



**ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ
ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΥΠΟΛΟΓΙΣΤΙΚΗ
ΒΙΟΙΑΤΡΙΚΗ»**

**Hidden Markov Model (HMM) και επεκτάσεις τους στην
Βιοπληροφορική
Hidden Markov Model (HMM) and their extensions in
Bioinformatics**

Ιωάννης Ταμπόσης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Υπεύθυνος

Παντελής Μπάγκος

Αναπληρωτής Καθηγητής

Λαμία, 2017



ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΔΙΑΤΜΗΜΑΤΙΚΟ ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΙΑΤΡΙΚΗ
ΚΑΤΕΥΘΥΝΣΗ

«ΥΠΟΛΟΓΙΣΤΙΚΗ ΙΑΤΡΙΚΗ ΚΑΙ ΒΙΟΛΟΓΙΑ»

Hidden Markov Model (HMM) και επεκτάσεις τους στην
Βιοπληροφορική
Hidden Markov Model (HMM) and their extensions in
Bioinformatics

Ιωάννης Ταμπόσης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων

Παντελής Μπάγκος

Αναπληρωτής Καθηγητής

Λαμία, 2017

«Υπεύθυνη Δήλωση μη λογοκλοπής και ανάληψης προσωπικής ευθύνης»

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, και γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα και ενυπογράφως ότι η παρούσα εργασία με τίτλο «Hidden Markov Model (HMM) και επεκτάσεις τους στην Βιοπληροφορική» αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές από τις οποίες χρησιμοποίησα δεδομένα, ιδέες, φράσεις, προτάσεις ή λέξεις, είτε επακριβώς (όπως υπάρχουν στο πρωτότυπο ή μεταφρασμένες) είτε με παράφραση, έχουν δηλωθεί κατάλληλα και ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής.

Ο ΔΗΛΩΝ

Ιωάννης Ταμπόσης

Ημερομηνία

Υπογραφή

**Hidden Markov Model (HMM) και επεκτάσεις τους στην
Βιοπληροφορική**
Hidden Markov Model (HMM) and their extensions in

Ιωάννης Ταμπόσης

Τριμελής Επιτροπή:

Παντελής Μπάγκος (επιβλέπων)
Αναπληρωτής Καθηγητής

Βασίλειος Πλαγιανάκος
Αναπληρωτής Καθηγητής

Μαρία Αδάμ
Επίκουρος Καθηγήτρια

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ	6
ΠΡΟΛΟΓΟΣ.....	8
ΠΕΡΙΛΗΨΗ.....	9
ABSTRACT	10
1. Hidden Markov Model	11
1.1 Αλυσίδα Markov.....	11
1.2 Επέκταση σε Hidden Markov Models	13
1.3 Παραδείγματα σε Hidden Markov Model	13
1.4 Ορισμός των HMMs	16
1.5 Γραφική απεικόνιση των HMMs.....	17
1.6 Τα τρία βασικά προβλήματα στην θεωρία των HMMs	19
1.7 Υπολογισμός πιθανοφάνειας	20
1.7.1 Αλγόριθμος Forward.....	21
1.7.2 Αλγόριθμος Backward	22
1.8 Αποκωδικοποίηση.....	22
1.8.1 Αλγόριθμος Viterbi	23
1.9 Εκτίμηση Παραμέτρων	23
1.9.1 Αλγόριθμος Baum-Welch	24
1.9.2 Μέθοδοι Gradient-descent	25
2. Class Hidden Markov Model.....	27
2.1 Ορισμός των CHMMs	27
2.2 Υπολογισμός πιθανοφάνειας	28
2.3 Εκτίμηση Παραμέτρων	29
3. Επεκτάσεις των HMM.....	31
3.1 High-Order HMM.....	32
3.2 Partially HMM.....	33
3.3 HMM with states depending on observations.....	33
3.4 Hidden Neural Network HNN	34
4. Στόχος.....	35
5. Μεθοδολογία	37
5.1 μέθοδος επέκτασης της Κωδικοποίησης.....	37
5.2 Μοντέλα.....	38
5.3 Κωδικοποίηση.....	42
5.4 Αλφάβητο	42
5.5 Πιθανότητες	44

5.6 Διαδικασία Αξιολόγησης.....	45
5.6.1 Μέθοδοι Ελέγχου	46
5.6.2 Μέτρα αξιοπιστίας	48
5.7 Υλοποίηση	50
6. Αποτελέσματα	51
6.1 Σύνολο Παραμέτρων και Παρατηρήσεων	51
6.2 PRED-TMBB.....	51
6.3 HMM-TM	53
6.4 PRED-TAT	55
6.5 PRED-LIPO	57
6.6 PRED-SIGNAL	59
6.7 Απόδοση	60
7. Συζήτηση-Συμπεράσματα	62
Βιβλιογραφία.....	64

ΠΡΟΛΟΓΟΣ

Η παρούσα μεταπτυχιακή εργασία εκπονήθηκε στα πλαίσια του διατμηματικού μεταπτυχιακού προγράμματος «Πληροφορική και Υπολογιστική Βιοϊατρική», κατεύθυνση «Υπολογιστική Ιατρική και Βιολογία», του τμήματος Πληροφορικής με εφαρμογές στην Βιοϊατρική της Σχολής Θετικών Επιστημών του Πανεπιστημίου Θεσσαλίας.

Μέσα από τη συγκεκριμένη εργασία μου δίνεται η δυνατότητα να ευχαριστήσω την γυναίκα μου Άννα και τους γιούς μου Αριστοτέλη και Δημήτρη για την τεράστια υπομονή που δείχνουν, για τον προσωπικό και ποιοτικό χρόνο που χάνουν από εμένα και που με υποστηρίζουν και θα συνεχίσουν να μου συμπαραστέκονται στα υπόλοιπα ακαδημαϊκά και ερευνητικά βήματα μου.

Τέλος, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Παντελή Μπάγκο για την ευκαιρία που μου έδωσε να ασχοληθώ με τον τομέα της έρευνας και της Βιοπληροφορικής καθώς και την απαιτούμενη γνώση, βοήθεια και καθοδήγηση που μου παρέχει στη μέχρι τώρα πορεία μου. Επίσης θα ήθελα να ευχαριστήσω την μεταδιδάκτορα Μαργαρίτα Θεοδωροπούλου για την βοήθεια που μου παρείχε σε όλη τη διάρκεια των δοκιμών και την άψογη συνεργασία την οποία καταφέραμε να οικοδομήσουμε. Η συνεργασία αυτή θα αποτελέσει και το έναυσμα για την συνέχεια της συνεργασίας μας.

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία θα ασχοληθούμε αποκλειστικά με ένα είδος Hidden Markov Model. Θα περιγράψουμε αρχικά τη θεωρία των κλασικών Hidden Markov Models με σκοπό να ορίσουμε και να θέσουμε τις βάσεις για να περιγράψουμε την επέκταση τους στην κατηγορία των Class Hidden Markov Models με την οποία θα εργαστούμε. Στη συνέχεια θα εστιάσουμε την προσοχή μας σχετικά με την ενσωμάτωση επιπρόσθετης πληροφορίας, λαμβάνοντας υπόψη το υδρόφοβο-πολικό μοντέλο, επεκτείνοντας μια σειρά από γνωστά HMM μοντέλα που χρησιμοποιούνται για την πρόβλεψη διαμεμβρανικών πρωτεϊνών με σκοπό να βελτιστοποιήσουμε την αποδοτικότητα των μοντέλων στην πρόγνωση πρωτεϊνών. Με κατάλληλες δοκιμές Jack-Knife και cross-validation τα αποτελέσματά μας δείχνουν ότι μπορεί να είναι δυνατόν να βελτιωθεί η ακρίβεια πρόβλεψης.

Λέξεις Κλειδιά: Hidden Markov Model, Βιοπληροφορική, πρόβλεψη, διαμεμβρανικές πρωτεΐνες,

ABSTRACT

In this dissertation, we will study one type of Hidden Markov Model. We will first review the theory of Hidden Markov Models and then extend the ideas to Class Hidden Markov models. We will then focus on the integration of additional information, taking into account the hydrophobic-polar model, extending a series of known HMM models used for predicting transmembrane proteins in order to optimize the efficiency of the models. With appropriate Jack-Knife and cross-validation tests, our results suggest that it may be possible to improve the prediction accuracy.

Keywords: Hidden Markov Model, Bioinformatics, prediction, transmembrane proteins

1. Hidden Markov Model

Αν και αρχικά εισήχθησαν και μελετήθηκαν στα τέλη της δεκαετίας του 1960 και στις αρχές της δεκαετίας του 1970, οι στατιστικές μέθοδοι της πηγής Markov ή Hidden Markov Models έχουν γίνει όλο και πιο δημοφιλείς τα τελευταία χρόνια. Η Μαρκοβιανή θεωρία πήρε το όνομά της από τον Andrei Markov στις αρχές του 20^{ου} αιώνα ο οποίος εμπνεύστηκε τα μαρκοβιανά μοντέλα, μελετώντας τις εναλλαγές των συμφώνων και φωνηέντων σε ένα ποίημα του Pushkin [1], αλλά η θεωρία των Hidden Markov Models στην πραγματικότητα αναπτύχθηκε από τον Baum και τους συνεργάτες του στην δεκαετία του '60 [2]. Τα HMMs είναι κατά βάση στοχαστικά, με ξεκάθαρη δηλαδή πιθανοθεωρητική δομή και ως εκ τούτου μπορούν να αποτελέσουν τη θεωρητική βάση για χρήση σε ένα ευρύ φάσμα εφαρμογών στους τομείς, όπως: αναγνώριση και σύνθεση φωνής και γραφής, αναγνώριση και ταξινόμηση σεισμικών σημάτων, αναγνώριση και ταξινόμηση μελωδιών, επεξεργασία φυσικής γλώσσας, στην βιοπληροφορική και σε άλλους τομείς. Επιπλέον, χρησιμοποιούνται σε εφαρμογές, όπου το στατιστικό μοντέλο χρησιμεύει στην γρήγορη μετάδοση ή αποθήκευση μεγάλου όγκου πληροφορίας.

Στον τομέα της βιοπληροφορικής, τα HMMs είναι γνωστά από την χρήση τους σε εφαρμογές, όπως η αναγνώριση γονιδίων, η μοντελοποίηση και στοίχιση βιολογικών ακολουθιών, η πρόβλεψη πρωτεϊνικής δομής στον χώρο, η εύρεση ορισμένων υπακολουθιών σε μεγαλύτερες, βιολογικές ακολουθίες, όπως π.χ η εύρεση των promoters, που καθορίζουν το σημείο έναρξης της μετάφρασης πάνω στο DNA και από άλλες εφαρμογές. Αξίζει να αναφερθεί στο συγκεκριμένο σημείο, ότι τα μαρκοβιανά μοντέλα θεωρούνται από πολλούς ερευνητές ως τα πιο φυσικά για να περιγράψουν ακολουθίες μακρομορίων όπως του DNA αλλά και των πρωτεϊνών καθώς προσεγγίζουν την εξάρτηση της πληροφορίας που εμπεριέχεται σε μια ακολουθία.

1.1 Αλυσίδα Markov

Ως μια αλυσίδα Markov (Markov chain) θεωρείται ένα μοντέλο πιθανοτήτων που μπορεί να χαρακτηριστεί σε οποιαδήποτε χρονική στιγμή ότι βρίσκεται σε μια κατάσταση από ένα σύνολο L διακριτών καταστάσεων x_1, x_2, \dots, x_L , και το οποίο μεταβαίνει από κατάσταση σε κατάσταση [3]. Στην περίπτωση των βιολογικών αλληλουχιών, ως σύνολο παρατηρούμενων συμβόλων ορίζονται τα σύμβολα της ακολουθίας τα οποία ανήκουν στο πεπερασμένο αλφάβητο δηλαδή τα 4 νουκλεοτίδια (AGCT ή U) στην περίπτωση του DNA ή τα 20 αμινοξέα στην περίπτωση των πρωτεϊνών (ACDEFGHIKLMNPQRSTVWY).

Συμβολίζουμε τις χρονικές στιγμές που σχετίζονται με τις αλλαγές της κατάστασης ως $i = 1, 2, \dots$ και την πραγματική κατάσταση τη χρονική στιγμή i ως x_i . Μια πλήρης πιθανοθεωρητική περιγραφή του παραπάνω συστήματος, σε γενικές γραμμές, θα απαιτούσε προδιαγραφή της τρέχουσας κατάστασης (τη χρονική στιγμή i), καθώς και όλες τις προηγούμενες καταστάσεις. Στην Μαρκοβιανή αλυσίδα αυτή η πιθανοθεωρητική περιγραφή περικόπτεται ακριβώς στην τρέχουσα και την κατάσταση αμέσως προηγούμενη κατάσταση της. Με άλλα λόγια, η γνώση της πιο πρόσφατης κατάστασης του συστήματος καθιστά τη λιγότερο πρόσφατη ιστορία άχρηστη.

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

Επιπλέον μια αλυσίδα Markov έχει και έναν πίνακα πιθανοτήτων μετάβασης τα στοιχεία του οποίου δίνονται από την παρακάτω σχέση

$$\alpha_{ij} = P(x_i = t | x_{i-1} = s), 1 \leq i, j \leq N$$

και έχει τις παρακάτω ιδιότητες

$$\alpha_{ij} \geq 0$$

$$\sum_{j=1}^L \alpha_{ij} = 1 \text{ για κάθε } j = 1, 2, \dots, N$$

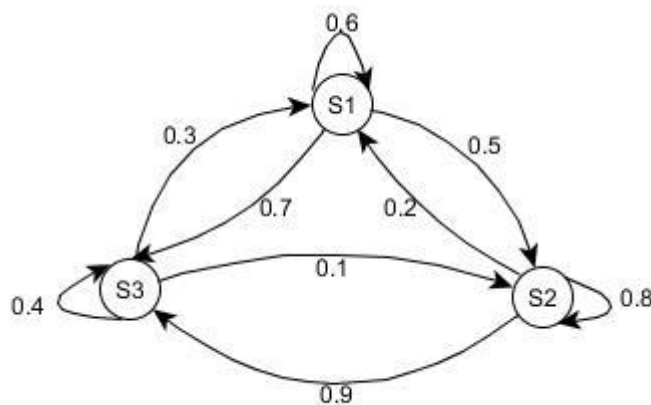
Παράδειγμα αλυσίδας Markov

Έστω σύνολο καταστάσεων $S = (S_1, S_2, S_3)$ και αρχικές πιθανότητες $P(X_B = i)$ που προσδιορίζουν τις πιθανότητες η πρώτη κατάσταση του συστήματος να είναι η i .

Κατάσταση	S1	S2	S3
Πιθανότητα Έναρξης	0.7	0.2	0.1

Πίνακας 1. Πίνακας πιθανοτήτων μετάβασης των αρχικών καταστάσεων

Και ένα σύνολο μεταβάσεων με τον αντίστοιχο πίνακα



Εικόνα 1. Μια αλυσίδα Markov με 3 Καταστάσεις (επισημασμένες από S_1 έως S_3) με τις αντίστοιχες πιθανότητες μετάβασης.

Κατάσταση	S1	S2	S3
S1	$P(S1 \rightarrow S1) = 0.6$	$P(S1 \rightarrow S2) = 0.5$	$P(S1 \rightarrow S3) = 0.7$
S2	$P(S2 \rightarrow S1) = 0.2$	$P(S2 \rightarrow S2) = 0.8$	$P(S2 \rightarrow S3) = 0.9$
S3	$P(S3 \rightarrow S1) = 0.3$	$P(S3 \rightarrow S2) = 0.1$	$P(S3 \rightarrow S3) = 0.4$

Πίνακας 2. Πίνακας πιθανοτήτων μετάβασης των καταστάσεων

Υποθέτουμε επίσης ότι το σύνολο συμβόλων εκπομπής είναι το $T = (0, 1)$. Η κατανομή πιθανότητας εκπομπής συμβόλων ανά κατάσταση δίνεται από τον παρακάτω πίνακα:

Κατάσταση	Πιθανότητα εκπομπής Συμβόλου 0	Πιθανότητα εκπομπής Συμβόλου 1
S1	0.8	0.2
S2	0.6	0.4
S3	0.3	0.7

Πίνακας 3. Πίνακας πιθανοτήτων εκπομπής συμβόλων των καταστάσεων

Η αλυσίδα Markov συνήθως περιγράφεται χρησιμοποιώντας την έννοια της χρονικής εξέλιξης, όπου οι μεταβάσεις πραγματοποιούνται σε διακριτές χρονικές στιγμές. Στην εφαρμογή του μοντέλου σε βιολογικές ακολουθίες, ο “διακριτός χρόνος” $i = 0, 1, 2, \dots$ αντιστοιχεί στο πρώτο, δεύτερο, τρίτο αμινοξύ της ακολουθίας κ.ο.κ. και οι έξοδοι του συστήματος είναι τα ίδια τα αμινοξέα.

1.2 Επέκταση σε Hidden Markov Models

Το μοντέλο Markov το οποίο εξετάσαμε προηγουμένως αντιστοιχεί με ένα παρατηρήσιμο (φυσικό) γεγονός και είναι περιοριστικό όταν εφαρμόζεται σε πολλά προβλήματα που παρουσιάζουν ενδιαφέρον. Το πρόβλημα έρχεται να λύσει η επέκταση τους σε HMMs (Hidden Markov Models) όπου οι παρατηρήσεις πλέον αποδεσμεύονται από τις καταστάσεις. Το νέο μοντέλο περιγράφεται ως μια διπλή ενσωμάτωση στοχαστικών διαδικασιών, σύμφωνα με την οποία η μία διαδικασία που δεν είναι δυνατόν να παρατηρηθεί (είναι κρυφή), πάρα μόνον διαμέσου ενός συνόλου στοχαστικών διαδικασιών, οι οποίες παράγουν ακολουθίες παρατηρήσεων [3]. Στην απλή περίπτωση, το HMM αποτελείται από δύο στοχαστικές διαδικασίες, εκ των οποίων η μία αποτελείται από ένα πεπερασμένο σύνολο τυχαίων μεταβλητών (καταστάσεων), των οποίων την στοχαστική συμπεριφορά δεν μπορούμε να παρατηρήσουμε. Η δεύτερη διαδικασία αποτελείται από ένα πεπερασμένο σύνολο τυχαίων μεταβλητών, που εξαρτώνται άμεσα από τις καταστάσεις, για τις οποίες γνωρίζουμε πως εξαρτώνται από αυτές και την στοχαστική συμπεριφορά τους μπορούμε να την παρατηρήσουμε. Συνεπώς, ο μόνος τρόπος για να αντλήσουμε πληροφορία για την «κρυφή» στοχαστική διαδικασία είναι, μέσω των ακολουθιών παρατηρήσεων, που παράγονται από την δεύτερη, στοχαστική διαδικασία.

Το Hidden Markov Model (HMM) είναι μια στατιστική μέθοδος που χρησιμοποιεί το μέτρο πιθανότητας για την μοντελοποίηση δεδομένων που αποτελούνται από ακολουθίες παρατηρήσεων.

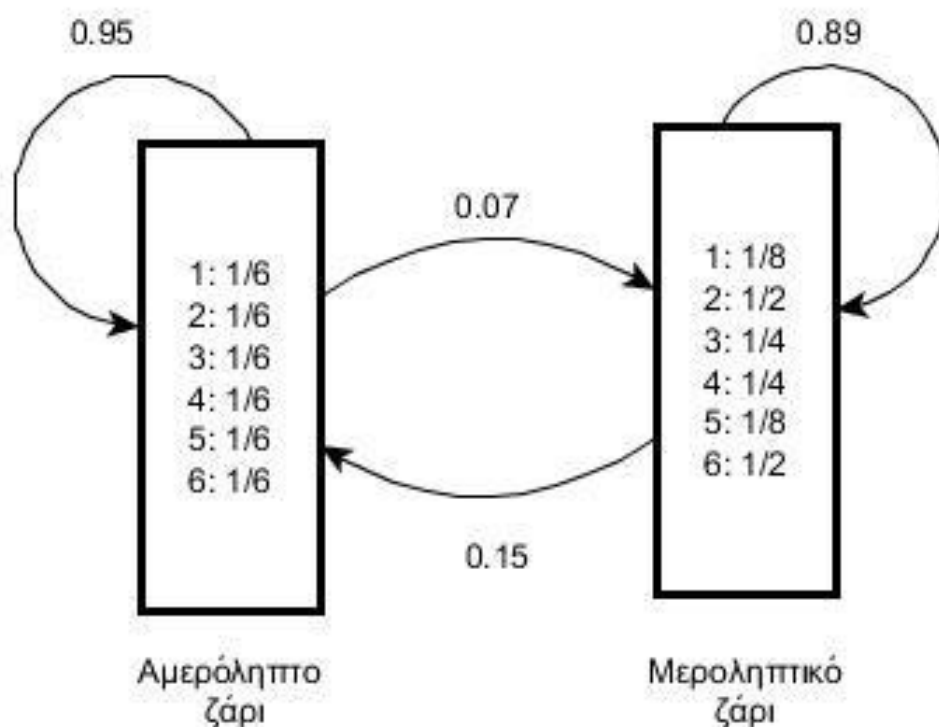
1.3 Παραδείγματα σε Hidden Markov Model

Στο σημείο αυτό, παρουσιάζουμε γνωστά παραδείγματα που μπορούν να μοντελοποιηθούν από ένα HMM και για τα οποία υπάρχει αναφορά στην βιβλιογραφία.

Παράδειγμα 1

Το πιο απλό παράδειγμα είναι το παράδειγμα του «ανέντιμου καζίνου» στο οποίο ένα καζίνο χρησιμοποιεί κατά περίπτωση δύο ειδών ζάρια, τα οποία μπορεί να τα αλλάξει με κάποιες πιθανότητες και να τα χρησιμοποιεί ανάλογα με τον τρόπο που ευνοείται σε κάποιο αποτέλεσμα [4]. Το ένα ζάρι είναι αμερόληπτο και οι πιθανότητες εμφάνισης ενός συμβόλου (1,2,3,4,5,6) καθορίζεται από τον παράγοντα τύχη και το δεύτερο ζάρι θεωρείται «πειραγμένο» με αποτέλεσμα να παράγει σύμβολα με προκαθορισμένο τρόπο οπότε έχουμε μεροληπτική κατανομή εμφάνισης των συμβόλων. Όπως διακρίνουμε στην Εικόνα 2 τα δύο διαφορετικά είδη ζαριών απεικονίζονται ως παραλληλόγραμμα και αποτελούν τις κρυφές καταστάσεις, τις οποίες ο παίκτης δεν μπορεί να παρατηρήσει. Επίσης επιλέχθηκε αυθαίρετα η πιθανότητα μετάβασης από το αμερόληπτο στο μεροληπτικό 0,07 και από το μεροληπτικό στο αμερόληπτο 0,15. Όπως είναι φυσικό ο παίκτης το μόνο που βλέπει είναι το αποτέλεσμα που έρχεται σε κάθε ρήξη του ζαριού χωρίς να γνωρίζει από πιο ζάρι έχει προέλθει η ρήξη.

Αντίστοιχα μέσα στα παραλληλόγραμμα εμφανίζονται οι πιθανότητες από τις οποίες μπορεί να προέλθει ένα σύμβολο σε κάποια από τις δύο καταστάσεις που μπορεί να βρίσκεται το ζάρι. Η διαφορά του HMM σε σχέση με το απλό μοντέλο Markov που είδαμε προηγουμένως είναι ότι σε αυτήν την περίπτωση δεν υπάρχει εξάρτηση μεταξύ των συμβόλων και των καταστάσεων του μοντέλου. Η λειτουργικότητα του νέου μοντέλου διαφαίνεται ακόμη καλύτερα στην περίπτωση που θα μπορούσαμε να χρησιμοποιήσουμε 3 κρυφές καταστάσεις στις οποίες θα μπορούσε να βρεθεί το ζάρι χωρίς βέβαια να επηρεάζονται οι παρατηρήσιμες τιμές. Σε αυτή την περίπτωση θα χρειαζόταν να επεκτείνουμε το μοντέλο που απεικονίζεται στην Εικόνα 2.



Εικόνα 2. Το παράδειγμα του ανέντιμου καζίνου.

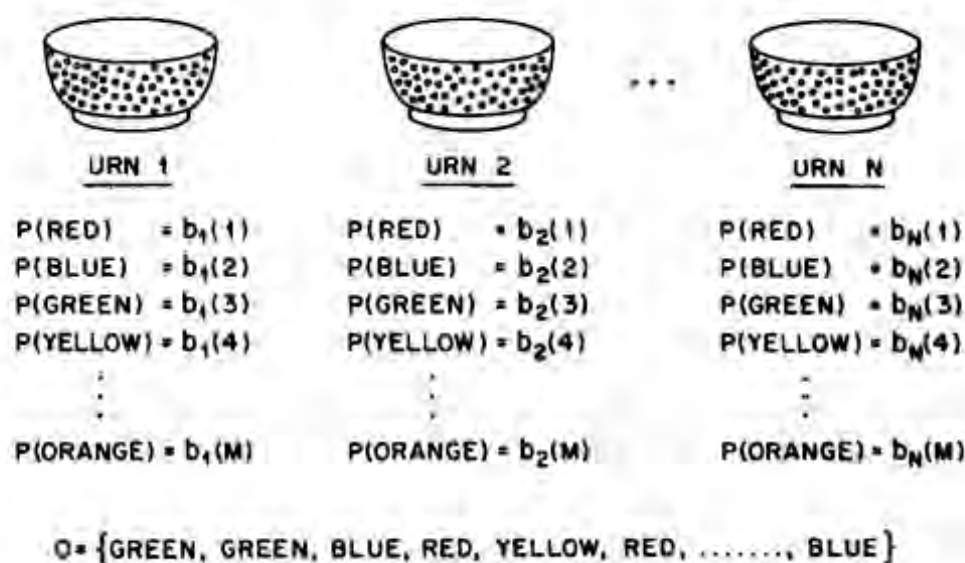
Παράδειγμα 2

Το επόμενο παράδειγμα πρωτοπαρουσιάστηκε από τον Jack Ferguson στις εισαγωγικές διαλέξεις του για τα HMMs [3].

Υποθέτουμε ότι υπάρχουν N δοχεία και σε κάθε δοχείο υπάρχει ένας μεγάλος αριθμός από χρωματιστές μπάλες διαφόρων χρωμάτων. Υποθέτουμε ότι υπάρχουν M διακριτά χρώματα των μπαλών. Η μέθοδος παραγωγής των παρατηρήσεων είναι η εξής : ένα τζίνι που υπάρχει στο χώρο, επιλέγει με τυχαία διαδικασία, ένα από τα N δοχεία. Επιλέγει τυχαία μια μπάλα, της οποίας το χρώμα καταγράφεται και την επανατοποθετεί στο δοχείο. Στην συνέχεια, επιλέγει πάλι ένα δοχείο, με τρόπο που καθορίζεται από την προηγούμενη επιλογή δοχείου ή των m προηγούμενων επιλογών του. Από το δοχείο που επιλέγει, βγάζει εκ νέου μια μπάλα, της οποίας το χρώμα καταγράφεται. Η διαδικασία επαναλαμβάνεται, παράγοντας τελικά μια ακολουθία παρατηρήσεων, που αποτελείται από τα χρώματα κάθε μπάλας που επιλέχθηκε.

Από την παραπάνω περιγραφή γίνεται προφανές ότι ένα απλό HMM που μπορεί να κατασκευαστεί για την μοντελοποίηση του παραπάνω συστήματος, από ένα ευρύ σύνολο

δυνατών HMMs, αποτελείται από ένα σύνολο N καταστάσεων, στο οποίο κάθε κατάσταση αναπαριστά ένα από τα N δοχεία του συστήματος. Σε κάθε κατάσταση, αντιστοιχεί μια τυχαία μεταβλητή που αναπαριστά το χρώμα της μπάλας που επιλέγεται και μια πιθανοτική συνάρτηση για την μεταβλητή του χρώματος, που εξαρτάται από την συγκεκριμένη κατάσταση (π.χ. το περιεχόμενο του συγκεκριμένου δοχείου σε χρωματιστές μπάλες, τον τρόπο που επιλέγει από ένα δοχείο τις μπάλες κ.τ.λ.). Ένας πίνακας μεταβάσεων μεταξύ των καταστάσεων περιγράφει τον τρόπο, με τον οποίο επιλέγει το τζίνι από ποιο δοχείο θα επιλέξει μια μπάλα την επόμενη χρονική στιγμή. Επομένως, στον πίνακα μετάβασης, η πιθανότητα μετάβασης σε μια κατάσταση την επόμενη χρονική στιγμή καθορίζεται από την παρούσα κατάσταση ή τις m προηγούμενες καταστάσεις όπως καθορίζει η ιδιότητα Markov, που ισχύει στα HMMs για τις καταστάσεις του μοντέλου.



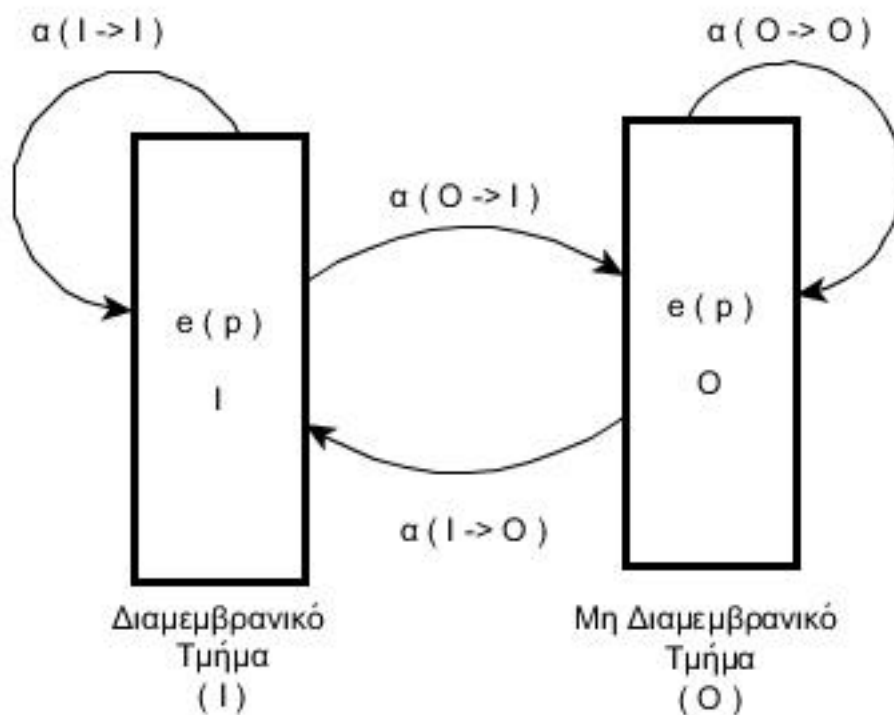
Εικόνα 3. Το παράδειγμα με το δοχείο και τις μπάλες[3]

Παράδειγμα 3

Μία ικανότητα των HMM, είναι να μοντελοποιεί μια γραμματική [5]. Πολλά προβλήματα στην ανάλυση βιολογικών ακολουθιών έχουν μια γραμματική όπως θα παρουσιάσουμε στο συγκεκριμένο παράδειγμα με την χρήση ενός μοντέλου HMM σε διαμεμβρανικές πρωτεΐνες. Σαν σύμβολα θα χρησιμοποιήσουμε τα 20 αμινοξέα (ACDEFGHIKLMNPQRSTVWY) και ως καταστάσεις αν ανήκουν σε διαμεμβρανικό τμήμα ή όχι.

A D E R P G G S D A D K L I L V C I G F V L I F V T Y T
 O O O O O O O O O O O I I I I I I I I I I I I O O O O

Στην πρώτη γραμμή έχουμε μία ακολουθία αμινοξέων διαμεμβρανικής πρωτεΐνης, και στην δεύτερη γραμμή παρουσιάζεται μία ακολουθία καταστάσεων του μοντέλου μας που χαρακτηρίζει ποια περιοχή αποτελεί διαμεμβρανικό τμήμα και ποια μη διαμεμβρανικό. Το διαμεμβρανικό τμήμα έχει επισημανθεί με γκρι χρωματισμό. Μια άλλη πιθανή εφαρμογή είναι για τη μοντελοποίηση DNA ή πρωτεϊνών [6].



Εικόνα 4. Το παράδειγμα των διαμεμβρανικών πρωτεϊνών

1.4 Ορισμός των HMMs

Τα παραπάνω παραδείγματα αποτελούν μια αρκετά καλή εισαγωγή, στο τι είναι ένα HMM και πώς μπορεί να εφαρμοστεί σε μερικές απλές περιπτώσεις, να ορίσουν τα στοιχεία που ενσωματώνει ένα HMM, και να εξηγήσουν πώς το μοντέλο παράγει τις ακολουθίες παρατηρήσεων.

Ένα Hidden Markov Model (HMM) είναι ένα στατιστικό-πιθανοθεωρητικό μοντέλο που ορίζεται από τα εξής 3 στοιχεία Σ , Q , θ

$$M = (V, Q, \theta)$$

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

Αναλύοντας περισσότερο τα αντικείμενα ενός HMM αποτελείται από

1. Ένα σύνολο κρυφών καταστάσεων (N). Παρά το γεγονός ότι οι καταστάσεις είναι κρυφές, σε πολλές πρακτικές εφαρμογές έχει κάποια φυσική σημασία που συνδέεται με τις καταστάσεις ή σε ομάδες καταστάσεων του μοντέλου όπως είδαμε στα παραπάνω παραδείγματα και πιο συγκεκριμένα στο παράδειγμα με τις διαμεμβρανικές πρωτεΐνες.

Συμβολίζουμε τις επιμέρους καταστάσεις μέλη ως $q = (q_1, q_2, \dots, q_N)$, και την κατάσταση τη χρονική στιγμή i ως π_i .

2. ένα σύνολο διακριτών παρατηρούμενων συμβόλων (Ω). Τα παρατηρούμενα σύμβολα αντιστοιχούν στο φυσικό αποτέλεσμα του προβλήματος που μοντελοποιείται. Σε μια πρωτεύουσα ακολουθία σαν σύμβολα θα χρησιμοποιήσουμε τα 20 αμινοξέα (ACDEFGHIKLMNPQRSTVWY) ενώ για το μοντέλο δοχείο και μπάλες θα είναι τα χρώματα των μπαλών που επιλέγονται από τα δοχεία. Συμβολίζουμε τα μεμονωμένα σύμβολα ως $V = (V_1, V_2, \dots, V_\Omega)$
3. οι πιθανότητες μετάβασης των καταστάσεων του μοντέλου (transitions) όπου ισχύει $a_{kl} = P(\pi_i = l | \pi_{i-1} = k) \quad 1 \leq i, j \leq N$
Για την ειδική περίπτωση όπου οποιαδήποτε κατάσταση μπορεί να μεταβεί σε οποιαδήποτε άλλη κατάσταση σε ένα μόνο βήμα ισχύει $a_{ij} > 0$ για κάθε i, j και σε κάθε άλλη περίπτωση ισχύει $a_{ij} = 0$.

Σε ένα μοντέλο HMM μπορούμε να ορίσουμε επιπλέον και μια πιθανότητα για την κατάσταση έναρξης B (begin)

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

και μια άλλη κατάσταση για τον τερματισμό E (End)

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

4. οι πιθανότητες εκπομπής ή εμφάνισης συμβόλων (emissions) για κάθε κατάσταση i , όπου ισχύει

$$e_k(b) = P(x_i = b | \pi_i = k), \quad 1 \leq j \leq N, \quad 1 \leq k \leq \Omega$$

Στο παράδειγμα με τις πρωτεΐνες, δηλώνουν την πιθανότητα εμφάνισης ενός συγκεκριμένου συμβόλου (αμινοξέα) b , στην θέση i της ακολουθίας, δεδομένου ότι το σύστημα βρίσκεται στην κατάσταση k (διαμεμβρανικό τμήμα ή μη).

5. Την αρχική κατανομή των καταστάσεων $\pi = (\pi_i)$ έως μια συγκεκριμένη θέση i και την ονομάζουμε «μονοπάτι» (path)

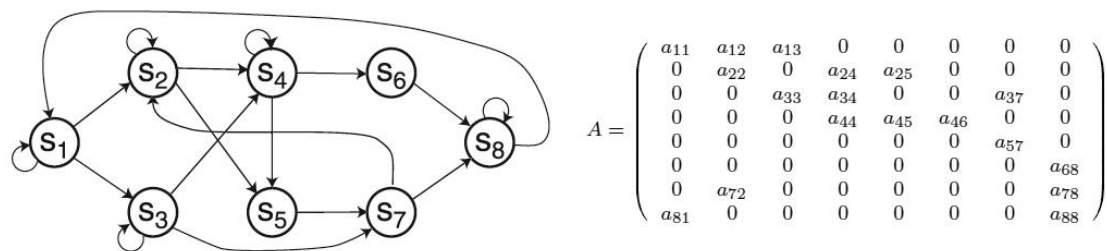
Ορίζοντας κατάλληλες τιμές στους παραπάνω ορισμούς, ένα HMM μπορεί να χρησιμοποιηθεί για να δώσει μια παρατηρούμενη ακολουθία $x = x_1, x_2, \dots, x_{L-1}, x_L$ όπου κάθε παρατήρηση αντιστοιχεί σε ένα από τα σύμβολα του αλφαβήτου που έχει οριστεί από το σύνολο V και το L αντιστοιχεί στο μήκος της ακολουθίας, υπολογίζοντας την από κοινού πιθανότητα μιας ακολουθίας x και του μονοπατιού π ως εξής:

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

1.5 Γραφική απεικόνιση των HMMs

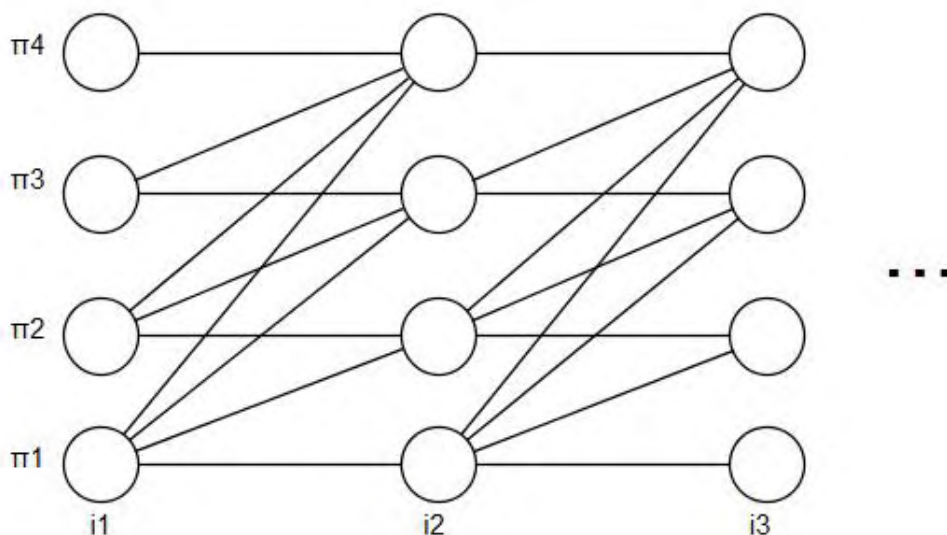
Σε αυτό το σημείο έχει ενδιαφέρον, να παρουσιάσουμε εν συντομία, και μερικούς τρόπους απεικόνισης των HMMs. Για τα HMMs χρησιμοποιούνται, κυρίως, τρεις τρόποι γραφικής απεικόνισης [7].

Όπως φαίνεται στην Εικόνα 5, ένα HMM απεικονίζεται, όπως ένα στοχαστικό αυτόματο πεπερασμένων καταστάσεων (stochastic FSA). Το γράφημα απεικονίζει μόνο τις επιτρεπόμενες μεταβάσεις στις οποίες μπορεί να βρεθεί το HMM. Κάθε κόμβος αντιστοιχεί σε μία από τις καταστάσεις και όπου μία ακμή μεταβαίνει από τον κόμβο i στον κόμβο j υποδεικνύει ότι $a_{ij} > 0$, και η έλλειψη μιας τέτοιας ακμής υποδεικνύει ότι $a_{ij} = 0$. Κάποιος μπορεί να προβλέψει ότι βρίσκεται σε μια συγκεκριμένη κατάσταση j σε μια ορισμένη χρονική στιγμή, παράγοντας ένα σύμβολο παρατήρησης από την κατανομή της παρατήρησης που αντιστοιχεί σε αυτή την κατάσταση $b_j(x)$, και στη συνέχεια να μεταβεί στην επόμενη κατάσταση, σύμφωνα με τις μη μηδενικές μεταβάσεις. Στοχεύει κυρίως στην παρουσίαση της τοπολογίας μιας Hidden Markov αλυσίδας.



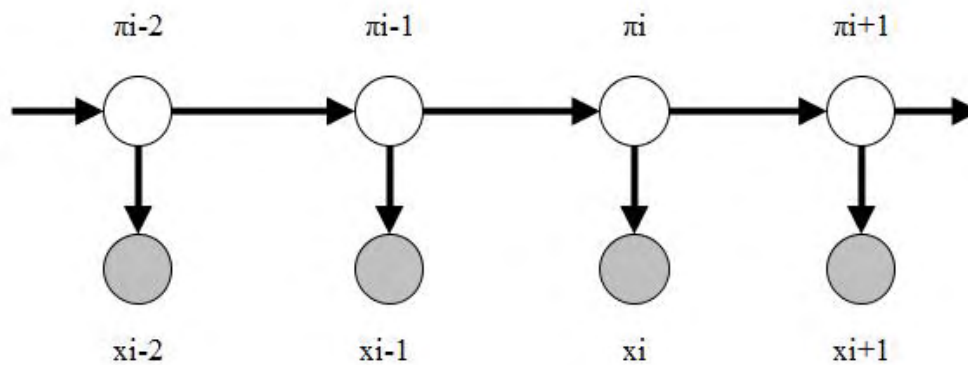
Εικόνα 5. Απεικόνιση HMM με FSA (finite stochastic automaton) τρόπο [7]

Ένας δεύτερος τρόπος απεικόνισης, όπως φαίνεται στην Εικόνα 6, απεικονίζει το σύνολο των καταστάσεων στις οποίες μπορεί να βρεθεί το μοντέλο και το σύνολο των επιτρεπτών μεταβάσεων μεταξύ των καταστάσεων σε κάθε χρονική στιγμή. Αυτό που γίνεται αντιληπτό είναι ότι με αυτόν τον τρόπο απεικόνισης είναι δυνατή η γραφική αναπαράσταση μη ομογενών HMMs, σε αντίθεση με τον πρώτο τρόπο απεικόνισης.



Εικόνα 6. Ένα 4-καταστάσεων HMM με τις πιθανές μεταβάσεις του πάνω από 3 χρονικά βήματα

Ένας τρίτος τρόπος απεικόνισης, όπως φαίνεται στην Εικόνα 7, απεικονίζει τις στατιστικές εξαρτήσεις που ισχύουν μεταξύ των τυχαίων μεταβλητών του HMM και όχι την συνολική εικόνα της τοπολογίας μιας Hidden Markov αλυσίδας. Οι στατιστικές εξαρτήσεις, είναι πανομοιότυπες με εκείνες που εκφράζονται στους ορισμούς του HMM όπως διατυπώθηκαν στην προηγούμενη ενότητα.



Εικόνα 7. Απεικόνιση ενός HMM ως μοντέλο γραφήματος. Οι κόμβοι του γραφήματος αντιπροσωπεύουν τυχαίες μεταβλητές (όχι καταστάσεις ή μεταβάσεις καταστάσεων), και οι άκρες κωδικοποιεί όλες τις υπό όρους ιδιότητες ανεξαρτησία ενός HMM

1.6 Τα τρία βασικά προβλήματα στην θεωρία των HMMs

Δεδομένου του ορισμού ενός τυπικού HMM όπως παρουσιάστηκε στις προηγούμενες ενότητες, δημιουργήθηκαν 3 βασικά προβλήματα τα οποία πρέπει να επιλυθούν για να μπορεί ένα HMM μοντέλο να επιλύσει ένα πραγματικό πρόβλημα [3, 4].

Πρόβλημα 1. Δεδομένου μιας ακολουθίας παρατηρήσεων $x = (x_1, x_2, \dots, x_L)$ και ενός καθορισμένου μοντέλου θ , πως μπορούμε να υπολογίσουμε την συνολική πιθανότητα μια ακολουθία x να έχει εμφανιστεί από αυτό το μοντέλο; Δηλαδή να υπολογίσουμε την ποσότητα $P(x/\theta)$.

Το συγκεκριμένο πρόβλημα είναι γνωστό ως πρόβλημα εκτίμησης (Evaluation problem) για το οποίο μπορούμε για παράδειγμα να λάβουμε υπόψη την περίπτωση κατά την οποία προσπαθούμε να επιλέξουμε ανάμεσα σε πολλά ανταγωνιστικά μοντέλα, το μοντέλο που ταιριάζει καλύτερα με τις παρατηρούμενες ακολουθίες.

Πρόβλημα 2. Δεδομένου μιας ακολουθίας παρατηρήσεων $x = (x_1, x_2, \dots, x_L)$ και ενός καθορισμένου μοντέλου θ πως μπορούμε να υπολογίσουμε το πιθανότερο μονοπάτι π , δηλαδή την ακολουθία καταστάσεων, με την καλύτερη πιθανότητα.

Στο πρόβλημα αυτό που είναι γνωστό ως πρόβλημα Αποκωδικοποίησης (decoding problem), αναζητάμε την βέλτιστη ακολουθία καταστάσεων του μοντέλου από την οποία προέκυψε η ακολουθία των παρατηρήσεων.

Στο παράδειγμα με τον παίκτη στο καζίνο, το αντίστοιχο πρόβλημα ορίζεται από το ακόλουθο ερώτημα: για μια δεδομένη ακολουθία ρίψεων ζαριού, είναι δυνατόν να βρούμε πότε ο παίκτης κλέβει ή εναλλακτικά, σε ποιες θέσεις της ακολουθίας ρίψεων ο παίκτης χρησιμοποίησε το κανονικό ζάρι και σε ποιες το «πειραγμένο». Υπάρχουν αρκετά μονοπάτια μέσα από τις κρυμμένες καταστάσεις (Αμερόληπτο, Μεροληπτικό) που οδηγούν στην δεδομένη αλληλουχία, αλλά δεν έχουν την ίδια πιθανότητα.

Πρόβλημα 3. Πως μπορούμε να τροποποιήσουμε τις παραμέτρους του μοντέλου θ , με νέα δεδομένα, ώστε να προκύψουν καλύτερα μοντέλα;

Στο πρόβλημα αυτό, γνωστό και ως πρόβλημα εκπαίδευσης (Training Problem) αναζητάμε τις κατάλληλες τιμές για τις παραμέτρους του μοντέλου, ώστε αυτό να περιγράφει με καλύτερη ακρίβεια το σύστημα που παρήγαγε την δεδομένη ακολουθία παρατηρήσεων, οπότε αναζητούμε τον υπολογισμό των παραμέτρων έτσι ώστε να μεγιστοποιείται το $P(x, \theta)$. Η ακολουθία παρατήρησης που χρησιμοποιείται για να ρυθμίσει τις παραμέτρους του μοντέλου ονομάζεται ακολουθία εκπαίδευσης, δεδομένου ότι χρησιμοποιείται για να "εκπαιδεύσει" το HMM. Το πρόβλημα εκπαίδευσης θεωρείται κρίσιμο για τις περισσότερες εφαρμογές των HMMs, αφού μας επιτρέπει να προσαρμόζουμε βέλτιστα τις παραμέτρους του μοντέλου με βάση τα παρατηρούμενα δεδομένα εκπαίδευσης με σκοπό να δημιουργήσουν καλύτερα μοντέλα για τα πραγματικά προβλήματα.

Για να λύσουμε οποιοδήποτε φυσικό πρόβλημα όπως αυτά που έχουν αναφερθεί και στα παραδείγματα θα πρέπει να εργαστούμε πάνω στα συγκεκριμένα προβλήματα. Για παράδειγμα στο πρόβλημα των πρωτεϊνών, η πρώτη εργασία είναι να οικοδομήσουμε μεμονωμένα μοντέλα πρωτεϊνών. Αυτή η εργασία γίνεται χρησιμοποιώντας μεθόδους επίλυσης στο πρόβλημά 3 για την εκτίμηση των παραμέτρων του μοντέλου για κάθε μοντέλο διαμεμβρανικών πρωτεϊνών. Στη συνέχεια για να αναπτυχθεί μια κατανόηση της φυσικής έννοιας των καταστάσεων του μοντέλου, χρησιμοποιούμε τη λύση στο πρόβλημα 2 για να τμηματοποιήσουμε κάθε μία από τις ακολουθίες εκπαίδευσης των διαμεμβρανικών πρωτεϊνών σε καταστάσεις, και στη συνέχεια να μελετήσουμε τις ιδιότητες τους σύμφωνα με τις παρατηρήσεις που παρουσιάζονται σε κάθε κατάσταση. Ο στόχος εδώ είναι να γίνουν βελτιώσεις στο μοντέλο (π.χ., περισσότερες καταστάσεις, διαφορετικό μέγεθος συμβόλων, κλπ) έτσι ώστε να βελτιστοποιηθεί η ικανότητα του μοντέλου διαμεμβρανικών πρωτεϊνών. Τέλος, εφόσον το μοντέλο διαμεμβρανικών πρωτεϊνών έχει σχεδιαστεί, βελτιστοποιηθεί και εκπαιδευτεί σε βάθος, η αναγνώριση μιας άγνωστης ακολουθίας γίνεται με την λύση στο πρόβλημά 1. Συγκεκριμένα, ορίζοντας ένα σκορ για κάθε μοντέλο διαμεμβρανικής πρωτεΐνης, με βάση τις δεδομένες παρατηρούμενες ακολουθίες της δοκιμής, επιλέγει τη πρωτεΐνη της οποίας το μοντέλο είχε το υψηλότερο σκορ [π.χ. μέγιστη Πιθανοφάνεια] [3].

1.7 Υπολογισμός πιθανοφάνειας

Όπως αναφέραμε παραπάνω θα επιθυμούσαμε να υπολογίσουμε την από κοινού πιθανότητα μιας ακολουθίας x και του μονοπατιού π . Ο πιο απλός τρόπος για να γίνει αυτό είναι απαριθμώντας κάθε δυνατή ακολουθία καταστάσεων μήκους L (ο αριθμός των παρατηρήσεων) από τις οποίες μπορεί να διέλθει το μοντέλο, μέσω της σχέσης

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

Η ερμηνεία του υπολογισμού στην παραπάνω εξίσωση είναι η ακόλουθη. Αρχικά (χρονική στιγμή $i = 1$) είμαστε στην κατάσταση π_1 με πιθανότητα $a_{B\pi_1}$, και θα δημιουργήσει το σύμβολο x_1 (σε αυτή την κατάσταση) με πιθανότητα $e_{\pi_1}(x_1)$. Στο επόμενο στάδιο αλλάζει η χρονική στιγμή i σε $i + 1$ ($i = 2$) και κάνουμε μια μετάβαση στην κατάσταση π_2 , από την κατάσταση π_1 με πιθανότητα $a_{\pi_1 \pi_2}$, και θα δημιουργήσει το σύμβολο x_2 με πιθανότητα $e_{\pi_2}(x_2)$. Η διαδικασία συνεχίζεται με αυτόν τον τρόπο μέχρι να γίνει η μετάβαση (σε χρόνο L) από την κατάσταση q_{i-1} στην κατάσταση π_L με πιθανότητα $a_{\pi_{L-1} \pi_L}$ και παράγει το σύμβολο x_L με πιθανότητα $e_{\pi_L}(x_L)$.

Η μέθοδος αυτή απαιτεί $O(NL^L)$ υπολογισμούς και $O(NL)$ αποθηκευτικό χώρο, γεγονός που την καθιστά μη αποδοτική, ακόμη και για μικρές τιμές, διότι αυξάνεται εκθετικά, όσο το μήκος της ακολουθίας αυξάνεται. Όπως αναφέραμε και στα παραδείγματα, αν έχουμε ένα μοντέλο με 70 καταστάσεις και μια ακολουθία μήκους 200 συμβόλων, θα προκύψει ο αριθμός των μονοπατιών 70^{200} πιθανών μονοπατιών. Είναι σαφές ότι είναι απαραίτητη μια πιο αποτελεσματική διαδικασία για την επίλυση του προβλήματος 1 και αναπτύχθηκε μια βέλτιστη λύση με χρήση δυναμικού προγραμματισμού για την εκτίμηση της $P(x|\theta)$ η οποία βελτιώνει και τα υπολογιστικά προβλήματα που είχαν δημιουργηθεί. Η μέθοδος αυτή καλείται forward-backward και αναλύεται συνοπτικά παρακάτω [3]. Για την εύρεση της χρησιμοποιείται είτε ο Forward αλγόριθμος, είτε ο Backward αλγόριθμος, στους οποίους αναφερόμαστε ακολούθως, είτε συνδυασμός των δύο.

1.7.1 Αλγόριθμος Forward

Από τους πιο γνωστούς αλγόριθμους που έχουν προταθεί για την βέλτιστη επίλυση του προβλήματος είναι ο Forward [3, 4].

Αλγόριθμος Forward

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

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

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

Ο Forward αλγόριθμος, ορίζει έναν πίνακα με διαστάσεις $N(L+1)$, όπου N είναι ο αριθμός των καταστάσεων του μοντέλου και L το μήκος της ακολουθίας και χρησιμοποιεί μια προσωρινή μεταβλητή $f_k(i)$ για κάθε θέση i και κατάσταση k της ακολουθίας. Η τιμή της μεταβλητής τη κάθε χρονική στιγμή αντιστοιχεί στην από κοινού πιθανότητα της ακολουθίας μέχρι το κατάλοιπο i και του μονοπατί που αντιστοιχεί στην κατάσταση k .

Αρχικά ο πίνακας με τις από κοινού πιθανότητες αρχικοποιείται για κάθε κατάσταση k και την αρχική παρατήρηση. Στο επαναληπτικό βήμα, που είναι και το πιο σημαντικό βήμα του αλγόριθμου, υπολογίζεται η από κοινού πιθανότητα για κάθε παρατηρούμενη τιμή της ακολουθίας και το οποίο απεικονίζεται και στην Εικόνα 8. Η κατάσταση $k1$ μπορεί να υπολογιστεί σε χρόνο $i+1$ (στη προκειμένη περίπτωση 2) με την αντίστοιχη συνεισφορά όλων των προηγούμενων τιμών και πολλαπλασιάζοντας την αθροιζόμενη ποσότητα με την πιθανότητα μετάβασης της αντίστοιχης κατάστασης.

Συνοψίζοντας την ίδια διαδικασία για όλες τις καταστάσεις για την ίδια χρονική στιγμή t έχουμε υπολογίσει την από κοινού πιθανότητα για την χρονική στιγμή i με όλες τις απαραίτητες τιμές των προηγούμενων τιμών. Αντίστοιχα συνεχίζουμε για τις υπόλοιπες χρονικές στιγμές. Στο τελευταίο βήμα αθροίζονται για να προκύψει η τελική πιθανοφάνεια.

Ο αλγόριθμος πλέον απαιτεί NL υπολογισμούς αντί για NL^L και αντίστοιχα αν έχουμε ένα μοντέλο με 70 καταστάσεις και μια ακολουθία μήκους 200 συμβόλων, θα χρειαστεί 200×70 υπολογισμούς.

	Παρατηρούμενη Ακολουθία							
Καταστάσεις	0	x_1	x_2	x_3	x_4	x_L
k1								
k2								
k3								
k4								
...								
...								
k_N								

Εικόνα 8: Απεικόνιση των βημάτων για τον υπολογισμό της από κοινού πιθανότητας για τον αλγόριθμο Forward για ένα μοντέλο με L το μήκος της παρατηρούμενης ακολουθίας και N καταστάσεις. Για τον υπολογισμό της τιμής ενός κελιού απεικονίζεται η συνεισφορά όλων των προηγούμενων κελιών.

1.7.2 Αλγόριθμος Backward

Κατά παρόμοιο τρόπο μπορούμε να θεωρήσουμε και τον αλγόριθμο Backward με τη μόνη διαφορά ότι διατρέχει την παρατηρούμενη ακολουθία ξεκινώντας από το τέλος [3, 4]. Δημιουργείται ένας πίνακας με διαστάσεις $N(L+1)$, όπου N ο αριθμός των καταστάσεων του μοντέλου και L το μήκος της ακολουθίας και πλέον μια προσωρινή μεταβλητή $b_k(i)$ για κάθε θέση i και κατάσταση k της ακολουθίας. Η τιμή της μεταβλητής τη κάθε χρονική στιγμή αντιστοιχεί στην από κοινού πιθανότητα της ακολουθίας από το κατάλοιπο $i+1$ έως το τέλος και του μονοπατιού που αντιστοιχεί στην κατάσταση k .

Αλγόριθμος Backward

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

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

$$P(x|\theta) = \sum_k e_k(x_1) b_k(1) a_{Bk}$$

Αντίστοιχα αν δεν υπάρχουν καταστάσεις λήξεως, κατά την αρχικοποίηση, οι αντίστοιχες πιθανότητες ορίζονται με την τιμή 1. Ο αλγόριθμος απαιτεί NT υπολογισμούς.

1.8 Αποκωδικοποίηση

Σε αντίθεση με το Πρόβλημα 1 για το οποίο μπορεί να δοθεί μια ακριβής λύση, υπάρχουν διάφοροι τρόποι για την επίλυση του δεύτερου προβλήματος, που αναζητάμε την βέλτιστη ακολουθία καταστάσεων του μοντέλου από την οποία προέκυψε η ακολουθία των παρατηρήσεων γιατί υπάρχουν αρκετά πιθανά κριτήρια βελτιστοποίησης και γι' αυτό έχουν αναπτυχθεί διάφορες μέθοδοι, που διαφέρουν ως προς το κριτήριο βελτιστοποίησης. Μια εκ των μεθόδων αυτών που αναπτύχθηκε αντίστοιχα με τους προηγούμενους αλγόριθμους αποτελεί και ο αλγόριθμος Viterbi.

1.8.1 Αλγόριθμος Viterbi

Ο αλγόριθμος Viterbi είναι ένας αλγόριθμος δυναμικού προγραμματισμού που μας επιτρέπει να υπολογίσουμε την πιο πιθανή διαδρομή. Η αρχή του είναι παρόμοια με τα προγράμματα που χρησιμοποιούνται για την στοίχιση ακολουθιών (π.χ. Needleman-Wunsch).

Αλγόριθμος Viterbi

$$\forall k \neq B, i = 0: V_B(0) = 1, V_k(0) = 0$$

$$\forall 1 \leq i \leq L: V_i(i) = e_i(x_i) \max_k \{V_k(i-1)a_{ki}\}$$

$$P(x, \pi^{\max} | \theta) = \max_k \{V_k(L)a_{kE}\}$$

Για να βρούμε τη βέλτιστη ακολουθία καταστάσεων, $k = (k_1, k_2, \dots, k_t)$, για τη δεδομένη ακολουθία παρατήρησης $x = (x_1, x_2, \dots, x_L)$, πρέπει να καθορίσουμε την ποσότητα $V_i(i) = e_i(x_i) \max_k \{V_k(i-1)a_{ki}\}$ που αντιστοιχεί στην μέγιστη πιθανότητα εμφάνισης μιας ακολουθίας παρατηρήσεων, που παράγεται κατά μήκος ενός μονοπατιού καταστάσεων μέχρι την χρονική στιγμή t και το οποίο τελειώνει στην κατάσταση i .

Ο αλγόριθμος στηρίζεται στην ιδέα, ότι η βέλτιστη ακολουθία καταστάσεων, μέχρι την θέση i στην ακολουθία x , είναι μέρος της βέλτιστης ακολουθίας που καλύπτει όλη την ακολουθία των παρατηρήσεων. Ο αλγόριθμος Viterbi, είναι παρόμοιος με τον αλγόριθμο Forward με μια μικρή διαφορά. Η διαφορά αυτή εντοπίζεται στην αντικατάσταση των αθροισμάτων από μεγιστοποιήσεις. Για να εφαρμόσουμε αυτή τη λύση, ορίζουμε με π^{\max} το μονοπάτι με τη μεγαλύτερη πιθανότητα η οποία συμβολίζεται με $P(x, \pi^{\max} | \theta)$ και ισχύει ότι $P(x, \pi^{\max} | \theta) \leq P(x | \theta)$. Η μεγαλύτερη πιθανότητα που λαμβάνεται για το σύμβολο στην τελευταία θέση είναι η πιθανότητα της πιο πιθανής διαδρομής. Αυτή η διαδρομή μπορεί να ανακτηθεί με αναδρομή (back-tracking).

$$q_t = V_{i+1}(P(x, \pi^{\max} | \theta)), \quad i = i-1, i-2, \dots, 1$$

Ο αλγόριθμος εκτελείται σε χρόνο $O(NL^2)$ και απαιτεί χώρο $O(NL)$.

1.9 Εκτίμηση Παραμέτρων

Το τρίτο, και μακράν το πιο δύσκολο, πρόβλημα του HMM είναι να καθοριστεί μια μέθοδος για την εκτίμηση των παραμέτρων του μοντέλου θ^{ML} για να μεγιστοποιηθεί η πιθανότητα της ακολουθίας παρατήρησης που δίνεται από το μοντέλο. Για την εκτίμηση των παραμέτρων ενός στατιστικού-πιθανοθεωρητικού μοντέλου συνήθως εφαρμόζεται η μέθοδος της Μέγιστης Πιθανοφάνειας (Maximum Likelihood). Αρχικά ορίζονται οι τιμές των παραμέτρων του μοντέλου θ^{ML} ως Εκτιμητές Μέγιστης Πιθανοφάνειας (ΕΜΠ) οι οποίες μεγιστοποιούν τη συνάρτηση πιθανοφάνειας η οποία είναι η από κοινού συνάρτηση κατανομής όλων των παρατηρήσεων, δεδομένου ενός συνόλου ακολουθιών παρατηρήσεων και των παραμέτρων του μοντέλου, θεωρώντας τις παραμέτρους ως τυχαίες μεταβλητές. Οπότε ισχύει :

$$\theta^{ML} = \arg \max_{\theta} P(x | \theta)$$

Όπου για λόγους υπολογιστικής απλότητας εφαρμόζουμε τον λογάριθμο της $l(x|\theta)$, ο οποίος μεγιστοποιείται στην ίδια περιοχή με αυτή.

$$l(x|\theta) = \log P(x|\theta)$$

Για την σωστή και γρήγορη σύγκλιση του αλγορίθμου, αλλά και για την εύρεση μιας προσεγγιστικής λύσης που να πλησιάζει το ολικό μέγιστο, είναι σημαντικό ο αλγόριθμος να ξεκινήσει από ένα καλό, αρχικό μοντέλο. Γίνεται αντιληπτό, ότι για τις παραμέτρους των A και π συνόλων, ακόμα και όταν τα στοιχεία τους παίρνουν τυχαίες αρχικές τιμές ή ίδιες μεταξύ τους αρχικές τιμές, οι προσεγγίσεις που δίνει ο αλγόριθμος πλησιάζουν αυτές που παίρνουμε, όταν χρησιμοποιούμε ένα καλό, αρχικό μοντέλο [3].

Όπως έχει αναφερθεί στην προκειμένη περίπτωση οι ακολουθίες είναι πρωτεΐνες και τα παρατηρηθέντα σύμβολα τα αμινοξικά κατάλοιπα. Οπότε από το σύνολο των ακολουθιών που συμμετέχουν για εκπαίδευση, τις θεωρούμε ανεξάρτητες και το γινόμενο των πιθανοφανειών τους θα αποτελεί και τη συνολική πιθανοφάνεια.

Στην πιο απλή περίπτωση, αρχικά υιοθετείται ως αρχικό μοντέλο, ένα τυχαίο μοντέλο ή ένα γνωστό μοντέλο που ταιριάζει με το σύνολο εκπαίδευσης που διαθέτουμε. Στην συνέχεια, χρησιμοποιώντας τον αλγόριθμο Viterbi, που αναφέρθηκε παραπάνω, γίνεται καταχώριση του συνόλου των παρατηρήσεων του συνόλου εκπαίδευσης, στις καταστάσεις του παρόντος μοντέλου (δηλαδή, τρέχει ο Viterbi αλγόριθμος για κάθε ακολουθία παρατηρήσεων του συνόλου εκπαίδευσης, βρίσκει την βέλτιστη ακολουθία καταστάσεων για κάθε μια και τέλος, εξάγονται στατιστικά στοιχεία για κάθε μια από τις M δυνατές τιμές παρατηρήσεων, για κάθε κατάσταση του μοντέλου). Από αυτήν την διαδικασία παίρνουμε μια εκτίμηση για την μέγιστη πιθανότητα εμφάνισης κάθε μιας από τις M παρατηρήσεις, ανά κατάσταση και δεν έχουμε παρά να καταμετρήσουμε πόσες φορές παρατηρήθηκε μια συγκεκριμένη μετάβαση από κάθε κατάσταση, και πόσες φορές ένα αμινοξύ εμφανίστηκε σε κάθε κατάσταση. Επομένως σύμφωνα με την παραπάνω περίπτωση οι ΕΜΠ, για τις πιθανότητες μετάβασης θα είναι:

$$a_{kl} = \frac{A_{kl}}{\sum_l A_{kl}}$$

και για τις πιθανότητες εμφάνισης συμβόλων,

$$e_k(b) = \frac{E_k(b)}{\sum_b E_k(b)}$$

Στη γενικότερη και πιο συνηθισμένη περίπτωση, που δεν γνωρίζουμε την ακολουθία των καταστάσεων του μοντέλου το πρόβλημα είναι πιο σύνθετο και επιλύονται με την χρήση άλλων μεθόδων όπως τη μέθοδο Baum-Welch (ειδική περίπτωση του αλγορίθμου EM (Expectation-Maximization), ή με τη χρήση μεθόδων Gradient-Descent.

1.9.1 Αλγόριθμος Baum-Welch

Σε αυτή την ενότητα θα συζητήσουμε μια επαναληπτική διαδικασία, η οποία βασίζεται κατά κύριο λόγο στο κλασικό έργο του Baum και των συναδέλφων του, για την επιλογή των παραμέτρων του μοντέλου και είναι γνωστή ως ο αλγόριθμος Baum-Welch [2, 8, 9] . Ο αλγόριθμος, ξεκινά από ένα αρχικό, υποθετικό μοντέλο θ , εκτελείται επαναληπτικά, δίνοντας, κάθε φορά, μια εκτίμηση για τις παραμέτρους του μοντέλου, δηλαδή ένα νέο θ' και σταματά,

όταν οι εκτιμήσεις των παραμέτρων συγκλίνουν ικανοποιητικά. Οι παράμετροι π , A , B του μοντέλου εκτιμώνται ως εξής:

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

Για τις μεταβάσεις έχουμε

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

και αντίστοιχα για τις πιθανότητες εκπομπής

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

όπου $f_k(i)$, $b_k(i)$ οι μεταβλητές που υπολογίζονται από τους αλγόριθμους forward και backward, αντίστοιχα και $i, j \in [1, N]$.

και ορίζοντας την σχέση

$$Q(\theta|\theta') = \sum_{k=1} \sum_b E_k(b) \log e_k(b) + \sum_{k=0} \sum_{l=1} A_{kl} \log a_{kl}$$

Όπως αποτυπώνεται παραπάνω ο αλγόριθμος Baum–Welch σε πρώτο βήμα εκτέλεσης υπολογίζεται το Q όπου υπολογίζει τις ποσότητες $f_k(i)$, $b_k(i)$ από τους αλγόριθμους forward και backward αντίστοιχα και στη συνέχεια υπολογίζονται οι τιμές A_{kl} και $E_k(b)$. Σε δεύτερο βήμα υπολογίζονται ξανά οι ΕΜΠ και η πιθανοφάνεια του μοντέλου μέχρι οι τιμές στην πιθανοφάνεια και αντίστοιχα στην συνάρτηση Q είναι μικρότερες από μια προκαθορισμένη τιμή (threshold). Ο αλγόριθμος εκπαίδευσης του μοντέλου, όπως προαναφέρθηκε, είναι προσεγγιστικός και συγκλίνει, αν όχι στο ολικό μέγιστο της πιθανότητας $P(x|\theta)$, στο τοπικό μέγιστο της $P(x|\theta)$.

1.9.2 Μέθοδοι Gradient-descent

Έχει αποδειχθεί, ότι η μεγιστοποίηση της πιθανότητας με τον αλγόριθμο Baum-Welch μπορεί να γίνει ισοδύναμα με μία μέθοδο Gradient-descent [10, 11] έχοντας όμως δύο πλεονεκτήματα, δεν υπάρχει περίπτωση να παρουσιάσει μηδενική πιθανότητα και η ενημέρωση των παραμέτρων γίνεται χωρίς να χρειαστεί να υπολογιστούν σε ενδιάμεσο βήμα οι βοηθητικές μεταβλητές.

Οι παράμετροι σύμφωνα με την μέθοδο Krogh και Riss [12] για τις πιθανότητες υπολογίζονται από την σχέση

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

Οπότε για τις πιθανότητες μετάβασης έχουμε

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

Και για τις πιθανότητες εκπομπής

$$e_k^{(t+1)} = \frac{e_k^{(t)} \exp(-\eta[E_k - e_k(b) \sum_b E_k(b)])}{\sum_b e_k^{(t)} \exp(-\eta[E_k - e_k(b) \sum_b E_k(b)])}$$

2. Class Hidden Markov Model

Στη βιολογία, έχουν προκύψει περιπτώσεις στις οποίες ένα βιολογικό πρόβλημα χρειάζεται ένα μοντέλο, το οποίο να χρησιμοποιεί διαφορετικές καταστάσεις για διαφορετικές μιας πρωτεϊνικής αλυσίδας. Ένα τέτοιο παράδειγμα αποτελεί η πρόγνωση διαμεμβρανικών περιοχών για την οποία χρειάζεται το μοντέλο να χρησιμοποιεί διαφορετικές καταστάσεις για κάθε περιοχή της πρωτεΐνης (εξωκυττάρια, διαμεμβρανική, κυτταροπλασματική) με αποτέλεσμα να αποτυπώσει βέλτιστα τη προ-υπάρχουσα βιολογική γνώση. Για την βελτιστοποίηση του ποσοστού σφάλματος του μοντέλου προτάθηκε ένας νέος τύπος HMM για επιστημονικά δεδομένα που ονομάζεται Class Hidden Markov Model (CHMM). Αποτελεί μια επέκταση ενός κλασσικού HMM, το οποίο επιτρέπει κάθε κατάσταση να περιλαμβάνει επίσης και μια κατανομή πιθανοτήτων για μια σήμανση [13]. Στην περίπτωση ενός CHMM, ο στόχος είναι να προβλεφθούν οι σημάνσεις που συνδέονται με τις ακολουθίες και να βελτιστοποιηθεί η πιθανότητα της επισήμανσης, $P(\text{σημάνσεις}|\text{μοντέλο}, \text{παρατηρήσεις})$, παρά την πιθανότητα των παρατηρήσεων $P(\text{παρατηρήσεις}|\text{μοντέλο})$, η οποία βελτιστοποιείται από ένα κλασσικό HMM. Η εκπαίδευση ενός κλασσικού μοντέλου HMM χαρακτηρίζεται ως «μέθοδος χωρίς επίβλεψη» (unsupervised learning) ενώ η νέα μέθοδος εκπαίδευσης χαρακτηρίζεται «μέθοδος με επίβλεψη» (supervised learning). Η επέκταση ενός HMM σε CHMM είναι πολύ απλή και στη συνέχεια του κεφαλαίου θα παρουσιάσουμε τις τροποποιήσεις των αλγόριθμων που παρουσιάστηκαν στο προηγούμενο κεφάλαιο.

2.1 Ορισμός των CHMMs

Όπως αναφέραμε κάθε ακολουθία συμβόλων

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

περιλαμβάνει και μια ακολουθία επισημάνσεων (labels)

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

Στο πρόβλημα των διαμεμβρανικών τμημάτων, οι σημάνσεις θα μπορούσαν να είναι μια για τα διαμεμβρανικά τμήματα (M), μια για την κυτταροπλασματική περιοχή (I) και μια για την εξωκυττάρια περιοχή (O) όπως αποτυπώνονται και στην Εικόνα 9. Ένα άλλο αντίστοιχο βιολογικό πρόβλημα είναι η μοντελοποίηση της πρωτεΐνης, όπου τα δεδομένα είναι η ακολουθία των αμινοξέων και οι σημάνσεις/ετικέτες θα μπορούσαν να είναι, για παράδειγμα, ο τύπος της τρισδιάστατης δομής (άλφα έλικα, βήτα βαρέλι, κ.α.).

Ένα CHMM μοντελοποιεί παρατηρήσεις με τη σχετική σήμανση κατηγορίας. Η βάση είναι ένα πρότυπο HMM με πιθανότητες εκπομπής συμβόλων παρατήρησης για κάθε κατάσταση και πιθανότητες για μεταβάσεις μεταξύ των καταστάσεων. Όπως γίνεται αντιληπτό από τα παραπάνω θα πρέπει να ορίσουμε πλέον μια κατανομή για την πιθανότητα ταύτισης μιας κατάστασης με μια δεδομένη σήμανση όπου ουσιαστικά ομαδοποιούνται οι καταστάσεις ανάλογα την βιολογική τους σημασία. Αυτές οι παράμετροι πιθανότητας που αναλογούν στην κατηγορία ορίζονται ως δ , και η πιθανότητα η κατάσταση k να έχει την σήμανση x είναι $\delta_k(x)$.

			Παρατηρούμενη Ακολουθία							
			I	I	M	M	O	O
Καταστάσεις	Σημάνσεις	0	x_1	x_2	x_3	x_4	x_{L-1}	x_L
k_1	I		$f = 0$							
k_2	I									
k_3	M		$f = 0$							
k_4	M									
...	...		$f = 0$							
...	...									
k_{N-1}	O		$f = 0$							
k_N	O									

Εικόνα 9: Απεικόνιση ενός μοντέλου CHMM για διαμεμβρανικές πρωτεΐνες όπου οι σημάνσεις είναι μια για τα διαμεμβρανικά τμήματα (M), μια για την κυτταροπλασματική περιοχή (I) και μια για την εξωκυττάρια (O)

Λαμβάνοντας υπόψη ένα μοντέλο (θ, x) μπορεί κανείς να υπολογίσει τις συνήθεις ποσότητες για το συγκεκριμένο HMM που χαρακτηρίζεται ως θ . Για παράδειγμα, η πιθανότητα μιας ακολουθίας συμβόλων παρατήρησης y , $P(x|\theta)$, υπολογίζεται από το αλγόριθμο Forward, και το πιο πιθανό μονοπάτι μέσω του μοντέλου μπορεί να βρεθεί από τον αλγόριθμο Viterbi. Δεδομένου ότι κάποιος ενδιαφέρεται για στην εύρεση των αντίστοιχων ποσοτήτων λαμβάνοντας υπόψη τις σημάνσεις για μια δεδομένη ακολουθία θα πρέπει να εφαρμόσει κάποιες τροποποιήσεις στους γνωστούς αλγόριθμους οι οποίοι θα παρουσιαστούν στη συνέχεια του κεφαλαίου.

2.2 Υπολογισμός πιθανοφάνειας

Όπως αναφέραμε σε ένα CHMM πρέπει να λαμβάνεται υπόψη και η ακολουθία των σημάνσεων όπου η πιθανότητα πλέον υπολογίζεται μόνο για τα μονοπάτια Π_y , των καταστάσεων τα οποία συνάδουν με τις παρατηρηθείσες σημάνσεις. Επομένως, ορίζεται ως αντικειμενική συνάρτηση η από κοινού πιθανότητα $P(x, y|\theta)$ των ακολουθιών x με τις σημάνσεις y , δεδομένου του μοντέλου θ , ως εξής:

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

Για τον υπολογισμό της από κοινού πιθανότητας χρησιμοποιούμε τους γνωστούς αλγόριθμους που παρουσιάστηκαν στο προηγούμενο κεφάλαιο με μικρές τροποποιήσεις. Η διαφορά είναι η χρήση μια δίτιμης συνάρτησης $(0,1)$ η οποία παίρνει την τιμή 1 αν η κατάσταση συμφωνεί με την σήμανση και 0 αν δεν συμφωνεί. Οπότε προκύπτει ο τροποποιημένος αλγόριθμος forward ως εξής:

Αλγόριθμος Forward

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

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

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

Και ο τροποποιημένος αλγόριθμος backward ως εξής:

Αλγόριθμος Backward

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

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

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

Όπου με την προσθήκη της δίτιμης συνάρτησης μηδενίζονται οι περιοχές των πινάκων Forward και Backward, όπου δεν υπάρχει συμφωνία μεταξύ καταστάσεων και σημάνσεων.

2.3 Εκτίμηση Παραμέτρων

Η εκπαίδευση του μοντέλου γίνεται δεδομένο ότι είναι διαθέσιμος ένας αριθμός ακολουθιών x με τις σχετικές σημάνσεις y .

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

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

και μπορούμε να επιλέξουμε να εκπαιδεύσουμε το μοντέλο σύμφωνα με το κριτήριο της Μέγιστης Πιθανοφάνειας (ML)

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

χρησιμοποιώντας τους αλγόριθμους που παρουσιάσαμε προηγουμένως με μικρές τροποποιήσεις αντίστοιχα. Για τις μεταβάσεις έχουμε

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

και αντίστοιχα για τις πιθανότητες εκπομπής

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

όπου $f_k(i)$, $b_k(i)$ οι μεταβλητές που υπολογίζονται από τους τροποποιημένους αλγόριθμους forward και backward, αντίστοιχα και $i, j \in [1, N]$. Η εκπαίδευση με ML θα μπορούσε να

πραγματοποιηθεί με τη χρήση τυποποιημένων τεχνικών όπως είναι ο αλγόριθμος Baum-Welch ή Gradient descent.

Δεσμευμένη Μέγιστη Πιθανοφάνεια

Σε ένα CHMM, ο στόχος είναι να προβλεφθούν οι σημάνσεις που συνδέονται με το x , και αντί της Μέγιστης Πιθανοφάνειας, αναζητούμε πλέον να μεγιστοποιήσουμε την πιθανότητα των σημάνσεων, με μια νέα μέθοδο η οποία ονομάζεται Δεσμευμένη Μέγιστη πιθανοφάνεια (CML) [13].

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

Δυστυχώς, ο αλγόριθμος Baum-Welch δεν μπορεί να εφαρμοστεί για την εκτίμηση της CML γιατί θα δώσει αρνητικές τιμές για τις παραμέτρους. Μπορούμε όμως να χρησιμοποιήσουμε με τη μέθοδο Gradient Descent, δουλεύοντας με τον ίδιο τρόπο όπως συζητήθηκε στο προηγούμενο κεφάλαιο. Για τον υπολογισμό των gradients, θα χρησιμοποιηθεί ο αρνητικός λογάριθμος της δεσμευμένης πιθανοφάνειας που ορίζεται ως εξής :

$$l = -\log P(y|x, \theta) = l_c - l_f$$

$$l_c = -\log P(x, y|\theta)$$

$$l_f = -\log P(x|\theta)$$

Με τους δείκτες f και c εκφράζεται αντίστοιχα η ποσότητα για την περίπτωση όπου οι σημάνσεις δεν λαμβάνονται υπόψη και την περίπτωση που λαμβάνονται υπόψη. Τα gradients αποτελούν τις αναμενόμενες τιμές και τις μερικές παραγώγους των παραμέτρων του μοντέλου και υπολογίζονται από τις σχέσεις :

$$\frac{\partial l}{\partial a_{kl}} = \frac{\partial l_c}{\partial a_{kl}} - \frac{\partial l_f}{\partial a_{kl}} = \frac{A_{kl}^c - A_{kl}^f}{a_{kl}}$$

$$\frac{\partial l}{\partial e_k(b)} = \frac{\partial l_c}{\partial e_k(b)} - \frac{\partial l_f}{\partial e_k(b)} = \frac{E_k^c(b) - E_k^f(b)}{e_k(b)}$$

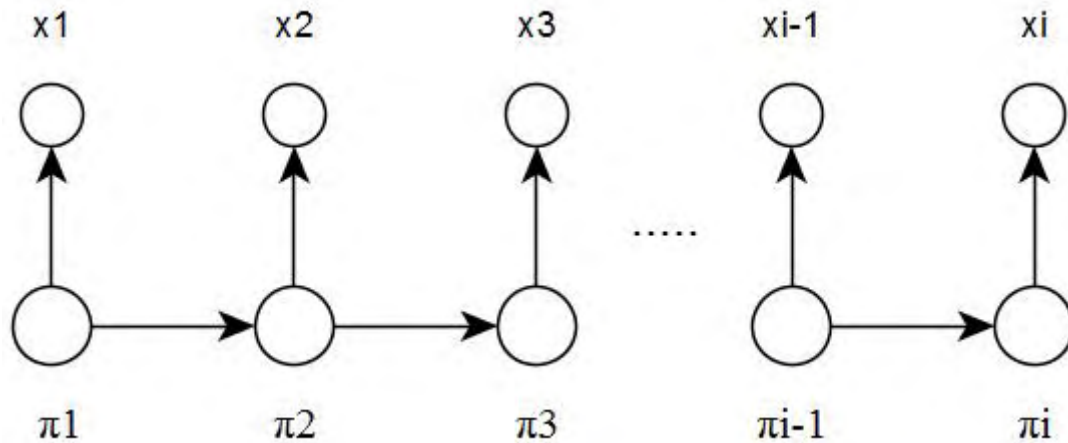
και τελικά οι ανανεωμένες τιμές σε κάθε επανάληψη θα προκύψουν από τις σχέσεις

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

$$e_k^{(t+1)} = \frac{e_k^{(t)} \exp(-\eta[E_k^c(b) - E_k^f(b) - e_k(b) \sum_{b'} (E_{b'}^c(b) - E_{b'}^f(b))])}{\sum_{b'} e_k^{(t)} \exp(-\eta[E_k^c(b) - E_k^f(b) - e_k(b) \sum_{b'} (E_{b'}^c(b) - E_{b'}^f(b))])}$$

3. Επεκτάσεις των HMM

Στα κλασσικά HMM αλλά και σε πολλές επεκτάσεις τους υπάρχει ένα κοινό χαρακτηριστικό όπως διαφαίνεται και στην Εικόνα 10, η τρέχουσα κατάσταση εξαρτάται μόνο από την αμέσως προηγούμενη κατάσταση, αλλά δεν συσχετίζεται με την αμέσως προηγούμενη παρατήρηση.



Εικόνα 10: Απεικόνιση ενός κλασσικού μοντέλου HMM

Όπως αναφέραμε στους ορισμούς ενός HMM συνήθως ένα HMM (x, θ) ορίζεται ως μια ακολουθία καταστάσεων $\pi = (\pi_1, \pi_2, \pi_3, \dots, \pi_{i-1}, \pi_i)$ που έχει τη μαρκοβιανή ιδιότητα και μια ακολουθία παρατηρήσεων $x = (x_1, x_2, x_3, \dots, x_{i-1}, x_i)$, όπου η σύνδεση των παρατηρηθέντων συμβόλων με τις καταστάσεις γίνεται μέσω των πιθανοτήτων εμφάνισης συμβόλων και ισχύει

$$P(x_i | \pi_i, \pi_{i-1}, \dots, \pi_1, x_{i-1}, x_{i-2}, \dots, x_1) = P(x_i | \pi_i)$$

$$P(\pi_i | \pi_{i-1}, \pi_{i-2}, \dots, \pi_1, x_{i-1}, x_{i-2}, \dots, x_1) = P(\pi_i | \pi_{i-1})$$

Για τις πιθανότητες μετάβασης των καταστάσεων του μοντέλου (transitions) ισχύει

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

Και για τις πιθανότητες εκπομπής ή εμφάνισης συμβόλων (emissions) για κάθε κατάσταση i , ισχύει

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

Ωστόσο, η υπόθεση ότι η κατάσταση μετάβασης και το παρατηρούμενο σύμβολο εξαρτάται μόνο από την τρέχουσα κατάσταση δεν ταιριάζει ακριβώς απόλυτα σε όλα τα βιολογικά προβλήματα. Αυτό έχει ως αποτέλεσμα, στις βιολογικές ακολουθίες όταν δεν χρησιμοποιείται πληροφορία από την συσχέτιση των παρατηρήσεων, να χάνεται αρκετή πληροφορία για την πρόβλεψη.

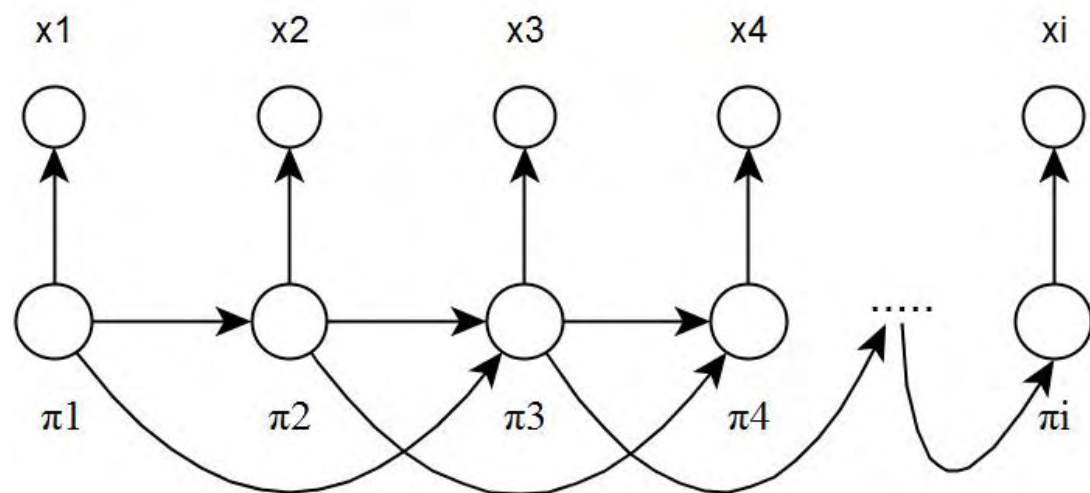
Σε κάποιες περιπτώσεις έγινε προσπάθεια να λάβουν υπόψη τους το συγκεκριμένο θέμα χρησιμοποιώντας την γενίκευση της Μαρκοβιανής ιδιότητας και την χρήση Ανώτερης τάξεως Μαρκοβιανά μοντέλα. Μια πρώτη προσπάθεια έγινε για την πρόβλεψη της υποκυτταρικής θέσης των βακτηριακών πρωτεϊνών χρησιμοποιώντας μεγαλύτερης τάξεως αλυσίδες [14]. Στη συγκεκριμένη μελέτη γίνεται αναφορά ότι με μεγαλύτερης τάξεως αλυσίδες, εκτός από το

πρόβλημα της υπερ-προσαρμογής (>6) και της έλλειψης δεδομένων, προκύπτει και το πρόβλημα της εισαγωγής θορύβου, από μη-σημαντικές μακρινές συσχετίσεις.

Σε αυτήν την ενότητα παρουσιάζεται μια σειρά από επεκτάσεις που έχουν εφαρμοστεί στα HMM και έχουν σκοπό να χρησιμοποιούν πληροφορία από προηγούμενες καταστάσεις ή παρατηρήσεις ή και τις δυο περιπτώσεις.

3.1 High-Order HMM

Μια προσπάθεια αποτέλεσε η μελέτη και χρήση των High-Order HMMs που έχουν βρει εφαρμογή σε συστήματα επεξεργασίας ομιλίας και ανάλυσης βιολογικών ακολουθιών [15, 16]. Μια k^{th} σειρά διαδικασία Markov είναι μια στοχαστική διαδικασία όπου κάθε περίπτωση εξαρτάται από τα προηγούμενα k γεγονότα. Η κύρια διαφορά μεταξύ των κλασικών HMM και High-Order HMM είναι ότι στο κρυφό επίπεδο, η πιθανότητα μετάβασης μιας κατάστασης διέπεται από τη διάταξη της k^{th} τάξης ενός High-Order HMM. Στις αντίστοιχες προσπάθειες χρειάστηκε να επεκταθούν οι γνωστοί αλγόριθμοι των κλασικών HMM που επιτρέπουν την αξιολόγηση των διαφόρων πιθανοτήτων, καθώς και την εκτίμηση των παραμέτρων για την χρήση τους σε ένα High-Order HMM.



Η σχέση πλέον με την εξάρτηση των καταστάσεων ορίζεται ως εξής:

$$P(x_i | \pi_i, \pi_{i-1}, \dots, \pi_1, x_{i-1}, x_{i-2}, \dots, x_1) = P(x_i | \pi_i)$$

$$P(\pi_i | \pi_i, \pi_{i-1}, \dots, \pi_1, x_{i-1}, x_{i-2}, \dots, x_1) = P(\pi_i | \pi_{i-1}, d_{i-i}(\pi_{i-1}))$$

Όπου $d_i(\pi_i)$ αντιστοιχεί στη σχέση που συμπεριλαμβάνει την εξάρτηση με τις k προηγούμενες καταστάσεις.

Για $k=2$ πλέον (2 προηγούμενες καταστάσεις) ισχύει για τις πιθανότητες μετάβασης των καταστάσεων του μοντέλου (transitions)

$$a_{ckl} = P(\pi_i = l | \pi_{i-1} = k, \pi_{i-2} = c)$$

3.2 Partially HMM

Σε μια δεύτερη περίπτωση παρουσιάστηκαν τα Partially Hidden Markov Model (PHMM) στα οποία η διαφορά από ένα κλασσικό HMM έγκειται στο ότι τόσο οι πιθανότητες μετάβασης από τις κρυφές καταστάσεις όσο και οι πιθανότητες εκπομπής συμβόλων εξαρτώνται από προηγούμενες παρατηρήσεις. Η συγκεκριμένη προσέγγιση εφαρμόστηκε για την συμπίεση σε ασπρόμαυρες εικόνες, όπου οι προηγούμενες παρατηρήσεις αντιπροσώπευαν γειτονικά pixels της εικόνας τα οποία ορίζονται ως “context” [17].

Ορίζοντας ως $x^t = x_{i-t}, x_{i-t-1}, \dots, x_{i-1}, x_i$ τις προηγούμενες παρατηρήσεις και τις συναρτήσεις $r_i = Fx(x^t)$ και $s_i = F\pi(x^t)$ ως “context” που αποτελεί το υποσύνολο προηγούμενων παρατηρήσεων.

Η σχέση πλέον με την εξάρτηση των καταστάσεων ορίζεται ως εξής:

$$P(x_i | \pi_i, \pi_{i-1}, \dots, \pi_{i-k}, x_{i-1}, x_{i-2}, \dots, x_{i-k}) = P(x_i | \pi_i, r_{i-1})$$

$$P(\pi_i | \pi_{i-1}, \pi_{i-2}, \dots, \pi_{i-k}, x_{i-1}, x_{i-2}, \dots, x_{i-k}) = P(\pi_i | \pi_{i-1}, s_{i-1})$$

Για $t=2$ πλέον (2 προηγούμενες καταστάσεις) ισχύει για τις πιθανότητες μετάβασης των καταστάσεων του μοντέλου (transitions)

$$a_{ckl} = P(\pi_i = l | \pi_{i-1} = k, \pi_{i-2} = c)$$

και για τις πιθανότητες εκπομπής ή εμφάνισης συμβόλων (emissions) για κάθε κατάσταση j

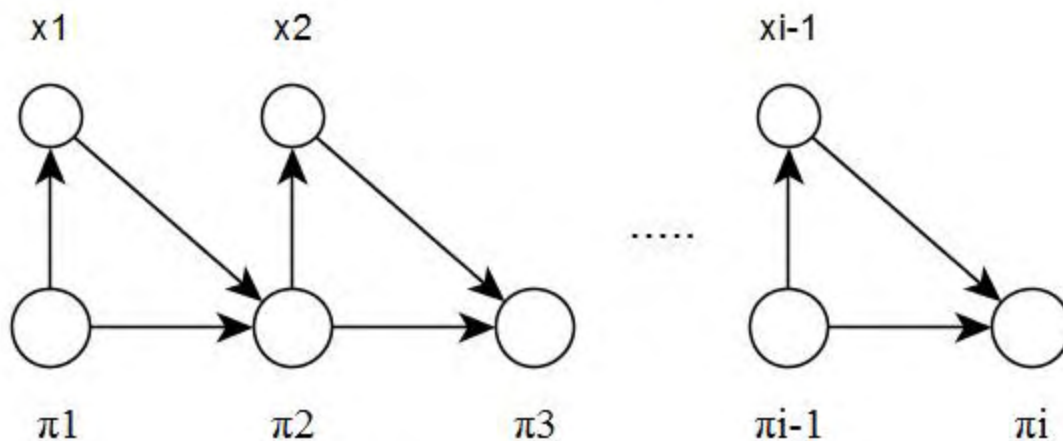
$$e_k(x_i = b') = P(x_i = b | \pi_i = k', x_{i-1} = c)$$

όπου k' συμβολίζει το νέο σύμβολο και c το προηγούμενο σύμβολο

Οι αποδοτικοί αλγόριθμοι που επιτρέπουν την αξιολόγηση των διαφόρων πιθανοτήτων, καθώς και την εκτίμηση των παραμέτρων των κλασσικών HMM επεκτείνονται αντίστοιχα για την χρήση τους στο PHMM.

3.3 HMM with states depending on observations

Μια ακόμη προσπάθεια ήταν η εξάρτηση της κατάστασης με προηγούμενες παρατηρήσεις [18]. Η κύρια διαφορά μεταξύ ενός μοντέλου που μια κατάσταση εξαρτάται από προηγούμενες παρατηρήσεις και ενός κλασσικού HMM, απεικονίζεται στην Εικόνα 12, παρουσιάζοντας την εξάρτηση που έχει μια κατάσταση τη χρονική στιγμή i από την προηγούμενη κατάσταση π_{i-1} από την οποία προήλθε αλλά και από την προηγούμενη παρατήρηση x_{i-1} . Η υλοποίησή τους έγινε με τροποποιήσεις και εφαρμογή των γνωστών αλγόριθμων που αναφέραμε σε προηγούμενα κεφάλαια.



Εικόνα 11: Απεικόνιση ενός μοντέλου που κάθε κατάσταση είναι εξαρτώμενη από το προηγούμενο παρατηρούμενο σύμβολο

Η σχέση πλέον με την εξάρτηση των καταστάσεων ορίζεται ως εξής:

$$P(x_i | \pi_i, \pi_{i-1}, \dots, \pi_1, x_{i-1}, x_{i-2}, \dots, x_1) = P(x_i | \pi_i)$$

$$P(\pi_i | \pi_{i-1}, \pi_{i-2}, \dots, \pi_1, x_{i-1}, x_{i-2}, \dots, x_1) = P(\pi_i | \pi_{i-1}, x_{i-1})$$

Και οι πιθανότητες μετάβασης των καταστάσεων του μοντέλου (transitions) όπου πλέον ισχύει

$$a_{kl} = P(\pi_i = l | \pi_{i-1} = k, x_{i-1} = c)$$

Χρήσιμα συμπεράσματα της συγκεκριμένης εργασίας αποτελούν η χρησιμότητα που έχει η εξάρτηση από προηγούμενες παρατηρήσεις στην πρόβλεψη πρωτεϊνών και η ανάγκη για μεγαλύτερο σύνολο εκπαίδευσης λόγω του μεγαλύτερου αριθμού παραμέτρων. Σε δοκιμές που έγιναν με μικρό αριθμό δεδομένων εκπαίδευσης, το κλασσικό HMM είχε καλύτερα αποτελέσματα αλλά όπως επισημαίνουν, αναμένουν να βελτιστοποιηθεί σε σχέση με το κλασσικό HMM σε κάποιες περιπτώσεις που θα χρησιμοποιηθεί ένας μεγάλος αριθμός συνόλου εκπαίδευσης και ελέγχου, που θα αποτελείται αντίστοιχα από ακολουθίες 15000 και 2000 αμινοξέων.

3.4 Hidden Neural Network HNN

Μια πιο γενική περίπτωση αποτελεί αυτή των Hidden Neural Network [12]. Το HNN είναι ένα υβριδικό μοντέλο, όπου ο συνδυασμός CHMM με νευρωνικά δίκτυα, μπορεί να οδηγήσει σε ένα πιο ευέλικτο και ισχυρό πρότυπο για την ταξινόμηση. Η βασική ιδέα του HNN είναι να αντικαταστήσει τις πιθανότητες των παραμέτρων του CHMM με χρήση κατάλληλων νευρώνων που λαμβάνουν τις παρατηρήσεις ως είσοδο και θα γίνει εκτίμηση των παραμέτρων χρησιμοποιώντας τον αλγόριθμο back-propagation. Κατ-αρχάς, τα νευρωνικά δίκτυα μπορούν να εφαρμόσουν πολύπλοκες λειτουργίες χρησιμοποιώντας πολύ λιγότερες παραμέτρους. Επιπλέον, ένα HNN μπορεί να χρησιμοποιήσει άμεσα ένα πλαίσιο παρατηρήσεων (context) ως είσοδος στα νευρωνικά δίκτυα και έτσι να εκμεταλλευτεί συσχετίσεις υψηλότερης τάξης (high-order) μεταξύ διαδοχικών παρατηρήσεων, το οποίο είναι δύσκολο να εφαρμοστεί στα κλασσικά HMM. Η είσοδος του δικτύου αντιστοιχεί στο “context” και συνήθως είναι ένα πλαίσιο παραθύρου k παρατηρήσεων $x^k = x_{i-k}, x_{i-k-1}, \dots, x_{i-1}, x_i$.

4. Στόχος

Η επέκταση ενός HMM σε μεγαλύτερης τάξεως αλυσίδες μπορεί θεωρητικά να είναι χρήσιμη όπως είδαμε και στο προηγούμενο κεφάλαιο. Όμως, λόγω της υψηλής υπολογιστικής πολυπλοκότητας τους, τα High-Order HMM έχουν βρει μικρή εφαρμογή σε υπάρχοντα συστήματα επεξεργασίας ομιλίας και ανάλυσης βιολογικών ακολουθιών. Ένα από τα προβλήματα είναι ο αριθμός των παραμέτρων. Για παράδειγμα, σε ένα High-Order HMM $K^{\text{ης}}$ τάξης με N καταστάσεις, χρειάζονται NK πιθανότητες μετάβασης. Αυτός ο μεγάλος αριθμός παραμέτρων απαιτεί επίσης ένα μεγάλο σύνολο δεδομένων για την κατάλληλη εκπαίδευση. Εκτός από αυτές τις δυσκολίες, οι γνωστοί αλγόριθμοι των κλασικών HMM δεν επαρκούν για την αντιμετώπιση των προβλημάτων μοντελοποίησης και είναι απαραίτητο να επεκταθούν κατάλληλα για την εφαρμογή τους στη συγκεκριμένη επέκταση του HMM.

Ωστόσο, σε μια εργασία παρουσιάστηκε ένας έξυπνος τρόπο εκπαίδευσης μεγαλύτερης τάξεως HMM χωρίς να χρειαστεί να τροποποιηθούν οι γνωστοί αλγόριθμοι [19]. Εξ ορισμού βασιστήκαν στην υπόθεση ότι κάθε κατάσταση διατηρεί την Μαρκοβιανή ιδιότητα (σε διακριτό χρόνο). Η πιθανότητα μετάβασης προς μια συγκεκριμένη κατάσταση στο επόμενο χρονικό βήμα εξαρτάται μόνο από την κατάσταση που βρίσκεται κατά την τρέχουσα χρονική στιγμή. Ως εκ τούτου, μόνο μία πιθανότητα μετάβασης συμβαίνει στη σχέση μεταξύ των δύο καταστάσεων. Αυτές οι πιθανότητες όμως μπορούν να παρουσιαστούν ως συλλογή (collection) και αναπτύσσοντας κατάλληλο αλγόριθμο μετασχηματίζουν την κωδικοποίηση των καταστάσεων, αποφεύγοντας όμως περιττές πιθανότητες μετάβασης. Με την νέα κωδικοποίηση των καταστάσεων, αλλάζει η δομή του μοντέλου και χρησιμοποιούνται οι γνωστοί αλγόριθμοι για την εκπαίδευση και τον υπολογισμό της απαιτούμενης πιθανότητας. Σε σύγκριση με τις προηγούμενες προσπάθειες, μειώνει σημαντικά τις υπολογιστικές απαιτήσεις.

Αφού μελετήσαμε όλες τις προηγούμενες προσπάθειες θεωρήσαμε ότι, στην περίπτωση των πρωτεϊνών, που τα μοντέλα είναι πιο σύνθετα, αναμένουμε τα high-order HMM να έχουν μεγάλη υπολογιστική πολυπλοκότητα και να μην έχουν τόσο καλά αποτελέσματα. Επιπλέον, επειδή στις προγνωστικές μεθόδους που χρησιμοποιούμε, η δομή των καταστάσεων της κάθε μεθόδου είναι σχεδιασμένη να αντιμετωπίζει ένα συγκεκριμένο βιολογικό πρόβλημα, δεν μπορούμε να τροποποιήσουμε την κωδικοποίηση των καταστάσεων.

Ο κύριος στόχος της παρούσας εργασίας είναι να επεκτείνουμε τυπικά HMM για να ενσωματώσουμε μεγαλύτερης τάξεως αλυσίδες στις πιθανότητες εμφάνισης συμβόλων. Βασικός σκοπός είναι να επεκτείνουμε μια σειρά από προγνωστικές μεθόδους που χρησιμοποιούνται στην πρόγνωση χαρακτηριστικών σε μια πρωτεϊνική ακολουθία όπως διαμεμβρανικά τμήματα. Είναι και βιολογικά σωστό βέβαια γιατί σύμφωνα με τις βασικές αρχές της βιοχημείας στη πρωτοταγή δομή μιας πρωτεΐνης εμφανίζει μεγάλη σημασία η σειρά που συντάσσονται τα αμινοξέα παρουσιάζοντας μια εξάρτηση με τα προηγούμενα.

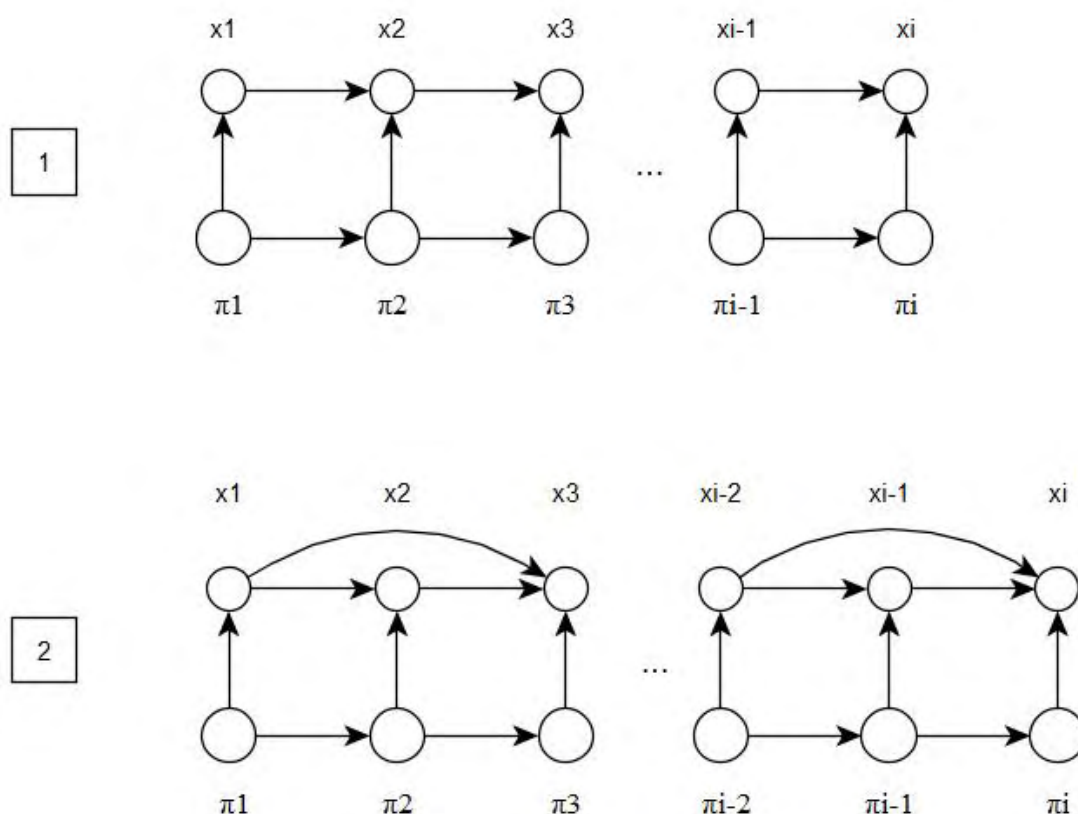
Προτείνουμε μια μέθοδο η οποία επεκτείνει τις κωδικοποιήσεις του αλφαβήτου που προσδιορίζει το σύνολο των συμβόλων των παρατηρήσεων. Η μέθοδος καθιστά δυνατή την επεξεργασία ενός HMM χρησιμοποιώντας την δομή του HMM και τους γνωστούς αλγόριθμους, επιτυγχάνοντας έτσι την ίδια υπολογιστική αποδοτικότητα και στην εκτεταμένη προσέγγιση. Με την επέκταση της κωδικοποίησης στα μοντέλα το τρέχον παρατηρούμενο σύμβολο (αμινοξύ) μιας ακολουθίας συσχετίζεται με ένα αριθμό προηγούμενων παρατηρούμενων συμβόλων και έχουμε τη δυνατότητα να ελέγξουμε κατά πόσο η χρήση

πρόσθετης πληροφορίας (υδροφοβικότητα, πολικότητα) μπορεί να βελτιώσει την απόδοση στην πρόβλεψη χαρακτηριστικών των πρωτεϊνών.

5. Μεθοδολογία

Για να εκτιμήσουμε τη συμβολή της πρόσθετης πληροφορίας, επιλέξαμε να εργαστούμε με γνωστές προγνωστικές μεθόδους που αναπτυχθήκαν από την ομάδα μας, και να τις επεκτείνουμε χρησιμοποιώντας 2^{ης} τάξεως HMM και λαμβάνοντας υπόψη μια σειρά από διαφορετικές κωδικοποιήσεις. Σε αυτή την ενότητα, συζητάμε την μέθοδο που ακολουθήσαμε για να τροποποιήσουμε το αλφάβητο, τις παραμέτρους και τα σύνολα εκπαίδευσης. Τέλος, για την αξιολόγηση πραγματοποιήσαμε μια σειρά από δοκιμές χρησιμοποιώντας κατάλληλες μεθόδους και μετρήσαμε την απόδοσή τους.

5.1 μέθοδος επέκτασης της Κωδικοποίησης



Εικόνα 12: Απεικόνιση νέων μοντέλων εξαρτώμενα από προηγούμενες παρατηρήσεις. 1. Μοντέλο 2^{ης} τάξης που ελέγχει διπεπτίδια και λαμβάνει υπόψη τη προηγούμενη παρατηρούμενη τιμή. 2. Μοντέλο 3^{ης} τάξης που ελέγχει τριπεπτίδια δηλαδή κάθε παρατηρούμενη τιμή λαμβάνει υπόψη τις 2 προηγούμενες παρατηρούμενες τιμές

Για την εφαρμογή της μεθόδου, ορίζουμε την μεταβλητή k η οποία χαρακτηρίζει τον αριθμό των αμινοξικών καταλοίπων που θα ελέγχονται και αντίστοιχα καθορίζει τον αριθμό (τάξη) των προηγούμενων παρατηρούμενων συμβόλων που θα λαμβάνει υπόψη η νέα κωδικοποίηση, για την τρέχουσα παρατηρούμενη τιμή. Ο αριθμός των προηγούμενων συμβόλων καθορίζεται ως $k-1$. Επομένως, για ένα μοντέλο 1^{ης} τάξης, που ουσιαστικά αντιστοιχεί σε ένα κλασικό HMM, οι συχνότητες στη θέση i εξαρτώνται από το υπόλειμμα στη θέση i . Το μοντέλο 2^{ης} τάξης λαμβάνει υπόψη τις συχνότητες της θέσης i και $i-1$. Ένα μοντέλο k ^{ης} τάξης υποθέτει ότι ένα σύμβολο εξαρτάται από k προηγούμενες συνεχόμενες παρατηρήσεις προς τα πίσω. Ένα χαρακτηριστικό παράδειγμα μοντέλων 2^{ης} και 3^{ης} τάξης παρουσιάζονται στην Εικόνα 13.

Στο παράδειγμα μας ελέγχουμε το τρέχον αμινοξικό κατάλοιπο με το προηγούμενο οπότε ελέγχουμε για διπεπτίδια, δηλαδή $k = 2$ και το είδος της τάξης θα είναι αντίστοιχα 2^{ης} τάξεως μοντέλα. Αντίστοιχα για $k=3$ ελέγχουμε με 3^{ης} τάξεως μοντέλα τριπεπτίδιο δηλαδή το τρέχον αμινοξικό κατάλοιπο και 2 προηγούμενα. Για $k=1$ ελέγχουμε με 1^{ης} τάξεως μοντέλο δηλαδή το τρέχον αμινοξικό κατάλοιπο που αντιστοιχεί στο κλασσικό HMM.

Στη δική μας περίπτωση η σχέση πλέον με την εξάρτηση των συμβόλων με τις παρατηρήσεις ορίζεται ως εξής:

2^{ης} τάξης

$$P(x_i | \pi_i, \pi_{i-1}, \dots, \pi_1, x_{i-1}, x_{i-2}, \dots, x_1) = P(x_i | \pi_i, x_{i-1})$$

$$P(\pi_i | \pi_{i-1}, \pi_{i-2}, \dots, \pi_1, x_{i-1}, x_{i-2}, \dots, x_1) = P(\pi_i | \pi_{i-1})$$

3^{ης} τάξης

$$P(x_i | \pi_i, \pi_{i-1}, \dots, \pi_1, x_{i-1}, x_{i-2}, \dots, x_1) = P(x_i | \pi_i, x_{i-1}, x_{i-2})$$

$$P(\pi_i | \pi_{i-1}, \pi_{i-2}, \dots, \pi_1, x_{i-1}, x_{i-2}, \dots, x_1) = P(\pi_i | \pi_{i-1})$$

k ^{ης} τάξης

$$P(x_i | \pi_i, \pi_{i-1}, \dots, \pi_1, x_{i-1}, x_{i-2}, \dots, x_1) = P(x_i | \pi_i, x_{i-1}, x_{i-2}, \dots, x_{i-k})$$

$$P(\pi_i | \pi_{i-1}, \pi_{i-2}, \dots, \pi_1, x_{i-1}, x_{i-2}, \dots, x_1) = P(\pi_i | \pi_{i-1})$$

και οι πιθανότητες εκπομπής ή εμφάνισης συμβόλων (emissions) για κάθε κατάσταση j , όπου για 2^{ης} τάξεως μοντέλο πλέον ισχύει

$$e_k(x_i = b') = P(x_i = b | \pi_i = k, x_{i-1} = c)$$

όπου k συμβολίζει το νέο σύμβολο και c το προηγούμενο σύμβολο

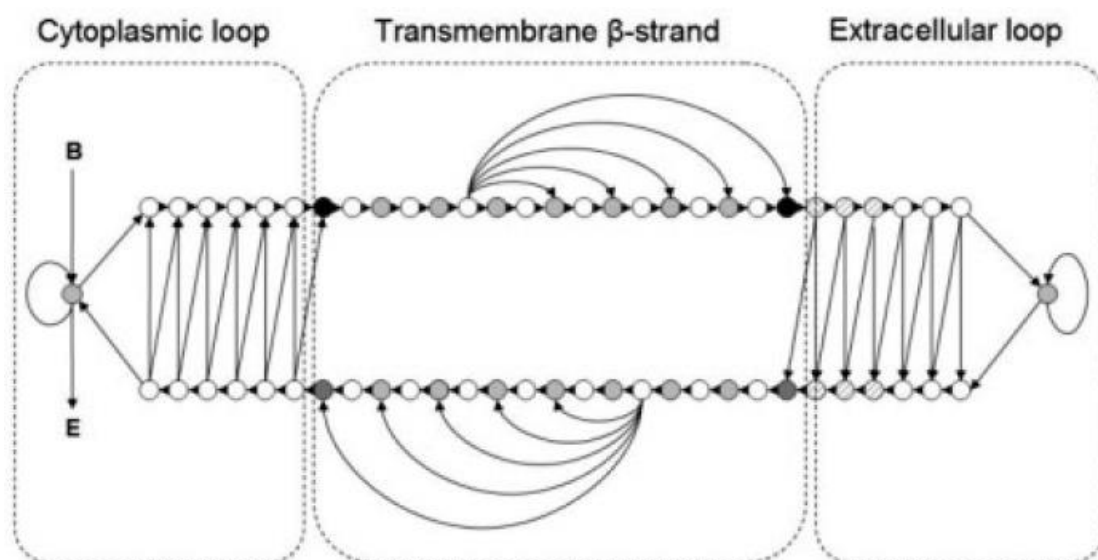
Σε γενικές γραμμές αναμένουμε ότι με μεγαλύτερης τάξεως αλυσίδες θα έχουμε και καλύτερη διαχωριστική ικανότητα των μοντέλων. Βέβαια όπως θα δούμε εφαρμόζοντας μεγάλης τάξεως αλυσίδες πέραν του προβλήματος της υπερ-προσαρμογής (over-fitting) που πιθανόν να προκύπτει επιπλέον ανακύπτει το πρόβλημα έλλειψης δεδομένων και περιορισμένου αριθμού συμβόλων που μπορούμε να χρησιμοποιήσουμε.

5.2 Μοντέλα

Προκειμένου να εξετάσουμε την απόδοση των νέων κωδικοποιήσεων πραγματοποιήσαμε δοκιμές σε γνωστές μεθόδους πρόβλεψης διαμεμβρανικών πρωτεϊνών που αναπτυχθήκαν από την ομάδα μας και τα οποία περιγράφονται εν συντομία παρακάτω.

PRED-TMBB

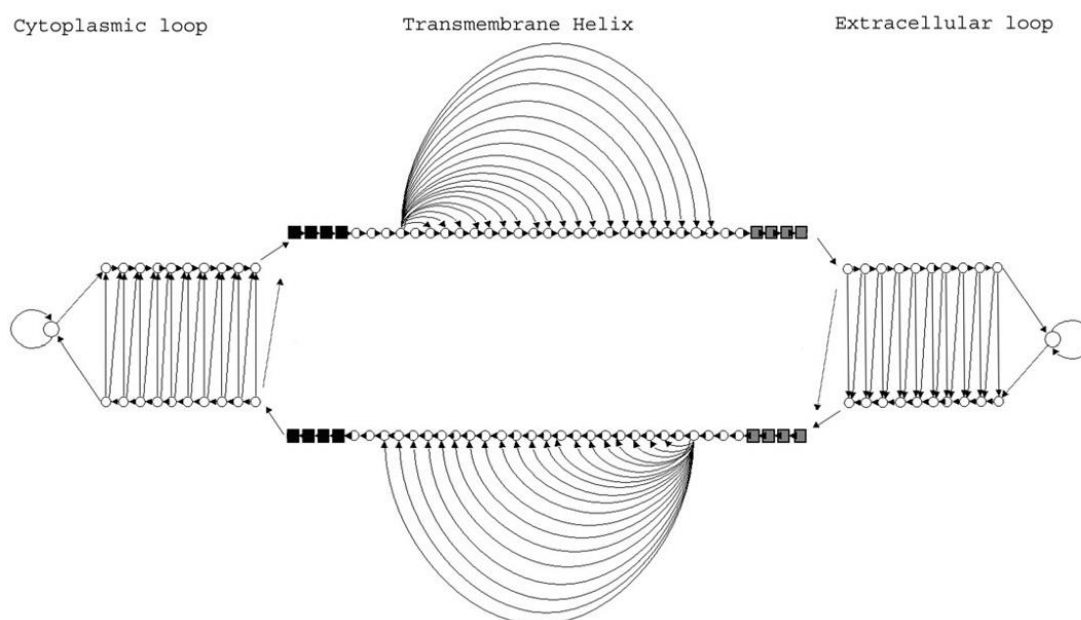
μέθοδος πρόγνωσης διαμεμβρανικών β-βαρελίων [20, 21]. Το μοντέλο χρησιμοποιεί ένα HMM 62 καταστάσεων του οποίου η αρχιτεκτονική απεικονίζεται στην Εικόνα 14 και ένα σετ 391 παραμέτρων. Για την εκπαίδευση χρησιμοποιήθηκε ένα σετ 49 πρωτεϊνών β-βαρελίων οι οποίες στο σύνολο τους περιλαμβάνουν 23422 αμινοξέα.



Εικόνα 13: Απεικόνιση της αρχιτεκτονικής του HMM που χρησιμοποιείται στη μέθοδο PRED-TMBB2.

HMM-TM

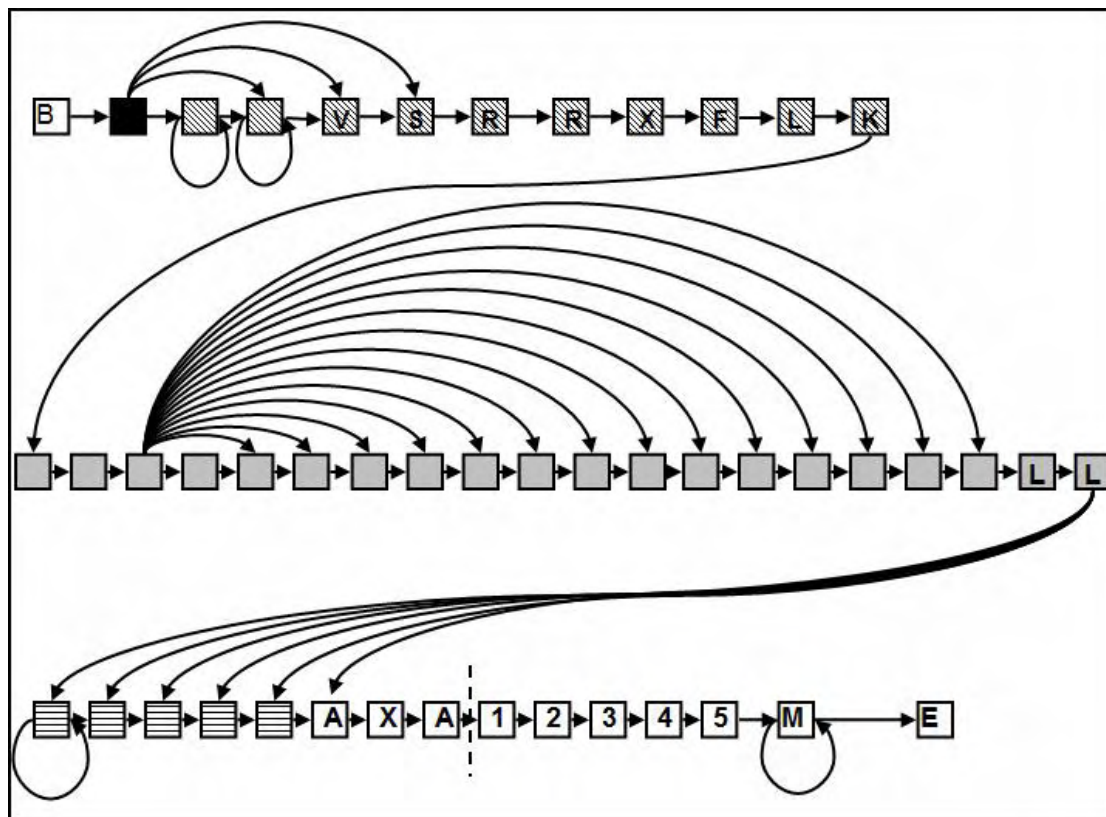
μέθοδος πρόγνωσης διαμεμβρανικών α -ελικών [22]. Το μοντέλο που χρησιμοποιείται είναι κυκλικό, αποτελείται από 114 καταστάσεις, συμπεριλαμβανομένων των καταστάσεων Έναρξης (B) και τερματισμού (E) του οποίου η αρχιτεκτονική απεικονίζεται στην Εικόνα 15 και ένα σετ 302 παραμέτρων. Για την εκπαίδευση χρησιμοποιήθηκε ένα σετ 72 α -ελικοειδών διαμεμβρανικών πρωτεϊνών με πειραματικά προσδιορισμένη τρισδιάστατη δομή(δεδομένα από την Protein Data Bank (PDB)) και στο σύνολο τους περιλαμβάνουν 17537 αμινοξέα.



Εικόνα 14: Απεικόνιση της αρχιτεκτονικής του HMM που χρησιμοποιείται στη μέθοδο HMM-TM.

PRED-TAT

μέθοδος πρόβλεψης των θέσεων αποκοπής των πεπτιδίων οδηγητών βακτηριών [23]. Η μέθοδος μπορεί να διαχωρίσει τα πεπτίδια οδηγητές (Sec και Tat) και να προβλέψει και τις θέσεις αποκοπής στις δύο κατηγορίες.

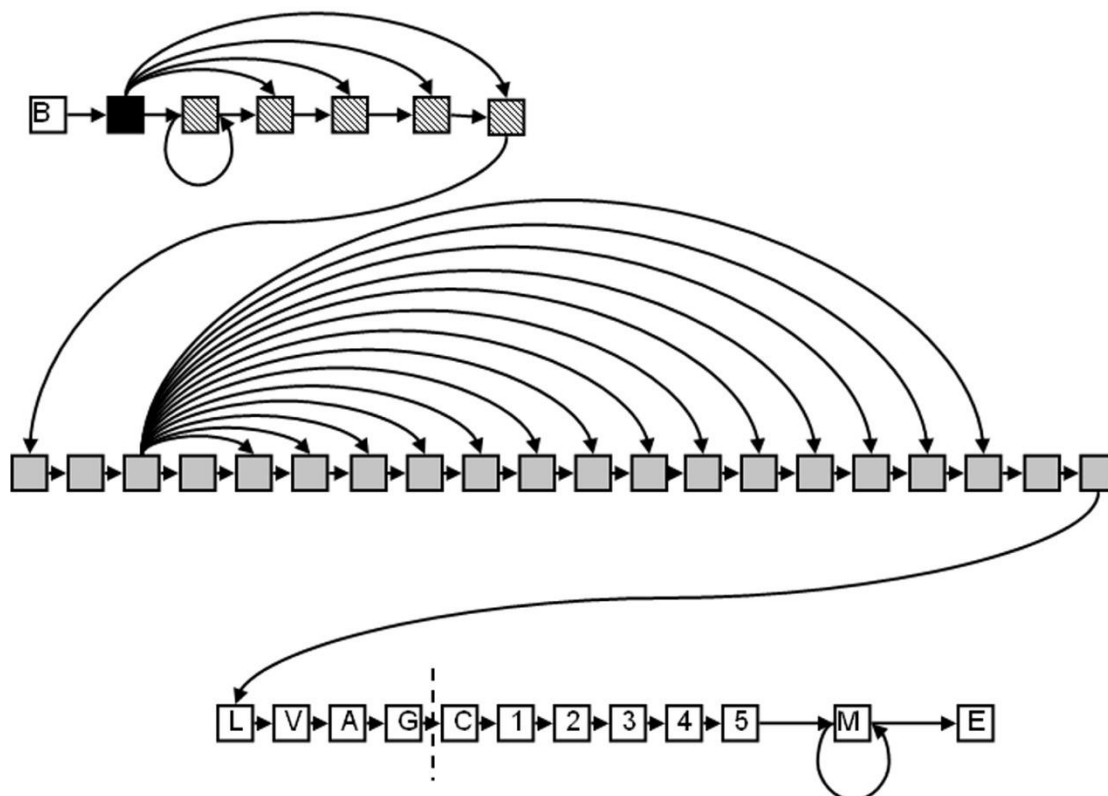


Εικόνα 15: Απεικόνιση της αρχιτεκτονικής του υπομοντέλου HMM που χρησιμοποιείται στη μέθοδο PRED-TAT και αντιστοιχεί στα πεπτίδια οδηγητές Tat. Το σημείο κοπής παρουσιάζεται με διακεκομμένη κάθετη γραμμή μεταξύ των καταστάσεων A και 1.

Το μοντέλο χρησιμοποιεί ένα σετ 632 παραμέτρων και το σύνολο δεδομένων που χρησιμοποιείται για την εκπαίδευση περιείχε ένα σετ 906 ακολουθιών από τις οποίες 150 είναι πρωτεΐνες που φέρουν πεπτίδια οδηγητές Tat (119 από Gram-αρνητικά βακτήρια και 31 από Gram-θετικά), 328 είναι εκκρινόμενες πρωτεΐνες που περιέχουν πεπτίδια οδηγητές Sec (216 από Gram-αρνητικά βακτήρια και 112 από Gram-θετικά), 288 είναι κυτταροπλασματικές πρωτεΐνες (183 από Gram-αρνητικά βακτήρια και 105 από Gram-θετικά) και 140 ακολουθίες αποτελούν τμήματα διαμεμβρανικών πρωτεϊνών (90 από Gram-αρνητικά βακτήρια και 50 από Gram-θετικά). Στο σύνολο του το σετ εκπαίδευσης περιλαμβάνει 65148 αμινοξέα.

PRED-LIPO

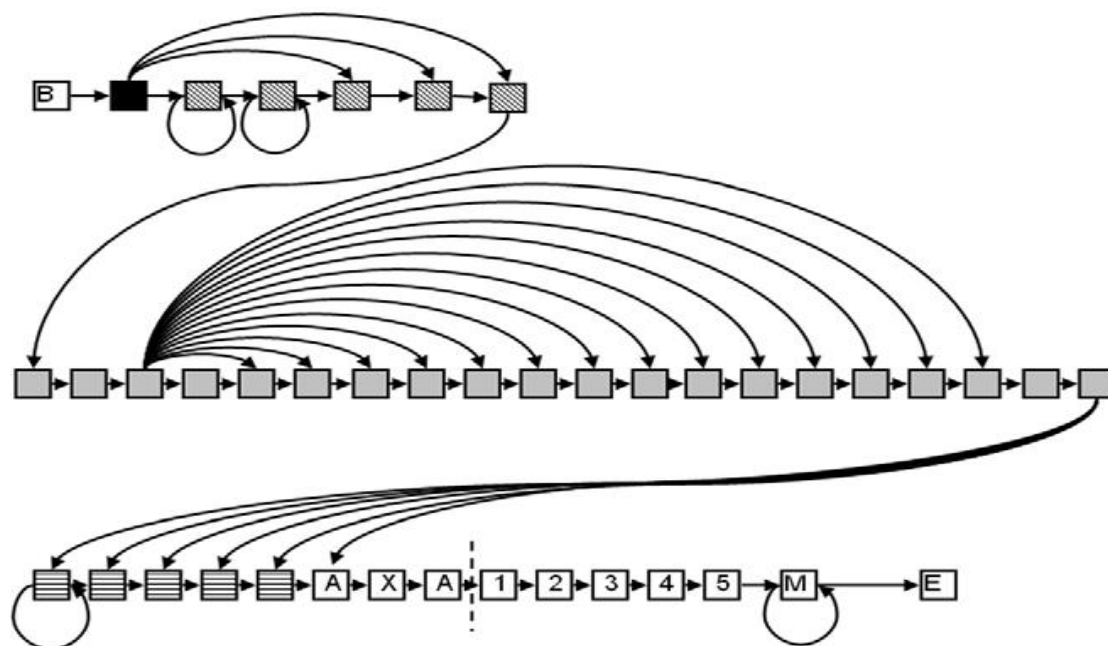
μέθοδος πρόβλεψης των λιποπρωτεϊνών από θετικά κατά Gram Βακτήρια [24]. Η μέθοδος προβλέπει τα πεπτίδια των εκκρινόμενων πρωτεϊνών και τις διαμεμβρανικές έλικες. Ο συνολικός αριθμός των καταστάσεων του μοντέλου είναι 134 (συμπεριλαμβανομένων των καταστάσεων έναρξης και τερματισμού) με ένα σετ 626 παραμέτρων. Για την εκπαίδευση χρησιμοποιήθηκε ένα σετ 355 ακολουθιών οι οποίες στο σύνολο τους περιλαμβάνουν 23422 αμινοξέα.



Εικόνα 16: Απεικόνιση της αρχιτεκτονικής του HMM που χρησιμοποιείται στη μέθοδο PRED-LIPO. Το σημείο αποκοπής παρουσιάζεται με διακεκομμένη κάθετη γραμμή μεταξύ G και C

PRED-SIGNAL

μέθοδος πρόβλεψης των πεπτιδίων οδηγητών στα αρχαία [25]. Η μέθοδος χρησιμοποιεί ένα σετ 414 παραμέτρων και για την εκπαίδευση χρησιμοποιήθηκε ένα σετ 321 ακολουθιών οι οποίες στο σύνολο τους περιλαμβάνουν 21479 αμινοξέα.



Εικόνα 17: Απεικόνιση της αρχιτεκτονικής του HMM που χρησιμοποιείται στη μέθοδο PRED-SIGNAL. Το σημείο κοπής παρουσιάζεται με διακεκομμένη κάθετη γραμμή μεταξύ των καταστάσεων A και 1.

5.3 Κωδικοποίηση

Ένα αρκετά σημαντικό θέμα αποτελεί η μέθοδος συσχέτισης των παρατηρήσεων που θα χρησιμοποιήσουμε, από βιολογικής σημασίας για την δική μας περίπτωση, για να επεκτείνουμε την κωδικοποίηση και να πετύχουμε την σύνδεση της εξάρτησης ενός συμβόλου με τις προηγούμενες παρατηρήσεις. Στην εργασία μας, λάβαμε υπόψη το υδρόφοβο-πολικό μοντέλο το οποίο πρώτος πρότεινε ο Ken Dill το 1985 [26] και αποτελεί μια απλοποιημένη λύση που εξετάζει την αναδίπλωση της πρωτεΐνης στο χώρο. Η κεντρική του ιδέα, πηγάζει από την παρατήρηση ότι οι υδρόφοβες αλληλεπιδράσεις μεταξύ των κατάλοιπων αμινοξέων αποτελεί τη κινητήρια δύναμη για την αναδίπλωση των πρωτεϊνών σε κανονικές δομές. Επίσης, όλοι οι τύποι αμινοξέων ταξινομούνται είτε ως υδρόφοβα (H) ή πολικά (P).

Λαμβάνοντας υπόψη τα παραπάνω, στην παρούσα εργασία επεκτείναμε την κωδικοποίηση των μοντέλων, για κάθε μια από τις παρακάτω περιπτώσεις:

1. Μοντέλο που ελέγχει και το προηγούμενο κατάλοιπο αν είναι υδρόφοβο (AFHILMVWY) ή μη υδρόφοβο (CDEGKNPQRST). Για συντομία το χαρακτηρίζουμε ως Μοντέλο (40).
2. Μοντέλο που ελέγχει αν το προηγούμενο ανήκει σε μία από τέσσερις ομάδες: Υδρόφοβα-Αρωματικά (FHYW), Υδρόφοβα-Μη Αρωματικά (AILMVG), Μη Υδρόφοβα-Φορτισμένα (DEKR), Μη Υδρόφοβα-Πολικά (CNPQST). Για συντομία το χαρακτηρίζουμε ως Μοντέλο (80).
3. Μοντέλο που ελέγχει αν το προηγούμενο ανήκει σε μία από οκτώ ομάδες: Υδρόφοβα-Μικρά (AG), Πολικά-Ειδικά (PC), Πολικά-OH (ST), Πολικά-NH (NQ), Φορτισμένα-Αρνητικά (DE), Φορτισμένα-Θετικά (KR), Υδρόφοβα-Μεγάλα (ILMV) και Υδρόφοβα-Αρωματικά (FHYW). Για συντομία το χαρακτηρίζουμε ως Μοντέλο (160).
4. Μοντέλο που λαμβάνει υπόψη τα όλα τα δυνατά διπεπτίδια. Για συντομία το χαρακτηρίζουμε ως Μοντέλο (400).

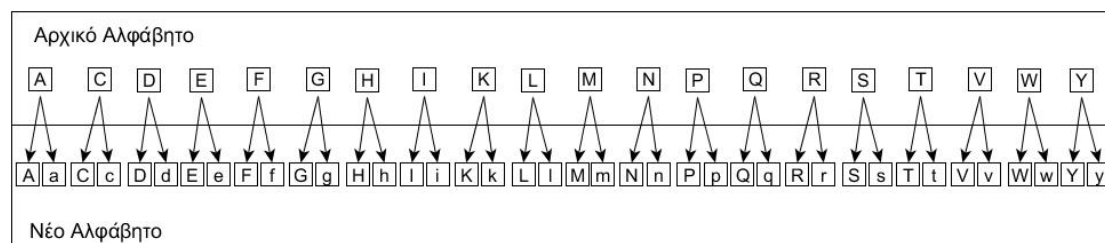
5.4 Αλφάβητο

Σε γενικές γραμμές, ένα υψηλότερης τάξης Markov μοντέλο για την ανάλυση ακολουθίας της πρωτεΐνης, απαιτεί πολύ περισσότερα σύμβολα από ένα μοντέλο χαμηλότερης τάξης για να υπολογιστεί. Για παράδειγμα, για την κωδικοποίηση που λαμβάνει υπόψη όλα τα δυνατά διπεπτίδια, επειδή τη χρονική στιγμή i , τα παρατηρούμενα σύμβολα x_i και x_{i-1} μπορεί να είναι οποιοδήποτε από τα 20 αμινοξέα, τα στατιστικά στοιχεία των διαδοχικών ζευγών-υπολειμμάτων θα δημιουργήσει ένα πλέγμα με 20×20 στοιχεία. Στην περίπτωση αυτή χρειαζόμαστε να υπολογίσουμε $20^2 = 400$ νέα σύμβολα για ένα μοντέλο 2^{ης} τάξεως, $20^3 = 8000$ νέα σύμβολα για ένα μοντέλο 3^{ης} τάξεως, αριθμός υπερβολικά μεγάλος ο οποίος θα απαιτούσε υπερβολικά μεγάλο αριθμό συμβόλων και ακολουθιών για την εκπαίδευση. Σε περιπτώσεις νουκλεοτιδίων θα μπορούσαμε να εφαρμόσουμε μέχρι και 5^{ης} τάξεως αλυσίδα δηλαδή $4^5 = 1024$ σύμβολα. Δεδομένου ότι οι συσχετίσεις μεταξύ των καταλοίπων είναι διαφορετικές ανάλογα με την μέθοδο συσχέτισης των παρατηρήσεων που θα ληφθεί υπόψη για την επέκταση ενός μοντέλου, θα πρέπει να ληφθεί υπόψη και το μέγεθος του αλφαβήτου.

Παράδειγμα

Θα δώσουμε ένα παράδειγμα για την παραγωγή των νέων συμβόλων που θα χρησιμοποιηθούν στην νέα κωδικοποίηση και την αντίστοιχη εφαρμογή τους σε μια ακολουθία. Για την αντιμετώπιση του 1^{ου} μοντέλου που ελέγχει αν το προηγούμενο κατάλοιπο είναι υδρόφοβο ή μη υδρόφοβο χρειάζεται να συμπεριλάβουμε για κάθε ένα σύμβολο του αρχικού Αλφαβήτου

που αποτελούν τα 20 αμινοξέα ένα νέο σύμβολο για την περίπτωση που το προηγούμενο αμινοξύ είναι υδρόφοβο (Θα το συμβολίσουμε με το ίδιο σύμβολο σε Κεφαλαία μορφή) και ένα νέο σύμβολο για την περίπτωση που το προηγούμενο αμινοξύ είναι μη υδρόφοβο (θα το συμβολίσουμε με το ίδιο γράμμα σε πεζή μορφή). Οπότε στο σύνολο θα χρειαστούμε πλέον τα διπλάσια σύμβολα από το αρχικό αλφάβητο δηλαδή $20 \times 2 = 40$ σύμβολα τα οποία δανειζόμαστε από το Αγγλικό αλφάβητο σε Κεφαλαία και πεζή μορφή όπως απεικονίζεται χαρακτηριστικά στην Εικόνα 18.



Εικόνα 18: Απεικόνιση δημιουργίας νέου αλφαβήτου που λαμβάνει υπόψη αν το προηγούμενο είναι υδρόφοβο ή όχι. Με κεφαλαία γράμματα απεικονίζονται τα σύμβολα που το προηγούμενο σύμβολο είναι υδρόφοβο και με μικρά γράμματα τα σύμβολα που το προηγούμενο σύμβολο δεν είναι υδρόφοβο.

Εφόσον ορίσαμε και δημιουργήσαμε το νέο αλφάβητο που θα χρησιμοποιήσουμε, θα τροποποιήσουμε την αναπαράσταση της κάθε ακολουθίας του συνόλου εκπαίδευσης με τη χρήση της τεχνικής του κινούμενου παραθύρου. Με την τεχνική αυτή, ένα κινούμενο παράθυρο ολισθαίνει κατά μήκος της ακολουθίας και κάθε φορά το παράθυρο καθορίζει το σύμβολο με το οποίο θα αντικατασταθεί το τελευταίο κατάλοιπο του παραθύρου, που αντιστοιχεί και στο τρέχον σύμβολο. Το μήκος του παραθύρου εξαρτάται από τον αριθμό των καταλοίπων που θα ελέγχονται (έχει οριστεί ως k) και το μέγεθος του αντιστοιχεί σε k . Στο παράδειγμα μας εφόσον κοιτάμε διπεπτίδια (δηλαδή $k = 2$) το μέγεθος του παραθύρου θα είναι 2.

Ένα θέμα που προκύπτει είναι με τις αρχικές θέσεις της ακολουθίας που δεν θα έχουν αντίστοιχα προηγούμενα σύμβολα. Η αντιμετώπιση του καθορίζεται ανάλογα με την φύση του προβλήματος. Στην περίπτωση μας λάβαμε υπόψη για τι αρχικές θέσεις της ακολουθίας που δεν υπήρχαν σύμβολα το σύμβολο A που αντιστοιχεί στο αμινοξύ Αλανίνη.

Παραπάνω δίνεται ένα παράδειγμα για την επέκταση της ακολουθίας

>P14319_1

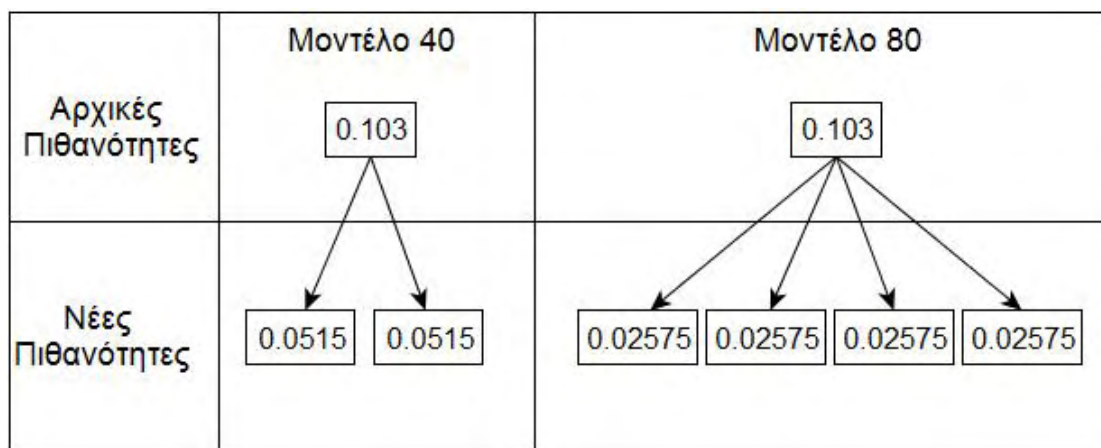
MPYIYLIPIASTEVIKSAFLKSSEG

για το μοντέλο που ελέγχει και το προηγούμενο κατάλοιπο αν είναι υδρόφοβο ή μη υδρόφοβο. Στο παράδειγμα βλέπουμε ότι στην αρχή της ακολουθίας προστίθεται η Αλανίνη (A) και την κατά την ολίσθηση του παραθύρου διακρίνουμε στο αριστερό τμήμα με έντονη γραφή το τρέχον κατάλοιπο και το προηγούμενο του. Λαμβάνοντας υπόψη αν το προηγούμενο κατάλοιπο είναι Υδρόφοβο ή μη παράγεται στο δεξιό τμήμα το νέο σύμβολο. Η διαδικασία ολοκληρώνεται με το τέλος της ακολουθίας.

Παρατηρώντας τις δύο ακολουθίες με γυμνό μάτι μπορούμε άμεσα να αντιληφθούμε την πρόσθετη πληροφορία που έχει ενσωματωθεί στο μοντέλο παρατηρώντας μόνο τα σύμβολα που απεικονίζονται με πεζή μορφή δηλαδή ότι το προηγούμενο σύμβολο χαρακτηρίζεται ως μη υδρόφοβο.

Τα αρχικά μοντέλα περιλαμβάνουν τις αρχικές πιθανότητες (με την εκπαίδευση θα αλλάξουν) για κάθε ένα από τα σύμβολα του αλφαβήτου, που στην περίπτωση μας αποτελείται από τα σύμβολα των 20 αμινοξέων, σε κάθε κατάσταση. Με την επέκταση του αλφαβήτου θα πρέπει αντίστοιχα να επεκτείνουμε και το σύνολο των πιθανοτήτων. Οπότε σε αντιστοιχία με την επέκταση του αλφαβήτου χρειαζόμαστε και την επέκταση των πιθανοτήτων. Το μέγεθος των πιθανοτήτων ακολουθεί την περιπλοκότητα της αύξησης του μεγέθους του αλφαβήτου. Επομένως, ένα υψηλότερης τάξης μοντέλο για την ανάλυση μιας πρωτεϊνικής ακολουθίας, απαιτεί πολύ περισσότερες πιθανότητες από ένα μοντέλο χαμηλότερης τάξης για να υπολογιστεί. Για παράδειγμα, ένα k^{th} τάξης μοντέλο απαιτεί 20^k πιθανοτήτων.

Στην εργασία ένας τρόπος με τον οποίο εργαστήκαμε για να επεκτείνουμε τις αρχικές πιθανότητες για κάθε αμινοξύ ήταν παράγοντας αντίστοιχα τόσες νέες πιθανότητες με ίση κατανομή της αρχικής πιθανότητας όπως απεικονίζεται στην Εικόνα 19. Για παράδειγμα στο μοντέλο 40 που ελέγχουμε αν το προηγούμενο κατάλοιπο είναι Υδρόφοβο ή μη από την αρχική πιθανότητα ενός συμβόλου δημιουργήθηκαν δύο νέες πιθανότητες μια για την περίπτωση το αντίστοιχο αμινοξύ να έχει προηγούμενο κατάλοιπο υδρόφοβο και μια για το αντίστοιχο αμινοξύ να έχει προηγούμενο κατάλοιπο μη υδρόφοβο.



Εικόνα 19: Απεικόνιση κατανομής πιθανοτήτων στα νέα μοντέλα.

Σε αντίστοιχη προσπάθεια να χρησιμοποιηθούν μεγαλύτερης τάξεως αλυσίδες σε μαρκοβιανά μοντέλα και αντίστοιχα μεγαλύτερος αριθμός παραμέτρων γίνεται αναφορά στο γεγονός ότι σε μεγαλύτερης τάξεως μοντέλα δεν μπορεί να γίνει καλή εκτίμηση των παραμέτρων, λόγω του μικρού σχετικά δείγματος ακολουθιών πρωτεϊνών που είναι διαθέσιμες [14]. Επίσης αναμένουμε κατά την εκπαίδευση να χρειαστεί μεγαλύτερος χρόνος να συγκλίνουν οι παράμετροι.

5.6 Διαδικασία Αξιολόγησης

Στόχος μας ήταν να συγκρίνουμε την επίδραση των νέων κωδικοποιήσεων στα μοντέλα, διαχωρίζοντας την επίδραση τους από το σύνολο εκπαίδευσης. Αρχικά, μια εργασία είναι να συγκεντρωθούν τα σύνολα εκπαίδευσης. Για να το πετύχουμε αυτό, χρησιμοποιήσαμε τα σύνολα εκπαίδευσης και τις παραμέτρους που είχαν χρησιμοποιηθεί και παλαιότερα από τα μοντέλα που εργαστήκαμε. Επιπλέον, πραγματοποιήσαμε και όλες τις αντίστοιχες εκτελέσεις για την επανεκπαίδευση των αρχικών μοντέλων προκειμένου να αξιολογηθεί η επιρροή της χρήσης πρόσθετης πληροφορίας.

Για το σκοπό της σύγκρισης της αποτελεσματικότητας διαφορετικών μοντέλων και των παραμέτρων τους, χρειάζεται να έχουμε κάποια αριθμητικά μέτρα της απόδοσης τους,

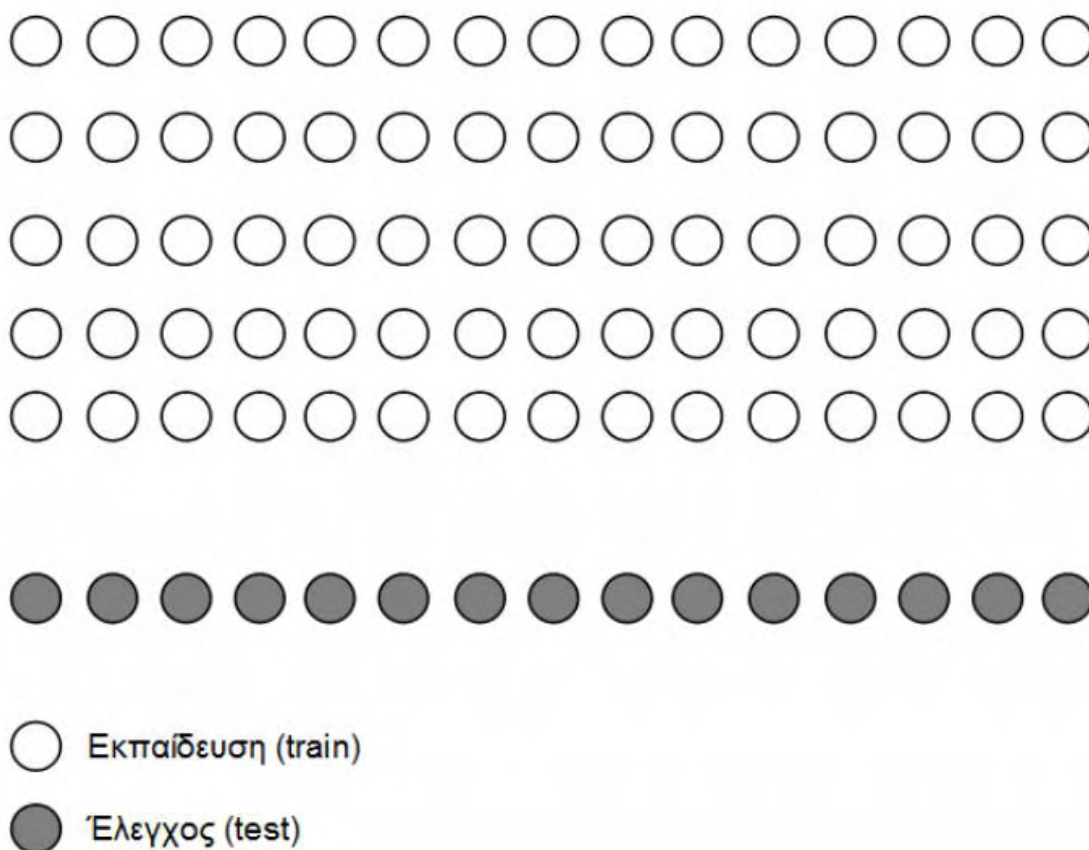
αξιολογώντας τη δυνατότητα να επιτελούν το σκοπό τους αποτελεσματικά. Για το σκοπό αυτό χρησιμοποιήσαμε κατάλληλες μεθόδους ελέγχου και στατιστικά μέτρα.

5.6.1 Μέθοδοι Ελέγχου

Οι δοκιμές βασίστηκαν στις παρακάτω μεθόδους ελέγχου.

Self-consistency

Ο πιο απλός τρόπος να ελεγχθεί η απόδοση είναι τα ταξινομημένα δεδομένα που διαθέτουμε να μοιραστούν σε δύο χωριστά σύνολα: το σύνολο εκπαίδευσης (training set) και το σύνολο ελέγχου (test set). Το μοντέλο εκπαιδεύεται με βάση τα δεδομένα του συνόλου εκπαίδευσης, και λαμβάνουμε τις εκτιμήσεις του για τα παραδείγματα του συνόλου ελέγχου. Συγκρίνοντας τις εκτιμήσεις του μοντέλου με τις σωστές προβλέψεις, υπολογίζουμε το μέτρο της απόδοσης του μοντέλου. Αυτή η μέθοδος ενδέχεται να δώσει μεροληπτικά αποτελέσματα καθώς υπάρχει ο κίνδυνος της υπερ-προσαρμογής (over-fitting) δηλαδή να «εκπαιδευτεί» αρκετά καλά από το σύνολο εκπαίδευσης αλλά να μην αποδίδει καλά σε νέα δεδομένα.



Εικόνα 20: Ένα παράδειγμα με το σύνολο εκπαίδευση και το σύνολο ελέγχου.¹

k-fold Cross-Validation

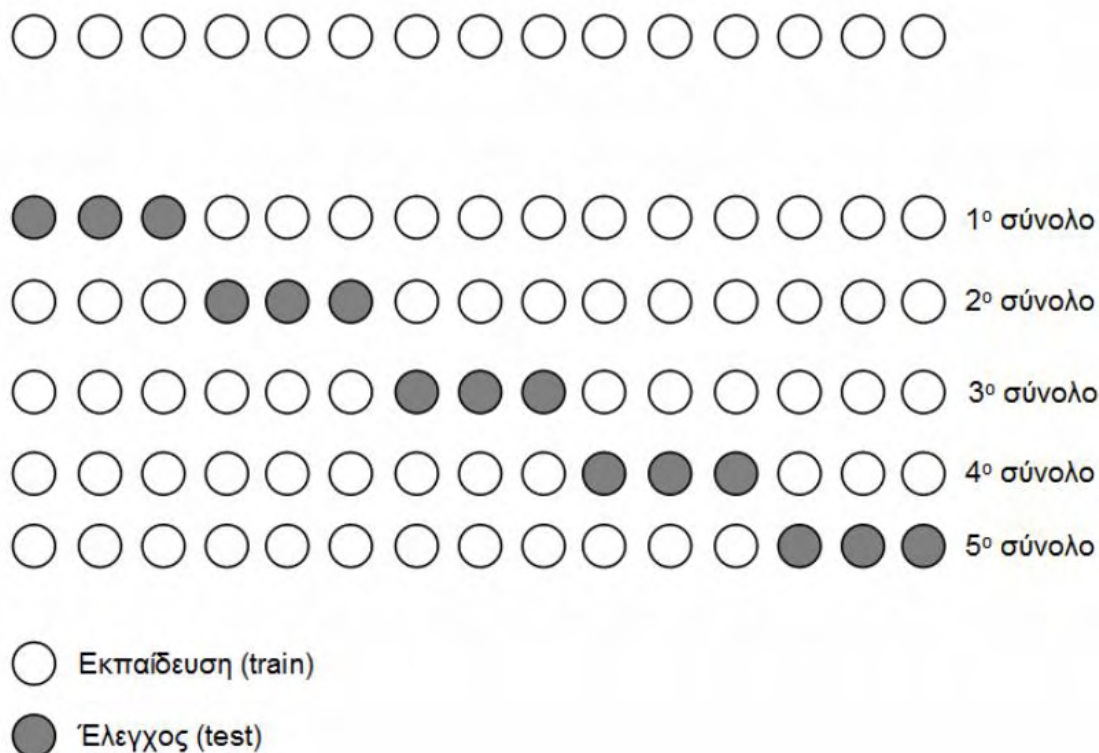
Μια εναλλακτική μέθοδος πετυχαίνει να χρησιμοποιούνται όλα τα δεδομένα του διαθέσιμου συνόλου δεδομένων ως δεδομένα εκπαίδευσης και δεδομένα ελέγχου, σε διαφορετικές φάσεις, εξασφαλίζοντας όμως ότι στο ίδιο πείραμα δεν θα χρησιμοποιηθούν γνωστά στο μοντέλο

¹ Μπάγκος, Π., 2015. Βιοπληροφορική. [ηλεκτρ. βιβλ.] Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. Διαθέσιμο στο: <http://hdl.handle.net/11419/5016>

δεδομένα ως δεδομένα ελέγχου. Αυτή η μέθοδος ονομάζεται k-fold cross-validation) [27] και λειτουργεί ως εξής:

1. Επιλέγεται θετικός ακέραιος k , και το σύνολο δεδομένων χωρίζεται σε k υποσύνολα ίσου μεγέθους.
2. Η διαδικασία επαναλαμβάνεται k φορές, όπου σε κάθε επανάληψη, ένα από τα k υποσύνολα του συνόλου δεδομένων χρησιμοποιείται ως σύνολο ελέγχου, ενώ τα εναπομείναντα υποσύνολα χρησιμοποιούνται ως σύνολο εκπαίδευσης. Σε κάθε επανάληψη μετριέται το μέτρο της απόδοσης χωρίς καμία ακολουθία να έχει χρησιμοποιηθεί στην κατασκευή της μεθόδου με την οποία έγινε η πρόβλεψη.
3. Εφόσον, ολοκληρωθούν οι k επαναλήψεις, το συνολικό μέτρο απόδοσης είναι ο μέσος όρος των μέτρων που προέκυψαν.

Ένα θέμα αποτελεί η κατάλληλη επιλογή της τιμής του k . Η τιμή αυτή θα πρέπει να μην είναι τόσο μεγάλη ώστε το μέγεθος του συνόλου ελέγχου να είναι πολύ μικρό (αυτό έχει μέγεθος το $1/k$ των συνολικών δεδομένων). Επίσης, όσο μεγαλύτερη είναι η τιμή του k τόσο περισσότερο χρόνο θα διαρκεί η εκτέλεση των δοκιμών, αφού γίνονται k δοκιμές.



Εικόνα 21: Ένα παράδειγμα με το σύνολο εκπαίδευσης και το σύνολο ελέγχου χωρισμένο κατάλληλα για την εφαρμογή της διαδικασίας ελέγχου cross-validation με $k=3$.²

Jackknife

Μια παραλλαγή της παραπάνω μεθόδου η οποία όμως απαιτεί περισσότερους υπολογισμούς, είναι η μέθοδος Jackknife όπου το k επιλέγεται να είναι ίσο με μέγεθος του συνόλου εκπαίδευσης, με αποτέλεσμα το κάθε υποσύνολο να έχει μέγεθος ίσο με ένα. Γενικά σε σύνολα με μέτριο μέγεθος η συγκεκριμένη μέθοδος επιλέγεται γιατί κάθε φορά το σύνολο εκπαίδευσης

² Μπάγκος, Π., 2015. Βιοπληροφορική. [ηλεκτρ. βιβλ.] Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. Διαθέσιμο στο: <http://hdl.handle.net/11419/5016>

είναι όσο μεγαλύτερο γίνεται. Σε περιπτώσεις που έχουμε μεγάλο σύνολο δεδομένων ή η μέθοδος είναι αργή δεν προτείνεται η εφαρμογή του.

5.6.2 Μέτρα αξιοπιστίας

Για να μετρήσουμε την αξιοπιστία της επίδοσης για κάθε μέθοδο χρησιμοποιήσαμε τα παρακάτω μέτρα. Αρχικά τα δεδομένα μπορούν να αναπαρασταθούν με ένα πίνακα συνάφειας (contingency tables) όπως αποτυπώνεται στην Εικόνα 23. Η πιο απλή περίπτωση ενός πίνακα συνάφειας είναι ο πίνακας 2x2 που προκύπτει από την ταξινόμηση των κατηγοριών δύο δίτιμων κατηγορικών μεταβλητών. Με αυτόν τον τρόπο, συμβολίζουμε με TP (True Positive) τον αριθμό των ορθώς θετικά προσδιορισμένων καταλοίπων, TN (True Negatives) τον αριθμό των ορθώς αρνητικά προσδιορισμένων καταλοίπων, FN (False Negatives) τον αριθμό των εσφαλμένων αρνητικά προσδιορισμένων καταλοίπων και FP (False Positives) τον αριθμό των εσφαλμένων θετικά προσδιορισμένων καταλοίπων. Στην περίπτωση που αναφερόμαστε στην κατάταξη των ακολουθιών, ως παρατηρήσεις χαρακτηρίζονται οι πρωτεΐνες [28].

		<u>True Class</u>		
		Positive	Negative	
<u>Predicted Class</u>	Positive	True Positive TP	False Positive FP	Positive Predictive Value (PPV) TP/(TP+FP)
	Negative	False Negative FN	True Negative TN	Negative Predictive Value (NPV) TN/(FN+TN)
		Sensitivity TP/(TP+FN)	Specificity TN/FP+TN)	Accuracy (TP+TN)/(TP+TN+FP+FN)

Εικόνα 22: Πίνακας συνάφειας για την κατάταξη των προβλεπόμενων πρωτεϊνών.³

Σε αυτές τις περιπτώσεις μας ενδιαφέρει να μελετήσουμε την ευαισθησία (Sensitivity), η οποία μετράει την το ποσοστό των ορθώς θετικά προβλέψεων και την ειδικότητα (specificity) η οποία μετράει το ποσοστό των ορθώς αρνητικά προβλέψεων.

Ποσοστό σωστά προβλεπόμενων καταλοίπων

Ένα από τα μέτρα το οποίο χρησιμοποιείτε αρκετά συχνά είναι το συνολικό ποσοστό των σωστά προβλεπόμενων καταλοίπων (Q)

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

³ Μπάγκος, Π., 2015. Βιοπληροφορική. [ηλεκτρ. βιβλ.] Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. Διαθέσιμο στο: <http://hdl.handle.net/11419/5016>

Στις μεθόδους πρόγνωσης διαμεμβρανικών τμημάτων αξιολογήσαμε το ποσοστό των σωστά προβλεπόμενων καταλοίπων σε λειτουργία δύο καταστάσεων όπου ένα αμινοξικό κατάλοιπο, στην πρόγνωση που μας ενδιαφέρει, έχει δύο δυνατές καταστάσεις, δηλαδή να είναι διαμεμβρανικό (T) ή όχι (-). Τότε ανάλογα μπορούμε να υπολογίσουμε το δείκτη

$$Q2 = 100 \times \frac{\sum_{j=1}^2 C^j}{N}$$

όπου C^j το πλήθος των καταλοίπων που προβλέφθηκαν σωστά σε μια δεδομένη κατάσταση j (T ή -) και N το συνολικό πλήθος καταλοίπων της ακολουθίας

Ποσοστό σωστά προσδιορισμένων τοπολογιών

Ακόμη πιο σημαντικό αποτελεί ο προσδιορισμός της θέσης αλλά και της τοπολογίας διαμεμβρανικών τμημάτων σε μεμβρανικές πρωτεΐνες. Αρχικά μετρήσαμε το συνολικό πλήθος διαμεμβρανικών τμημάτων στην πρωτεΐνη και στη συνέχεια υπολογίσαμε την κατανομή του πλήθους διαμεμβρανικών τμημάτων συναρτήσει με το τον συνολικό αριθμό των ακολουθιών ως το ποσοστό των σωστά προσδιορισμένων τοπολογιών (Correct Ori).

Segment Overlap Score (SOV)

Το μέτρο των επικαλυπτόμενων τμημάτων (SOV) θεωρείται ένας από τους πιο αξιόπιστους δείκτες για την μέτρηση της αξιοπιστίας των αλγορίθμων πρόγνωσης δευτεροταγούς δομής [29]. Το μέτρο SOV βασίζεται στη μέση επικάλυψη μεταξύ των παρατηρούμενων και των προβλεπόμενων τμημάτων. Για παράδειγμα, ο ορισμός του μέτρου SOV για α-έλικες είναι η εξής:

$$Sov(i) = \frac{1}{N(i)} \sum_{s(i)} \left[\frac{\min OV(s_1, s_2) + \delta(s_1, s_2)}{\max OV(s_1, s_2)} \times \text{len}(s_1) \right]$$

Όπου, s_1 και s_2 είναι τα τμήματα από τις παρατηρούμενες και τις προβλέψιμες ακολουθίες για κάθε κατάσταση i του α-έλικα (H, E ή C), S είναι ο αριθμός όλων των ζευγών τμήματος (s_1, s_2), όπου s_1 και s_2 έχουν τουλάχιστον ένα κατάλοιπο κοινό σε μια κατάσταση α-έλικα, $\min OV(s_1, s_2)$ είναι το μήκος της πραγματικής επικάλυψης των s_1 και s_2 και $\max OV(s_1, s_2)$ είναι το μήκος της συνολικής έκτασης για την οποία για κάθε ένα από τα τμήματα s_1 ή s_2 έχει ένα κατάλοιπο σε μια κατάσταση α-έλικα. $N(i)$ είναι ο συνολικός αριθμός κατάλοιπων αμινοξέων που παρατηρείται στην διαμόρφωση μιας κατάστασης α-έλικας. Ο ορισμός της $\delta(s_1, s_2)$ έχει ως εξής

$$\delta(s_1, s_2) = \min \begin{cases} \max OV(s_1, s_2) - \min OV(s_1, s_2) \\ \min OV(s_1, s_2) \\ \text{int}(\text{len}(s_1) \times 0.5) \\ \text{int}(\text{len}(s_2) \times 0.5) \end{cases}$$

Το μέτρο Συνολικά για όλες τις τρεις καταστάσεις της α-έλικας (H, E ή C), ορίζεται ως εξής

$$Sov = 100 \times \frac{1}{N} \sum_{i \in \{H, E, C\}} \sum_{s(i)} \left[\frac{\min OV(s_1, s_2) + \delta(s_1, s_2)}{\max OV(s_1, s_2)} \times \text{len}(s_1) \right]$$

Το μέτρο που ορίζεται στον προηγούμενο ορισμό μπορεί εύκολα να επεκταθεί για την αξιολόγηση πολλών καταστάσεων που περιέχει μιας δευτεροταγής δομή.

ποσοστό σωστά ταξινομημένων σημείων αποκοπής

Εξίσου σημαντικό αποτελεί και το ποσοστό των σωστά ταξινομημένων σημείων αποκοπής (Cleavage site)

Συντελεστής συσχέτισης του Matthews

Ο συντελεστής συσχέτισης του Matthews (C), συνοψίζει σε ένα ενιαίο μέτρο τα αληθώς θετικά (TP), ψευδώς θετικά (FP), αληθώς αρνητικά (TN) και ψευδώς αρνητικά (FN)[30].

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

5.7 Υλοποίηση

Κατά την εκπαίδευση ενός HMM με σημασμένες ακολουθίες (labeled sequences), μπορούμε είτε να επιλέξουμε να εκπαιδεύσουμε το μοντέλο σύμφωνα με το κριτήριο της Μέγιστης Πιθανοφάνειας (ML), ή να εφαρμόσουμε αυτήν της Δεσμευμένης Μέγιστης Πιθανοφάνειας (Conditional ML) εκπαίδευση που φαίνεται να έχει καλύτερες επιδόσεις σε διάφορες εφαρμογές. Η εκπαίδευση των μοντέλων με τις νέες κωδικοποιήσεις, έγινε με δύο τρόπους, είτε χρησιμοποιώντας τον αλγόριθμο Baum-Welch βασισμένο στο κριτήριο ML για επισημασμένες αλληλουχίες είτε χρησιμοποιώντας τον αλγόριθμο gradient-descent βασισμένο στο κριτήριο CML για επισημασμένες αλληλουχίες. Για την αποκωδικοποίηση χρησιμοποιήσαμε τη μέθοδο Viterbi.

Όλες οι παραπάνω μέθοδοι εφαρμόστηκαν σε προϋπάρχουσα υλοποίηση η οποία αποτελεί ένα ολοκληρωμένο πακέτο που έχει διαθέσιμες όλες τις γνωστές μεθόδους που καλύπτουν τα HMM. Για την κωδικοποίηση του αλφαβήτου οι υλοποιημένες μέθοδοι χρησιμοποιούσαν την ASCII κωδικοποίηση. Λόγω περιορισμών μεγέθους όμως, χρησιμοποιήσαμε το διεθνές πρότυπο κωδικοποίησης Unicode για να καλύψουμε το μεγαλύτερο εύρος αλφαβήτου που θα χρειαστεί να επεκτείνουμε.

Για την εφαρμογή της κωδικοποίησης λάβαμε υπόψη θέματα χωρητικότητας, συμβατότητας πηγαίου κώδικα και διαλειτουργικότητας με άλλα συστήματα. Για το συγκεκριμένο λόγο, χρησιμοποιήσαμε τη κωδικοποίηση UTF-16 η οποία έχει εκτεταμένη χρήση και έχει υιοθετηθεί σε πολλά λειτουργικά συστήματα και γλώσσες προγραμματισμού. Ως σημείο έναρξης εισαγωγής του πρώτου χαρακτήρα ορίσαμε το Αγγλικό Κεφαλαίο Γράμμα Α που αντιστοιχεί στην ακέραια θέση 65 του Πίνακα Χαρακτήρων. Για την δημιουργία των νέων κωδικοποιήσεων αλλά την επέκταση των παραμέτρων και συνόλων εκπαίδευσης χρησιμοποιήσαμε την Perl. Η διαδικασία είναι πλήρως αυτοματοποιημένη και μας δίνει τη δυνατότητα να κάνουμε περισσότερες δοκιμές και σε μεγαλύτερης τάξεως μοντέλα.

6. Αποτελέσματα

Όπως αναφέρθηκε στην προηγούμενη ενότητα για να εξεταστεί η αποτελεσματικότητα των νέων μοντέλων πραγματοποιήθηκε ένα σύνολο δοκιμών για κάθε μία από τις πέντε προγνωστικές μεθόδους οι οποίες βασίστηκαν σε διαφορετικές μεθόδους δοκιμών και διαφορετικά σύνολα παραμέτρων χρησιμοποιώντας 2^{ης} τάξεως μοντέλα.

Σε αυτήν την ενότητα θα παρουσιάσουμε τα αποτελέσματα που παρουσίασαν το μεγαλύτερο ενδιαφέρον ως προς την απόδοσή τους. Όπως αναφέραμε, οι δοκιμές βασίστηκαν στις μεθόδους self-consistency, Jackknife και Cross-Validation. Οι δοκιμές με self-consistency παρουσίασαν σε όλες τις περιπτώσεις βελτίωση για όλους τους δείκτες αποδοτικότητας. Τα αποτελέσματα αυτά πιθανόν δείχνουν συμπτώματα υπερ-προσαρμογής (over-fitting), που αποτελεί και χαρακτηριστικό της συγκεκριμένης μεθόδου. Η μέθοδος αν και δεν αποτελεί αντικειμενικό κριτήριο, είναι χρήσιμη για τη δοκιμή ενός νέου αλγορίθμου. Όλα τα αποτελέσματα καταγράφηκαν και παρουσιάζονται, μετά από μια σειρά δοκιμών που εφαρμοστήκαν στα μοντέλα χρησιμοποιώντας την προκαθορισμένη και τις νέες κωδικοποιήσεις.

6.1 Σύνολο Παραμέτρων και Παρατηρήσεων

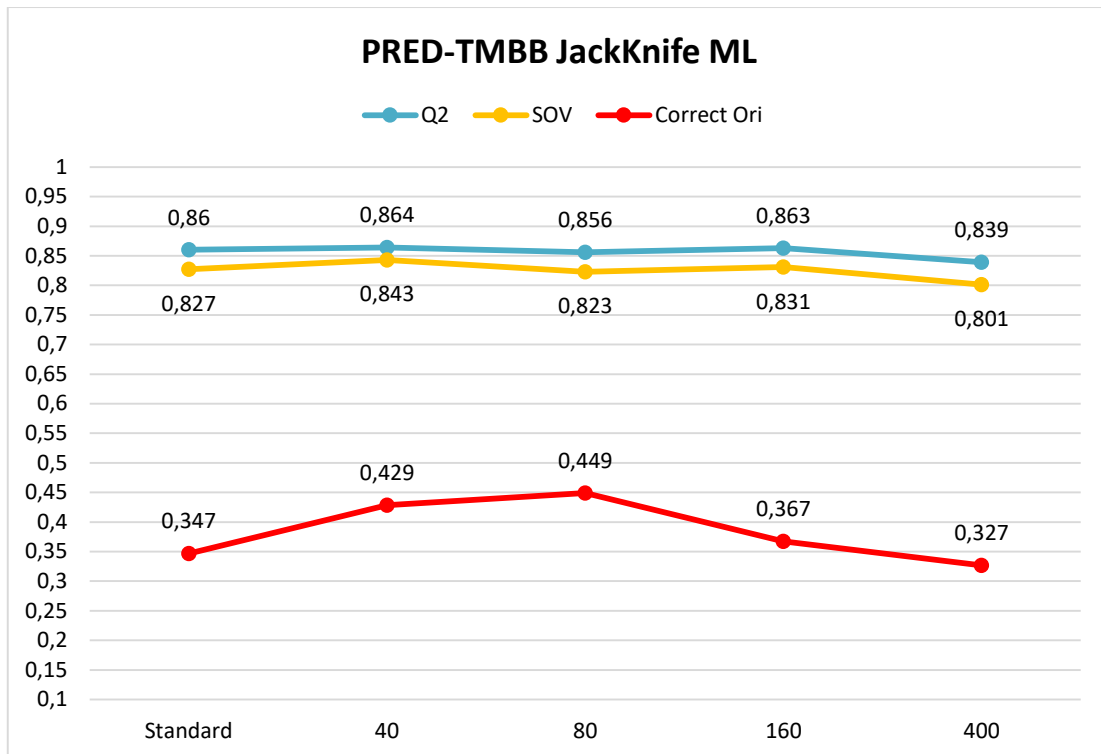
Λαμβάνοντας υπόψη από προηγούμενες αναφορές, τη σημαντικότητα του μεγέθους του συνόλου εκπαίδευσης και τον αριθμό των παραμέτρων, τις καταγράψαμε και τις παρουσιάζουμε αντίστοιχα για να τις εξετάσουμε. Ο πίνακας 1 δείχνει τα σύνολα παραμέτρων (S) και συνόλων εκπαίδευσης/παρατηρήσεων (T) για τα αρχικά και τα νέα μοντέλα. Το σύνολο των παραμέτρων αποτελείται από τις παραμέτρους που έχουν πιθανότητα (όχι τις μηδενικές).

Μοντέλο	PRED-TMBB		HMM-TM		PRED-TAT		PRED-LIPO		PRED-SIGNAL	
	Παράμετροι (S)	T/S	Παράμετροι (S)	T/S	Παράμετροι (S)	T/S	Παράμετροι (S)	T/S	Παράμετροι (S)	T/S
Παρατηρήσεις (T)	17295		17537		65148		23422		21479	
Standard	184	93.995	283	61.968	632	103.082	626	37.415	414	51.882
40	324	53.380	403	43.516	1192	54.654	1186	19.749	774	27.751
80	604	28.634	643	27.274	2312	28.178	2306	10.157	1494	14.377
160	1164	14.858	1123	15.616	4552	14.312	4546	5.152	2934	7.321
400	2844	6.081	2563	6.842	11272	5.780	11266	2.079	7254	2.961

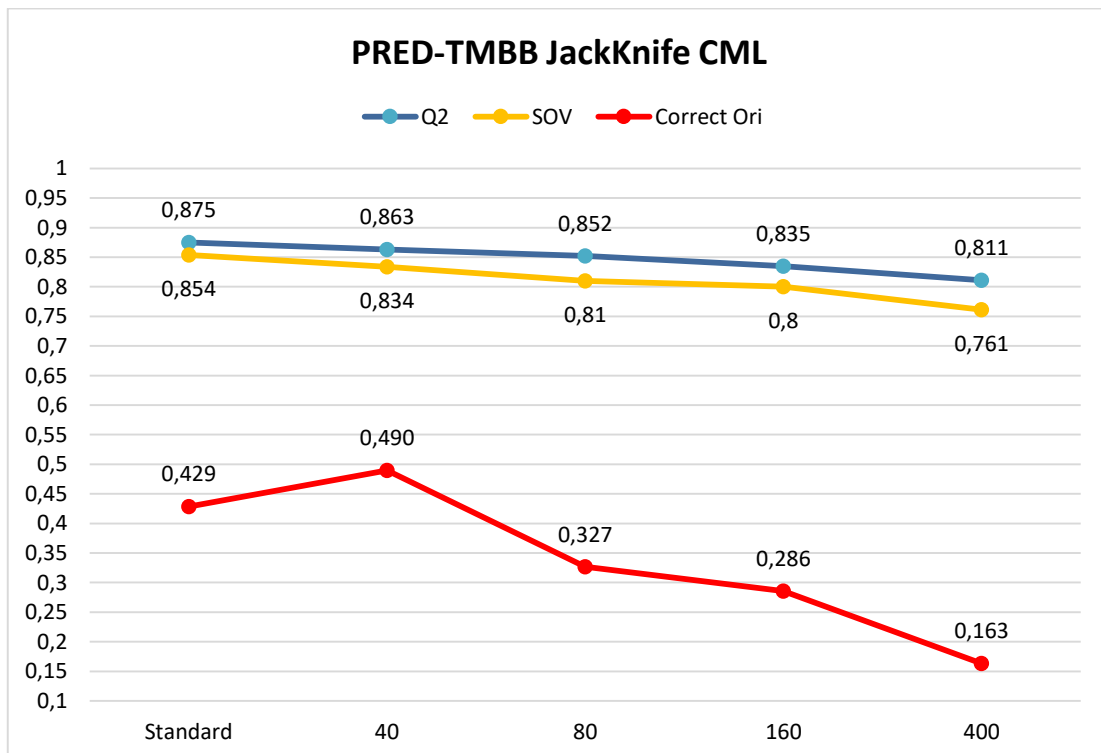
Πίνακας 1. Το σύνολο των παραμέτρων και των δεδομένων εκπαίδευσης των μοντέλων.

6.2 PRED-TMBB

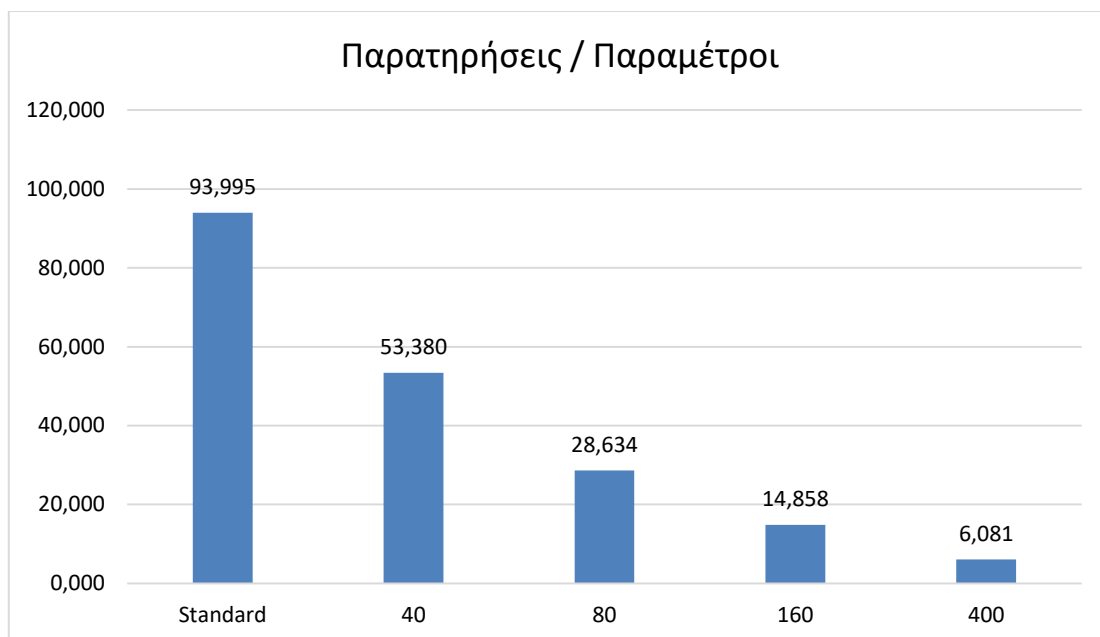
Τα αποτελέσματα που προέκυψαν από τη μέθοδο (PRED-TMBB2) για τα μοντέλα και τις αντίστοιχες κωδικοποιήσεις εφαρμόζοντας τη διαδικασία Jackknife, για το κριτήριο της Μέγιστης Πιθανοφάνειας (ML) και της Δεσμευμένης Μέγιστης Πιθανοφάνειας (Conditional ML) παρουσιάζονται στα παρακάτω γραφήματα μαζί με τα σύνολα των παρατηρήσεων προς των παραμέτρων. Η αποδοτικότητα των μοντέλων με τις νέες κωδικοποιήσεις αυξήθηκε ελάχιστα. Η πιο αξιοσημείωτη βελτίωση είναι στο ποσοστό των σωστά προσδιορισμένων τοπολογιών. Είναι και λογικό από βιολογικής πλευράς επειδή στο συγκεκριμένο βιολογικό πρόβλημα συχνά συναντάμε εναλλαγές υδρόφοβων-υδρόφιλων καταλοίπων. Παρατηρείται βέβαια, ειδικά για το κριτήριο ML, μια σταθερή απόδοση με σχετικά μικρό αριθμό παραμέτρων και παρατηρήσεων.



Εικόνα 23: Η σύγκριση του αρχικού μοντέλου PRED-TMBB2 με επανεκπαίδευση και των νέων μοντέλων που χρησιμοποιούν πρόσθετη πληροφορία, σχετικά με την ακρίβεια των πιο σημαντικών μέτρων αξιοπιστίας (Q2, SOV, Correct Ori) των jackknife δοκιμών για το μοντέλο HMM-TM για το κριτήριο της Μέγιστης Πιθανοφάνειας (ML).



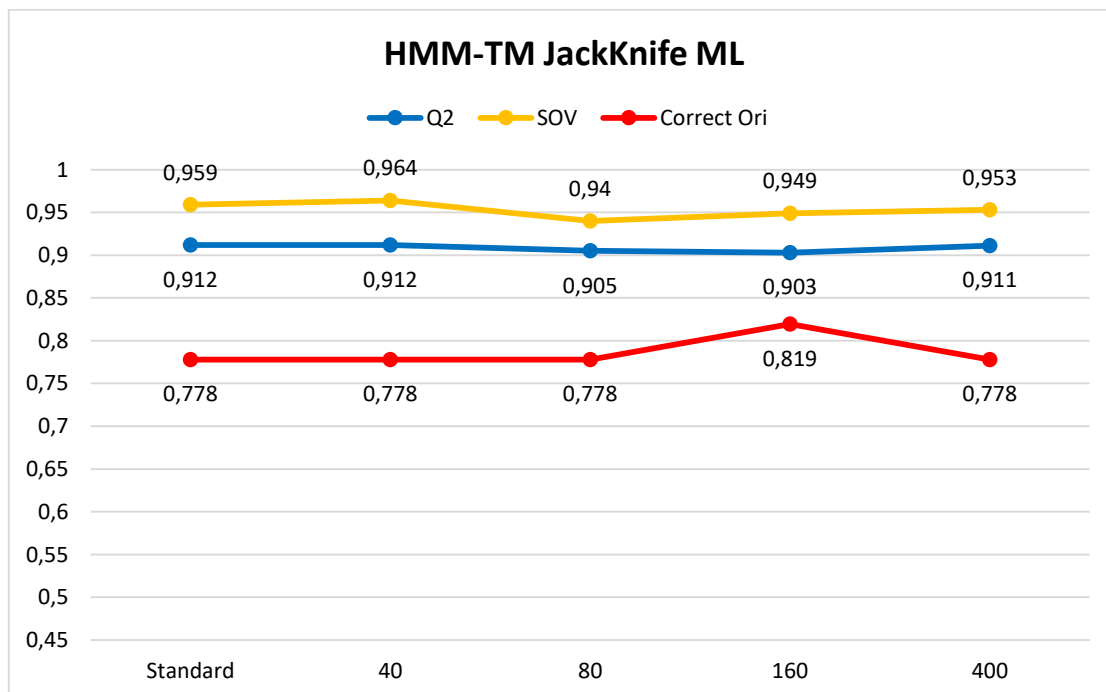
Εικόνα 24: Η σύγκριση του αρχικού μοντέλου PRED-TMBB2 με επανεκπαίδευση και των νέων μοντέλων που χρησιμοποιούν πρόσθετη πληροφορία, σχετικά με την ακρίβεια των πιο σημαντικών μέτρων αξιοπιστίας (Q2, SOV, Correct Ori) των jackknife δοκιμών για το μοντέλο HMM-TM για το κριτήριο της Μέγιστης Πιθανοφάνειας (CML).



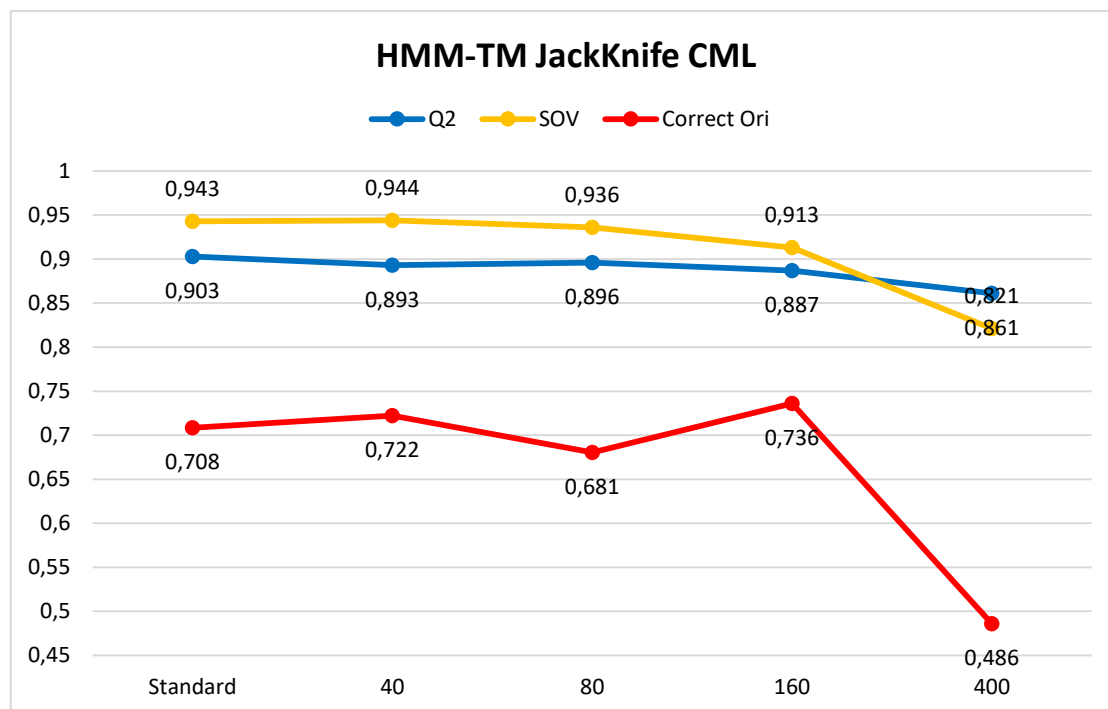
Εικόνα 25: Απεικόνιση γραφικής παράστασης αποτυπώνοντας το λόγο του συνόλου παρατηρήσεων προς το σύνολο παραμέτρων τα οποία χρησιμοποιήθηκαν στις αντίστοιχες δοκιμές για το μοντέλο PRED-TMBB2.

6.3 HMM-TM

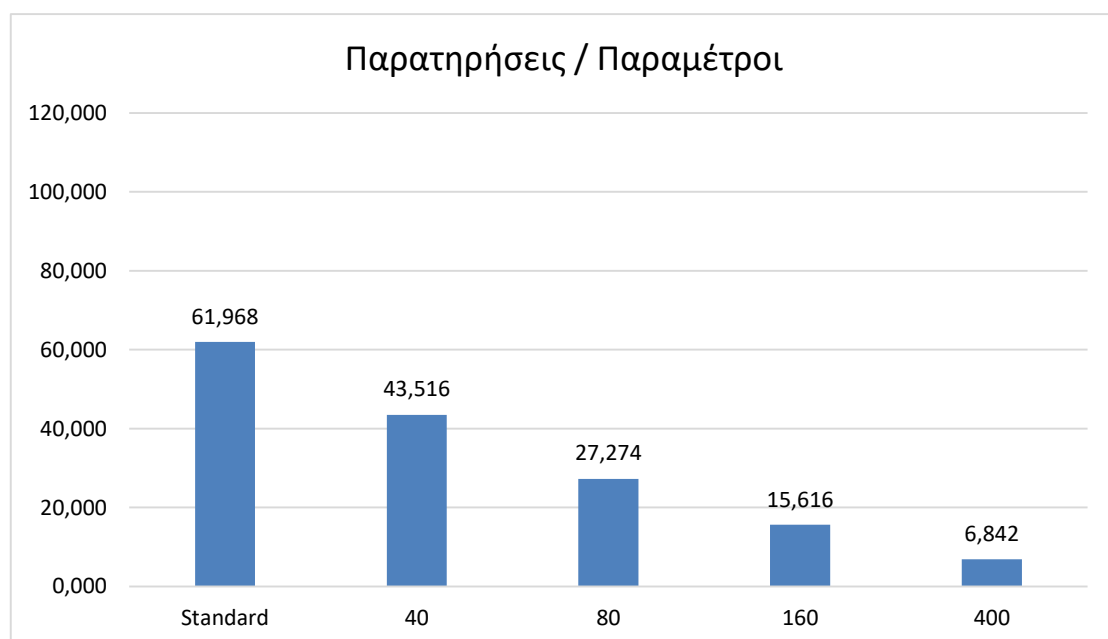
Τα αποτελέσματα που προέκυψαν από τη μέθοδο (HMM-TM) για τα μοντέλα και τις αντίστοιχες κωδικοποιήσεις εφαρμόζοντας τη διαδικασία Jackknife αντίστοιχα, για το κριτήριο της Μέγιστης Πιθανοφάνειας (ML) και της Δεσμευμένης Μέγιστης Πιθανοφάνειας (Conditional ML) παρουσιάζονται στα παρακάτω γραφήματα μαζί με τα σύνολα των παρατηρήσεων προς των παραμέτρων. Η αποδοτικότητα των μοντέλων με τις νέες κωδικοποιήσεις φαίνεται να έχει σταθερή απόδοση και κάποια μέτρα δείχνουν να βελτιώνονται. Τα αποτελέσματα για το κριτήριο ML φαίνεται να έχουν πιο σταθερή απόδοση.



Εικόνα 26: Η σύγκριση του αρχικού μοντέλου HMM-TM με επανεκπαίδευση και των νέων μοντέλων που χρησιμοποιούν πρόσθετη πληροφορία, σχετικά με την ακρίβεια των πιο σημαντικών μέτρων αξιοπιστίας (Q2, SOV, Correct Ori) των jackknife δοκιμών για το μοντέλο HMM-TM για το κριτήριο της Μέγιστης Πιθανοφάνειας (ML).



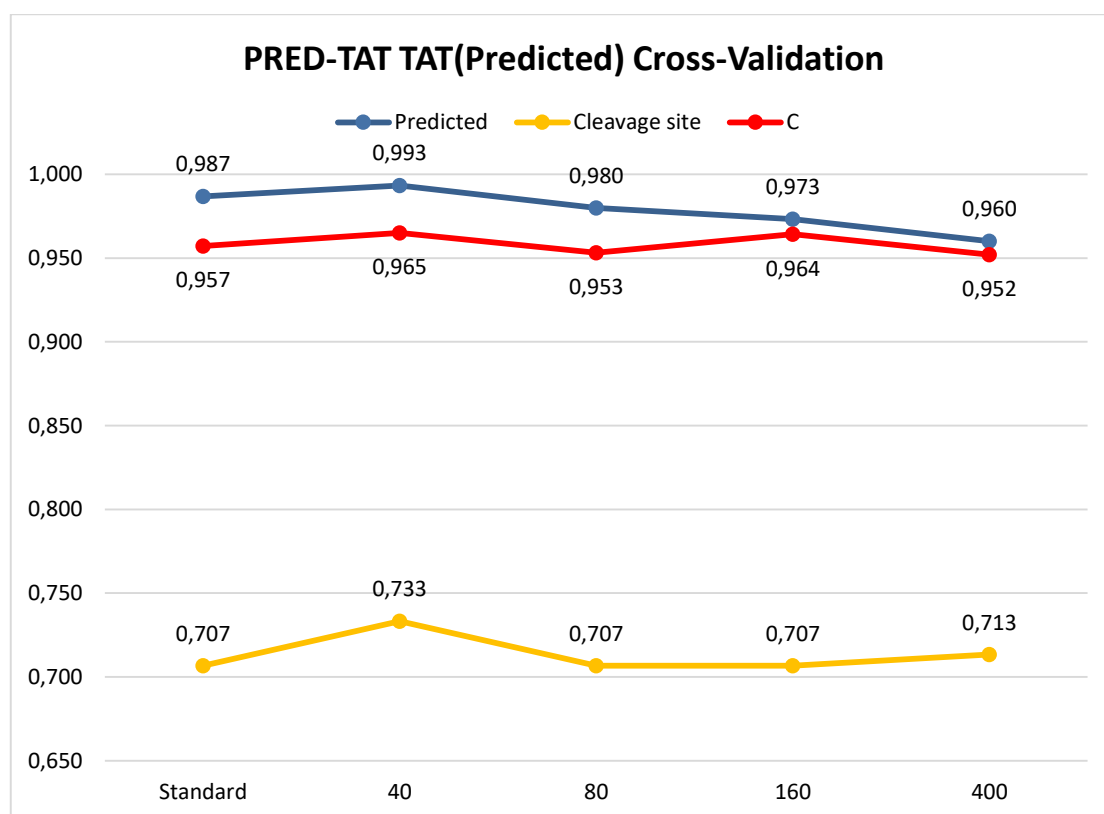
Εικόνα 27: Η σύγκριση του αρχικού μοντέλου HMM-TM με επανεκπαίδευση και των νέων μοντέλων που χρησιμοποιούν πρόσθετη πληροφορία, σχετικά με την ακρίβεια των πιο σημαντικών μέτρων αξιοπιστίας (Q2, SOV, Correct Ori) των jackknife δοκιμών για το μοντέλο HMM-TM για το κριτήριο της Μέγιστης Πιθανοφάνειας (CML).



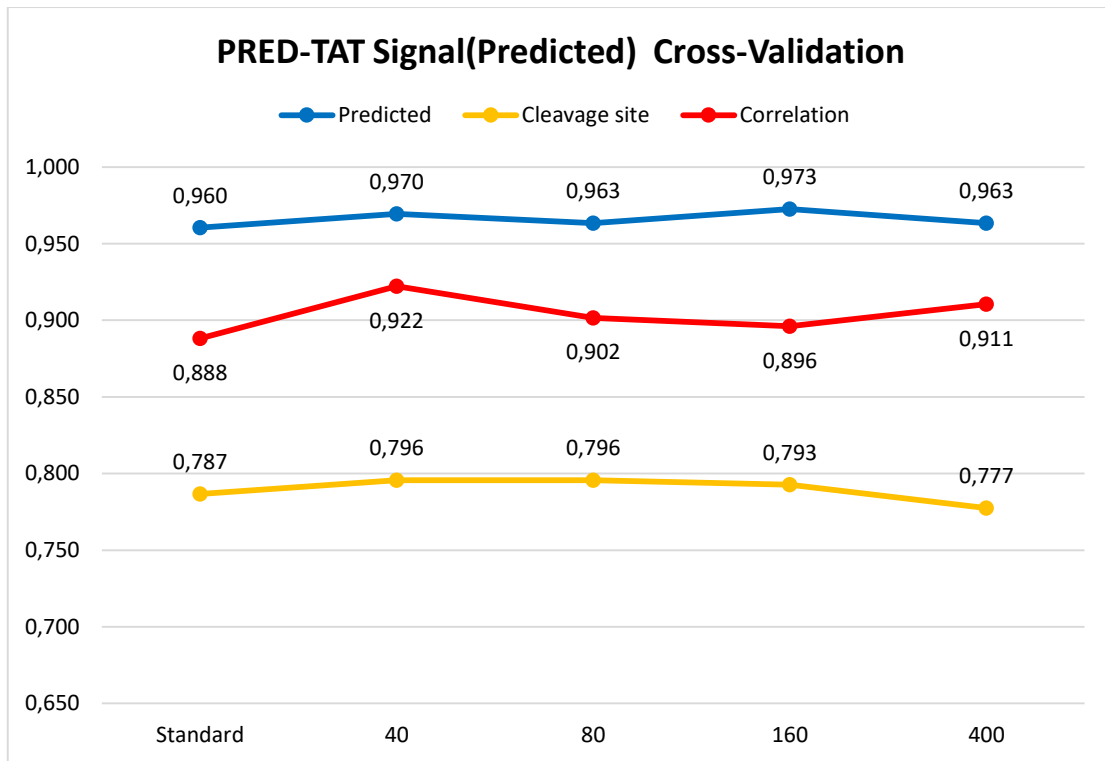
Εικόνα 28: Απεικόνιση γραφικής παράστασης αποτυπώνοντας το λόγο του συνόλου παρατηρήσεων προς το σύνολο παραμέτρων τα οποία χρησιμοποιήθηκαν στις αντίστοιχες δοκιμές για το μοντέλο HMM-TM.

6.4 PRED-TAT

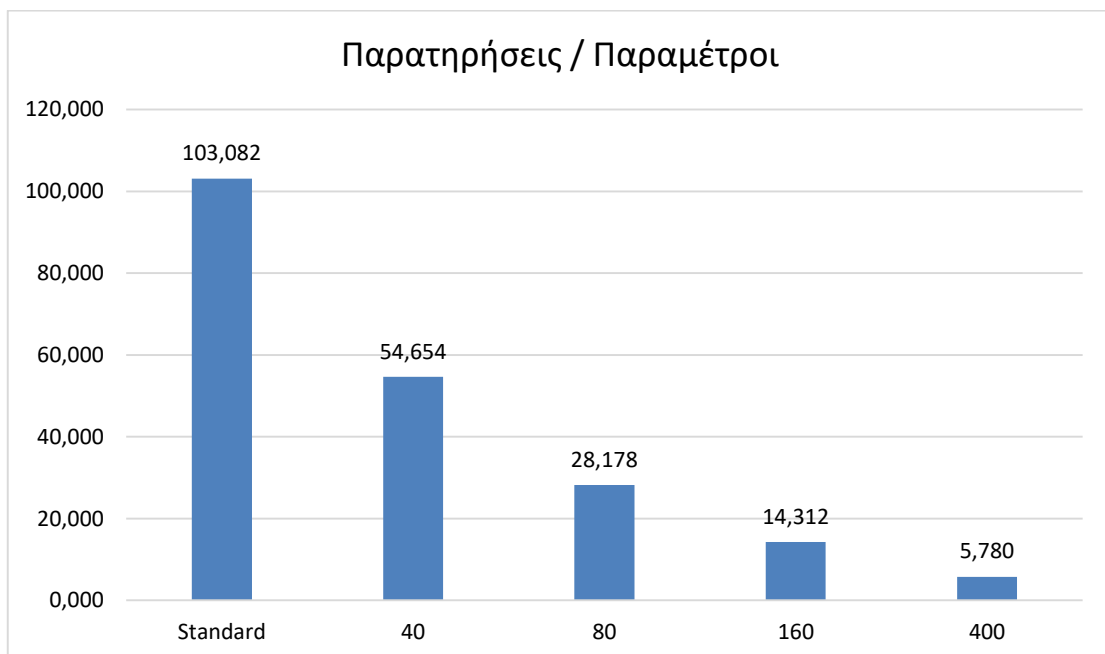
Τα αποτελέσματα που προέκυψαν από τη μέθοδο (PRED-TAT) για τα μοντέλα και τις αντίστοιχες κωδικοποιήσεις εφαρμόζοντας τη διαδικασία cross-validation 31-υποσυνόλων παρουσιάζονται στα παρακάτω γραφήματα μαζί με τα σύνολα των παρατηρήσεων προς τα σύνολα των παραμέτρων. Η επίδοση των μοντέλων με τις νέες κωδικοποιήσεις αυξήθηκε ελάχιστα. Επίσης παρουσιάζει μια σταθερή απόδοση στην πρόβλεψη των πεπτιδίων οδηγητών Tat και Signal. Ειδικά στην πρόβλεψη των σημείων αποκοπής (Cleavage site) φαίνεται ότι έχουν με μικρό ποσοστό καλύτερη απόδοση. Να σημειωθεί ότι το συγκεκριμένο μοντέλο χρησιμοποιεί τις λιγότερες παρατηρήσεις και το μεγαλύτερο σύνολο εκπαίδευσης.



Εικόνα 29: Η σύγκριση του αρχικού μοντέλου PRED-TAT με επανεκπαίδευση και των νέων μοντέλων που χρησιμοποιούν πρόσθετη πληροφορία, σχετικά με την ακρίβεια των πιο σημαντικών μέτρων αξιοπιστίας πρόβλεψης των πεπτιδίων οδηγητών Tat (ποσοστό των σωστά ταξινομημένων, ποσοστό των σωστά ταξινομημένων σημείων αποκοπής και του συντελεστή συσχέτισης του Matthews).



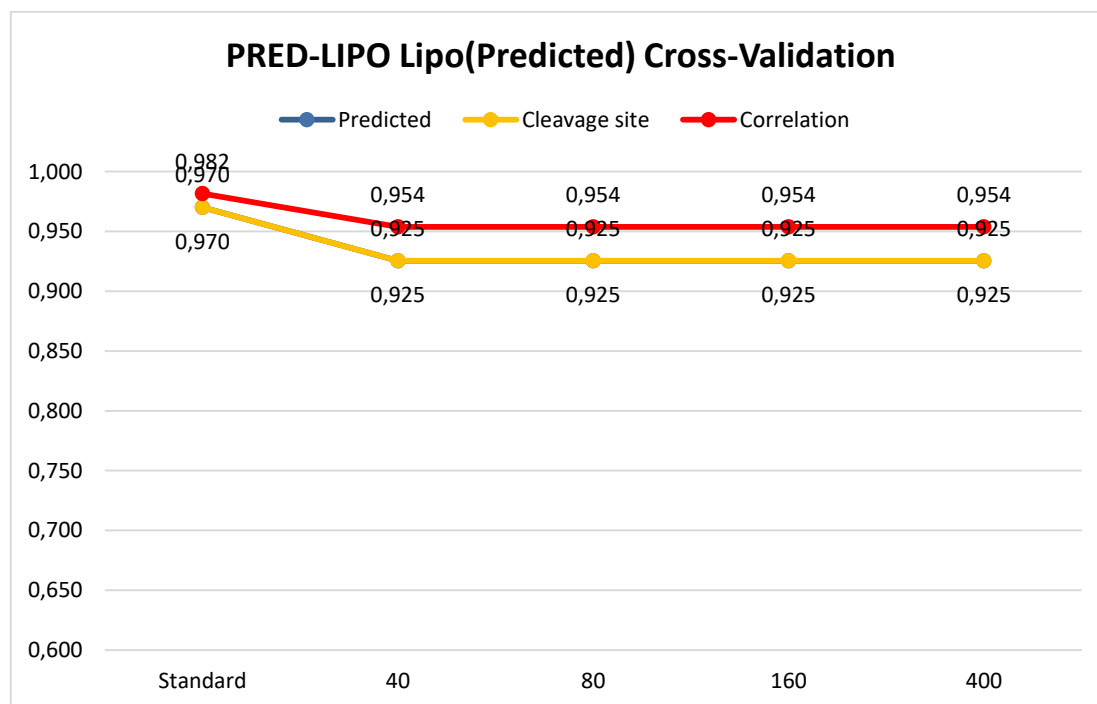
Εικόνα 30: Η σύγκριση του αρχικού μοντέλου PRED-TAT με επανεκπαίδευση και των νέων μοντέλων που χρησιμοποιούν πρόσθετη πληροφορία, σχετικά με την ακρίβεια των πιο σημαντικών μέτρων αξιοπιστίας πρόβλεψης των πεπτιδίων οδηγητών Sec Signal (ποσοστό των σωστά ταξινομημένων, ποσοστό των σωστά ταξινομημένων σημείων αποκοπής και του συντελεστή συσχέτισης του Matthews).



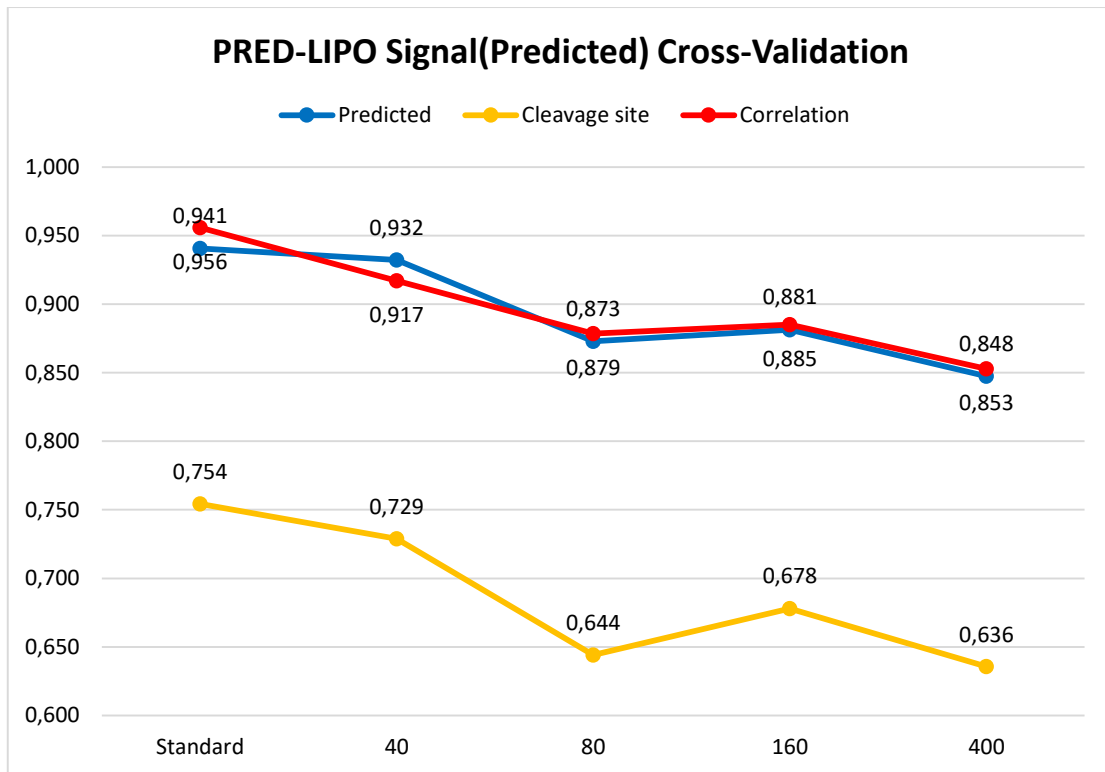
Εικόνα 31: Απεικόνιση γραφικής παράστασης αποτυπώνοντας το λόγο του συνόλου παρατηρήσεων προς το σύνολο παραμέτρων τα οποία χρησιμοποιήθηκαν στις αντίστοιχες δοκιμές για το μοντέλο PRED-TAT.

6.5 PRED-LIPO

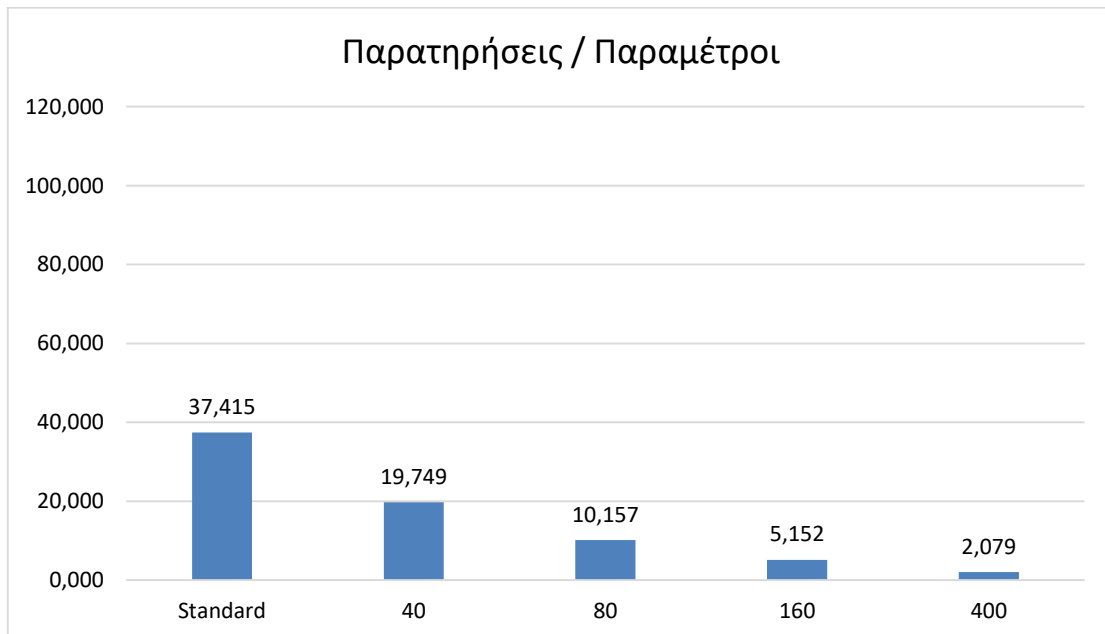
Τα αποτελέσματα που προέκυψαν από τη μέθοδο (PRED-LIPO) για τα μοντέλα και τις αντίστοιχες κωδικοποιήσεις εφαρμόζοντας τη διαδικασία cross-validation 11-υποσυνόλων παρουσιάζονται στα παρακάτω γραφήματα μαζί με τα σύνολα των παρατηρήσεων προς τα σύνολα των παραμέτρων. Η αποδοτικότητα των μοντέλων με τις νέες κωδικοποιήσεις μειώθηκε ελάχιστα. Αυτό ίσως μπορεί να εξηγηθεί εν μέρει από το μικρό αριθμό παρατηρήσεων που χρησιμοποιούνται για την εκπαίδευση, εν αντίθεσή με τις παραμέτρους που χρησιμοποιεί το μοντέλο.



Εικόνα 32: Η σύγκριση του αρχικού μοντέλου PRED-LIPO με επανεκπαίδευση και των νέων μοντέλων που χρησιμοποιούν πρόσθετη πληροφορία, σχετικά με την ακρίβεια των πιο σημαντικών μέτρων αξιοπιστίας πρόβλεψης των πεπτιδίων οδηγητών Lipo (ποσοστό των σωστά ταξινομημένων, ποσοστό των σωστά ταξινομημένων σημείων αποκοπής και του συντελεστή συσχέτισης του Matthews).



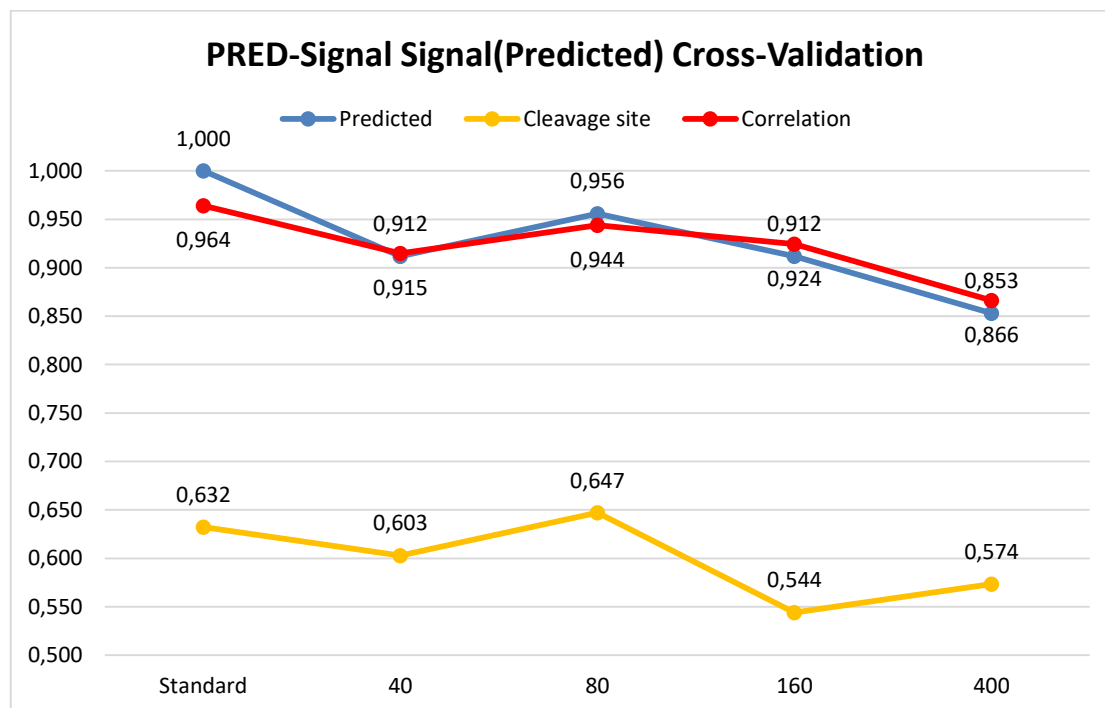
Εικόνα 33: Η σύγκριση του αρχικού μοντέλου PRED-TAT με επανεκπαίδευση και των νέων μοντέλων που χρησιμοποιούν πρόσθετη πληροφορία, σχετικά με την ακρίβεια των πιο σημαντικών μέτρων αξιοπιστίας πρόβλεψης των πεπτιδίων οδηγητών Signal (ποσοστό των σωστά ταξινομημένων, ποσοστό των σωστά ταξινομημένων σημείων αποκοπής και του συντελεστή συσχέτισης του Matthews).



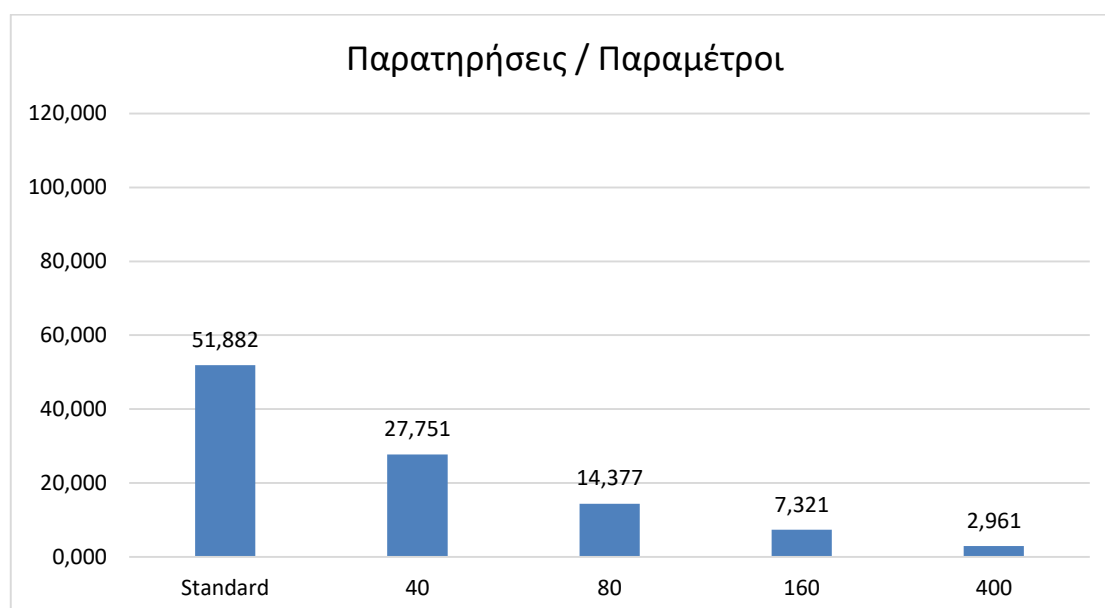
Εικόνα 34: Απεικόνιση γραφικής παράστασης αποτυπώνοντας το λόγο του συνόλου παρατηρήσεων προς το σύνολο παραμέτρων τα οποία χρησιμοποιήθηκαν στις αντίστοιχες δοκιμές για το μοντέλο PRED-LIPO.

6.6 PRED-SIGNAL

Τα αποτελέσματα που προέκυψαν από τη μέθοδο (PRED-SIGNAL) για τα μοντέλα και τις αντίστοιχες κωδικοποιήσεις εφαρμόζοντας τη διαδικασία cross-validation 9-υποσυνόλων παρουσιάζονται στα παρακάτω γραφήματα μαζί με τα σύνολα των παρατηρήσεων προς τα σύνολα των παραμέτρων. Η αποδοτικότητα των μοντέλων με τις νέες κωδικοποιήσεις μειώθηκε ελάχιστα. Αυτό ίσως μπορεί να εξηγηθεί εν μέρει από το μικρό αριθμό παρατηρήσεων που χρησιμοποιούνται για την εκπαίδευση του μοντέλου.



Εικόνα 35: Η σύγκριση του αρχικού μοντέλου PRED-TAT με επανεκπαίδευση και των νέων μοντέλων που χρησιμοποιούν πρόσθετη πληροφορία, σχετικά με την ακρίβεια των πιο σημαντικών μέτρων αξιοπιστίας πρόβλεψης των πεπτιδίων οδηγτών Tat (ποσοστό των σωστά ταξινομημένων, ποσοστό των σωστά ταξινομημένων σημείων αποκοπής και του συντελεστή συσχέτισης του Matthews).

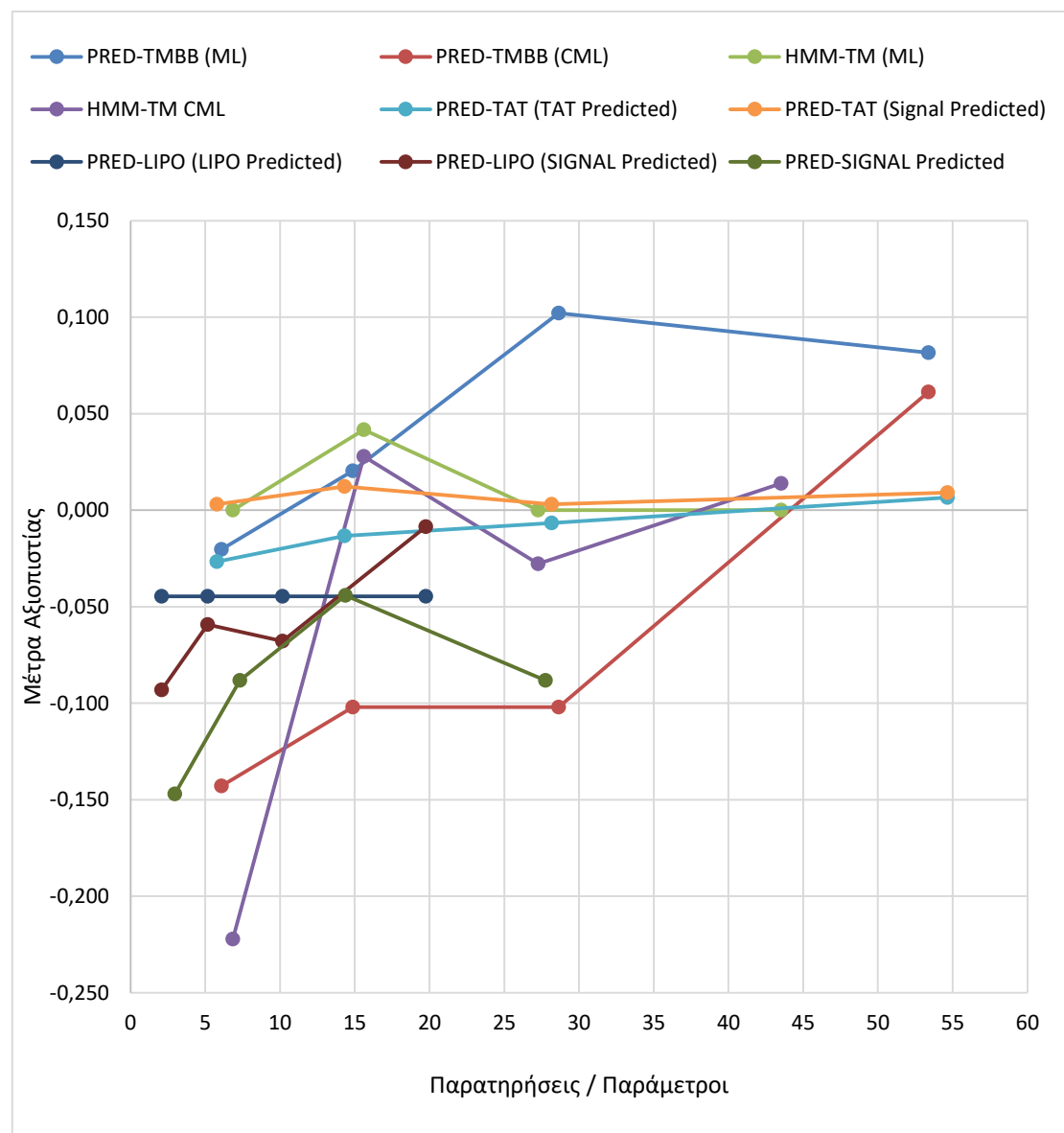


Εικόνα 36: Απεικόνιση γραφικής παράστασης αποτυπώνοντας το λόγο του συνόλου παρατηρήσεων

προς το σύνολο παραμέτρων τα οποία χρησιμοποιήθηκαν στις αντίστοιχες δοκιμές για το μοντέλο PRED-SIGNAL.

6.7 Απόδοση

Τέλος, χρησιμοποιώντας τα αποτελέσματα των δοκιμών, εκφράσαμε τη διαφορά του πιο σημαντικού μέτρου αξιοπιστίας μεταξύ των νέων μοντέλων με το αρχικό, προκειμένου να διερευνηθεί η συνολική εκτίμηση της απόδοσης των νέων μοντέλων σε σχέση με την απόδοση των αρχικών μοντέλων. Επίσης για να διερευνήσουμε τη σημαντικότητα του μεγέθους του συνόλου εκπαίδευσης και των παραμέτρων αποτυπώσαμε τη διασπορά της προηγούμενης διαφοράς με το λόγο μεταξύ παρατηρήσεων προς παραμέτρους το οποίο αποτυπώνεται στο παρακάτω γράφημα.



Εικόνα 37: Απεικόνιση γραφικής παράστασης της διασποράς που αναπαριστά την σχέση της διαφοράς των μέτρων απόδοσης μεταξύ των νέων μοντέλων με το αρχικό και το μέγεθος των αντίστοιχων παρατηρήσεων/παραμέτρων.

Τα αποτελέσματα αποδεικνύουν ότι η πρόβλεψη διαμεμβρανικών περιοχών έχει βελτιωθεί στα περισσότερα μοντέλα με τη χρήση πρόσθετης πληροφορίας. Τα μοντέλα PRED-TAT και PRED-TMBB2 που έχουν και τα πιο μεγάλα σύνολα εκπαίδευσης φαίνεται να έχουν την πιο σταθερή απόδοση για όλες τις κωδικοποιήσεις. Το μοντέλο με τα υδρόφοβα (Μοντέλο 40) φαίνεται να είναι το πιο αποδοτικό, σαν μια λύση με σχετικά μικρό αριθμό παραμέτρων, ενώ είναι και λογικό από βιολογική πλευράς.

Έχει ενδιαφέρον να σημειωθεί ότι η ακρίβεια πρόβλεψης ήταν ελαφρώς μειωμένη όταν χρησιμοποιήσαμε μικρό αριθμό δεδομένων εκπαίδευσης. Επιπλέον, σε σχετικές μελέτες γίνεται αναφορά στο γεγονός ότι όταν χρησιμοποιούμε μεγαλύτερο αριθμό παραμέτρων φαίνεται να υπάρχει συσχέτιση με το μέγεθος του συνόλου εκπαίδευσης που χρησιμοποιείται. Επομένως, αν και μερικές δοκιμές δείχνουν ότι υπάρχει βελτίωση στα περισσότερα μοντέλα, θα μπορούσαν να αποδίδουν καλύτερα αν χρησιμοποιηθεί μεγαλύτερος αριθμός δεδομένων εκπαίδευσης. Αυτό ίσως μπορεί να εξηγήσει εν μέρει ότι απέδωσαν καλύτερα τα μοντέλα που είχαν το μεγαλύτερο σύνολο εκπαίδευσης με σχετικά μικρό αριθμό παραμέτρων.

7. Συζήτηση-Συμπεράσματα

Οι πρωτεΐνες αποτελούν τα πιο διαδεδομένα και πολυδιάστατα, τόσο στη μορφή όσο και στη λειτουργία τους, μακρομόρια και αποτελούν είτε δομικά συστατικά των μεμβρανών του κυττάρου, είτε συνεργούν σε κάποια συγκεκριμένη λειτουργία. Οι μεμβρανικές πρωτεΐνες είναι εξαιρετικά σημαντικές για την εύρυθμη λειτουργία των οργανισμών, παίζουν καθοριστικό ρόλο στην εμφάνιση πολλών ασθενειών, ενώ, συχνά, αποτελούν κύριους στόχους πολλών ουσιών-φαρμάκων. Βασικός στόχος στην παρούσα έρευνα ήταν να εισαγάγει τις τροποποιήσεις των κλασσικών μοντέλων και αλγορίθμων που χρησιμοποιούνται στα HMM, λαμβάνοντας υπόψη τη βιβλιογραφία, και να προτείνει μια καινοτόμο μέθοδο που θα επιτρέπει την ενσωμάτωση βιολογικής πληροφορίας. Συμπεραίνουμε, ότι η μέθοδος που παρουσιάζεται είναι αποτελεσματική και δείχνει να βελτιώνει σε κάποιες περιπτώσεις την πρόγνωση των αρχικών μοντέλων.

Εξίσου σημαντικό ότι τα HMM ενσωματώνουν και αξιοποιούν βιολογική πληροφορία η οποία ανάλογα το βιολογικό πρόβλημα που προσπαθεί να λύσει ένα HMM, καθορίζει και την απόδοση του. Για παράδειγμα στα διαμεμβρανικά μπορεί να έχει καλύτερα αποτελέσματα κάποιο μοντέλο λόγω της καλύτερης συσχέτισης που μπορεί να έχει με το προηγούμενο αμινοξύ όπως για παράδειγμα και από τον σχεδιασμό του μοντέλου πρόγνωσης των διαμεμβρανικών β-βαρελίων όπου συχνά συναντάμε εναλλαγές υδρόφοβων-υδρόφιλων καταλοίπων. Οι αλγόριθμοι αυτοί μπορούν επίσης να είναι χρήσιμοι και σε άλλες εφαρμογές των HMM, εκτός από την πρόβλεψη διαμεμβρανικών πρωτεϊνών, δεδομένου ότι θα μπορούσαν να εφαρμοστούν σε οποιοδήποτε HMM, ανεξάρτητα από τη διαδικασία εκπαίδευσης που χρησιμοποιείται.

Αν και μερικές δοκιμές δείχνουν ότι υπάρχει βελτίωση στα περισσότερα μοντέλα, οι επεκτάσεις των μοντέλων θα μπορούσαν να αποδίδουν καλύτερα όταν χρησιμοποιείται μεγαλύτερο σύνολο εκπαίδευσης. Αυτό ίσως μπορεί να εξηγηθεί εν μέρει από το μεγαλύτερο αριθμό παραμέτρων του συγκεκριμένου μοντέλου. Συνοψίζοντας, τα αποτελέσματά μας δείχνουν ότι είναι δυνατόν να βελτιωθεί η ακρίβεια πρόβλεψης και δεδομένου ότι όλο και περισσότερες ακολουθίες προστίθενται σε βάσεις δεδομένων, αλλά και ότι η εύχρηστη αρχιτεκτονική του HMM που επιτρέπει εύκολα την επέκταση ενός μοντέλου, αποτελεί αξιοσημείωτης σημασίας μια τέτοια προσπάθεια να ληφθεί υπόψη μελλοντικά. Μια προσπάθεια αποτελεί να εστιάσουμε σε δοκιμές χρησιμοποιώντας Hidden Neural Networks (HNN).

Η μέθοδος που παρουσιάστηκε σε αυτή την εργασία επιτρέπει σε κάθε τυπικό HMM να επεκταθεί σε ένα ισοδύναμο μοντέλο ανώτερης τάξης χρησιμοποιώντας πρόσθετη πληροφορία. Αυτό σημαίνει ότι όλοι οι υπάρχοντες αλγόριθμοι που χρησιμοποιούνται για την επεξεργασία των τυπικών HMMs μπορούν να εφαρμοστούν άμεσα για HMMs ανώτερης τάξης. Σε αυτή την εργασία, για τις δοκιμές της επίδοσης των μεθόδων με νέες κωδικοποιήσεις αλλά και οι επεκτάσεις των παραμέτρων και συνόλων εκπαίδευσης, αναπτυχθήκαν με πλήρως αυτοματοποιημένες διαδικασίες. Στόχος μας είναι η χρήση τέτοιων μοντέλων να είναι απλή και να έχουμε τη δυνατότητα να κάνουμε περισσότερες δοκιμές. Αυτό από μόνο του θα πρέπει να ενθαρρύνει περισσότερο τους ερευνητές να πειραματιστούν με τις βελτιωμένες δυνατότητες.

Στόχος μας αποτελεί η σχετική υλοποίηση να αναπτυχθεί και να ενσωματωθεί στις υπάρχουσες μεθόδους που έχουν αναπτυχθεί από την ομάδα μας, η οποία αποτελεί ένα ολοκληρωμένο πακέτο που έχει διαθέσιμες όλες τις γνωστές μεθόδους που καλύπτουν τα HMM. Τα εργαλεία που υλοποιούν αυτά τα μοντέλα, χρησιμοποιούν μια σταθερή αρχιτεκτονική και έτσι είναι

ικανά να παρέχουν στο χρήστη ένα ελάχιστο επίπεδο παρέμβασης στη διαδικασία σχεδιασμού του μοντέλου. Η μέθοδος, που παρουσιάσαμε σε αυτή την εργασία επιτρέπει σε κάθε HMM να επεκταθεί σε ένα ισοδύναμο μοντέλο που χρησιμοποιεί εμπλουτισμένη πληροφορία μέσω της επέκτασης της κωδικοποίησης των παρατηρήσεων. Αυτό σημαίνει ότι όλοι οι υπάρχοντες αλγόριθμοι που χρησιμοποιούνται για την επεξεργασία ενός τυπικού HMM μπορεί να εφαρμοστεί άμεσα και για HMMs ανώτερης τάξης. Αυτό από μόνο του θα πρέπει να ενθαρρύνει περισσότερο τους ερευνητές να πειραματιστούν με τις βελτιωμένες δυνατότητες. Στόχος μας είναι το λογισμικό αυτό να τύχει ευρείας αποδοχής από την ερευνητική κοινότητα και να χρησιμοποιηθεί στην πράξη από επιστημονικές ομάδες σαν ερευνητικό εργαλείο.

Βιβλιογραφία

1. Markov, A., *An example of statistical study on text of Eugeny Onegin illustrating the linking of events to a chain*. Izvestiya Akademii Nauk, 1913. **6**: p. 153-162.
2. Baum, L.E., et al., *A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains*. The annals of mathematical statistics, 1970. **41**(1): p. 164-171.
3. Rabiner, L.R., *A tutorial on hidden Markov models and selected applications in speech recognition*. Proceedings of the IEEE, 1989. **77**(2): p. 257-286.
4. Durbin, R., et al., *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. 1998: Cambridge university press.
5. Krogh, A., *An introduction to hidden Markov models for biological sequences*. New Comprehensive Biochemistry, 1998. **32**: p. 45-63.
6. Krogh, A., et al., *Hidden Markov models in computational biology: Applications to protein modeling*. Journal of molecular biology, 1994. **235**(5): p. 1501-1531.
7. Bilmes, J., *What HMMs can do*. 2002.
8. Baum, L.E., *An equality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes*. Inequalities, 1972. **3**: p. 1-8.
9. Baum, L.E. and J.A. Eagon, *An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology*. Bull. Amer. Math. Soc, 1967. **73**(3): p. 360-363.
10. Bagos, P.G., T.D. Liakopoulos, and S.J. Hamodrakas. *Faster gradient descent training of hidden Markov models, using individual learning rate adaptation*. in *International Colloquium on Grammatical Inference*. 2004. Springer.
11. Baldi, P. and Y. Chauvin, *Smooth on-line learning algorithms for hidden Markov models*. Neural Computation, 1994. **6**(2): p. 307-318.
12. Krogh, A. and S.K. Riis, *Hidden neural networks*. Neural Computation, 1999. **11**(2): p. 541-563.
13. Krogh, A. *Hidden Markov models for labeled sequences*. in *Pattern Recognition, 1994. Vol. 2-Conference B: Computer Vision & Image Processing., Proceedings of the 12th IAPR International. Conference on*. 1994. IEEE.
14. Yuan, Z., *Prediction of protein subcellular locations using Markov chain models*. FEBS letters, 1999. **451**(1): p. 23-26.
15. Ching, W.K., E.S. Fung, and M.K. Ng. *Higher-order hidden Markov models with applications to DNA sequences*. in *International Conference on Intelligent Data Engineering and Automated Learning*. 2003. Springer.
16. Lee, L.-M. and J.-C. Lee. *A study on high-order hidden Markov models and applications to speech recognition*. in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. 2006. Springer.
17. Forchhammer, S. and J. Rissanen, *Partially hidden Markov models*. IEEE Transactions on Information Theory, 1996. **42**(4): p. 1253-1256.
18. Li, Y., *Hidden Markov models with states depending on observations*. Pattern Recognition Letters, 2005. **26**(7): p. 977-984.

19. du Preez, J.A., *Efficient training of high-order hidden Markov models using first-order representations*. Computer speech & language, 1998. **12**(1): p. 23-39.
20. Bagos, P.G., et al., *A Hidden Markov Model method, capable of predicting and discriminating β -barrel outer membrane proteins*. BMC bioinformatics, 2004. **5**(1): p. 29.
21. Tsirigos, K.D., A. Elofsson, and P.G. Bagos, *PRED-TMBB2: improved topology prediction and detection of beta-barrel outer membrane proteins*. Bioinformatics, 2016. **32**(17): p. i665-i671.
22. Bagos, P.G., T.D. Liakopoulos, and S.J. Hamodrakas, *Algorithms for incorporating prior topological information in HMMs: application to transmembrane proteins*. BMC bioinformatics, 2006. **7**(1): p. 189.
23. Bagos, P.G., et al., *Combined prediction of Tat and Sec signal peptides with hidden Markov models*. Bioinformatics, 2010. **26**(22): p. 2811-2817.
24. Bagos, P.G., et al., *Prediction of lipoprotein signal peptides in Gram-positive bacteria with a Hidden Markov Model*. Journal of proteome research, 2008. **7**(12): p. 5082-5093.
25. Bagos, P., et al., *Prediction of signal peptides in archaea*. Protein Engineering Design and Selection, 2009. **22**(1): p. 27-35.
26. Dill, K.A., *Theory for the folding and stability of globular proteins*. Biochemistry, 1985. **24**(6): p. 1501-1509.
27. Kohavi, R. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. in *Ijcai*. 1995. Stanford, CA.
28. Vihinen, M., *How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis*. BMC genomics, 2012. **13**(4): p. S2.
29. Zemla, A., et al., *A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment*. Proteins: Structure, Function, and Bioinformatics, 1999. **34**(2): p. 220-223.
30. Baldi, P., et al., *Assessing the accuracy of prediction algorithms for classification: an overview*. Bioinformatics, 2000. **16**(5): p. 412-424.