ΠΑΝΕΠΙΣΤΗΜΙΟ ΣΤΕΡΕΑΣ ΕΛΛΑΔΑΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Πρόγνωση σηματοδοτικών αλληλουχιών βακτηριακών πρωτεϊνών που εκκρίνονται με το σύστημα των δίδυμων αργινινών (twin arginine – TAT)

Νικολάου Ελισάνθη

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ
Υπεύθυνος
Μπάγκος Παντελής
Επίκουρος Καθηγητής

Λαμία, 2010

Πρόλογος

Η παρούσα εργασία πραγματοποιήθηκε εξ ολοκλήρου στο τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική του Πανεπιστημίου Στερεάς Ελλάδας κατά την διάρκεια του ακαδημαϊκού έτους 2009-2010, υπό την επίβλεψη του Επίκουρου καθηγητή κ. Παντελή Μπάγκου.

Στο πρόλογο αυτής της πτυχιακής θα ήθελα να εκφράσω την εκτίμηση και την ευγνωμοσύνη μου στα άτομα που με βοήθησαν ώστε να ολοκληρώσω αυτή την εργασία.

Θα ήθελα αρχικά να ευχαριστήσω τον κύριο Μπάγκο, πρώτον γιατί μου εμπιστεύτηκε το θέμα του και θέλησε να συνεργαστούμε. Ακόμη, κατά την διάρκεια της συνεργασίας μας, πάντα ήταν στο πλευρό μου, να μου εξηγεί και επιλύει τις απορίες μου. Θα ήθελα να του εκφράσω την ευγνωμοσύνη μου για την απέραντη υπομονή που έδειξε μαζί μου, άσχετα με τον φόρτο εργασίας που είχε. Το κυριότερο σε αυτή την εργασία που έκανε για μένα ήταν να με βοηθήσει να δημοσιεύσω την δουλεία μου, κάτι που θα το κουβαλάω για το υπόλοιπο της ζωής μου. Είναι μια μεγάλη αρχή για τις μελλοντικές μου σπουδές.

Ένα ευχαριστώ οφείλω και στα υπόλοιπα δύο μέλη της Τριμελούς Συμβουλευτικής Επιτροπής, τον Επίκουρο Καθηγητή κ. Πλαγιανάκο Βασίλειο και το Λέκτορα κ. Μάρκου Ευριπίδη, οι οποίοι αφιέρωσαν το πολύτιμο τους χρόνο για να μελετήσουν και να παρακολουθήσουν την διπλωματική μου εργασία.

Επίσης θα ήθελα να ευχαριστήσω το άλλο μέλος της ερευνητικής μας ομάδας, κ. Κώστα Τσιρίγο ο οποίος από την αρχή της συνεργασίας μας, με βοήθησε σε κάθε μου πρόβλημα.

Κλείνοντας θα ήθελα να ευχαριστήσω τους γονείς μου και τα τρία μου αδέλφια, για την ψυχολογική τους στήριξη, την οικονομική τους ενίσχυση και για όλα όσα κάνουν για μένα. Θα αφιερώσω σ’αυτά τα άτομα αυτήν την πτυχιακή εργασία, γιατί είναι ότι πιο σημαντικό έχω.

Στον μπαμπά μου, Πολύβιο
την μαμά μου Φλώρα
και στα τρία μου αδέλφια Ανδρέα, Κώστα, Φάνο.

ΠΕΡΙΕΧΟΜΕΝΑ **ΣΕΛ**

Περίληψη
Η πρόγνωση των πεπτιδίων οδηγητών (signal peptides - Sec) έχει μεγάλη σημασία σε θέματα υπολογιστικής βιολογίας. Εκτός από το σύστημα μεταφοράς Sec, τα βακτήρια, τα αρχαία και οι χλωροπλάστες, διαθέτουν και ένα άλλο σημαντικό σύστημα μεταφοράς, την οδό twin arginine translocase (Tat), η οποία αναγνωρίζει σηματοδοτικές ακολουθίες με μειωμένη υδροφοβικότητα και ένα μοτίβο με δυο διαδοχικές Αργινίνες (RR) που έχουν στην n-περιοχή. Μια σημαντική λειτουργική διαφοροποίηση μεταξύ των συστημάτων μεταφοράς Sec και Tat πεπτιδίων οδηγητών έγκειται στο γεγονός ότι από την πρώτη οδό εκκρίνονται ξετυλιγμένες πρωτεΐνες, ενώ από τη δεύτερη οδό, μεταφέρονται πλήρως διπλωμένες οι πρωτεΐνες χρησιμοποιώντας έναν ακόμα άγνωστο μηχανισμό. Για την πρόγνωση των πρωτεϊνών με το Tat και Sec πεπτίδιο οδηγητή, έχουν αναπτυχθεί διάφορες μέθοδοι, οι οποίες εκπαιδεύτηκαν με δεδομένα που αντλήθηκαν από τη βάση δεδομένων της UniProt. Ο σκοπός της παρούσας εργασίας είναι να αναπτυχθεί ένα καινούργιο Hidden Markov Model (HMM) για την πρόβλεψη και την διάκριση των Sec και Tat πεπτιδίων οδηγητών με μεγαλύτερη ακρίβεια. Αναφέρουμε την ανάπτυξη δύο profiles Hidden Markov Models (pHMMs) τα οποία έχουν την ικανότητα να διακρίνουν Sec και Tat πεπτίδια οδηγητές και να προβλέπουν το σημείο αποκοπής. Τα pHMMs κατασκευάστηκαν με την βοήθεια του πακέτου HMMER 2.3.2. Το μοντέλο μας εκπαιδεύτηκε με μια συγκεκριμένη στοίχιση, που δημιουργήθηκε μετά από εξονυχιστική έρευνα. Τα αποτελέσματα που πήραμε με την χρήση των pHMMs με τα δεδομένα μας, μας οδήγησαν στο συμπέρασμα ότι η πρόγνωση που κάναμε για τις πρωτεΐνες με το Tat πεπτίδιο οδηγητή ήταν βέλτιστη σε σχέση με μεθόδους άλλων εργαλείων που είχαν παρόμοιο σκοπό (TatP και TATFIND). Επιπλέον η στοίχιση για τα Sec πεπτίδια οδηγητές μας έδωσε πολύ ικανοποιητικά αποτελέσματα, παρ' όλο που υπολείπεται του SignalP και του Phobius. Ταυτόχρονα με αυτή την εργασία, με τις συγκεκριμένες στοιχίσεις, δημιουργήθηκε ένα «χειροποίητο» HMM από μέλη της ερευνητικής μας ομάδας. Το εργαλείο PRED-TAT που στηρίζεται σ'αυτό το HMM, θα το βρείτε στην ιστοσελίδα http://www.compgen.org/tools/PRED-TAT/. Η όλη δουλειά που έχει γίνει για το μοντέλο αυτό δημοσιεύθηκε στο περιοδικό Bioinformatics (Bagos, Nikolaou et al. 2010).

Abstract

Computational prediction of signal peptides is of great importance in computational biology. In addition to the general secretory pathway (Sec), Bacteria, Archaea and chloroplasts, possess another major pathway that utilizes the Twin-Arginine translocase (Tat), which recognizes longer and less hydrophobic signal peptides carrying a distinctive pattern of two consecutive Arginines (RR) in the n-region. A major functional differentiation between the Sec and Tat export pathways lies in the fact that the former translocates secreted proteins unfolded through a protein-conducting channel, whereas the latter, translocates completely folded proteins using an unknown mechanism. The purpose of this work was to develop a novel method for predicting and discriminating Sec from Tat signal peptides at better accuracy. We constructed two profiles Hidden Markov Models (pHMMs), which have the ability to distinguish Sec from Tat signal peptides using the HMMER 2.3.2 package. The method we propose, PRED-TAT_{HMMER}, is capable of discriminating Sec from Tat signal peptides and predicting their cleavage sites at the same time. The method and the associated profile HMMs are freely available for academic users at http://www.compgen.org/tools/PRED-TAT/. On an independent test set of experimentally verified Tat signal peptides, PRED-TAT_{HMMER} clearly outperforms the previously proposed methods TatP and TATFIND, whereas, when evaluated as a Sec signal peptide predictor compares favorably to top-scoring predictors such as SignalP and Phobius. The results obtained in this work are presented in a recent Bioinformatics paper (Bagos, Nikolaou et al. 2010)

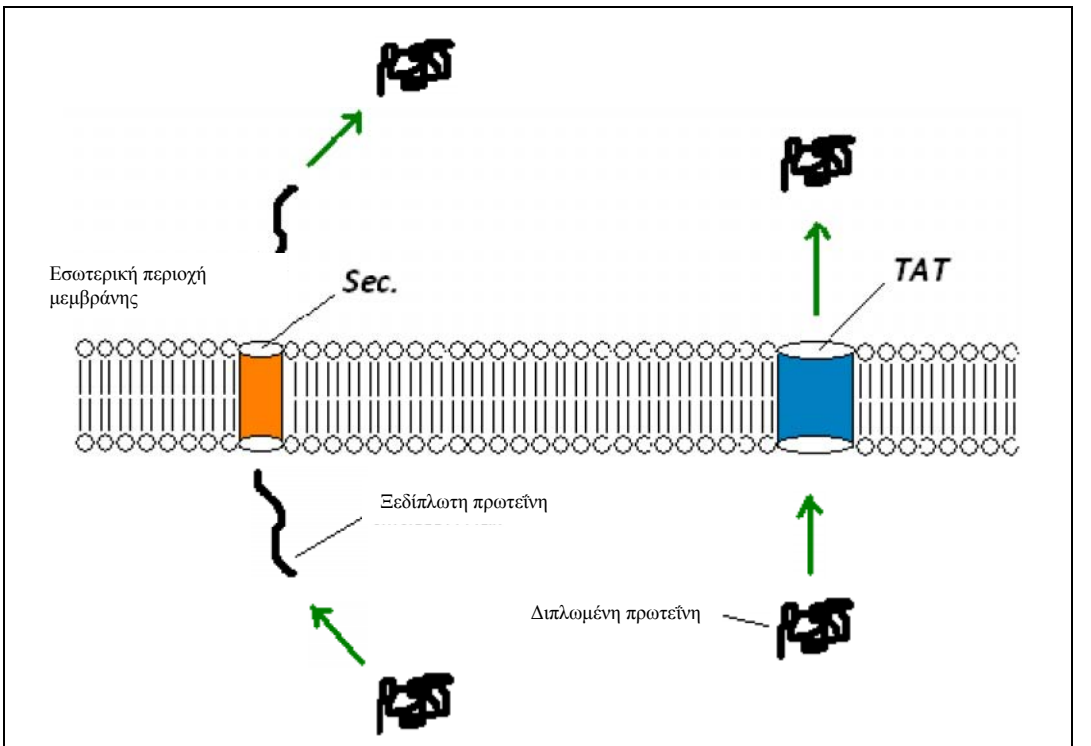
1. Εισαγωγή

Το 1999 ο Γκούντερ Μπλόμπελ, Γερμανός επιστήμονας, καθηγητής στο αμερικάνικο Πανεπιστήμιο Ροκφέλερ βραβεύτηκε με νόμπελ για την εργασία του, στην υπόθεση ότι οι πρωτεΐνες συντίθενται μέσα από μια αλληλουχία αμινοξέων, η οποία ονομάζεται σηματοδοτικό πεπτίδιο (signal peptide) ή πεπτίδιο οδηγητή. Διαπίστωσε ότι οι πρωτεΐνες περιέχουν ενδογενή σήματα, τα οποία καθορίζουν τη μετακίνηση και την τοποθέτησή τους μέσα στα κύτταρα. Σε όλους τους μικροοργανισμούς (βακτήρια ,αρχαία και ευκάρυα), μια θεμελιώδης διαδικασία είναι η μεταφορά των πρωτεϊνών στις βιολογικές μεμβράνες. Σχεδόν σε όλες τις περιπτώσεις, οι πρωτεΐνες που βγαίνουν πέραν από τη κυτταρική μεμβράνη (van Roosmalen, Geukens et al. 2004; Tuteja 2005) έχουν στη δομή τους ένα αμινοτελικό πεπτίδιο οδηγητή (N-terminal). Στους ευκαριωτικούς οργανισμούς, η μεταφορά των περισσοτέρων εξωκυττάριων πρωτεϊνών που χαρακτηρίζονται από το αμινοτελικό πεπτίδιο οδηγητή, πραγματοποιείται από το γενικό σύστημα έκκρισης (Sec). Παρόλα αυτά, αυτό που συναντά κανείς στις εκκριτικές πρωτεΐνες στους ευκαριωτικούς οργανισμούς και στα βακτήρια είναι εντελώς διαφορετικό (von Heijne, Steppuhn et al. 1989; Habib, Neupert et al. 2007). Εκτός από την οδό Sec, πολλά βακτήρια έχουν στο σύστημά τους και μια άλλη πρωτεϊνική εκκριτική οδό, την Tat οδό, που παρέχει την δυνατότητα διακίνησης μέσω της δίδυμης αργινίνης. Αυτή η οδός αναγνωρίζεται με το χαρακτηριστικό μοτίβο RR στην n-περιοχή (Teter and Klionsky 1999; Berks, Palmer et al. 2005; Lee, Tullman-Ercek et al. 2006). Επίσης το διαφορετικό στην λειτουργία των οδών Sec και Tat είναι ότι η πρώτη οδός μεταφέρει τις εκκριμένες πρωτεΐνες που ξετυλίγονται μέσω ενός πρωτεϊνικού καναλιού ενώ η δεύτερη μεταφέρει εντελώς διπλωμένες τις πρωτεΐνες χρησιμοποιώντας ένα άγνωστο ακόμα μηχανισμό (Berks, Palmer et al. 2005). Σημαντικό ενδιαφέρον υπάρχει στα haloarchaea, όπου η παρουσία της οδού Tat δεν παραμένει απλά στην λειτουργία μεταφοράς διπλωμένων πρωτεϊνών, αλλά παίζει σημαντικό ρόλο και στη βιωσιμότητα (Dilks, Gimenez et al. 2005; Thomas and Bolhuis 2006). Υπάρχουν επίσης στοιχεία ότι το σύστημα μεταφοράς Tat, χρησιμοποιείται και ως τμήμα ενός μηχανισμού για την προσαρμογή σε περιβάλλον με περίσσιο άλας (Rose, Bruser et al. 2002).

Πολλές είναι οι μέθοδοι οι οποίοι αναπτύχθηκαν για την πρόβλεψη του Tat και Sec πεπτιδίου οδηγητή. Όσον αφορά την υπολογιστική πρόβλεψη των πεπτιδίων οδηγητών πραγματοποιήθηκε αρχικά με την μελέτη ενός πίνακα βαρών (von Heijne 1986). Σήμερα για τον σχεδιασμό των μοντέλων των εργαλείων πρόγνωσης γίνεται χρήση, των νευρωνικών δικτύων (NN) (Nielsen, Engelbrecht et al. 1997; Nielsen, Brunak et al. 1999) καθώς επίσης και των Hidden Markov Models (Nielsen and Krogh 1998). Ένα από τα πιο αξιόλογα εργαλεία πρόβλεψης είναι το SignalP, το οποίο πρόσφατα επανεκπαιδεύθηκε λαμβάνοντας υπόψη περισσότερες πληροφορίες που προστέθηκαν στις βιολογικές βάσεις δεδομένων όσον αφόρα τις πρωτεΐνες, και έτσι παρατηρείται να έχει καλύτερη ακρίβεια στην πρόβλεψη (Bendtsen, Nielsen et al. 2004). Η μέθοδος του Phobius (Kall, Krogh et al. 2004; Kall, Krogh et al. 2007), και η μέθοδος του Philius (Reynolds, Kall et al. 2008) ακολούθησαν ένα διαφορετικό τρόπο σχεδιασμού του μοντέλου (Hidden Markov Model and Bayesian network, αντίστοιχα). Άλλα εργαλεία όπως το LipoP (Juncker, Willenbrock et al. 2003) και το PRED-LIPO (Bagos, Tsirigos et al. 2008) αναπτύχθηκαν κατά τη διάρκεια των τελευταίων ετών για την πρόβλεψη των πεπτιδίων οδηγητών στις λιποπρωτεΐνες. Τα πεπτίδια αυτά κατέχουν στην διακριτή περιοχή αποκοπής τους ένα αμινοξύ κυστεΐνης για να ενσωματωθούν στη μεμβράνη (Sankaran and Wu 1994; Sankaran, Gupta et al. 1995).

Οι μέθοδοι που αναπτύχθηκαν τα τελευταία χρόνια θεωρούνται αρκετά ικανές για την πρόβλεψη των πεπτιδίων οδηγητών Sec, όμως πολύ πιο λίγες ήταν οι μέθοδοι για την πρόβλεψη των Tat πεπτιδίων οδηγητών. Το TATFIND παρουσιάστηκε αρχικά συνδυάζοντας τα πρότυπα κανονικών εκφράσεων και ανάλυση της υδροφοβικότητας (Rose, Bruser et al. 2002), ενώ μερικά χρόνια αργότερα, παρουσιάστηκε το TatP χρησιμοποιώντας έναν συνδυασμό από πρότυπα κανονικών εκφράσεων και νευρωνικών δικτύων (Bendtsen, Nielsen et al. 2005). Το εργαλείο TatP έχει αποδειχθεί ότι είναι πιο αξιόπιστο, ενώ το TATFIND δεν είναι σε θέση να

αναγνωρίσει την περιοχή αποκοπής, αλλά μόνο την ύπαρξη της n- και της h-περιοχής. Κύριο πρόβλημα σε αυτή την έρευνα είναι ότι καμία από αυτές της μεθόδους δεν εκπαιδεύθηκε για να κάνει διακρίσεις ταυτόχρονα για Tat και Sec πεπτίδια οδηγητές.



Εικόνα 1.1: Είναι μια χαρακτηριστική απεικόνιση Sec και Tat οδών, για την πρωτεϊνική έκκριση σε όλη την κυτταροπλασματική μεμβράνη των βακτηρίων. Όπου χαρακτηριστικά από την οδό Sec διαπερνά ξεδιπλωμένη η πρωτεΐνη, ενώ από την Tat οδό περνάει διπλωμένη διαμέσου της κυτταρικής μεμβράνης.

1.1. Βακτήρια και μηχανισμοί

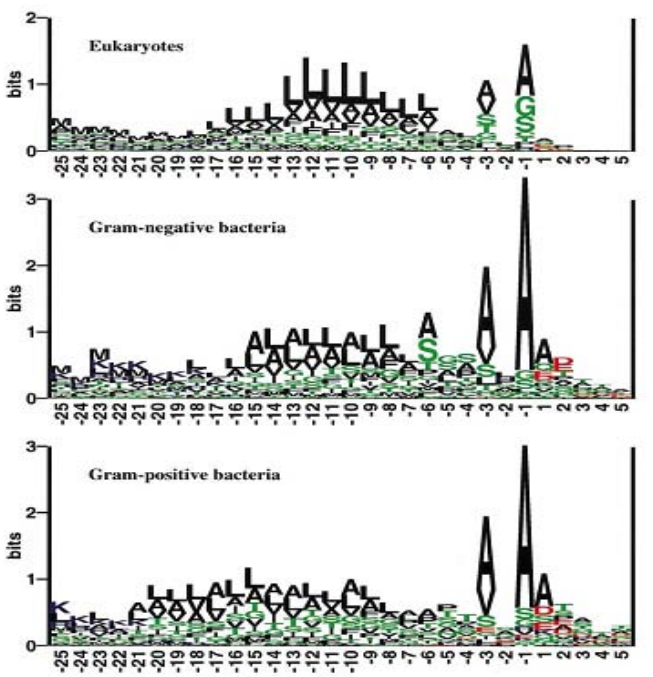
Τα βακτήρια έχουν την απλούστερη δομή, ουσιαστικά δεν περιέχουν πυρήνα. Για τον λόγο αυτό οι οργανισμοί αυτοί ονομάζονται προκαρύωτες. Συνήθως είναι σφαιρικά, ραβδόσχημα ή σπειροειδή κύτταρα μήκους λίγων μικρομέτρων. Στις αρχικές μελέτες οι προκαρύωτες εθεωρείτο ότι ανήκουν στην ίδια μεγάλη ομάδα. Στην συνέχεια όμως ανατράπηκε αυτή η εντύπωση, με αποτέλεσμα να χωριστούν σε δύο μεγάλες υποομάδες: τα ευβακτήρια (eubacteria) και τα αρχαιοβακτήρια ή αρχαία (archaebacteria). Τα περισσότερα βακτήρια, τα είδη που ζουν στο χώμα ή ευθύνονται για διάφορα νοσήματα, είναι τα ευβακτήρια. Τα αρχαιοβακτήρια συνήθως βρίσκονται σε περιβάλλον εχθρικό για τα περισσότερα είδη κυττάρων. Πρόκειται για κύτταρα που ζουν σε πολύ αλμυρό νερό, σε θερμές όξινες ηφαιστειακές πηγές, στον πυθμένα της θάλασσας και σε άλλα τέτοια μέρη. Γενικά τα βακτήρια διακρίνονται σε δύο βασικές κατηγορίες: Gram Positive και Gram Negative. Η διάκριση αυτή πραγματοποιήθηκε για πρώτη φορά από ένα επιστήμονα με το όνομα Gram, όπου επινόησε μια τεχνική ονομαζόμενη χρωστικότητα Gram, με την οποία τα βακτήρια μπορούν να αλλάζουν χρώμα και να χωρίζονται σε αυτές τις δύο ομάδες. Με την μέθοδο του Gram τα βακτήρια που παίρνουν το μπλε-μοβ χρώμα είναι τα Gram Positive και τα βακτήρια που παίρνουν κόκκινο χρώμα είναι τα Gram Negative. Η διαφορά στην αντίδραση από τις δύο αυτές ομάδες βακτηρίων πιστεύεται ότι οφείλεται στην διαφορά της δομής του κυτταρικού τοιχώματος.

Πίνακας 1.1: Σύγκριση σύμφωνα με τα χαρακτηριστικά τους Gram Positive & Gram Negative βακτηρίων

Χαρακτηριστικά	**Gram Positive**	**Gram Negative**
Αντίδραση στην μέθοδο ανίχνευσης Gram	Διατηρούν το κρυσταλλικό χρωματισμό ιωδίου και χρώση σκούρο βιολετί ή μοβ	κόκκινο
Στρώματα πεπτιδογλυκάνης (peptidoglycan)	Πάχους (έχει δηλαδή πολλά στρώματα)	Πολύ λεπτή στρώση
Teichoic οξέα	Υπάρχει σε πολλά	Δεν υπάρχει σε κανένα
Περιπλασματικός χώρος έξω από την μεμβράνη	Δεν έχει	Έχει
Εξωτερική μεμβράνη	Δεν έχει	Έχει
Περιέχει λιποπολυσακχαρίτες (LPS)	Σχεδόν σε κανένα	Στα περισσότερα
Οικογένειες πρωτεϊνών	*Actinobacteria, Actinobacteridae, Actinomycetales*	*Proteobacteria, Chlorobi, Verrucomicrobia, Acidobacteria, Aquificae, Bacteroidetes, Chlamydiae, Cyanobacteria, Spirochaetes, Thermodesulfobacteria, Nitrospirae, Thermotogae, Deferribacteres, Fusobacteria, Planctomycetes, Thermus thermophilus.*

Έξω από την κυτταροπλασματική μεμβράνη στο περιπλασματικό χώρο υπάρχει στρώμα πεπτιδογλυκάνης, και πέρα από αυτό το στρώμα βρίσκεται μια εξωτερική μεμβράνη, η οποία περιέχει τα φωσφολιπίδια και τους λιποπολυσακχαρίτες (Driessen and Nouwen 2008). Η σημαντικότερη οδός για την πρωτεϊνική μεταφορά πέρα, από, και στην κυτταροπλασματική μεμβράνη είναι η οδός Sec.

Οι διαφορές μεταξύ των πεπτιδίων οδηγητών στους διαφορετικούς οργανισμούς είναι προφανείς. Και αυτό φαίνεται στην εικόνα 1.2.

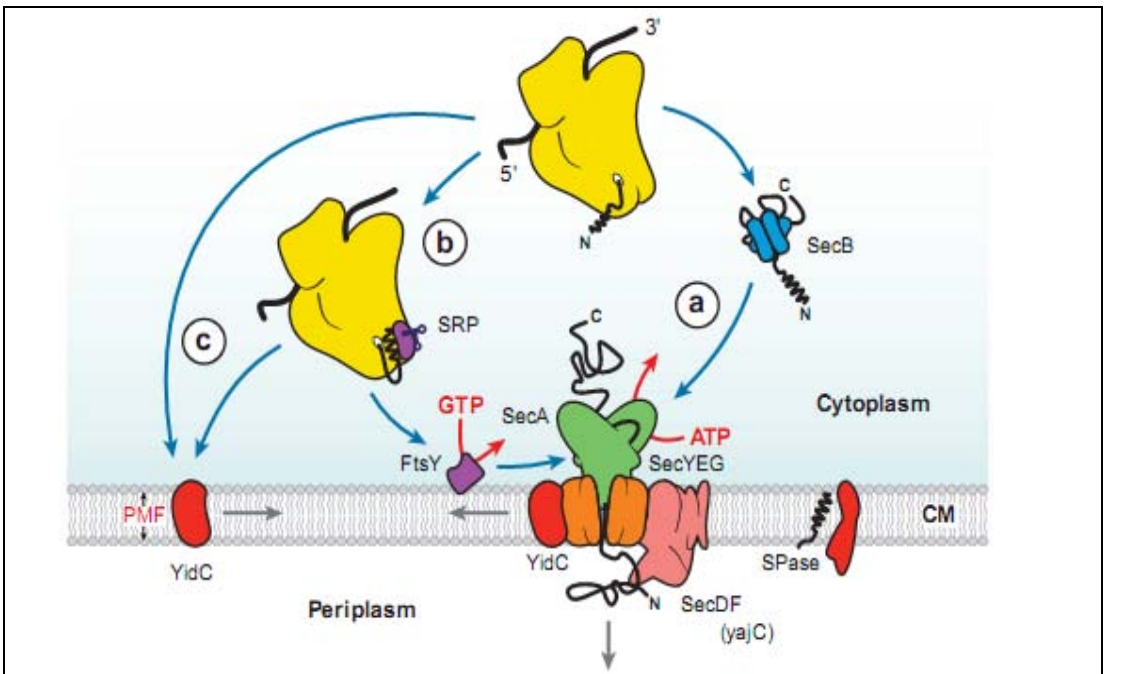


Εικόνα 1.2: Παρουσιάζει τα logos των ακολουθιών (Schneider and Stephens 1990) ανάμεσα σε διαφορετικούς οργανισμούς. Αντιπροσωπεύουν την στοίχιση του σημείου αποκοπής στο κάθε οργανισμό, και παρατηρούμε να υπάρχει μια διαφορά. Το συνολικό ύψος στην κάθε θέση του logo αντιπροσωπεύει το γενικό σύνολο πληροφοριών που βρίσκονται στην θέση αυτή της ακολουθίας της πρωτεΐνης. Το ύψος του γράμματος αντιπροσωπεύει το πλήθος παρουσίας του συγκεκριμένου αμινοξέος. Θετικά και αρνητικά φορτισμένα αμινοξέα παρουσιάζονται σε μπλε και κόκκινο αντίστοιχα, ενώ τα πολικά αμινοξέα είναι πράσινα και τα υδρόφοβα αμινοξέα είναι μαύρα. (Nielsen, Engelbrecht et al. 1997)

Τα πεπτίδια οδηγητές από τα Gram-positive βακτήρια έχουν μεγαλύτερο μήκος ακολουθίας από ακολουθίες άλλων οργανισμών, στο μήκος της h-περιοχής οφείλεται η διαφορά, όπως παρατηρείται και στη δημοσίευση του Heijne (von Heijne, Steppuhn et al. 1989). Οι h-περιοχές στους προκαρυωτικούς οργανισμούς έχουν περισσότερα αμινοξέα Leu [L] και Ala [A], στα ίδια ποσοστά εμφάνισης και η περιοχή αυτή στους ευκαριωτικούς οργανισμούς έχει Leu[L] και συνοδεύεται με λιγοστά αμινοξέα Val [V], Ala [A], Phe [F] και Ile [I]. Κοντά στην περιοχή αποκοπής, (-3,-1) περιοχή στις ακολουθίες και των τριών οργανισμών, παρατηρείται να υπερισχύει η παρουσία της αλανίνης [A]. Στις πρώτες θέσεις της ώριμης πρωτεΐνης (προς το τέλος της περιοχής αποκοπής) παρουσιάστηκε σε ορισμένους προκαρυωτικούς οργανισμούς η αλανίνη [A], τα αρνητικά φορτισμένα αμινοξέα [D ή E], και τα υδρόξυαμινοξέα [S ή T]. Στη περιοχή (-16,-8) της στοίχισης στους προκαρυωτικούς οργανισμούς, υπάρχουν θετικά φορτισμένα αμινοξέα Lys [L] και αμινοξέα Arg [R], ενώ στους ευκαριωτικούς οργανισμούς υπάρχουν με μικρότερο βαθμό εμφάνισης αμινοξέα Arg [R] και σχεδόν καθόλου Lys [L].

1.2. Πεπτίδιο οδηγητή Sec

Ένα θεμελιώδες σημαντικό χαρακτηριστικό της κυτταρικής ζωής είναι η δυνατότητα διακίνησης των πρωτεϊνών πέρα από την κυτταροπλασματική μεμβράνη μέσω της καθιερωμένης οδού Sec που βρίσκεται και στα ευκαρυωτικά και προκαρυωτικά κύτταρα. Το πεπτίδιο οδηγητή αποκόπτεται συνήθως μεταξύ του 15^{ου} και 40^{ου} αμινοξέος της ακολουθίας. Η οδός Sec αποτελείται από ένα σύνθετο σύστημα μεταφοράς που συμπεριλαμβάνει SecY, SecE και SecG και ένα σύστημα μεταφοράς SecA. Η μεταφορά των πρωτεϊνών διαμέσου της κυτταρικής μεμβράνης γίνεται μέσω του συστήματος μεταφοράς SecA, που παίρνει την αναδιπλωμένη πρωτεΐνη από το πρωτεϊνικό κανάλι SecYEG και την οδηγεί με την βοήθεια της ενέργειας που ελευθερώνεται από την υδρόλυση της ATP. Οι πρωτεΐνες που προορίζονται για την εξαγωγή μέσο της οδού Sec είναι συνδεδεμένες με αμινοτελικό πεπτίδιο οδηγητή της ακολουθίας το οποίο αφαιρείται σε ένα προχωρημένο στάδιο κατά τη διαδικασία εξαγωγής από την πεπτιδάση οδηγητή (SPase), στη περιοχή του περιπλασματικού χώρου έξω από την μεμβράνη.



Εικόνα 1.3: Μοντέλο της πρωτεΐνης που στοχεύει στην οδό Sec. Το βακτηριακό σύστημα μεταφοράς Sec είναι μια οδός μέσα στην κυτταροπλασματική μεμβράνη (CM), η οποία περιλαμβάνει ένα σύστημα μεταφοράς SecA (πράσινο), το πρωτεϊνικό κανάλι, SecYEG (πορτοκάλι), και τις βοηθητικές πρωτεΐνες SecDF (yajC) (ροζ) και YidC (κόκκινο). Το πεπτίδιο οδηγητή (SPase) είναι πεπτίδιο το οποίο διασπά την ακολουθία οδηγητών από τις προπρωτεΐνες στην περιοχή του περιπλασματικού χώρου έξω από την μεμβράνη. (α) οι εκκριτικές πρωτεΐνες (κίτρινες) στοχεύουν στη μεταμεταφραστική οδό Sec της ακολουθίας οδηγητών, η οποία αναγνωρίζεται από το σύστημα μεταφοράς SecA, το σωματίδιο αναγνώρισης οδηγητή (SRP), και από το μοριακό συνοδό SecB (μπλε). (β) Οι πρωτεΐνες μεμβρανών και μερικές προπρωτεΐνες στοχεύουν στο σύστημα μεταφοράς Sec. (γ) Μερικές διαμεμβρανικές πρωτεΐνες μεταφέρονται πέραν της κυτταροπλασματικής μεμβράνης μέσω YidC. (Driessen and Nouwen 2008)

Τα πεπτίδια οδηγητές στα βακτήρια διαιρούνται κυρίως σε εκκριτικά πεπτίδια οδηγητές που διασπόνται από Signal Peptidase I (SPase I), και εκείνα που διασπόνται από Signal Peptidase II (SPase II ή Lsp). Τα πεπτίδια οδηγητές λιποπρωτεϊνών Sec είναι αυτά που υποβάλλονται σε επεξεργασία από την πεπτιδάση II. Τα πεπτίδια αυτά έχουν μελετηθεί εκτενώς για χρόνια, αποκαλύπτοντας μια δομή που αποτελείται από τις ακόλουθες τρεις περιοχές: (i) θετικά φορτισμένη n-περιοχή, (ii) μια υδροφοβική h-περιοχή που εκτείνεται στη μεμβράνη και (iii) μια συνήθως μικρού μήκους πολική c-περιοχή. Η περιοχή αποκοπής για το πεπτίδιο οδηγητή βρίσκεται στην c-περιοχή (von Heijne 1990), γνωστή με το μοτίβο AXA, στο οποίο το

Α αντιπροσωπεύει την αλανίνη και το Χ οποιοδήποτε αμινοξύ, και αναγνωρίζεται από την πεπτιδάση Ι (SPase I) που διασπά και απελευθερώνει την ώριμη πρωτεΐνη. Εντούτοις, το μήκος της ακολουθίας οδηγητή, καθώς επίσης και η θέση της περιοχής αποκοπής, ποικίλλει σημαντικά μεταξύ των διαφόρων πρωτεϊνών. Οι μεγαλύτερες διαφορές παρατηρήθηκαν μεταξύ των ευκαριωτικών και βακτηριακών ακολουθιών οδηγητών. Για διάφορους λόγους είναι επιθυμητό να προσδιοριστούν τα πεπτίδια οδηγητές και οι αντίστοιχες θέσεις αποκοπής τους.

1.3. Πεπτίδιο οδηγητή Tat

Η ανακάλυψη της οδού Tat έγινε στις αρχές της δεκαετίας του '90, όταν παρατηρήθηκε ότι ένα υποσύνολο πολυπεπτιδίων στους χλωροπλάστες θα μπορούσε να μεταφερθεί ανεξάρτητα από την υδρόλυση ATP (Mould and Robinson 1991; Cline, Ettinger et al. 1992). Για αυτόν τον λόγο, ονομάστηκε αρχικά οδός ΔpH ή ακόμη και οδός cpTat (οδός χλωροπλαστών Tat/ΔpH). Το 1995 ο Creighton και οι συνεργάτες του, (Creighton, Hulford et al. 1995) παρουσίασαν τα πρώτα στοιχεία ότι η οδός cpTat επιτρέπει τη δυνατότητα διακίνησης των πρωτεϊνών. Σύντομα, ο Berks (Berks 1996) παρατήρησε ότι μια ομάδα βακτηριακών πρωτεϊνών που βρίσκονται στην περιοχή του περιπλασματικού χώρου, περιέχει διάφορους συμπαράγοντες με μοναδικό τύπο πεπτιδίου οδηγητή που χαρακτηρίζεται με ένα μοτίβο διαδοχικής «δίδυμης αργινίνης», S/T-**R-R**-x-F-L-K (Berks 1996; Stanley, Palmer et al. 2000) το οποίο βρίσκεται επίσης στα υποστρώματα των χλωροπλαστών. Το μοτίβο δίδυμης αργινίνης, Arg-Arg, βρίσκεται στη περιοχή μεταξύ των n- και h-περιοχών του Tat πεπτιδίου οδηγητή. Η ύπαρξη μιας βακτηριακής οδού ανάλογης με αυτήν στους χλωροπλάστες καθιερώθηκε και κλήθηκε αρχικά mtt (membrane targeting and translocation) (Weiner, Bilous et al. 1998) και αργότερα Tat (twin-arginine translocation) (Sargent, Bogsch et al. 1998).

Το ελάχιστο σύνολο συστατικών που απαιτείται για τη δυνατότητα διακίνησης στην οδό Tat στο *Escherichia coli* αποτελείται από τρεις διαμεμβρανικές πρωτεΐνες, τις TatA,TatB και TatC (Tha4, Hcf106, και cpTatC αντίστοιχα στους χλωροπλάστες) (Bolhuis 2002; de Leeuw, Granjon et al. 2002). Η πλειοψηφία των πρωτεϊνών στους προκαρυωτικούς οργανισμούς μεταφέρεται μέσω του συστήματος μεταφοράς Sec. Μια εξαίρεση βρίσκεται στα αλόφιλα αρχαία (haloarchaea), τα οποία προβλέπονται να μεταφέρουν τις περισσότερες από τις πρωτεΐνες τους μέσω του συστήματος μεταφοράς Tat (Bolhuis 2002; Rose, Bruser et al. 2002). Εντούτοις, σε άλλα βακτήρια η σύνθεση του συστήματος μεταφοράς Tat ποικίλει. Μια πρωτεΐνη TatB φαίνεται να μην είναι σημαντική για την μεταφορά, δεδομένου ότι μερικά γονίδια κωδικοποιούν μόνο ενιαίες TatA και TatC πρωτεΐνες (π.χ., στο *Rickettsia prowazekii* και στο *Staphyloccocus aureus*). Κάποια Gram-positive βακτήρια και αρχαία περιέχουν πολλά γονίδια TatC καθώς επίσης και πολλά γονίδια TatA/B. Παραδείγματος χάριν, στο *Bacillus subtilis* έχει δύο γονίδια TatC και τρία γονίδια TatA (Jongbloed, Martin et al. 2000), κάθε ένα από τα οποία φαίνεται να αποτελεί τα χωριστά υποστρώματα του συστήματος μεταφοράς Tat (Jongbloed, Martin et al. 2000; Jongbloed, Antelmann et al. 2002; Pop, Martin et al. 2002). Οι εκτιμήσεις από αναλύσεις βασισμένες στη πρωτεομική και τη βιοπληροφορική δείχνουν ότι 5-8% των εκκριμένων πρωτεϊνών στα βακτήρια όπως στο *Escherichia coli* και στο *Bacillus subtilis* μεταφέρεται μέσω της Tat οδού.

Έχει καθοριστεί ότι η δυνατότητα διακίνησης της πρωτεΐνης πέρα από την κυτταροπλασματική μεμβράνη των Gram-negative βακτηρίων μεσολαβεί από τουλάχιστον τέσσερις ευδιάκριτες διαβάσεις: 1) η γενική εκκριτική διάβαση (Sec), 2) το σωματίδιο αναγνώρισης οδηγητή (SRP-signal recognition particle), 3) το σύστημα YidC, 4) και το σύστημα μεταφοράς, δίδυμης αργινίνης (Tat).

Οι ακολουθίες με το Tat πεπτίδιο οδηγητή έχουν μεγαλύτερο μήκος από τις αντίστοιχες ακολουθίες με το Sec πεπτίδιο οδηγητή (λόγω μιας επιπλέον n-περιοχής), με μέσα μήκη 38 και 24 αμινοξέα, αντίστοιχα (Cristobal, de Gier et al. 1999). Επιπλέον, η h-περιοχή των Tat πεπτιδίων οδηγητών είναι λιγότερο υδροφοβική από τα κλασσικά πεπτίδια οδηγητές Sec εξαιτίας της παρουσίας περισσότερων καταλοίπων γλυκίνης και θρεονίνης (Cristobal, de Gier et

al. 1999). Τα πεπτίδια αυτά δεν περιορίζονται στα βακτήρια, έχουν βρεθεί επίσης και σε αρχαία (Rose, Bruser et al. 2002) και στη μεμβράνη θυλακοειδών των χλωροπλαστών. Στα βακτήρια το σύστημα μεταφοράς Tat βρίσκεται στην κυτταροπλασματική μεμβράνη και χρησιμεύει για την εξαγωγή των πρωτεϊνών στο περιβάλλον των κυττάρων ή στον εξωκυττάριο χώρο. Η Tat οδός είναι ένα πρωτεϊνικό σύστημα για την μεταφορά διπλωμένων πρωτεϊνών διαμέσου της μεμβράνης, πολλές από τις οποίες περιέχουν οξειδοαναγωγικά ενεργούς συμπαράγοντες που πρέπει να ενσωματωθούν στο κυτταρόπλασμα πριν από την εξαγωγή (Berks 1996; Berks, Sargent et al. 2000). Στους χλωροπλάστες, το σύστημα μεταφοράς Tat βρίσκεται στη μεμβράνη των θυλακοειδών και κατευθύνει την εισαγωγή των πρωτεϊνών από το στρώμα.

Οι μελέτες για τα Tat πεπτίδια οδηγητές των βακτηρίων και χλωροπλαστών έχουν καταδείξει ότι τα δύο αμετάβλητα κατάλοιπα αργινίνης του μοτίβου είναι αυστηρά σημαντικά για τη μεταφορά των πρωτεϊνών διαμέσου της οδού. Ακόμη και με την αντικατάσταση της μιας από τις αργινίνες με λυσίνη εμποδίζει την κανονική δυνατότητα διακίνησης. Το σύστημα Tat είναι τώρα γνωστό ότι εκτός από την μεταφορά διπλωμένων πρωτεϊνών (Wickner and Schekman 2005) είναι απαραίτητο σε πολλές βακτηριακές διαδικασίες συμπεριλαμβανομένου του μεταβολισμού ενέργειας, της βιοσύνθεσης του κυτταρικού τοιχώματος και της βακτηριακής παθογένειας. Αν και το μοτίβο ακολουθίας δίδυμης αργινίνης έχει καθοριστεί σε προηγούμενες έρευνες, μερικές παραλλαγές μπορούν ακόμα να γίνουν αποδεκτές από τα προγράμματα αναγνώρισης Tat πεπτιδίου οδηγητή.

1.3.1. Συντηρητικότητα αργινίνης

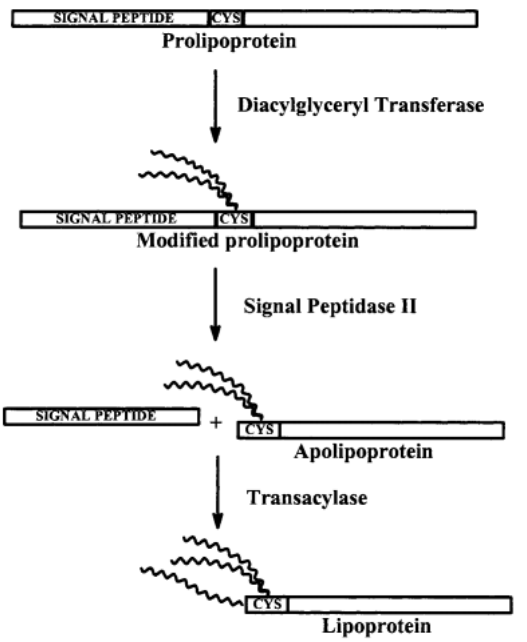
Σε κάποιους οργανισμούς η πρώτη αργινίνη του διπεπτιδικού μοτίβου Arg-Arg μπορεί να αντικατασταθεί με Lys (λυσίνη), και η δεύτερη αργινίνη μπορεί να αντικατασταθεί με Gln ή Asn καθώς επίσης και με Lys, με αποτέλεσμα να φέρουν διαφορετική απόδοση κατά την μεταφορά τις εκκριτικής πρωτεΐνης (Stanley, Palmer et al. 2000; Buchanan, Sargent et al. 2001; DeLisa, Samuelson et al. 2002; Ize, Gerard et al. 2002). Η αντικατάσταση και στις δύο αργινίνες με τη λυσίνη εμποδίζει την εξαγωγή των πρωτεϊνών από την Tat οδό (Stanley, Palmer et al. 2000). Μέχρι σήμερα, μόνο σε δύο πρότυπα Tat παρατηρήθηκε η απουσία της δίδυμης αργινίνης, στο πρότυπο pre-pro-penicillin amidase του *Escherichia coli* που περιέχει Asn ενδιάμεσα από τις αργινίνες (Arg-Asn-Arg) (Ignatova, Hornle et al. 2002) και στο πρότυπο της tetrathionate reductase από το *Salmonella enterica*, όπου αντικαθίσταται η μια αργινίνη από λυσίνη (Lys-Arg) (Hinsley, Stanley et al. 2001). Τα άλλα κατάλοιπα μέσα στο μοτίβο συναίνεσης είναι πιο ελαστικά στην αντικατάσταση του αμινοξέος. Παραδείγματος χάριν, τα Phe και Leu κατάλοιπα μπορούν να αντικατασταθούν με άλλα ιδιαίτερα υδρόφοβα αμινοξέα με αποτέλεσμα να μειωθεί ο χρόνος μεταφοράς της πρωτεΐνης (Stanley, Palmer et al. 2000). Επίσης, ούτε η Ser ούτε η Lys μέσα στο μοτίβο είναι τόσο σημαντικά για το σύστημα μεταφοράς Tat (Stanley, Palmer et al. 2000).

1.4. Αρχαία

Τα *Haloarchaea* αναπτύσσονται σε περιβάλλον με μεγάλες συγκεντρώσεις από άλας που πλησιάζουν τον κορεσμό (δηλαδή το άλας φθάνει περίπου 36%). Εντούτοις, σε λίγα αρχαία είναι γνωστός ο τρόπος με τον οποίο σταθεροποιούν τις εκκριμένες πρωτεΐνες τους σε τέτοιο εχθρικό περιβάλλον. Έχει αποδειχτεί, (Rose, Bruser et al. 2002) ότι η χρησιμοποίηση των πρωτεϊνικών συστημάτων μετακίνησης για την πρωτεϊνική έκκριση από το *Halobacteriaceae* διαφέρει σημαντικά από αυτή των μη-haloarchaea, και αντιπροσωπεύει μια προσαρμογή σε περιβάλλον με υψηλά επίπεδα άλατος. Αν και οι περισσότερες πρωτεΐνες εκκρίνονται μέσω των γενικών μηχανημάτων έκκρισης (Sec), η οδός δυνατότητας διακίνησης μέσω της δίδυμης αργινίνης (Tat) χρησιμοποιείται κυρίως για την έκκριση των οξειδοαναγωγικών πρωτεϊνών. Λόγω των υψηλών επιπέδων άλατος ενδοκυτταρικά στα *Haloarchaea*, οι πρωτεΐνες πρέπει να διπλωθούν πολύ γρήγορα για να αποτρέψουν τη συσσωμάτωση. Επομένως, πολλές πρωτεΐνες στα haloarchaeal διπλώνονται πριν φθάσουν στη μεμβράνη. Αυτές οι πρωτεΐνες μπορούν να εξαχθούν μόνο μέσω του Tat συστήματος, επειδή το σύστημα Sec μπορεί να χειριστεί μόνο τις ξετυλιγμένες πρωτεΐνες. Τα γονίδια TatA και TatC βρίσκονται σε όλα τα haloarchaeal που αναλύονται μέχρι σήμερα, δηλ. στο *Halobacterium salinarum* {SP. Nrc-1} (Ng, Kennedy et al. 2000), στο *Haloarcula marismortui* (Baliga, Bonneau et al. 2004), στο *Natromonas pharaonis* (Falb, Pfeiffer et al. 2005), και στο *Haloferax volcanii*.

1.5. Λιποπρωτεΐνες

Οι βακτηριακές λιποπρωτεΐνες χαρακτηρίζονται από ένα αμινοτελικό πεπτίδιο οδηγητή (N-acyl) (Inouye, Wang et al. 1977) που συνδέεται με μια κυστεΐνη, και είναι οι βασικές μεμβρανικές πρωτεΐνες στα ομοιοστατικά βακτήρια. Η κυστεΐνη στην τελευταία θέση του προτύπου και στα Gram-positive και Gram-negative βακτήρια είναι απαραίτητη. Η βιοσύνθεση των λιποπρωτεϊνών στα Gram-negative και Gram-positive βακτήρια αποτελείται από τρία στάδια, όπως φαίνεται στην εικόνα 1.4. (Sankaran and Wu 1994). Το πεπτίδιο οδηγητή των βακτηριακών λιποπρωτεϊνών κατέχει μια παρόμοια δομή, με τις κύριες διαφορές, ότι το μήκος της ακολουθίας είναι συγκριτικά πιο μικρό και το μοναδικό μοτίβο στην c-περιοχή με την μορφή [LVI][AST]-[GA]-C φέρει την ονομασία «lipobox». Η lipobox περιέχει την κυστεΐνη (C) που τροποποιεί τα λιπίδια, στη -3 θέση υπάρχει κυρίως λευκίνη (L), στην θέση -2 μια αλανίνη (A) και στην θέση -1 μια γλυκίνη (G) ή μια αλανίνη (A) (von Heijne, Steppuhn et al. 1989). Αυτά τα χαρακτηριστικά χρησιμοποιούνται για την ταυτοποίηση των λιποπρωτεϊνών στα δεδομένα πρωτεϊνικής βάσης δεδομένων. Η ακολουθία οδηγητή διασπάται από signal peptidase II (SPase II), αποκαλούμενο επίσης πεπτίδιο οδηγητή λιποπρωτεϊνών (Lsp) και αναγνωρίζει το σημείο αποκοπής στην c-περιοχή. Είναι αρκετά παρόμοιο με το πεπτίδιο οδηγητή των εκκριτικών πρωτεϊνών, οι οποίες διασπόνται από signal peptidase I (SPase I).



Εικόνα 1.4: Βιοσύνθεση των λιποπρωτεϊνών. Λιπίδια, συνδέονται στην κυστεΐνη. Πεπτίδια εμφανίζονται στα αριστερά και στα δεξιά της κυστεΐνης. Τα καταλυτικά ένζυμα καταγράφονται δίπλα από τα βέλη της αντίδρασης. (Juncker, Willenbrock et al. 2003)

1.6. Αλγόριθμοι πρόγνωσης Sec πεπτιδίου οδηγητή

1.6.1. PrediSi

Το PrediSi είναι εργαλείο πρόβλεψης του πεπτιδίου οδηγητή και της θέσης αποκοπής του, στις βακτηριακές και ευκαριωτικές ακολουθίες αμινοξέων. Σε σύγκριση με άλλα εργαλεία είναι χρήσιμο για την ανάλυση μεγάλων συνόλων δεδομένων, σε πραγματικό χρόνο με υψηλή ακρίβεια. Επιτρέπει επίσης την αξιολόγηση των πρωτεομικών συνόλων. Η μέθοδος του είναι βασισμένη στην μελέτη του πίνακα βαρών (von Heijne 1986). Η προσέγγιση του πίνακα βαρών (von Heijne 1986) εφαρμόζεται εύκολα ακόμη και σε ένα μικρο-υπολογιστή. Η οργάνωση μιας μεθόδου πρόβλεψης, παρέχει μια σωστή διάκριση μεταξύ των ακολουθιών οδηγητών και του αμινοτελικού πεπτιδίου οδηγητή στις κυτταροπλασματικές πρωτεΐνες, και μπορεί να δώσει σωστές προβλέψεις για το 75-80% την φορά όταν εφαρμόζεται στις νέες ακολουθίες (στους προκαρυωτικούς και ευκαριωτικούς οργανισμούς). Αυτό αντιπροσωπεύει ένα ουσιαστικό κέρδος πέρα από την παλαιά μέθοδο, που είχε ποσοστό επιτυχίας περίπου 65% και 45% για τις ευκαριωτικές και προκαρυωτικές πρωτεΐνες, αντίστοιχα. Αυτή η μέθοδος ήταν από τις πρώτες για την εύρεση του σημείου αποκοπής των πεπτιδίων οδηγητών και μη πεπτιδίων οδηγητών. Οι αρχικοί πίνακες χρησιμοποιούνται ακόμη και σήμερα, ασχέτως αν οι πληροφορίες για το πεπτίδιο οδηγητή έχουν αυξηθεί από το 1986.

Για την κατασκευή της συγκεκριμένης μεθόδου πρόβλεψης του PrediSi παρήχθησαν τρεις διαφορετικοί πίνακες συχνότητας που στηρίχτηκαν στα κατασκευασμένα και στοιχισμένα σύνολα δεδομένων που αναφέρονται στη δημοσίευση του Hiller (Hiller, Grote et al. 2004). Οι πίνακες βαρών είναι βασισμένοι στη συχνότητα εμφάνισης του αμινοξέος στις περιοχές των ακολουθιών με πεπτίδιο οδηγητή, επίσης λαμβάνει υπόψη του μέχρι τέσσερα αμινοξέα από το αμινοτελικό πεπτίδιο οδηγητή. Υπολογίστηκε το βέλτιστο μέγεθος της θέση του πίνακα βαρών (PWMs, position weight matrices) με τον υπολογισμό της ακρίβειας όλων των σημαντικών συνδυασμών. Στην συνέχεια προτού υπολογιστεί το αποτέλεσμα πραγματοποιήθηκε διόρθωση της συχνότητας για να ρυθμιστεί το σφάλμα δείγματος του αμινοξέος στις συγκεκριμένες πρωτεΐνες (Schreiber and Brown 2002). Το αποτέλεσμα στην εργασία αυτή υπολογίστηκε με την εξίσωση:

$$S = \sum_{i=1}^{l_{\mathrm{PWM}}} \log\left(P_i \frac{P_{\mathrm{ideal}}}{P_{\mathrm{obs}}}\right)$$

Τα χαρακτηριστικά γνωρίσματα του λογισμικού είναι:

- ότι εκπαιδεύτηκε από ακολουθίες που είναι καταχωρημένες στην βάση δεδομένων Swiss Prot.
- η ταχύτητα που το διακατέχει, και η δυνατότητα του να λειτουργεί σε Linux.
- καλύτερη αποδοτικότητα ακόμη και χωρίς την διευκρίνιση κατηγορίας οργανισμών

Είναι το γρηγορότερο διαθέσιμο δημόσιο εργαλείο για την πρόβλεψη των πεπτιδίων οδηγητών (Hiller, Grote et al. 2004). Χρησιμοποιώντας το PrediSi δεν είναι απαραίτητο να παραδοθούν τα αποτελέσματα μέσο ηλεκτρονικού ταχυδρομείου, επειδή τα αποτελέσματα παρουσιάζονται άμεσα στη μηχανή αναζήτησης στο διαδίκτυο. Στην συνέχεια παρουσιάζεται ένα παράδειγμα, με τα δεδομένα που μας βγάζει το πρόγραμμα δίνοντας τις ακόλουθες πρωτεΐνες (εικόνες 1.5, 1.6).

Settings:

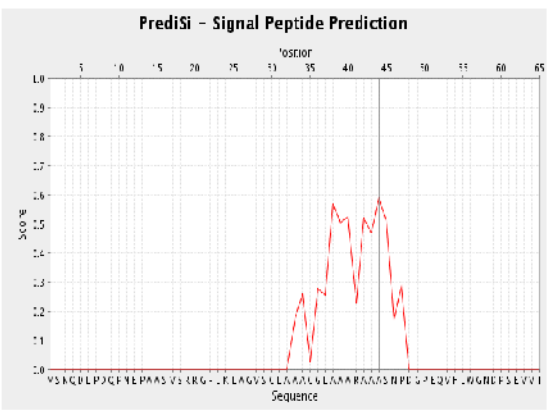
Matrix:	null
Truncation:	70 residues
Output:	null

FASTA-ID	Score	Cleavage Position	Signal Peptide ?	Chart
Q02J64	0.3956	25	N	Details
E4EKR2	0.5899	44	Y	Details
A4QFQ3	0.7375	35	Y	Details
Q8RJN8	0.5421	21	Y	Details
P31224	0.5533	26	Y	Details
P0ABD5	0.0000	17	N	Details

Εικόνα 1.5: Αποτελέσματα που μας δίνει η εφαρμογή του PrediSi, με την εισαγωγή των συγκεκριμένων ακολουθιών. Βλέπουμε να μας δίνει το ID της ακολουθίας, το σκορ συσχέτισης, το σημείο αποκοπής και αν είναι πεπτίδιο οδηγητή ή όχι.

Details:

Matrix:	Eukarya
Truncation:	70 residues
Cleavage position:	44
Score:	0.5899
Secreted protein:	predicted for secretion

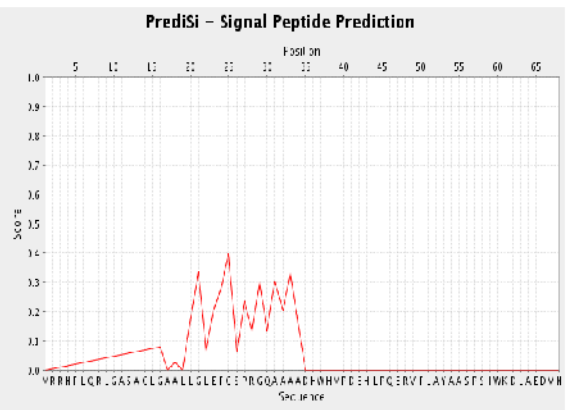


B4EKR2:

α

Details:

Matrix:	Eukarya
Truncation:	70 residues
Cleavage position:	25
Score:	0.3956
Secreted protein:	not predicted for secretion



Q02J64:

β

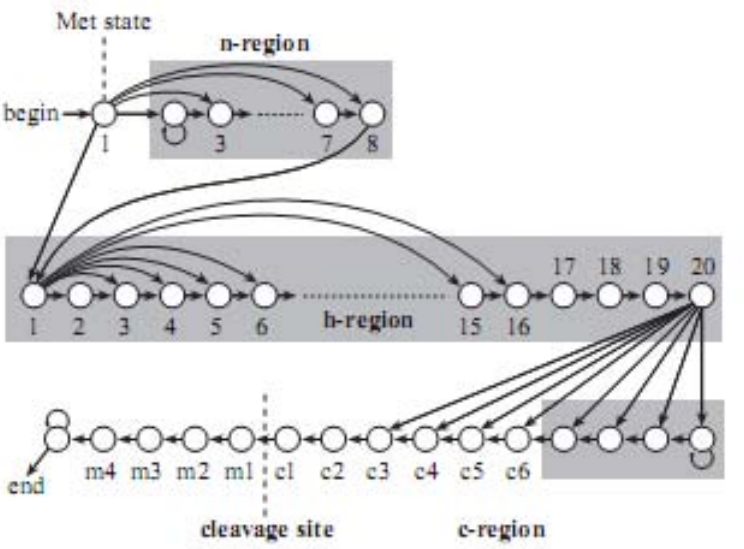
Εικόνα 1.6: Παρατηρούμε τα αποτελέσματα που παίρνουμε αν ακολουθήσουμε την σύνδεση που έχει μετά από κάθε πρωτεΐνη, στην εικόνα 1.5 (details).(α) B4EKR2 αναγνωρίζει ότι είναι ακολουθία με σημείο αποκοπής που φαίνεται στη γραφική παράσταση, αλλά και στο σκιασμένο κομμάτι στην ακολουθία που βρίσκεται από κάτω. (β) Q02364, δεν αναγνωρίζει την ακολουθία με πεπτίδιο οδηγητή.

1.6.2. SignalP

Το SignalP είναι ένα εργαλείο πρόβλεψης πεπτιδίου οδηγητή Sec, με βάση τη μέθοδο των νευρωνικών δικτύων. Οι τρεις εκδόσεις του εκπαιδεύονται σε διαφορετικά σύνολα δεδομένων (ευκαρυωτικά κύτταρα, Gram-negative και Gram-positive βακτήρια), που αντικατοπτρίζουν τις σημαντικές διαφορές στα χαρακτηριστικά των πεπτιδίων οδηγητών από αυτές τις ομάδες των οργανισμών, καθώς κάθε μια δίνει καλύτερη απόδοση από ό,τι ένα εργαλείο που είναι εκπαιδευμένο για όλες τις ομάδες μαζί. Το SignalP v.1.1 (Nielsen, Engelbrecht et al. 1997) συνδυάζει δύο διαφορετικά νευρωνικά δίκτυα, (i) το ένα είναι υπεύθυνο να αναγνωρίσει τις περιοχές αποκοπής στα πλαίσια όλων των άλλων θέσεων ακολουθίας, και (ii) το άλλο να ταξινομήσει τα αμινοξέα σε σχέση με αυτά που ανήκουν στο πεπτίδιο οδηγητή ή όχι. Τα σύνολα δεδομένων για εκπαίδευση και δοκιμή των μεθόδων του SignalP περιέχουν μόνο το αμινοτελικό πεπτίδιο οδηγητή (μέχρι 70 αμινοξέα) κάθε πρωτεΐνης, και οι διαμεμβρανικές πρωτεΐνες δεν περιλήφθηκαν στο αρνητικό σύνολο (Bendtsen, Nielsen et al. 2004). Η απόφαση να χρησιμοποιηθεί μόνο το πεπτίδιο οδηγητή κάθε πρωτεΐνης βασίστηκε στην ιδέα ότι το SignalP πρέπει να κάνει αναγνώριση του πεπτιδίου που καλύπτεται από το κύτταρο σε ζώντα οργανισμό, όπου το σημείο αποκοπής του πεπτιδίου οδηγητή παρουσιάζεται μόνο μέσα σε μια συγκεκριμένη περιοχή από το αμινοτελικό πεπτίδιο οδηγητή. Ο λόγος της απουσίας των διαμεμβρανικών ελίκων (TM) στο αρνητικό σύνολο, είναι ότι δεν υπάρχουν πειραματικά στοιχεία που αποδεικνύουν την απουσία περιοχής αποκοπής σε μια διαμεμβρανική πρωτεΐνη (Nielsen, Brunak et al. 1999).

Στην δεύτερη επανέκδοση του SignalP (Nielsen and Krogh 1998) πραγματοποιήθηκε προσθήκη ενός hidden Markov model. Περιλαμβάνει δύο μεθόδους πρόβλεψης πεπτιδίου οδηγητή, το SignalP-NN που έχει καλύτερη απόδοση στη πρόβλεψη (με βάση τα νευρωνικά δίκτυα, που αντιστοιχούν στο SignalP v1.1) (Menne, Hermjakob et al. 2000) και το SignalP-HMM (με βάση τα HMMs). Για τα ευκαρυωτικά δεδομένα, το SignalP-HMM πραγματοποιεί μια πιο βελτιωμένη αναγνώριση των πεπτιδίων οδηγητών και των πεπτιδίων αγκυροβόλησης χωρίς σημείο αποκοπής (Nielsen and Krogh 1998), σε σχέση με το νευρωνικό δίκτυο του SignalP, αλλά έχει μικρότερη ακρίβεια στην πρόβλεψη της σωστής τοποθεσίας της θέσης αποκοπής (Zhang and Henzel 2004). Ο χρήστης μπορεί να επιλέξει εάν θα τρέξει το SignalP-NN, το SignalP- HMM, ή και τα δύο. Το μοντέλο του νευρωνικού δικτύου στο SignalP v2.0 είναι εκπαιδευμένο με νέα δεδομένα που προκύπτουν από Swiss-prot. Το SignalP-HMM παρέχει μια πρόβλεψη της παρουσίας ενός πεπτιδίου οδηγητή και τη θέση αποκοπής του, αλλά και μια κατά προσέγγιση κατανομή των n- h- και c-περιοχών εντός του πεπτιδίου οδηγητή. Όλα αυτά φαίνονται σε μια γραφική παράσταση που μας δίνει αυτή η έκδοση, όπως και τις πιθανότητες για κάθε θέση σε μία από αυτές τις τρεις περιοχές. Στην τρίτη και τελευταία μέχρι τώρα έκδοση του SignalP (Bendtsen, Nielsen et al. 2004), σαφώς πιο βελτιωμένη, λαμβάνονται υπόψη οι μέθοδοι που αναπτύχθηκαν στις προηγούμενες εκδόσεις, αλλά στη προκειμένη περίπτωση υπάρχει η εξάλειψη των ψευδών θετικών λόγο της ανανέωσης λεπτομερειών στα δεδομένα. Η νέα έκδοση του SignalP-HMM 3.0 φαίνεται να είναι καλύτερη στην αναγνώριση της περιοχής αποκοπής από το SignalP-HMM 2.0 (Zhang and Henzel 2004).

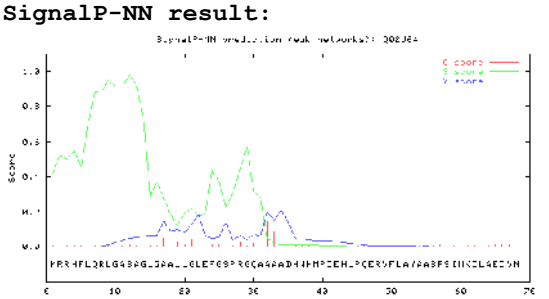
Ταυτόχρονα αναπτύχθηκαν και δύο άλλα εργαλεία βασισμένα στη μέθοδο του SignalP για την πρόγνωση πεπτιδίων οδηγητών στα αρχαία και στους χλωροπλάστες. Όσον αφορά τα εκκριτικά πεπτίδια οδηγητές στα ευκαρυωτικά κύτταρα και στα βακτήρια έχουμε αρκετές πληροφορίες στην περιγραφή τους στην βάση. Από την άλλη όμως οι πληροφορίες για τα αρχαία που είναι θετικά καταχωρημένες στην βάση είναι περιορισμένες. Στην εικόνα 1.7 βλέπουμε το μοντέλο που δημιουργήθηκε γι'αυτό το εργαλείο πρόβλεψης. Επίσης παρουσιάζεται ένα παράδειγμα από την χρήση του εργαλείου με την ακολουθίας Q02J64, και τα αποτελέσματα που μας δίνει είναι από τα μοντέλα που δημιουργήθηκαν με την μέθοδο των νευρωνικού δικτύου και των hidden Markov model (εικόνα 1.8).



Εικόνα 1.7: Το μοντέλο που χρησιμοποιείται για τα πεπτίδια οδηγητές (Nielsen and Krogh 1998).

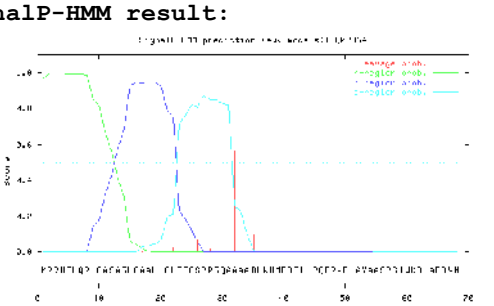
Το εργαλείο SignalP για την πρόβλεψη των πεπτιδίων οδηγητών Sec, μπορεί να προβλέπει και τα Tat πεπτίδια οδηγητή σωστά αν και οι h-περιοχές των Tat πεπτιδίων οδηγητών τείνουν να είναι λιγότερο υδροφοβικές από τα κλασσικά πεπτίδια οδηγητή Sec. Εντούτοις, τα ποσοστά επιτυχίας από το SignalP για την πρόβλεψη των Tat πεπτιδίων οδηγητών είναι χαμηλά, και η θέση της προβλεφθείσας περιοχής αποκοπής συχνά ορίζεται λανθασμένα. Αυτό οφείλεται πιθανώς στο μεγάλο μήκος της ακολουθίας του Tat πεπτιδίου οδηγητή και του γεγονότος ότι συχνά οι c-περιοχές των Tat πεπτιδίων περιέχουν βασικά αμινοξέα. Για τον παραπάνω λόγο αναπτύχθηκαν καινούργιες μέθοδοι αναγνώρισης TAT πεπτιδίων οδηγητών, που αναφέρονται στο υποκεφάλαιο 1.7.

SignalP-NN result:



```
>Q02J64                    length = 68
# Measure   Position   Value   Cutoff
signal peptide?
  max. C     34         0.167   0.32    NO
  max. Y     34         0.209   0.33    NO
  max. S     12         0.979   0.87    YES
  mean S      1-33      0.466   0.48    NO
       D      1-33      0.338   0.43    NO
# Most likely cleavage site between
pos. 33 and 34: AAA-AD → περιοχή αποκοπής
```

SignalP-HMM result:

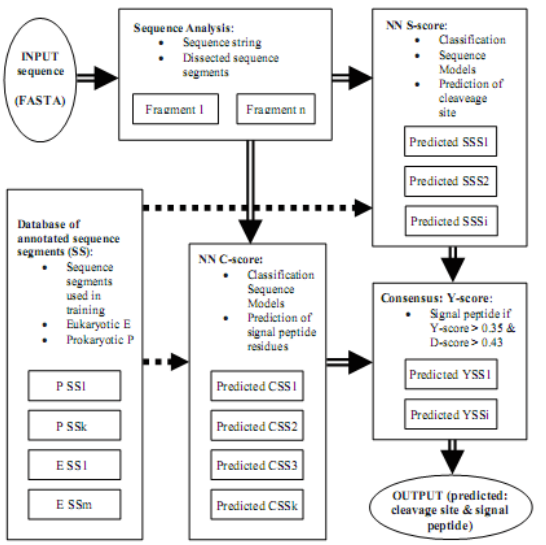


```
>Q02J64
Prediction: Signal peptide
Signal peptide probability: 0.971
→εμφανίζει το ποσοστό αξιοπιστίας παρουσίας του
πεπτιδίου οδηγητή, είναι 0.971
Signal anchor probability: 0.022
Max cleavage site probability: 0.567
between pos. 31 and 32 → περιοχή αποκοπής
```

Εικόνα 1.8: Εμφάνιση αποτελεσμάτων από το SignalP με τις δύο μεθόδους, NN και HMM, για την αναγνώριση Sec πεπτιδίου στη πρωτεΐνη Q02J64.

1.6.3. RPSP

Το εργαλείο αναγνώρισης πεπτιδίων οδηγητών στις πρωτεΐνες RPSP, είναι από τα πιο γρήγορα και έχει δημιουργηθεί με βάση δύο νευρωνικά δίκτυά που εκπαιδεύτηκαν από πρωτεΐνες που είναι καταχωρημένες στην SwissProt. Μπορούν να εκτελεσθούν για τρείς τύπους προβλέψεων: ένα για τις προκαρυωτικές ακολουθίες, ένα για τις ευκαριωτικές και ένα που δεν προσδιορίζει τον τύπο πρωτεΐνης. Το RPSP είναι βασισμένο στην ακολουθία και γι'αυτό είναι σε θέση να παρέχει προβλέψεις για μικρού μεγέθους ακολουθίες που φωσφορυλιώνονται, δεσμεύουν ligands, αλληλεπιδρούν με άλλες πρωτεΐνες και μόρια RNA (Plewczynski, Slabinski et al. 2008). Η ακρίβεια τις πρόβλεψης του RPSP είναι συγκρίσιμη, αφού για την πρόβλεψη περιοχών αποκοπής φθάνει στο 73% και περιέχει και ευκαριωτικές και προκαρυωτικές ακολουθίες, ενώ η ακρίβεια για την διάκριση των πεπτιδίων οδηγητών και των μη πεπτιδίων οδηγητών είναι πάνω από 93% για οποιοδήποτε σύνολο δεδομένων (Plewczynski, Slabinski et al. 2008). Στην εικόνα 1.9 παρουσιάζεται το διάγραμμα ροής των στοιχείων στο RPSP και στην συνέχεια στον πίνακα 1.2 βλέπουμε τα αποτελέσματα που παίρνουμε από την μέθοδο δίνοντας τις τα παρακάτω δεδομένα.



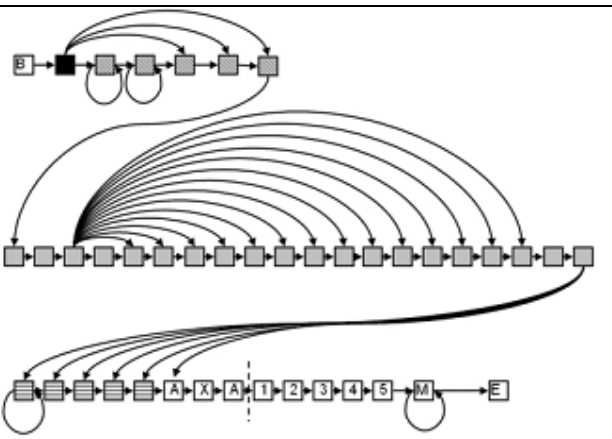
Εικόνα 1.9: Διάγραμμα ροής στοιχείων στο RPSP. Τα διπλά βέλη δείχνουν τη ροή πληροφοριών κατά τη διάρκεια της πρόβλεψης, οι διακεκομμένες γραμμές παρουσιάζουν σύνδεση στη βάση δεδομένων των σχολιασμένων τμημάτων ακολουθίας από τη βάση δεδομένων Swiss-Prot. (Plewczynski, Slabinski et al. 2008)

Πίνακας 1.2: Παρουσίαση αποτελεσμάτων που δίνει η χρήση του RPSP για τις ακόλουθες πρωτεΐνες

Sequence ID	Signal peptide	Length
Q02J64	non-signal peptide	0
B4EKR2	non-signal peptide	0
A4QFQ3	MGKHRRNNSNATRKAVAASAVALGATAAIASPAQA	35
Q8RJN8	MKVFFKITTLLLILISYQSLA	21
P31224	MPNFFIDRPIFAWVIAIIIMLAGGLA	26
P0ABD5	non-signal peptide	0

1.6.4. PRED-SIGNAL

Το PRED-SIGNAL είναι από τα λίγα εργαλεία που έχουν σκοπό την αναγνώριση του πεπτιδίου οδηγητή (SPs) στα αρχαία. Εκτελείται με συνέπεια και είναι καλύτερο σε σύγκριση με άλλα διαθέσιμα εργαλεία που εκπαιδεύθηκαν στις ακολουθίες ευκαριωτικής ή βακτηριακής προέλευσης. Η δημιουργία μοντέλου hidden Markov model του PRED-SIGNAL, είναι παρόμοια με του SignalP (Nielsen and Krogh 1998), προβλέπει την παρουσία του πεπτιδίου οδηγητή και των περιοχών αποκοπής των αρχαίων. Κάνει επίσης διακρίσεις από κυτταροπλασματικές και διαμεμβρανικές πρωτεΐνες. Αποτελείται από τρία διαφορετικά υπομοντέλα, το υπομοντέλο SP που αντιστοιχεί στην χαρακτηριστική ακολουθία SPs, το υπομοντέλο του TM που αντιστοιχεί στο τμήμα του αμινοτελικού της διαμεμβρανικής πρωτεΐνης και είναι ίδιο με αυτό που χρησιμοποιείται από το εργαλείο αναγνώρισης HMM-TM για τις α-ελικοειδείς διαμεμβρανικές πρωτεΐνες (Bagos, Liakopoulos et al. 2006), και ένα υπομοντέλο που χρησιμοποιείται για να αναγνωρίσει το αμινοτελικό πεπτίδιο οδηγητή των σφαιρικών πρωτεϊνών. Ο κεντρικός πυρήνας του μοντέλου είναι το υπομοντέλο SP. Το μοντέλο HMM του εργαλείου αυτού φαίνεται στην εικόνα 1.10.



Εικόνα 1.10: Αρχιτεκτονική του HMM χρησιμοποιείται για να μοντελοποιηθεί η εκκριτική ακολουθία SP. Κάθε γραμμή (επάνω προς τα κάτω) αντιστοιχεί στην n- h- και c- περιοχή, αντίστοιχα. Οι καταστάσεις για την n-και h-περιοχή που μοιράζονται τις ίδιες πιθανότητες εκπομπής, απεικονίζονται με το ίδιο σύμβολο. Το σημείο διάσπασης υποδεικνύεται χρησιμοποιώντας μια διακεκομμένη κατακόρυφη γραμμή μεταξύ A και 1 (το πρώτο αμινοξύ της ώριμης πρωτεΐνης). Οι καταστάσεις μεταβάσεις απεικονίζονται με βέλη. B και E αντιστοιχούν στην κατάσταση έναρξης και λήξης, αντίστοιχα, ενώ οι καταστάσεις μετά το σημείο αποκοπής (1-5 και M) χρησιμοποιούνται για να διαμορφώσουν τα πρώτα κατάλοιπα της ώριμης πρωτεΐνης. (Bagos, Tsirigos et al. 2009)

Το μοντέλο εκπαιδεύθηκε χρησιμοποιώντας τον αλγόριθμο Baum-Welch (Krogh 1994). Για τις ακολουθίες η αποκωδικοποίηση εκτελέσθηκε χρησιμοποιώντας τον τυποποιημένο αλγόριθμο Viterbi. Εκτός από την αποκωδικοποίηση Viterbi που παράγει τη βέλτιστη πορεία των καταστάσεων μετάβασης του μοντέλου, και ως εκ τούτου προβλέπει ταυτόχρονα τον τύπο της ακολουθίας (SP, TM ή Globular) καθώς επίσης και τη περιοχή αποκοπής. Παρουσιάζεται επίσης ο S1 δείκτη αξιοπιστίας (Melen, Krogh et al. 2003), ο οποίος παίρνει τις τιμές από το διάστημα (0-1) και παρέχει ένα χρήσιμο μέτρο της αξιοπιστίας της πρόβλεψης. Το εργαλείο αυτό αποδίδει ικανοποιητικά, (Bagos, Tsirigos et al. 2009), με ευαισθησία 100%, ειδικότητα 98.41% και με τον συντελεστή συσχέτισης Matthews ίσο με 0.964.

Αποτελέσματα που εκπονούμε με την εισαγωγή των παρακάτω πρωτεϊνών στο PRED-SIGNAL:

```
Q02J64        Signal 1-21
B4EKR2        Signal 1-44
A4QFQ3        Signal 1-35
```

1.6.5. PSORTb

Η πρώτη έκδοση αυτού του εργαλείου πρόβλεψης κάλυπτε μόνο τα Gram-negative βακτήρια. Αυτό το είδος βακτηρίων έχει πέντε βασικές τοποθεσίες εντοπισμού, στο κυτταρόπλασμα, στην εσωτερική μεμβράνη, στο περιπλασματικό χώρο, στην εξωτερική μεμβράνη και στον εξωκυττάριο χώρο. Στην πρωταρχική έκδοση του PSORTb1 (Nakai and Kanehisa 1991) εξετάστηκαν τέσσερις τοποθεσίες εντοπισμού: στο κυτταρόπλασμα, στο εσωτερικό της κυτταροπλασματικής μεμβράνη, του περιπλασματικού χώρου και της εξωτερικής μεμβράνης. Οι περισσότεροι κανόνες προήλθαν από πειραματικές παρατηρήσεις. Για παράδειγμα ο κανόνας για να αναγνωριστεί μια πρωτεΐνη στην εσωτερική μεμβράνη είναι η παρουσία της υδρόφοβης περιοχής στην προβλεπόμενη ώριμη πρωτεΐνη, με το χαρακτηριστικό αμινοτελικό πεπτίδιο οδηγητή ή χωρίς το σημείο αποκοπής. Επίσης σε αυτή την έρευνα αναγνωρίστηκε για πρώτη φορά το μοτίβο λιποπρωτεϊνών και στην συνέχεια μελετήθηκε η παρουσία του στο εσωτερικό και εξωτερικό της μεμβράνης. Αυτό έγινε με την μελέτη για την ύπαρξη όξινης περιοχής στο τμήμα του αμινοτελικού πεπτιδίου της ώριμης πρωτεΐνης. Αργότερα μια μικρή επανέκδοση της αρχικής έρχεται να λάβει υπόψη της την πρόγνωση στην πέμπτη τοποθεσία, στον εξωκυττάριο χώρο των πρωτεϊνών, πρωτεΐνες που μπορούν να αποτελέσουν σημαντικούς παράγοντες παθογένειας σε παθογόνους μικροοργανισμούς (Gardy, Spencer et al. 2003). Αργότερα είχε αναπτυχθεί μια επανέκδοση, το PSORTb v.2.0, που είχε ως σκοπό την βελτίωση του προηγούμενου μοντέλου διατηρώντας το υπάρχον επίπεδο ακρίβειας, αλλά επεκτάθηκε και στα Gram-positive βακτήρια. Το PSORTbv.2.0 περιλαμβάνει επίσης μια ενότητα profile, στην οποία τα συγκεκριμένα profiles HMMs εντοπισμού του πεπτιδίου οδηγητή, προήλθαν από το PROSITE v.1.8 επιλέχτηκαν για να παραγάγουν 100.0% ακριβείς προβλέψεις ενάντια του PSORTdb. Έξι profiles επιλέχτηκαν (Gardy, Laird et al. 2005), τέσσερα από τα οποία προσδιορίζουν τις Gram-negative και Gram-positive βακτηριακές κυτταροπλασματικές πρωτεΐνες και τις κυτταροπλασματικές πρωτεΐνες μεμβρανών, και τα άλλα δυο προσδιορίζουν τις Gram-positive βακτηριακές πρωτεΐνες του κυτταρικού τοιχώματος και των διαμεμβρανικών περιοχών. Το πρόγραμμα επιτυγχάνει μια ακρίβεια πρόβλεψης 96% για τα Gram-positive και Gram-negative βακτήρια (Gardy, Laird et al. 2005).

1.6.6. LipoP

Στο LipoP για την αναγνώριση των λιποπρωτεϊνών εξελίχτηκε ένα HMM, το μοντέλο αυτό χωρίζεται σε τέσσερις κλάδους (εικόνα 1.11):

1. το μοντέλο αναγνώρισης του σημείου αποκοπής από την SPase I

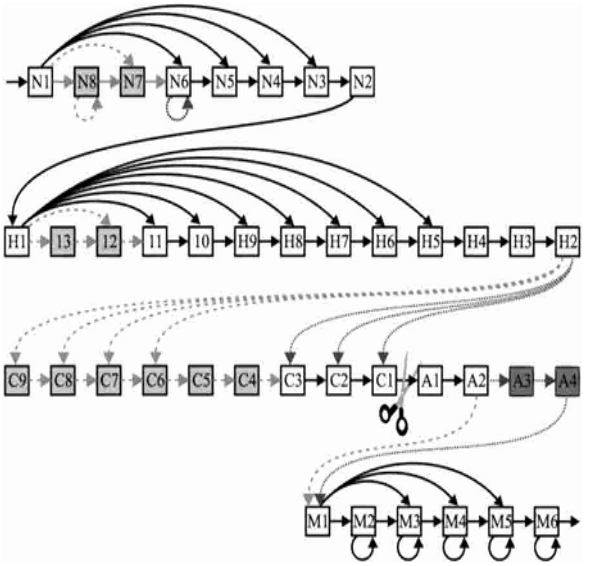
Έχει αναφέρει την μοντελοποίηση της n-περιοχής, της h και c-περιοχής. Το μοντέλο πεπτιδίου οδηγητή είναι πολύ παρόμοιο με εκείνο των Nielsen και Krogh (Nielsen and Krogh 1998), αλλά είναι λίγο πιο απλοποιημένο.

2. το υπομοντέλο αποκοπής του πεπτιδίου οδηγητή από την SPase II (λιποπρωτεϊνών)
3. το υπομοντέλο αναγνώρισης αμινοτελικού της διαμεμβρανικής πρωτεΐνης.

Πρόκειται ουσιαστικά για ένα μέρος του μοντέλου του TMHMM (Krogh, Larsson et al. 2001), στο οποίο μοντελοποιεί μόνο μια διαμεμβρανική έλικα. Ο σκοπός αυτού του τμήματος του μοντέλου είναι κατά κύριο λόγο στον περιορισμό του αριθμού των ψευδών θετικών αποτελεσμάτων από τις προβλέψεις πεπτιδίου οδηγητή και όχι να προβλέψει αν μια πρωτεΐνη έχει αμινοτελικό πεπτίδιο οδηγητή στην ακολουθία.

4. ένα υπομοντέλο αναγνώρισης κυτταροπλασματικών πρωτεϊνών

Αυτό το υπομοντέλο αποτελείται από δύο καταστάσεις: μια κατάσταση για το πρώτο αμινοξύ και μια κατάσταση για τη μετάβαση στο υπόλοιπο μοντέλο. Μόνο τα πρώτα 70 αμινοξέα χρησιμοποιήθηκαν τόσο για την εκπαίδευση όσο και για την δοκιμή του μοντέλου.

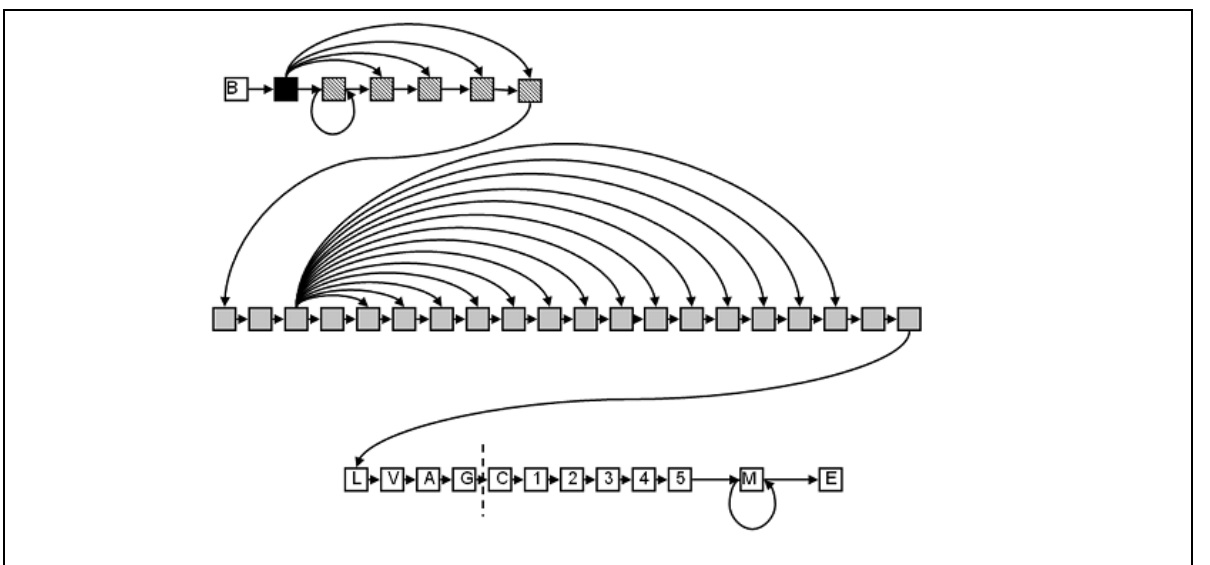


Εικόνα 1.11: Η αρχιτεκτονική των μοντέλων του SPaseI και SPaseII. Οι N-καταστάσεις διαμορφώνουν την n-περιοχή. Οι H-καταστάσεις διαμορφώνουν την h-περιοχή. Οι C-καταστάσεις και A-καταστάσεις διαμορφώνουν τις περιοχές πριν και μετά από την περιοχή αποκοπής αντίστοιχα, και οι M-καταστάσεις διαμορφώνουν τα υπόλοιπα κατάλοιπα. (Juncker, Willenbrock et al. 2003)

1.6.7. PRED-LIPO

Το PRED-LIPO είναι ένα εργαλείο πρόγνωσης πεπτιδίου οδηγητή στις λιποπρωτεΐνες και αναπτύχτηκε μετά από τη μέθοδο του LipoP με σκοπό να ανιχνεύσει περιοχές που δεν κάλυπτε το προηγούμενο εργαλείο, δηλαδή να κάνει έλεγχο λιποπρωτεϊνών στα Gram positive βακτήρια. Αφού αναλύθηκαν οι πληροφορίες, που πάρθηκαν από τις βάσεις δεδομένων, συγκρίνοντας τις πρωτεΐνες Gram negative και Gram positive των βακτηρίων, βρέθηκαν κάποιες μικρές διαφορές και εκεί στηρίχθηκε η δημιουργία μοντέλου HMM, για την εύρεση SPs στις λιποπρωτεΐνες. Το εργαλείο επίσης κάνει διάκριση στα εκκριτικά πεπτίδια οδηγητές, δηλαδή στην περιοχή αποκοπής τους, καθώς επίσης και στο αμινοτελικό πεπτίδιο στις ακολουθίες των διαμεμβρανικών πρωτεϊνών.

Το μοντέλο του PRED-LIPO αποτελείται από τέσσερα υπομοντέλα, το υπομοντέλο των λιποπρωτεϊνών που αντιστοιχεί στο πεπτίδιο οδηγητή που διασπάται από το SPase II (εικόνα 1.12), το υπομοντέλο πεπτίδιο οδηγητή που αντιστοιχεί στα εκκριτικά πεπτίδια οδηγητές που διασπόνται από το SPase I, το υπομοντέλο για την αναγνώριση του αμινοτελικού στη διαμεμβρανική πρωτεΐνη (TM), και ένα υπομοντέλο για το αμινοτελικό πεπτίδιο στις σφαιρικές πρωτεΐνες (Bagos, Tsirigos et al. 2008).

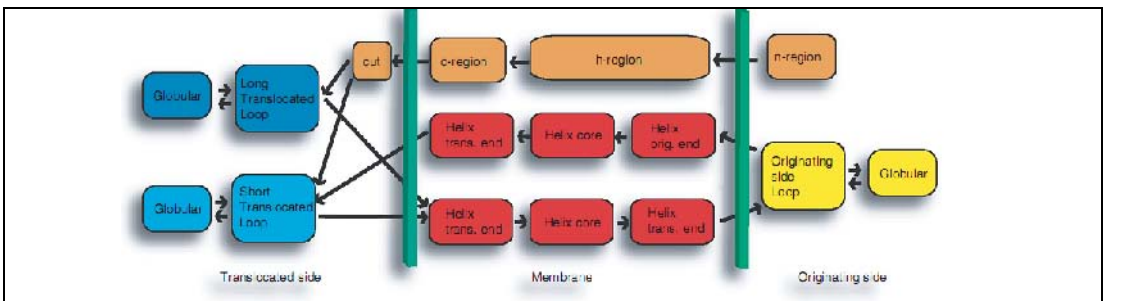


Εικόνα 1.12: Το υπομοντέλο που αντιστοιχεί στα πεπτίδια οδηγητών λιποπρωτεϊνών. Οι καταστάσεις στις n και h-περιοχές που μοιράζονται τις ίδιες πιθανότητες εκπομπής απεικονίζονται με το ίδιο χρώμα. Η περιοχή αποκοπής παρουσιάζεται με μια κάθετη διακεκομμένη γραμμή μεταξύ του G και του C. Οι συνδέσεις απεικονίζονται με τα βέλη (Bagos, Tsirigos et al. 2008).

Τα αποτελέσματα που πάρθηκαν από το PRED-LIPO όταν δοκιμάστηκε στις λιποπρωτεΐνες των Gram-Positive βακτηρίων είναι καλύτερα από αυτά του LipoP (Bagos, Tsirigos et al. 2008). Για τον λόγο αυτό θα πρέπει να χρησιμοποιείται αποκλειστικά για την πρόβλεψη αυτών των αλληλουχιών. Ένα από τα κύρια πλεονεκτήματα της μεθόδου πρόβλεψης είναι η υψηλή εξειδίκευση, δεδομένου ότι προβλέπει ελάχιστα (<0,3%), ψευδώς θετικά. Παρόμοια αποτελέσματα δίνει και στην πρόβλεψη παρουσίας των εκκριτικών πεπτιδίων οδηγητών.

1.6.8. Phobius

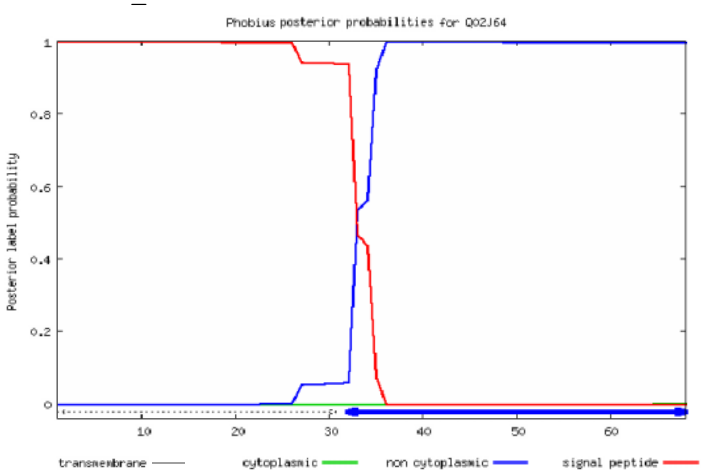
Το Phobius είναι ένα εργαλείο που δημιουργήθηκε για την πρόβλεψη της διαμεμβρανικής τοπολογίας. Τα πιο πολλά εργαλεία συχνά προβλέπουν τα πεπτίδια οδηγητές ως διαμεμβρανικά τμήματα και αντίστροφα. Για την επίλυση αυτού του προβλήματος, αφαιρούνται οι πρωτεΐνες με το πεπτίδιο οδηγητή πριν πραγματοποιηθεί η πρόβλεψη διαμεμβρανικών πρωτεϊνών. Για να επιλυθεί καλύτερα το πρόβλημα σχεδιάστηκε ένα Hidden Markov Model (Kall, Krogh et al. 2007), που περιέχει υπομοντέλα για τα πεπτίδια οδηγητές και για τα διαμεμβρανικά τμήματα (εικόνα 1.13).



Εικόνα 1.13:Το μοντέλο του Phobius. Το μοντέλο περιλαμβάνει τα υπομοντέλα για τα πεπτίδια οδηγητές, τις έλικες διαμεμβρανικών πρωτεϊνών, κυτταροπλασματικούς βρόχους και δύο διαφορετικά υπομοντέλα για τους μη-κυτταροπλασματικούς βρόχους.(Kall, Krogh et al. 2007)

Ένα πλεονέκτημα είναι ότι το εργαλείο αύξησε ακόμη και την υψηλή ακρίβεια που είχε το εργαλείο TMHMM (Kall, Krogh et al. 2007) στην σωστή πρόβλεψη των διαμεμβρανικών τοπολογιών από 44.5% ποσοστό αναγνώρισης σε 53.9%. Η ακρίβεια των προβλέψεων μπορεί να βελτιωθεί κατά πολύ, εάν μπορούμε να λάβουμε υπόψη πληροφορίες σχετικά με τη θέση ενός τμήματος της ακολουθίας σε μια περιορισμένη πρόβλεψη (Melen, Krogh et al. 2003) .
Αποτελέσματα που παίρνουμε από τη μέθοδο αυτή με την πρωτεΐνη Q02J64.

```
Prediction of Q02J64
ID   Q02J64
FT   SIGNAL        1      31       → αναγνώριση περιοχής πεπτιδίου οδηγητή Sec
FT   REGION        1       8       N-REGION. → n-περιοχή
FT   REGION        9      20       H-REGION. → h-περιοχή
FT   REGION       21      31       C-REGION. → c-περιοχή
FT   TOPO_DOM     32      68       NON CYTOPLASMIC.
```



1.6.9. Philius

Το μοντέλο του Philius στηρίχτηκε στο προηγούμενο δημοσιευμένο HMM του Phobius, και συνδυάζει ένα υπομοντέλο για την ακολουθία του πεπτιδίου οδηγητή και ένα για τις διαμεμβρανικές πρωτεΐνες. Το διαφορετικό όμως εδώ είναι ότι το μοντέλο βασίζεται στην ισχυρότερη κατηγορία δυναμικών Μπεϋζιανών δικτύων (DBNs). Επίσης παρέχει τον πρωτεϊνικό τύπο, το τμήμα και τους βαθμούς εμπιστοσύνης της τοπολογίας για την ενίσχυση και την ερμηνεία των προβλέψεων. Στον έλεγχο αξιοπιστίας του μοντέλου, παρατηρήθηκε μια βελτίωση 13% καλύτερη από το Phobius, όσον αφορά την ακρίβεια πρόβλεψης πλήρης τοπολογίας στις διαμεμβρανικές πρωτεΐνες, με ευαισθησία και ειδικότητα 0.96 στην αναγνώριση του πεπτιδίου οδηγητή (Reynolds, Kall et al. 2008). Το εργαλείο αυτό αντιμετωπίζει το πρόβλημα μεταξύ τεσσάρων βασικών τύπων πρωτεϊνών: σφαιρικές πρωτεΐνες (G), σφαιρικές πρωτεΐνες με πεπτίδιο οδηγητή (SP+G), διαμεμβρανικών πρωτεϊνών (TM) και διαμεμβρανικών πρωτεϊνών με πεπτίδιο οδηγητή (SP+TM). Προβλέπει επίσης την περιοχή αποκοπής των πεπτιδίων οδηγητών και την πλήρη τοπολογία για τις διαμεμβρανικές πρωτεΐνες.

1.7. Αλγόριθμοι πρόγνωσης Tat πεπτιδίου οδηγητή

1.7.1. TATFIND

Με αφορμή την έρευνα αν οι διαμεμβρανικές πρωτεΐνες που βρίσκονται στα Halobacteriaceae φέρουν το Tat και Sec πεπτίδιο οδηγητή (Rose, Bruser et al. 2002), αναπτύχθηκε ένα εργαλείο αναγνώρισης του συστήματος Tat με την ονομασία TATFIND για να ανιχνευτούν τα υποθετικά υποστρώματα Tat μέσα στα *Halobacterium sp.* NRC-1. Αυτό το πρόγραμμα είναι βασισμένο στη θέση και την ακολουθία του μοτίβου Tat, καθώς επίσης και στη θέση, το μήκος και την υδροφοβικότητα της περιοχής μετά από το μοτίβο της δίδυμης αργινίνης. Παρά το γεγονός ότι το μοτίβο που αναγνωρίζεται από το TATFIND είναι βασισμένο απλώς στις ακολουθίες των πεπτιδίων οδηγητών ορισμένων υποθετικών υποστρωμάτων Tat είναι επίσης πιθανό ότι ένας ακόμα μεγάλος αριθμός δεδομένων από τα haloarchaeal θα αναγνωριστεί από το μοντέλο να φέρει το Tat πεπτίδιο οδηγητή. Η επανέκδοση του TATFIND ορίζει ένα υπόστρωμα Tat που βρίσκεται στις πρωτεΐνες και πληροί τα δύο παρακάτω κριτήρια: (i) η παρουσία ενός μοτίβου (X^{-1}) R0R1(X^2) (X^3) (X^4) εντός των πρώτων 35 αμινοξέων των πρωτεϊνών, όπου σε κάθε θέση X αντιπροσωπεύει ένα καθορισμένο σύνολο των επιτρεπόμενων αμινοξέων, και (ii) η παρουσία μιας επέκτασης 13 αμινοξέων μετά από το R0R1. Στην συνέχεια μια ταυτοποίηση νέων Tat πεπτιδίων οδηγητών από το *P.aeruginosa*, οδήγησε στην επέκταση μοντέλου του TATFIND για να καταστεί δυνατή η μεθειονίνη στη θέση X^{-1} και μια γλουταμίνη στη θέση X^4 (Dilks, Rose et al. 2003).
Στην συνέχεια παρουσιάζονται τα αποτελέσματα που παίρνουμε από το TATFIND, για τις πρωτεΐνες A4QFQ3, B4EKR2, P0ABD5, P31224, Q02J64, Q8RJN8.

```
TATFIND - version 1.4 December, 2005; Rose, Brueser, Dilks, Kissinger
& Pohlschroder
SEARCH RESULTS for ../../html/data/src/seqData7366.faa:
```

- Results for A4QFQ3: FALSE
  No TAT recognition site in the first 35 aa.
- Results for B4EKR2: TRUE
  The twin arg pattern has been found: SRRGFL
  The hydrophobic region is LAGVSGLAAAGGLAAA
  The length of the hydrophobic region is 16
  The hydrophobicity score is 0.46
  Four residues after the twin argenine there is a charged residue.
  Data is:
  MSNQDLPDQPNEPAASVSRRGFLKLAGVSGLAAAGGLAAARAAASNPDGPEQVHLWGNDPSEVVI
- Results for P0ABD5: FALSE
  The twin arg pattern has been found: SRRGFL
  No TAT recognition site in the first 35 aa.
- Results for P31224: FALSE
  The twin arg pattern has been found: SRRGFL
  No TAT recognition site in the first 35 aa.
- Results for Q02J64: FALSE
  The twin arg pattern has been found: SRRGFL
  No TAT recognition site in the first 35 aa.
- Results for Q8RJN8: FALSE
  The twin arg pattern has been found: SRRGFL
  No TAT recognition site in the first 35 aa.

1.7.2. TatP

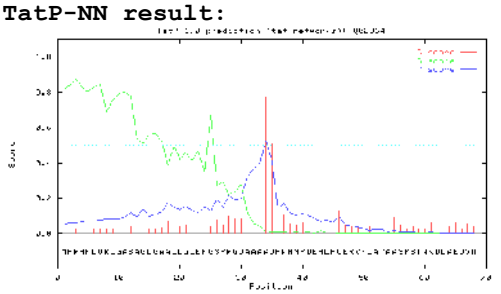
Στο εργαλείο TatP, για την πρόβλεψη Tat πεπτιδίου οδηγητή, ανακτήθηκαν ακολουθίες που φέρουν την ακολουθία αναγνώρισης του Tat πεπτιδίου οδηγητή προκειμένου να εκπαιδευτεί μια υπολογιστική μέθοδος για την διάκριση των Sec και Tat πεπτιδίων οδηγητών. Χρησιμοποιεί δύο διαφορετικά νευρωνικά δίκτυα για την αντιμετώπιση των προβλημάτων πρόβλεψης Tat πεπτιδίων οδηγητών. Ένα δίκτυο για την αναγνώριση του σημείου αποκοπής, και ένα δίκτυο για τον καθορισμό εάν ένα δεδομένο αμινοξύ ανήκει στο Tat πεπτίδιο οδηγητή ή όχι. Τα NN που αναπτύχθηκαν είναι ίδια με αυτά που αναφέρθησαν σε προηγούμενη βιβλιογραφία (Nielsen, Engelbrecht et al. 1997; Bendtsen, Nielsen et al. 2004). Το εργαλείο αυτό (Bendtsen, Nielsen et al. 2005) είναι σε θέση να ταξινομήσει 91% από τις 35 ακολουθίες με το Tat πεπτίδιο οδηγητή και να προβλέψει το σημείο αποκοπής των Tat πεπτιδίων οδηγητών από το 84% αυτού του συνόλου. Παράγει τις πιο λίγες ψευδώς θετικές προβλέψεις σε διαφορετικά σύνολα δεδομένων. Η περαιτέρω προσθήκη ενός νευρωνικού δικτύου βοηθά στη σωστή ταξινόμηση των Tat και Sec πεπτιδίων οδηγητών. Δίνει την δυνατότητα διάκρισης των Tat και Sec πεπτιδίων οδηγητών από τις κυτταροπλασματικές πρωτεΐνες που φέρουν ένα παρόμοιο μοτίβο. Επίσης παρέχει την δυνατότητα στο χρήστη να αλλάξει το τρόπο διεξαγωγής της αναγνώρισης, αλλάζοντας το μοτίβο, μέσω μιας ευκολίας που δίνει το πρόγραμμα τις Perl, πριν την χρήση του.

Στο μοντέλο του εφαρμόστηκε ένα νευρωνικό δίκτυο για να προσδιορίσει τις λιγότερο υδροφοβικές h-περιοχές των Tat πεπτιδίων οδηγητών και το ποσοστό των ψευδώς θετικών μειώθηκε στο 0,4%. Πράγματι, αυτό δείχνει τη δύναμη του νευρωνικού δικτύου, το οποίο είναι ικανό να κάνει διακρίσεις μεταξύ της υδροφοβικότητας των Sec / SRP και των Tat πεπτιδίων οδηγητών (Bendtsen, Nielsen et al. 2005).

Στην συνέχεια βλέπουμε τα αποτελέσματα που παίρνουμε με την χρήση του TatP για τις πρωτεΐνες Q02J64, B4EKR2 (εικόνα 1.14).

>Q02J64

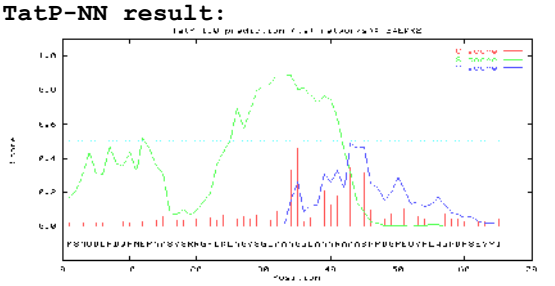
TatP-NN result:



```
>Q02J64                    length = 68
# Measure   Position  Value   Cutoff   Tat signal
peptide?
  max. C     34        0.776   0.51     YES
  max. Y     34        0.525   0.35     YES
  max. S      3        0.871   0.75     YES
  mean S      1-33     0.532   0.30     YES
  max. D      1-33     0.528   0.36     YES
# Most likely cleavage site between pos. 33 and
34: AAA-AD →περιοχή αποκοπής
# Found RRHFL as Tat motif starting at pos. 2
# Used regex: RR.[FGAVML][LITMVF]
//
```

>B4EKR2

TatP-NN result:



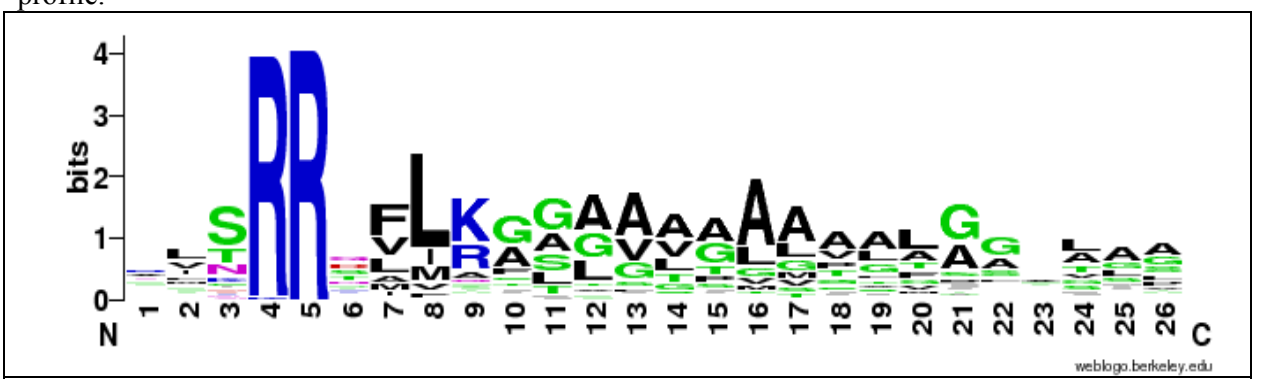
```
>B4EKR2                    length = 65
# Measure   Position  Value   Cutoff   Tat signal
peptide?
  max. C     35        0.461   0.51     NO
  max. Y     43        0.487   0.35     YES
  max. S     33        0.889   0.75     YES
  mean S      1-42     0.480   0.30     YES
  max. D      1-42     0.484   0.36     YES
# Most likely cleavage site between pos. 42 and
43: ARA-AA
# Found RRGFL as Tat motif starting at pos. 19
# Used regex: RR.[FGAVML][LITMVF]
//
```

Εικόνα 1.14: Παρουσίαση αποτελεσμάτων από το TatP για τις πρωτεΐνες Q02J64, B4EKR2

Τα profiles PF10518 (PFAM) και TIGR01409 (TIGR), ακολουθούν ένα μοντέλο HMM αναγνώρισης Tat πεπτιδίου οδηγητή. Η στοίχιση από την οποία προέκυψε το μοτίβο (S/T)-R-R-X-F-L-K. Όπως παρατηρείται στις εικόνες 1.15 και 1.16 αντίστοιχα δεν λαμβάνουν υπόψη τους την θετικά φορτισμένη n-περιοχή, με αποτέλεσμα να υστερεί στην ανεύρεση του σημείου αποκοπής. Τα pHMMs αυτά, χαρακτηρίζονται από (i) την περιοχή με το χαρακτηριστικό μοτίβο της δίδυμης αργινίνης RR, (ii) μια υδροφοβική h-περιοχή, και (iii) μια πολική περιοχή c-περιοχή, με το χαρακτηριστικό μοτίβο πεπτίδιο αποκοπής AXA. Εντούτοις το ποσοστό επιτυχίας τους για την πρόγνωση του Tat πεπτιδίου οδηγητή είναι πολύ υψηλά όπως θα δούμε παρακάτω.

1.7.3. PF10518

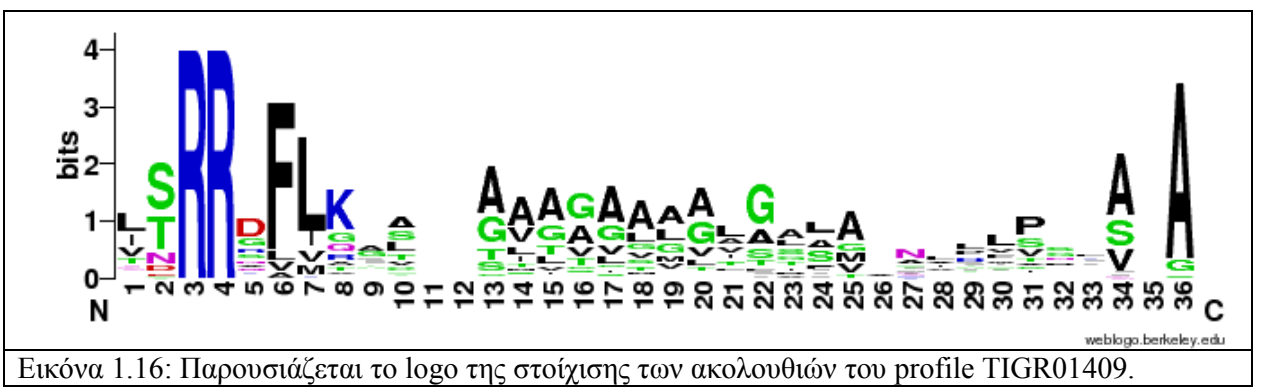
Για την αναγνώριση Tat πεπτιδίου οδηγητή, δίνει την δυνατότητα το εργαλείο στην ιστοσελίδα http://myhits.vital-it.ch/cgi-bin/msa_hub, να επιλέξει ο χρήστης αν θα κάνει την αναγνώριση του με το profile στο HMMER 2.0 ή HMMER 3.0. Στην συνέχεια ο χρήστης μπορεί να επιλέξει κάποιες παραμέτρους και να διαμορφώσει όπως ο ίδιος επιθυμεί την πρόγνωση του Tat πεπτιδίου. Ακόμη ο χρήστης έχει την επιλογή να κάνει μια απλή στοίχιση των ακολουθιών με βάση το profile αυτό. Στην αρχική ιστοσελίδα του PF10518, εκτός από τις στοιχίσεις ακολουθιών που εκπαιδεύτηκε το πρόγραμμα, παρουσιάζεται το hmmer logo του profile.



Εικόνα 1.15: Παρουσιάζεται το logo της στοίχισης των ακολουθιών του profile PF10518

1.7.4. TIGR01409

Στην αρχική ιστοσελίδα του TIGR01409 έχει διαθέσιμες τις ακολουθίες και την στοίχιση τους. Επίσης έχει μια παραπομπή http://blast.jcvi.org/web-hmm/ στο εργαλείο που λαμβάνει υπόψη του το profile. Στο εργαλείο αυτό δίνεται η δυνατότητα στο χρήστη να κάνει αναζήτηση σε μια ακολουθία πρωτεϊνών, για να αναγνώριση του Tat πεπτιδίου οδηγητή. Οι επιλογές που δίνει στο χρήστη είναι να προσδιορίσει το cutoff, και σε ποια βάση θα κάνει αναγνώριση.



Εικόνα 1.16: Παρουσιάζεται το logo της στοίχισης των ακολουθιών του profile TIGR01409.

Στην συνέχεια ακολουθεί ένα παράδειγμα τις εξόδου που παίρνουμε από το profile αυτό, κάνοντας εισαγωγή της ακολουθίας των πρωτεϊνών Q02J64, B4EKR2.

```
hmmpfam - search one or more sequences against HMM database
HMMER 2.3.2 (Oct 2003)
Copyright (C) 1992-2003 HHMI/Washington University School of Medicine
Freely distributed under the GNU General Public License (GPL)
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
HMM file:                 ALL_LIB_bin.HMM
Sequence file:            hmmpfam-search-381-1285501355.in
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query sequence: Q02J64
Accession:      [none]
Description:    [none]

Scores for sequence family classification (score includes all domains):
Model     Description                                       Score    E-value  N
--------  -----------                                       -----    ------- ---
        [no hits above thresholds]

Parsed for domains:
Model     Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
--------  ------- ----- -----    ----- -----      -----  -------
        [no hits above thresholds]

Alignments of top-scoring domains:
        [no hits above thresholds]
//
Query sequence: B4EKR2
Accession:      [none]
Description:    [none]

Scores for sequence family classification (score includes all domains):
Model     Description                                       Score    E-value  N
--------  -----------                                       -----    ------- ---
TIGR01409 TAT_signal_seq: Tat (twin-arginine transloc    19.5      0.016   1
TIGR02811 formate_TAT: formate dehydrogenase region T   -14.6       0.25   1

Parsed for domains:
Model     Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
--------  ------- ----- -----    ----- -----      -----  -------
TIGR02811   1/1       9    55 ..     1    69 []   -14.6     0.25
TIGR01409   1/1      17    42 ..     1    34 []    19.5    0.016

Alignments of top-scoring domains:
TIGR02811: domain 1 of 1, from 9 to 55: score -14.6, E = 0.25
                   *->MsddqkadlsRRdLLKglgvgaaAGavaAaTgrvaPqsAsvasepse
                      +  + +a +sRR +LK +gv    +++aAa g++a  + ++as p+
      B4EKR2     9    QPNEPAASVSRRGFLKLAGV----SGLAAAGGLAA--ARAAASNPDG 49

                      paekkkkGYRETqHIRdYYrtL<-*
                      p         E  H       L
      B4EKR2    50 P---------EQVH-------L    55

TIGR01409: domain 1 of 1, from 17 to 42: score 19.5, E = 0.016
                   *->lsRRdFLkaaaaaagaaaalgalanlllpsparA<-*
                      +sRR+FLk a ++ g+aaa g++        arA
      B4EKR2    17    VSRRGFLKLA-GVSGLAAAGGLA-------AARA    42
```

1.8. Σκοπός της εργασίας

Είναι να αναπτυχτεί ένα καινούργιο Hidden Markov Model (HMM) για την αναγνώριση Tat και Sec πεπτιδίου οδηγητή. Το μοντέλο αυτό θα λαμβάνει υπόψη τα μοτίβα που θα στηρίζονται στις στοιχίσεις Sec και Tat πεπτιδίου οδηγητή. Οι στοιχίσεις θα πραγματοποιηθούν σε δεδομένα που ήταν πειραματικά θετικά σχολιασμένα από τη βάση, για την παρουσία του Sec και Tat πεπτιδίου οδηγητή. Με τις στοιχίσεις αυτές και με την βοήθεια του πακέτου HMMER 2.3.2 θα αναπτυχθούν μοντέλα pHMMs.

Τέλος με τα profiles αυτά θα τρέξουμε μια σειρά από δεδομένα (Tat, Sec, TM, Cyto), και θα παρουσιάσουμε τα αποτελέσματα όσον αφορά την πρόγνωση πεπτιδίου οδηγητή αλλά και την πρόβλεψη σημείου αποκοπής. Επίσης θα συγκρίνουμε τα ποσοστά πρόβλεψης με τις προγενέστερες μεθόδους, του PrediSi, SignalPv3 και άλλων εργαλείων.

2. Υλικά και Μέθοδοι

2.1. Δεδομένα

Το σύνολο των πρωτεϊνών με το Tat πεπτίδιο οδηγητή συλλέχθηκε με συγκεκριμένο ερώτημα από την βάση δεδομένων UniProt (Wu, Apweiler et al. 2006). Η βάση δεδομένων UniProt **(U**niversal **P**rotein **R**esource**)**, είναι ο πιο γνωστός παροχής της πρωτεϊνικής ακολουθίας, ο οποίος δημιουργείται από την συμμετοχή τριών βάσεων δεδομένων, SwissProt, TrEMBL και PIR-PSD. Περιέχει ένα μεγάλο όγκο πληροφοριών σχετικά με τη βιολογική λειτουργία των πρωτεϊνών που προέρχονται από την ερευνητική βιβλιογραφία. Η βάση δεδομένων UniProtKB/SwissProt, περιλαμβάνει το σχολιασμό πρωτεϊνών με χειροκίνητη εισαγωγή, ενώ η UniProtKB/ TrEMBL περιέχει τον αυτόματο σχολιασμό πρωτεϊνών. Η ύπαρξη παραπομπών σε δημοσιευμένα αρχεία που αφορούν τις ακολουθίες των πρωτεϊνών, αλλά και ο σχολιασμός τους, επιτρέπουν στους επιστήμονες την μελέτη πρωτεϊνών αλλά και την επιβολή ερωτημάτων.

Επειδή από το σύνολο των πρωτεϊνών που συλλέξαμε ήταν λίγες οι πειραματικά ελεγμένες (είχαν δηλαδή στο πεδίο FT το χαρακτηριστικό Tat-signal peptide) επιλέξαμε να λάβουμε υπόψη μας και τις πρωτεΐνες που σχολιάζονται “Putative” ή “Potential” Tat πεπτίδιο οδηγητή στην ακολουθία τους. Το αρχικό σύνολο δεδομένων υποβλήθηκε στη μείωση πλεονασμού, μετά από τις διαδικασίες που χρησιμοποιήθηκαν στις δημοσιεύσεις για το SignalP (Nielsen, Engelbrecht et al. 1997; Nielsen, Brunak et al. 1999). Στην συνέχεια με την βοήθεια των εργαλείων BLAST (Altschul, Madden et al. 1997) δημιουργήσαμε ένα αρχείο με τα δεδομένα μας στοιχισμένα. Ακολούθως το αρχείο αυτό το χρησιμοποιήσαμε και με την βοήθεια κάποιον προγραμμάτων που δημιουργήθηκαν στη perl, τα δεδομένα απέκτησαν μια κατάλληλη μορφή. Με την μορφή αυτήν που είχαν τα δεδομένα μου, με την βοήθεια ενός άλλου προγράμματος, και με κριτήριο ένα κατώτατο όριο ομοιότητας σε 20 ίδια αμινοξέα στις ακολουθίες, μειώσαμε και άλλο τα δεδομένα μας.

Οι κυτταροπλασματικές και διαμεμβρανικές πρωτεΐνες με το Sec πεπτίδιο οδηγητή, συλλέχθηκαν όπως περιγράφονται σε προηγούμενες δημοσιεύσεις στην ανάπτυξη του PRED-LIPO (Bagos, Tsirigos et al. 2008) και του CW-PRED (Litou, Bagos et al. 2008) με δύο κύριες διαφορές. Αρχικά, δεδομένου ότι το PRED-LIPO και το CW-PRED εκπαιδεύθηκαν στα Gram-positive βακτήρια, επαναλάβαμε την ίδια διαδικασία προκειμένου να περιληφθούν οι πρόσθετες πρωτεΐνες από τα Gram-negative βακτήρια. Και κατά δεύτερο λόγο, αφαιρέσαμε από το σύνολο δεδομένων τις πρωτεΐνες πεπτιδίων οδηγητών Sec που βρέθηκε σύμφωνα με τον σχολιασμό της UniProt ότι εξάγονται από την Tat οδό. Εν συντομία, οι πρωτεΐνες με το πεπτίδιο οδηγητή Sec εξήχθησαν από το σύνολο εκπαίδευσης του SignalPv2 (Nielsen, Engelbrecht et al. 1997; Nielsen, Brunak et al. 1999), οι κυτταροπλασματικές πρωτεΐνες από το σύνολο Menne (Menne, Hermjakob et al. 2000) και οι διαμεμβρανικές πρωτεΐνες (Dilks, Rose et al.) από τα σύνολα δεδομένων που σχολιάστηκαν σε προηγούμενες δημοσιεύσεις (Moller, Kriventseva et al. 2000; Jayasinghe, Hristova et al. 2001; Chen and Rost 2002; Ikeda, Arai et al. 2003).

Όλες οι πρωτεΐνες ήταν βακτηριακής προέλευσης (Gram-positive ή Gram-negative) και αποκλείσαμε σκόπιμα τις ακολουθίες αρχαίων, δεδομένου ότι ο αριθμός πρωτεϊνών με την πειραματικά ελεγμένη περιοχή αποκοπής πεπτιδίων οδηγητών, είναι πολύ μικρός (Bagos, Tsirigos et al. 2009).

Τα δεδομένα που χρησιμοποιήσαμε για να εκπαιδεύσουμε το μοντέλο μας ήταν:

- 150 ακολουθίες πρωτεϊνών με Tat πεπτίδιο οδηγητή, από τις οποίες οι 119 ήταν Gram-negative και οι 31 Gram-positive βακτηρίων (πίνακας A_1.2).
- 328 ακολουθίες πρωτεϊνών με Sec πεπτίδιο οδηγητή, από τις οποίες οι 216 ήταν Gram-negative και οι 112 Gram-positive βακτηρίων (πίνακας A_1_1).
- 140 ακολουθίες διαμεμβρανικών πρωτεϊνών (TM) όπου αναγνωρίζεται το αμινοτελικό τους, από τις οποίες οι 90 ήταν Gram-negative και οι 50 Gram-positive βακτηρίων (πίνακας A_1.3).
- 288 κυτταροπλασματικές πρωτεΐνες, από τις οποίες οι 183 ήταν Gram-negative και οι 105 Gram-positive βακτηρίων (πίνακας A_1.4).

Ταυτόχρονα με σκοπό να αποκτήσουμε και ένα σύνολο δεδομένων ανεξάρτητο με αυτό που αποκτήσαμε από την βάση δεδομένων, για να χρησιμοποιήσουμε για την αξιολόγηση του μοντέλου μας, πραγματοποιήσαμε μια λεπτομερή αναζήτηση σε πρόσφατες δημοσιεύσεις. Προσδιορίσαμε τα πειραματικά προσδιορισμένα Tat πεπτίδια οδηγητές που προέρχονται από τα Gram-negative, Gram-positive βακτήρια και τα αρχαία. Οι πρωτεΐνες που βρίσκονταν ήδη στο σύνολο δεδομένων εκπαίδευσης για την δημιουργία του μοντέλου τις αφαιρέσαμε από το σύνολο δοκιμής. Με αυτό τον τρόπο στο πλήθος των πρωτεϊνών με το Tat πεπτίδιο οδηγητή, δεδομένων που θα χρησιμοποιούσαμε για την δοκιμή, δεν θα ήταν γνωστή η θέση αποκοπής τους. Αφαιρέθηκαν επίσης και μερικές πιθανές λιποπρωτεΐνες καθώς επίσης και αρκετές πρωτεΐνες με πεπτίδιο αγκυροβόλησης (Hatzixanthis, Palmer et al. 2003; Bachmann, Bauer et al. 2006; Bachmann , Brigitte et al. 2006; Aldridge, Spence et al. 2008). Επιπλέον αφαιρέθηκαν και πρωτεΐνες που ήταν λανθασμένα σχολιασμένες, στη βάση δεδομένων Uniprot, που ανακαλύφθηκαν κατά τη διάρκεια της αναζήτησης των δημοσιεύσεων, όπως Q3L8N0 (Yikmis, Arenskotter et al. 2008).

Στην συνέχεια ανασύρθηκαν από την UniProt βακτηριακές πρωτεΐνες, που αναφέρθηκαν στο Menne και τους συνεργάτες του (Menne, Hermjakob et al. 2000) και έχουν ένα πειραματικά ελεγχόμενο πεπτίδιο οδηγητή Sec. Αφαιρέσαμε έπειτα τις πρωτεΐνες που ήταν ήδη παρούσες στο σύνολο του SignalPv2, δηλαδή βρίσκονταν στο σύνολο εκπαίδευσης του μοντέλου. Οι πρωτεΐνες που φέρουν ένα Sec καθώς επίσης και ένα Tat πεπτίδιο οδηγητή υποβλήθηκαν άλλη μια φορά στη μείωση πλήθους, που αυτό μας εξασφαλίζει ότι δεν υπήρχε καμία ομοιότητα με τις πρωτεΐνες του συνόλου που χρησιμοποιήθηκε στη δημιουργία του μοντέλου. Ανακτήσαμε επίσης τις βακτηριακές κυτταροπλασματικές πρωτεΐνες από την UniProt και αποκλείσαμε τις καταχωρήσεις που χαρακτηρίστηκαν “Potential”, “Putative” και “By Similarity”. Δεδομένου ότι ο αριθμός ακολουθιών ήταν μεγάλος, αυτές υποβλήθηκαν στη μείωση του συνόλου με όριο 30% μιας στοίχισης τουλάχιστον 80 αμινοξέων χρησιμοποιώντας τις πλήρεις ακολουθίες και άλλη μια φορά αφαιρέσαμε τις ομόλογες ή τις πρωτεΐνες που βρίσκονταν και στο σύνολο εκπαίδευσης.

Τέλος, προκειμένου να εξεταστεί η μέθοδός μας στις διαμεμβρανικές πρωτεΐνες (TM), χρησιμοποιήσαμε τις πειραματικά προσδιορισμένες διαμεμβρανικές πρωτεΐνες από τα βακτήρια που χρησιμοποιήθηκαν για την ανάπτυξη της μεθόδου του PSORTb (Gardy, Laird et al. 2005). Από αυτό το σύνολο, αφαιρέσαμε, επίσης τις πρωτεΐνες που βρίσκονταν και στο σύνολο εκπαίδευσης, τις πρωτεΐνες με ένα υποθετικό πεπτίδιο οδηγητή (βασισμένο στο σχολιασμό από την βάση) και τέλος εκτελέσαμε τη μείωση του συνόλου με όριο 30% μιας στοίχισης τουλάχιστον 80 αμινοξέων.

Τα αποτελέσματα από τις παραπάνω διαδικασίες καταγράφονται στη συνέχεια:

- 273 ακολουθίες με το Sec πεπτίδιο οδηγητή, από τις οποίες οι 193 ήταν Gram-negative και οι 80 Gram-positive βακτηρίων (πίνακας A_2.1).
- 75 ακολουθίες με το Tat πεπτίδιο οδηγητή, από τις οποίες οι 18 ακολουθίες ήταν Gram-negative, οι 45 τα Gram-positive βακτηρίων και 12 από τα αρχαία (πίνακας A_2.2).
- 192 ακολουθίες διαμεμβρανικών πρωτεϊνών (Dilks, Rose et al.), από τις οποίες 136 ήταν Gram-negative και οι 56 Gram-positive βακτηρίων (πίνακας A_2.3).

- 601 κυτταροπλασματικές πρωτεΐνες (cyto), από τις οποίες οι 407 ήταν Gram-negative και 194 Gram-positive βακτηρίων (πίνακας A_2.4).

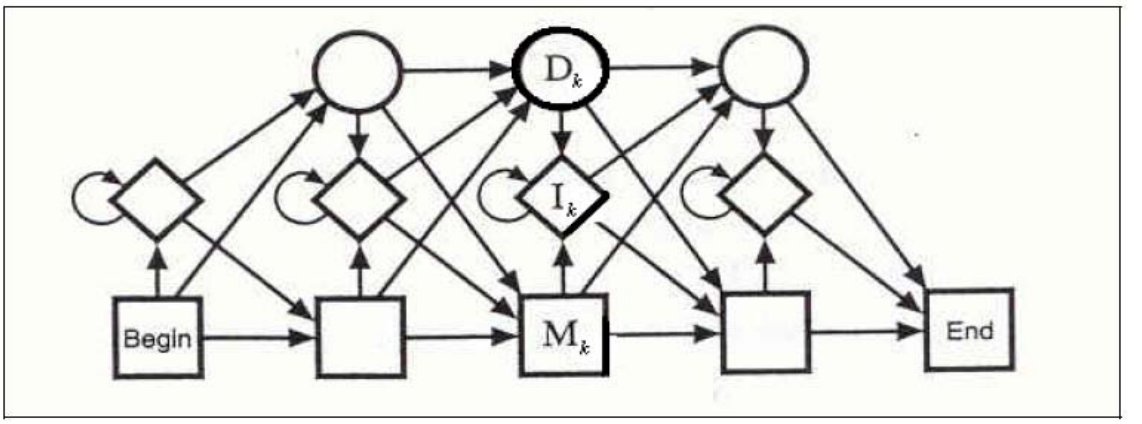
Τέλος, με την έρευνα στις δημοσιεύσεις προσδιορίσαμε μερικές πρόσθετες πρωτεΐνες με τα πεπτίδια οδηγητές που κατέχουν στην n-περιοχή του μοτίβου τους, το RR, αλλά επαληθεύονται πειραματικά πως δεν φέρουν το Tat πεπτίδιο οδηγητή (Palmer, Sargent et al. 2005; Widdick, Dilks et al. 2006; Widdick, Eijlander et al. 2008). Αυτές οι πρωτεΐνες χρησιμοποιήθηκαν ως πρόσθετο αρνητικό σύνολο δοκιμής για την αξιολόγηση της ειδικότητας των μεθόδων που αναπτύχθηκαν στην εργασία.

2.2. Hidden Markov Model (HMM)

Τα Hidden Markov Models έχουν μια δομή που αποτελείται από ένα σύνολο κρυφών καταστάσεων, ένα σύνολο παρατηρούμενων συμβόλων και δύο σύνολα πιθανοτήτων. Οι πιθανότητες μετάβασης από την μια κατάσταση στην άλλη, ακόμα και αν στο μοντέλο οι παράμετροι είναι γνωστοί, το μοντέλο εξακολουθεί να είναι «hidden» και οι πιθανότητες εκπομπής. Σε ένα απλό μοντέλο HMM έχουμε δύο καταστάσεις που συμβολίζουν την έναρξη και τον τερματισμό του μοντέλου B(Begin) και E(End), αντίστοιχα. Οι καταστάσεις του μοντέλου συνδέονται μεταξύ τους, λαμβάνοντας υπόψη κάποιες πιθανότητες συσχέτισης. Μπορούν επίσης να αναπαραστήσουν στατιστικές παραστάσεις των πρωτεϊνικών οικογενειών που προέρχονται από τα μοντέλα πολλαπλών στοιχίσεων των ακολουθιών και έχουν χρησιμοποιηθεί για τον εντοπισμό ομόλογων πρωτεϊνών με σημαντική επιτυχία.

Με το σκεπτικό της λειτουργίας των HMMs, αναπτύχθηκε από τον Eddy ένα λογισμικό πακέτο, με την ονομασία HMMER, που σκοπό έχει να πραγματοποιεί πρωτεϊνική ανάλυση της ακολουθίας (Eddy 1998). Γενικά χρησιμοποιείται για την αναζήτηση βάσεων δεδομένων για την αλληλουχία των ομόλογων ακολουθιών πρωτεϊνών, καθώς και για την κατασκευή στοίχισης ακολουθίας πρωτεϊνών. Υλοποιεί τις μεθόδους που χρησιμοποιούν πιθανολογίες μοντέλων που ονομάζεται "profiles hidden Markov model» (pHMMs). Σε σύγκριση με το BLAST, το FASTA και άλλα εργαλεία στοίχισης ακολουθιών, το HMMER στοχεύει να είναι πολύ πιο ακριβής μέθοδος και πιο ικανή να ανιχνεύει εξ αποστάσεως ομόλογα.

Το πακέτο HMMER, είναι ελεύθερα διαθέσιμο στο διαδίκτυο, εύκολο στη χρήση και δεν απαιτεί τη γνώση του πώς λειτουργούν τα profiles HMMs (εικόνα 2.1). Στην συνέχεια ακολουθεί μια σύντομη περιγραφή των βασικών προγραμμάτων και τη χρήση τους. Για να δημιουργηθεί ένα profile HMM υπάρχει στο πρόγραμμα η εντολή hmmbuild, το οποίο λαμβάνει μια πολλαπλή στοίχιση ακολουθίας εισόδου και παράγει ένα αρχείο που αντιπροσωπεύει το profile HMM. Αυτό το πρόγραμμα θα διαρκέσει για ένα δευτερόλεπτο που τρέχει για όλες εκτός από τις μεγαλύτερες στοιχίσεις. Το profile HMM μπορεί να χρησιμοποιηθεί για να ψάξει τις ακολουθίες. Εντούτοις, το πακέτο HMMER παρέχει επίσης ένα άλλο πρόγραμμα αποκαλούμενο hmmcalibrate. Αυτό το πρόγραμμα παίρνει ένα profile και το ψάχνει ενάντια σε ένα ίδιο σύνολο 5.000 πρωτεϊνών. Αυτό χρησιμοποιείται για να υπολογίσει τις τιμές mu και lambda που επιτρέπουν την ακριβή εκτίμηση των e-value για το profile HMM. Το πακέτο HMMER περιέχει ένα πρόγραμμα αποκαλούμενο hmmsearch που χρησιμοποιείται για να ανιχνεύσει το profile HMM ενάντια σε μια βάση δεδομένων των ακολουθιών.



Εικόνα 2.1: Σχηματική αναπαράσταση ενός τυπικού profile Hidden Markov Model. Οι καταστάσεις που παρατηρούνται σε ένα τέτοιο μοντέλο (εκτός αυτών της εκκίνησης και του τερματισμού) χωρίζονται σε 3 κατηγορίες: Καταστάσεις Σύμπτωσης (Match states) Mk →τετράγωνα, Καταστάσεις Εισαγωγής (Insertion states) Ik → ρόμβοι, Καταστάσεις Απαλοιφής (Deletion states) Dk → κύκλοι. Συνδέονται με τις αντίστοιχες πιθανότητες μεταβάσεως, που συμβολίζονται με βέλη.

Η δημιουργία δύο profiles hidden Markov models (pHMMs) έγινε χρησιμοποιώντας το πακέτο HMMER 2.3.2 (Durbin, Eddy et al. 1998). Είναι κατάλληλο για την μοντελοποίηση πρωτεϊνικών οικογενειών, χρησιμοποιήθηκε στην κατασκευή των profiles, δεδομένου ότι έχει αποδειχθεί ότι υπό ορισμένες συνθήκες, μπορεί επίσης να χρησιμοποιηθεί για να μοντελοποιήσει τα χαρακτηριστικά γνωρίσματα ακολουθίας των πεπτιδίων οδηγητών (Zhang and Wood 2003; Zhang and Henzel 2004). Αρχικά, δημιουργήσαμε τις πολλαπλές στοιχίσεις όπως αναφέρονται στο υποκεφάλαιο 2.1. Δηλαδή στοιχίσαμε τις ακολουθίες με πεπτίδια οδηγητές Tat σε τέσσερις περιοχές:

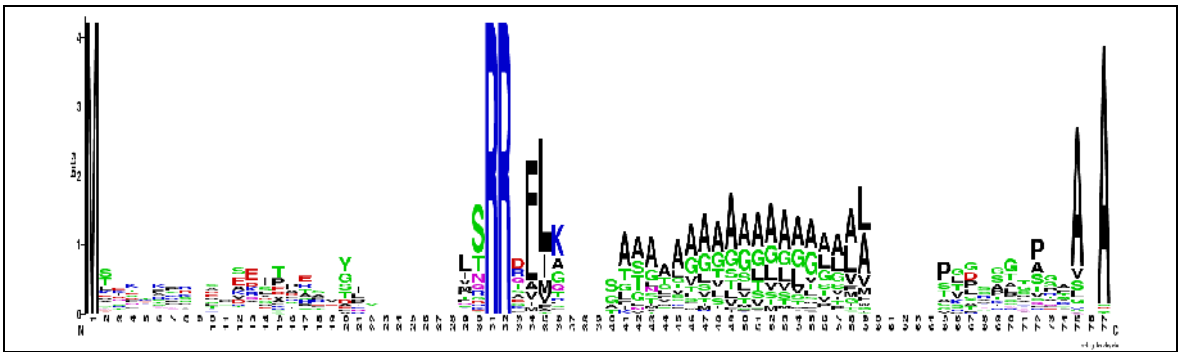
(i) θετικά φορτισμένη n-περιοχή, που φαίνεται με χρώμα μπλε στον πίνακα 2.1,

(ii) ενδιάμεση περιοχή με το χαρακτηριστικό μοτίβο της δίδυμης αργινίνης RR, διακρίνεται με χρώμα κόκκινο στον πίνακα 2.1,

(iii) μια λιγότερο υδροφοβική h-περιοχή που εκτείνεται στη μεμβράνη, στον πίνακα 2.1 με χρώμα πράσινο και

(iv) μια συνήθως μικρού μήκους πολική c-περιοχή με το χαρακτηριστικό μοτίβο πεπτίδιο αποκοπής AXA, με χρώμα πορτοκαλί στο πίνακα 2.1. Μετά από τις στοιχίσεις των ακολουθιών στην εικόνα 2.2. βλέπουμε το χαρακτηριστικό logo από τις στοίχισεις.

Πίνακας 2.1: Ακολουθίες με Tat πεπτίδιο οδηγητή				
P76342	MKKNQFLKESDVTAESVFF	MKRRQVLK	ALGISATALSL	PHAAHA
Q57366	MTKLSGQELHAE	LSRRAFLS	YTAAVGALGLCGTSLLA	QGARA
P39185	MK	ISRRDFIK	QTAITATASVAGVTL	PAGA
Q9RK81	MSQTPA	VSRRLLLG	SAAATGALATGIGSAA	PVAAA
Q9RI72	MQQDGTQQDRIKQSPAPLNG	MSRRGFLG	GAGTLALATASGLLL	PGTAHA

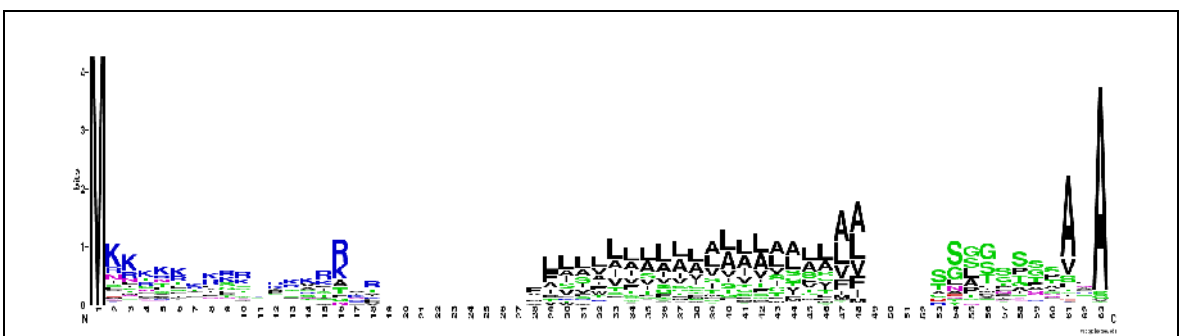


Εικόνα 2.2: Παρουσιάζεται η στοίχιση του πεπτιδίου οδηγητή Tat σε logo

Ενώ τα δεδομένα από της πρωτεΐνες που έχουν μόνο το χαρακτηριστικό πεπτίδιο οδηγητή Sec τις στοιχίσαμε σε τρεις χαρακτηριστικές περιοχές:

(i) θετικά φορτισμένη n-περιοχή, που φαίνεται με χρώμα μπλε,
(ii) μια υδροφοβική h-περιοχή που εκτείνεται στη μεμβράνη, με χρώμα πράσινο και
(iii) μια c-περιοχή μικρού μήκους με χρώμα πορτοκαλί, η στοίχιση φαίνεται στον πίνακα 2.2. Ακολουθεί το logo των στοιχίσεων (εικόνα 2.3).

Πίνακας 2.2: Ακολουθίες με Sec πεπτίδιο οδηγητή			
P0AEB2	MNTIFSARIMKR	LALTTALCTAFI	SAAHA
P31716	MLKNK	ILATTLSVSLLAPLANPLL	ENAKA
P20041	MNHRYT	LLALAAAAL	SAGAHA
P11278	MKKR	ALLLSMSVLAML	YIPAGQA
P31133	MTALNKK	WLSGLVAGALMAV	SVGTLA



Εικόνα 2.3: Παρουσιάζεται η στοίχιση του Sec πεπτιδίου οδηγητή σε logo

Οι στοιχίσεις όλων το δεδομένων φαίνονται στον πίνακα (B_1.1, B_1.2, παράρτημα Β). Παίρνοντας ξεχωριστά αυτά τα δυο αρχεία στοίχισης δημιουργήσαμε τα pHMMs (πίνακες B_1.3, B_1.4, παράρτημα Β) χρησιμοποιώντας την εντολή hmmbuild, βάζοντας παράμετρο 0.95, σύμφωνα με το εμπειρικό αποτέλεσμα από το (Zhang and Wood 2003), του πακέτου HMMER. Και στην συνέχεια για την τελειοποίηση του μοντέλου μας κάναμε και βαθμονόμηση με την εντολή hmmcalibrate των δυο καινούργιων αρχείων hmm που δημιουργήσαμε. Στη συνέχεια έγινε αναζήτηση του Tat πεπτιδίου οδηγητή με το αρχείο TAT.hmm και την εντολή hmmsearch από το HMMER στα δεδομένα μας: ακολουθίες με Tat και Sec πεπτίδιο οδηγητή που χρησιμοποιήθηκαν στην δημιουργία του hmm σε μορφή FASTA και τα δεδομένα από τις διαμεμβρανικές και κυτταροπλασματικές πρωτεΐνες, πάλι σε μορφή FASTA.

Τα αποτελέσματα που αναφέρονται για τα δεδομένα των πρωτεϊνών με Tat και Sec πεπτιδίων οδηγητών αντιστοιχούν σε μια διαδικασία cross validation 25 ακολουθιών, όπου κάθε σύνολο περιέχει έναν εξίσου ισορροπημένο αριθμό Tat πεπτιδίων οδηγητών, Sec πεπτιδίων οδηγητών, των διαμεμβρανικών (TM) και κυτταροπλασματικών (cyto) πρωτεϊνών. Η διαδικασία εκπαίδευσης του μοντέλου αποτελείτο από όλα τα δεδομένα μας αφαιρώντας κάθε φορά από 25 ακολουθίες, οι υπόλοιπες αποτελούσαν σύνολο εκπαίδευσης, και εκτέλεση της δοκιμής έγινε με τις 25 πρωτεΐνες του συνόλου που αφαιρούνταν κάθε φορά. Αυτή η διαδικασία επαναλαμβάνεται αφαιρώντας από το αρχικό σύνολο δεδομένων κάθε φορά 25 πρωτεΐνες, για όλα τα υποσύνολα στο σύνολο εκπαίδευσης, και η τελική ακρίβεια πρόβλεψης συνοψίζει τα αποτελέσματα όλων των ανεξάρτητων δοκιμών. Στην συνέχεια λαμβάνοντας υπόψη αν το score που πάρθηκε από το cross validation, για κάθε συγκεκριμένη πρωτεΐνη ήταν θετικό μεγαλύτερο από το score που πάρθηκε από την αναζήτηση με το αντίθετο pHMM παρέμενε στο σύνολο που προϋπήρχε, αν ήταν μικρότερο τότε εντασσόταν στο αντίθετο σύνολο, αν πάλι και στις δυο περιπτώσεις το score ήταν αρνητικό τότε δεν βρισκόταν σε κανένα σύνολο.

Για τα μέτρα της ακρίβειας σε κάθε δυαδικό πρόβλημα ταξινόμησης (Tat υποστρώματα έναντι των μη-Tat υποστρωμάτων, πεπτίδια οδηγητές Sec έναντι των μη πεπτιδίων οδηγητών), χρησιμοποιήσαμε το ποσοστό των σωστά ταξινομημένων θετικών παραδειγμάτων (ευαισθησία, 1), το ποσοστό των σωστά ταξινομημένων αρνητικών παραδειγμάτων (ειδικότητα, 2) και το συντελεστή συσχέτισης Mathews που συνοψίζει σε ένα γενικό μέτρο αληθώς θετικά (tp), τα ψευδώς θετικά (fp), τα αληθώς αρνητικά (Delahanty, Kang et al.) και τα ψευδώς αρνητικά (fn) (Baldi, Brunak et al. 2000).

$$\textbf{sensitivity} = \frac{tp}{tp + fn} \qquad [1]$$

$$\textbf{specificity} = \frac{tn}{tp + fp} \qquad [2]$$

$$\textbf{mcc} = \frac{tp * tn - fp * fn}{\sqrt{(tp + fp) * (tp + fn)(tn + fp)(tn + fn)}} \qquad [3]$$

2.3. Σύγκριση με άλλες μεθόδους πρόβλεψης

Πίνακας 2.3: Παρουσιάζονται τα εργαλεία αναγνώρισης Sec και Tat πεπτιδίου, που συγκρίθηκαν με τα profiles HMMs.

Εργαλεία πρόβλεψης	Ιστοσελίδα	Δημοσίευση
TatP [1]	http://www.cbs.dtu.dk/services/TatP/	(Bendtsen, Nielsen et al. 2005)
TATFIND[1]	http://signalfind.org/tatfind.html	(Rose, Bruser et al. 2002)
PF10518[1]	http://pfam.sanger.ac.uk/family?acc=PF10518#tabview=tab0	
TIGR01409[1]	http://cmr.jcvi.org/cgi-bin/CMR/HmmReport.cgi?hmm_acc=TIGR01409	(Haft, Selengut et al. 2003)
SignalPv3[2]	http://www.cbs.dtu.dk/services/SignalP/	(Bendtsen, Nielsen et al. 2004)
Phobius[2]	http://phobius.sbc.su.se/	(Kall, Krogh et al. 2004)
Philius[2]	http://www.yeastrc.org/philius/	(Reynolds, Kall et al. 2008)
RPSP[2]	http://rpsp.bioinfo.pl/	(Plewczynski, Tkacz et al. 2008)
PrediSi[2]	http://www.predisi.de/	(Hiller, Grote et al. 2004)

[1] Εργαλεία αναγνώρισης Tat πεπτιδίου οδηγητή
[2] Εργαλεία αναγνώρισης Sec πεπτιδίου οδηγητή

3. Αποτελέσματα

3.1. Δεδομένα Κατάρτισης

Τα αναλυτικά αποτελέσματα που πάρθηκαν από την διαδικασία αναζήτησης με τα pHMMs, πίνακες (για το SEC.hmm, A_1.5, A_1.6), (για το TAT.hmm, A_1.7, A_1.8) επεξεργάστηκαν και με την βοήθεια των τύπων που αναφέρονται στο (Baldi, Brunak et al. 2000) οι ευαισθησίες και οι ειδικότητες, (τύποι, 1-3 σελ. 35). Επίσης αναφέρονται τα αποτελέσματα από τον έλεγχο των ακολουθιών με το cross validation, στον πίνακα (3.1, 3.2). Παρουσιάζονται τα αποτελέσματα ανάλογα με την προτεραιότητα που δίνουμε στο score ή στο αποτέλεσμα του TAT.hmm.

Πίνακας 3.1:HMMER 3X3 (προτεραιότητα στο score)

	Predicted			
	Tat	**Non-signal**	**Signal Peptide**	**Total**
Tat	148 (98.66%)	2 (1.33%)	0 (0.00%)	150 (100%)
Non-signal	1 (0.24%)	414 (96.73%)	13 (3.03%)	428 (100%)
Signal Peptide	3 (0.92%)	41 (12.5%)	284 (86.59%)	328 (100%)

Πίνακας 3.2 :HMMER 3X3 (προτεραιότητα στο Tat.hmm)

	Predicted			
	Tat	**Non-signal**	**Signal Peptide**	**Total**
Tat	150 (100.00%)	0 (0.00%)	0 (0.00%)	150 (100%)
Non-signal	1 (0.24%)	414 (96.73%)	13 (3.03%)	428 (100%)
Signal Peptide	14 (4.27%)	41 (12.5%)	271 (82.62%)	328 (100%)

Δίνοντας προτεραιότητα στο score, που μας δίνει το cross validation, η ακολουθία δηλαδή που έχει μεγαλύτερο από το μηδέν για score από το pHMM τότε ανήκει στο σύνολο με τα θετικά. Αν πάλι έχει score μικρότερο του μηδέν τότε ανήκει στο σύνολο δεδομένων του αντίθετου πεπτιδίου οδηγητή, εκτός βέβαια αν είναι και σε αυτό κάτω του μηδενός τότε δεν ανήκει σε κανένα σύνολο. Τα profiles HMMs με την συγκεκριμένη στοίχιση, φαίνεται να είναι πολύ καλή, αφού κάνει αναγνώριση τα 148 από τα 150 Tat πεπτίδια οδηγητές (98.67%) και προβλέπει ψευδώς μόνο 2 Tat πεπτίδια (1.33 %, όλα είναι πραγματικά πεπτίδια οδηγητές Sec), και ο παράγοντας MCC είναι ίσος με 0.93 (πίνακας 3.4). Συμπερασματικά το μοντέλο μας υπέρεχει στην πρόβλεψη πεπτιδίου οδηγητή Tat από τα άλλα αντίστοιχα εργαλεία (το TatP, το TATFIND, το PF10518, τοTIGR01409)

Στο cross validation δίνοντας προτεραιότητα στο Tat.hmm, δηλαδή αν το score είναι πάνω από το μηδέν, μετράμε την ακολουθία στο σύνολο με το Tat πεπτίδιο οδηγητή, άσχετα αν στον έλεγχο με το Sec.hmm έδωσε θετικό αποτέλεσμα. Παρατηρώντας τα αποτελέσματα στον πίνακα 3.2, έχουμε 100% επιτυχία στην αναγνώριση του Tat πεπτιδίου αλλά τα ψευδώς θετικά μας, δηλαδή ακολουθίες που ήταν σίγουρα με Sec πεπτίδιο έχουν αναγνωρισθεί ως Tat πεπτίδιο με ποσοστό 4.27%, κατάσταση που μας κάνει να επιλέγουμε την μέθοδο cross validation δίνοντας σημασία στο score, με ποσοστό ψευδώς θετικό 0.92%.

Τα αποτελέσματα που φαίνονται στον πίνακα 3.3, παρουσιάζουν την σύγκριση του pHMM SEC.hmm με τα υπόλοιπα εργαλεία. Όσον αφορά τα αποτελέσματα με το cross validation το μοντέλο κάνει αναγνώριση του πεπτιδίου Sec με ποσοστό επιτυχίας 86.89%, υστερεί δηλαδή ελάχιστα σε σχέση με άλλες γνωστές μεθόδους οι οποίες έχουν μεγαλύτερα ποσοστά επιτυχίας. Γενικά ο παράγοντας MCC είναι 0.88, μια τιμή ικανοποιητική. Όσον αφορά τώρα την πρόγνωση στις κυτταροπλασματικές πρωτεΐνες έχει βέλτιστα αποτελέσματα σε σχέση με τα υπόλοιπα εργαλεία εκτός από το RPSP. Τέλος για τις διαμεμβρανικές πρωτεΐνες τα ποσοστά επιτυχίας του pHMM είναι πολύ καλύτερα από όλα τα εργαλεία εκτός από το Philius.

Πίνακας 3.3: Σύγκριση του PRED-TAT$_{HMMER}$ (pHMM SEC.hmm) με άλλα εργαλεία, για την εύρεση της πρωτεΐνης με το Sec πεπτίδιο οδηγητή.

Method	**Sec Sps**	**Cyto**	**TMs**	**MCC**
PRED-TAT	315/328 (96.04%)	265/288 (92.01%)	130/140 (92.86%)	0.88
PRED-TAT$_{HMMER}$	285/328 * (86.89%)	285/288 (98.96%)	130/140 (92.86%)	0.88
RPSP	303/328 (92.38%)	287/288 (99.65%)	116/140 (82.86%)	0.87
PrediSi	317/328 (96.65%)	280/288 (97.22%)	108/140 (77.14%)	0.87
SignalPv3 (NN)	323/328 (98.48%)	280/288 (97.22%)	117/140 (83.57%)	0.91
SignalPv3 (HMM)	325/328 (99.09%)	283/288 (98.26%)	114/140 (81.43%)	0.91
Phobius	318/328 (96.95%)	281/288 (97.57%)	129/140 (92.14%)	0.93
Philius	318/328 (96.95%)	274/288 (95.14%)	132/140 (94.29%)	0.91

*Αποτελέσματα που πάρθηκαν με cross validation ανά 25 ακολουθίες.

Επίσης σύμφωνα με τα ποσοστά που καταγράφονται στον πίνακα 3.4 παρατηρούμε ότι το profile TAT.hmm, όσον αφορά την απλή αναζήτηση σε πρωτεΐνες που δεν υπάρχει το Tat πεπτίδιο οδηγητή, παρατηρούνται πολύ μικρές αποκλίσεις. Όσον αφορά τώρα τις διαμεμβρανικές πρωτεΐνες (TM) οι λάθος προγνώσεις είναι 0.7%, ποσοστό καλύτερο από την απόδοση του TatP. Τέλος έχουμε άριστη πρόβλεψη σε πρωτεΐνες που βρίσκονται σε κυτταροπλασματικές πρωτεΐνες (cyto).

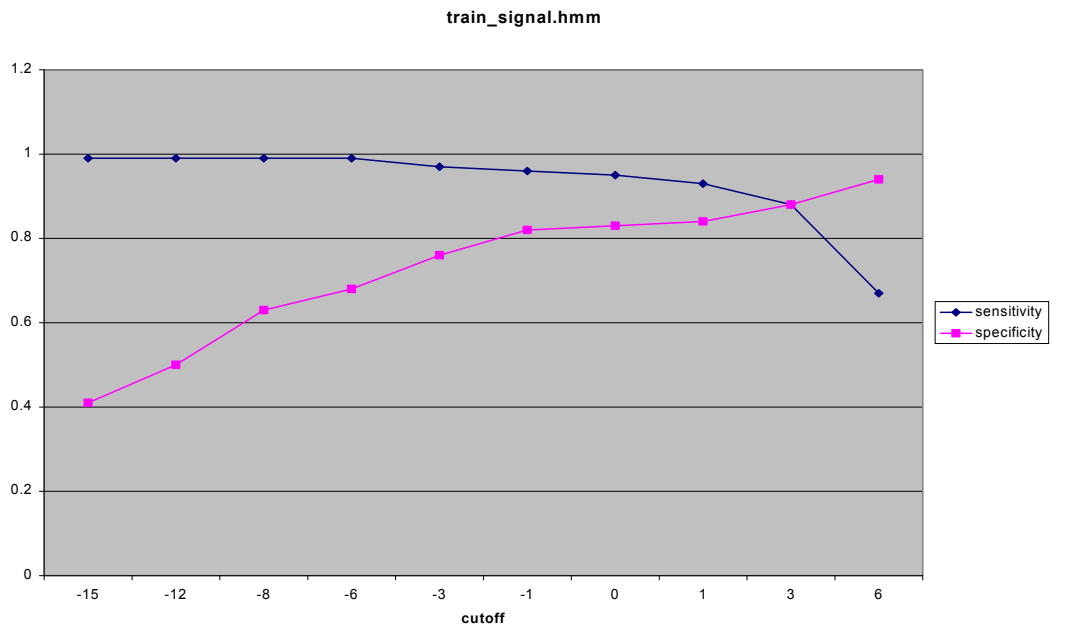
Πίνακας 3.4: Σύγκριση του PRED-TAT$_{HMMER}$ (pHMM TAT.hmm), με άλλα εργαλεία, για την εύρεση της πρωτεΐνης με το TAT πεπτίδιο οδηγητή.					
Method	**Tat SPs**	**Sec Sps**	**Cyto**	**TMs**	**MCC**
PRED-TAT	148/150 (98.67%)	319/328 (97.26%)	288/288 (100.00%)	140/140 (100.00%)	0.96
PRED-TAT$_{HMMER}$	148/150* (98.67%)	312/328 (95.12%)	288/288 (100.00%)	139/140 (99.3%)	0.93
TATFIND	134/150 (89.33%)	326/328 (99.39%)	287/288 (99.65%)	140/140 (100.00%)	0.92
TatP	130/150 (89.33%)	284/328 (86.59%)	283/288 (98.26%)	133/140 (95.00%)	0.73
PF10518	15/150 (10.00%)	328/328 (100.00%)	288/288 (100.00%)	140/140 (100.00%)	0.29
TIGR01409	105/150 (70.00%)	327/328 (99.70%)	288/288 (100.00%)	140/140 (100.00%)	0.81

*Αποτελέσματα που πάρθηκαν με cross validation ανά 25 ακολουθίες.

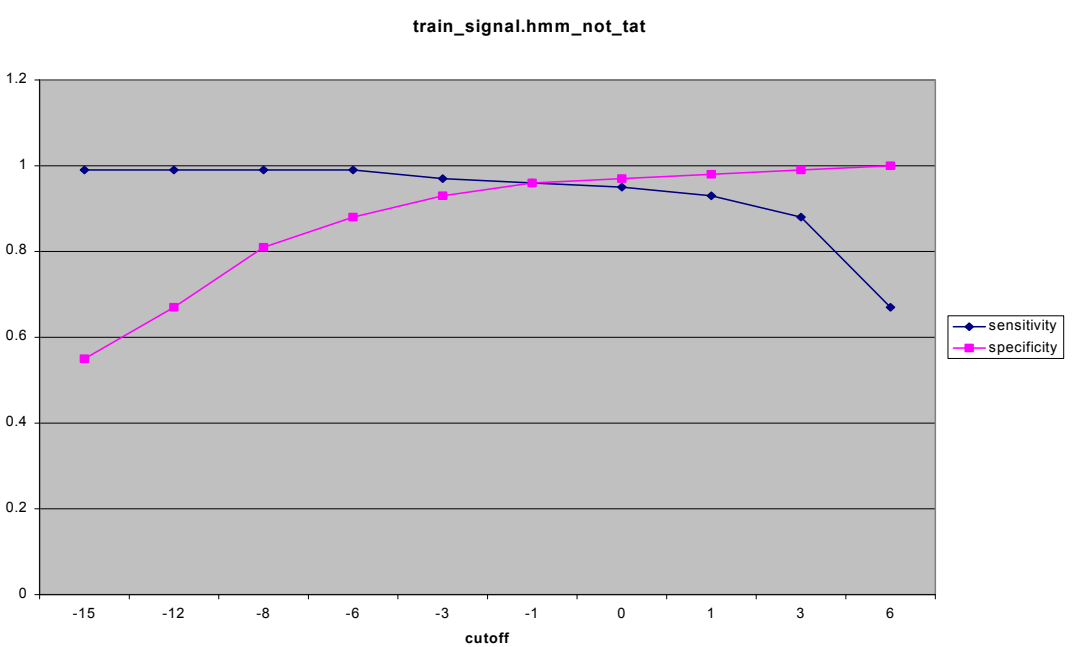
Πίνακας 3.5: Συνολικά αποτελέσματα για τις 44 πρωτεΐνες που περιέχουν στο μοτίβο τους το RR, αλλά επαληθεύονται πειραματικά ότι δεν είναι πρωτεΐνες με Tat πεπτίδιο οδηγητή.						
Uniprot	PRED-TAT	TatP	TatFind	TIGR01409	PF10518	PRED-TAT$_{HMMER}$
Συνολικά	39/44 (86.64%)	22/44 (50%)	44/44 (100%)	42/44 (95.45%)	44/44 (100%)	37/44 (84.1%)

Στον πίνακα 3.5 βλέπουμε τα αποτελέσματα που παίρνουμε κάνοντας χρήση των profiles με τις 44 πρωτεΐνες που έχουν το ενδεικτικό RR στο μοτίβο τους, αλλά είναι πειραματικά αποδεδειγμένο ότι δεν έχουν το πεπτίδιο Tat. Παρατηρούμε ότι τα pHMMs που δημιουργήσαμε έχουν ένα ποσοστό επιτυχίας 84.1%. Παρόλο που η πρόγνωση δεν είναι τέλεια, υπερέχει το μοντέλο μας κατά πολύ σε σχέση με το TatP.

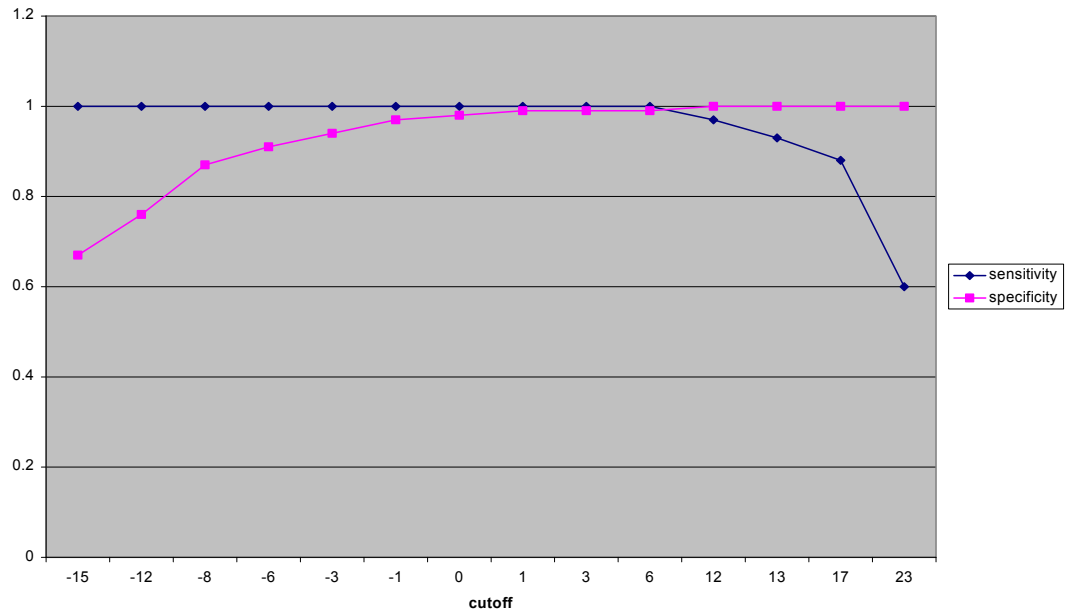
Στη συνέχεια παρουσιάζονται αναλυτικά τα αποτελέσματα από τις ευαισθησίες και τις ειδικότητες από τα δεδομένα εκπαίδευσης, σε μορφή γραφικών παραστάσεων που δημιουργήθηκαν από τους πίνακας (A_1.5, A_1.6, A_1.7, A_1.8, παράρτημα Α1). Σύμφωνα με την ένδειξη της γραφικής 3.1 το ικανοποιητικό cutoff για την πρόγνωση Sec πεπτιδίου οδηγητή είναι ίσο με +3 με ευαισθησία και ειδικότητα 0.88 και mcc 0.75. Ενώ cutoff ίσο με -1 στον ίδιο έλεγχο αφαιρόντας τις ακολουθίες με το Tat πεπτίδιο οδηγητή, με ευαισθησία και ειδικότητα 0.96 με mcc 0.93, γράφημα 3.2. Για την πρόγνωση Tat πεπτιδίου οδηγητή βρίσκεται περίπου στο σημείο 6 με ευαισθησία ίση με 1, ειδικότητα ίση με 0.99 και με mcc 0.98, γράφημα 3.3. Ενώ στον υπολογισμό χωρίς να λάβουμε υπόψη τις πρωτεΐνες με Sec πεπτίδιο οδηγητή είναι όλα τα αποτελέσματα μου ίσα με 1 στο cutoff ίσον με 6, γράφημα 3.4.



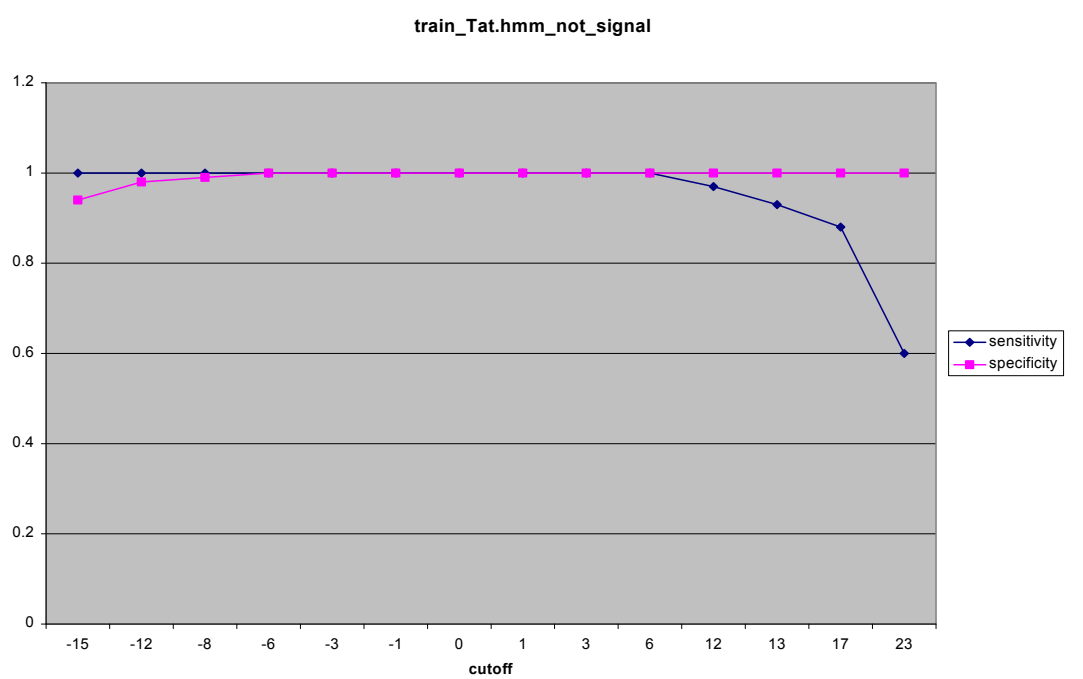
Γράφημα 3.1: Είναι η γραφική παράσταση που αντλούμε από τις ευαισθησίες και τις ειδικότητες, κάνοντας αναζήτηση το pHMM, Sec.hmm με τα δεδομένα εκπαίδευσης. Δεδομένα φαίνονται στον πίνακα A_1.5 (Παράρτημα Α1)



Γράφημα 3.2: Είναι η γραφική παράσταση που αντλούμε από τις ευαισθησίες και τις ειδικότητες, κάνοντας αναζήτηση το pHMM, Sec.hmm με τα δεδομένα εκπαίδευσης χωρίς να λάβει υπόψη του τις πρωτεΐνες με το Tat πεπτίδιο οδηγητή. Δεδομένα φαίνονται στον πίνακα A_1.6 (Παράρτημα Α1)



Γράφημα 3.3: Είναι η γραφική παράσταση που αντλούμε από τις ευαισθησίες και τις ειδικότητες, κάνοντας αναζήτηση το pHMM, Tat.hmm με τα δεδομένα εκπαίδευσης. Δεδομένα φαίνονται στον πίνακα A_1.7(Παράρτημα Α1)



Γράφημα 3.4: Είναι η γραφική παράσταση που αντλούμε από τις ευαισθησίες και τις ειδικότητες, κάνοντας αναζήτηση το pHMM, Tat.hmm με τα δεδομένα εκπαίδευσης. Χωρίς να λάβει υπόψη του τις πρωτεΐνες με το Sec πεπτίδιο οδηγητή. Δεδομένα φαίνονται στον πίνακα A_1.8 (Παράρτημα Α1)

Επίσης πραγματοποίησα την αναζήτηση Tat & Sec πεπτιδίων οδηγητών στα δεδομένα δοκιμής με τα profiles TAT.hmm, SEC.hmm που δημιούργησα.

3.2. Δεδομένα δοκιμής

Χρησιμοποιώντας το profile SEC.hmm στο ανεξάρτητο σύνολο δοκιμής, παρατηρούμε το συντελεστή συσχέτισης Mathews ίσο με 0.86 (πίνακας 3.6), τιμή η οποία επαληθεύει ότι σε γενικές γραμμές το μοντέλο μας είναι καλό. Όσον αφορά τώρα το profile αυτό, για την αναγνώριση στις κυτταροπλασματικές πρωτεΐνες παρατηρούμε να έχει καλά αποτελέσματα, με ποσοστό επιτυχίας 99.33%. Τέλος όσον αφορά στις διαμεμβρανικές πρωτεΐνες το ποσοστό επιτυχίας 90.62%, είναι βέλτιστο σε σχέση με τα υπόλοιπα εργαλεία εκτός από το Philius.

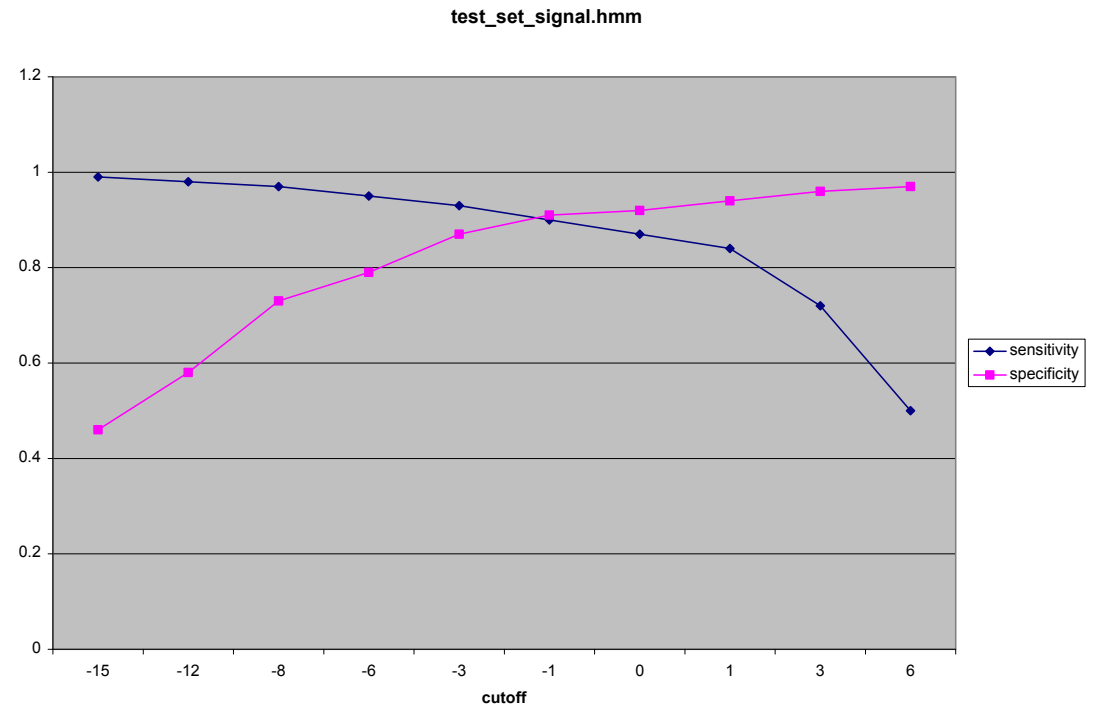
Πίνακας 3.6: Σύγκριση PRED-TAT$_{HMMER}$ (pHMM SEC.hmm), με τα άλλα εργαλεία, για την εύρεση της πρωτεΐνης με το Sec πεπτίδιο οδηγητή.

Method	Sec Sps	Cyto	TMs	MCC
PRED-TAT	252/273 (92.31%)	570/601 (94.84%)	167/192 (86.98%)	0.82
PRED-TAT$_{HMMER}$	238/273 (87.18%)	597/601 (99.33%)	174/192 (90.62%)	0.86
RPSP	249/273 (91.21%)	601/601 (100.00%)	146/192 (76.04%)	0.83
PrediSi	260/273 (95.24%)	579/601 (96.34%)	114/192 (59.38%)	0.76
SignalPv3 (NN)	252/273 (92.31%)	599/601 (99.67%)	150/192 (78.12%)	0.85
SignalPv3 (HMM)	264/273 (96.70%)	593/601 (98.67%)	134/192 (69.79%)	0.83
Phobius	249/273 (91.21%)	594/601 (98.84%)	154/192 (80.21%)	0.84
Philius	253/273 (92.67%)	582/601 (96.84%)	181/192 (94.27%)	0.88

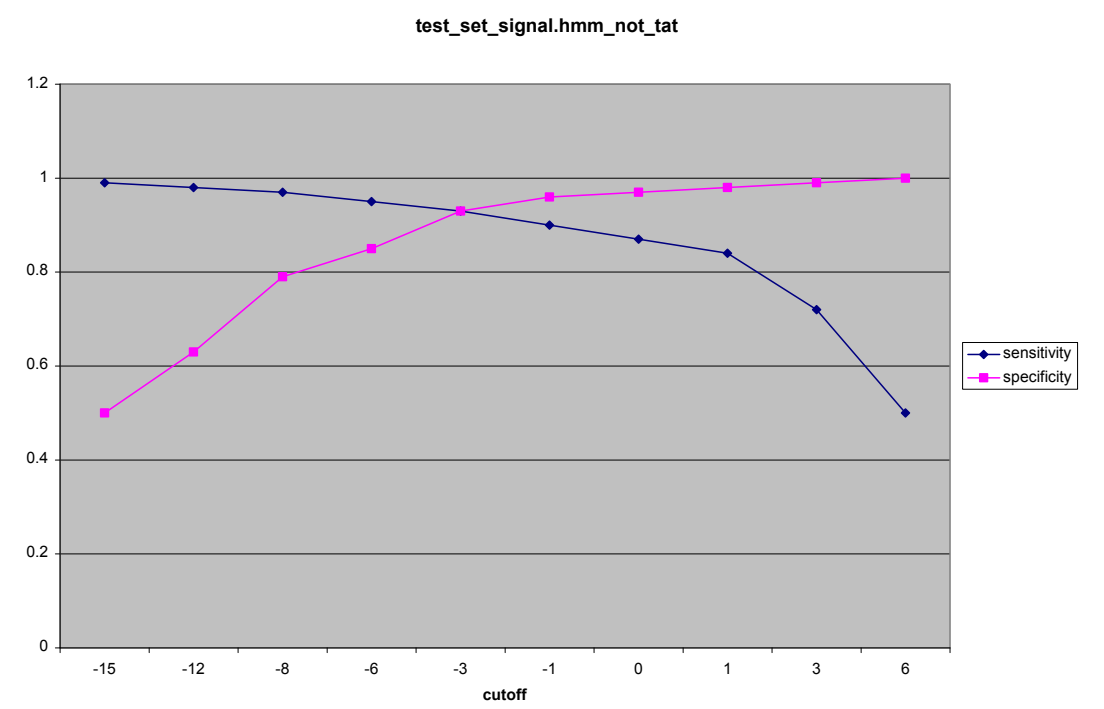
Χρησιμοποιώντας το profile TAT.hmm στις ακολουθίες του συνόλου δοκιμής (πίνακας 3.7), παρατηρούμε ότι το ποσοστό αναγνώρισης του Tat πεπτιδίου είναι 96.00%, επαληθεύεται και εδώ ότι το μοντέλο μας είναι καλύτερο σε σχέση με άλλα εργαλεία. Επίσης επαληθεύουμε για ακόμη μια φορά το μικρό ποσοστό 5.13%, αποτέλεσμα τις αναγνώρισης του Tat πεπτιδίου σε πρωτεΐνες με το Sec πεπτίδιο οδηγητή, ποσοστό που δεν είναι και τόσο μικρό όσο θα επιθυμούσαμε. Τώρα όσον αφορά τα ποσοστά επιτυχίας στις κυτταροπλασματικές και διαμεμβρανικές πρωτεΐνες επαληθεύονται με τα αποτελέσματα που πήραμε και με τα δεδομένα εκπαίδευσης.

Πίνακας 3.7: Σύγκριση PRED-TAT$_{HMMER}$ (pHMM TAT.hmm,) με τα άλλα εργαλεία, για την εύρεση της πρωτεΐνης με το Tat πεπτίδιο οδηγητή.					
Method	**Tat SPs**	**Sec Sps**	**Cyto**	**TMs**	**MCC**
PRED-TAT	71/75 (94.67%)	265/273 (97.07%)	596/601 (99.50%)	190/192 (98.96%)	0.89
PRED-TAT$_{HMMER}$	72/75 (96.00%)	259/273 (94.87%)	601/601 (100.00%)	190/192 (98.96%)	0.88
TATFIND	60/75 (80.00%)	270/273 (98.90%)	599/601 (99.67%)	192/1192 (100.00%)	0.85
TatP	62/75 (82.67%)	231/273 (84.62%)	594/601 (98.84%)	177/192 (92.19%)	0.61
PF10518	9/75 (12.00%)	273/273 (100.00%)	601/601 (100.00%)	192/192 (100.00%)	0.34
TIGR01409	47/75 (62.67%)	272/273 (99.63%)	601/601 (100.00%)	192/192 (100.00%)	0.77

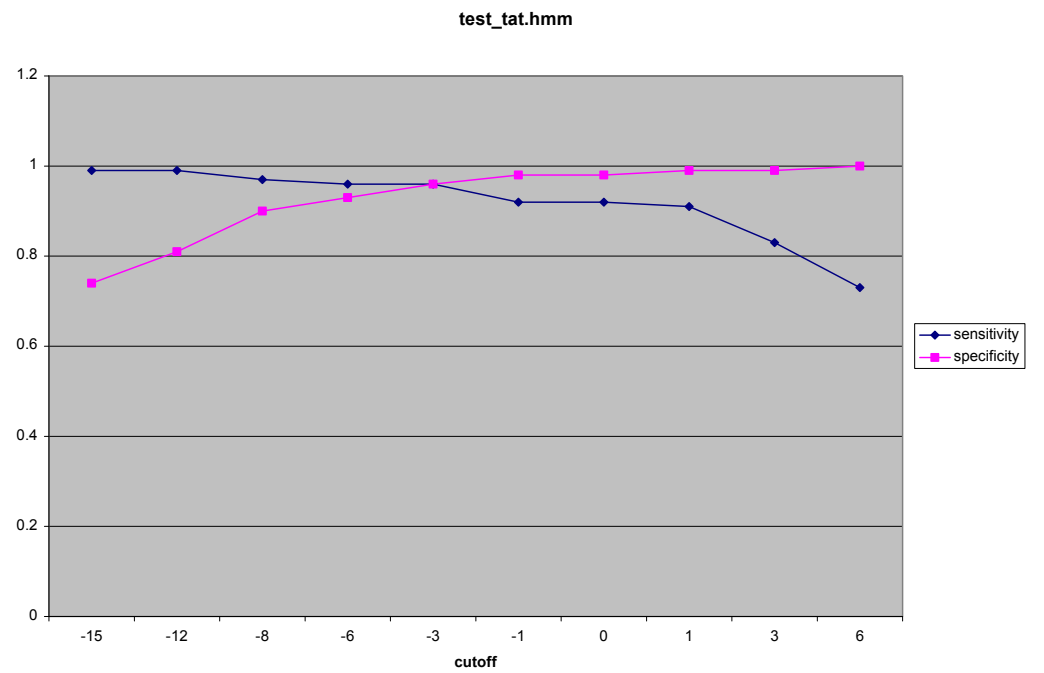
Στην συνέχεια παρουσιάζονται αναλυτικά τα αποτελέσματα από τις ευαισθησίες και τις ειδικότητες από τα δεδομένα δοκιμής, σε μορφή γραφικών παραστάσεων που δημιουργήθηκαν από τους πίνακας (A_2.5,A_2.6, A_2.7, A_2.8, παράρτημα Α2). Σύμφωνα με τις ενδείξεις της γραφικής 3.5 το ικανοποιητικό cutoff για την αναγνώριση Sec πεπτίδιο οδηγητή είναι ίσο με -1 με ευαισθησία και ειδικότητα περίπου 0.90 και mcc 0.77. Επίσης αφαιρόντας από το σύνολο τις ακολουθίες με Tat πεπτίδιο οδηγητή το ικανοποιητικό cutoff στη γραφική 3.6 είναι ισό με -3 με ευαισθησία και ειδικότητα 0.93 και mcc 0.84. Ενώ για την αναγνώριση του Tat πεπτιδίου οδηγητή σύμφωνα με το γράφημα 3.7 το ικανοποιητικό cutoff είναι ίσο με -3 με ευαισθησία και ειδικότητα 0.96 και mcc 0.75.βρίσκεται περίπου στο σημείο (-10, -3). Χωρίς τις ακολουθίες με το πεπτίδιο οδηγητή Sec το ικανοποιητικό cutoff είναι περίπου -10 γράφημα 3.8 με ευαισθησία και ειδικότητα 0.96 και mcc 0.75.



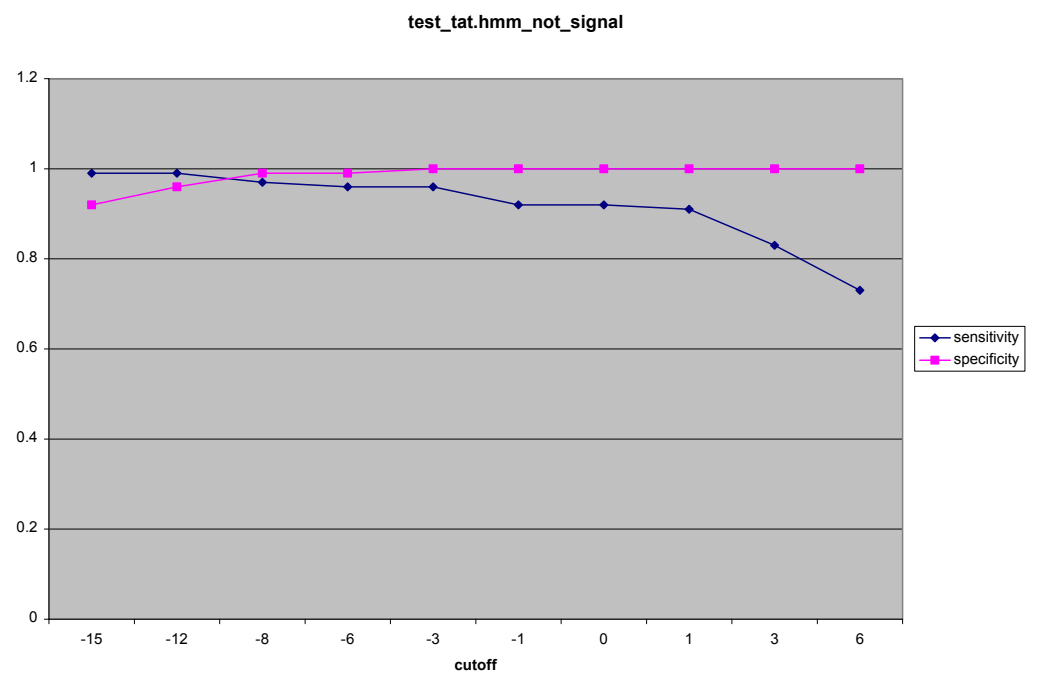
Γράφημα 3.5: Είναι η γραφική παράσταση που αντλούμε από τις ευαισθησίες και τις ειδικότητες, κάνοντας αναζήτηση το pHMM, SEC.hmm με τα δεδομένα δοκιμής. Δεδομένα φαίνονται στον πίνακα A_2.5 (Παράρτημα Α2)



Γράφημα 3.6: Είναι η γραφική παράσταση που αντλούμε από τις ευαισθησίες και τις ειδικότητες, κάνοντας αναζήτηση το pHMM, SEC.hmm με τα δεδομένα δοκιμής χωρίς να λάβει υπόψη του τις πρωτεΐνες με το Tat πεπτίδιο οδηγητή. Δεδομένα φαίνονται στον πίνακα A_2.6 (Παράρτημα Α2)



Γράφημα 3.7: Είναι η γραφική παράσταση που αντλούμε από τις ευαισθησίες και τις ειδικότητες, κάνοντας αναζήτηση το pHMM, TAT.hmm με τα δεδομένα δοκιμής. Δεδομένα φαίνονται στον πίνακα A_2.7 (Παράρτημα Α2)



Γράφημα 3.8: Είναι η γραφική παράσταση που αντλούμε από τις ευαισθησίες και τις ειδικότητες, κάνοντας αναζήτηση το pHMM, TAT.hmm με τα δεδομένα δοκιμής. Χωρίς να λάβει υπόψη του τις πρωτεΐνες με το Sec πεπτίδιο οδηγητή. Δεδομένα φαίνονται στον πίνακα A_2.8 (Παράρτημα Α2)

3.3. PRED-TAT $_{hmmer}$

Σύμφωνα με τα αποτελέσματα των παραπάνω γραφικών παραστάσεων βγάλαμε το συμπέρασμα ότι το ιδανικό cutoff είναι κοντά στο 0. Σε αυτή την ενότητα παρουσιάζονται αποτελέσματα που παίρνουμε κάνοντας χρήση των δύο profiles, Sec.hmm και Tat.hmm, με τις παρακάτω ακολουθίες. Τις ακολουθίες τις πήραμε από το σύνολο δεδομένων δοκιμής. Μερικές από αυτές ή και όλες χρησιμοποιήθηκαν και στο υποκεφάλαιο 1.6 και 1.7 για να δούμε αποτελέσματα άλλων εργαλείων. Αποτελέσματα που παίρνουμε από τα δύο pHMMs στον πίνακα 3.8, παρατηρούμε ότι πολύ σωστά προέβλεψε για όλες τις πρωτεΐνες.

>**Q02J64**
MRRHFLQRLGASAGLGAALLGLEFGSPRGQAAAADHWHMPDEHLPQERVFLAYAASPSIWKDLA
EDVN
>**B4EKR2**
MSNQDLPDQPNEPAASVSRRGFLKLAGVSGLAAAGGLAAARAAASNPDGPEQVHLWGNDPSEVV
I
>**A4QFQ3**
MGKHRRNNSNATRKAVAASAVALGATAAIASPAQAAEVVVPGTGISVDIAGIETTPGLNNVPGI
DQWIPSLSSQAAPTAYAAVIDAPAAEAQAAPAASTG
>**Q8RJN8**
MKVFFKITTLLLLILISYQSLAAFNNNPSSVGAYSSGTYRNLAQEMGKTNIQQKVNSTFDNMFGY
NNTQQLYYPYTENGVYKAHYIKAINPDEGDDIRTEG
>**P31224**
MPNFFIDRPIFAWVIAIIIMLAGGLAILKLPVAQYPTIAPPAVTISASYPGADAKTVQDTVTQV
IEQNMNGIDNLMYMSSNSDSTGTVQITLTFESGTDA
>**P0ABD5**
MSLNFLDFEQPIAELEAKIDSLTAVSRQDEKLDINIDEEVHRLREKSVELTRKIFADLGAWQIA
QLARHPQRPYTLDYVRLAFDEFDELAGDRAYADDKA

Πίνακας 3.8: Παρουσιάζονται κάποια συγκριτικά αποτελέσματα σε σχέση με το Sec και Tat profiles.

Πρωτεΐνη	Sec.hmm (score)	Tat.hmm (score)	Αποτέλεσμα
A4QFQ3	13.6	2.6	Sec
Q8RJN8	1.9	-18.3	Sec
Q02J64	1.4	7.5	Tat
P31224	-3.2	-17.4	Τίποτα απ'τα δύο
B4EKR2	-11.3	29.4	Tat
P0ABD5	-24.9	-36.1	Τίποτα απ'τα δύο

```
hmmsearch - search a sequence database with a profile HMM
```
HMMER 2.3.2 (Oct 2003) →πακέτο HMMER
```
Copyright (C) 1992-2003 HHMI/Washington University School of Medicine
Freely distributed under the GNU General Public License (GPL)
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```
HMM file: signal.hmm [Apla_signals_ALL] → όνομα αρχείου με τις στοιχισμένες ακολουθίες που πάρθηκε για την δημιουργία του hmmer.
```
Sequence database:          test.fasta
per-sequence score cutoff:  >= -100.0  → κατώτατο όριο αποκοπής -100
per-domain score cutoff:    [none]
per-sequence Eval cutoff:   <= 1e+002
per-domain Eval cutoff:     [none]
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query HMM:   Apla_signals_ALL
Accession:   [none]
Description: [none]
  [HMM has been calibrated; E-values are empirical estimates]
```
Ποσοστό που χαρακτηρίζει το βαθμό συσχέτισης με το profile
```
Scores for complete sequences (score includes all domains):
Sequence Description                          Score    E-value     N
-------- -----------                          -----    -------   ---
A4QFQ3                                        13.6     0.0005     1
Q8RJN8                                         1.9      0.056     1
Q02J64                                         1.4      0.064     1
P31224                                        -3.2       0.25     1
B4EKR2                                       -11.3        2.2     1
P0ABD5                                       -24.9          6     1
Parsed for domains:
Sequence Domain  seq-f seq-t    hmm-f hmm-t       score  E-value
-------- ------- ----- -----    ----- -----       -----  -------
A4QFQ3     1/1     1     35 [.     1    55 []        13.6   0.0005
Q8RJN8     1/1     1     22 [.     1    55 []         1.9    0.056
Q02J64     1/1     1     33 [.     1    55 []         1.4    0.064
P31224     1/1     1     35 [.     1    55 []        -3.2     0.25
B4EKR2     1/1    19     44 ..     1    55 []        -11.3     2.2
P0ABD5     1/1     1     17 [.     1     55 []       -24.9       6
```
Περιοχή αποκοπής της ακολουθίας
```
Alignments of top-scoring domains:
A4QFQ3: domain 1 of 1, from 1 to 35: score 13.6, E = 0.0005
             *->MkkkkkkkrrkrkkrrrrnkRlimafllllllllllalllaallaas
                M k +++ +   +k                  ++++++++++aa
A4QFQ3     1    MGKHRRNNSNATRK-----------------AVAASAVALGATAA- 28

                     gssssAlA<-*
                    +s+A+A
      A4QFQ3    29 -IASPAQA    35

Q8RJN8: domain 1 of 1, from 1 to 22: score 1.9, E = 0.056
             *->MkkkkkkkrrkrkkrrrrnkRlimafllllllllllalllaallaas
                Mk + k                               +++++l+++
Q8RJN8     1    MKVFFK------------------------------ITTLLLILI- 15

                   gssssAlA<-*
                    s +s +A
      Q8RJN8    16 -SYQSLAA    22

Q02J64: domain 1 of 1, from 1 to 33: score 1.4, E = 0.064
             *->MkkkkkkkrrkrkkrrrrnkRlimafllllllllllalllaallaa.
                M++    +                           l+a++ +++++++
Q02J64     1    MRRHFLQR-------------------------LGASAGLGAALLg 21
```

```
                  ...sgssssAlA<-*
                   + +++ ++A+A
      Q02J64    22 lefGSPRGQAAA    33

P31224: domain 1 of 1, from 1 to 35: score -3.2, E = 0.25
            *->MkkkkkkkrrkrkkrrrrnkRlimafllllllllllalllaallaas
               M ++   ++               ++++++++++++  +a l +
P31224     1    MPNFFIDRP----------------IFAWVIAIIIMLAGGLAILKL- 30

                  gssssAlA<-*
                     + A+
      P31224    31 ---PVAQY    35

B4EKR2: domain 1 of 1, from 19 to 44: score -11.3, E = 2.2
            *->MkkkkkkkrrkrkkrrrrnkRlimafllllllllllalllaallaas
                ++   k                       l+ + +++++++l+aa
B4EKR2    19    -RRGFLK----------------------LAGVSGLAAAGGLAAA- 40

                  gssssAlA<-*
                       A+A
      B4EKR2    41 ----RAAA    44

P0ABD5: domain 1 of 1, from 1 to 17: score -24.9, E = 6
            *->MkkkkkkkrrkrkkrrrrnkRlimafllllllllllalllaallaas
               M  +                                  +l +   +
P0ABD5     1    MSLN---------------------------------FLDFEQPI- 12

                  gssssAlA<-*
                     +   A
      P0ABD5    13 ---AELEA    17


Histogram of all scores:
score    obs    exp  (one = represents 1 sequences)
-----    ---    ---
<   2      5      -|=====


% Statistical details of theoretical EVD fit:
              mu =   -13.9218
          lambda =     0.2949
chi-sq statistic =     0.0000
  P(chi-square)  =          0

Total sequences searched: 6

Whole sequence top hits:
tophits_s report:
     Total hits:           6
     Satisfying E cutoff:  6
     Total memory:         17K

Domain top hits:
tophits_s report:
     Total hits:           6
     Satisfying E cutoff:  6
     Total memory:         19K
```

```
hmmsearch - search a sequence database with a profile HMM
HMMER 2.3.2 (Oct 2003)
Copyright (C) 1992-2003 HHMI/Washington University School of Medicine
Freely distributed under the GNU General Public License (GPL)
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```

HMM file: tat.hmm [Tat_signals_ALL] → όνομα αρχείου με τις στοιχισμένες ακολουθίες που πάρθηκε για την δημιουργία του hmmer

```
Sequence database:          test.fasta
per-sequence score cutoff:  >= -100.0
per-domain score cutoff:    [none]
per-sequence Eval cutoff:   <= 1e+002
per-domain Eval cutoff:     [none]
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query HMM:   Tat_signals_ALL
Accession:   [none]
Description: [none]
  [HMM has been calibrated; E-values are empirical estimates]

Scores for complete sequences (score includes all domains):
Sequence Description                        Score      E-value     N
-------- -----------                        -----     -------     ---
B4EKR2                                      29.4       8.3e-009     1
Q02J64                                       7.5        0.00068     1
A4QFQ3                                       2.6         0.0026     1
P31224                                     -17.4           0.55     1
Q8RJN8                                     -18.3            0.7     1
P0ABD5                                     -36.1              6     1

Parsed for domains:
Sequence Domain  seq-f seq-t  hmm-f hmm-t       score      E-value
-------- ------- ----- -----    ----- -----      -----      -------
B4EKR2       1/1     1    44 [.     1    61 []      29.4     8.3e-009
Q02J64       1/1     1    34 [.     1    61 []       7.5      0.00068
A4QFQ3       1/1     1    35 [.     1    61 []       2.6       0.0026
P31224       1/1     1    33 [.     1    61 []     -17.4         0.55
Q8RJN8       1/1     1    22 [.     1    61 []     -18.3          0.7
P0ABD5       1/1    27    67 ..     1    61 []     -36.1            6

Alignments of top-scoring domains:
B4EKR2: domain 1 of 1, from 1 to 44: score 29.4, E = 8.3e-009
             *->MskktksrlspseitpesvylsRRdFLksaaaaaaaaaaaaaaaaaa
                Ms+++ ++++ + + +     sRR FLk     +a+++++aaa+++a
B4EKR2     1    MSNQDLPDQPNEPAASV----SRRGFLK-----LAGVSGLAAAGGLA 38

             lpvgssgtpgpAaA<-*
             +         +AaA
B4EKR2    39 A--------ARAAA    44

Q02J64: domain 1 of 1, from 1 to 34: score 7.5, E = 0.00068
             *->MskktksrlspseitpesvylsRRdFLksaaaaaaaaaaaaaaaaaa
                M                    RR FL+       +a+a+++aa++
Q02J64     1    M---------------------RRHFLQ-----RLGASAGLGAALLG 21

             lpvgssgtpgpAaA<-*
             l   +s+  ++AaA
Q02J64    22 L-EFGSPRGQAAAA    34

A4QFQ3: domain 1 of 1, from 1 to 35: score 2.6, E = 0.0026
             *->MskktksrlspseitpesvylsRRdFLksaaaaaaaaaaaaaaaaaa
                M k +++ + +            R   +         a+a+a++a+a
A4QFQ3     1    MGKHRRNNSNAT-----------RKAVA---------ASAVALGATA 27
```

```
             lpvgssgtpgpAaA<-*
             +       ++pA+A
A4QFQ3    28 A------IASPAQA    35

P31224: domain 1 of 1, from 1 to 33: score -17.4, E = 0.55
             *->MskktksrlspseitpesvylsRRdFLksaaaaaaaaaaaaaaaaaa
                M+++               + R  F         a    +a+++a
P31224     1    MPNFF---------------IDRPIFAW-----VIAIIIMLAGGLAI 27

             lpvgssgtpgpAaA<-*
             l           A
P31224    28 L--------KLPVA    33

Q8RJN8: domain 1 of 1, from 1 to 22: score -18.3, E = 0.7
             *->MskktksrlspseitpesvylsRRdFLksaaaaaaaaaaaaaaaaaa
                M+                      F k          +++++ ++
Q8RJN8     1    MKV----------------------FFK---------ITTLLLILIS 16

            lpvgssgtpgpAaA<-*
                     ++ aA
Q8RJN8    17 Y--------QSLAA    22

P0ABD5: domain 1 of 1, from 27 to 67: score -36.1, E = 6
            *->Mskktksrlspseitpesvy....lsRRdFLksaaaaaaaaaaaaaa
                  ++k +   +e +++   ++ +l+R    k              a
P0ABD5    27    -RQDEKLDINIDEEVHRLREksveLTR----K------------IFA 56

             aaaalpvgssgtpgpAaA<-*
              ++a+       + +  A
P0ABD5    57 DLGAW-------QIAQLA    67


Histogram of all scores:
score    obs    exp  (one = represents 1 sequences)
-----    ---    ---
> -37      6      -|======
% Statistical details of theoretical EVD fit:
              mu =   -25.9520
          lambda =     0.2716
chi-sq statistic =     0.0000
  P(chi-square)  =          0

Total sequences searched: 6

Whole sequence top hits:
tophits_s report:
     Total hits:           6
     Satisfying E cutoff:  6
     Total memory:         17K

Domain top hits:
tophits_s report:
     Total hits:           6
     Satisfying E cutoff:  6
     Total memory:         19K
```

3.4. PRED-TAT

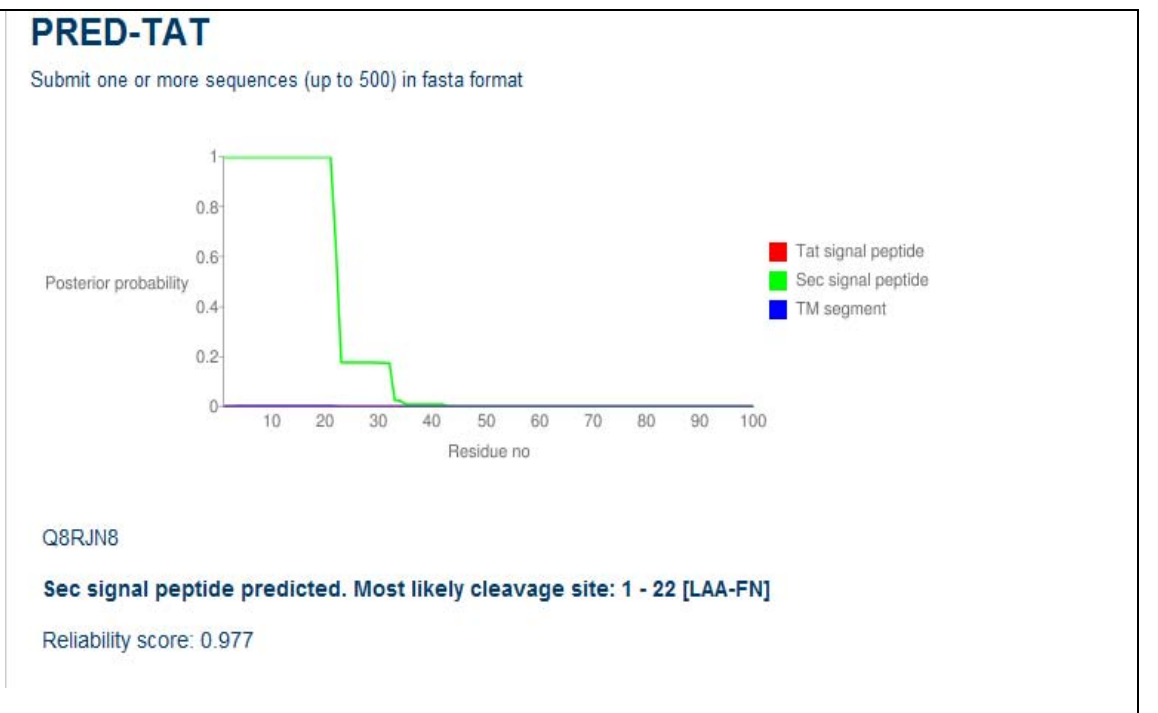
Παράλληλα με την μελέτη της πτυχιακής εργασίας αναπτύχθηκε ένα καινούργιο εργαλείο από την ερευνητική μας ομάδα. Λαμβάνοντας υπόψη τις στοιχίσεις για την αναγνώριση των συγκεκριμένων πεπτιδίων οδηγητών, δημιουργήθηκε ένα μοντέλο «χειροποίητο» (custom) HMM. Το εργαλείο αυτό είχε ως κύριο σκοπό την αναζήτηση Tat και Sec πεπτιδίων οδηγητών στο σύνολο των πρωτεϊνών. Το HMM του PRED-TAT, έχει μια δομή αρχιτεκτονικής αρκετά παρόμοια με αυτές που χρησιμοποιούνται σε άλλα εργαλεία όπως το PRED-LIPO, το CW-PRED και το PRED-SIGNAL. Αποτελείται από τέσσερα διαφορετικά υπομοντέλα,

- το υπομοντέλο Tat, που αντιστοιχεί στις ακολουθίες πρωτεϊνών που μεταφέρονται από τα Tat πεπτίδια οδηγητές,
- το υπομοντέλο πεπτιδίων οδηγητών Sec, που αντιστοιχεί στις ακολουθίες πρωτεϊνών που μεταφέρονται τα πεπτίδια οδηγητές Sec,
- το υπομοντέλο TM: που αντιστοιχεί στο αμινοτελικό των διαμεμβρανικών πρωτεϊνών,
- ένα υπομοντέλο για το αμινοτελικό πεπτίδιο οδηγητή των σφαιρικών πρωτεϊνών. Χρησιμοποιείται για να αναγνωρίσει τις σφαιρικές πρωτεΐνες των κυτταροπλασματικών και διαμεμβρανικών πρωτεϊνών.

Το PRED-TAT βρίσκεται στην ιστοσελίδα: http://www.compgen.org/tools/PRED-TAT/. Συμπεριλαμβανομένου και του άρθρου δημοσίευσης (Bagos, Nikolaou et al. 2010) της μελέτης και τον δεδομένων εκπαίδευσης και δοκιμής. Η μέθοδος μπορεί να εκτελεστεί με δύο τρόπους, ο πρώτος τρόπος είναι όταν ο χρήστης υποβάλει μια ακολουθία και λαμβάνει μια λεπτομερή έξοδο (γραφικές παραστάσεις, αξιοπιστία, κ.λπ.) (γράφημα 3.9), και ο δεύτερος τρόπος είναι όταν ο χρήστης μπορεί να υποβάλει μέχρι 500 ακολουθίες κάθε φορά και λαμβάνει τα συνοπτικά αποτελέσματα (τύπος πεπτιδίου οδηγητή, σημείο διάσπασης) σε εύκολα αναγνώσιμη μορφή (πίνακας 3.8). Στην συνέχεια θα παρουσιαστεί μια επίδειξη της συγκεκριμένης εφαρμογής. Κάνοντας χρήση του εργαλείου PRED-TAT με τις ακολουθίες που χρησιμοποιήσαμε και στο υποκεφάλαιο 3.3, λήφθηκαν τα ακόλουθα αποτελέσματα:

Με την χρήστη του εργαλείου για δυο ακολουθίες και πάνω δίνει πληροφορίες για τις ακολουθίες που καταγράφονται στον πίνακα 3.9:

Πίνακας 3.9: Παρουσία αποτελεσμάτων μετά την εισαγωγή δεδομένων στο πρόγραμμα PRED-TAT.
A4QFQ3 - **Sec signal peptide predicted. Most likely cleavage site: 1 - 35 [AQA-AE]** - Reliability score: 1.000
Q8RJN8 - **Sec signal peptide predicted. Most likely cleavage site: 1 - 22 [LAA-FN]** - Reliability score: 0.977
Q02J64 - **Sec signal peptide predicted. Most likely cleavage site: 1 - 33 [AAA-AD]** - Reliability score: 0.993
B4EKR2 - **Tat signal peptide predicted. Most likely cleavage site: 1 - 44 [AAA-SN]** - Reliability score: 0.988
P31224 - **TM segment predicted: 10 - 28** - Reliability score: 0.780
P24181 - **Sec signal peptide predicted. Most likely cleavage site: 1 - 54 [ADA-QT]** - Reliability score: 0.962



Γράφημα 3.9: Παρουσιάζεται το αποτέλεσμα που μας δίνει η μέθοδος PRED-TAT για την ακολουθία πρωτεΐνης Q8RJN8. Παρατηρούμε την αναγνώριση του Sec πεπτιδίου οδηγητή στην περιοχή από 1-22. Με βαθμό αξιοπιστίας 0.977.

4. Συμπεράσματα

Σκοπός της εργασίας αυτής είναι να αναπτυχθεί ένα καινούργιο Hidden Markov Model (HMM) για την αναγνώριση Tat και Sec πεπτιδίων οδηγητών. Στο μοντέλο αυτό λήφθηκαν υπόψη τα μοτίβα που στηρίχθηκαν στις στοιχίσεις Sec και Tat πεπτιδίου οδηγητή. Οι στοιχίσεις αυτές πάρθηκαν από δεδομένα που ήταν πειραματικά θετικά σχολιασμένα στη βάση δεδομένων, για την παρουσία του πεπτιδίου οδηγητή. Η στοίχιση που λάβαμε υπόψη μας για τις ακολουθίες με το Tat πεπτίδιο οδηγητή, χαρακτηρίζεται από τέσσερα διαφορετικά μέρη: (i) θετικά φορτισμένη n-περιοχή, (ii) ενδιάμεση περιοχή με το χαρακτηριστικό μοτίβο της δίδυμης αργινίνης RR, (iii) μια λιγότερο υδρόφοβη h-περιοχή που εκτείνεται στη μεμβράνη, και (iv) μια συνήθως μικρού μήκους πολική c-περιοχή με το χαρακτηριστικό μοτίβο πεπτίδιο αποκοπής AXA. Η στοίχιση για την αναγνώριση του Sec πεπτιδίου οδηγητή στις ακολουθίες χαρακτηρίζεται μόνο από τις (i), (iii) και (iv) περιοχές.

Με τις στοιχίσεις αυτές και με την βοήθεια του πακέτου HMMER 2.3.2 αναπτύχθηκαν δυο profiles hmms. Σύμφωνα με τα αποτελέσματα που πάρθηκαν χρησιμοποιώντας το Tat.hmm με τα δεδομένα, σε όλες τις περιπτώσεις παρουσιάστηκαν να ξεπερνούν το TatP και το TATFIND σχεδόν σε όλα τα μέτρα της ακρίβειας (ευαισθησία, ειδικότητας και MCC, και στην πρόβλεψη της περιοχής αποκοπής). Όσον αφορά τώρα την σύγκριση των pHMMs, του PRED-$TAT_{HMMER,}$ φαίνεται να ξεπερνά κατά πολύ στην πρόβλεψη του Tat πεπτιδίου οδηγητή, από τα profiles PFAM και TIGR. Όσον αφορά το pHMM, Sec.hmm, παρατηρήσαμε ότι και αυτό κάνει πάρα πολύ καλές προγνώσεις πεπτιδίων οδηγητών Sec, εντούτοις βλέπουμε να υστερεί κάπου και τα υπόλοιπα εργαλεία πρόγνωσης Sec πεπτιδίου να υπερέχουν στα αποτελέσματά τους. Τα profiles HMMs μπορούν να χρησιμοποιηθούν ως εναλλακτική λύση για τις αναλύσεις μεγάλης κλίμακας.

Ανεξαρτήτως αυτού πρέπει να πούμε ότι δεν λάβαμε υπόψη τις αλλαγές που πρόκειται να γίνουν στο μέλλον ή ακόμη και τις σπάνιες περιπτώσεις που υπάρχουν. Μια από τις σπάνιες περιπτώσεις που αναφέραμε είναι η αντικατάσταση της μιας από τις δύο αργινίνες στο μοτίβο RR, όπου μπορεί να αντικατασταθεί από την λυσίνη χωρίς να επηρεάσει βέβαια την TAT οδό (Hinsley, Stanley et al. 2001). Ενώ υπάρχουν επίσης παραλλαγές με την επέμβαση της ασπαραγίνης (Caspers, Brockmeier et al.) και το μοτίβο να τύχει μιας τέτοιας διαμόρφωσης RNR (Ignatova, Hornle et al. 2002). Λόγω της συγκεκριμένης εκπαίδευσης που υπέστη το HMM, όπου δεν λήφθηκαν υπόψη στο σύνολο εκπαίδευσης αυτές οι εξαιρέσεις, θα είχαμε αύξηση των ψευδώς θετικών.

Το HMM εκπαιδεύτηκε με τον συνδυασμό Gram-positive και Gram-negative βακτηριακών ακολουθιών. Ακόμα κι αν υπάρχουν μικρές διαφορές μεταξύ των περιοχών αποκοπής πεπτιδίων οδηγητών μεταξύ αυτών των κατηγοριών, το περιορισμένο πλήθος του συνόλου εκπαίδευσης μας ανάγκασε αν ακολουθήσουμε αυτή την επιλογή. Επίσης στο σύνολο μας υπάρχουν διάφορες ακολουθίες αρχαίων με θετική ένδειξη για την παρουσία Tat πεπτιδίου οδηγητή στην ακολουθία τους. Λόγω του ότι ο αριθμός λιποπρωτεϊνών που εξάγονται χρησιμοποιώντας την οδό Tat, είναι πολύ μικρός δεν τις λάβαμε υπόψη μας στη μελέτη (Gimenez, Dilks et al. 2007; Shruthi, Anand et al. 2010).

Παράρτημα Α

A1: Δεδομένα εκπαίδευσης

Πίνακας A_1.1: ID δεδομένων εκπαίδευσης Sec πεπτιδίων οδηγητών								
P16397	P15320	P02943	P0C0T5	P14738	P04635	P25447	P0C0J0	P15279
P17953	P0AD64	P09794	P06886	P30920	P0A3T3	P06971	P16454	P11439
Q00971	P06279	P32722	P27035	P15321	P28623	P31697	P12061	P20149
P23135	P14892	P04979	P31835	P21175	P18429	P0ABE7	P18956	P0A0Y3
P24735	P05818	P09545	P09489	P0AG78	P0ADA1	P18473	P14542	P24059
P0ABK9	P13430	P06546	P27195	P11701	P13429	P27951	P20910	Q02760
P31797	P0AG80	P0C2E9	P07110	P22629	P13720	P17543	P0AGC3	P0A1V8
P26501	P11220	P24305	P13470	P22751	P04960	P09333	P00083	P22865
P16216	P13794	P0A921	P33590	P13717	P23598	P04977	P05695	P24092
P13810	P30141	P07986	P0AFK9	P61153	P15704	P09331	P06608	P19424
P00099	P15452	P13734	P18477	P00775	P22391	P0AFH8	P13482	P0C1C0
P0AG82	P19843	P24474	P33781	Q02307	P06717	P15319	P33406	P17266
P11000	P09790	P16869	P35150	P17315	P13626	P07941	P0C2T2	P18336
P15917	P14005	P68588	P15930	P04957	P12616	P31830	Q01996	P09394
P00809	Q03011	P21543	P14191	P00131	P35804	P19571	P17137	P01077
P07254	P33665	P00648	P12608	P22364	P17855	P04845	P02971	P30692
P18958	P22940	P00446	P36924	P14283	P07103	P11797	P04816	P33673
P04127	P08704	P14488	P62605	P15488	P29822	P27032	P30705	P23827
P05825	P08506	P29957	P22340	P19369	P27755	P32823	P00694	P06202
P07111	P23549	P24040	Q70YI1	P13036	P06111	P24093	P26514	P14212
P11889	P10549	P13507	P05458	P26827	Q00499	P02930	P32520	P10520
P15922	P10477	P21982	P04377	P0A3R5	P0A917	P25394	P21948	P20861
P04981	P13650	P02910	Q01269	P19576	P21171	P31715	P07528	
P19926	Q01786	P20723	P06597	P20862	P14090	P05655	Q05156	

Πίνακας A_1.2: ID δεδομένων εκπαίδευσης Tat πεπτιδίων οδηγητών							
P76342	Q9PA38	Q9HYL2	P0AAK9	Q8PLY8	Q9PIC3	P31075	Q7NZY0
Q57366	P94127	P22222	P94170	P69741	Q9CK41	Q06650	P33225
P39185	Q59746	P44847	P21853	P18775	B1Y6A6	P26648	P55048
Q9RK81	Q8XT53	Q01578	P52320	Q8XV50	P0AAL5	P46923	P07822
Q9RI72	P81040	Q21AR4	Q44292	B3PWV0	Q8GR90	Q8GGJ7	Q888N2
P96465	Q8XUX6	Q8ECL7	P45004	P12676	B2FLK4	P37600	P15713
P17201	Q8GPG4	P07883	P35393	P46448	Q8XQB8	B4SRN1	P77554
P39597	P22641	P73452	Q93HX3	Q5LNE0	P82594	Q9F0W4	O87948
P63882	Q59517	B2UL75	Q8DLH9	Q2IGN2	P69739	Q5N0R0	Q92Z36
Q63Q46	P45015	Q8YBC6	P19573	A6LB54	B1W5J7	P05448	Q51705
P07984	P38043	Q01537	P81594	P31884	P50500	Q7VJT5	P38501
Q55460	Q8FX16	P13063	Q7M962	A4FN60	P07603	O34870	P36548
A6VQE0	C3MB06	Q50644	Q07982	P55047	C1D9G3	Q9S1H0	Q59543
Q59634	Q9A4T2	P22637	P12374	P0A5E1	P36649	O34213	Q9HVA4
P35392	P40120	P55669	Q5FQ05	Q60AK7	P77374	P0AAJ8	Q01S58
Q7MR39	Q44052	Q88DZ2	P17687	P63884	P71244	P81186	P95246
P14559	P06200	B2UQL7	P0A4R1	Q89BU4	Q7VKI8	P44652	P0AB24
Q8X6I9	Q53239	P10509	P55046	Q89XJ6	P44798	Q44018	Q8XAS4
Q06530							

Πίνακας A_1.3: ID δεδομένων εκπαίδευσης TM						
P02980	P15030	P80107	P19934	P16336_1	P43457_1	P0AEL0
P08005	P0AEX3	P02948	P02983_11	P16336_3	P43457_2	Q02761
P08400	P23889	P20672	P02983_13	P16336_5	A5U127_1	P0AEY8
P0AG96	P0AGA2	P10955	P02983_9	P16336_7	P0A0J9_1	P31602
P15877	P0AD68	Q45247	P02983_7	P16336_9	P0A3G5	P0AEC8
P09348	P0AGC0	P95673	P02983_3	P08064_1	P16655	Q51575
P0AAD2	P0ABN1	P0A8Q3	P02983_5	P08064_3	P12287_2	P02942
P06009	P0ABJ9	P0A8Q0	P02983_1	P08064_5	P12287_4	P52002
P0AEA5	P0A6M2	P77921	P0A0J9_13	P0A334_1	P14319_1	P07117
P0ABV2	P11551	P0C0Y8	P0A0J9_11	Q10762_1	P14319_3	P02981
P31706	P02920	P97253	P0A0J9_9	Q10762_3	P14319_2	P0C2U2_1
P26789	P08006	P17413	P0A0J9_7	P67643_1	P0A5K8_1	P0C2U2_3
P68183	P23516	P0C0Y1	P0A0J9_5	P67643_3	Q99T13_1	P0C2U2_4
P0AF06	P05701	P25737	P0A0J9_3	P53380_1	Q99T13_2	P0C2U2_6
P0ABV6	P22519	P0ABJ6	P02916	P69786	Q99T13_3	P0C2U2_8
P07654	P0AEN4	P0C2U2_10	P69801	P12691	P02950	P0AB93
P0ACG4	P00845_2	P24207	P0ABJ3	P50600	P35106	P09130
P0AGI1	P02921	P0AG90	P0AA82	P23462	P35099	P0AE82
P0A742	P26790	P0AAI3	P00804	P0AG93	P0ABK2	P08194
P06030	P19568	P08305	P98002	P25714	P0AEJ4	P31710

Πίνακας A_1.4: ID δεδομένων εκπαίδευσης CYTO									
P63490	P31101	P22320	P29930	P18134	P23238	P07395	P07859	P75504	P75247
P16250	P23536	P22318	P24251	P0A705	P0AFJ5	P07395	P56206	Q11066	P50866
P47477	P69791	P22319	P0A9X9	P05827	P0A9K1	P27002	P36238	P47486	P26209
P46321	P69796	P24404	P06614	P24323	P22608	P00960	P13537	P75437	P39751
P37887	P17127	P13266	P29847	P0A715	P0AFK0	P00961	P09961	O34942	P39211
Q59961	P23537	P27828	P27369	P30125	P05055	P60906	P14898	P46394	P22818
P75559	P69797	Q02047	P0A9D8	Q00412	P17856	P00956	P14899	P32724	P53558
P39486	P16114	P09373	P27236	P61495	P0A249	P0A8N5	P0C1U9	O34399	P63642
P46906	P27000	P09403	P0A1F0	P36774	P0A283	P07813	P13419	P19064	Q59111
P78021	P35636	P14697	P07862	P0A9M0	P05706	P00959	P78032	P39125	Q59112
P22317	P23395	P23608	P14776	P31858	P16481	P16659	P47587	P39630	P0A006
Q04944	P34945	P23869	P13551	P03030	P37080	P00962	O34526	Q04718	P35146
P07800	Q01551	P09063	P0A6P1	P15977	P37081	P11875	P47634	P0C0G0	P78030
Q06774	P33788	P17052	P33171	P21517	P09378	P0A8L1	P78011	P94550	Q46171
P33393	Q44444	P28629	P33675	P08997	P0A944	P0A8M3	P38022	P71153	P75289
P52974	P29442	P00509	P28181	P00935	P0A948	P07118	P52157	P53526	Q53634
P0A9N4	P26997	P0A9G6	P13039	P18949	P13857	P00954	P37870	P80879	P63498
P37571	P27888	Q00594	P0A6T1	P0A2Q4	P30850	P0AGJ9	P66773	P09122	Q01360
P27001	P0A962	P13016	P28718	P29901	P0A7Y0	P0C6D6	P66785	P27903	P47345
P47277	P0A9J0	P26613	P11886	P24519	P09155	Q01911	O87085	O34623	P13243
P0A277	P17725	P69503	P0A9F6	P10183	P21513	P07464	P35901	P75080	P15339
Q05428	P31570	P26474	P25821	P0A9G2	P27217	P75948	P47299	O34996	O30931
O31139	P30327	P26475	P28605	P0A9I8	P13394	P07097	P12880	P63962	P54569
P54374	P28894	P0A1Y2	P13035	P04674	P23721	P04693	P00467	Q59549	Q02141
P47597	P34917	P0A9H7	P24531	P04338	P0A9K9	P26602	P10021	O67987	P33744
P54965	P34918	P09384	P34895	P02963	P0A9E2	P10908	P04766	Q45421	P17889
P35489	Q02154	P04042	P23189	P19844	P00957	P18317	P78024	Q49700	P53002
Q44297	P15214	P06110	P33771	P0AA19	P21888	P0AED0	P30338	P63873	P47707
Q10401	P33770	P0A2D5	P23619	P08308	P36419	P13041	P42103	P47507	

Πίνακας Α_1.5: Παρουσιάζονταιι τα αποτελέσματα από τον έλεγχο των δεδομένων εκπαίδευσης με το profile SEC.hmm

cutoff	TP[1]	TN[2]	FP[3]	FN[4]	Sensitivity	Specificity	mcc
-15	326	236	342	2	0.99	0.41	0.44
-12	326	291	287	2	0.99	0.5	0.51
-8	325	364	214	3	0.99	0.63	0.61
-6	324	396	182	4	0.99	0.68	0.65
-3	319	442	136	9	0.97	0.76	0.71
-1	316	473	105	12	0.96	0.82	0.75
0	312	480	98	16	0.95	0.83	0.75
1	305	487	91	23	0.93	0.84	0.75
3	287	511	67	41	0.88	0.88	0.75
6	221	546	32	107	0.67	0.94	0.66

[1]: αντιπροσωπεύει τα αληθώς θετικά, δηλαδή το σύνολο των πρωτεϊνών με Sec που βρίσκονται πάνω από το συγκεκριμένο cutoff.
[2]: αντιπροσωπεύει τα αληθώς αρνητικά, δηλαδή το σύνολο πρωτεϊνών με TAT,TM,CYTO ,που βρίσκονται κάτω από το συγκεκριμένο cutoff.
[3]: αντιπροσωπεύει τα ψευδώς θετικά, δηλαδή το σύνολο πρωτεϊνών με TAT,TM,CYTO , που βρίσκονται πάνω από το συγκεκριμένο cutoff.
[4]: αντιπροσωπεύει τα ψευδώς αρνητικά , δηλαδή το σύνολο των πρωτεϊνών με Sec που βρίσκονται κάτω από το συγκεκριμένο cutoff.

Πίνακας Α_1.6: Παρουσιάζονται τα αποτελέσματα από τον έλεγχο των δεδομένων εκπαίδευσης με το profile SEC.hmm, χωρίς να λάβουμε υπόψη στις μετρήσεις τις πρωτεΐνες με το TAT πεπτίδιο οδηγητών

cutoff	TP[1]	TN[2]	FP[3]	FN[4]	Sensitivity	Specificity	mcc
-15	326	234	194	2	0.99	0.55	0.58
-12	326	286	142	2	0.99	0.67	0.68
-8	325	348	80	3	0.99	0.81	0.8
-6	324	375	53	4	0.99	0.88	0.86
-3	319	399	29	9	0.97	0.93	0.9
-1	316	413	15	12	0.96	0.96	0.93
0	312	415	13	16	0.95	0.97	0.92
1	305	418	10	23	0.93	0.98	0.91
3	287	424	4	41	0.88	0.99	0.88
6	221	428	0	107	0.67	1	0.73

[1]: αντιπροσωπεύει τα αληθώς θετικά, δηλαδή το σύνολο των πρωτεϊνών με Sec που βρίσκονται πάνω από το συγκεκριμένο cutoff.
[2]: αντιπροσωπεύει τα αληθώς αρνητικά, δηλαδή το σύνολο πρωτεϊνών TM,CYTO ,που βρίσκονται κάτω από το συγκεκριμένο cutoff.
[3]: αντιπροσωπεύει τα ψευδώς θετικά, δηλαδή το σύνολο πρωτεϊνών TM,CYTO , που βρίσκονται πάνω από το συγκεκριμένο cutoff.
[4]: αντιπροσωπεύει τα ψευδώς αρνητικά , δηλαδή το σύνολο των πρωτεϊνών με Sec που βρίσκονται κάτω από το συγκεκριμένο cutoff.

Πίνακας Α_1.7: Παρουσιάζονται τα αποτελέσματα από τον έλεγχο των δεδομένων εκπαίδευσης με το profile TAT.hmm.

cutoff	TP[1]	TN[2]	FP[3]	FN[4]	Sensitivity	Specificity	mcc
-15	150	505	251	0	1	0.67	0.5
-12	150	575	181	0	1	0.76	0.59
-8	150	654	102	0	1	0.87	0.72
-6	150	689	67	0	1	0.91	0.79
-3	150	714	42	0	1	0.94	0.86
-1	150	730	26	0	1	0.97	0.91
0	150	739	17	0	1	0.98	0.94
1	150	745	11	0	1	0.99	0.96
3	150	748	8	0	1	0.99	0.97
6	150	751	5	0	1	0.99	0.98
12	146	755	1	4	0.97	1	0.98
13	140	756	0	10	0.93	1	0.96
17	132	756	0	18	0.88	1	0.93
23	90	756	0	60	0.6	1	0.75

[1]: αντιπροσωπεύει τα αληθώς θετικά, δηλαδή το σύνολο των πρωτεϊνών με Tat που βρίσκονται πάνω από το συγκεκριμένο cutoff.
[2]: αντιπροσωπεύει τα αληθώς αρνητικά, δηλαδή το σύνολο πρωτεϊνών SEC,TM,CYTO ,που βρίσκονται κάτω από το συγκεκριμένο cutoff.
[3]: αντιπροσωπεύει τα ψευδώς θετικά, δηλαδή το σύνολο πρωτεϊνών SEC,TM,CYTO , που βρίσκονται πάνω από το συγκεκριμένο cutoff.
[4]: αντιπροσωπεύει τα ψευδώς αρνητικά , δηλαδή το σύνολο των πρωτεϊνών με Tat που βρίσκονται κάτω από το συγκεκριμένο cutoff.

Πίνακας Α_1.8: Παρουσιάζονται τα αποτελέσματα από τον έλεγχο των δεδομένων εκπαίδευσης με το profile TAT.hmm,χωρίς να λάβουμε υπόψη τα δεδομένα από τις πρωτεΐνες με το Sec πεπτίδιο οδηγητή.

cutoff	TP[1]	TN[2]	FP[3]	FN[4]	Sensitivity	Specificity	mcc
-15	150	403	25	0	1	0.94	0.9
-12	150	419	9	0	1	0.98	0.96
-8	150	424	4	0	1	0.99	0.98
-6	150	427	1	0	1	1	1
-3	150	427	1	0	1	1	1
-1	150	427	1	0	1	1	1
0	150	427	1	0	1	1	1
1	150	427	1	0	1	1	1
3	150	427	1	0	1	1	1
6	150	428	0	0	1	1	1
12	146	428	0	4	0.97	1	0.98
13	140	428	0	10	0.93	1	0.95
17	132	428	0	18	0.88	1	0.92
23	90	428	0	60	0.6	1	0.73

[1]: αντιπροσωπεύει τα αληθώς θετικά, δηλαδή το σύνολο των πρωτεϊνών με Tat που βρίσκονται πάνω από το συγκεκριμένο cutoff.
[2]: αντιπροσωπεύει τα αληθώς αρνητικά, δηλαδή το σύνολο πρωτεϊνών TM,CYTO ,που βρίσκονται κάτω από το συγκεκριμένο cutoff.
[3]: αντιπροσωπεύει τα ψευδώς θετικά, δηλαδή το σύνολο πρωτεϊνών TM,CYTO , που βρίσκονται πάνω από το συγκεκριμένο cutoff.
[4]: αντιπροσωπεύει τα ψευδώς αρνητικά , δηλαδή το σύνολο των πρωτεϊνών με Tat που βρίσκονται κάτω από το συγκεκριμένο cutoff.

Α2 Δεδομένα δοκιμής

Πίνακας Α_2.1: ID δεδομένων δοκιμής Sec πεπτιδίων οδηγητών								
A4QFQ3	A2TJI4	Q06529	P0AEU0	P77747	Q46444	P0AAX8	P04190	Q93MW7
Q8RJN8	P96156	P16700	Q99289	Q04064	Q8FDW4	P0A8X2	P15926	P54421
P44569	P42517	P37735	O50319	P33136	Q8VSL2	P33219	Q06110	O07532
P84888	P08331	Q8GPG1	P45354	Q0P9X8	Q9S1G7	P76520	P00733	O33599
P39180	P0A1E7	P38540	Q05918	Q6CZT4	O52057	P0ADU5	P43379	P16947
Q9L7P2	P0ABW7	P0AEG4	P0AD59	O68900	O52179	P0ADV1	P30921	P02977
Q59478	P09787	P0AEG6	P40601	P28581	O52055	P37648	P09121	P08089
P22365	Q55210	P36655	Q05489	P50057	P81238	P76172	P17692	Q53599
P46883	P0A387	P39176	P61316	P80672	Q9Z4N3	P76344	P0C2S1	P0A5Q2
P49250	Q55013	Q9EZE7	P54300	Q51485	P35823	P77269	Q06851	P0A618
Q59544	P07625	Q7BSW5	P29399	P37726	Q8KIL1	P39187	Q8GFE2	P38577
P30859	Q5SME3	O32591	Q9Z4N4	P35483	Q82S91	P39325	P40136	P39046
P30860	P74917	Q07792	P0A908	P35482	P15453	P39172	P08750	P11657
Q47898	P00120	Q9Z4J7	P37329	P10858	P0AGD1	P0AAA9	P39844	P39116
P36560	P00091	P35755	Q46203	P0C6R1	P77754	Q8GMV7	P39042	P96740
P19567	P94690	Q05202	P77348	P0C1A4	Q55025	Q9RB42	P39045	P19406
P80401	P94691	P37921	Q9Z4I3	P31519	Q7BQ98	P82593	P0C6P1	Q81HT3
P12334	P00132	P08191	P39186	P0C6E9	P02906	Q06316	P34071	P68802
P0A321	O33731	P45523	Q9I596	P35077	P0C918	P0C1D6	P54583	P49051
P81549	P43302	P0AEM9	P46739	P08038	P0ABZ6	P12690	P22541	P38538
P48982	P00106	P42512	Q06006	P0AFL5	Q06987	P00644	Q02934	P38059
P52682	P07497	Q9Z4P0	P80604	Q2RVM4	Q9XD84	P38939	P0C2S5	P0A3V1
P28585	P46445	Q07WU7	P38006	Q05433	P0A855	Q7M1T6	P50401	Q8NKX2
P23954	P81894	P07662	Q9S3R8	P30899	P96116	P00692	P80172	P50470
Q44642	Q9RQB9	O34214	Q9S3R9	Q06304	Q48245	Q05884	P0A074	P35706
P37028	P00154	P37902	P80603	P31091	P37387	Q50906	P60158	P01006
P04369	P00149	P37696	P38368	Q9X7I4	P44544	P01551	Q9LAB5	P0C1S6
P00150	P00144	O88093	P38399	P31083	P44276	Q44856	P83674	P00779
P55929	P00148	P48823	P76045	Q8DJE2	Q57449	P10424	P81715	Q9RU24
P14532	P95522	P00260	P29739	P23857	P0A940	P10425	P37957	P54423
P83513	P40943	P26515						

Πίνακας Α_2.2: ID δεδομένων δοκιμής Tat πεπτιδίων οδηγητών				
Q02J64	Q88AA1	P42251	Q8CJI8	Q3L8N0
B4EKR2	Q888K8	O07242	O88050	B9W4V4
Q7M827	P37049	O05816	Q9X9W6	A0QVQ7
Q5ZS14	B9DJC1	P71829	Q8CK08	HAH4_HALME
Q52522	Q8GJ31	A4Q9Z7	Q9RDG9	HMEA_ARCFU
Q87YZ1	P81453	A4QA36	Q9F2J1	Q60224_9EURY
Q87TY9	P0C5C1	A4QCC6	Q9L0J3	Q97V37_SULSO
Q9HYB6	Q3ZAB8	A4QCH6	O69840	HLY_HAL17
Q6W8A3	P95028	A4QDY7	Q9Z5A4	C3108_HALVO
Q8KU06	Q9EX55	A4QFX3	Q93S08	C3156_HALVO
Q7B469	Q9RJC7	A4QI82	Q9ZBF6	C3082_HALVO
A9CH43	Q9ZBW1	Q9EWT5	Q8CJM4	C2996_HALVO
A3ZF85	Q9RD58	Q9RJZ8	O87849	Q2ME8_HALSA
Q5DYT6	Q9Z510	Q9RJX4	Q9F3Q6	Q4A3E0_HALHI
P26290	Q9RJ88	Q9RKZ0	Q9EWQ0	Q6JSL9_HALAS

Πίνακας A_2.3: ID δεδομένων δοκιμής TM								
P31224	P23895	P13514	O06873	P50598	Q59835	P37019	Q56992	P41036
P24181	P41406	P18814	P31135	P50599	Q04943	Q8ZRP8	P16429	P11350
P40812	P33692	P05528	P40862	P22708	P94633	P32482	P77858	P33591
Q03024	Q02729	P07017	P23596	P22709	P31141	P08369	P15701	Q59647
P18275	P21409	P02941	O32521	Q05807	P16645	P30000	Q9XB52	P13156
P15993	P15030	P07018	O52788	P18006	P30345	Q09049	P29823	Q47421
P03959	P23876	Q02755	P10502	P23849	P0C0N4	P16701	P29824	P16256
P03960	P23877	P21823	P16552	P10905	Q04442	P31601	P73157	P42194
Q08120	P40729	O07366	P19057	P10906	Q04733	P10047	P31712	Q04804
P19072	Q56887	Q02581	P51760	P27668	P30145	P37739	P22610	P18200
P21627	P20923	P32166	P10717	P95730	P39843	P51055	P22729	P37433
P21628	O32617	P23200	P26280	P45562	O07380	P41086	P30296	P12681
P25185	P37147	Q55282	Q01854	P77173	P35865	P18777	P23930	P33639
P71345	P41083	Q53174	P26406	Q5SJ80	P39755	Q03203	P26381	P27611
P06609	P21345	P55891	P27125	P0ABN5	P26235	Q6G9L1	P0A4N3	P65566
P31801	P15643	Q9KQW9	P27135	P0AEK7	P39823	Q99027	O06754	P43454
P16482	P52094	P24205	Q07396	P39694	P12667	P22821	P16449	P45706
P21608	O31652	Q04664	P0C0H1	P28008	O06493	P24400	P46912	P34956
P46832	P97046	P0A4I5	A5U7G8	P0A4N1	P39886	O52733	P49022	P23648
P23173	P25396	P02982	P76350	P28539	Q9PJN1	Q05207		
Q8XB33	P51563	P33951	P96169	Q9Z7S5	P45130	A5U8T5		
P43440	P23054	P46913	P46921	P54146	Q04452	P28785		

Πίνακας A_2.4: ID δεδομένων δοκιμής CYTO								
P0ABD5	P21244	P07862	O06644	P44521	Q89UH1	O06925	P68739	P0A9K1
P0A9G6	P0A9H7	P0A6K6	P28181	P15214	Q9ZM46	P23996	P0A9G2	P0A9K7
P11071	P0A6G1	P0A6L0	P13024	P33012	Q3APW0	P07623	P0A9I8	P22608
P44406	P22034	Q59516	P13039	P31658	Q9ZMV2	P00935	Q51480	P0A7A5
P0A2W5	P09384	P14776	P0A8S9	P31101	P33547	P44527	Q07739	O32449
P72324	Q2KV65	P50970	Q46604	P36553	P0A2U4	P0A8U6	P24150	P54737
P28629	Q92YM4	P03004	Q56415	P57777	Q6XVY3	Q1R0L1	Q46877	P0AFK0
P23872	P0A964	P08622	P43500	P23871	P51064	P0A9F9	P0A9N8	P44584
P38448	P0AE67	P10443	P78055	P77915	P22939	P45131	P0C0L2	P80354
Q00594	P96126	P0A988	Q46106	P60757	P0A796	Q1DB04	P11724	P22259
P30013	Q51434	P43313	P69922	Q9S5G3	Q9PHM8	P0A731	P54893	P23869
P15034	P75726	P45573	P28894	Q9X0C8	Q9X1I8	O86956	P0A790	P0A9L5
P24734	P69330	Q72VM5	P0A9A9	P60664	P0A6I0	P32173	P04744	P56601
P26612	P0A9I1	Q822X6	O83816	P06987	P0A715	P0A9G8	P42193	P23916
Q8PJY6	O87444	Q0SNU8	P52983	P23619	P60546	P84308	P32427	P0A8T1
P38434	P0A6G9	Q3YSG0	P28718	P33393	P0A717	P30621	P42673	P07004
P0A6D0	P0A6I6	Q1ID35	P11886	P39662	P76008	P60390	P0A794	P0A7B5
P07639	P29930	Q057T0	P0A9I3	P22317	P16115	P55818	P42195	P95435
P00963	P24251	Q2GCI5	P0A6V8	P22320	P54354	P14900	Q05526	P23536
P50286	Q9X0E6	A1WHU3	P17169	P22318	P0A8P1	P11880	Q9RF52	P69811
P26474	Q59685	P04994	P04772	P22319	P60716	P06722	P36936	P69819
P26475	P27111	P0A8G9	P36205	P37171	P36774	P11458	P26311	P69783
P0A1Y2	P0A9F3	P33694	O83004	P0A6Y5	P32099	P10902	P0A9N4	P23537
Q9X758	P17854	P33700	P0A6V5	P0A6H5	P31858	P75949	P09373	P0A7D1
P12995	P27369	P0A6Q3	P0A9C9	P0A7B8	P32199	P10183	P57525	P69791
P13000	P0A6L2	P0A953	P0AC69	P25080	P43341	P44539	P50203	P69795
Q8GHL1	P04036	P0A6R0	O83349	P0C058	P15977	O33732	P50176	P37081

P80193	P0A9D8	P0A6Q6	P63224	P16100	P08997	P94212	Q02286	P30335
P31572	P0A6K1	Q9FA38	P48638	P80046	P37330	P74494	P43336	Q8PGR7
P42321	P24171	P32055	P82998	Q46822	O85014	P0A0Z7	P0AFJ5	P08178
P07913	P16525	Q05335	Q97II1	P75393	O86963	P54383	P75361	P45618
Q01911	P43902	P12994	P29422	P13267	P50434	P80019	Q53596	O50515
Q46M57	P44694	P76407	P0A5P0	P39043	Q9AGY7	O33007	Q8E0C9	Q9WXK8
P07464	P44420	P61417	Q60023	Q7A338	Q60151	Q9PQP0	Q10744	Q6YP15
P30139	P26602	Q7UP77	Q60024	P29541	P32396	P00553	Q02VB0	P54322
P77718	P12758	Q7URP3	Q8EQW0	P64034	P71756	O32765	P55179	P56968
P45369	P10908	Q7UX42	P39802	Q9PQJ3	P75119	P95313	P22346	Q9AG29
P0AGG4	P12295	P53627	P19080	P61337	P0A5T0	O31147	O68575	Q8NS78
P52197	Q9ZMZ5	P41554	O53079	Q48QY5	P0A672	P42450	Q5HJF4	Q8DR09
P36662	P18317	P75245	O53077	Q83MV5	Q9KWG2	Q53062	Q5XA18	P81101
Q9APM5	P0A8F4	Q9ZN78	O53078	Q93PU8	Q1IX94	A2RM21	P50849	Q7TWW7
O86262	P0AED0	P0A5N2	P39127	P37454	Q9RSK1	Q83GE4	P37487	Q9F0R1
P10026	P0A8F8	P50846	Q9CJ45	P0A574	P38037	Q8Y960	Q9AGJ6	Q2G1N7
P28904	P0A8G0	Q59118	P0A512	P0A0A5	Q5M5V5	P23342	Q59568	P68575
P62601	P24417	Q11010	P77985	P07343	P47438	Q9S4K9	P77949	P81100
P25745	P27828	Q741P3	P56220	P0C0G6	P65145	P0A5L8	P0A5S6	P80734
P39440	P44818	Q59280	P47722	P0A4E2	P39804	P50589	P32397	Q46337
Q89AJ2	P43859	P17893	Q08352	Q97R46	O05724	P0A4V0	Q9Z4P6	P40875
P21762	Q84H41	P54264	P39071	P19432	Q2FZE2	O34777	Q9Z4Q7	Q46336
P0A890	P26997	P38645	P06632	Q8CSR8	P11018	O34767	P45595	Q46338
O34559	P0A9K9	Q5NR44	Q6A8J7	Q72GE2	P13857	Q83GS9	P59573	Q9ABY6
P0A7D4	P52934	Q2W3D5	Q67LN9	Q0I902	P30850	Q9PQ32	Q986B5	Q1MBK9
P31473	P06535	Q7VRU3	Q10765	P96142	P51837	Q5FIY7	Q9WY60	O83650
P0A9J6	Q08788	Q01WJ8	Q9X895	P67586	P44442	Q601U6	Q9Z6W0	Q07UN5
P69506	Q9L524	Q9Z6R6	Q8P1X6	Q8Y0A1	P45175	Q9CC16	O83678	Q9ZKG9
P0A7G6	Q9EYW6	O83938	Q8RCT6	P0AGJ9	P16114	Q6MT18	Q5LAE0	Q72MR4
P59456	P75368	Q5LC76	P59076	P41256	P0A805	Q8EUJ9	P60919	Q5LTL1
P0C348	Q83GD1	P59505	Q50319	P76145	P46849	O87386	P75563	O83618
Q7U3W6	Q83GH3	Q5L6W7	P67577	P27859	P42454	P0A9E2	Q4A6A2	Q6N5P6
Q5N192	Q8G7G5	Q72MG8	Q9L4Q8	Q01551	P17052	Q59967	Q6YPX7	Q5SM28
P32169	Q9PR21	Q7VBX6	Q38VQ9	P0C6D6	P35636	P27306	Q9PQ33	Q8Y199
P09378	Q67LP0	Q68WW1	O83647	Q7NC67	P44853	P77499	Q8DP65	Q7VQX7
P0A944	P0A636	Q30S83	Q8D1Y2	Q9WX29	P0A821	Q9EXP1	Q8DP66	O51540
P0A948	Q9PPP0	O83059	P56690	Q83GE9	P43927	Q9EXP2	Q9PPZ7	Q14IZ0
P47372	P39795	P47469	P81102	P75225	P23721	P61709	Q6HG86	Q7VI99
P67598	O32271	Q9RC92	Q06539	Q83GK5	O85300	Q2JKL2	Q4AA96	Q2GEZ6
P45362	P19368	Q56559	Q93PS3	P77832	P24299	P37552		

Πίνακας Α_2.5: Παρουσιάζονται τα αποτελέσματα από τον έλεγχο των δεδομένων δοκιμής με το profile SEC.hmm

cutoff	TP[1]	TN[2]	FP[3]	FN[4]	Sensitivity	Specificity	mcc
-15	269	402	466	4	0.99	0.46	0.4
-12	267	504	364	6	0.98	0.58	0.48
-8	264	636	232	9	0.97	0.73	0.6
-6	260	684	184	13	0.95	0.79	0.65
-3	255	759	109	18	0.93	0.87	0.74
-1	247	791	77	26	0.9	0.91	0.77
0	238	801	67	35	0.87	0.92	0.77
1	230	813	55	43	0.84	0.94	0.77
3	196	829	39	77	0.72	0.96	0.71
6	137	845	23	136	0.5	0.97	0.58

[1]: αντιπροσωπεύει τα αληθώς θετικά, δηλαδή το σύνολο των πρωτεϊνών με Sec που βρίσκονται πάνω από το συγκεκριμένο cutoff.
[2]: αντιπροσωπεύει τα αληθώς αρνητικά, δηλαδή το σύνολο πρωτεϊνών με TAT,TM,CYTO ,που βρίσκονται κάτω από το συγκεκριμένο cutoff.
[3]: αντιπροσωπεύει τα ψευδώς θετικά, δηλαδή το σύνολο πρωτεϊνών με TAT,TM,CYTO , που βρίσκονται πάνω από το συγκεκριμένο cutoff.
[4]: αντιπροσωπεύει τα ψευδώς αρνητικά , δηλαδή το σύνολο των πρωτεϊνών με Sec που βρίσκονται κάτω από το συγκεκριμένο cutoff.

Πίνακας Α_2.6 : Παρουσιάζονται τα αποτελέσματα από τον έλεγχο των δεδομένων δοκιμής με το profile SEC.hmm, χωρίς να λάβουμε υπόψη στις μετρήσεις τις πρωτεΐνες με το TAT πεπτίδιο οδηγητή

cutoff	TP[1]	TN[2]	FP[3]	FN[4]	Sensitivity	Specificity	mcc
-15	269	394	399	4	0.99	0.5	0.44
-12	267	502	291	6	0.98	0.63	0.53
-8	264	628	165	9	0.97	0.79	0.68
-6	260	672	121	13	0.95	0.85	0.73
-3	255	740	53	18	0.93	0.93	0.84
-1	247	765	28	26	0.9	0.96	0.87
0	238	771	22	35	0.87	0.97	0.86
1	230	780	13	43	0.84	0.98	0.86
3	196	789	4	77	0.72	0.99	0.8
6	137	791	2	136	0.5	1	0.65

[1]: αντιπροσωπεύει τα αληθώς θετικά, δηλαδή το σύνολο των πρωτεϊνών με Sec που βρίσκονται πάνω από το συγκεκριμένο cutoff.
[2]: αντιπροσωπεύει τα αληθώς αρνητικά, δηλαδή το σύνολο πρωτεϊνών TM,CYTO ,που βρίσκονται κάτω από το συγκεκριμένο cutoff.
[3]: αντιπροσωπεύει τα ψευδώς θετικά, δηλαδή το σύνολο πρωτεϊνών TM,CYTO , που βρίσκονται πάνω από το συγκεκριμένο cutoff.
[4]: αντιπροσωπεύει τα ψευδώς αρνητικά , δηλαδή το σύνολο των πρωτεϊνών με Sec που βρίσκονται κάτω από το συγκεκριμένο cutoff.

Πίνακας Α_2.7 : Παρουσιάζονται τα αποτελέσματα από τον έλεγχο των δεδομένων δοκιμής με το profile TAT.hmm.

cutoff	TP[1]	TN[2]	FP[3]	FN[4]	Sensitivity	Specificity	mcc
-15	74	785	281	1	0.99	0.74	0.39
-12	74	864	202	1	0.99	0.81	0.46
-8	73	957	109	2	0.97	0.9	0.59
-6	72	990	76	3	0.96	0.93	0.66
-3	72	1022	44	3	0.96	0.96	0.75
-1	69	1044	22	6	0.92	0.98	0.82
0	69	1050	16	6	0.92	0.98	0.85
1	68	1054	12	7	0.91	0.99	0.87
3	62	1059	7	13	0.83	0.99	0.85
6	55	1064	2	20	0.73	1	0.83

[1]: αντιπροσωπεύει τα αληθώς θετικά, δηλαδή το σύνολο των πρωτεϊνών με Tat που βρίσκονται πάνω από το συγκεκριμένο cutoff.
[2]: αντιπροσωπεύει τα αληθώς αρνητικά, δηλαδή το σύνολο πρωτεϊνών SEC,TM,CYTO ,που βρίσκονται κάτω από το συγκεκριμένο cutoff.
[3]: αντιπροσωπεύει τα ψευδώς θετικά, δηλαδή το σύνολο πρωτεϊνών SEC,TM,CYTO , που βρίσκονται πάνω από το συγκεκριμένο cutoff.
[4]: αντιπροσωπεύει τα ψευδώς αρνητικά , δηλαδή το σύνολο των πρωτεϊνών με Tat που βρίσκονται κάτω από το συγκεκριμένο cutoff.

Πίνακας Α_2.8 : Παρουσιάζονται τα αποτελέσματα από τον έλεγχο των δεδομένων δοκιμής με το profile TAT.hmm,χωρίς να λάβουμε υπόψη τα δεδομένα από τις πρωτεΐνες με το Sec πεπτίδιο οδηγητή.

cutoff	TP[1]	TN[2]	FP[3]	FN[4]	Sensitivity	Specificity	mcc
-15	74	729	64	1	0.99	0.92	0.7
-12	74	758	35	1	0.99	0.96	0.8
-8	73	784	9	2	0.97	0.99	0.92
-6	72	787	6	3	0.96	0.99	0.94
-3	72	790	3	3	0.96	1	0.96
-1	69	791	2	6	0.92	1	0.94
0	69	791	2	6	0.92	1	0.94
1	68	791	2	7	0.91	1	0.93
3	62	793	0	13	0.83	1	0.9
6	55	793	0	20	0.73	1	0.85

[1]: αντιπροσωπεύει τα αληθώς θετικά, δηλαδή το σύνολο των πρωτεϊνών με Tat που βρίσκονται πάνω από το συγκεκριμένο cutoff.
[2]: αντιπροσωπεύει τα αληθώς αρνητικά, δηλαδή το σύνολο πρωτεϊνών TM,CYTO ,που βρίσκονται κάτω από το συγκεκριμένο cutoff.
[3]: αντιπροσωπεύει τα ψευδώς θετικά, δηλαδή το σύνολο πρωτεϊνών TM,CYTO , που βρίσκονται πάνω από το συγκεκριμένο cutoff.
[4]: αντιπροσωπεύει τα ψευδώς αρνητικά , δηλαδή το σύνολο των πρωτεϊνών με Tat που βρίσκονται κάτω από το συγκεκριμένο cutoff.

Παράρτημα Β

Πίνακας B_1.1: Ακολουθίες πρωτεϊνών με Sec πεπτίδιο οδηγητή, σε στοίχιση.

ID	SEQ		
P0AEB2	MNTIFSARIMKR	LALTTALCTAFI	SAAHA
P31716	MLKNK	ILATTLSVSLLAPLANPLL	ENAKA
P20041	MNHRYT	LLALAAAAL	SAGAHA
P11278	MKKR	ALLLSMSVLAML	YIPAGQA
P31133	MTALNKK	WLSGLVAGALMAV	SVGTLA
P01549	MLQNTSRFLAR	AGATVGVAAGLAF	SLPADRDG
P06129	MIKK	ASLLTACSVTAF	SAWA
P23659	MRK	FYSFAIIISLLVTGLFI	HTPKAEA
P31746	MNDLNDFLKT	ILLSFIFFLLL	SLPTVAEA
P0AES9	MKK	VLGVILGGLLLLPVV	SNA
P33379	MGLNRFMR	AMMVVFITANCITI	NPDIIFA
P18278	MTRPASAKRRS	LLGILAAGTICAAALPYAAV	PARA
P15555	MVSGTVGRGT	ALGAVLLALLAV	PAQAGTAAA
P09332	MDKNMFKK	IILAASIFTISLPVI	PFESTLQA
P06970	MKKYVTTKSVQPVAFR	LTTLSLVMSAVL	GSASVIA
P10641	MKKFNQS	LLATAMLLAA	GGANA
P18539	MTKMNRC	AALIAALIL	PTAHA
P0A402	MRR	FLALLLVLTLWL	GFTPLASA
Q04681	MKLSK	IALAAALVF	GINSVATA
P16699	MKVYKK	VAFVMAFIMFFSVL	PTISMS
P09888	MQPAKN	LLFSSLLFSSLLF	SSAARA
P20713	MNKK	AMAAAVSMILA	GGAHA
P01092	MRVRALR	LAALVGAGAALALSPLAA	GPASA
P24228	MRFSR	FIIGLTSCIAF	SVQA
P08954	MSNKKLILK	LFICSTIFITFVFA	LHDKRVVAA
P19405	MKKFPKK	LLPIAVLSSIAFSSLA	SGSVPEASA
P31956	MKMK	WLISVIILFVFIFPQNLVF	AGE
P18895	MNSSRSVNPRPSFAPR	ALSLAIALLL	GAPAFA
P13423	MKKRK	VLIPLMALST	ILVSSTGNLEVIQA
P23031	MINHNKTPNILAKVFKRT	CGLVSTGAALAIL	SQAASA
P14775	MKTT	LIAAAIVALSGLAA	PALA
P19584	MIGAFKRLGQK	LFLTLLTASLIFASSIV	TANA
P05364	MMRKS	LCCALLL	GISCSALA

P17901	MKKTT	AFLLCFLMIFTALL	PMQNANA
P21697	MSKK	FLTILAGLLLVVSSFFL	SVSPAAA
P04738	MKK	AFLLACVFFL	TGGGVSHA
P11312	MQK	IQFILGILAAA	SSSATLA
P12267	MKK	VLLSAAMATAFFGMAA	ANA
P22266	MSSAMRKTTNSPVVRR	LTAAAVALGSCLALA	GPAGSAGA
P35811	MKTFSVTKSS	VVFAMALGMA	STAFA
P23903	MKPSHFTEKRFMKK	VLGLFLVVVMLASVGVL	PTSKVQA
Q05523	MRRRKQN	WLFWLLSICLCLTFGPF	QQTVKA
P07024	MKLLQRG	VALALLTTFTLA	SETALA
P17989	MNIKKTAVKS	ALAVAAAAAAL	TTNVSA
P26877	MKKKS	LLPLGLAIGLASLAA	SPLIQA
P04172	MR	ALAFAAALAAF	SATAALA
P13812	MSSKK	IIGAFVLM	TGILSG
P16952	MKNKKEVYGFRKSK	VAKTLCGAVLGTALIAF	ADKAVFA
P14756	MKKVSTLD	LLFVAIM	GVSPAAFA
P21338	MKAFWRN	AALLAVSLL	PFSSANA
P16630	MQTVNTQPHRIFR	VLLPAVFSSLLL	SSLTVSAA
P27050	MNQAVRFRP	VITFALAFILIITWFA	PRADA
P20379	MIR	LSAAAALGLAAALAA	SPALA
Q3J164	MKFQVK	ALAAIAAFA	ALPALA
P62593	MSIQHFR	VALIPFFAAF	CLPVFA
P27458	MKKISK	AGLGLALVCALA	TIGGNA
P15292	MQRKKKG	LSILLAGTVALGALAVL	PVGEIQAKA
P0AET2	MNISSLRK	AFIFMGAVAALSLV	NAQSALA
P0AEQ3	MKSVLK	VSLAALTLAFAV	SSHAA
P12625	MVRRLWRR	IAGWLAACVAILCAF	PLHAA
P23550	MKKRRSSK	VILSLAIVVALLAAV	EPNAALA
P01552	MYKR	LFISHVILIFALILVI	STPNVLA
P18103	MKKT	LLAIILGGMAFA	TTNASA
P0A3R9	MVPISIIRNRVAK	VAVGSAAVLGLAV	GFQTPAVA
P19247	MKK	MTLFTLSLLATAV	QVGA
P19361	MPNRRRCK	LSTAISTVA	TLAIA
P16397	MRKKTKNR	LISSVLSTVVISSLLF	PGAAGA
P17953	MKQQTEVKKRFKMYKAKK	HWVVAPILFIGVLGVVGL	ATDDVQA

Q00971	MNKTQRH	INWLLAVSAATAL	PVTAA
P23135	MTTIVKR	ALVAAGMVLAI	GGAAQA
P24735	MRDTR	FPCLCGIAASTLLFA	TTPAIA
P0ABK9	MTRIKINARR	IFSLLIPFFFF	TSVHA
P31797	MRR	WLSLVLSMSFVFSAIFIV	SDTQKVTVEA
P26501	MKCPK	ILAALLGCAVLAGVPAM	PAHA
P16216	MNSKK	IGAMIAAAVLSLIVMTPAATRKIV	QRQTRNSSTAVENSA
P13810	MIKH	VLLFFVFI	SFSVSA
P00099	MKPY	ALLSLLATGTLLA	QGAWA
P0AG82	MKVMRTT	VATVVAATLSMSAF	SVFA
P11000	MKMKRK	LLSLVSVLTILLGAFWV	TKIVKA
P15917	MNIKKEFIK	VISMSCLVTAITLSGPVFI	PLVQGA
P00809	MILKNKRMLK	IGICVGIL	GLSITSLEA
P07254	MRKFNKP	LLALLI	GSTLCSAAQA
P18958	MKQ	LIIATLLSAL	SGGCMA
P04127	MIKS	VIAGAVAMAVV	SFGVNNAA
P05825	MNKK	IHSLALLV	NLGIYGVAQA
P07111	MRLR	FSVPLFFFGCVFV	HGVFA
P11889	MKGKLLK	GVLSLGVGLGAL	YSGTSAQA
P15922	MKVITFSRRS	ALASIVATCLM	STPALA
P04981	MQRQAGLPLK	ANPMHTIASILLSVL	GIYSPADVA
P19926	MNKT	LIAAAVAGIVLLA	SNAQA
P28307	MKLLKVA	AIAAIVF	SGSALA
P20845	MKGKK	WTALALTLPLAA	SLSTGVDAET
P14768	MRTAMAKS	LGAAAFLGAALFA	HTLAA
P14774	MMNRVK	IGTALLGLTLAGIAL	PALA
P31831	MKK	IMLLLMTLLLVSLPLA	QEAQA
P0AEE5	MNKK	VLTLSAVMASMLFGAA	AHA
P15320	MKNNNFRLSAAGK	LAAALAIILAA	SAGAYA
P0AD64	MRYIR	LCIISLLATLPLA	VHA
P06279	MLTFHRIIRKG	WMFLLAFLLTALLF	CPTGQPAKA
P14892	MEIKQ	MLVPVSRYSVLCPYEMNPTEITF	HNTYNDAPA
P05818	MNLKK	IAIASSVFAGITMAL	TCHA
P13430	MKLK	AIILATGLINCIAF	SAQA
P0AG80	MKPLHYT	ASALALGLALM	GNAQA

P11220	MRFRHK	AAALAATLALPLAGLVGLA	SPAQA
P13794	MKLKNT	LGVVIGSLVAASAM	NAFA
P30141	MSKTNPNKLYSLRK	LKTGTASVAVDLTVLG	TGLANTTDVKA
P15452	MKK	FLLVAVVGLA	GITFA
P19843	MFKAQATFSRYS	AAVSLLLLF	SGAAQA
P09790	MKAKRFK	INAISLSIFLAYAL	TPYSEAA
P14005	MKR	FALSLLAGLVAL	QASA
Q03011	MKLNK	LALVLGLGLSVVA	GSALA
P33665	MHSQHRTAR	IALAVVLTAIPASLATAGVGYA	STQASTAVKAGA
P22940	MKK	IMLIASAMSAL	SLPFSASA
P08704	MKRNRFFNTS	AAIAISIALNTFF	CSMQTIAA
P08506	MTQYSSLLRG	LAAGSAFLFLFA	PTAFA
P23549	MKRS	ISIFITCLLITLLTMGGMLA	SPASA
P10549	MRFRPS	IVALLSVCFGLLTFL	YSGSAFA
P10477	MKK	IVSLVCVLVMLVSILGSFSVV	AASPVKGFQV
P13650	MNKHLLAK	IALLSAVQLV	TLSAFA
Q01786	MAKSPAARKGXPP	VAVAVTAALALLIALL	SPGVAQA
P62560	MENNKKVLKK	MVFFVLVTFL	GLTISQEVFA
P0A2C5	MNMKK	LATLVSAVAL	SATVSANAMA
P32890	MNKVKCY	VLFTALL	SSLYAHG
P05149	MKK	LAILGVTVYSFA	QLANA
P19487	MSIFRT	ASTLALATALALAA	GPAFS
Q05044	MQSSLKKS	LYLGLAALSFAGVAAV	STTASA
P02943	MMITLRK	LPLAVAVAAGVM	SAQAMA
P09794	MARR	LATASLAVLAAAATAL	TAPTPAAA
P32722	MKVMK	WSAIALAV	SAGSTQFAVA
P04979	MLINNKK	LLHHILPILVLALL	GMRTAQA
P09545	MPKLNRCA	IAIFTIL	SAISSPTLLA
P06546	MKK	VVNSVLASALALTVA	PMAFA
P0C2E9	MIRFKKTK	LIASIAMALCLFSQPVI	SFS
P24305	MKKS	LIALAVLAA	SGAAMA
P0A921	MRT	LQGWLLPVFMLPMA	VYA
P07986	MPRTTPAPGHPARGAR	TALRTTRRRAATLVVGATVVL	PAQA
P13734	MLK	VIPWLLVTSSLVAI	PTYIHA
P24474	MPFGKP	LVGTLLASLTLLGLA	TAHA

P16869	MLSTQFNRDNQYQAITKP	SLLAGCIALALL	PSAAFA
P68588	MSHCVVLNK	LESVLII	GDSRYALS
P21543	MTLYRSLWKKGC	MLLLSLVLSLTAFI	GSPSNTASA
P00648	MMKMEGIALKKR	LSWISVCLLVLVSAAGMLF	STAA
P00446	MNKAKT	LLFTALAF	GLSHQALA
P14488	MKNTLLK	LGVCVSLLGITPFV	STISSVQA
P29957	MKLNK	IITTAGLSLGLLL	PSIATA
P24040	MSNVGKP	ILAGLIAGLSLLGLAV	AQA
P13507	MSHILR	AAVLAAMLL	PLPSMA
P21982	MKR	FFAILGAALFV	GNSGAFA
P02910	MKK	LALSLSLVLAF	SSATAAFA
P20723	MKK	FNILIALLFFTSLVI	SPLNVKA
P33682	MSRKLRT	LMAALCALPLAFAAA	PPAHA
P08306	MMAIATKRRG	VAAVMSLGVATMTAV	PALA
P21413	MRMKKS	ALTLAVLSSLF	SGYSLAAPA
P19401	MAKNTTNRH	YSLRKLKTGTASVAVALTVVGAGLV	AGQTVRA
P22390	MFKKRGRQT	VLIAAVLAFF	TASSPLLA
Q02192	MFRRSKNNSYDTSQTKQR	FSIKKFKFGAASVLIGLSFL	GGV
P0C0T5	MNKT	AIALLALLA	SSASLA
P06886	MNKK	LLMNFFIVSPLLLATTATDFTPVPL	SSNQIIKTAKA
P27035	MKR	LLALLATGVSIVGLTALA	GPPAQA
P31835	MKKQVK	WLTSVSMSVGIALGAAL	PVWA
P09489	MILNKRLK	LAYCVFL	GCYGLSIHSSLA
P27195	MKYKK	LSVAVAAFAFAA	VSA
P07110	MKDR	IPFAVNNITCVILLSLF	CNA
P13470	MEKKVRFKLRKVKKR	WVTVSVASAVV	TLTSLSGS
P33590	MLSTLRRT	LFALLACASFI	VHAAA
P0AFK9	MKKWSRH	LLAAGALAL	GMSAAHA
P18477	MHSLNTRR	GLGLAAAMTLAAGALVA	PTGAA
P33781	MKKN	LLITSVLAMA	TVSGSVLA
P35150	MRIFKK	AVFVIMISFLIA	TVNVNTAHA
P15930	MFK	ALAGIVLALVA	TLAHA
P14191	MKKT	LIALAIAASAA	SGMAHA
P12608	MIVTGSQVRQGLNTWFVLPLRR	TAIGLGCAGVATLFSAC	GQTQA
P36924	MKNQFQYCC	IVILSVVMLFVSLLI	PQASSA

P62605	MKLK	FISMAVFSALTLGVA	TNAS
P22340	MYRKST	LAMLIALLTSAA	SAHA
Q70YI1	MKMK	LVTAAVMGLAM	STAMA
P05458	MPRSTWFK	ALLLLVALWA	PLSQA
P04377	MRNIAIK	FAAAGILAMLAA	PALA
Q01269	MPKSFRH	LVQALACLALLA	SASLQA
P06597	MKKLLKS	VLVFAAL	SSASSLQA
P04032	MKR	VLLLSSLCAALSFGLAV	SGVA
P20626	MREK	VVLFLSIIMAIML	PVGNAAA
P15644	MIIANVIRS	FSLTLLIFAALLF	RPAAA
P0C0V0	MKKTTLALSA	LALSLGLAL	SPLSATA
P02924	MHKFTK	ALAAIGLAAVM	SQSAMA
P02931	MMKRN	ILAVIVPALLVA	GTANA
P14738	MKNNLRYGIRKHK	LGAASVFLGTMIVV	GMGQDKEAA
P30920	MFQMAKRAFLSTT	LTLGLLAGSALPFL	PASAVYA
P15321	MIKK	ITALTLLV	STALSA
P21175	MKKGTQRLSR	LFAAMAIAGFA	SYSMA
P0AG78	MNK	WGVGLTFLLAA	TSVMA
P11701	METKVRKKMYKKGK	FWVVATITTAML	TGIGLSSV
P22629	MRK	IVVAAIAVSLTTVSITA	SASA
P22751	MKY	AASGLLSVALNSLLLL	GSNQRFA
P13717	MRFNNKM	LALAALLFAA	QASA
P61153	MRNTAR	WAATLGLTATAVCGPLAGASLAS	PATAPA
P00775	MKHFLRALKRCS	VAVATVAIAVV	GLQPVTASA
Q02307	MKKNLVKS	LAIASAVISIYSIVNIV	SPTNVIA
P17315	MFRLNPFVR	VGLCLSAISCAW	PVLA
P04957	MPYLKR	VLLLLVTGLFMSLFAV	TATASA
P00131	MRK	LFFCGVLALAVAFAL	PVVA
P22364	MISATKIRSC	LAACVLAAF	GATGALA
P14283	MNMSLSRIVKAAPLRRTT	LAMALGALGAA	PAAHA
P15488	MLKIKY	LLIGLSLSAM	SSYSLAAA
P19369	MPNFFRNGC	IALVGSVAAM	GAAHA
P13036	MTPLRVFRKTTPLVNTIR	LSLLPLAGLSF	SAFA
P26827	MKKTFK	LILVLMLSLTLVF	GLTAPIQA
P0A3R5	MLRRFPTRTTAPGQGGARRSRVR	ALAWLLASGAMTHL	SPALA

P19576	MKIKTG	VGILALSALTTMMI	SAPALA
P20862	MIKK	VPVLLFFMA	SISITHA
P12993	MKKT	AFILLLFIALTLTTSPLV	NGSEK
P19909	MEKEKKVKYFLRKS	AFGLASVSAAFLV	GSTVFA
P09616	MKTR	IVSSVTTTLLLGSILM	NPVAGA
P27477	MSVRSLRWPRQK	AFLAVISLVVAVLLAVPGWL	TPATA
P19250	MKYRYFAKKS	FLFISMLAAF	KTFA
P06278	MKQQKRLYAR	LLTLLFALIFLL	PHSAAAA
P04635	MKETKHQHTFSIRKS	AYGAASVMVASCIFVI	GGGVAEA
P0A3T3	MKFGSKIRR	LAVAAVAGAIAL	GASFAVA
P28623	MIKHLLSRGK	LLLFVSVMATSSIIA	GGNAYG
P18429	MFKFKKN	FLVGLSAALMSISLF	SATASA
P0ADA1	MMNFNNVFR	WHLPFLFLVLL	TFRAAA
P13429	MVKDIIKT	VTFSCMLAGSMFV	TCHVCAA
P13720	MKK	WFPAFLFL	SLSGGNDALA
P04960	MKNTRVRSIGTKS	LLAAVVTAALMA	TSAYA
P23598	MRRK	AVLLTVVL	SLSGGSAQAMG
P15704	MFSKIKKINFFKKT	FSFLIAVVMMLFTVL	GTNTY
P22391	MLKSSWRKT	ALMAAAAVPLLL	ASGSLWASADA
P06717	MKN	ITFIFFILLA	SPLYA
P13626	MRNASVTARLTRSVRAIVKT	LLIAIATVTFYFSCDLAL	PQSAAA
P12616	MKYNTSTLGRRA	AAAAGVLTLAVLGLA	PMAQA
P35804	MRTVGIGAGVRRLGR	AVVMAAAVGGLVL	GSAGASNA
P17855	MKKQ	IISLGALAVASSLF	TWDNKADA
P07103	MPLSYLDKNPVIDSKKHALRKK	LFLSCAYFGLSLACL	SSNAWA
P29822	MDYSRLLKRS	VSAALTAAALL	CSTAAFA
P27755	MNAIKT	AVAAVTAAASLVAF	SPAEA
P06111	MKK	IALFITASLIA	GNALA
Q00499	MK	ISIYATLAALSLAL	PAVA
P0A917	MKK	IACLSALAAVLAF	TAGTSVAA
P21171	MNMKK	ATIAATAGIAVTAFAA	PTIASA
P14090	MVSRRSSQARG	ALTAVVATLALALA	GSGTALA
P26221	MSVTEPPPRRRGRHSRARR	FLTSLGATAALTAGMLGVPLA	TGTAHA
P19531	MKKKT	LSLFVGLMLLIGLLFSGSL	PYNPNAAEA
P00805	MEFFKKT	ALAALVM	GFSGAALA

P24121	MKNKLLFK	IFLSLSLALSVYSINDKII	EVSNTSLA
P11073	MKS	LITPITAGLLLAL	SQPLLA
P23847	MRISLKKSGMLK	LGLSLVAMTVAA	SVQA
P25447	MKKT	MMAAALVLSAL	SIQSALA
P06971	MARSKTAQPKHSLRK	IAVVVATAV	SGMSVYAQA
P31697	MSNKNVNVRKSQE	ITFCLLAGILMFMAMMVA	GRAEA
P0ABE7	MRKS	LLAILAVSSLVF	SSASFAA
P18473	MKLSMKS	LAALLMML	NGAVMA
P27951	MFKSNYERKMRYSIRK	FSVGVASVAVASLFM	GSVAHA
P17543	MVVNKTT	AVLYLIALSLSGFI	HTFLRA
P09333	MNKK	VVLSVLSTTLVASVAA	SAFA
P04977	MRCTRAIRQTARTG	WLTWLAILAV	TAPVTSPAWA
P09331	MNNSKIISK	VLLSLSLFTVGASAFVI	QDELMQKNHAKA
P0AFH8	MTMTRLKISKT	LLAVMLTSAVA	TGSAYA
P15319	MIRKK	ILMAAIPLFVI	SGADA
P07941	MKNRNR	MIVNGIVTSLI	CCSSLSALA
P31830	MK	IFGLVIMSLLFV	SLPITQQPEAR
P19571	MKMRTGKKG	FLSILLAFLLVITSIPFTLV	DVEA
P04845	MKKT	AIALAVALAGFA	TVAQA
P11797	MSTRKAVIGYYFIPTNQINN	YTETDTSVVPFPV	SNITPAKA
P27032	MGKPMWRC	WALMLMVWF	SASATA
P32823	MKLNKITSY	IGFALL	SGGALA
P24093	MKK	LAIMAAASMVFAV	SSAHA
P02930	MKK	LLPILIGLSLSGF	SSLSQA
P25394	MKR	LVFISFVAL	SMTAGSAMA
P31715	MKMNKLVKSS	VATSMALLLL	SGTANA
P05655	MNIKKFAKQ	ATVLTFTTALLA	GGATQAFA
P09169	MRAK	LLGIVLTTPIAI	SSFA
P0AFM2	MRHS	VLFATAFATLI	STQTFAA
P07102	MK	AILIPFLSLLI	PLTPQSAFA
P24037	MKKT	LMASAVGAVIAF	GTHGAMA
P12293	MNRNTPKARG	ASSLAMAVAMGLAVL	TTAPATA
P00777	MRIKRTSNRSNAARR	VRTTAVLAGLAAVAALAV	PTANA
P0C0J0	MNKKKLGIR	LLSLLALGGFVLA	NPVFA
P16454	MKKT	LLASSLIACLSIA	SVNVYA

P12061	MRKS	ASAVAVLALIA	CGSAHA
P18956	MIKPTFLRR	VAIAALL	SGSCFSAAA
P14542	MMISKKYT	LWALNPLLLTMMA	PAVA
P20910	MPMFRIRLPKP	AALIAAGGIGACIATVAV	PSAYA
P0AGC3	MEKAKQVTWR	LLAAGVCLL	TVSSVARA
P00083	MRK	LVFGLFVLAASVA	PAAA
P05695	MIRRHSCKGVGSS	VAWSLLGLAI	SAQSLA
P06608	MERWFKS	LFVLVLFFVF	TASA
P13482	MKSPAPSRPQK	MALIPACIFLCFAAL	SVQA
P33406	MNMKKFVKKP	LAIAVLMLASGGMV	NMVHA
P0C2T2	MTDVSRKIRAWGRR	LMIGTAAAVVLPGLVGLA	GGAATAGA
Q01996	MQQQHLFR	LNILCLSLMTAL	PAYA
P17137	MLRRK	VIFTVLATLVMTSLTIV	DNTAFA
P02971	MKFKKT	IGAMALTTMFVAV	SASA
P04816	MKRNAKT	IIAGMIALAI	SHTAMA
P30705	MNKST	LAIVVSIIA	SASVHAA
P00694	MNLRKLR	LLFVMCIGLTLILTAV	PAHA
P26514	MGSYALPRSGVRRSIR	VLLLALVVGVLGTATALIA	PPGAHA
P32520	MKK	LFVVLTSIFIAA	SAYG
P21948	MKQS	AIALALLSCLI	TPVSQA
P07528	MITLFRKP	FVAGLAISLLV	GGGIGNVAAA
Q05156	MKRRTT	AVLTLTALLGTALTAL	PVQQAGA
P06874	MNKR	AMLGAIGLAFGLLAA	PIGASA
P17835	MSKFSYPALR	AALILAASPVL	PALA
P0C1A8	MLKTISGT	LALSLIIAA	SVHQAQA
P01559	MKK	LMLAIFISVL	SFPSFS
P00282	MLRK	LAAVSLLSLL	SAPLLA
P29725	MGRY	IVPALLCVA	GMGFAHA
P15279	MSR	FVTSVSALAMLALAPAAL	SSVAYA
P11439	MH	LTPHWIPLVASLGLLA	GGSFASA
P20149	MKKT	LAALIVGAFAA	SAANA
P0A0Y3	MKTSIRY	ALLAAALTAA	TPALA
P24059	MHLHLRG	ICLVLAVA	SSSSSALA
Q02760	MIRK	LTLTAATALAL	SGGAAMA
P0A1V8	MAIR	IFAILFSIFSLA	TFAHA

P22865	MKKK	IISAILMSTVILSAAA	PLSGVYA
P24092	MRNGRTLLR	WAGVLAATAIIGVGGFW	SQGTT
P19424	MKIKQIKQS	LSLLLIITLIMSLFV	PMASA
P0C1C0	MKY	LLPSAAAGLLLLAA	QPTMA
P17266	MKQT	ICGLAVLAALSSA	PVFA
P18336	MPLR	ALVAVIVTTAVMLV	PRAWA
P09394	MKLTLKN	LSMAIMMSTIVM	GSSAMA
P01077	MVRKR	ALGLAGSALTLVLGAV	GFTAPAQA
P30692	MKKS	LIALTLAALPVA	AMA
P33673	MHMSNARPSKSRTK	FLLAFLCFTLMASLFGATALF	GPSKAAA
P23827	MKT	ILPAVLFAAFA	TTSAWA
P06202	MSNITKKS	LIAAGILTALIAA	SAATA
P14212	MKKT	LLGSLILLAFA	GNVQA
P10520	MKN	YLSIGVIALLFALTF	GTVKSVQAIA
P20861	MKKLYK	AITVICILM	SNLQSA

Πίνακας Β_1.2: Ακολουθίες πρωτεϊνών με ΤΑΤ πεπτίδιο οδηγητή, σε στοίχιση.

ID	SEQ			
P76342	MKKNQFLKESDVTAESVFF	MKRRQVLK	ALGISATALSL	PHAAHA
Q57366	MTKLSGQELHAE	LSRRAFLS	YTAAVGALGLCGTSLLA	QGARA
P39185	MK	ISRRDFIK	QTAITATASVAGVTL	PAGA
Q9RK81	MSQTPA	VSRRLLLG	SAAATGALATGIGSAA	PVAAA
Q9RI72	MQQDGTQQDRIKQSPAPLNG	MSRRGFLG	GAGTLALATASGLLL	PGTAHA
P96465	M	LARRRFLQ	FSGAAVASSLALPLLARAA	GKTAASA
P17201	MGRLNRFRLGKDGRREQAS	LSRRGFLV	TSLGAGVMF	GFARPSSA
P39597	MSDEQKKPEQ	IHRRDILK	WGAMAGA	AVA
P63882	MNDL	LTRRLLTM	GAAAAMLAAVLLL	TPITVPA
Q63Q46	MLIKKTLRAALAGDDIPRSEITPRAVF	EHRRRILQ	AAGAAAAGGLVGAHGLALA	AYA
P07984	M	STRRTAAA	LLAAAAVAVGGLTAL	TTTAAQA
Q55460	MGS	FNRRKFLL	TSAATATGALFL	KGCAGNPPDPNA
A6VQE0	MNKFTKTDVTPEKLF	IQRRKIIQ	GMSVLSAAAAF	PNLAAA
Q59634	MFGT	PSRRTFLT	ASALSAMALAA	SPTVTDAIA
P35392	MDRTTAR	PNRRAVLA	TGVGAALAATAAAA	GPAHA
Q7MR39	MKPKPSDVTPKELF	DQRRSFLK	LGAASVVATGSVSNLLAAL	KTKA
P14559	MHPSTSR	PSRRTLLT	ATAGAALAAATLV	PGTAHA
Q8X6I9	MSNQGEYPEDNRVGKHEPHDLS	LTRRDLIK	VSAATAAAAVVY	PHSTLAASVPA
Q06530	MT	LNRRDFIK	TSGAAVAAVGIL	GFPHLAFG
P36548	MSTFKPLKTL	TSRRQVLK	AGLAALTLSGM	SQAIA
P15713	MIS	KSRRSFIR	LAAGTVGATVATSML	PSSIQAALA
Q9HYL2	MSDDTKSPHEETHG	LNRRGFLG	ASALTGAAALVGASALGSAVV	GREARA
P22222	MPHDRKN	SSRRAWAA	LCAAVLAVSGALVGVAA	PASA
P44847	MPR	LSRRQLLK	TAAISTAL	STVPAPLLAA
Q01578	MTTGR	MSRRECLS	AAVMVPIAAM	TATATITGSAQA
Q21AR4	MTTPK	LDRRQVLK	LEAAAMAALAGGIAM	PAAAA
Q8ECL7	MHN	IHRRHFLK	AAGAVTAGLVTANIAL	NANA
P07883	M	VNRRDLIK	WSAVALGAGAGLAG	PAPAAHA
P73452	MSNFSR	STRRKFMF	TAGAAAIGGVVL	HGCTSPTTTS
B2UL75	MDNS	SSRRRFLQ	TLGLATGALAA	GSFANA
Q9PA38	ML	MYRRDFLK	SVTAAWVAFGL	PNPLGGPFATNRVIA
P94127	MESKEHKG	LSRRALFS	ATAGSAILAGTVGPAAL	SLGAAGLATPARA
Q59746	MSNEETKMR	LNRRQMLG	TTAFMAAAGAVGAGGAL	TLSGGTATPARA
Q8XT53	M	VTRRHLLA	SASLSATLAAL	GITPEALA
P81040	MN	IGRRDLIC	GLGGLAVGGAML	GLGSVEARA
Q8XUX6	M	HRRDLLK	QLAAGFLALA	PGLTPSTASA

Q8GPG4	MLR	TTRRTLMQ	GASLVGAGLFAA	GRGWA
P22641	MLGNFRFDDMVEKLSRRVAGR	TSRRGAIG	RLGTVLAGAALV	PLLPVDRRGRVSRANA
Q59517	MTG	LSRRNVLI	GSLVAAAAVGAGVGGAA	PAFA
P45015	MTV	CSRRNFVS	GMGAVILM	TGTSLPAFA
P38043	MSQ	FSRRKFLL	TAGGTAAAALWL	NACGSNNSST
Q8FX16	MTEETRKSQ	LSRRQLLG	TSAFVAAAGASGLGGALLA	GSSETALA
C3MB06	MPVYRPPRIAASEITPERFF	LDRRSFLA	AAGGLVLGGTGMA	HAAA
Q9A4T2	MLIRHAPDLTDNDVTDHSLY	LKRRTLMA	GVAGLGVAGA	SASHAQA
P40120	M	DRRRFIK	GSMAMAAVCGTSGIASLF	SQAAFA
Q44052	MMN	LSRRTLLT	TGSAATLAYALGMA	GSAQA
P06200	MTENWK	FRRRTFLK	HGAQAATLAGLSGLF	PETLRRALA
Q53239	M	FTRRAALV	GAAALASAPLVI	RTAGAEEAPA
Q8XAS4	MQYEDKNGVNE	PSRRRLLK	GIGALALA	GSCPVAHA
P38501	MAEQMQ	ISRRTILA	GAALAGALAPVLA	TTSAWG
Q888N2	MLIKLPSASGSKESDVTPESIY	LSRRTLLA	SSLAGLAVTAL	PRWASAADA
Q8YBC6	MTEETRKSQ	LSRRQLLG	TSAFVAAAGASGLGGALLA	GSSETALA
Q01537	MSEQFR	LTRRSMLA	GAAVAGALAPVV	TSVAHA
P13063	MS	LSRREFVK	LCSAGVAGLGI	SQIYHPGIVHA
Q50644	MGADLKQPQDADSPPKG	VSRRRFLT	TGAAAVVGTGVGAGGTALL	SSHPRGPA
P22637	MTDSRANRADATRGVAS	VSRRRFLA	GAGLTAGAIAL	SSMSTSASA
P55669	MT	ISRRDLFK	AGLAAGAALSVPSLL	RAQTAVA
Q88DZ2	MLIKLPRSSECKASEITPEGIY	LSRRTLLG	GSLAGLALGAL	PGGVGAAQMSRYA
B2UQL7	MSIF	SSRRQFLK	SLGLAAGAAAAGNAL	PGKA
P10509	MRLTQAP	PSRRTLMT	LGAGATMAALL	PAGGAAYA
Q8PLY8	MSFRDALNLPSSEITDESVY	RDRRRLLQ	LLALTPALGVAGCAEA	DPPPPPKTVVTPAQA
P69741	MTGDNTLIHSHG	INRRDFMK	LCAALAATMGL	SSKAAA
P18775	MKTKIPDAVLAAE	VSRRGLVK	TTAIGGLAMASSALTL	PFSRIAHA
Q8XV50	MLIKTDRWLRGDDIPASEITPQHLF	DQRRRLLA	AAALGAAGAAL	SPWAARRAFA
B3PWV0	MPSYRPPKIASSEITPRRIY	MRRREFLG	AAALGAVALY	GAGKASA
P12676	MTAQQH	LSRRRMLG	MAAFGAAALAGGTTIAA	PRAAAAAKSAA
P46448	MQ	VSRRKFFK	ICAGGMAGTSAAMLGFA	PANVLA
Q5LNE0	MAHRWINDLTPADITPRGAW	MNRRQVMA	GMAGAGLAAFA	GSAQA
Q2IGN2	MARWRPDTAEREATPEALY	LRRREFLA	LGAAGAVGLLL	PRGARA
A6LB54	ME	NTRRSFLK	KVSAAGIGAAGLAMA	GNAGA
P31884	MLEEKG	IERRDFMK	WAGAMTAML	SLPATFTPLTAKA
A4FN60	MAGDESRSNP	FSRRTLLR	TSAAAGAGLGVAGLSTGYGAA	QPVRPA
P55047	MPE	LSRRRALG	AAAALAAAAGTQAVAA	PAATA
P0A5E1	MTTMIT	LRRRFAVA	VAGVATAAATTVTLA	PAPANA

Q60AK7	MRKTSSPRIAPSEITPRDLY	HDRRRFMQ	AAAGAAAAALW	PHWLSA	
P63884	MSGSNTA	ISRRRLLQ	GAGAMWLLSV	SQVSLA	
Q89BU4	M	NRRQLLR	GSVAVSAAAVL	PRAADA	
Q89XJ6	MSDSDNIKG	VSRRTLLG	TTAAAAGVGLAGGAVV	TKDGAGFVSTADA	
Q01S58	MDK	TSRRDLLK	LASLAGIGAGLA	RSQGSSKSMA	
Q934F5	ML	IKRRAFLK	LTAAGATLSAFGGLGVDLA	PAKAQA	
P25006	MTEQLQ	MTRRTMLA	GAALAGAVAPLL	HTAQAHAAGAAA	
P0AB24	MTI	NFRRNALQ	LSVAALFSSAFM	ANA	
Q51705	MESKQEKG	LSRRALLG	ATAGGAAVAGAFGGRLALGPAAL	GLGTAGVATVAGSGAALA	
P07822	MSGLPL	ISRRRLLT	AMALSPLLW	QMNTAHA	
P84887	MRWLDKFGESLSRSVAHK	TSRRSVLR	SVGKLMVGSAFVL	PVLPVARA	
Q7CI09	MRDKTGPKFGPYQPDDEAVS	PSRRRLIL	GMGMVSGALVL	GGAKTAQA	
P0AAK9	MT	WSRRQFLT	GVGVLAAV	SGTAGRVVA	
P94170	MSST	LYRRQLLK	LLGMSVL	GTSFSSCVTSPARA	
P21853	MKISIGLGKEGVEERLAERG	VSRRDFLK	FCTAIAVTMGMGPAFA	PEVARA	
P52320	MERTT	LRRRALVA	GTATVAVGALALA	GLTGVASA	
B2FLK4	MFA	MKRREFIA	ASAAVAASSLL	PQTPAWA	
Q8XQB8	MMSKHPHSPSTPQDETPSV	PGRRRFMN	SAALAGLATVVA	CTDKGAPAGSAAA	
P82594	MCTREAVRMSREHDLPEI	PSRRLLLK	GAAAAGALTAV	PGVAHA	
P69739	MNNEETFYQAMRRQG	VTRRSFLK	YCSLAATSLGLGAGMA	PKIAWA	
B1W5J7	MNDDARPAPEPQDMPPHSGAADEPARQD	PSRRSVLW	TTAGVAGAGIGLGALG	TGNASA	
P50500	MSEKDKM	ITRRDALR	NIAVVVGSVATTTMM	GVGVADA	
P07603	MQIAS	ITRRGFLK	VACVTTGAALIGIRM	TGKAVA	
C1D9G3	MT	LTRRDFIK	ANAAAAAATAAAVNLPLV	PSMAQA	
P36649	M	QRRDFLK	YSVALGVASALPLW	SRAVFA	
P77374	MSKNERMVG	ISRRTLVK	STAIGSLALAAGGF	SLPFTLRNAAAA	
P71244	MPFKK	LSRRTFLT	ASSALAFL	HTPFARA	
Q7VKI8	MKQLMMSDVTPEEIF	NQRRQIIK	SMGLGIATLGL	PNIAFA	
P44798	MKKNNVN	EQRRDFLK	KTSLGVAGSALSGGMVGVV	SKSAVA	
Q9HVA4	MLIKIPSRSDCSESEVTSETLY	LSRRRLLG	ASFAGLALA	SGLPRLGFA	
O87948	M	NRRDFLK	GIASSSFVVLGGSSVL	TPLNALA	
P33225	MNNNDLFQ	ASRRRFLA	QLGGLTVAGMLGPSLL	TPRRATA	
Q9PIC3	MLITPEKLY	KQRRNFLK	LGAGALISSSVLA	SKLSA	
Q9CK41	MK	QSRRQFLK	NMSAMAATFAMPNFLIA	QNAFA	
B1Y6A6	MQ	SNRRDFLK	AQALAASAAAAGIPIV	VEA	
P0AAL5	MSRSAKPQ	NGRRRFLR	DVVRTAGGLAAVGVAL	GLQQQTARA	
P0AB09	MDKFD	ANRRKLLA	LGGVALGAAIL	PTPAFA	
Q44292	MTH	VSRRKFLF	TTGAAAAASILV	HGCTSNGSQSA	

P45004	MSNFNQ	ISRRDFVK	ASSAGAALAV	SNLTLPFNVMA
P35393	MGTTGAR	PSRRAVLT	AAAGAAVA	GIPLGGSTAFA
Q93HX3	MS	LTRRDFIK	ANAVPATAAAAGLATPAIA	QPAKA
Q8DLH9	MEK	VGRRVFLG	MGAAATAYVTHHLW	NQNAESSYA
P19573	MSDKDSKNTPQVPEKLG	LSRRGFLG	ASAVTGAAVAATALGGAVM	TRESWAQA
P81594	MENNQKRQQSG	MSRRSFLK	VGAAATTMGVIGAI	KAPAKVANA
Q7M962	MA	FSRREFLK	SAAAASAASAVGMSVPSQLL	AQA
Q07982	MTNKISSSDNLSNAVSATDDNASRTPN	LTRRALVG	GGVGLAAAGALASGLQAATL	PAGA
P31075	METT	MTRRDFLK	SAGAAGAAGLVW	SQTIPGTLGA
Q06650	MRKPTSS	LTRRSVLG	AGLGLGGALAL	GSTTASAASA
P26648	MS	LSRRQFIQ	ASGIALCAGAV	PLKASA
P46923	MT	LTRREFIK	HSGIAAGALVVTSAA	PLPAWA
Q8GGJ7	MEHQ	TSRRNFLK	IAGSSAAVAGAGLV	SGNANAAPA
P37600	MS	ISRRSFLQ	GVGIGCSACALGAF	PPGALA
B4SRN1	M	QRRDFIR	NASLALAAF	GLPSLPACA
Q9F0W4	MSDKKDQVPGAVEAPRG	VSRRSFLG	TGAVTGAVLAGATAL	GAGTFTRESWAAAA
Q5N0R0	MSESM	FSRRDFLL	GGTALAGTLLLDSFG	DWRRRAEA
P05448	MTDT	LNRRAAMA	LGLASAAGAALATPAL	SQDAAPA
Q7VJT5	MTSKIQGKKPT	LSRRDFIK	SAAAASAAA	SVGLSIPSVMSAEA
O34870	MKK	MSRRQFLK	GMFGALAAGALTA	GGGYGYA
Q9S1H0	MRKVMNSPDDG	NGRRRFLQ	FSMAALASAAA	PSSVWA
O34213	MSEHKNG	HTRRDFLL	RTITLAPAMAVGSTAMGALVA	PMAAGA
P0AAJ8	M	NRRNFIK	AASCGALLTGAL	PSVSHAAA
P81186	MS	TSRRDFLK	YFAMSAAVAAA	SGAGFGSLALA
P44652	MKVRLKSKKKMKKPALN	PERRKFLK	EATRTAGGLAGVGILL	GLQQNQSLA
Q44018	M	QRRHFIA	RAGIAAATAALGLAAM	PAQAQA
Q59543	MKKDTGFDSKIEKLARTTASK	TGRRGFIG	RLGGFLVGSALLPLL	PVDRRSRLGGEVQ
P77554	MKES	NSRREFLS	QSGKMVTAAALFG	TSVPLAHA
Q7NZY0	MLIRKPADHLPSEITSESVY	FNRRQFMA	GAAGLLLSAETLAGLAA	KKSPLSQLAA
P12374	MESR	TSRRTFVK	GLAAAGVLGGLGLW	RSPSWA
Q5FQ05	MALFRYPRPLPSEITPRDMY	LSRRSLIG	GAAALGAV	SATADA
P17687	MPE	LTRRRALG	AAAVVAAGVPLVAL	PAARA
P0A4R1	MSSFKPSRFSTARLTGDAVTPKSIY	LRRREFMI	GLGAIAATGAASSAFA	DPLEA
P55046	MPDI	TRRRAYTT	AAAVAATASAAAPTAA	PAATA
P95246	MGSEHPVDG	MTRRQFFA	KAAAATTAGAFMSLA	GPIIEKA
Q92Z36	MTGE	LTRREMLK	AHAAGIAAATAGIAL	PAAA
P55048	MPR	LTRRRALT	AAAALASGAGAGAGAQAAAA	PGAAA
Q8GR90	MH	KRRRLLAF	ATVGAVICTAGFTPSVSQA	ASSGD

Πίνακας Β_1.3: pHMM TAT.hmm

```
HMMER2.0  [2.3.2]
NAME  Tat_signals_ALL
LENG  61
ALPH  Amino
RF    no
CS    no
MAP   yes
COM   c:\Program Files\CBSU\hmmer\v2.3.2\hmmbuild.exe --archpri 0.95 tat.hmm Tat_signals_ALL.stoixisi
COM   c:\Program Files\CBSU\hmmer\v2.3.2\hmmcalibrate.exe tat.hmm
NSEQ  150
DATE  Mon Jul 05 08:40:05 2010
CKSUM 8398
XT      -8455     -4  -1000  -1000  -8455     -4  -8455     -4
NULT      -4  -8455
NULE     595  -1558     85    338   -294    453  -1158    197    249    902  -1085   -142    -21   -313     45    531    201
384  -1998   -644
EVD   -25.951990   0.271570
HMM        A      C      D      E      F      G      H      I      K      L      M      N      P      Q      R      S      T
V      W      Y
         m->m   m->i   m->d   i->m   i->i   d->m   d->d   b->m   m->e
          -10      *  -7218
     1 -10802  -9221 -10165 -10551 -10585  -9181  -9784 -11691 -10738 -10942   5406 -10579  -9617 -10630 -10149 -11377 -10953 -
11468  -9344 -10603     1
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -130 -12377  -3535   -894  -1115   -701  -1378    -10      *
     2   -680  -1295   -828    114  -1458   -833     92  -2877    629     49   -610     22    251    -75   -170   1520   1479  -
5339  -5901  -5218     2
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -228 -12247  -2774   -894  -1115  -3980    -94      *      *
     3  -1409  -5499    788    674  -1200    -17    648    795    959  -2324  -4589   1029  -1485   -129    416    414    255  -
1325   -637  -1824     3
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -235 -12019  -2738   -894  -1115  -5237    -39      *      *
     4  -2280  -5276    651    360    932  -2386   -976  -2486   1484   -599  -1111    704  -1302    570    517    318    922  -
2392   -334   -703     4
```

```
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -185 -11785  -3055   -894  -1115  -5830    -26      *      *
     5  -1464  -5104    774    152   -849   -968    724    207    839   -871    995    427   -105    775    643   -495    967  -
4725    847  -4605     5
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -123 -11600  -3618   -894  -1115  -6129    -21      *      *
     6    387  -4989   -929   -475  -1652   -130    109  -2070   1325  -1805   -832   -180   1521    359    802    286    568  -
2077  -5173  -1282     6
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -211 -11478  -2884   -894  -1115  -6281    -19      *      *
     7  -1078  -4734   -182  -2656   1076  -1211   -425  -1971    763   -126    316    693   1312    592    877    821  -1558  -
1120  -4934  -1035     7
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -229 -11268  -2770   -894  -1115  -6487    -16      *      *
     8   -502  -4584   1398  -1204  -4904    202  -2747  -1723    983  -2155   -637   -253    738    779   1510    220  -1353   -
384    270   -957     8
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -100 -11039  -3904   -894  -1115  -6656    -14      *      *
     9   -378  -4475    514    309   -260   -361   1291   -182      9    209   -441   -884    504    858  -1241      4    -19
104  -4664   -660     9
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -166 -10940  -3211   -894  -1115  -6717    -14      *      *
    10    401  -4349    922    700  -4670    151     -7  -4421   -201   -211    -46    -54    960   -545   -234    921    546  -
3971  -4532  -3849    10
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -    -95 -10775  -3988   -894  -1115  -6804    -13      *      *
    11    730   1113    -96   -236  -4571    246     37   -384    -70   -589    713   -659   1703   -520   -149   -106    229  -
1309  -4441  -3761    11
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -133 -10682  -3517   -894  -1115  -6847    -13      *      *
    12    374  -4150    480   1582  -4471  -1405  -2309  -4222    616  -4166  -3239   -480   -960   -294     -6   1485   -195
404  -4333   -511    12
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
```

```
     -    -71 -10551  -4392   -894  -1115  -6901    -12      *      *
    13   -527  -4089    905   2137  -4410   -364   1090  -4161    873  -1737  -3178   -552  -1041    930    866  -1047   -840  -
1233  -4272  -3589    13
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -    -39 -10481  -5269   -894  -1115  -6927    -12      *      *
    14     76  -3290    492   -239   -569     87  -2579   1902  -2412    210    238  -2731    488   -449   -640    675   -837   -
910  -3650  -3170    14
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -    -80 -10443  -4224   -894  -1115  -6941    -12      *      *
    15   -292  -3964    526    792    282  -1280  -2155  -4017    103  -1747  -3056  -2135   1552  -1698    214   -845   1987
141  -4154  -3478    15
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -123 -10364  -3626   -894  -1115  -6967    -12      *      *
    16    542  -3858    769    208  -4167  -1080    499   -170   -412   -274  -2950  -2031   1856  -1595    863    361   -623   -
311  -4048  -3373    16
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -2 -10243 -11285   -894  -1115  -7005    -11      *      *
    17    222  -3884   -522   1450  -4205    231   1894  -3956  -1624  -3900  -2973   -383    729     59   1444   -229    901  -
1052  -4067  -3384    17
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -222 -10243  -2820   -894  -1115  -7005    -11      *      *
    18    478  -3615    385    264   -348   -904  -1885   -866   -291   -525  -2714  -1873    741  -1438    931   1470    346   -
724  -3821  -3162    18
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -    -99 -10023  -3937   -894  -1115  -7063    -11      *      *
    19    646  -2607   -531    589    753   -845    727    335  -2334   -186    696   -235  -3634  -2205   -433    153  -2226
1452  -3007    203    19
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -244  -9927  -2692   -894  -1115  -7085    -11      *      *
    20   -452  -3072    658    -92    -96   1088  -1721  -2913  -1384  -3028  -2202  -1752  -3129  -1328  -1856    740   1065  -
2618   1564   2811    20
     -   -146   -506    232     39   -377    393    105   -618    209   -465   -726    274    399     46    101    358    116   -
372   -300   -236
     -   -945  -1059 -10727  -1919   -443    -99  -3917      *      *
```

```
    21  -2160  -1100  -1881  -1974    561  -5994   -648   1079  -1271   1624   1851    -66    599  -2277  -2549   -785    344
496   -761  -4607    29
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12298 -13340   -894  -1115    -90  -4041      *      *
    22  -2974  -5843   -587  -1830  -2465   -709    266  -5915   -877  -5859  -4932   1106  -5437    631    128   2599   1416  -
5465  -6026  -1073    30
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12377 -13419   -894  -1115   -701  -1378      *      *
    23 -10937  -9251 -10166 -10521 -10722  -9198  -9758 -11855 -10480 -11123 -11132 -10573  -9626 -10530   4276 -11530 -11042 -
11583  -9362 -10716    31
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12377 -13419   -894  -1115   -701  -1378      *      *
    24 -10937  -9251 -10166 -10521 -10722  -9198  -9758 -11855 -10480 -11123 -11132 -10573  -9626 -10530   4276 -11530 -11042 -
11583  -9362 -10716    32
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12377 -13419   -894  -1115   -701  -1378      *      *
    25     95  -5843   1737    143  -2507   -284   -192  -5915    -58  -1690  -4932    -98  -5437   1486   1434    249    948  -
2953  -6026  -5343    33
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12377 -13419   -894  -1115   -701  -1378      *      *
    26   -220  -1109  -7405  -6771   3642  -6626  -5487   -644  -6373   1479    363  -6273  -6637  -5958  -6161  -5717  -4995
63   -823  -2176    34
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12377 -13419   -894  -1115   -701  -1378      *      *
    27  -1766  -6699  -9656  -9113   -565  -9296  -8215   1367  -8892   2777   1832  -8980  -8708  -8012  -8557  -8571  -2061
-59  -7126  -7182    35
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12377 -13419   -894  -1115   -701  -1378      *      *
    28    988  -1295  -5675  -5100   -776    948  -4721  -1858   2567  -1335  -1451  -2489  -6002    993   -167   -984    329  -
1834   -818  -4764    36
     -   -143   -501    232     42   -381    398    105   -627    210   -467   -721    275    393     44     98    358    121   -
370   -295   -250
     -  -5379  -5194    -76   -581  -1592   -701  -1378      *      *
    29    816  -1777  -3801  -3805  -4211   1898  -3350  -3990  -3652  -4227  -3296  -2645  -2817  -3288  -3600   2491   1271  -
2856  -4418  -4258    40
```

```
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -8  -8095  -9138   -894  -1115  -3665   -118      *      *
    30   2349  -1773  -3832  -3223    399    460  -2232    239    -51    435   -976  -2890  -3431  -2594  -2835  -2441    880  -
1189  -2228  -1876    41
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -3  -9329 -10371   -894  -1115  -5517    -32      *      *
    31   1865  -2632  -3588  -3103  -3207   1124  -2863  -2793  -2936   -697  -2365     97  -3685  -2772  -3178    954   1882   -
246  -3576  -3184    42
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -3  -9479 -10522   -894  -1115  -3745   -112      *      *
    32   2174  -2299  -4626  -4001  -2256   1185  -2838   -314  -3632   -192    571    441  -4027  -3286  -3491     64    253   -
148  -2754    194    43
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -2 -10007 -11049   -894  -1115  -2747   -233      *      *
    33   1864   1299  -1328  -1450     85    694  -3223  -2581  -3565    234   1006  -3700  -4477   -718  -1212   -727   1100
59  -3487   -512    44
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -1 -10682 -11724   -894  -1115  -2472   -287      *      *
    34   2294    744  -3457  -2903  -4296    207    480  -1626   -658    164  -3313  -1115   -717   -307  -1241   -164    423
-58  -4482  -3943    45
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -1 -11156 -12198   -894  -1115  -2887   -210      *      *
    35   2169  -3537  -6051  -2348   -664   1586   -876   -771  -5011   -669    406  -4901  -5306  -4634  -4812   -526    199
933    154  -1018    46
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -1 -11381 -12423   -894  -1115  -2915   -205      *      *
    36   2229   -344  -6191  -5557  -3651   1576  -4282   -412  -5156    118    600  -5051  -5462  -1518  -1167    -22   -169
207  -4154  -3811    47
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -1 -11548 -12590   -894  -1115  -1876   -459      *      *
    37   2098   -347  -6467  -5831  -1768   1272  -4543   -439  -5427    -26   -170  -5317  -5722  -1863  -2084    132   1027
656  -4410  -4068    48
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
```

```
     -     -1 -11817 -12859   -894  -1115   -851  -1167      *      *
    38   2573  -4278  -6796  -6160   -641    941  -4870  -1127  -5755   -164  -3481  -5644  -2441  -5378  -5555    500    617
1  -4736  -1843    49
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12153 -13195   -894  -1115  -1694   -534      *      *
    39   2137    450  -6852  -6216  -1299   1816  -4932  -2480  -1892   -125    663  -5705  -6111  -2323  -5616    368     62
99  -4800  -4458    50
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12221 -13263   -894  -1115  -1251   -786      *      *
    40   2191   -278  -6922  -6286  -1283   1503  -4996  -2459  -5881    366   -465  -2646  -6175  -5504  -2731    101    739
303   -703  -4520    51
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12285 -13328   -894  -1115   -414  -2003      *      *
    41   2220  -4475  -6992  -2961  -1967   1414  -5067  -1775  -5952    663   1209  -5841  -6246  -5575  -5752    -56    150
416  -4933  -4591    52
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12356 -13398   -894  -1115   -251  -2647      *      *
    42   2325  -1208  -7012  -6376  -2247   1131  -5085  -1253  -5971    619    600  -5860  -2885  -5593  -5770    169    -27
425   -878  -1940    53
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12377 -13419   -894  -1115   -701  -1378      *      *
    43   2050  -1286  -7012  -6375  -2218   1565  -1582  -1262  -5970   1001   -164  -2585  -1643  -5593  -5770   -232   -153
168   -680  -1985    54
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12377 -13419   -894  -1115   -701  -1378      *      *
    44   2078  -1286  -7009  -6373  -1416   1966  -5084   -894  -5968   -540    589  -1777  -1375  -1240  -5769   -339    236
197   -843  -4608    55
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12377 -13419   -894  -1115   -701  -1378      *      *
    45   1773  -1100  -2782  -6368  -2247   1171  -1544  -1321  -5965    921    -74  -2697    415  -5589  -5767    313    472
283  -4951  -1966    56
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12377 -13419   -894  -1115   -701  -1378      *      *
```

```
    46   1963  -4519  -2795  -6159   -871    910  -1540     50  -5802    922  -1529  -2750   -165  -2438  -2875    330     52
196  -4973  -4625    57
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12377 -13419   -894  -1115   -701  -1378      *      *
    47   2083  -4493  -7010  -2996    615   -324  -5085   -718  -5969   1535   1377  -5859  -6263  -2121  -2711  -2258  -1044
823   -818  -4608    58
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12377 -13419   -894  -1115   -701  -1378      *      *
    48   2090  -4566  -7089  -6453    252  -1240  -5163  -1272  -6049   1879   1382  -5940  -6337  -5666  -5847  -5380  -4683
711   1695  -1151    59
     -   -149   -502    233     41   -383    403    103   -628    208   -461   -723    276    399     43     94    359    117   -
371   -297   -252
     -  -4462  -4322   -145  -1089   -917   -701  -1378      *      *
    49    190   2088  -4619  -4431  -4063    114  -3779  -3707  -4168  -4008  -3225  -3399   2868  -3833  -4118   1055   1356
163  -4413  -4156    65
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -4  -8988 -10030   -894  -1115  -4564    -62      *      *
    50  -1838  -2385  -2307  -1749  -2448    870    943  -2037    764    685  -1559  -1928     58  -1524  -1992    531   1303
1227  -2759  -2291    66
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -3  -9411 -10453   -894  -1115  -4080    -88      *      *
    51    147  -3109   1468  -1714    -53   1264    785  -2910  -1616    117  -2261  -1973   1455  -1554  -2074    619    353  -
2647  -3420  -2862    67
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -2  -9855 -10897   -894  -1115  -3464   -137      *      *
    52    120    585  -2370  -1821   -655    899   1326  -1127    212  -3918  -3008    364    637   -159   1395   1099     64  -
1362  -4111  -3443    68
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -2 -10340 -11382   -894  -1115  -2874   -212      *      *
    53    808    176  -3049  -2497  -1058   1189  -2719   -552    186  -1978  -3121   -817   1113     19   -125   1245    815  -
1509  -4275  -3705    69
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -1 -10829 -11871   -894  -1115  -2639   -253      *      *
    54    424   1452  -1433  -1557   -303   1816  -3251  -3673  -3070   -116  -3165  -3394    795   -244   -765    123   1390  -
1644    154   -986    70
```

```
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -1 -11203 -12245   -894  -1115  -2305   -326      *      *
    55   -307  -4010  -4583  -2038    571   1030    619   -530  -3850   -311   -844   -242    766   -152    -17    732   1449   -
631    877  -4000    71
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -1 -11519 -12561   -894  -1115   -895  -1114      *      *
    56   1203  -1019  -4059  -2280  -5476   -474  -1185  -2612  -2188  -1346  -1347  -1974   2476    317   -838    989    713
92  -5515  -4905    72
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -1 -12040 -13083   -894  -1115   -827  -1197      *      *
    57   -140  -5708  -1632   -866  -2411   1282     19  -1812   -825  -1093   -196    419   1316    584    695   1078    327  -
2949   -782  -5210    73
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12239 -13281   -894  -1115   -659  -1447      *      *
    58    856  -5775  -1684   -971  -1619   -762    -50   -722    142  -1122  -1826    311   1500   -407    365    486    831
166     50  -2160    74
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12328 -13370   -894  -1115   -133  -3503      *      *
    59   3244  -5154  -5404  -1929  -5343  -1030  -4801  -4927  -4726  -1097  -4476  -4970  -2723  -2273  -1895   -179  -1820
89  -5682  -5240    75
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12377 -13419   -894  -1115   -701  -1378      *      *
    60    348  -1436   -657  -1042    820   -732   1800  -1405   -677   -391   -797    100    -36   1018    582    395  -1077   -
632   1881     28    76
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12377 -13419   -894  -1115   -701  -1378      *      *
    61   3652  -6123  -2557  -7501  -8763  -2254  -7663  -8606  -8278  -8830  -7887  -6732  -7096  -2366  -8237  -3036  -2881  -
7305  -8969  -8805    77
     -      *      *      *      *      *      *      *      *      *      *      *      *      *      *      *      *      *
*      *      *
     -      *      *      *      *      *      *      *      *      0
```

Πίνακας Β_1.4: pHMM SEC.hmm

```
HMMER2.0  [2.3.2]
NAME  Apla_signals_ALL
LENG  55
ALPH  Amino
RF    no
CS    no
MAP   yes
COM   c:\Program Files\CBSU\hmmer\v2.3.2\hmmbuild.exe --archpri 0.95 signal.hmm Apla_signals_ALL.stoixisi
COM   c:\Program Files\CBSU\hmmer\v2.3.2\hmmcalibrate.exe signal.hmm
NSEQ  328
DATE  Mon Jul 05 08:39:00 2010
CKSUM 5695
XT      -8455     -4  -1000  -1000  -8455     -4  -8455     -4
NULT      -4  -8455
NULE     595  -1558     85    338   -294    453  -1158    197    249    902  -1085   -142    -21   -313     45    531    201
384  -1998   -644
EVD   -13.921825   0.294921
HMM        A      C      D      E      F      G      H      I      K      L      M      N      P      Q      R      S      T
V      W      Y
         m->m   m->i   m->d   i->m   i->i   d->m   d->d   b->m   m->e
           -5      *  -8307
     1 -12042 -10349 -11274 -11662 -11791 -10299 -10905 -12941 -11855 -12195   5406 -11716 -10732 -11754 -11261 -12652 -12166 -
12686 -10463 -11804     1
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -13472 -14514   -894  -1115   -701  -1378     -5      *
     2  -2203  -6901  -2748  -1150   -586  -2297   -288    101   2641  -1198     99   1180   -535   -811   1099   -321   -833  -
1490  -7087  -2404     2
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -    -22 -13472  -6033   -894  -1115   -701  -1378      *      *
     3  -1894  -1425  -1891  -2874   -660  -2853   -401    -75   2444   -498    790    846  -1391   -289   1263   -393   -262  -
1176  -7026   -377     3
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -203 -13450  -2934   -894  -1115  -2801   -224      *      *
```

```
     4   -920  -1987  -3262  -3338    537  -1860   -296   -352   1827   -968   -322    836  -1443   -106   1163    512    952  -
1297   -859   -305     4
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -290 -13247  -2457   -894  -1115  -5705    -28      *      *
     5   -783  -6148  -4945  -2387    -70  -2216    670  -1662   2031    -78    -75   1162  -1848   -486   1056    -14    717   -
642  -1529   -836     5
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -183 -12957  -3074   -894  -1115  -6740    -14      *      *
     6  -1226  -1596  -2878  -1691    703  -1402  -1946   -326   1840    -81  -5220  -1911   -268   -632   1534     62    771   -
360    647   -854     6
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -227 -12775  -2782   -894  -1115  -7088    -11      *      *
     7   -126    417  -2672  -2749    589   -727   -484   -112   1336   -606   -914    677   -822   -983   1302   -341    772    -
75  -6057   -108     7
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -362 -12548  -2174   -894  -1115  -7388     -9      *      *
     8   -109   -194  -2441  -2320   -780  -1997  -1384    167   1750    -24  -1415   -172    102   -349   1835   -230   -281  -
1079    244  -1091     8
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -331 -12187  -2288   -894  -1115  -7704     -7      *      *
     9  -1783    130  -2093  -3321  -1996    603   -121   -980   1404   -446  -1032   -112    217   -840   1953    783    -45  -
2570   -530   -786     9
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -354 -11857  -2202   -894  -1115  -7894     -6      *      *
    10   -572   -734  -3905  -3350    290     10   -661  -1924   1425  -1710    613   -790   -947    -62   2081    354    141   -
753  -4716    743    10
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -478 -11504  -1829   -894  -1115  -8038     -5      *      *
    11   -145     95  -1213  -2422    399    255   -160  -1523   1091   -829  -3087  -1077    105    432    960    704    261   -
331    469    194    11
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -218 -11027  -2841   -894  -1115  -8170     -5      *      *
    12    578    241  -4138  -3565    192   -284    -16   -364   1217   -943    -50  -3584    878   -777   1569    413   -390  -
1099  -3690    232    12
```

```
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -247 -10810  -2670   -894  -1115  -8214     -5      *      *
    13  -1299  -4145   -795   -748   -864   -492  -2326   -957   1656   -417    120   -585   -258   -376   1621    871    290   -
519  -4333  -3655    13
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -334 -10564  -2278   -894  -1115  -8256     -5      *      *
    14    235  -3718  -2336  -1787  -3972    321    454  -1043   1762  -1611  -2826  -2084   -427    648    755    354    213
754  -3943    -21    14
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -473 -10232  -1843   -894  -1115  -8301     -5      *      *
    15   -497  -3226  -2010  -1459    795   -760  -1732    343   1124  -3196  -2344     86   -350  -1316   2258    218    664   -
2790   1484  -2864    15
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -343  -9761  -2250   -894  -1115  -8347     -4      *      *
    16    533  -3252  -1766  -1207    141   -482  -1488  -3297   1696  -3265  -2366    589  -2937  -1040   2555  -1768    774   -
2882  -3446  -2804    16
     -   -146   -501    231     41   -382    397    110   -620    213   -468   -722    274    392     51     97    357    118   -
368   -296   -251
     -  -4586   -626  -1688    -15  -6566  -8372     -4      *      *
    17    168  -1696  -2454  -1880  -1699  -2807  -1578    733    444    -82   -883  -1922    299  -1544   2207  -1812   1175
440  -2111  -1707    18
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -892  -8887  -1124   -894  -1115  -8400     -4      *      *
    18  -1408  -2626  -1393   -743  -3023  -2307   -764  -2658   1437    285  -1739   1668   1102   -329   2166  -1307  -1300   -
2299  -2683  -2187    19
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -   -226  -8003  -2823   -894  -1115  -8429     -4      *      *
    19   -771  -1822   -833   -288  -2061  -1797   -506  -1717   -123    335  -1004   1687  -1906   -114   1316    848   1174   -
1414  -2136  -1564    20
     -   -149   -500    233     43   -381    398    105   -627    212   -466   -721    275    393     45    101    359    117   -
370   -295   -250
     -  -2951  -1022  -1402    -60  -4624  -8433     -4      *      *
    20  -1323  -2131  -1662   -817  -2555  -2160   -424  -1999   2153  -2019  -1297   -894  -2181    -28   2238  -1290  -1162
1447  -2152  -1831    22
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
```

```
     -  -1128  -7106   -902   -894  -1115  -8444     -4      *      *
    21  -1135  -1575  -1380   -857  -2018  -1595   -410  -1891    588  -1820  -1278   -856  -1836   -120   3398  -1179  -1102  -
1665  -1711  -1485    23
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -  -1175  -6011   -884   -894  -1115  -6793    -13      *      *
    22  -1675  -1417  -2952  -2621    780  -2882   -557   -390  -2192   1820     15  -2068  -2844  -1802  -2088  -2065  -1610   -
623      1   3302    24
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -    -22  -6644  -7686   -894  -1115  -7275     -9      *      *
    23   -903   -804  -2749  -2209   -662  -2410  -1413   2104  -1901   1314    228  -1946  -2478  -1640  -1897   1288   -878
361  -1399  -1051    25
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -    -15  -7136  -8178   -894  -1115  -6620    -15      *      *
    24  -1169  -1033  -3196  -2600   -664    957  -1476   -324  -2214    964   3470  -2220  -2700  -1890  -2108  -1761  -1114   -
373   3234   -910    26
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -    -10  -7801  -8843   -894  -1115  -7308     -9      *      *
    25   1585  -1171  -1924  -1350  -1184  -2264  -1043   1148  -1092    711   -360   1132  -2340   -998   1045  -1280   -898   -
585  -1597  -1195    27
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -8  -8029  -9071   -894  -1115  -5975    -23      *      *
    26   1134  -1228  -3620  -2993   2319  -2926  -1794    882    322  -1070   1948  -2534  -2976  -2259  -2455  -2006    658
920  -1692  -1347    28
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -5  -8666  -9708   -894  -1115  -5735    -27      *      *
    27    687  -1658  -4178  -3546   1997  -3394  -2271   1151  -3146   1577   -840  -3038    430  -2769  -2950  -2480    258
407  -2131  -1791    29
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -4  -9181 -10223   -894  -1115  -5031    -45      *      *
    28    690  -2085  -4591  -3956    110  -3803  -2674   1168   -576   1537  -1288  -3446  -3853  -3178  -3357    647   -281
1102   1736  -2201    30
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -2  -9781 -10823   -894  -1115  -4613    -60      *      *
```

```
    29    846  -2557  -5056  -4421    503   -349    442   1007  -4020   1412  -1760  -3914  -4324  -3646   -784    972   -752
469   1298  -2672    31
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -2 -10319 -11362   -894  -1115  -3704   -115      *      *
    30    488  -3162  -5677  -5041   1001   -914  -3753    949  -1639   1349   -128  -4527  -4932  -4260   -440   -908     -7
1159   2876  -3277    32
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -1 -10982 -12024   -894  -1115  -3532   -130      *      *
    31    870   -225  -6122  -5486     -6  -2307  -4196   1273  -2081   1674   -495  -4970   -914  -4704  -4881   -334    411
965  -4061   -114    33
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -1 -11452 -12495   -894  -1115  -2956   -199      *      *
    32    908    124  -6571  -5935   1324   -777  -4645   1220  -1595   1370   -442  -5419  -5823  -5153  -2282   -298    322
762  -4510   -202    34
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -1 -11922 -12964   -894  -1115  -2825   -220      *      *
    33   1360   -198  -6898  -6262   -132   -928  -4973    498  -2859   1319    -97  -1743  -1205  -5481  -2611    699    476
798    320  -4497    35
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12262 -13304   -894  -1115  -2441   -294      *      *
    34   1091     94  -7198  -6562    739    148  -1850    695  -6157   1647   -194  -1928  -1546  -2627  -2910    135  -1009
587   -927  -1919    36
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12569 -13611   -894  -1115  -1879   -458      *      *
    35   1493    564  -7495  -6859     33   -289  -1191    766  -6454   1452    403  -2121  -1724  -6077  -6254   -124    -67
624    665  -5092    37
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12872 -13914   -894  -1115  -1420   -675      *      *
    36   1427    931  -7743  -3530    144    185  -1403    539  -6702   1340    572  -2356  -1612  -6325  -6502   -583   -321
755   1398  -1874    38
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -13124 -14166   -894  -1115  -1392   -692      *      *
    37   1412     31  -7882  -7246      7   -796  -2400    708  -6840   1096   1322  -6730  -2139  -2628  -6640    456    277
1173  -1701  -2624    39
```

```
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -13265 -14307   -894  -1115  -1047   -954      *      *
    38   1679   -228  -3574  -3794    -20   -476  -6056    536  -6942   1573   -253  -3471  -1938  -6565  -6742     54    113
629  -1835  -2957    40
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -13368 -14410   -894  -1115  -1025   -976      *      *
    39   1436     87  -8034  -7398    296   -145  -6107   1048  -6993   1291    220  -3710  -1219  -6615  -6792   -173    198
839   -364  -2052    41
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -13419 -14461   -894  -1115   -280  -2503      *      *
    40   1520   -364  -4060  -7442    741   -618  -2875    398  -7037   1533    315  -3405  -1137  -6660  -6837   -642    395
859  -1897  -2996    42
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -13464 -14506   -894  -1115   -283  -2490      *      *
    41   1483   -592  -3914  -4020    386    373  -6159    721  -7044   1065    915  -3607  -1535  -6667  -6844    424    314
650  -6024  -5682    43
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -13472 -14514   -894  -1115   -701  -1378      *      *
    42   1590    873  -8086  -7450    522    306  -6159    283  -7045   1102    634  -2111  -1950  -3356  -6844    589    403
177  -6024  -3024    44
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -13472 -14514   -894  -1115   -701  -1378      *      *
    43   1291    428  -2918  -3894   1177   -157  -6153   -595  -7024   1260    627  -2140   -832  -3350  -3694    621    461
395  -6027  -2282    45
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -13472 -14514   -894  -1115   -701  -1378      *      *
    44   1067    167  -8086  -7450    670    103  -6159    252  -2978   1543    712  -3599   -232  -3521  -6844    356    267
263  -1877  -3059    46
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -13472 -14514   -894  -1115   -701  -1378      *      *
    45   2064  -2248  -8086  -7450   1548  -3397  -2761    102  -7045   1385    674  -6934  -2218  -6667  -6844  -2437  -1968
1185    288  -3019    47
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
```

```
     -      0 -13472 -14514   -894  -1115   -701  -1378      *      *
    46   1895  -2275  -8090  -7454   1651  -4186  -6164    699  -7049   1477   1014  -6938  -7342  -6672  -6849  -4127  -3751
1326   -737  -5687    48
     -   -151   -493    229     39   -387    402     99   -629    204   -466   -727    274    387     54     95    375    121   -
373   -301   -250
     -  -3955  -3758   -215   -931  -1073   -701  -1378      *      *
    47    845   1317   -784  -2367  -4130   1320    378  -3790  -2273    715  -3080   -759   1141  -2215  -2742   1257    295  -
3483  -4226  -3645    55
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -1 -10619 -11661   -894  -1115  -2868   -212      *      *
    48  -1335    516  -5560  -2030  -3690   1949   -648   -425  -4668   -714    779  -1464   -346  -1614  -4668   1576   1404   -
598  -4171   -453    56
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -     -1 -11496 -12538   -894  -1115  -2281   -332      *      *
    49   -697  -1057  -1539  -2696    153   1171  -1404   -296  -2587   -611  -1327    369    711   -222  -4730   1627    782   -
456   -916    -77    57
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12171 -13213   -894  -1115  -1384   -697      *      *
    50    113    475  -1755  -2296  -2917    405  -1052   -308  -1645  -2523     60  -1206   1442   -816  -4906   2003   1186   -
891  -1491  -1083    58
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -12811 -13853   -894  -1115   -883  -1127      *      *
    51     91  -2183  -1903  -2156  -1117   1183    446  -2345  -1724  -1336  -5725    474   1026   1019  -1528   1333   1104   -
587  -6821  -2900    59
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -13205 -14247   -894  -1115   -362  -2171      *      *
    52    425  -6826  -1045  -2046  -1048   -242  -1598  -1124  -1259  -1328   -678    283   1734    255  -1816   1272   1209    -
48  -7018   -896    60
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -13418 -14460   -894  -1115    -64  -4535      *      *
    53   3007   -352  -3978  -2530  -1963  -1809  -1185  -1528  -6316  -1194  -4903  -1641  -7173  -2517  -2784   -143   -975
1081  -6145  -5773    61
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -13472 -14514   -894  -1115   -701  -1378      *      *
```

```
    54    485  -6854  -1689   -779   1335   -939   2031  -1840   -776    362   1209   -110  -2213   1490  -1632    313   -995  -
1398    736    283    62
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -
369   -294   -249
     -      0 -13472 -14514   -894  -1115   -701  -1378      *      *
    55   3611  -7150  -8912  -4097  -9591  -1787  -8577  -9348  -4276  -9558  -8662  -7952  -8175  -8501  -3727  -1585  -2960  -
2658  -9721  -3109    63
     -      *      *      *      *      *      *      *      *      *      *      *      *      *      *      *      *      *
*      *      *
     -      *      *      *      *      *      *      *      *      0
```

Παράρτημα Γ

Αναφορές

Aldridge, C., E. Spence, et al. (2008). "Tat-dependent targeting of Rieske iron-sulphur proteins to both the plasma and thylakoid membranes in the cyanobacterium Synechocystis PCC6803." Mol Microbiol **70**(1): 140-150.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-3402.

Apel, A. K., A. Sola-Landa, et al. (2007). "Phosphate control of phoA, phoC and phoD gene expression in Streptomyces coelicolor reveals significant differences in binding of PhoP to their promoter regions." Microbiology **153**(Pt 10): 3527-3537.

Bachmann, J., B. Bauer, et al. (2006). "The Rieske protein from Paracoccus denitrificans is inserted into the cytoplasmic membrane by the twin-arginine translocase." FEBS J **273**(21): 4817-4830.

Bachmann , J., B. Brigitte, et al. (2006). "The Rieske protein from Paracoccus denitrificans is inserted into the cytoplasmic membrane by the twin-arginine translocase." FEBS J **273**: 4817-4830.

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

Bagos, P. G., E. P. Nikolaou, et al. (2010). "Combined prediction of Tat and Sec signal peptides with Hidden Markov Models." Bioinformatics.

Bagos, P. G., K. D. Tsirigos, et al. (2008). "Prediction of lipoprotein signal peptides in Gram-positive bacteria with a Hidden Markov Model." J Proteome Res **7**(12): 5082-5093.

Bagos, P. G., K. D. Tsirigos, et al. (2009). "Prediction of signal peptides in archaea." Protein Eng Des Sel **22**(1): 27-35.

Baldi, P., S. Brunak, et al. (2000). "Assessing the accuracy of prediction algorithms for classification: an overview." Bioinformatics **16**(5): 412-424.

Baliga, N. S., R. Bonneau, et al. (2004). "Genome sequence of Haloarcula marismortui: a halophilic archaeon from the Dead Sea." Genome Res **14**(11): 2221-2234.

Bendtsen, J. D., H. Nielsen, et al. (2004). "Improved prediction of signal peptides: SignalP 3.0." J Mol Biol **340**(4): 783-795.

Bendtsen, J. D., H. Nielsen, et al. (2005). "Prediction of twin-arginine signal peptides." BMC Bioinformatics **6**: 167.

Berks, B. C. (1996). "A common export pathway for proteins binding complex redox cofactors?" Mol Microbiol **22**(3): 393-404.

Berks, B. C., T. Palmer, et al. (2005). "Protein targeting by the bacterial twin-arginine translocation (Tat) pathway." Curr Opin Microbiol **8**(2): 174-181.

Berks, B. C., F. Sargent, et al. (2000). "The Tat protein export pathway." Mol Microbiol **35**(2): 260-274.

Bidwell, G. L., 3rd and D. Raucher (2009). "Therapeutic peptides for cancer therapy. Part I - peptide inhibitors of signal transduction cascades." Expert Opin Drug Deliv **6**(10): 1033-1047.

Bolhuis, A. (2002). "Protein transport in the halophilic archaeon Halobacterium sp. NRC-1: a major role for the twin-arginine translocation pathway?" Microbiology **148**(Pt 11): 3335-3346.

Buchanan, G., F. Sargent, et al. (2001). "A genetic screen for suppressors of Escherichia coli Tat signal peptide mutations establishes a critical role for the second arginine within the twin-arginine motif." Arch Microbiol **177**(1): 107-112.

Caspers, M., U. Brockmeier, et al. (2010). "Improvement of Sec-dependent secretion of a heterologous model protein in Bacillus subtilis by saturation mutagenesis of the N-domain of the AmyE signal peptide." Appl Microbiol Biotechnol **86**(6): 1877-1885.

Chen, C. P. and B. Rost (2002). "Long membrane helices and short loops predicted less accurately." Protein Sci **11**(12): 2766-2773.

Cline, K., W. F. Ettinger, et al. (1992). "Protein-specific energy requirements for protein transport across or into thylakoid membranes. Two lumenal proteins are transported in the absence of ATP." J Biol Chem **267**(4): 2688-2696.

Creighton, A. M., A. Hulford, et al. (1995). "A monomeric, tightly folded stromal intermediate on the delta pH-dependent thylakoidal protein transport pathway." J Biol Chem **270**(4): 1663-1669.

Cristobal, S., J. W. de Gier, et al. (1999). "Competition between Sec- and TAT-dependent protein translocation in Escherichia coli." EMBO J **18**(11): 2982-2990.

de Leeuw, E., T. Granjon, et al. (2002). "Oligomeric properties and signal peptide binding by Escherichia coli Tat protein transport complexes." J Mol Biol **322**(5): 1135-1146.

Delahanty, R. J., J. Q. Kang, et al. (2009). "Maternal transmission of a rare GABRB3 signal peptide variant is associated with autism." Mol Psychiatry.

DeLisa, M. P., P. Samuelson, et al. (2002). "Genetic analysis of the twin arginine translocator secretion pathway in bacteria." J Biol Chem **277**(33): 29825-29831.

DeLisa, M. P., D. Tullman, et al. (2003). "Folding quality control in the export of proteins by the bacterial twin-arginine translocation pathway." Proc Natl Acad Sci U S A **100**(10): 6115-6120.
Dilks, K., M. I. Gimenez, et al. (2005). "Genetic and biochemical analysis of the twin-arginine translocation pathway in halophilic archaea." J Bacteriol **187**(23): 8104-8113.
Dilks, K., R. W. Rose, et al. (2003). "Prokaryotic utilization of the twin-arginine translocation pathway: a genomic survey." J Bacteriol **185**(4): 1478-1483.
Driessen, A. J. and N. Nouwen (2008). "Protein translocation across the bacterial cytoplasmic membrane." Annu Rev Biochem **77**: 643-667.
Durbin, R., S. R. Eddy, et al. (1998). Biological Sequence Analysis. , Cambridge University Press.
Eddy, S. R. (1998). "Profile hidden Markov models." Bioinformatics **14**(9): 755-763.
Falb, M., F. Pfeiffer, et al. (2005). "Living with two extremes: conclusions from the genome sequence of Natronomonas pharaonis." Genome Res **15**(10): 1336-1343.
Gardy, J. L., M. R. Laird, et al. (2005). "PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis." Bioinformatics **21**(5): 617-623.
Gardy, J. L., C. Spencer, et al. (2003). "PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria." Nucleic Acids Res **31**(13): 3613-3617.
Gimenez, M. I., K. Dilks, et al. (2007). "Haloferax volcanii twin-arginine translocation substates include secreted soluble, C-terminally anchored and lipoproteins." Mol Microbiol **66**(6): 1597-1606.
Habib, S. J., W. Neupert, et al. (2007). "Analysis and prediction of mitochondrial targeting signals." Methods Cell Biol **80**: 761-781.
Haft, D. H., J. D. Selengut, et al. (2003). "The TIGRFAMs database of protein families." Nucleic Acids Res **31**(1): 371-373.
Hatzixanthis, K., T. Palmer, et al. (2003). "A subset of bacterial inner membrane proteins integrated by the twin-arginine translocase." Mol Microbiol **49**(5): 1377-1390.
Hiller, K., A. Grote, et al. (2004). "PrediSi: prediction of signal peptides and their cleavage positions." Nucleic Acids Res **32**(Web Server issue): W375-379.
Hinsley, A. P., N. R. Stanley, et al. (2001). "A naturally occurring bacterial Tat signal peptide lacking one of the 'invariant' arginine residues of the consensus targeting motif." FEBS Lett **497**(1): 45-49.
Hobohm, U., M. Scharf, et al. (1992). "Selection of representative protein data sets." Protein Sci **1**(3): 409-417.
Ignatova, Z., C. Hornle, et al. (2002). "Unusual signal peptide directs penicillin amidase from Escherichia coli to the Tat translocation machinery." Biochem Biophys Res Commun **291**(1): 146-149.
Ikeda, M., M. Arai, et al. (2003). "TMPDB: a database of experimentally-characterized transmembrane topologies." Nucleic Acids Res **31**(1): 406-409.
Inouye, S., S. Wang, et al. (1977). "Amino acid sequence for the peptide extension on the prolipoprotein of the Escherichia coli outer membrane." Proc Natl Acad Sci U S A **74**(3): 1004-1008.
Ize, B., F. Gerard, et al. (2002). "In vivo assessment of the Tat signal peptide specificity in Escherichia coli." Arch Microbiol **178**(6): 548-553.
Jayasinghe, S., K. Hristova, et al. (2001). "MPtopo: A database of membrane protein topology." Protein Sci **10**(2): 455-458.
Jongbloed, J. D., H. Antelmann, et al. (2002). "Selective contribution of the twin-arginine translocation pathway to protein secretion in Bacillus subtilis." J Biol Chem **277**(46): 44068-44078.
Jongbloed, J. D., U. Martin, et al. (2000). "TatC is a specificity determinant for protein secretion via the twin-arginine translocation pathway." J Biol Chem **275**(52): 41350-41357.
Juncker, A. S., H. Willenbrock, et al. (2003). "Prediction of lipoprotein signal peptides in Gram-negative bacteria." Protein Sci **12**(8): 1652-1662.
Kall, L., A. Krogh, et al. (2004). "A combined transmembrane topology and signal peptide prediction method." J Mol Biol **338**(5): 1027-1036.
Kall, L., A. Krogh, et al. (2007). "Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server." Nucleic Acids Res **35**(Web Server issue): W429-432.
Krogh , A. (1994). Hidden Markov Models for labelled sequences. . Proceedings of the 12th IAPR International Conference on Pattern Recognition, Israel.
Krogh, A., B. Larsson, et al. (2001). "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes." J Mol Biol **305**(3): 567-580.
Lee, P. A., D. Tullman-Ercek, et al. (2006). "The bacterial twin-arginine translocation pathway." Annu Rev Microbiol **60**: 373-395.
Litou, Z. I., P. G. Bagos, et al. (2008). "Prediction of cell wall sorting signals in gram-positive bacteria with a hidden markov model: application to complete genomes." J Bioinform Comput Biol **6**(2): 387-401.
Melen, K., A. Krogh, et al. (2003). "Reliability measures for membrane protein topology prediction algorithms." J Mol Biol **327**(3): 735-744.

Menne, K. M., H. Hermjakob, et al. (2000). "A comparison of signal sequence prediction methods using a test set of signal peptides." Bioinformatics **16**(8): 741-742.
Moller, S., E. V. Kriventseva, et al. (2000). "A collection of well characterised integral membrane proteins." Bioinformatics **16**(12): 1159-1160.
Mould, R. M. and C. Robinson (1991). "A proton gradient is required for the transport of two lumenal oxygen-evolving proteins across the thylakoid membrane." J Biol Chem **266**(19): 12189-12193.
Nakai, K. and M. Kanehisa (1991). "Expert system for predicting protein localization sites in gram-negative bacteria." Proteins **11**(2): 95-110.
Ng, W. V., S. P. Kennedy, et al. (2000). "Genome sequence of Halobacterium species NRC-1." Proc Natl Acad Sci U S A **97**(22): 12176-12181.
Nielsen, H., S. Brunak, et al. (1999). "Machine learning approaches for the prediction of signal peptides and other protein sorting signals." Protein Eng **12**(1): 3-9.
Nielsen, H., J. Engelbrecht, et al. (1997). "Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites." Protein Eng **10**(1): 1-6.
Nielsen, H., J. Engelbrecht, et al. (1997). "A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites." Int J Neural Syst **8**(5-6): 581-599.
Nielsen, H. and A. Krogh (1998). "Prediction of signal peptides and signal anchors by a hidden Markov model." Proc Int Conf Intell Syst Mol Biol **6**: 122-130.
Palmer, T., F. Sargent, et al. (2005). "Export of complex cofactor-containing proteins by the bacterial Tat pathway." Trends Microbiol **13**(4): 175-180.
Plewczynski, D., L. Slabinski, et al. (2008). "Prediction of signal peptides in protein sequences by neural networks." Acta Biochim Pol **55**(2): 261-267.
Plewczynski, D., A. Tkacz, et al. (2008). "AutoMotif Server for prediction of phosphorylation sites in proteins using support vector machine: 2007 update." J Mol Model **14**(1): 69-76.
Pop, O., U. Martin, et al. (2002). "The twin-arginine signal peptide of PhoD and the TatAd/Cd proteins of Bacillus subtilis form an autonomous Tat translocation system." J Biol Chem **277**(5): 3268-3273.
Reynolds, S. M., L. Kall, et al. (2008). "Transmembrane topology and signal peptide prediction using dynamic bayesian networks." PLoS Comput Biol **4**(11): e1000213.
Rose, R. W., T. Bruser, et al. (2002). "Adaptation of protein secretion to extremely high-salt conditions by extensive use of the twin-arginine translocation pathway." Mol Microbiol **45**(4): 943-950.
Sankaran, K., S. D. Gupta, et al. (1995). "Modification of bacterial lipoproteins." Methods Enzymol **250**: 683-697.
Sankaran, K. and H. C. Wu (1994). "Lipid modification of bacterial prolipoprotein. Transfer of diacylglyceryl moiety from phosphatidylglycerol." J Biol Chem **269**(31): 19701-19706.
Sargent, F., E. G. Bogsch, et al. (1998). "Overlapping functions of components of a bacterial Sec-independent protein export pathway." EMBO J **17**(13): 3640-3650.
Schneider, T. D. and R. M. Stephens (1990). "Sequence logos: a new way to display consensus sequences." Nucleic Acids Res **18**(20): 6097-6100.
Schreiber, M. and C. Brown (2002). "Compensation for nucleotide bias in a genome by representation as a discrete channel with noise." Bioinformatics **18**(4): 507-512.
Shruthi, H., P. Anand, et al. (2010). "Twin arginine translocase pathway and fast-folding lipoprotein biosynthesis in E. coli: interesting implications and applications." Mol Biosyst **6**(6): 999-1007.
Stanley, N. R., T. Palmer, et al. (2000). "The twin arginine consensus motif of Tat signal peptides is involved in Sec-independent protein targeting in Escherichia coli." J Biol Chem **275**(16): 11591-11596.
Stephenson, K. (2005). "Sec-dependent protein translocation across biological membranes: evolutionary conservation of an essential protein transport pathway (review)." Mol Membr Biol **22**(1-2): 17-28.
Teter, S. A. and D. J. Klionsky (1999). "How to get a folded protein across a membrane." Trends Cell Biol **9**(11): 428-431.
Thomas, J. R. and A. Bolhuis (2006). "The tatC gene cluster is essential for viability in halophilic archaea." FEMS Microbiol Lett **256**(1): 44-49.
Tuteja, R. (2005). "Type I signal peptidase: an overview." Arch Biochem Biophys **441**(2): 107-111.
van Roosmalen, M. L., N. Geukens, et al. (2004). "Type I signal peptidases of Gram-positive bacteria." Biochim Biophys Acta **1694**(1-3): 279-297.
von Heijne, G. (1986). "A new method for predicting signal sequence cleavage sites." Nucleic Acids Res **14**(11): 4683-4690.
von Heijne, G. (1990). "The signal peptide." J Membr Biol **115**(3): 195-201.
von Heijne, G., J. Steppuhn, et al. (1989). "Domain structure of mitochondrial and chloroplast targeting peptides." Eur J Biochem **180**(3): 535-545.
Weiner, J. H., P. T. Bilous, et al. (1998). "A novel and ubiquitous system for membrane targeting and secretion of cofactor-containing proteins." Cell **93**(1): 93-101.

Wickner, W. and R. Schekman (2005). "Protein translocation across biological membranes." Science **310**(5753): 1452-1456.
Widdick, D. A., K. Dilks, et al. (2006). "The twin-arginine translocation pathway is a major route of protein export in Streptomyces coelicolor." Proc Natl Acad Sci U S A **103**(47): 17927-17932.
Widdick, D. A., R. T. Eijlander, et al. (2008). "A facile reporter system for the experimental identification of twin-arginine translocation (Tat) signal peptides from all kingdoms of life." J Mol Biol **375**(3): 595-603.
Wu, C. H., R. Apweiler, et al. (2006). "The Universal Protein Resource (UniProt): an expanding universe of protein information." Nucleic Acids Res **34**(Database issue): D187-191.
Yikmis, M., M. Arenskotter, et al. (2008). "Secretion and transcriptional regulation of the latex-clearing protein, Lcp, by the rubber-degrading bacterium Streptomyces sp. strain K30." Appl Environ Microbiol **74**(17): 5373-5382.
Zhang, Z. and W. J. Henzel (2004). "Signal peptide prediction based on analysis of experimentally verified cleavage sites." Protein Sci **13**(10): 2819-2824.
Zhang, Z. and W. I. Wood (2003). "A profile hidden Markov model for signal peptides generated by HMMER." Bioinformatics **19**(2): 307-308.