

Υπολογιστική Βιολογία : BIO315
[3] Εύρεση μοτίβων σε αλληλουχίες

Στόχοι του μαθήματος

Να περιγράψει την έννοια των “μοτίβων” αλληλουχίας και τη βιολογική τους σημασία

Να περιγράψει τρόπους για:

- Τον προσδιορισμό των μοτίβων
- Την αξιολόγηση μοτίβων από πλευράς πληροφορίας
- Την εύρεση γνωστών μοτίβων σε μεγάλου μήκους αλληλουχίες
- Την “ανακάλυψη” νέων, μέχρι πρότινος άγνωστων μοτίβων

Στο τέλος του μαθήματος θα πρέπει να μπορείτε:

- Να διακρίνετε μεταξύ συναινετικών αλληλουχιών και μοτίβων αλληλουχίας και να αναγνωρίζετε τις μεταξύ τους διαφορές.
- Να δημιουργήσετε πίνακες μοτίβων αλληλουχιών από σύνολα ολιγονουκλεοτιδίων και να περιγράφετε μοτίβα αλληλουχιών μέσω αυτών.
- Να αξιολογήσετε το πληροφοριακό περιεχόμενο μοτίβων με βάση την Εντροπία Shannon.
- Να εντοπίσετε σημεία πρόσδεσης μεταγραφικών παραγόντων σε μια γονιδιωματική αλληλουχία με βάση ένα γνωστό μοτίβο.

Τι είναι μοτίβο;

- Οι γλώσσες με νόημα περιέχουν:
 - Περιορισμούς στη χρήση συμβόλων/φωνημάτων (φωνητική)
 - Περιορισμούς στη διαδοχή λέξεων/φθόγγων (γραμματική)
 - Διαμόρφωση φράσεων με σκοπό τη δημιουργία ανώτερων νοημάτων όπως έμφαση, συνεκτικότητα μηνύματος, σύνδεση νοημάτων, αναφορά) (δομή).

Μοτίβο είναι ένα αυτόνομο στοιχείο με βαρύνουσα συμβολική αξία που επαναλαμβάνεται μέσα σε ένα έργο με σκοπό να διαμορφώσει γενικότερα χαρακτηριστικά ρυθμού και διάθεσης, να δημιουργήσει αναφορές σε άλλα σημεία του μηνύματος καθώς και για να διασυνδέσει μέρη του μηνύματος.

"I have a dream"

MLK

Έμφαση/Επανάληψη

"Who controls the past, controls the future: who controls the present controls the past."

*Από το "1984" του George Orwell
Σύνδεση νοήματος*

"Fair is foul and foul is fair"

*Από το "Macbeth" του William Shakespeare
Συνεκτικότητα*

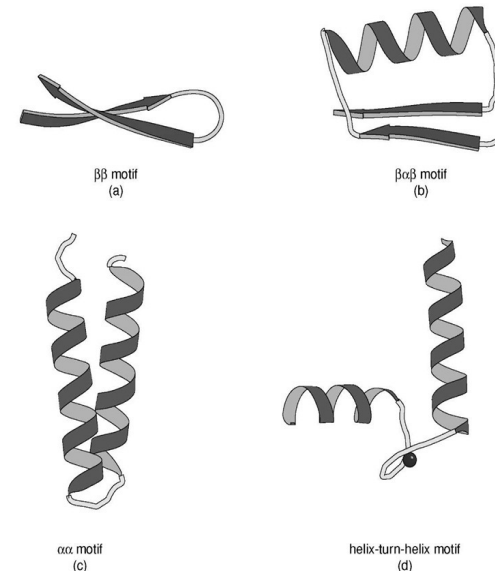
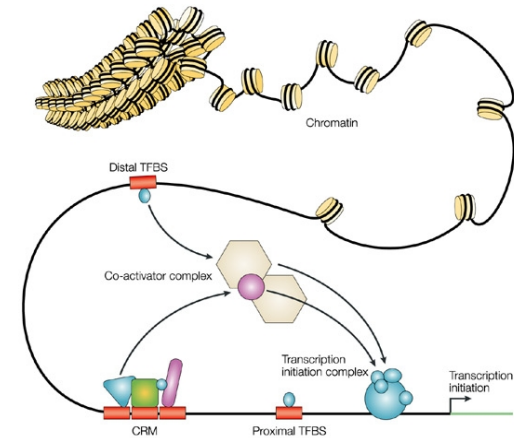
"The Rhine"



*Από το "Rheingold" του Richard Wagner
Εσωτερική αναφορά*

Μοτίβα στη Βιολογία

- Σε γονιδιωματικές αλληλουχίες είναι στοιχεία της αλληλουχίας που (επαν)εμφανίζονται σε συγκεκριμένες θέσεις με σκοπό να δημιουργήσουν αναφορές → πρόσδεση πρωτεϊνών στο DNA.
- Σε πρωτεϊνικές αλληλουχίες είναι συχνά τα τμήματα εκείνα που καθορίζουν τις περιοχές που επιτελούν συγκεκριμένες λειτουργίες (π.χ. ενεργά κέντρα ενζύμων) και που αλληλεπιδρούν με άλλες πρωτεΐνες ή στοιχεία του κυττάρου

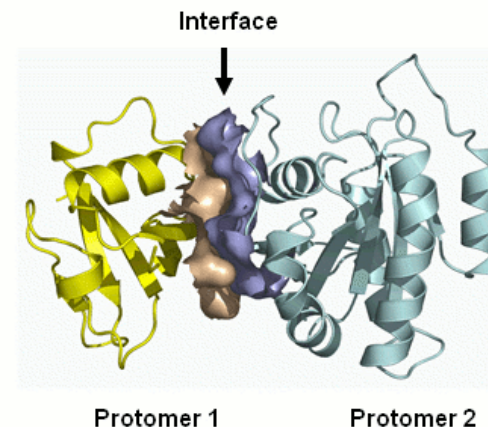
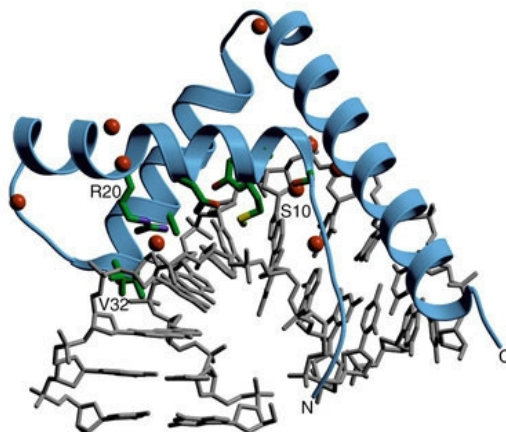


Ποιο είναι το νόημα; Φυσική σημασία μοτίβων

Ολιγονουκλεοτιδικές “λέξεις” καθοδηγούν την πρόσδεση πρωτεϊνών στο DNA.

Στο DNA/RNA καθορίζουν α) τη θέση β) την ισχύ της πρόσδεσης καθώς και την πιθανή αναδιαμόρφωση (κάμψη, συστροφή κλπ) του μηνύματος-υποστρώματος μέσω χημικών αλληλεπιδράσεων

Στις πρωτεΐνες καθορίζουν τις περιοχές που επιτελούν συγκεκριμένες λειτουργίες (π.χ. ενεργά κέντρα ενζύμων) και που αλληλεπιδρούν με άλλες πρωτεΐνες ή στοιχεία του κυττάρου



Τα βιολογικά ερωτήματα

Πώς ορίζουμε ένα μοτίβο;

Πώς εντοπίζουμε ένα γνωστό μοτίβο σε μια αλληλουχία;

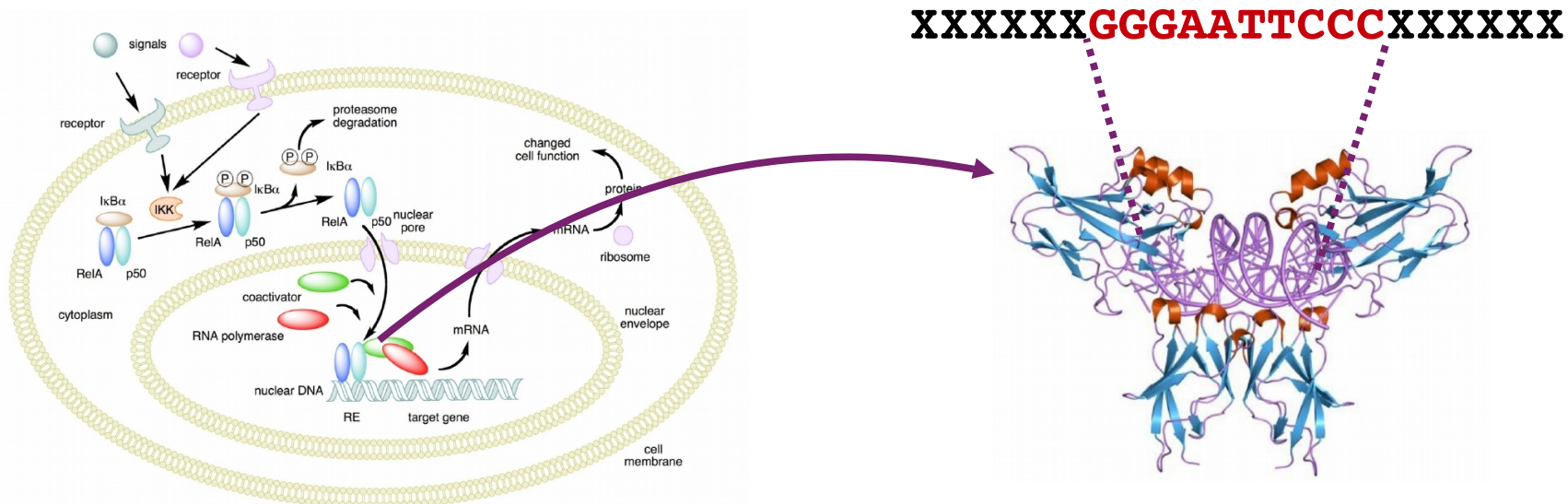
Πώς αξιολογούμε κατά πόσο ένα μοτίβο είναι “ισχυρό” ή όχι;

Πώς ανακαλύπτουμε μέχρι πρότινος άγνωστα μοτίβα;

Μελετώντας ένα μοτίβο

Ο μεταγραφικός παράγοντας NF-κB είναι μια από τις βασικές πρωτεΐνες που ρυθμίζουν τη μεταγραφή γονιδίων στην περίπτωση ανοσοαπόκρισης σε διάφορα ευκαρυωτικά συστήματα.

Είναι μια διμερής πρωτεΐνη που γνωρίζουμε ότι προσδένεται στο DNA σε μια αλληλουχία 10 νουκλεοτιδίων που πολύ συχνά είναι: **GGGAATTCCC**



Τα μοτίβα έρχονται σε “παραλλαγές”

Όπως στον Wagner ή τον Edgar Allan Poe, έτσι και στα βιολογικά “κείμενα”, τα μοτίβα εμφανίζονται παραλλαγμένα, άλλα σε μικρότερο και άλλα σε μεγαλύτερο βαθμό.

Τα παρακάτω 8 10-νουκλεοτίδια είναι όλα σημεία πρόσδεσης του NF-κΒ ακόμα κι αν δεν ταυτίζονται απόλυτα με την “ιδανική” αλληλουχία που είδαμε νωρίτερα.

```

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

```

Πώς μπορούμε να περιγράψουμε συνολικά ένα μοτίβο λαμβάνοντας υπ' όψιν τις διαφορετικές παραλλαγές του;

Then this ebony bird beguiling my sad fancy into smiling, By the grave and stern decorum of the countenance it wore, “Though thy crest be shorn and shaven, thou,” I said, “art sure no craven, Ghastly grim and ancient Raven wandering from the Nightly shore—Tell me what thy lordly name is on the Night’s Plutonian shore!”

Quoth the Raven “Nevermore.”

Much I marvelled this ungainly fowl to hear discourse so plainly, Though its answer little meaning—little relevancy bore; For we cannot help agreeing that no living human being Ever yet was blessed with seeing bird above his chamber door— Bird or beast upon the sculptured bust above his chamber door,

With such name as “Nevermore.”

But the Raven, sitting lonely on the placid bust, spoke only. That one word, as if his soul in that one word he did outpour. Nothing farther then he uttered—not a feather then he fluttered— Till I scarcely more than muttered “Other friends have flown before— On the morrow he will leave me, as my Hopes have flown before.”

Then the bird said “Nevermore.”

Edgar Allan Poe, “The Raven”, (excerpt)

Αναπαράσταση Μοτίβων Συναινετικές Αλληλουχίες (consensus)

Ονομάζουμε Συναινετική Αλληλουχία την αλληλουχία εκείνη που:

α) περιγράφει με το συνολικότερο δυνατό τρόπο το μοτίβο μέσω γραμματικών/λογικών κανόνων

ή

β) αναπαριστά την κοινότερη έκφραση του μοτίβου

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

GGG [AG] [AC] TT [TC] CC

GGGAATTTC



Πότε οι συναινετικές αλληλουχίες δεν έχουν νόημα;

Ο παρακάτω πίνακας περιέχει 104 διαφορετικά σημεία πρόσδεσης του NF-κΒ.
Πόσο πολύ διαφέρουν μεταξύ τους;
Τι μας λέει η “συναινετική” (consensus) αλληλουχία τους;

GGGGCATTCC	GGGATATCCC	GGGAATTCCC	GGGAATGTCC	GGGATATTTT	GGGGCCTCCC	GGGAATTTCC	GGGACTGCCC
GGGAAATTCC	GGGAAATCCC	GGGAATTCCC	GGGACTTACC	GGGGATTTC	GGGAATTTCC	GGGACATTCC	GGGAATTTCC
GGAAATTTCC	GGGAATTCCC	GGGGATTTC	GGGGTTTCAC	GGGAAGGTCC	GGGGCTTCCC	GGGGCTTTCC	GGGAAATTCC
GGGGCTTTCC	GGGACTTTCC	GGGACATTCC	GGGAATTTCC	GGGACATTCT	GGGACAGCCC	GGGGCTTTAC	GGGACTTCCC
GGGAATTCAC	GGGAAATCCC	GGAGCTTTCC	GGGACTTTCC	GGGAAACCCC	GGGGCTTCCC	GGGAATTTCC	GGGAAATTCC
GGGACTTCCC	GGGAATTTCT	GGGAATTCCC	GGGACTTCCC	GGGACTTTCC	GGGGATTTC	GGGACATCCC	GGGAAATCCC
GGGATGTTCC	GGGGTCTCCC	GGGACTGTCC	GGGAATTCCC	GGGACTTTAC	GGGAATTTCC	GGGACTTTCC	GGGGCGTCCC
GGGGTTTCCC	GGGAATTTCC	GGGAATTTCC	GGGGATTTC	GGGAATGCCC	GGGGATTTC	GGGAATTTCC	GGGATTTTCC
GGGGAAATCC	GGGACTTCCC	GGGATTTTCC	GGGAAGTCCC	GGGAAATTCC	GGGAATTTCC	GGGAATTTAC	GGGAAATTCC
GGGGGTTTAC	GGGACTTTCC	GGGAATTTCC	GGGAATTTCC	GGGACATCCC	GGGAATTCAC	GGGACTTCCC	GGGACTTTCC
GGGAATTTCC	GGGACTTTCC	GGGGACTTCC	GGGACTTTAC	GGGACTTTCC	GGGATACTCC	GGGGATGTAC	GGGATATCCC
GGGAATTTCC	GGGACTTCCC	GGGACTTCAC	GGGGTTACCC	GGGAATCTCC	GGGAATTTCC	GGGACATCTC	GGAAATTTCC
GGGAAACTCT	GGGGTTTCCC	GGGATTTTCC	GGGGCGTTCC	GGGAAACTCT	GGGGTTTCCC	GGGATTTTCC	GGGGCGTTCC

Πίνακας 3.1: 104 σημεία πρόσδεσης του NF-κΒ από το γονιδίωμα του ποντικίου (*Mus musculus*)

GG [AG] [AG] [AGCT] [AGCT] [AGCT] [ACT] [ACT] [CT]

Συναινετικές αλληλουχίες: Η πιο κοινή αλληλουχία

GGGGCATTCC	GGGATATCCC	GGGAATTCCC	GGGAATGTCC	GGGATATTTT	GGGGCCTCCC	GGGAATTTCC	GGGACTGCCC
GGGAAATTCC	GGGAAATCCC	GGGAATTCCC	GGGACTTACC	GGGGATTTCC	GGGAATTTCC	GGGACATTCC	GGGAATTTCC
GGAAATTTCC	GGGAATTCCC	GGGGATTTCC	GGGGTTTCAC	GGGAAGGTCC	GGGGCTTCCC	GGGGCTTTCC	GGGAAATTCC
GGGGCTTTCC	GGGACTTTCC	GGGACATTCC	GGGAATTTCC	GGGACATTCT	GGGACAGCCC	GGGGCTTTAC	GGGACTTCCC
GGGAATTCAC	GGGAAATCCC	GGAGCTTTCC	GGGACTTTCC	GGGAAACCCC	GGGGCTTCCC	GGGAATTTCC	GGGAAATTCC
GGGACTTCCC	GGGAATTTCT	GGGAATTCCC	GGGACTTCCC	GGGACTTTCC	GGGGATTTCC	GGGACATCCC	GGGAAATCCC
GGGATGTTCC	GGGGTCTCCC	GGGACTGTCC	GGGAATTTCC	GGGACTTTAC	GGGAATTTCC	GGGACTTTCC	GGGGCGTCCC
GGGGTTTCCC	GGGAATTTCC	GGGAATTTCC	GGGGATTTCC	GGGAATGCCC	GGGGATTTCC	GGGAATTTCC	GGGATTTTCC
GGGGAATTCC	GGGACTTCCC	GGGATTTTCC	GGGAAGTCCC	GGGAAATTCC	GGGAATTTCC	GGGAATTTAC	GGGAAATTCC
GGGGGTTTAC	GGGACTTTCC	GGGAATTTCC	GGGAATTTCC	GGGACATCCC	GGGAATTCAC	GGGACTTCCC	GGGACTTTCC
GGGAATTTCC	GGGACTTTCC	GGGGACTTCC	GGGACTTTAC	GGGACTTTCC	GGGATACTCC	GGGGATGTAC	GGGATATCCC
GGGAATTTCC	GGGACTTCCC	GGGACTTCAC	GGGGTTACCC	GGGAATCTCC	GGGAATTTCC	GGGACATCTC	GGAAATTTCCC
GGGAAACTCT	GGGGTTTCCC	GGGATTTTCC	GGGGCGTTCC	GGGAAACTCT	GGGGTTTCCC	GGGATTTTCC	GGGGCGTTCC

Πίνακας 3.1: 104 σημεία πρόσδεσης του NF-κΒ από το γονιδίωμα του ποντικίου (*Mus musculus*)

Η πιο κοινή αλληλουχία από τον παραπάνω πίνακα προκύπτει ότι είναι η:

GGGAATTTCC

Μπορείτε να σκεφτείτε έναν τρόπο για να την εξάγετε από μια δεδομένη συλλογή αλληλουχιών;

Αναζήτηση γνωστού μοτίβου σε αλληλουχία

Δεδομένου ενός μοτίβου με μήκος l , πώς θα το εντοπίσουμε σε μια μεγάλη αλληλουχία;

Μπορούμε να “σαρώσουμε” την αλληλουχία ελέγχοντας όλες τις πιθανές υπο-αλληλουχίες μήκους l και να κρατήσουμε μόνο εκείνες που “μοιάζουν” με το μοτίβο. Χρειαζόμαστε συνεπώς ένα μέτρο “ομοιότητας” μεταξύ αλληλουχιών.

Απόσταση Hamming δύο σειρών χαρακτήρων ίδιου μήκους ορίζουμε το άθροισμα των χαρακτήρων στους οποίους διαφέρουν.

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

Έχουν απόσταση Hamming $d=1$ ενώ οι:

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

Έχουν απόσταση Hamming $d=3$.

Το πρόβλημα με την απόσταση Hamming

Το πρόβλημα με την απόσταση Hamming (και άλλα μέτρα απόστασης) είναι ότι δε λαμβάνει υπ' όψιν τη σημασία που μπορεί να έχουν διαφορετικές θέσεις μέσα στο μοτίβο.

AAAAATTCCC => d=3

GGGGGTTTCC => d=3

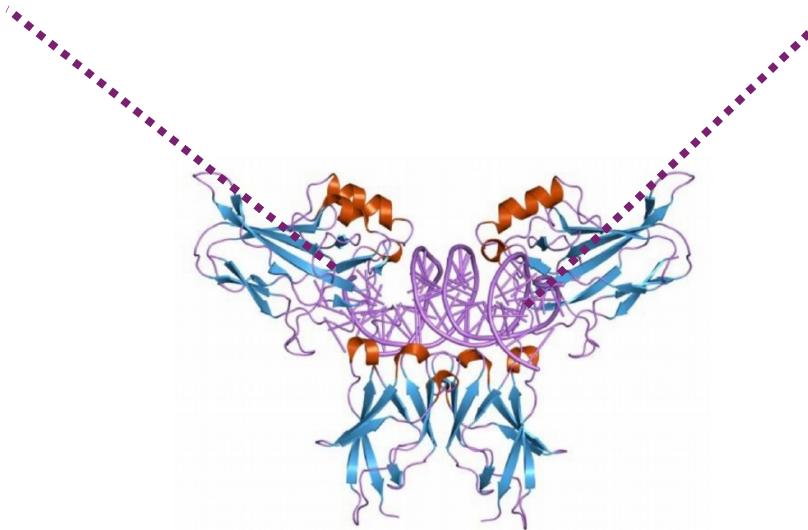
GGGAACCCCC => d=2

XXXXXXXXGGGAATTCCCXXXXXXXX

Όμως το μοτίβο είναι αυτό:

GGG[AG][AC]TT[TC]CC

Οι θέσεις δεν είναι όλες το ίδιο "κρίσιμες"



Αξιολόγηση μοτίβων: Πίνακες Ειδικοί ανά Θέση

Positional Weight Matrices (PWM)

Αξιολογούμε την κάθε θέση στο μοτίβο με διαφορετικό βάρος που αντιστοιχεί στην πιθανότητα εμφάνισης του κάθε καταλοίπου στη συγκεκριμένη θέση.

Για μοτίβο DNA μήκους 10 νουκλεοτιδίων έχουμε έναν 10x4 PWM.

GGGGCATTCC	GGGATATCCC	GGGAATTCCC	GGGAATGTCC	GGGATATTTT	GGGGCCTCCC	GGGAATTTCC	GGGACTGCCC
GGGAAATTCC	GGGAAATCCC	GGGAATTCCC	GGGACTTACC	GGGGATTTCC	GGGAATTTCC	GGGACATTCC	GGGAATTTCC
GGAAATTTCC	GGGAATTCCC	GGGGATTTCC	GGGGTTTTCAC	GGGAAGGTCC	GGGGCTTCCC	GGGGCTTTCC	GGGAAATTCC
GGGGCTTTCC	GGGACTTTCC	GGGACATTCC	GGGAATTTCC	GGGACATTCT	GGGACAGCCC	GGGGCTTTAC	GGGACTTCCC
GGGAATTCAC	GGGAAATCCC	GGAGCTTTCC	GGGACTTTCC	GGGAAACCCC	GGGGCTTCCC	GGGAATTTCC	GGGAAATTCC
GGGACTTCCC	GGGAATTTCT	GGGAATTCCC	GGGACTTCCC	GGGACTTTCC	GGGGATTTCC	GGGACATCCC	GGGAAATCCC
GGGATGTTCC	GGGGTCTCCC	GGGACTGTCC	GGGAATTTCC	GGGACTTTAC	GGGAATTTCC	GGGACTTTCC	GGGGCGTCCC
GGGGTTTCCC	GGGAATTTCC	GGGAATTTCC	GGGGATTTCC	GGGAATGCCC	GGGGATTTCC	GGGAATTTCC	GGGATTTTCC
GGGGAATTCC	GGGACTTCCC	GGGATTTTCC	GGGAAGTCCC	GGGAAATTCC	GGGAATTTCC	GGGAATTTAC	GGGAAATTCC
GGGGGTTTAC	GGGACTTTCC	GGGAATTTCC	GGGAATTTCC	GGGACATCCC	GGGAATTCAC	GGGACTTCCC	GGGACTTTCC
GGGAATTTCC	GGGACTTTCC	GGGGACTTCC	GGGACTTTAC	GGGACTTTCC	GGGATACTCC	GGGGATGTAC	GGGATATCCC
GGGAATTTCC	GGGACTTCCC	GGGACTTCAC	GGGGTTACCC	GGGAATCTCC	GGGAATTTCC	GGGACATCTC	GGAAATTTCC
GGGAAACTCT	GGGGTTTCCC	GGGATTTTCC	GGGGCGTTCC	GGGAAACTCT	GGGGTTTCCC	GGGATTTTCC	GGGGCGTTCC

Νουκλεοτίδιο	1	2	3	4	5	6	7	8	9	10
A	0.00	0.00	0.03	0.76	0.49	0.23	0.01	0.01	0.10	0.00
C	0.00	0.00	0.00	0.00	0.37	0.03	0.04	0.38	0.88	0.97
G	1.00	1.00	0.97	0.24	0.01	0.05	0.07	0.00	0.00	0.00
T	0.00	0.00	0.00	0.00	0.13	0.69	0.88	0.61	0.02	0.03

Δημιουργία PWM

Τι χρειαζόμαστε σαν input;

Πώς θα εργαστούμε;

Αλγόριθμος :: PWM

Δήλωση Πίνακα n Σημείων Πρόσδεσης μήκους l , $TFBS[n,l]$;

Δήλωση Πίνακα N τεσσάρων νουκλεοτιδίων (A,G,C,T);

Δήλωση Πίνακα P Πιθανοτήτων νουκλεοτιδίων (A,G,C,T);

Δήλωση Πίνακα $PWM[4, l]$;

Απαρίθμηση 1: Για θέση $i = 1$ έως $i = l$ ανά 1;

 Δημιούργησε τη σειρά $C=TFBS[1:n,i]$; # όλα τα στοιχεία κάθε στήλης

 Απαρίθμηση 2: Για θέση $j=1$ έως $j=n$ ανά 1;

 Διάβασε $s=C[j]$;

 Αύξησε το πλήθος πίνακα νουκλεοτιδίων $N[s]++$;

 Τέλος: Απαρίθμηση 2

 #

 Συχνότητα: Για κάθε νουκλεοτίδιο s ;

 Υπολόγισε τη συχνότητα $P[s]=N[s]/n$; # διαίρεση με πλήθος σημείων

 Απόδοση στον $PWM[i,s]=P[s]$

$N=0$; $P=0$; # αρχικοποίηση πινάκων πλήθους συχνοτήτων

Τέλος Απαρίθμηση 1

Απόδωσε αποτέλεσμα: Πίνακας PWM

Τερματισμός

Αξιολόγηση μοτίβων: Πίνακες Ειδικοί ανά Θέση Positional Weight Matrices (PWM)

Με βάση τον παρακάτω πίνακα

Νουκλεοτίδιο	1	2	3	4	5	6	7	8	9	10
A	0.00	0.00	0.03	0.76	0.49	0.23	0.01	0.01	0.10	0.00
C	0.00	0.00	0.00	0.00	0.37	0.03	0.04	0.38	0.88	0.97
G	1.00	1.00	0.97	0.24	0.01	0.05	0.07	0.00	0.00	0.00
T	0.00	0.00	0.00	0.00	0.13	0.69	0.88	0.61	0.02	0.03

Ποια από τις τρεις παρακάτω αλληλουχίες συμφωνεί καλύτερα με το μοτίβο;

AAAAATTCCC => d=3

GGGGGTTTCC => d=3

GGGAACCCCC => d=2

Ποια είναι η επίδραση της νουκλεοτιδικής σύστασης του μοτίβου;

Πίνακες Βαθμονόμησης ανά θέση Position-specific Scoring Matrices (PSSM)

Είναι πίνακες που προκύπτουν από το συνδυασμό PWM που λαμβάνουν όμως υπ' όψιν τη σύσταση καταλοίπων του υποβάθρου (τυχαία εκτίμηση).

Μοτίβο NF-κΒ (*P*)

Νουκλεοτίδιο	1	2	3	4	5	6	7	8	9	10
A	0.00	0.00	0.03	0.76	0.49	0.23	0.01	0.01	0.10	0.00
C	0.00	0.00	0.00	0.00	0.37	0.03	0.04	0.38	0.88	0.97
G	1.00	1.00	0.97	0.24	0.01	0.05	0.07	0.00	0.00	0.00
T	0.00	0.00	0.00	0.00	0.13	0.69	0.88	0.61	0.02	0.03

Πίνακας Υποβάθρου (*Q*)

Νουκλεοτίδιο	1	2	3	4	5	6	7	8	9	10
A	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16
C	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.27
G	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33
T	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23

$$R = \log_2(P_{i,j}/Q_{i,j})$$

Νουκλεοτίδιο	1	2	3	4	5	6	7	8	9	10
A	-7.3	-7.3	-2.4	2.2	1.6	0.5	-3.9	-3.9	-0.7	-7.3
C	-8.1	-8.1	-8.1	-8.1	0.5	-3.1	-2.7	0.5	1.7	1.8
G	1.6	1.6	1.6	-0.5	-4.9	-2.7	-2.2	-8.4	-8.4	-8.4
T	-7.8	-7.8	-7.8	-7.8	-0.8	1.6	1.9	1.4	-3.5	-2.9

Position-Specific Scoring Matrix, PSSM

Το αποτέλεσμα είναι ένας πίνακας που μπορεί να χρησιμοποιηθεί πιο αξιόπιστα για την αναζήτηση/αξιολόγηση μοτίβων.

Αναζήτηση μοτίβων σε αλληλουχίες

Αλγόριθμος :: PWM Αναζήτηση

Δήλωση Αλληλουχίας S μήκους n ;

Δήλωση Πίνακα $PWM[4, l]$;

Δήλωση Πίνακα $Score[n-l+1]$

Απαρίθμηση 1: Για θέση $i = 1$ έως $i = n-l+1$ ανά 1;

 Δημιούργησε την υποαλληλουχία $s \leftarrow S[i:i+l-1]$; # μήκους= l

 Απαρίθμηση 2: Για κάθε θέση $j=1$ έως $j=l$ ανά 1;

$Score[i] = Score[i] + PWM[s[j], j]$


 Τέλος: Απαρίθμηση 2

Τέλος Απαρίθμηση 1

Απόδωσε αποτέλεσμα: Πίνακας $Score$

Τερματισμός

GAGTTACCCTAGCGGGTACATGGGA



$$Score = P[G, 1] + P[A, 2] + P[G, 3] + P[T, 4] + P[T, 5] + P[A, 6] + P[C, 7] + P[C, 8] + P[C, 9] + P[T, 10]$$

$$Score = 1 + 0 + 0.97 + 0 + 0.13 + 0.23 + 0.04 + 0.38 + 0.88 + 0.03 = 3.66$$

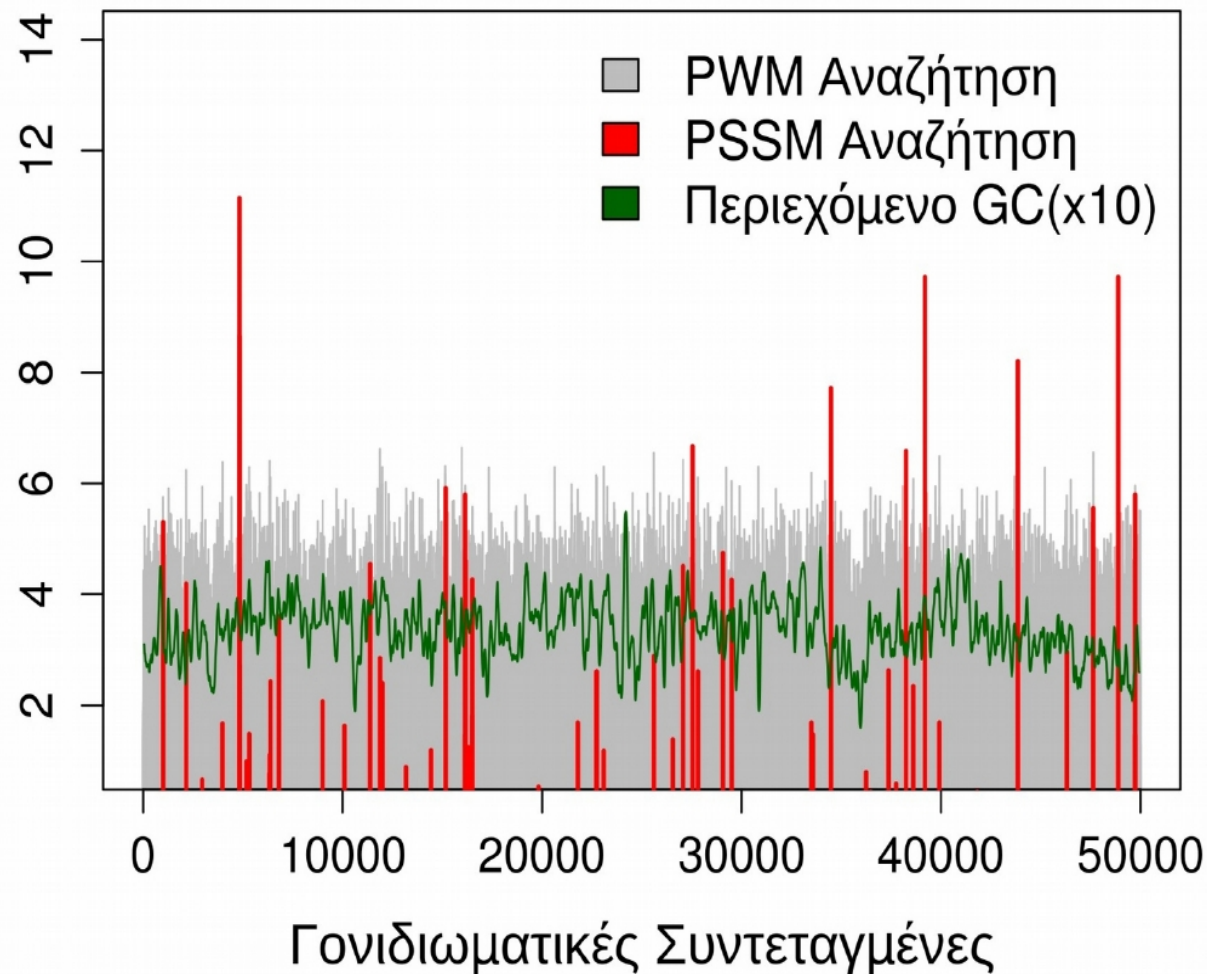
GAGTTACCCTAGCGGGTACATGGGA \longrightarrow

GAGTTACCCTAGCGGGTACATGGGA \longrightarrow

GAGTTACCCTAGCGGGTACATGGGA \longrightarrow

Scores

Αναζήτηση με PWM/PSSM



Εικόνα 3.5: Αναζήτηση σημείων πρόσδεσης του NF- κ B μέσω PWM(γκρι) και PSSM(κόκκινο). Η αναζήτηση με το PSSM εντοπίζει σημεία πρόσδεσης με μεγαλύτερη εξειδίκευση, ενώ η αναζήτηση μέσω PWM φαίνεται να σχετίζεται σε μεγάλο βαθμό με το GC περιεχόμενο της αλληλουχίας (πράσινο).

Το επόμενο πρόβλημα

Μέχρι τώρα έχουμε δει ότι:

1. Ένα μοτίβο μπορεί να εμφανίζεται σε παραλλαγές
2. Παρότι τείνει να είναι γενικά σπάνιο, εμφανίζεται αρκετές φορές, τόσες ώστε να υποψιαζόμαστε ότι δεν αντιστοιχούν όλες οι εμφανίσεις του σε λειτουργικά φαινόμενα.

Το ερώτημα που προκύπτει είναι:

Πώς μπορούμε να αξιολογήσουμε την “ισχύ” ενός μοτίβου, την πιθανότητα δηλαδή:

- Να κωδικοποιεί πληροφορία
- Να επιτελεί μια συγκεκριμένη λειτουργία

Τι είναι “πληροφορία”; Ένα παιχνίδι



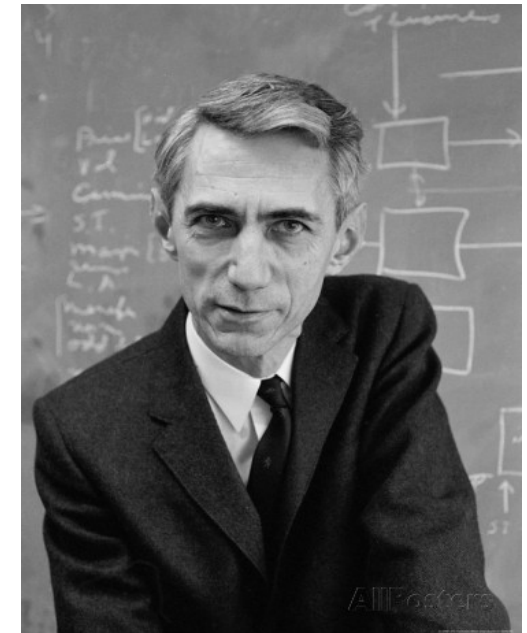
Πληροφορία και Εντροπία

Από φυσική άποψη, η πληροφορία ορίζεται ως το μέτρο της αβεβαιότητας που χάνεται κατά την μετάδοση ενός μηνύματος.

Στο παράδειγμα του παιχνιδιού με τον χάρτη, οι απαντήσεις με τη μεγαλύτερη πληροφορία είναι αυτές που μειώνουν σε μεγαλύτερο ποσοστό τις πιθανές λύσεις του “γρίφου”.

Από αυτήν την άποψη μπορούμε να φανταστούμε την πληροφορία ως ένα μέτρο της μείωσης της αταξίας ή αλλιώς, της Εντροπίας ενός συστήματος μετάδοσης μηνυμάτων.

Το 1948 ο Claude Shannon διατύπωσε μια “Μαθηματική θεωρία της Πληροφορίας”, σύμφωνα με την οποία οι έννοιες τόσο της Εντροπίας όσο και της Πληροφορίας μπορούν να ορίσουν ποσοτικά.



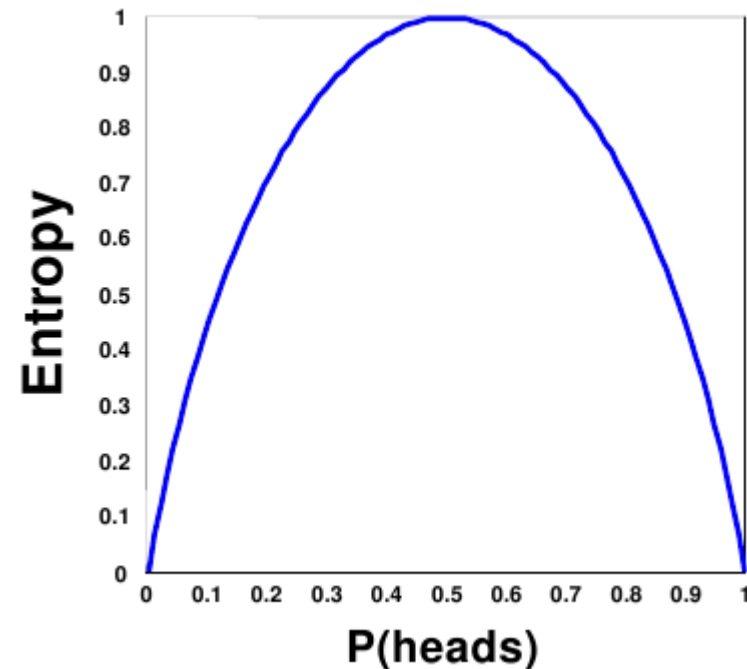
Εντροπία Shannon (Shannon Entropy)

Ορίζουμε ως Εντροπία Shannon μιας “πηγής μηνυμάτων” το παρακάτω άθροισμα:

$$H(X) = -\sum_i p_i \log_2 p_i$$

Όπου p_i είναι η πιθανότητα του ενδεχομένου i .

Προκύπτει πως η εντροπία είναι μέγιστη για ισοπίθανα ενδεχόμενα, δηλαδή στην περίπτωση της μέγιστης αβεβαιότητας



Ως Πληροφορία που προκύπτει από μια διαδικασία ορίζεται ο βαθμός μείωσης της Εντροπίας κατά τη διαδικασία αυτή.

Shannon Information

Ορίζουμε ως πληροφοριακό περιεχόμενο τη μεταβολή της Εντροπίας

$$I(X) = H_{\text{πριν}} - H_{\text{μετά}}$$

Η πληροφορία, όπως και η Εντροπία Shannon μετριέται σε bits.

1 bit πληροφορίας ισούται με το ισοδύναμο μιας απάντησης ΝΑΙ ή ΟΧΙ, κορώνα ή γράμματα κλπ.

Έχοντας αυτό υπ' όψιν πόσα bit πληροφορίας μπορεί **δυνητικά** να φέρει ένα νουκλεοτίδιο σε μια αλληλουχία;

Πληροφορία Shannon σε μοτίβα βιολογικών αλληλουχιών

Μια θέση στο μοτίβο όπου παρατηρείται πάντα το ίδιο κατάλοιπο έχει ελάχιστη αβεβαιότητα. Αν σε έναν αριθμό N εμφανίσεων του μοτίβου παρατηρούμε 100%A, 0%G, 0%C και 0%T, τότε η Εντροπία Shannon θα είναι ίση με:

$$H_{fin} = \log(1) + 0 + 0 + 0 = 0$$

(θεωρούμε το $0\log(0)=0$)

Η εντροπία του λοιπόν (για τη συγκεκριμένη θέση) είναι 0.

Είδαμε νωρίτερα ότι η μέγιστη εντροπία προκύπτει όταν όλα τα ενδεχόμενα είναι ισοπίθανα, που στην περίπτωση αυτή είναι:

$$H_{init} = -4(0.25\log(0.25)) = -2$$

Τότε μπορούμε να υπολογίσουμε την πληροφορία του μοτίβου για τη συγκεκριμένη θέση ως τη διαφορά των δύο τιμών:

$$I = H_{fin} - H_{init} = 0 - (-2) = 2$$

Εντροπία/Πληροφορία Shannon σε μοτίβα βιολογικών αλληλουχιών

Νουκλεοτίδιο	1	2	3	4	5	6	7	8	9	10
A	0.00	0.00	0.03	0.76	0.49	0.23	0.01	0.01	0.10	0.00
C	0.00	0.00	0.00	0.00	0.37	0.03	0.04	0.38	0.88	0.97
G	1.00	1.00	0.97	0.24	0.01	0.05	0.07	0.00	0.00	0.00
T	0.00	0.00	0.00	0.00	0.13	0.69	0.88	0.61	0.02	0.03

$$H(X) = -\sum_i p_i \log_2 p_i$$
$$I(X) = H_{\text{πριν}} - H_{\text{μετά}}$$

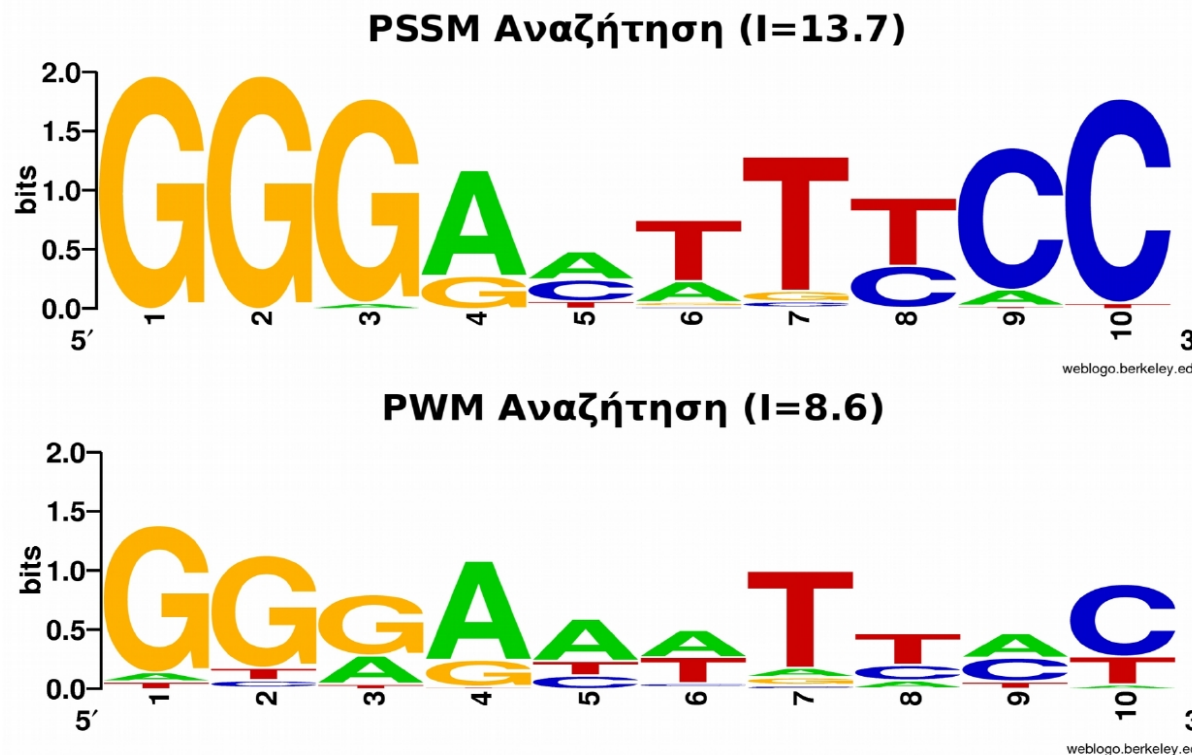
Θέση	1	2	3	4	5	6	7	8	9	10
A	0.00	0.00	0.05	0.92	0.25	0.18	0.01	0.01	0.14	0.00
C	0.00	0.00	0.00	0.00	0.19	0.02	0.05	0.37	1.23	1.75
G	2.00	2.00	1.75	0.29	0.01	0.04	0.09	0.00	0.00	0.00
T	0.00	0.00	0.00	0.00	0.07	0.53	1.16	0.59	0.03	0.05
I(θέσης)	2.00	2.00	1.81	1.20	0.52	0.78	1.32	0.97	1.39	1.81

Αναπαράσταση Πληροφορίας Μοτίβου. Weblogo

Το συνολικό ύψος της κάθε στήλης αντιστοιχεί στην πληροφορία του μοτίβου για την αντίστοιχη θέση.

Τα επιμέρους ύψη του κάθε συμβόλου αποτυπώνουν τη “συνεισφορά” (σε τιμές $P\log(P)$) του κάθε καταλοίπου.

Ποια είναι η μέγιστη πληροφορία I για ένα 10-νουκλεοτιδικό μοτίβο; Πότε θα παίρνουμε τη μέγιστη αυτή τιμή;



Το πιο δύσκολο πρόβλημα. Ανακάλυψη μοτίβων *de novo*

Τι συμβαίνει όταν δεν γνωρίζουμε τίποτα για το μοτίβο και δεν έχουμε μια στοίχιση που να μας βοηθά.

Ακόμα χειρότερα, στις περισσότερες περιπτώσεις το μοτίβο δεν εμφανίζεται πάντα με την ίδια εκδοχή αλλά με ελαφρές “παραλλαγές”

```
CGGGGCTATGCAACTGGGTCGTCACATTCCCCTTTTCGATA
TTTGAGGGTGCCCAATAAATGCAACTCCAAAGCGGACAAA
GGATGCAACTGATGCCGTTTGACGACCTAAATCAACGGCC
AAGGATGCAACTCCAGGAGCGCCTTTGCTGGTTCTACCTG
AATTTTCTAAAAAGATTATAATGTCGGTCCATGCAACTTC
CTGCTGTACAACCTGAGATCATGCTGCATGCAACTTTCAAC
TACATGATCTTTTGATGCAACTTGGATGAGGGAATGATGC
```

```
CGGGGCTATcCAgCTGGGTCGTCACATTCCCCTTTTCGATA
TTTGAGGGTGCCCAATAAaggGCAACTCCAAAGCGGACAAA
GGATGgAtCTGATGCCGTTTGACGACCTAAATCAACGGCC
AAGGAaGCAACcCCAGGAGCGCCTTTGCTGGTTCTACCTG
AATTTTCTAAAAAGATTATAATGTCGGTCCtTGgAACTTC
CTGCTGTACAACCTGAGATCATGCTGCATGCcAtTTCAAC
TACATGATCTTTTGATGgcACTTGGATGAGGGAATGATGC
```

Στην περίπτωση αυτή θα πρέπει να ανακαλύψουμε το μοτίβο *de novo* αναλύοντας τη σύσταση των αλληλουχιών και αναζητώντας τοπικές ανομοιομορφίες στη σύστασή τους.

Προσεγγίσεις

- 1. Τυχαιοποιημένοι αλγόριθμοι: π.χ. Gibbs sampling
- 2. Προσεγγίσεις συγκριτικής γονιδιωματικής



Συντηρημένα μοτίβα



Χαρτογράφηση ρυθμιστικών μοτίβων 3 διαφορετικών μεταγραφικών παραγόντων σε μια διαγονιδιακή περιοχή του *S. cerevisiae*

Προσεγγίσεις για την de novo ανακάλυψη μοτίβων

Greedy search. Αναζήτηση όλων των πιθανών μοτίβων
(CONSENSUS: <http://stormo.wustl.edu/consensus/html/Html/main.html>)

Gibbs Sampling. Τυχαιοποιημένος αλγόριθμος που ξεκινά με μια τυχαία επιλογή μοτίβου και στη συνέχεια το τροποποιεί ώστε να μεγιστοποιήσει την πληροφορία. Αποδίδει ένα μόνο μοτίβο.
(GibbsSampler: http://ccmbweb.ccv.brown.edu/cgi-bin/gibbs.12.pl?data_type=DNA)

Expectation Maximization (EM). Βασίζεται σε μια επαναληπτική διαδικασία που διαρκώς μεταβάλλει όχι μόνο τη σύσταση των μοτίβων αλλά και τη σχετική πυκνότητά τους μέσα σε μια αλληλουχία. Αποδίδει περισσότερα από ένα μοτίβα. (MEME: <http://meme.nbcr.net/meme/>)

Comparative Genomics

(Weeder: <http://pesolelab.ibbe.cnr.it/index.php/weederweb>)

Διαβάστε περισσότερα

Για τα PWM, PSSM

Computational Genome Analysis (Deonier, Tavare & Waterman) (Chapter 9)

Για την εντροπία και το πληροφοριακό περιεχόμενο

Computational Genome Analysis (Deonier, Tavare & Waterman) (Chapter 9)

<http://weblogo.berkeley.edu/>

Για την χρήση του βαθμού συντήρησης αλληλουχιών στην ανακάλυψη μοτίβων (Manolis Kellis PhD Thesis, MIT)

<http://web.mit.edu/manoli/www/thesis/Chapter3.html>