιας ακολουθίας ενδεχομένων ως αλυσίδα Markov στηρίζεται, πολύ απλά, 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 επόμενό του. Αν αυτή η εξάρτηση επεκταθεί και σε *2,3,...,k* προηγούμενα ενδεχόμενα τότε μιλάμε για αλυσίδες Markov $2^{ης}$,$3^{ης}$, ...,$k^{ης}$ τάξης.

Πρέπει να τονιστεί εδώ, ότι το μοντέλο Markov θεωρείται από πολλούς ερευνητές ως το πιο φυσικό για να περιγράψει αλληλουχίες μεγαλομορίων όπως του DNA αλλά και των πρωτεϊνών, και αυτό φαίνεται διαισθητικά φυσικό καθώς αυτή η εξάρτηση φαίνεται να προσεγγίζει την έννοια της πληροφορίας που εμπεριέχεται σε μια αλληλουχία. Ήδη από τη δεκαετία του 1970 τα μοντέλα αυτά χρησιμοποιούνταν και χρησιμοποιούνται ακόμα με σκοπό την αναγνώριση και επεξεργασία εικόνας, ήχου κ.α. και υπάρχει πλούσια βιβλιογραφία πάνω στα θέματα αυτά. Η πιο απλή εξήγηση για τα παραπάνω είναι το γεγονός ότι σε οποιοδήποτε κωδικοποιημένο σύστημα επικοινωνίας όπως στις φυσικές γλώσσες, υπάρχει μια εσωτερική δομή που καθορίζει κάποιο είδος εξάρτησης των συμβόλων. Για παράδειγμα, στην αγγλική γλώσσα το γράμμα Q ακολουθείται σχεδόν πάντοτε από το U, άρα η πιθανότητα να εμφανιστεί το U σε μια θέση δεν είναι πάντα ίδια αλλά εξαρτάται από το αν προηγήθηκε το Q. Για την ακρίβεια, ο ίδιος ο Ρώσος Μαθηματικός Andrey Markov (1856-1922) οδηγήθηκε στην σύλληψη της έννοιας των ομώνυμων αλυσίδων, μελετώντας τις εναλλαγές φωνηέντων και συμφώνων σε κάποιο ποίημα του Pushkin (Markov, 1913). Θα προχωρήσουμε στην συνέχεια στον τυπικό ορισμό του μοντέλου Markov (Markov Model-MM) αλλά και του «κρυμμένου» μοντέλου Markov (Hidden Markov Model-HMM) και θα εξετάσουμε τις κυριότερες εφαρμογές τους.

8.1. Αλυσίδες Markov

8.1.1. Ορισμοί

Μια αλυσίδα Markov $1^{ης}$ τάξης ορίζεται ως μια στοχαστική ανέλιξη διακριτών καταστάσεων σε διακριτό χρόνο. Στην περίπτωση των βιολογικών αλληλουχιών, ως καταστάσεις ορίζονται τα σύμβολα της ακολουθίας τα οποία ανήκουν σε ένα πεπερασμένο αλφάβητο, Ω (τα τέσσερα νουκλεοτίδια στην περίπτωση του DNA ή

τα 20 αμινοξέα στην περίπτωση των πρωτεϊνών). Αν θεωρήσουμε μια πρωτεϊνική αλληλουχία μήκους L καταλοίπων, και την ονομάσουμε **x**, έτσι ώστε:

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

και θεωρήσουμε την κατανομή των αμινοξέων σε κάθε θέση i κατά μήκος της αλληλουχίας ως τυχαία μεταβλητή, τότε μπορούμε να ορίσουμε την αλυσίδα Markov, ως μια στοχαστική ανέλιξη η οποία διαθέτει τη λεγόμενη «Μαρκοβιανή Ιδιότητα». Στη διακριτή περίπτωση (όπως η συγκεκριμένη), η ανέλιξη αποτελείται από την ακολουθία των τυχαίων μεταβλητών **x**, η οποία παίρνει τιμές σε ένα «χώρο καταστάσεων» οριζόμενο από το συγκεκριμένο αλφάβητο. Όπως είδαμε, οι τιμές των x_i συμβολίζουν την «κατάσταση στην οποία βρίσκεται το σύστημα την χρονική στιγμή *i*». Η Μαρκοβιανή ιδιότητα (σε διακριτό χρόνο) ορίζει ότι η δεσμευμένη κατανομή των «μελλοντικών» παρατηρήσεων x_{i+1}, x_{i+2}, x_{i+3} δεδομένου του «παρελθόντος» x_1, x_2, ...x_{i-1}, x_i, εξαρτάται από το παρελθόν μόνο μέσω του x_i. Με άλλα λόγια, η γνώση της πιο πρόσφατης κατάστασης του συστήματος καθιστά τη λιγότερο πρόσφατη ιστορία άχρηστη. Αυτό τυπικά διατυπώνεται ως εξής:

$$P(x_i \mid x_{i-1}, ..., x_1) = P(x_i \mid x_{i-1}) \tag{8.1}$$

Μια συγκεκριμένη Αλυσίδα Markov χαρακτηρίζεται από τον πίνακα των «πιθανοτήτων μετάβασης» (transition probabilities), ο οποίος πιο απλά ονομάζεται πίνακας μεταβάσεων. Τα στοιχεία αυτού του πίνακα, δίνονται από την παρακάτω σχέση:

$$a_{st} = P(x_i = t \mid x_{i-1} = s) = \alpha_{x_{i-1}x_i} \tag{8.2}$$

η οποία δηλώνει, την πιθανότητα το κατάλοιπο t να εμφανιστεί στη θέση i της αλληλουχίας, δεδομένου ότι το προηγούμενο κατάλοιπο (i-1) είναι s. Αν αναλογιστούμε ότι μπορούμε να γενικεύσουμε την εξάρτηση στα k προηγούμενα κατάλοιπα, είναι φυσικό η δεδομένη αλυσίδα να ονομάζεται Αλυσίδα Markov $1^{ης}$ τάξεως. Η συνολική πιθανότητα μιας αλληλουχίας υπολογίζεται ως εξής:

$$P(\mathbf{x}) = P(x_1, x_2, ..., x_{L-1}, x_L) = P(x_L \mid x_{L-1}, ..., x_1) P(x_{L-1} \mid x_{L-2}, ..., x_1) ... P(x_1)$$

και από τη σχέση (8.1), έχουμε:

$$P(\mathbf{x}) = P(x_L \mid x_{L-1}) P(x_{L-1} \mid x_{L-2}) ... P(x_2 \mid x_1) P(x_1) = P(x_1) \prod_{i=2}^{L} P(x_i \mid x_{i-1}) = P(x_1) \prod_{i=2}^{L} \alpha_{x_{i-1}x_i} \tag{8.3}$$

όπου $P(x_1)$ είναι η πιθανότητα για την εμφάνιση του πρώτου συμβόλου. Σύμφωνα με τον ορισμό αυτό, βλέπουμε ότι οι πιθανότητες μεταβάσεως είναι ίδιες, ανεξαρτήτως της θέσης τους στην αλυσίδα δηλαδή:

$$p_{ab}(n-1, n) = P(x_i = b \mid x_{i-1} = a) = p_{ab} \text{ για κάθε } n=1,2,..L.$$

Η αλυσίδα αυτή λεμε ότι έχει στάσιμες πιθανότητες μεταβάσεως, ή, ισοδύναμα, ότι η αλυσίδα αυτή είναι *ομογενής χρονικά*. Ο περιορισμός αυτός χρησιμοποιείται πολλές φορές στις περιπτώσεις μακρομορίων (αν και υπάρχουν εξαιρέσεις, όπως θα αναφέρουμε), κυρίως μεν γιατί προσφέρει υπολογιστική απλότητα αλλά και γιατί δεν έχουμε, στις περισσότερες περιπτώσεις, καμία ένδειξη που να υποστηρίζει μια τέτοια εξάρτηση από τη θέση στην αλυσίδα. Ο πίνακας ο οποίος περιέχει τις πιθανότητες μεταβάσεως, όπως είδαμε, λέγεται πίνακας *πιθανοτήτων μεταβάσεως* ή *πίνακας μεταβάσεως* $1^{ης}$ *τάξης* και πρέπει για ένα αλφάβητο με πλήθος k να ικανοποιεί τα παρακάτω:

$$p_{a,b} \geq 0 \text{ για } a, b = 1,2,..., k$$

και $$\sum_{b=1}^{k} p_{a,b} = 1 \text{ για κάθε } b=1,2,...,k$$

Γενικότερα κάθε τετραγωνικός πίνακας που ικανοποιεί τις δυο αυτές σχέσεις, λέγεται *στοχαστικός*. Όπως είδαμε από τις παραπάνω, σχέσεις ορίζεται πλήρως μια αλυσίδα Markov, αρκεί να ορίσουμε επιπλέον μια πιθανότητα για την κατάσταση της έναρξης της αλυσίδας (B=Begin). Η πιθανότητα αυτή ονομάζεται αρχική πιθανότητα και ορίζεται ως:

$$P(x_1 = a) = p_{Ba} \tag{8.4}$$

Όμοια μπορούμε να ορίσουμε (χωρίς όμως και να είναι απαραίτητο) μια άλλη τελική κατάσταση (E=End) για τον τερματισμό της αλυσίδας με πιθανότητα:

$$P(E \mid x_n = b) = p_{bE} \tag{8.5}$$

Έτσι πλέον μια πλήρης σχηματική αναπαράσταση του μοντέλου Markov φαίνεται στην Εικόνα 8.1 παρακάτω. Παραδοσιακά η λήξη της αλληλουχίας δεν συμπεριλαμβάνεται στο μοντέλο, θεωρούμε δηλαδή

ότι η αλυσίδα μπορεί να τελειώνει οπουδήποτε. Το πλεονέκτημα του να συμπεριληφθεί αυτή η κατάσταση στο μοντέλο, είναι όταν θέλουμε να μελετήσουμε την κατανομή του μήκους της αλυσίδας. Έτσι αν

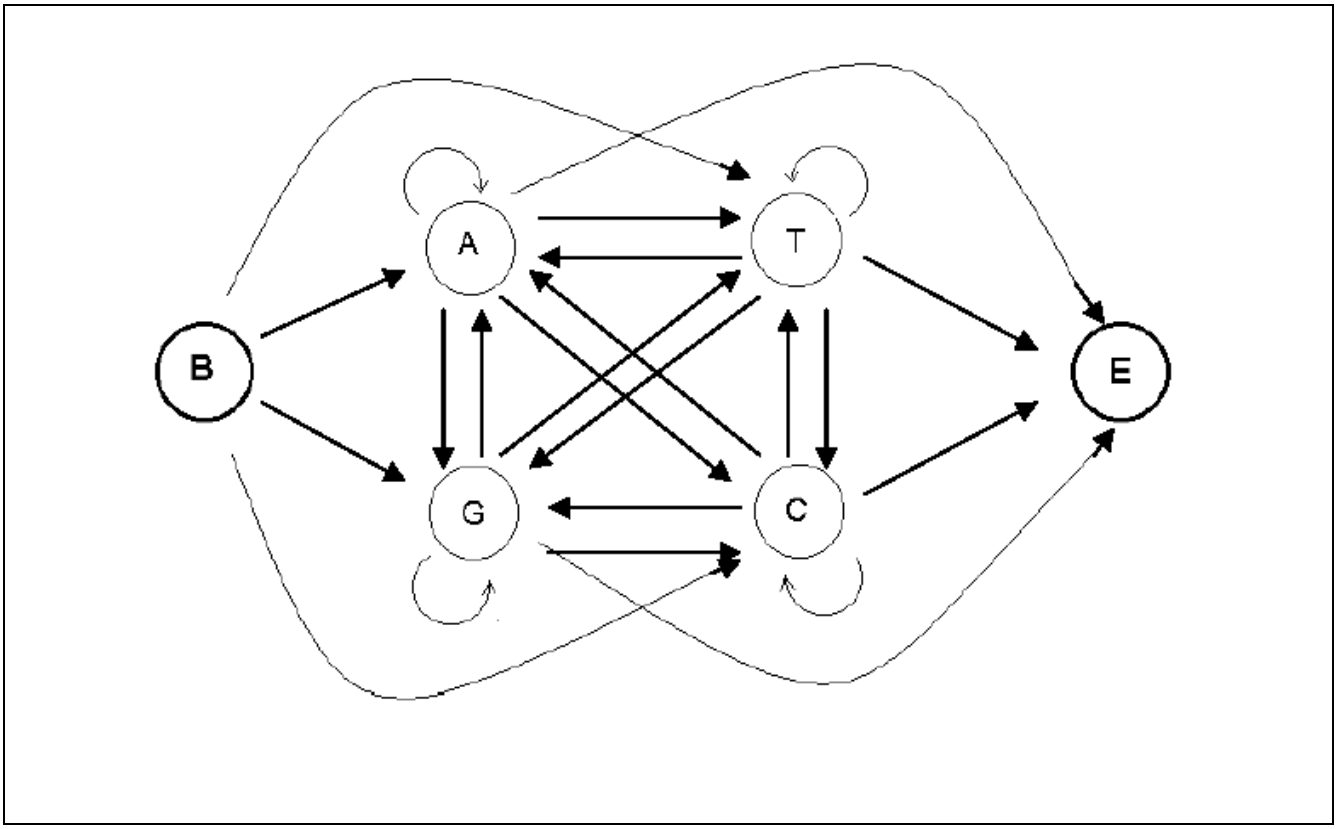
$$P(E \mid x_n = b) = p_{bE} = q$$

τότε η κατανομή του αθροίσματος των πιθανοτήτων της σχέσης (8.3) για μια αλληλουχία μήκους L είναι:

$$p_{o\lambda} = q(1-q)^{L-1} \tag{8.6}$$

δηλαδή η κατανομή του αθροίσματος των πιθανοτήτων για όλες τις αλληλουχίες μήκους L ακολουθεί γεωμετρική κατανομή. Αντίστοιχα το άθροισμα των πιθανοτήτων όλων των πιθανών ακολουθιών είναι (Durbin, Eddy, Krogh, & Mithison, 1998):

$$p_{o\lambda} = \sum_{\{X\}} P(\mathrm{x}) = \sum \sum \ldots \sum P(x_1) \prod_{i=2}^{n} P(x_i \mid x_{i-1}) = 1 \tag{8.7}$$



Εικόνα 8.1: *Ένα τυπικό μοντέλο αλυσίδας Markov, με καταστάσεις τις 4 βάσεις του DNA. Τα βέλη συμβολίζουν τις επιτρεπτές μεταβάσεις.Με B και E, συμβολίζονται οι καταστάσεις έναρξης και τερματισμού του μοντέλου, αντίστοιχα.*

8.1.2. Εκτίμηση Παραμέτρων

Οι εκτιμητές μέγιστης πιθανοφάνειας (Maximum Likelihood Estimates-MLEs) των πιθανοτήτων μεταβάσεως, υπολογίζονται σύμφωνα με τη σχέση:

$$\hat{\alpha}_{x_{i-1}x_i} = \frac{n_{st}}{\sum_{t'} n_{st'}} \tag{8.8}$$

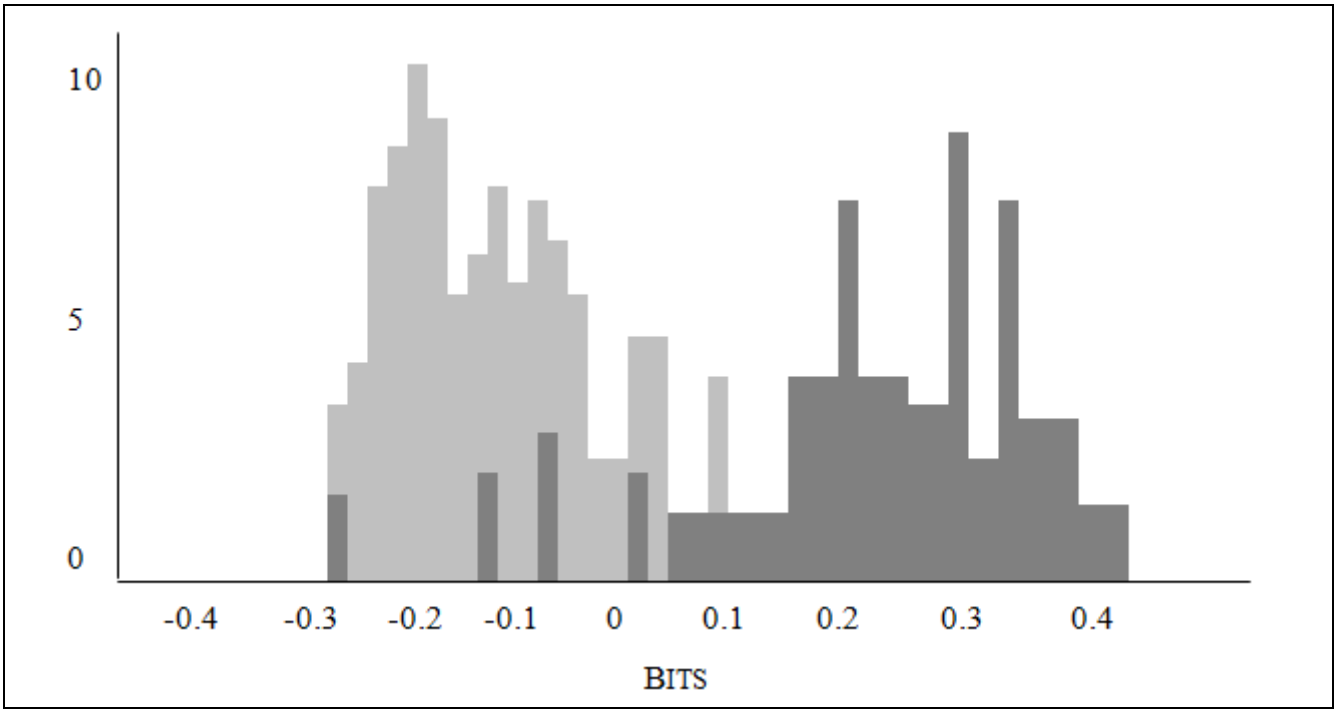
όπου n_{st} είναι οι παρατηρούμενες εμφανίσεις του καταλοίπου s ακολουθούμενο από το κατάλοιπο t στις αλληλουχίες εκπαίδευσης, με το άθροισμα στον παρονομαστή να εκτείνεται σε όλο το αλφάβητο των 20 αμινοξέων (ή των νουκλεοτιδίων αν μιλάμε για DNA). Θεωρώντας δυο διαφορετικά μοντέλα, με τη χρήση δυο πινάκων μεταβάσεων (για παράδειγμα, ένα μοντέλο + για τα λεγόμενα θετικά παραδείγματα και ένα μοντέλο – για τα λεγόμενα αρνητικά), μπορούμε να ορίσουμε ένα *log-odds score, S*(**x**) για ολόκληρη την αλληλουχία, το οποίο είναι χρήσιμο για διαχωριστικούς σκοπούς:

$$S(\mathbf{x}) = \log \frac{P(\mathbf{x}|+)}{P(\mathbf{x}|-)} = \sum_{i=1}^{L} \log\left(\frac{\alpha^{+}_{x_{i-1}x_i}}{\alpha^{-}_{x_{i-1}x_i}}\right) = \sum_{i=1}^{L} \beta_{x_{i-1}x_i} \tag{8.9}$$

όπου $\beta_{x_{i-1}x_i}$, είναι το log-odds για την πιθανότητα μετάβασης από το κατάλοιπο x_{i-1} στο x_i, και είναι ένα σχετικό μέτρο της τάσης των πιθανοτήτων μετάβασης να εμφανίζονται πιο συχνά στο ένα ή το άλλο μοντέλο. Το σκορ αυτό, είναι εντελώς ανάλογο με τα αντίστοιχα που είδαμε στο κεφάλαιο 2, όπου και μελετούσαμε τις αλληλουχίες, κάτω από τις προϋποθέσεις του μοντέλου της ανεξαρτησίας. Τιμές των $\beta_{x_{i-1}x_i}$ μεγαλύτερες από το 0, υποδηλώνουν προτιμήσεις των συγκεκριμένων μεταβάσεων για το μοντέλο (+), ενώ τιμές μικρότερες από το 0 προτίμηση για το μοντέλο (-). Για να εκμηδενίσουμε την επιρροή του μήκους των ακολουθιών στο συνολικό σκορ, κανονικοποιούμε περαιτέρω τις τιμές, διαιρώντας με το μήκος L της αλληλουχίας έτσι ώστε να πάρουμε ένα log-odds score ανά κατάλοιπο.

$$S^{norm}(\mathbf{x}) = \frac{S(\mathbf{x})}{L} = \frac{\sum_{i=1}^{L} \beta_{x_{i-1}x_i}}{L} \tag{8.10}$$

Χαρακτηριστικό παράδειγμα, $1^{ης}$ τάξης μοντέλου με την παραπάνω διατύπωση, αναφέρεται στην εύρεση νησίδων CG στα ευκαρυωτικά γονιδιώματα (Durbin, et al., 1998).



Εικόνα 8.2: *Ένα παράδειγμα μαρκοβιανής αλυσίδας για την εύρεση νησίδων CG στα ευκαρυωτικά γονιδιώματα. Ένας αριθμός πραγματικών γονιδίων και ένας αριθμός μη-γονιδιών, αναλύθηκαν με τη βοήθεια των σχέσεων* (8.8) *και* (8.9) *και τα αποτελέσματα για το Σκορ παρατίθενται σε ένα απλό ιστόγραμμα συχνοτήτων. Παρατηρούμε ότι οι δύο κατανομές διαχωρίζονται ικανοποιητικά. Τα bits στις τιμές του σκορ, αναφέρονται σε λογάριθμο με βάση το 2.*

8.1.3. Αλυσίδες ανώτερης τάξεως

Μια $k^{ης}$ τάξεως αλυσίδα Markov, μπορεί να προκύψει αυτόματα από γενίκευση της Μαρκοβιανής ιδιότητας της εξισώσεως (8.1). Συγκεκριμένα, η σχέση αυτή τροποποιείται έτσι ώστε να συμπεριλάβει εξάρτηση στις k προηγούμενες παρατηρήσεις:

$$P(x_i | x_{i-1},...,x_1) = P(x_i | x_{i-1}, x_{i-2},..., x_{i-k}) = \alpha_{x_k \ldots x_{i-1}x_i} \tag{8.11}$$

Δεδομένου ότι ισχύει:

$$P(x_i | x_{i-1}, x_{i-2},..., x_{i-k}) = P(x_i, x_{i-1},..., x_{i-k+1} | x_{i-1}, x_{i-2},..., x_{i-k}) \tag{8.12}$$

η $k^{ης}$ τάξεως αλυσίδα Markov, είναι ισοδύναμη με μια αλυσίδα $1^{ης}$ τάξεως, αλλά με ένα αλφάβητο της τάξης του 20^k. Κατά συνέπεια, απαιτεί τον υπολογισμό πινάκων μεταβάσεων μεγέθους 20^kx20^k. Άρα, στην

περίπτωση των πρωτεϊνών, ενώ για μια αλυσίδα $1^{ης}$ τάξεως χρειαζόμαστε να υπολογίσουμε 20^2=400 πιθανότητες για κάθε μοντέλο, για ένα μοντέλο $2^{ης}$ τάξεως χρειαζόμαστε 20^3=8000 παραμέτρους, αριθμός υπερβολικά μεγάλος ο οποίος στην περίπτωση αλληλουχιών πρωτεϊνών θα απαιτούσε υπερβολικά μεγάλο αριθμό ακολουθιών για να χρησιμοποιηθούν ως παραδείγματα για την εκπαίδευση των μοντέλων. Περιπτώσεις αλυσίδων ανώτερης τάξης είναι δυνατόν να εφαρμοστούν πιο εύκολα σε αλληλουχίες νουκλεοτιδίων, όπου το αλφάβητο είναι μικρότερο και οι αλληλουχίες πολύ μεγαλύτερες (Ellrott, Yang, Sladek, & Jiang, 2002; Phillips, Arnold, & Ivarie, 1987). Σε μια ενδιαφέρουσα εργασία, οι Audic και Claverie (Audic & Claverie, 1998), χρησιμοποίησαν αλυσίδες ανώτερης τάξης σε συνδυασμό με μια υπολογιστικά εντατική μεθοδολογία έτσι ώστε να ανιχνεύσουν διαφορετικής σύστασης περιοχές σε βακτηριακά γονιδιώματα. Με αυτό, τον τρόπο, διαχώρισαν χωρίς την ανάγκη συνόλου εκπαίδευσης, περιοχές οι οποίες κωδικοποιούν πρωτεΐνες, περιοχές που δεν κωδικοποιούν τίποτα και περιοχές που κωδικοποιούν πρωτεΐνες αλλά στη συμπληρωματική τους αλυσίδα, σε ποσοστό που έφτανε και το 90% (Audic & Claverie, 1998).

Σε περιπτώσεις πρωτεϊνών, έχει προταθεί η προσέγγιση των πιθανοτήτων μετάβασης μεγαλύτερης τάξης. Συγκεκριμένα, η πιθανότητα μετάβασης για μια $k^{ης}$ τάξης αλυσίδα θα μπορούσε να προσεγγισθεί (Yuan, 1999), από τη σχέση:

$$P\left(x_i \mid x_{i-1}, x_{i-2}, ..., x_{i-k}\right) \approx \prod_{j=1}^{k} P\left(x_i \mid x_{i-j}\right) \tag{8.13}$$

Η σχέση (8.13) χρησιμοποιήθηκε (Yuan, 1999) στην προσπάθεια να προβλεφθεί η υποκυτταρική τοποθεσία των βακτηριακών πρωτεϊνών, με αρκετή επιτυχία. Σε γενικές γραμμές, αναμένουμε ότι με μεγαλύτερης τάξεως αλυσίδες θα έχουμε και καλύτερη διαχωριστική ικανότητα των μοντέλων, γεγονός που επιβεβαιώνεται και από τη μελέτη αυτή (Yuan, 1999). Από την άλλη μεριά, μεγαλώνοντας πάρα πολύ την τάξη (>6), ακόμα και για νουκλεοτιδικές αλληλουχίες, πέραν του προβλήματος υπερ-προσαρμογής (over-fitting) και της έλλειψης δεδομένων, ανακύπτει και το πρόβλημα της εισαγωγής θορύβου, από μη-σημαντικές μακρινές αλληλεπιδράσεις (Ellrott, et al., 2002; Phillips, et al., 1987; Yuan, 1999). Άλλη μέθοδος, χρήσιμη κυρίως σε αλληλουχίες νουκλεοτιδίων, είναι αυτή της χρήσης μη-ομογενών αλυσίδων (non-homogenous Markov chains), με την οποία χρησιμοποιούνται διαφορετικοί πίνακες μεταβάσεων, έτσι ώστε να εντοπιστούν καλύτερα οι στατιστικές προτιμήσεις στις διάφορες θέσεις της τριπλέτας βάσεων μιας κωδικής περιοχής (Borodovsky & Peresetsky, 1994).

Μεγάλο ενδιαφέρον, τόσο πρακτικό όσο και θεωρητικό, παρουσιάζουν τα μοντέλα Markov μεταβλητού μήκος (Variable length Markov Models-VMM), τα οποία όπως διατυπώθηκαν από τον Bejerano (Bejerano, 2004) είναι μια επέκταση της διατύπωσης των Στοχαστικών Πεπερασμένων Αυτομάτων (Probabilistic Finite Automata-PFA), από τον Ron και τους συνεργάτες του (Ron, Singer, & Tishby, 1996). Το μοντέλο αυτό, αντί να υπολογίζει όλα τα C πλαίσια παραθύρων μήκους n, τα οποία θα καθορίσουν τις παραμέτρους της αλυσίδας $k^{ης}$ τάξης, υπολογίζει παραμέτρους μόνο για ένα υποσύνολο $C^* \subset C$, το οποίο προσδιορίζεται με εκπαίδευση από τα δεδομένα και προβλέπει μεγαλύτερες εξαρτήσεις στα προηγηθέντα κατάλοιπα, όταν είναι απαραίτητο, ενώ μικρότερες όταν δεν είναι. Έτσι, η προσεγγιστική σχέση που χρησιμοποιείται, είναι η εξής:

$$P\left(x_i \mid x_{i-1}, x_{i-2}, ..., x_{i-k}\right) \approx P\left(x_i \mid \max_{k_i \geq 0}\left\{x_{i-k_i}, ..., x_{i-1} \in C\right\}\right) \tag{8.14}$$

Με τη χρήση αυτής της σχέσης (και μιας πολύπλοκης διαδικασίας εκπαίδευσης που δεν θα αναφερθεί εδώ), ο Bejerano και οι συνεργάτες του (Bejerano, Seldin, Margalit, & Tishby, 2001; Bejerano & Yona, 2001), κατάφεραν να κατασκευάσουν μοντέλα τα οποία διακρίνουν με αρκετά μεγάλη ακρίβεια σχεδόν όλες τις οικογένειες πρωτεϊνών που είναι κατατεθειμένες στην βάση δεδομένων PFAM (Bateman et al., 2004). Η προσέγγιση αυτή, έχει ενδιαφέρον γιατί έδειξε ότι απλούστερα αλλά καλής προγνωστικής αξίας μοντέλα, μπορούν να κατασκευαστούν, και να συναγωνίζονται σε επιτυχία τα πιο πολύπλοκα Hidden Markov Models (βλ. παρακάτω).

Μια άλλη προσέγγιση, είναι το λεγόμενο Mixture Transition Distribution (MTD) model το οποίο προτάθηκε αρχικά από τον (Raftery, 1985a), και στο οποίο οι πιθανότητες μετάβασης της σχέσης (8.11) προσεγγίζονται από τη σχέση:

$$a_{s_k \ldots s_1 s_0} = P\left(x_i \mid x_{i-1}, x_{i-2}, ..., x_{i-k}\right) = \sum_{j=1}^{k} \lambda_j \alpha_{s_j s_0} \tag{8.15}$$

Έτσι, η επίδραση κάθε παλιάς παρατήρησης (j=1,2…k) λαμβάνεται υπόψη ξεχωριστά και τελικά η ανώτερης τάξης πιθανότητα μετάβασης υπολογίζεται σαν ένας γραμμικός συνδυασμός πιθανοτήτων

μετάβασης πρώτης τάξης. Στην πιο γενική μορφή του μοντέλου (MTDg), η οποία προτάθηκε αργότερα (Raftery, 1985b), κάθε παλιά θέση *j* συνοδεύεται από διαφορετικό πίνακα μεταβάσεων, α^j και έτσι έχουμε:

$$a_{s_k \ldots s_1 s_0} = P\left(x_i \mid x_{i-1}, x_{i-2}, \ldots, x_{i-k}\right) = \sum_{j=1}^{k} \lambda_j \alpha^j_{s_j s_0} \tag{8.16}$$

Προφανώς, για να διατηρεί το μοντέλο την πιθανοθεωρητική του ερμηνεία θα πρέπει να ισχύει:

$$0 \le \sum_{j=1}^{k} \lambda_j \alpha^j_{s_j s_0} \le 1 \tag{8.17}$$

και $\sum_{s_k,\ldots,s_1,s_0,\forall s_0 \in Q} \left(\sum_{j=1}^{k} \lambda_j \alpha^j_{s_j s_0} \right) = 1$, και κατά συνέπεια οι παρακάτω περιορισμοί θα πρέπει να ισχύουν:

$$\sum_{j=1}^{k} \lambda_j = 1, \lambda_j \ge 0 \ \forall \ j = 1, 2, \ldots, k \tag{8.18}$$

Ο Raftery, μελέτησε τις ασυμπτωτικές ιδιότητες του μοντέλου αυτού και έδειξε ότι προσεγγίζει την ανώτερης τάξης αλυσίδα Markov. Παρ' όλα αυτά, αναλυτικές εκφράσεις για τους εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων δεν μπορούν να βρεθούν, και κατά συνέπεια χρειάζεται κάποιου είδους επαναληπτική διαδικασία. Ο Raftery, στην αρχική εργασία (Raftery, 1985a), βελτιστοποίησε την πιθανοφάνεια με μια ρουτίνα γραμμικής βελτιστοποίησης με περιορισμούς (NAG). Ο Bercthold, πρότεινε μια μορφή ευριστικού αλγορίθμου που χρησιμοποιεί το gradient (Berchtold, 2001). Τέλος, μια προσέγγιση που βασίζεται στον αλγόριθμο Expectation-Maximization έγινε πρόσφατα από τους (Lebre & Bourguignon, 2008). Το μοντέλο αυτό είναι ιδιαίτερα υποσχόμενο γιατί έχει μια σειρά από συγκριτικά πλεονεκτήματα (απλότητα, ευκολία στην ερμηνεία των παραμέτρων κ.ο.κ.), αλλά Παρ' όλα αυτά, δεν έχει χρησιμοποιηθεί ακόμα αρκετά σε εφαρμογές στη βιοπληροφορική.

Τυπικά παραδείγματα εφαρμογής των μαρκοβιανών αλυσίδων αφορούν στην εύρεση γονιδίων (gene finding), είτε σε επιβλεπόμενη (Borodovsky & McIninch, 1993; Borodovsky & Peresetsky, 1994) είτε σε μη-επιβλεπόμενη διαδικασία (Audic & Claverie, 1998). Επεκτάσεις του βασικού μοντέλου, όπως οι λεγόμενες interpolated Markov chains ή οι αλυσίδες μεταβλητού μήκους, έχουν επίσης χρησιμοποιηθεί για τον ίδιο σκοπό σε Βακτήρια (Salzberg, Delcher, Kasif, & White, 1998) και σε Ευκαρυωτικούς οργανισμούς (Ohler, Harbeck, Niemann, Noth, & Reese, 1999; Salzberg, Pertea, Delcher, Gardner, & Tettelin, 1999), για την εύρεση μοτίβων σε βιολογικές αλληλουχίες (Barash, Elidan, Friedman, & Kaplan, 2003), για τον εντοπισμό οριζόντιας γονιδιακής μεταφοράς (Dalevi, Dubhashi, & Hermansson, 2006), πρόγνωση της κυτταρικής θέσης των πρωτεϊνών (Yuan, 1999), για ταξινόμηση πρωτεϊνικών αλληλουχιών (Bejerano, et al., 2001), για την ανακατασκευή απλοτύπων στη γενετική (Eronen, Geerts, & Toivonen, 2004) και για τη λεγόμενη ανάλυση πολλών σημείων σε μελέτες γενετικής συσχέτισης (Browning, 2006).

8.2. Hidden Markov Models

8.2.1. Ορισμοί

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 εμφάνισης συμβόλων (emissions). Θεωρώντας μια πρωτεϊνική αλληλουχία **x** μήκους *L* καταλοίπων:

$$\mathbf{x} = x_1, x_2, \ldots, x_{L-1}, x_L \tag{8.19}$$

όπου με x_i συμβολίζουμε τις παρατηρήσεις αποτελούμενες από ένα εκ των 20 αμινοξέων (ή γενικότερα, ενός διακριτού αλφάβητου Ω), στο HMM οι παρατηρήσεις πλέον αποδεσμεύονται από τις καταστάσεις. Συνηθίζεται να συμβολίζουμε την αλληλουχία των καταστάσεων έως μια συγκεκριμένη θέση *i* στην αλληλουχία, με π_i, και να την ονομάζουμε «μονοπάτι» (path). Έτσι, δυο καταστάσεις *k, l* συνδέονται μέσω των πιθανοτήτων μετάβασης a_{kl}, σχηματίζοντας μια αλυσίδα Markov 1ης τάξης. Ο τυπικός ορισμός αυτών των πιθανοτήτων μεταβάσεως, έχει ως εξής:

$$a_{kl} = P(\pi_i = l \mid \pi_{i-1} = k) \tag{8.20}$$

και συμβολίζει πολύ απλά την πιθανότητα, η κατάσταση *k* να δώσει μετάβαση (να προηγείται δηλαδή) προς την κατάσταση *l*. Αντίστοιχα με το απλό μοντέλο Markov, ορίζονται και εδώ ειδικές καταστάσεις ενάρξεως και τερματισμού της αλληλουχίας, οι οποίες για συντομία ονομάζονται B (Begin)

$$a_{Bk} = P(\pi_1 = k \mid B) \tag{8.21}$$

και E (End) αντίστοιχα

$$a_{kE} = P(E \mid \pi_i = k) \tag{8.22}$$

Η σύνδεση των παρατηρηθέντων συμβόλων με τις καταστάσεις, γίνεται μέσω των πιθανοτήτων εμφάνισης συμβόλων:

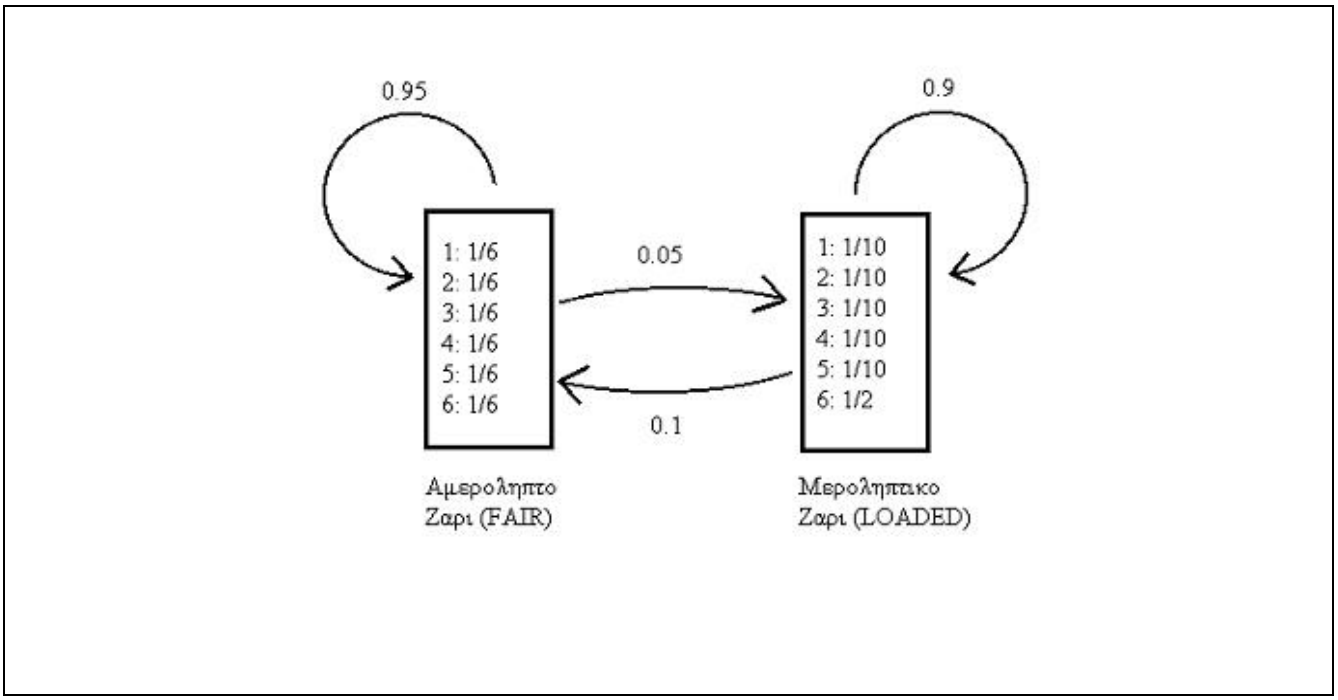
$$e_k(b) = P(x_i = b \mid \pi_i = k) \tag{8.23}$$

οι οποίες, δηλώνουν την πιθανότητα εμφάνισης στη θέση *i* της αλληλουχίας, ενός συγκεκριμένου συμβόλου *b*, δεδομένου ότι το σύστημα βρίσκεται στην κατάσταση *k*. Η από κοινού πιθανότητα μιας αλληλουχίας **x** και του μονοπατιού *π*, υπολογίζεται ως εξής:

$$P(\mathbf{x}, \pi) = P(x_L, x_{L-1}, ..., x_1, \pi) = a_{B\pi_1} \prod_{i=1}^{L} e_{\pi i}(x_i) a_{\pi_i \pi_{i+1}} \tag{8.24}$$

Ένα χαρακτηριστικό παράδειγμα HMM, το οποίο αναφέρεται στη βιβλιογραφία (Durbin, et al., 1998) είναι αυτό του «ανέντιμου καζίνο» (dishonest casino). Στο παράδειγμα αυτό, το καζίνο χρησιμοποιεί κατά περίσταση «κανονικά» αμερόληπτα ζάρια, αλλά έχει τη δυνατότητα να τα αλλάζει (π.χ. με πιθανότητα 0.05 για κάθε ζαριά) και να χρησιμοποιεί κάποια άλλα μεροληπτικά, δηλαδή ζάρια τα οποία ευνοούν κάποιο συγκεκριμένο αποτέλεσμα. Από την μεριά του ο παίκτης, το μόνο που μπορεί να δει είναι τα αποτελέσματα του ζαριού, αλλά δεν μπορεί να ξέρει ποιο ζάρι χρησιμοποιείται κάθε φορά.

Όπως βλέπουμε (Εικόνα 8.3) η πιθανότητα με την οποία αλλάζει το ζάρι από αμερόληπτο σε μεροληπτικό είναι 0.05 (επιλέχθηκε αυθαίρετα σε αυτό το παράδειγμα) και από μεροληπτικό πίσω σε αμερόληπτο, 0.1 (επίσης αυθαίρετη επιλογή). Το μοντέλο αυτό δίκαια ονομάζεται «κρυμμένο» (hidden) γιατί η πραγματική κατάσταση στην οποία βρίσκεται το ζάρι είναι κρυμμένη από τον παίκτη. Προφανώς η μετάβαση από τη μια κατάσταση του ζαριού (μεροληπτικό) στην άλλη (αμερόληπτο) και πάλι πίσω, είναι μια ανέλιξη Markov. Η σημαντική διαφορά του μοντέλου αυτού (HMM) από το απλό Μοντέλο Markov (MM) είναι το ότι σε αυτή την περίπτωση δεν υπάρχει μια προς μια αντιστοίχιση ανάμεσα στα σύμβολα και στις καταστάσεις του μοντέλου. Δηλαδή βλέποντας ένα σύμβολο (πχ. το ζάρι να έχει φέρει αποτέλεσμα 4), δεν μπορούμε να πούμε από ποια κατάσταση έχει διέλθει το μοντέλο για να δώσει το αποτέλεσμα αυτό.

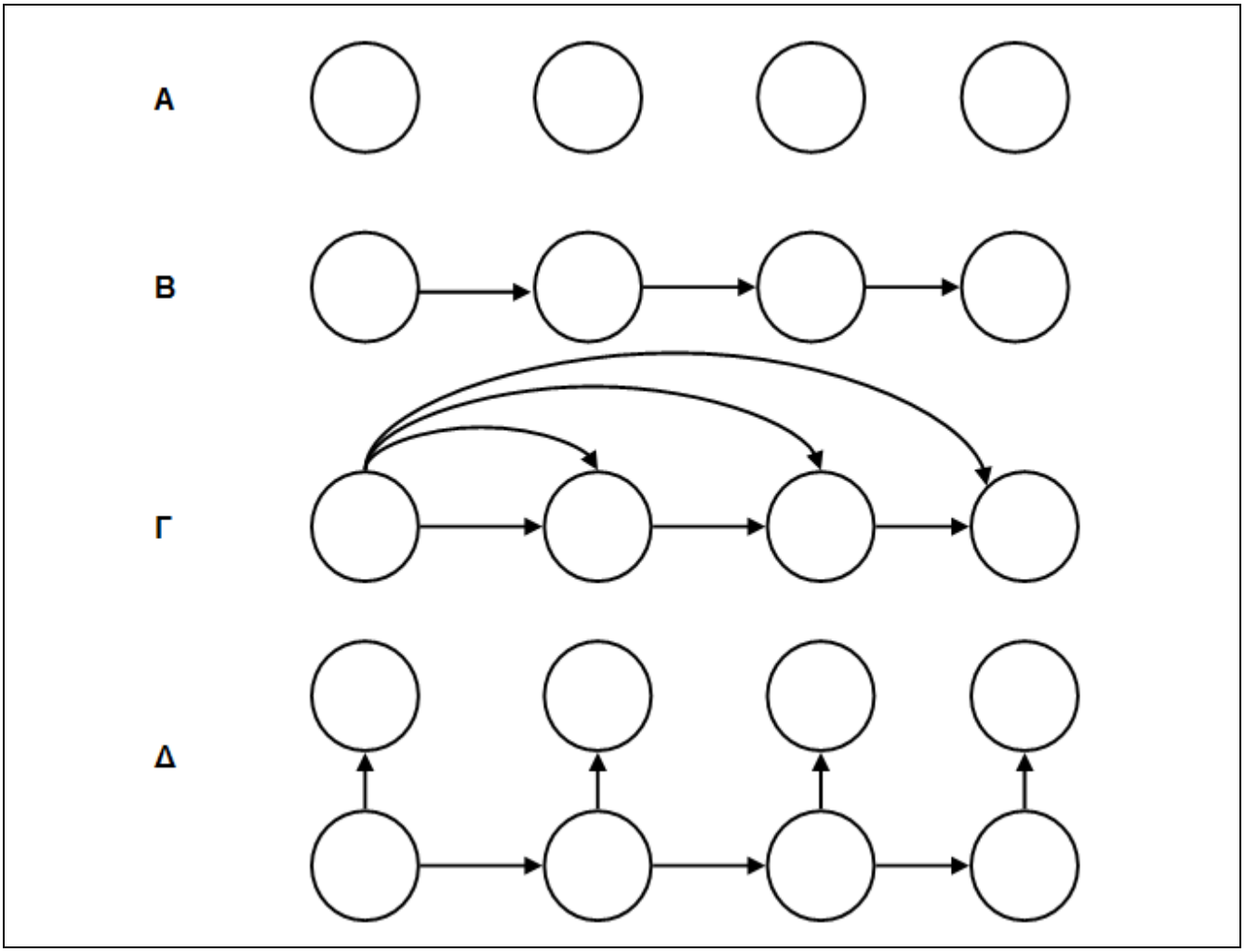


Εικόνα 8.3: *Το παράδειγμα του 'ανέντιμου καζίνο'. Τα δυο παραλληλόγραμμα συμβολίζουν τις δυο καταστάσεις του ζαριού (αμερόληπτο-μεροληπτικό), και τα βέλη τις επιτρεπτές μεταβάσεις. Μέσα σε κάθε κατάσταση, αναγράφονται οι πιθανότητες εμφανίσεως των συμβόλων.*

8.2.2. Τα 3 βασικά ερωτήματα σε ένα HMM

Στην ενότητα αυτή αναπτύσσονται τα 3 βασικά ερωτήματα, τα οποία μπορούν να τεθούν σε ένα HMM όπως διατυπώνονται από τον Rabiner (Rabiner, 1989).

- Δεδομένου ενός μοντέλου θ, πώς μπορούμε να υπολογίσουμε τη συνολική πιθανότητα μια αλληλουχία **x** να έχει εμφανιστεί από αυτό το μοντέλο; Δηλαδή, πώς μπορούμε να υπολογίσουμε την ποσότητα $P(\mathbf{x}|\theta)$;
- Δεδομένου ενός μοντέλου θ και μιας αλληλουχίας **x**, πώς μπορούμε να υπολογίσουμε το μονοπάτι, δηλαδή την αλληλουχία καταστάσεων, με την καλύτερη πιθανότητα; Με άλλα λόγια, πώς μπορούμε να υπολογίσουμε το μονοπάτι π, έτσι ώστε: $\pi^{\max} = \arg\max_{\pi} P(\mathbf{x}, \pi)$;
- Πώς μπορούμε να τροποποιήσουμε τις παραμέτρους του μοντέλου θ, υπό το φως νέων δεδομένων, ώστε να έχουμε καλύτερα μοντέλα; Το πρόβλημα αυτό, ανάγεται στην εκτίμηση παραμέτρων με τη μέθοδο της μέγιστης πιθανοφάνειας. Συγκεκριμένα, ζητάμε τον υπολογισμό των παραμέτρων θ, έτσι ώστε $\theta^{ML} = \arg\max_{\theta} P(\mathbf{x} \mid \theta)$.



Εικόνα 8.4: *Γραφική αναπαράσταση των πιθανοθεωρητικών μοντέλων που έχουμε συναντήσει έως τώρα. Α. Το βασικό μοντέλο της ανεξαρτησίας. Β. Το μαρκοβιανό μοντέλο 1ης τάξης. Γ. Ένα μαρκοβιανό μοντέλο 3ης τάξης. Δ. Το Hidden Markov Model. Στα μοντέλα Α-Γ, οι καταστάσεις αντιστοιχούν σε παρατηρήσιμα σύμβολα. Στο Δ οι καταστάσεις (στην κάτω γραμμή) ακολουθούν μια μαρκοβιανή αλυσίδα 1ης τάξης, κάθε κατάσταση της οποίας «παράγει» με διαφορετική πιθανότητα τα παρατηρήσιμα σύμβολα.*

8.2.3. Υπολογισμός Πιθανοφάνειας

Η σχέση (8.24), είναι όπως είδαμε, η από κοινού πιθανότητα μιας αλληλουχίας **x** και του μονοπατιού π, και δεν μας είναι ιδιαίτερα χρήσιμη γιατί δεν είναι δυνατόν να γνωρίζουμε ποια αλληλουχία καταστάσεων έδωσε γέννηση στην αλληλουχία των παρατηρήσεων. Για να υπολογίσουμε τη συνολική πιθανότητα μιας αλληλουχίας **x** δεδομένου του μοντέλου, θα πρέπει να υπολογίσουμε το άθροισμα τις για όλες τις πιθανές αλληλουχίες καταστάσεων, δηλαδή να αθροίσουμε τη συνεισφορά στη συνολική πιθανότητα, όλων των πιθανών μονοπατιών π.

$$P(\mathbf{x}\mid\theta)=\sum_{\pi}P(\mathbf{x},\pi\mid\theta)=\sum_{\pi}a_{B\pi_1}\prod_{i=1}^{L}e_{\pi i}(x_i)a_{\pi_i\pi_{i+1}} \tag{8.25}$$

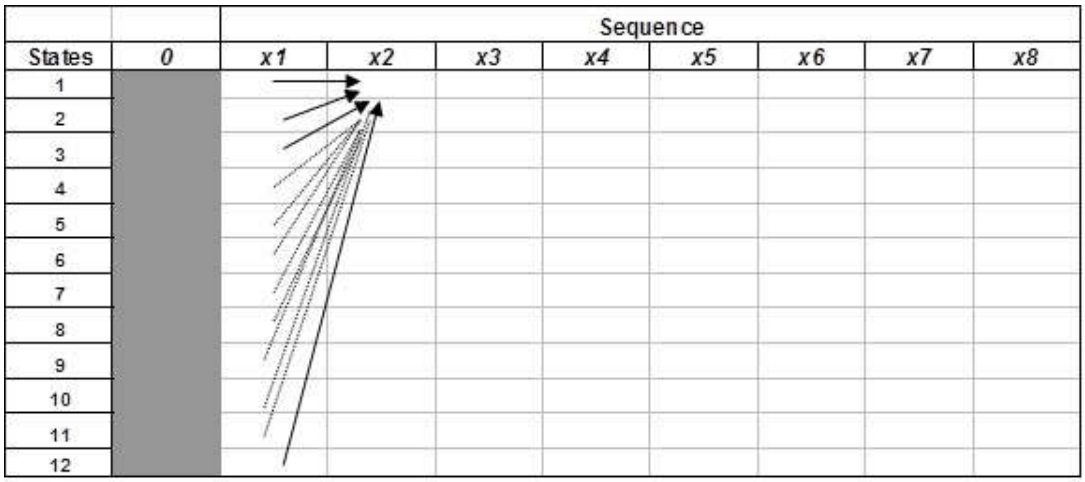
Η ποσότητα αυτή, πρακτικά δεν μπορεί να υπολογιστεί, γιατί μια απλή απαρίθμηση του πλήθους των πιθανών μονοπατιών αυξάνει εκθετικά καθώς αυξάνει το μήκος της αλληλουχίας. Έτσι, π.χ. αν έχουμε ένα μοντέλο με 50 καταστάσεις και μια αλληλουχία μήκους 300 καταλοίπων, τα πιθανά μονοπάτια είναι 50^{300}, αριθμός αστρονομικά μεγάλος. Η συνηθισμένη τακτική σε τέτοιου είδους υπολογιστικά προβλήματα, όπως είδαμε και στο κεφάλαιο 3, είναι ο δυναμικός προγραμματισμός (dynamic programming). Με τη μέθοδο αυτή, το μεγάλο πρόβλημα σπάει σε αρκετά μικρότερα, οι λύσεις των οποίων υπολογίζονται πολύ πιο εύκολα. Ο πιο γνωστός αλγόριθμος δυναμικού προγραμματισμού που έχει προταθεί για το παραπάνω πρόβλημα, είναι ο αλγόριθμος Forward (Εικόνα 8.4), ο οποίος σκιαγραφείται παρακάτω (Durbin, et al., 1998; Rabiner, 1989).

Αλγόριθμος Forward

$$\begin{aligned}&\forall k\neq B,\ i=0: f_B(0)=1,\ f_k(0)=0\\ &\forall 1\le i\le L: f_l(i)=e_l(x_i)\sum_k f_k(i-1)a_{kl}\\ &P(\mathbf{x}|\theta)=\sum_k f_k(L)a_{kE}\end{aligned} \tag{8.26}$$

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

$$f_k(i)=P(x_1,x_2,...,x_i,\pi_i=k) \tag{8.27}$$



Εικόνα 8.5: *Διαγραμματική απεικόνιση του πίνακα Forward, για ένα υποθετικό μοντέλο με 12 καταστάσεις (states) και μια αλληλουχία 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 αλληλουχίας (βέλη).*

Στο πρώτο βήμα, ο πίνακας των $f_k(i)$ αρχικοποιείται, στο δεύτερο συμπληρώνονται οι τιμές του διαδοχικά από την αρχή ως το τέλος της αλληλουχίας, και στο τελευταίο βήμα αθροίζονται για να προκύψει η τελική πιθανοφάνεια. Αν το μοντέλο δεν έχει κατάσταση λήξεως, τότε στο τελευταίο βήμα οι αντίστοιχες πιθανότητες απλώς απαλείφονται. Ο αλγόριθμος αυτός απαιτεί NL υπολογισμούς, γι' αυτό και λέμε ότι είναι της τάξης O(NL).

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 αλληλουχίας από τη θέση i+1 έως το τέλος, δεδομένου ότι στη θέση i συναντάμε την κατάσταση k. Δηλαδή:

$$b_k(i) = P(x_{i+1},...,x_L \mid \pi_i = k) \tag{8.28}$$

Άρα ο αλγόριθμος, διατυπώνεται ως εξής:

Αλγόριθμος Backward

$$\begin{aligned} &\forall k,\ i = L:\ b_k(L) = a_{kE} \\ &\forall\ 1 \le i < L:\ b_k(i) = \sum_l a_{kl} e_l(x_{i+1}) b_l(i+1) \\ &P(\mathbf{x}|\theta) = \sum_l a_{Bl} e_l(x_1) b_l(1) \end{aligned} \tag{8.29}$$

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.

8.2.4. Αποκωδικοποίηση

Στο δεύτερο ερώτημα, θέλουμε να βρούμε ποια είναι η πιο πιθανή αλληλουχία καταστάσεων από την οποία προέκυψε η αλληλουχία των παρατηρήσεων. Αυτό, αναφέρεται στην ουσία, στην αποκωδικοποίηση (decoding) ενός μοντέλου. Ένας παρόμοιος αλγόριθμος με αυτούς που είδαμε παραπάνω, είναι ο αλγόριθμος του Viterbi (Durbin, et al., 1998; Rabiner, 1989).

Αλγόριθμος Viterbi

$$\begin{aligned} &\forall k \ne B,\ i = 0:\ u_B(0) = 1,\ u_k(0) = 0 \\ &\forall\ 1 \le i \le L:\ u_l(i) = e_l(x_i) \max_k \{u_k(i-1) a_{kl}\} \\ &P(\mathbf{x}, \pi^{\max} \mid \theta) = \max_k \{u_k(L) a_{kE}\} \end{aligned} \tag{8.30}$$

Ο αλγόριθμος του Viterbi, είναι στην ουσία όμοιος με τον Forward, με τη μόνη διαφορά να εντοπίζεται 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 $P(\mathbf{x}, \pi^{max}|\theta) \le P(\mathbf{x}|\theta)$. Ένα επιπλέον χαρακτηριστικό του αλγόριθμου αυτού, στο οποίο μοιάζει με τους αλγόριθμους στοίχισης αλληλουχιών, είναι το ότι απαιτεί την ύπαρξη ενός ξεχωριστού πίνακα στον οποίο θα κρατούνται δείκτες (pointers), για την καλύτερη (πιθανότερη) κατάσταση σε κάθε θέση της αλληλουχίας. Με αναδρομή (back-tracking), σε αυτόν τον πίνακα, ανακτά κανείς στο τέλος, το ίδιο το πιθανότερο μονοπάτι.

Εκ των υστέρων αποκωδικοποίηση

Πολλές φορές μπορεί να χρειαστούμε κάτι παραπάνω από τον απλό υπολογισμό του πιο πιθανού μονοπατιού. Μπορεί για παράδειγμα, να θέλουμε να υπολογίσουμε την πιο πιθανή κατάσταση για μια συγκεκριμένη παρατήρηση x_i, ή πιο γενικά μπορεί να θέλουμε να βρούμε την πιθανότητα η παρατήρηση x_i να προέρχεται από μια κατάσταση k, δεδομένης ολόκληρης της αλληλουχίας **x**. Αναζητούμε δηλαδή, την ποσότητα $P(\pi_i=k|\mathbf{x})$. Στην αρχή υπολογίζουμε την από κοινού πιθανότητα της αλληλουχίας **x** και του ενδεχομένου η i παρατήρηση να προέρχεται από την κατάσταση k. Άρα:

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

Ο πρώτος όρος του τελευταίου γινομένου, προκύπτει από τη σχέση (8.27), ενώ ο τελευταίος από τη σχέση (8.28). Άρα, προκύπτει ότι:

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

Τέλος, σύμφωνα με το θεώρημα Bayes θα έχουμε:

$$P(\pi_i = k \mid \mathbf{x}) = \frac{f_k(i) b_k(i)}{P(\mathbf{x})} \tag{8.31}$$

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:

$$\hat{\pi}_i = \arg\max_k P(\pi_i = k \mid \mathbf{x}) \tag{8.32}$$

Η σχέση αυτή μπορεί να είναι πιο χρήσιμη όταν ενδιαφερόμαστε περισσότερο για τον καθορισμό της κατάστασης μιας συγκεκριμένης παρατήρησης και όχι για ολόκληρο το μονοπάτι. Με τον τρόπο αυτό, ορίζουμε το μονοπάτι που μεγιστοποιεί την εκ των υστέρων πιθανότητα.

Μπορούμε επίσης, να ομαδοποιήσουμε κατά κάποιο τρόπο τις καταστάσεις των οποίων η ύπαρξη έχει την ίδια βιολογική σημασία. Με αυτόν το δεύτερο τρόπο, σε γενικές γραμμές, δεν μας ενδιαφέρει η ίδια η αλληλουχία καταστάσεων αλλά κάποια άλλη ιδιότητα που προκύπτει από αυτήν. Για παράδειγμα, στην περίπτωση που έχουμε ένα μοντέλο, στο οποίο οι καταστάσεις, ομαδοποιούνται σε δυο κατηγορίες (π.χ. διαμεμβρανικές-μη διαμεμβρανικές), αν έχουμε μια συνάρτηση $g(k)$ που να ορίζεται πάνω στις ίδιες τις καταστάσεις, με

$$g(k) = \begin{cases} 1, \textbf{αν } k \in C^{TM} \\ 0, \textbf{αν } k \in C^{NTM} \end{cases}$$

τότε

$$G(i \mid \mathrm{x}) = \sum_k P(\pi_i = k \mid x) g(k) \tag{8.33}$$

και αυτή είναι ακριβώς η εκ των υστερών πιθανότητα το αμινοξύ *i* να ανήκει σε μια διαμεμβρανική περιοχή σύμφωνα με το μοντέλο.

Η βασική αδυναμία των παραπάνω τεχνικών αποκωδικοποίησης, οι οποίες ονομάζονται τεχνικές «εκ των υστέρων αποκωδικοποίησης» (posterior decoding), είναι ότι είναι δυνατόν να δώσουν πρόγνωση ασύμβατη με το μοντέλο. Με άλλα λόγια, μπορεί να προβλέψουν ως την πιο πιθανή αλληλουχία καταστάσεων, μια αλληλουχία η οποία δεν θα μπορούσε να προκύψει μέσω του μοντέλου (μη επιτρεπτές μεταβάσεις). Η σοβαρή αυτή αδυναμία, η οποία μπορεί πρακτικά να ακυρώσει τα πλεονεκτήματα του HMM, μπορεί να αντιμετωπιστεί μέσω μιας διαδικασίας «φιλτραρίσματος» και επεξεργασίας των εκ των υστέρων πιθανοτήτων, με έναν αλγόριθμο δυναμικού προγραμματισμού.

Παράδειγμα 8.2.4.1

Έστω ότι έχουμε ένα υποθετικό μοντέλο το οποίο να περιγράφει τις αλληλουχίες DNA, και το οποίο περιγράφεται παρακάτω (είναι ανάλογο με το παράδειγμα με το ζάρι): υπάρχουν 2 διακριτές περιοχές στις αλληλουχίες οι οποίες υποθέτουμε ότι έχουν κάποια λειτουργική σημασία, και οι οποίες διαφέρουν στις πιθανότητες εμφάνισης των 4 βάσεων (emission probabilities). Έτσι στην περιοχή «1» ισχύουν :

$$e_1(A) = P(x_i = A \mid \pi_i = 1) = 0.6,$$
$$e_1(T) = P(x_i = T \mid \pi_i = 1) = 0.1$$
$$e_1(G) = P(x_i = G \mid \pi_i = 1) = 0.1$$
$$e_1(C) = P(x_i = C \mid \pi_i = 1) = 0.1$$

ενώ στην περιοχή «0» ισχύουν αντίστοιχα:

$$e_0(A) = P(x_i = A \mid \pi_i = 0) = 0.25$$
$$e_0(T) = P(x_i = T \mid \pi_i = 0) = 0.25$$

$$e_0(G) = P(x_i = G \mid \pi_i = 0) = 0.25$$

$$e_0(C) = P(x_i = C \mid \pi_i = 0) = 0.25$$

Οι πιθανότητες μεταβάσεως (από τη μία περιοχή στην άλλη) είναι:

$$a_{11} = P(\pi_i = 1 \mid \pi_{i-1} = 1) = 0.9$$

$$a_{10} = P(\pi_i = 0 \mid \pi_{i-1} = 1) = 0.1$$

$$a_{00} = P(\pi_i = 0 \mid \pi_{i-1} = 0) = 0.9$$

$$a_{01} = P(\pi_i = 1 \mid \pi_{i-1} = 0) = 0.1$$

Παρακάτω φαίνεται μια αλληλουχία DNA μήκους 200 βάσεων η οποία προήλθε από το παραπάνω μοντέλο. Στην πρώτη σειρά φαίνεται η αλληλουχία του DNA, ενώ από κάτω φαίνεται η αλληλουχία των καταστάσεων (1/0). Στην Εικόνα 8.7 φαίνεται η διαδικασία αποκωδικοποίησης και οι εκ των υστέρων πιθανότητες για αυτήν την αλληλουχία.

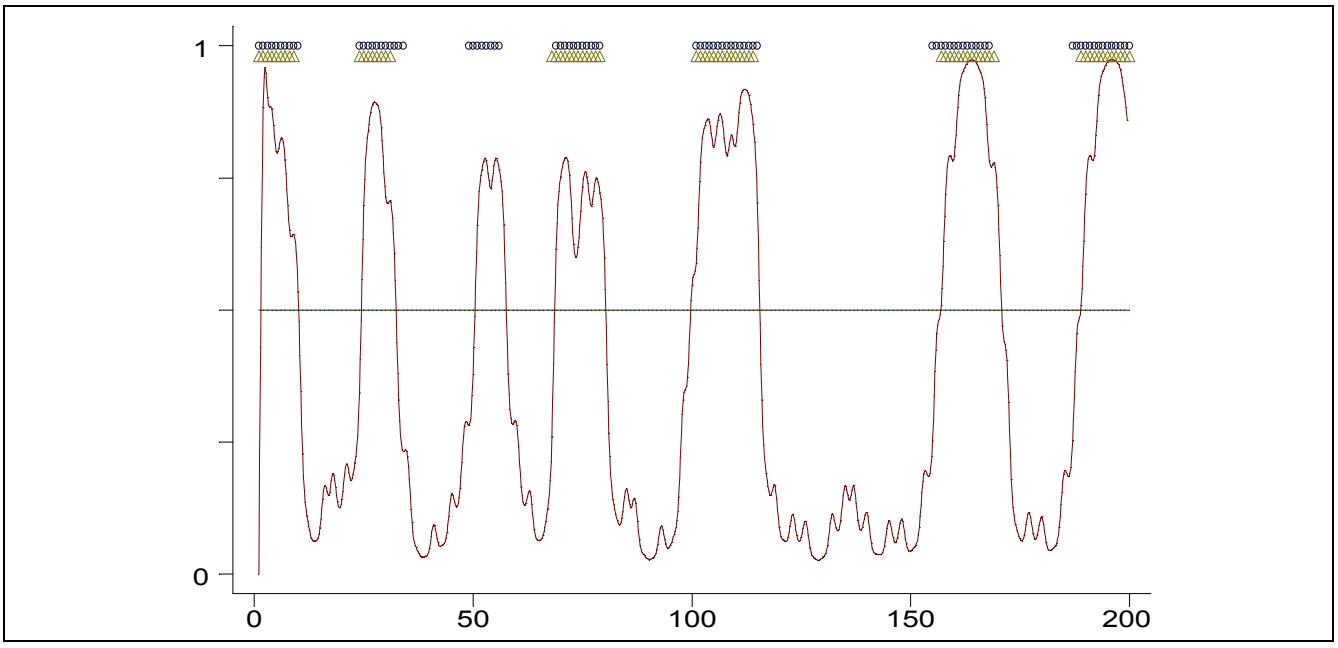
```
AAACAAGAATGCGCACACTACGCAAAAACAATTAGTCGCACTCACGATGAAACAAATTACCACGGTGAA
1111111111000000000000011111111110000000000000111111110000000000001

AACGAATAAACCTCAGAGGCCCAGCGTATATAAACAAGATAAAAACCTAGTCAGCACTCTGACCAGACG
1111111111000000000000000000000011111111111111000000000000000000000000

AGCTCACGACTTGAGGATAAGAAAAAAACAACAGCTCACGACTTGAGGATAAGAAAAAAACA
0000000000000000111111111111110000000000000000001111111111111
```

Εικόνα 8.6: *Μια τυχαία αλληλουχία DNA από το παραπάνω μοντέλο. Κάτω από την αλληλουχία δίνεται και το αντίστοιχο μονοπάτι (0/1)*

Στην παρακάτω εικόνα (Εικόνα 8.7) φαίνεται η αποκωδικοποίηση με τους δυο δυνατούς τρόπους (αποκωδικοποίηση Viterbi και εκ των υστέρων αποκωδικοποίηση) που αναφέρθηκαν παραπάνω.



Εικόνα 8.7: *Οι εκ των υστέρων πιθανότητες και η αποκωδικοποίηση Viterbi για την αλληλουχία που δόθηκε στην Εικόνα 8.6. Όπως φαίνεται και οι δυο μέθοδοι δουλεύουν καλά, προβλέποντας την κατάσταση στην οποία βρίσκονται τα νουκλεοτίδια του DNA. Παρατηρούμε ότι η «εκ των υστέρων αποκωδικοποίηση», είναι λίγο πιο αποτελεσματική καθώς έχει προβλέψει σωστά και τις 7 περιοχές τύπου «1», ενώ η «αποκωδικοποίηση Viterbi» έχει αποτύχει να εντοπίσει μία από αυτές.*

Αν σε μια παραλλαγή του υποθετικού αυτού μοντέλου, αλλάξουν οι πιθανότητες μεταβάσεως ως εξής:

$$a_{11} = P(\pi_i = 1 \mid \pi_{i-1} = 1) = 0.98$$

$$a_{10} = P(\pi_i = 0 \mid \pi_{i-1} = 1) = 0.02$$

$$a_{00} = P(\pi_i = 0 \mid \pi_{i-1} = 0) = 0.97$$

$$a_{01} = P(\pi_i = 1 \mid \pi_{i-1} = 0) = 0.03$$

Τότε, όταν στην παρακάτω αλληλουχία που προέρχεται από το μοντέλο αυτό, εφαρμόσουμε τις δύο παραπάνω μεθόδους αποκωδικοποίησης :

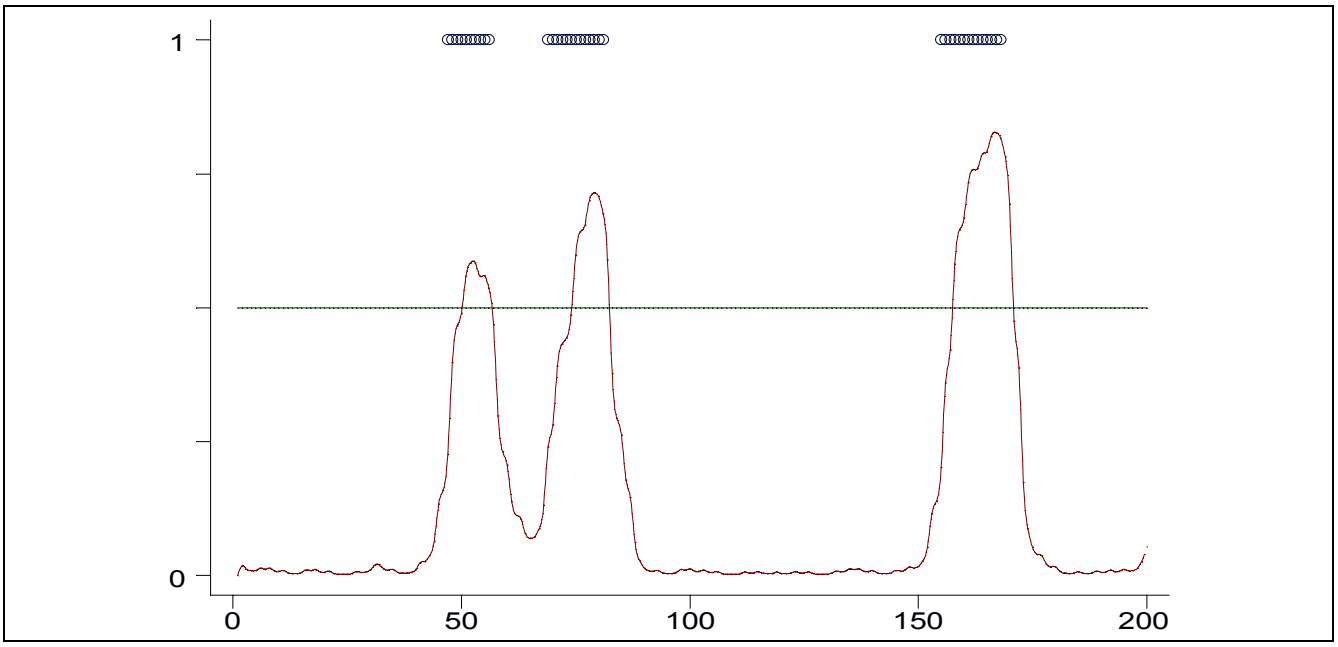
```
GCGCACACTAGCGCACACTACGCCTACGCAATTAGTCGCACTCACGAAGAAACAAATTACCACGGTGAG
000000000000000000000000000000000000000000001111111110000000000001

AACGAATAAAAATCAGAGGCCCAGCGTATATCAGCACTCTGACCACCTAGTCAGCACTCTGACCAGACG
111111111110000000000000000000000000000000000000000000000000000000

AGCTCACGACTTGAGGATAAGAATAGAAAAACAGCTCACGACTTGAGGCACGACTAGCTCAG
000000000000000011111111111111000000000000000000000000000000000
```

Εικόνα 8.8: *Μια τυχαία αλληλουχία DNA από το μοντέλο, όταν οι πιθανότητες αλλάξουν. Κάτω από την αλληλουχία δίνεται και το αντίστοιχο μονοπάτι (0/1)*

Θα έχουμε το επόμενο διάγραμμα:



Εικόνα 8.9: *Οι εκ των υστέρων πιθανότητες και η αποκωδικοποίηση Viterbi για την αλληλουχία από το αλλαγμένο μοντέλο. Όπως είναι φανερό σ' αυτή την περίπτωση ο αλγόριθμος του Viterbi, δεν προβλέπει καμία περιοχή να είναι τύπου «1», καθώς το πιο πιθανό μονοπάτι καταστάσεων δεν επισκέπτεται καμία φορά την κατάσταση «1». Αντίθετα η «εκ των υστέρων αποκωδικοποίηση» δίνει μια αρκετά πιο καλή προσέγγιση της αληθινής κατάστασης. Η «εκ των υστέρων αποκωδικοποίηση» είναι σε γενικές γραμμές καλύτερη μέθοδος, ιδιαίτερα σε περιπτώσεις όπως αυτές (που είδαμε παραπάνω) στις οποίες έχουμε πολύ μικρές (ή μεγάλες) πιθανότητες μεταβάσεως. Πρέπει να τονίσουμε εδώ, ότι στο συγκεκριμένο παράδειγμα, δεν έγινε ανάλυση για να υπολογιστεί το κατώφλι (cut-off value) της πιθανότητας πάνω από την οποία θα θεωρούμε ότι η μέθοδος προβλέπει περιοχές τύπου «1». Το κατώφλι που δεχθήκαμε εδώ (αυθαίρετα) είναι το 0.5 (πράσινη γραμμή), και αν ακολουθήσει ανάλυση ευαισθησίας, θα επιτύχουμε καλύτερη πρόβλεψη.*

Ακόμα πιο δύσκολες καταστάσεις αναμένεται να συναντήσουμε σε πραγματικά προβλήματα, όταν το μοντέλο περιέχει αρκετές καταστάσεις που συνδέονται μεταξύ τους με μια σειρά που έχει κάποιο βιολογικό

νόημα. Για παράδειγμα, στην εύρεση γονιδίων, μπορεί η κατάσταση 1 να αντιστοιχεί σε εξώνιο, η κατάσταση 2 στο σημείο αποκοπής, η κατάσταση 3 σε εσώνιο, ενώ στις διαμεμβρανικές πρωτεΐνες οι καταστάσεις μπορεί να συμβολίζουν αντίστοιχα τις εξωκυττάριες, τις διαμεμβρανικές και τις κυτοπλασμικές περιοχές. Σε όλες αυτές τις περιπτώσεις, η εκ των υστέρων αποκωδικοποίηση είναι δυνατόν, υπό συνθήκες, να δώσει ένα πιθανό μονοπάτι του τύπου 1-3-2, κάτι που όμως είναι αδύνατο από βιολογική σκοπιά. Το όλο επιχείρημα πίσω από τη χρήση κρυπτομαρκοβιανών μοντέλων σε αυτά τα προβλήματα, είναι ότι με τα μοντέλα αυτά μπορούμε να μοντελοποιήσουμε καλύτερα το υπό μελέτη βιολογικό σύστημα. Κατά συνέπεια, πρέπει να αναζητηθούν πιο αποδοτικοί αλγόριθμοι αποκωδικοποίησης οι οποίοι (όπως και ο Viterbi) να διαφυλάσσουν τη γραμματική του μοντέλου.

Εκ των υστέρων αποκωδικοποίηση με χρήση δυναμικού προγραμματισμού

Με τη μέθοδο αυτή, οι εκ των υστέρων πιθανότητες φιλτράρονται μέσα από έναν αλγόριθμο δυναμικού προγραμματισμού, ο οποίος περιορίζει τις πιθανές λύσεις μέσα σε κάποια προαποφασισμένα όρια (π.χ. τα ελάχιστα και μέγιστα μήκη των διαμεμβρανικών τμημάτων) και βρίσκει την ολική καλύτερη τοπολογία για την πρωτεΐνη. Η μέθοδος αυτή προτάθηκε αρχικά από τον Jones και τους συνεργάτες του (Jones, Taylor, & Thornton, 1994).

Το συνολικό πρόβλημα του εντοπισμού της βέλτιστης θέσεως και του μήκους των n διαμεμβρανικών τμημάτων σε μια αλληλουχία m καταλοίπων, υποδιαιρείται σε n μικρότερα προβλήματα αντιμετωπίζοντας κάθε διαμεμβρανική περιοχή ξεχωριστά. Έτσι, με s^{il} συμβολίζουμε το συνολικό σκορ (άθροισμα των εκ των υστέρων πιθανοτήτων που αντιστοιχούν σε διαμεμβρανική περιοχή), για το διαμεμβρανικό τμήμα με μήκος l στη θέση i μιας αλληλουχίας. Τότε, το συνολικό σκορ $S_j^i\left(i:1,2,...,n;\, j:1,2,...,m\right)$, θα υπολογίζεται από την αναδρομική σχέση:

$$S_j^i = \max_{l=l_{\min}\to l_{\max}} \left\{ s_j^{il} + \max_{k=1+l+A\to n} \left\{ S_{j-1}^k \right\} \right\} \tag{8.34}$$

όπου j είναι ο συνολικός αριθμός διαμεμβρανικών τμημάτων, l_{min} και l_{max} τα ελάχιστα και μέγιστα επιτρεπόμενα μήκη των διαμεμβρανικών τμημάτων, και A το ελάχιστο επιτρεπόμενο μήκος στροφής. Μια πιο γενική μορφή αυτού του αλγόριθμου προτάθηκε αργότερα (Fariselli et al., 2003), σαν μια γενικότερη λύση του προβλήματος εντοπισμού υπο-περιοχών με συγκεκριμένα χαρακτηριστικά.

Posterior-Viterbi

Πολύ πρόσφατα, ο Fariselli και οι συνεργάτες του, πρότειναν έναν αλγόριθμο αποκωδικοποίησης ο οποίος συνδυάζει χαρακτηριστικά του αλγορίθμου Viterbi και της εκ των υστέρων αποκωδικοποίησης (Fariselli, Martelli, & Casadio, 2005). Ο αλγόριθμος, βρίσκει το μονοπάτι π^{PV}:

$$\pi^{PV} = \arg\max_{\pi\in\Pi_p} \prod_{i=1}^{L} P\left(\pi_i \mid \mathbf{x}\right)$$

όπου Π_p είναι το σύνολο των επιτρεπτών από το μοντέλο μονοπατιών, και $P(\pi_i=k|\mathbf{x})$ η εκ των υστέρων πιθανότητα για μια κατάσταση, όπως ορίστηκε στη σχέση (8.31). Για να οριστούν τα επιτρεπτά μονοπάτια, χρειαζόμαστε μια δίτιμη συνάρτηση η οποία να παίρνει τιμή 1 για μια επιτρεπτή μετάβαση και 0 για μια μη-επιτρεπτή. Έτσι:

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

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

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

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

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

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

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

Αποκωδικοποίηση Forward

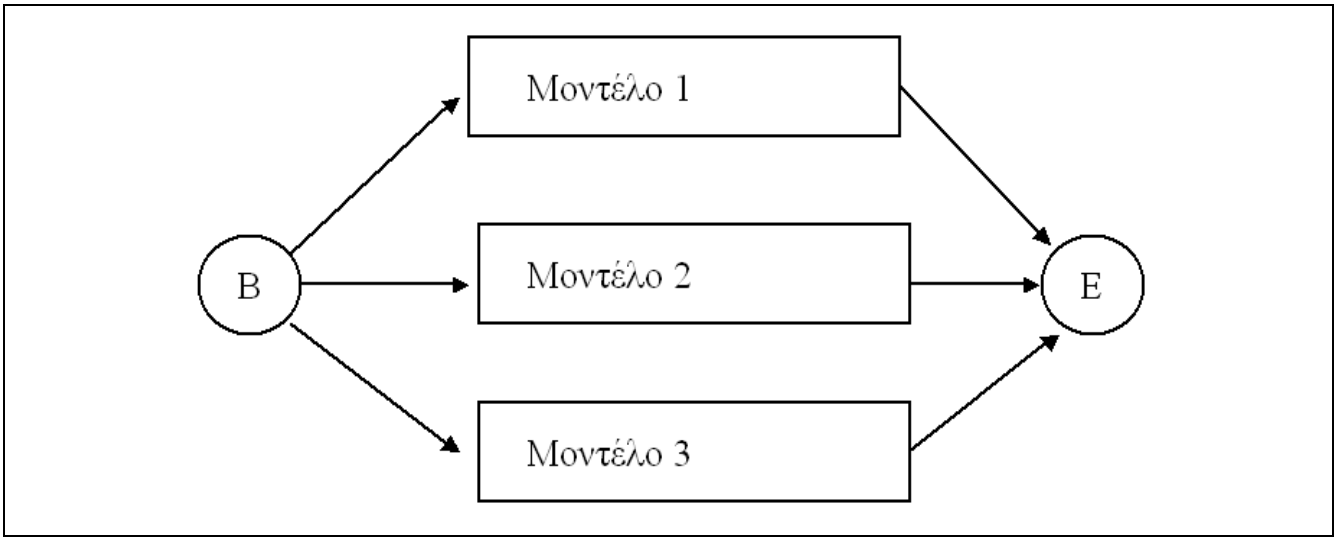
Ένα άλλο είδος αποκωδικοποίησης, είναι επίσης δυνατόν να πραγματοποιηθεί με τη χρήση του αλγόριθμου Forward. Η μέθοδος αυτή, η οποία ονομάζεται πολλές φορές 'Forward Decoding method', είναι χρήσιμη όταν για παράδειγμα έχουμε ένα μοντέλο το οποίο περιέχει πολλαπλούς κλάδους (Εικόνα 8.10). Σε αυτή την περίπτωση, ο κλάδος που δίνει την μεγαλύτερη πιθανότητα, θα είναι ο προτιμώμενος αλλά δεν θα μπορούμε με αυτό τον τρόπο να βρούμε την ακριβή αλληλουχία των καταστάσεων (μπορεί και να μην μας ενδιαφέρει).

Η χρήση της ολικής πιθανότητας από τη σχέση (8.26), δεν αρκεί γιατί μεγαλύτερες αλληλουχίες, κατά κανόνα θα δίνουν και μεγαλύτερη πιθανότητα. Έτσι χρησιμοποιήσαμε μια κανονικοποιημένη σχέση ανάλογη με αυτή της σχέσης (8.10). Συγκεκριμένα, θα έχουμε μια τιμή σκορ:

$$S(\mathbf{x}|\theta) = -\frac{\log P\left(\mathbf{x}|\theta\right)}{L} \tag{8.35}$$

όπου L, θα είναι το μήκος της πρωτεΐνης. Γενικά, μεγάλες τιμές αυτής της συνάρτησης θα είναι ένδειξη ότι η πρωτεΐνη δεν παράγεται από το μοντέλο, ενώ μικρές (που αντιστοιχούν σε μεγάλη πιθανότητα), ένδειξη ότι η πρωτεΐνη ταιριάζει με το μοντέλο. Εναλλακτικές σχέσεις, έχουν προταθεί (Eddy, 1998), οι οποίες χρησιμοποιούν έναν λόγο πηλίκου πιθανοφανειών, συγκρίνοντας για παράδειγμα την πιθανότητα της αλληλουχίας δεδομένου του μοντέλου, με την πιθανότητα της αλληλουχίας δεδομένου ενός 'τυχαίου' μοντέλου θ_0, ενός μοντέλου δηλαδή, το οποίο προϋποθέτει 'τυχαίες' κατανομές των αμινοξέων (μηδενικό – null μοντέλο).

$$S(\mathbf{x}|\theta) = -\frac{\log P\left(\mathbf{x}|\theta\right)}{\log P\left(\mathbf{x}|\theta_{\mathbf{0}}\right)} \tag{8.36}$$



Εικόνα 8.10: *Υποθετική περίπτωση ενός μοντέλου με τρεις κλάδους. Με εφαρμογή του αλγορίθμου Forward, μπορούμε να βρούμε τον κλάδο με τη μεγαλύτερη πιθανότητα.*

Οι πιθανότητες για το τυχαίο αυτό μοντέλο, προκύπτουν από κάποια βάση δεδομένων, και για κάθε κατάλοιπο στην αλληλουχία αντιστοιχούν στα ποσοστά εμφάνισης του καταλοίπου στη βάση. Αθροίζοντας τις πιθανότητες για κάθε κατάλοιπο της αλληλουχίας, έχουμε τελικά την πιθανότητα της αλληλουχίας δεδομένου του μηδενικού μοντέλου.

8.2.5. Εκτίμηση Παραμέτρων στα HMM

Μέγιστη Πιθανοφάνεια

Η εκτίμηση των παραμέτρων ενός στατιστικού-πιθανοθεωρητικού μοντέλου, συνηθίζεται να πραγματοποιείται με τη διαδικασία της Μέγιστης Πιθανοφάνειας (Maximum Likelihood). Οι Εκτιμητές Μέγιστης Πιθανοφάνειας (ΕΜΠ), ορίζονται ως οι τιμές των παραμέτρων θ^{ML} του μοντέλου, οι οποίες μεγιστοποιούν τη συνάρτηση πιθανοφάνειας. Η τελευταία, είναι απλά η από κοινού συνάρτηση κατανομής όλων των παρατηρήσεων, δεδομένων των παραμέτρων του μοντέλου αν θεωρήσουμε τις παραμέτρους σαν τυχαίες μεταβλητές. Άρα:

$$\theta^{ML} = \arg\max_{\theta} P(\mathbf{x} \mid \theta) \tag{8.37}$$

Για λόγους υπολογιστικής απλότητας, συνήθως δουλεύουμε με το λογάριθμο της πιθανοφάνειας $l(\mathbf{x}|\theta)$, ο οποίος μεγιστοποιείται στα ίδια σημεία με αυτή.

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

Αν εργαζόμαστε με τον λογάριθμο της πιθανοφάνειας ο σκοπός είναι να μεγιστοποιήσουμε την τιμή του, ενώ αν εργαζόμαστε με το αντίθετό του (αρνητική λογαριθμική πιθανοφάνεια) ο σκοπός είναι να ελαχιστοποιήσουμε την αντίστοιχη τιμή. Πρέπει να τονιστεί εδώ, ότι οι ατομικές παρατηρήσεις είναι τα αμινοξικά κατάλοιπα. Κατά συνέπεια, αν διαθέτουμε για εκπαίδευση, περισσότερες αλληλουχίες, τις θεωρούμε ανεξάρτητες και η συνολική πιθανοφάνεια είναι το γινόμενο των πιθανοφανειών τους. Άρα, ο λογάριθμός της, είναι το άθροισμα των λογαρίθμων των πιθανοφανειών κάθε αλληλουχίας. Σε όλα τα παρακάτω, θα θεωρούμε ως δεδομένα του συνόλου εκπαίδευσης μια αλληλουχία **x**, με τη γενίκευση σε περίπτωση πολλαπλών ακολουθιών, να είναι τετριμμένη περίπτωση.

Στην ιδανική (όσο και ανέφικτη) περίπτωση, κατά την οποία γνωρίζουμε τα ακριβή μονοπάτια για τις αλληλουχίες 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:

$$\hat{a}_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}} \tag{8.38}$$

και για τις πιθανότητες εμφάνισης συμβόλων,

$$\hat{e}_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b)} \tag{8.39}$$

όπου τα αθροίσματα στους παρονομαστές, εκτείνονται σε όλο το εύρος των παραμέτρων. Πρόβλημα με αυτή την προσέγγιση, μπορεί να υπάρξει, αν κάποια παράμετρος δεν παρατηρηθεί ούτε μια φορά στο σύνολο εκπαίδευσης, με αποτέλεσμα να εμφανιστούν μηδενικές πιθανότητες. Στην περίπτωση αυτή, μπορούμε να προσθέσουμε κάποιες ψευδοτιμές (pseudo-counts), ως εξής:

$$\hat{a}_{kl} = \frac{A_{kl} + r_{kl}}{\sum_{l'} A_{kl'} + \sum_{l'} r_{kl'}} \tag{8.40}$$

$$\hat{e}_k(b) = \frac{E_k(b) + r_k(b)}{\sum_{b'} E_k(b') + \sum_{b'} r_k(b')} \tag{8.41}$$

Αν κάποιος υιοθετήσει μια Μπεϋζιανή προσέγγιση στην ανάλυση των δεδομένων, οι ποσότητες αυτές r_{kl}, $r_k(b)$ αντιστοιχούν στις παραμέτρους μιας κατανομής Dirichlet. Η κατανομή αυτή χρησιμοποιείται ως η εκ των προτέρων (prior) κατανομή, λόγω του ότι είναι συζυγής με την πολυωνυμική κατανομή που ακολουθούν οι πιθανότητες μεταβάσεως και γεννήσεως.

Ο αλγόριθμος Baum-Welch

Στη γενικότερη και πιο συνηθισμένη περίπτωση, κατά την οποία δεν γνωρίζουμε την αλληλουχία των καταστάσεων, το πρόβλημα είναι πιο σύνθετο γιατί πρέπει να εκτιμηθούν οι παράμετροι ταυτόχρονα με τα μονοπάτια. Η λύση, προτάθηκε κατά τη δεκαετία του 1970 από τον Baum και την ερευνητική του ομάδα, και

έγινε γνωστή ως ο αλγόριθμος Baum-Welch (Baum, 1972). Στην πράξη, έχει αποδειχθεί, ότι ο αλγόριθμος αυτός, είναι μια ειδικότερη περίπτωση του αλγορίθμου Expectation-Maximisation (EM) (Dempster, Laird, & Rubin, 1977), ο οποίος προτάθηκε σαν μια γενική προσέγγιση για την εκτίμηση παραμέτρων από δεδομένα με ελλείπουσες τιμές (missing values). Είναι ενδιαφέρον ότι, ο Baum και οι συνεργάτες του, πρότειναν τον αλγόριθμο, χωρίς να γνωρίζουν την γενικευμένη προσέγγιση η οποία προτάθηκε αργότερα (Dempster, et al., 1977). Στα παρακάτω, για λόγους στατιστικής συνέπειας θα παραθέσουμε την παρουσίαση του αλγορίθμου υπό το πρίσμα του αλγορίθμου EM.

Γενικά ο αλγόριθμος EM χρησιμοποιείται για εκτιμήσεις μέγιστης πιθανοφάνειας όταν υπάρχουν ελλείπουσες τιμές (missing values). Εδώ οι ελλείπουσες τιμές είναι οι άγνωστες καταστάσεις π. Σε ολόκληρη την παράγραφο αυτή θ, θ^t,θ^{t+1} κλπ, εννοούμε το σετ των παραμέτρων του μοντέλου σε κάθε επανάληψη (iteration) t και $\mathbf{x}$ γενικά τα δεδομένα μας, δηλαδή τις αλληλουχίες. Ο λογάριθμος της πιθανοφάνειας θα είναι:

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

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:

$$Q(\theta \mid \theta^{t}) = \sum_{\pi} P(\pi|\mathrm{x},\theta^{\mathrm{t}}) \log P(\mathrm{x},\pi \mid \theta) \tag{8.42}$$

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:

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

η πιθανοφάνεια του νέου μοντέλου θα είναι πάντα μεγαλύτερη. Ας δούμε αναλυτικά τώρα τον αλγόριθμο:
Η σχέση (8.24) γίνεται:

$$P(\mathbf{x},\pi \mid \theta) = \prod_{k=1} \prod_{b} [e_k(b)]^{E_k(b,\pi)} \prod_{k=0} \prod_{l=1} a_{kl}^{A_{kl}(\pi)} \tag{8.43}$$

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 τώρα στην (8.42) έχουμε:

$$Q(\theta \mid \theta^{t}) = \sum_{\pi} P(\pi|\mathrm{x},\theta^{\mathrm{t}}) \left[\sum_{k=1} \sum_{b} E_k(b,\pi) \log e_k(b) + \sum_{k=0} \sum_{l=1} A_{kl}(\pi) \log \alpha_{\mathrm{kl}} \right] \tag{8.44}$$

Θα δείξουμε παρακάτω, ότι οι αναμενόμενες τιμές $E_k(b,\pi)$ και $A_{kl}(\pi)$ 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. Πράγματι,

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

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

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

Από τη σχέση (8.27), βλέπουμε ότι:

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

Επιπλέον,

$$P\left(x_{i+1},x_{i+2},...,x_L,\pi_{i+1}=l\,|\,\pi_i=k,\theta\right)=$$
$$P\left(x_{i+1},\pi_{i+1}=l\,|\,\pi_i=k,\theta\right)P\left(x_{i+2},...,x_L\,|\,x_{i+1},\pi_{i+1}=l,\pi_i=k,\theta\right) \tag{8.45}$$

Ο πρώτος όρος του γινομένου, γίνεται:

$$P\left(x_{i+1},\pi_{i+1}=l\,|\,\pi_i=k,\theta\right)=$$
$$=P\left(\pi_{i+1}=l\,|\,\pi_i=k\right)P\left(x_{i+1}\,|\,\pi_{i+1}=l\right)= \tag{8.46}$$
$$=a_{kl}e_l\left(x_{i+1}\right)$$

ενώ ο δεύτερος, με τη βοήθεια της σχέσης (24):

$$P\left(x_{i+2},...,x_L\,|\,x_{i+1},\pi_{i+1}=l,\pi_i=k,\theta\right)=$$
$$=P\left(x_{i+2},...,x_L\,|\,\pi_{i+1}=l\right)=b_l\left(i+1\right) \tag{8.47}$$

Αντικαθιστώντας τις σχέσεις (8.46) και (8.47) στην (8.45), έχουμε:

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

και τελικά:

$$P\left(\mathbf{x},\pi_i=k,\pi_{i+1}=l\,|\,\theta\right)=f_k\left(i\right)a_{kl}e_l\left(x_{i+1}\right)b_l\left(i+1\right) \tag{8.48}$$

απ' όπου με χρήση του θεωρήματος του Bayes:

$$P(\pi_i=k,\pi_{i+1}=l\,|\,\mathrm{x},\theta)=\frac{f_k(i)a_{kl}e_l(x_{i+1})b_l(i+1)}{P(\mathrm{x})} \tag{8.49}$$

όπου $f_k(i)$, $b_k(i)$, είναι οι ποσότητες που υπολογίζονται από τους αλγορίθμους forward και backward. Με όμοιο τρόπο, είδαμε στη σχέση (8.31) ότι:

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

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

$$A_{kl}=\sum_{\pi}P(\pi|\mathbf{x},\theta^t)A_{kl}(\pi)=\frac{1}{P(\mathbf{x})}\sum_i f_k\,(i)a_{kl}e_l(x_{i+1})b_l\,(i+1) \tag{8.50}$$

και αντίστοιχα για τις πιθανότητες γεννήσεως:

$$E_k(b)=\sum_{\pi}P(\pi|\mathbf{x},\theta^t)E_k(b,\pi)=\frac{1}{P(\mathbf{x})}\sum_{\{i|x_i=b\}}f_k^j(i)b_k^j(i) \tag{8.51}$$

και αντικαθιστώντας στη σχέση (8.41) έχουμε:

$$Q(\theta\,|\,\theta^t)=\sum_{k=1}\sum_b E_k(b)\log e_k(b)+\sum_{k=0}\sum_{l=1}A_{kl}\log\alpha_{kl} \tag{8.52}$$

Και αυτή η συνάρτηση τελικά μεγιστοποιείται από τους εκτιμητές των εξισώσεων (8.38) και (8.39). Έτσι σε γενικές γραμμές ο αλγόριθμος των Baum-Welch αποτελείται από το *E*-βήμα (Expectation) στο οποίο υπολογίζονται οι ποσότητες $f_k(i)$, $b_k(i)$, από τους αλγορίθμους forward και backward αντίστοιχα και κατόπιν υπολογίζονται οι αναμενόμενες τιμές για τις πιθανότητες A_{kl}, και $E_k(b)$ από τις σχέσεις (8.50) και (8.51). Έτσι καθορίζεται μονοσήμαντα η συνάρτηση Q. Το *M*-βήμα (Maximization) περιορίζεται στο να τεθούν οι παραπάνω τιμές των A_{kl} και $E_k(b)$ στις σχέσεις (8.38) και (8.39), να υπολογιστούν ξανά οι Ε.Μ.Π. και η πιθανοφάνεια του μοντέλου. Ο αλγόριθμος τερματίζεται απλά όταν οι μεταβολές στην πιθανοφάνεια (log-likelihood) και αντίστοιχα στην συνάρτηση Q μετά από κάποια βήματα είναι μικρότερές από μια προκαθορισμένη τιμή (threshold).

Μέθοδοι Gradient-Descent

Ο αλγόριθμος Baum-Welch, παρουσιάζει όπως είδαμε μια σειρά από θαυμαστά προτερήματα. Το βασικό, είναι ότι είναι μαθηματικά εγγυημένος ότι θα συγκλίνει, ενώ δευτερευόντως, αποδεικνύεται και αρκετά γρήγορος. Το βασικό του μειονέκτημα, είναι ότι απαιτεί η ανανέωση (update) των παραμέτρων να γίνεται, αφού όλο το σύνολο εκπαίδευσης έχει παρουσιαστεί (batch mode of learning). Επιπλέον, είναι «απορροφητικός», υπό την έννοια ότι αν μια παράμετρος μηδενιστεί, δεν υπάρχει περίπτωση να αποκτήσει πλέον διαφορετική τιμή. Αυτά τα δυο μειονεκτήματα, προσπάθησαν να αντιμετωπίσουν οι Baldi και Chauvin (Baldi & Chauvin, 1994).

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

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

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

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

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:

$$\omega^{t+1}=\omega^t-\eta\frac{\partial\ell\left(\mathbf{x}\,|\,\theta\right)}{\partial\omega} \tag{8.53}$$

Άρα το πρόβλημα έγκειται στον υπολογισμό των μερικών παραγώγων του λογαρίθμου της πιθανοφάνειας, ως προς τις παραμέτρους του μοντέλου. Αφού ορίσουμε

$$\ell=-\log P\left(\mathbf{x}\,|\,\theta\right)$$

για τον υπολογισμό των μερικών παραγώγων κινούμαστε ως εξής:

$$\begin{aligned}
\frac{\partial\log P\left(\mathbf{x}\,|\,\theta\right)}{\partial\omega}&=\frac{1}{P\left(\mathbf{x}\,|\,\theta\right)}\frac{\partial P\left(\mathbf{x}\,|\,\theta\right)}{\partial\omega}\\
&=\frac{1}{P\left(\mathbf{x}\,|\,\theta\right)}\frac{\partial P\left(\mathbf{x},\pi\,|\,\theta\right)}{\partial\omega}\\
&=\frac{1}{P\left(\mathbf{x}\,|\,\theta\right)}\sum_{\pi}P\left(\mathbf{x},\pi\,|\,\theta\right)\frac{\partial\log P\left(\mathbf{x},\pi\,|\,\theta\right)}{\partial\omega}\\
&=\sum_{\pi}P\left(\pi\,|\,\mathbf{x},\theta\right)\frac{\partial\log P\left(\mathbf{x},\pi\,|\,\theta\right)}{\partial\omega}
\end{aligned} \tag{8.54}$$

Με χρήση της σχέσης (8.43), και αφού λογαριθμοποιήσουμε, έχουμε για τη μερική παράγωγο ως προς τις πιθανότητες μεταβάσεως:

$$\begin{aligned}
\frac{\partial\log P\left(\mathbf{x},\pi\,|\,\theta\right)}{\partial a_{kl}}&=\frac{\partial\left(\sum_{k=0}\sum_{l=1}A_{kl}\left(\pi\right)\log a_{kl}\right)}{\partial a_{kl}}\\
&=A_{kl}\left(\pi\right)\frac{\partial\left(\sum_{k=0}\sum_{l=1}\log a_{kl}\right)}{\partial a_{kl}}\\
&=\frac{A_{kl}\left(\pi\right)}{a_{kl}}
\end{aligned} \tag{8.55}$$

Άρα, η σχέση (8.54) τελικά γίνεται:

$$\frac{\partial \log P(\mathbf{x} \mid \theta)}{\partial a_{kl}} = \sum_{\pi} P(\pi \mid \mathbf{x}, \theta) \frac{A_{kl}(\pi)}{a_{kl}} = \frac{A_{kl}}{a_{kl}} \tag{8.56}$$

Με εντελώς όμοιο συλλογισμό, υπολογίζουμε και την αντίστοιχη μερική παράγωγο για μια πιθανότητα γεννήσεως:

$$\frac{\partial \log P(\mathbf{x} \mid \theta)}{\partial e_k(b)} = \sum_{\pi} P(\pi \mid \mathbf{x}, \theta) \frac{E_k(\pi, b)}{e_k(b)} = \frac{E_k(b)}{e_k(b)} \tag{8.57}$$

Παρατηρούμε, ότι η μερική παράγωγος του λογαρίθμου της πιθανοφάνειας ως προς τις παραμέτρους του μοντέλου, είναι ίση με την αντίστοιχη μερική παράγωγο της βοηθητικής συνάρτησης Q από τη σχέση (8.52). Μπορούμε δηλαδή, να συμπεράνουμε ότι και οι δυο μέθοδοι κινούνται προς την ίδια κατεύθυνση στην ελαχιστοποίηση ενέργειας, και οδηγούν τελικά στο ίδιο αποτέλεσμα καθώς στο τοπικό ελάχιστο και οι δυο μηδενίζονται (Baldi & Chauvin, 1994).

Για να ολοκληρωθεί ο αλγόριθμος, χρειάζεται ένα ακόμη απαραίτητο βήμα. Πρέπει με κάποιο τρόπο, να περιορίσουμε τον αλγόριθμο, έτσι ώστε οι τιμές των παραμέτρων να είναι πραγματικές πιθανότητες. Αν εφαρμόσουμε τη σχέση (8.53), πραγματοποιώντας Gradient Descent, πάνω στις πραγματικές τιμές των παραμέτρων, είναι πιθανόν να πάρουμε ακόμα και αρνητικούς εκτιμητές για τις πιθανότητες αυτές. Κατά συνέπεια, είναι απαραίτητο να ορίσουμε κάποιες βοηθητικές μεταβλητές, οι οποίες να παίρνουν πάντα τιμές μεταξύ 0 και 1, να πραγματοποιήσουμε την ελαχιστοποίηση και κατόπιν να ανακτήσουμε τις τιμές των πιθανοτήτων. Στη συγκεκριμένη περίπτωση, χρησιμοποιήσαμε τη μέθοδο των Krogh και Riss (Krogh & Riis, 1999), η οποία θεωρεί τον λεγόμενο «soft-max» μετασχηματισμό. Για την ακρίβεια, για τις πιθανότητες μετάβασης a_{kl}, ορίζονται μια σειρά από βοηθητικές παραμέτρους $z_{kl,}$ έτσι ώστε:

$$a_{kl} = \frac{\exp(z_{kl})}{\sum_{l'} \exp(z_{kl'})} \tag{8.58}$$

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

$$z_{kl}^{t+1} = z_{kl}^{t} - \eta \frac{\partial \ell^{t}}{\partial z_{kl}} \tag{8.59}$$

παίρνουμε τις ανανεωμένες παραμέτρους για τις πιθανότητες μετάβασης:

$$a_{kl}^{(t+1)} = \frac{z_{kl}^{(t)} \exp\left(-\eta \frac{\partial \ell^{(t)}}{\partial z_{kl}}\right)}{\sum_{l'} z_{kl'}^{(t)} \exp\left(-\eta \frac{\partial \ell^{(t)}}{\partial z_{kl'}}\right)} \tag{8.60}$$

Με αλλαγή μεταβλητής στη σχέση (8.56), μπορούμε να υπολογίσουμε επίσης τις μερικές παραγώγους του αντίθετου του λογαρίθμου της πιθανοφάνειας ως προς τις βοηθητικές παραμέτρους z_{kl}:

$$\frac{\partial \ell}{\partial z_{kl}} = -\left[A_{kl} - a_{kl} \sum_{l'} A_{kl'}\right] \tag{8.61}$$

Αντικαθιστώντας, τη σχέση (8.61) στη σχέση (8.60), παίρνουμε μια έκφραση η οποία εξαρτάται πλέον μόνο από τις τιμές των παραμέτρων στην προηγούμενη επανάληψη και από τις αναμενόμενες τιμές τους:

$$\alpha_{kl}^{(t+1)} = \frac{\alpha_{kl}^{(t)} \exp\left(-\eta\left[A_{kl} - a_{kl} \sum_{l'} A_{kl'}\right]\right)}{\sum_{l'} \alpha_{kl'}^{(t)} \exp\left(-\eta\left[A_{kl} - a_{kl} \sum_{l'} A_{kl'}\right]\right)} \tag{8.62}$$

Με αυτόν τον τρόπο, η ανανέωση των παραμέτρων επιτυγχάνεται χωρίς να χρειαστεί να υπολογιστούν σε κάποιο ενδιάμεσο βήμα οι βοηθητικές μεταβλητές. Προφανώς, με όμοιο τρόπο εργαζόμαστε και για τις πιθανότητες γέννησης. Με τη μέθοδο αυτή, μπορούμε πλέον να πραγματοποιήσουμε «ομαλή» (smooth) εκπαίδευση, χωρίς τον κίνδυνο μηδενισμού κάποιων παραμέτρων ο οποίος ενυπάρχει στον αλγόριθμο Baum-Welch, αλλά και να πραγματοποιήσουμε τη λεγόμενη διαδικασία «online training», κατά την οποία οι παράμετροι μπορούν να ανανεώνονται κατάλοιπο-κατάλοιπο, και όχι με την παρουσίαση

ολόκληρου του συνόλου εκπαίδευσης. Η μέθοδος αυτή, θα φανεί χρήσιμη παρακάτω, καθώς μόνο με αυτή μπορεί να πραγματοποιηθεί εκπαίδευση Δεσμευμένης Μέγιστης Πιθανοφάνειας (Conditional Maximum Likelihood-CML). Παρ' όλα αυτά το βασικό της μειονέκτημα, καθώς πρόκειται περί ευριστικής μεθόδου, είναι η αυθαίρετη επιλογή της παραμέτρου η (ρυθμός μάθησης), η οποία οδηγεί σε αστάθεια στη διαδικασία εκπαίδευσης και η μικρή της ταχύτητα σύγκλισης στον πολυδιάστατο χώρο των παραμέτρων.

Viterbi training

Τέλος, πρέπει να αναφέρουμε και έναν άλλο, αρκετά απλό τρόπο εκτίμησης παραμέτρων, ο οποίος μάλιστα παρουσιάστηκε σχετικά αργά στη βιβλιογραφία. Ο αλγόριθμος αυτός ονομάζεται Viterbi training ή αλλιώς Segmental *k*-means algorithm και παρουσιάστηκε από τους Juang και Rabiner το 1990 (Juang & Rabiner, 1990). Ο αλγόριθμος στηρίζεται στην απλή ιδέα, να εναλλάσσονται διαδοχικά δύο βήματα: α) εύρεση του καλύτερου μονοπατιού με τον αλγόριθμο Viterbi, και β) θεώρηση του μονοπατιού αυτού ως πραγματικό και χρήση των τύπων (8.34) και (8.35) για την εύρεση των παραμέτρων του μοντέλου.

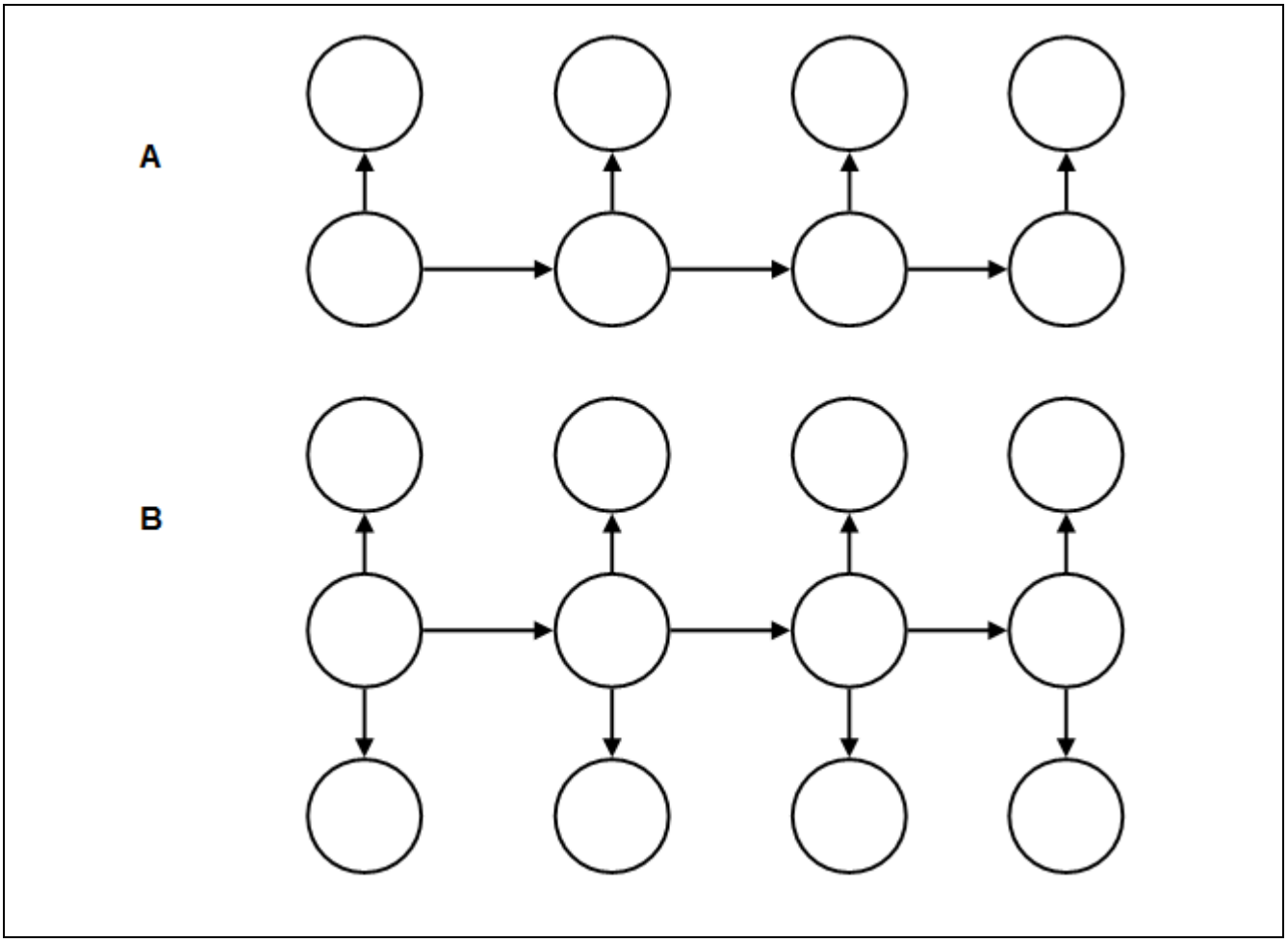
Η ιδέα πίσω από τον αλγόριθμο, είναι αρκετά απλή και έλκει την καταγωγή της από τη βιβλιογραφία της στατιστικής ομαδοποίησης (clustering), όπου και ο αλγόριθμος προτάθηκε για πρώτη φορά το 1968 με το όνομα k-means algorithm (MacQueen, 1967). Με την εύρεση του βέλτιστου μονοπατιού και την παραδοχή ότι αυτό είναι «παρατηρηθέν», η από κοινού πιθανοφάνεια μιας αλληλουχίας και του μονοπατιού αυτού, δίνεται από ένα απλό γινόμενο (8.20), οπότε είναι ανοικτός ο δρόμος για την απλή εφαρμογή των τύπων που δίνουν τους εκτιμητές μέγιστης πιθανοφάνειας. Οι Juang και Rabiner έδειξαν ότι η διαδοχική εφαρμογή των δύο αυτών βημάτων, οδηγεί σε μονοτονική αύξηση της τιμής της πιθανοφάνειας, την οποία και ονόμασαν πιθανοφάνεια βελτιστοποιημένη για τις καταστάσεις (state-optimized likelihood). Ένα επιπλέον ενδιαφέρον σημείο αυτού του αλγορίθμου, είναι το γεγονός ότι, επειδή τα πιθανά μονοπάτια είναι πεπερασμένα, ο αλγόριθμος συγκλίνει σε ένα τοπικό μέγιστο της πιθανοφάνειας σε πεπερασμένο αριθμό βημάτων.

Φυσικά, όπως και όλοι οι αλγόριθμοι βελτιστοποίησης που αναφέρουμε εδώ, δεν υπάρχει εγγύηση ότι ο αλγόριθμος θα εντοπίσει ένα ολικό ελάχιστο της πιθανοφάνειας. Έτσι, εξαρτάται από τις αρχικές τιμές αν το τοπικό αυτό ελάχιστο θα είναι κάποιο το οποίο θα δίνει μοντέλα με σημαντική προγνωστική αξία (το ίδιο φυσικά ισχύει και για τον αλγόριθμο Baum-Welch αλλά και για τους αλγόριθμους gradient). Διαισθητικά, ο αλγόριθμος αυτός είναι μια διακριτή εκδοχή του αλγορίθμου Baum-Welch. Εκεί που ο τελευταίος βρίσκει την αναμενόμενη τιμή αθροίζοντας τη συνεισφορά όλων των πιθανών μονοπατιών, ο πρώτος επιλέγει να κρατήσει τη συνεισφορά μόνο του μονοπατιού με την καλύτερη πιθανότητα. Αυτό έχει σαν αποτέλεσμα, να δίνει, θεωρητικά πάντα, λίγο χειρότερα αποτελέσματα, αλλά από την άλλη έχει το μεγάλο πλεονέκτημα ότι απαιτεί μόνο ένα πέρασμα του αλγορίθμου Viterbi εκεί που ο αλγόριθμος Baum-Welch απαιτεί ένα πέρασμα του Forward και ένα του Backward (κατά συνέπεια, απαιτεί το μισό χρόνο υπολογισμού, πράγμα σημαντικό σε περίπτωση μεγάλων μοντέλων ή/και μεγάλου όγκου δεδομένων).

8.3. Class Hidden Markov Model

8.3.1. Ορισμοί

Το βασικό χαρακτηριστικό των παραπάνω κλασικών προσεγγίσεων στην εκπαίδευση ενός μοντέλου, αποτελεί το γεγονός ότι είναι «μέθοδοι χωρίς επίβλεψη» (unsupervised methods). Το μόνο που έχει να κάνει ο χρήστης, είναι να προσφέρει κάποιες αλληλουχίες-παραδείγματα, και οι αλγόριθμοι θα βρουν την βέλτιστη κατανομή των παραμέτρων για να περιγράψουν τα δεδομένα αυτά. Ένα βασικό πρόβλημα στη βιολογία, ανακύπτει στην περίπτωση κατά την οποία θέλουμε να εκπαιδεύσουμε ένα μεγάλο μοντέλο, το οποίο να περιγράφει με μεγάλη ακρίβεια (χρησιμοποιώντας διαφορετικές καταστάσεις) διαφορετικές περιοχές μέσα σε μια πρωτεϊνική αλυσίδα.



Εικόνα 8.11: *Γραφική αναπαράσταση του κλασικού HMM και του CHMM. Α. Το βασικό HMM όπως το έχουμε περιγράψει έως τώρα. Οι καταστάσεις (στην κάτω γραμμή) ακολουθούν μια μαρκοβιανή αλυσίδα 1ης τάξης, κάθε κατάσταση της οποίας «παράγει» με διαφορετική πιθανότητα τα παρατηρήσιμα σύμβολα Β. Το CHMM στην πιο γενική του μορφή, μπορεί να θεωρηθεί ως ένα μοντέλο που σε κάθε κατάσταση «παράγει» ταυτόχρονα δύο σειρές από παρατηρήσιμα σύμβολα. Η μία είναι τα σύμβολα της αλληλουχίας (όπως ακριβώς το HMM), ενώ η άλλη είναι η αλληλουχία των σημάνσεων. Η αλληλουχία των καταστάσεων του μοντέλου, εξακολουθεί να είναι μαρκβιανή 1ης τάξης.*

Χαρακτηριστικότατο παράδειγμα, είναι αυτό της πρόγνωσης των διαμεμβρανικών περιοχών. Είναι φυσικό, να θέλουμε να κατασκευάσουμε ένα πολύπλοκο μοντέλο, το οποίο να περιέχει αρκετές και διαφορετικές καταστάσεις για την διαμεμβρανική περιοχή, άλλες για την εξωτερική περιοχή της μεμβράνης και άλλες για την εσωτερική, έτσι ώστε να μπορούμε να μοντελοποιήσουμε καλύτερα το πρόβλημα και να αποτυπώσουμε πιο αποτελεσματικά την πρότερη βιολογική γνώση. Στην περίπτωση αυτή, θα έπρεπε να χωρίσουμε τις περιοχές, να εκπαιδεύσουμε 3 διαφορετικά μοντέλα και κατόπιν να τα ενώσουμε αυθαίρετα με κάποιες πιθανότητες μετάβασης. Το πρόβλημα αυτό παρακάμπτεται, αν χρησιμοποιήσουμε μια μέθοδο εκμάθησης «με επίβλεψη» (supervised learning). Η μέθοδος αυτή βασίζεται στις σημασμένες αλληλουχίες (labeled sequences), προτάθηκε από τον Krogh και αντιστοιχεί στο λεγόμενο Class Hidden Markov Model (Anders. Krogh, 1994). Συγκεκριμένα, με την τεχνική αυτή, κάθε αλληλουχία συμβόλων

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

συνοδεύεται, και από μια αλληλουχία σημάνσεων ή ετικετών (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 ταύτισης 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.

8.3.2. Πιθανοφάνεια

Όπως γίνεται πλέον φανερό, με την εισαγωγή των σημάνσεων, ένας τρόπος να επιτύχουμε «επιβλεπόμενη μάθηση», 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:

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

Η μοναδική διαφορά της σχέσης αυτής, με τη σχέση (8.25), είναι ότι η άθροιση πρέπει να γίνει μόνο για αυτά τα μονοπάτια Π_y, τα οποία διέρχονται μέσω καταστάσεων οι οποίες είναι συμβατές με τη σήμανση. Η ολική από κοινού πιθανότητα των ακολουθιών και των σημάνσεων δεδομένου του μοντέλου, μπορεί να υπολογιστεί με κάποιες τετριμμένες τροποποιήσεις των γνωστών αλγορίθμων Forward και Backward, που συναντήσαμε παραπάνω. Η μόνη διαφορά, έγκειται στον πολλαπλασιασμό των ενδιάμεσων μεταβλητών, με τη δίτιμη συνάρτηση (0,1), η οποία δείχνει την συμφωνία καταστάσεων και σημάνσεων. Κατά συνέπεια, ο τροποποιημένος αλγόριθμος Forward, είναι:

Τροποποιημένος αλγόριθμος Forward

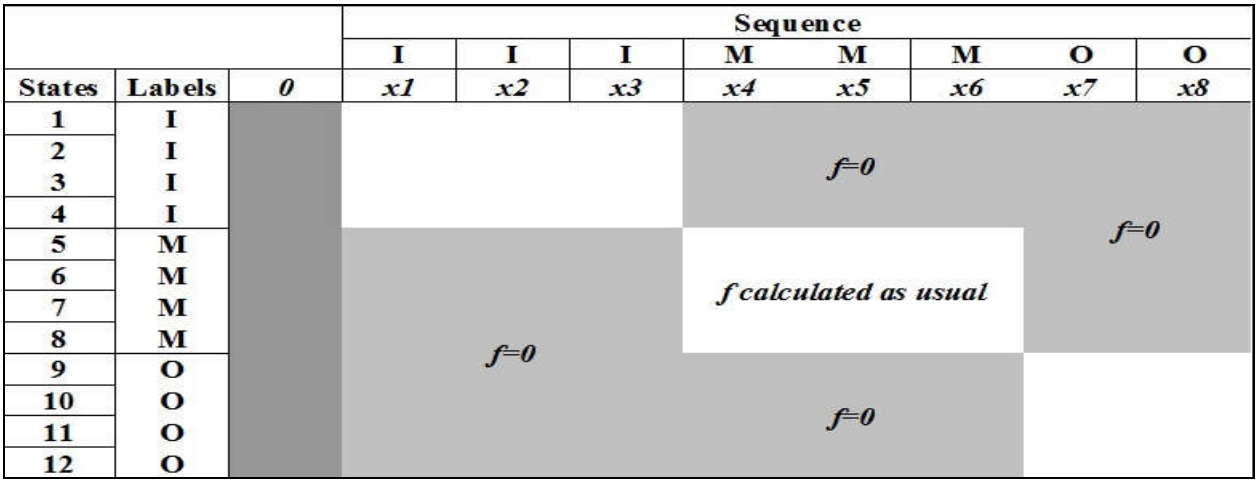
$$\begin{aligned}&\forall k\neq B,\ i=0:\ f_B(0)=1,\ f_k(0)=0\\&\forall 1\le i\le L:\ f_l(i)=e_l(x_i)\delta_l\left(y_i\right)\sum_k f_k(i-1)a_{kl}\\&P(\mathbf{x},\mathbf{y}|\theta)=\sum_k f_k(L)a_{kE}\end{aligned}\qquad(8.63)$$

Εντελώς όμοια, ο τροποποιημένος αλγόριθμος Backward, θα είναι:

Τροποποιημένος αλγόριθμος Backward

$$\begin{aligned}&\forall k,\ i=L:\ b_k(L)=a_{kE}\\&\forall\ 1\le i<L:\ b_k(i)=\sum_l a_{kl}e_l(x_{i+1})\delta_l\left(y_{i+1}\right)b_l(i+1)\\&P(\mathbf{x},\mathbf{y}|\theta)=\sum_l a_{Bl}e_l(x_1)b_l(1)\end{aligned}\qquad(8.64)$$

Διαισθητικά, οι αλγόριθμοι αυτοί απλώς μηδενίζουν τις περιοχές των πινάκων Forward και Backward, στις οποίες δεν υπάρχει συμφωνία καταστάσεων και σημάνσεων (Εικόνα 8.12). Λόγω του ότι το σύνολο των επιτρεπτών μονοπατιών Πy, είναι υποσύνολο του συνόλου όλων των πιθανών μονοπατιών Π, αντιλαμβανόμαστε ότι $P(\mathbf{x},\mathbf{y}|\theta)\le P(\mathbf{x}|\theta)$.



Εικόνα 8.12: *Διαγραμματική απεικόνιση του πίνακα Forward για σημασμένες αλληλουχίες. Έχουμε ένα μοντέλο με 12 υποθετικές καταστάσεις, και μια αλληλουχία 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.*

8.3.3. Εκτίμηση παραμέτρων

Μέγιστη Πιθανοφάνεια

Με την εισαγωγή των αλγορίθμων για σημασμένες αλληλουχίες, 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 αλληλουχίες, ισχύουν με μικρές παραλλαγές και εδώ. Λόγω του ότι, οι αλληλουχίες και οι σημάνσεις είναι ανεξάρτητες, οι σχέσεις (8.65) και (8.66) γίνονται αντίστοιχα:

$$A_{kl} = \frac{1}{P(\mathbf{x},\mathbf{y} \mid \theta)} \sum_i f_k(i) a_{kl} e_l(x_{i+1}) \delta_l\left(y_{i+1}\right) b_l(i+1) \tag{8.67}$$

$$E_k(b) = \frac{1}{P(\mathbf{x},\mathbf{y} \mid \theta)} \sum_{\{i \mid x_i^j = b\}} f_k(i) b_k(i) \tag{8.68}$$

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.

Δεσμευμένη Μέγιστη Πιθανοφάνεια

Με την εισαγωγή της έννοιας των σημασμένων ακολουθιών, ο Krogh (Anders. Krogh, 1994) πρότεινε και μια νέα μέθοδο εκπαίδευσης, αυτήν της Δεσμευμένης Μέγιστης Πιθανοφάνειας (Conditional Maximum Likelihood). Με το κριτήριο αυτό, αναζητούμε πλέον να μεγιστοποιήσουμε την πιθανότητα των σημάνσεων, δεδομένων των ακολουθιών και του μοντέλου:

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

Ο Krogh, έδειξε επίσης (Anders. Krogh, 1994), ότι η προσέγγιση αυτή αποτελεί γενική περίπτωση της από παλιότερα γνωστής διαδικασίας εκπαίδευσης με το κριτήριο της Μέγιστης Αμοιβαίας Πληροφορίας (Maximum Mutual Information) όπως αναφέρεται στον (Rabiner, 1989). Ο αρνητικός λογάριθμος αυτής της δεσμευμένης πιθανοφάνειας, μπορεί να εκφραστεί ως η διαφορά:

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

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). Από τις σχέσεις (8.56) και (8.57), μπορούμε εύκολα να υπολογίσουμε τις αναμενόμενες τιμές και τις μερικές παραγώγους των παραμέτρων του μοντέλου:

$$\frac{\partial \ell}{\partial a_{kl}} = \frac{\partial \ell_c}{\partial a_{kl}} - \frac{\partial \ell_f}{\partial a_{kl}} = -\frac{A^c_{kl} - A^f_{kl}}{a_{kl}} \tag{8.69}$$

$$\frac{\partial \ell}{\partial e_k(b)} = \frac{\partial \ell_c}{\partial e_k(b)} - \frac{\partial \ell_f}{\partial e_k(b)} = -\frac{E^c_k(b) - E^f_k(b)}{e_k(b)} \tag{8.70}$$

Όπως είναι φανερό, ο αλγόριθμος Baum-Welch δεν μπορεί να χρησιμοποιηθεί, λόγω του ότι η διαφορά των αναμενόμενων τιμών θα δώσει αρνητικές εκτιμήσεις για τις παραμέτρους. Παρ' όλα αυτά, με τη χρήση των μερικών παραγώγων μπορούμε να προχωρήσουμε σε εκπαίδευση με τη μέθοδο Gradient Descent. Δουλεύοντας με τον ίδιο μετασχηματισμό όπως και στις προηγούμενες παραγράφους για τις πιθανότητες μεταβάσεως, η μερική παράγωγος της πιθανοφάνειας ως προς τις βοηθητικές μεταβλητές, θα είναι:

$$\frac{\partial \ell}{\partial z_{kl}} = -\left(A^c_{kl} - A^f_{kl} - a_{kl} \sum_{l'} \left(A^c_{kl'} - A^f_{kl'} \right) \right) \tag{8.71}$$

και τελικά η σχέση η οποία θα δώσει τις ανανεωμένες τιμές των παραμέτρων στην επανάληψη *t*, θα είναι, εντελώς ανάλογα:

$$\alpha^{(t+1)}_{kl} = \frac{\alpha^{(t)}_{kl} \exp\left(-\eta \left[A^c_{kl} - A^f_{kl} - a_{kl} \sum_{l'} \left(A^c_{kl'} - A^f_{kl'} \right) \right] \right)}{\sum_{l'} \alpha^{(t)}_{kl'} \exp\left(-\eta \left[A^c_{kl} - A^f_{kl} - a_{kl} \sum_{l'} \left(A^c_{kl'} - A^f_{kl'} \right) \right] \right)} \tag{8.72}$$

Μια βασική αδυναμία αυτής της μεθόδου, η οποία θεωρείται και «μέθοδος εκπαίδευσης προσανατολισμένη στο διαχωρισμό» (Discriminative Training), είναι ότι απαιτεί διπλάσιο υπολογιστικό χρόνο και μνήμη στον υπολογιστή, καθώς χρειάζονται δυο «περάσματα» των αλγορίθμων για τις δυο διαφορετικές πιθανοφάνειες που υπολογίζονται. Παρ' όλα αυτά, τα πλεονεκτήματα τα οποία προσδίδει μια τέτοια διαδικασία υπερτερούν, κυρίως λόγω της καλύτερης ικανότητας πρόγνωσης που προσφέρει. Το μεγάλο μειονέκτημα του αλγορίθμου, είναι η μικρή του ταχύτητα και η ανάγκη εύρεσης μιας βέλτιστης τιμής για την επιλογή της παραμέτρου η (ρυθμός μάθησης). Το πρόβλημα αυτό, λύθηκε σε μεγάλο βαθμό με τη δημιουργία ενός αλγορίθμου ο οποίος χρησιμοποιεί διαφορετικούς ρυθμούς μάθησης για κάθε παράμετρο, αλλά έχει και την ιδιότητα να τους αναπροσαρμόζει κατά τη διάρκεια της διαδικασίας εκπαίδευσης (Bagos, Liakopoulos, & Hamodrakas, 2004).

8.3.4. Αποκωδικοποίηση

1-best Decoding

Καθώς είδαμε ότι με την προσέγγιση των σημασμένων ακολουθιών, αποκτάμε μια αντιστοίχιση των καταστάσεων με τις σημάνσεις, μια λογική απορία θα ήταν αν θα μπορεί να έχει κανείς μια μέθοδο αποκωδικοποίησης η οποία να βρίσκει την πιο πιθανή αλληλουχία (μονοπάτι) των σημάνσεων και όχι των καταστάσεων. Μια αλληλουχία καταστάσεων σύμφωνα με τον ορισμό που δώσαμε, αντιστοιχεί μονοσήμαντα σε μια αλληλουχία σημάνσεων αλλά το αντίθετο δεν ισχύει, καθώς είναι δυνατόν να έχουμε περισσότερες από μια αλληλουχίες καταστάσεων οι οποίες δίνουν την ίδια σήμανση κατά μήκος της αλληλουχίας συμβόλων. Ακριβής αλγόριθμος, ο οποίος να υπολογίζει την *ολικά καλύτερη* αλληλουχία σημάνσεων δεν υπάρχει, αλλά έχουν προταθεί προσεγγιστικές λύσεις.

Ο αλγόριθμος 1-best (Krogh, 1997), είναι μια τροποποίηση του αλγορίθμου N-best, ο οποίος είχε προταθεί παλαιότερα για αναγνώριση ομιλίας (Schwartz & Chow, 1990). Στην ουσία, πρόκειται για έναν ευριστικό αλγόριθμο δυναμικού προγραμματισμού, ο οποίος αναζητά την εύρεση της πιο πιθανής αλληλουχίας σημάνσεων ymax αντί αυτή της πιο πιθανής αλληλουχίας καταστάσεων. Ο αλγόριθμος, για κάθε

θέση i της αλληλουχίας, αποθηκεύει όλες τις πιθανές «ενεργές υποθέσεις» hi-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 yi και διαλέγει την καλύτερη. Η όλη διαδικασία επαναλαμβάνεται ως το τέλος της αλληλουχίας. 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

$$
\begin{aligned}
&i = 1: \ \gamma_l(h_1) = a_{Bl} e_l(x_1) \\
&\forall 1 < i \le L: \ \gamma_l(h_i \mathrm{y}_i) = e_l(x_i) \sum_k \gamma_k(h_{i-1}) a_{kl} \\
&P\left(\mathbf{x}, \mathbf{y}^{\max} \mid \theta\right) = \sum_k \gamma_k(h_L) a_{kE}
\end{aligned}
\tag{8.73}
$$

Η πιθανότητα της βέλτιστης αυτή σήμανσης, είναι πάντα μεγαλύτερη ή ίση από την πιθανότητα του βέλτιστου μονοπατιού καταστάσεων, καθώς πολλά διαφορετικά μονοπάτια καταστάσεων συνεισφέρουν σε αυτή, άρα:

$$P\left(\mathbf{x}, \pi^{\max} \mid \theta\right) \le P\left(\mathbf{x}, \mathbf{y}^{\max} \mid \theta\right) \le P\left(\mathbf{x} \mid \theta\right)$$

Optimal Accuracy Posterior Decoder

Πριν από μερικά χρόνια, ο Kall και συνεργάτες παρουσίασαν έναν εναλλακτικό αλγόριθμο, τον λεγόμενο Optimal Accuracy Posterior Decoder (Kall, Krogh, & Sonnhammer, 2005). Ο αλγόριθμος αυτός μοιάζει πάρα πολύ με τον Posterior-Viterbi, αλλά διαθέτει κάποιες διαφοροποιήσεις οι οποίες εμφανίζονται ειδικά στην περίπτωση του CHMM (αν και γενικά, οι δύο αλγόριθμοι αποδίδουν σχεδόν ταυτόσημα στα περισσότερα προβλήματα που έχουν εφαρμοστεί). Ο αλγόριθμος, λειτουργεί ως εξής: για κάθε θέση στην αλληλουχία, αθροίζει τις εκ των υστέρων πιθανότητες της κάθε σήμανσης (posterior label probabilities -PLPs), χρησιμοποιώντας τη σχέση (8.29), και μετά υπολογίζοντας μόνο τις επιτρεπτές μεταβάσεις, υπολογίζει με έναν αλγόριθμο τύπου Viterbi τη βέλτιστη αλληλουχία των σημάνσεων που μεγιστοποιεί την ποσότητα:

$$\pi^{OAPD} = \arg\max_{\pi} \sum_{i=1}^{L} \left\{ \delta\left(\pi_i, \pi_{i+1}\right) \left(\sum_k P\left(\pi_i \mid \mathbf{x}\right) \lambda_k\left(c\right) \right) \right\} \tag{8.74}$$

Αλγόριθμος Optimal Accuracy Posterior Decoder

$$
\begin{aligned}
&\forall k \ne B, \ i = 0: \ A_B(0) = 0, \ A_k(0) = -\infty \\
&\forall \ 1 \le i \le L: \ A_l(i) = P\left(y_i = c^l \mid \mathbf{x}, \theta\right) + \max_k \left\{ A_k(i-1) \delta\left(k, l\right) \right\} \\
&P(\mathbf{x}, \pi^{OAPD} \mid \theta) = \max_k \left\{ A_k(L) \delta\left(k, E\right) \right\}
\end{aligned}
\tag{8.75}
$$

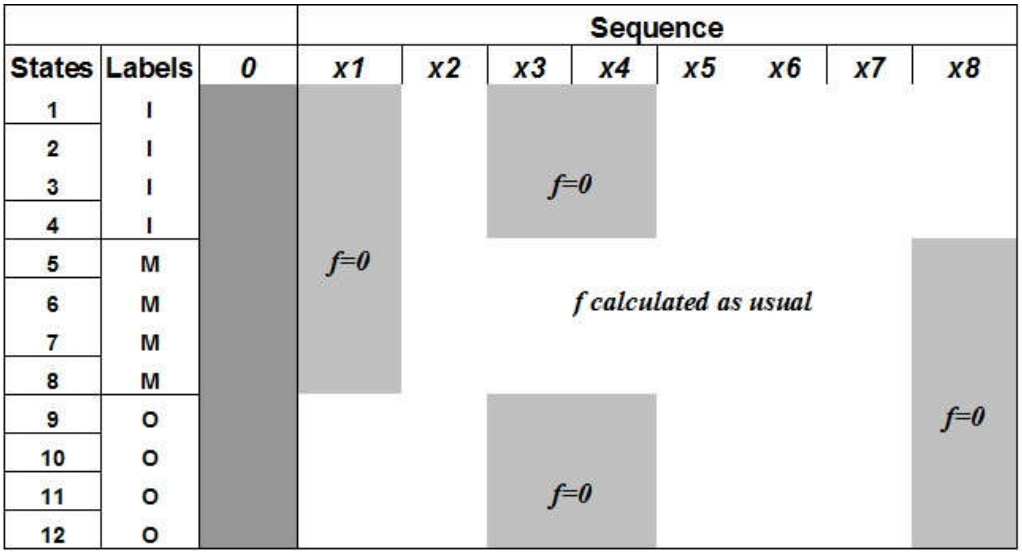
Όπως είπαμε, ο αλγόριθμος αυτός μοιάζει πολύ με τον Posterior-Viterbi και οι μόνες διαφορές τους είναι ότι, α) χρησιμοποιεί τις εκ των υστέρων πιθανότητες των σημάνσεων και όχι των καταστάσεων, και β) αντί για το γινόμενο των πιθανοτήτων αυτών, μεγιστοποιεί το άθροισμά τους. Αυτό, είναι αναγκαίο γιατί ο Posterior-Viterbi υπολογίζει τελικά ένα μονοπάτι καταστάσεων, ενώ ο Optimal Accuracy Posterior Decoder μια αλληλουχία από σημάνσεις που είναι όμως συμβατές με το μοντέλο, αλλά ενδέχεται να περιέχουν πολλά εναλλακτικά μονοπάτια. Για το λόγο αυτό, οι τελική πιθανότητα που αποδίδει ο αλγόριθμος αυτός, δεν είναι συγκρίσιμη σε απόλυτες τιμές με τις πιθανότητες των άλλων αλγορίθμων.

8.3.5. Δεσμευμένη πρόγνωση και αλγόριθμοι για ενσωμάτωση πειραματικής πληροφορίας

Σε διάφορα βιολογικά προβλήματα, όπως για παράδειγμα στην περίπτωση της πρόγνωσης των διαμεμβρανικών πρωτεϊνών, είναι γνωστό ότι η ενσωμάτωση μιας ακόμα και περιορισμένης πειραματικά

προσδιορισμένης πληροφορίας σχετικά με την τοπολογία θα βελτίωνε κατά ένα μεγάλο μέρος την απόδοση ακόμα και των καλύτερων μεθόδων. Με την ανάπτυξη εύκολων και γρήγορων πειραματικών τεχνικών βασισμένων σε συντήξεις γονιδίων (gene fusions), με την οποία καθορίζεται η θέση του αμινοτελικού άκρου μιας πρωτεΐνης, προτάθηκε ότι αυτές οι τεχνικές συνδυαζόμενες μαζί θα βελτιώσουν κατά ένα μεγάλο μέρος την απόδοση των προγνωστικών μεθόδων και την εφαρμογή τους σε πλήρως προσδιορισμένα γονιδιώματα (Drew et al., 2002; Melen, Krogh, & von Heijne, 2003). Δεδομένα στην βιβλιογραφία υπάρχουν αρκετά τα οποία δείχνουν και άλλους εναλλακτικούς τρόπους προσδιορισμού της θέσης διαφόρων τμημάτων της αλληλουχίας (αντισώματα, πρωτεόλυση κλπ) αλλά οι πιο ολοκληρωμένες πειραματικές αποδείξεις σε μεγάλη κλίμακα γι' αυτήν την βελτίωση, ήρθαν από μελέτες που αφορούν πρωτεΐνες της *E. coli* (Rapp et al., 2004) και του *S. cerevisiae* (Kim, Melen, & von Heijne, 2003).

Από τις ήδη διαθέσιμες προγνωστικές μεθόδους, το TMHMM και το HMMTOP (Tusnady & Simon, 2001), προσφέρουν επιλογή στο χρήστη έτσι ώστε να ενσωματώσει στην πρόγνωσή του πειραματικά προσδιορισμένη πληροφορία για την τοπολογία, όπως επίσης τέτοια επιλογή, προσφέρεται και από την συνδυασμένη πρόγνωση διαμεμβρανικών ελίκων και πεπτιδίων οδηγητών από την μέθοδο Phobius (Kall, Krogh, & Sonnhammer, 2004). Η πρώτη όμως προσπάθεια να αναλυθούν αυτοί οι αλγόριθμοι, και να ενσωματωθούν με γενική μορφή σε κάθε αλγόριθμο αποκωδικοποίησης, έγινε από τους (Bagos, Liakopoulos, & Hamodrakas, 2006) και η εφαρμογή τους στον αλγόριθμο HMMTM.



Εικόνα 8.13: *Διαγραμματική απεικόνιση του πίνακα Forward, όταν γίνεται η ενσωμάτωση της εκ των προτέρων πληροφορίας. Έχουμε ένα (υποθετικό) μοντέλο από 12 καταστάσεις, και μια αλληλουχία* ***x*** *από 8 κατάλοιπα. Στον υπολογισμό της πιθανοφάνειας της αλληλουχίας* ***x,*** *ενσωματώνεται η πληροφορία ότι τα κατάλοιπα 3,4 είναι διαμεμβρανικά, ότι το κατάλοιπο 1 βρίσκεται στην εξωκυττάρια πλευρά και ότι το κατάλοιπο 8 βρίσκεται στο κυτταρόπλασμα.*

Οι τροποποιήσεις αυτές, είναι εντελώς ανάλογες με τις τροποποιήσεις που επιτρέπουν την εκπαίδευση με σημασμένες αλληλουχίες, με τη διαφορά ότι εδώ χρησιμοποιούνται στο πλαίσιο της αποκωδικοποίησης. Οι συγγραφείς έδειξαν, ότι οι πιθανοφάνειες που προκύπτουν με αυτό τον τρόπο, μπορούν να εκφραστούν σαν εκ των υστέρων πιθανότητες των πειραματικών πληροφοριών δεδομένης της αλληλουχίας και του μοντέλου, διατηρώντας έτσι την πιθανοθεωρητική ερμηνεία των αποτελεσμάτων. Παρόμοιας φύσεως τροποποιήσεις εισάγονται σε όλους τους αλγόριθμους αποκωδικοποίησης τους οποίους αναφέραμε ήδη, και με αυτόν τον τρόπο, είμαστε σε θέση να πραγματοποιήσουμε δεσμευμένη πρόγνωση για οποιασδήποτε μορφής πειραματική πληροφορία, και με οποιαδήποτε μέθοδο αποκωδικοποίησης.

Κατ' αρχάς ορίζουμε την έννοια της *Πληροφορίας (Information)* **ω,** η οποία αποτελείται από $1 \le r \le L$, αμοιβαίως αποκλειόμενα, μη-μηδενικού μήκους τμήματα στην αλληλουχία, τα οποία συμβολίζονται με ω_1, ω_2, ...ω_r, και για τα οποία γνωρίζουμε την ακριβή πειραματικώς προσδιορισμένη τοπολογία και κατά συνέπεια τη σήμανση. Όμοια με την περίπτωση της εκπαίδευσης με σημασμένες αλληλουχίες, μπορούμε να ορίσουμε μια δίτιμη συνάρτηση που να δείχνει τη συμφωνία της σήμανσης με την κατάσταση:

$$d_k(i) = \begin{cases} 0, \text{ if } \lambda_k(\omega_i) \neq 0 \text{ and } i \in \omega^r \\ 1, \text{ otherwise} \end{cases}$$

Οι απλές τροποποιήσεις οι οποίες προτάθηκαν για την περίπτωση του αλγόριθμου forward, συνίστανται στο να θέσουμε τη forward μεταβλητή f ίση με το μηδέν για κάθε θέση i και κατάσταση k η οποία δε συμφωνεί με την πληροφορία (Εικόνα 8.13). Αυτό είναι ακριβώς όμοιο με τη διαδικασία εκπαίδευσης με σημασμένες αλληλουχίες, όπου επιτρέπουμε μόνο τα μονοπάτια Π_y τα οποία είναι σε συμφωνία με τη σήμανση **y**, να συνεισφέρουν στην ολική πιθανοφάνεια. Στη συγκεκριμένη περίπτωση, επιτρέπουμε απλά, τη συνεισφορά μόνο των μονοπατιών Π_ω τα οποία είναι σε συμφωνία με την εκ των προτέρων πληροφορία **ω**. Εκτός από τους αλγόριθμους forward και backward που υπολογίζουν την πιθανοφάνεια, με τον ίδιο ακριβώς τρόπο τροποποιούνται και όλοι οι αλγόριθμοι αποκωδικοποίησης που είδαμε σε προηγούμενες ενότητες. Με τους αλγόριθμους αυτούς, μπορούμε να κάνουμε δεσμευμένες προγνώσεις ενσωματώνοντας όλη τη διαθέσιμη εκ των προτέρων πληροφορία για μια αλληλουχία. Τέτοια παραδείγματα, που αφορούν τις διαμεμβρανικές πρωτεΐνες, θα δούμε σε επόμενο κεφάλαιο. Λεπτομέρειες για τους τροποποιημένους αλγόριθμους, μπορούν να βρεθούν στην αντίστοιχη δημοσίευση (Bagos, et al., 2006).

8.3.6. Λεπτομέρειες της αλγοριθμικής υλοποίησης

Ένα σημαντικό πρόβλημα στην υλοποίηση των αλγορίθμων, προκύπτει καθώς οι διαδοχικοί πολλαπλασιασμοί μικρών πιθανοτήτων οδηγούν με μαθηματική ακρίβεια σε τελικό μηδενισμό των πιθανοτήτων, λόγω της μαθηματικής ακρίβειας των υπολογισμών (underflow error). Η χρησιμοποίηση του λογαρίθμου της πιθανοφάνειας λύνει εν μέρει αυτό το πρόβλημα, καθώς τα γινόμενα των πιθανοτήτων εύκολα μετατρέπονται σε αθροίσματα λογαριθμικών πιθανοτήτων.

Παρ' όλα αυτά, χρειάζεται μια επιπλέον τροποποίηση στην περίπτωση κατά την οποία πρέπει να υπολογιστεί ο λογάριθμος ενός αθροίσματος από τους λογαρίθμους των προσθετέων. Έτσι, λειτουργούμε ως εξής:

$$\log(a+b) = \log\left(a\left(1+{}^{a}\!/_{b}\right)\right) = \log(a) + \log\left(1+{}^{a}\!/_{b}\right)$$
$$= \log(a) + \log\left(1+\exp\left(\log(a)-\log(b)\right)\right) \quad (8.76)$$

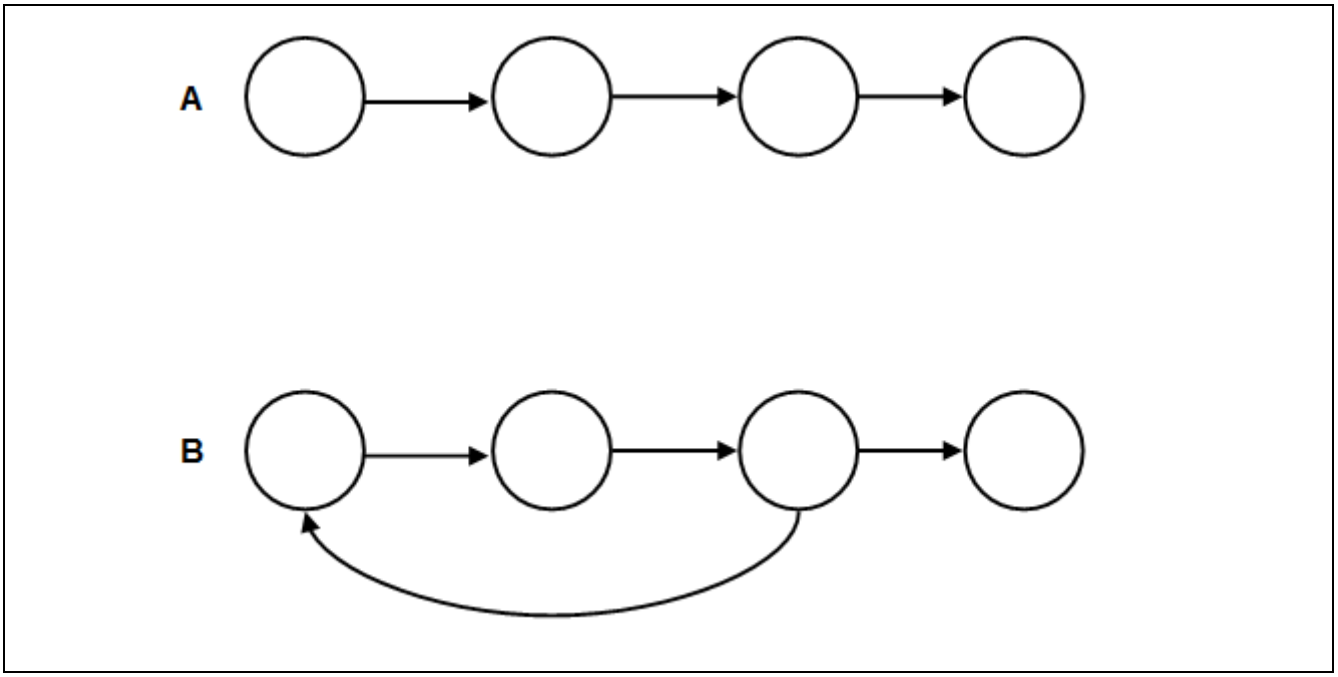
Όταν η διαφορά $|\log(a)-\log(b)|$ είναι μικρότερη από –37, το οποίο είναι το όριο για τους αριθμούς διπλής ακρίβειας, τότε ο δεύτερος προσθετέος της παραπάνω σχέσης θα είναι περίπου ίσος με 0, αλλιώς χρησιμοποιείται αυτούσια η σχέση (8.76).

8.4. Σχεδιασμός της δομής των μοντέλων

Ίσως το πιο ενδιαφέρον αλλά και πιο επίπονο στάδιο στην κατασκευή ενός προγνωστικού αλγορίθμου βασισμένο σε HMM, είναι η διαδικασία σχεδιασμού του κατάλληλου μοντέλου. Παρ΄ όλο που διάφορες μέθοδοι έχουν προταθεί για την εύρεση της Βέλτιστης δομής ενός μοντέλου είτε με γενικότερες τεχνικές Μέγιστης Πιθανοφάνειας (Ostendorf & Singer, 1997; Vasko, El-Jaroudi, & Boston, 1996), είτε με χρήση Γενετικών αλγορίθμων (Won, Prugel-Bennett, & Krogh, 2004; Yada, Ishikawa, H., & Asai, 1994), γενικώς, σε πολύπλοκα μοντέλα αυτές δεν αποδίδουν τόσο καλά και το καλύτερο μοντέλο προκύπτει πάντα από ανθρώπινο χέρι. Η διαδικασία, απαιτεί άριστη γνώση των αλγορίθμων που χρησιμοποιούνται όσο και βαθιά κατανόηση του βιολογικού προβλήματος το οποίο καλούμαστε να αντιμετωπίσουμε. Η μεγάλη δύναμη του HMM είναι ότι μπορεί να μοντελοποιήσει πολλά βιολογικά προβλήματα, τουλάχιστον όταν αυτά αφορούν αλληλουχίες στις οποίες υπάρχει «διαμερισματοποίηση», δηλαδή περιοχές με ξεκάθαρες διαφορές στη σύσταση των αμινοξέων. Όταν πηγαίνουμε στα πιο σύνθετα CHMM, όταν δηλαδή υπάρχουν σημάνσεις, κάθε σήμανση αντιστοιχεί σε ένα επιπλέον υπο-μοντέλο.

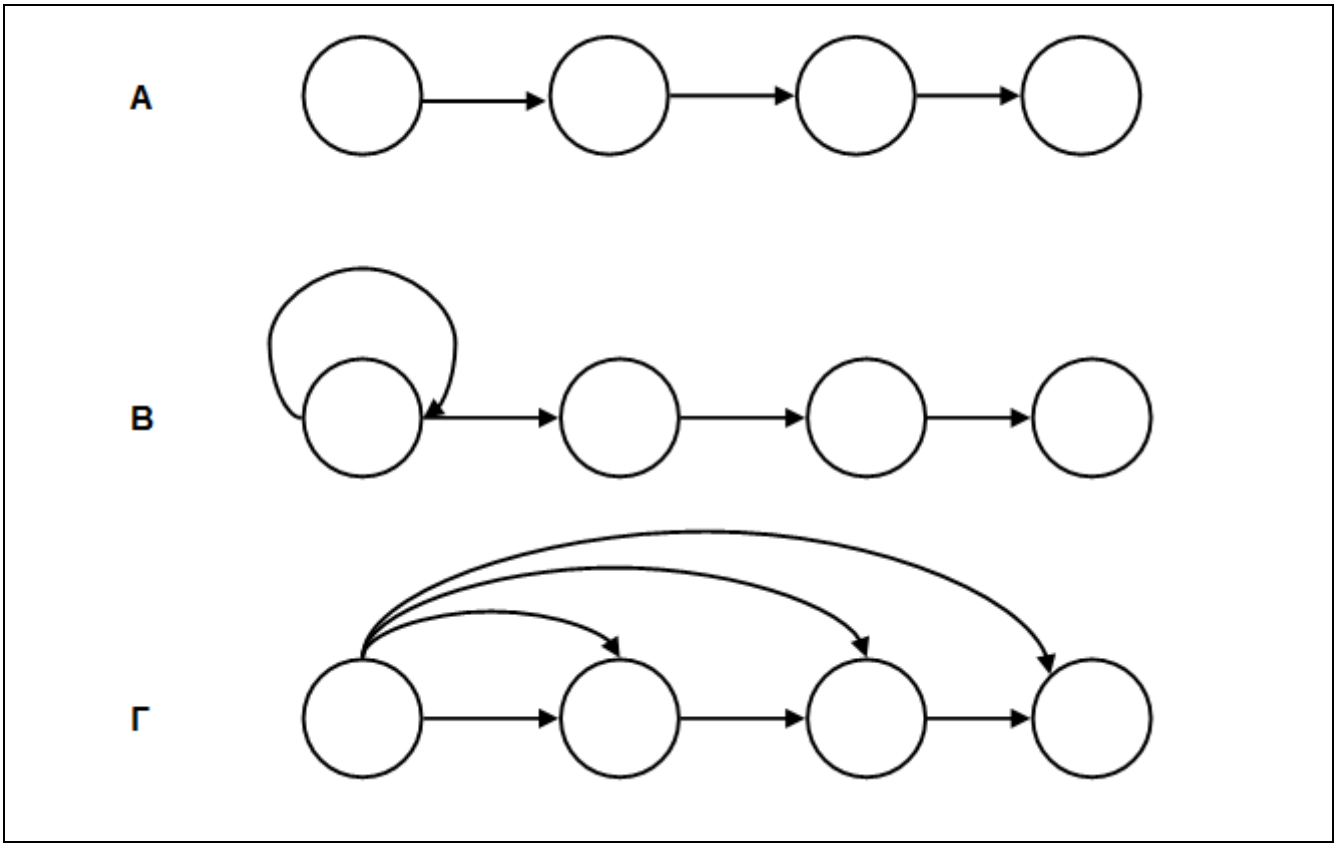
Γενικά, για να μοντελοποιήσουμε σύνθετες περιοχές (πχ δευτεροταγή δομή, διαμεμβρανικά τμήματα, εσώνια-εξώνια κ.ο.κ.) πρέπει να λάβουμε υπόψη μας δύο πράγματα: τις πιθανότητες εμφάνισης συμβόλων και τις μεταβάσεις μεταξύ των καταστάσεων. Πολλές φορές, ειδικά όταν έχουμε περιοχές με μεγάλο μήκος, χρειάζεται να χρησιμοποιήσουμε «παρόμοιες» καταστάσεις. Δηλαδή, καταστάσεις που είναι μεν διαφορετικές αλλά περιμένουμε να έχουν τις ίδιες ιδιότητες. Για παράδειγμα, στα διαμεμβρανικά τμήματα πρέπει να έχουμε μεγάλο αριθμό καταστάσεων για να μοντελοποιήσουμε τα διαμεμβρανικά τμήματα τα οποία έχουν

μήκος 15 έως 35 αμινοξικά κατάλοιπα. Σε αυτές τις περιπτώσεις, μια καλή πρακτική είναι να κάνουμε το λεγόμενο «parameter tying» ή αλλιώς «sharing». Πρακτικά, αυτό σημαίνει ότι θα έχουμε περισσότερες από μία καταστάσεις οι οποίες όμως θα έχουν τις ίδιες πιθανότητες εμφάνισης συμβόλων. Αυτό επιτυγχάνεται με το να αθροίζονται οι αναμενόμενες τιμές *E* στους αλγόριθμους forward και backward. Το βασικό πλεονέκτημα μια τέτοιας στρατηγικής, είναι ότι ελαττώνονται κατά πολύ οι παράμετροι του μοντέλου (αν ενώσουμε 10 καταστάσεις, θα έχουμε τελικά μόνο 20 πιθανότητες εμφάνισης συμβόλων αντί για 10*20).



Εικόνα 8.14: *Α. Τυπική εικόνα ενός γραμμικού μοντέλου (left to right). Το μοντέλο αν αφήσει μια δεδομένη κατάσταση, δεν επιστρέφει ποτέ. Β. Ένα τυπικό παράδειγμα κυκλικού μοντέλου. Το μοντέλο μπορεί να επιστρέψει (απεριόριστες φορές) σε μια δεδομένη κατάσταση.*

Το άλλο μεγάλο πρόβλημα, είναι τι είδους πιθανότητες μετάβασης θα επιτρέπουμε. Όπως είδαμε, αν και υπάρχουν προτάσεις για αυτόματη επιλογή του βέλτιστου μοντέλου, συνήθως στις περισσότερες εφαρμογές χρειάζεται ανθρώπινη παρέμβαση. Γενικά, εδώ χρειάζεται και κάποια φαντασία και γνώση κάποιων βασικών κανόνων. Το βασικό που χρειάζεται να λάβουμε υπόψη μας, είναι το μήκος των περιοχών και το πόσο περιοριστικό θέλουμε να είναι το μοντέλο. Ένα απλό μοντέλο με λίγες καταστάσεις, μπορεί να ταιριάζει σε κάθε είδους αλληλουχία που μπορεί να συναντήσει, αλλά μάλλον δεν θα έχει μεγάλη προβλεπτική αξία. Από την άλλη, ένα μοντέλο με πολλούς περιορισμούς, ενδέχεται να αποτύχει σε κάποιες αλληλουχίες που δεν ταιριάζουν σε αυτό. Συνήθως υπάρχουν 3 γενικοί τρόποι για να μοντελοποιηθεί μια περιοχή από *k* καταστάσεις (Εικόνα 8.14). Η πιο απλή περίπτωση είναι όταν η αλληλουχία είναι τελείως γραμμική και η μία κατάσταση ακολουθεί υποχρεωτικά την άλλη. Με αυτόν τον τρόπο, ο οποίος οδηγεί σε μοντέλα ανάλογα με τα profile τα οποία μελετήσαμε σε προηγούμενο κεφάλαιο, αναγκάζουμε το μοντέλο να περάσει ακριβώς μία φορά (ούτε περισσότερες, ούτε λιγότερες) από τις καταστάσεις της περιοχής αυτής (από το υπο-μοντέλο αυτό). Μια άλλη περίπτωση, συναντάμε όταν η περιοχή που θέλουμε να μοντελοποιήσουμε έχει ένα ελάχιστο μήκος, αλλά δεν είναι εύκολο να υπολογίσουμε κάποιο μέγιστο. Σε αυτή την περίπτωση, εισάγουμε μια κατάσταση η οποία έχει μετάβαση προς τον εαυτό της, με συνέπεια να μπορεί να επαναληφθεί επ' άπειρον. Η κατανομή που μπορεί να δώσει μια τέτοια τοπολογία, είναι η γεωμετρική κατανομή. Φυσικά, με κατάλληλη αλλαγή των πιθανοτήτων μετάβασης μπορεί το αναμενόμενο μήκος μιας τέτοιας αλληλουχίας καταστάσεων που θα παράξει αυτό το μοντέλο, να αλλάξει. Τέλος, σε περιπτώσεις που η περιοχή που θέλουμε να μοντελοποιήσουμε έχει ξεκάθαρα ελάχιστα και μέγιστα όρια, μπορούμε να επιλέξουμε μια τοπολογία κατά την οποία από μια δεδομένη κατάσταση επιτρέπονται μεταβάσεις προς όλες τις επόμενές της. Τέτοια περίπτωση θα δούμε ότι προκύπτει στην πρόγνωση των διαμεμβρανικών τμημάτων.



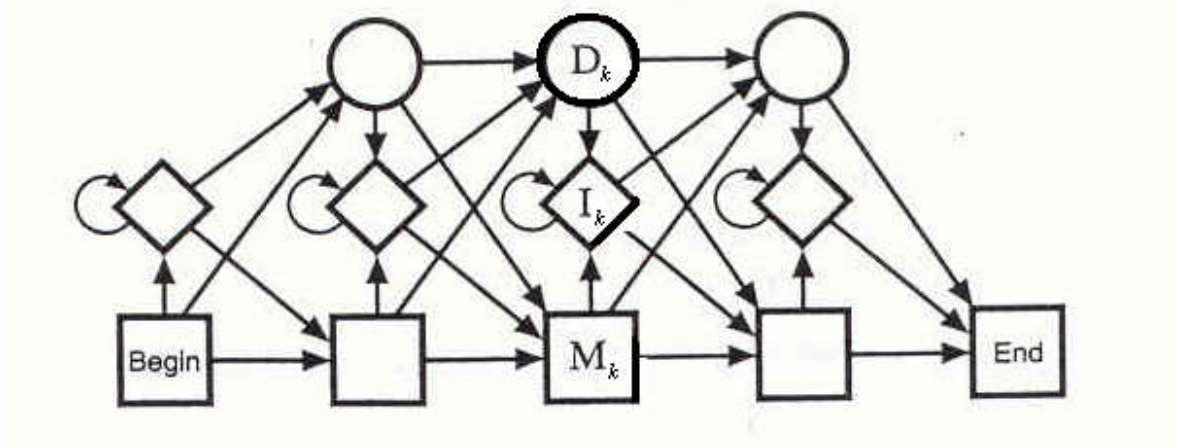
Εικόνα 8.15: *Α. Ένα μοντέλο το οποίο έχει πάντα καταστάσεις οι οποίες διαδέχονται πάντα η μία την άλλη. Αν το μοντέλο φτάσει στην κατάσταση 1, θα περάσει αναγκαστικά και από τις 2,3 και 4. Β. Ένα μοντέλο με 4 καταστάσεις, από τις οποίες όμως η μία επαναλαμβάνεται με μία πιθανότητα. Αυτό το μοντέλο μπορεί να μοντελοποιήσει περιοχές με μήκος πάνω από 4 κατάλοιπα (δεν υπάρχει όμως ανώτατο όριο). Γ. Ένα μοντέλο με ενδιάμεσες μεταβάσεις. Το μοντέλο αυτό μπορεί να περιγράψει περιοχές με μήκος από 1 έως 4 κατάλοιπα.*

Ένα άλλο θέμα που συχνά προκύπτει στον καθορισμό της δομής του μοντέλου είναι ο καθορισμός κάποιων «σιωπηλών καταστάσεων» (silent states). Σιωπηλές καταστάσεις ονομάζονται οι καταστάσεις που δεν παράγουν κάποιο σύμβολο, όπως για παράδειγμα οι καταστάσεις που συνδέονται με την έναρξη και τον τερματισμό του μοντέλου. Μια πιθανή χρησιμότητα τέτοιων καταστάσεων είναι όταν θέλουμε να επιτρέπουμε σε κάθε κατάσταση να συνδέεται με κάποιες από τις επόμενές της. Αν αυτό γίνει χωρίς τη χρήση silent states τότε αυξάνεται εκθετικά ο αριθμός των παραμέτρων (πιθανότητες μεταβάσεως) του μοντέλου, ενώ αν χρησιμοποιηθούν, το πιθανό μειονέκτημα είναι η αύξηση της πολυπλοκότητας των αλγορίθμων οι οποίοι χρειάζονται τροποποίηση.

8.5. Profile Hidden Markov Models

Μια πολύ ειδική κατηγορία Hidden Markov Models, είναι τα λεγόμενα προφίλ (profile) Hidden Markov Models (Eddy, 1998), τα οποία επέκτειναν την έννοια του προφίλ αλληλουχιών (Gribskov, Luthy, & Eisenberg, 1990; Gribskov, McLachlan, & Eisenberg, 1987) τα οποία συναντήσαμε σε προηγούμενα κεφάλαια και την επένδυσαν με πιθανοθεωρητικό χαρακτήρα. Ένα Profile Hidden Markov Model (pHMM), είναι στην ουσία ένα HMM το οποίο περιγράφει με ακρίβεια μια πολλαπλή στοίχιση αλληλουχιών. Η βασική διαφορά του profile από τα κλασικά μοντέλα που αναφέραμε παραπάνω, είναι ότι σε αυτό κάθε κατάσταση περιγράφει μια συγκεκριμένη θέση (στήλη) στην πολλαπλή στοίχιση. Κατά συνέπεια, το μοντέλο έχει ειδικές παραμέτρους ανά θέση (position specific) και ως εκ τούτου η κατεύθυνση των μεταβάσεων θα πρέπει να είναι πάντα μονόδρομη. Γι' αυτόν ακριβώς το λόγο, τα μοντέλα αυτά ονομάζονται μοντέλα *left-to-right* σε αντιδιαστολή με τα κυκλικά που είδαμε παραπάνω τα οποία επιτρέπουν στο μοντέλο να επισκεφθεί μια κατάσταση περισσότερες από μια φορές. Από την πλευρά των κλασικών προφίλ αλληλουχιών που είδαμε σε προηγούμενα κεφάλαια, το προφίλ HMM είναι μια γενίκευση, στην οποία δεν μοντελοποιούνται μόνο οι πιθανότητες εμφάνισης συμβόλων σε κάθε θέση της πολλαπλής στοίχισης, αλλά μοντελοποιούνται με

πιθανοθεωρητικό τρόπο και οι πιθανότητες εισαγωγής κενού και απαλοιφής (insert/deletion). Αυτό είναι πολύ σημαντικό, καθώς το πρόβλημα της επιλογής της ποινής για τα κενά, αποτελούσε μέχρι τώρα ένα πρόβλημα στο οποίο η απάντηση δινόταν με ξεκάθαρα εμπειρικό τρόπο, χωρίς καμία θεωρητική τεκμηρίωση.



Εικόνα 8.16: *Σχηματική αναπαράσταση ενός τυπικού profile Hidden Markov Model*

Μια άλλη σημαντική διαφορά των μοντέλων αυτών, είναι η ύπαρξη ειδικών καταστάσεων οι οποίες δεν εκπέμπουν κανένα σύμβολο, οι οποίες ονομάζονται σιωπηρές καταστάσεις (silent states). Οι καταστάσεις αυτές, χρησιμοποιούνται για να πραγματοποιήσουν μεταβάσεις από ένα κατάλοιπο σε κάποιο άλλο αρκετά κατάλοιπα μακριά του χωρίς να υπάρχει ανάγκη για πολλαπλές μεταβάσεις από την κατάσταση αυτή. Η ύπαρξη τέτοιων καταστάσεων, είναι αναγκαία για να μειώσει τον αριθμό των παραμέτρων του μοντέλου οι οποίες μεγαλώνουν υπερβολικά καθώς κάθε στήλη στην πολλαπλή στοίχιση αντιστοιχεί πλέον σε μια κατάσταση. Όπως είδαμε στην προηγούμενη παράγραφο, μια τοπολογία στην οποία μια κατάσταση επιτρέπεται να κάνει μετάβαση σε *k* επόμενες, θα μπορούσε να χρησιμοποιηθεί για να μοντελοποιήσει τα κενά σε μια πολλαπλή στοίχιση. Παρ' όλα αυτά, μια τέτοια στρατηγική θα αύξανε πάρα πολύ τον αριθμό των παραμέτρων του μοντέλου και θα δημιουργούσε δυσκολία στην εκπαίδευση. Ένα τυπικό pHMM, φαίνεται στην Εικόνα 8.16.

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

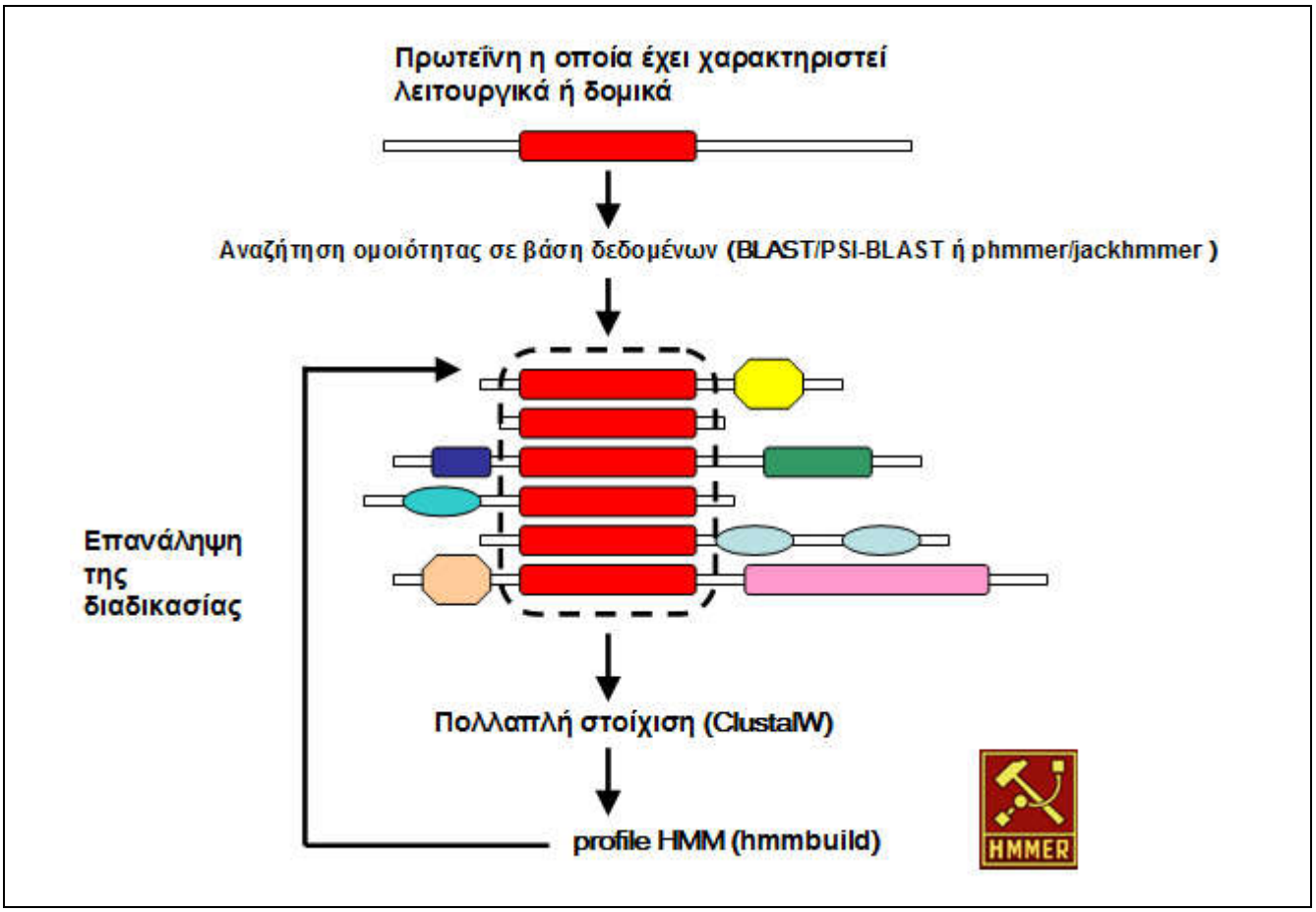
- Καταστάσεις Ταύτισης (Match states) M_k τετράγωνα
- Καταστάσεις Εισαγωγής (Insertion states) I_k ρόμβοι
- Καταστάσεις Απαλοιφής (Deletion states) D_k κύκλοι

και συνδέονται με τις αντίστοιχες πιθανότητες μεταβάσεως, που συμβολίζονται με βέλη. Αντίστοιχα ορίζονται οι πιθανότητες εμφάνισης συμβόλων 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 ταύτισης και εισαγωγής, είναι κανονικές καταστάσεις οι οποίες συνδέονται μέσω των πιθανοτήτων γέννησης με την εμφάνιση συμβόλων. Η διαφορά τους είναι η εξής: οι μεν καταστάσεις ταύτισης 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 αλληλουχίες, εμφανίζονται ως κενά τα οποία μοντελοποιούνται μέσω των σιωπηρών καταστάσεων απαλοιφής.

Στο αλγοριθμικό κομμάτι, μετατροπές χρειάζονται για την ενσωμάτωση των σιωπηρών καταστάσεων, καθώς και για να μην καταμετρώνται μεταβάσεις σε καταστάσεις που προηγούνται, σε όλους τους παραπάνω αλγορίθμους. Έτσι είναι δυνατόν με χρήση όλων των βασικών αλγόριθμων δυναμικού προγραμματισμού (Viterbi, backward, forward, Baum-Welch κλπ) που αναφέραμε πριν, να υπολογίσουμε τις παραμέτρους του HMM που περιγράφει μια πολλαπλή στοίχιση.

8.6. Εφαρμογές των profile HMM

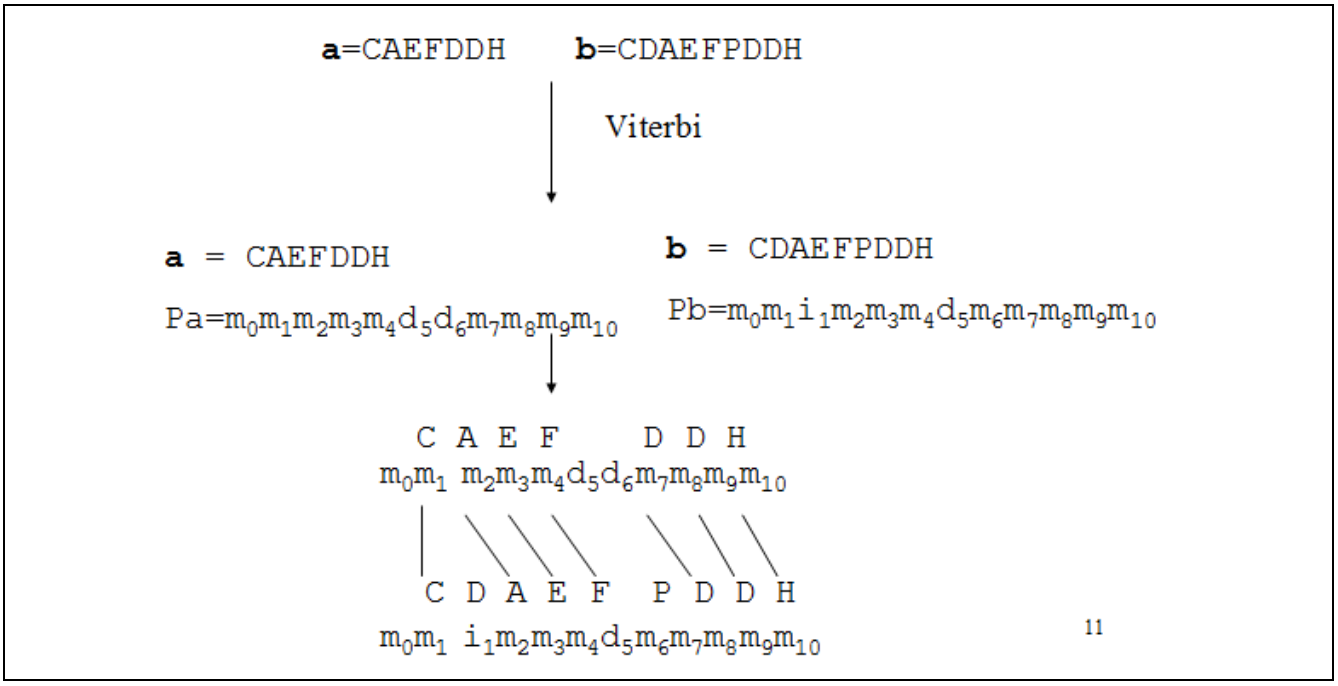
Με την εισαγωγή των διαφορετικών καταστάσεων ταύτισης και εισαγωγής, γίνεται μια σημαντική τομή με τις κλασικές μεθόδους στοίχισης, οι οποίες καθώς δεν προϋποθέτουν ένα μοντέλο δεν διαχωρίζουν τις πληροφοριακές θέσεις στη στοίχιση από τις απλές τυχαίες εισαγωγές. Επιπλέον, για πρώτη φορά οι ποινές για 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.



Εικόνα 8.17: *Σχηματική αναπαράσταση της διαδικασίας χαρακτηρισμού μιας νέας πρωτεϊνικής οικογένειας. Για λεπτομέρειες, δείτε το κείμενο.*

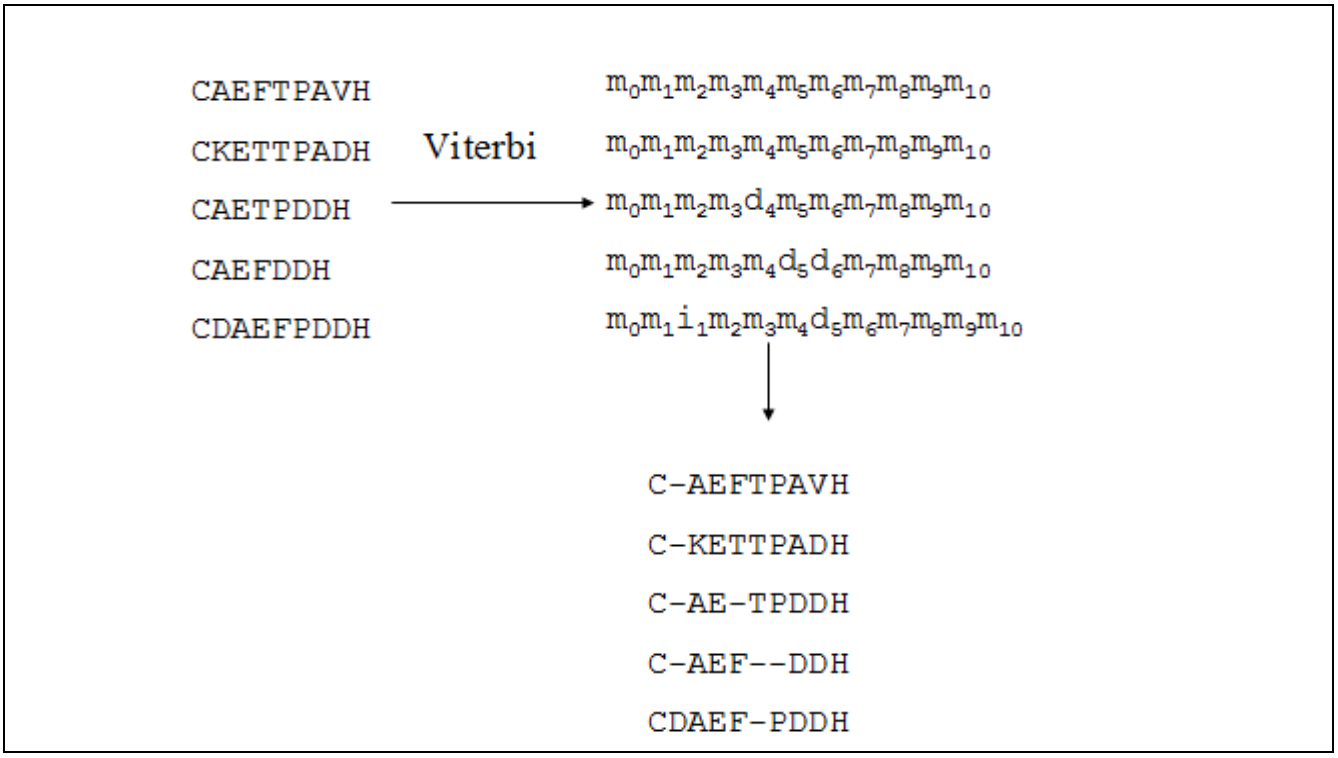
Η χρησιμότητα των pHMM, ξεκινάει από δημιουργία πολλαπλών στοιχίσεων, οι οποίες πολλές φορές είναι εφάμιλλες αντίστοιχων δομικών πολλαπλών στοιχίσεων (Eddy, 1995) και φτάνει στη δημιουργία μοντέλων μέσω των οποίων μπορεί να γίνει ευαίσθητη αναζήτηση απομακρυσμένων ομολογιών (Krogh, Brown, Mian, Sjolander, & Haussler, 1994). Ιδιαίτερα με την τελευταία μέθοδο και τη δημιουργία όλο και περισσότερων μοντέλων για την ταξινόμηση πρωτεϊνικών οικογενειών, έχουν κατασκευαστεί ειδικές βάσεις δεδομένων όπως η PFAM (Bateman, et al., 2004), μέσω της οποίας ταξινομείται μεγάλο μέρος των άγνωστης λειτουργίας πρωτεϊνών που προσδιορίζονται καθημερινά, π.χ. με την αποκωδικοποίηση των γονιδιωμάτων. Παρόλο που υπάρχουν και άλλες αντίστοιχες βάσεις, όπως για παράδειγμα η TIGRFAM, η PFAM θεωρείται σήμερα η κορυφαία βάση δεδομένων για πρωτεϊνικές οικογένειες. Αυτό οφείλεται, τόσο στο σχεδιασμό της, βάσει του οποίου μία πρωτεϊνική περιοχή μπορεί να ανήκει μόνο σε μία οικογένεια, όσο και στο λογισμικό το οποίο χρησιμοποιεί, το οποίο είναι ιδιαίτερα αποδοτικό και θα περιγραφεί στην επόμενη ενότητα. Τέλος, πρέπει να σημειωθεί ότι μια σειρά θεωρητικές και αλγοριθμικές βελτιώσεις που έχουν προκύψει τα τελευταία χρόνια, έχουν επιτρέψει την υλοποίηση αλγορίθμων που επιτελούν ακόμα και κλασικές αναζητήσεις ομοιότητας σε μια βάση δεδομένων, με ένα profile HMM (Eddy, 2011). Γενικά μια διαδικασία χαρακτηρισμού μιας πρωτεϊνικής οικογένειας έχει τα εξής βήματα:

- Στην αρχή, ξεκινάμε με μια αλληλουχία για την οποία υπάρχουν πειραματικές ενδείξεις για τη λειτουργία ή τη δομή της
- Γίνεται αναζήτηση σε βάσεις δεδομένων (BLAST, PSI-BLAST ή πλέον, με το HMMER)
- Συλλογή ομολόγων, επιλογή και ξεσκαρτάρισμα
- Γίνεται μια πολλαπλή στοίχιση (μπορεί και τροποποίηση αυτής με το χέρι)
- Ανάλογα με την περίπτωση, πραγματοποιούνται προγνώσεις (δευτεροταγούς δομής, διαμεμβρανικών τμημάτων ή οποιουδήποτε άλλου χρήσιμου χαρακτηριστικού)
- Γίνεται κατασκευή HMM και αξιολόγηση του (HMMER)
- Αναζήτηση εκ νέου σε βάσεις δεδομένων, μέχρι να μην προκύπτουν νέα μέλη της οικογένειας.



Εικόνα 8.18: *Στοίχιση δύο αλληλουχιών με ένα profile HMM.*

Ειδικά η περίπτωση της πολλαπλής στοίχισης αλληλουχιών με ένα profile HMM, είναι μια σημαντική διαδικασία, καθώς όπως είδαμε στο αντίστοιχο κεφάλαιο, αλγόριθμοι δυναμικού προγραμματισμού είναι δύσκολο να υλοποιηθούν, και στην πράξη όλα τα αντίστοιχα προγράμματα ακόμα και τα πιο πετυχημένα, βασίζονται σε ευριστικούς αλγόριθμους. Κατά συνέπεια, το profile HMM είναι μια ιδιαίτερα αξιόπιστη εναλλακτική, η οποία μάλιστα έχει αποδειχθεί ότι λειτουργεί εξαιρετικά καλά (Eddy, 1995). Η βασική αρχή της μεθόδου, φαίνεται στις Εικόνες 8.18 και 8.19. Το βασικό χαρακτηριστικό της μεθόδου, είναι ότι κάθε αλληλουχία στοιχίζεται με το μοντέλο, ανεξάρτητα από τις υπόλοιπες, και κατά συνέπεια, υπάρχει μόνο γραμμική εξάρτηση από τον αριθμό των αλληλουχιών. Για κάθε αλληλουχία, ο αλγόριθμος Viterbi βρίσκει το μονοπάτι των καταστάσεων από τις οποίες έχει περάσει. Αυτές οι καταστάσεις, όπως είδαμε είναι για κάθε θέση *i* στην αλληλουχία, τριών ειδών (*m*, *d*, *i*), ταύτιση, απαλοιφή και εισαγωγή. Επειδή όμως οι καταστάσεις αντιστοιχούν στη θέση *i* με αυτόν τον τρόπο μπορούμε άμεσα να στοιχίσουμε τις αλληλουχίες, αντιστοιχίζοντας απλά τις ταυτίσεις σε κάθε θέση (m_i).



Εικόνα 8.19: *Πολλαπλή στοίχιση αλληλουχιών με ένα profile HMM.*

8.7. Το πακέτο λογισμικού HMMER

Το πιο γνωστό πακέτο λογισμικού για κατασκευή Profile Hidden Markov Models, είναι το **HMMER** (Eddy, 2000). Το πακέτο αυτό, είναι μια συλλογή προγραμμάτων, διανεμόμενων ελεύθερα με την άδεια 'ανοικτού κώδικα' GPL (GNU Public License), η οποία επιτρέπει ελεύθερη πρόσβαση στον πηγαίο κώδικα, και το οποίο έχει αποδειχθεί το καλύτερο ίσως πακέτο τέτοιου είδους με πολλές εφαρμογές σε μεγάλο εύρος βιολογικών δεδομένων.

Το HMMER, στην έκδοση 3, περιέχει μεταξύ άλλων τα παρακάτω προγράμματα:

- **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 αλληλουχιών οι οποίες προέρχονται από ένα HMM, στοιχίζονται σε μια πολλαπλή στοίχιση. Η πολλαπλή στοίχιση, επιτυγχάνεται μέσω διαδοχικών στοιχίσεων των αλληλουχιών με το μοντέλο.
- **hmmsearch:** Πρόγραμμα το οποίο, πραγματοποιεί αναζητήσεις ενός μοντέλου HMM έναντι μιας βάσης αλληλουχιών πρωτεϊνών.
- **phmmer:** Πρόγραμμα το οποίο πραγματοποιεί αναζήτηση μια πρωτεϊνικής αλληλουχίας έναντι μιας βάσης δεδομένων πρωτεϊνών (ανάλογο με το BLASTP)
- **jackhmmer:** Πρόγραμμα το οποίο πραγματοποιεί επαναληπτικές αναζητήσεις μια πρωτεϊνικής αλληλουχίας έναντι μιας βάσης δεδομένων πρωτεϊνών (ανάλογο με το PSI-BLAST)
- **hmmscan:** Πρόγραμμα με το οποίο πραγματοποιούνται αναζητήσεις μιας η περισσότερων αλληλουχιών έναντι μιας βάσης δεδομένων από μοντέλα HMM. Πρέπει να τονιστεί εδώ, ότι αν έχουμε μια αλληλουχία και ένα HMM, τα δυο παραπάνω προγράμματα επιστρέφουν ακριβώς το ίδιο αποτέλεσμα. Αν διαφέρουν, είτε οι αλληλουχίες είτε τα μοντέλα, τότε δίνουν άλλο αποτέλεσμα, λόγω του διαφορετικού τρόπου υπολογισμού της στατιστικής σημαντικότητας.

- **nhmmer:** Πρόγραμμα που πραγματοποιεί αναζήτηση μιας αλληλουχίας DNA, μιας στοίχισης ή ενός pHMM, έναντι μιας βάσης αλληλουχιών DNA. (ανάλογο με το BLASTN)
- **nhmmscan:** Πρόγραμμα που πραγματοποιεί αναζήτηση μιας αλληλουχίας 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) αλληλουχία η οποία θα μπορούσε να παραχθεί από το μοντέλο.
- **hmmpress:** Μετατρέπει μια βάση δεδομένων HMM σε δυαδικό κώδικα για το hmmscan.
- **hmmstat:** Δείχνει συνοπτικά στατιστικά για μια βάση δεδομένων HMM.

Τα παραπάνω προγράμματα, περιέχουν μια σειρά από βελτιστοποιήσεις με σκοπό την επιτάχυνση των διαδικασιών. Για παράδειγμα, υπάρχουν βελτιστοποιήσεις για την ταχύτητα κατά την εκτέλεση των αλγορίθμων με τον μη υπολογισμό των προηγούμενων καταστάσεων, βελτιστοποιήσεις στον υπολογισμό των κατανομών συχνοτήτων των αμινοξέων του μηδενικού (null) μοντέλου με την εισαγωγή μίξεων από εκ των προτέρων κατανομές, διαφορικό στάθμισμα των αλληλουχιών με διαφορετικό βαθμό ομοιότητας, και μια σειρά από βελτιστοποιήσεις στη δομή του μοντέλου. Το τελευταίο, είναι πολύ σημαντικό, καθώς με αυτή τη διαδικασία, ο χρήστης δεν ασχολείται καθόλου με τη δομή και το μέγεθος του μοντέλου. Με επαναλήψεις ενός βασικού μοτίβου το οποίο αποτελεί παραλλαγή του κλασικού μοντέλου, επιτυγχάνεται, αναλόγως και του μήκους της πολλαπλής στοίχισης, η τελική διαμόρφωση του μοντέλου. Το παραπάνω μοντέλο, διαφέρει από την τυπική έκδοση του Profile Hidden Markov Model, που είδαμε παραπάνω, στο ότι δεν επιτρέπει μεταβάσεις από μια κατάσταση εισαγωγής κενού (I) σε κατάσταση απαλοιφής (D) και το αντίστροφο. Το πρόγραμμα, είναι διαθέσιμο στην ηλεκτρονική διεύθυνση: http://hmmer.janelia.org/.

Ιστορικά, αξίζει να αναφερθεί ότι υπάρχουν μεγάλες διαφορές μεταξύ των εκδόσεων του HMMER. Οι εκδόσεις μέχρι τη 1.8 επέτρεπαν τη δημιουργία μοντέλου ακόμα και από μη στοιχισμένες αλληλουχίες. Από την έκδοση 2.0 και μετά, η ύπαρξη μιας πολλαπλής στοίχισης έγινε απαραίτητη, καθώς το λογισμικό εξειδικεύτηκε σε αναλύσεις πρωτεϊνών. Παρ' όλα αυτά, το λογισμικό ήταν εξαιρετικά πετυχημένο και χρησιμοποιήθηκε για χιλιάδες αναλύσεις. Σε μερικές μάλιστα περιπτώσεις, χρησιμοποιήθηκε και για αναλύσεις που φαινομενικά δεν προϋπέθεταν κάποια «ομοιότητα» στις υπό μελέτη αλληλουχίες (Zhang & Wood, 2003). Εκτός αυτού, υπήρξαν και βελτιώσεις του, από τρίτους επιστήμονες, έτσι ώστε να μπορεί να πραγματοποιεί «διαχωριστική εκπαίδευση» (discriminative training) (Srivastava, Desai, Nandi, & Lynn, 2007). Από την έκδοση 3.0 και μετά, το λογισμικό, βασισμένο σε μια σειρά αλγοριθμικές και θεωρητικές βελτιώσεις, έγινε πολύ καλύτερο (Eddy, 2011). Καταρχάς, έγινε πολύ πιο γρήγορο. Επίσης, δεν χρειάζεται πλέον να πραγματοποιούνται προσομοιώσεις για να υπολογιστούν οι παράμετροι της κατανομής του Gumbel, αλλά αυτές προκύπτουν θεωρητικά. Τέλος, βασισμένο εν μέρει και στα προηγούμενα, το λογισμικό μπορεί πλέον να επιτελέσει και αναζητήσεις μιας αλληλουχίας έναντι μιας βάσης δεδομένων αλληλουχιών, -όπως ακριβώς το BLAST και το PSI-BLAST-, τις οποίες πραγματοποιεί καλύτερα και μάλιστα σε συγκρίσιμο χρόνο. Με αυτόν τον τρόπο, φαίνεται πως αργά αλλά σταθερά οδηγούμαστε σε καθολική αποδοχή των πιθανοθεωρητικών μοντέλων, τα οποία θα αντικαταστήσουν τις ευριστικές τεχνικές ομοιότητας.

Βιβλιογραφία

Audic, S., & Claverie, J. M. (1998). Self-identification of protein-coding regions in microbial genomes. *Proc Natl Acad Sci U S A, 95*(17), 10026-10031.

Bagos, P. G., Liakopoulos, T. D., & Hamodrakas, S. J. (2004). Faster Gradient Descent Conditional Maximum Likelihood Training of Hidden Markov Models, Using Individual Learning Rate Adaptation. In G. Paliouras & Y. Sakakibara (Eds.), *Grammatical Inference: Algorithms and Applications* (Vol. 3264, pp. 40-52): Spinger Berlin/Heidelberg.

Bagos, P. G., Liakopoulos, T. D., & Hamodrakas, S. J. (2006). Algorithms for incorporating prior topological information in HMMs: application to transmembrane proteins. *BMC Bioinformatics, 7*, 189.

Baldi, P., & Chauvin, Y. (1994). Smooth On-Line Learning Algorithms for Hidden Markov Models. *Neural Comput, 6*(2), 305-316.

Barash, Y., Elidan, G., Friedman, N., & Kaplan, T. (2003). *Modeling dependencies in protein-DNA binding sites*. Paper presented at the Proceedings of the seventh annual international conference on Computational molecular biology. RECOMB '03., New York, NY, USA.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., . . . Eddy, S. R. (2004). The Pfam protein families database. *Nucleic Acids Res, 32*(Database issue), D138-141.

Baum, L. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities, 3*, 1-8.

Bejerano, G. (2004). Algorithms for variable length Markov chain modeling. *Bioinformatics, 20*(5), 788-789.

Bejerano, G., Seldin, Y., Margalit, H., & Tishby, N. (2001). Markovian domain fingerprinting: statistical segmentation of protein sequences. *Bioinformatics, 17*(10), 927-934.

Bejerano, G., & Yona, G. (2001). Variations on probabilistic suffix trees: statistical modeling and prediction of protein families. *Bioinformatics, 17*(1), 23-43.

Berchtold, A. (2001). Estimation in the Mixture Transition Distribution Model. *Journal of Time Series Analysis, 22*(4), 379-397.

Borodovsky, M., & McIninch, J. (1993). GeneMark: parallel gene recognition for both DNA strands. *Comput Chem, 17*(19), 123-133.

Borodovsky, M., & Peresetsky, A. (1994). Deriving non-homogeneous DNA Markov chain models by cluster analysis algorithm minimizing multiple alignment entropy. *Comput Chem, 18*(3), 259-267.

Browning, S. R. (2006). Multilocus association mapping using variable-length Markov chains. *Am J Hum Genet, 78*(6), 903-913.

Dalevi, D., Dubhashi, D., & Hermansson, M. (2006). Bayesian classifiers for detecting HGT using fixed and variable order markov models of genomic signatures. *Bioinformatics, 22*(5), 517-522.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J Royal Stat Soc B, 39*, 1-38.

Drew, D., Sjostrand, D., Nilsson, J., Urbig, T., Chin, C. N., de Gier, J. W., & von Heijne, G. (2002). Rapid topology mapping of Escherichia coli inner-membrane proteins by prediction and PhoA/GFP fusion analysis. *Proc Natl Acad Sci U S A, 99*(5), 2690-2695.

Durbin, R., Eddy, S. R., Krogh, A., & Mithison, G. (1998). *Biological sequence analysis, probabilistic models of proteins and nucleic acids.*: Cambridge University Press.

Eddy, S. R. (1995). Multiple alignment using hidden Markov models. *Proc Int Conf Intell Syst Mol Biol, 3*, 114-120.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics, 14*(9), 755-763.

Eddy, S. R. (2000). *HMMER: profile hidden Markov models for biological sequence analysis*. St Louis, MO: Washington University school of medicine.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Comput Biol, 7*(10), e1002195.

Ellrott, K., Yang, C., Sladek, F. M., & Jiang, T. (2002). Identifying transcription factor binding sites through Markov chain optimization. *Bioinformatics, 18 Suppl 2*, S100-S109.

Eronen, L., Geerts, F., & Toivonen, H. (2004). A Markov chain approach to reconstruction of long haplotypes. *Pac Symp Biocomput*, 104-115.

Fariselli, P., Finelli, M., Marchignoli, D., Martelli, P. L., Rossi, I., & Casadio, R. (2003). MaxSubSeq: an algorithm for segment-length optimization. The case study of the transmembrane spanning segments. *Bioinformatics, 19*(4), 500-505.

Fariselli, P., Martelli, P. L., & Casadio, R. (2005). A new decoding algorithm for hidden Markov models improves the prediction of the topology of all-beta membrane proteins. *BMC Bioinformatics, 6 Suppl 4*, S12.

Gribskov, M., Luthy, R., & Eisenberg, D. (1990). Profile analysis. *Methods Enzymol, 183*, 146-159.

Gribskov, M., McLachlan, A. D., & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A, 84*(13), 4355-4358.

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1994). A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry, 33*(10), 3038-3049.

Juang, B. H., & Rabiner, L. R. (1990). The Segmental K-Means Algorithm for Estimating Parameters of Hidden Markov Models. *IEEE Transactions on Acoustics, Speech and Signal Processing, 38*(9), 1639-1641.

Kall, L., Krogh, A., & Sonnhammer, E. L. (2004). A combined transmembrane topology and signal peptide prediction method. *J Mol Biol, 338*(5), 1027-1036.

Kall, L., Krogh, A., & Sonnhammer, E. L. (2005). An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics, 21 Suppl 1*, i251-i257.

Kim, H., Melen, K., & von Heijne, G. (2003). Topology models for 37 Saccharomyces cerevisiae membrane proteins based on C-terminal reporter fusions and predictions. *J Biol Chem, 278*(12), 10208-10213.

Krogh, A. (1994). Hidden Markov models for labelled sequences. *Proceedings of the12th IAPR International Conference on Pattern Recognition*, 140-144.

Krogh, A. (1997). Two methods for improving performance of an HMM and their application for gene finding. *Proc Int Conf Intell Syst Mol Biol, 5*, 179-186.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K., & Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol, 235*(5), 1501-1531.

Krogh, A., & Riis, S. K. (1999). Hidden neural networks. *Neural Comput, 11*(2), 541-563.

Lebre, S., & Bourguignon, P. Y. (2008). An EM algorithm for estimation in the mixture transition distribution model. *Journal of Statistical Computation and Simulation, 78*(8), 713-729.

MacQueen, B. (1967). *Some Methods for classification and Analysis of Multivariate Observations*. Paper presented at the Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability.

Markov, A. A. (1913). An example of statistical study on text of Eugeny Onegin illustrating the linking of events to a chain. *Izvestija Imp. Akad. nauk, 6*(3), 153-162.

Melen, K., Krogh, A., & von Heijne, G. (2003). Reliability measures for membrane protein topology prediction algorithms. *J Mol Biol, 327*(3), 735-744.

Ohler, U., Harbeck, S., Niemann, H., Noth, E., & Reese, M. G. (1999). Interpolated markov chains for eukaryotic promoter recognition. *Bioinformatics, 15*(5), 362-369.

Ostendorf, M., & Singer, H. (1997). HMM topology design using maximum likelihood successive state splitting. *Computer Speech & Language, 11*(1), 17-41.

Phillips, G. J., Arnold, J., & Ivarie, R. (1987). Mono- through hexanucleotide composition of the Escherichia coli genome: a Markov chain analysis. *Nucleic Acids Res, 15*(6), 2611-2626.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE, 77*(2), 257-286.

Raftery, A. E. (1985a). A model for high-order Markov chains. *Journal of the Royal Statistical Society. Series B (Methodological), 47*(3), 528-539.

Raftery, A. E. (1985b). A new model for discrete-valued time series: Autocorrelations and extensions. *Rassegna di Metodi Statistici ed Applicazioni, 3-4* 149-162.

Rapp, M., Drew, D., Daley, D. O., Nilsson, J., Carvalho, T., Melen, K., . . . Von Heijne, G. (2004). Experimentally based topology models for E. coli inner membrane proteins. *Protein Sci, 13*(4), 937-945.

Ron, D., Singer, Y., & Tishby, N. (1996). The power of amnesia: learning probabilistic automata with variable memory length. *Machine Learning, 25*, 117-149.

Salzberg, S. L., Delcher, A. L., Kasif, S., & White, O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Res, 26*(2), 544-548.

Salzberg, S. L., Pertea, M., Delcher, A. L., Gardner, M. J., & Tettelin, H. (1999). Interpolated Markov models for eukaryotic gene finding. *Genomics, 59*(1), 24-31.

Schwartz, R., & Chow, Y. L. (1990). The N-Best Algorithm: An Efficient and Exact Procedure for Finding the N Most Likely Sentence Hypotheses. *Proc IEEE Int Conf Acoust, Speech, Sig Proc, 1*, 81-84.

Srivastava, P. K., Desai, D. K., Nandi, S., & Lynn, A. M. (2007). HMM-ModE--improved classification using profile hidden Markov models by optimising the discrimination threshold and modifying emission probabilities with negative training sequences. *BMC Bioinformatics, 8*, 104.

Tusnady, G. E., & Simon, I. (2001). The HMMTOP transmembrane topology prediction server. *Bioinformatics, 17*(9), 849-850.

Vasko, R. C. J., El-Jaroudi, A., & Boston, J. R. (1996). *An algorithm to determine hidden Markov model topology.* Paper presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-96.

Won, K. J., Prugel-Bennett, A., & Krogh, A. (2004). Training HMM structure with genetic algorithm for biological sequence analysis. *Bioinformatics, 20*(18), 3613-3619.

Yada, T., Ishikawa, M., H., T., & Asai, K. (1994). DNA Sequence Analysis using Hidden Markov Model and Genetic Algorithm. *Genome Informatics, 5*, 178-179.

Yuan, Z. (1999). Prediction of protein subcellular locations using Markov chain models. *FEBS Lett, 451*(1), 23-26.

Zhang, Z., & Wood, W. I. (2003). A profile hidden Markov model for signal peptides generated by HMMER. *Bioinformatics, 19*(2), 307-308.

Ερωτήσεις

1) Στο μοντέλο με το «μεροληπτικό ζάρι» που αναφέρεται στην παράγραφο 8.3.1, δίνεται η ακολουθία συμβόλων **x**= `2145264366365616662321 45` και το μονοπάτι π= `----------++++++++++++++------`. Εκτιμήστε τις παραμέτρους του μοντέλου με βάση τις σχέσεις (8.34) και (8.35).

2) Σε ένα Hidden Markov Model (HMM) δίνονται ο πίνακας των πιθανοτήτων μετάβασης (*a*, transitions), και αυτός των πιθανοτήτων εμφάνισης των συμβόλων (*e*, emissions), αντίστοιχα:

$$a=\begin{bmatrix} 0.7 & 0.3 & 0 & x_1 \\ 0 & 0 & 0.8 & x_2 \\ 0 & 0 & 0.9 & 0.1 \\ x_3 & 0 & 0 & 0 \end{bmatrix},\ e=\begin{bmatrix} 0.2 & 0.1 & x_4 & 0.5 \\ 0.5 & x_5 & 0.1 & 0.1 \\ 0.1 & 0.2 & 0.1 & 0.6 \\ 0.4 & 0.4 & 0.1 & 0.1 \end{bmatrix}$$

Α) Από πόσες καταστάσεις αποτελείται αυτό το μοντέλο και πόσα είναι τα γράμματα του αλφαβήτου του; Εξηγήστε.
Β) Υπολογίστε τα x_1, x_2, x_3, x_4 και x_5. Εξηγείστε.
Γ) Απεικονίστε γραφικά το παραπάνω μοντέλο. Το μοντέλο αυτό είναι γραμμικό ή κυκλικό;
Δ) Πως θα τροποποιηθεί το μοντέλο αν προστεθούν καταστάσεις έναρξης (Begin) και τερματισμού (End);

3) Η κανονική έκφραση (regular expression) που περιγράφει την περιοχή συρραφής (splicing) σε ευκαρυωτικά γονίδια είναι η ακόλουθη:
[AC]AGGT[AG]AGT
1 2 3 4 5 6 7 8 9,
όπου οι θέσεις αριθμούνται με 1-9 και η αποκοπή γίνεται μεταξύ των θέσεων 3 και 4.
Α) Κατασκευάστε (και σχεδιάστε) ένα Hidden Markov Model, το οποίο να είναι *εντελώς ανάλογο* με την παραπάνω κανονική έκφραση.
Β) Αν σας δοθεί η επιπλέον πληροφορία ότι στη θέση 1, η Αδενίνη (A) εμφανίζεται με ποσοστό 10%, ενώ στη θέση 6 η πιθανότητα εμφάνισης της Γουανίνης (G) είναι τριπλάσια από αυτή της Αδενίνης (A), τροποποιήστε κατάλληλα το μοντέλο.
Γ) Δίνονται δύο αλληλουχίες DNA
x: AAACAGGTGAGTAAA
y: TTTAAGGTAAGTGGG
Ποια από τις δυο έχει μεγαλύτερες πιθανότητες να εμφανιστεί κάτω από τις προϋποθέσεις του μοντέλου, σύμφωνα με τον αλγόριθμο Forward? Εξηγήστε *ποιοτικά* τα αποτελέσματα.
Δ) Ποια είναι τα πλεονεκτήματα του HMM που κατασκευάσατε στο ερώτημα (Β), και κάποιων ακόμα καλύτερων που μπορεί να κατασκευαστούν υπό το φως περισσότερων δεδομένων (αλληλουχιών), σε σχέση με το απλό regular expression?

Σημείωση: για όλα τα παραπάνω θεωρήστε ότι σε τυχαίες περιοχές (εκτός της περιοχής συρραφής) οι πιθανότητες εμφάνισης κάθε βάσης (A, C, T, G) είναι ίσες.

4) Δίνεται η παρακάτω πρωτεϊνική αλληλουχία και η δευτεροταγής δομή της:

```
  1     GSAPSRKFFVGGNWKMNGRKQSLGELIGTLNAAKVPADTEVVCAPPTAYI
        CCCCCCEEEEEEEECCCCCHHHHHHHHHHHHCCCCCCEEEEEEECCCCCH
 51     DFARQKLDPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATWVVLGHSER
        HHHHHHCCCCEEEEEEECCCCCCCCCCCHHHHHHHCCCCEEEEECHHHH
101     RHVFGESDELIGQKVAHALAEGLG
        HCCCCCCHHHHHHHHHHHHCCCC
```

Α) Σε ποια κατηγορία πρωτεϊνικού διπλώματος πιστεύετε ότι κατατάσσεται η εν λόγω πρωτεΐνη στη βάση δεδομένων SCOP και γιατί?
Β) Κατασκευάστε ένα όσο το δυνατό πιο απλό Hidden Markov Model το οποίο να προβλέπει τη

δευτεροταγή δομή πρωτεϊνών σε 3 καταστάσεις (H, E, C) και εξηγήστε. Πώς θα μπορούσε το μοντέλο αυτό να γίνει καλύτερο?
Γ) Αν θεωρήσετε σαν δεδομένα εκπαίδευσης την παραπάνω αλληλουχία και τη δομή της, να υπολογίσετε τις παραμέτρους του εν λόγω μοντέλου (πιθανότητες μετάβασης και πιθανότητες εμφάνισης συμβόλων).

Σημείωση: για λόγους απλότητας θεωρήστε ότι στο μοντέλο δεν υπάρχουν καταστάσεις έναρξης και τερματισμού

5) Δίνεται το παρακάτω τμήμα μιας αλληλουχίας DNA η οποία προέρχεται από ένα γονίδιο με τον αντίστοιχο σχολιασμό (E=εξόνιο, I=εσώνιο, N=μη-κωδική περιοχή εκτός γονιδίου, 5=5'-άκρο συρραφής, 3=3'-άκρο συρραφής)

```
…AGCATATGCGTAGTCTGACTGCCAAGATATAAAGTTATAATCTATATACGATCGCTGTCAATGCT…
…NNNNNEEEEEEEEEEEEEEEEEEEE5IIIIIIIIIIIIIIIIIIII3EEEEEEEEEEEEEEE…
```

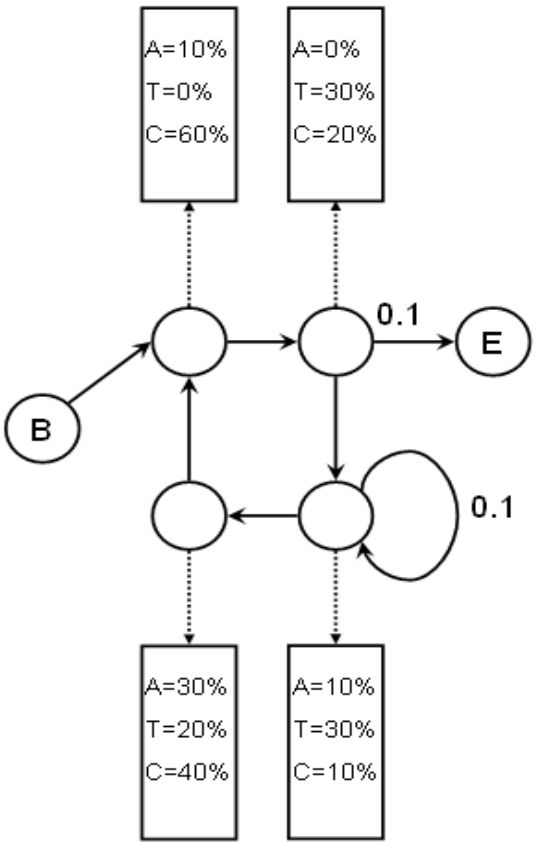
Α) Από τι είδους οργανισμό (ευκαρυωτικό ή προκαρυωτικό) προέρχεται το παραπάνω γονίδιο; Εξηγήστε την απάντησή σας.
Β) Κατασκευάστε ένα όσο το δυνατό πιο απλό Hidden Markov Model το οποίο να προβλέπει την ύπαρξη γονιδίων σε 5 καταστάσεις (E, I, 5, 3, N) και εξηγήστε. Με ποιους τρόπους θα μπορούσε το μοντέλο αυτό να γίνει καλύτερο;
Γ) Αν θεωρήσετε σαν δεδομένα εκπαίδευσης την παραπάνω αλληλουχία, να υπολογίσετε τις παραμέτρους του εν λόγω μοντέλου (πιθανότητες μετάβασης και πιθανότητες εμφάνισης συμβόλων).

Σημείωση: Για λόγους απλότητας θεωρήστε ότι στο μοντέλο δεν υπάρχουν καταστάσεις έναρξης και τερματισμού. Για όσες παραμέτρους δεν έχετε στοιχεία, βάλτε κάποια υποθετικά νούμερα.

6) Δίνεται το παρακάτω σχήμα στο οποίο απεικονίζεται η γραφική αναπαράσταση ενός HMM.
Α) Περιγράψτε το μοντέλο αναλυτικά (Πίνακες μεταβάσεων και εμφανίσεων συμβόλων, αλφάβητο, καταστάσεις).
Β) Δίνονται οι παρακάτω αλληλουχίες:

$\mathbf{x}_1$=CGGCCG
$\mathbf{x}_2$=AT
$\mathbf{x}_3$=ACCGAT
$\mathbf{x}_4$=TACGAT
$\mathbf{x}_5$=GGCCG

Υπολογίστε τη συνολική πιθανότητα κάθε μιας με βάση το παρακάτω μοντέλο. Ποιες από τις αλληλουχίες δεν μπορούν να παραχθούν από το μοντέλο αυτό?
Γ) Πως θα τροποποιούσαμε το μοντέλο έτσι ώστε να μπορεί να παράγει και τις αλληλουχίες του ερωτήματος Β?



7) Θεωρήστε ένα όσο το δυνατόν πιο απλό HMM, το οποίο να περιέχει 2 καταστάσεις και 2 σύμβολα. Σχεδιάστε το μοντέλο και δώστε μια γενική έκφραση για τους πίνακες *a* και *e*. Δείξτε ότι το ίδιο μοντέλο, μπορεί να αναπαρασταθεί με ένα κλασικό μαρκοβιανό μοντέλο, του οποίου τις παραμέτρους θα ορίσετε. Δοκιμάστε να αυξήσετε την πολυπλοκότητα του HMM, αυξάνοντας για παράδειγμα των αριθμό των καταστάσεων σε 3, 4, κ.ο.κ. Τι παρατηρείτε για το αντίστοιχο μαρκοβιανό μοντέλο;

8) Αποδείξτε ότι σε ένα HMM, ισχύει :

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

Σημείωση: Ξεκινήστε από την πιθανότητα $P(\mathbf{x}, \pi_i = k \mid \theta)$ και θυμηθείτε ότι

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

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

9) Πώς πραγματοποιείται μια πολλαπλή στοίχιση με χρήση profile HMM?
Δίνονται οι αλληλουχίες:

x_a=WAYDDR, και

x_b=WDAYPDDR

και τα αντίστοιχα μονοπάτια (paths) από τον αλγόριθμο του Viterbi:

$p_a = m_0 m_1 m_2 m_3 d_4 d_5 m_6 m_7 m_8 m_9$, και $p_b = m_0 m_1 i_1 m_2 m_3 d_4 m_5 m_6 m_7 m_8 m_9$

Δώστε την στοίχιση των δυο αλληλουχιών και εξηγείστε.