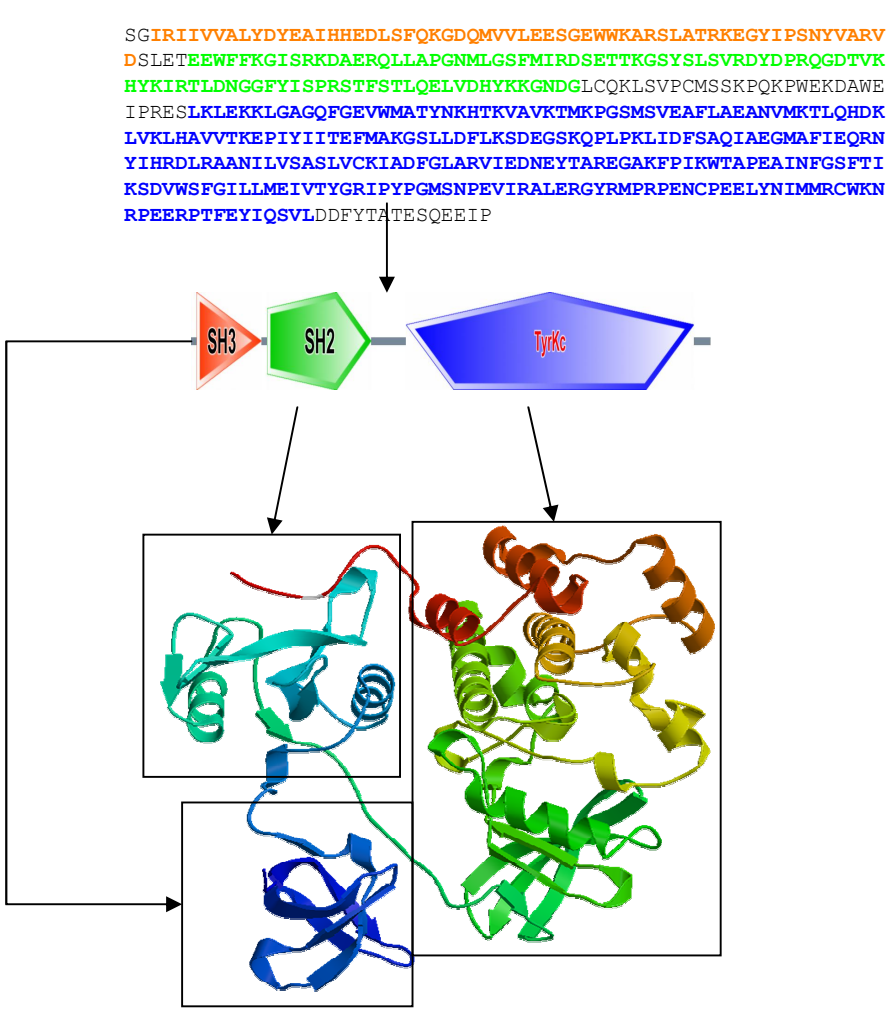
Ειδικά Θέματα Βιοπληροφορικής

Παντελής Μπάγκος
Αναπληρωτής Καθηγητής

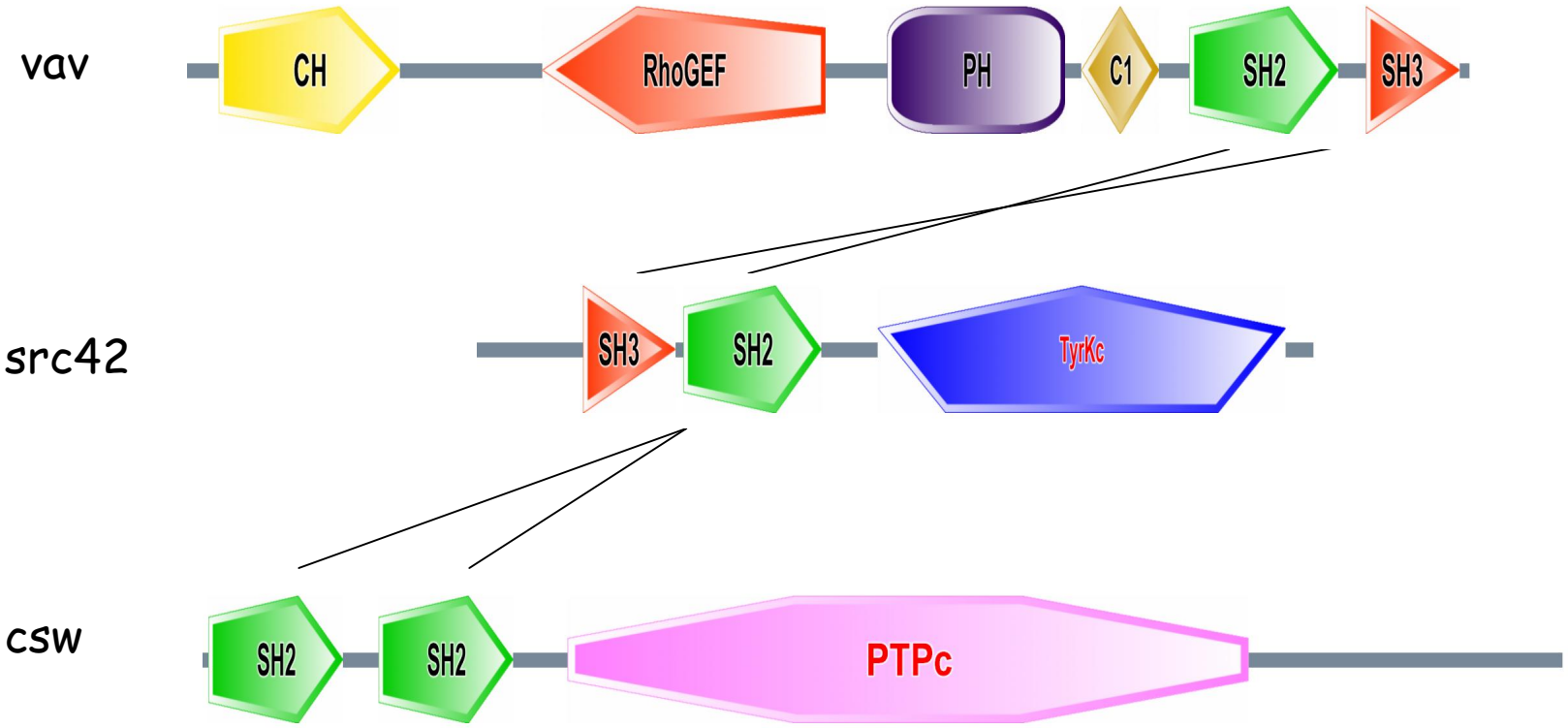
Πανεπιστήμιο Θεσσαλίας
Λαμία, 2015

The modular nature of proteins



Protein Domains

Local Similarity



Patterns

A T A G A C A C A A

A T T G T C A C T A

A T T G A C G C T A

A T A G G C A C G A

A T T G A C C C C A

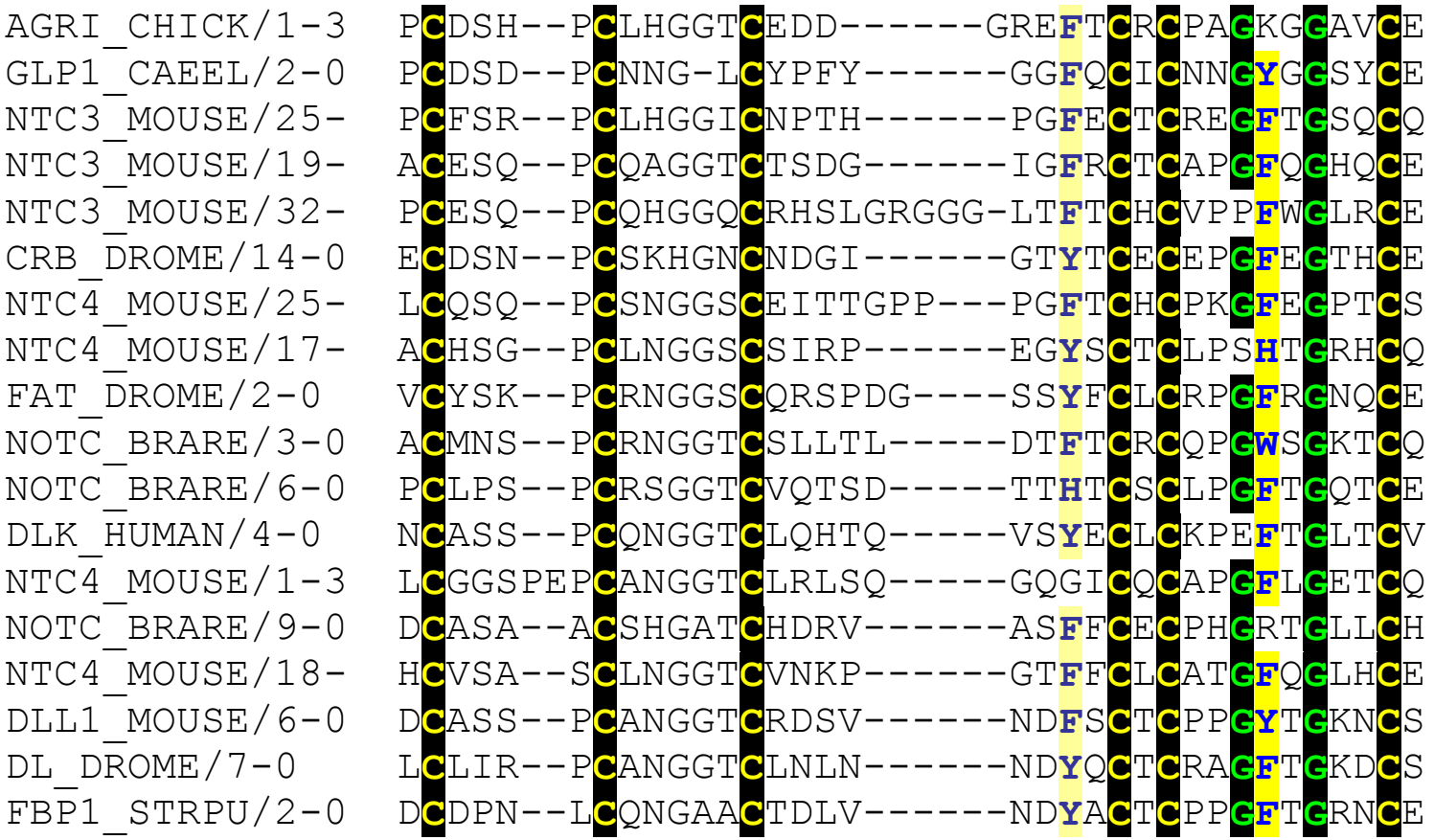
A T T G C C A C G A

A-T-[AT]-G-x-C-[AGC]-C-x-A

```
A T A G A C A C - - - A A
A T T G T C A C A - - T A
A T T G A C G C G C A T A
A T A G G C A C - - - G A
A T T G A C C C - - - C A
A T T G C C A C - - - G A
```

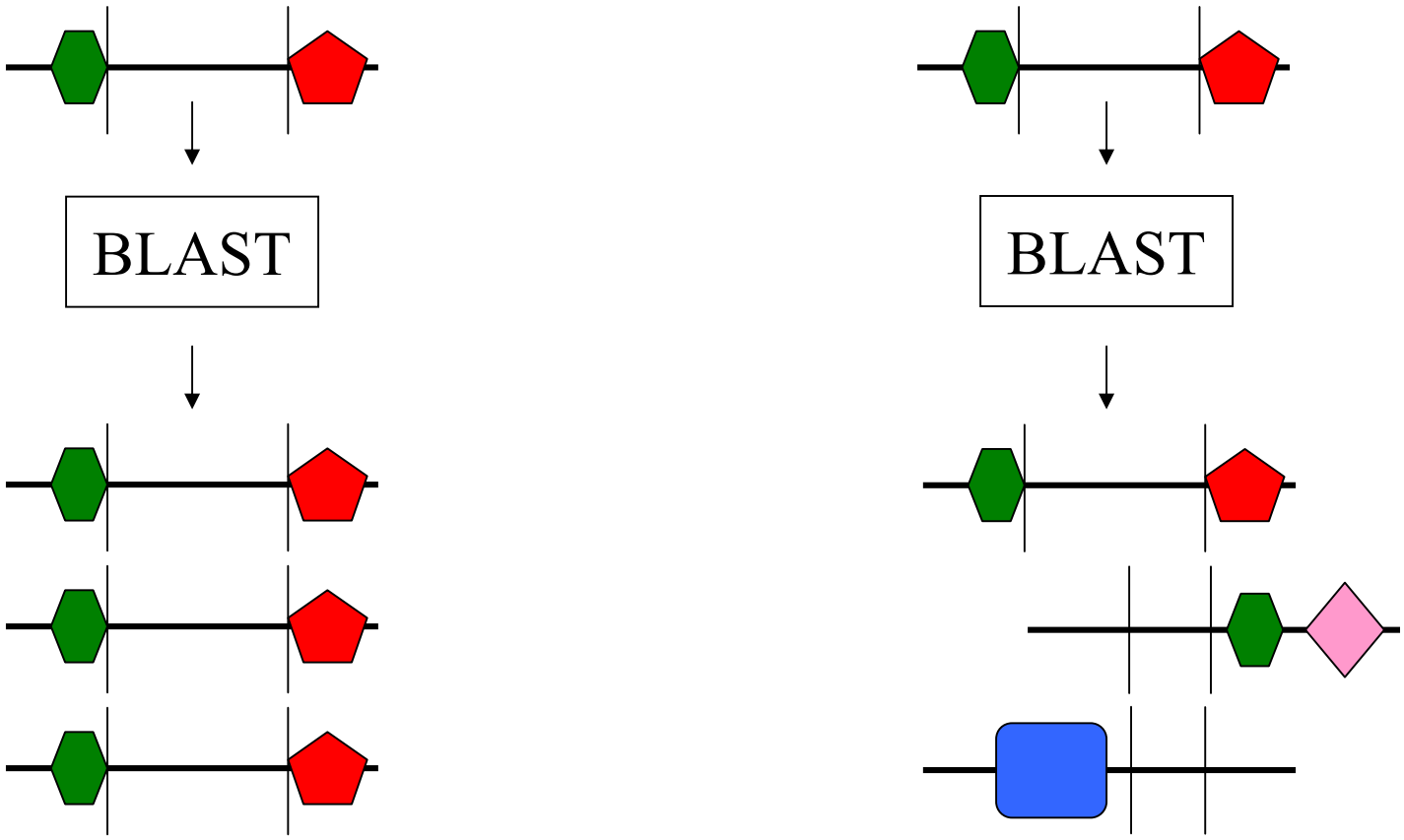
A-T-[AT]-G-x-C-[AGC]-C-x(1,3)-A

EGF domain



–C-x-C-x(5)-G-x(2)-C

Finding New Domains



Boring

Interesting

Κανόνες

- 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 άγκυστρο, 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.
- Επαναλήξεις συμβολίζονται με παρένθεση (). 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 άγκυστρα για να χαρακτηρίσει την πιθανή ύπαρξη καρβοξυτελικού άκρου. Έτσι, το P-R-L-[G>] σημαίνει είτε P-R-L-G ή P-R-L>.

Prosite

- 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.

Regular Expressions

- 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 άγκυστρο {}.
- 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]`

`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

`[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

`[CG]-A-G-G-T-[AG]-A-G`	Exon/Intron splice site
`[CT]-x-[CT]-A-G-[AG]`	Intron/Exon splice site
`A-[AU]-U-A-A-A`	Poly-A signal
`C-G-G-x(11)-C-C-G`	GAL4 binding site
`T-G-A-[GC]-T-C-[AT]-[TC]`	GCN4 binding site
`[TC]-T-A-A-T-T`	YOX1 binding site
`A-C-C-[CT]-T-[CAT]-A-A-G-G-G-x-[GAC]-T`	ZAP1 binding site
`T-C-A-C-T-G-x(80,100)-G-T` `-T-G-T-C-C-G-A-A-A-A`	Centromere

Πλεονεκτήματα

- 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 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. 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).

Μειονεκτήματα

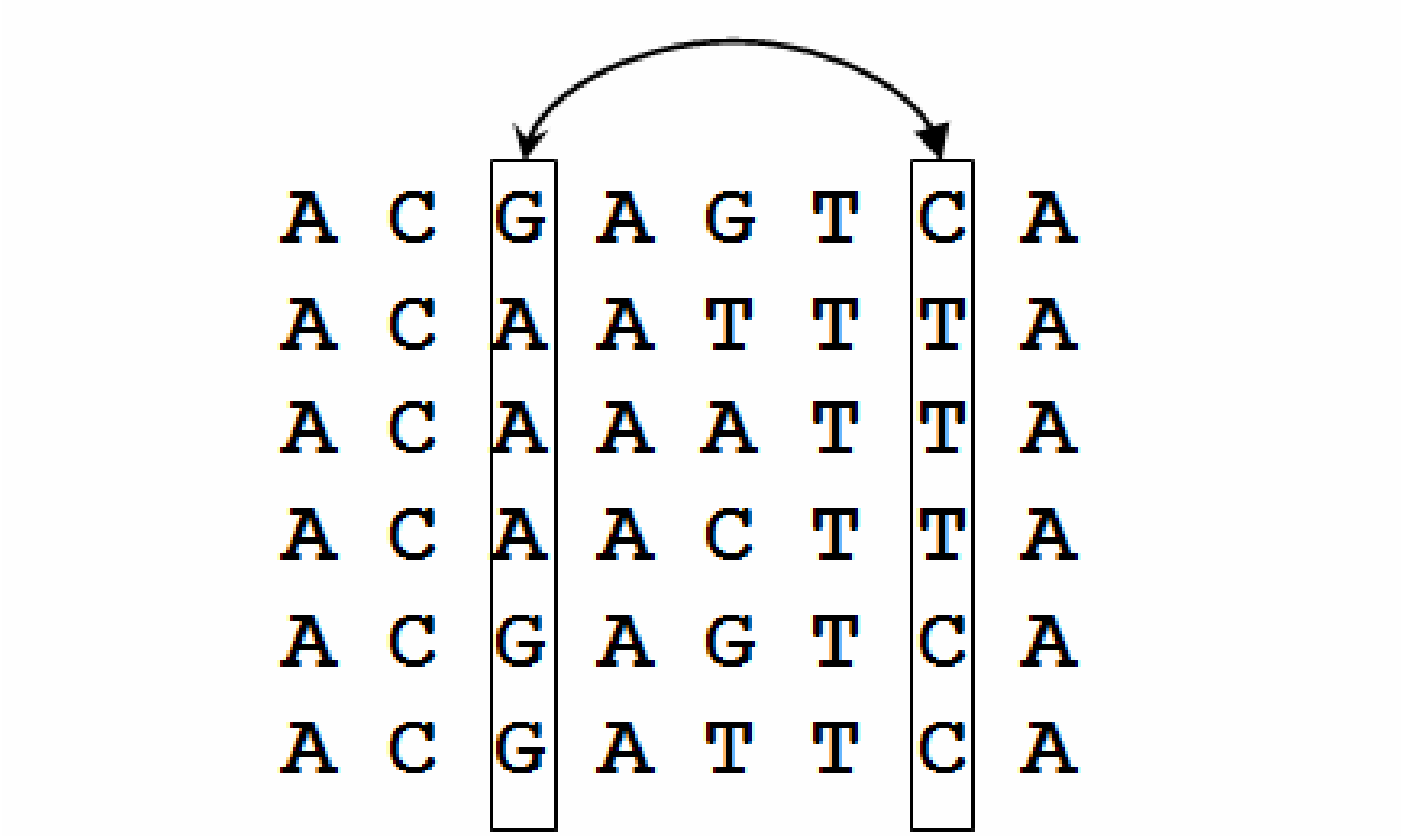
- Από την άλλη μεριά όμως, αυτά ακριβώς τα χαρακτηριστικά που κάνουν τα πρότυπα ιδιαίτερα επιτυχημένα, περιέχουν και το σπόρο με τις αδυναμίες τους. Το βασικό μειονέκτημα είναι ότι χάνεται μεγάλο μέρος της πληροφορίας της πολλαπλής στοίχισης. Για παράδειγμα στην Εικόνα 5.1 στην 3η στήλη της στοίχισης το πρότυπο προβλέπει [AT], δηλαδή A ή T, αλλά δεν μας δίνει τη σχετική πιθανότητα για το καθένα, παρόλο που από την πολλαπλή στοίχιση βλέπουμε ότι η Θυμίνη (T) έχει διπλάσια πιθανότητα από την Αδενίνη (A). Φανταστείτε ότι έχουμε τώρα την ίδια περίπτωση αλλά σε μια στοίχιση με 100 αλληλουχίες, και εκεί έχουμε 65 T και 35 A. 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.

συνέχεια

- 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 1η, 4η, 5η και 6η αλληλουχία αντιστοιχούν απλά στις εισαγωγές νουκλεοτιδίων στην 2η και στην 3η αλληλουχία. Τι θα γινόταν όμως αν λ.χ. στην πρώτη αλληλουχία στην 8η 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.

συνέχεια

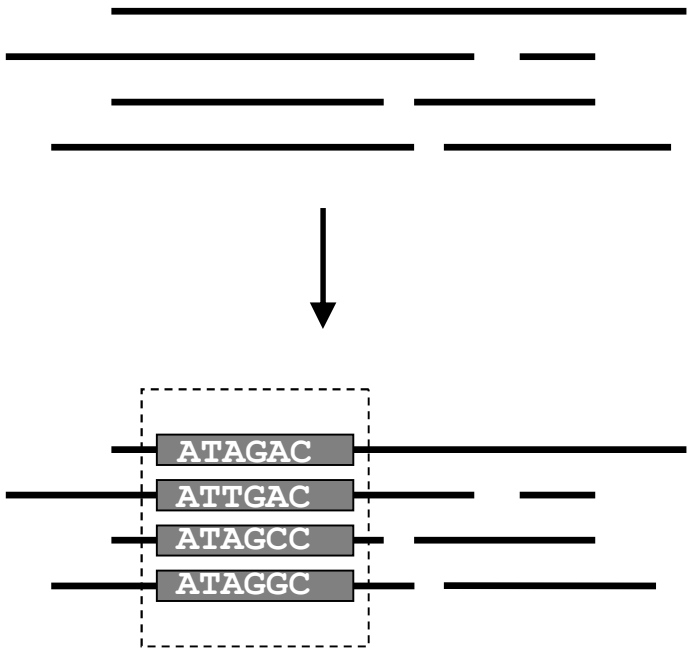
- Ένα άλλο πρόβλημα, είναι ότι τα πρότυπα με τον τρόπο που τα ορίσαμε δεν μπορούν να ενσωματώσουν εύκολα τα κενά στην πολλαπλή στοίχιση. Στην Εικόνα 5.1 είδαμε μια πολλαπλή στοίχιση που περιέχει κενά, αλλά όλα προέρχονται από εισαγωγές (τυχαίων) νουκλεοτιδίων σε κάποιες από τις αλληλουχίες της στοίχισης. Έτσι, τα κενά στην 1η, 4η, 5η και 6η αλληλουχία αντιστοιχούν απλά στις εισαγωγές νουκλεοτιδίων στην 2η και στην 3η αλληλουχία. Τι θα γινόταν όμως αν λ.χ. στην πρώτη αλληλουχία στην 8η 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.


A C G A G T C A
A C A A T T T A
A C A A A T T A
A C A A C T T A
A C G A G T C A
A C G A T T C A
A-C-[AG]-A-x-T-[CT]-A

Κατασκευή των προτύπων

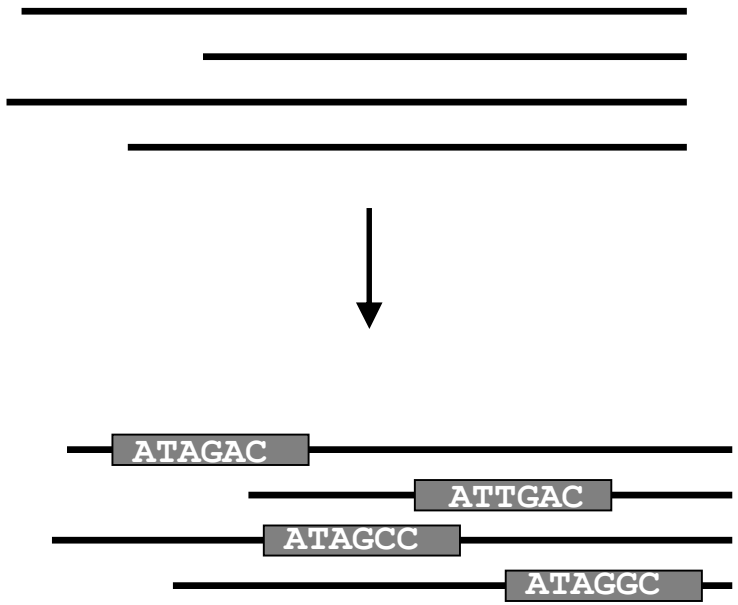
- Επιλογή της γλώσσας: στο πρώτο στάδιο θα πρέπει να επιλεγεί ο τρόπος περιγραφής των προτύπων. Μπορεί δηλαδή να χρησιμοποιηθεί η σύνταξη της PROSITE, αλλά υπάρχουν και περιπτώσεις στις οποίες επιλέγονται και πιο απλές περιγραφές (π.χ. πρότυπα τα οποία περιέχουν εκφράσεις χωρίς πολλαπλές ταυτίσεις σε 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.

Πολλαπλή στοίχιση



A-T-[AT]-G-x-C

Μη στοιχισμένες αλληλουχίες



A-T-[AT]-G-x-C

Λογισμικό

- Το πιο παλιό και ευρέως χρησιμοποιούμενο εργαλείο για την κατασκευή προτύπων, είναι το **PRATT** (http://web.expasy.org/pratt/). 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.
- 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 για εύρεση θέσεων πρόσδεσης μεταγραφικών παραγόντων ή επαναληπτικές αλληλουχίες, ενώ είναι διαθέσιμη και ως αυτόνομο λογισμικό.
- 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. 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 όλες[u3] 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.

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 pb(i) του κάθε συμβόλου

A T A G A C A C A A

A T T G T C A C T A

A T T G A C G C T A

A T A G G C A C G A

A T T G A C C C C A

A T T G C C A C G A

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

	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

```
A T A G A C A C A A
A T T G T C A C T A
A T T G A C G C T A
A T A G G C A C G A
A T T G A C C C C A
A T T G C C A C G A
```

$$s_b(i) = \log\left(p_b(i) / p_b \right)$$

	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	-∞

Πρωτεϊνες

- 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.

- 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:

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

		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

	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

Λογισμικό

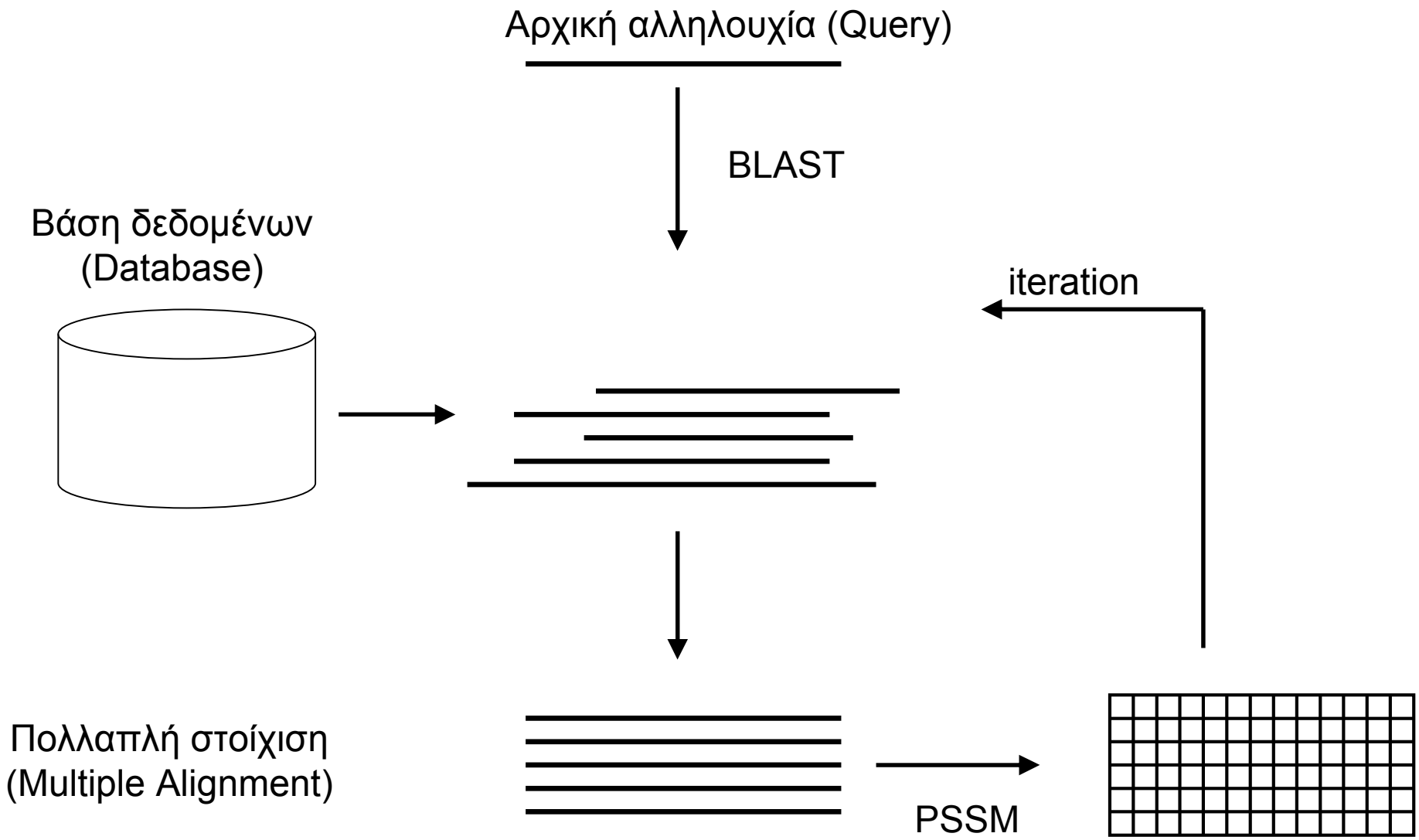
- Το πιο γνωστό πρόγραμμα της πρώτης κατηγορίας είναι το **ScanProsite** (http://prosite.expasy.org/scanprosite/). Το ScanProsite είναι κατασκευασμένο για να εντοπίζει πρότυπα και προφίλ της PROSITE, σε οποιαδήποτε αλληλουχία, είτε του χρήστη, είτε κάποια που έχει επιλεγεί από μια βάση δεδομένων. Είναι 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)

Λογισμικό

- 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, pfscale, pfw, pfsearch, pfscan

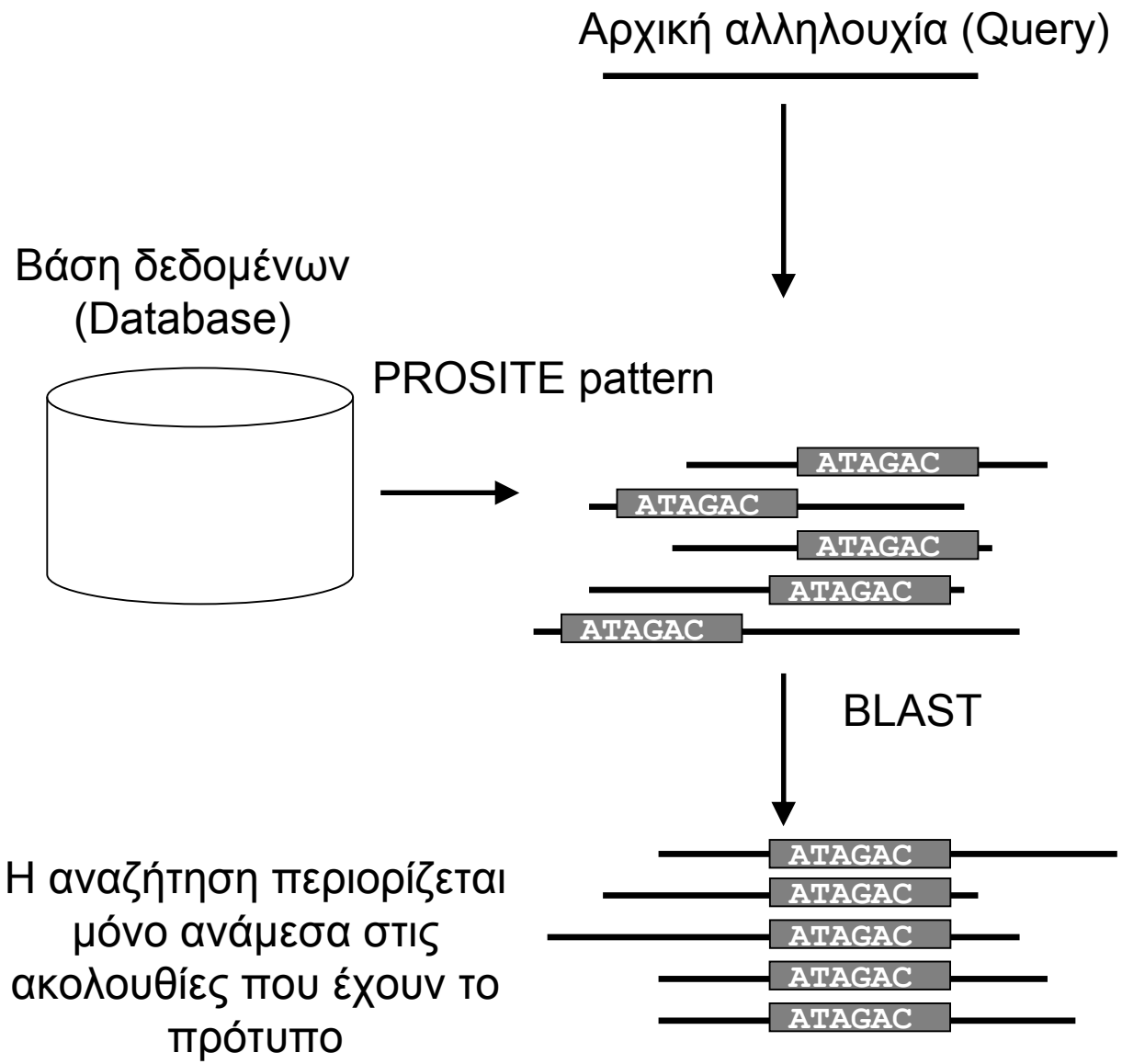
PSI-BLAST (Position-specific-iterated BLAST)

- Είναι μια επέκταση του γνωστού αλγορίθμου BLAST και χρησιμοποιείται για την εύρεση μακρινών ομολόγων. Η μέθοδος δουλεύει ως εξής.
- 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).


Αρχική αλληλουχία (Query)
BLAST
Βάση δεδομένων
(Database)
iteration
Πολλαπλή στοίχιση
(Multiple Alignment)
PSSM

PHI-BLAST (pattern-hit initiated BLAST)

- 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

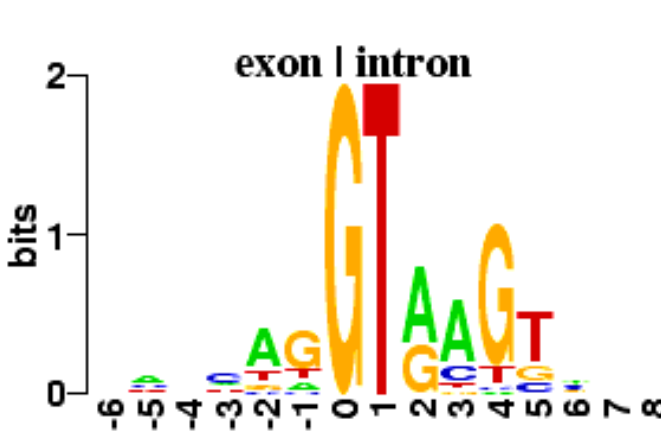

Αρχική αλληλουχία (Query)
Βάση δεδομένων (Database)
PROSITE pattern
ATAGAC
ATAGAC
ATAGAC
ATAGAC
ATAGAC
BLAST
Η αναζήτηση περιορίζεται μόνο ανάμεσα στις ακολουθίες που έχουν το πρότυπο
ATAGAC
ATAGAC
ATAGAC
ATAGAC
ATAGAC

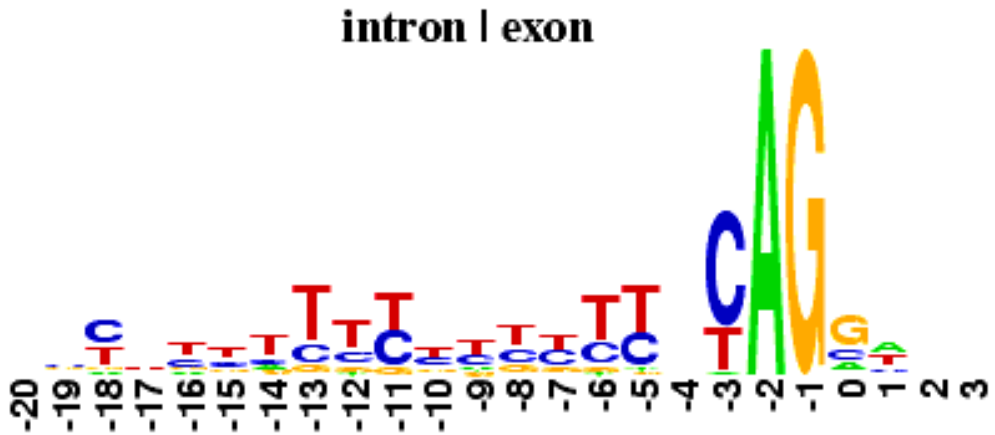
WebLogo

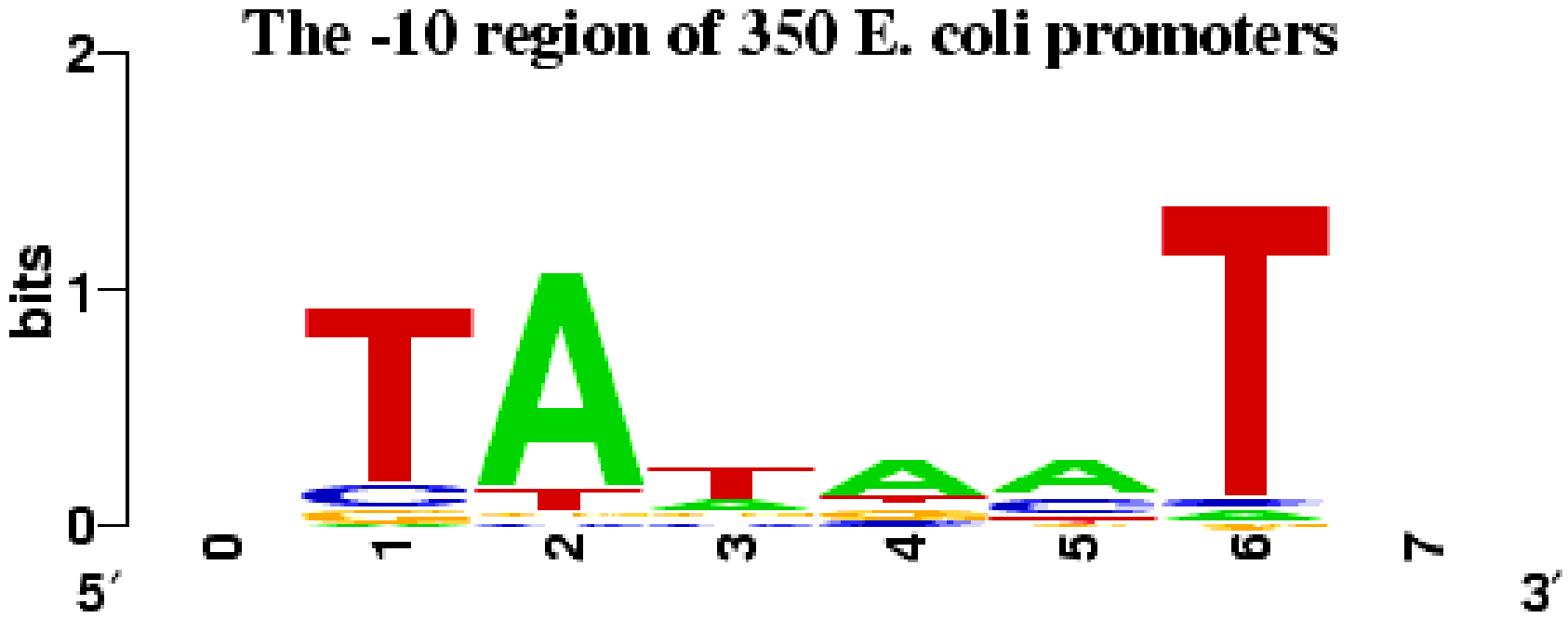
- Τέλος, μια πολύ σημαντική εφαρμογή που χρησιμοποιείται για την οπτικοποίηση των περιοχών που απεικονίζονται σε ένα πρότυπο ή προφίλ, είναι το **WebLogo** (http://weblogo.berkeley.edu/) (Crooks, Hon, Chandonia, & Brenner, 2004). Το WebLogo βασίζεται στην απλή ιδέα των Λογότυπων Αλληλουχιών (Sequence 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:

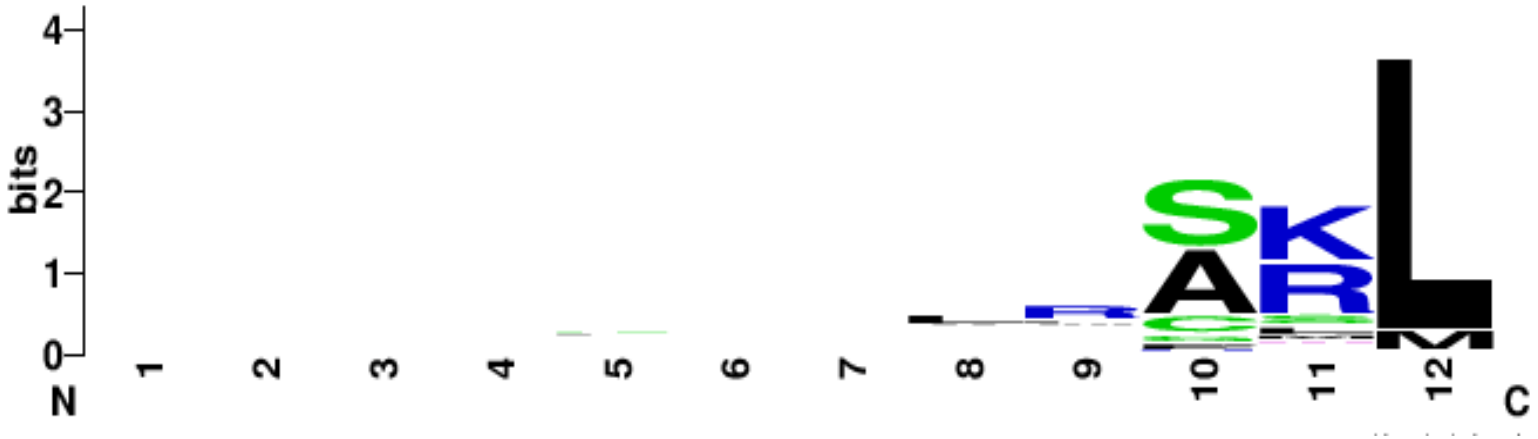
$$R = S_{\max} - S_{obs} = \log_2 k - \left(-\sum_{\forall b \in \Omega} n_b(i) \log p_b(i) \right)$$

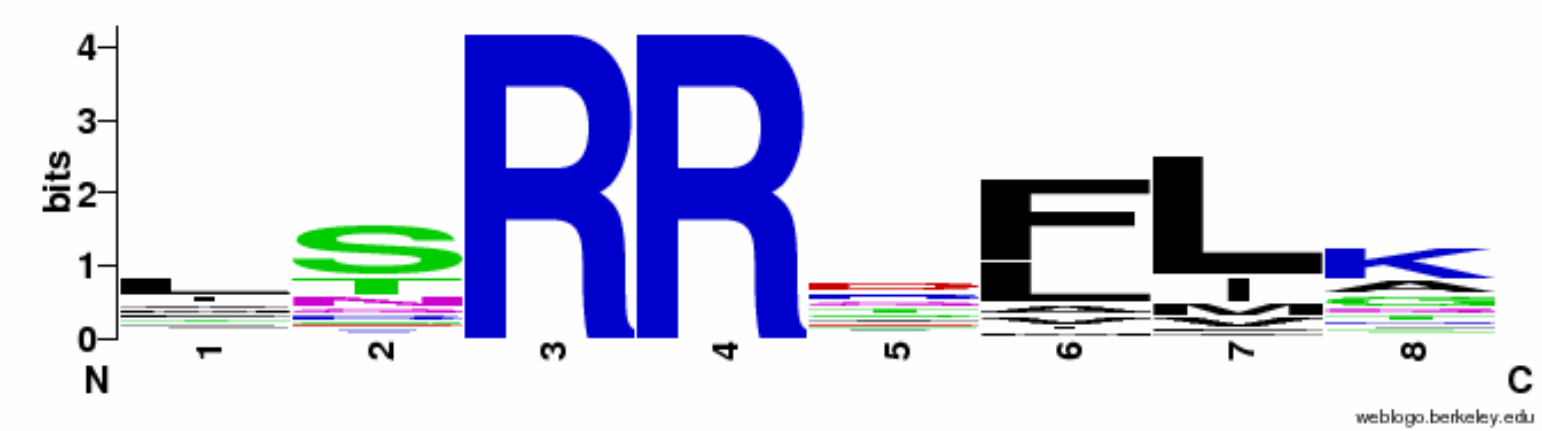
Παραδείγματα






The -10 region of 350 E. coli promoters
bits
2
1
0
0
1
2
3
4
5
6
7
5′
3′


bits
4
3
2
1
0
N
1
2
3
4
5
6
7
8
9
10
11
12
C


bits
4
3
2
1
0
N
1
2
3
4
5
6
7
8
C
weblogo.berkeley.edu