Κεφάλαιο 5: Αναζήτηση Προτύπων σε Αλληλουχίες

Σύνοψη
Στο κεφάλαιο αυτό θα μελετήσουμε τα πρότυπα αλληλουχιών και θα εξετάσουμε τη χρησιμότητά τους. Θα δούμε τον τρόπο ορισμού των προτύπων της PROSITE και τη σχέση τους με τα πρότυπα κανονικών εκφράσεων και θα συζητήσουμε τα πλεονεκτήματα και τα μειονεκτήματά τους. Κατόπιν, θα αξιολογήσουμε πώς κάποια από αυτά τα μειονεκτήματα αντιμετωπίζονται με τους πίνακες του σκορ ειδικούς ανά θέση (PSSMs) και τα προφίλ αλληλουχιών (profiles), τα οποία είναι πιο ευέλικτες στατιστικές περιγραφές των συντηρημένων περιοχών σε μια πολλαπλή στοίχιση. Τέλος, θα μιλήσουμε και για τα πιο γνωστά εργαλεία λογισμικού που χρησιμοποιούνται για την κατασκευή αλλά και για την αναγνώριση τέτοιων προτύπων και προφίλ σε αλληλουχίες.

Προαπαιτούμενη γνώση
Το κεφάλαιο απαιτεί κατανόηση των μεθόδων του κεφαλαίου 3 και του κεφαλαίου 4.

5. Εισαγωγή

Στο κεφάλαιο αυτό, αφού έχουμε ήδη μελετήσει τη στοίχιση και την πολλαπλή στοίχιση αλληλουχιών, θα μελετήσουμε το επόμενο πρόβλημα που προκύπτει: τον τρόπο με τον οποίο θα περιγράψουμε μαθηματικά μια πολλαπλή στοίχιση, με σκοπό να πάρουμε μια πιο συμπυκνωμένη αναπαράσταση της πληροφορίας που περιέχεται στις συντηρημένες περιοχές. Έτσι, θα δούμε στην αρχή τα πρότυπα ακολουθιών (patterns) και τον τρόπο με τον οποίο αυτά περιγράφονται στη λεγόμενη μορφή της PROSITE, ενώ παράλληλα θα δούμε και τις αναλογίες με τις κανονικές εκφράσεις (regular expressions) του UNIX. Αφού μελετήσουμε αναλυτικά τα πρότυπα, θα ασχοληθούμε και με τις αδυναμίες τους, και κατά συνέπεια την ανάγκη για πιο ακριβείς περιγραφές στις οποίες δεν θα υπάρχει απώλεια πληροφορίας. Έτσι, θα μιλήσουμε και για τα προφίλ (profiles) και τους πίνακες σκορ ειδικούς ανά θέση (Position Specific Scoring Matrices). Θα δούμε, ότι με αυτές τις περιγραφές δεν διευκολύνεται μόνο η αναζήτηση συντηρημένων περιοχών σε βάσεις δεδομένων, αλλά επιπλέον, ανοίγει και δρόμος για πιο ευαίσθητες αναζητήσεις και εντοπισμό μακρινών ομολόγων. Για όλα τα παραπάνω θέματα, θα μελετήσουμε επίσης τους αλγόριθμους που χρησιμοποιούνται για την κατασκευή αυτών των αναπαραστάσεων (προτύπων ή προφίλ), αλλά και τα εργαλεία λογισμικού που υπάρχουν διαθέσιμα για το σκοπό αυτό.

5.1. Πρότυπα και μοτίβα αλληλουχιών

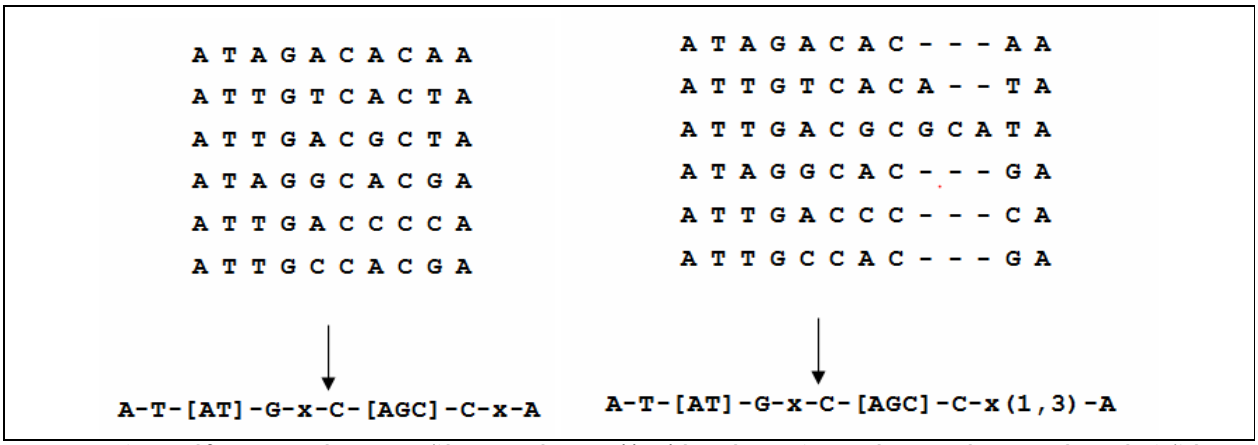
5.1.1 Τι είναι τα πρότυπα

Όταν έχουμε μια καλά προσδιορισμένη πολλαπλή στοίχιση αλληλουχιών (είτε πρωτεϊνών, είτε νουκλεϊκών οξέων), ένα θέμα που μας ενδιαφέρει είναι να μπορούμε να εξάγουμε μια πιο πληροφοριακή περιγραφή της. Είδαμε, για παράδειγμα, ότι μια τέτοια πολλαπλή στοίχιση μπορεί να περιγράφει μια πρωτεϊνική οικογένεια, δηλαδή μια ομάδα πρωτεϊνών με κοινή εξελικτική ιστορία, οι οποίες έχουν κοινά δομικά και πιθανώς και λειτουργικά χαρακτηριστικά. Έτσι, θα μας ενδιέφερε να βρούμε έναν τρόπο να περιγράψουμε τα κοινά χαρακτηριστικά όλων αυτών των αλληλουχιών, και να έχουμε μια εύκολη και κατανοητή περιγραφή, χωρίς να χρειάζεται κάθε φορά να ανατρέχουμε στην ίδια την πολλαπλή στοίχιση η οποία μπορεί να είναι μεγάλη αλλά και δυσνόητη. Επίσης, θα μας ενδιέφερε να βρούμε με βάση αυτή την περιγραφή, έναν εύκολο και γρήγορο τρόπο για να πραγματοποιήσουμε μια αναζήτηση στις βάσεις δεδομένων, για αλληλουχίες που έχουν αυτό το κοινό χαρακτηριστικό, χωρίς όμως να χρειάζεται να πραγματοποιήσουμε εκ νέου στοίχιση.

Οι περιγραφές αυτές, ονομάζονται πρότυπα (patterns) και έχουν μεγάλη ιστορία στην επιστήμη των υπολογιστών για την περιγραφή κειμένων και άλλων ειδών ακολουθιών από χαρακτήρες (είναι τα γνωστά regular expressions). Με τις περιγραφές αυτές, μπορούμε να δούμε σε ποια θέση μιας πολλαπλής στοίχισης υπάρχει μεγάλη ή μικρότερη συντήρηση και έτσι να χαρακτηρίσουμε και να εντοπίσουμε μεταξύ άλλων ενεργά κέντρα, περιοχές δράσης των ενζύμων και θέσεις δισουλφιδικών δεσμών (στις πρωτεΐνες) ή υποκινητές, θέσεις έναρξης γονιδίων και σημεία συρραφής εξωνίων (στα γονίδια). Στη βιοπληροφορική ο παραδοσιακός τρόπος χρήσης τέτοιων εκφράσεων είναι με τα λεγόμενα πρότυπα της PROSITE, τα οποία είναι μεν εντελώς ανάλογα με τις κανονικές εκφράσεις του UNIX αλλά έχουν μια σύνταξη λίγο πιο «εύκολη» και κατανοητή.

Στα πρότυπα αυτά (Εικόνα 5.1), ολόκληρη η πολλαπλή στοίχιση ή για την ακρίβεια, οι στήλες της που εμφανίζουν το μεγαλύτερο ενδιαφέρον, περιγράφονται με μία συμπυκνωμένη έκφραση. Αφενός μεν αυτό προσδίδει μια τεράστια ευκολία καθώς μια πολλαπλή στοίχιση πιθανά εκατοντάδων αλληλουχιών συνοψίζεται σε μία γραμμή, αφετέρου δε, αυτή η διαδικασία αναπόφευκτα οδηγεί σε απώλεια πληροφορίας (θα επανέλθουμε σε αυτό). Τα βασικά χαρακτηριστικά της σύνταξης PROSITE είναι τα παρακάτω:

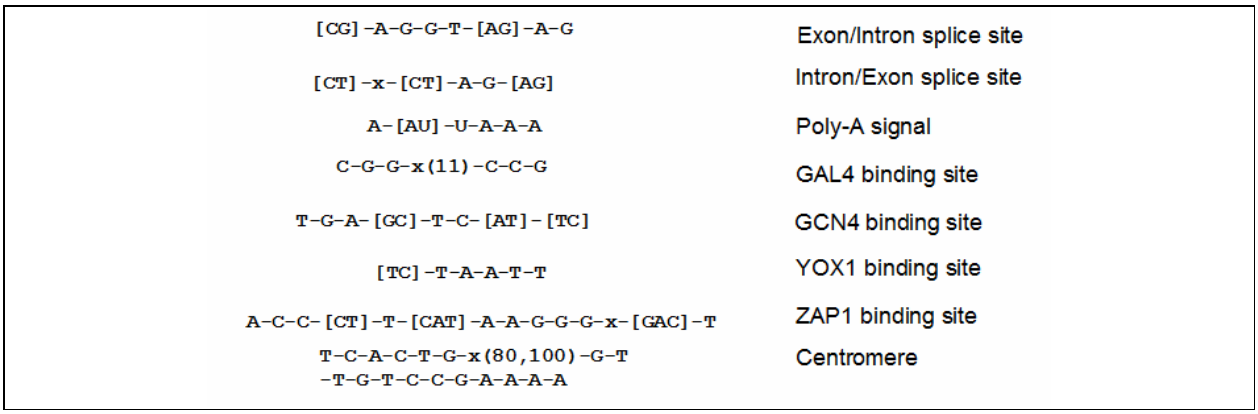
- Τα αμινοξέα ή τα νουκλεοτίδια αναπαρίστανται με τον τυπικό κωδικό του ενός γράμματος της IUPAC.
- Κάθε θέση της πολλαπλής στοίχισης αντιστοιχεί σε μια θέση στο πρότυπο, η οποία διαχωρίζεται από τις υπόλοιπες με μία παύλα (-).
- Οι θέσεις είναι ανεξάρτητες μεταξύ τους.
- Αν σε κάποια θέση εμφανίζεται μόνο ένας χαρακτήρας, τότε στο πρότυπο χρησιμοποιείται αυτούσιος (π.χ. A, T κ.ο.κ.)
- Αν σε κάποια θέση εμφανίζονται δύο ή περισσότεροι χαρακτήρες τότε αυτοί εμφανίζονται μέσα σε άγκιστρο, για παράδειγμα [AT] σημαίνει ότι επιτρέπεται A ή T, ενώ [ACG] σημαίνει ότι επιτρέπεται είτε A, είτε G, είτε C.
- Αν σε κάποια θέση επιτρέπεται να εμφανιστεί οποιοδήποτε σύμβολο, τότε αυτή η θέση συμβολίζεται με x.
- Αν σε κάποια θέση επιτρέπεται να εμφανιστεί οποιοδήποτε σύμβολο εκτός από κάποιο/α, τότε τη θέση τη συμβολίζουμε με {}. Για παράδειγμα, για να πούμε «οποιοδήποτε νουκλεοτίδιο εκτός από A» γράφουμε {A} το οποίο στην περίπτωση του DNA είναι ισοδύναμο με το [CGT]. Προφανώς, αυτός ο κανόνας είναι περισσότερο χρήσιμος στην περίπτωση των πρωτεϊνών με το μεγάλο αλφάβητο.
- Επαναλήψεις συμβολίζονται με παρένθεση () μετά από ένα σύμβολο. Για παράδειγμα το A(3) σημαίνει A-A-A, ενώ το x(3) σημαίνει x-x-x (δηλαδή 3 οποιαδήποτε σύμβολα). Επίσης, μέσα στην παρένθεση μπορεί να μπει και ένα εύρος τιμών. Έτσι, το x(2,4) σημαίνει x-x, ή x-x-x, ή x-x-x-x.
- Η αρχή και το τέλος της αλληλουχίας συμβολίζονται με τα σύμβολα < και > αντίστοιχα. Έτσι, για να πούμε ότι η αλληλουχία αρχίζει με A και μετά ακολουθεί οποιοδήποτε σύμβολο γράφουμε <A-x
- Σε κάποιες ειδικές περιπτώσεις το σύμβολο '>' μπορεί να εμφανιστεί μέσα στα άγκιστρα για να χαρακτηρίσει την πιθανή ύπαρξη καρβοξυτελικού άκρου. Έτσι, το P-R-L-[G>] σημαίνει είτε P-R-L-G ή P-R-L>.



Εικόνα 5.1: *Παράδειγμα προτύπου που εξάγεται από μια πολλαπλή στοίχιση. Αριστερά μια στοίχιση χωρίς κενά. Δεξιά μια στοίχιση με κενά. Στην τελευταία περίπτωση θα πρέπει να αποφασίσουμε ποιες στήλες δεν θα αντιπροσωπευθούν στο πρότυπο.*

Συνήθως στις αλληλουχίες των γονιδίων, τέτοια πρότυπα, δηλαδή πολύ συντηρημένες περιοχές, συναντάμε στις αλληλουχίες των υποκινητών, στις θέσεις αποκοπής των εσωνίων από τα εξώνια κ.ο.κ (Εικόνα 5.2). Στις αλληλουχίες πρωτεϊνών, τέτοιες περιοχές χαρακτηρίζουν τις θέσεις δράσης ενζύμων, τα

ενεργά κέντρα των ενζύμων ή κάποιες πολύ χαρακτηριστικές περιοχές της δευτεροταγούς δομής όπως για παράδειγμα τις θέσεις κυστεϊνών που σχηματίζουν δισουλφιδικούς δεσμούς (Εικόνα 5.3).



Εικόνα 5.2: *Μερικά παραδείγματα γνωστών προτύπων που εμφανίζονται σε αλληλουχίες DNA.*

Όπως είδαμε στο Κεφάλαιο 2, η **PROSITE** (http://www.expasy.ch/prosite/) αποτελεί μια βάση ταξινόμησης πρωτεϊνικών ακολουθιών και αυτοτελών περιοχών ακολουθιών (sequence domains) σε οικογένειες (Sigrist et al., 2010). Ο παραδοσιακός τρόπος καταχώρησης μιας οικογένειας στη βάση αυτή, γίνεται με τους ομώνυμους κανόνες που περιγράψαμε παραπάνω και είναι ο πιο παλιός και εύκολος στη δημιουργία, ενώ ο άλλος βασίζεται στην κατασκευή προφίλ, μέθοδος η οποία είναι πιο σύνθετη αλλά και πιο ευαίσθητη (θα μελετηθεί στη συνέχεια). Μέχρι σήμερα η PROSITE περιέχει καταχωρήσεις για περισσότερες από 1700 οικογένειες. Συνολικά, υπάρχουν στη βάση 1308 πρότυπα, 1107 προφίλ και 1105 "κανόνες" (αφορούν κυρίως πληροφορίες για το πού θα πρέπει να βρίσκεται το πρότυπο για να θεωρηθεί έγκυρο αλλά και πληροφορίες για συνδυασμούς από πρότυπα). Προφανώς, υπάρχουν οικογένειες για τις οποίες υπάρχουν διαθέσιμα και πρότυπα και προφίλ (συνήθως, οι παλαιότερες καταχωρήσεις αφορούσαν το πρότυπα). Στη βάση υπάρχουν επίσης αναλύσεις τόσο για τις πρωτεΐνες της Uniprot που ανήκουν σε κάθε οικογένεια, όσο και για τις πρωτεΐνες στις οποίες εμφανίζεται ένα "αποτύπωμα" (κυρίως όταν έχουμε να κάνουμε με πρότυπα) αλλά είναι γνωστό ότι δεν ανήκουν λειτουργικά στην οικογένεια αυτή. Τέλος, υπάρχουν εργαλεία για την αναζήτηση των προτύπων και των προφίλ σε ακολουθίες, όσο και εργαλεία αναπαράστασης της "σπονδυλωτής" δομής των πρωτεϊνών, δηλαδή της αναπαράστασης των περιοχών αυτών και την αποτύπωση της διάταξής τους πάνω σε μια δεδομένη ακολουθία.

Όπως αναφέραμε ήδη, οι κανονικές εκφράσεις (regular expressions) και οι εκφράσεις της PROSITE είναι ισοδύναμες. Οι διαφορές στη σύνταξη είναι οι εξής:

- Η κάθε θέση αναγράφεται συνεχόμενα χωρίς να μεσολαβεί η παύλα (-).
- Το σύμβολο για «οποιοδήποτε» χαρακτήρα είναι η τελεία (.) αντί για το x
- Το σύμβολο για το «οποιοδήποτε χαρακτήρα εκτός από» είναι το ^ μέσα στην αγκύλη, και όχι το άγκιστρο {}.

Για παράδειγμα, αν θεωρήσουμε το πρότυπο της PROSITE που δίνεται από την έκφραση:

```
[RK]-G-{EDRKHPCG}-[AGSCI]-[FY]-[LIVA]-x-[FYM]
```

τότε η αντίστοιχη κανονική έκφραση θα είναι:

```
[RK]G[^EDRKHPCG][AGSCI][FY][LIVA].[FYM]
```

Κάτι που επίσης πρέπει να τονιστεί, είναι ότι αν και τα πρότυπα αυτά χρησιμοποιούνται εντατικά από τη δεκαετία του 1980 για τον χαρακτηρισμό οικογενειών πρωτεϊνών, και παρά το γεγονός ότι η PROSITE περιέχει πλήθος τέτοιων καταχωρίσεων, η υπόθεση εύρεσης και χαρακτηρισμού προτύπων τα οποία θα μπορούν να χρησιμοποιηθούν για την πρόγνωση δομικών και λειτουργικών χαρακτηριστικών των πρωτεϊνών δεν έχει σταματήσει καθόλου, καθώς τέτοια πρότυπα ανακαλύπτονται συνεχώς. Στην Εικόνα 5.3 βλέπουμε μόνο μερικά από τα εκατοντάδες σχετικά πρότυπα που είναι γνωστά εδώ και πολλά χρόνια. Παρ' όλα αυτά, στην Εικόνα 5.4 βλέπουμε κάποια άλλα πρότυπα, τα οποία έχουν ανακαλυφθεί μέσα στα τελευταία 15 χρόνια.

Για παράδειγμα, τα σήματα πυρηνικού εντοπισμού (nuclear localization signals - NLSs) είναι μικρές αλληλουχίες, γνωστές από παλιά, πλούσιες σε Αργινίνη και Λυσίνη, οι οποίες είναι υπεύθυνες για τη μεταφορά των πρωτεϊνών στον πυρήνα του κυττάρου. Οι Cocol, Nair και Rost, πραγματοποίησαν μια

εκτεταμένη ανάλυση στις γνωστές πυρηνικές πρωτεΐνες και εντόπισαν 214 επιπλέον τέτοια πρότυπα (πέραν των 91 που ήταν ήδη γνωστά) (Cokol, Nair, & Rost, 2000). Ένα άλλο παράδειγμα, αφορά τα σήματα μεταφοράς των πρωτεϊνών στα υπεροξεισώματα. Το πρώτο είδος σήματος στόχευσης των υπεροξεισωμάτων (peroxisomal targeting signal -PTS1) ήταν γνωστό εδώ και χρόνια (το καρβοξυτελικό S-K-L). Το 2004 όμως, ανακαλύφθηκε και ένας δεύτερος μηχανισμός ο οποίος έκανε χρήση ενός αμινοτελικού πεπτιδίου και οι Petriv και συνεργάτες εντόπισαν το πρότυπο που το περιγράφει, το οποίο και ονόμασαν PTS2 (Petriv, Tang, Titorenko, & Rachubinski, 2004).

C-x-C-x(2)-{V}-x(2)-G-{C}-x-C	EGF-like 1 domain
[RK]-x(2,3)-[DE]-x(2,3)-Y	Tyrosine kinase phosphorylation site
N-{P}-[ST]	N-linked glycosylation
[LIVMA]-G-[EQ]-H-G-[DN]-[ST]	L-lactate dehydrogenase active site
P-[LIVM]-C-T-[LIVM]-[KRH]-x-[FT]-P	Ubiquitin-activating enzyme signature
S-K-L>	Peroxisomal Target Sequence 1 (PTS1)
{DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C	Bacterial Lipoprotein signal peptide

Εικόνα 5.3: *Μερικά από τα γνωστά παραδείγματα προτύπων που εμφανίζονται σε πρωτεΐνες.*

Μια άλλη πολύ γνωστή περίπτωση, είναι αυτή των βακτηριακών λιποπρωτεϊνών. Οι πρωτεΐνες αυτές έχουν μια σηματοδοτική αλληλουχία (signal peptide) η οποία μοιάζει αρκετά με αυτή των εκκρινόμενων πρωτεϊνών, αλλά στο καρβοξυτελικό της άκρο φέρει μια χαρακτηριστική αλληλουχία η οποία αναγνωρίζεται από ειδικό ένζυμο, το οποίο αποκόπτει το πεπτίδιο αυτό και ακολούθως η ώριμη πρωτεΐνη προσκολλάται στα λιπίδια της μεμβράνης με ομοιοπολικό δεσμό. Η αλληλουχία που αναγνωρίζει το ένζυμο, έχει μια συντηρημένη κυστεΐνη στο καρβοξυτελικό της άκρο (περίπου στη θέση 17-30 της πρόδρομης πρωτεΐνης, εκεί που γίνεται και η τροποποίηση), ενώ στις προηγούμενες θέσεις υπάρχουν κυρίως Αλανίνες και Βαλίνες. Τέτοια πρότυπα είχαν περιγραφεί από τη δεκαετία του 1980, αλλά το πιο γνωστό είναι το λεγόμενο PS00013, όπως ήταν γνωστό από τον κωδικό της PROSITE (Εικόνα 5.3). Παρ' όλα αυτά, έχουν περιγραφεί και εναλλακτικά πρότυπα πολλές φορές ακόμα και χρόνια αργότερα, όπως το [LVI]-[ASTVI]-[GAS]-C το οποίο χρησιμοποιήθηκε για να κατασκευαστεί η βάση των βακτηριακών λιποπρωτεϊνών (DOLOP). Το 2002 επίσης, οι Sutcliffe και Harrington μελετώντας λιποπρωτεΐνες από βακτήρια θετικά κατά Gram, κατέληξαν σε ένα πιο αυστηρό αλλά ταυτόχρονα και πιο περιεκτικό πρότυπο, το οποίο περιγράφει καλύτερα τις λιποπρωτεΐνες αυτών των βακτηρίων (Sutcliffe & Harrington, 2002). Το πρότυπο αυτό δίνεται (μαζί με άλλα παραδείγματα) στην Εικόνα 5.4.

Μια πιο πρόσφατη εργασία όμως, αφορά τις λιποπρωτεΐνες που εκκρίνονται με το σύστημα TAT (twin-arginine translocation). Το σύστημα αυτό, υπάρχει σε όλα τα βακτήρια αλλά και τους χλωροπλάστες και εκκρίνει πρωτεΐνες με ένα σύστημα διαφορετικό από το γνωστό εκκριτικό μονοπάτι SEC. Για την ακρίβεια, οι πρωτεΐνες εκκρίνονται διπλωμένες στην τρισδιάστατη δομή τους, μέσω ενός διαμεμβρανικού υποδοχέα που λειτουργεί με έναν άγνωστο προς το παρόν μηχανισμό. Από άποψη αλληλουχίας, οι πρωτεΐνες αυτές φέρουν ένα αμινοτελικό πεπτίδιο (σηματοδοτική αλληλουχία), που μοιάζει πάρα πολύ με το κλασικό πεπτίδιο έκκρισης αλλά έχει ένα συντηρημένο πρότυπο που αποτελείται από 2 συνεχόμενες Αργινίνες (R-R-x-[FGAVML]-[LITMVF]). Για τις πρωτεΐνες αυτές έχουν αναπτυχθεί βέβαια πιο ειδικές μέθοδοι πρόγνωσης. Παρ' όλα αυτά, τα τελευταία χρόνια υπήρξαν πειραματικά δεδομένα που έδειχναν ότι υπάρχουν και περίεργες περιπτώσεις, δηλαδή πρωτεΐνες που εκκρίνονται με το TAT αλλά η ώριμη πρωτεΐνη δεν απελευθερώνεται, αντιθέτως προσκολλάται στη μεμβράνη, όπως μια λιποπρωτεϊνη. Με άλλα λόγια, υπάρχουν λιποπρωτεΐνες που χρησιμοποιούν το σύστημα TAT για την έκκριση και όχι το SEC. Έτσι, το 2010 οι Shruthi, Babu και Sankaran, χρησιμοποίησαν αυτές τις απλές παρατηρήσεις και με χρήση μόνο αυτών των δύο απλών προτύπων (R-R-x-[FGAVML]-[LITMVF] για τις πρωτεΐνες TAT και [LVI]-[ASTVI]-[GAS]-C για τις λιποπρωτεΐνες), εντόπισαν όλες τις πιθανές TAT-λιποπρωτεΐνες που υπάρχουν στα βακτηριακά γονιδιώματα και μελέτησαν τις πιθανές λειτουργίες τους (Shruthi, Babu, & Sankaran, 2010).

Τέλος, ένα άλλο σχετικά πρόσφατο παράδειγμα, προέρχεται από την πρόγνωση των διαμεμβρανικών β-βαρελιών των αρνητικών κατά Gram βακτηρίων. Το 2004, οι Berven και συνεργάτες, εντόπισαν ένα συντηρημένο πρότυπο που περιγράφει τις περισσότερες καρβοξυτελικές περιοχές των β-βαρελιών (Berven, Flikka, Jensen, & Eidhammer, 2004). Το πρότυπο αυτό, δεν είναι βέβαια ικανό από μόνο του να διαχωρίσει όλες τις αντίστοιχες πρωτεΐνες, αλλά χρησιμοποιήθηκε σε συνδυασμό με άλλες μεθοδολογίες για την κατασκευή του αλγορίθμου BOMP.

Σε κάθε περίπτωση, αυτό που πρέπει να γίνει κατανοητό από τα παραπάνω, είναι ότι ναι μεν η μεθοδολογία αυτή είναι εξαιρετικά απλή και εύκολη στη χρήση, αλλά ακόμα και σήμερα υπάρχουν περιθώρια για τη χρήση της και πολλά χρήσιμα βιολογικά συμπεράσματα μπορούν να εξαχθούν από αυτή.

[RK]-[LVI]Q]-x(2)-[LVIHQ] -[LSGAK]-x-[HQ]-[LAF]	Peroxisomal Target Sequence 2 (PTS2)
K-R-K-x{11}-K-K-K-S-K-K	Nuclear localization signal (*)
[LVI]-[ASTVI]-[GAS]-C	Alternative bacterial Lipoprotein signal peptide pattern
<[MV]-X(0,13)-[RK]-{DERKQ}(6,20) -[LIVMFESTAG]-[LVIAM]-[IVMSTAFG]-[AG]-C	Pattern specific for lipoproteins of Gram+ bacteria
R-R-x-[FGAVML]-[LITMVF]	Twin-arginine (TAT) signal peptide
x(100,)-{C}-[YFWKLHVITMAD]-{C} -[YFWKLHVITMAD]-{C}-[YFWKLHVITMAD] -{C}-[YFWKLHVITMAD]-{C}-[FYW]	C-terminal beta-strand pattern of bacterial OMPs

Εικόνα 5.4: *Κάποια επιλεγμένα παράδειγμα προτύπων σε πρωτεϊνικές αλληλουχίες, τα οποία ανακαλύφθηκαν μέσα στα τελευταία 15 χρόνια.*

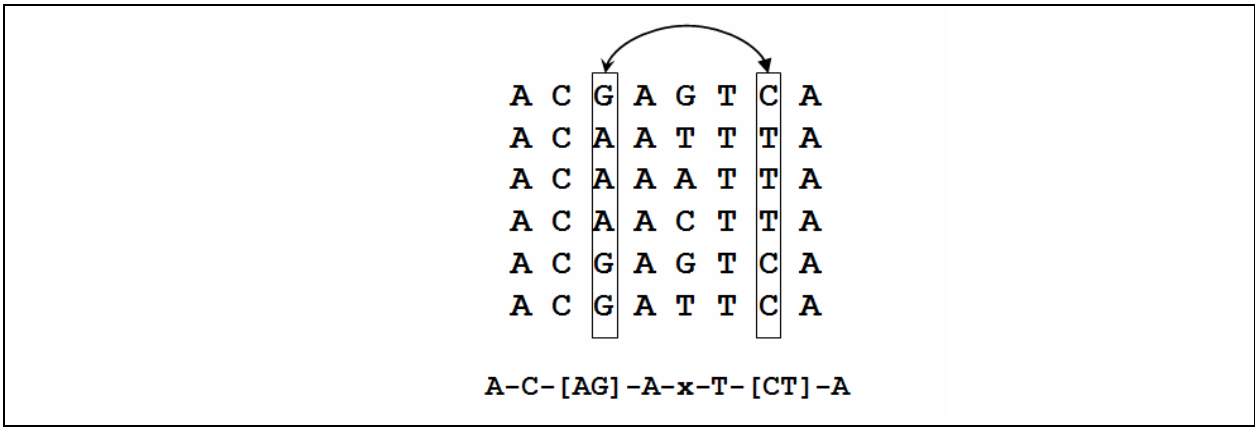
5.1.2 Πλεονεκτήματα και μειονεκτήματα των προτύπων

Τα πρότυπα, έχουν κάποια μοναδικά πλεονεκτήματα. Καταρχάς, είναι κατανοητά στο ανθρώπινο μάτι. Διαβάζοντας μια τέτοια έκφραση, καταλαβαίνουμε αμέσως την πληροφορία που περιέχει. Έτσι, είναι πολύ περιεκτικά και συμπυκνώνουν την πληροφορία μιας πιθανά μεγάλης πολλαπλής στοίχισης, μέσα σε μερικούς μόνο χαρακτήρες. Μας βοηθούν με αυτόν τον τρόπο να ταξινομήσουμε και να κατανοήσουμε φαινόμενα που είναι γενικά δύσκολα. Επίσης, είναι ιδιαίτερα αποδοτικά από υπολογιστικής πλευράς για πρακτικές χρήσης. Τα πρότυπα PROSITE καθώς είναι ισοδύναμα με τις κανονικές εκφράσεις (regular expressions), μπορούν να βασιστούν στις υλοποιήσεις που κάνουν χρήση πεπερασμένων αυτομάτων με συνέπεια να είναι ιδιαίτερα εύκολο και γρήγορο το να αποτελέσουν τμήμα μια υπολογιστικής μεθοδολογίας για ταχείες αναζητήσεις σε μεγάλες βάσεις δεδομένων. Στο κεφάλαιο 12 θα δούμε ότι η υλοποίηση τέτοιων εκφράσεων σε μια γλώσσα προγραμματισμού όπως η Perl είναι κάτι ιδιαίτερα εύκολο, ενώ αντίστοιχες δυνατότητες δίνουν ακόμα και οι βασικές εντολές του UNIX (grep, egrep).

Από την άλλη μεριά όμως, αυτά ακριβώς τα χαρακτηριστικά που κάνουν τα πρότυπα ιδιαίτερα επιτυχημένα, περιέχουν και το σπόρο με τις αδυναμίες τους. Το βασικό μειονέκτημα είναι ότι χάνεται μεγάλο μέρος της πληροφορίας της πολλαπλής στοίχισης. Για παράδειγμα στην Εικόνα 5.1 στην $3^{η}$ στήλη της στοίχισης το πρότυπο προβλέπει [AT], δηλαδή Α ή Τ, αλλά δεν μας δίνει τη σχετική πιθανότητα για το καθένα, παρόλο που από την πολλαπλή στοίχιση βλέπουμε ότι η Θυμίνη (Τ) έχει διπλάσια πιθανότητα από την Αδενίνη (Α). Φανταστείτε ότι έχουμε τώρα την ίδια περίπτωση αλλά σε μια στοίχιση με 100 αλληλουχίες, και εκεί έχουμε 65 Τ και 35 Α. Αν τώρα γίνει γνωστή μια επιπλέον αλληλουχία που ανήκει σίγουρα (με βάση βιολογικά κριτήρια) στη συγκεκριμένη οικογένεια, αλλά στη θέση αυτή έχει G, τι θα πρέπει να γίνει σε αυτή την περίπτωση; Αν το πρότυπο διευρυνθεί για να περιλαμβάνει και τη νέα αλληλουχία (γίνει δηλαδή [AGT]), τότε θα έχουμε χάσει ακόμα μεγαλύτερο μέρος της προβλεπτικής δύναμης. Αν επιλέξουμε να μην κάνουμε αυτή τη διεύρυνση, τότε θα είμαστε αναγκασμένοι να έχουμε ένα πρότυπο το οποίο «χάνει» κάποια από τα πραγματικά μέλη της οικογένειας. Αυτό είναι ένα πραγματικό πρόβλημα, και υπάρχουν και στη βάση

PROSITE πρότυπα τα οποία αδυνατούν να χαρακτηρίσουν το 100% των μελών μιας πρωτεϊνικής οικογένειας. Προφανώς, στην περίπτωση των πρωτεϊνών το πρόβλημα είναι πολύ πιο έντονο καθώς όπως είδαμε στα προηγούμενα κεφάλαια, σε πρωτεϊνικές οικογένειες με πολλά μέλη είναι σχεδόν αδύνατο να βρεις στήλες στην πολλαπλή στοίχιση με απόλυτη ομοφωνία καθώς αυτό που συντηρείται τις περισσότερες φορές είναι οι φυσικοχημικές ιδιότητες (πχ υδρόφοβα αμινοξέα, θετικά φορτισμένα αμινοξέα κ.ο.κ.). Με λίγα λόγια, είναι πολύ συνηθισμένο μια καλή στοίχιση να περιλαμβάνει σε μια στήλη αρκετά, διαφορετικά μεταξύ τους, αμινοξέα. Το πρόβλημα αυτό, θα το λύσουν εν μέρει τα προφίλ αλληλουχιών (sequence profiles) και οι ειδικοί ανά θέση πίνακες σκορ (PSSMs), τους οποίους θα δούμε στην επόμενη ενότητα.

Ένα άλλο πρόβλημα, είναι ότι τα πρότυπα με τον τρόπο που τα ορίσαμε δεν μπορούν να ενσωματώσουν εύκολα τα κενά στην πολλαπλή στοίχιση. Στην Εικόνα 5.1 είδαμε μια πολλαπλή στοίχιση που περιέχει κενά, αλλά όλα προέρχονται από εισαγωγές (τυχαίων) νουκλεοτιδίων σε κάποιες από τις αλληλουχίες της στοίχισης. Έτσι, τα κενά στην $1^{η}$, $4^{η}$, $5^{η}$ και $6^{η}$ αλληλουχία αντιστοιχούν απλά στις εισαγωγές νουκλεοτιδίων στην $2^{η}$ και στην $3^{η}$ αλληλουχία. Τι θα γινόταν όμως αν λ.χ. στην πρώτη αλληλουχία στην $8^{η}$ θέση δεν είχε την Κυτοσίνη (C); Με την υπάρχουσα ορολογία, απλά δεν θα ταίριαζε στο μοντέλο. Το πρόβλημα αυτό το λύνουν εν μέρει τα προφίλ, αντιμετωπίζοντάς το με τον κλασικό τρόπο που είδαμε στη στοίχιση αλληλουχιών (με δυναμικό προγραμματισμό και ποινές για τα κενά), αλλά την πιο ολοκληρωμένη λύση τη δίνουν τα Hidden Markov Models (HMMs) που θα δούμε στο κεφάλαιο 8.



Εικόνα 5.5: *Ένα παράδειγμα πολλαπλής στοίχισης με εξάρτηση μεταξύ 2 γειτονικών θέσεων.*

Τέλος, υπάρχει και ένα μεγαλύτερο πρόβλημα, το οποίο όμως είναι και πιο δύσκολο να εντοπιστεί αλλά και να διορθωθεί. Ο τρόπος που αντιμετωπίζουν τα πρότυπα τις θέσεις της πολλαπλής στοίχισης, είναι σαν να πρόκειται για ανεξάρτητες παρατηρήσεις. Στην Εικόνα 5.5 βλέπουμε μια πολλαπλή στοίχιση με το αντίστοιχο πρότυπο στην οποία υπάρχει ισχυρή συσχέτιση (δηλαδή, αλληλεπίδραση) μεταξύ της στήλης 3 και της στήλης 7. Αν εξετάσουμε κάθε στήλη ξεχωριστά, βλέπουμε ότι στην 3 έχουμε 50% G και 50% A, ενώ στην 7 έχουμε 50% T και 50% C. Το πρότυπο PROSITE θα έδινε για παράδειγμα την ίδια πιθανότητα να εμφανιστεί G ($3^{η}$) και C ($7^{η}$), και να εμφανιστεί G ($3^{η}$) και T ($7^{η}$). Παρατήρηση όμως των συχνοτήτων των δινουκλεοτιδίων, μας δείχνει ότι όταν υπάρχει G ($3^{η}$) υπάρχει πάντα C ($7^{η}$), ενώ όταν υπάρχει A ($3^{η}$) πάντα ακολουθείται από T ($7^{η}$). Αυτή η εξάρτηση, είναι κάτι ιδιαίτερα δύσκολο να μοντελοποιηθεί, καθώς όλες οι μεθοδολογίες που έχουμε δει μέχρι τώρα κάνουν λόγο για ανεξάρτητες θέσεις, ενώ το ίδιο ισχύει και για τα προφίλ που θα δούμε παρακάτω αλλά και για τα HMM που είναι η γενίκευση τους. Μεθοδολογίες που θα μπορούσαν με διαφορετικό τρόπο η καθεμιά να αντιμετωπίσουν αυτό το πρόβλημα, περιλαμβάνουν τα μαρκοβιανά μοντέλα εξάρτησης (Κεφάλαιο 8), τα Νευρωνικά Δίκτυα (Κεφάλαιο 7), αλλά και τις στοχαστικές γραμματικές χωρίς συμφραζόμενα (Κεφάλαιο 10).

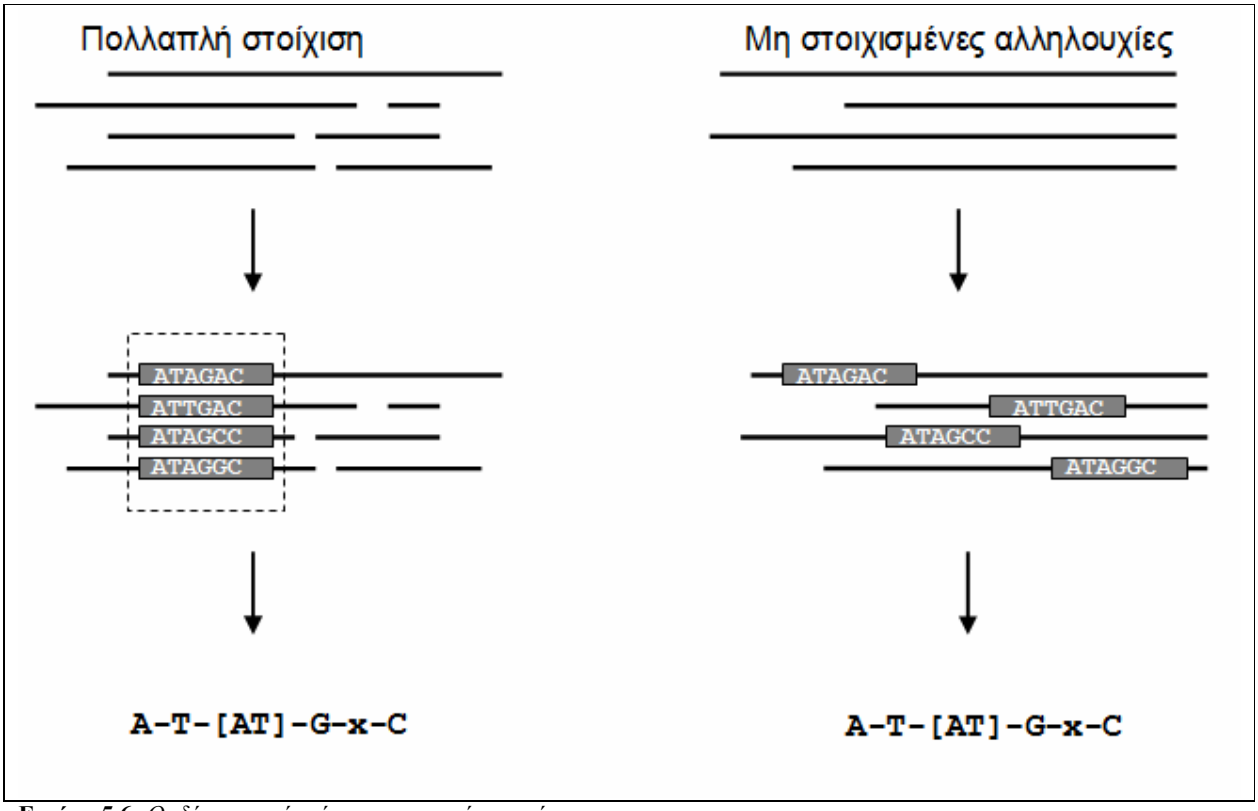
5.1.3 Κατασκευή των προτύπων και λογισμικό

Υπάρχουν δύο γενικοί τρόποι για την κατασκευή προτύπων: είτε ξεκινώντας από στοιχισμένες αλληλουχίες, είτε από μη στοιχισμένες (Εικόνα 5.6). Στην πρώτη περίπτωση, τα πράγματα είναι πιο απλά καθώς έχουμε εντοπίσει τη στοίχιση και η εύρεση των συντηρημένων περιοχών είναι μια τετριμμένη διαδικασία η οποία μπορεί να διεκπεραιωθεί με μια απλή καταμέτρηση. Η δεύτερη περίπτωση όμως, έχει μεγαλύτερο ενδιαφέρον

καθώς αποδεσμεύει το πρόβλημα από τη στοίχιση, αλλά επιπλέον προσφέρει το πλεονέκτημα ότι μπορεί να εντοπίσει πολλαπλές επαναλήψεις του ίδιου προτύπου στην αλληλουχία, αλλά και να εντοπίσει πρότυπα σε μη ομόλογες αλληλουχίες. Το μειονέκτημα βέβαια είναι ότι απαιτείται ειδικός αλγόριθμος.

Γενικά, η εύρεση προτύπων αποτελείται από 3 διακριτά μέρη (Brazma, Jonassen, Eidhammer, & Gilbert, 1998):

- *Επιλογή της γλώσσας*: στο πρώτο στάδιο θα πρέπει να επιλεγεί ο τρόπος περιγραφής των προτύπων. Μπορεί δηλαδή να χρησιμοποιηθεί η σύνταξη της PROSITE, αλλά υπάρχουν και περιπτώσεις στις οποίες επιλέγονται και πιο απλές περιγραφές (π.χ. πρότυπα τα οποία περιέχουν εκφράσεις χωρίς πολλαπλές ταυτίσεις σε κάποια θέση, δηλαδή είτε ένα σύμβολο είτε οποιοδήποτε).
- *Κριτήριο καταλληλότητας*: αυτό είναι το μέτρο με το οποίο θα αξιολογήσουμε ένα πρότυπο ως καλό. Μπορεί να περιλαμβάνει απλές εκφράσεις, όπως τον αριθμό ή το ποσοστό των συντηρημένων θέσεων, μέχρι πιο σύνθετες όπως το συνολικό πληροφοριακό περιεχόμενο ή την πιθανοφάνεια.
- *Αλγόριθμος:* το τελευταίο κομμάτι αφορά τον τρόπο αναζήτησης και είναι περισσότερο σχετικό στην περίπτωση μη στοιχισμένων αλληλουχιών, στις οποίες το πρόβλημα είναι NP-complete, οπότε συνήθως χρησιμοποιούνται ευριστικές τεχνικές (heuristic) ή άπληστοι (greedy) αλγόριθμοι, στους οποίους περιορίζεται το εύρος αναζήτησης (π.χ. αναζήτηση όλων των προτύπων με μέγεθος μέχρι ένα ορισμένο σημείο). Επίσης, χρησιμοποιούνται ευρέως και στατιστικές τεχνικές, όπως ο αλγόριθμος EM (Expectation-Maximization) και ο Gibbs sampler.



Εικόνα 5.6: *Οι δύο γενικοί τρόποι κατασκευής προτύπων.*

Το πιο παλιό και ευρέως χρησιμοποιούμενο εργαλείο για την κατασκευή προτύπων, είναι το **PRATT** (http://web.expasy.org/pratt/). Το PRATT χρησιμοποιεί μια αναπαράσταση γράφων για τα πρότυπα, λειτουργεί με μη στοιχισμένες αλληλουχίες και δέχεται πρότυπα στη μορφή PROSITE (Jonassen, Collins, & Higgins, 1995). Ο χρήστης δίνει σαν δεδομένα εισόδου τις αλληλουχίες και τις γενικές απαιτήσεις των προτύπων, π.χ. το εύρος του μήκους τους, τον αριθμό με τις μη συντηρημένες θέσεις που μπορεί να

περιέχουν, και τον αριθμό των πρωτεϊνών στις οποίες πρέπει να εμφανίζονται. Το PRATT έχει μπορέσει να ανακατασκευάσει αρκετά ήδη γνωστά πρότυπα, ενώ είναι μια ιδιαίτερα εύχρηστη και γρήγορη εφαρμογή που υπάρχει και σε διαδικτυακή έκδοση.

Το **MEME** (http://meme-suite.org/tools/meme) είναι μια επίσης πολύ γνωστή μέθοδος που βασίζεται στον αλγόριθμο EM (Multiple EM For Motif Elicitation). Το MEME διαθέτει πολλές εφαρμογές, κάποιες εκ των οποίων χρησιμοποιούν και προφίλ αλληλουχιών (θα τα εξετάσουμε παρακάτω). Στη γενική περίπτωση, ο αλγόριθμος χρησιμοποιεί μια στατιστική περιγραφή των προτύπων και βασίζεται στο γνωστό πρόβλημα της μίξης των κατανομών (αντιμετωπίζει τις στήλες σαν ανεξάρτητες παρατηρήσεις από πολυωνυμικές κατανομές με διαφορετικές πιθανότητες). Ο αλγόριθμος δέχεται επίσης κάποιες αρχικές παραδοχές για το μήκος του προτύπου και με μια επαναληπτική διαδικασία μέγιστης πιθανοφάνειας εντοπίζει τις βέλτιστες περιοχές πάνω στις αλληλουχίες οι οποίες φέρουν κάποιο χαρακτηριστικό (Bailey & Elkan, 1994).

Μια άλλη παρόμοια εφαρμογή, είναι ο **Gibbs Motif Sampler** ο οποίος όπως λέει το όνομα, βασίζεται στη στατιστική μεθοδολογία του Gibbs sampler (http://ccmbweb.ccv.brown.edu/gibbs/gibbs.html). Η μέθοδος αυτή έχει διάφορες παραλλαγές εστιασμένες σε διαφορετικές απαιτήσεις, όπως για παράδειγμα την εύρεση θέσεων πρόσδεσης μεταγραφικών παραγόντων ή τις επαναληπτικές αλληλουχίες, ενώ είναι διαθέσιμη και ως αυτόνομο λογισμικό (Thompson, Rouchka, & Lawrence, 2003).

Τέλος, ο **TEIRESIAS** ο οποίος αναπτύχθηκε από τον Έλληνα επιστήμονα Ισίδωρο Ριγούτσο όταν αυτός εργαζόταν στην IBM, είναι ίσως ο πιο ενδιαφέρων από τους διαθέσιμους αλγορίθμους (Rigoutsos & Floratos, 1998). Ο αλγόριθμος είναι συνδυαστικός (combinatorial) και εντοπίζει πρότυπα που εμφανίζονται περισσότερες φορές από έναν επιλεγμένο από τον χρήστη αριθμό, αλλά το επιτυγχάνει αυτό χωρίς να απαριθμεί όλα τα ενδεχόμενα. Επιπλέον δε, τα πρότυπα που ανακαλύπτει είναι τα βέλτιστα δυνατά, με την έννοια ότι είναι αδύνατο να γίνουν πιο ειδικά και ταυτόχρονα να εμφανίζονται στις ίδιες ακριβώς θέσεις σε όλες τις αλληλουχίες. Ο TEIRESIAS είναι διαθέσιμος στη διεύθυνση https://cm.jefferson.edu/Teiresias/, ενώ ενδιαφέρον έχει ότι εκτός από τις εφαρμογές του στην ανακάλυψη προτύπων σε αλληλουχίες DNA, έχει χρησιμοποιηθεί και σε άλλου είδους προβλήματα όπως στον εντοπισμό ύποπτων συμπεριφορών στα δίκτυα υπολογιστών.

5.2. Weight Matrices, Profiles και PSSMs

Είδαμε στην προηγούμενη ενότητα τις βασικές αδυναμίες των προτύπων. Η πιο σημαντική από αυτές, είναι ότι σε κάθε θέση «χάνεται» η πληροφορία για τη σχετική αναλογία των συμβόλων του αλφαβήτου, και η αδυναμία να ποσοτικοποιήσει την ταύτιση μιας δεδομένης αλληλουχίας. Τα προβλήματα αυτά, άρχισαν να γίνονται φανερά και πιο έντονα όσο τα δεδομένα συσσωρεύονταν με αποτέλεσμα να εμφανίζονται όλο και περισσότερες περιπτώσεις αλληλουχιών που για μία ή δύο αλλαγές στην αλληλουχία τους, δεν ταίριαζαν στο γνωστό πρότυπο. Τις αδυναμίες αυτές, έρχονται να αντιμετωπίσουν οι σταθμισμένοι πίνακες (weight matrices) και τα προφίλ (profiles). Με τη μεθοδολογία αυτή, κατασκευάζεται ένας πίνακας k x p, όπου k είναι το μέγεθος του αλφαβήτου και p το μέγεθος της περιοχής που μοντελοποιούμε (οι στήλες της πολλαπλής στοίχισης). Έτσι, σε κάθε θέση i της πολλαπλής στοίχισης αντιστοιχίζουμε ένα διάνυσμα με τις πιθανότητες εμφάνισης $p_b(i)$ του κάθε συμβόλου (Εικόνα 5.7). Αν ονομάσουμε $n_b(i)$ τον αριθμό των εμφανίσεων του συμβόλου b στη στήλη i, τότε $p_b(i)$ θα είναι η πιθανότητα του συμβόλου b στη στήλη i, οποία θα δίνεται από τη σχέση:

$$p_b(i) = \frac{n_b(i)}{\sum_{\forall b' \in \Omega} n_{b'}(i)} \tag{5.1}$$

Με αυτόν τον τρόπο μπορούμε αμέσως να αντιμετωπίσουμε και τα δύο προβλήματα που προκύπτουν από την απώλεια πληροφορίας των προτύπων. Μπορούμε να καταλάβουμε ποιο σύμβολο εμφανίζεται με μεγαλύτερη πιθανότητα σε μια θέση, ενώ μπορούμε και να ποσοτικοποιήσουμε την ταύτιση μιας αλληλουχίας με το μοντέλο. Για παράδειγμα στα δεδομένα της Εικόνας 5.1, και αν θυμηθούμε το κεφάλαιο 3, θα δούμε ότι η αλληλουχία ATTGAACTA έχει συνολική πιθανότητα εμφάνισης ίση με:

$$P(\mathbf{x}) = \prod_{i-1}^{p} p_b(i) = P(x_1 = A)P(x_2 = T)P(x_3 = T)...P(x_{10} = A) = 0.074 \tag{5.2}$$

ενώ αντίστοιχα, η πιθανότητα της αλληλουχίας ATAGTTCAA θα είναι ίση με 0.00155 (η δε αλληλουχία AATGAACTA θα έχει πιθανότητα 0, καθώς έχει μια μη αποδεκτή αλλαγή στη θέση 1). Γενικά όμως, η απλή

αυτή μέθοδος δεν είναι πολύ πρακτική κυρίως λόγω των πολύ μικρών πιθανοτήτων που μπορεί να εμφανιστούν. Συνήθως σε τέτοιες περιπτώσεις παίρνουμε το λογάριθμο των πιθανοτήτων, αλλά ακόμα πιο αξιόπιστα αποτελέσματα θα έχουμε αν πάρουμε ένα λογαριθμικό σκορ όπως αυτά που συναντήσαμε στο κεφάλαιο 3. Το προσθετικό αυτό σκορ, θα αντικατοπτρίζει και τη σχετική πιθανότητα ενός συμβόλου και θα είναι της μορφής:

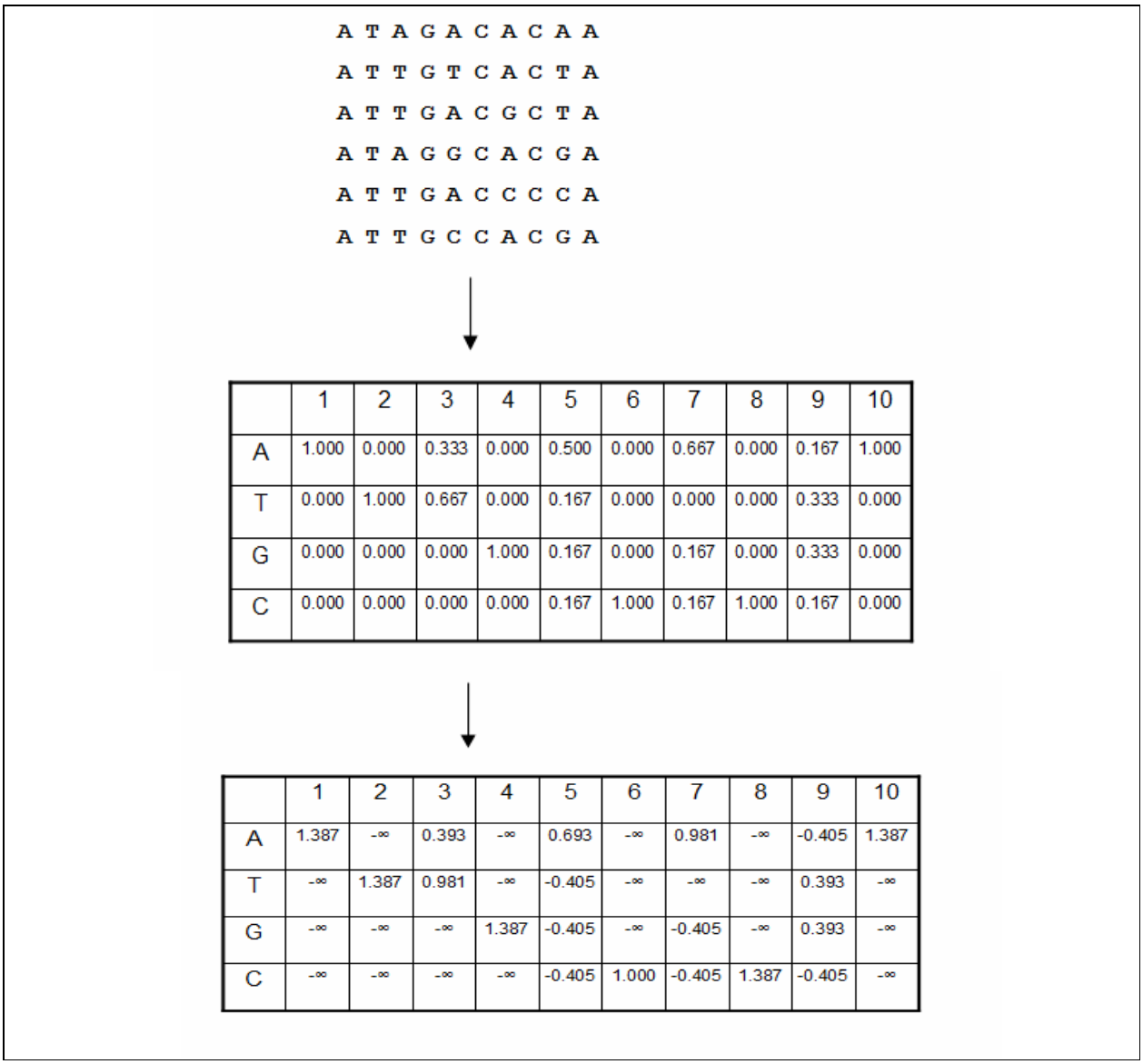
$$s_b(i) = \log\left(p_b(i) / p_b\right) \tag{5.3}$$

όπου p_b είναι η πιθανότητα εμφάνισης ενός συμβόλου (αμινοξέος ή νουκλεοτιδίου) γενικά (στο υπόβαθρο όπως λέμε) και $p_b(i)$ η πραγματική πιθανότητα εμφάνισης του ίδιου συμβόλου στη συγκεκριμένη θέση του πίνακα.

Με τον τρόπο αυτό, έχουμε ένα αθροιστικό σκορ το οποίο λαμβάνει επίσης υπόψη και τις συνολικές πιθανότητες εμφάνισης του κάθε συμβόλου. Ειδικά στις πρωτεΐνες, οι διαφορές μπορεί να είναι μεγάλες καθώς δεν είναι το ίδιο να έχουμε 100% συντήρηση ενός κοινού αμινοξέος με την συντήρηση ενός σπανίου (στη δεύτερη περίπτωση το σκορ θα είναι μεγαλύτερο). Όπως σε όλες τις περιπτώσεις με τα αντίστοιχα σκορ, ένα μικρό πρόβλημα μπορεί να προκύψει στις περιπτώσεις που ένα σύμβολο δεν εμφανίζεται καθόλου σε μια θέση, οπότε η σχέση (3.19) δεν ορίζεται και το σκορ γίνεται -∞. Τότε, υπάρχουν δύο εναλλακτικές. Αν δεν θέλουμε να επιτρέψουμε αυτό το σύμβολο να εμφανιστεί ποτέ, απλά αντικαθιστούμε την τιμή αυτή με έναν ιδιαίτερα μικρό αριθμό (π.χ. -10,000) και όλα λειτουργούν κανονικά, καθώς έστω και μια τέτοια εμφάνιση θα δώσει αρνητικό σκορ. Η άλλη εναλλακτική είναι να προσθέσουμε μικρές ψευδοτιμές, έτσι ώστε να καλύψουμε τη θεωρητική πιθανότητα το σύμβολο αυτό να έχει εμφανιστεί. Έτσι, η σχέση (3.19) θα γίνει:

$$s_b(i) = \log\left(\frac{p_b(i) + z_b}{p_b + \sum_{s=1}^{k} z_s}\right) \tag{5.4}$$

και με αυτόν τον τρόπο θα δοθούν μεγάλες αρνητικές τιμές στα σύμβολα που δεν εμφανίζονται σε κάποια θέση. Εναλλακτικά, οι ψευδοτιμές μπορούν να προστεθούν ευκολότερα στην εξίσωση (5.1) ως ακέραιες τιμές εμφάνισης των συμβόλων. Συνήθως, τέτοιοι πίνακες στρογγυλοποιούνται σε ακέραιες τιμές, για ακόμα μεγαλύτερη ευκολία στους υπολογισμούς. Με αυτόν τον ορισμό, αλληλουχίες με θετικό σκορ έχουν καλή ταύτιση με το μοντέλο, ενώ αλληλουχίες με αρνητικό σκορ θεωρείται ότι δεν έχουν.

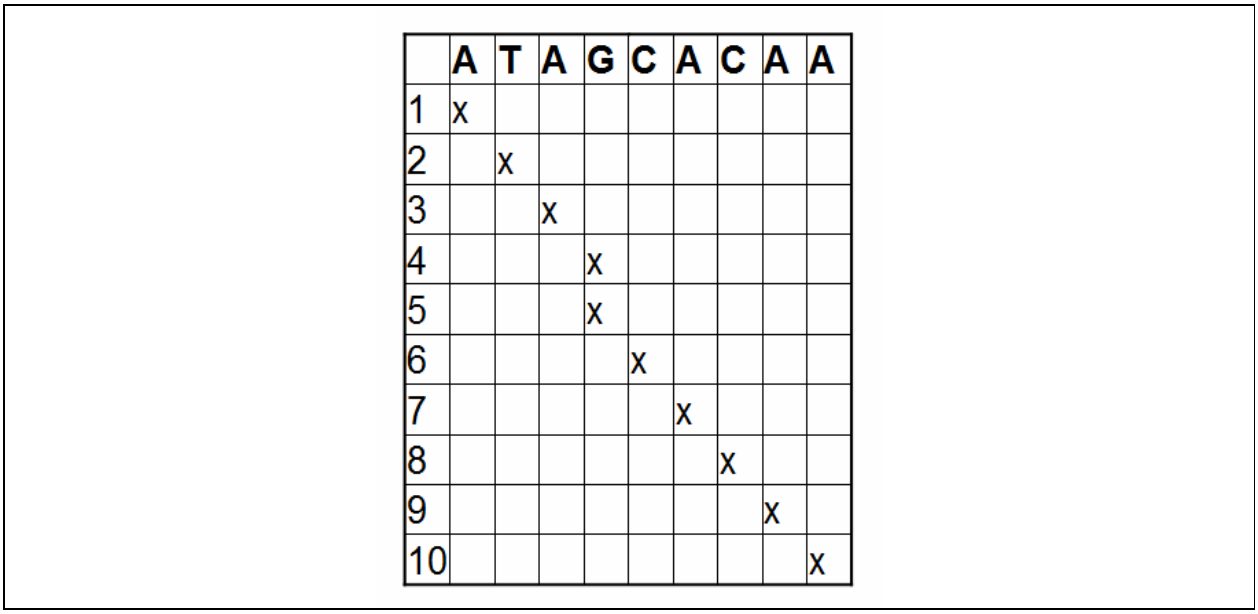


	1	2	3	4	5	6	7	8	9	10
A	1.000	0.000	0.333	0.000	0.500	0.000	0.667	0.000	0.167	1.000
T	0.000	1.000	0.667	0.000	0.167	0.000	0.000	0.000	0.333	0.000
G	0.000	0.000	0.000	1.000	0.167	0.000	0.167	0.000	0.333	0.000
C	0.000	0.000	0.000	0.000	0.167	1.000	0.167	1.000	0.167	0.000

	1	2	3	4	5	6	7	8	9	10
A	1.387	-∞	0.393	-∞	0.693	-∞	0.981	-∞	-0.405	1.387
T	-∞	1.387	0.981	-∞	-0.405	-∞	-∞	-∞	0.393	-∞
G	-∞	-∞	-∞	1.387	-0.405	-∞	-0.405	-∞	0.393	-∞
C	-∞	-∞	-∞	-∞	-0.405	1.000	-0.405	1.387	-0.405	-∞

Εικόνα 5.7: *Ένα παράδειγμα δημιουργίας σταθμισμένου πίνακα (weight matrix) και πίνακα σκορ ειδικού ανά θέση (PSSM), από μια πολλαπλή στοίχιση.*

Οι πίνακες αυτοί, έχουν πάρα πολλές εφαρμογές και σε πολλές περιπτώσεις έχουν αντικαταστήσει τα κλασικά πρότυπα ακριβώς λόγω της ευελιξίας τους. Ανάλογα με το πρόβλημα, μπορεί να υπάρχουν και επιπλέον διαφοροποιήσεις. Για παράδειγμα, η πιο απλή επιλογή είναι να έχουμε κατασκευάσει έναν τέτοιον πίνακα και απλά να κάνουμε μια αναζήτηση ελέγχοντας διαδοχικά τα επικαλυπτόμενα παράθυρα κατά μήκος της αλληλουχίας (Staden, 1990). Αυτό ισοδυναμεί με την υπόθεση ότι το προφίλ που αναζητάμε αναμένουμε να έχει ακριβώς τις ίδιες θέσεις με το αρχικό (μια συνηθισμένη υπόθεση όταν ψάχνουμε για μια καλά χαρακτηρισμένη από λειτουργικής άποψης περιοχή, π.χ. το ενεργό κέντρο ενός ενζύμου ή τη θέση πρόσδεσης ενός μεταγραφικού παράγοντα). Σε άλλες περιπτώσεις μπορεί να ενδιαφερόμαστε για κάτι πιο γενικό, οπότε μπορεί να μας ενδιαφέρει να έχουμε ευελιξία και να επιτρέπουμε κενά (τόσο στην αλληλουχία, όσο και στο προφίλ) (Barton & Sternberg, 1990). Αυτό επιτυγχάνεται με μια μικρή επέκταση των κλασικών αλγορίθμων δυναμικού προγραμματισμού που έχουμε γνωρίσει για την περίπτωση στοίχισης δύο αλληλουχιών. Η διαφορά είναι ότι σε αυτή την εκδοχή αντί να έχουμε στοίχιση αλληλουχίας με αλληλουχία, θα έχουμε τη στοίχιση της αλληλουχίας με το προφίλ. Προφανώς, χρειάζεται και σε αυτή την περίπτωση μια καλά υπολογισμένη, εμπειρικά, ποινή για τα κενά.



	A	T	A	G	C	A	C	A	A
1	x								
2		x							
3			x						
4				x					
5				x					
6					x				
7						x			
8							x		
9								x	
10									x

Εικόνα 5.8: *Στοίχιση μια αλληλουχίας με ένα προφίλ.*

Ειδικά στις πρωτεΐνες, είναι δυνατό να κατασκευαστεί ένα ακόμα πιο ευαίσθητο σύστημα για το σκορ, ικανό να εντοπίζει και μακρινές ομοιότητες. Η μέθοδος αυτή ονομάζεται profile analysis και ήταν μια από τις πρώτες και πολύ ικανοποιητικές προσεγγίσεις στον εντοπισμό μακρινών ομολόγων (Gribskov, McLachlan, & Eisenberg, 1987). Η ιδέα είναι να φτιαχτεί ένας ειδικός ανά θέση πίνακας του σκορ (position specific scoring matric-PSSM), ο οποίος θα μπορεί να χρησιμοποιηθεί αντί των κλασικών πινάκων ομοιότητας (PAM, BLOSUM κλπ) σε μια κλασική μέθοδο στοίχισης. Αρχικά, ξεκινάμε με μια αλληλουχία και εντοπίζουμε τις ομόλογες. Από αυτές, κατασκευάζουμε μια πολλαπλή στοίχιση από την οποία κατασκευάζουμε όμοια με προηγουμένως τον πίνακα με τις πιθανότητες εμφάνισης κάθε καταλοίπου. Βασικό σημείο που χρειάζεται προσοχή εδώ, είναι το γεγονός ότι ο πίνακας έχει τόσες θέσεις, όσο είναι και το μήκος της αρχικής αλληλουχίας. Αυτό συμβαίνει γιατί στήλες στην πολλαπλή στοίχιση που περιέχουν τυχόν κενά στην αρχική αλληλουχία, αγνοούνται. Με άλλα λόγια, η αλληλουχία «μετατρέπεται» σε έναν πίνακα που περιέχει πληροφορίες από όλες τις ομόλογές της και με τον τρόπο αυτόν πετυχαίνουμε μεγαλύτερη ευαισθησία στις αναζητήσεις. Φυσικά, η μέθοδος είναι πιο γενική και μπορεί να χρησιμοποιηθεί και για κατασκευή μοντέλου από μια οποιαδήποτε πολλαπλή στοίχιση, μόνο που τότε θα πρέπει να αποφασιστεί ποιες στήλες δεν θα συμπεριληφθούν στο μοντέλο (αυτές που έχουν κενά περισσότερα από μια προκαθορισμένη τιμή).

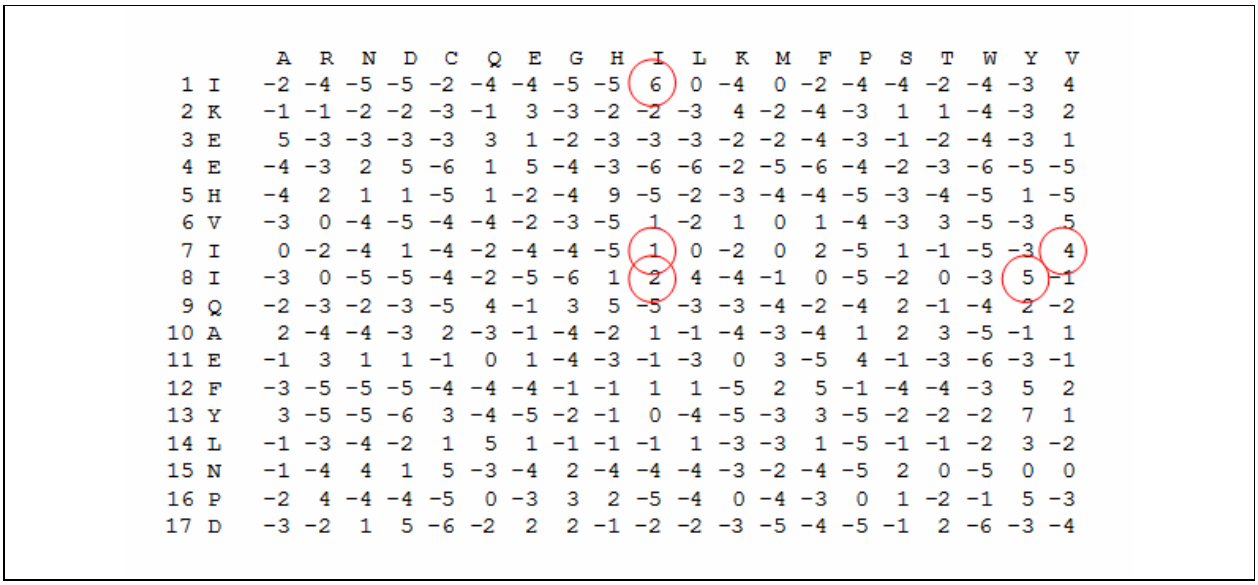
Στον υπολογισμό του σκορ, η βασική διαφορά από την κλασική μέθοδο, έγκειται στο ότι σε κάθε θέση η τιμή του σκορ δίνεται από ένα μέσο όρο όλων των τιμών που προβλέπει ένας κλασικός πίνακας του σκορ για τις συγκρίσεις αλληλουχιών. Έτσι, θα έχουμε:

$$s_b(i) = \sum_{j=1}^{k} p_j(i) S_{bj} \tag{5.5}$$

όπου $p_j(i)$ είναι όμοια με παραπάνω η πιθανότητα εμφάνισης του αμινοξέος j στη θέση i της πολλαπλής στοίχισης (που θα αντιστοιχεί στην ομόλογη θέση της αρχικής αλληλουχίας), ενώ το S_{bj} είναι η τιμή που προβλέπει ο επιλεγμένος πίνακας ομοιότητας (πχ BLOSUM62) για τη σύγκριση των αμινοξέων b και j. Από τον παραπάνω τρόπο υπολογισμού του σκορ, καταλαβαίνουμε ότι ακόμα και αν ένα αμινοξύ δεν εμφανίζεται καθόλου σε μια δεδομένη θέση της πολλαπλής στοίχισης, θα μπορεί να έχει Παρ' όλα αυτά θετική τιμή του σκορ, καθώς θα δεχτεί θετικές συνεισφορές από τα αμινοξέα με τα οποία έχει θετική τιμή στον επιλεγμένο πίνακα ομοιότητας. Στην αρχική εργασία, οι συγγραφείς χρησιμοποίησαν πίνακα ομοιότητας της οικογένειας PAM, αλλά περαιτέρω αναλύσεις έδειξαν ότι ο BLOSUM45 είναι καλύτερος, ενώ επιπλέον η διαφορική στάθμιση των αλληλουχιών, έτσι ώστε οι πολύ όμοιες να συνεισφέρουν λιγότερο στον πίνακα, βελτιώνει τη μεθοδολογία (Lüthy, Xenarios, & Bucher, 1994).

Το σκορ από τις αναζητήσεις με προφίλ, ακολουθεί την κατανομή του Gumbel, όμοια με τη στοίχιση αλληλουχιών. Στις περισσότερες περιπτώσεις, οι παράμετροι της κατανομής υπολογίζονται με

προσομοιώσεις. Σε μερικές απλές περιπτώσεις όμως, όπως για παράδειγμα σε περιπτώσεις χρήσης σταθμισμένου πίνακα και στοίχισης χωρίς κενά, χρησιμοποιούνται και εμπειρικοί κανόνες (π.χ. το σκορ πρέπει να είναι μεγαλύτερο από το 60% της μέγιστης τιμής που προβλέπει ο πίνακας).



		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	I	-2	-4	-5	-5	-2	-4	-4	-5	-5	6	0	-4	0	-2	-4	-4	-2	-4	-3	4
2	K	-1	-1	-2	-2	-3	-1	3	-3	-2	-2	-3	4	-2	-4	-3	1	1	-4	-3	2
3	E	5	-3	-3	-3	-3	3	1	-2	-3	-3	-3	-2	-2	-4	-3	-1	-2	-4	-3	1
4	E	-4	-3	2	5	-6	1	5	-4	-3	-6	-6	-2	-5	-6	-4	-2	-3	-6	-5	-5
5	H	-4	2	1	1	-5	1	-2	-4	9	-5	-2	-3	-4	-4	-5	-3	-4	-5	1	-5
6	V	-3	0	-4	-5	-4	-4	-2	-3	-5	1	-2	1	0	1	-4	-3	3	-5	-3	5
7	I	0	-2	-4	1	-4	-2	-4	-4	-5	1	0	-2	0	2	-5	1	-1	-5	-3	4
8	I	-3	0	-5	-5	-4	-2	-5	-6	1	2	4	-4	-1	0	-5	-2	0	-3	5	-1
9	Q	-2	-3	-2	-3	-5	4	-1	3	5	-5	-3	-3	-4	-2	-4	2	-1	-4	2	-2
10	A	2	-4	-4	-3	2	-3	-1	-4	-2	1	-1	-4	-3	-4	1	2	3	-5	-1	1
11	E	-1	3	1	1	-1	0	1	-4	-3	-1	-3	0	3	-5	4	-1	-3	-6	-3	-1
12	F	-3	-5	-5	-5	-4	-4	-4	-1	-1	1	1	-5	2	5	-1	-4	-4	-3	5	2
13	Y	3	-5	-5	-6	3	-4	-5	-2	-1	0	-4	-5	-3	3	-5	-2	-2	-2	7	1
14	L	-1	-3	-4	-2	1	5	1	-1	-1	-1	1	-3	-3	1	-5	-1	-1	-2	3	-2
15	N	-1	-4	4	1	5	-3	-4	2	-4	-4	-4	-3	-2	-4	-5	2	0	-5	0	0
16	P	-2	4	-4	-4	-5	0	-3	3	2	-5	-4	0	-4	-3	0	1	-2	-1	5	-3
17	D	-3	-2	1	5	-6	-2	2	2	-1	-2	-2	-3	-5	-4	-5	-1	2	-6	-3	-4

Εικόνα 5.9: *Ένα παράδειγμα PSSM. Συνήθως για ευκολία αλλάζουμε τις στήλες με τις γραμμές έτσι ώστε όλοι οι πίνακες να έχουν 20 στήλες αλλά τόσες γραμμές όσα είναι και τα αμινοξέα της πρωτεΐνης. Παρατηρήστε ότι το ίδιο αμινοξύ, π.χ. η Ισολευκίνη μπορεί να έχει σε διαφορετικές θέσεις τελείως διαφορετικό διάνυσμα, καθώς στις αντίστοιχες στήλες της πολλαπλής στοίχισης υπήρχαν διαφορετικά αμινοξέα. Ειδικά στις θέσεις 7 και 8, η πρωτεΐνη μας έχει Ισολευκίνη, αλλά οι περισσότερες πρωτεΐνες της στοίχισης έχουν Βαλίνη και Τυροσίνη, αντίστοιχα.*

5.3. Λογισμικό

Στην ενότητα αυτή, θα παρουσιάσουμε τα πιο γνωστά πακέτα λογισμικού που χρησιμοποιούνται είτε για να κατασκευάζουν PSSMs, είτε για να κάνουν αναζητήσεις. Το πιο γνωστό πρόγραμμα της πρώτης κατηγορίας είναι το **ScanProsite** (http://prosite.expasy.org/scanprosite/). Το ScanProsite είναι κατασκευασμένο για να εντοπίζει πρότυπα και προφίλ της PROSITE, σε οποιαδήποτε αλληλουχία, είτε του χρήστη, είτε κάποια που έχει επιλεγεί από μια βάση δεδομένων. Είναι το εργαλείο που χρησιμοποιείται επίσημα στις αναζητήσεις στην PROSITE και έχει πολλές βελτιστοποιήσεις για να αυξάνεται η ταχύτητα, όπως προϋπολογισμένες ταυτίσεις για τις γνωστές αλληλουχίες κ.ο.κ. (De Castro et al., 2006)

Το **PFTOOLS** (http://web.expasy.org/pftools/) ειναι ένα εργαλείο κατάλληλο τόσο για κατασκευή όσο και για αναζήτηση προφίλ από στοιχισμένες αλληλουχίες (Bucher, Karplus, Moeri, & Hofmann, 1996). Το PFTOOLS είναι πολύ γενικό, και περιλαμβάνει όλες τις περιπτώσεις προφίλ που αναφέραμε στην προηγούμενη ενότητα (πρότυπα, weight matrices, PSSMs), ενώ ενσωματώνει και την πιο γενική περίπτωση στην οποία όλες οι ποινές για τα κενά είναι επίσης ειδικές ανά θέση (generalized profile). Η τελευταία περίπτωση, απέχει ένα μόνο βήμα πριν από το Hidden Markov Model το οποίο θα εξετάσουμε στο κεφάλαιο 8. Το PFTOOLS χρησιμοποιείται κυρίως για την κατασκευή μοντέλων για πρωτεϊνικές οικογένειες, χρησιμοποιώντας μια πολλαπλή στοίχιση των μελών της οικογένειας και διαθέτει διάφορες ρουτίνες, όπως:

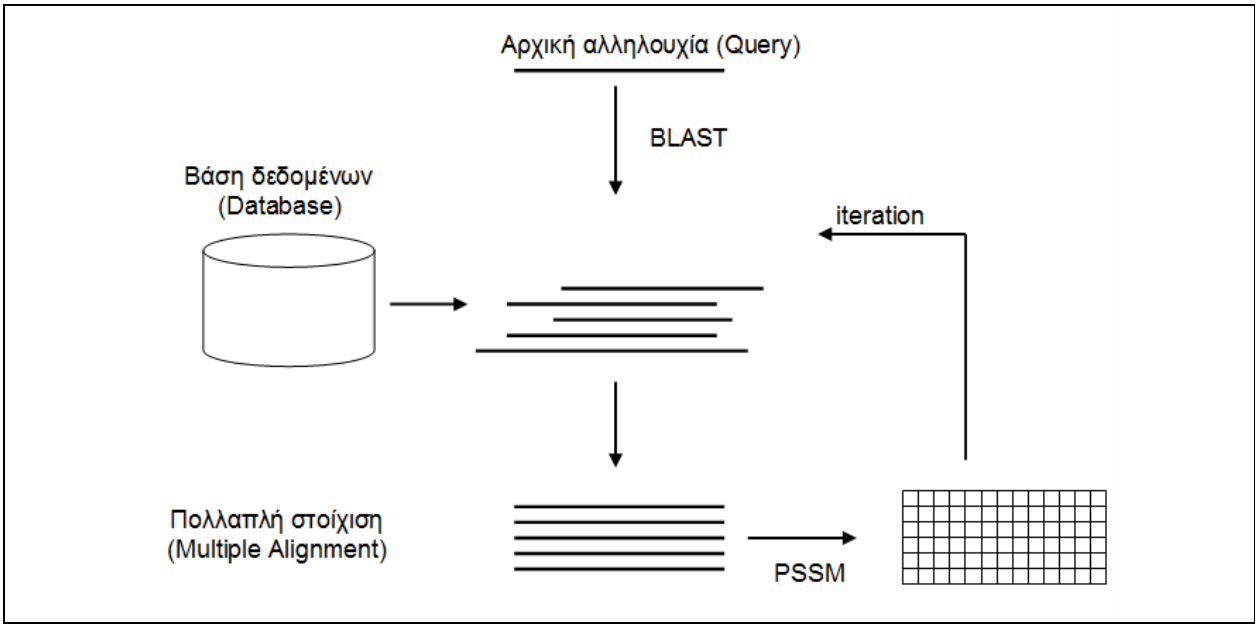
- **pfmake**: κατασκευάζει ένα προφίλ από μια δεδομένη πολλαπλή στοίχιση
- **pfscale**: βρίσκει τις παραμέτρους της κατανομής του Gumbel για να υπολογιστεί η στατιστική σημαντικότητα
- **pfw**: εφαρμόζει τη μέθοδο της διαφορικής στάθμισης των αλληλουχιών για να διορθώσει το συστηματικό σφάλμα από την υπερ-αντιπροσώπευση κάποιων μελών της οικογένειας.
- **pfsearch**: πραγματοποιεί αναζήτηση σε μια βάση δεδομένων αλληλουχιών πρωτεϊνών ή DNA έναντι σε ένα προφίλ.
- **pfscan**: πραγματοποιεί αναζήτηση μιας αλληλουχίας DNA ή πρωτεΐνης έναντι σε μια βιβλιοθήκη με προφίλ.

Επίσης, υπάρχουν μια σειρά από βοηθητικά προγράμματα που μετατρέπουν τα μοντέλα και τις ακολουθίες από και προς διάφορες άλλες γνωστές μορφές, μεταξύ των οποίων συμπεριλαμβάνεται και η μορφή HMMER που θα συναντήσουμε στο κεφάλαιο 8 (**psa2msa**, **gtop**, **htop**, **ptoh**) ή μετατρέπουν τις αλληλουχίες από DNA σε πρωτεϊνικές και το αντίστροφο (**ptof**, **2ft**, **6ft**).

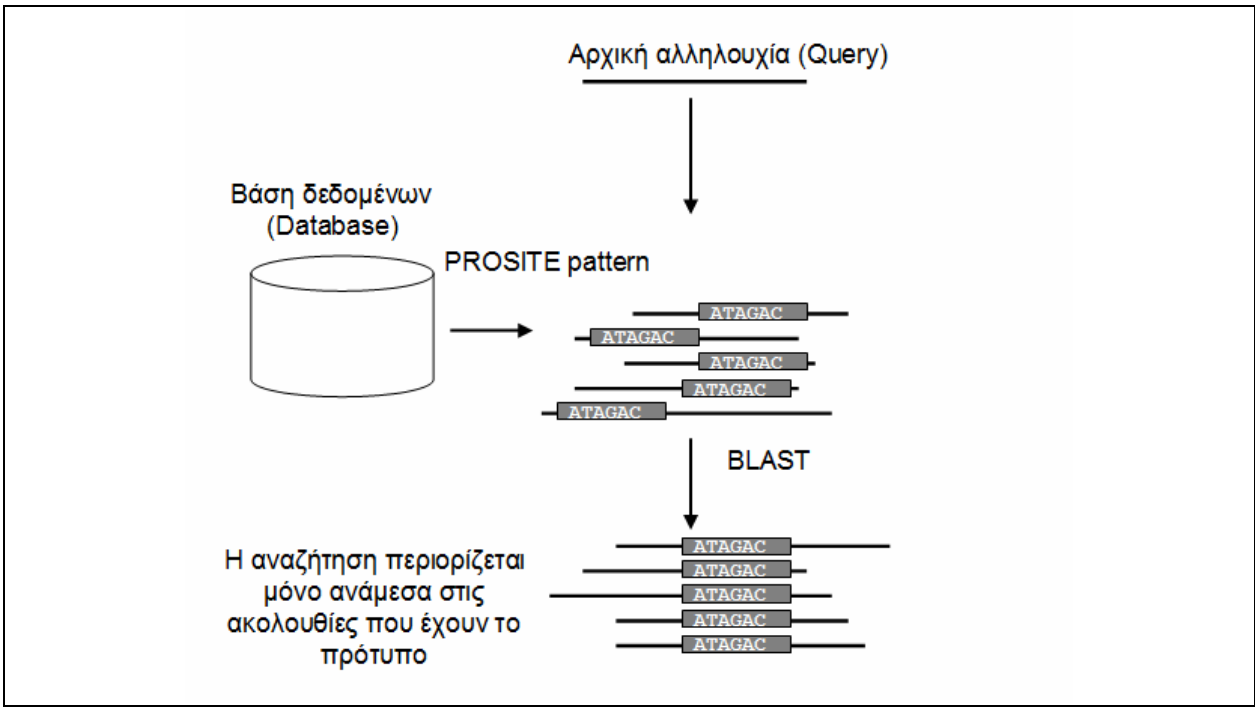
Ίσως η πιο ευρέως χρησιμοποιούμενη εφαρμογή που κάνει χρήση PSSM είναι το **PSI-BLAST** (Position-specific-iterated BLAST) (Altschul et al., 1997). Είναι μια επέκταση του γνωστού αλγορίθμου BLAST και χρησιμοποιείται για την εύρεση μακρινών ομολόγων. Η μέθοδος δουλεύει ως εξής (Εικόνα 5.10): Στην αρχή πραγματοποιείται μια κανονική αναζήτηση με το BLAST και συλλέγονται οι αλληλουχίες με E-value μικρότερο από κάποιο όριο που ορίζεται από τον χρήστη. Αυτές θεωρείται ότι είναι οι «σίγουρες» ομόλογες και χρησιμοποιούνται για να κατασκευαστεί ένας PSSM όπως περιγράψαμε παραπάνω, χωρίς όμως κενά καθώς κάθε στήλη του αντιστοιχεί σε μια θέση της αλληλουχίας της αρχικής πρωτεΐνης. Με αυτόν τον πίνακα, πραγματοποιείται εκ νέου αναζήτηση στη βάση δεδομένων, η οποία πλέον θα δώσει περισσότερες ομόλογες με E-value μικρότερο από το αρχικό όριο. Η διαδικασία αυτή επαναλαμβάνεται αρκετές φορές, είτε μέχρι να σταματήσουν να προστίθενται νέες αλληλουχίες, είτε μέχρι να ξεπεραστεί ένας συγκεκριμένος αριθμός επαναλήψεων (συνήθως 3 ή 4). Η μέθοδος είναι εξαιρετικά αποδοτική και εντοπίζει μεγάλο αριθμό ομόλογων πρωτεϊνών (μακρινών ομολόγων), οι οποίες δεν θα μπορούσαν να εντοπιστούν με μια συμβατική αναζήτηση. Η επαναληπτική αυτή διαδικασία, θυμίζει τον αλγόριθμο EM, και οι μόνες περιπτώσεις στις οποίες μπορεί να αποτύχει είναι είτε όταν δεν βρεθούν καθόλου ομόλογες στην πρώτη αναζήτηση, είτε όταν το όριο είναι αρκετά ψηλά με συνέπεια να συμπεριληφθούν και πρωτεΐνες που δεν έχουν πραγματική ομολογία, οπότε και το προφίλ δεν θα είναι πλέον ειδικό αρκετά (contamination).

Μια ενδιαφέρουσα επέκταση του PSI-BLAST είναι το **DELTA-BLAST** (domain enhanced lookup time accelerated BLAST), το οποίο αντί να κατασκευάσει το PSSM από την αρχή, πραγματοποιεί αναζήτηση σε μια βάση δεδομένων με ήδη χαρακτηρισμένες οικογένειες έτσι ώστε να πετύχει καλύτερη ακρίβεια στην αναγνώριση. Για το σκοπό αυτό, χρησιμοποιεί τη βάση Conserved Domain Database (CDD) του NCBI, και τα αποτελέσματα δείχνουν ότι με τη μέθοδο αυτή, πετυχαίνουμε καλύτερα αποτελέσματα από το PSI-BLAST, καθώς συνδυάζονται τα πλεονεκτήματα της επαναληπτικής διαδικασίας με αυτά της χρήσης της καλά χαρακτηρισμένης βάσης δεδομένων (Boratyn et al., 2012).

Το **PHI-BLAST** (pattern-hit initiated BLAST) είναι άλλη μια παραλλαγή του BLAST, η οποία όμως χρησιμοποιεί πρότυπα κανονικών εκφράσεων (Zhang et al., 1998). Η ιδέα εδώ είναι διαφορετική και συνίσταται στη χρησιμοποίηση γνωστών πρότυπων, τα οποία υπάρχουν στην αλληλουχία επερώτησης και τα καθορίζει ο χρήστης, για να καθοδηγήσουν την αναζήτηση. Με τον τρόπο αυτό, το εύρος της αναζήτησης περιορίζεται και σε πολλές περιπτώσεις εντοπίζονται ομόλογες πρωτεΐνες οι οποίες δεν μπορούσαν να εντοπιστούν με το συμβατικό τρόπο αναζήτησης (Εικόνα 5.11).



Εικόνα 5.10: *Σχηματικό διάγραμμα αναπαράστασης της λειτουργίας του PSI-BLAST.*



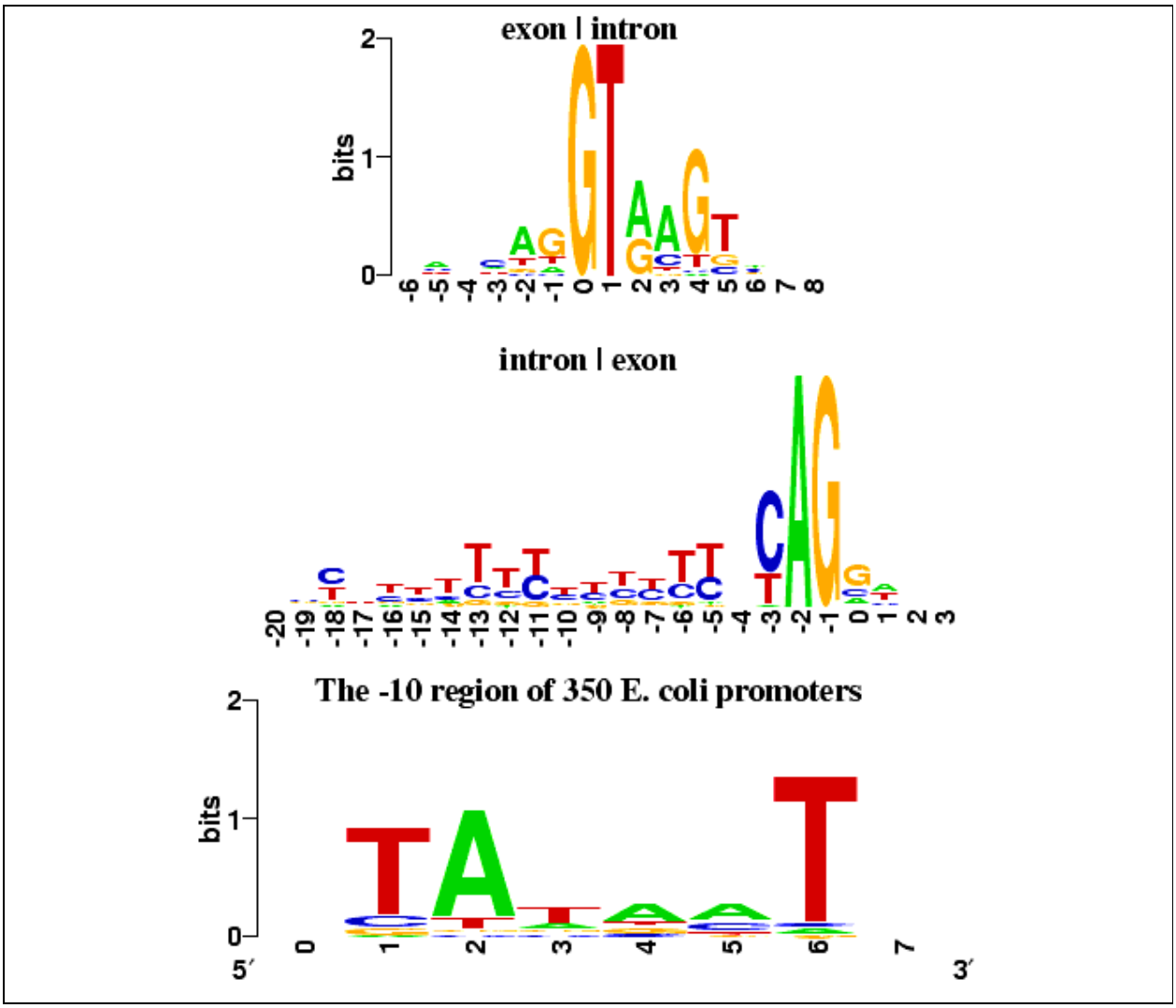
Εικόνα 5.11: *Σχηματικό διάγραμμα αναπαράστασης της λειτουργίας του PHI-BLAST.*

Τέλος, μια πολύ σημαντική εφαρμογή που χρησιμοποιείται για την οπτικοποίηση των περιοχών που απεικονίζονται σε ένα πρότυπο ή προφίλ, είναι το **WebLogo** (http://weblogo.berkeley.edu/) (Crooks, Hon, Chandonia, & Brenner, 2004). Το WebLogo βασίζεται στην απλή ιδέα των Λογότυπων Αλληλουχιών (Sequence Logo) των Schneider και Stephens (Schneider & Stephens, 1990) και απεικονίζει μια πολλαπλή στοίχιση σε μια γραφική αναπαράσταση, με στήλες στις οποίες εμφανίζονται τοποθετημένα κάθετα τα σύμβολα που εμφανίζονται σε αυτή. Το ύψος της στήλης αντιστοιχεί στη συνολική πληροφορία που φέρει η στήλη αυτή, και δίνεται από τον τύπο:

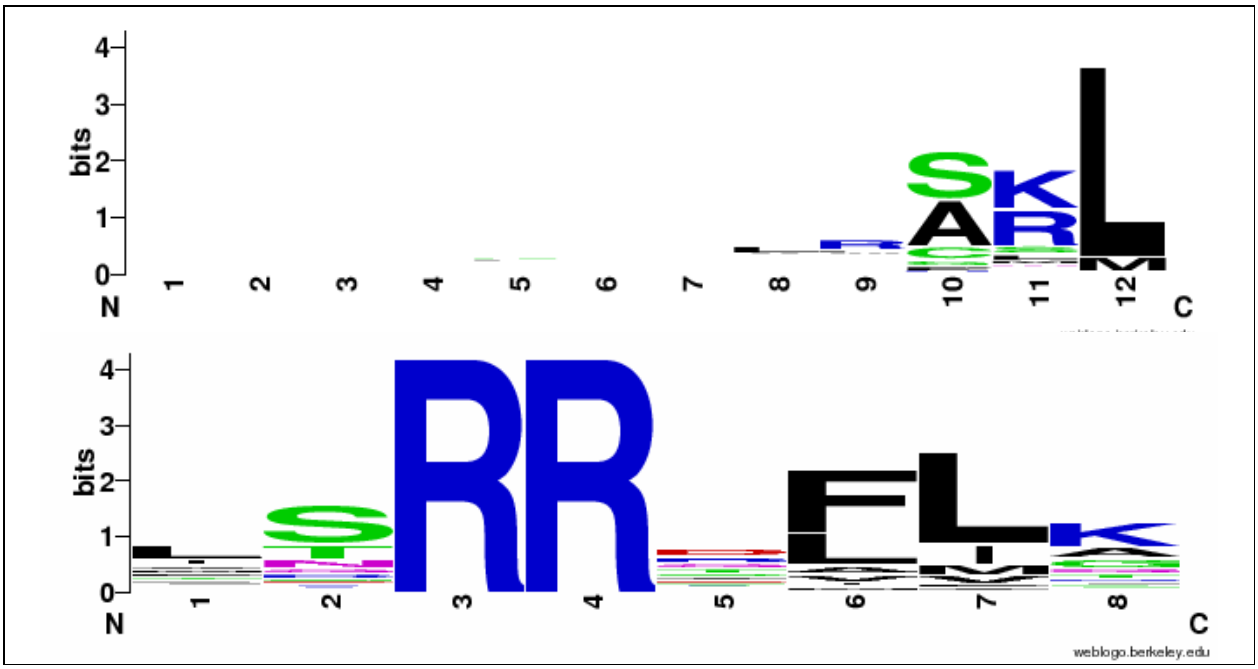
$$R = S_{\max} - S_{obs} = \log_2 k - \left(-\sum_{\forall b \in \Omega} n_b(i) \log p_b(i) \right) \tag{5.6}$$

Στη σχέση αυτή, το S_{max} είναι η μέγιστη εντροπία που μπορεί να έχει η στήλη και S_{obs} η παρατηρηθείσα εντροπία που είδαμε στο κεφάλαιο 3. Το k είναι το μέγεθος του αλφάβητου, το οποίο και καθορίζει τη μέγιστη τιμή (2.2 bits για DNA/RNA και ~4.32 για πρωτεΐνες). Το σχετικό ύψος του κάθε συμβόλου σε κάθε στήλη, δίνεται από τη συχνότητα εμφάνισής του. Το λογισμικό δέχεται σαν είσοδο μια πολλαπλή στοίχιση και παράγει τη γραφική παράσταση, η οποία είναι ιδιαίτερα κατατοπιστική καθώς μας δείχνει με μια γρήγορη ματιά ποιες στήλες είναι συντηρημένες, αλλά και ποια σύμβολα επικρατούν σε κάθε μια από αυτές. Οι στήλες που δεν έχουν ιδιαίτερη συντήρηση, εμφανίζονται σύμφωνα με τη σχέση (5.6) με μικρό ύψος.

Εικόνα 5.12: *Το Sequence Logo της πολλαπλής στοίχισης από την Εικόνα 5.1.*



Εικόνα 5.13: *Παραδείγματα Sequence Logo από αλληλουχίες DNA. Πάνω, απεικονίζονται τα λογότυπα των περιοχών εναλλαγής εσωνίων-εξωνίων, όπως προκύπτουν από τις πειραματικά προσδιορισμένες αλληλουχίες της EID (Exon-Intron database). Κάτω, απεικονίζεται η περιοχή του υποκινητή από 350 γονίδια της E. coli. Παρατηρήστε ότι παρόλο που οι περιοχές αυτές περιγράφονται και από πρότυπα, σε ένα μεγάλο σύνολο δεδομένων, λίγες είναι οι στήλες με απόλυτη συντήρηση.*



Εικόνα 5.14: *Παραδείγματα Sequence Logo από αλληλουχίες πρωτεϊνών. Πάνω απεικονίζονται τα λογότυπα των καρβοξυτελικών περιοχών που περιέχουν το σήμα στόχευσης για το υπεροξείσωμα (PTS1). Κάτω, απεικονίζεται η περιοχή των σηματοδοτικών αλληλουχιών από τις βακτηριακές πρωτεΐνες που εκκρίνονται με το μονοπάτι TAT. Παρατηρήστε ότι παρόλο που οι περιοχές αυτές περιγράφονται και από πρότυπα PROSITE, σε ένα μεγάλο σύνολο δεδομένων λίγες είναι οι στήλες με απόλυτη συντήρηση.*

Βιβλιογραφία

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research, 25*(17), 3389-3402.

Bailey, T. L., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol, 2*, 28-36.

Barton, G. J., & Sternberg, M. J. (1990). Flexible protein sequence patterns: A sensitive method to detect weak structural similarities. *Journal of molecular biology, 212*(2), 389-402.

Berven, F. S., Flikka, K., Jensen, H. B., & Eidhammer, I. (2004). BOMP: a program to predict integral b-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. *Nucleic Acids Res, 32*(Web Server Issue), W394-W399.

Boratyn, G. M., Schaffer, A. A., Agarwala, R., Altschul, S. F., Lipman, D. J., & Madden, T. L. (2012). Domain enhanced lookup time accelerated BLAST. *Biol Direct, 7*(1), 12.

Brazma, A., Jonassen, I., Eidhammer, I., & Gilbert, D. (1998). Approaches to the automatic discovery of patterns in biosequences. *Journal of computational biology, 5*(2), 279-305.

Bucher, P., Karplus, K., Moeri, N., & Hofmann, K. (1996). A flexible motif search technique based on generalized profiles. *Computers & chemistry, 20*(1), 3-23.

Cokol, M., Nair, R., & Rost, B. (2000). Finding nuclear localization signals. *EMBO reports, 1*(5), 411-415.

Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res, 14*(6), 1188-1190.

De Castro, E., Sigrist, C. J., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P. S., Gasteiger, E., . . . Hulo, N. (2006). ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Research, 34*(suppl 2), W362-W365.

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

Jonassen, I., Collins, J. F., & Higgins, D. G. (1995). Finding flexible patterns in unaligned protein sequences. *Protein Science, 4*(8), 1587-1595.

Lüthy, R., Xenarios, I., & Bucher, P. (1994). Improving the sensitivity of the sequence profile method. *Protein Science, 3*(1), 139-146.

Petriv, I., Tang, L., Titorenko, V. I., & Rachubinski, R. A. (2004). A new definition for the consensus sequence of the peroxisome targeting signal type 2. *Journal of molecular biology, 341*(1), 119-134.

Rigoutsos, I., & Floratos, A. (1998). Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics, 14*(1), 55-67.

Schneider, T. D., & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res, 18*(20), 6097-6100.

Shruthi, H., Babu, M. M., & Sankaran, K. (2010). TAT-pathway-dependent lipoproteins as a niche-based adaptation in prokaryotes. *Journal of Molecular Evolution, 70*(4), 359-370.

Sigrist, C. J., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., & Hulo, N. (2010). PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res, 38*(Database issue), D161-166.

Staden, R. (1990). Searching for patterns in protein and nucleic acid sequences. *Methods in enzymology, 183*, 193-211.

Sutcliffe, I. C., & Harrington, D. J. (2002). Pattern searches for the identification of putative lipoprotein genes in Gram-positive bacterial genomes. *Microbiology, 148*(Pt 7), 2065-2077.

Thompson, W., Rouchka, E. C., & Lawrence, C. E. (2003). Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Research, 31*(13), 3580-3585.

Zhang, Z., Miller, W., Schäffer, A. A., Madden, T. L., Lipman, D. J., Koonin, E. V., & Altschul, S. F. (1998). Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Research, 26*(17), 3986-3990.