ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικό και Καποδιστριακό
Πανεπιστήμιο Αθηνών

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ
ΤΟΜΕΑΣ ΒΙΟΧΗΜΕΙΑΣ ΚΑΙ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΥΠΟΛΟΓΙΣΤΙΚΗ ΜΕΛΕΤΗ ΤΗΣ ΔΟΜΗΣ ΚΑΙ ΤΗΣ ΟΡΓΑΝΩΣΗΣ ΤΩΝ ΣΥΝΤΗΡΗΜΕΝΩΝ ΜΗ ΕΚΦΡΑΖΟΜΕΝΩΝ ΣΤΟΙΧΕΙΩΝ (CNE) ΣΤΑ ΕΥΚΑΡΥΩΤΙΚΑ ΓΟΝΙΔΙΩΜΑΤΑ ΩΣ ΕΡΓΑΛΕΙΟ ΔΙΕΡΕΥΝΗΣΗΣ ΤΗΣ ΠΙΘΑΝΗΣ ΛΕΙΤΟΥΡΓΙΑΣ ΚΑΙ ΤΗΣ ΕΞΕΛΙΚΤΙΚΗΣ ΔΥΝΑΜΙΚΗΣ ΤΟΥΣ

ΔΗΜΗΤΡΙΟΣ Κ. ΠΟΛΥΧΡΟΝΟΠΟΥΛΟΣ

Α.Μ 298007

ΑΘΗΝΑ 2014

HELLENIC REPUBLIC
National and Kapodistrian
University of Athens
SCHOOL OF SCIENCES
FACULTY OF BIOLOGY
DEPARTMENT OF BIOCHEMISTRY AND MOLECULAR BIOLOGY

Ph.D THESIS

COMPUTATIONAL STUDY OF THE STRUCTURE AND ORGANIZATION OF CONSERVED NON-CODING ELEMENTS (CNE) IN EUKARYOTIC GENOMES AS A TOOL IN ORDER TO ELUCIDATE THEIR POTENTIAL FUNCTION AND EVOLUTIONARY DYNAMICS

DIMITRIOS C. POLYCHRONOPOULOS

ATHENS 2014

Η έγκριση Διδακτορικής Διατριβής από τη Σχολή Θετικών Επιστημών του Πανεπιστημίου Αθηνών δε δηλώνει αποδοχή των γνωμών του συγγραφέα.

Ν. 5343/1932 Άρθρο 202

Συμβουλευτική επιτροπή

1)	Καθηγητής Γ.Κ. Ροδάκης (επιβλέπων)	Τμήμα Βιολογίας, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
2)	Καθηγητής Σ.Ι. Χαμόδρακας	Τμήμα Βιολογίας, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
3)	Ερευνητής Α' Δρ. Γ. Αλμυράντης	Ινστιτούτο Βιοεπιστημών και Εφαρμογών, Ε.Κ.Ε.Φ.Ε «ΔΗΜΟΚΡΙΤΟΣ»

Εξεταστική επιτροπή

1)	Καθηγητής Γ.Κ. Ροδάκης	Τμήμα Βιολογίας, Εθνικό & Καποδιστριακό Πανεπιστήμιο Αθηνών
2)	Καθηγητής Σ.Ι. Χαμόδρακας	Τμήμα Βιολογίας, Εθνικό & Καποδιστριακό Πανεπιστήμιο Αθηνών
3)	Ερευνητής Α' Δρ. Γ. Αλμυράντης	Ινστιτούτο Βιοεπιστημών & Εφαρμογών, Ε.Κ.Ε.Φ.Ε «ΔΗΜΟΚΡΙΤΟΣ»
4)	Ερευνητής Α' Δρ. Γ. Παλιούρας	Ινστιτούτο Πληροφορικής & Τηλεπικοινωνιών, Ε.Κ.Ε.Φ.Ε «ΔΗΜΟΚΡΙΤΟΣ»
5)	Επίκουρος Καθηγητής Χ. Νικολάου	Τμήμα Βιολογίας, Πανεπιστήμιο Κρήτης
6)	Επίκουρος Καθηγητής Π. Μπάγκος	Τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική, Πανεπιστήμιο Θεσσαλίας
7)	Λέκτορας Τ. Θηραίου	Τμήμα Βιοτεχνολογίας, Γεωπονικό Πανεπιστήμιο Αθηνών

“ Do. Or do not. There is no try”

— Master Joda

Αφιερωμένο στην Οικογένεια και στους συναδέλφους μου

Πρόλογος

Η παρούσα Διδακτορική Διατριβή πραγματοποιήθηκε στο Ινστιτούτο Βιοεπιστημών και Εφαρμογών (πρώην Ινστιτούτο Βιολογίας) του Εθνικού Κέντρου Έρευνας Φυσικών Επιστημών «Δημόκριτος», υπό την επίβλεψη του Δρ. *Γιάννη Αλμυράντη* και μέλη της Τριμελούς Συμβουλευτικής Επιτροπής τον Καθηγητή κ. *Γεώργιο Ροδάκη* και τον Καθηγητή κ. *Σταύρο Χαμόδρακα*.

Κατ' αρχήν, οφείλω πραγματικά ένα μεγάλο ευχαριστώ στον Γιάννη Αλμυράντη, για την εμπιστοσύνη που έχει δείξει στο πρόσωπο μου όλα αυτά τα χρόνια. Ήταν ένα μεσημέρι τον Σεπτέμβριο του 2009 όταν ήρθα στον «Δημόκριτο» πρώτη φορά και μιλήσαμε για πιθανά πράγματα που μπορούσαμε να δούμε στα πλαίσια της μεταπτυχιακής διατριβής του Μ.Δ.Ε «Βιοπληροφορική» του Τμήματος Βιολογίας του Πανεπιστημίου Αθηνών. Η ευγένειά του, η καλοσύνη του και η ανεξάντλητη υπομονή του αποτέλεσαν για μένα μαθήματα ζωής όλα αυτά τα χρόνια. Τον ευχαριστώ θερμά γιατί με την αγάπη του για γνώση και διαρκή αναζήτηση με δίδαξε τον τρόπο με τον οποίο γίνεται η έρευνα, μου έδωσε σημαντικά εφόδια και γνώσεις για να αντιλαμβάνομαι και να επεξεργάζομαι τα διάφορα ερωτήματα που εγείρονται κατά τη διάρκεια μιας επιστημονικής μελέτης και με έκανε να γνωρίσω και να αγαπήσω την Βιοπληροφορική / Υπολογιστική Γονιδιωματική ως ένα πολυθεματικό, διεπιστημονικό πεδίο έρευνας. Σε κάθε στάδιο της παρούσας διατριβής ήταν διαρκώς δίπλα μου με πολύτιμες συμβουλές, αφιερώνοντας μου πολύτιμο προσωπικό του χρόνο και όντας πάντα «ανοικτός» και άριστος συνομιλητής. Πραγματικά αισθάνομαι πολύ τυχερός που οι δρόμοι μας συναντήθηκαν.

Μεγάλη ευγνωμοσύνη οφείλω επίσης στον Καθηγητή κ. *Γεώργιο Ροδάκη*. Τον ευχαριστώ θερμά γιατί με την υποστήριξη του και τις εύστοχες συμβουλές του και ιδέες, με βοήθησε σε όλα τα στάδια της διδακτορικής διατριβής. Ήταν πάντα διαθέσιμος για συζητήσεις σε ευρύτερα θέματα που αφορούν την Εξέλιξη και την Γονιδιωματική. Τον ευχαριστώ θερμά ειδικά για τη βοήθειά του ώστε να οργανωθεί σωστά η παρούσα διατριβή και για παρόμοια βοήθεια κατά τη σύνταξη των εκθέσεων προόδου. Οι συζητήσεις μας για τις αποδόσεις βιολογικών όρων στα Ελληνικά ήταν πάντα γόνιμες και αιτιολογημένες και τον ευχαριστώ για το λόγο ότι ήταν πάντα διαθέσιμος.

Ένα μεγάλο ευχαριστώ οφείλω στον Καθηγητή κ. *Σταύρο Χαμόδρακα* τον οποίο είχα την τύχη και χαρά να γνωρίσω πρώτη φορά κατά τη διάρκεια των συνεντεύξεων του Μ.Δ.Ε «Βιοπληροφορική», του οποίου είναι Διευθυντής και Επιστημονικός Υπεύθυνος. Καταρχήν, τον ευχαριστώ επειδή με επέλεξε για το συγκεκριμένο μεταπτυχιακό και διότι μέσα από τις διαλέξεις του, αλλά και των άλλων υπευθύνων μαθημάτων του προγράμματος, έδωσε σε εμάς όλους τη δυνατότητα να δούμε και πέρα από τα «στενά» όρια της ειδικότητας που ο καθένας είχε

σπουδάσει, ως βασικό πτυχίο, και μας εισήγαγε στον κόσμο της Βιοπληροφορικής. Όποτε τον χρειάστηκα ήταν πάντα διαθέσιμος και τον ευχαριστώ θερμά και γι'αυτό.

Συνολικά, οφείλω ένα μεγάλο ευχαριστώ και στα τρία μέλη της Συμβουλευτικής Επιτροπής για την τιμή που μου έκαναν να συμμετέχουν σε αυτή, για την αμέριστη συμπαράσταση τους, το χρόνο που μου αφιέρωσαν και γιατί από την πρώτη μας γνωριμία, αποτέλεσαν πρότυπο για εμένα.

Η παρουσία και μόνο των αξιότιμων μελών της Επταμελούς Εξεταστικής Επιτροπής, του Ερευνητή Α' κ. *Γιώργου Παλιούρα*, του Επικουρου Καθηγητή κ. *Χριστόφορου Νικολάου*, του Επίκουρου Καθηγητή κ. *Παντελή Μπάγκου* και, τέλος, της Λέκτορος κας *Τριάδας Θηραίου* με τιμά ιδιαίτερα και τους ευχαριστώ θερμά.

Στο σημείο αυτό θα ήθελα να ευχαριστήσω τα μέλη της ερευνητικής ομάδας του εργαστηρίου *Θεωρητικής Βιολογίας και Υπολογιστικής Γονιδιωματικής*, υπεύθυνος του οποίου είναι ο Δρ. *Γιάννης Αλμυράντης*. Πιο συγκεκριμένα τον *Κωνσταντίνο Αποστόλου*, με τον οποίο μοιραστήκαμε το ίδιο γραφείο τα τελευταία 4 χρόνια, καθώς και πολλές συζητήσεις για ποικίλα θέματα επιστημονικής και μη φύσεως. Ένα μεγάλο ευχαριστώ οφείλω επίσης στον *Διαμαντή Σελλή*, με τον οποίο συνεργαστήκαμε από την πρώτη στιγμή που ήρθα στο εργαστήριο, παρά το γεγονός ότι ο ίδιος βρίσκεται πολλά χιλιόμετρα μακριά. Ευχαριστίες οφείλω και σε όλους τους ανθρώπους που γνώρισα και συνεργάστηκα αυτά τα χρόνια στα πλαίσια της διατριβής αυτής: στους συναδέλφους Δρ. *Γεώργιο Παλιούρα*, Δρ. *Αναστασία Κριθαρά* και Δρ. *Γιώργο Γιαννακόπουλο*, από το Εργαστήριο Τεχνολογίας Γνώσεων και Λογισμικού του Ινστιτούτου Πληροφορικής, που με εισήγαγαν στον κόσμο των Γραφημάτων (Γράφων) Ν – γραμμάτων (N-gram graphs). Τους δύο πρώτους τους ευχαριστώ και για τον επιπλέον λόγο ότι συνεργαστήκαμε μαζί για το έργο FP7 BioASQ – A challenge on large scale biomedical semantic indexing and question answering, του οποίου ο Δρ. *Γ. Παλιούρας* είναι επιστημονικός υπεύθυνος. Οφείλω επίσης πολλές ευχαριστίες στον Δρ. *Philipp Bucher* για την ευκαιρία που μου έδωσε να επισκεφτώ το εργαστήριό του στο Πολυτεχνείο της Λωζάνης (EPFL) για 3 μήνες, στα πλαίσια μιας υποτροφίας μικρής διάρκειας της Ευρωπαϊκής Οργάνωσης Μοριακής Βιολογίας (EMBO). Βεβαίως είμαι ευγνώμων και στην ίδια την EMBO για τη χρηματοδότηση και την υποστήριξη που μου παρείχε για όλο εκείνο το διάστημα. Στο EPFL γνώρισα αξιόλογους ανθρώπους με τους οποίους συνεργαστήκαμε σε κοινά ερευνητικά θέματα. Ήταν ιδιαίτερη τιμή και χαρά που γνώρισα την Δρ. *Slavica Dimitrieva* και την ευχαριστώ θερμά για τη βοήθειά της κατά την παραμονή μου εκεί και τις πολύωρες συζητήσεις, μαζί και με τον Δρ. *Philipp Bucher*, για την πιθανή προέλευση αυτών των υπερσυντηρημένων στοιχείων (CNE). Ευχαριστώ θερμά και τους συνεργάτες μας από το Εθνικό Συμβούλιο Έρευνας της Ιταλίας στη Ρώμη. Πιο συγκεκριμένα τους Δρ. *Emanuel Weitschek* και Δρ. *Giovanni Felici* με τους οποίους ολοκληρώσαμε μια

προσπάθεια κατάταξης των CNE με τη βοήθεια αλγορίθμων μηχανικής μάθησης. Με τον Emanuel γνωριστήκαμε σε ένα Practical Course της EMBO το 2012 και από τότε γίναμε φίλοι και ήταν ιδιαίτερη χαρά και τιμή για μένα που μπορέσαμε να κάνουμε μια δουλειά μαζί από απόσταση και χωρίς να έχουμε καμία χρηματοδότηση, μόνο από αγάπη και μεράκι.

Θα ήθελα να ευχαριστήσω και τα παλαιότερα μέλη του εργαστηρίου και φίλους μου Δρ. *Χριστόφορο Νικολάου*, κ. *Ιωάννη Τσιάγκα* και κα *Λαμπρινή Αθανασοπούλου*, με τους οποίους έχουμε συνεργαστεί όλα αυτά τα χρόνια και έχουμε μοιραστεί πολλές ευχάριστες στιγμές. Η συνεργασία μαζί τους ήταν άψογη και τους ευχαριστώ θερμά για αυτό. Τον *Χριστόφορο* τον ευχαριστώ ιδιαίτερα για τον επιπλέον λόγο ότι με βοήθησε πάρα πολύ με το να μου δείξει εργαλεία και μεθόδους που χρησιμοποιούνται στην υπολογιστική ανάλυση του γονιδιώματος και κάνουν την προσαρμογή ενός βιολόγου στα της Βιοπληροφορικής (σταδιακά) πιο εύκολη. Το chat του gmail σου, Χριστόφορε, ήταν πάντα ανοικτό και το εκτιμώ πολύ που πάντα ήσουν εκεί να μου γράψεις μια γραμμή (κώδικα και μη...).

Ευχαριστίες πρέπει να απευθύνω στο *Τμήμα Βιολογίας* του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών και στον *Τομέα Βιοχημείας και Μοριακής Βιολογίας* και όλα τα *μέλη* του (*ΔΕΠ, ερευνητές και διοικητικό προσωπικό*). Η εύρυθμη λειτουργία του Τομέα αποτέλεσε ένα καθοριστικής σημασίας στοιχείο για την απρόσκοπτη διεξαγωγή της διατριβής.

Τέλος, και κάθε άλλο παρά ήσσονος σημασίας, θα ήθελα να ευχαριστήσω την *Οικογένεια μου* και συγκεκριμένα τους *Γονείς μου Κώστα και Γρηγορία* και *τον αδερφό μου Γιώργο* που είναι πάντοτε κοντά μου και μου παρέχουν αγάπη, κατανόηση και ουσιαστική και πρακτική συμπαράσταση σε οτιδήποτε και αν επιχειρώ. Η παρούσα διατριβή δεν θα ήταν εφικτή χωρίς την απεριόριστη αγάπη, υπομονή, κατανόηση και στήριξή τους όλα αυτά τα χρόνια και τους ευχαριστώ από τα βάθη της καρδιάς μου για αυτό.

Η παρούσα έρευνα δεν έλαβε κάποια συγκεκριμένη χρηματοδότηση. Ωστόσο συμμετείχα σε ερευνητικά προγράμματα με συντονιστές το Ε.Κ.Ε.Φ.Ε «Δημόκριτος» και το Γεωπονικό Πανεπιστήμιο Αθηνών (BioASQ και ΘΑΛΗΣ αντίστοιχα). Οφείλω να ευχαριστήσω τον Δρ. *Γ. Παλιούρα* και την Καθηγήτρια κα *Ε. Τσακαλίδου* από τον «Δημόκριτο» και το Γεωπονικό Πανεπιστήμιο Αθηνών αντίστοιχα, που μου έκαναν την τιμή να συμμετέχω σε ερευνητικά προγράμματα.

Πίνακας Περιεχομένων

1. ΕΙΣΑΓΩΓΗ 1

1.1 Περιγραμμα Διατριβης 2

2. ΣΥΝΤΗΡΗΜΕΝΕΣ ΜΗ ΚΩΔΙΚΟΠΟΙΟΥΣΕΣ ΑΛΛΗΛΟΥΧΙΕΣ (CNE) ΣΤΑ ΓΟΝΙΔΙΩΜΑΤΑ ΜΕΤΑΖΩΩΝ 5

2.1 Τι ποσοστο του γονιδιωματος ειναι λειτουργικο; 5
2.2 Αναζητησεις λειτουργικων μη κωδικοποιουσων αλληλουχιων χρησιμοποιωντας κριτηρια συντηρησης αλληλουχιας 6
2.3 Χαρακτηριστικα Συντηρημενων Μη Κωδικοποιουσων Αλληλουχιων 10
2.4 Λειτουργιες Συντηρημενων Μη Κωδικοποιουσων Αλληλουχιων 12
2.5 Τα CNE ακολουθουν κατανομες τυπου νομου δυναμης σε μια ποικιλια γονιδιωματων ως αποτελεσμα της δυναμικης του γονιδιωματος 16

3. ΚΑΤΑΤΑΞΗ ΑΛΛΗΛΟΥΧΙΩΝ ΤΟΥ ΓΟΝΙΔΙΩΜΑΤΟΣ ΠΟΥ ΒΡΙΣΚΟΝΤΑΙ ΥΠΟ ΕΠΙΛΕΚΤΙΚΟ ΠΕΡΙΟΡΙΣΜΟ ΜΕ ΜΕΘΟΔΟΥΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ 44

3.1 Καταταξη CNE και εξωνιων με τη χρηση διανυσματων χαρακτηριστικων αλληλουχιας και ταξινομητων λογικων κανονων 44
3.2 Αναλυση και ταξινομηση αλληλουχιων DNA, που βρισκονται υπο επιλεκτικο περιορισμο, με τη χρηση γραφων (γραφηματων) N – γραμματων (N – gram Graphs, NGG) και γονιδιωματικων υπογραφων (GS) 66

4. ΑΝΑΛΥΣΗ ΤΗΣ ΧΡΩΜΟΣΩΜΙΚΗΣ ΚΑΤΑΝΟΜΗΣ ΤΩΝ CNE ΜΕ ΧΡΗΣΗ ΜΕΘΟΔΩΝ ΚΛΙΜΑΚΩΣΗΣ ΕΝΤΡΟΠΙΑΣ (ENTROPIC SCALING) ΚΑΙ ΕΓΚΙΒΩΤΙΣΜΟΥ (BOX COUNTING) 79

5. ΣΥΝΟΨΗ 86

6. ΒΙΒΛΙΟΓΡΑΦΙΑ 89

7. ΣΥΝΤΟΜΟ ΒΙΟΓΡΑΦΙΚΟ ΣΗΜΕΙΩΜΑ ΜΕ ΔΗΜΟΣΙΕΥΣΕΙΣ 101

Περίληψη

Η αλληλούχηση και συγκριτική ανάλυση πολλών γονιδιωμάτων θηλαστικών κατέδειξε ότι τουλάχιστον ένα 5.5% του ανθρώπινου γονιδιώματος βρίσκεται κάτω από επιλεκτική πίεση κατά την εξελικτική πορεία του. Από αυτό, το 1.5% εκτιμάται ότι κωδικοποιεί πολυπεπτιδικές αλυσίδες ενώ το 3.5% φαίνεται πως παίζει ρυθμιστικό ρόλο. Εν τούτοις, ο βαθμός κατανόησής μας για τους ρόλους που επιτελεί μεγάλο μέρος του συντηρημένου DNA που δεν κωδικοποιεί πρωτεΐνες ποικίλει. Μία από τις σημαντικότερες ανακαλύψεις που προέκυψαν από την ολική στοίχιση γονιδιωμάτων θηλαστικών ήταν η ταυτοποίηση εκατοντάδων «υπερσυντηρημένων» στοιχείων (UltraConserved Elements, UCE) μήκους άνω των 200 bp τα οποία δείχνουν απόλυτη (100%) συντηρητικότητα μεταξύ των γονιδιωμάτων του ανθρώπου, του ποντικού και του αρουραίου. Ένα στα τέσσερα από αυτά τα στοιχεία επικαλύπτουν μερικώς γνωστά γονίδια που κωδικοποιούν πρωτεΐνες. Παρόλα αυτά, τόσο υψηλό βαθμό συντηρητικότητας (100%) δεν αναμένουμε ούτε για εξώνια γονιδίων, λόγω του εκφυλισμού του γενετικού κώδικα. Από τότε που ανακαλύφθηκαν τα UCE έγιναν προσπάθειες για τον εντοπισμό συντηρημένων στοιχείων σε ολικές στοιχίσεις γονιδιωμάτων δύο ή περισσοτέρων ειδών, με κριτήριο χαμηλότερο κατώφλι ομοιότητας και διαφορετικά κατώφλια ελάχιστου μήκους της συντηρημένης ακολουθίας. Επιπλέον, χρησιμοποιήθηκε ως κριτήριο αποκλεισμού στοιχείων, η παρουσία τους μέσα σε γονίδια που κωδικοποιούν πρωτεΐνες. Στην παρούσα διατριβή χρησιμοποιούμε συγκεντρωτικά τον όρο Συντηρημένα Μη Εκφραζόμενα Στοιχεία (Conserved Noncoding Elements, CNE) παρά το γεγονός ότι στη βιβλιογραφία αναφέρονται και ως UCEs, UCNEs, HCNEs, LCNEs, CNGs, κλπ. Όταν αναφερόμαστε σε μια συγκεκριμένη τάξη στοιχείων τότε χρησιμοποιούμε την εκάστοτε ονομασία.

Τα CNE δεν είναι καινοτομία των σπονδυλωτών γιατί ανάλογα στοιχεία ανιχνεύονται και στα γονιδιώματα ασπονδύλων, καθώς και φυτών μέσω στοιχίσεων μεταξύ μελών της εκάστοτε ταξινομικής ομάδας. Εντούτοις, στο σχετικά πρόσφατο εξελικτικό παρελθόν των γονιδιωμάτων των σπονδυλωτών, το μέσο μήκος και ο βαθμός συντηρητικότητας των CNE παρατηρούνται να έλαβαν μεγαλύτερες τιμές, σχετικά με τις άλλες ταξινομικές ομάδες, ενώ οι ρόλοι που φαίνεται ότι απέκτησαν είναι ιδιαίτερα σημαντικοί.

Τα CNE φαίνεται πως δεν κατανέμονται τυχαία στο ανθρώπινο και σε άλλα γονιδίωματα. Μάλιστα, κατά ένα ποσοστό, συναθροίζονται κοντά σε γονίδια που εμπλέκονται στη ρύθμιση της μεταγραφής ή/και γενικότερα, στην ανάπτυξη. Χρησιμοποιώντας ανάλυση μικροσυστοιχιών έγινε γνωστό ότι ένα μεγάλο ποσοστό μη κωδικοποιούντων UCE εμφανίζουν ιστοειδικά επίπεδα έκφρασης (σε επίπεδο λειτουργικού RNA), ενώ απορρυθμίζονται σε ορισμένα είδη καρκίνου. Οι

γονιδιακές έρημοι είναι συνήθως εμπλουτισμένες σε CNE ενώ, στα γονιδιώματα θηλαστικών, η πλειοψηφία αυτών των στοιχείων ευρίσκεται σε μεγάλες αποστάσεις από τα πλησιέστερα γονίδια. Έχει δημοσιευτεί πληθώρα μελετών που προτείνουν ότι τα CNE βρίσκονται όντως υπό επιλεκτική πίεση κατά την εξέλιξή τους και δεν αποτελούν σημεία με χαμηλότερο ρυθμό μεταλλάξεων (mutational cold spots). Παρά ταύτα, λίγα είναι γνωστά για το ποιά είναι η λειτουργία τους σε κυτταρικό επίπεδο. Μελέτες δείχνουν ότι ενδεχομένως δρουν ως ρυθμιστές της μεταγραφής, δηλαδή ως ενισχυτές ή μονωτές, ωστόσο τα περισσότερα (με μία εξαίρεση) *in vivo* πειράματα σε ποντίκια, όπου γίνεται αφαίρεση κάποιων από αυτά τα στοιχεία, δε δίνουν κάποιο ορατό φαινοτυπικό αντίκτυπο, κάνοντας ακόμα πιο πολύπλοκη την όποια ερμηνεία βιοχημικών και υπολογιστικών πειραμάτων. Έχει επίσης διατυπωθεί μια εναλλακτική υπόθεση, σύμφωνα με την οποία τα CNE μεταφέρονται οριζόντια μεταξύ γενεών και συσσωρεύονται κατά τη μακρά εξελικτική πορεία. Σε μια μελέτη, επιπλέον, προτάθηκε ότι κάποια CNE ενδεχομένως δρουν ως περιοχές πρόσδεσης στον πυρηνικό φάκελο (Matrix Attachment Regions, MARs) διαδραματίζοντας το ρόλο αλληλουχιών που ρυθμίζουν την αρχιτεκτονική της χρωματίνης μέσω εξειδικευμένης πρόσδεσης συγκεκριμένων πρωτεϊνών. Τα CNE έχει αναφερθεί, μάλιστα, ότι εμπλέκονται στη φαινοτυπική ποικιλομορφία και σε ποικιλία ασθενειών κυρίως σχετιζόμενων με αναπτυξιακές διαδικασίες.

Στην παρούσα διατριβή επιχειρήσαμε να αναλύσουμε την χωροταξική οργάνωση των Συντηρημένων Μη Εκφραζομένων Στοιχείων (CNE) σε γονιδιώματα σπονδυλωτών και ασπόνδυλων, με σκοπό να διαπιστώσουμε αν μπορούμε να εξάγουμε κάποια συμπεράσματα για το πώς εξελίχθησαν αυτές οι αλληλουχίες με βάση την κατανομή τους στα χρωμοσώματα. Διαπιστώσαμε ότι οι αποστάσεις αυτών ακολουθούν κατανομές τύπου νόμου δύναμης σε μια ποικιλία γονιδιωμάτων. Τέτοιου τύπου κατανομές συνδέονται με συσχετίσεις μακράς εμβέλειας και μορφοκλασματικότητα (έννοιες που έχουν προταθεί για τη στερεοδιαμόρφωση της δομής της χρωματίνης του πυρήνα) και φαίνεται ότι απαντώνται πολύ συχνά στο γονιδίωμα, όπως προκύπτει από τη μελέτη διαφόρων στοιχείων του, σε πληθώρα οργανισμών. Δεδομένου ότι τα CNE σχετίζονται χωρικά με γονίδια, ειδικά με αυτά που ρυθμίζουν αναπτυξιακές διαδικασίες, επιβεβαιώσαμε ότι ένα πρότυπο νόμου δύναμης διατηρείται ανεξάρτητα από το εάν συμπεριληφθούν στοιχεία που βρίσκονται εντός ή εκτός γονιδίων. Όσο πιο «αρχαία» είναι αυτά τα στοιχεία τόσο πιο εκτεταμένες γραμμικότητες δίνουν σε διπλή λογαριθμική κλίμακα, δηλαδή τόσο πιο πολύ συμβάλουν στις παρατηρούμενες κατανομές. Προτείναμε ένα εξελικτικό μοντέλο για την κατανόηση αυτών των ευρημάτων που περιλαμβάνει γεγονότα τμηματικών διπλασιασμών ή διπλασιασμών ολόκληρου του γονιδιώματος και απαλοιφές των περισσοτέρων από τα διπλασιασμένα CNE. Προσομοιώσεις που πραγματοποιήσαμε αναπαράγουν τα κύρια χαρακτηριστικά των παρατηρουμένων κατανομών μεγέθους.

Με βάση τα παραπάνω ευρήματα, προχωρήσαμε και σε έναν άλλο τύπο ανάλυσης της χρωμοσωμικής κατανομής των CNE, με χρήση μεθόδων κλιμάκωσης εντροπίας *Shannon* (Shannon entropy scaling) και εγκιβωτισμού (box counting) που έχουν αναπτυχθεί στο εργαστήριο. Οι συγκεκριμένες μέθοδοι κάνουν εκτίμηση των χαρακτηριστικών μορφοκλασματικότητας σε ένα σύνολο δεδομένων και έχουν χρησιμοποιηθεί για τη μελέτη της κατανομής άλλων στοιχείων του γονιδιώματος, όπως είναι οι κωδικοποιούσες αλληλουχίες και τα μεταθετά στοιχεία. Ενδείκνυνται για τη μελέτη της κατανομής των CNE ειδικότερα, διότι τα τελευταία έχει προταθεί μέσω πειραμάτων 3C (Chromosome Conformation Capture) ότι αλληλεπιδρούν μεταξύ τους και συνεπώς ενέχονται πιθανόν σε συσχετίσεις μακράς εμβέλειας. Παρατηρήσαμε ενδιαφέροντα πρότυπα κατανομής, χαρακτηριστικά των διαφόρων κλάσεων CNE, που διαφοροποιούνται σύμφωνα με το εξελικτικό βάθος συντηρητικότητας των εκάστοτε στοιχείων.

Τα CNE παρουσιάζουν ενδιαφέρουσες ιδιότητες σύστασης και γι'αυτό επιχειρήσαμε να δούμε αν μπορούν να κατηγοριοποιηθούν με βάση αυτές τους τις ιδιότητες. Πιο συγκεκριμένα είναι γενικά αλληλουχίες πλούσιες σε A+T ενώ περιβάλλονται από περιοχές χαμηλού A+T. Προσπαθήσαμε, λοιπόν, να ταξινομήσουμε στοιχεία που βρίσκονται υπό επιλεκτική πίεση (εξώνια και CNE) με δύο μεθόδους μηχανικής μάθησης: «Γραφήματα Ν-γραμμάτων» (N-Gram Graphs, NGGs) και «Ανάλυση κ-μερών» (Logic Alignment Free, LAF). Διαπιστώσαμε ότι και με τις δύο μεθόδους, που για πρώτη φορά εφαρμόστηκαν στα πλαίσια ανάλυσης γονιδιωματικών δεδομένων, είναι εφικτή η κλασμάτωση αλληλουχιών του γονιδιώματος (CNE, εξώνια) σε διαφορετικές κατηγορίες μεταξύ γονιδιωμάτων ή εντός του ίδιου γονιδιώματος. Χρησιμοποιήσαμε στις αναλύσεις / συγκρίσεις μας κατάλληλες αναπληρωματικές αλληλουχίες που απομονώνονταν από το εκάστοτε γονιδίωμα έτσι ώστε να έχουν ίδιο μήκος και ποσοστό GC% με τις υπό μελέτη αλληλουχίες μας (CNE / εξώνια). Συγκρίναμε τα αποτελέσματα ταξινόμησης που πήραμε και από τις δύο μεθόδους με μια άλλη ευρέως διαδεδομένη προσέγγιση διαχωρισμού ολόκληρων γονιδιωμάτων που αναφέρεται ως «Γονιδιωματικές Υπογραφές» (Genomic Signatures, GS). Η μελέτη μας αυτή ήταν η πρώτη εφαρμογή των «Γονιδιωματικών Υπογραφών» στην κατάταξη μικρών βιολογικών αλληλουχιών μεγέθους < 50 kb.

Για τις ανάγκες όλων των προαναφερθέντων πειραματικών προσεγγίσεων προχωρήσαμε και σε ταυτοποίηση καινούριων στοιχείων CNE στα γονιδιώματα του ανθρώπου (*H. sapiens*), του σκώληκα (*C. elegans*) και της μύγας (*D. melanogaster*). Τα στοιχεία αυτά ταυτοποιήθηκαν έτσι ώστε να προέρχονται από οργανισμούς που να έχουν αποκλίνει από τον κοινό τους εξελικτικό πρόγονο παρόμοιες χρονικές περιόδους. Ενδιαφέρουσες συσχετίσεις και διαφοροποιήσεις μεταξύ αυτών των στοιχείων παρατηρήθηκαν με τη χρήση μεθόδων μηχανικής μάθησης που αναφέρθηκαν πιο πριν. Πιο συγκεκριμένα είδαμε ότι αλληλουχίες CNE που

παρουσιάζουν υψηλή ομοιότητα (> 95% και έως 100%) μεταξύ στοιχίσεων γονιδιωμάτων ανθρώπου / κοτόπουλου φαίνεται πως συνιστούν μια διακριτή κατηγορία υπερσυντηρημένων στοιχείων που επιτελεί λειτουργίες που μένει να ανακαλυφθούν. Το εντυπωσιακό αυτό ποσοστό συντηρητικότητας είναι ακόμα μεγαλύτερο από αυτό που παρατηρείται στα εξώνια (συγκρίνοντας τους δύο αυτούς οργανισμούς, άνθρωπο - κοτόπουλο), ενώ δεν είναι γνωστή κάποια λειτουργία στη φύση, που να απαιτεί τόσο υψηλό βαθμό ομοιότητας σε επίπεδο αλληλουχίας.

Λέξεις κλειδιά: Συντηρημένες μη κωδικοποιούσες αλληλουχίες; υπερσυντηρημένες αλληλουχίες; κατάταξη βιολογικών αλληλουχιών; συγκριτική γονιδιωματική

Abstract

The sequencing and comparative analysis of many mammalian genomes has indicated that at least 5.5% of the human genome is under selective constraint; of that, 1.5% is estimated to code for proteins, 3.5% displays known regulatory functions, while for the function of the rest there is little or no information available. One of the most interesting discoveries that have arisen from comparative analysis among mammalian genomes is the discovery of hundreds of ultraconserved elements (UCEs) of more than 200bp in length that show absolute conservation among the human, mouse and rat genomes. One out of four of UCEs overlaps known protein-coding genes. However, we do not expect such a high degree of conservation (100%) even in exons, due to the degeneration of the genetic code. Since the discovery of UCEs, there have been efforts to identify conserved elements based on lower thresholds of sequence identity over whole genome alignments of two or more species. Several thresholds of minimal length of conserved sequence have been used as well as the exclusion of elements inside protein-coding genes. Throughout this thesis, we use the term CNE(s) for Conserved Noncoding Elements to describe all such elements despite their specific characterization as UCEs, UCNEs, HCNEs, CNGs, CNEs etc in the related literature. We have used the specific name only when we refer to the corresponding class of elements.

CNEs are not a vertebrate innovation, but are also found in invertebrate and plant genomes. The vertebrate, insect and nematode CNEs are not related to each other at the sequence level. In the relatively recent evolution of vertebrate genomes, the mean length and conservation of CNEs found therein are the highest observed, regarding all taxonomic groups, while the conjectured roles they have acquired are particularly important.

CNEs are not randomly distributed in the human and other genomes. They are often clustered in the vicinity of genes involved in transcriptional regulation and/or development. Using microarray analysis it was reported that a large fraction of noncoding UCEs have tissue-specific expression levels and are deregulated in human cancer. Gene deserts are usually enriched in CNEs while, in mammalian genomes, the vast majority of those elements are found at long distances from the closest genes, exceeding in some cases 2Mb, which is the limit for any known cis regulatory element. There is a corpus of literature suggesting that CNEs are selectively constrained and not mutational cold spots. Despite this fact, little is known about what their function could be at the cellular level. Studies showing that CNEs might act as transcriptional regulators, e.g. enhancers or insulators, have been published, although most (with one exception) *in vivo* experiments of elimination of some of these elements yield viable mice, rendering more

difficult the interpretation of any biochemical or computational experiment. The alternative hypothesis that CNEs are horizontally transferred between lineages and accumulate during the course of long-term evolution has also been expressed. Furthermore, a study has suggested that CNEs might act as Matrix-Attachment Regions (MARs) by serving as sequences that regulate the architecture of chromatin through specific binding of particular proteins. An association between CNEs and phenotypic variation and disease has also been reported.

In the present thesis, we attempted to analyse the spatial organization of Conserved Noncoding Elements (CNEs) in vertebrate and invertebrate genomes with the aim to investigate whether we could deduce how those sequences evolved. We found out that the distances of consecutive CNEs follow power law-like distributions in a variety of genomes. Such kinds of distributions are associated with long range correlations and fractality (notions that have been proposed for the conformation of the chromatin inside the nucleus) and seem to occur frequently in the genome as evidenced by the study of different genomic elements in a variety of organisms. Given that CNEs are spatially associated with genes, especially with those that regulate developmental processes, we verified by appropriate gene masking that a power-law-like pattern emerges irrespectively of whether elements found inside protein-coding genes are excluded or not. In addition, we found that the more ancient elements form the most extended linearities in log log plots, when the distances between ancient CNEs are plotted. An evolutionary model was put forward for the understanding of these findings that includes *segmental or whole genome duplication* events and *eliminations (loss)* of most of the duplicated CNEs. Simulations reproduce the main features of the observed size distributions. Power-law-like patterns in the genomic distributions of CNEs are in accordance with current knowledge about their evolutionary history in several genomes.

Based on our initial findings we proceeded to another type of analysis of the chromosomal distribution of CNEs using the methods of Shannon entropy scaling and box counting, adopted from the field of *Information Theory*. Those methods are used to measure the fractality features in a dataset and have been employed in our lab for the study of the distribution of other genomic elements, such as coding sequences and transposable elements. They are especially suited in the case of CNEs, as the latter have been shown via Chromosome Conformation Capture (3C) to be involved in long range correlations. We observed interesting distributional patterns, characteristic of the different classes of CNEs studied, that differentiate according to the evolutionary depth of CNEs.

CNEs display interesting DNA composition preferences. This prompted us to investigate whether we could classify them by means of their sequence characteristics alone. More specifically, CNEs are generally AT rich sequences while they are surrounded by regions of low

AT content. We attempted to classify constrained elements in general (exons and CNEs) using two machine learning approaches: N-Gram Graphs (NGGs) and Logic Alignment Free (LAF). The application of those of two methodologies in the field of genomics is presented for the first time in this thesis. Overall, we managed to effectively classify genomic sequences of functional (or presumably functional) roles into different categories between genomes or inside the same genome. We used pairwise comparisons to do our analysis and naturally – occuring surrogate sequences that are of the same length and GC content with each one of the sequences comprising the studied dataset (CNEs / exons). We compared the classification rates obtained using both these approaches (NGGs and LAF) with another methodology, widely implemented in disciminating whole genomes, that is called «Genomic Signatures» (GS). Our study is the first one demonstrating the applicability of the GS approach in disciminating short biological sequences of length < 50 kb.

For the sake of all the above mentioned approaches, we also proceeded to the identification of new Conserved Noncoding Elements in the human (*H. sapiens*), worm (*C. elegans*) and insect (*D. melanogaster*) genomes. In those case, the species selected for CNE identification are characterized by the fact that evolutionary distances with every pair of whole genome alignments are close. We managed to discriminate those sequences efficiently and proposed biological interpretations. More specifically, CNE that display high sequence similarity (> 95% and up to 100%) between human / chicken whole genome alignments are thought to compose a distinct category of ultraconserved elements that probably play roles in processes that are yet to be determined. This remarkable percentage of sequence similarity is even greater than the one observed for exonic sequences (comparing the two organisms, human / chicken) while there is no known function that requires such a high degree of conservation.

Keywords: Conserved noncoding elements (CNEs); ultraconserved elements (UCEs); classification of biological sequences; comparative genomics

Δημοσιεύσεις που έχουν προκύψει στα πλαίσια της παρούσας διδακτορικής διατριβής:

1. Polychronopoulos D., Sellis D., Almirantis Y. **Conserved noncoding elements follow power-law-like distributions in several genomes as a result of genome dynamics.** 2014, *PLoS One*, 9(5):e95437.

2. Polychronopoulos D., Krithara A., Nikolaou C., Paliouras G., Almirantis Y., Giannakopoulos G. **Analysis and Classification of Constrained DNA Elements with N – gram Graphs and Genomic Signatures.** 2014 in *Algorithms Comput. Biol. SE - 18* (Dediu, A.-H., Martín-Vide, C. & Truthe, B.) 8542, 220–234 (Springer International Publishing, 2014).

3. Polychronopoulos D., Tsiangas G., Athanasopoulou L., Sellis D., Almirantis Y. **Long-range order and fractality in the structure and organization of eukaryotic genomes**. 2014, *World Scientific Publishing*, (in press).

4. Polychronopoulos D.*, Weitschek E.*, Dimitrieva S., Bucher P., Felici G., Almirantis Y. **Classification of selectively constrained DNA elements using feature vectors and rule-based classifiers.** 2014, *Genomics*, 104(2), 79 – 86 (*equally contributing authors*).

Άλλες Δημοσιεύσεις:

5. Gambus A.*, van Deursen F.*, Polychronopoulos D., Foltman M., Jones R.C., Edmondson R.D., Calzada A., Labib K. **A key role for Ctf4 in coupling the MCM2-7 helicase to DNA polymerase alpha within the eukaryotic replisome.** 2009, *EMBO J.*, Oct 7;28(19):2992-3004. Epub 2009 Aug 6 (**equally contributing authors*).
6. Tsatsaronis G., Partalas I., Balikas G., Zschunke M., Alvers M.R., Weissenborn D., Krithara A., Petridis S., Polychronopoulos D., Almirantis Y., Malakasiotis P., Androutsopoulos I., Pavlopoulos J., Baskiotis N., Gallinari P., Artieres T., Ngonga A., Heino N., Gaussier E., Alvers L.B., Schroeder M., Paliouras G. **An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition.** 2014, *BMC Bioinformatics* (revisions).
7. Zegos D., Papageorgiou D., Amaral-Psarris A., Xiucheng Fan A., Kastrissianaki E., Polychronopoulos D., Kerdidani D., Bungert J., Strouboulis J. **Mammalian expression vectors for the N- or C-terminal metabolic biotinylation tandem affinity tagging of proteins by the E. coli BirA biotin ligase.** 2014, *Journal of Biotechnology* (revisions).

1. Εισαγωγή

Οι γενετικές οδηγίες για την ανάπτυξη και λειτουργία όλων των ζώντων οργανισμών είναι κωδικοποιημένες στο μόριο του δεοξυριβονουκλεϊκού οξέος (DNA). Η πληροφορία μεταφέρεται από το DNA με τη μορφή μιας αλληλουχίας μικρών βασικών δομικών ενοτήτων που ονομάζονται νουκλεοτίδια. Το σύνολο της γενετικής πληροφορίας ενός κυττάρου είναι γνωστό ως γονιδίωμα. Το ανθρώπινο γονιδίωμα αποτελείται από 3.2 δισεκατομμύρια ζευγών βάσεων που είναι οργανωμένα σε γραμμικά χρωμοσώματα (Human Genome Sequencing Consortium International 2004). Από μια υπολογιστική σκοπιά, ολόκληρο το γονιδίωμα μπορεί να θεωρηθεί ως μια μεγάλη συμβολοσειρά (string) τεσσάρων χαρακτήρων: {A, C, G και T}, όπου ο κάθε χαρακτήρας αντιπροσωπεύει ένα από τα τέσσερα νουκλεοτίδια. Το γονιδίωμα είναι (σε μεγάλο βαθμό) το ίδιο σε όλα τα κύτταρα ενός οργανισμού, ενώ διαφορετικά κύτταρα είναι σε θέση να επιτελέσουν πολύ διαφορετικές λειτουργίες. Ένα από τα αναπάντητα και μεγάλα ερωτήματα στο χώρο της Βιολογίας είναι η κατανόηση του τρόπου με τον οποίο οι γενετικές πληροφορίες που είναι κωδικοποιημένες στο DNA καθοδηγούν την ανάπτυξη του οργανισμού και τη λειτουργία του και δίνουν γένεση στην ποικιλομορφία που παρατηρείται στα κύτταρα κάθε οργανισμού. Ένα σημαντικό βήμα που έγινε προς αυτήν την κατεύθυνση είναι η αποκρυπτογράφηση του ανθρώπινου γονιδιώματος το 2003, η οποία πήρε πάνω από 10 χρόνια ώσπου να ολοκληρωθεί. Τα τελευταία χρόνια έχει σημειωθεί ραγδαία πρόοδος στις τεχνολογίες αλληλούχησης που οδήγησε στην εκθετική αύξηση του αριθμού των γονιδιωμάτων που έχουν αλληλουχηθεί. Έως σήμερα έχει διαβαστεί η αλληλουχία 1153 γονιδιωμάτων, ενώ είναι διαθέσιμο το γονιδίωμα 1285 οργανισμών σε πρόχειρη μορφή και 889 γονιδιώματα βρίσκονται σε διάφορα στάδια αλληλούχησης ή συναρμολόγησης (assembly)[1].

Στη σημερινή εποχή παρατηρείται υπέρμετρη αύξηση βιολογικών δεδομένων τα οποία παράγονται κατά έναν μαζικό ρυθμό και έχουν αλλάξει τον τρόπο με τον οποίο προσεγγίζει κανείς προβλήματα βιολογικής φύσεως σήμερα. Το πεδίο της υπολογιστικής γονιδιωματικής έχει συμβάλλει σε σημαντικό βαθμό στον μετασχηματισμό των ακατέργαστων (raw) βιολογικών δεδομένων σε χρήσιμη πληροφορία. Όντας ένα πολυθεματικό πεδίο, η υπολογιστική γονιδιωματική ενσωματώνει εργαλεία και μεθόδους από την επιστήμη των υπολογιστών, τη μοριακή βιολογία και τη στατιστική, με απώτερο

[1] http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html

σκοπό να απαντήσει σε βιολογικά ερωτήματα, χρησιμοποιώντας υπολογιστικές προσεγγίσεις. Τα πεδία εφαρμογής της υπολογιστικής γονιδιωματικής εκτείνονται σε ποικίλους τομείς της γονιδιωματικής: εύρεση ομοιοτήτων μεταξύ αλληλουχιών διαφορετικών οργανισμών, πρόβλεψη και χαρακτηρισμός ρυθμιστικών αλληλουχιών, συναρμολόγηση αλληλουχιών DNA, ανάλυση γονιδιώματος με σκοπό την εύρεση γονιδίων, κ.ά.

Στην παρούσα διατριβή αναλύουμε γονιδιωματικά δεδομένα από διάφορα σπονδυλωτά και ασπόνδυλα με σκοπό τον χαρακτηρισμό μιας ειδικής κατηγορίας στοιχείων του γονιδιώματος που είναι εντυπωσιακά συντηρημένα στην πλειοψηφία των σπονδυλωτών. Αυτά τα στοιχεία ενδεχομένως δρουν ως κύριοι ρυθμιστές του γονιδιώματος κατά την εμβρυική ανάπτυξη, η λειτουργία τους όμως παραμένει αινιγματική. Μια καλύτερη κατανόηση της λειτουργίας αυτών των στοιχείων, του μηχανισμού δράσης τους, καθώς και της εξελικτικής τους προέλευσης καθίσταται επιτακτική στην προσπάθεια διελεύκανσης των μηχανισμών που διέπουν την ανάπτυξη των οργανισμών.

1.1 Περίγραμμα Διατριβής

Στα παρακάτω κεφάλαια πραγματοποιούμε μια υπολογιστική ανάλυση των Συντηρημένων Μη Κωδικοποιουσών Αλληλουχιών (Conserved Noncoding Elements, CNE), τα οποία είναι συντηρημένα σε υψηλό ποσοστό στα σπονδυλωτά και περιγράφουμε τη συνεισφορά της παρούσας διατριβής στην κατανόηση των ιδιοτήτων σύστασης αυτών των στοιχείων, στη μελέτη του τρόπου οργάνωσής τους σε διάφορα γονιδιώματα, καθώς και σε πιθανά σενάρια εξελικτικής δυναμικής τους.

Στο Κεφάλαιο 2 δίνουμε γενικές πληροφορίες για τα CNE, τα χαρακτηριστικά αλληλουχίας τους, τους πιθανούς λειτουργικούς τους ρόλους, καθώς και για τον τρόπο με τον οποίο αυτά οργανώνονται στο ανθρώπινο και σε άλλα γονιδιώματα. Παρατηρούμε ότι αυτά κατανέμονται ακολουθώντας κατανομές τύπου νόμου δύναμης. Τα αποτελέσματά μας ενσωματώνονται σε ένα μοντέλο, το οποίο προτείνουμε ότι περιγράφει την κατανομή στο γονιδίωμα των στοιχείων που βρίσκονται υπό επιλεκτική πίεση (selectively constrained DNA elements) συνολικότερα και περιλαμβάνουν εξώνια και CNE.

Στο Κεφάλαιο 3 περιγράφουμε μεθοδολογίες που υιοθετούμε από το πεδίο της Επιστήμης Υπολογιστών, με σκοπό την ταξινόμηση στοιχείων που βρίσκονται υπό επιλεκτικό περιορισμό (εξώνια και CNE). Εισάγουμε τη μεθοδολογία ανάλυσης

συχνοτήτων k-mers χωρίς στοίχιση με τη χρήση λογικών κανόνων (Logic Alignment Free – LAF). Μέσω ζευγών συγκρίσεων στοιχείων του γονιδιώματος οδηγούμαστε στην εξαγωγή χρήσιμων βιολογικών συμπερασμάτων σχετικά με τη σύσταση διαφόρων κατηγοριών υπερσυντηρημένων στοιχείων. Συγκρίνουμε τα αποτελέσματά μας με την κλασική μεθοδολογία των γονιδιωματικών υπογραφών (Genomic Signatures, GS) και παρουσιάζουμε και γι'αυτήν την περίπτωση ποσοστά επιτυχούς ταξινόμησης κλάσεων αλληλουχιών. Στη συνέχεια περιγράφουμε την εφαρμογή, για πρώτη φορά στην ανάλυση γονιδιωμάτων, μιας μεθοδολογίας, που αρχικά αναπτύχθηκε για εξαγωγή περιλήψεων σε κείμενα από ομάδα ερευνητών Πληροφορικής του Ε.Κ.Ε.Φ.Ε «Δημόκριτος». Οι γράφοι (γραφήματα) N-γραμμάτων (N-Gram Graphs, NGG), όπως ονομάζονται, χρησιμοποιούνται εδώ για την αναπαράσταση αλληλουχιών (και ομάδων αλληλουχιών) διαφόρων πιθανών λειτουργικοτήτων και φαίνεται πως είναι αποτελεσματικοί στην ταξινόμηση μικρών αλληλουχιών που βρίσκονται εντός του ίδιου γονιδιώματος αλλά και μεταξύ γονιδιωμάτων. Και εδώ συγκρίνουμε τα αποτελέσματά μας με τις γονιδιωματικές υπογραφές.

Στο Κεφάλαιο 4 εισάγουμε τον αναγνώστη στην έννοια της θεωρίας της πληροφορίας και στις εφαρμογές που έχει στην ανάλυση γονιδιωμάτων. Πιο συγκεκριμένα χρησιμοποιούμε την εντροπία Shannon (Shannon block entropy) και τη μέθοδο του εγκιβωτισμού (boxcounting) για την ανάλυση της χρωμοσωμικής κατανομής των CNE, με σκοπό τον προσδιορισμό του βαθμού της μορφοκλασματικότητας (fractality). Τα αποτελέσματα που λαμβάνουμε παρουσιάζουν ενδιαφέρον, ιδιαίτερα όταν συγκρίνουμε διαφορετικές κατηγορίες CNE.

Στο Κεφάλαιο 5 συνοψίζουμε τα αποτελέσματα της διατριβής, συζητούμε τις πιθανές συνεισφορές της και προτείνουμε περαιτέρω κατευθύνσεις για μελλοντική έρευνα και προβληματισμό.

Στο Κεφάλαιο 6 παρατίθεται η σχετική βιβλιογραφία που χρησιμοποιείται για τη συγγραφή της παρούσας διατριβής.

Στο Κεφάλαιο 7 αναπαράγονται οι δημοσιεύσεις, σε διεθνή περιοδικά με κριτές, που έχουν προκύψει στα πλαίσια αυτής της διατριβής.

2. Συντηρημένες Μη Κωδικοποιούσες Αλληλουχίες (CNE) στα γονιδιώματα μετάζωων

2.1 Τι ποσοστό του γονιδιώματος είναι λειτουργικό;

Στις αρχές του ’70, ο Susumu Ohno έγραψε μια σειρά άρθρων σε μια προσπάθεια να ερμηνεύσει την ποικιλομορφία σε μέγεθος γονιδιωμάτων θηλαστικών (Ohno & others 1970). Πρότεινε ότι εκτεταμένα γεγονότα γονιδιακού διπλασιασμού, ακολουθούμενα από απώλεια λειτουργίας διπλασιασμένων αντιγράφων, ευθύνονται για το μεγάλο μέρος «άχρηστου» (“junk”) DNA στο γονιδίωμά μας. Αυτή η ιδέα ακόμα ισχύει ως ένα βαθμό (στο ανθρώπινο γονιδίωμα υπάρχουν ψευδογονίδια, ήτοι μη λειτουργικά αντίγραφα γονιδίων), αλλά τώρα με την ανάλυση πολλών γονιδιωμάτων έχει γίνει σαφές ότι τα πράγματα είναι πιο περίπλοκα. Για παράδειγμα, το γονιδίωμα του fugu[2] (*Takifugu rubripes*) έχει μέγεθος που αντιστοιχεί στο 1/8 του μεγέθους του ανθρώπινου γονιδιώματος (390 Mb), παρά το γεγονός ότι και σε αυτό έχει εντοπιστεί ότι συνέβησαν σχετικά πρόσφατα (μετά την απόσχιση των τετραπόδων) γεγονότα γονιδιωματικού διπλασιασμού. Η συμπαγής φύση του γονιδιώματός του οφείλεται στο μικρό ποσοστό επαναλαμβανόμενων αλληλουχιών και στο μειωμένο μέγεθος των εσωνικών και εξωγονιδιακών (intergenic) αλληλουχιών. Και στα δύο γονιδιώματα όμως, ανθρώπου και fugu, η πλειοψηφία του DNA δεν κωδικοποιεί πρωτεΐνες.

Ένα από τα προβλήματα που αντιμετωπίζει Μοριακή Βιολογία είναι ότι πολλές από τις ιδέες και ορισμούς που έστεκαν για πολύ καιρό έπρεπε να αλλάξουν καθώς ανέκυπταν νέα στοιχεία από το πεδίο της γονιδιωματικής ανάλυσης. Έτσι λοιπόν η ιδέα για το τί είναι ένα γονίδιο έχει αλλάξει, με το μοντέλο ένα γονίδιο – μια πολυπεπτιδική αλυσίδα να είναι πλέον απαρχαιωμένο, αφού έχουν βρεθεί γονίδια που επικαλύπτουν το ένα το άλλο, φτιάχνουν πολυάριθμα εναλλακτικά μετάγραφα ή έχουν αμοιβαία ρυθμιστικό ρόλο ευρισκόμενα σε γειτονικές θέσεις.

[2] fugu: αναφέρεται ως σύντμηση του *Takifugu rubripes*. Ο συγκεκριμένος οργανισμός είναι ένας τελεόστεος ιχθύς, του οποίου το γονιδίωμα δημοσιεύτηκε το 2002 και ήταν το πρώτο γονιδίωμα σπονδυλωτών που παρουσιάστηκε (μετά το ανθρώπινο). Ο κοινός πρόγονος ανθρώπου – Fugu υπολογίζεται ότι έζησε 450 MYA (Million Years Ago - Εκατομμύρια Χρόνια Πριν), συνεπώς CNE κοινά μεταξύ Fugu – Human είναι εξαιρετικά αρχαίας προέλευσης (ancient CNEs)

Στο επίπεδο της ρύθμισης της μεταγραφής, η πολυπλοκότητα είναι επίσης μεγάλη. Η διαρκής ανακάλυψη νέων υποκινητών, καταστολέων και ενισχυτών γονιδίων είναι ενδεικτική. Γενικά, δεν έχουμε κάποιο γενικό τρόπο αναγνώρισης ενός ενισχυτή και πράγματι δεν είναι ξεκάθαρο ότι αυτές οι αλληλουχίες που ονομάζουμε ενισχυτές σχετίζονται μεταξυ τους, δηλαδή ενδέχεται να δρουν χρησιμοποιώντας εντελώς διαφορετικούς μοριακούς μηχανισμούς (Harmston et al. 2013). Επιπλέον, είναι σχεδόν αδύνατο να πούμε με ακρίβεια πού ξεκινούν και πού τελειώνουν τα ρυθμιστικά στοιχεία του γονιδιώματος (ενισχυτές, υποκινητές, καταστολείς, μονωτές). Συνήθως τα δεδομένα μας εξαρτώνται από την ακρίβεια που παρέχουν τα λειτουργικά πειράματα που χρησιμοποιήθηκαν για τον καθορισμό του ρυθμιστικού στοιχείου.

Πρόσφατα αποτελέσματα για το 1% του ανθρώπινου γονιδιώματος από το ENCODE (ENCyclopedia Of DNA Elements) έχουν μεν εμπλουτίσει τη γνώση μας για το λειτουργικό τμήμα του γονιδιώματός μας αλλά γέννησαν και νέα ερωτήματα (The ENCODE Project Consortium 2004). Για παράδειγμα, ένα από τα πιο συναρπαστικά ευρήματα ήταν ότι > 90% των περιοχών που ελέγχθησαν μεταγράφονταν σε πρωτογενή μετάγραφα και ότι γενικά το γονιδίωμα είναι περισσότερο ενεργό μεταγραφικά από ότι πιστευόταν αρχικά. Αλλά τι σημαίνει αυτό; Πώς ορίζουμε μια λειτουργική αλληλουχία; Στο παρελθόν, μια λειτουργική αλληλουχία, όταν μεταλλασσόταν, είχε κάποιο αποτέλεσμα στην προσαρμοστικότητα του οργανισμού.

Η αλληλούχηση και συγκριτική ανάλυση πολλών γονιδιωμάτων θηλαστικών κατέδειξε ότι τουλάχιστον ένα 5.5% του ανθρώπινου γονιδιώματος βρίσκεται υπό επιλεκτική πίεση κατά την εξελικτική πορεία του. Από αυτό, περίπου 1.5% εκτιμάται ότι κωδικοποιεί πολυπεπτιδικές αλυσίδες, περίπου 3.5% φαίνεται πως παίζει ρυθμιστικό ρόλο, ενώ ο βαθμός κατανόησής μας για τους ρόλους που επιτελεί το υπόλοιπο (0.5 – 1.5%) ποικίλει (Lindblad-Toh et al. 2011). Για όλους αυτούς τους λόγους καθίσταται επιτακτική η μελέτη του τμήματος του γονιδιώματος που φέρεται πως διαδραματίζει ρυθμιστικό ρόλο.

2.2 Αναζητήσεις λειτουργικών μη κωδικοποιουσών αλληλουχιών χρησιμοποιώντας κριτήρια συντήρησης αλληλουχίας

Συγκρίνοντας ομοιότητες και διαφορές μεταξύ γονιδιωμάτων υπαρχόντων ειδών, η συγκριτική γονιδιωματική παρέχει τα εργαλεία για την ταυτοποίηση λειτουργικών γονιδιωματικών αλληλουχιών. Σύμφωνα με την ουδέτερη θεωρία της μοριακής εξέλιξης (Kimura 1984), «οι περισσότερες μεταλλάξεις συμβαίνουν τυχαία στα γονιδιώματα και είναι ουδέτερες για τον οργανισμό», δηλαδή δεν επηρεάζουν την ικανότητά του να

επιβιώσει και να αναπαραχθεί. Οι μεταλλάξεις, που επηρεάζουν αρνητικά τον οργανισμό, είναι λιγότερο πιθανό να επικρατήσουν και σταδιακά απομακρύνονται από τον πληθυσμό με την πάροδο του χρόνου. Με άλλα λόγια, οι μεταλλάξεις στο λειτουργικό DNA είναι λιγότερο πιθανό να γίνονται ανεκτές και, συνεπώς, αλληλουχίες που είναι υψηλά συντηρημένες μεταξύ απομακρυσμένων οργανισμών ενδεχομένως να είναι λειτουργικές. Επομένως, υποθετικά λειτουργικά στοιχεία, κωδικοποιούντα και μη, μπορούν να ταυτοποιηθούν επιτυχώς με συγκριτική ανάλυση γονιδιωμάτων.

Η ύπαρξη πολυάριθμων υψηλά συντηρημένων μη κωδικοποιουσών αλληλουχιών στα σπονδυλωτά αποκαλύφθηκε ακόμα και πριν από τη διαθεσιμότητα αλληλουχιών ολόκληρων γονιδιωμάτων. Συγκεκριμένα, πριν 20 χρόνια, οι Duret και συνεργάτες ανέλυσαν τα μη κωδικοποιούντα τμήματα των γονιδίων σε διαφορετικές τάξεις σπονδυλωτών και ταυτοποίησαν μεγάλες και υψηλά συντηρημένες αλληλουχίες στα μη κωδικοποιούντα τμήματα γονιδίων, ειδικότερα στις μη μεταφραζόμενες περιοχές (UnTranslated Regions, UTR) εκείνων των γονιδίων που είναι απαραίτητα για την κυτταρική ζωή (Duret et al. 1993). Δέκα χρόνια αργότερα, όταν ολοκληρώθηκε η αλληλούχηση του γονιδιώματος του ποντικού, παρουσιάστηκαν οι πρώτες αναλύσεις μεγάλης κλίμακας του γονιδιώματος που δεν περιορίζονταν μόνο σε περιοχές που σχετίζονται με γονίδια. Δεδομένου ότι, < 2% του ανθρώπινου γονιδιώματος κωδικοποιεί πολυπεπτιδικές αλυσίδες, έγινε σαφές ότι πολλά στοιχεία με πιθανούς λειτουργικούς ρόλους ευρίσκονται εκτός περιοχών που κωδικοποιούν πρωτεΐνες.

Για την ταυτοποίηση συντηρημένων αλληλουχιών που είναι πιθανόν να είναι λειτουργικές εφαρμόστηκαν διαφορετικά κριτήρια από διαφορετικές ερευνητικές ομάδες. Αυτό είχε ως αποτέλεσμα να ποικίλει ο ακριβής αριθμός και το μέγεθος αυτών των αλληλουχιών και να εξαρτάται από τα γονιδιώματα που συγκρίνονται κάθε φορά καθώς και από το ελάχιστο ποσοστό ομοιότητας σε επίπεδο αλληλουχίας. Η πιο συνηθισμένη προσέγγιση εύρεσης συντηρημένων αλληλουχιών βασίζεται σε υπολογισμούς ποσοστού ομοιότητας σε επίπεδο αλληλουχίας, χρησιμοποιώντας πολλαπλές ή κατά ζεύγη στοιχίσεις γονιδιωμάτων επιλεγμένων ειδών. Κάποιες εναλλακτικές και πιο πολύπλοκες μέθοδοι έχουν επίσης αναπτυχθεί, όπως για παράδειγμα η κατασκευή φυλογενετικών HMM (Hidden Markov Models) (Siepel et al. 2005), καθώς και μοντέλα φειδωλότητας που εφαρμόζονται σε στοιχίσεις πολλαπλών ειδών (Margulies et al. 2003). Όσον αφορά την επιλογή ειδών προς σύγκριση, διάφορες πρώιμες μελέτες συνέκριναν τα γονιδιώματα ανθρώπου και τρωκτικών, με σκοπό την ταυτοποίηση ρυθμιστικών στοιχείων στα θηλαστικά (Dermitzakis et al. 2002; Dermitzakis et al. 2003; Bejerano, Pheasant, et al.

2004). Περίπου 327000 συντηρημένες μη κωδικοποιούσες αλληλουχίες, με πιθανό ρυθμιστικό ρόλο, ταυτοποιήθηκαν από συγκρίσεις γονιδιωμάτων ανθρώπου – ποντικού, χρησιμοποιώντας ένα αυθαίρετο κριτήριο 70% ομοιότητας σε επίπεδο αλληλουχίας για πάνω από 100 ζεύγη βάσεων (bp) μιας στοίχισης χωρίς κενά (Dermitzakis et al. 2003; Dermitzakis et al. 2005). Χρησιμοποιώντας περισσότερο αυστηρά κριτήρια, ο αριθμός των ανιχνευόμενων συντηρημένων μη κωδικοποιούντων στοιχείων μειώνεται δραματικά, αλλά η πιθανότητα να είναι λειτουργικά αυξάνει. Οι πρώτοι που εφάρμοσαν πολύ αυστηρά κριτήρια ταυτοποίησης CNE ήταν οι Bejerano *et al.* (Bejerano, Pheasant, et al. 2004). Ταυτοποίησαν 481 αλληλουχίες DNA μεγαλύτερες από 200 νουκλεοτίδια που είναι απολύτως (100% ομοιότητα σε επίπεδο αλληλουχίας) συντηρημένες μεταξύ των γονιδιωμάτων του ανθρώπου, του ποντικού και του αρουραίου. Αυτές οι αλληλουχίες έχουν παραμείνει συντηρημένες για περισσότερα από 75 – 80 εκατομμύρια χρόνια (Million Years, Myr), που είναι ο χρόνος απόσχισης, προσεγγιστικά, των γενεών του ανθρώπου και των τρωκτικών (Hedges et al. 2006), και ονομάστηκαν «υπερσυντηρημένα στοιχεία» (UltraConserved Elements, UCE). Διάφορες άλλες μελέτες επικεντρώθηκαν στη σύγκριση πιο απομακρυσμένων ειδών: για παράδειγμα, συγκρίσεις μεταξύ των γονιδιωμάτων θηλαστικών και ψαριών έχουν οδηγήσει στην ταυτοποίηση αλληλουχιών που παρέμειναν συντηρημένες για πάνω από 450 Myr εξέλιξης των σπονδυλωτών. Οι Sandelin και συνεργάτες ταυτοποίησαν 3583 υποθετικές ρυθμιστικές περιοχές, αναζητώντας αλληλουχίες που είναι μεγαλύτερες από 50 νουκλεοτίδια και παρουσιάζουν ομοιότητα σε επίπεδο αλληλουχίας 95% μεταξύ των γονιδιωμάτων του ανθρώπου και του ποντικού, αλλά είναι ανιχνεύσιμες και στον fugu (Sandelin et al. 2004). Βασισμένοι σε στοιχίσεις γονιδιωμάτων ανθρώπου – fugu, οι Woolfe *et al.* ταυτοποίησαν 1373 στοιχεία που εμφανίζουν ομοιότητα για τουλάχιστον 100 νουκλεοτίδια (Woolfe et al. 2005). Σε μια προσπάθεια να βρεθούν στοιχεία ακόμα πιο «αρχαία», δηλαδή συντηρημένα πιο βαθειά στη φυλογένεση των σπονδυλωτών, οι Venkatesh και συνεργάτες σύγκριναν τα γονιδιώματα του ανθρώπου και του elephant shark. Ταυτοποίησαν 4782 CNE με ομοιότητα σε επίπεδο αλληλουχίας που ποικίλει από 71% έως 98% και μέσο μήκος 210 νουκλεοτίδια (Venkatesh et al. 2006). Αυτό το αποτέλεσμα ήταν συναρπαστικό και μη αναμενόμενο, καθώς η ομάδα των χονδριχθύων, όπου ανήκει ο elephant shark (*Callorhinchus milii*), απέκλινε από τον κοινό πρόγονο των σπονδυλωτών με οστά (bony vertebrates) περίπου 530 Myr πριν (Kumar & Hedges 1998). Το γεγονός ότι το γονιδίωμα του elephant shark περιέχει διπλάσια στοιχεία CNE από αυτά που έχουν ταυτοποιηθεί μέσω στοιχίσεων ανθρώπου και τελεόστεων ιχθύων καταδεικνύει ότι τα CNE στους τελεόστεους ιχθύες

έχουν εξελιχθεί με πιο ταχείς ρυθμούς ή έχουν χαθεί μετά την απόσχιση από τη γενεαλογία των θηλαστικών. Επιπλέον, αυτή η μελέτη έδειξε ότι ένα μεγάλος αριθμός CNE υπήρχε ακόμα πιο πριν από την απόσχιση χονδριχθύων και οστεϊχθύων και παρέμεινε πρακτικά αναλλοίωτος έκτοτε.

Στη βιβλιογραφία, οι συντηρημένες αλληλουχίες που δεν κωδικοποιούν πρωτεΐνες περιγράφονται με διαφορετικά ονόματα, από διαφορετικές ομάδες, που τις περιέγραψαν (*Πίνακας 1*), δίχως να υπάρχει κάποια απόδειξη λειτουργικής διαφοροποίησης αυτών των στοιχείων, πέρα από τα κριτήρια επιλογής και ταυτοποίησής τους. Παρόλου που έχουν δοθεί διαφορετικοί ορισμοί, ανάλογα με τα γονιδιώματα ειδών υπό σύγκριση, το ελάχιστο μήκος αλληλουχίας όπου παρατηρείται ομοιότητα, καθώς και το ελάχιστο κατώφλι ομοιότητας μεταξύ δύο ή περισσοτέρων οργανισμών, οι αλληλουχίες που εξάγονται κάθε φορά φαίνεται πως μοιράζονται πολλά κοινά χαρακτηριστικά (Elgar & Vavouri 2008; Nelson & Wardle 2013; Harmston et al. 2013). Στην παρούσα διατριβή αναφερόμαστε με τον όρο CNE στις Συντηρημένες Μη Κωδικοποιούσες Αλληλουχίες γενικότερα. Όπου εξετάζεται κάποια κλάση συντηρημένων στοιχείων ξεχωριστά, αυτή αναφέρεται με το όνομά της, π.χ. UCE, UCNE, CNG, κλπ.

Πίνακας 1: Ονοματολογία για την περιγραφή των μη κωδικοποιουσών αλληλουχιών που βρίσκονται υπό επιλεκτικό περιορισμό

Ακρωνύμιο	Περιγραφή	Ταυτοποίηση	Αναφορά
ANCOR	Ancestral non-coding conserved region	Διάφορες υπολογιστικές μέθοδοι	(Aloni & Lancet 2005)
CNC	Conserved non-coding	≥ 70% ομοιότητα για ≥ 100 νουκλεοτίδια στα γονιδιώματα ανθρώπου και ποντικού	(Couronne et al. 2003)
CNE	Conserved non-coding elements	Χρήση τοπικής ομολογίας αλληλουχίας (MegaBLAST) μεταξύ ανθρώπου και fugu για ≥ 100 νουκλεοτίδια.	(Woolfe et al. 2005)
CNEE	Conserved non-exonic elements	Συντηρημένες αλληλουχίες από ένα φυλογενετικό HMM (PhastCons) που εφαρμόστηκε σε πολλαπλές στοιχίσεις 40 γονιδιωμάτων θηλαστικών	(Lowe et al. 2011)
CNG	Conserved non-genic	≥ 70% ομοιότητα για ≥ 100 νουκλεοτίδια στα γονιδιώματα ανθρώπου και ποντικού	(Dermitzakis et al. 2002)
CNS	Conserved non-coding sequences	Διάφορες υπολογιστικές μέθοδοι	(Dubchak et al. 2000)
EC	Extremely conserved elements	≥ 80% ομοιότητα για ≥ 100 νουκλεοτίδια σε ένα υποσύνολο τουλάχιστον 40 γονιδιωμάτων θηλαστικών (44 στο σύνολο)	(Tseng & Tompa 2009)
HCE	Highly Conserved Elements	Ένωση αλληλουχιών που ταυτοποιήθηκαν σε άλλες μελέτες: UCE (Bejerano, Pheasant, et al. 2004), UCR (Sandelin et al. 2004) και CNE (Woolfe et al. 2005)	(Sun et al. 2006)
HCR	Highly conserved regions	≥ 70% ομοιότητα για ≥ 50 νουκλεοτίδια στα γονιδιώματα ανθρώπου και κοτόπουλου	(Duret & Bucher 1997)

HCNE	Highly conserved non-coding elements	Περιοχές ≥ 50 νουκλεοτίδια για τις οποίες η πιθανότητα να βρίσκονται υπό επιλεκτική πίεση, δεδομένου του σκορ συντήρησης στον άνθρωπο και στον σκύλο, είναι ≥ 95%	(Lindblad-Toh et al. 2005)
HCNR	Highly conserved non-coding regions	≥ 70% ομοιότητα για ≥ 100 νουκλεοτίδια στα γονιδιώματα ανθρώπου, ποντικού, κοτόπουλου, xenopus και fugu	(de la Calle-Mustienes et al. 2005)
LCNS	Long conserved non-coding sequences	≥ 95% ομοιότητα για ≥ 500 νουκλεοτίδια στα γονιδιώματα ανθρώπου και ποντικού	(Sakuraba et al. 2008)
MCS	Multispecies conserved sequences	Το κατώφλι συντήρησης τέθηκε ώστε το 5% του ανθρώπινου γονιδιώματος να βρίσκεται μέσα σε MCS, 12 είδη σπονδυλωτών συγκρίθηκαν	(Thomas et al. 2003)
PCNE	Phylogenetically Conserved non-coding elements	Τοπική ομοιότητα σε επίπεδο αλληλουχίας για ≥ 45 νουκλεοτίδια στα γονιδιώματα του amphioxus, zebrafish, ποντικού και ανθρώπου	(Hufton et al. 2009)
UCE	Ultraconserved elements	100% ομοιότητα για ≥ 200 νουκλεοτίδια στα γονιδιώματα του ανθρώπου, του ποντικού και του αρουραίου. 1 στα 4 από αυτά τα στοιχεία επικαλύπτονται με γνωστές κωδικοποιούσες αλληλουχίες	(Bejerano, Pheasant, et al. 2004)
UCNE	Ultraconserved non-coding elements	≥ 95% ομοιότητα για ≥ 200 νουκλεοτίδια στα γονιδιώματα ανθρώπου και κοτόπουλου	(Dimitrieva & Bucher 2013)
UCR	Ultraconserved regions	≥ 95% ομοιότητα για ≥ 50 νουκλεοτίδια στα γονιδιώματα του ανθρώπου και του ποντικού που στοιχίζονται εν μέρει με το γονιδίωμα του fugu	(Sandelin et al. 2004)

2.3 Χαρακτηριστικά Συντηρημένων Μη Κωδικοποιουσών Αλληλουχιών

Τα CNE παρουσιάζουν κάποιες χαρακτηριστικές ιδιότητες σύστασης που τα καθιστούν ένα σχετικά ομογενές σύνολο αλληλουχιών DNA και τα διακρίνουν από άλλες μη κωδικοποιούσες αλληλουχίες. Η εντυπωσιακή συντήρηση που παρατηρείται στα CNE για μεγάλες εξελικτικές περιόδους υποδηλώνει ότι έχουν εξελιχθεί υπό αυστηρή επιλεκτική πίεση (Drake et al. 2006; Katzman et al. 2007). Αυτό έχει ως αποτέλεσμα να θεωρούνται σήμερα οι πιο συντηρημένες αλληλουχίες DNA που υπάρχουν στη φύση, καθώς για τα εξώνια γνωρίζουμε ότι, λόγω του εκφυλισμού της τρίτης θέσης, δεν απαιτείται τόσο υψηλός βαθμός συντήρησης. Πράγματι αυτές οι αλληλουχίες στην πλειοψηφία τους επιδεικνύουν υψηλότερη συντήρηση από την πλειοψηφία των γονιδίων που κωδικοποιούν πρωτεΐνες (Bejerano, Pheasant, et al. 2004). Αναλύοντας τις κατανομές συχνοτήτων αλληλομόρφων, ερευνητές έδειξαν ότι αυτές οι περιοχές δεν αντιπροσωπεύουν περιοχές με χαμηλότερους ρυθμούς μεταλλάξεων αλλά πράγματι υπόκεινται σε επιλεκτική πίεση (Drake et al. 2006; Katzman et al. 2007).

Τα CNE δεν είναι μια καινοτομία των σπονδυλωτών αλλά ευρίσκονται επίσης σε γονιδιώματα ασπόνδυλων και φυτών (Vavouri et al. 2007; Glazov et al. 2005; Lockton & Gaut 2005). Τα CNE των σπονδυλωτών, των εντόμων και των νηματώδων δε σχετίζονται

γενικά μεταξύ τους από άποψη αλληλουχίας (Glazov et al. 2005; Siepel et al. 2005; Vavouri et al. 2006). Παρόλα αυτά, μια πρόσφατη μελέτη οδήγησε στην ταυτοποίηση δύο στοιχείων που είναι συντηρημένα μεταξύ σπονδυλωτών και ασπόνδυλων (Clarke et al. 2012) και αναμένεται ότι, με την πρόοδο στις μεθοδολογίες αλληλούχησης και την αυξανόμενη διαθεσιμότητα γονιδιωμάτων, θα ταυτοποιηθούν περισσότερα στο μέλλον. Κατά την πρόσφατη εξέλιξη των σπονδυλωτών, το μέσο μήκος και η συντήρηση των CNE έχουν τις μεγαλύτερες τιμές σε σχέση με όλες τις ταξινομικές ομάδες (Retelska et al. 2007), ενώ οι υποτιθέμενοι ρόλοι που απέκτησαν είναι ιδιαίτερα σημαντικοί στα σπονδυλωτά (Mikkelsen et al. 2007).

Τα CNE βρίσκονται συχνά σε συστάδες κοντά σε περιοχές του γονιδιώματος που περιέχουν γονίδια τα οποία εμπλέκονται στην εμβρυογένεση (Sandelin et al. 2004; Woolfe et al. 2005; Siepel et al. 2005). Πολλά από αυτά τα γονίδια κωδικοποιούν μεταγραφικούς παράγοντες και σηματοδοτικά μόρια που ενέχονται στη ρύθμιση της εμβρυικής ανάπτυξης (*trans – dev* γονίδια) και περιέχουν επικράτειες που επιτρέπουν αλληλεπιδράσεις πρωτεϊνών – DNA (Harmston et al. 2013). Αυτό οδηγεί στην υπόθεση ότι τα CNE ενδεχομένως είναι υπεύθυνα για την ακριβή εκτέλεση του περίτεχνου αναπτυξιακού προγράμματος, δρώντας ως *cis* ρυθμιστικά στοιχεία κατά τη μεταγραφή (Nelson & Wardle 2013).

Εξ’ ορισμού, τα CNE ευρίσκονται στις μη κωδικοποιούσες περιοχές του γονιδιώματος και μπορούν να βρεθούν σε περιοχές μεταξύ γονιδίων και σε περιοχές γονιδίων, μη κωδικοποιούσες, όπως τα εσώνια και οι μη μεταφραζόμενες περιοχές (UTR). Οι γονιδιακές έρημοι είναι πλούσιες σε CNE (Kim & Pritchard 2007; Stephen et al. 2008), ενώ στα γονιδιώματα θηλαστικών η μεγάλη πλειοψηφία αυτών των στοιχείων συναντώνται σε μεγάλες αποστάσεις από τα πλησιέστερα γονίδια που ξεπερνούν σε ορισμένες περιπτώσεις τα 2 εκατομμύρια βάσεις (Mb), που είναι το όριο για κάθε γνωστό *cis* ρυθμιστικό στοιχείο (Bishop et al. 2000; Lettice et al. 2003; Woolfe & Elgar 2008). Λίγα είναι γνωστά για το ποιά λειτουργία επιτελούν αυτά τα CNE. Δημοσιευμένες μελέτες συνηγορούν στην άποψη ότι ενδέχεται να αποτελούν περιοχές ρύθμισης γονιδίων (Gene Regulatory Blocks, GRB) και ότι μπορούν να λειτουργούν συνεργιστικά μαζί με τα γονίδια στόχους τους (Harmston et al. 2013; Kikuta et al. 2007; Nelson & Wardle 2013; Viturawong et al. 2013).

Η πλειοψηφία των CNE απαντώνται σε ένα αντίγραφο στο γονιδίωμα (Bejerano, Haussler, et al. 2004). Τα λίγα εκείνα στοιχεία, που εμφανίζουν μεταξύ τους ομοιότητα σε επίπεδο αλληλουχίας, έχουν προέλθει κυρίως από αρχαία γεγονότα διπλασιασμών μέσα

στη γενεαλογία των σπονδυλωτών και σχετίζονται με παράλογα γονίδια (Vavouri et al. 2006; McEwen et al. 2006). Τα CNE είναι πλούσια σε AT σε σχέση με το γονιδιωματικό περιβάλλον στο οποίο βρίσκονται, ενώ περιέχουν συχνά συστάδες πανομοιότυπων νουκλεοτιδίων (ομοπουρίνες / ομοπυριμιδίνες), καθώς και χαρακτηριστικές παλίνδρομες αλληλουχίες. Οι Walter και συνεργάτες ανέλυσαν τη νουκλεοτιδική σύσταση CNE ανθρώπου και fugu σε επίπεδο νουκλεοτιδίου και βρήκαν ότι είναι πλούσια σε A + T, πολύ περισσότερο από τις περιβάλλουσες αλληλουχίες που τα οριοθετούν, οι οποίες παρουσιάζουν σημαντική πτώση στο περιεχόμενο σε A + T, δίνοντας χαρακτηριστικό πρότυπο (Walter et al. 2005). Χαρακτηριστικό των CNE αποτελεί επίσης το γεγονός ότι πολλά από αυτά αποτελούν θέσεις πρόσδεσης, πολλές φορές επικαλυπτόμενες, μεταγραφικών παραγόντων (Xie et al. 2007). Με βάση όλα τα παραπάνω συμπεραίνει κανείς ότι το λεξιλόγιο και το συντακτικό των οδηγιών που ενδεχομένως είναι κωδικοποιημένες στα CNE παρουσιάζουν εξαιρετικό ενδιαφέρον και μένει να αποκαλυφθούν.

2.4 Λειτουργίες Συντηρημένων Μη Κωδικοποιουσών Αλληλουχιών

Παρά το γεγονός, ότι η υψηλή συντήρηση σε επίπεδο αλληλουχίας των CNE είναι ενδεικτική κάποιου πιθανού λειτουργικού ρόλου, η λειτουργία και ο λόγος συντήρησής τους παραμένει ένα μυστήριο. Όπως αναφέρθηκε πιο πάνω είναι γενικά αποδεκτό ότι τα CNE δρουν ως *cis* ρυθμιστικές αλληλουχίες παρέχοντας υψηλή εξειδίκευση στη χωρική, ποσοτική και χρονική ρύθμιση της γονιδιακής έκφρασης κατά τη διάρκεια της εμβρυικής ανάπτυξης. Ωστόσο μέχρι σήμερα δεν έχει περιγραφεί κάποιος μοριακός μηχανισμός που να απαιτεί τόσο υψηλό επίπεδο συντήρησης σε επίπεδο αλληλουχίας.

Από τη στιγμή που ανακαλύφθηκαν τα CNE, πολλές πειραματικές μελέτες πραγματοποιήθηκαν ώστε να ελεγθεί *in vivo* η ικανότητά τους να δρουν ως ρυθμιστικές αλληλουχίες, κυρίως χρησιμοποιώντας δοκιμασίες ελέγχου γονιδίων (reporter gene assays) σε διάφορα σπονδυλωτά, αλλά κυρίως στο ποντίκι (Nobrega et al. 2003; Pennacchio et al. 2006; Visel et al. 2008), στο zebrafish (Woolfe et al. 2005; Kikuta et al. 2007; Shin et al. 2005; Li et al. 2010), στο βάτραχο (de la Calle-Mustienes et al. 2005) και στο κοτόπουλο (Sabherwal et al. 2007; Maas et al. 2011). Οι προηγούμενες μελέτες έδειξαν ότι όντως πολλά CNE δρουν ως ενισχυτές. Πιο συγκεκριμένα, χρησιμοποιώντας έμβρυα ποντικών στην εμβρυική μέρα 11.5, οι Pennacchio και συνεργάτες έλεγξαν 167 αλληλουχίες, που είναι απολύτως συντηρημένες στα γονιδιώματα του ανθρώπου, του ποντικού και του αρουραίου και σημαντικά συντηρημένες στον fugu (Pennacchio et al. 2006). Βρήκαν ότι

45% από τις αλληλουχίες που ελέγθησαν λειτουργούν ως ενισχυτές, ενώ στην πλειοψηφία τους καθοδηγούν την έκφραση περιοχών του αναπτυσσόμενου νευρικού συστήματος. Παρόλο που οι περισσότερες πειραματικές μελέτες έχουν εστιαστεί στη μελέτη της υποτιθέμενης λειτουργίας των CNE ως ενισχυτές, ορισμένα CNE δρουν επίσης ως μονωτές. Πραγματοποιώντας *in vivo* μελέτες ενεργότητας μονωτή στο zebrafish για 13 CNE που δεν παρουσιάζουν δραστηριότητα ενισχυτή, οι Royo και συνεργάτες βρήκαν ότι 3 εξ' αυτών έχουν ενεργότητα παρεμπόδισης δράσης ενισχυτή (Royo et al. 2011). Επίσης οι Xie και συνεργάτες ταυτοποίησαν περιοχές πρόσδεσης του γνωστού μονωτή CTCF σε αλληλουχίες CNE, γεγονός που κάνει ακόμα πιο δύσκολη και ασαφή την ακριβή απόδοση λειτουργικοτήτων σε διάφορες κατηγορίες CNE (Xie et al. 2007).

Παρά τις μελέτες που αναφέρθηκαν πιο πάνω, ενεργότητες ενισχυτή / μονωτή δεν έχουν επιβεβαιωθεί ευθέως για όλα τα CNE, ενώ ακόμα και αυτές οι μελέτες δεν έδωσαν απαντήσεις για τις εξειδικευμένες ιδιότητες σύστασης των CNE που είναι υπεύθυνες για τις ρυθμιστικές τους λειτουργίες, ούτε ήταν σε θέση να εξηγήσουν τους λόγους της υπερβολικής συντήρησης αυτών των στοιχείων. Επιπλέον, παρόλο που τα CNE είναι τόσο καλά συντηρημένα σε απομακρυσμένα είδη και τα στοιχεία που προέρχονται από ένα είδος μπορούν να επάγουν την έκφραση σε άλλα είδη (Matsunami & Saitou 2013; Matsunami et al. 2010; Woolfe et al. 2005), τα πρότυπα έκφρασης που επάγονται από ορθόλογα CNE στα διάφορα είδη δεν είναι συντηρημένα στην πλειοψηφία των περιπτώσεων που εξετάστηκαν (Ritter et al. 2010).

Κάποιες άλλες μελέτες έχουν προτείνει την ιδέα ότι τα CNE ενδεχομένως φέρουν κωδικοποιημένες πολλαπλές λειτουργικότητες με επικαλυπτόμενες ομάδες οδηγιών. Για παράδειγμα, οι Feng και συνεργάτες χαρακτήρισαν ένα CNE που δρα ως ενισχυτής σε επίπεδο DNA ενώ σε επίπεδο RNA δρα ως ένα μόριο που δεν κωδικοποιεί και στρατολογεί την πρωτεΐνη *Dlx2* πίσω στον ενισχυτή σε ένα σύμπλοκο DNA – RNA – πρωτεΐνης (Feng et al. 2006). Έχει επίσης προταθεί, όπως αναφέρθηκε ελάχιστα πρωτύτερα, ότι τα CNE αποτελούνται από πολλαπλές θέσεις πρόσδεσης μεταγραφικών παραγόντων (Transcription Factor Binding Sites, TFBS) (Boffelli et al. 2004; Vavouri & Elgar 2005; Viturawong et al. 2013). Σε αυτήν την περίπτωση, κάθε νουκλεοτίδιο ενδεχομένως φέρει αρκετούς λειτουργικούς περιορισμούς στους οποίους υπόκειται και μικρές αλλαγές στην αλληλουχία θα είχαν καταστροφικά αποτελέσματα στην πρόσδεση συμπλόκων μεταγραφικών παραγόντων στην αλληλουχία CNE (Elgar & Vavouri 2008). Παρόλα αυτά, η συγκεκριμένη υπόθεση δεν έχει επιβεβαιωθεί ούτε από υπολογιστικά, ούτε από πειραματικά αποτελέσματα.

Τα CNE παίζουν ρόλο στην πρόκληση ασθενειών. Πιο συγκεκριμένα, σημειακές μεταλλάξεις στα CNE μπορούν να προκαλέσουν ασθένειες, όπως προαξονική πολυδακτυλία (Lettice et al. 2003), σύνδρομο Pierre Robin (Benko et al. 2009), χειλεοσχιστία (Rahimov et al. 2008) και ολοπροσεγκεφαλία (Jeong et al. 2008). Και άλλες, όμως, ασθένειες έχουν συσχετισθεί με αλλαγές στην αλληλουχία των CNE, όπως η κώφωση (de Kok et al. 1996), η Leri-Weill δυσχονδροστέωση (Sabherwal et al. 2007), ο κίνδυνος εμφάνισης της νόσου Hirschsprung (Emison et al. 2005) και άλλες. Εντυπωσιακό, ωστόσο, παραμένει το γεγονός ότι η απαλοιφή συγκεκριμένων CNE στο ποντίκι δε δίνει κάποιο ανιχνεύσιμο, ορατό φαινότυπο (Nóbrega et al. 2004; Ahituv et al. 2007). Συγκεκριμένα, πρώτοι οι Nobrega και συνεργάτες αφαίρεσαν δύο μεγάλες, μη κωδικοποιούσες, περιοχές, μεγέθους 1 Mb, στα χρωμοσώματα 3 και 19 του ποντικού (Nóbrega et al. 2004). Αυτές οι περιοχές περιέχουν ένα σύνολο 1243 ταυτοποιημένων CNE (> 70% ομοιότητα για πάνω από 100 νουκλεοτίδια σε στοιχίσεις ανθρώπου – ποντικού). Δεν παρατηρήθηκαν φαινοτυπικές αλλαγές ακόμα και μετά την απαλοιφή 4 μη κωδικοποιούντων UCE από τη συλλογή του Bejerano (100% ομοιότητα για ≥ 200 νουκλεοτίδια σε στοιχίσεις ανθρώπου – τρωκτικών). Ενδιαφέρον είναι επίσης ότι τα UCE, που αφαιρέθηκαν, είχε δειχθεί προηγουμένως ότι δρουν ως ενισχυτές και εντοπίζονται κοντά σε γονίδια που παρουσιάζουν αλλαγές στην έκφραση όταν απενεργοποιούνται (Ahituv et al. 2007). Αυτές οι δύο μελέτες έγειραν το ερώτημα της λειτουργικής σημασίας των CNE. Είναι πολύ πιθανό, βέβαια, οι απαλοιφές CNE ενδεχομένως να οδηγούν σε φαινοτύπους που δεν μπορούν να ανιχνευθούν με βάση τις εκάστοτε περιβαλλοντικές / εργαστηριακές συνθήκες. Αυτό έχει δειχθεί σε πρόσφατες μελέτες στη *Drosophila*, όπου ένας φαινότυπος που σχετίζεται με την απαλοιφή ενός CNE γίνεται εμφανής μόνο κάτω από συνθήκες εσωτερικού ή εξωτερικού stress (Frankel et al. 2010; Perry et al. 2010). Επιπλέον, μπορεί να υποστηριχθεί ότι οι φαινότυποι που προκύπτουν μπορεί να είναι πολύ ήπιοι ώστε να ανιχνεύονται από τους ανθρώπους, αλλά αρκετά έντονοι ώστε να αφαιρούνται από τη φυσική επιλογή. Μόνο πολύ πρόσφατα, οι Nolte και συνεργάτες έδειξαν ότι η αφαίρεση του ενισχυτή Μ280 οδηγεί σε σημαντικά μικρότερα, σε μέγεθος, ποντίκια. Η περιοχή, όπου βρίσκεται ο Μ280, περιλαμβάνει μια υπερσυντηρημένη αλληλουχία που είναι πανομοιότυπη μεταξύ ανθρώπου, ποντικού και αρουραίου, συνεπώς αυτή είναι η πρώτη αναφορά φαινότυπου που προκύπτει από αφαίρεση ενός υπερσυντηρημένου στοιχείου (Nolte et al. 2014).

Με απώτερο σκοπό να σχολιαστούν όλα τα λειτουργικά στοιχεία του ανθρώπινου γονιδιώματος, το πρόγραμμα ENCODE χρησιμοποίησε έναν αριθμό διαφορετικών

πειραματικών τεχνικών, όπως ανοσοκατακρήμνιση χρωματίνης και αλληλούχηση (ChIP – Seq), υπερευαισθησία σε DNAάση, RNA – seq και πειράματα μεθυλίωσης DNA, ώστε να χαρτογραφηθούν ενεργές ρυθμιστικές περιοχές σε διαφορετικές κυτταρικές σειρές (The ENCODE Project Consortium 2004). Τα δεδομένα αυτά δείχνουν ότι σχεδόν οι μισές συντηρημένες περιοχές του γονιδιώματος θηλαστικών δε φέρουν βιοχημική ενεργότητα (Ward & Kellis 2012). Παρόλα αυτά, οι πειραματικές τεχνικές που χρησιμοποιήθηκαν στο ENCODE δεν είναι σε θέση να απομονώσουν τις αλληλεπιδράσεις DNA – πρωτεΐνών που συμβαίνουν σε ένα μικρό αριθμό κυττάρων και σε ένα στενό παράθυρο χρόνου κατά την ανάπτυξη του εμβρύου. Γι' αυτό το λόγο, εάν τα CNE δρουν μόνο κατά τη διάρκεια συγκεκριμένων σταδίων κατά την εμβρυική ανάπτυξη και σε μικρό αριθμό ιστών, δε θα αποκαλυφθεί η λειτουργία τους χρησιμοποιώντας δεδομένα που προκύπτουν από το ENCODE. Οι πιθανές λειτουργίες κάποιων CNE (ξεχωριστών ή σε ομάδες), μέσω των αλληλεπιδράσεών τους με συγκεκριμένα γονίδια μέσα στον πυρήνα και με τη βοήθεια της αναδίπλωσης της χρωματίνης, έλαβαν πρόσφατα πειραματική επιβεβαίωση με την εργασία των Viturawong και συνεργάτες (Viturawong et al. 2013). Χρησιμοποιώντας υψηλής ανάλυσης φασματομετρία μάζας σε συνδυασμό με άλλες τεχνικές, οι συγγραφείς της συγκεκριμένης μελέτης έδειξαν, σε μια συλλογή 193 υπερσυντηρημένων αλληλουχιών, ότι αυτά δρουν ως *cis* ρυθμιστικές αλληλουχίες, μέσω συγκεκριμένης διαμόρφωσης που λαμβάνει η χρωματίνη σε εκείνες τις περιοχές, όπου αλληλεπιδρούν. Περαιτέρω πειράματα τέτοιου είδους είναι αναγκαία ώστε να διελευκανθεί ο ρόλος του πλήθους των αλληλεπιδράσεων CNE με διάφορες περιοχές του γονιδιώματος.

Τελικά όμως, γιατί είναι τόσο δύσκολος ο ακριβής χαρακτηρισμός των CNE; Όπως αναφέραμε πιο πριν, εκτός ελαχίστων εξαιρέσεων, οι υπάρχουσες πειραματικές τεχνικές ευρείας κλίμακας δεν είναι κατάλληλες για τη μελέτη της λειτουργίας των CNE, σε διαφορετικά βιολογικά πλαίσια, καθώς επίσης και σε διαφορετικά αναπτυξιακά στάδια. Πειραματικές προσεγγίσεις, όπως η απαλοιφή διαφόρων CNE και οι κατευθυνόμενη μεταλλαξιγένεση αλληλουχίας, αποδεικνύονται χρονοβόρες και πραγματοποιούνται συνήθως κάτω από πολύ αυστηρές συνθήκες, κατά τις οποίες δεν μπορεί να αποκαλυφθεί η ενδεχόμενη πολυπλοκότητα των στοιχείων αυτών (και των αλληλεπιδράσεων στις οποίες μεσολαβούν). Από υπολογιστική σκοπιά, η μελέτη των CNE καθίσταται δύσκολη και γεμάτη προκλήσεις διότι, αντίθετα με τις κωδικοποιούσες αλληλουχίες όπου έχουμε έναν καθολικό (ή σχεδόν καθολικό) γενετικό κώδικα, η γλώσσα και το συντακτικό που είναι κωδικοποιημένα στα CNE παραμένουν ένα αίνιγμα. Έχουν γίνει προσπάθειες για λειτουργικό σχολιασμό μεγάλης κλίμακας CNE (Bejerano, Haussler, et al. 2004; Pedersen

et al. 2006), αλλά δεν έδωσαν πολλές πληροφορίες, ενώ ο λειτουργικός σχολιασμός των CNE αποτελεί μία από τις πιο σημαντικές προκλήσεις που αντιμετωπίζει το πεδίο της ανάλυσης αλληλουχιών και της Βιοπληροφορικής γενικότερα.

2.5 Τα CNE ακολουθούν κατανομές τύπου νόμου δύναμης σε μια ποικιλία γονιδιωμάτων ως αποτέλεσμα της δυναμικής του γονιδιώματος

Στις ακόλουθες παραγράφους περιγράφουμε μια προσπάθεια μελέτης της κατανομής των αποστάσεων μεταξύ διαδοχικών CNE, διαφόρων κατηγοριών και εξελικτικής προέλευσης. Ιδιαίτερη μνεία γίνεται στις έννοιες των νόμων δύναμης και της μορφοκλασματικότητας, ενώ με βάση τις παρατηρούμενες κατανομές προτείνεται στο τέλος ένα μοντέλο το οποίο προσπαθεί να ερμηνεύσει τις παρατηρούμενες κατανομές και συμφωνεί με την εξελικτική δυναμική των CNE στο ανθρώπινο και σε άλλα γονιδιώματα.

Εισαγωγή

Συσχετίσεις μακράς εμβέλειας έχουν αναφερθεί για τη νουκλεοτιδική αλληλουχία του μη κωδικοποιούντος τμήματος του γονιδιώματος, αμέσως μετά τη διαθεσιμότητα ολόκληρων γονιδιωμάτων (Li & Kaneko 1992; Peng et al. 1992; Voss 1992). Στο παρελθόν μελέτησε η ομάδα μας τα χαρακτηριστικά μακράς εμβέλειας διαφόρων στοιχείων του γονιδιώματος, όπως είναι οι αλληλουχίες που κωδικοποιούν πρωτεΐνες (Sellis & Almirantis 2009; Athanasopoulou et al. 2010) και τα μεταθετά στοιχεία (Sellis et al. 2007; Klimopoulos et al. 2012; Athanasopoulou et al. 2014) μέσω της μελέτης της κατανομής των αποστάσεων μεταξύ εξωνίων και επαναλαμβανόμενων στοιχείων αντίστοιχα. Στις περισσότερες περιπτώσεις διαπιστώθηκαν κατανομές τύπου νόμου δύναμης, μορφοκλασματικότητα και αυτο-ομοιότητα, που εκτείνονται σε αρκετές τάξεις μεγέθους. Εφαρμόζουμε και εδώ την ίδια μεθοδολογία (που αναπτύσσεται και πιο κάτω) για την ανάλυση των αποστάσεων μεταξύ διαδοχικών CNE. Γι' αυτόν το σκοπό, χρησιμοποιούμε δημοσιευμένα σύνολα δεδομένων CNE, τα οποία χαρακτηρίζονται από διαφορετικούς βαθμούς εξελικτικής συντήρησης και έχουν ταυτοποιηθεί σε μια πληθώρα οργανισμών, από τα ασπόνδυλα έως στα σπονδυλωτά. Ανιχνεύουμε κατανομές τύπου νόμου δύναμης μεταξύ των αποστάσεων των CNE στην πλειονότητα των περιπτώσεων που μελετήθηκαν (Polychronopoulos, Sellis, et al. 2014). Μία προγενέστερη μελέτη από τους Salerno και συνεργάτες (Salerno et al. 2006) ανέφερε την ύπαρξη κατανομής νόμου δύναμης στο μήκος «τέλειας συντηρημένης» αλληλουχίας από στοίχιση γονιδιωμάτων ανθρώπου – ποντικού. Η μελέτη, που

προτείνουμε, εστιάζει στις αποστάσεις μεταξύ διαδοχικών CNE σε ποικιλία γονιδιωμάτων και επιπλέον παρουσιάζουμε ένα εξελικτικό μοντέλο που επιβεβαιώνει τις παρατηρούμενες χρωμοσωμικές κατανομές.

Δεδομένης της ανίχνευσης, όπως αναφέραμε λίγο πιο πάνω, συσχετίσεων μακράς εμβέλειας στις μη συντηρημένες αλληλουχίες του γονιδιώματος ευκαρυωτικών, χρησιμοποιήθηκαν απλά μοντέλα μοριακής δυναμικής με την προοπτική να προσπαθήσουν να εξηγήσουν τα παρατηρούμενα πρότυπα. Ένα απλό σύστημα επέκτασης / τροποποίησης έχει δειχθεί πως δημιουργεί συσχετίσεις μακράς εμβέλειας μέσω της εναλλαγής διπλασιασμού συμβόλων και γεγονότων απαλοιφής συμβόλων (Li 1991). Πιο πρόσφατα ευρήματα πάνω στην ολίσθηση κλώνων (strand slippage), κατά τη διάρκεια της αντιγραφής, σε συνδυασμό με σημειακές μεταλλάξεις, έριξαν φως στις ομονουκλεοτιδικές αλληλουχίες και στην εξέλιξη των μικροδορυφορικών αλληλουχιών, ενώ ενδεχομένως αποτελούν μια ρεαλιστική απεικόνιση του πιο πάνω μοντέλου στη δυναμική του γονιδιώματος (δείτε (Athanasopoulou et al. 2010) όπου περιλαμβάνεται μια μικρή συζήτηση για εξελικτικά σενάρια που μπορούν να δώσουν συσχετίσεις μακράς εμβέλειας). Το μοντέλο, που ενσωματώνει τις ιδέες πάνω στις οποίες θεωρούμε ότι βασίζεται η κατανομή των CNE, προτάθηκε αρχικά από τους Sellis και συνεργάτες (Sellis & Almirantis 2009) για την κατανομή των αποστάσεων μεταξύ των εξωνίων και εμείς επεκτείναμε την εφαρμογή του σε γονιδιωματικές αλληλουχίες που βρίσκονται υπό επιλεκτική πίεση γενικότερα (εξώνια και CNE). Αυτό το εξελικτικό σενάριο βασίζεται, με τη σειρά του, σε ένα απλούστερο μοντέλο που προσπαθεί να εξηγήσει τις κατανομές μεγεθών τύπου νόμου δύναμης που εμφανίζονται στην ανάπτυξη μέσω συσσωματωμάτων των σωματιδίων σε φυσικοχημικά συστήματα (Takayasu et al. 1991). Ο μηχανισμός, όπως εφαρμόζεται στο γονιδίωμα στην περίπτωσή μας, περιλαμβάνει κυρίως τμηματικό διπλασιασμό (συμπεριλαμβανομένων γεγονότων διπλασιασμού ολόκληρου του γονιδιώματος) και απώλεια των περισσότερων από τα διπλασιασμένα CNE, μαζί με μια απώλεια, μέτριου βαθμού, μη διπλασιασμένων CNE σε ορισμένες περιπτώσεις.

Υλικά και Μέθοδοι

Σύνολα δεδομένων

Αναλύουμε τη χρωμοσωμική κατανομή διαφόρων κατηγοριών CNE (ορισμένες εκ των οποίων αναφέρονται στον *Πίνακα 1* πιο πάνω). Περιλαμβάνουμε στις αναλύσεις μας μια

φυλογενετικά ευρεία συλλογή δεδομένων, από τον άνθρωπο μέχρι τον elephant shark και από τα σπονδυλωτά έως τα ασπόνδυλα:

- **i.** 13736 CNE, που είναι χαρτογραφημένα πάνω στο ανθρώπινο γονιδίωμα (hg18) και είναι πανομοιότυπα για πάνω από 100 νουκλεοτίδια σε τουλάχιστον 3 από τα 5 πλακουντοφόρα θηλαστικά (Stephen et al. 2008). Ολόκληρο το σύνολο ονομάζεται EU100+. Συγκεκριμένα υποσύνολα λαμβάνονται υπόψη για τις ανάγκες των υπολογιστικών μας πειραμάτων ως ακολούθως (τα δεδομένα αυτά παραχωρήθηκαν ευγενικά από τον J.S. Mattick): **(ia)** 8332 στοιχεία από το EU100+ που δεν υπάρχουν στο ψάρι (fugu). Αυτά τα CNE εμφανίστηκαν κατά την εξέλιξη των τετράποδων (είναι παρόντα στον βάτραχο, στο κοτόπουλο και / ή στα θηλαστικά) και ονομάζονται EU-FR. **(ib)** 5404 στοιχεία από το EU100+ που έχουν ορθόλογα στο ψάρι («αρχαία» στοιχεία) και ονομάζονται FR. **(ic)** 1665 στοιχεία που είναι παρόντα στον βάτραχο, αλλά όχι στο ψάρι (εξειδίκευση τετράποδων) και ονομάζονται XT-FR. **(id)** 980 στοιχεία που είναι παρόντα στο κοτόπουλο αλλά όχι στον βάτραχο ή στο fugu (εξειδίκευση αμνιωτικών) και ονομάζονται GG-XT-FR. **(ie)** 600 στοιχεία που δεν είναι παρόντα στο κοτόπουλο ή στο βάτραχο ή στο fugu (εξειδίκευση θηλαστικών) και ονομάζονται EU-GG-XT-FR.
- **ii.** 82335 CNE θηλαστικών (συντηρημένα στα θηλαστικά αλλά όχι στο κοτόπουλο ή στο ψάρι) και 16575 CNE αμνιωτικών (συντηρημένα στα θηλαστικά και στο κοτόπουλο αλλά όχι στο ψάρι) τα οποία είναι χαρτογραφημένα πάνω στο ανθρώπινο γονιδίωμα (hg17) (Kim & Pritchard 2007).
- **iii.** 4386 UCNE (Υπερσυντηρημένα Μη Κωδικοποιούντα Στοιχεία, Ultraconserved Noncoding Elements, μεγαλύτερα από 200 νουκλεοτίδια), που είναι χαρτογραφημένα στο ανθρώπινο γονιδίωμα (hg19) και επιδεικνύουν ομοιότητα σε επίπεδο αλληλουχίας, που είναι μεγαλύτερη ή ίση με 95%, μεταξύ ολικών στοιχίσεων γονιδιωμάτων ανθρώπου και κοτόπουλου (Dimitrieva & Bucher 2013).
- **iv.** 3124 συντηρημένες μη κωδικοποιούσες αλληλουχίες μεταξύ ανθρώπου – fugu, που είναι χαρτογραφημένες στο ανθρώπινο γονιδίωμα (hg17) και παρουσιάζουν ομοιότητα σε ποσοστό 70% μεταξύ των γονιδιωμάτων των δύο αυτών οργανισμών (Pennacchio et al. 2006).

v. 2833 CNE, που είναι συντηρημένα μεταξύ ανθρώπου – zebrafish (*D. rerio*), είναι χαρτογραφημένα πάνω στο ανθρώπινο γονιδίωμα (hg17) και επιδεικνύουν ομοιότητα μεγαλύτερη από 70%, για πάνω από 80 νουκλεοτίδια, στα συγκεκριμένα γονιδιώματα (Shin et al. 2005).

vi. 4782 CNE που είναι συντηρημένα μεταξύ ανθρώπου – elephant shark (*C. milii*), είναι χαρτογραφημένα πάνω στο ανθρώπινο γονιδίωμα (hg17) και επιδεικνύουν ομοιότητα μεταξύ των δύο γονιδιωμάτων, που κυμαίνεται μεταξύ 71% και 98% (Venkatesh et al. 2006).

vii. 4519 PCNE (Phylogenetically CNEs), που είναι χαρτογραφημένα στο γονιδίωμα του zebrafish (Ensembl 42) και είναι συντηρημένα σε amphioxus[3], zebrafish, ποντίκι και fugu (Hufton et al. 2009).

viii. 23651 CNE εντόμων, που εμφανίζουν 100% ομοιότητα, σε επίπεδο αλληλουχίας, μεταξύ *D. melanogaster* / *D. pseudoobscura* για περισσότερα από 50 νουκλεοτίδια και είναι χαρτογραφημένα πάνω στο γονιδίωμα της *D.melanogaster* (dm1) (Glazov et al. 2005).

ix. 2082 CNE νηματώδων, που παρουσιάζουν μια μέση ομοιότητα 96% σε επίπεδο αλληλουχίας μεταξύ των γονιδιωμάτων των *C. elegans* / *C. briggsae* και είναι χαρτογραφημένα στο γονιδίωμα του *C. elegans* (WS140) (Vavouri et al. 2007).

x. 2614 μη κωδικοποιούσες αλληλουχίες, που χαρακτηριστικό τους είναι η εντυπωσιακή συντήρηση μεταξύ των γονιδιωμάτων του ανθρώπου, του ποντικού και του αρουραίου και είναι χαρτογραφημένα στο ανθρώπινο γονιδίωμα (hg17). Ένα υποσύνολο από αυτά τα στοιχεία έχει επιβεβαιωθεί πειραματικά ότι δρουν ως ενισχυτές κατά την ανάπτυξη (Visel et al. 2008).

Στις περισσότερες περιπτώσεις χρησιμοποιούμε τη σουίτα εργαλείων *BEDTools* για τις υπολογιστικές αναλύσεις και το χειρισμό αρχείων τύπου BED[4] (Quinlan & Hall 2010).

Μασκάρισμα γονιδίων και κωδικοποιουσών αλληλουχιών

[3] amphioxus (ή lancelet, lancet: νυστέρι): τάξη θαλάσσιων χορδωτών, τα οποία αποτελούν αντικείμενο έρευνας στη ζωολογία, διότι η μελέτη του γονιδιώματός τους παρέχει πληροφορίες για την εξέλιξη των σπονδυλωτών, την προσαρμογή τους στην ξηρά, καθώς και τον τρόπο με τον οποίο τα σπονδυλωτά χρησιμοποίησαν «παλαιά» γονίδια για νέες λειτουργίες.

[4] Τα αρχεία BED (Browser Extensible Data) είναι αρχεία συντεταγμένων διαφόρων στοιχείων γονιδιωμάτων. Δηλαδή αρχεία τριών στηλών του τύπου: chr αρχή τέλος.

Προχωράμε σε ένα πλήρες μασκάρισμα των περιοχών του ανθρώπινου γονιδιώματος (hg17 και hg18) που χαρακτηρίζονται ως γονιδιακές. Επιπλέον μασκάρουμε περιβάλλουσες αλληλουχίες γύρω από κάθε γονίδιο: 5 kb στο 5' άκρο και 2 kb στο 3' άκρο, ώστε να αποκλείσουμε *cis* – ρυθμιστικά στοιχεία, των οποίων ο εντοπισμός ενδεχομένως καθορίζεται από την χωροταξική οργάνωση του εκάστοτε γονιδίου υπό ρύθμιση. Η περιοχή που βρίσκεται ανοδικά από τις περιοχές έναρξης της μεταγραφής (Transcription Start Sites, TSS) είναι ιδιαιτέρως εμπλουτισμένη σε τέτοιες ρυθμιστικές αλληλουχίες. Εκτεταμένες μάσκες των 10 kb και 100 kb χρησιμοποιούνται επίσης κατά έναν παρόμοιο τρόπο (δείτε «Αποτελέσματα»), ενώ σε όλες τις περιπτώσεις κάνουμε χρήση των *BEDTools* και δικών μας scripts. Στην περίπτωση της *D. melanogaster*, όταν αναφερόμαστε σε μασκαρισμένα CNE εντόμων, εννοούμε στοιχεία που δεν επικαλύπτονται με εξώνια και περιοχές ματίσματος και τα λαμβάνουμε από το πρόσθετο υλικό της μελέτης των Glazov και συνεργάτες (Glazov et al. 2005). Δεν προχωρούμε σε μασκάρισμα άλλων στοιχείων του γονιδιώματος, όπως είναι τα μεταθετά στοιχεία (Transposable Elements - TE), για τα οποία υπάρχουν ενδείξεις ότι ακολουθούν και αυτά κατανομές τύπου νόμους δύναμης, διότι δεν υπάρχουν ενδείξεις λειτουργικής αλληλεπίδρασης ή συστηματικού συνεντοπισμού μεταξύ CNE και TE. Μόνο ένα μικρό ποσοστό από TE έχει αναφερθεί πως έχουν εξελιχθεί σε CNE, ωστόσο είναι πολύ λίγα ώστε να επηρεάσουν και να επαναδιαμορφώσουν την όλη κατανομή των CNE (Xie et al. 2006).

Κατανομές μεγεθών αποστάσεων

Ας υποθέσουμε ότι έχουμε μια μεγάλη συλλογή n αντικειμένων (στην περίπτωσή μας οι αποστάσεις, στο εξής spacers, μεταξύ των CNE) και ότι καθένα χαρακτηρίζεται από το μήκος του S. Σε τυπικές, τυχαίες τέτοιες συλλογές (όπως διαδοχές κορώνας σε ένα πείραμα ρίψης νομίσματος) μπορούμε να προσεγγίσουμε την κατανομή των μεγεθών με μια εκθετική κατανομή. Έστω ότι p(S) είναι η πιθανότητα ένας spacer να έχει μήκος μεταξύ S-s/2 και S+s/2 (οπου s είναι το μέγεθος του πλάτους του διαστήματος) και N*(S) ο αριθμός των spacers:

$$\mathbf{N^*(S) = np(S) \propto e^{-\alpha S} \quad \alpha > 0} \qquad (1)$$

Όταν εμφανίζεται ομαδοποίηση ανεξάρτητη κλίμακας (scale – free clustering), συσχετίσεις μακράς εμβέλειας εκτείνονται σε διάφορες κλίμακες μηκών (ιδανικά, για όλο το

εξεταζόμενο μήκος της γονιδιωματικής αλληλουχίας στην περίπτωσή μας) και οι κατανομές μεγέθους των spacers ακολουθούν νόμο δύναμης, ο οποίος αντιστοιχεί σε ένα γραμμικό γράφημα σε διπλή λογαριθμική κλίμακα:

$$\mathbf{N^*(S) = np(S) \propto S^{-\zeta} = S^{-1-\mu} \quad \mu > 0} \qquad (2)$$

Σε αυτή τη διατριβή χρησιμοποιούμε την αθροιστική κατανομή μεγέθους και πιο συγκεκριμένα τη συμπληρωματική αθροιστική κατανομή μεγέθους (Clauset et al. 2009) που ορίζεται ως εξής:

$$\mathbf{P(S) = \int_S^{\infty} p(r)dr} \qquad (3)$$

όπου p(r) είναι η κατανομή μεγέθους των αρχικών spacer (αποστάσεων). Η αθροιστική κατανομή έχει γενικά καλύτερες στατιστικές ιδιότητες, καθώς σχηματίζει ομαλότερες ουρές, που επηρεάζονται λιγότερο από στατιστικές διακυμάνσεις. Επιπλέον, εξ' ορισμού, είναι ανεξάρτητη από την επιλογή του διαστήματος: σε μια αθροιστική καμπύλη, η τιμή του P(S) για μήκος S δε σχετίζεται με το υποσύνολο των spacer, των οποίων το μήκος πέφτει στο ίδιο διάστημα (bin), όπως στην αρχική κατανομή, αλλά αντιστοιχεί στον αριθμό των spacer που είναι μεγαλύτεροι από S. Για άρθρα ανασκόπησης πάνω στις κατανομές μεγεθών που ακολουθούν νόμους δύναμης και τις ιδιότητές τους, δείτε (Clauset et al. 2009; Adamic & Huberman 2002; Li 2002; Stumpf & Porter 2012; Newman 2005).

Η αθροιστική μορφή της κατανομής μεγέθους τύπου νόμου δύναμης είναι επίσης ένας νόμος δύναμης, που χαρακτηρίζεται από έναν εκθέτη (κλίση) ίσο με αυτόν της αρχικής κατανομής αυξημένο κατά 1: εάν $\mathbf{p(r) \propto r^{-1-\mu}}$, τότε:

$$\mathbf{N(S) = nP(S) \propto \int_S^{\infty} (r^{-1-\mu})dr \propto S^{-\mu}} \qquad (4)$$

όπου N(S) είναι ο αριθμός των spacer που είναι μεγαλύτεροι ή ίσοι με S. Όλα τα διαγράμματα κατανομών που παρουσιάζονται εδώ αποτυπώνουν συμπληρωματικές

αθροιστικές κατανομές μεγεθών αποστάσεων (spacer) μεταξύ διαδοχικών CNE. Οι λογάριθμοι των μηκών αυτών των spacer (S) απεικονίζονται στον οριζόντιο άξονα, ενώ οι λογάριθμοι του αριθμού N(S) όλων των spacer μεγαλύτερων ή ίσων με S απεικονίζονται στον κατακόρυφο άξονα.

Η κλίση για έναν τυπικό νόμο δύναμης δεν υπερβαίνει την τιμή του μ = 2, ενώ το μ < 2 είναι μια συνθήκη που οδηγεί σε μια μη συγκλίνουσα τυπική απόκλιση. Στις γραμμικότητες τύπου νόμου δύναμης, που αναφέρονται πιο κάτω, η τιμή του μ είναι πάντα μικρότερη από 2. Οι κατανομές τύπου νόμου δύναμης που παρατηρούνται στη φύση έχουν πάντα ένα ανώτατο και ένα κατώτατο κατώφλι (inner and outer cutoff), το οποίο καθορίζει τη γραμμική περιοχή σε διπλή λογαριθμική κλίμακα, όπου εκτείνεται η αυτο-ομοιότητα και μορφοκλασματικότητα (fractality). Η έκταση της γραμμικής περιοχής (E) αντικατοπτρίζει τις τάξεις μεγέθους, όπου εκτείνεται η μορφοκλασματικότητα. Η γραμμικότητα καθορίζεται με γραμμική παλινδρόμηση και η σχετιζόμενη τιμή του r^2 είναι σε όλες τις περιπτώσεις μεγαλύτερη από 0.97 και σε ποσοστό, μεγαλύτερο από 90% των περιπτώσεων, υψηλότερη από 0.98.

Όλες οι εικόνες που παρουσιάζονται πιο κάτω περιλαμβάνουν, εκτός από τις κατανομές μεγεθών των αποστάσεων των CNE στα διάφορα γονιδιώματα, ένα σύνολο δέκα αναπληρωματικών κατανομών μεγεθών στοιχείων, που έχουμε προσομοιώσει (συνεχόμενες γραμμές), έτσι ώστε τα στοιχεία, που αναπαριστούν τα CNE, να τοποθετούνται τυχαία σε μια αλληλουχία. Ο αριθμός των τυχαία τοποθετημένων στοιχείων και το μήκος της αλληλουχίας που προσομοιώνονται είναι ίσα με το πλήθος των CNE και το μέγεθος του εξεταζόμενου χρωμοσώματος αντίστοιχα, κάθε φορά. Η ενσωμάτωση αυτών των τυχαίων (αναπληρωματικών) συνόλων δεδομένων στις εικόνες μάς επιτρέπει μια οπτικοποίηση της διαφοράς μεταξύ των παρατηρούμενων γονιδιωματικών κατανομών και εκείνων που αναμένονται στα πλαίσια της τυχαιότητας. Σημειώνουμε επίσης ότι, όπου εφαρμόζεται η μεθοδολογία του μασκαρίσματος των γονιδίων, κατασκευάζονται αντίστοιχα αναπληρωματικά σύνολα δεδομένων, που αποκλείουν την τυχαία τοποθέτηση στοιχείων, στο χώρο εκείνο, όπου εφαρμόζεται η μάσκα.

Προσομοιώσεις χρησιμοποιώντας το μοντέλο γονιδιωματικών διπλασιασμών και απαλοιφών CNE

Οι παρατηρούμενες γονιδιωματικές κατανομές αναπαράγονται μέσω προσομοιώσεων, χρησιμοποιώντας μια ευρεία σειρά παραμέτρων. Στην τελευταία εικόνα (*Εικόνα 5*) αυτού του κεφαλαίου δείχνουμε χαρακτηριστικές περιπτώσεις. Αρχικά, 1000 στοιχεία (που

αντιπροσωπεύουν CNE) εισάγονται τυχαία σε μια αλληλουχία μήκους 2 Mb. Στο τμήμα (α) της τελευταίας εικόνας φαίνονται στιγμιότυπα του αναδυόμενου προτύπου νόμου δύναμης, όπως αυτό εξελίσσεται στο χρόνο. Κάθε 50 γεγονότα τμηματικού διπλασιασμού υπολογίζονται συμπληρωματικές αθροιστικές κατανομές μεγεθών αποστάσεων (spacer) μεταξύ διαδοχικών CNE. Κάθε γεγονός τμηματικού διπλασιασμού (Segmental Duplication, SD) περιλαμβάνει μια περιοχή με μήκος που λαμβάνεται από μια ομοιόμορφη κατανομή, με μέγιστο το 5% του εκάστοτε μήκους κάθε φορά της αλληλουχίας που προσομοιώνεται. Σε όλες αυτές τις προσομοιώσεις, μετά από κάθε γεγονός SD, ένα πλήθος CNE ίσο με 90% του αριθμού των διπλασιασμένων CNE απαλοίφεται (αυτό υποδηλώνεται ως fr = 0.9). Στο τμήμα (β) της τελευταίας εικόνας παρουσιάζονται τρεις καμπύλες κατανομών, που έχουν παραχθεί μετά από αριθμητικές προσομοιώσεις και όπου το κλάσμα fr παίρνει τις τιμές 0.8, 0.9 και 1. Στο τμήμα (γ) της ίδιας εικόνας παρουσιάζονται ξανά τρεις καμπύλες κατανομών. Σε αυτές τις προσομοιώσεις, το κλάσμα fr παραμένει σταθερό και ίσο με 0.9, ενώ σε δύο εξ' αυτών επιτρέπονται περαιτέρω απαλοιφές, μία και δύο μετά από κάθε γεγονός τμηματικού διπλασιασμού (SD) αντίστοιχα.

Αποτελέσματα

Εμφάνιση κατανομών τύπου νόμου δύναμης στις αποστάσεις μεταξύ των CNE

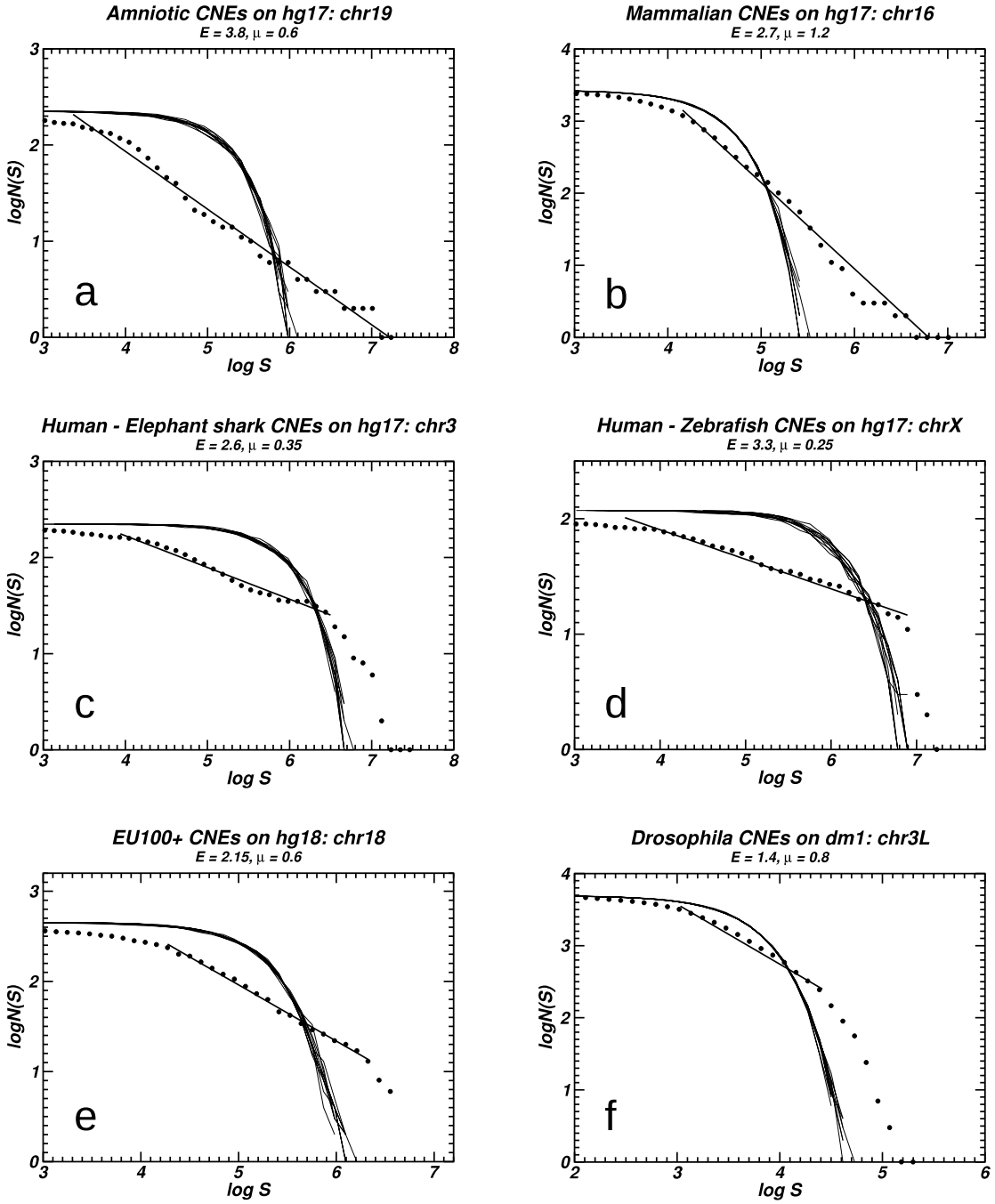
Το κύριο εύρημα της μελέτης που περιγράφουμε σε αυτό το κεφάλαιο είναι η εκτεταμένη εμφάνιση κατανομών τύπου νόμου δύναμης στις αποστάσεις μεταξύ διαδοχικών CNE. Στην ανάλυσή μας περιλαμβάνουμε σύνολα δεδομένων CNE από διαφορετικές ταξινομικές ομάδες και συγκρίνουμε πληθυσμούς CNE που φαίνεται να έχουν αποκτήσει κάποια λειτουργία σε διαφορετικά εξελικτικά στάδια. Τα CNE, που μελετούμε, είναι επίσης χαρτογραφημένα σε διαφορετικά γονιδιώματα (άνθρωπος, *D. melanogaster*, *C. elegans*, *D. rerio*).

Στην *Εικόνα 1* παρουσιάζουμε τις κατανομές μεγεθών των αποστάσεων μεταξύ διαδοχικών CNE σε διαγράμματα διπλής λογαριθμικής κλίμακας. Αναφέρουμε, επίσης, τη γραμμική περιοχή E της εκάστοτε κατανομής και την κλίση μ. Στον *Πίνακα 2* συνοψίζουμε τα αποτελέσματα για κάθε οργανισμό και αναφέρουμε τη μέση τιμή για τη γραμμική περιοχή E σε διπλή λογαριθμική κλίμακα για όλα τα χρωμοσώματα (avg E) και για τα πέντε χρωμοσώματα με τη μεγαλύτερη παρατηρούμενη έκταση E (avg E-5). Η έκταση E αντικατοπτρίζει τις τάξεις μεγέθους, όπου εκτείνεται η κατανομή τύπου νόμου δύναμης. Σε αυτό το κεφάλαιο και, πιο συγκεκριμένα στα αποτελέσματα που βλέπετε, χρησιμοποιούμε

την ποσότητα E, ώστε να μετρήσουμε την ύπαρξη μιας αυτο-ομοιότητας στη γεωμετρία των χρωμοσωμάτων και για να εκτιμήσουμε τη συμφωνία των παρατηρούμενων γονιδιωματικών κατανομών με το εξελικτικό μοντέλο που προτείνουμε (δείτε «Συζήτηση»).

Τα πρότυπα τύπου νόμου δύναμης δεν ανιχνεύονται μόνο σε στοιχίσεις μεταξύ στενά σχετιζόμενων γονιδιωμάτων, αλλά είναι διαδεδομένα παντού. Τα στοιχεία, που ταυτοποιήθηκαν από ολικές στοιχίσεις γονιδιωμάτων θηλαστικών και αμνιωτικών, ήταν από τα πρώτα συντηρημένα μη κωδικοποιούντα στοιχεία, που απαντώνται σε αριθμούς ικανούς ώστε να μας επιτρέψουν στατιστική ανάλυση της χρωμοσωμικής κατανομής τους. Το πλήρες σύνολο περιλαμβάνει 16575 αμνιωτικά CNE και 82335 CNE θηλαστικών (Kim & Pritchard 2007). Παρατηρούμε πρότυπα τύπου νόμου δύναμης επίσης σε συλλογές CNE που προέρχονται από στοιχίσεις που περιλαμβάνουν θηλαστικά μαζί με τελεόστεους, των οποίων ο κοινός πρόγονος αποσχίστηκε ±450 MYA και χονδριχθύες (elephant shark) που απέκλιναν ±530 MYA (Kumar & Hedges 1998). Η κατανομή μεγέθους των αποστάσεων μεταξύ των CNE στα γονιδιώματα ασπόνδυλων, όπως η *D. melanogaster* και ο *C. elegans*, ακολουθεί επίσης παρόμοιο πρότυπο. Είναι γνωστό από τη βιβλιογραφία ότι τα CNE σπονδυλωτών και ασπόνδυλων μοιράζονται κάποια κοινά χαρακτηριστικά αλληλουχίας αλλά δεν είναι πανομοιότυπα (Vavouri et al. 2007), συνεπώς φαίνεται, σε συνδυασμό με τα αποτελέσματά μας, ότι οι κατανομές τους είναι πιθανόν να διαμορφώνονται από κοινούς μηχανισμούς.

Εικόνα 1: Ενδεικτικά παραδείγματα κατανομών αποστάσεων μεταξύ διαδοχικών CNE, ανά χρωμόσωμα, στο ανθρώπινο γονιδίωμα (a-e) και στο γονιδίωμα της *D. melanogaster* (f). Παρατηρούμε εκτενή γραμμικότητα σε διπλή λογαριθμική κλίμακα. Περισσότερες πληροφορίες στο κείμενο.



Πίνακας 2: Τάση σχηματισμού νόμων δύναμης για διάφορα σύνολα δεδομένων CNE, όπως ποσοτικοποιείται από την έκταση E. Για περισσότερες πληροφορίες, δείτε κείμενο.

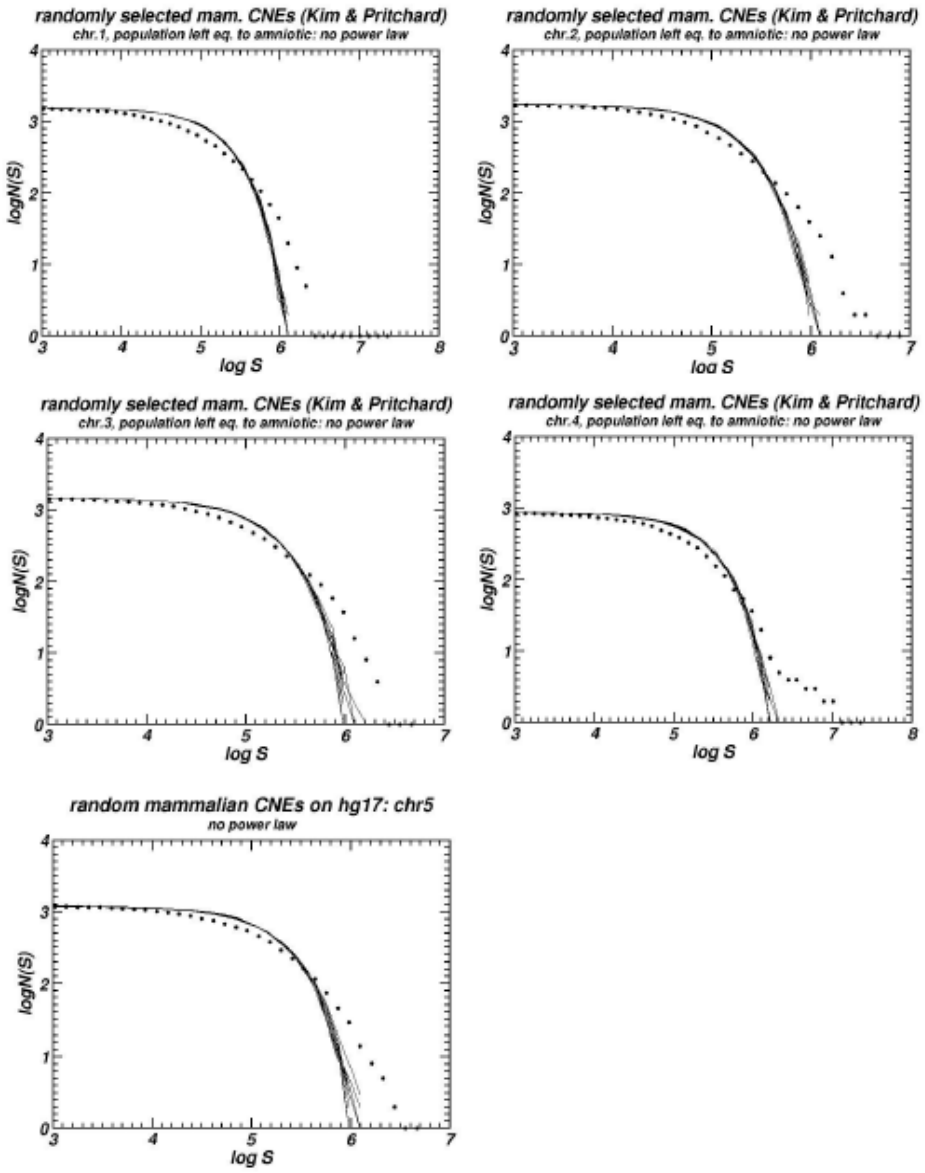
Dataset	Class of CNEs	Unmasked		Masked		Reference genome
		Average Extent (avg E)	Average Extent of five 'best' chr. (avg E-5)	Average Extent (avg E)	Average Extent of five 'best' chr. (avg E-5)	
i	EU100+	2	2.38	2.48	2.98	hg18
ia	EU-FR	1.97	2.46			hg18
ib	FR	2.2	2.72			hg18
ic	XT-FR	2.32	2.32			hg18
id	GG-XT-FR	2.14	2.14			hg18
ie	EU-GG-XT-FR	1.96	1.96			hg18
iia	Mammalian	1.49	1.9	1.59	2.04	hg17
iib	Amniotic	2.2	2.86	2.25	2.91	hg17
iii	Human/Chicken	2.35	2.63			hg19
iv	Human/Fugu	2.78	3.26			hg17
v	Human/Zebrafish	2.46	2.98	2.31	2.31	hg17
vi	Human/El. shark	2.36	2.69	2.42	2.68	hg17
vii	*D. rerio* PCNEs	2.43	3.01			#
viii	Insect CNEs	1.23	1.23	1.42	1.42	dm1
ix	Worm CNEs	1.7	1.7			WS140
x	Human/Rodents	2.15	2.43			hg17

Propensity for the formation of power-law-like size distributions of the inter-CNE distances as quantified by the extent (E) of linearity in log-log scale. Average values of E for all chromosomes (avg E) and average values of E for 5 chromosomes with the largest E (avg E-5) in each genome are presented. Gene-masked genomes are also included when available (for details see in the text).
#: genome-build Ensembl 42 (zebrafish).
doi:10.1371/journal.pone.0095437.t001

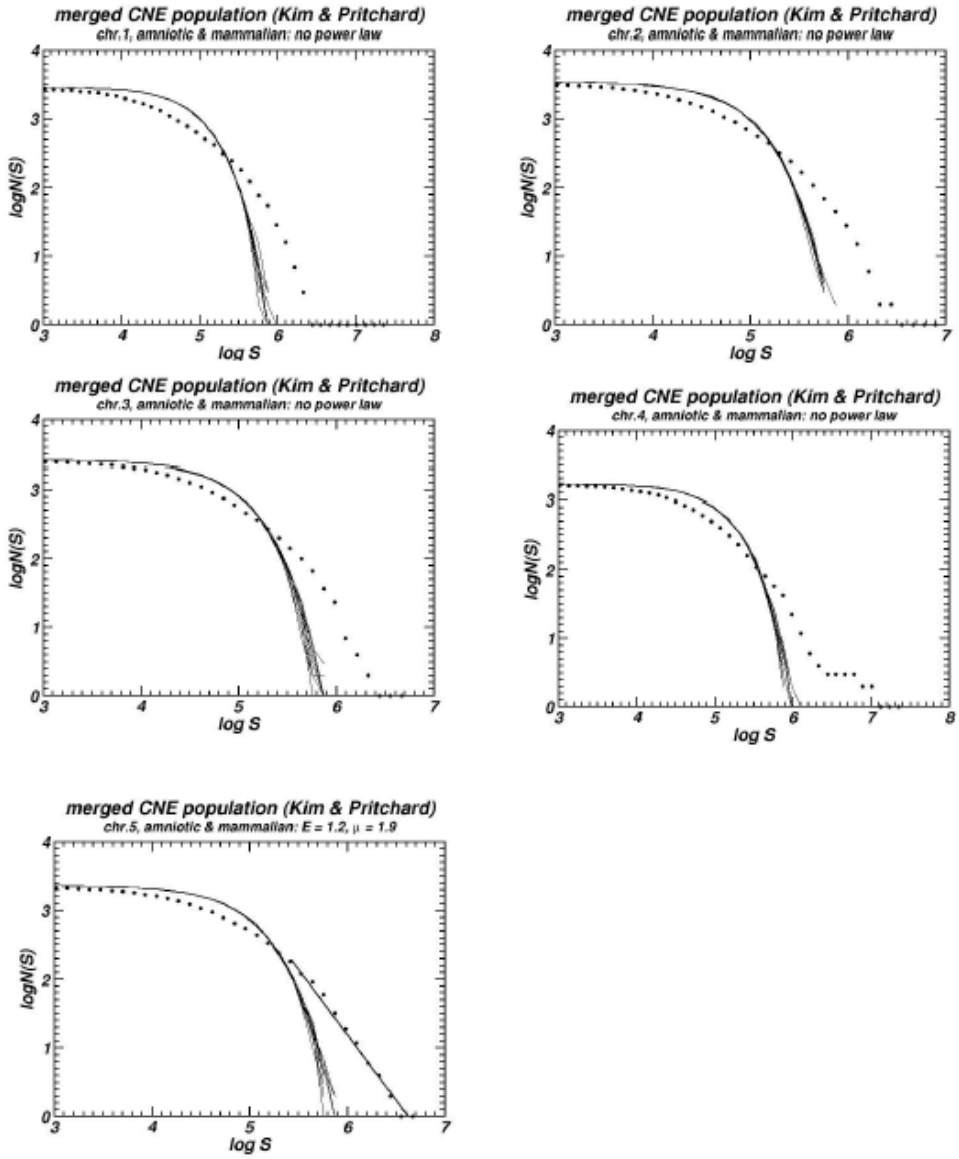
Οι κατανομές τύπου νόμου δύναμης χαρακτηρίζονται από την υπερεκπροσώπηση μεγάλων spacer. Γι' αυτό το λόγο, θα περίμενε κανείς, ότι εκτεταμένες γραμμικότητες ευνοούνται σε μικρότερα σύνολα δεδομένων. Αποδεικνύουμε ότι αυτό δεν ισχύει. Βασισμένοι στα δεδομένα μας, συμπεραίνουμε ότι, οι κατανομές μεγέθους των αποστάσεων μεταξύ των CNE είναι εγγενείς του συστήματος που μελετούμε και δεν εξαρτώνται από τους πληθυσμούς των στοιχείων. Αυτό φαίνεται από το γεγονός ότι όταν μειώνουμε τους αριθμούς των CNE θηλαστικών (± 80000), μέσω τυχαίας δειγματοληψίας, σε παρόμοιους αριθμούς με αυτούς των αμνιωτικών (± 16000), τα οποία χαρακτηρίζονται από πιο εκτεταμένες γραμμικότητες, η έκταση της γραμμικότητας δεν αυξάνεται. Απεναντίας, η γραμμικότητα, σε διπλή λογαριθμική κλίμακα, εξαφανίζεται ως συνέπεια της αλλαγής του υπό μελέτη γονιδιωματικού τοπίου (δείτε *Εικόνα 2, Τμήμα Α*). Παρομοίως, η γραμμικότητα εξαφανίζεται όταν μελετούμε πληθυσμούς αμνιωτικών CNE και CNE θηλαστικών που έχουν συγχωνευθεί, δηλαδή ανακατευτεί μεταξύ τους χωρίς φυσικά να αλλάξουμε τη χωροταξική δομή και το μέγεθος των στοιχείων (δείτε *Εικόνα 2, Τμήμα Β*). Επομένως, ισχυριζόμαστε οτι τα αποτελέσματά μας είναι χαρακτηριστικά και αντικατοπτρίζουν τη δυναμική κάθε κλάσης CNE που μελετάται και δεν εξαρτώνται από τους πληθυσμούς των στοιχείων, δεδομένου βέβαια ότι το πλήθος των CNE είναι επαρκές για στατιστική ανάλυση.

Εικόνα 2: Ενδεικτικά παραδείγματα κατανομών αποστάσεων μεταξύ διαδοχικών CNE για τα πέντε πρώτα χρωμοσώματα στο ανθρώπινο γονιδίωμα. Α) τυχαία επιλεγμένων (randomly selected) CNE θηλαστικών ίδιου αριθμού με τα CNE αμνιωτικών και Β) CNE αμνιωτικών και θηλαστικών που έχουν ανακατευτεί (merged), απαρτίζοντας έναν ενιαίο πληθυσμό.

A.



B.



Οι παρατηρούμενες κατανομές των αποστάσεων των CNE δεν αποτελούν συνέπεια του εντοπισμού των γονιδίων στο ίδιο χρωμόσωμα

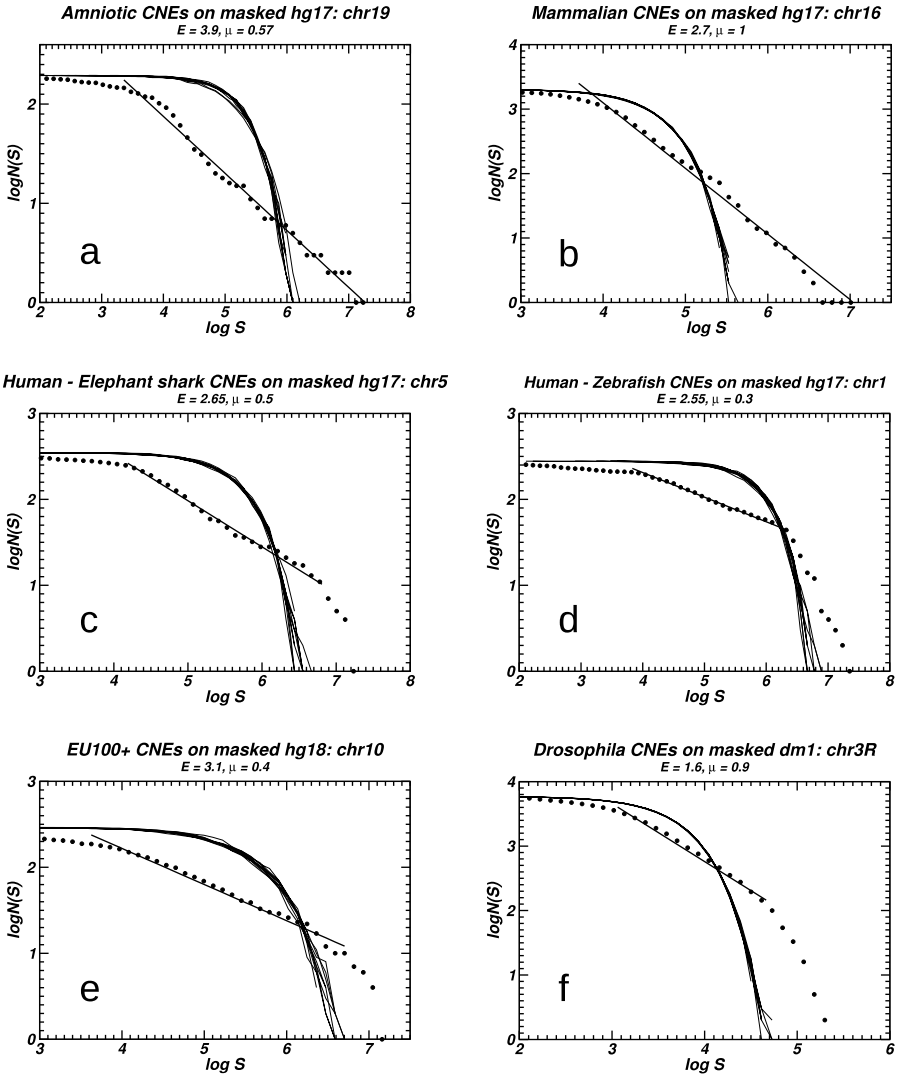
Κατανομές τύπου νόμου δύναμης παρατηρούνται και στη χρωμοσωμική κατανομή των κωδικοποιουσών περιοχών (Sellis & Almirantis 2009). Όπως είναι γνωστό από τη βιβλιογραφία και αναφέραμε πιο πάνω, τα CNE σχετίζονται χωρικά με γονίδια που κωδικοποιούν μεταγραφικούς παράγοντες και ρυθμιστές της ανάπτυξης (επίσης γνωστά ως *trans – dev* γονίδια) (Sandelin et al. 2004; Sanges et al. 2013; Woolfe & Elgar 2008). Με σκοπό να αποκλείσουμε την πιθανότητα οι παρατηρούμενες χρωμοσωμικές κατανομές να είναι αποτέλεσμα των προτύπων τύπου νόμου δύναμης, που ακολουθούνται από τις κατανομές των αποστάσεων μεταξύ γονιδίων, εφαρμόζουμε μασκάρισμα όλων των γονιδίων που κωδικοποιούν πρωτεΐνες καθώς και εκτεταμένων περιοχών εκατέρωθεν κάθε

γονιδίου, όπου συνήθως εντοπίζονται οι περισσότερες ρυθμιστικές αλληλουχίες. Με τον όρο μασκάρισμα εννοούμε τον αποκλεισμό όλων των στοιχείων που βρίσκονται μέσα σε γονίδια και στις περιβάλλουσες περιοχές και όχι αφαίρεση των ίδιων των γονιδίων, καθώς το τελευταίο θα άλλαζε την κατανομή μεγέθους των αποστάσεων μεταξύ των CNE. Διαπιστώνουμε ότι η γραμμικότητα, στα διαγράμματα διπλής λογαριθμικής κλίμακας, όχι μόνο διατηρείται, αλλά στις περισσότερες περιπτώσεις βελτιώνεται, όπως φαίνεται από την αύξηση της έκτασης της γραμμικής περιοχής. Αυτό δείχνει ότι, ακόμα και αν αποκλείσουμε από τη μελέτη μας τα CNE, τα οποία εντοπίζονται κοντά σε γονίδια (και επομένως ακολουθούν την κατανομή τους), η χρωμοσωμική κατανομή των CNE, που απομένουν, παρουσιάζει και πάλι πρότυπα τύπου νόμου δύναμης. Ο βασικός μας στόχος, εδώ, είναι να δείξουμε ότι η δυναμική, που δημιουργεί το πρότυπο τύπου νόμου δύναμης, δεν είναι μια απλή απόρροια της κατανομής των γονιδίων, παρόλο που οι δύο κατανομές είναι αναμενόμενο να επηρεάζουν η μία την άλλη, ένα γεγονός το οποίο δεν λαμβάνουμε υπόψη στο απλό μοντέλο μας εδώ. Παραδείγματα τέτοιων διαγραμμάτων δίνονται στην *Εικόνα 3*, ενώ στον *Πίνακα 2* δίνονται, για μια σύγκριση, τα αποτελέσματα, που αφορούν τα χρωμοσώματα, όπου έχουν μασκαριστεί τα γονίδια, ανά οργανισμό.

Διαλέγουμε να εφαρμόσουμε τη μεθοδολογία του μασκαρίσματος στο ανθρώπινο γονιδίωμα και για τα πιο πολυπληθή σύνολα δεδομένων CNE (Kim & Pritchard 2007; Stephen et al. 2008), καθώς και για τα πιο «αρχαία» στοιχεία που είναι συντηρημένα μεταξύ ανθρώπου και zebrafish (Shin et al. 2005) ή ανθρώπου και elephant shark (Venkatesh et al. 2006), δείτε σύνολα δεδομένων iia,b, i, v και vi αντίστοιχα. Σε όλες τις περιπτώσεις, που μελετούμε, παρατηρούμε κατανομές τύπου νόμου δύναμης μεταξύ των αποστάσεων των CNE που εκτείνονται σε αρκετές τάξεις μεγέθους (δείτε *Πίνακα 2*). Μια παρόμοια μεθοδολογία εφαρμόζεται και στο γονιδίωμα της *D. melanogaster*, δείτε σύνολο δεδομένων viii. Ο σκοπός της μεθοδολογίας μασκαρίσματος, που εφαρμόζουμε, δεν είναι να αποκλείσει τα CNE, που δρουν ως απομακρυσμένα ρυθμιστικά στοιχεία, μέσω π.χ. του μηχανισμού της αναδίπλωσης της χρωματίνης (Viturawong et al. 2013). Κάτι τέτοιο θα ήταν πολύ δύσκολο, διότι οι αλληλεπιδράσεις των CNE με τα γονίδια στόχους τους παραμένουν ασαφείς, ενώ όταν μιλάμε για λειτουργικές αλληλεπιδράσεις τα πράγματα γίνονται ακόμα πιο δύσκολα, καθώς δεν υπάρχουν, όπως αναφέραμε και στην αρχή της διατριβής, πειραματικές διατάξεις που να μας επιτρέπουν τέτοια μελέτη. Επομένως, επιλέγουμε να παρουσιάσουμε μια μετριοπαθή (5 kb ανοδικά του 5’ άκρου και 2 kb καθοδικά του 3’ άκρου) μέθοδο μασκαρίσματος, έτσι ώστε να αποκλείσουμε μόνο στοιχεία, τα οποία δρουν ως ρυθμιστές της μεταγραφής, λόγω εγγύτητας σε υποκινητές και

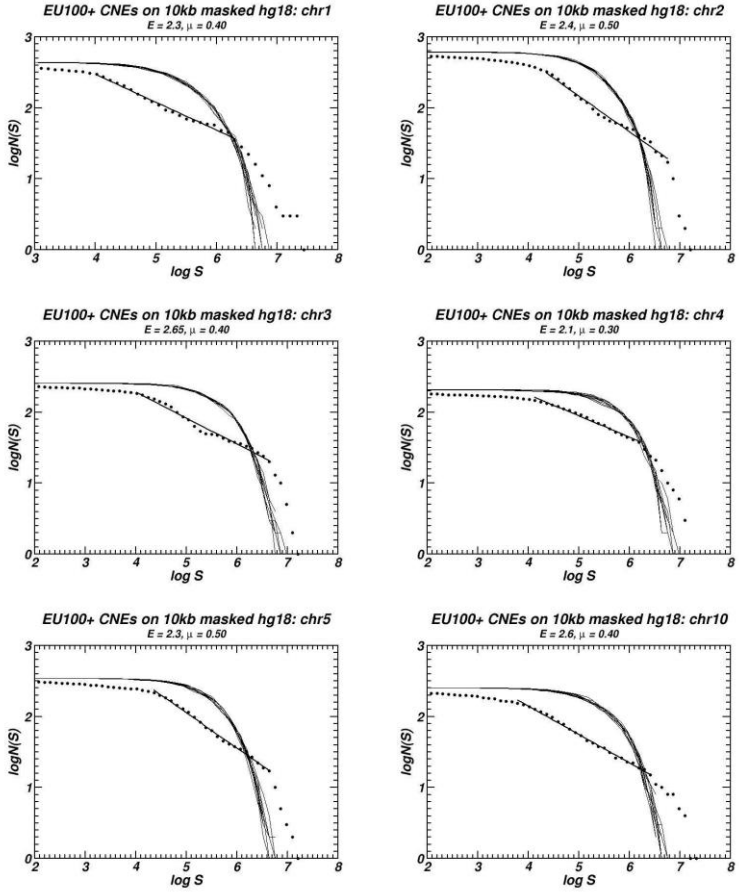
που ενδεχομένως σχετίζονται χωρικά με γειτονικά γονίδια. Για να επιβεβαιώσουμε τον ισχυρισμό μας, ωστόσο, εφαρμόζουμε μεθοδολογία μασκαρίσματος, θεωρώντας εκτεταμένες παρεμβάλλουσες αλληλουχίες (10 kb και 100 kb, σε ξεχωριστά πειράματα) στο σύνολο δεδομένων EU100+ (ευθηρίων, δείτε και *Πίνακα 2*) για έξι χρωμοσώματα (τα πέντε μεγαλύτερα και το χρωμόσωμα 10, το οποίο είναι εμπλουτισμένο σε CNE). Παρατηρούμε ότι η γραμμικότητα σε διπλή λογαριθμική κλίμακα είναι ακόμα εμφανής στις κατανομές των αποστάσεων μεταξύ διαδοχικών CNE και επιπλέον εκτείνεται σε αρκετές τάξεις μεγέθους. Ακόμη και στην περίπτωση του μασκαρίσματος στα 100 kb, οι γραμμικότητες παραμένουν, παρά το γεγονός ότι απομένουν λίγα CNE μετά το μασκάρισμα σε τόσο μεγάλη κλίμακα (δείτε *Εικόνα 4*).

Εικόνα 3: Ενδεικτικά παραδείγματα κατανομών αποστάσεων μεταξύ διαδοχικών CNE, ανά χρωμόσωμα, στο ανθρώπινο γονιδίωμα (a-e) και στο γονιδίωμα της *D. melanogaster* (f), μετά τον αποκλεισμό στοιχείων CNE, που ανιχνεύονται σε περιοχές γονιδίων (εσωνικές ή εξωγονιδιακές), καθώς και σε περιβάλλουσες των γονιδίων περιοχές. Η έκταση των παρατηρούμενων νόμων δύναμης όχι μόνο διατηρείται, αλλά αυξάνει, όπως προκύπτει από την αύξηση γραμμικότητας σε διπλή λογαριθμική κλίμακα, πρβλ. *Εικόνα 1*. Για την κατασκευή των δέκα τυχαίων κατανομών μεγεθών στοιχείων, σε αυτήν την περίπτωση, αποκλείεται και η θεωρηθείσα «περιοχή αποκλεισμού» κάθε φορά, κατά την τυχαία κατανομή στοιχείων σε τεχνητά χρωμοσώματα, μεγέθους ίδιου με το πραγματικό.

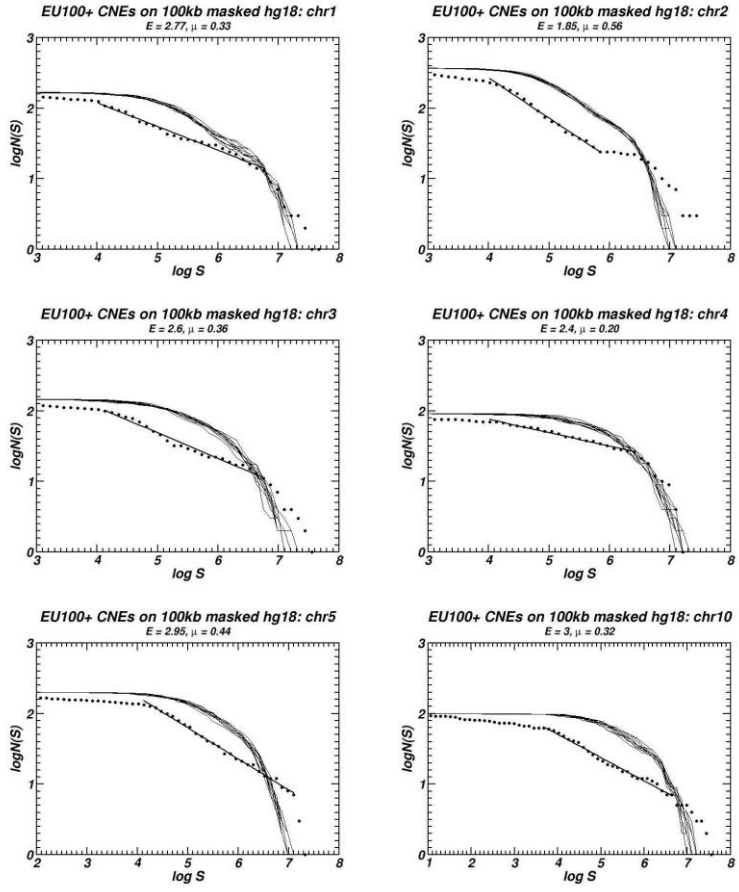


Εικόνα 4: Ενδεικτικά παραδείγματα κατανομών αποστάσεων μεταξύ διαδοχικών CNE, για τα έξι χρωμοσώματα στο ανθρώπινο γονιδίωμα (πέντε μεγαλύτερα και χρωμόσωμα 10), στο σύνολο δεδομένων EU100+, μετά τον αποκλεισμό στοιχείων CNE, που ανιχνεύονται σε περιοχές γονιδίων (εσωνικές ή εξωγονιδιακές), καθώς και σε περιβάλλουσες των γονιδίων περιοχές. Α) αποκλεισμός στοιχείων CNE που βρίσκονται μέσα σε γονίδια καθώς και 10 kb εκατέρωθεν από κάθε γονίδιο και Β) αποκλεισμός στοιχείων CNE που βρίσκονται μέσα σε γονίδια καθώς και 100 kb εκατέρωθεν από κάθε γονίδιο.

Α.


EU100+ CNEs on 10kb masked hg18: chr1
E = 2.3, μ = 0.40
EU100+ CNEs on 10kb masked hg18: chr2
E = 2.4, μ = 0.50
EU100+ CNEs on 10kb masked hg18: chr3
E = 2.65, μ = 0.40
EU100+ CNEs on 10kb masked hg18: chr4
E = 2.1, μ = 0.30
EU100+ CNEs on 10kb masked hg18: chr5
E = 2.3, μ = 0.50
EU100+ CNEs on 10kb masked hg18: chr10
E = 2.6, μ = 0.40
logN(S)
log S

B.


EU100+ CNEs on 100kb masked hg18: chr1
E = 2.77, μ = 0.33
EU100+ CNEs on 100kb masked hg18: chr2
E = 1.85, μ = 0.56
EU100+ CNEs on 100kb masked hg18: chr3
E = 2.6, μ = 0.36
EU100+ CNEs on 100kb masked hg18: chr4
E = 2.4, μ = 0.20
EU100+ CNEs on 100kb masked hg18: chr5
E = 2.95, μ = 0.44
EU100+ CNEs on 100kb masked hg18: chr10
E = 3, μ = 0.32
logN(S)
log S

Συζήτηση

Ένα προτεινόμενο εξελικτικό μοντέλο που αναπαράγει τις παρατηρούμενες κατανομές τύπου νόμου δύναμης και βασίζεται σε διπλασιασμούς (τμηματικούς ή ολόκληρου του γονιδιώματος) και σε απώλεια CNE

Γεγονότα τμηματικού διπλασιασμού συνέβησαν διαρκώς κατά την εξελικτική ιστορία σχεδόν όλων των ευκαρυωτικών οργανισμών (De Grassi et al. 2008; Kehrer-Sawatzki & Cooper 2008; Kirsch et al. 2008; McLysaght et al. 2000). Περίπου ένα 10% του μη επαναλαμβανόμενου ανθρώπινου γονιδιώματος αποτελείται από αναγνωρίσιμα (δηλαδή σχετικά πρόσφατα) γεγονότα τμηματικών διπλασιασμών (Bailey et al. 2002). Υπολογίζεται ότι ένα 50% όλων των γονιδίων, σε ένα γονιδίωμα, αναμένεται πως διπλασιάζεται, δίνοντας «απογόνους» μία φορά τουλάχιστον, σε κλίμακες χρόνου από 35 έως 350 εκατομμύρια χρόνια (Lynch 2000). Επιπλέον, οι περισσότεροι οργανισμοί ταξινομικών ομάδων έχουν υποστεί παλαιοπολυπλοειδισμό κατά τη διάρκεια της εξέλιξής τους, δηλαδή διπλασιασμό ολόκληρου του γονιδιώματός τους και μετέπειτα ταυτόχρονη μείωση σε διπλοειδισμό (Gibson & Spring 2000; Sémon & Wolfe 2007). Τμηματικοί διπλασιασμοί και παλαιοπολυπλοειδισμοί παράγουν αντίγραφα ορισμένων ή όλων των γονιδίων ενός οργανισμού, αλλά επίσης και άλλων λειτουργικών ή πιθανά λειτουργικών στοιχείων του γονιδιώματος, όπως τα CNE. Όπως συμφωνούν όλοι οι ερευνητές (Lynch 2000; Sémon & Wolfe 2007; Kasahara 2007), η μοίρα των περισσότερων διπλασιασμένων γονιδίων είναι ότι ένα αντίγραφο αποσιωπάται, χάνοντας την ικανότητα να μεταγραφεί και μετά σταδιακά αποσυντίθεται με τη συσσώρευση τυχαίων μεταλλάξεων, ενώ εκτίθεται επίσης στην πιθανότητα εκτομής, λόγω των απαλοιφών που καθοδηγούνται από τον ανασυνδυασμό. Η μοίρα των διπλασιασμένων CNE αναμένεται να είναι παρόμοια, επομένως ένα διπλασιασμένο CNE ενδεχομένως να καθίσταται περιττό και να σταματά να βρίσκεται κατώ από επιλεκτική πίεση, με αποτέλεσμα σταδιακά να συσσωρεύει μεταλλάξεις, να χάνεται και να μην είναι πλέον αναγνωρίσιμο. Η ύπαρξη μιας άλλης πηγής απώλειας CNE υποστηρίζεται από πρόσφατα ευρήματα της συγκριτικής γονιδιωματικής, όπως το γεγονός ότι πολλά από τα CNE, που είναι συντηρημένα μεταξύ elephant shark και ανθρώπου, δεν αναγνωρίζονται πλέον στο γονιδίωμα του fugu (δείτε αμέσως επόμενη ενότητα για περαιτέρω συζήτηση). Σε αυτήν την περίπτωση, όχι μόνο χάνονται τα διπλασιασμένα αντίγραφα των CNE, αλλά επιπλέον τα CNE που είναι παρόντα σε έναν προγονικό οργανισμό μειώνονται αισθητά.

Η περιστασιακή απώλεια λειτουργίας CNE και η μετέπειτα αποσιώπησή τους, οι διπλασιασμοί του γονιδιώματος και οι επαναλαμβανόμενοι τμηματικοί διπλασιασμοί, μαζί με άλλες εκφάνσεις της δυναμικής του γονιδιώματος (π.χ. εισδοχές μεταθετών στοιχείων και άλλων ειδών παρασιτικών αλληλουχιών) μπορούν να συνδυαστούν σε ένα εξελικτικό μοντέλο, του οποίου η ικανότητα να παράγει χρωμοσωμικές κατανομές τύπου νόμου δύναμης μπορεί να ελεγχθεί με τη βοήθεια υπολογιστικών προσομοιώσεων. Ενσωματώνουμε λοιπόν όλες αυτές τις παραμέτρους σε ένα μοντέλο, το οποίο ονομάζουμε «μοντέλο διπλασιασμών του γονιδιώματος – απώλειας CNE». Τα γεγονότα που συμβαίνουν στο γονιδίωμα και περιλαμβάνονται στο μοντέλο μας είναι:

- ***i.*** Τμηματικοί διπλασιασμοί εκτεταμένων περιοχών των χρωμοσωμάτων. Αυτό το βήμα ενδεχομένως περιλαμβάνει, ως περιοριστικό παράγοντα, διπλασιασμούς ολόκληρου του γονιδιώματος, παρόλο που δεν λαμβάνονται υπόψη στα παραδείγματα που δείχνονται εδώ.
- ***ii.*** Τυχαίες απαλοιφές ενός αριθμού CNE, ο οποίος είναι χαμηλότερος ή ίσος με τον αριθμό εκείνων που διπλασιάζονται.
- ***iii.*** Περιστασιακά, επιπλέον απαλοιφές μη διπλασιασμένων CNE.
- ***iv.*** Εισδοχές αλληλουχιών, που αυξάνουν το συνολικό μήκος της χρωμοσωμικής αλληλουχίας. Αυτές μπορεί να είναι μεταθετά στοιχεία, ρετροϊοί, επεκτάσεις μικροδορυφόρων, κ.ά.
- ***v.*** Απαλοιφές τμημάτων αλληλουχίας, που βρίσκονται, συνήθως, υπό ασθενή ή καθόλου επιλεκτική πίεση.

Το προτεινόμενο εξελικτικό μοντέλο αναπαράγει τις κατανομές τύπου νόμου δύναμης των μεγεθών των αποστάσεων μεταξύ των CNE, δείτε *Εικόνα 5*. Αυτή η ιδιότητα αποδεικνύεται αριθμητικά ότι ευσταθεί απέναντι στις ποσοτικές τροποποιήσεις όλων των τύπων των εμπλεκόμενων μοριακών γεγονότων. Μόνο τα γεγονότα τύπου ***i*** και ***ii*** είναι *αναγκαία* για την εμφάνιση του προτύπου νόμου δύναμης. Αυτή η δυναμική είναι πολύ κοντινή στη σύλληψή της με αυτή που έχει περιγραφεί νωρίτερα για την εξήγηση ανάλογων προτύπων κατανομής που ακολουθούνται από τα γονίδια, που κωδικοποιούν πρωτεΐνες (Sellis & Almirantis 2009). Σε ένα εντελώς διαφορετικό πλαίσιο μελέτης άλλων στοιχείων του γονιδιώματος, δηλαδή όταν εξετάζονται μη συντηρημένες αλληλουχίες, όπως επαναμβανόμενες αλληλουχίες ή μικροδορυφόροι, απαιτούνται τα γεγονότα τύπου ***iii*** και ***iv***, αντί για τα ***i*** και ***ii*** (Sellis et al. 2007; Klimopoulos et al. 2012). Τα γεγονότα, τύπου ***iv*** και ***v***, έχει δειχθεί αριθμητικά, σε προσομοιώσεις υπολογιστών, ότι δεν απαιτούνται για την ανάδυση των προτύπων τύπου νόμου δύναμης. Η ενσωμάτωση των γεγονότων του

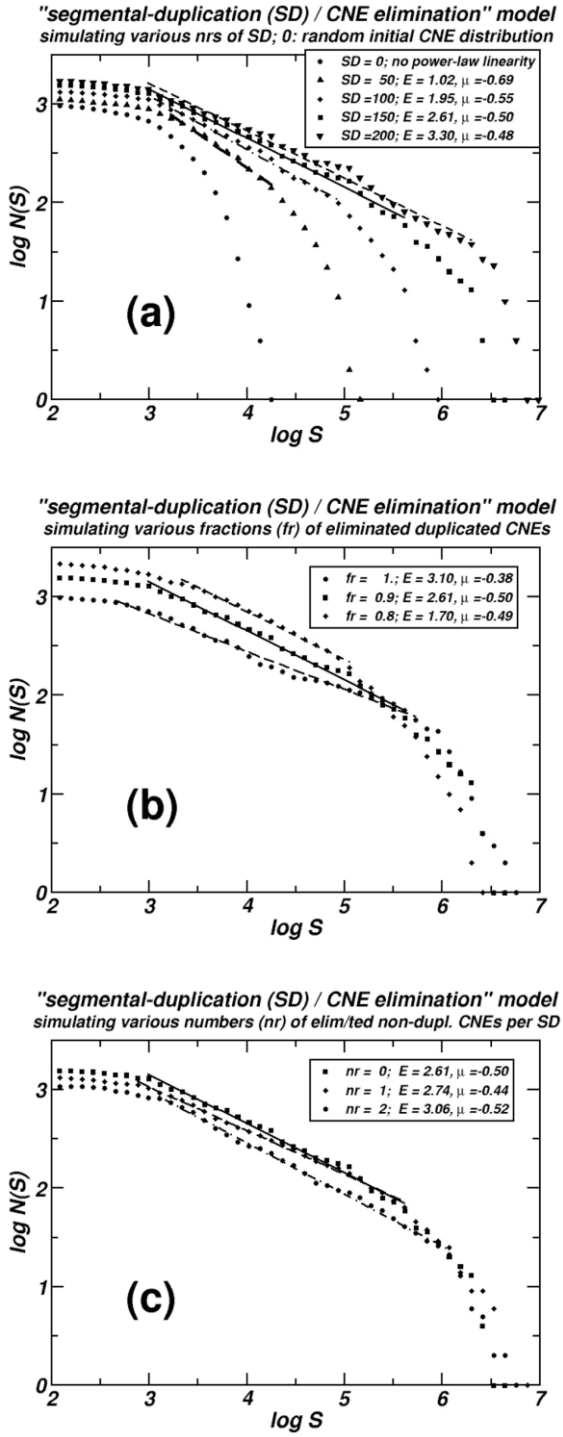
τύπου ***iv*** δοκιμάζει την ακεραιότητα του μοντέλου, καθώς ο πολλαπλασιασμός των επαναλαμβανόμενων στοιχείων αποτελεί, για πολλούς οργανισμούς, σημαντικό μέρος και κινητήρια δύναμη επέκτασης του γονιδιώματός τους. Γεγονότα του τύπου ***v*** αντιπροσωπεύουν, είτε διαγραφή περιοχών αλληλουχίας, συνήθως εξ' αιτίας άνισου επιχιασμού, είτε σταδιακή συρρίκνωση από γεγονότα εισαγωγής / διαγραφής (*indel* events), τα οποία ευνοούν τη μείωση του συνολικού μήκους της αλληλουχίας του χρωμοσώματος. Αυτοί οι τύποι γεγονότων αποδεικνύονται μεγάλης σημασίας στα γονιδιώματα που έχουν καταστεί εξαιρετικά συμπαγή (π.χ. κατά την εξέλιξη της *Drosophila melanogaster* ή στην περίπτωση του *Takifugu rubripes*). Παραδείγματα προσομοιώσεων, χρησιμοποιώντας όλους αυτούς τους τύπους δυναμικής του γονιδιώματος, δίνονται στην πρόσφατη μελέτη των Sellis και συνεργάτες (Sellis & Almirantis 2009).

Τα προαναφερθέντα εξελικτικά σενάρια βασίζονται σε ένα αναλυτικά επιλύσιμο μοντέλο, που προτάθηκε από τους Takayasu και συνεργάτες (Takayasu et al. 1991), για την εμφάνιση των κατανομών μεγεθών τύπου νόμου δύναμης κατά την αύξηση, μέσω συσσωμάτωσης, σφαιριδίων σε φυσικοχημικά συστήματα. Σημειώνουμε ότι το μοντέλο, που παρουσιάζουμε εδώ και το οποίο περιγράφει τη δυναμική στο γονιδίωμα των CNE, δεν είναι αναλυτικά επιλύσιμο και επομένως δεν υπάρχουν σταθεροί εκθέτες (κλίσεις για το γραμμικό τμήμα σε διπλή λογαριθμική κλίμακα). Αυτό επιβεβαιώνεται από όλες τις υπολογιστικές μας προσομοιώσεις και βρίσκεται σε συμφωνία με την ποικιλία τιμών της κλίσης μ, που συναντώνται στη μελέτη των κατανομών των CNE του γονιδιώματος. Επομένως, τα αποτελέσματά μας αποκλίνουν από τον τυπικό / φορμαλιστικό ορισμό του νόμου δύναμης, αφενός επειδή η γραμμικότητα σε διπλή λογαριθμική κλίμακα λαμβάνει, στην περίπτωσή μας, ένα κατώτατο και ένα ανώτατο κατώφλι (στην πραγματικότητα, αυτό είναι ένα χαρακτηριστικό που είναι κοινό σε όλες τις περιπτώσεις νόμων δύναμης, που απαντώνται στη φύση) και αφετέρου κυρίως επειδή απουσιάζει ένας σταθερός / καθολικός εκθέτης (κλίση). Αυτός είναι και ο κύριος λόγος που ονομάζουμε τα πρότυπα που παρατηρούμε «*τύπου* νόμου δύναμης», σε όλη την παρούσα διατριβή. Για μια, εις βάθος, ανασκόπηση των μαθηματικών απαιτήσεων ώστε να έχουμε νόμους δύναμης, δείτε εδώ (Stumpf & Porter 2012). Οι συγκεκριμένοι συγγραφείς προτείνουν ότι αυτές οι απαιτήσεις περιλαμβάνουν έναν στατιστικά ισχυρό νόμο δύναμης (εκτεταμένη γραμμικότητα σε διπλή λογαριθμική κλίμακα, με ενδείξεις σύγκλισης προς έναν καθολικό εκθέτη) και μια ισχυρή θεωρία από πίσω, που να τον υποστηρίζει. Στην περίπτωσή μας δείχνουμε ξεκάθαρα ότι απουσιάζει ένας καθολικός εκθέτης (κλίση) από τις γραμμικότητες σε διπλή λογαριθμική

κλίμακα, που παρατηρούμε στα γονιδιωματικά μας δεδομένα και που είναι συχνά αρκετά εκτεταμένες. Από την άλλη, αυτή η απουσίας καθολικότητας αποτελεί ένα χαρακτηριστικό γνώρισμα, που μοιράζονται τα πρότυπα κατανομής των αποστάσεων μεταξύ των CNE που περιγράφουμε εδώ, καθώς και οι προσομοιώσεις του προτεινόμενου εξελικτικού μοντέλου. Αυτό το χαρακτηριστικό, μαζί με την κοινή εξάρτηση στις εξελικτικές παραμέτρους που μοιράζονται μεταξύ τους οι γονιδιωματικές κατανομές και οι κατανομές του μοντέλου, ενισχύουν την υπόθεση ότι τα γεγονότα εξελικτικής δυναμικής που προτείνουμε (και περιγράφονται από το μοντέλο) ενδεχομένως αποτελούν την αιτία των παρατηρούμενων προτύπων στο γονιδίωμα.

Κάτω από ένα εύρος συνδυασμών παραμέτρων, το μοντέλο μας είναι σε θέση να αναπαράγει τις παροδικές κατανομές τύπου νόμου δύναμης που παρατηρούμε στα γονιδιωματικά δεδομένα. Στις προσομοιώσεις της *Εικόνας 5a & b*, περιλαμβάνονται μόνο γεγονότα του τύπου ***i*** και ***ii***, δηλαδή τμηματικοί διπλασιασμοί (SD) ακολουθούμενοι από απώλειες CNE. Στην *Εικόνα 5a* παρακολουθείται ένα χρωμόσωμα υπό προσομοίωση, λαμβάνοντας διαδοχικά στιγμιότυπα κάθε 50 SD, ξεκινώντας από μια αρχικά (στη στιγμή 0) τυχαία κατανομή στοιχείων, που αντιπροσωπεύουν τα CNE του γονιδιώματος. Βλέπουμε ότι σταδιακά αρχίζει να εμφανίζεται μια γραμμική περιοχή τύπου νόμου δύναμης στις αθροιστικές κατανομές των αποστάσεων μεταξύ διαδοχικών CNE, σε διπλή λογαριθμική κλίμακα, παρόμοια με αυτή που παρατηρείται για τα πραγματικά χρωμοσώματα. Αυτό το γράφημα δείχνει επίσης τη θετική εξάρτηση της παρατηρούμενης έκτασης της γραμμικότητας, σε σχέση με τον αριθμό των SD που έχουν συμβεί. Στην *Εικόνα 5b* απεικονίζεται η θετική εξάρτηση της έκτασης της γραμμικότητας σε σχέση με τον αριθμό των CNE που έχουν απαλοιφθεί (έχουν χαθεί μετά από κάθε γεγονός τμηματικού διπλασιασμού). Εδώ ο αριθμός των CNE, που έχουν απαλοιφθεί, εκφράζεται ως ένα κλάσμα (fraction, fr) εκείνων που έχουν διπλασιαστεί εξ' αιτίας των γεγονότων τμηματικού διπλασιασμού. Τέλος, στην *Εικόνα 5c*, προσομοιώνονται επιπρόσθετες απαλοιφές μη διπλασιασμένων CNE (γεγονότα του τύπου ***iii***) και δείχνεται η θετική εξάρτηση της έκτασης της γραμμικότητας, σε σχέση με τον αριθμό των μη διπλασιασμένων CNE που έχουν χαθεί. Όπως συζητούμε με περισσότερη λεπτομέρεια πιο κάτω, αυτό το εύρημα είναι συμβατό με τις εκτεταμένες γραμμικότητες που παρατηρούνται στις κατανομές των CNE τελεόστεων, όπου έχουν αναφερθεί σημαντικές απώλειες προγονικών («αρχαίων») CNE. Τέτοια προγονικά CNE δε βρίσκονται στο γονιδίωμα των τελεόστεων, ενώ διατηρούνται στη γενιά των τετράποδων.

Εικόνα 5: Προσομοιώσεις χρησιμοποιώντας το μοντέλο γονιδιωματικών διπλασιασμών και απώλειας CNE. Απεικονίζεται η εξάρτηση της έκτασης της γραμμικότητας, σε log-log κλίμακα, για τις αποστάσεις μεταξύ διαδοχικών CNE υπό προσομοίωση και για διαφορετικές παραμέτρους από: **(a)** Τον αριθμό των τμηματικών διπλασιασμών (Segmental Duplications, SD), **(b)** Το ποσοστό των διπλασιασμένων CNE (fr) που απαλοίφονται μετά από κάθε γεγονός SD, **(c)** Τον αριθμό των επιπρόσθετων απαλοιφών, μη διπλασιασμένων, CNE. Στο **(a)** είμαστε σε θέση να παρακολουθήσουμε την εξέλιξη του αναδυόμενου προτύπου τύπου νόμου δύναμης, καθώς οι τέσσερις καμπύλες ανταποκρίνονται σε διαδοχικά στιγμιότυπα, που πάρθηκαν από το ίδιο αριθμητικό πείραμα. Η καμπύλη που αποτυπώνεται με τετράγωνα είναι κοινή και στα τρία διαγράμματα και αντιπροσωπεύει μια προσομοίωση που περιλαμβάνει 150 SD, όπου το 90% (fr=0.9) του αριθμού των διπλασιασμένων CNE χάνονται. Τα γραμμικά τμήματα υπολογίζονται με γραμμική παλινδρόμηση και σε όλες τις περιπτώσεις $r^2 > 0.98$



Εξελικτική προέλευση και προεκτάσεις των χρωμοσωμικών κατανομών των CNE

Στο τμήμα «Αποτελέσματα», πιο πριν, είδαμε ότι το πρότυπο κατανομής τύπου νόμου δύναμης εφαρμόζεται για τα CNE, UCE και άλλες υψηλά συντηρημένες αλληλουχίες, δίχως να αποτελεί απλή απόρροια της χωρικής κατανομής των γονιδίων (δείτε *Εικόνα 3* και *Πίνακα 2*). Η πλήρης μελέτη των χρωμοσωμάτων, που έχουν υποστεί μασκάρισμα των γονιδίων τους, πραγματοποιήθηκε στο ανθρώπινο γονιδίωμα για τα πιο πολυπληθή σύνολα δεδομένων CNE, όπως επίσης και για τα πιο «αρχαία» στοιχεία. Σε πέντε από τις έξι περιπτώσεις που εξετάστηκαν (δείτε *Πίνακα 2*, σύνολα δεδομένων: *i, iia, iib, vi, viii*), η μελέτη της μέσης έκτασης της γραμμικότητας για όλα τα χρωμοσώματα (avg E) και για τα πέντε χρωμοσώματα με την πιο εκτεταμένη γραμμικότητα (avg E-5) αποκαλύπτει ότι η έκταση της γραμμικής περιοχής σε διπλή λογαριθμική κλίμακα ακολουθεί μια αυξητική τάση (ή τουλάχιστον διατηρείται), αφού εφαρμοστεί η μεθοδολογία αποκλεισμού CNE που βρίσκονται κοντά ή μέσα σε γονίδια (για λεπτομέρειες, δείτε πιο πάνω στις «Μεθόδους»). Στη μία περίπτωση που απομένει (σύνολο δεδομένων *v*, άνθρωπος / zebrafish), καλά σχηματισμένες κατανομές τύπου νόμου δύναμης διατηρούνται, ακόμα και μετά το μασκάρισμα, με συνεπακόλουθη μείωση, όμως, του μήκους της γραμμικότητας. Η διατήρηση της γραμμικότητας σε διπλή λογαριθμική κλίμακα, μετά τον αποκλεισμό στοιχείων, δείχνει την ανεξαρτησία των δύο προτύπων. Το γεγονός ότι, στις περισσότερες περιπτώσεις, η μέση έκταση γραμμικότητας όχι μόνο διατηρείται, αλλά αυξάνεται, ισχυροποιεί αυτό το συμπέρασμα. Η συχνή αυτή βελτίωση των χαρακτηριστικών του νόμου δύναμης, όταν εφαρμόζεται το μασκάρισμα, ενδεχομένως αποτυπώνει το ότι, όταν μελετάται ολόκληρος ο χρωμοσωμικός πληθυσμός των CNE, τα παρατηρούμενα χαρακτηριστικά κατανομής αντανακλούν μια *υπέρθεση δύο διακριτών κατανομών*. Και οι δύο περιλαμβάνουν γεγονότα μοριακής δυναμικής που ανήκουν στους ίδιους τύπους, αλλά με διαφορετικούς ρυθμούς που ανταποκρίνονται στις διακριτές εξελικτικές διαφοροποιήσεις των CNE, που δεν είναι συσχετισμένα με γονίδια και των γονιδίων (με τα οποία τα CNE που είναι κοντά σε γονίδια είναι χωρικά συνδεδεμένα). Αυτή η υπέρθεση των κατανομών με διαφορετικά χαρακτηριστικά (την κλίση και την κλίμακα μήκους παρεμβαλλουσών αλληλουχιών) αναμένεται να ελαττώσει την παρατηρούμενη γραμμικότητα εξ' αιτίας ενός μετασχηματισμού τμήματος των υπερτιθέμενων γραμμικών log – log κατανομών σε ένα καμπυλωτό σχήμα. Ένα άλλο στοιχείο, που υποστηρίζει την απόκλιση μεταξύ των δύο προτύπων κατανομών, που ακολουθούνται από τα γονίδια και τα CNE, προέρχεται από την παρατήρηση, που έχει αναφερθεί από αρκετές ερευνητικές

ομάδες, ότι τα UCE και τα CNE είναι άφθονα στις γονιδιακές ερήμους (Kim & Pritchard 2007; Stephen et al. 2008).

Σε μια μελέτη του βαθμού συντήρησης των αποστάσεων μεταξύ των UCE στα σπονδυλωτά (Sun et al. 2006), πραγματική συντήρηση των αποστάσεων ευρίσκεται μόνο μεταξύ στοιχείων κοντινών αποστάσεων, ένα εύρος αποστάσεων που δύσκολα συνεισφέρει στις γραμμικότητες διπλης λογαριθμικής κλίμακας που αναφέρονται εδώ. Η απουσία σημαντικής διατήρησης αποστάσεων μεταξύ συντηρημένων στοιχείων είναι συμβατή και ενισχύει τον ισχυρισμό μας ότι ένα μοντέλο συσσωματώσεων (όπως αυτό που περιγράφηκε πιο πάνω) είναι κατάλληλο για την ερμηνεία της διαδεδομένης εμφάνισης γραμμικοτήτων τύπου νόμου δύναμης σε κατανομές αποστάσεων μεταξύ διαδοχικών CNE.

Ο Mattick και οι συνεργάτες του, σε ένα άρθρο πάνω στις υπερσυντηρημένες αλληλουχίες (UCE) στα γονιδιώματα των τετράποδων, παρατήρησαν μια εντυπωσιακή διαφορά στους πληθυσμούς των UCE μεταξύ των γονιδιωμάτων των τετράποδων, τα οποία είναι πλούσια σε UCE, και των γονιδιωμάτων των ψαριών, όπου βρέθηκαν σημαντικά χαμηλότεροι αριθμοί UCE (Stephen et al. 2008). Πρότειναν, ως την πιο φειδωλή εξήγηση, ότι στη γενιά των τετράποδων συνέβη μια μαζική προσαρμογή λειτουργικών στοιχείων, η οποία είναι πιθανό να ήταν απαραίτητη για την πιο πολύπλοκη μορφολογία των τετράποδων και την προσαρμογή / επιβίωση στις ποικίλες περιβαλλοντικές προκλήσεις που αντιμετώπιζαν εκείνοι οι οργανισμοί. Από την άλλη, οι ίδιοι συγγραφείς ανέφεραν ότι αυτό το εύρημα ενδεχομένως να εξηγείται από το λιγότερο πιθανό σενάριο μαζικής απώλειας τέτοιων στοιχείων στο γονιδίωμα των τελεόστεων ιχθύων. Η μετέπειτα αλληλούχηση του πρώτου γονιδιώματος χονδριχθύος (elephant shark) οδήγησε στην επιβεβαίωση του τελευταίου σεναρίου. Συγκεκριμένα, οι Lee και συνεργάτες βρήκαν ότι ο πρόγονος των γναθόστομων σπονδυλωτών είχε έναν σημαντικό αριθμό CNE, τα οποία έχουν απαλοιφθεί (έχει αλλάξει η αλληλουχία τους με αποτέλεσμα να μην είναι πλέον αναγνωρίσιμα) σε μεγάλο βαθμό στους τελεόστεους, ενώ έχουν διατηρηθεί στα τετράποδα (Lee et al. 2011). Αυτό το εύρημα ταιριάζει καλά με την παρατήρησή μας ότι οι τρεις μεγαλύτερες εκτάσεις γραμμικότητας, όπως φαίνεται από τον *Πίνακα 2* (σύνολα δεδομένων *iv*, *v*, *vii*), παρατηρούνται σε περιπτώσεις στοιχίσεων που συμπεριλαμβάνουν γονιδιώματα τελεόστεων ισχύων. Επιπροσθέτως, οι Wang και συνεργάτες συσχέτισαν ευθέως τις εκτεταμένες απαλοιφές CNE στους τελεόστεους ιχθύες με τον διπλασιασμό ολόκληρου του γονιδιώματος που είχε συμβεί στη γενεαλογία των ακτινοπτερύγιων (Wang et al. 2009). Σημειώστε τη σημασία των διπλασιασμών (ολόκληρου του γονιδιώματος αλλά και των τμηματικών) για το προτεινόμενο μοντέλο μας. Ο ρόλος τους είναι διπλός, καθώς

καθιστούν εφικτή την μετέπειτα απαλοιφή των διπλασιασμένων CNE λόγω πλεονασμού και, ταυτοχρόνως, παρέχουν την αναγκαία επέκταση αλληλουχίας για το σχηματισμό ευμεγέθων spacer μεταξύ των CNE.

Σε μια πρόσφατη εργασία των Dimitrieva και συνεργάτες (Dimitrieva & Bucher 2012) παρουσιάστηκαν στοιχεία που συνηγορούν υπέρ μιας εκτεταμένης ομαδοποίησης των UCNE (στοιχεία με μεγαλύτερη από 95% ομοιότητα μεταξύ ανθρώπου / κοτόπουλου για πάνω από 200 νουκλεοτίδια) στα γονιδιώματα διαφόρων σπονδυλωτών, σε ομάδες που σχετίζονται με τη ρύθμιση ενός γονιδίου κάθε φορά (γονίδιο – στόχος). Όταν πραγματοποιήθηκε απώλεια διπλασιασμένων UCNE στην εξέλιξη, η διατήρηση υπερσυντηρημένων αλληλουχιών αναφέρθηκε πως κάθε άλλο παρά τυχαία ήταν. Πιο συγκεκριμένα πρότειναν ότι ισχύει ένα πρότυπο διατήρησης στοιχείων που το ονόμασαν «ο νικητής τα παίρνει όλα» (“winner takes all”), δηλαδή ένα γονίδιο – στόχος διατηρεί πολλά UCNE ενώ το άλλο, παράλογο, γονίδιο χάνει όλα τα UCNE πιο συχνά από ότι θα αναμενόταν στα πλαίσια καθαρής τύχης. Βρέθηκε, επιπλέον, μια υπερεκπροσώπηση γονιδίων ψαριού που επιβίωσαν και τα οποία έχουν χάσει όλα τα προγονικά τους CNE. Αυτό το εύρημα είναι σε συμφωνία με την παρατήρησή μας για εκτεταμένη γραμμικότητα τύπου νόμου δύναμης που παρατηρείται στα πρότυπα κατανομής στοιχείων που έχουν προκύψει από στοιχίσεις με γονιδιώματα τελεόστεων ιχθύων. Οι σχετικά συχνές απαλοιφές ολόκληρων ομάδων UCNE ενισχύουν την εμφάνιση μεγάλων σε μήκος αποστάσεων (spacer) μεταξύ των UCNE, οι οποίες συνεισφέρουν στις μακριές ουρές και επομένως στο σχηματισμό εκτεταμένης γραμμικότητας σε διπλή λογαριθμική κλίμακα, για τις εκάστοτε κατανομές αποστάσεων (Newman 2005). Η λειτουργική αυτή ομαδοποίηση των UCNE με βάση τα γονίδια – στόχους, που περιγράφηκε από την ομάδα του Philipp Bucher, φαίνεται πως αντιπροσωπεύει μια διακριτή συνιστώσα στη χωρική κατανομή των UCNE. Η συγκεκριμένη ομαδοποίηση εκτείνεται σε κλίμακες μεγέθους της τάξης της γειτνίασης με ένα γονίδιο – στόχο. Τα αποτελέσματα που παρουσιάζονται εδώ έχουν να κάνουν με μια πιο εκτεταμένη κλίμακα μεγέθους, την κατανομή των CNE σε επίπεδο χρωμοσώματος, που έχει ως αποτέλεσμα μια πληθώρα επικρατειών πλούσιων σε CNE και φτωχών σε CNE, λόγω της δυναμικής συσσωματώσεων που περιγράφηκε πιο πριν. Αυτές οι επικράτειες ακολουθούν ένα πρότυπο τύπου μορφοκλασματικού σφαιριδίου (‘fractal globule’), όπως γίνεται εμφανές μέσω των κατανομών τύπου νόμου δύναμης που παρατηρούνται στις αποστάσεις μεταξύ των CNE. Η γνώση που έχουμε, ωστόσο, για τους ρόλους των CNE και τις ποσοτικές αλληλεπιδράσεις τους με γονίδια παραμένει αρκετά περιορισμένη.

Η σύγκριση της έκτασης της γραμμικότητας τύπου νόμου δύναμης μεταξύ συνόλων δεδομένων CNE Αμνιωτικών και Θηλαστικών (Kim & Pritchard 2007) βρίσκεται σε συμφωνία με το προτεινόμενο μοντέλο (*Πίνακας 2*, σύνολα δεδομένων *iia, iib*). Βλέπουμε ότι η παλαιότερη, στον εξελικτικό χρόνο, συλλογή δεδομένων Αμνιωτικών CNE σχηματίζει τις πιο εκτεταμένες γραμμικές περιοχές, όπως προβλέπεται από τις υποθέσεις του μοντέλου συσσωμάτωσης – απαλοιφών, όπου, όσο πιο πολύ εκτίθεται το σύστημα (σε εξελικτικό χρόνο) στη δυναμική εισδοχών και απαλοιφής αλληλουχιών CNE, τόσο πιο εκτεταμένη αναμένεται να είναι η γραμμικότητα σε διπλή λογαριθμική κλίμακα. Η ίδια τάση είναι εμφανής και σε τέσσερα σύνολα δεδομένων που έχουν παρθεί από μια άλλη μελέτη (Stephen et al. 2008). Αυτά τα σύνολα δεδομένων (*Πίνακας 2, ib – ie*) συνίστανται από στοιχεία που έχουν προσαρμοστεί σε διαδοχικές εξελικτικές περιόδους κατά την ειδογένεση των τετράποδων, των αμνιωτικών και των θηλαστικών. Παρατηρούμε ότι οι μέσες εκτάσεις γραμμικότητας για τα πέντε καλύτερα ("5 best") χρωμοσώματα ακολουθούν αυστηρά την αναμενόμενη αυξητική τάση (από το πιο πρόσφατο στο παλιότερο), ενώ το ίδιο ισχύει όταν εξετάζουμε τις μέσες εκτάσεις για πλήρη σύνολα χρωμοσωμάτων, με μόνο μία εξαίρεση (μεταξύ *ic* και *ib*).

Σε ότι αφορά όχι *τις αποστάσεις μεταξύ* λειτουργικών γονιδιωματικών εντοπισμών, αλλά *τα μεγέθη των εντοπισμών αυτά καθεαυτά*, πρέπει να αναφέρουμε ξανά εδώ ότι έχει παρατηρηθεί μια κατανομή νόμου δύναμης «τέλεια συντηρημένων» αλληλουχιών μεταξύ των γονιδιωμάτων ανθρώπου και ποντικού (Salerno et al. 2006). Ένα σχετιζόμενο εύρημα είναι ότι οι κατανομές μεγεθών «συντηρημένων ομάδων» μεταξύ των γονιδιωμάτων της *Drosophilia melanogaster* και *Drosophila virilis* ταιριάζουν καλά σε μια λογαριθμοκανονική (lognormal) κατανομή (Clark 2001). Σημειώνουμε εδώ ότι αυτή η κατανομή παρουσιάζει επίσης γραμμικές περιοχές σε log – log κλίμακα (Newman 2005). Μια κατανομή τύπου νόμου δύναμης έχει αναφερθεί επίσης για τις 3' μη μεταφραζόμενες περιοχές (Martignetti & Caselle 2007). Αυτά τα ευρήματα, όσον αφορά τα μεγέθη των τμημάτων της γονιδιωματικής αλληλουχίας, πρέπει να είναι το αποτέλεσμα της δράσης μηχανισμών εντελώς διαφορετικών από το εξελικτικό σενάριο που προτείνουμε εδώ, καθώς εκτείνονται σε πολύ μικρές κλίμακες μεγεθών (δεκάδες έως εκατοντάδες νουκλεοτιδίων), ενώ οποιαδήποτε τυχαία διαδικασία συσσωμάτωσης είναι ακατάλληλη για το σχηματισμό λειτουργικών στοιχείων αλληλουχίας. Η αύξηση μεγέθους μέσω συσσωματώσεων είναι περισσότερο κατάλληλη για τη διαμόρφωση μεγάλων περιοχών του γονιδιώματος (π.χ. αλληλουχίες που παρεμβάλλονται μεταξύ γονιδίων) με μικρές απαιτήσεις συντήρησης (Sellis & Almirantis 2009; Takayasu et al. 1991). Παρόλα αυτά, η

σύμπραξη αυτών των γραμμικοτήτων στη μορφή των νόμων δύναμης, για αρκετές τάξεις μεγέθους και για διάφορα λειτουργικά (ή πιθανόν λειτουργικά) στοιχεία του γονιδιώματος ή για τις αποστάσεις μεταξύ αυτών, είναι πιθανό να σχετίζεται με τη δομή μορφοκλασματικού σφαιριδίου (fractal globule), που έχει αναφερθεί για το ανθρώπινο γονιδίωμα, όταν αυτό βρίσκεται στη μορφή στενά πακεταρισμένης χρωματίνης (Lieberman-Aiden et al. 2009). Αυτή η δομή χαρακτηρίζεται από εκτεταμένες κατανομές μεγεθών τύπου νόμου δύναμης στις αποστάσεις μεταξύ των σημείων του χρωμοσωμικού νήματος, τα οποία έρχονται σε αμοιβαία επαφή λόγω της τρισδιάστατης αναδίπλωσης της χρωματίνης.

3. Κατάταξη αλληλουχιών του γονιδιώματος που βρίσκονται υπό επιλεκτικό περιορισμό με μεθόδους μηχανικής μάθησης

3.1 Κατάταξη CNE και εξωνίων με τη χρήση διανυσμάτων χαρακτηριστικών αλληλουχίας και ταξινομητών λογικών κανόνων

Εισαγωγή

Η ομοιότητα αλληλουχιών θεωρείται γενικά ως μια ένδειξη ομοιότητας σε επίπεδο λειτουργίας. Οι περισσότερες από τις υπάρχουσες μεθόδους ανάλυσης αλληλουχιών βασίζονται σε στοιχίσεις και πιο συγκεκριμένα στοιχίσεις περιοχών αλληλουχιών σε διάφορες κλίμακες μήκους, από γονίδια μέχρι ολόκληρα γονιδιώματα. Κάθε στοίχιση αξιολογείται με βάση ένα σκορ που εξαρτάται από τον αριθμό των ίδιων και επαναλαμβανόμενων χαρακτήρων στις αλληλουχίες. Οι βέλτιστοι μέθοδοι στοίχισης αλληλουχιών βασίζονται σε τεχνικές δυναμικού προγραμματισμού, εκ των οποίων οι πιο γνωστές είναι των Needleman και Wunsch (Needleman & Wunsch 1970) και των Smith και Waterman (Smith & Waterman 1981). Οι συγκεκριμένοι αλγόριθμοι είναι απαιτητικοί από υπολογιστικής σκοπιάς και η πολυπλοκότητά τους αυξάνει εκθετικά με το μήκος των αλληλουχιών. Έχουν προταθεί ευριστικές λύσεις που λύνουν το πρόβλημα της στοίχισης αλληλουχιών, όπως το BLAST (Altschul 1997) και το FASTA (Pearson 1990). Άλλοι αλγόριθμοι, όπως ο ClustalW (Thompson et al. 2002), ο Muscle (Edgar 2004), ο Mafft (Katoh 2002) και ο Motalign (Mokaddeml & Elloumi 2013) έχουν προταθεί για τη στοίχιση πολλαπλών αλληλουχιών με μεγαλύτερη αποτελεσματικότητα.

Από τις πρώτες δεκαετίες συστηματικής αλληλούχησης περιοχών του γονιδιώματος που κωδικοποιούν πρωτεΐνες και μη κωδικοποιουσών περιοχών παρατηρήθηκε ότι, ενώ οι στοιχίσεις κωδικοποιουσών περιοχών μεταξύ οργανισμών και μέσα στο ίδιο γονιδίωμα παρέχουν πολλές πληροφορίες, δεν έχουν (ή έχουν περιορισμένη) εφαρμογή στο μη κωδικοποιούν τμήμα του γονιδιώματος. Αυτό συμβαίνει επειδή το τελευταίο, κατά το μεγαλύτερο μέρος του, δεν είναι συντηρημένο κατά την εξέλιξη. Παρόλα αυτά, οι μέθοδοι στοίχισης ενδέχεται να είναι χρήσιμοι όταν εφαρμόζονται σε μικρές μη κωδικοποιούσες αλληλουχίες DNA. Σε μεγαλύτερες χρωμοσωμικές περιοχές είναι περιορισμένη η χρησιμότητά τους, διότι σε μεγάλες

περιοχές η νουκλεοτιδική σύνθεση δεν είναι συντηρημένη μεταξύ οργανισμών, οι οποίοι είναι σχετικά απομακρυσμένοι εξελικτικά. Η στοίχιση αλληλουχιών αποδεικνύεται ιδιαίτερα χρήσιμη κατά τη μελέτη των μεταθετών στοιχείων, τα οποία ανιχνεύονται σε πολλαπλά αντίγραφα στους περισσότερους οργανισμούς και χαρακτηρίζονται από ποικίλους βαθμούς ομολογίας μεταξύ τους, ανάλογα με την ηλικία τους στο γονιδίωμα. Στοιχίσεις, άλλωστε, ολικών γονιδιωμάτων μεταξύ δύο ή περισσότερων ειδών οδήγησε στην ανακάλυψη των διαφόρων κατηγοριών CNE.

Οι μέθοδοι που δε βασίζονται στη στοίχιση είναι χρήσιμες όταν θέλουμε να εξάγουμε χαρακτηριστικά σύστασης διαφόρων αλληλουχιών. Εφαρμόζονται και σε συγκρίσεις ολόκληρων γονιδιωμάτων μεταξύ οργανισμών αλλά και στην, εντός του ίδιου γονιδιώματος, ανίχνευση τμημάτων που επιδεικνύουν ιδιαιτερότητες στη σύστασή τους, οι οποίες συχνά σχετίζονται με τη λειτουργικότητά τους. Ένα κλασικό παράδειγμα μεθοδολογίας ανάλυσης του γονιδιώματος, που δε χρησιμοποιεί στοιχίσεις, βασίζεται στη μελέτη των αποστάσεων μεταξύ των γονιδιωματικών υπογραφών των αλληλουχιών, παρουσιάζεται στις «Μεθόδους» πιο κάτω και χρησιμοποιείται για σύγκριση με τις μεθόδους ανάλυσης του γονιδιώματος που εισάγουμε εδώ. Βασισμένοι σε μεθοδολογίες που δε χρησιμοποιούν στοίχιση είναι και διάφοροι αλγόριθμοι για τον εντοπισμό νησίδων CpG (“CpG islands”), οι οποίες είναι μικρές αλληλουχίες του γονιδιώματος με καμία αμοιβαία ομοιότητα αλλά με διάφορα, διακριτά, χαρακτηριστικά σύστασης. Τα τμήματα που κωδικοποιούν πρωτεΐνες είναι δυνατόν να καθορισθούν με μεθοδολογίες που βασίζονται σε στοιχίσεις, αλλά και με αυτές που δε βασίζονται, όπως η χρήση του γενετικού κώδικα και οι ιδιαιτερότητες του μηχανισμού παραγωγής πρωτεϊνών (πρόσδεση mRNA – ριβοσώματος, πλεονασμοί tRNA, κ.ά.), γνωρίσματα που αποδίδουν στο τμήμα του γονιδιώματος που κωδικοποιεί πρωτεΐνες χαρακτηριστικές ιδιαιτερότητες σύστασης. Γενικά θα μπρούσαμε να πούμε ότι οι μέθοδοι που βασίζονται σε στοιχίσεις και αυτές που δε βασίζονται είναι συμπληρωματικές μεταξύ τους και ότι συγκεκριμένες συνιστώσες του γονιδιώματος θα μπορούσαν να μελετηθούν με κατάλληλους συνδυασμούς των δύο.

Οι μέθοδοι, που δε βασίζονται σε στοιχίσεις, αποδεικνύονται ιδιαίτερα επιτυχημένες κατά τη φυλογένεση και την ανάλυση αλληλουχιών (Vinga & Almeida 2003). Όσον αφορά αυτές τις μεθόδους, η ομοιότητα δύο αλληλουχιών εκτιμάται με βάση αποκλειστικά το λεξικό των υποακολουθιών που εμφανίζονται μέσα στις αλληλουχίες υπό μελέτη, μη λαμβάνοντας υπόψη τη σχετική τους θέση. Μια υποσχόμενη μέθοδος, που δε χρησιμοποιεί στοιχίσεις, βασίζεται στην αναπαράσταση

μιας αλληλουχίας μέσω διανύσματος χαρακτηριστικών, όπου μια αλληλουχία κωδικοποιείται σε ένα διάνυσμα που περιέχει τις εμφανίσεις (ή συχνότητες εμφάνισης) των υποακολουθιών που την απαρτίζουν (κ – μερή, k – mers) (Kudenko & Hirsh 1998; Kuksa & Pavlovic 2009; Xing et al. 2010). Ένα κ – μερές (k – mer) ορίζεται ως μια υποακολουθία της αρχικής αλληλουχίας, μήκους κ (k) χαρακτήρων. Στην ανάλυση συχνοτήτων κ – μερών εξάγεται κάθε δυνατό κ – μερές από το νουκλεοτιδικό αλφάβητο {A, C, G, T}, απαριθμούνται οι εμφανίσεις του στην αρχική αλληλουχία και υπολογίζεται η συχνότητά του. Υπολογίζεται, στη συνέχεια, ένα διάνυσμα που περιέχει τη σχετική συχνότητα κάθε κ – μερούς, για κάθε αλληλουχία στο σύνολο δεδομένων που εξετάζεται.

Τα CNE απαντώνται συνήθως σε ένα αντίγραφο στο γονιδίωμα (Bejerano, Pheasant, et al. 2004; McLean & Bejerano 2008; Stephen et al. 2008) και έχουν προταθεί ως δείκτες σε μελέτες μοριακής φυλογένεσης (McCormack et al. 2012). Στη συνέχεια προτείνουμε μια προσέγγιση, που δε βασίζεται στο πληροφοριακό περιεχόμενο μιας στοίχισης μεταξύ διαφορετικών περιοχών DNA, από διαφορετικά γονιδιώματα, και την εφαρμόζουμε στην ταξινόμηση λειτουργικών (ή πιθανόν λειτουργικών) αλληλουχιών, που προέρχονται από διαφορετικά γονιδιώματα σπονδυλωτών και ασπόνδυλων. Προχωράμε επίσης σε συγκρίσεις ώστε να ελέγξουμε τη δυναμική της μεθοδολογίας μας, συγκρίνοντας την απόδοσή της στην ταξινόμηση μικρών γονιδιωματικών αλληλουχιών διαφορετικών λειτουργικοτήτων και εξελικτικού βάθους, όχι μόνο μεταξύ ειδών, αλλά και μέσα στον ίδιο οργανισμό. Γι’ αυτό το σκοπό χρησιμοποιούμε μια ευρεία συλλογή CNE σπονδυλωτών που είναι δημοσιευμένα στη βιβλιογραφία, καθώς και νέα σύνολα δεδομένων CNE ανθρώπου και ασπόνδυλων, που ταυτοποιούμε σε αυτή τη μελέτη (Polychronopoulos, Weitschek, et al. 2014).

Μέθοδοι

Δημοσιευμένα σύνολα δεδομένων

Σε αυτή τη μελέτη θεωρούμε διάφορα δημοσιευμένα σύνολα δεδομένων αλληλουχιών, που βρίσκονται υπό επιλεκτικό περιορισμό και έχουν εξαχθεί από τα γονιδιώματα του ανθρώπου (*Homo sapiens*), της μύγας (*Drosophila melanogaster*) και του σκώληκα (*Caenorhabditis elegans*). Επιλέγουμε, τυχαία, 1000 αλληλουχίες από κάθε υπερσύνολο για μετέπειτα αναλύσεις, δεδομένης της ετερογένειας, όσον αφορά τους αριθμούς, των συνόλων δεδομένων. Μόνο οι αλληλουχίες του σκώληκα μελετώνται στο σύνολό τους,

καθώς αυτό το σύνολο περιέχει μόνο 1869 στοιχεία. Οι εξωνικές αλληλουχίες του ανθρώπου, του σκώληκα και της μύγας λαμβάνονται από το *UCSC* (University of California at Santa Cruz) (Kent et al. 2002). Εκτός από εξωνικές αλληλουχίες, διάφορες άλλες κλάσεις μη εξωνικών αλληλουχιών, υπό επιλεκτική πίεση, χρησιμοποιούνται και πιο συγκεκριμένα (δείτε επίσης dmb.iasi.cnr.it/cnes.php):

(a) Υπερσυντηρημένες Μη Κωδικοποιούσες Αλληλουχίες (UltraConserved Noncoding Elements, UCNE): Αυτές είναι ανθρώπινες αλληλουχίες που δεν κωδικοποιούν πρωτεΐνες (εκδοχή γονιδιώματος: hg19) και οι οποίες επιδεικνύουν ≥ 95% ομοιότητα για πάνω από 200 νουκλεοτίδια μεταξύ των γονιδιωμάτων του ανθρώπου και του κοτόπουλου (Dimitrieva & Bucher 2013).

(b) EU100 μη εξωνικές αλληλουχίες CNE (EU100nx CNE): Αυτές είναι ανθρώπινες αλληλουχίες (εκδοχή γονιδιώματος: hg18) που είναι πανομοιότυπες για πάνω από 100 νουκλεοτίδια σε τουλάχιστον 3 από 5 πλακουντοφόρα θηλαστικά (άνθρωπος, ποντίκι, αρουραίος, σκύλος και αγελάδα) (Stephen et al. 2008). Ολόκληρο το σύνολο ονομάζεται EU100+, όπως αναφέραμε και στο προηγούμενο κεφάλαιο, και επειδή αφαιρούμε στοιχεία που επικαλύπτονται με εξώνια τις ονομάζουμε EU100nx CNE.

(c) CNE αμνιωτικών και θηλαστικών: Αυτές είναι αλληλουχίες που ταυτοποιήθηκαν από τους Kim και Pritchard (Kim & Pritchard 2007) και οι μεν θηλαστικών είναι αλληλουχίες που είναι συντηρημένες στα θηλαστικά αλλά δε βρίσκονται στο κοτόπουλο ή στο ψάρι, ενώ εκείνες των αμνιωτικών είναι συντηρημένες στα θηλαστικά και στο κοτόπουλο, αλλά δε βρίσκονται στο ψάρι. Το LiftOver χρησιμοποιείται για τη μετατροπή των συντεταγμένων στην πιο πρόσφατη εκδοχή του ανθρώπινου γονιδιώματος (από hg17 σε hg19).

Ταυτοποίηση νέων συνόλων δεδομένων CNE σε σπονδυλωτά, έντομα και σκώληκες

Για τις ανάγκες της μελέτης μας προχωρούμε στην ταυτοποίηση συνόλων συντηρημένων στοιχείων στα σπονδυλωτά, στα έντομα και στους σκώληκες χρησιμοποιώντας ομοιόμορφα κριτήρια, όσον αφορά τα κατώφλια ελάχιστης ομοιότητας που εφαρμόζονται μεταξύ των συγκρινόμενων αλληλουχιών καθώς και το θεωρούμενο ελάχιστο μήκος. Διαλέγουμε να συγκρίνουμε ολόκληρα γονιδιώματα *D. melanogaster* / *D. virilis* και *C. elegans* / *C. japonica*, διότι τα είδη που απαρτίζουν αυτά τα ζευγάρια είναι αρκετά απομακρυσμένα (εξελικτικά) μεταξύ τους, ώστε να

επιτρέπουν έναν ξεκάθαρο διαχωρισμό μεταξύ λειτουργικά συντηρημένων και ουδέτερα εξελισσόμενων γονιδιωματικών αλληλουχιών, ενώ επιλέγονται έτσι ώστε οι εξελικτικές αποστάσεις μέσα σε κάθε ζεύγος να είναι κοντινές (Retelska et al. 2007). Ολικές στοιχίσεις γονιδιωμάτων μεταξύ *D. melanogaster* (dm3) / *D. virilis* (droVir3) και *C. elegans* (ce10) / *C. japonica* (caeJap4) ανακτώνται από το *UCSC Genome Browser* (Kent et al. 2002). Ως συντηρημένα στοιχεία θεωρούνται περιοχές αλληλουχίας, όπου το ποσοστό ομοιότητας σε επίπεδο αλληλουχίας είναι διαρκώς ≥ 90% και το μήκος μεγαλύτερο από 60 νουκλεοτίδια. Η ομοιότητα σε επίπεδο αλληλουχίας υπολογίζεται κατά έναν ασύμμετρο τρόπο, λαμβάνοντας ως αναφορά τα γονιδιώματα της *D. melanogaster* και του *C. elegans* και μετρώντας τον αριθμό των συντηρημένων νουκλεοτιδίων στα είδη – στόχους με ένα κυλιόμενο παράθυρο 61 νουκλεοτιδίων. Ο αριθμός, που λαμβάνεται για κάθε παράθυρο, χρησιμοποιείται για την απόδοση μιας τιμής ποσοστού ομοιότητας στο νουκλεοτίδιο που βρίσκεται στη μέση του παραθύρου, όπως περιγράφεται από τους Retelska και συνεργάτες (Retelska et al. 2007). Σε σύνολο, ταυτοποιήθηκαν 45345 αλληλουχίες ως συντηρημένες μεταξύ *D. melanogaster* και *D. virilis* στο γονιδίωμα αναφοράς dm3 και 1869 αλληλουχίες μεταξύ *C. elegans* και *C. japonica* στο γονιδίωμα αναφοράς ce10. Βασισμένοι στους σχολιασμούς γονιδίων του Ensembl για τα γονιδιώματα της *D. melanogaster* και *C. elegans*, κάθε μία από τις συντηρημένες αλληλουχίες κατηγοριοποιείται ως κωδικοποιούσα κάποια πρωτεΐνη ή μη κωδικοποιούσα (εσωνική, σχετιζόμενη με UTR ή μεταξύ γονιδίων αλληλουχία).

Προχωρούμε επίσης στην ταυτοποίηση CNE σπονδυλωτών, βασισμένοι σε συγκρίσεις μεταξύ ανθρώπου και κοτόπουλου, που επιδεικνύουν ποικίλους βαθμούς ομοιότητας (με ένα σταθερό βήμα 5%, ήτοι 75 – 80%, 80 - 85%, 85 – 90%, 90 – 95%). Με σκοπό να πάρουμε CNE που έχουν ομοιότητα μεταξύ π.χ. 75% και 80%, εξάγουμε πρώτα ένα σύνολο CNE που επιδεικνύουν ομοιότητα 75% (χρησιμοποιώντας την ίδια μεθοδολογία όπως περιγράφτηκε πιο πάνω) και από αυτό το σύνολο αφαιρούμε CNE που επιδεικνύουν ομοιότητα 80%. Από τις αλληλουχίες που απομένουν, κρατούμε μόνο εκείνες που είναι μεγαλύτερες από 199 νουκλεοτίδια. Επαναλαμβάνουμε την ίδια διαδικασία για τα άλλα σύνολα δεδομένων CNE, κατά τον ίδιο τρόπο.

Εξαγωγή αναπληρωματικών αλληλουχιών

Αλληλουχίες τύπου FASTA λαμβάνονται από αρχεία τύπου BED, χρησιμοποιώντας το πακέτο *fastaFromBed* των *BEDTools* (Quinlan & Hall 2010). Επικαλυπτόμενα τμήματα μεταξύ διαφορετικών συνόλων δεδομένων υπολογίζονται με τη χρήση του πακέτου *intersectBed* από την ίδια σουίτα εργαλείων. Επιπλέον εφαρμόζουμε τα προγράμματα που ενσωματώνει το πακέτο *EMBOSS* ώστε να υπολογίζουμε κάθε φορά το περιεχόμενο σε GC των αλληλουχιών (Rice et al. 2000). Δεδομένου ενός CNE ή ενός άλλου λειτουργικού στοιχείου κάθε συλλογής υπό μελέτη, ένα ανάλογο κάθε στοιχείου εξάγεται από το αντίστοιχο γονιδίωμα. Κάθε τέτοιο στοιχείο (αναπληρωματική αλληλουχία) διασφαλίζεται να είναι του ίδιου μήκους και περιεχομένου σε GC (±1%) με το αντίστοιχο στοιχείο στην, υπό μελέτη, συλλογή. Τα στατιστικά των συνόλων δεδομένων (μέσο μήκος αλληλουχιών και περιεχόμενο σε GC) είναι διαθέσιμα στο «Συμπληρωματικό Excel» στη σελίδα dmb.iasi.cnr.it/cnes/php. Λόγω του ότι τα CNE σπονδυλωτών και ασπόνδυλων διαφέρουν στα μήκη τους (αυτά που ανήκουν στην τελευταία κατηγορία είναι σημαντικά μικρότερα, δείτε (Retelska et al. 2007)), διασφαλίζουμε ότι παίρνουμε στοιχεία ίδιων μηκών ως ακολούθως (σε όλες τις περιπτώσεις τα σύνολα δεδομένων που συγκρίνονται περιλαμβάνουν το κάθε ένα 1000 αλληλουχίες):
(i) υπολογίζουμε κάθε φορά το μήκος ενός στοιχείου από τη «συλλογή μικρών αλληλουχιών», (ii) έπειτα παίρνουμε ένα στοιχείο από τη «συλλογή μεγάλων αλληλουχιών» και το τροποποιούμε στις άκρες του διατηρώντας μόνο το κεντρικό κομμάτι ίσο με το μήκος της εκάστοτε «μικρής αλληλουχίας», (iii) επαναλαμβάνουμε τη διαδικασία διεξοδικά (1000 φορές σε όλες τις περιπτώσεις μας). Επομένως καταλήγουμε σε ένα νέο σύνολο δεδομένων για τις «μεγάλες αλληλουχίες» (συλλογή αλληλουχιών σε μορφή αρχείου BED), των οποίων τα μέλη έχουν μήκη ίσα με τα μέλη της «συλλογής μικρών αλληλουχιών».

Η μέθοδος ταξινόμησης αλληλουχιών χωρίς στοίχιση (Logic Alignment Free, LAF)

Με σκοπό να δείξουμε ότι μπορούμε να διαχωρίσουμε CNE από άλλες λειτουργικές αλληλουχίες αλλά και από αναπληρωματικές αλληλουχίες, με κριτήριο μόνο τις ενδογενείς ιδιότητες αλληλουχίας τους, εφαρμόζουμε μια μέθοδο χωρίς στοιχίσεις που βασίζεται σε:

- μια αναπαράσταση αλληλουχιών με διανύσματα χαρακτηριστικών και

- ταξινόμηση με τεχνικές εποπτευόμενης μηχανικής μάθησης που βασίζονται σε λογικούς κανόνες.

και την ονομάζουμε LAF (Logic Alignment Free).

Το διάνυσμα χαρακτηριστικών είναι μια καθιερωμένη τεχνική για την αναπαράσταση βιολογικών αλληλουχιών και για την μετέπειτα ταξινόμησή τους. Η μεθοδολογία περιγράφεται από τους Kudenko και Hirsch (Kudenko & Hirsh 1998), Vinga και Almeida (Vinga & Almeida 2003) και Xing και συνεργάτες (Xing et al. 2010), όπου εισάγεται η έννοια της αναπαράστασης αλληλουχίας μέσω διανυσμάτων χαρακτηριστικών και συνδυάζεται με εποπτευόμενες μεθόδους μηχανικής μάθησης, όπως Μηχανές Υποστήριξης Διανυσμάτων (Support Vector Machines - SVM), με σκοπό την ταξινόμηση δειγμάτων σε είδη. Μια άλλη πρόσφατη εργασία είναι αυτή των Kuksa και Pavlovic (Kuksa & Pavlovic 2009), οι οποίοι εφάρμοσαν αναπαραστάσεις διανυσμάτων χαρακτηριστικών για την ταξινόμηση ειδών (DNA Barcode[5]).

Η αναπαράσταση μέσω διανύσματος χαρακτηριστικών μιας αλληλουχίας S βασίζεται στον υπολογισμό των εμφανίσεων υποακολουθιών δεδομένου μήκους κ στην αρχική αλληλουχία, εφαρμόζοντας ένα κυλιόμενο παράθυρο στην S. Αυτές οι υποακολουθίες ονομάζονται κ – μερή (k – mers). Το πλήθος των κ – μερών μιας αλληλουχίας αντιπροσωπεύεται από ένα διάνυσμα χαρακτηριστικών, όπου κάθε συνιστώσα του διανύσματος σχετίζεται με τις εμφανίσεις ενός δεδομένου κ – μερούς.

Οι Vinga και Almeida δίνουν τον εξής ορισμό (Vinga & Almeida 2003): Έστω S μια αλληλουχία n χαρακτήρων για ένα αλφάβητο A, π.χ. A = {A, C, G και T}, και έστω $\kappa \epsilon I$, $\kappa < n$, $\kappa > 0$. Εάν το K είναι μια γενική υποακολουθία της S μεγέθους κ, το K καλείται κ – μερές. Έστω το σύνολο V = {κ-μερές1, κ-μερές2, ..., κ-μερέςr} ότι αντιπροσωπεύει όλα τα πιθανά κ-μερή στο A και ότι το V έχει μέγεθος t = |A|κ. Τα κ – μερή υπολογίζονται μετρώντας τις εμφανίσεις όλων των υποακολουθιών στην S με ένα κυλιόμενο παράθυρο μεγέθους κ για όλη την S, ξεκινώντας από τη θέση 1 και τελειώνοντας στη θέση ν-κ+1. Ένα διάνυσμα συχνοτήτων C περιέχει, για κάθε κ – μερές, τις εμφανίσεις (ή αριθμούς) του $C = \{c_{\kappa\text{-μερές1}}, c_{\kappa\text{-μερές2}}, \ldots, c_{\kappa\text{-μερέςt}}\}$. Στη συνέχεια

[5] Το 2003, ο Paul Hebert, ερευνητής στο Πανεπιστήμιο του Guelph στο Οντάριο πρότεινε το «DNA barcoding» ως μία μέθοδο για την ταυτοποίηση των ειδών. Το Barcoding χρησιμοποιεί μια μικρή γενετική ακολουθία 648 νουκλεοτιδίων που βρίσκεται στο γονίδιο της κυτοχρωμικής c οξειδάσης 1 (“CO1”), ενώ ονομάστηκε έτσι, λόγω της διάκρισης που πραγματοποιεί με τρόπο ίδιο με τις μαύρες γραμμές (barcodes) των υπερμάρκετ. Δύο αντικείμενα μπορεί να φαίνονται ίδια στο μη εκπαιδευμένο μάτι, ωστόσο και στις δύο περιπτώσεις, τα barcodes είναι διακριτά.

υπολογίζονται οι συχνότητες, σύμφωνα με τις εμφανίσεις, και αποθηκεύονται σε ένα διάνυσμα $F = \{f_{κ\text{-μερές}1}, f_{κ\text{-μερές}2}, ..., f_{κ\text{-μερές}t}\}$, ενώ για ένα γενικό κ – μερές, έστω j, η συχνότητά του ορίζεται ως $f_j = c_{κ\text{-μερές}j} / (ν\text{-}κ+1)$. Για παράδειγμα, θεωρώντας το αλφάβητο του DNA {A, C, G, T}, 2 – μερή, και την αλληλουχία ACGACT, το διάνυσμα χαρακτηριστικών C είναι:

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	2	0	0	0	0	1	1	1	0	0	0	0	0	0	0

Και το διάνυσμα συχνοτήτων F είναι:

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	2/5	0	0	0	0	1/5	1/5	1/5	0	0	0	0	0	0	0

Οι αλληλουχίες αντιπροσωπεύονται κατά τέτοιο τρόπο σε ένα χώρο συντεταγμένων, ώστε να είναι μαθηματικά ανιχνεύσιμες με γραμμική άλγεβρα και στατιστική, δηλαδή, π.χ. θεωρώντας την αναπαράσταση διανυσμάτων για τις αλληλουχίες είναι εφικτό να υπολογίσουμε διαφορετικές μετρικές αποστάσεων μεταξύ δύο αλληλουχιών ή να δώσουμε το αντιπροσωπευτικό διάνυσμα ως είσοδο σε έναν αλγόριθμο εποπτευόμενης μηχανικής μάθησης, δηλαδή σε έναν ταξινομητή (classifier).

Η προσέγγιση εποπτευόμενης μηχανικής μάθησης καλείται επίσης ταξινόμηση (classification). Κάθε συλλογή των αλληλουχιών, που αναλύονται, πρέπει να περιέχει μια εκ των προτέρων γνωστή ετικέτα κλάσης, δηλαδή κάθε αλληλουχία σχετίζεται με μια δεδομένη κλάση, π.χ. Σπονδυλωτά - Ασπόνδυλα ή Αμνιωτικά - Θηλαστικά. Κάθε τέτοια συλλογή καλείται σύνολο δεδομένων εκπαίδευσης (training dataset). Βασισμένη στο σύνολο εκπαίδευσης, η μέθοδος εξάγει ένα μοντέλο ταξινόμησης (classification model), για παράδειγμα, κανόνες «εάν – τότε» (“if-then”), για το διαχωρισμό των διαφορετικών αλληλουχιών που υπάρχουν στο σύνολο δεδομένων. Το μοντέλο αυτό, στη συνέχεια, μπορεί να εφαρμοστεί (και να εκτιμηθεί) πάνω σε ένα σύνολο δεδομένων δοκιμασίας (test dataset), το οποίο αποτελείται από άγνωστες αλληλουχίες ή αλληλουχίες που ανήκουν σε μια γνωστή κλάση, έτσι ώστε να πάρουμε αποδόσεις ταξινόμησης. Η μέθοδος, που υιοθετούμε, συνδυάζει την αναπαράσταση μέσω διανυσμάτων χαρακτηριστικών μαζί με μεθόδους εποπτευόμενης μηχανικής μάθησης, που βασίζονται σε λογικούς κανόνες (Lehr et al. 2011; Weitschek et al. 2014).

Οι ταξινομητές που βασίζονται σε λογικούς κανόνες (rule based classifiers) έχουν προταθεί, στο παρελθόν, ως μια τεχνική για την ανάλυση βιολογικών αλληλουχιών και πιο συγκεκριμένα για την κατάταξη οργανισμών (χρησιμοποιώντας

αλληλουχίες DNA Barcode) και για την ταυτοποίηση ειδών (Bertolazzi et al. 2009; Weitschek et al. 2012). Σε αυτές τις εργασίες, οι αλληλουχίες του γονιδιώματος αναλύονται με μια προσέγγιση που λάμβανε υπόψη τη θέση: κάθε νουκλεοτίδιο μελετάται ξεχωριστά, αναφέροντας μόνο στη θέση του. Επομένως ήταν αναγκαία μια στοίχιση των αλληλουχιών ή μιας επικαλυπτόμενης γονιδιακή περιοχής. Πραγματοποιήθηκε μια ανάλυση των χαρακτηριστικών νουκλεοτιδίων, που είναι παρόντα σε μια δεδομένη θέση για κάθε κλάση, οδηγώντας σε λογικούς κανόνες του τύπου, π.χ., εάν θέση90 = A τότε η αλληλουχία ανήκει στην κλάση X. Η στοίχιση ήταν απαραίτητη, διότι μια ανάλυση θέσης είναι εφικτή μόνο όταν οι αλληλουχίες προέρχονται από τις ίδιες περιοχές γονιδίων και στοιχίζονται ως προς μια θέση αναφοράς.

Η έξοδος ενός ταξινομητή, που βασίζεται σε λογικούς κανόνες, είναι μια συλλογή κανόνων "if-then", εκχωρώντας μια αλληλουχία σε μια συγκεκριμένη κλάση (είδος), π.χ., εάν θέση30 = A και θέση40 = C τότε η αλληλουχία ανήκει στον οργανισμό squalus edmundsi. Το κύριο πλεονέκτημα της ταξινόμησης με βάση λογικούς κανόνες είναι η επιπρόσθετη γνώση που παρέχεται από τα συμπαγή μοντέλα – κανόνες (κανόνες "if then"). Στην παρούσα εργασία, οι ακόλουθοι αλγόριθμοι χρησιμοποιούνται, ώστε να πραγματοποιήσουμε ανάλυση ταξινόμησης, με βάση λογικούς κανόνες, των αναπαραστάσεων μέσω διανυσμάτων χαρακτηριστικών, αλληλουχιών που βρίσκονται υπό επιλεκτικό περιορισμό (CNE ή εξώνια): RIPPER (Cohen & Singer 1999), RIDOR (Gaines & Compton 1995) και PART (Frank & Witten 1998).

Ο RIPPER είναι ένας αλγόριθμος ταξινόμησης που χρησιμοποιεί μια άμεση μέθοδο για εξαγωγή κανόνων: συνάγει τους κανόνες με την άμεση επεξεργασία των δεδομένων. Ο RIPPER συνίσταται από τα ακόλουθα υπολογιστικά βήματα:

1) επέκταση κανόνων (rule growth)
2) απλοποίηση / «ψαλίδισμα» κανόνων (rule pruning)
3) βελτιστοποίηση μοντέλου (model optimization)
4) επιλογή μοντέλου (model selection).

Στο πρώτο βήμα υπολογίζονται οι κανόνες με ένα γρήγορο τρόπο μέσω της επεξεργασίας των χαρακτηριστικών των δεδομένων. Οι κανόνες, στη συνέχεια, απλοποιούνται στο δεύτερο βήμα, σύμφωνα με τα στατιστικά μέτρα στο σύνολο εκπαίδευσης. Στο τελευταίο βήμα επιλέγονται οι κανόνες με την υψηλότερη επίδοση και εκείνοι που απομένουν απορρίπτονται από το μοντέλο ταξινόμησης.

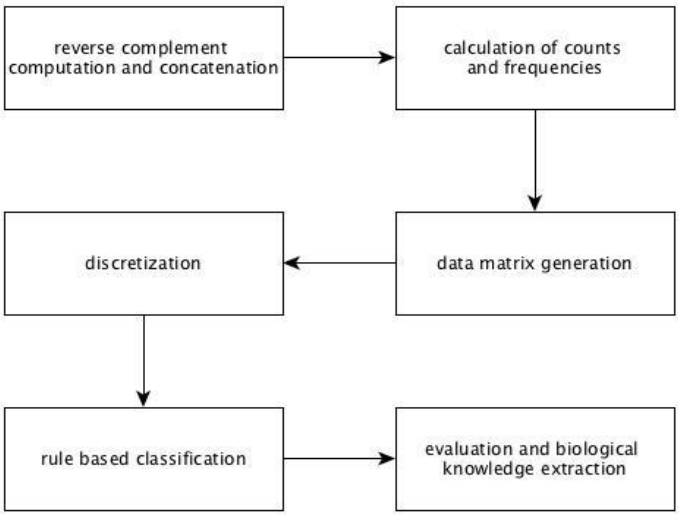
Ο αλγόριθμος ταξινόμησης RIDOR εξάγει επίσης απευθείας τους κανόνες από τα δεδομένα. Υπολογίζει πρώτα ένα γενικό κανόνα που καλύπτει την περισσότερο εκπροσωπημένη κλάση (π.χ., "όλες οι αλληλουχίες είναι σπονδυλωτών") και έπειτα κανόνες, που αντιπροσωπεύουν εξαιρέσεις (exception rules) και οι οποίοι καλύπτουν τις άλλες κλάσεις (π.χ., "εκτός εάν η συχνότητα(ACG) > 250 και η συχνότητα(AGT) < 200", αυτές οι αλληλουχίες είναι ασπόνδυλων).

Απ' την άλλη, ο PART αποτελεί μια έμμεση μέθοδο εξαγωγής κανόνων. Επεξεργάζεται τα δεδομένα, χρησιμοποιώντας τον ταξινομητή C4.5, που βασίζεται σε δένδρα (Quinlan 1993), δίνοντας ένα απλουστευμένο δέντρο απόφασης (decision tree) για κάθε επανάληψη. Τελικά, ο αλγόριθμος επιλέγει το καλύτερο δέντρο ταξινόμησης και χρησιμοποιεί τα φύλλα ως λογικούς κανόνες.

Σύμφωνα με τις πιο πάνω μεθόδους, η προσέγγιση LAF, που υιοθετούμε σε αυτήν την διατριβή, επιτελεί για κάθε αλληλουχία s σε ένα σύνολο δεδομένων, η οποία σχετίζεται (ανήκει) σε μια κλάση (π.χ. σπονδυλωτά, ασπόνδυλα, κ.ο.κ), τα ακόλουθα υπολογιστικά βήματα (δείτε και *Εικόνα 6*):

1. Υπολογίζεται η αντίστροφη συμπληρωματική κάθε αλληλουχίας s.
2. Οι αλληλουχίες s και οι αντίστροφες συμπληρωματικές τους s_r συνδυάζονται, με συνένωση, σε μια αλληλουχία S.
3. Στην αλληλουχία S, υπολογίζονται οι μετρήσεις των κ – μερών (για κ = 3, 4, 5) και παίρνουμε το διάνυσμα αναπαράστασης C = {$c_{k\text{-}mer1}$, $c_{k\text{-}mer2}$, …, $c_{k\text{-}mert}$}. Το πλήθος των κ – μερών εξάγεται με το λογισμικό Jellyfish (Marçais & Kingsford 2011). Το κ έχει επιλεγεί να λαμβάνει τιμές μεταξύ 3 και 5, με βάση τις αναφορές των (Teeling et al. 2004; Brendel et al. 1986). Οι ερευνητές αυτών των μελετών δηλώνουν ότι οι τιμές κ σε αυτό το εύρος ενδεχομένως να είναι οι βέλτιστες, όσον αφορά την ανάλυση αλληλουχιών, καθώς παρέχουν μια ιδανική ισορροπία μεταξύ του μήκους των υποακολουθιών και του αριθμού τους, όταν οι αλληλουχίες εκφράζονται με τη μορφή του αλφάβητου τεσσάρων γραμμάτων του DNA.
4. Υπολογίζονται οι συχνότητες για κάθε κ – μερές: $f_j = c_{κ\text{-}μερές j} / (ν\text{-}κ+1)$.

Εικόνα 6: Διάγραμμα ροής της μεθόδου LAF. Για περισσότερες πληροφορίες, δείτε στο κείμενο



5. Λαμβάνεται ένα διάνυσμα συχνοτήτων χαρακτηριστικών για κάθε αλληλουχία: $F = \{f_{κ\text{-μερές}1}, f_{κ\text{-μερές}2}, ..., f_{κ\text{-μερές}t}\}$.
6. Τα διανύσματα χαρακτηριστικών συνδυάζονται σε μια μήτρα δεδομένων (data matrix), όπου κάθε γραμμή αντιπροσωπεύει μια αλληλουχία και κάθε στήλη τη συχνότητα κ – μερούς. Μια επικεφαλίδα δηλώνει το όνομα των αλληλουχιών και την κλάση όπου ανήκουν. Ένα παράδειγμα της μήτρας δεδομένων δίνεται στον ακόλουθο πίνακα (*Πίνακας 3*).

Πίνακας 3: Παράδειγμα μήτρας δεδομένων για τη μέθοδο LAF. Η μήτρα δεδομένων αποτελείται από δύο στήλες επικεφαλίδων, η πρώτη με τα αναγνωριστικά (identifiers) των αλληλουχιών και η δεύτερη με τις ετικέτες κλάσεων. Οι ακόλουθες σειρές περιέχουν τις τιμές συχνοτήτων για τα κ – μερή στις αλληλουχίες.

	Seq 1	**Seq 2**	**….**	**Seq N-1**	**Seq N**
	Vertebrate	**Vertebrate**	**….**	**Invertebrate**	**Invertebrate**
AAA	0.46	0.26	…	0.24	0.54
AAC	0.12	0.16	…	0.23	0.24
AAG	0.12	0.23	…	0.23	0.23
….	…	…	…	….	…

7. Τα αριθμητικά σύνολα δεδομένων διακριτοποιούνται, δηλαδή οι συχνότητες μετατρέπονται από αριθμητικές σε κατηγορικές, σύμφωνα με τη μέθοδο MDL των Fayyad και Irani (Fayyad & Irani 1993). Η

διαδικασία διακριτοποίησης βελτιώνει την απόδοση των αλγορίθμων μηχανικής μάθησης, που βασίζονται σε λογικούς κανόνες.

8. Η μήτρα δεδομένων δίνεται, ως είσοδος, σε τρεις αλγορίθμους (εποπτευόμενης) μηχανικής μάθησης, που βασίζονται σε λογικούς κανόνες: RIPPER, RIDOR και PART.
9. Οι αλγόριθμοι ταξινόμησης εκτελούνται ύστερα με την τεχνική της διασταυρωμένης επικύρωσης (cross validation). Συνοπτικά, τα δεδομένα χωρίζονται σε ν ίσα μέρη και τα πειράματα επαναλαμβάνονται n φορές. Σε κάθε επανάληψη χρησιμοποιείται ένα διαφορετικό μέρος των δεδομένων για την αξιολόγηση του συστήματος, ενώ τα υπόλοιπα ν μέρη χρησιμοποιούνται ως δεδομένα εκπαίδευσης. Τα τελικά αποτελέσματα είναι ο μέσος όρος των αποτελεσμάτων από τις ν επαναλήψεις. Με τη μέθοδο αυτή χρησιμοποιούνται τελικά όλα τα δεδομένα τόσο για εκπαίδευση όσο και για αξιολόγηση, ενώ δεν υπάρχει περίπτωση το σύστημα να αξιολογηθεί σε δεδομένα, τα οποία έχουν χρησιμοποιηθεί ταυτόχρονα για την εκπαίδευσή του. Μια συνηθισμένη τιμή που παίρνει το ν είναι 10 και αυτή είναι που υιοθετούμε (10 – fold cross validation).
10. Στη συνέχεια εξάγονται τα μοντέλα ταξινόμησης, π.χ. «εάν συχνότητα(AAAC) < 0.195 τότε ο οργανισμός είναι σπονδυλωτός» ενώ «εάν συχνότητα(AAAC) > 0.195 τότε ο οργανισμός είναι ασπόνδυλος».
11. Τα μοντέλα ταξινόμησης αξιολογούνται, μετέπειτα, με βάση τα ποσοστά επιτυχούς ταξινόμησης.

Σε όλα τα υπολογιστικά πειράματα χρησιμοποιείται η σουίτα εργαλείων WEKA, καθώς και οι εκδοχές των αλγορίθμων ταξινόμησης, που βασίζονται σε λογικούς κανόνες, που είναι ενσωματωμένες σε αυτό το σύνολο προγραμμάτων (Hall et al. 2009).

Η μέθοδος των γονιδιωματικών υπογραφών (Genomic Signatures, GS)

Εφαρμόζουμε και μια άλλη, διαφορετική μεθοδολογία για την κατάταξη βιολογικών ακολουθιών μικρού μήκος, με σκοπό τη σύγκριση με τα αποτελέσματα που λαμβάνουμε με τη μέθοδο που προτείνουμε εδώ (LAF). Μια κλασική μέθοδος για την ποσοτικοποίηση των προτιμήσεων γειτόνων σε μια αλληλουχία DNA ενός οργανισμού βασίζεται στον υπολογισμό του διανύσματος των λόγων πιθανοτήτων για τα

δινουκλεοτίδια (Karlin 1998). Ο λόγος πιθανότητας για κάθε δινουκλεοτίδιο είναι η ποσότητα

$$\rho_{ij} = \frac{f_{ij}}{f_i f_j}$$

όπου f_{ij} και f_i, f_j δηλώνουν τις συχνότητες εμφάνισης, στις αλληλουχίες υπό μελέτη, ενός δινουκλεοτιδίου και των συστατικών του νουκλεοτιδίων αντίστοιχα. Επομένως οι δείκτες i και j αντιπροσωπεύουν κάθε ζεύγος A, G, C και T. Αυτή είναι η αναλογία της «παρατηρούμενης» συχνότητας δινουκλεοτιδίου προς την «αναμενόμενη», χωρίς προτίμηση γείτονα και έτσι εκφράζει τις πραγματικές προτιμήσεις γειτόνων του δεδομένου ζεύγους νουκλεοτιδίων. Προτού υπολογιστούν οι λόγοι πιθανοτήτων για μια δεδομένη αλληλουχία, αυτή (η αλληλουχία) ενώνεται με την αντίστροφη συμπληρωματική της. Συνεπώς, οι σχετικές αναλογίες είναι μόνο δέκα, δηλαδή τέσσερις για τα ομοίως συμπληρωματικά δινουκλεοτίδια και έξι για τα αμοιβαίως συμπληρωματικά ζεύγη. Ο Karlin και οι συνεργάτες του βρήκαν ότι αυτές οι ποσότητες διαφοροποιούνται μεταξύ διαφορετικών γονιδιωμάτων, σύμφωνα, προσεγγιστικά, με την εξελικτική τους απόσταση (Karlin & Mrázek 1997). Επομένως απέδωσαν στο διάνυσμα αυτών των δέκα «προτιμήσεων πρώτων γειτόνων» τον όρο Γονιδιωματική Υπογραφή (Genomic Signature, GS). Στα αποτελέσματα, που ακολουθούν, εφαρμόζουμε ταξινόμηση μέσω γονιδιωματικών υπογραφών, για άμεση σύγκριση με την ταξινόμηση μέσω της μεθόδου LAF, η οποία δε βασίζεται σε στοιχίσεις αλληλουχιών και περιγράφηκε στην προηγούμενη ενότητα. Θεωρούμε επίσης τους RIPPER, PART και RIDOR, ως ταξινομητές για όλα τα πειράματα, και πραγματοποιούμε τη διαδικασία της διακριτοποίησης, όπως την περιγράψαμε πριν για τη LAF, ομοίως και σε όλες αυτές τις συγκρίσεις. Όλα τα αρχεία (σε μορφή BED) που χρησιμοποιούμε βρίσκονται στο «Πρόσθετο Υλικό» (Supplementary Material) στη διεύθυνση dmb.iasi.cnr.it/cnes.php.

Αποτελέσματα και συζήτηση

Σε αυτήν την ενότητα παρουσιάζουμε μια συστηματική ανάλυση ταξινόμησης μικρών αλληλουχιών του γονιδιώματος, που επιδεικνύουν διαφορετικές λειτουργικότητες και

βρίσκονται στο ανθρώπινο και σε άλλα γονιδιώματα, χρησιμοποιώντας μια προσέγγιση εποπτευόμενης μηχανικής μάθησης (LAF). Για κάθε πείραμα ταξινόμησης υιοθετούμε τη μεθοδολογία της αναπαράστασης βιολογικών αλληλουχιών με διανύσματα χαρακτηριστικών και μετέπειτα ταξινόμησή τους με τη χρήση λογικών κανόνων, όπως περιγράφηκε στις προηγούμενες ενότητες. Κάθε πείραμα ταξινόμησης πραγματοποιείται χρησιμοποιώντας αναπαραστάσεις διανυσμάτων χαρακτηριστικών με κ – μερή μήκους 3, 4 και 5, τα οποία δίνονται ως είσοδοι σε τρεις διαφορετικούς αλγορίθμους που βασίζονται σε λογικούς κανόνες (RIPPER, RIDOR, PART), υιοθετώντας την τεχνική της 10πλής διασταυρωμένης επικύρωσης (10 – fold cross validation). Όλη η λίστα των αποτελεσμάτων ακρίβειας (accuracy) ταξινόμησης μαζί με όλες τις παραμέτρους συνοψίζονται στο «Συμπληρωματικό Excel» που είναι διαθέσιμο στο dmb.iasi.cnr.it/cnes.php. Παραλείπουμε τα ποσοστά των ψευδώς θετικών και ψευδώς αρνητικών, διότι τα σφάλφατα είναι ομοιογενώς κατανεμημένα στις διαφορετικές κλάσεις στα σύνολα δεδομένων. Αναφέρουμε εδώ και συζητάμε τα αποτελέσματά μας, βασισμένοι στο καλύτερο αποτέλεσμα (υψηλότερο ποσοστό επιτυχούς ταξινόμησης) μεταξύ κάθε διαφορετικού αλγορίθμου, που βασίζεται σε λογικούς κανόνες και εφαρμόζεται σε διανύσματα χαρακτηριστικών που προκύπτουν από χρήση κ – μερών μεγέθους 3, 4 και 5. Ο λόγος, που επιλέγουμε το συνολικά καλύτερο αποτέλεσμα γι' αυτό το σκοπό, είναι ότι θεωρούμε πως αντανακλά πιο άμεσα τη δυναμική της μεθόδου ταξινόμησης που βασίζεται σε κ – μερή και επομένως σχετίζεται, ενδεχομένως, με αρκετά βιολογικά χαρακτηριστικά και αποκαλύπτει λειτουργικές (ή πιθανά λειτουργικές) ιδιαιτερότητες των αλληλουχιών υπό μελέτη. Στο «Συμπληρωματικό Excel», οι ενδιαφερόμενοι αναγνώστες μπορούν να βρουν τους μέσους όρους των ποσοστών επιτυχούς ταξινόμησης, όπως υπολογίστηκαν βάση των 26 πειραμάτων, που πραγματοποιήθηκαν. Βασισμένοι στα αποτελέσματά μας προτείνουμε τιμή του κ = 3 και τον PART αλγόριθμο, ως ένα βέλτιστο συνδυασμό τιμής του κ και αλγορίθμου ταξινόμησης αντίστοιχα, για ένα *ab initio* πείραμα ταξινόμησης (δείτε «Συμπληρωματικό Excel»). Για όλα τα πειράματα που πραγματοποιήθηκαν, εφαρμόσαμε ξεχωριστά τη μέθοδο των γονιδιωματικών υπογραφών (GS), ως μια εναλλακτική της μεθόδου LAF, που προτείνουμε εδώ. Οι GS, δίνοντας χαρακτηριστικές «υπογραφές σύστασης» για διαφορετικά γονιδιώματα, αντιπροσωπεύουν μια κλασική μέθοδο, με βάση την οποία θέλαμε να συγκρίνουμε τη LAF. Και στην περίπτωση των GS, λαμβάνεται υπόψη το καλύτερο, συνολικά, αποτέλεσμα, όσον αφορά τον αλγόριθμο που χρησιμοποιείται.

Οι ακόλουθες αναλύσεις ταξινόμησης παρουσιάζονται, προκειμένου να εκτιμήσουμε την πιθανή χρησιμότητα της LAF, στο διαχωρισμό μεταξύ ποικίλων γονιδιωματικών στοιχείων, που βρίσκονται στο ίδιο ή σε διαφορετικά γονιδιώματα μετάζωων. Οι κανόνες ταξινόμησης, που εξάγονται με τη μέθοδο LAF, όπως αυτή περιγράφεται πιο πάνω στις «Μεθόδους», είναι διαθέσιμοι στην τοποθεσία dmb.iasi.cnr.it/cnes.php. Ο ενδιαφερόμενος αναγνώστης παραπέμπεται εκεί, όπου δίνονται τα χαρακτηριστικά κ – μερή για κάθε πείραμα ανά ζεύγη και κλάση των συνόλων δεδομένων που μελετώνται.

Συγκρίσεις αλληλουχιών υποβάθρου μεταξύ ειδών

Σε αυτές τις συγκρίσεις μεταξύ αλληλουχιών υποβάθρου ανθρώπου, σκώληκα και εντόμου, χρησιμοποιούμε, ως αντιπροσωπευτικά δείγματα των διαφορετικών γονιδιωμάτων, αναπληρωματικά σύνολα δεδομένων (surrogate sets), που λαμβάνονται από το εκάστοτε γονιδίωμα, έχοντας ως σημείο αναφοράς, είτε τα εξώνια, είτε τα CNE. Στις συγκρίσεις ανά ζεύγη, όπου περιλαμβάνεται ο *H. sapiens*, παίρνουμε πάντα τα καλύτερα ποσοστά ταξινόμησης, χρησιμοποιώντας και τις δύο μεθόδους (LAF και GS, δείτε *Πίνακα 4*). Αυτό το εύρημα μπορεί να κατανοηθεί, στα πλαίσια της υψηλής διαφοράς στις προτιμήσεις γειτόνων, κυρίως στα CpG και TpA, μεταξύ *H. sapiens* και ασπόνδυλων, ενώ αυτές οι προτιμήσεις βρίσκονται να είναι αρκετά κοντά σε είδη ασπόνδυλων: *D. melanogaster* (έντομο) και *C. elegans* (σκώληκας). Η GS συνιστά αποκλειστικά μια ποσοτικοποίηση της προτίμησης πρώτων γειτόνων, ενώ στη LAF συμπεριλαμβάνονται και διάφορες άλλες συνιστώσες σύστασης αλληλουχίας, όπως η σύσταση μονονουκλεοτιδίων και οι προτιμήσεις γειτόνων ανώτερης τάξης (πέρα της πρώτης). Σε συμφωνία με την πιο πάνω περιγραφή, στις περιπτώσεις όπου περιλαμβάνονται ανθρώπινες αλληλουχίες στη σύγκριση, οι GS αποδίδουν συστηματικά καλύτερα από τη LAF, ενώ η τελευταία αποδίδει καλύτερα στις εναπομείνουσες δύο περιπτώσεις.

Ένα άλλο συμπέρασμα, που προκύπτει από απλή επιθεώρηση του *Πίνακα 4*, είναι ότι στις συγκρίσεις γονιδιωματικών αλληλουχιών υποβάθρου μεταξύ ειδών, τα καλύτερα αποτελέσματα παρατηρούνται για τις εξωνικές αναπληρωματικές αλληλουχίες (exonic surrogates), σε σύγκριση με τις αναπληρωματικές αλληλουχίες CNE (CNE surrogates). Αυτό εξηγείται από το γεγονός ότι τα μέσα μήκη των εξωνίων, σε όλες τις περιπτώσεις, υπερβαίνουν τα μέσα μήκη των CNE (και το ίδιο ισχύει για τις

αναπληρωματικές τους αλληλουχίες, από «κατασκευής»). Τα υψηλότερα μέσα μήκη επιτρέπουν ξεκάθαρα καλύτερη στατιστική και επομένως υψηλότερα ποσοστά επιτυχούς ταξινόμησης και για τις δύο μεθόδους (LAF και GS). Σημειώνουμε εδώ ότι, όταν πρόκειται για συγκρίσεις μεταξύ CNE διαφορετικών ειδών, τροποποιούμε τις αλληλουχίες CNE των σπονδυλωτών, έτσι ώστε να φτάσουν στο μέσο μήκος των CNE ασπονδύλων (τα οποία είναι μικρότερα, για λεπτομέρειες δείτε στις «Μεθόδους»).

Πίνακας 4: Σύγκριση γονιδιωματικών αλληλουχιών υποβάθρου από διαφορετικούς οργανισμούς. Συγκρίσεις μεταξύ ζευγών συλλογών αλληλουχιών, που αποτελούνται από 1000 αλληλουχίες υποβάθρου η καθεμιά (όχι CNE, όχι εξωνικές), από 3 διαφορετικά γονιδιώματα. Αυτές οι αλληλουχίες λαμβάνονται ως αναπληρωματικές για τα CNE και τα εξώνια, που μελετούμε στη συνέχεια. Κάθε τέτοια αλληλουχία έχει το ίδιο μήκος και ποσοστό GC% με ένα μέλος από κάθε συλλογή CNE ή εξωνίων. Η υψηλότερη τιμή ακρίβειας ταξινόμησης για τη LAF και τη GS, σε κάθε πείραμα ταξινόμησης, δείχνεται με έντονη εκτύπωση. Για περισσότερες πληροφορίες πάνω στις χρησιμοποιούμενες μεθόδους, δείτε στο κείμενο και στο «Συμπληρωματικό υλικό».

Experiment No	Description	Average length	Average GC	LAF, accuracy	GS, accuracy
#1	surrogate for human exons surrogates for worm exons	167.84 169.32	0.40 0.52	84.21	**86.93**
#14	surrogates for humanUCNEs surrogates for insect UCNEs	86.09 86.58	0.36 0.39	81.45	**84.30**
#20	surrogates for human exons surrogates for insect exons	169.82 169.32	0.51 0.52	84.66	**87.89**
#22	surrogates for worm exons surrogates for insect exons	213.37 212.86	0.42 0.52	**78.60**	70.95
#23	surrogates for worm UCNEs surrogates for hUCNEs	83.18 82.93	0.43 0.36	82.10	**84.48**
#13	surrogates for worm UCNEs surrogates for insect UCNEs	83.41 86.58	0.43 0.39	**64.70**	63.65
Average				79.29	**79.70**

Συγκρίση αλληλουχιών που βρίσκονται υπό επιλεκτικό περιορισμό έναντι των αντίστοιχων αναπληρωματικών τους αλληλουχιών

Τα ακόλουθα πειράματα αναφέρονται σε συγκρίσεις αλληλουχιών που βρίσκονται υπό επιλεκτικό περιορισμό (constrained sequences) έναντι των αναπληρωματικών τους αλληλουχιών (*Πίνακας 5*). Σημειώνουμε ότι οι αναπληρωματικές αλληλουχίες (surrogates) μοιράζονται με τις αρχικές αλληλουχίες το ίδιο ποσοστό σε GC% και μήκος (δείτε «Μεθόδους»). Όπως αποδεικνύεται, οι συγκρίσεις που περιλαμβάνουν αλληλουχίες ασπόνδυλων, που βρίσκονται υπό επιλεκτικό περιορισμό, δεν ταξινομούνται με τόσο μεγάλη επιτυχία όσο οι αντίστοιχες ανθρώπινες, χρησιμοποιώντας την LAF. Το συγκεκριμένο εύρημα μπορεί να κατανοηθεί στα πλαίσια αρκετών ιδιαιτεροτήτων των θερμόαιμων ζώων (και συνήθως όλων των

σπονδυλωτών), ιδιαίτερα όσον αφορά το μη λειτουργικό, μη συντηρημένο τμήμα του γονιδιώματος υποβάθρου. Αυτές περιλαμβάνουν υψηλό εμπλουτισμό σε επαναλαμβανόμενες αλληλουχίες. Επιπλέον, τα γονιδιώματα των σπονδυλωτών παρουσιάζουν ένα τυπικό προφίλ σπάνιων δινουκλεοτιδίων (ειδικά CpG και TpA) τα οποία αποφεύγονται λιγότερο στις αλληλουχίες που υπόκεινται σε επιλεκτική πίεση (εξώνια και CNE). Αυτές οι αλληλουχίες, έχοντας λειτουργικούς ρόλους, δεν ακολουθούν αυστηρά τις τάσεις του υπόλοιπου γονιδιώματος όσον αφορά τη σύσταση. Τα γονιδιώματα ασπόνδυλων δεν περιέχουν πολλές επαναλαμβανόμενες αλληλουχίες και επιπλέον οι υποεκπροσωπήσεις συγκεκριμένων δινουκλεοτιδίων δεν είναι τόσο συχνές. Στις συγκρίσεις με τη μέθοδο των γονιδιωματικών υπογραφών ακολουθείται η ίδια τάση, αλλά το ποσοστό επιτυχούς ταξινόμησης είναι σημαντικά ελαττωμένο, λόγω της απλούστερης δομής αυτού του μέτρου σύγκρισης γονιδιωματικών αλληλουχιών. Γνωρίζουμε, μάλιστα, από προηγούμενες μελέτες ότι οι γονιδιωματικές υπογραφές δεν τα πηγαίνουν καλά σε συγκρίσεις στοιχείων που προέρχονται από το ίδιο γονιδίωμα, διότι οι προτιμήσεις γειτόνων παραμένουν σχετικά σταθερές μέσα στο ίδιο γονιδίωμα. Γι'αυτόν το λόγο, σε όλες τις συγκρίσεις που παρουσιάζονται στον πίνακα πιο κάτω, η LAF αποδεικνύεται καλύτερη.

Οι τελευταίες έξι σειρές του *Πίνακα 5* δείχνουν συγκρίσεις διαφόρων συλλογών CNE αλληλουχιών έναντι αναπληρωματικών αλληλουχιών (CNE surrogates). Γενικά παρατηρούμε ότι, μεταξύ των αλληλουχιών που βρίσκονται υπό επιλεκτική πίεση, οι ανθρώπινες αλληλουχίες CNE έναντι των δικών τους αναπληρωματικών αλληλουχιών επιδεικνύουν συγκριτικά υψηλότερα ποσοστά επιτυχούς ταξινόμησης, εάν συγκριθούν με τις εξωνικές αλληλουχίες έναντι των δικών τους αναπληρωματικών αλληλουχιών. Παρόλο που δεν είμαστε σε θέση να ερμηνεύσουμε πλήρως αυτό το εύρημα, είναι πιθανό να σχετίζεται με την υπόθεση ότι τα CNE ρυθμίζουν πολύπλοκες αναπτυξιακές διαδικασίες όπως είναι ο σχηματισμός του εγκεφάλου (Matsunami & Saitou 2013). Οι οδηγίες που ρυθμίζουν αυτές τις διαδικασίες μπορεί να είναι κωδικοποιημένες στις ίδιες τις αλληλουχίες. Έχει δειχθεί στη βιβλιογραφία ότι ορισμένες κατηγορίες CNE δρουν ως περιοχές πρόσδεσης μεταγραφικών παραγόντων και επιπλέον φέρουν διάφορα μοτίβα (Xie et al. 2007; Viturawong et al. 2013), τα οποία η ανάλυση κ – μερών είναι πιθανόν αρκετά ευαίσθητη ώστε να εντοπίζει και αυτό αντικατοπτρίζεται στα υψηλότερα ποσοστά επιτυχούς ταξινόμησης των CNE εναντίον όμοιων με CNE (CNE-like) αλληλουχιών σε σχέση με τα εξώνια εναντίον των όμοιων με εξώνια (exon-like) αλληλουχιών.

Πίνακας 5: Σύγκριση αλληλουχιών που βρίσκονται υπό επιλεκτικό περιορισμό έναντι των αντίστοιχων αναπληρωματικών τους. Συγκρίσεις μεταξύ ζευγών συλλογών αλληλουχιών CNE ή εξωνίων έναντι των αντίστοιχων συμπληρωματικών τους, μέσα στο ίδιο γονιδίωμα. Η υψηλότερη τιμή ακρίβειας ταξινόμησης για τη LAF και τη GS, σε κάθε πείραμα ταξινόμησης, δείχνεται με έντονη εκτύπωση. Εδώ τα CNE του σκώληκα μελετώνται στο σύνολό τους έναντι των αναπληρωματικών τους (1869 αλληλουχίες). Για περισσότερες πληροφορίες πάνω στις χρησιμοποιούμενες μεθόδους, δείτε στο κείμενο και στο «Συμπληρωματικό υλικό».

Experiment No	Description	Average length	Average GC	LAF, accuracy	GS, accuracy
#2	worm exons	213.37	0.42	**65.63**	63.75
	surrogates	213.37	0.42		
#3	human exons	169.82	0.52	**73.31**	65.81
	surrogates	169.82	0.52		
#17	insect exons	388.82	0.54	**63.72**	61.95
	surrogates	381.56	0.54		
#4	worm UCNEs	82.88	0.43	**61.00**	57.71
	surrogates	82.88	0.43		
#5a	human UCNEs	326.92	0.37	**81.15**	73.55
	surrogates	326.92	0.37		
#5b	human EU100nx CNEs	155.50	0.38	**75.95**	65.15
	surrogates	155.50	0.38		
#5c	amniotic CNEs	289.06	0.38	**76.25**	66.10
	surrogates	289.06	0.38		
#5d	mammalian CNEs	246.49	0.40	**73.65**	60.00
	surrogates	246.49	0.40		
#12	insect UCNEs	86.58	0.39	**67.95**	66.95
	surrogates	86.58	0.39		
Average				**70.96**	64.55

Συγκρίση αλληλουχιών που βρίσκονται υπό επιλεκτικό περιορισμό μέσα στο ίδιο είδος (λειτουργικές αλληλουχίες έναντι λειτουργικών αλληλουχιών διαφορετικού τύπου)

Η περίπτωση των εξωνίων έναντι των UCNE: Προχωρούμε στη μελέτη χαρακτηριστικών αλληλουχίας στοιχείων, τα οποία έχουν χαρακτηριστεί ως λειτουργικά: εξώνια, τα οποία είναι γνωστό πως βρίσκονται κάτω από επιλεκτική πίεση και κωδικοποιούν πολυπεπτιδικές αλυσίδες, καθώς και CNE, τα οποία είναι άγνωστης, εν πολλοίς, λειτουργικότητας, η οποία όμως υπονοείται από τον υψηλό βαθμό συντήρησης μεταξύ δύο ή περισσότερων οργανισμών. Με μια πρώτη ματιά, οι διαφοροποιήσεις στα ποσοστά ταξινόμησης μεταξύ LAF και GS, όπως συνοψίζονται στον *Πίνακα 6*, ενδεχομένως σχετίζονται με την ποσόστωση στο περιεχόμενο σε GC μεταξύ των αναλυόμενων κλάσεων. Πράγματι, είναι γνωστό από τη βιβλιογραφία ότι τα CNE είναι γενικά πλούσια σε AT (Walter et al. 2005), ενώ τα γονίδια που κωδικοποιούν πρωτεΐνες είναι σχετικά πλούσια σε GC (δείτε επίσης τιμές μέσου περιεχομένου σε GC στην τέταρτη στήλη του *Πίνακα 6*). Αυτό, ωστόσο, ενδέχεται να

μην είναι απόλυτα σωστό, καθώς η επιτυχία ταξινόμησης στην περίπτωση των εξωνίων του σκώληκα σε σχέση με τα UCNE του σκώληκα είναι μέτρια (δείτε πείραμα #6, 68.65%), ενώ η διαφορά στο ποσοστό GC είναι ελάχιστη. Το γεγονός αυτό καταδεικνύει, ενδεχομένως, ότι οι λειτουργικές διαφορές των εξωνίων και των CNE είναι κωδικοποιημένες στη σύσταση της αλληλουχίας τους. Τέτοιες τροπικότητες συνιστούν οι προτιμήσεις στη χρήση κωδικονίων και οι συχνότητες των αμινοξέων σε κωδικοποιούσες αλληλουχίες ή η παρουσία διαφόρων υποακολουθιών πρόσδεσης πρωτεϊνικών παραγόντων και άλλων συναινετικών (consensus) αλληλουχιών που βρίσκονται στα CNE.

Πίνακας 6: Συγκρίσεις διαφορετικών τύπων αλληλουχιών που βρίσκονται υπό επιλεκτικό περιορισμό. Συγκρίσεις εντός του ίδιου γονιδιώματος, μεταξύ συλλογών CNE έναντι συλλογών εξωνίων, αποτελούμενες από 1000 αλληλουχίες η καθεμιά. Με έντονη εκτύπωση φαίνεται η υψηλότερη τιμή, για κάθε πείραμα ταξινόμησης, για κάθε μέθοδο (LAF: Logic Alignment-Free, GS: Genomic Signature). Για περισσότερες πληροφορίες δείτε και στο κείμενο.

Experiment No	Description	Average length	Average GC	LAF, accuracy	GS, accuracy
#6	worm exons	82.64	0.42	**68.65**	61.88
	worm UCNEs	83.11	0.43		
#7	human exons	169.82	0.52	**82.79**	77.66
	human UCNEs	169.32	0.37		
#18	insect exons	86.09	0.52	**82.80**	81.90
	insect UCNEs	86.58	0.39		
Average				**78.08**	73.81

Η περίπτωση σύγκρισης διαφορετικών κατηγοριών CNE: Εδώ συγκρίνουμε διαφορετικές κατηγορίες CNE, που χαρακτηριστικό τους είναι ότι έχουν εξαχθεί με τα ίδια κριτήρια αλλά με σταδιακά αυξανόμενο κατώφλι ελάχιστης ομοιότητας κάθε φορά, χρησιμοποιώντας ένα βήμα 5 ποσοστιαίων μονάδων (δείτε «Μεθόδους» στο τμήμα «***Ταυτοποίηση νέων συνόλων δεδομένων CNE σε σπονδυλωτά, έντομα και σκώληκες***»). Σημειώνεται εδώ ότι η διαφοροποίηση βάσει σύστασης σε GC μεταξύ αυτών των συνόλων δεδομένων είναι χαμηλή. Παρατηρούμε καλά ποσοστά επιτυχούς ταξινόμησης χρησιμοποιώντας τη μέθοδο LAF, σαφώς καλύτερα από αυτά που δίνουν οι GS. Πιο συγκεκριμένα, λαμβάνουμε ιδιαίτερα υψηλές τιμές ακρίβειας όταν στις συγκρίσεις περιλαμβάνεται η κλάση των UCNE (υπερσυντηρημένες αλληλουχίες με 95-100% ομοιότητα μεταξύ ανθρώπου – κοτόπουλου, πειράματα #27, #30, #32, #33, #5a). Τα UCNE εικάζεται ότι ενδεχομένως απαρτίζουν μια ξεχωριστή κατηγορία, διακριτή από CNE άλλων συνόλων δεδομένων, που περιλαμβάνονται σε συγκρίσεις στον *Πίνακα 7*, και τα οποία έχουν ταυτοποιηθεί χρησιμοποιώντας λιγότερο αυστηρά κριτήρια ομοιότητας αλληλουχίας στους ίδιους οργανισμούς (άνθρωπο – κοτόπουλο).

Επιπλέον, όταν συγκρίνουμε τα CNE 75-80%, CNE 85-90% και CNE 95-100% με τις αναπληρωματικές τους αλληλουχίες στο ανθρώπινο γονιδίωμα, παρατηρούμε μια σταδιακή αύξηση στην απόδοση (συγκρίνετε #34 και #35 με το #5a). Συνολικά λαμβάνουμε τα υψηλότερα ποσοστά επιτυχούς ταξινόμησης όταν περιλαμβάνονται UCNE (Πείραμα #5a). Το πληροφοριακό περιεχόμενο που υπάρχει σε αυτές τις αλληλουχίες τις καθιστά ανιχνεύσιμες από την μέθοδό μας (LAF) με υψηλή ακρίβεια. Η διαφορετική προσέγγιση των γονιδιωματικών υπογραφών (GS) επίσης δίνει καλά αποτελέσματα, γεγονός το οποίο υποδηλώνει πως πράγματι αυτές οι αλληλουχίες (UCNE) σχηματίζουν μια ξεχωριστή κατηγορία υπερσυντηρημένων αλληλουχιών και υπό το πρίσμα των προτιμήσεων «πρώτων γειτόνων».

Πίνακας 7: Συγκρίσεις διαφορετικών τύπων αλληλουχιών CNE, που έχουν ταυτοποιηθεί χρησιμοποιώντας τα ίδια κριτήρια, αλλά με διαφορετικά κατώφλια ελάχιστης ομοιότητας. Συγκρίσεις, εντός του ίδιου γονιδιώματος, μεταξύ ζευγών συλλογών ανθρώπινων CNE (από 1000 CNE η καθεμιά κλάση). Αυτά τα σύνολα δεδομένων συνίστανται από αλληλουχίες που έχουν ταυτοποιηθεί μεταξύ στοιχίσεων ανθρώπου / κοτόπουλου με σταδιακά αυξανόμενο κατώφλι ελάχιστης ομοιότητας. Με έντονη εκτύπωση φαίνεται η υψηλότερη τιμή, για κάθε πείραμα ταξινόμησης, για κάθε μέθοδο (LAF: Logic Alignment-Free, GS: Genomic Signature). Για περισσότερες πληροφορίες δείτε και στο κείμενο.

Experiment No	Description	Average length	Average GC	LAF, accuracy	GS, accuracy
#24	CNEs 75-80% CNEs 80-85%	248.92 248.39	0.39 0.38	**55.45**	53.95
#25	CNEs 75-80% CNEs 85-90%	248.92 250.63	0.39 0.38	**57.45**	50.00
#26	CNEs 75-80% CNEs 90-95%	248.92 268.64	0.39 0.37	**60.95**	59.05
#27	CNEs 75-80% CNEs 95-100%	248.92 319.70	0.39 0.37	**68.50**	63.25
#28	CNEs 80-85% CNEs 85-90%	248.39 250.63	0.38 0.38	**50.00**	**50.00**
#29	CNEs 80-85% CNEs 90-95%	248.39 268.64	0.38 0.37	**58.90**	54.30
#30	CNEs 80-85% CNEs 95-100%	248.39 319.70	0.38 0.37	**68.60**	61.80
#31	CNEs 85-90% CNEs 90-95%	250.63 268.64	0.38 0.37	**58.00**	50.00
#32	CNEs 85-90% CNEs 95-100%	250.63 319.70	0.38 0.37	**65.65**	59.30
#33	CNEs 90-95% CNEs 95-100%	268.64 319.70	0.37 0.37	**61.55**	50.00
#34	CNEs 75-80% surrogates	248.92 248.92	0.39 0.39	**76.20**	64.10
#35	CNEs 85-90% surrogates	250.63 250.63	0.38 0.38	**78.45**	65.25
#5a	CNEs 95-100% surrogates	326.92 326.92	0.37 0.37	**81.15**	73.55
Average				**64.68**	58.04

Η ταξινόμηση, που βασίζεται στην ανάλυση κ – μερών και σε ταξινομητές λογικών κανόνων, αποδεικνύεται αποτελεσματική στο διαχωρισμό αλληλουχιών από το ίδιο είδος (ενδοειδικές συγκρίσεις)

Οι γονιδιωματικές υπογραφές (GS) χρησιμοποιούνται ευρέως στην ταυτοποίηση διαφορετικών ειδών (Karlin & Mrázek 1997). Χρησιμοποιώντας μια γραφική αναπαράσταση της σύστασης του γονιδιώματος, που ονομάζεται Chaos Game Representation (CGR), οι Deschavanne και συνεργάτες ανέφεραν ότι οι υποακολουθίες ενός γονιδιώματος επιδεικνύουν τα κύρια χαρακτηριστικά ολόκληρου του γονιδιώματος, συγκλίνοντας έτσι στην ιδέα της γονιδιωματικής υπογραφής (Deschavanne et al. 1999). Οι συγγραφείς της συγκεκριμένης μελέτης θεώρησαν ότι μπορούν να περιγράψουν τα κύρια χαρακτηριστικά της ποικιλομορφίας, που παρατηρείται μεταξύ αλληλουχιών μέσω συγκεντρώσεων / ποσοστώσεων βάσεων, διαδοχής συμπληρωματικών βάσεων ή πουρινών / πυριμιδινών, καθώς και από τις υπερ- ή υποεκπροσωπήσεις λέξεων ποικίλων μηκών, με κύριο σκοπό να μετρήσουν τη φυλογενετική εγγύτητα ειδών. Πιο κάτω προχωράμε σε μια κατηγοριοποίηση των πειραμάτων μας σε διαειδικά και ενδοειδικά, βασισμένοι στο εάν οι αλληλουχίες που μελετούμε προέρχονται από διαφορετικούς ή από τον ίδιο οργανισμό αντίστοιχα. Παρουσιάζουμε στον *Πίνακα 8* συγκεντρωτικά τα αποτελέσματα και για τα 26 πειράματα. Φαίνεται ότι η μέθοδος που προτείνουμε (LAF) αποδεικνύεται εξίσου καλή με τις γονιδιωματικές υπογραφές (GS), όταν έχουμε να κάνουμε με αλληλουχίες που προέρχονται από διαφορετικά είδη (80.20% έναντι 80.19% για τις GS). Όταν όμως συγκρίνουμε αλληλουχίες που προέρχονται από τον ίδιο οργανισμό, η LAF τα πηγαίνει σημαντικά καλύτερα σε σχέση με τις GS (70.90% έναντι 65.83%). Αυτό το εύρημα καταδεικνύει ότι η ανάλυση κ – μερών σχετικά μικρών αλληλουχιών, σε συνδυασμό με ταξινομητές που βασίζονται σε λογικούς κανόνες, είναι μια υποσχόμενη μέθοδος, καθώς η LAF λαμβάνει υπόψη, επιπλέον των προτιμήσεων πρώτων γειτόνων, την ύπαρξη στις υπό μελέτη αλληλουχίες διαφόρων συνδυασμών ολιγονουκλεοτιδίων, όπως ομο-πουρίνες και ομο-πυριμιδίνες και παλίνδρομων αλληλουχιών, οι οποίες φαίνεται πως υπερεκπροσωπούνται σε συγκεκριμένα σύνολα δεδομένων CNE (μη δημοσιευμένα αποτελέσματα από το εργαστήριο του Philipp Bucher). Άλλες τέτοιες χαρακτηριστικές «υπογραφές», που έχουν προταθεί στη βιβλιογραφία για τα CNE και που ενδεχομένως ανιχνεύονται από τη μέθοδο LAF που προτείνουμε, είναι η παρουσία στα CNE διαφόρων επικαλυπτόμενων περιοχών πρόσδεσης μεταγραφικών παραγόντων (Viturawong et al. 2013). Η ύπαρξη αυτών των περιοχών είναι πιθανόν να διαφοροποιεί

συστηματικά τις συγκεκριμένες αλληλουχίες από τις υπόλοιπες μη-συντηρημένες αλληλουχίες.

Πίνακας 8: Γενικά στατιστικά στοιχεία που περιγράφουν την αποτελεσματικότητα της LAF. LAF (Best): Ο καλύτερος συνδυασμός για κάθε πείραμα, όσον αφορά την τιμή του k και του ταξινομητή που χρησιμοποιείται, παρουσιάζεται και λαμβάνεται υπόψη για τον υπολογισμό του αριθμητικού μέσου, LAF (average of 9): Αριθμητικός μέσος όλων των διαφορετικών συνδυασμών για κάθε πείραμα κ – μερών και ταξινομητών που βασίζονται σε λογικούς κανόνες, LAF (κ = 3, PART): Με βάση όλα τα πειράματα που πραγματοποιήθηκαν, ο συγκεκριμένος, πιο κατάλληλος συνδυασμός τιμής κ και αλγορίθμου ταξινόμησης, επιλέγεται για την εξαγωγή συμπερασμάτων σε κάθε πείραμα, GS (without discretization): οι γονιδιωματικές υπογραφές όπως έχουν υπολογιστεί ως διάνυσμα συχνοτήτων χωρίς το επιπλέον βήμα διακριτοποίησης που κάνουμε στη LAF και περιγράφεται πιο πάνω.

	LAF, accuracy, Best	LAF, accuracy, average of 9	LAF, accuracy, k = 3, PART	GS, accuracy, without discretization	GS, accuracy, Best, with discretization	GS, accuracy, mean of 3, with discretization
Overall	**75.19**	71.33	74.39	70.86	72.45	71.78
Inter-species	**80.20**	75.56	79.81	79.62	80.19	79.53
Intra-Species	**70.90**	67.69	69.74	63.35	65.83	65.14

Συμπεράσματα

Παρουσιάζουμε και εφαρμόζουμε μια προσέγγιση, τη LAF (Logic Alignment Free), που βασίζεται στην ανάλυση κ – μερών και στην ταξινόμηση με χρήση λογικών κανόνων, με σκοπό την αναπαράσταση και τη μετέπειτα κατάταξη βιολογικών αλληλουχιών διαφορετικών λειτουργικοτήτων καθώς και διαφορετικής προέλευσης. Συγκρίνουμε την προσέγγισή μας *vis-à-vis* με τη μέθοδο των γονιδιωματικών υπογραφών (GS) του Karlin και παρουσιάζουμε ποσοστά ταξινόμησης. Από όσο γνωρίζουμε, η μελέτη μας είναι η πρώτη που εφαρμόζει αυτές τις μεθοδολογίες σε μικρές βιολογικές αλληλουχίες, καθώς μέχρι τώρα οι GS είχαν χρησιμοποιηθεί μόνο σε περιπτώσεις διαχωρισμού περισσότερο εκτεταμένων περιοχών του γονιδιώματος, μεγέθους 50 kb ή και περισσότερο (Karlin 1998). Βασισμένοι στα αποτελέσματά μας, συμπεραίνουμε ότι η LAF αποδίδει ιδιαιτέρως καλά, όταν μελετούμε αλληλουχίες από το ίδιο είδος, ξεπερνώντας την απόδοση των GS, ενώ σε συγκρίσεις μεταξύ ειδών, όπου η δυναμική των GS έχει ήδη επιβεβαιωθεί μέσω της σύγκρισης μεγάλων γονιδιωματικών περιοχών, οι δύο μέθοδοι αποδίδουν εξίσου καλά. Επομένως, προτείνουμε ότι η LAF μπορεί να χρησιμοποιηθεί περαιτέρω σε εφαρμογές που ενέχουν ανάλυση αλληλουχίας, όπως η κατηγοριοποίηση διαφορετικών, πιθανών,

λειτουργικοτήτων μέσα σε ομάδες αλληλουχιών CNE ή η ταυτοποίηση και ο διαχωρισμός διαφορετικών συνιστωσών του μεταγονιδιώματος.

3.2 Ανάλυση και ταξινόμηση αλληλουχιών DNA, που βρίσκονται υπό επιλεκτικό περιορισμό, με τη χρήση γράφων (γραφημάτων) N – γραμμάτων (N – gram Graphs, NGG) και γονιδιωματικών υπογραφών (GS)

Αναλύοντας τη σύσταση αλληλουχίας μέσω ενός συνδυασμού περιεχομένου λέξεων και πληροφορίας σχετικής θέσης

Μια αλληλουχία μπορεί να θεωρηθεί ισοδύναμη με ένα κείμενο φυσικής γλώσσας, υπό την προϋπόθεση ότι το λεξιλόγιο είναι πολύ περιορισμένο. Παραδοσιακά, μέθοδοι επεξεργασίας φυσικής γλώσσας που βασίζονται σε n – γράμματα (n – νουκλεοτίδια αντίστοιχα) έχουν εφαρμοσθεί σε βιολογικές αλληλουχίες, στοχεύοντας στη βελτιστοποίηση μεθόδων αντιστοίχισης αλληλουχίας (Kim & Shawe-Taylor 1994), στη δημιουργία ευρετηρίων (indexing) (Kim et al. 2005), στην ανάλυση πρωτεϊνικών αλληλουχιών (Ganapathiraju et al. 2002), καθώς και στην ανάλυση κωδικοποιουσών και μη αλληλουχιών (Mantegna et al. 1995). Τα μοντέλα n – γραμμάτων αλληλουχιών δείχνουν πώς μικρές υποακολουθίες μήκους n εμφανίζονται σε όλη την αλληλουχία υπό ανάλυση. Άλλες εναλλακτικές μέθοδοι ανάλυσης, όπως τα HMM (Hidden Markov Models) ή τα CRM (Conditional Random Fields), μοντελοποιούν πιθανοτικά τους πιθανούς συνδυασμούς στοιχείων σε μια αλληλουχία.

Στις επόμενες παραγράφους, προτείνουμε και περιγράφουμε την εφαρμογή της μεθοδολογίας αναπαράστασης των γράφων / γραφημάτων N – γραμμάτων (NGG), η οποία καταφέρνει να αποτυπώσει τοπικά και καθολικά χαρακτηριστικά των αναλυόμενων αλληλουχιών (Giannakopoulos G. Vouros G. and Stamatopoulos, P. 2008; Polychronopoulos, Krithara, et al. 2014). Η κύρια ιδέα πίσω από τους γράφους είναι ότι η γειτονιά μεταξύ των υποακολουθιών σε μια αλληλουχία περιέχει ένα κρίσιμο μέρος της ίδιας της πληροφορίας της αλληλουχίας. Ο γράφος N – γραμμάτων (NGG), όπως προκύπτει από κάθε αλληλουχία, είναι αναγκαστικά ένα ιστόγραμμα συνεντοπισμών / συνεμφανίσεων συμβόλων (στην περίπτωσή μας, νουκλεοτιδίων). Τα σύμβολα θεωρούνται ότι συνεντοπίζονται όταν βρίσκονται σε μια μέγιστη απόσταση (παράθυρο, window) μεταξύ τους. Το μέγεθος του παραθύρου, το οποίο αποτελεί μια

παράμετρο των NGG, επιτρέπει ευελιξία στην αναπαράσταση των συνεντοπισμών μέσα σε μια αλληλουχία. Το γεγονός ότι οι γράφοι N – γραμμάτων λαμβάνουν υπόψη συνεντοπισμούς προσφέρει μια δυναμική περιγραφής της τοπικότητας, ενώ το γεγονός ότι δρουν ως ένα ιστόγραμμα αυτών των συνεντοπισμών παρέχει μια δυναμική καθολικής αναπαράστασης του συνόλου της αλληλουχίας. Σημειώνουμε εδώ ότι, στο όριο του μεγέθους του παραθύρου (άπειρο μέγεθος παραθύρου), οι NGG χάνουν τη δυνατότητα περιγραφής της τοπικότητας.

Σε αντίθεση με τα πιθανοτικά μοντέλα (όπως τα HMM), οι NGG είναι ντετερμινιστικοί. Επιπλέον, οι NGG δε βασίζονται σε πολυάριθμα παραδείγματα ώστε να συνάγουν τις παραμέτρους του μοντέλου. Τρίτον, μεταχειρίζονται ισότιμα, φαινόμενα που υποεκπροσωπούνται, κάτι το οποίο μας απαλλάσσει από τις προτιμήσεις των πιθανοτικών προσεγγίσεων για συχνά απαντώμενα πρότυπα. Σε αντίθεση με τα μοντέλα N – γραμμάτων, οι NGG προσφέρουν περισσότερη πληροφορία, βασισμένοι στην αναπαράσταση των συνεντοπισμών. Γενικά, μπορούμε να πούμε, ότι παρέχουν μια μέση λύση μεταξύ εκφραστικότητας και γενίκευσης.

Το πλαίσιο εφαρμογής των NGG διαθέτει επίσης ένα σύνολο σημαντικών τελεστών. Αυτοί οι τελεστές μάς επιτρέπουν το συνδυασμό μεμονωμένων γράφων σε ένα γράφο πρότυπο / μοντέλο (πιο συγκεκριμένα, ο τελεστής της ενημέρωσης, update operator) και τη σύγκριση ζευγών γράφων, με την επιπλέον δυνατότητα διαβαθμισμένων μετρήσεων ομοιότητας (τελεστές ομοιότητας). Στο πεδίο της ανάλυσης σύστασης αλληλουχίας, η αναπαράσταση και το σύνολο των τελεστών παρέχουν ένα ακόμα μέσο ανάλυσης και σύγκρισης αλληλουχιών, με χαρακτηριστικά τα οποία απουσιάζουν από ευρέως διαδεδομένα πιθανοτικά μοντέλα, όπως τα HMM.

Οι NGG μπορούν να συνδυαστούν με διανυσματική αναπαράσταση αλληλουχιών, ώστε να καταστεί δυνατή η εφαρμογή αλγορίθμων μηχανικής μάθησης για την ταξινόμηση αλληλουχιών, με την ίδια λογικά όπως περιγράψαμε στην προηγούμενη ενότητα αυτού του κεφαλαίου. Στις επόμενες παραγράφους μελετούμε συντηρημένες μη κωδικοποιούσες αλληλουχίες και εξώνια με την προσέγγιση των γράφων N – γραμμάτων και συγκρίνουμε τα αποτελέσματά μας με τις γονιδιωματικές υπογραφές (GS) (Karlin 1998). Επιπροσθέτως, μελετούμε σύνολα δεδομένων CNE και εξωνίων, μαζί με κατάλληλες αναπληρωματικές αλληλουχίες, και δίνουμε, σε όλες τις περιπτώσεις, ποσοστά επιτυχούς ταξινόμησης.

Μέθοδοι

Ανάκτηση συνόλων δεδομένων

Για τους σκοπούς της μελέτης μας θεωρούμε πληθώρα δημοσιευμένων συνόλων δεδομένων αλληλουχιών που βρίσκονται υπό εξελικτικό περιορισμό. Δεδομένου ότι τα σύνολα δεδομένων διαφέρουν σε αριθμούς, διαλέξαμε τυχαία 1000 στοιχεία, από αρχεία μορφής BED, για περαιτέρω ανάλυση. Μόνο τα CNE του σκώληκα (worm CNE) μελετώνται στο σύνολό τους, καθώς αυτό το σύνολο περιέχει μόνο 1869 στοιχεία, τα οποία δημοσιεύτηκαν πρόσφατα (Polychronopoulos, Weitschek, et al. 2014). Οι εξωνικές αλληλουχίες των γονιδιωμάτων του ανθρώπου, του σκώληκα και των εντόμων (human, worm και insect exons αντίστοιχα) ανακτήθηκαν από το αποθετήριο του *UCSC*, βάσει των σχολιασμών του *RefSeq*, που αναφέρονται στις τελευταίες εκδόσεις των εκάστοτε γονιδιωμάτων (hg19, ce10, dm3 για άνθρωπο, σκώληκα και έντομο αντίστοιχα) (Pruitt et al. 2007). Τα σύνολα δεδομένων που περιγράφονται πιο κάτω, μαζί με τις αναπληρωματικές τους αλληλουχίες, χρησιμοποιούνται σε 26, στο σύνολο, πειράματα ταξινόμησης ανά ζεύγη, τα οποία στο κείμενο και στους πίνακες αναφέρονται ως #1, #2, κ.ο.κ. Στο «Συμπληρωματικό Excel[6]» υπάρχουν όλες οι πληροφορίες και τα στατιστικά για τους ενδιαφερόμενους αναγνώστες. Εκτός από τις εξωνικές αλληλουχίες, χρησιμοποιούνται οι ακόλουθες κλάσεις μη κωδικοποιουσών αλληλουχιών που βρίσκονται υπό επιλεκτικό περιορισμό (και έχουν αναφερθεί αναλυτικά και πιο πάνω στην αρχή αυτής της διατριβής):

I. UCNE (UltraConserved Noncoding Elements) (Dimitrieva & Bucher 2013),

II. EU100 μη εξωνικά CNE (EU100nx CNE) (Stephen et al. 2008),

III. CNE Αμνιωτικών και Θηλαστικών (Amniotic και Mammalian CNE) (Kim & Pritchard 2007),

IV. CNE σκώληκα και εντόμου (Worm και Insect UCNE) (Polychronopoulos, Weitschek, et al. 2014).

[6] Μπορείτε να κατεβάσετε το «Συμπληρωματικό Excel» στο σύνδεσμο: http://users.iit.demokritos.gr/~ggianna/Publications/alcob2014/

Χειρισμός αλληλουχιών και εξαγωγή αναπληρωματικών αλληλουχιών

Χρησιμοποιείται η σουίτα εργαλείων *BEDTools* (Quinlan & Hall 2010) για την εξαγωγή αλληλουχιών τύπου FASTA από BED αρχεία και για τον υπολογισμό επικαλυπτόμενων στοιχείων. Επιπροσθέτως, χρησιμοποιούμε τη σουίτα *EMBOSS* για τον υπολογισμό του GC περιεχομένου των αλληλουχιών (Rice et al. 2000). Για να λαμβάνουμε κάθε φορά αλληλουχίες ίδιου μήκους χρησιμοποιούμε scripts, τα οποία περιγράφησαν αναλυτικά πιο πριν στην προηγούμενη ενότητα αυτού του κεφαλαίου και είναι διαθέσιμα στους αναγνώστες κατόπιν σχετικού αιτήματος.

Από τις αλληλουχίες στο χώρο διανυσμάτων ομοιότητας των γράφων N – γραμμάτων (NGG)

Ακολουθούμε μια σειρά από βήματα ώστε να αναπαραστήσουμε τις αλληλουχίες, χρησιμοποιώντας τους NGG. Η ιδέα είναι ότι, από γνωστές (σημασμένες) αλληλουχίες, δημιουργούμε αντιπροσωπευτικούς NGG για κάθε κλάση αλληλουχιών. Έπειτα, περιγράφουμε όλες τις αλληλουχίες, βάσει της ομοιότητάς τους με τους αντιπροσωπευτικούς NGG, για κάθε κλάση. Επομένως, η εφαρμογή των NGG συνίσταται από τα παρακάτω βήματα:

- Αντιπροσώπευση κάθε αλληλουχίας χρησιμοποιώντας τους NGG.
- Υπολογισμός του αντιπροσωπευτικού (πρότυπου) γράφου NGG για κάθε κλάση εκπαίδευσης που χρησιμοποιείται.
- Υπολογισμός της ομοιότητας μεταξύ των στιγμιότυπων εκπαίδευσης (training instances) και των πρότυπων γράφων.
- Αναπαράσταση των στιγμιότυπων εκπαίδευσης χρησιμοποιώντας αποκλειστικά και μόνο τις ομοιότητές τους, δηλαδή σε ένα χώρο διανυσμάτων ομοιότητας.

Για περισσότερες, αναλυτικές λεπτομέρειες, τύπους και μαθηματικές εξισώσεις που περιγράφουν τα βήματα πιο πάνω, δείτε τις ακόλουθες εργασίες (Giannakopoulos G. Vouros G. and Stamatopoulos, P. 2008; Polychronopoulos, Krithara, et al. 2014). Η πρώτη περιγράφει τη μαθηματική και υπολογιστική θεμελίωση των NGG με σκοπό την εξαγωγή περιλήψεων, ενώ η δεύτερη την εφαρμογή αυτών στο πεδίο της ανάλυσης του γονιδιώματος και της Βιοπληροφορικής.

Η μέθοδος των γονιδιωματικών υπογραφών (Genomic Signatures, GS)

Όπως περιγράφηκε στην προηγούμενη ενότητα αυτού του κεφαλαίου, χρησιμοποιούμε και εδώ τη μέθοδο των γονιδιωματικών υπογραφών για τα δινουκλεοτίδια (GS), με σκοπό τη σύγκριση των αποτελεσμάτων ταξινόμησης στοιχείων του γονιδιώματος με χρήση και των δύο μεθόδων. Στην περίπτωσή μας, χρησιμοποιούμε την προσέγγιση των GS για δύο λόγους: α. Οι αλληλουχίες υπό εξέταση είναι όλες μικρού μήκους. Τα εξώνια αλλά και τα CNE έχουν μέσο μήκος που δεν ξεπερνά τα 200 νουκλεοτίδια, επομένως η χρήση μεγαλύτερης κλίμακας GS (π.χ. τρι- ή τετρανουκλεοτιδίων) θα έδινε αβέβαια αποτελέσματα, λόγω της επίδρασης του πεπερασμένου μεγέθους (finite size effect), β. Τα εξώνια, μια από τις κύριες κατηγορίες λειτουργικών στοιχείων του γονιδιώματος που χρησιμοποιούνται εδώ, είναι γνωστό ότι έχουν προτιμήσεις σε ορισμένα τρινουκλεοτίδια (και πολλαπλάσιά τους, όπως έξα- ή εννιαμερή), λόγω της εγγενούς δομής του γενετικού κώδικα και εξ' αιτίας των προτιμήσεων που συνεπάγονται από τη διαδικασία της μετάφρασης. Αυτή η ιδιότητα, γνωστή ως προτίμηση κωδικονίων (codon bias), είναι πολύ πιο έντονη στα τρινουκλεοτίδια παρά στα δινουκλεοτίδια, επομένως επιλέγουμε τα τελευταία ως βάση για την ανάλυσή μας με τη μέθοδο των GS.

Όταν πραγματοποιούμε τα πειράματα ταξινόμησης στις περιπτώσεις και των NGG και των GS, χρησιμοποιούμε το αποτέλεσμα (την έξοδο, output) της ανάλυσης (διανύσματα ομοιοτήτων στην περίπτωση των NGG και ποσοστά εμφάνισης νουκλεοτιδίων στην περίπτωση των GS) ως διανύσματα εισόδου ώστε να εκπαιδεύσουμε ταξινομητές που βασίζονται σε λογικούς κανόνες, όπως π.χ. την υλοποίηση JRIP (Hall et al. 2009) του RIPPER (Cohen 1995). Σημειώνουμε ότι τα αποτελέσματα που παρατίθενται εδώ, με βάση τον RIPPER, είναι συγκρίσιμα με αυτά που παίρνουμε από άλλους ευρέως διαδεδομένους ταξινομητές που δοκιμάζουμε (όπως τα Support Vector Machines – SVM, Random Forest). Στις ακόλουθες παραγράφους αναφέρουμε τα ευρήματά μας με βάση τα πειράματα που πραγματοποιήθηκαν και με τις δύο μεθόδους (NGG και GS).

Αποτελέσματα και συζήτηση

Εδώ περιγράφουμε μια συστηματική ανάλυση μικρών τμημάτων του γονιδιώματος, τα οποία επιδεικνύουν διαφορετικές λειτουργικότητες, και προέρχονται από τα γονιδιώματα του ανθρώπου, του σκώληκα και του εντόμου. Για κάθε πείραμα

ταξινόμησης υιοθετούμε τις τεχνικές των NGG και των GS, όπως εξηγήσαμε νωρίτερα, ώστε να εκτιμήσουμε πώς διαφορετικές τροπικότητες μπορούν να διαχωριστούν βάσει της εγγενούς σύστασης αλληλουχίας.

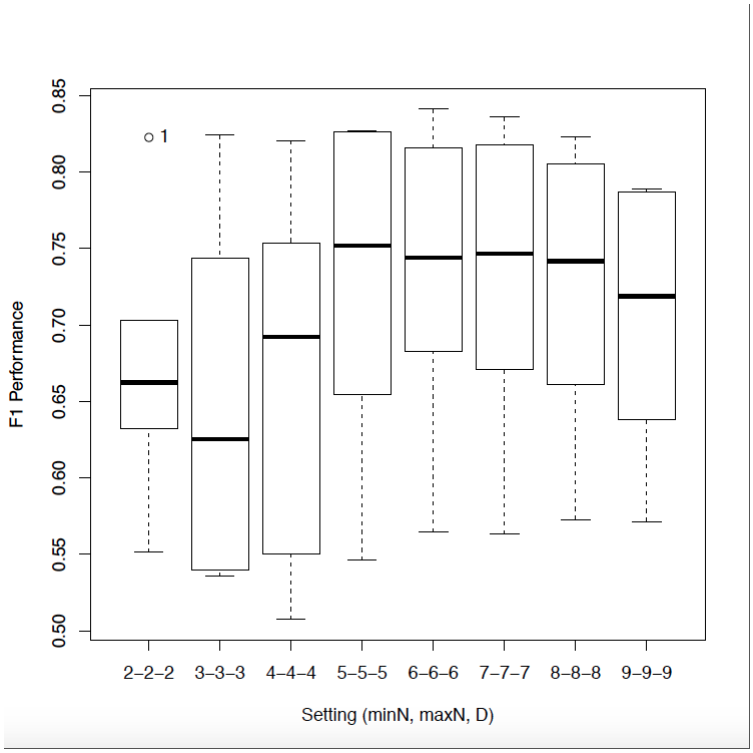
Σημειώνουμε ότι η αναπαράσταση μέσω NGG εμπεριέχει τρεις παραμέτρους. Η πρώτη είναι η τιμή του *m*, η οποία καθορίζει το ελάχιστο μήκος των N – γραμμάτων, στο οποίο χωρίζεται μια αλληλουχία. Η δεύτερη είναι η τιμή του *M*, η οποία καθορίζει το μέγιστο μήκος των N – γραμμάτων, στο οποίο χωρίζεται μια αλληλουχία. Η τρίτη είναι η τιμή του *D*, η οποία αντιπροσωπεύει τη μέγιστη απόσταση μέσα στην οποία θεωρούμε τα N – γράμματα (N – βάσεις) ότι είναι γείτονες. Η διατήρηση $m = M = D$ απλουστεύει την ανάλυση ελαχιστοποιώντας τον απαιτούμενο υπολογιστικό χρόνο, ενώ δεν τροποποιούνται σημαντικά τα αποτελέσματα στην πλειονότητα των περιπτώσεων (Giannakopoulos G. Vouros G. and Stamatopoulos, P. 2008).

Χρησιμοποιούμε έναν εκτιμητή, που είχε περιγραφεί αρχικά στην πρώτη εργασία με NGG, ώστε να καθοριστούν οι προαναφερθείσες παράμετροι. Δεδομένου ότι οι NGG αρχικά σχεδιάστηκαν για μια διαφορετική, γλωσσολογική ανάλυση (εκτίμηση υπολογιστικών συστημάτων εξαγωγής περιλήψεων), μη λαμβάνοντας υπόψη την ενδεχόμενη δυναμική πρόβλεψης στα πλαίσια πειραμάτων ταξινόμησης, προχωρούμε σε δοκιμασίες βελτιστοποίησης των παραμέτρων βασισμένοι σε διεξοδικά πειράματα (για $m \in [2, 9]$), χρησιμοποιώντας την τεχνική της 10πλής διασταυρωμένης επικύρωσης (10 – fold cross validation). Η βελτιστοποίηση πραγματοποιείται χρησιμοποιώντας 5 διαφορετικά σύνολα δεδομένων.

Υπολογίζουμε την επίδοση στα πειράματα ταξινόμησης με το μέτρο *F – measure*. Το *F – measure*, ένα μέτρο το οποίο χρησιμοποιείται συχνά στη μηχανική μάθηση σε πειράματα ταξινόμησης, είναι ο γεωμετρικός μέσος της ακρίβειας *P* (Precision: το ποσοστό των αλληλουχιών που έχουν εκχωρηθεί σε μια κλάση και ανήκουν πραγματικά σε αυτήν την κλάση) και της ανάκλησης *R* (Recall: το ποσοστό του συνόλου των αλληλουχιών που πράγματι ανήκουν σε μια κλάση και αποδόθηκαν σε εκείνη την κλάση) ενός ταξινομητή. Επομένως, $\mathbf{F1} = \frac{\mathbf{2 \times P \times R}}{\mathbf{P + R}}$. Στην *Εικόνα 7* απεικονίζουμε, μέσω του F1, την επίδοση της ταξινόμησης με τη βοήθεια των NGG. Κάθε στήλη ανταποκρίνεται σε ένα συνδυασμό τιμών παραμέτρων. Η έντονη οριζόντια γραμμή, σε κάθε στήλη, καταδεικνύει τη μέση επίδοση του συγκεκριμένου συνδυασμού για το για το σετ των 6 συνόλων δεδομένων. Οι υπόλοιπες γραμμές αντιπροσωπεύουν τις ποσοστιαίες τιμές της επίδοσης. Συμπεραίνουμε ότι υπάρχει ένα συστηματικά

υψηλό πλατώ επίδοσης, όταν $5 \leq m \leq 8$, το οποίο δείχνει ότι η μέθοδός μας δε βελτιώνεται συστηματικά / σημαντικά, όταν το m ξεπεράσει την τιμή 5. Μια άλλη παρατήρηση είναι ότι ο συνδυασμός παραμέτρων 2 – 2 – 2 αποδίδει σημαντικά καλύτερα από τον συνδυασμό 3 – 3 – 3 και σχεδόν το ίδιο καλά με τον 4 – 4 – 4. Το γεγονός αυτό ενδεχομένως σχετίζεται με τη σημασία των προτιμήσεων πρώτων γειτόνων, όπως αντανακλώνται στη μεθοδολογία των NGG.

Εικόνα 7: Θηκογράμματα (Boxplots) της επίδοσης ταξινόμησης των NGG με χρήση διαφορετικών παραμέτρων



Στους πίνακες, που περιλαμβάνονται στη συνέχεια, θεωρούμε την επίδοση των NGG για κάθε βέλτιστο συνδυασμό τιμών παραμέτρων. Για παράδειγμα, στο πείραμα 1 (Exp #1), *Πίνακας 9*, η τιμή 83.86 δίνεται για τους NGG, η οποία ανταποκρίνεται στην επιλογή συνδυασμού παραμέτρων 7 – 7 – 7. Προτείνουμε (m, M, D) = (5, 5, 5), ως τις βέλτιστες ρυθμίσεις παραμέτρων, για την ανάλυση μικρών αλληλουχιών του γονιδιώματος, βασισμένοι και στα 26 πειράματα ταξινόμησης που πραγματοποιήθηκαν (δείτε «Συμπληρωματικό Excel», καρτέλα "Overal Statistics"). Οι γονιδιωματικές υπογραφές (GS) χρησιμοποιούνται ως μια εναλλακτική μέθοδος ταξινόμησης σε

αντιπαραβολή με τους NGG. Από όσο γνωρίζουμε, για τόσο μικρές γονιδιωματικές αλληλουχίες, λειτουργικής (ή πιθανά λειτουργικής) σημασίας, αυτή είναι η πρώτη εργασία που εξετάζει την εφαρμογή και των δύο μεθόδων (NGG, GS).

Συγκρίσεις αλληλουχιών υποβάθρου (background sequences) μεταξύ ειδών

Για συγκρίσεις μεταξύ αλληλουχιών υποβάθρου του ανθρώπου, του σκώληκα και των εντόμων χρησιμοποιούμε αναπληρωματικά σύνολα δεδομένων (surrogate sets, όπως περιγράφουμε στις "Μεθόδους") ως αντιπροσωπευτικά δείγματα των διαφορετικών γονιδιωμάτων. Οι συγκρίσεις που περιλαμβάνουν τον *H. sapiens* δίνουν πάντα τα καλύτερα ποσοστά ταξινόμησης, χρησιμοποιώντας και τους γράφους N – γραμμάτων (NGG) και τις γονιδιωματικές υπογραφές (GS), δείτε *Πίνακα 9*. Αυτό μπορεί να εξηγηθεί στα πλαίσια της υψηλής διαφοράς των προτιμήσεων γειτόνων, κυρίως σε CpG και TpA μεταξύ *H. sapiens* και ασπόνδυλων, ενώ αυτές οι προτιμήσεις βρίσκονται αρκετά κοντά μεταξύ ειδών ασπόνδυλων, όπως η *D. melanogaster* (έντομο) και ο *C. elegans* (σκώληκας). Οι GS είναι αποκλειστικά μια ποσοτικοποίηση των προτιμήσεων πρώτων γειτόνων και ως συνέπεια δεν επηρεάζονται από το περιεχόμενο σε GC. Αντιθέτως, οι NGG είναι σε θέση να ενσωματώσουν εννοιολογικά διάφορα συστατικά / πτυχές της σύστασης αλληλουχίας, όπως είναι η μονονουκλεοτιδική σύσταση και οι προτιμήσει γειτόνων ανώτερης τάξης (ένα κριτήριο το οποίο παρέχεται ως μια τιμή παραμέτρου στη μεθοδολογία μας μεταβάλλοντας το *D*).

Σε περιπτώσεις όπου περιλαμβάνονται ανθρώπινες αλληλουχίες στις συγκρίσεις, οι GS αποδίδουν συστηματικά καλύτερα από τους NGG. Αυτή η διαφορά είναι επίσης στατιστικά σημαντική με ένα p – value μικρότερο από 0.10 (paired t – test). Αξίζει να σημειωθεί ότι οι δύο περιπτώσεις με τις υψηλότερες διαφορές σε περιεχόμενο GC μεταξύ των συνόλων που περιλαμβάνονται σε πειράματα ταξινόμησης είναι: το πείραμα #22 (Exp #22), το οποίο είναι το μοναδικό (σε αυτή την ομάδα συγκρίσεων), όπου οι NGG αποδίδουν καλύτερα από τις GS και το πείραμα #1 (Exp #1), όπου οι GS αποδίδουν καλύτερα από τους NGG, αλλά με τη χαμηλότερη διαφορά στις επιδόσεις. Οι περιπτώσεις με την υψηλότερη (σε μεγαλύτερο βαθμό), σχετικά, επικράτηση των NGG είναι εκείνες με τις υψηλότερες διαφορές σε ποσοστό GC, όπως αναμένεται λόγω της σχετικής ευαισθησίας των δύο μεθόδων σε αυτήν την παράμετρο σύστασης.

Πίνακας 9: Συγκρίσεις μεταξύ ειδών γονιδιωματικών αλληλουχιών υποβάθρου (background)

Exp	Description	Average length	Average GC	NGG	GS
#1	surrogates for human exons surrogates for worm exons	167.837 169.318	0.5155 0.4049	83.86	85.98
#14	surrogates for human UCNEs surrogates for insect UCNEs	86.094 86.582	0.3651 0.3949	79.38	84.05
#20	surrogates for insect exons surrogates for human exons	169.318 169.816	0.5202 0.5087	80.48	87.49
#22	surrogates for worm exons surrogates for insect exons	213.365 212.858	0.4194 0.5194	73.50	70.35
#23	surrogates for human UCNEs surrogates for worm UCNEs	82.932 82.875	0.3648 0.4297	80.35	83.75
#13	surrogates for worm UCNEs surrogates for insect UCNEs	83.407 86.582	0.4265 0.3949	58.79	64
	Average			**76.06**	**79.27**

Πειράματα ταξινόμησης αλληλουχιών DNA που βρίσκονται υπό επιλεκτικό περιορισμό έναντι των αναπληρωματικών αλληλουχιών υποβάθρου (Ταξινόμηση εντός του ίδιου είδους)

Τα ακόλουθα πειράματα αναφέρονται σε συγκρίσεις αλληλουχιών που βρίσκονται υπό επιλεκτικό περιορισμό έναντι των αναπληρωματικών τους (*Πίνακας 10*). Σημειώνουμε εδώ ότι οι αναπληρωματικές αλληλουχίες μοιράζονται το ίδιο ποσοστό GC% και μήκος με τις αρχικές αλληλουχίες (δείτε «Μεθόδους»). Όπως προκύπτει από επιθεώρηση του *Πίνακα 10*, οι αλληλουχίες ασπόνδυλων που βρίσκονται υπό επιλεκτικό περιορισμό δεν ταξινομούνται με τόση μεγάλη επιτυχία όπως οι αντίστοιχες ανθρώπινες χρησιμοποιώντας τους NGG (δείτε πειράματα #2 και #17). Μόνο οι UCNE εντόμων (insect UCNEs) φαίνεται πως αντιβαίνουν αυτόν τον κανόνα, δείτε πείραμα #12. Το συγκεκριμένο εύρημα μπορεί να κατανοηθεί στα πλαίσια διαφόρων ιδιαιτεροτήτων των θερμόαιμων ζώων (συνήθως όλων των σπονδυλωτών), ιδιαιτέρως όσον αφορά το μη λειτουργικό, μη συντηρημένο ποσοστό του γονιδιώματος. Αυτές περιλαμβάνουν υψηλό εμπλουτισμό σε μεταθετά στοιχεία, μικροδορυφόρους, ομοπουρινικά / ομοπυριμιδινικά κατάλοιπα και διάφορα άλλα χαρακτηριστικά μοτίβα σύστασης. Τέτοια γονιδιώματα παρουσιάζουν επίσης ένα τυπικό προφίλ δινουκλεοτιδίων προς αποφυγή (ειδικά CpG και TpA), τα οποία αποφεύγονται σε μικρότερο ποσοστό στις αλληλουχίες που βρίσκονται υπό επιλεκτικό περιορισμό (εξώνια, CNE), οι οποίες, έχοντας λειτουργικούς ρόλους, δεν ακολουθούν αυστηρά τις μέσες τάσεις σύστασης του γονιδιώματος. Σημειώνουμε εδώ ότι τα γονιδιώματα ασπόνδυλων είναι πολύ λιγότερο άφθονα σε επαναλαμβανόμενα στοιχεία και δε παρουσιάζουν υποεκπροσώπηση ειδικών δινουκλεοτιδίων. Στις συγκρίσεις με τη μέθοδο των GS ακολουθείται η ίδια τάση αλλά

οι διαφορές είναι ελάχιστες, λόγω της ευαισθησίας αυτών των ποσοτήτων αποκλειστικά στις προτιμήσεις πρώτων γειτόνων. Γνωρίζουμε, από πρότερες μελέτες, ότι οι GS δεν αποδίδουν καλά σε συγκρίσεις μέσα στο ίδιο είδος (intra – species comparisons), διότι οι προτιμήσεις γειτόνων παραμένουν σχετικά σταθερές μέσα στο ίδιο γονιδίωμα. Επομένως, στις περισσότερες περιπτώσεις που παρατίθενται, οι NGG αποδίδουν καλύτερα από τις GS (συγκρίνετε τους μέσους όρους, 68.76% έναντι 61.84% στον πίνακα που ακολουθεί). Η διαφορά είναι στατιστικά σημαντική (p – value μικρότερο από 0.10, paired t – test).

Πίνακας 10: Ταξινόμηση αλληλουχιών που βρίσκονται υπό επιλεκτικό περιορισμό έναντι αναπληρωματικών αλληλουχιών υποβάθρου (background surrogates)

Exp	Description	Average length	Average GC	NGG	GS
#2	worm exons	213.365	0.4243	57.7	61.05
	surrogates	213.365	0.4239		
#3	human exons	169.816	0.5190	74.3	63.41
	surrogates	169.816	0.5183		
#17	insect exons	388.822	0.5412	52.36	59.45
	surrogates	381.557	0.5389		
#4	worm UCNEs	82.875	0.4309	56.76	55.62
	surrogates	82.875	0.4308		
#5a	human UCNEs	326.923	0.3676	82.63	72.00
	surrogates	326.923	0.3676		
#5b	human EU100nx CNEs	155.499	0.3783	76.43	63.75
	surrogates	155.499	0.3783		
#5c	amniotic CNEs	289.061	0.3756	78.62	63.00
	surrogates	289.061	0.3756		
#5d	mammalian CNEs	246.488	0.4015	75.85	55.65
	surrogates	246.488	0.4018		
#12	insect UCNEs	86.582	0.3949	64.15	62.65
	surrogates	86.582	0.3949		
	Average			**68.76**	**61.84**

Οι τελευταίες έξι σειρές του *Πίνακα 10* δηλώνουν συγκρίσεις αρκετών συλλογών CNE αλληλουχιών έναντι των αναπληρωματικών τους (surrogates). Γενικά παρατηρούμε ότι, μεταξύ αλληλουχιών που βρίσκονται υπό επιλεκτικό περιορισμό, ανθρώπινες CNE αλληλουχίες έναντι των αναπληρωματικών τους παρουσιάζουν σχετικά υψηλότερα ποσοστά ταξινόμησης, σε σύγκριση με εξωνικές αλληλουχίες έναντι των αντίστοιχων αναπληρωματικών τους. Αυτό μπορεί να αποδοθεί στο γεγονός ότι τα CNE (και ειδικά τα ανθρώπινα UCNE που λαμβάνουμε υπόψη εδώ) είναι πιο συντηρημένα ακόμα και από τα εξώνια. Επιπρόσθετα, είναι γνωστό από τη βιβλιογραφία ότι τα CNE δρουν ως θέσεις πρόσδεσης μεταγραφικών παραγόντων και φέρουν αρκετά μοτίβα, τα οποία η ανάλυση μέσω NGG είναι αρκετά ευαίσθητη να ανιχνεύσει.

Οι γονιδιωματικές υπογραφές αποδίδουν καλύτερα στην ταξινόμηση τμημάτων του γονιδιώματος λειτουργικής σημασίας και διαφορετικής προέλευσης

Στα πειράματα, που περιλαμβάνονται στον *Πίνακα 11*, μελετούμε την απόδοση των NGG και των GS, όταν εφαρμόζονται σε σύνολα δεδομένων λειτουργικών αλληλουχιών (CNE ή εξώνια) διαφορετικών γονιδιωμάτων. Επιβεβαιώνουμε το γεγονός ότι οι GS αποδίδουν καλύτερα στην ταξινόμηση στοιχείων διαφορετικών γονιδιωμάτων και η απόδοσή τους βελτιώνεται ελάχιστα στον *Πίνακα 11*, συγκριτικά με τον *Πίνακα 9*. Αυτό σημαίνει ότι δεν αλλοιώνεται το αποτέλεσμα λόγω του «συντηρημένου» χαρακτήρα των αλληλουχιών, δηλαδή η προτίμηση πρώτων γειτόνων, που χαρακτηρίζουν τα διαφορετικά γονιδιώματα και φιλτράρονται από τις GS, διατηρείται ξεκάθαρα στα εξώνια και στα CNE. Ελέγχουμε, ξανά, τη στατιστική σημαντικότητα στην επίδοση μεταξύ των μεθόδων NGG και GS (paired t – test) και βλέπουμε ότι υπάρχει μια στατιστικά σημαντική διαφορά ($p - value < 0.05$).

Απ' την άλλη μεριά, η επίδοση των NGG μειώνεται. Το γεγονός αυτό ενδεχομένως σχετίζεται με το πολύπλοκο σύνολο ιδιοτήτων σύστασης που είναι χαρακτηριστικές της μεθόδου των NGG: νουκλεοτιδική σύσταση της αλληλουχίας, προτιμήσεις πρώτων γειτόνων και, επιπλέον, σχέσεις γειτνίασης διαφόρων επιπέδων / τάξεων. Επομένως, σε αυτό το πλαίσιο, οι διαφορές μεταξύ των γονιδιωμάτων που οδηγούν σε μια αποτελεσματική ταξινόμηση αλληλουχιών υποβάθρου (background sequences) με χρήση των NGG, αλλοιώνονται από την παρουσία κοινών περιορισμών σε επίπεδο σύστασης αλληλουχίας, οι οποίοι σχετίζονται με λειτουργία, όπως φαίνεται στον *Πίνακα 11* (M.O ή average = 74.47). Ένα τέτοιο παράδειγμα περιορισμού σύστασης είναι η παρατηρούμενη υπερεκπροσώπηση αδενίνης και γουανίνης στη σύσταση των εξωνίων που κωδικοποιούν πρωτεΐνες (purine loading).

Πίνακας 11: Συγκρίσεις λειτουργικών αλληλουχιών μεταξύ γονιδιωμάτων

Exp	Description	Average length	Average GC	NGG	GS
#9	worm exons human exons	169.318 169.816	0.3886 0.5190	74.68	82.21
#10	worm UCNEs human UCNEs	83.113 82.640	0.4277 0.3728	77.77	82.29
#15	worm UCNEs insect UCNEs	83.086 86.582	0.4285 0.3949	70.89	74.95
#16	human UCNEs insect UCNEs	86.094 86.582	0.3704 0.3949	82.08	86.70
#19	insect exons human exons	169.318 169.816	0.5148 0.5089	70.03	81.29
#21	worm exons insect exons	213.365 212.858	0.4196 0.5093	71.39	72.35
	Average			**74.47**	**79.97**

Συμπεράσματα

Το πρόβλημα της απόδοσης λειτουργικότητας σε αλληλουχίες, που δεν έχουν σχολιαστεί (unannotated sequences), αποτελεί πρόκληση για τη Γονιδιωματική, ειδικά όταν πρόκειται για χαρακτηρισμό των πολύπλοκων γονιδιωμάτων των θηλαστικών. Με την πραγματοποίηση μιας σειράς συγκρίσεων μεταξύ αντιπροσωπευτικών αλληλουχιών, διαφορετικής προέλευσης και λειτουργικότητας, δείχνουμε εδώ τη δυναμική των γράφων N – γραμμάτων (NGG) στην αποτελεσματική διάκριση μεταξύ τμημάτων αλληλουχιών του γονιδιώματος με αναμενόμενη λειτουργία έναντι του σωρού (bulk) του γονιδιώματος. Οι NGG είναι σε θέση να ποσοτικοποιήσουν το αποτέλεσμα των περιορισμών σε επίπεδο αλληλουχίας, επομένως η εφαρμογή τους στην περιγραφή άλλων λειτουργικών αλληλουχιών των γονιδιωμάτων θηλαστικών (όπως είναι περιοχές με δομικές ιδιότητες, περιοχές πρόσδεσης στον πυρηνικό φάκελο και διάφορα είδη μη κωδικοποιούντων RNA) ενδεχομένως δώσει πολύτιμες πληροφορίες πάνω στη διελεύκανση των αλληλεπιδράσεων και της σχέσης μεταξύ σύστασης και λειτουργίας. Όταν πρόκειται για συγκρίσεις αλληλουχιών διαφορετικής μεταξύ γονιδιωμάτων, οι NGG είναι λιγότερο αποτελεσματικοί, σε σύγκριση με τη μέθοδο των GS. Η αποτελεσματικότητα της τελευταίας δείχνεται για πρώτη φορά εδώ (σε συνδυασμό με τη μελέτη που δείξαμε στην προηγούμενη ενότητα αυτού του κεφαλαίου, δείτε **3.1**). Οι GS είχαν χρησιμοποιηθεί, έως σήμερα, για τη μελέτη περισσότερο εκτεταμένων περιοχών του γονιδιώματος, μήκους > 50 kb (Karlin 1998). Περισσότερη δουλειά απαιτείται για την περαιτέρω αξιοποίηση των αποτελεσμάτων που παρουσιάζονται, ενώ ένας συνδυασμός των δύο μεθόδων (GS και NGG) ενδεχομένως να οδηγήσει σε υψηλότερα ποσοστά ταξινόμησης.

4. Ανάλυση της χρωμοσωμικής κατανομής των CNE με χρήση μεθόδων κλιμάκωσης εντροπίας (entropic scaling) και εγκιβωτισμού (box counting)

Εισαγωγή

Η μελέτη της χρωμοσωμικής κατανομής των CNE με μεθόδους κλιμάκωσης εντροπίας και εγκιβωτισμού έχει ως σκοπό την ανάλυση του βαθμού μορφοκλασματικότητας (fractality). Οι συγκεκριμένες μέθοδοι έχουν χρησιμοποιηθεί για τη μελέτη της κατανομής άλλων στοιχείων του γονιδιώματος, όπως είναι οι κωδικοποιούσες αλληλουχίες (Athanasopoulou et al. 2010) και τα μεταθετά στοιχεία (Athanasopoulou et al. 2014), ενώ ενδείκνυνται για τη μελέτη της κατανομής των CNE ειδικότερα, διότι τα τελευταία έχει προταθεί μέσω πειραμάτων 3C (Chromosome Conformation Capture) ότι αλληλεπιδρούν μεταξύ τους (Akalin et al. 2009; Viturawong et al. 2013; Dimitrieva & Bucher 2012) και συνεπώς ενέχονται σε συσχετίσεις μακράς εμβέλειας (long – range correlations). Σε αυτό συνηγορεί το γεγονός, επιπλέον, ότι δείξαμε πως οι αποστάσεις μεταξύ αυτών των στοιχείων στο ανθρώπινο και σε άλλα γονιδιώματα ακολουθούν κατανομή τύπου νόμου δύναμης, ανεξάρτητα από το εάν βρίσκονται κοντά σε γονίδια ή όχι. Οι νόμοι δύναμης συνδέονται και αυτοί άμεσα με φαινόμενα μορφοκλασματικότητας (fractality) και συσχετίσεις μακράς εμβέλειας γενικότερα (Feder 1998; Mandelbrot 1982), ενώ, σύμφωνα με σχετικά πρόσφατα πειραματικά αποτελέσματα, βρίσκουν εφαρμογή και στο γονιδίωμα (Lieberman-Aiden et al. 2009).

Στη θεωρία της πληροφορίας, η έννοια της εντροπίας αναπτύχθηκε από τον Claude Shannon (Shannon 1948), με σκοπό την εκτίμηση της ποσότητας της πληροφορίας που μεταφέρεται σε ένα μεταδιδόμενο μήνυμα. Κατά τις τελευταίες δεκαετίες έχουν παρατηρηθεί φαινόμενα που δεν εξαρτώνται από κάποια κλίμακα, καθώς και φαινόμενα μορφοκλασματικότητας σε χρονοσειρές από μετάδοση σήματος στην ηλεκτρονική μηχανική, σεισμούς, οικονομία, κοινωνικές επιστήμες και σε πολλά άλλα πεδία. Συχνά τέτοιες μελέτες διεξάγονται χρησιμοποιώντας την καθιερωμένη μεθοδολογία του εγκιβωτισμού και σε αρκετές άλλες περιπτώσεις συστημάτων που

χαρακτηρίζονται από συσχετίσεις μακράς εμβέλειας χρησιμοποιείται η εντροπία Shannon.

Πρόσφατα από το εργαστήριό μας μελετήθηκαν οι ιδιότητες κλιμάκωσης εντροπίας Shannon (block entropy) της κατανομής των γονιδίων και των μεταθετών στοιχείων στα ευκαρυωτικά γονιδιώματα μέσω της αναγωγής της αλληλουχίας DNA σε μια αλληλουχία συμβόλων (Athanasopoulou et al. 2010; Athanasopoulou et al. 2014). Η σύμβαση που ακολουθήθηκε για τα γονίδια, π.χ., ήταν ότι τα '0', στην αλληλουχία συμβόλων, αναπαριστούν νουκλεοτίδια που δεν αντιστοιχούν σε κωδικοποιούσες περιοχές (γονίδια) και τα '1', νουκλεοτίδια που ανήκουν σε αλληλουχίες που κωδικοποιούν πρωτεΐνες (εξώνια). Πολλές μελέτες έχουν δείξει ότι, μια γραμμική κλιμάκωση της εντροπίας (Shannon-like ή block εντροπία) $H(n)$ με το μήκος n της λέξης (εφεξής την καλούμε n-λέξη) σε ημι-λογαριθμικά διαγράμματα, αποτελεί ξεκάθαρη ένδειξη κλίμακας μακράς εμβέλειας και μορφοκλασματικότητας (Ebeling & Nicolis 1991; Ebeling & Nicolis 1992).

Πιο κάτω μελετούμε την κατανομή διαφόρων συνόλων δεδομένων CNE με τον ακόλουθο τρόπο: Νουκλεοτίδια του χρωμοσώματος αντικαθίστανται από '0', εάν δεν ανήκουν στον τύπο CNE που μελετούμε και από '1' εάν ανήκουν. Επομένως, οι αλληλοδιαδοχές μικρών νησίδων, που σχηματίζονται από τις μονάδες και διακόπτουν το συνεχές των μηδενικών, αντικατοπτρίζουν το πρότυπο της χωρικής χρωμοσωμικής διαμόρφωσης για αυτόν το συγκεκριμένο τύπο CNE. Στη συνέχεια προχωρούμε με τη συγκεκριμένη αλληλουχία δυαδικών συμβόλων για κάθε χρωμόσωμα ως εξής: (i) Η κλιμάκωση της εντροπίας κατά ομάδες $H(n)$ *vs.* n σε αυτή την αλληλουχία συμβόλων μελετάται και ποσοτικοποιείται μέσω της έκτασης της γραμμικότητας σε διαγράμματα ημι-λογαριθμικής κλίμακας, και (ii) Μια μέθοδος εγκιβωτισμού εφαρμόζεται και υπολογίζεται η διάσταση ομοιότητας μαζί με την έκταση της γραμμικότητας σε διπλή λογαριθμική κλίμακα.

Μέθοδος κλιμάκωσης της εντροπίας Shannon (Block entropy scaling)

Ας υποθέσουμε ότι έχουμε μια αλληλουχία συμβόλων μεγέθους N, με τα σύμβολα να προέρχονται από ένα δυαδικό αλφάβητο {0,1} και έστω $p_n(A_1, \ldots, A_n)$ η πιθανότητα να βρούμε μια ομάδα ή μια n – λέξη $(A_1, \ldots, A_n)$ σε αυτήν την αλληλουχία. Η εντροπία Shannon για τις n – λέξεις, ή εντροπία κατά ομάδες (block entropy), ορίζεται ως:

$$H(n)=-\sum p_n(A_1,\ldots,A_n)\ln p_n(A_1,\ldots,A_n) \quad (1)$$

Η ποσότητα *H(n)* μπορεί να ερμηνευθεί ως ένα μέτρο της μέσης αβεβαιότητας για την πρόβλεψη μιας *n* – λέξης. Στη βιβλιογραφία μπορεί να συναντήσει κάποιος δύο τρόπους ανάγνωσης της αλληλουχίας συμβόλων και εξαγωγής της κατανομής της πιθανότητας για τις *n* – λέξεις: 'gliding' και 'lumping'. Στην παρούσα μελέτη οι αλληλουχίες συμβόλων διαβάζονται με 'lumping'. Αυτό σημαίνει ότι, αντί να διαβάζονται διεξοδικά όλες οι πιθανές λέξεις μήκους *n* (gliding), θεωρούνται μόνο οι *n* – λέξεις που λαμβάνονται με ένα σταθερό βήμα ίσο με *n*. Ισοδύναμα, μπορούμε να πούμε ότι αφού διαβαστεί η αρχική n-λέξη μιας αλληλουχίας, η επόμενη n-λέξη που μετράμε είναι αυτή που ξεκινά από τη θέση *n* + 1 και έπειτα μέχρι το τέλος της αλληλουχίας. Επομένως, κάθε γράμμα της αλληλουχίας ανήκει μόνο σε μια *n* – λέξη που μετράται. Η μεθοδολογία αυτή υιοθετήθηκε γιατί, ενώ λαμβάνει δειγματοληπτικά επαρκώς τη συμβολοσειρά, είναι πολύ πιο γρήγορη από άποψη υπολογιστικού χρόνου και δίνει τη δυνατότητα ανίχνευσης μη-μονοτονικής εντροπικής κλιμάκωσης (Karamanos & Nicolis 1999; Karamanos 2001).

Οι ιδιότητες κλιμάκωσης της εντροπίας κατά ομάδες έχουν χρησιμοποιηθεί ως ένα μέτρο για την ταξινόμηση των αλληλουχιών συμβόλων. Σημαντικά χαρακτηριστικά κλιμάκωσης του *H(n)* έχουν ερευνηθεί από διάφορους συγγραφείς. Οι Ebeling και Nicolis πρότειναν την εξής μορφή για την κλιμάκωση του *H(n)*:

$$H(n)=e+gn^{\mu_0}(\ln n)^{\mu_1}+nh \quad (2)$$

για αλληλουχίες συμβόλων που προέρχονται από δυναμική μη-γραμμικότητας συμπεριλαμβανομένων διαδικασιών επεξεργασίας γλώσσας (Ebeling & Nicolis 1991; Ebeling & Nicolis 1992). Πιο συγκεκριμένα, στην περίπτωση του ελκυστή Feigenbaum της λογιστικής απεικόνισης (logistic map) και για $n=2^k$ (k = 2, 3, 4 …), ο Grassberger(Grassberger 1986) έδειξε ότι για την ανάγνωση μιας αλληλουχίας με gliding ισχύει:

$$H(n)=\log_2(3n/2) \quad (3)$$

Σε αυτό το σύστημα συναντάται γραμμικότητα σε ημι-λογαριθμική κλίμακα που στην περίπτωση της Εξίσωσης (2) ανταποκρίνεται σε $g \neq 0$, $h = 0$, $\mu_0 = 0$, $\mu_1 > 0$ (Nicolis & Gaspard 1994). Ο συγκεκριμένος τύπος κλιμάκωσης εικάζεται ότι ισχύει για μια μεγάλη κλάση αλληλουχιών συμβόλων με ιδιότητες fractal. Γι'αυτό η γραμμικότητα *H(n)* – log*n* σχετίζεται με τη δομή ανεξάρτητη κλίμακας τέτοιων αλληλουχιών που συνεπάγεται την ύπαρξη συσχετίσεων μακράς εμβέλειας.

Για όλα τα γονιδιωματικά σύνολα δεδομένων και τις προσομοιώσεις κατασκευάσαμε αναπληρωματικές τυχαίες αλληλουχίες με την ίδια σύσταση 0/1 και οι οποίες δεν έχουν, από κατασκευής, καμία εσωτερική δομή. Συγκεκριμένα, κάναμε ανασύσταση μιας αλληλουχίας με το ίδιο μήκος όπως η γονιδιωματική διασπείρωντας τις βιολογικές λειτουργικές μονάδες (στην περίπτωσή μας τα CNE) σε τυχαίες θέσεις. Οι καμπύλες που δείχνουν την κλιμάκωση της εντροπίας (*H(n) vs. n*) της αρχικής αλληλουχίας και της αναπληρωματικής της παρουσιάζονται στο ίδιο διάγραμμα.

Ποσοτικοποιούμε τη μορφοκλασματικότητα μιας αλληλουχίας μέσω της έκτασης της γραμμικότητας E σε ημι-λογαριθμική κλίμακα και της ανάλογης κλίσης S. Όταν υπάρχει πάνω από ένα γραμμικό τμήμα, συμβολίζουμε με E* το άθροισμα των μηκών τους. Μια επιπλέον ποσότητα που εισάγουμε εδώ και χρησιμοποιείται ευριστικά ως ένας εκτιμητής του βαθμού οργάνωσης μιας αλληλουχίας είναι ο λόγος R της τιμής της εντροπίας της αναπληρωματικής αλληλουχίας προς την τιμή της εντροπίας της αναλυόμενης (γονιδιωματικής ή από προσομοίωση) αλληλουχίας. Αυτός ο λόγος υπολογίζεται πάντοτε για την τιμή του *n*, όπου η αναπληρωματική αλληλουχία παρουσιάζει τη μέγιστη τιμή της εντροπίας της, προτού η επίδραση του πεπερασμένου μεγέθους (finite size effect) στρεβλώσει το σχήμα της. Μεγάλες τιμές R υποδηλώνουν υψηλό βαθμό τάξης και μορφοκλασματικότητας της αλληλουχίας υπό μελέτη.

Μέθοδος εγκιβωτισμού (box counting) για την εκτίμηση του βαθμού της μορφοκλασματικότητας και της fractal διάστασης

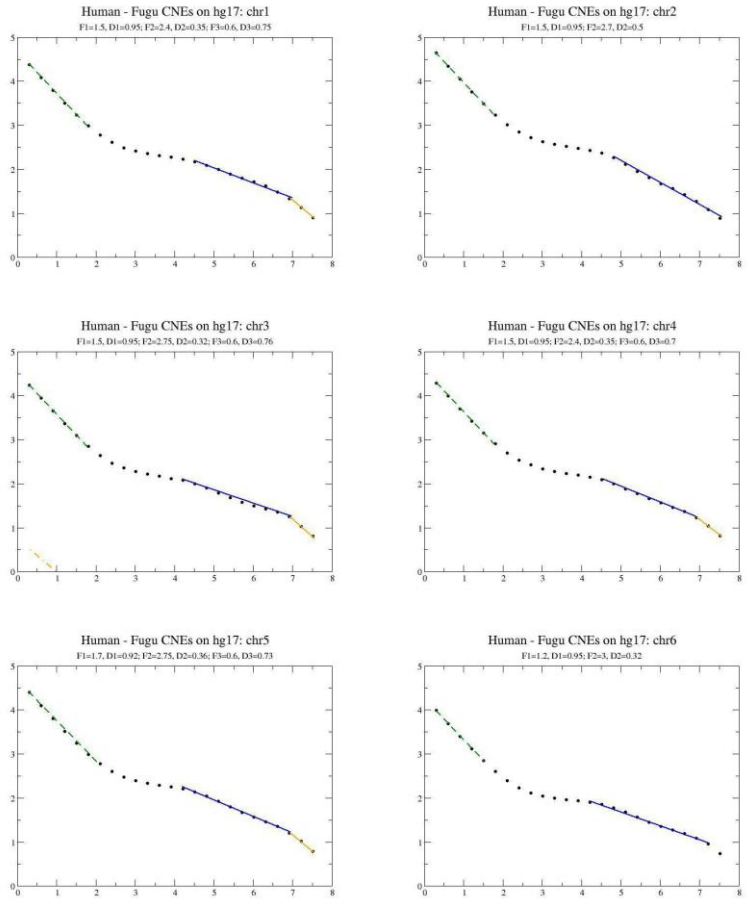
Η μέθοδος εγκιβωτισμού είναι μια κλασική μέθοδος για την εκτίμηση των χαρακτηριστικών μορφοκλασματικότητας σε ένα σύνολο δεδομένων (Feder 1998; Mandelbrot 1982). Χρησιμοποιούμε μια απλή μονοδιάστατη εκδοχή κατά την οποία ολόκληρο το μήκος του χρωμοσώματος καλύπτεται από μονοδιάστατα 'κυτία' μήκους *δ*. Ο αριθμός αυτών των κυτίων που επικαλύπτουν (πλήρως ή εν μέρει) τουλάχιστον ένα αντίγραφο του CNE θεωρείται ότι αντιπροσωπεύει το μήκος του χρωμοσώματος

L(δ) που καλύπτεται από CNE. Όταν υπάρχει μορφοκλασματικότητα, το μετρούμενο μήκος δε φαίνεται να φθάνει σε μια καθορισμένη τιμή, όσο το δ μικραίνει. Το μετρούμενο μήκος επίσης κλιμακώνεται ως $L(\delta) \sim \delta^D$, με τον εκθέτη D να παριστάνει την αρνητική fractal διάσταση D_f του αντικειμένου υπό μελέτη. Τα διαγράμματα που βλέπουμε στη συνέχεια, τα οποία με ποιο τρόπο το L(δ) κλιμακώνεται ως συνάρτηση του δ, παρουσιάζονται σε διπλή λογαριθμική κλίμακα. Μας ενδιαφέρει και η κλίση του γραμμικού τμήματος της καμπύλης αλλά και η έκταση της γραμμικότητας. Εδώ η μορφοκλασματικότητα θεωρούμε ότι ισχύει εάν η διάσταση D_f παίρνει τιμές μικρότερες από 0.9 για μια έκταση γραμμικότητας (F) που ξεπερνά τη μία τάξη μεγέθους. Οι τιμές στα όρια της γραμμικής περιοχής καθορίζουν τα χαμηλότερα και υψηλοτερα κατώφλια μεταξύ των οποίων το μελετούμενο χωρικό πρότυπο παρουσιάζει στατιστικά σημαντική αυτο-ομοιότητα. Υπολογίζουμε το L(δ) δέκα φορές, για κάθε τιμή του δ, με μια μετατόπιση πλαισίου ίση με 1/40 του συνολικού μήκους της αλληλουχίας και μετά παίρνουμε τον μέσο όρο ώστε να πάρουμε αποτελέσματα ανεξάρτητα από την επιλογή του σημείου εκκίνησης της μέτρησης.

Προσέξτε ότι σε μια γονιδιωματική κατανομή CNE παρατηρούνται δύο περιοχές γραμμικότητας στην πλειονότητα των περιπτώσεων: μία στην περιοχή μικρού μήκους που σχετίζεται με το μέγεθος των CNE υπό μελέτη και μία στην περιοχή μεγάλου μήκους. Η τελευταία είναι η μόνη για την οποία η κλίση βρίσκεται να αποκλίνει σημαντικά από το -1 (στις περιπτώσεις που δείχνουν μορφοκλασματικότητα).

Μορφοκλασματικότητα σε γονιδιωματικές κατανομές CNE όπως μετράται από μια μέθοδο εγκιβωτισμού: Στην *Εικόνα 8* παρουσιάζουμε διαγράμματα που παριστάνουν δεδομένα CNE που έχουν αναλυθεί με τη μέθοδο εγκιβωτισμού ανά χρωμόσωμα. Σε όλα τα διαγράμματα παρατηρούμε δύο (συνήθως) γραμμικά τμήματα, στις περιοχές μικρών και μεγάλων τιμών του μεγέθους του κυτίου. Οι κλίσεις για τα τμήματα μικρών μεγεθών είναι πάντοτε μεταξύ -0.9 και -1. Στην περιοχή μεγάλου μήκους, η έκταση και η κλίση του γραμμικού τμήματος ποικίλουν σημαντικά, και σε αυτήν την περιοχή ακριβώς είναι που αναμένουμε μορφοκλασματικότητα. Η έκταση της γραμμικότητας (F_1, F_2, F_3) και η fractal (ομοιότητα) διάσταση που μετρώνται από τις αντίστοιχες κλίσεις των δύο γραμμικών τμημάτων (D_1, D_2, D_3) φαίνονται στην πιο κάτω εικόνα:

Εικόνα 8: Διαγράμματα με τη μέθοδο του εγκιβωτισμού (box – counting) για τα έξι μεγαλύτερα χρωμοσώματα του ανθρώπου (hg17). Η κατηγορία CNE που μελετάται εδώ είναι εκείνη των «αρχαίων» CNE που είναι συντηρημένα μεταξύ ανθρώπου και *Takifugu rubripes*. Τα γραμμικά τμήματα που φαίνονται παράγονται με τη μέθοδο της γραμμικής παλινδρόμησης. Οι παχιές και στικτές γραμμές απεικονίζουν παρουσία και απουσία μορφοκλασματικότητας αντίστοιχα.



Συμπερασματικά, εδώ μπορούμε να πούμε ότι όσο πιο «αρχαία» είναι τα στοιχεία (CNE) που μελετούμε, τόσο καλύτερες γραμμικότητες παρατηρούμε. Αυτό το αποτέλεσμα έρχεται σε συμφωνία με τα ευρήματά μας ότι τα CNE παρουσιάζουν κατανομές τύπου νόμου δύναμης και ότι ο μακρύς εξελικτικός χρόνος ευνοεί την ωρίμανση και τη δημιουργία, μέσω διαφόρων γεγονότων γονιδιωματικής δυναμικής, των παρατηρούμενων κατανομών (Polychronopoulos, Sellis, et al. 2014).

5. Σύνοψη

Στην παρούσα διατριβή επιχειρούμε να μελετήσουμε υπολογιστικά τις συντηρημένες μη κωδικοποιούσες αλληλουχίες (Conserved Noncoding Elements, CNE). Τα CNE είναι αλληλουχίες που παρουσιάζουν εντυπωσιακή συντήρηση μεταξύ οργανισμών, που συχνά ξεπερνά αυτήν που παρατηρείται στα γονίδια που κωδικοποιούν πρωτεΐνες. Η λειτουργία τους, ωστόσο, ο μηχανισμός που οδήγησε στη δημιουργία τους, καθώς και, εν γένει, ο λόγος της συντήρησής τους σε τόσο υψηλό βαθμό, ανάμεσα στους οργανισμούς, παραμένει ένα μυστήριο που καλείται να λύσει η σύγχρονη βιολογία, χρησιμοποιώντας συνδυασμό μεθόδων, υπολογιστικών και πειραματικών. Αρχικά, αναλύσουμε την χωροταξική οργάνωση των CNE σε γονιδιώματα σπονδυλωτών και ασπόνδυλων με σκοπό να διαπιστώσουμε αν μπορούμε να βγάλουμε κάποια συμπεράσματα για το πώς εξελίχθησαν αυτές οι αλληλουχίες με βάση την κατανομή τους στα χρωμοσώματα. Διαπιστώνουμε ότι οι αποστάσεις αυτών ακολουθούν κατανομές τύπου νόμου δύναμης σε μια ποικιλία γονιδιωμάτων. Δεδομένου ότι τα CNE σχετίζονται χωρικά με γονίδια, ειδικά με αυτά που ρυθμίζουν αναπτυξιακές διαδικασίες, επιβεβαιώσαμε ότι ένα πρότυπο νόμου δύναμης διατηρείται ανεξάρτητα από το εάν συμπεριληφθούν στοιχεία που βρίσκονται μέσα ή εκτός γονιδίων. Προτείνουμε ένα εξελικτικό μοντέλο για την κατανόηση αυτών των ευρημάτων που περιλαμβάνει γεγονότα τμηματικών αναδιπλασιασμών ή αναδιπλασιασμών ολόκληρου του γονιδιώματος και απαλοιφές των περισσοτέρων από τα διπλασιασμένα CNE. Προσομοιώσεις που κάναμε αναπαράγουν τα κύρια χαρακτηριστικά των παρατηρουμένων κατανομών μεγέθους.

Στη συνέχεια παρουσιάζουμε προσπάθειες κατάταξης των στοιχείων αυτών με μεθόδους μηχανικής μάθησης (Γραφήματα N-γραμμάτων, NGG και ανάλυσης κ – μερών, LAF). Τα CNE παρουσιάζουν ενδιαφέρουσες ιδιότητες σύστασης και γι’αυτό προσπαθούσαμε να δούμε αν μπορούν να κατηγοριοποιηθούν με βάση αυτές τους τις ιδίοτητες. Στην προσπάθειά μας αυτή, ταυτοποιούμε νέα σύνολα δεδομένων CNE σπονδυλωτών και ασπόνδυλων. Διαπιστώνουμε ότι και με τις δύο μεθόδους, που για πρώτη φορά αναπτύχθηκαν και εφαρμόστηκαν στα πλαίσια ανάλυσης γονιδιωματικών δεδομένων, είναι εφικτή η κλασμάτωση στοιχείων του γονιδιώματος (CNE, εξώνια) σε διαφορετικές κατηγορίες μεταξύ γονιδιωμάτων ή εντός του ίδιου γονιδιώματος. Περαιτέρω συνεργασίες αναπτύσσονται στη βάση αυτών των μεθοδολογιών γονιδιωματικής, με σκοπό την ανάλυση και ταξινόμηση μεταγονιδιωματικών

δεδομένων, την εύρεση μοτίβων και θέσεων πρόσδεσης μεταγραφικών παραγόντων καθώς και την ανάλυσης ενδεχόμενης διαφορικής πρόσδεσης αυτών στα διάφορα σύνολα δεδομένων CNE σπονδυλωτών και ασπόνδυλων.

Με τη χρήση εργαλείων από τη θεωρία της πληροφορίας (εντροπία Shannon και εγκιβωτισμός) προχωρούμε σε ανάλυση βασικών ιδιοτήτων διαφόρων κατηγοριών CNE, με σκοπό να διαπιστώσουμε την ύπαρξη μορφοκλασματικότητας (fractality). Η συγκεκριμένη εργασία βρίσκεται σε εξέλιξη και, με βάση τα πρώτα αποτελέσματα, επιβεβαιώνεται η υπόθεση ότι η αρχαιότητα αυτών των στοιχείων του γονιδιώματος παίζει σημαντικό ρόλο στην ερμηνεία των προτύπων κατανομής τους και, ενδεχομένως, στην εξέλιξή τους.

Μελετούμε επίσης και άλλα ιδιαίτερα χαρακτηριστικά σύστασης διαφορετικών κατηγοριών CNE που έχουν δημοσιευτεί στη βιβλιογραφία, καθώς και συσχετίσεις που ενδέχεται να προκύπτουν με διάφορες κλάσεις γονιδίων, όπως π.χ. με γονίδια που κωδικοποιούν παράγοντες που παίζουν ρόλο στην ανάπτυξη, νησίδες CpG και γονίδια που κωδικοποιούν ncRNA. Επιπλέον, σχεδιάζουμε πειραματικές προσεγγίσεις ανάλυσης υπερευαισθησίας σε DNAάση, σε συνεργασία με άλλες ομάδες, ώστε να διαπιστώσουμε εάν η κατάσταση της χρωματίνης στα CNE είναι επίσης διακριτή, δηλαδή εάν υπάρχουν διακριτά αποτυπώματα μεταγραφικών παραγόντων, πιο πυκνά σε σχέση με ενισχυτές ελέγχου (control enhancers). Επίσης, δεδομένου ότι τα CNE παίζουν ρυθμιστικό ρόλο σε διάφορες αναπτυξιακές διαδικασίες, παρατηρούνται διαφορές στις καταστάσεις χρωματίνης των διαφόρων κατηγοριών CNE, σε διαφορετικές κυτταρικές σειρές ή σε διαφορετικά αναπτυξιακά στάδια της μύγας; Η χρήση των ποικίλων συνόλων δεδομένων, με διαφορετικά κατώφλια ελάχιστης συντήρησης και ελάχιστου μήκους μεταξύ δύο ή περισσοτέρων οργανισμών, σε διαφορετικά πλαίσια και πειραματικές διατάξεις, ενδεχομένως δώσουν κάποια στοιχεία για τη βιολογία των στοιχείων αυτών του γονιδιώματος.

6. Βιβλιογραφία

Adamic, L.A. & Huberman, B.A., 2002. Zipf's law and the Internet. *Glottometrics*, 3(1), pp.143–150.

Ahituv, N. et al., 2007. Deletion of ultraconserved elements yields viable mice. *PLoS Biol*, 5(9), p.e234.

Akalin, A. et al., 2009. Transcriptional features of genomic regulatory blocks. *Genome Biol*, 10(4), p.R38.

Aloni, R. & Lancet, D., 2005. Conservation anchors in the vertebrate genome. *Genome biology*, 6(7), p.115.

Altschul, S., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17), pp.3389–3402.

Athanasopoulou, L. et al., 2010. Scaling properties and fractality in the distribution of coding segments in eukaryotic genomes revealed through a block entropy approach. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 82(5 Pt 1), p.051917.

Athanasopoulou, L., Sellis, D. & Almirantis, Y., 2014. A Study of Fractality and Long-Range Order in the Distribution of Transposable Elements in Eukaryotic Genomes Using the Scaling Properties of Block Entropy and Box-Counting. *Entropy*, 16(4), pp.1860–1882.

Bailey, J. a et al., 2002. Recent segmental duplications in the human genome. *Science (New York, N.Y.)*, 297(5583), pp.1003–7.

Bejerano, G., Pheasant, M., et al., 2004. Ultraconserved elements in the human genome. *Science*, 304(5675), pp.1321–1325.

Bejerano, G., Haussler, D. & Blanchette, M., 2004. Into the heart of darkness: large-scale clustering of human non-coding DNA. *Bioinformatics*, 20 Suppl 1, pp.i40–8.

Benko, S. et al., 2009. Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. *Nature genetics*, 41(3), pp.359–64.

Bertolazzi, P., Felici, G. & Weitschek, E., 2009. Learning to classify species with barcodes. *BMC bioinformatics*, 10 Suppl 1(Suppl 14), p.S7.

Bishop, C.E. et al., 2000. A transgenic insertion upstream of sox9 is associated with dominant XX sex reversal in the mouse. *Nature genetics*, 26(4), pp.490–4.

Boffelli, D., Nobrega, M.A. & Rubin, E.M., 2004. Comparative genomics at the vertebrate extremes. *Nature reviews. Genetics*, 5(6), pp.456–65.

Brendel, V., Beckmann, J.S. & Trifonov, E.N., 1986. Linguistics of nucleotide sequences: morphology and comparison of vocabularies. *Journal of biomolecular structure & dynamics*, 4(1), pp.11–21.

Clark, A.G., 2001. The search for meaning in noncoding DNA. *Genome research*, 11(8), pp.1319–20.

Clarke, S.L. et al., 2012. Human developmental enhancers conserved between deuterostomes and protostomes. J. Brosius, ed. *PLoS genetics*, 8(8), p.e1002852.

Clauset, A., Shalizi, C.R. & Newman, M.E.J., 2009. Power-Law Distributions in Empirical Data. *SIAM Review*, 51(4), pp.661–703.

Cohen, W.W., 1995. Fast Effective Rule Induction. In *In Proceedings of the Twelfth International Conference on Machine Learning*. Morgan Kaufmann, pp. 115–123.

Cohen, W.W. & Singer, Y., 1999. A simple, fast, and effective rule learner. , pp.335–342.

Couronne, O. et al., 2003. Strategies and tools for whole-genome alignments. *Genome research*, 13(1), pp.73–80.

Dermitzakis, E.T. et al., 2003. Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science*, 302(5647), pp.1033–1035.

Dermitzakis, E.T. et al., 2002. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature*, 420(6915), pp.578–82.

Dermitzakis, E.T., Reymond, A. & Antonarakis, S.E., 2005. Conserved non-genic sequences - an unexpected feature of mammalian genomes. *Nat Rev Genet*, 6(2), pp.151–157.

Deschavanne, P.J. et al., 1999. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Molecular biology and evolution*, 16(10), pp.1391–9.

Dimitrieva, S. & Bucher, P., 2012. Genomic context analysis reveals dense interaction network between vertebrate ultraconserved non-coding elements. *Bioinformatics*, 28(18), pp.i395–i401.

Dimitrieva, S. & Bucher, P., 2013. UCNEbase--a database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic acids research*, 41(Database issue), pp.D101–9.

Drake, J.A. et al., 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat Genet*, 38(2), pp.223–227.

Dubchak, I. et al., 2000. Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome research*, 10(9), pp.1304–6.

Duret, L. & Bucher, P., 1997. Searching for regulatory elements in human noncoding sequences. *Current opinion in structural biology*, 7(3), pp.399–406.

Duret, L., Dorkeld, F. & Gautier, C., 1993. Strong conservation of non-coding sequences during vertebrates evolution: potential involvement in post-transcriptional regulation of gene expression. *Nucleic acids research*, 21(10), pp.2315–22.

Ebeling, W. & Nicolis, G., 1991. Entropy of Symbolic Sequences: The Role of Correlations. *Europhysics Letters (EPL)*, 14(3), pp.191–196.

Ebeling, W. & Nicolis, G., 1992. Word frequency and entropy of symbolic sequences: a dynamical perspective. *Chaos, Solitons & Fractals*, 2(6), pp.635–650.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5), pp.1792–7.

Elgar, G. & Vavouri, T., 2008. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends Genet*, 24(7), pp.344–352.

Emison, E.S. et al., 2005. A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. *Nature*, 434(7035), pp.857–63.

Fayyad, U.M. & Irani, K.B., 1993. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. In R. Bajcsy, ed. *IJCAI*. Morgan Kaufmann, pp. 1022–1029.

Feder, J., 1998. *Fractals*, New York: Plenum Press.

Feng, J. et al., 2006. The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Genes & development*, 20(11), pp.1470–84.

Frank, E. & Witten, I.H., 1998. Generating accurate rule sets without global optimization.

Frankel, N. et al., 2010. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature*, 466(7305), pp.490–3.

Gaines, B.R. & Compton, P., 1995. Induction of ripple-down rules applied to modeling large databases. *Journal of Intelligent Information Systems*, 5(3), pp.211–228.

Ganapathiraju, M. et al., 2002. Comparative n-gram analysis of whole-genome protein sequences. , pp.76–81.

Giannakopoulos G. Vouros G. and Stamatopoulos, P., K. V, 2008. Summarization system evaluation revisited: N-gram graphs. . *ACM Trans. Speech Lang. Process.*, 5(3), pp.1–39.

Gibson, T.J. & Spring, J., 2000. Evidence in favour of ancient octaploidy in the vertebrate genome. *Biochemical Society transactions*, 28(2), pp.259–64.

Glazov, E. a et al., 2005. Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. *Genome research*, 15(6), pp.800–8.

Grassberger, P., 1986. Toward a quantitative theory of self-generated complexity. *International Journal of Theoretical Physics*, 25(9), pp.907–938.

De Grassi, A., Lanave, C. & Saccone, C., 2008. Genome duplication and gene-family evolution: the case of three OXPHOS gene families. *Gene*, 421(1-2), pp.1–6.

Hall, M. et al., 2009. The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, 11(1), p.10.

Harmston, N., Baresic, A. & Lenhard, B., 2013. The mystery of extreme non-coding conservation. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1632), p.20130021.

Hedges, S.B., Dudley, J. & Kumar, S., 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics (Oxford, England)*, 22(23), pp.2971–2.

Hufton, A.L. et al., 2009. Deeply conserved chordate noncoding sequences preserve genome synteny but do not drive gene duplicate retention. *Genome research*, 19(11), pp.2036–51.

Human Genome Sequencing Consortium International, 2004. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), pp.931–45.

Jeong, J. et al., 2008. Dlx genes pattern mammalian jaw primordium by regulating both lower jaw-specific and upper jaw-specific genetic programs. *Development (Cambridge, England)*, 135(17), pp.2905–16.

Karamanos, K., 2001. Entropy analysis of substitutive sequences revisited. *Journal of Physics A: Mathematical and General*, 34(43), pp.9231–9241.

Karamanos, K. & Nicolis, G., 1999. Symbolic Dynamics and Entropy Analysis of Feigenbaum Limit Sets. *Chaos, Solitons & Fractals*, 10(7), pp.1135–1150.

Karlin, S., 1998. Global dinucleotide signatures and analysis of genomic heterogeneity. *Current Opinion in Microbiology*, 1(5), pp.598–610.

Karlin, S. & Mrázek, J., 1997. Compositional differences within and between eukaryotic genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 94(19), pp.10227–32.

Kasahara, M., 2007. The 2R hypothesis: an update. *Current opinion in immunology*, 19(5), pp.547–52.

Katoh, K., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14), pp.3059–3066.

Katzman, S. et al., 2007. Human genome ultraconserved elements are ultraselected. *Science (New York, N.Y.)*, 317(5840), p.915.

Kehrer-Sawatzki, H. & Cooper, D.N., 2008. Molecular mechanisms of chromosomal rearrangement during primate evolution. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 16(1), pp.41–56.

Kent, W.J. et al., 2002. The human genome browser at UCSC. *Genome research*, 12(6), pp.996–1006.

Kikuta, H. et al., 2007. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res*, 17(5), pp.545–555.

Kim, J.Y. & Shawe-Taylor, J., 1994. Fast string matching using an n-gram algorithm. *Software: Practice and Experience*, 24(1), pp.79–88.

Kim, M.-S. et al., 2005. n-gram/2L: a space and time efficient two-level n-gram inverted index structure. , pp.325–336.

Kim, S.Y. & Pritchard, J.K., 2007. Adaptive evolution of conserved noncoding elements in mammals. *PLoS genetics*, 3(9), pp.1572–86.

Kimura, M., 1984. *The Neutral Theory of Molecular Evolution*, Cambridge University Press.

Kirsch, S. et al., 2008. Evolutionary dynamics of segmental duplications from human Y-chromosomal euchromatin/heterochromatin transition regions. *Genome research*, 18(7), pp.1030–42.

Klimopoulos, A., Sellis, D. & Almirantis, Y., 2012. Widespread occurrence of power-law distributions in inter-repeat distances shaped by genome dynamics. *Gene*, 499(1), pp.88–98.

De Kok, Y.J. et al., 1996. Identification of a hot spot for microdeletions in patients with X-linked deafness type 3 (DFN3) 900 kb proximal to the DFN3 gene POU3F4. *Human molecular genetics*, 5(9), pp.1229–35.

Kudenko, D. & Hirsh, H., 1998. Feature generation for sequence categorization. , pp.733–738.

Kuksa, P. & Pavlovic, V., 2009. Efficient alignment-free DNA barcode analytics. *BMC bioinformatics*, 10 Suppl 1(Suppl 14), p.S9.

Kumar, S. & Hedges, S.B., 1998. A molecular timescale for vertebrate evolution. *Nature*, 392(6679), pp.917–20.

De la Calle-Mustienes, E. et al., 2005. A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. *Genome research*, 15(8), pp.1061–72.

Lee, A.P. et al., 2011. Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes. *Mol Biol Evol*, 28(3), pp.1205–1215.

Lehr, T. et al., 2011. Rule based classifier for the analysis of gene-gene and gene-environment interactions in genetic association studies. *BioData mining*, 4(1), p.4.

Lettice, L.A. et al., 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet*, 12(14), pp.1725–1735.

Li, Q. et al., 2010. A systematic approach to identify functional motifs within vertebrate developmental enhancers. *Developmental biology*, 337(2), pp.484–95.

Li, W., 1991. Expansion-modification systems: A model for spatial 1/f spectra. *Physical review. A*, 43(10), pp.5240–5260.

Li, W., 2002. Zipf's law everywhere. *Glottometrics*, 5, pp.14–21.

Li, W. & Kaneko, K., 1992. DNA correlations. *Nature*, 360(6405), pp.635–6.

Lieberman-Aiden, E. et al., 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, 326(5950), pp.289–93.

Lindblad-Toh, K. et al., 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370), pp.476–82.

Lindblad-Toh, K. et al., 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438(7069), pp.803–819.

Lockton, S. & Gaut, B.S., 2005. Plant conserved non-coding sequences and paralogue evolution. *Trends Genet*, 21(1), pp.60–65.

Lowe, C.B. et al., 2011. Three periods of regulatory innovation during vertebrate evolution. *Science (New York, N.Y.)*, 333(6045), pp.1019–24.

Lynch, M., 2000. The Evolutionary Fate and Consequences of Duplicate Genes. *Science*, 290(5494), pp.1151–1155.

Maas, S.A., Suzuki, T. & Fallon, J.F., 2011. Identification of spontaneous mutations within the long-range limb-specific Sonic hedgehog enhancer (ZRS) that alter Sonic hedgehog expression in the chicken limb mutants oligozeugodactyly and silkie breed. *Developmental dynamics : an official publication of the American Association of Anatomists*, 240(5), pp.1212–22.

Mandelbrot, B.B., 1982. *The fractal geometry of nature*, San Francisco: W.H. Freeman.

Mantegna, R.N. et al., 1995. Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics. *Physical review. E, Statistical physics, plasmas, fluids, and related interdisciplinary topics*, 52(3), pp.2939–50.

Marçais, G. & Kingsford, C., 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics (Oxford, England)*, 27(6), pp.764–70.

Margulies, E.H. et al., 2003. Identification and characterization of multi-species conserved sequences. *Genome Res*, 13(12), pp.2507–2518.

Martignetti, L. & Caselle, M., 2007. Universal power law behaviors in genomic sequences and evolutionary models. *Physical Review E*, 76(2), p.021902.

Matsunami, M. & Saitou, N., 2013. Vertebrate paralogous conserved noncoding sequences may be related to gene expressions in brain. *Genome biology and evolution*, 5(1), pp.140–50.

Matsunami, M., Sumiyama, K. & Saitou, N., 2010. Evolution of conserved non-coding sequences within the vertebrate Hox clusters through the two-round whole genome duplications revealed by phylogenetic footprinting analysis. *J Mol Evol*, 71(5-6), pp.427–436.

McCormack, J.E. et al., 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome research*, 22(4), pp.746–54.

McEwen, G.K. et al., 2006. Ancient duplicated conserved noncoding elements in vertebrates: a genomic and functional analysis. *Genome Res*, 16(4), pp.451–465.

McLean, C. & Bejerano, G., 2008. Dispensability of mammalian DNA. *Genome research*, 18(11), pp.1743–51.

McLysaght, A. et al., 2000. Estimation of synteny conservation and genome compaction between pufferfish (Fugu) and human. *Yeast (Chichester, England)*, 17(1), pp.22–36.

Mikkelsen, T.S. et al., 2007. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. *Nature*, 447(7141), pp.167–177.

Mokaddeml, A. & Elloumi, M., 2013. Motalign: A Multiple Sequence Alignment Algorithm Based on a New Distance and a New Score Function. In *2013 24th International Workshop on Database and Expert Systems Applications*. IEEE, pp. 81–84.

Needleman, S.B. & Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3), pp.443–453.

Nelson, A.C. & Wardle, F.C., 2013. Conserved non-coding elements and cis regulation: actions speak louder than words. *Development (Cambridge, England)*, 140(7), pp.1385–95.

Newman, M.E.J., 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5), pp.323–351.

Nicolis, G. & Gaspard, P., 1994. Toward a probabilistic approach to complex systems. *Chaos, Solitons & Fractals*, 4(1), pp.41–57.

Nobrega, M.A. et al., 2003. Scanning human gene deserts for long-range enhancers. *Science (New York, N.Y.)*, 302(5644), p.413.

Nóbrega, M.A. et al., 2004. Megabase deletions of gene deserts result in viable mice. *Nature*, 431(7011), pp.988–93.

Nolte, M.J. et al., 2014. Functional analysis of limb transcriptional enhancers in the mouse. *Evolution & development*, 16(4), pp.207–23.

Ohno, S. & others, 1970. *Evolution by gene duplication.*, London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag.

Pearson, W.R., 1990. *[5] Rapid and sensitive sequence comparison with FASTP and FASTA*,

Pedersen, J.S. et al., 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS computational biology*, 2(4), p.e33.

Peng, C.K. et al., 1992. Long-range correlations in nucleotide sequences. *Nature*, 356(6365), pp.168–70.

Pennacchio, L.A. et al., 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118), pp.499–502.

Perry, M.W. et al., 2010. Shadow enhancers foster robustness of Drosophila gastrulation. *Current biology : CB*, 20(17), pp.1562–7.

Polychronopoulos, D., Krithara, A., et al., 2014. Analysis and Classification of Constrained DNA Elements with N-gram Graphs and Genomic Signatures. In A.-H. Dediu, C. Martín-Vide, & B. Truthe, eds. *Algorithms for Computational Biology SE - 18*. Lecture Notes in Computer Science. Springer International Publishing, pp. 220–234.

Polychronopoulos, D., Weitschek, E., et al., 2014. Classification of selectively constrained DNA elements using feature vectors and rule-based classifiers. *Genomics*, 104(2), pp.79–86.

Polychronopoulos, D., Sellis, D. & Almirantis, Y., 2014. Conserved Noncoding Elements Follow Power-Law-Like Distributions in Several Genomes as a Result of Genome Dynamics C. A. Ouzounis, ed. *PLoS ONE*, 9(5), p.e95437.

Pruitt, K.D., Tatusova, T. & Maglott, D.R., 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 35(Database issue), pp.D61–5.

Quinlan, A.R. & Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), pp.841–842.

Quinlan, J.R., 1993. C4.5: programs for machine learning.

Rahimov, F. et al., 2008. Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. *Nature genetics*, 40(11), pp.1341–7.

Retelska, D. et al., 2007. Vertebrate conserved non coding DNA regions have a high persistence length and a short persistence time. *BMC Genomics*, 8, p.398.

Rice, P., Longden, I. & Bleasby, A., 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics : TIG*, 16(6), pp.276–7.

Ritter, D.I. et al., 2010. The importance of being cis: evolution of orthologous fish and mammalian enhancer activity. *Molecular biology and evolution*, 27(10), pp.2322–32.

Royo, J.L. et al., 2011. Dissecting the transcriptional regulatory properties of human chromosome 16 highly conserved non-coding regions. B. Mellone, ed. *PloS one*, 6(9), p.e24824.

Sabherwal, N. et al., 2007. Long-range conserved non-coding SHOX sequences regulate expression in developing chicken limb and are associated with short stature phenotypes in human patients. *Hum Mol Genet*, 16(2), pp.210–222.

Sakuraba, Y. et al., 2008. Identification and characterization of new long conserved noncoding sequences in vertebrates. *Mamm Genome*, 19(10-12), pp.703–712.

Salerno, W., Havlak, P. & Miller, J., 2006. Scale-invariant structure of strongly conserved sequence in genomic intersections and alignments. *Proc Natl Acad Sci U S A*, 103(35), pp.13121–13125.

Sandelin, A. et al., 2004. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics*, 5(1), p.99.

Sanges, R. et al., 2013. Highly conserved elements discovered in vertebrates are present in non-syntenic loci of tunicates, act as enhancers and can be transcribed during development. *Nucleic acids research*, 41(6), pp.3600–18.

Sellis, D. & Almirantis, Y., 2009. Power-laws in the genomic distribution of coding segments in several organisms: an evolutionary trace of segmental duplications, possible paleopolyploidy and gene loss. *Gene*, 447(1), pp.18–28.

Sellis, D., Provata, A. & Almirantis, Y., 2007. Alu and LINE1 distributions in the human chromosomes: evidence of global genomic organization expressed in the form of power laws. *Molecular biology and evolution*, 24(11), pp.2385–99.

Sémon, M. & Wolfe, K.H., 2007. Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. *Trends in genetics*, 23(3), pp.108–12.

Shannon, C.E., 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3), pp.379–423.

Shin, J.T. et al., 2005. Human-zebrafish non-coding conserved elements act in vivo to regulate transcription. *Nucleic Acids Res*, 33(17), pp.5437–5445.

Siepel, A. et al., 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8), pp.1034–1050.

Smith, T.F. & Waterman, M.S., 1981. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1), pp.195–197.

Stephen, S. et al., 2008. Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. *Mol Biol Evol*, 25(2), pp.402–408.

Stumpf, M.P.H. & Porter, M.A., 2012. Mathematics. Critical truths about power laws. *Science (New York, N.Y.)*, 335(6069), pp.665–6.

Sun, H., Skogerbø, G. & Chen, R., 2006. Conserved distances between vertebrate highly conserved elements. *Human molecular genetics*, 15(19), pp.2911–22.

Takayasu, H. et al., 1991. Statistical properties of aggregation with injection. *J. Stat. Phys*, 65, pp.725–745.

Teeling, H. et al., 2004. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC bioinformatics*, 5(1), p.163.

The ENCODE Project Consortium, 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (New York, N.Y.)*, 306(5696), pp.636–40.

Thomas, J.W. et al., 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, 424(6950), pp.788–93.

Thompson, J.D., Gibson, T.J. & Higgins, D.G., 2002. Multiple sequence alignment using ClustalW and ClustalX. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, Chapter 2, p.Unit 2.3.

Tseng, H.-H.E. & Tompa, M., 2009. Algorithms for locating extremely conserved elements in multiple sequence alignments. *BMC bioinformatics*, 10(1), p.432.

Vavouri, T. et al., 2006. Defining a genomic radius for long-range enhancer action: duplicated conserved non-coding elements hold the key. *Trends in genetics : TIG*, 22(1), pp.5–10.

Vavouri, T. et al., 2007. Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. *Genome Biol*, 8(2), p.R15.

Vavouri, T. & Elgar, G., 2005. Prediction of cis-regulatory elements using binding site matrices--the successes, the failures and the reasons for both. *Current opinion in genetics & development*, 15(4), pp.395–402.

Venkatesh, B. et al., 2006. Ancient noncoding elements conserved in the human genome. *Science*, 314(5807), p.1892.

Vinga, S. & Almeida, J., 2003. Alignment-free sequence comparison--a review. *Bioinformatics*, 19(4), pp.513–523.

Visel, A. et al., 2008. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet*, 40(2), pp.158–160.

Viturawong, T. et al., 2013. A DNA-Centric Protein Interaction Map of Ultraconserved Elements Reveals Contribution of Transcription Factor Binding Hubs to Conservation. *Cell reports*, 5(2), pp.531–45.

Voss, R., 1992. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Physical review letters*, 68(25), pp.3805–3808.

Walter, K. et al., 2005. Striking nucleotide frequency pattern at the borders of highly conserved vertebrate non-coding sequences. *Trends in genetics : TIG*, 21(8), pp.436–40.

Wang, J. et al., 2009. Large number of ultraconserved elements were already present in the jawed vertebrate ancestor. *Molecular biology and evolution*, 26(3), pp.487–90.

Ward, L.D. & Kellis, M., 2012. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science (New York, N.Y.)*, 337(6102), pp.1675–8.

Weitschek, E. et al., 2012. Human polyomaviruses identification by logic mining techniques. *Virology journal*, 9, p.58.

Weitschek, E., Fiscon, G. & Felici, G., 2014. Supervised DNA Barcodes species classification: analysis, comparisons and results. *BioData mining*, 7(1), p.4.

Woolfe, A. et al., 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol*, 3(1), p.e7.

Woolfe, A. & Elgar, G., 2008. Chapter 12 Organization of Conserved Elements Near Key Developmental Regulators in Vertebrate Genomes. *Adv Genet*, 61, pp.307–338.

Xie, X. et al., 2007. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci U S A*, 104(17), pp.7145–7150.

Xie, X., Kamal, M. & Lander, E.S., 2006. A family of conserved noncoding elements derived from an ancient transposable element. *Proc Natl Acad Sci U S A*, 103(31), pp.11659–11664.

Xing, Z., Pei, J. & Keogh, E., 2010. A brief survey on sequence classification. *ACM SIGKDD Explorations Newsletter*, 12(1), p.40.

7. Σύντομο βιογραφικό σημείωμα με δημοσιεύσεις

Dimitris Polychronopoulos
Date and place of birth: 30/5/1984, Athens, Greece
13-15 Delfon street, Gerakas, Athens, 15344, Greece
Citizenship: Greek – Australian
URL: http://scholar.google.com/citations?user=LsI4gg0AAAAJ&hl=en
Phone: +30 6933 258 658 – +30 210 66 13 347, E-mail: dpolychr@gmail.com

Education

University of Athens & NCSR "Demokritos" 2014
DPhil in Biology, Department of Biology
National and Kapodistrian University of Athens 2011
MSc in Bioinformatics, Department of Biology
Democritus University of Thrace 2006
4-year Diploma in Molecular Biology and Genetics

Training

European Molecular Biology Organisation (EMBO)

- EMBO Practical Course "Computational Biology: Genomes, Cells & Systems", 6-13 August 2011, Iceland (also supported by an EMBO travel grant)
- EMBO Practical Course "Bioinformatics and Comparative Genome Analysis", 7-19 May 2012, Naples, Italy (supported by EMBO)
- EMBO Short term fellowship (ASTF No: 453 - 2012) to conduct research at the Swiss Federal Institute of Technology, EPFL/ISREC, Switzerland for 3 months (January – April 2013, Philipp Bucher's lab)
- FEBS Short term fellowship to conduct research at the Max Delbruck Center for Molecular Medicine, Berlin, Germany for 2 months (Uwe Ohler's lab)

Funding

- BioASQ (Funded by FP7 EC; 2012 – 2014; www.bioasq.org) – A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. EUROPEAN PROJECT FP7 ICT (2012-14). **Total Budget:** 1.2M Euros; **Role:** Coordinator of the team of biomedical experts in collaboration with my PhD supervisor Dr.Y.Almirantis.
- THALIS (Co-Funded by European Union & National Funds; 2012-2015) – Genome sequencing and characterization of Streptococcus macedonicus, Streptococcus thermophilus, Lactobacillus delbrueckii subsp. lactis and Lactobacillus acidipiscis. Physiological, evolutionary and technological implications. **Total Budget:** 0.5M Euros; **Role:** Collaborating researcher
- EMBO short term fellowship (Funded by the European Molecular Biology Organization; Jan – Apr 2013; ASTF No 453 – 2012) – Computational Analysis of Conserved Non-coding Elements. **Budget:** 12K Euros; **Role:** Scholar; Philipp Bucher's group (http://bucher-lab.epfl.ch/)
- FEBS short term fellowship (Funded by the Federation of European Biochemical Societies) – Transcription Factor Occupancy and Chromatin States Analysis of UltraConserved Noncoding Elements (UCNEs). **Budget:** 3.8K Euros; **Role:** Scholar; Uwe Ohler's group (https://ohlerlab.mdc-berlin.de/)

Publications

Journals (3)

- Gambus A*, van Deursen F*, **Polychronopoulos D**, Foltman M, Jones RC, Edmondson RD, Calzada A and Labib K: A key role for Ctf4 in coupling the MCM2-7 helicase to DNA polymerase alpha within the eukaryotic replisome. **EMBO J**. 2009 Oct 7;28(19):2992-3004. Epub 2009 Aug 6 (*equal contribution)
- **Polychronopoulos D**, Sellis D and Almirantis Y: Conserved Noncoding Elements Follow Power-Law-Like Distributions in Several Genomes as a Result of Genome Dynamics. **PLoS One** 2014 May 2;9(5):e95437. doi: 10.1371/journal.pone.0095437. ECollection 2014.
- **Polychronopoulos D*,** Weitschek E*, Dimitrieva S, Bucher P, Felici G and Almirantis Y: Classification of selectively constrained DNA elements using feature vectors and rule-based classifiers. **Genomics** 2014 Jul 22. pii: S0888-7543(14)00119-0. doi: 10.1016/j.ygeno.2014.07.004. [Epub ahead of print] (*equal contribution)

Peer-reviewed conferences (1)

- **D. Polychronopoulos**, A. Krithara, C. Nikolaou, G. Paliouras, Y. Almirantis and G. Giannakopoulos. Analysis and Classification of Constrained DNA Elements with N-gram Graphs and Genomic Signatures. 1st International Conference on Algorithms for Computational Biology (AlCoB), Tarragona, Spain, **Lecture Notes in Computer Science (LNCS/LNBI)**, July 2014.

Under peer-review (2)

- George Tsatsaronis, Ioannis Partalas, George Balikas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, **Dimitris Polychronopoulos**, Yannis Almirantis, Prodromos Malakasiotis, Ion Androutsopoulos, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artieres, Axel Ngonga, Norman Heino, Eric Gaussier, Lilianna Barrio-Alvers, Michael Schroeder and Georgios Paliouras: An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition. (revisions, **BMC Bioinformatics**)
- Zegos D., Papageorgiou D., Amaral-Psarris A., Xiucheng Fan A., Kastrissianaki E., **Polychronopoulos D.**, Kerdidani D., Bungert J., Strouboulis J. Mammalian expression vectors for the N- or C-terminal metabolic biotinylation tandem affinity tagging of proteins by the E. coli BirA biotin ligase. (revisions, **Journal of Biotechnology**).

Book Chapters (1)

- **Polychronopoulos D**, Tsiangas G, Athanasopoulou L, Sellis D and Almirantis Y: Long-Range order and fractality in the structure and organization of eukaryotic genomes. **Chaos, Information Processing and Paradoxical Games** (The Legacy of John S Nicolis) URL: http://www.worldscientific.com/worldscibooks/10.1142/9145

Programming Languages – Computational skills

- Microsoft Office, Windows 3.1/2000/XP/Vista/7, Linux, MacOS
- C, Perl, Python, JAVA, awk, R programming language, shell scripting
- EMBOSS Utilities, UCSC utilities, BEDTools, samtools, biosquid, WEKA

Language Skills

- Greek (Native), English (fluent), German (fair), Italian (beginner)

Reviewer for international journals

- Artificial Intelligence in Medicine (Elsevier, IF: 1.356)
- PLoS One (Public Library of Science, IF: 3.534)

OPEN ACCESS Freely available online

PLOS ONE

Conserved Noncoding Elements Follow Power-Law-Like Distributions in Several Genomes as a Result of Genome Dynamics

Dimitris Polychronopoulos[1,2], Diamantis Sellis[3], Yannis Almirantis[1]*

1 Institute of Biosciences and Applications, National Center for Scientific Research "Demokritos", Athens, Greece, **2** Department of Biochemistry and Molecular Biology, Faculty of Biology, National and Kapodistrian University of Athens, Athens, Greece, **3** Department of Biology, Stanford University, Stanford, California, United States of America

Abstract

Conserved, ultraconserved and other classes of constrained elements (collectively referred as CNEs here), identified by comparative genomics in a wide variety of genomes, are non-randomly distributed across chromosomes. These elements are defined using various degrees of conservation between organisms and several thresholds of minimal length. We here investigate the chromosomal distribution of CNEs by studying the statistical properties of distances between consecutive CNEs. We find widespread power-law-like distributions, i.e. linearity in double logarithmic scale, in the inter-CNE distances, a feature which is connected with fractality and self-similarity. Given that CNEs are often found to be spatially associated with genes, especially with those that regulate developmental processes, we verify by appropriate gene masking that a power-law-like pattern emerges irrespectively of whether elements found close or inside genes are excluded or not. An evolutionary model is put forward for the understanding of these findings that includes *segmental or whole genome duplication* events and *eliminations (loss)* of most of the duplicated CNEs. Simulations reproduce the main features of the observed size distributions. Power-law-like patterns in the genomic distributions of CNEs are in accordance with current knowledge about their evolutionary history in several genomes.

Citation: Polychronopoulos D, Sellis D, Almirantis Y (2014) Conserved Noncoding Elements Follow Power-Law-Like Distributions in Several Genomes as a Result of Genome Dynamics. PLoS ONE 9(5): e95437. doi:10.1371/journal.pone.0095437

Editor: Christos A. Ouzounis, The Centre for Research and Technology, Hellas, Greece

Received November 20, 2013; **Accepted** March 26, 2014; **Published** May 2, 2014

Funding: Publication fees for this work were covered in the framework of "Target Identification for Disease Diagnosis and Treatment (DIAS)" project within General Secretariat for Research and Technology (GSRT) Development Proposals of Research Organisations-KRIPIS action, funded by Greece and the European Regional Development Fund of the European Union under the O.P. Competitiveness and Entrepreneurship, National Strategic Reference Framework (NSRF) 2007-2013. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: yalmir@bio.demokritos.gr

Introduction

The sequencing and comparative analysis of many mammalian genomes has indicated that at least 5.5% of the human genome is under selective constraint; of that, 1.5% is estimated to code for proteins, 3.5% displays known regulatory functions, while for the function of the rest there is little or no information available [1]. One of the most interesting findings that have arisen from comparative analysis among mammalian genomes is the discovery of hundreds of ultraconserved elements (UCEs) of more than 200 bp in length that show absolute conservation among human, mouse and rat genomes [2]. One out of four of UCEs overlaps known protein-coding genes. However, such a high degree of conservation (100%) is not expected even in exons, due to the degeneration of the genetic code. Since the discovery of UCEs, there have been efforts to identify conserved elements based on lower thresholds of sequence similarity over whole genome alignments of two or more species. Several thresholds of minimal length of conserved sequence have been used as well as the exclusion of elements inside protein-coding genes [3,4]. Throughout this article, we use the term CNE(s) for Conserved Noncoding Elements to describe all such elements despite their specific characterization as UCEs, UCNEs, HCNEs, CNGs, CNEs etc in the related literature. We here use the specific name only when we refer to the corresponding class of elements.

CNEs are not a vertebrate innovation but are also found in invertebrate and plant genomes [5–7]. The vertebrate, insect and nematode CNEs are not related to each other at the sequence level [6,8,9]. However, a recent study has identified two elements conserved between vertebrates and invertebrates [10] and it is possible that more will be identified in the near future with the advent on new sequencing methodologies and the increasing availability of sequenced genomes. In the relatively recent evolution of vertebrates, the mean length and conservation of CNEs found therein are the highest observed [11] regarding all taxonomic groups, while the conjectured roles they have acquired are particularly important [12].

CNEs are often clustered in the vicinity of genes involved in transcriptional regulation and/or development [13–15]. Using microarray analysis it was reported that a large fraction of noncoding UCEs have tissue-specific expression levels and are deregulated in human cancer [16,17]. When such elements are located in the vicinity of genes, these genes are invariably found in conserved synteny in all vertebrates, possibly due to the fact that the surrounding genomic environment of a regulation-dependent gene has to be maintained intact [18]. Gene deserts are usually

enriched in CNEs [19,20] while, in mammalian genomes, the vast majority of those elements are found at long distances from the closest genes, exceeding in some cases 2 Mb, which is the limit for any known cis regulatory element [18,21,22]. Little is known or could be speculated about what those distant CNEs actually do. Published studies tend to support the idea that they might be an essential part of Gene Regulatory Blocks (GRBs) and that they could function in a cooperative way alongside with their target genes [4,23–25].

There is a corpus of literature suggesting that CNEs are selectively constrained and not mutational cold spots [26,27]. Studies showing that CNEs might act as transcriptional regulators, e.g. enhancers or insulators, have been published [28,29], although *in vivo* experiments of elimination of some of these elements yield viable mice [30]. A CNE from one species may drive expression in another species as shown by transgenics experiments [31,32], although this is not a demonstration of whether a particular CNE drives conserved expression. Experiments that have been performed in order to test the same CNE in multiple species or the same CNE from multiple species in one species, have shown that although CNEs can be identified using sequence conservation criteria, the expression patterns they drive across species may show little conservation [33–36]. Another aspect not directly addressed herein is the existence of paralogous CNEs in vertebrate genomes. These are believed to often remain conserved having the possibility of controlling overlapping expression patterns of their adjacent paralogous protein-coding genes [37]. Paralogous CNEs are involved in the gene expression pattern of the vertebrate brain [38].

The alternative hypothesis that CNEs are horizontally transferred between lineages and accumulate during the course of long-term evolution has also been expressed [39]. Furthermore, a study has suggested that CNEs might act as Matrix-Attachment Regions (MARs) by serving as sequences that regulate the architecture of chromatin through specific binding of particular proteins [40]. An association between CNEs and phenotypic variation and disease has also been reported [41–43].

Long-range correlations were reported in the nucleotide sequence of the non-protein-coding part of eukaryotic genome soon after such large sequences became available [44–46]. In previous works, we explored the large-scale features of several classes of genomic elements, such as protein coding segments [47,48] and transposable elements [49,50], by studying the size distribution of inter-exon and inter-repeat distances. In most cases we found power-law-like size distributions, fractality and self-similarity, often spanning several orders of magnitude. We here apply the same methodology for the analysis of inter-CNE distances. We use published datasets, which are characterized by different degrees of evolutionary conservation, identified in a wide variety of organisms spanning vertebrates and invertebrates. We detect power-law-like inter-CNE size distributions in most cases studied. A previous study from Salerno *et al.* [51] reported the existence of a power-law distribution in the length of "perfectly conserved" sequence from mouse/human whole-genome intersection and alignment. The work we present here focuses on the *distances* of consecutive CNEs (inter-CNE spacers) for which we also propose an explanatory model. The model that we propose cannot apply to the length distribution of CNEs themselves, thus the finding of Salemo *et al.* appears to be the expression of an independent phenomenon.

Given the aforementioned detection of long-range correlations in the nucleotide juxtaposition in non-constrained sequences of the eukaryotic genome, simple molecular dynamics have been used in attempts to explain this emergent pattern. A simple expansion - modification system is shown to generate long-range correlations through the interplay of symbol duplication and symbol elimination events [52]. More recent findings on strand slippage during replication combined with point mutations shed light into homonucleotide tracts and microsatellites' evolution and may be a realistic implementation of the above model to genome dynamics (see Athanasopoulou *et al.* [48], where a short discussion about evolutionary scenaria generating long-range correlations is included). Here we implement a model (initially proposed in [47]) for the generation of the observed power-law-like distribution pattern of distances between evolutionary constrained genomic elements in general (protein-coding segments and CNEs). This evolutionary scenario is based on an earlier model accounting for the explanation of power-law size distributions appearing in aggregative growth of particles in physicochemical systems [53]. This mechanism, as applied in genome evolution herein, mainly involves segmental duplication (including whole genome duplication events) and loss of most of the duplicated CNEs, alongside a moderated loss of non-duplicated CNEs in some cases.

Methods and Materials

Datasets

We systematically investigate the chromosomal distribution of various CNEs. We include in our analysis a phylogenetically wide collection of datasets, ranging from human to elephant shark and from vertebrates to invertebrates:

(i) 13,736 CNEs mapped on the human genome (hg18), of various lengths, that are identical over at least 100 bp in at least 3 of 5 placental mammals (human, mouse, rat, dog and cow) [20]. The whole set is named EU100+. Specific subsets are also considered for our purposes as follows (data kindly provided by J.S. Mattick, see also Table 1): **(ia)** 8,332 EU100+ elements that are not present in fish (Fugu). These appeared during tetrapod evolution (present in frog, chicken and/or mammals) and are named EU-FR. **(ib)** 5,404 elements from EU100+ set with orthologs in fish (ancient). These are named FR. **(ic)** 1,665 elements that are present in frog but not in fish (tetrapod speciation). These are named XT-FR. **(id)** 980 elements that are present in chicken but not in frog or fugu (amniote speciation). These are named GG-XT-FR. **(ie)** 600 elements that are not present in chicken, or frog, or fugu (mammalian speciation). These are named EU-GG-XT-FR.

(ii) 82,335 Mammalian CNEs (conserved within mammals but not found in chicken or fish) and 16,575 Amniotic CNEs (conserved in mammals and chicken but not found in fish) respectively, mapped on the human genome (hg17) [19].

(iii) 4,386 UCNEs (Ultraconserved Noncoding Elements, longer than 200bp) mapped on the human genome (hg19) that display sequence identity which is consistently greater or equal to 95% between human and chicken whole genome alignments [24].

(iv) 3,124 Human – Fugu conserved noncoding elements mapped on the human genome (hg17) with 70% identity and a score of match-mismatch up to 60 [54].

(v) 2,833 Human – Zebrafish CNEs mapped on the human genome (hg17) that display identity greater than 70% over at least 80 bp [32].

(vi) 4,782 Human – Elephant Shark CNEs mapped on the human genome (hg17), with identity ranging from 71% to 98% [55].

(vii) 4,519 PCNEs (Phylogenetically CNEs) mapped on the zebrafish genome (genome-build Ensembl 42) that are conserved across amphioxus, zebrafish, mouse and fugu [56]. These elements are unique due to the way of their identification, which is not biased by rearrangement and duplication. In addition to that, local similarity searches (versus whole genome alignments) in the genomic regions surrounding phylogenetically defined gene families have been employed in order to detect them.

(viii) 23,651 *D. melanogaster* – *D. pseudoobscura* (insect) CNEs of 50 bp or more that are 100% conserved between these two species, mapped on the *D. melanogaster* genome (dm1) [6].

(ix) 2,082 Nematode (worm) CNEs with mean identity of 96% between *C. elegans* and *C. briggsae* mapped on genome WS140 [5].

(x) 2,614 Noncoding elements marked by extreme human-mouse-rat constraint (mapped on hg17), a subset of which act as developmental enhancers [57].

For specific details about the used data sets and the subsequent treatment see File S1. In most cases, the suite of utilities BEDTools has been used for the computational analysis [58]

Gene and CDS masking

We proceed to a complete masking of the regions characterized as genic in the human genome (hg17 and hg18). In addition, we mask flanks surrounding every gene: 5 kb at the 5' end and 2 kb at the 3' end, in order to exclude cis-regulatory elements the localization of which may be principally determined by the positioning of the regulated gene. The region located upstream of transcription start sites is usually particularly enriched in such regulatory sequences. Extended flanks of 10 kb and 100 kb have also been masked in a similar manner (see Results section). We use custom scripts and BEDTools in order to perform the masking. In the case of *D. melanogaster*, when we refer to masked CNEs of insect origin, we refer to elements that do not overlap exonic sequences and splice sites, as adopted from the supplementary material of Glazov *et al.* [6]. For masked genes' genomic coordinate data (file format and availability) see in the File S1.

We do not proceed to the masking of other genomic components, such as transposable elements (TEs), for which there are indications that they do follow power-law-like distributions, because there is no evidence about CNE – TE functional interaction or systematic co-localization. Only a tiny proportion of TEs is reported to have been exapted to the role of a CNE, but they are too few to influence and reshape the whole CNE distribution [59].

Size distributions

Suppose there is a large collection of n objects (in our case spacers between CNEs), each characterized by its length S. In typically random such collections (like runs of heads in a coin tossing experiment) we can approximate the distribution of sizes with an exponential distribution. Let p(S) the probability of a spacer having length between S-s/2 and S+s/2 (where s is the size of the bin width) and N*(S) the number of spacers:

Table 1. Summary characterization of genomes for several datasets: Power-law-like distributions of inter-CNE distances at chromosomal scale.

Dataset	Class of CNEs	Unmasked		Masked		Reference genome
		Average Extent (avg E)	Average Extent of five 'best' chr. (avg E-5)	Average Extent (avg E)	Average Extent of five 'best' chr. (avg E-5)	
i	EU100+	**2**	**2.38**	**2.48**	**2.98**	hg18
ia	EU-FR	**1.97**	**2.46**			hg18
ib	FR	**2.2**	**2.72**			hg18
ic	XT-FR	**2.32**	**2.32**			hg18
id	GG-XT-FR	**2.14**	**2.14**			hg18
ie	EU-GG-XT-FR	**1.96**	**1.96**			hg18
iia	Mammalian	**1.49**	**1.9**	**1.59**	**2.04**	hg17
iib	Amniotic	**2.2**	**2.86**	**2.25**	**2.91**	hg17
iii	Human/Chicken	**2.35**	**2.63**			hg19
iv	Human/Fugu	**2.78**	**3.26**			hg17
v	Human/Zebrafish	**2.46**	**2.98**	**2.31**	**2.31**	hg17
vi	Human/El. shark	**2.36**	**2.69**	**2.42**	**2.68**	hg17
vii	*D. rerio* PCNEs	**2.43**	**3.01**			#
viii	Insect CNEs	**1.23**	**1.23**	**1.42**	**1.42**	dm1
ix	Worm CNEs	**1.7**	**1.7**			WS140
x	Human/Rodents	**2.15**	**2.43**			hg17

Propensity for the formation of power-law-like size distributions of the inter-CNE distances as quantified by the extent (E) of linearity in log-log scale. Average values of E for all chromosomes (avg E) and average values of E for 5 chromosomes with the largest E (avg E-5) in each genome are presented. Gene-masked genomes are also included when available (for details see in the text).
#: genome-build Ensembl 42 (zebrafish).
doi:10.1371/journal.pone.0095437.t001

$$\mathbf{N^*(S) = np(S) \propto e^{-\alpha S} \quad \alpha > 0} \tag{1}$$

When scale-free clustering appears, long-range correlations extend to several length scales (ideally, in our case for the whole examined genomic length) and the spacers' size distributions follow a so-called power-law, which corresponds to a linear graph in a double logarithmic scale:

$$\mathbf{N^*(S) = np(S) \propto S^{-\zeta} = S^{-1-\mu} \quad \mu > 0} \tag{2}$$

In this article we use the "cumulative size distribution", more precisely: the complementary cumulative distribution function [60], defined as follows:

$$\mathbf{P(S) = \int_S^\infty p(r)dr} \tag{3}$$

where p(r) is the original spacers' size distribution. The cumulative distribution has in general better statistical properties, as it forms smoother "tails", less affected by statistical fluctuations. Also, by definition it is independent of any binning choice: in a cumulative curve the value of P(S) for length S is not associated with the subset of spacers whose length falls in the same bin, as in the original distribution, but it corresponds to the number of all spacers longer than S. For reviews on power-law size distributions, their properties and alternative forms see e.g. [60–64].

The cumulative form of a power-law size distribution is again a power-law characterized by an exponent (slope) equal to that of the original distribution minus 1: if $\mathbf{p(r) \propto r^{-1-\mu}}$, then

$$\mathbf{N(S) = nP(S) \propto \int_S^\infty (r^{-1-\mu})dr \propto S^{-\mu}} \tag{4}$$

where N(S) is the number of spacers longer or equal to S. All the distribution plots presented in this article depict complementary cumulative size distributions of distances (spacers) between consecutive CNEs. The logarithms of these spacers' length (S) are shown in the horizontal axis and the logarithms of the number N(S) of all the spacers longer or equal to S are shown on the vertical axis.

The slope for a typical power-law does not exceed the value of $\mu = 2$, as $\mu < 2$ is a condition leading to a non-convergent standard deviation. In the power-law-like linearities reported in what follows the value of μ is always below 2. Power-law-like distributions in nature always have an upper and a lower cutoff, which determine the linear region in log-log scale, where self-similarity and fractality is observed. The extent of the linear region (E) measures the orders of magnitudes that the fractal geometry spans. Linearity has been determined by linear regression and the associated value of r^2 is in all cases higher than 0.97 and in more than 90% of the cases higher than 0.98.

Additionally to genomic spacers' size distributions, all figures also include a bundle of ten surrogate simulated size distributions (continuous lines) where markers representing CNEs are randomly positioned in a sequence. The number of the randomly positioned markers and the length of the simulated sequence are equal to the number of CNEs and the size of the considered chromosome respectively. The inclusion of these random (surrogate) data sets in the figures visualizes the difference between observed distribution patterns and the ones expected on the grounds of pure randomness. Note that, whenever gene-masking methodology is applied, corresponding surrogates are being made that exclude the masked space from the random positioning of markers.

Simulations using the genomic duplications – CNE loss model

Simulations using an ample choice of parameter values reproduce the observed genomic distributions. In the last figure, we show characteristic cases, while in the appendix of Plot S1, some more examples are also included. Initially, 1000 markers (representing CNEs) are randomly inserted in a sequence 2 Mbp long. Part (a) of the last figure shows snapshots of the emerging power-law-like pattern as it develops through time. Complementary cumulative size distributions of distances (spacers) between consecutive CNEs are computed every 50 segmental duplication events. Each segmental duplication (SD) event involves a region with length sampled from a uniform distribution with maximum the 5% of the actual length of the simulated sequence. In all these simulations, after each SD event, a number of CNEs equal to 90% of the number of the duplicated CNEs are eliminated (denoted as: fr = 0.9). In the part (b) of last figure, three distribution curves are presented produced after numerical simulations where the fraction fr takes the values 0.8, 0.9 and 1. In part (c) of the same figure, three distribution curves are presented again. In these simulations the fraction fr remains constant and equal to 0.9, while, in two of them, additional eliminations of CNEs are allowed, one and two after each event of segmental duplication respectively.

Results

Occurrence of power-law-like size distribution between inter-CNEs' distances

The main finding of this study is the widespread occurrence of power-law-like size distribution of the distances between consecutive CNEs. In our analysis we include CNE datasets from various taxonomic groups and also compare CNE populations exapted at different evolutionary stages. The studied CNEs are mapped on different genomes (human, *D. melanogaster*, *C. elegans*, *D. rerio*).

In Figure 1 we present the size distributions of distances between consecutive CNEs in a double logarithmic plot in some typical cases. We also report the linear region E of the distribution, and the slope μ. The full set of plots is presented in Plot S1, while a complete quantitative description of our results is given in Table S1. In Table 1 we summarize the results per organism and report the average value of the linear extent E in log-log scale for all chromosomes (avg E) and for the five chromosomes with the largest E (avg E-5) (including only linear regressions with $r^2 > 0.97$, see in the Methods). The extent E captures the orders of magnitude that the power-law-like distribution spans. Throughout this work we use the quantity E for measuring the existence of a self-similar chromosomal geometry and for assessing the accordance of the observed genomic distributions with the evolutionary model we propose (see Discussion).

Power-law-like patterns are not only found in alignments of closely related genomes but are widespread. Elements identified from mammalian and amniotic whole genome alignments [19] were among the first non-coding constrained elements found in quantities allowing statistical analysis of their chromosomal distribution. The complete set includes 16,575 Amniotic and 82,335 Mammalian CNEs (see Datasets iia,b, File S1 & Table S1). We observe power-law-like patterns also in collections of CNEs

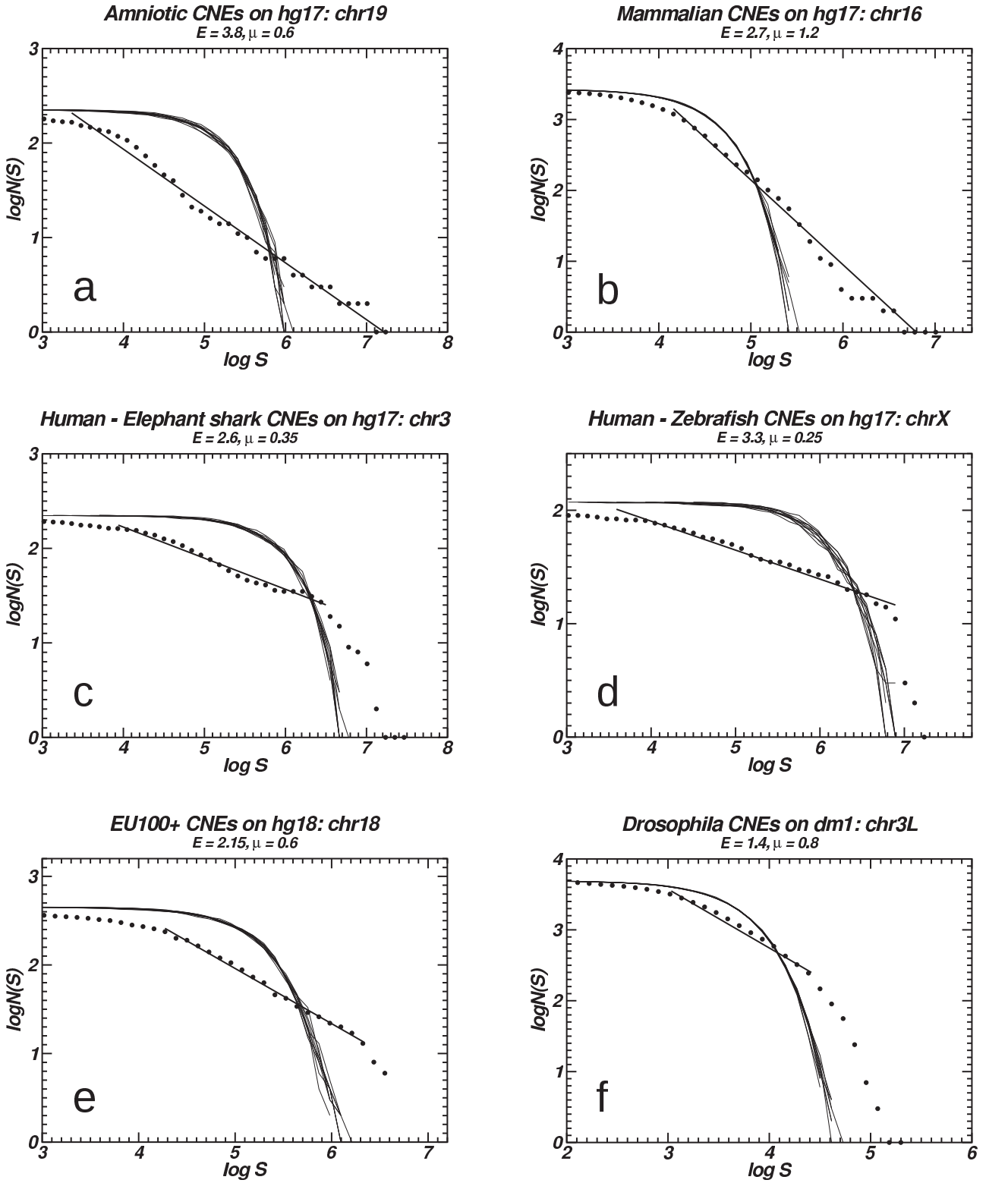


Figure 1. Examples of power-law-like size distributions. Twelve plots of inter-CNE spacers' cumulative size distributions in whole chromosomes. Genomic curves are accompanied in each plot by 10 curves of surrogate data (continuous lines), corresponding to randomly distributed markers. The linear segments are inferred by linear regression. Whenever we mention CNEs in the plots, we refer to the distances between consecutive CNEs.

doi:10.1371/journal.pone.0095437.g001

derived from alignments including mammals along with teleosts, their last common ancestor dated ~450 MYA and cartilaginous fish (elephant shark) that diverged ~530 MYA [65]. The size distribution of inter-CNE spacers in invertebrate genomes, such as *D. melanogaster* and *C. elegans* also follow a similar pattern. It is known that vertebrate and invertebrate CNEs share similar sequence characteristics but are not identical [5], hence indicating in combination with our results that their distributions are shaped by common mechanisms.

Power-law-like distributions are typically characterized by overrepresentation of large spacers. Thus, one could expect that extended power-law-like linearity would be favored in scarce data sets. However, this is not the case. Based on our data, we deduce that the power-law-like size distribution of inter-CNEs's spacers is inherent to the studied system and is not dependent on the population sizes (instances of CNEs). This is evidenced by the fact that when we reduce the numbers of mammalian CNEs (~80,000), by random downsampling to similar numbers as the amniotic ones (~16,000), which are characterized by more extended power-law-like linearity, the extent of linearity is not increased. Instead, linearity in double-log scale disappears as a consequence of the alteration of the studied genomic landscape. Similarly, linearity disappears when we study the merged populations of amniotic and mammalian CNEs. A description of this methodology and the related plots are included in the last section of Plot S1. Thus, we argue that our results are characteristic of each CNE class studied and are not dependent on CNE population sizes provided that the existing populations of constrained elements are sufficient for statistical analysis.

The observed distribution of CNEs is not a mere consequence of the localization of genes in the same chromosome

Power-law-like distributions are found in the chromosomal distribution of protein-coding segments [47]. As it is known from the literature, CNEs are somehow spatially associated with genes coding for transcription factors and developmental regulators (also known as trans-dev genes) [13,15,18]. To rule out the possibility that the observed CNE distributions are a consequence of power-law-like patterns followed by inter-genic distance distributions, we mask all protein coding genes and extended flanking regions, where usually most of the known regulatory elements are located. By masking, we mean excluding all the elements that fall within genes and flanking regions and not removing the genes themselves, as the latter would alter the inter-CNE distance size distribution. Linearity in log-log plots is not only preserved but in most cases improved, as shown by the increase of the linear region extent. This shows that even if we exclude from our study the CNEs that might be bound to be close to genes (thus following their distribution), the remaining CNEs still follow a power-law-like chromosomal distribution. Our principal aim here is to show that the dynamics creating the power-law-like pattern is not a mere consequence of the genic distribution, although the two distributions are expected to influence one another, a fact which is not taken into account in our simple model. Examples of such plots are given in Figure 2. The full set of these plots and the related quantitative description are also included in Plot S1, Plot S2 and Table S1. In Table 1 the results concerning "gene-masked" chromosomes per organism are also given for a direct comparison.

We choose to perform the masking methodology in the human genome for the most abundant sets of CNEs [19,20] as well as for the most ancient elements conserved between human and zebrafish [32] or elephant shark [55]; datasets **(iia,b)**, **(i)**, **(v)** and **(vi)** respectively. In all cases studied, we observe power-law-like size distributions of inter-CNEs' spacers that are extended over several orders of magnitude (see Table S1 & Table 1). A similar methodology is applied to the genome of *D. melanogaster*; dataset **(viii)**. The possible functions of many CNEs (individually or in blocks) through their interactions with specific genes within the nucleus, by means of chromatin looping and other conformations, has recently received a direct experimental verification through the work of Viturawong *et al.* [66]. These authors have demonstrated, in a collection of 193 UltraConserved Elements, the frequent cis action of (distant to a gene) CNEs through chromatin looping. The scope of our gene masking applied herein is not to exclude CNEs acting as distant regulatory elements through such a mechanism. Thus, we have chosen to present a moderated (5 kb upstream of the 5′ end and 2 kb downstream of the 3′ end) gene-flank masking, in order to only exclude elements, which are probably limited to act as close (e.g. promoter-like) regulators and consequently may be spatially linked to nearby genes. To further validate our claim we also performed gene-masking with extensive flanks (10 kb and 100 kb) in the EU dataset for six chromosomes (the five largest ones and chromosome 10, which is particularly abundant in CNEs). Linearity in log-log scale in the distributions of inter-CNE distances is still evident and extends at several length scales (see various statistics and plots in Table S1/sheets EU100+_masked10/100 kb and Plot S2 correspondingly). Even in the case of 100 kb flanks, such linearities are preserved, despite the few CNEs left after masking at such a large scale.

Discussion

An evolutionary model reproducing the observed power-law-like distributions based on genomic (segmental or whole-genome) duplications and CNE loss

Segmental duplication events occurred continuously in the evolutionary past of virtually all eukaryotes [67–70]. At least 10% of the non-repetitive human genome consists of identifiable (i.e. relatively recent) segmental duplication events [71]. It is estimated that 50% of all genes in a genome are expected to duplicate, giving an "offspring" at least once on time scales of 35 to 350 million years [72]. Additionally, most extant taxa have experienced paleopolyploidy during their evolution (i.e. duplication of the whole genome and subsequent reduction to diploidy), see e.g. [73,74] and references therein. Segmental duplication and polyploidization generate copies of some or all the genes of an organism, but also of other functional genomic elements, such as CNEs. As all authors agree, see e.g. [72,74,75], the fate of most duplicated genes is that one copy is silenced, losing the ability to be transcribed, and then disintegrates progressively by random mutations, while it is also exposed to the possibility of excision due to recombination driven eliminations. The fate of duplicated CNEs is expected to be similar, therefore a duplicated CNE often can become superfluous and stop to be under purifying selection, being thus gradually decomposed and lost. The existence of another source of CNE loss can be supported by current findings of comparative genomics, as many of the CNEs found to be conserved between elephant shark and human are not recognized in the fugu genome (see next section for further discussion). In that case, not only duplicates of CNEs are lost but also the population of CNEs present in an ancestral organism is considerably reduced.

Occasional CNE loss of function and subsequent degradation, complete genome duplications and repeated segmental duplications alongside with other forms of genomic dynamics (e.g. insertions of transposable elements and of other parasite sequences) can be combined in an evolutionary model the propensity of which to generate power-law-like chromosomal distributions is

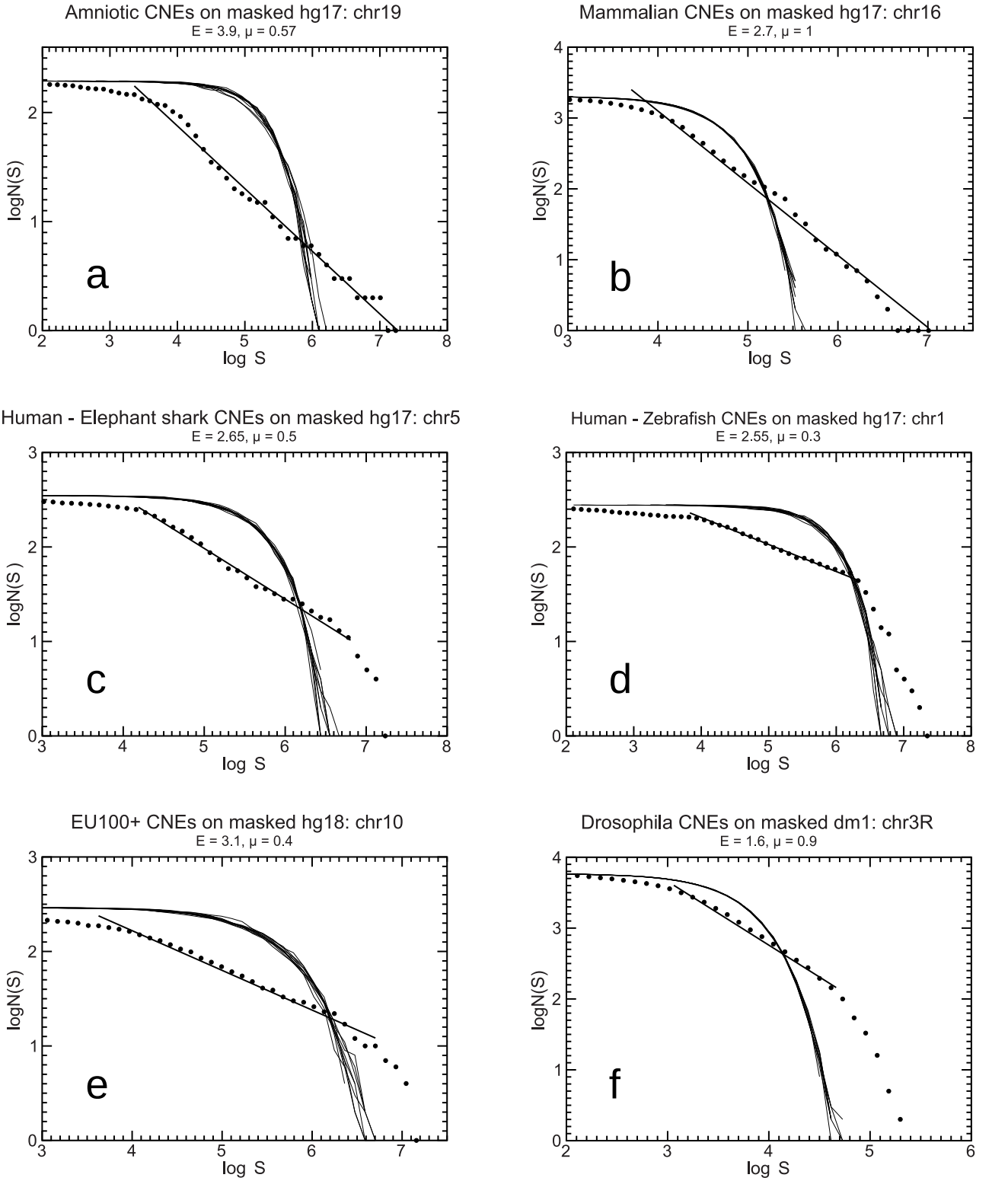


Figure 2. Examples of power-law-like size distributions after gene-masking. Six plots of inter-CNE spacers' cumulative size distributions in whole chromosomes after masking genes and flanks (for further details see in the text). Surrogate curves and regression as in figure 1. Whenever we mention CNEs in the plots, we refer to the distances between consecutive CNEs.
doi:10.1371/journal.pone.0095437.g002

testable through computer simulations. We implement such a model and name it "genomic duplications – CNE loss model" (see link at the bottom of File S1 where we provide the code in Fortran and a detailed description of the model). The genomic events included are:

- ***i*** Segmental duplications of extended regions of chromosomes. This step may include as limiting case whole genome duplications, although not considered in the examples shown.
- ***ii*** Random eliminations of a number of CNEs which is lower or equal to the number of the duplicated ones.
- ***iii*** Occasionally, additional eliminations of non-duplicated CNEs.
- ***iv*** Insertions of sequences increasing the total chromosomal length (these could be transposable elements, retroviruses, microsatellite expansions etc).
- ***v*** Deletions of sequence stretches (which usually are under weak or no purifying selection).

The proposed evolutionary model reproduces power-law-like distributions of the sizes of inter-CNE distances, see Figure 3 and additional examples of simulations in the appendix of Plot S1. This property is proven numerically to be robust to quantitative modifications of all the involved types of molecular events. Only events ***i*** and ***ii*** are indispensable for the appearance of the power-law-like pattern. This dynamics has close parallels with the one described earlier for the explanation of an analogous distributional pattern followed by protein-coding genes [47]. In a completely different genomic framework (i.e. when non-conserved elements, e.g. interspersed repeats or microsatellites are being studied), event types ***iii*** and ***iv*** (i.e. insertions of TE families more recent than the studied one) are required instead of ***i*** and ***ii*** [49,50]. Events ***iv*** and ***v*** are numerically shown not to be required for the emergence of power-law patterns in computer simulations described herein. Inclusion of events of the type ***iv*** tests the robustness of the model, as for many organisms important parts of the genome represent repeat proliferation. Events of type ***v*** represent either deletion of sequence regions, usually due to unequal recombination or gradual shrinkage by a balance of indel events, favoring decrease of the sequence length. These types of events are of importance in genomes getting more compact (evolution occurred e.g. in the recent past of *Drosophila melanogaster* or in the case of *Takifugu rubripes*). Examples of simulations including all these types of genomic dynamics can be found in [47].

These evolutionary scenarios are based on an analytically solvable model introduced by Takayasu *et al.* [53] for the appearance of power-law size distributions in aggregative growth of particles in physicochemical systems. Notice that the model presented herein, conceived to describe the genomic dynamics of CNEs, is not analytically solvable and thus no universal exponents (slopes for the linear segment in log-log scale) may be reached. This is verified by all our computer simulations and is in accordance with the variety of slope values met in the study of genomic CNE distributions. Thus, our data deviate from the typical power-law not only because they always have the linearity in log-log scale truncated at a lower and an upper cut-off (in fact, this is a feature common to all cases of naturally occurring "power-laws") but mainly because they lack any universal exponent (slope). This is the principal reason why we call the pattern we have found "power-law-like" throughout this article. For a recent in depth view of the requirements for having a power-law, see Stumpf and Porter [63]. These authors state that these requirements include a statistically sound power-law (extended linearity in log-log scale

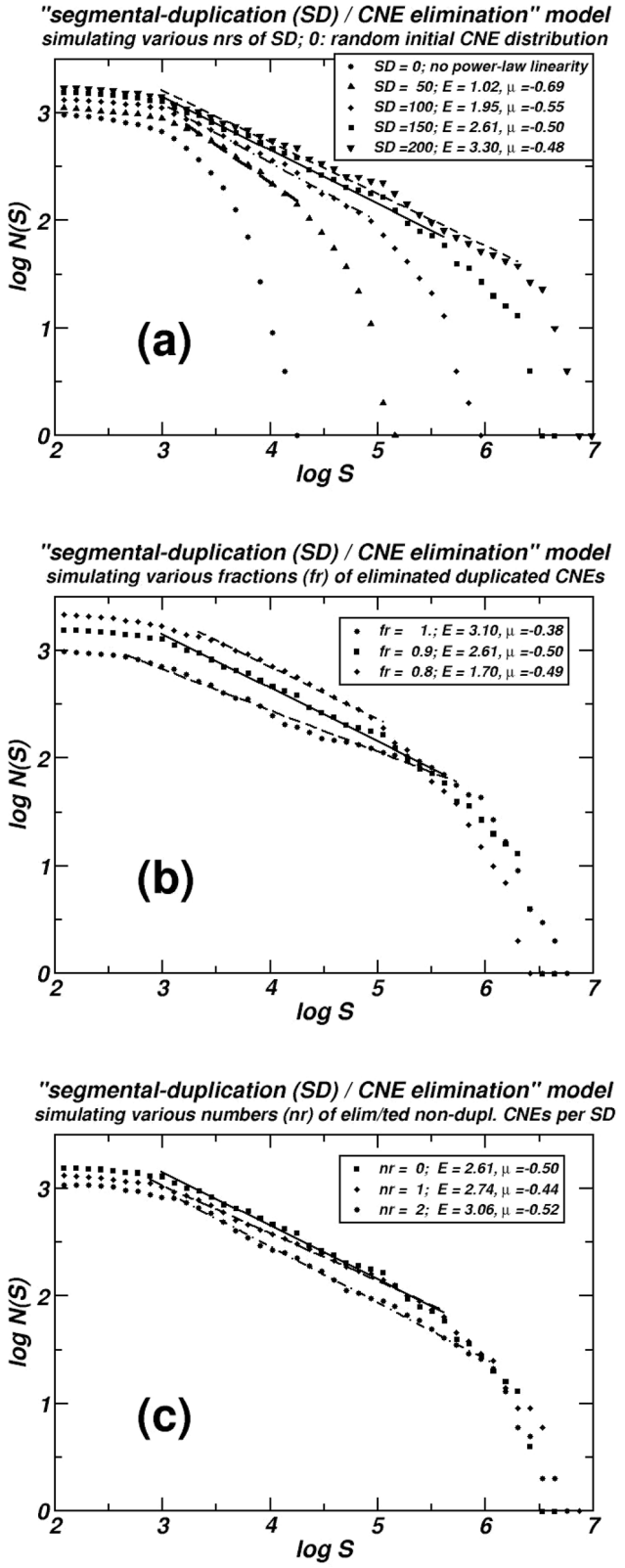


Figure 3. Simulations using the "genomic duplications – CNE loss model". The dependence of the extent of the linearity in log-log scale for the distances between consecutive simulated CNEs on several parameters is shown: **(a)** The number of Segmental Duplications (SD). **(b)** The fraction of the duplicated CNEs eliminated after each SD (fr). **(c)** The number of additional, non-duplicated, CNE eliminations. In **(a)** we

are able to follow the evolution of the emerging power-law-like pattern, as the four curves correspond to consecutive snapshots taken from the same numerical experiment. The curve depicted by squares (▪) is common in all three plots, representing a simulation including 150 segmental duplications, where 90% (fr = 0.9) of the number of duplicated CNEs are lost. No additional eliminations are supposed here. Linear segments are computed by linear regression and in all cases $r^2 > 0.98$.
doi:10.1371/journal.pone.0095437.g003

with indications of convergence to a universal exponent) and a concrete underlying theory to support it. In our case we clearly show that the log-log linearities we observe in our genomic data (often being quite extended) lack a universal exponent (slope). On the other hand, this deviation from universality is a characteristic feature shared between the genomic inter-CNE distributional patterns we describe herein and the simulations of the proposed evolutionary model. This feature, along with the common dependence on the evolutionary parameters shared between model and genomic distributions, corroborates the hypothesis that the evolutionary dynamics described by this model is at the origin of the observed genomic patterns.

Under a wide range of parameter combinations, our model reproduces the transient power-law-like distributions we observe in genomic data. In the simulation of Figure 3a & b, events of the types ***i*** and ***ii*** are only included: i.e. segmental duplications (SD) followed by CNE losses. In Figure 3a, a simulated chromosome is monitored using consecutive snapshots taken every 50 SD, starting from an initial (at "time zero") random distribution of markers representing genomic CNEs. We see that, gradually, a power-law-like linear region in log-log scale appears in the cumulative distributions of inter-CNE distances, as the ones observed in real chromosomes. This plot also shows the positive dependence of the observed extent of the linearity on the number of the occurred SD. In Figure 3b, the positive dependence of the extent of linearity on the number of the CNEs eliminated (lost after each segmental duplication) is shown. Here, the number of the eliminated CNEs is expressed as a fraction (fr) of the duplicated ones, because of the segmental duplication events. Finally, in Figure 3c additional eliminations of not duplicated CNEs are simulated (events of type ***iii***), and the positive dependence of the extent of linearity on the number of non-duplicated lost CNEs is also shown. As we discuss in more detail later, this finding is compatible with the extended linearities found in the distributions of teleosts' CNEs, where important losses of ancestral CNEs are reported. Such ancient CNEs are absent in the teleost genome while retained in the tetrapod lineage.

Evolutionary origin and implications of CNE chromosomal distributions

In the Results section we have seen that the power-law-like distributional pattern reported in the present work is proper to CNEs, UCEs and other highly conserved elements, not being a mere consequence of genic spatial distributions (see Figure 2 and Table 1). The complete study of gene-masked chromosomes has been conducted in the human genome for the most abundant sets of CNEs, as well as for the most ancient elements. In five out of six examined cases (see Table 1, datasets. **i**, **iia**, **iib**, **vi**, **viii**), inspection of both average extent of linearity for all chromosomes (avg E) and for the five chromosomes with the more extended linearity (avg E-5) reveals that the extent of the linear region in log-log scale of the original distribution follows an increasing trend (or is at least preserved) after masking of the CNEs positioned next to, or inside genes (for details see "Methods"). In the remaining one case (dataset **v**, Human/Zebrafish) well-shaped power-law-like distributions are still present after masking, with reduction of the length of the linearity. In Table S1 more information on the related statistics is given. The persistence of the linearity in log-log plots after gene-masking shows the independence of the two patterns. The fact that, in most cases, the average extent is not only preserved, but increased, further strengthens this conclusion. The frequent improvement of the power-law features when gene-masking applies, might indicate that, when the whole CNE chromosomal population is studied, the observed distributional features reflect *a superposition of two distinct dynamics*. Both include molecular events belonging to the same types but with different rates, corresponding to the distinctive evolutionary modalities of gene-uncorrelated CNEs and of genes (with which gene-proximal CNEs are spatially associated). This superposition of distributions with different features (the slope and the length scale of intervening sequences) is expected to reduce the observed linearity, because of a transformation of part of the superposed linear log-log distributions into a curved shape. Another evidence in support of the divergence between the distributional patterns followed by genes and CNEs stems from the observation reported by several research groups that UCEs and CNEs are abundant in gene deserts, see e.g. [19,20].

A link between CNEs and Segmental Duplication, and an additional insight about the fate of duplicated CNEs is provided in the work of Derti *et al.* [76]. These authors found that segmental duplications (SD) are depleted in UCEs (100% conserved elements). They explained this finding as a result of counter selection of duplications when they contain UCEs, probably due to dosage effects. If their result is valid for constrained elements independently of degree of conservation (denoted herein as CNEs in general), this implies that we should expect a low rate of duplication of CNEs, as is the case for genes under strong dosage dependence. In our proposed evolutionary scenario, we model these rare events that over long evolutionary time may have significantly contributed to the observed distributions. Note that in most of the cases of CNEs studied herein, the time from the divergence of the studied species is sufficient for accumulation of random mutations beyond recognition for a duplicated CNE, which is no longer under purifying selection. The finding of Derti *et al.* about counter selection of duplicated CNEs may drive to the inference that, when a SD containing a CNE is fixed in a population, we may have not only relaxed purifying selection on the second copy, but additionally, due to a deleterious dosage effect a fast rate of accumulation of mutations until all the functions of the duplicated CNE are lost.

In a study of the degree of conservation of distances between UCEs in vertebrates [77], real conservation of distances is found only between closely spaced elements, a range of distances which hardly contribute to the log-log linearity reported herein (see Figure 1 and Plot S1). The absence of considerable interspecies retention of distances between conserved elements is compatible with our claim that an aggregative model (like the one described in the previous section) is suitable for the explanation of the widespread occurrence of power-law type linearities in inter-CNE distance distributions.

Mattick and co-workers, in their article on ultraconserved elements (UCEs) in tetrapod genomes, observed a striking difference in UCE populations between tetrapod genomes, which are rich in UCEs, and fish genomes, where considerably lower UCE numbers have been found [20]. They proposed as the most parsimonious explanation that in the tetrapod lineage a massive exaptation of functional elements occurred, which would probably be required for the more complex morphology and different

environmental challenges met by these organisms. On the other hand, the same authors mentioned that this finding might also be explained by the less probable assumption of a massive loss of such elements in the teleost fish genome. The subsequent sequencing of the first cartilaginous fish genome (elephant shark) led to the verification of this latter scenario, as Lee *et al.* found that the jawed vertebrate ancestor had an important number of UCEs which have been eliminated (diverged beyond recognition) at great extent in teleosts, while retained in tetrapods [78]. This finding fits well with our observation that the three globally best scores in linearity extent, as deduced by inspection of Table 1 (datasets **iv**, **v**, **vii**), are all cases of alignments including teleost fish genomes. Additionally, Wang *et al.* directly correlates the observed extended eliminations of UCEs in the teleost fish with the whole genome duplication that had occurred in the ray-fish lineage [79]. Note the significance of duplications (both whole genome and segmental ones) for our proposed model. Their role is twofold, as they make possible the subsequent elimination of duplicated CNEs due to redundancy, and simultaneously provide the necessary sequence extension for the formation of lengthy inter-CNE spacers (see previous subsection).

Evidence for extensive clustering of UCNEs in the vertebrate genome, in groups which are related to the regulation of one gene each, has been presented in a recent work [24]. When loss of duplicated UCNEs occurred, the retention of UCNEs is reported to be far from the expected on the basis of a random retention model. A "winner-takes-all retention pattern" applies, i.e. one gene retains many UCNEs whereas the other paralog loses all of them more often than expected on the grounds of pure chance. An over-representation of survived fish genes, which have lost all their ancestral UCNEs, has been found. This finding is in line with our observation of extended power-law-like linearity observed in the distributional pattern emerging in alignments including teleost fish genomes. Therein, the relatively frequent eliminations of whole UCNE clusters promote the appearance of large inter – UCNEs spacers, which contribute to the long tails and thus to the formation of extended linearity in log-log scale for the corresponding distance distributions [64]. Gene-centered functional clustering described by Bucher and co-workers of CNEs seems to represent one distinct component in the spatial distribution of CNEs. This clustering extends at length scales of the order of single gene neighborhoods. The findings presented herein deals with a broader length scale, the distribution of CNEs at the chromosomal level, resulting in a variety of CNE-rich and CNE-poor domains due to the described aggregative dynamics. These domains follow a fractal-like pattern, as witnessed by the power-law-like inter-CNE distance distributions. The model we propose here consists a better null model than the random positioning of CNEs at a chromosomal level, which then, has to converge with local, gene-specific organization trends. However, the state of our knowledge about the roles of CNEs and their quantitative interactions with genes is currently limited. Further research on evolutionary dynamics and functional roles of CNEs is required in order to better understand the interweaving of the two distributional trends.

The comparison of the extent of power-law-like linearity between Amniotic and Mammalian data sets of Kim and Prichard [19] is also in accordance with the proposed model (Table 1, datasets **iia**, **iib**). We see that the older in evolutionary time Amniotic CNE collection forms the most extended linear segments, as predicted by the hypotheses of the aggregation-elimination model, where, the more the system is exposed (in evolutionary time) to the CNE elimination – sequences insertion dynamics, the more extended the linearity in log-log scale is expected to get. The same trend is clearly present in four data sets extracted from another study [20]. These datasets (Table 1, **ib** - **ie**) consist of elements exapted in consecutive evolutionary periods: tetrapod, amniote and mammalian speciations. We observe that the mean extents of linearity for the "5 best" chromosomes strictly follow the expected increasing order (from the more recent to the older), while the same holds true when we examine the mean extents for complete chromosomal sets with only one inversion found (between **ic** and **ib**).

In a very interesting work, Martinez-Mekler *et al.* [80] have shown that a broad range of systems consist of elements which, when plotted in rank *vs.* frequency or size diagrams, at semi-logarithmic scale, fit closely to a functional form including two fitting parameters. We have not further elaborated on the relation of the fitness of our data to the rank-ordering distribution approach. The range of application of this approach is quite large, and as these authors suggest, there must be an underlying explanation, possibly of a statistical nature [80]. On the other hand, our principal aim is to focus on the specific events of molecular dynamics origin, which may have caused the linearity in double logarithmic scale, and the evolutionary model proposed herein serves this purpose. A question that remains still open is the molecular dynamics impact of the fitting parameters of the functional form proposed by Martinez-Mekler *et al.*, which however lies beyond the scope of the present work. This task might be more straightforward in the case of the genomic evolution of non-constrained elements [49,50], where the evolutionary modeling is simpler.

In what concerns not *the distances between* functional genomic localizations, but *the sizes of the localizations themselves*, we have to mention again here that a power-law size distribution of "perfectly conserved" sequences between human and mouse (repeat-masked) genomes have been observed [51]. A related finding is that the size distribution of "conserved blocks" between *Drosophila melanogaster* and *Drosophila virilis* genomes fits well to a lognormal distribution [81]. Note that this distribution also presents linear regions in log-log scale [64]. A power-law-like distribution has been reported for 3′ untranslated regions earlier [82]. These findings, concerning the sizes of functional genomic sequence stretches, have to be the result of the action of mechanisms entirely different than the evolutionary scenario applied herein, as they extend in very short length scales (tenths to hundreds of nucleotides) while any random-aggregation procedure of the type proposed herein is unsuitable for the formation of functional sequence elements. Aggregative length growth is more suitable for the shaping of large genomic regions (e.g. intervening sequences) with low conservation requirements (see also [47,53]). However, the interweaving of linearities in the form of power-laws at several orders of magnitude and for several functional elements or the distances between them, is probably related to the fractal globule structure reported for the whole human genome when it is in the form of the tightly packed chromatin, which is characterized by extended power-law-type size distributions of the distances between points of the chromosomal thread which come in close mutual contact due to the 3D chromatin folding (see Figure 4A in [83]).

Supporting Information

Table S1 "Supplementary Table" (including all the examined cases).
(XLS)

File S1 Description of the data sets that include the chromosomal coordinates of conserved noncoding elements (BED file-format) and the corresponding data sets

after suitable gene masking as described in the ''Materials and Methods''. All these data sets are also available in a compressed form via the provided URL.
(TXT)

Plot S1 ''Plots'' (including all the examined cases) and additional examples of model simulations.
(PDF)

Plot S2 ''Plots'' of EU100+ CNEs on 10 kb and 100 kb gene-masked genome (hg18).
(PDF)

Acknowledgments

The authors wish to thank the academic editor and the two anonymous reviewers for their useful comments. We would like to acknowledge Dr Christoforos Nikolaou and Giannis Tsiagkas for all of their help and support during the preparation of this study and Professors Stavros J. Hamodrakas and George C. Rodakis (both at the Faculty of Biology, University of Athens) for serving as academic advisors to D.P. We are also indebted to Dr John S. Mattick, Dr Slavica Dimitrieva and Dr Philipp Bucher for sharing unpublished datasets of CNEs.

Author Contributions

Conceived and designed the experiments: DP DS YA. Performed the experiments: DP. Analyzed the data: DP DS YA. Contributed reagents/materials/analysis tools: DP DS YA. Wrote the paper: DP DS YA.

References

1. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, et al. (2011) A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478: 476–482. doi:10.1038/nature10530.
2. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, et al. (2004) Ultraconserved elements in the human genome. Science 304: 1321–1325. doi:10.1126/science.1098119.
3. Elgar G, Vavouri T (2008) Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. Trends Genet 24: 344–352. doi:10.1016/j.tig.2008.04.005.
4. Harmston N, Baresic A, Lenhard B (2013) The mystery of extreme non-coding conservation. Philos Trans R Soc Lond B Biol Sci 368: 20130021. doi:10.1098/rstb.2013.0021.
5. Vavouri T, Walter K, Gilks WR, Lehner B, Elgar G (2007) Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. Genome Biol 8: R15. doi:10.1186/gb-2007-8-2-r15.
6. Glazov EA, Pheasant M, McGraw EA, Bejerano G, Mattick JS (2005) Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. Genome Res 15: 800–808. doi:10.1101/gr.3545105.
7. Lockton S, Gaut BS (2005) Plant conserved non-coding sequences and paralogue evolution. Trends Genet 21: 60–65. doi:10.1016/j.tig.2004.11.013.
8. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034–1050. doi:10.1101/gr.3715005.
9. Vavouri T, McEwen GK, Woolfe A, Gilks WR, Elgar G (2006) Defining a genomic radius for long-range enhancer action: duplicated conserved non-coding elements hold the key. Trends Genet 22: 5–10. doi:10.1016/j.tig.2005.10.005.
10. Clarke SL, VanderMeer JE, Wenger AM, Schaar BT, Ahituv N, et al. (2012) Human developmental enhancers conserved between deuterostomes and protostomes. PLoS Genet 8: e1002852. doi:10.1371/journal.pgen.1002852.
11. Retelska D, Beaudoing E, Notredame C, Jongeneel CV, Bucher P (2007) Vertebrate conserved non coding DNA regions have a high persistence length and a short persistence time. BMC Genomics 8: 398. doi:10.1186/1471-2164-8-398.
12. Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL, et al. (2007) Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature 447: 167–177. doi:10.1038/nature05805.
13. Sandelin A, Bailey P, Bruce S, Engstrom PG, Klos JM, et al. (2004) Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. BMC Genomics 5: 99. doi:10.1186/1471-2164-5-99.
14. Sanges R, Kalmar E, Claudiani P, D'Amato M, Muller F, et al. (2006) Shuffling of cis-regulatory elements is a pervasive feature of the vertebrate lineage. Genome Biol 7: R56. doi:10.1186/gb-2006-7-7-r56.
15. Sanges R, Hadzhiev Y, Gueroult-Bellone M, Roure A, Ferg M, et al. (2013) Highly conserved elements discovered in vertebrates are present in non-syntenic loci of tunicates, act as enhancers and can be transcribed during development. Nucleic Acids Res 41: 3600–3618. doi:10.1093/nar/gkt030.
16. Baira E, Greshock J, Coukos G, Zhang L (2008) Ultraconserved elements: genomics, function and disease. RNA Biol 5: 132–134.
17. Calin GA, Liu CG, Ferracin M, Hyslop T, Spizzo R, et al. (2007) Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. Cancer Cell 12: 215–229. doi:10.1016/j.ccr.2007.07.027.
18. Woolfe A, Elgar G (2008) Chapter 12 Organization of Conserved Elements Near Key Developmental Regulators in Vertebrate Genomes. Adv Genet 61: 307–338. doi:10.1016/s0065-2660(07)00012-0.
19. Kim SY, Pritchard JK (2007) Adaptive evolution of conserved noncoding elements in mammals. PLoS Genet 3: 1572–1586. doi:10.1371/journal.pgen.0030147.
20. Stephen S, Pheasant M, Makunin I V, Mattick JS (2008) Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. Mol Biol Evol 25: 402–408. doi:10.1093/molbev/msm268.
21. Lettice LA, Heaney SJH, Purdie LA, Li L, de Beer P, et al. (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet 12: 1725–1735. doi:Doi 10.1093/Hmg/Ddg180.
22. Bishop CE, Whitworth DJ, Qin Y, Agoulnik AI, Agoulnik IU, et al. (2000) A transgenic insertion upstream of sox9 is associated with dominant XX sex reversal in the mouse. Nat Genet 26: 490–494. doi:10.1038/82652.
23. Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engstrom PG, et al. (2007) Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. Genome Res 17: 545–555. doi:10.1101/gr.6086307.
24. Dimitrieva S, Bucher P (2012) Genomic context analysis reveals dense interaction network between vertebrate ultraconserved non-coding elements. Bioinformatics 28: i395–i401. doi:10.1093/bioinformatics/bts400.
25. Nelson AC, Wardle FC (2013) Conserved non-coding elements and cis regulation: actions speak louder than words. Development 140: 1385–1395. doi:10.1242/dev.084459.
26. Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, et al. (2006) Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genet 38: 223–227. doi:10.1038/ng1710.
27. Sakuraba Y, Kimura T, Masuya H, Noguchi H, Sezutsu H, et al. (2008) Identification and characterization of new long conserved noncoding sequences in vertebrates. Mamm Genome 19: 703–712. doi:10.1007/s00335-008-9152-7.
28. Paparidis Z, Abbasi AA, Malik S, Goode DK, Callaway H, et al. (2007) Ultraconserved non-coding sequence element controls a subset of spatiotemporal GLI3 expression. Dev Growth Differ 49: 543–553. doi:10.1111/j.1440-169X.2007.00954.x.
29. Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, et al. (2007) Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. Proc Natl Acad Sci U S A 104: 7145–7150. doi:10.1073/pnas.0701811104.
30. Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, et al. (2007) Deletion of ultraconserved elements yields viable mice. PLoS Biol 5: e234. doi:10.1371/journal.pbio.0050234.
31. Poulin F, Nobrega MA, Plajzer-Frick I, Holt A, Afzal V, et al. (2005) In vivo characterization of a vertebrate ultraconserved enhancer. Genomics 85: 774–781. doi:10.1016/j.ygeno.2005.03.003.
32. Shin JT, Priest JR, Ovcharenko I, Ronco A, Moore RK, et al. (2005) Human-zebrafish non-coding conserved elements act in vivo to regulate transcription. Nucleic Acids Res 33: 5437–5445. doi:10.1093/nar/gki853.
33. McEwen GK, Goode DK, Parker HJ, Woolfe A, Callaway H, et al. (2009) Early evolution of conserved regulatory sequences associated with development in vertebrates. PLoS Genet 5: e1000762. doi:10.1371/journal.pgen.1000762.
34. Navratilova P, Fredman D, Hawkins TA, Turner K, Lenhard B, et al. (2009) Systematic human/zebrafish comparative identification of cis-regulatory activity around vertebrate developmental transcription factor genes. Dev Biol 327: 526–540. doi:10.1016/j.ydbio.2008.10.044.
35. Ritter DI, Li Q, Kostka D, Pollard KS, Guo S, et al. (2010) The importance of being cis: evolution of orthologous fish and mammalian enhancer activity. Mol Biol Evol 27: 2322–2332. doi:10.1093/molbev/msq128.
36. Sato S, Ikeda K, Shioi G, Nakao K, Yajima H, et al. (2012) Regulation of Six1 expression by evolutionarily conserved enhancers in tetrapods. Dev Biol 368: 95–108. doi:10.1016/j.ydbio.2012.05.023.
37. Matsunami M, Sumiyama K, Saitou N (2010) Evolution of conserved non-coding sequences within the vertebrate Hox clusters through the two-round whole genome duplications revealed by phylogenetic footprinting analysis. J Mol Evol 71: 427–436. doi:10.1007/s00239-010-9396-1.

38. Matsunami M, Saitou N (2013) Vertebrate paralogous conserved noncoding sequences may be related to gene expressions in brain. Genome Biol Evol 5: 140–150. doi:10.1093/gbe/evs128.
39. Hickey DA (2008) Highly similar noncoding genomic DNA sequences: ultraconserved, or merely widespread? Genome 51: 396–397. doi:10.1139/G08-011.
40. Glazko GV, Koonin EV, Rogozin IB, Shabalina SA (2003) A significant fraction of conserved noncoding DNA in human and mouse consists of predicted matrix attachment regions. Trends Genet 19: 119–124.
41. Lettice LA, Horikoshi T, Heaney SJH, van Baren MJ, van der Linde HC, et al. (2002) Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. Proc Natl Acad Sci U S A 99: 7548–7553. doi:DOI 10.1073/pnas.112212199.
42. Loots GG, Kneissel M, Keller H, Baptist M, Chang J, et al. (2005) Genomic deletion of a long-range bone enhancer misregulates sclerostin in Van Buchem disease. Genome Res 15: 928–935.
43. Sagai T, Hosoya M, Mizushina Y, Tamura M, Shiroishi T (2005) Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. Development 132: 797–803. doi:Doi 10.1242/Dev.01613.
44. Li W, Kaneko K (1992) DNA correlations. Nature 360: 635–636. doi:10.1038/360635b0.
45. Peng CK, Buldyrev S V, Goldberger AL, Havlin S, Sciortino F, et al. (1992) Long-range correlations in nucleotide sequences. Nature 356: 168–170. doi:10.1038/356168a0.
46. Voss R (1992) Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Phys Rev Lett 68: 3805–3808.
47. Sellis D, Almirantis Y (2009) Power-laws in the genomic distribution of coding segments in several organisms: An evolutionary trace of segmental duplications, possible paleopolyploidy and gene loss. Gene 447: 18–28. doi:10.1016/j.gene.2009.04.028.
48. Athanasopoulou L, Athanasopoulos S, Karamanos K, Almirantis Y (2010) Scaling properties and fractality in the distribution of coding segments in eukaryotic genomes revealed through a block entropy approach. Phys Rev E Stat Nonlin Soft Matter Phys 82: 051917.
49. Klimopoulos A, Sellis D, Almirantis Y (2012) Widespread occurrence of power-law distributions in inter-repeat distances shaped by genome dynamics. Gene. doi:10.1016/j.gene.2012.02.005.
50. Sellis D, Provata A, Almirantis Y (2007) Alu and LINE1 distributions in the human chromosomes: evidence of global genomic organization expressed in the form of power laws. Mol Biol Evol 24: 2385–2399. doi:10.1093/molbev/msm181.
51. Salerno W, Havlak P, Miller J (2006) Scale-invariant structure of strongly conserved sequence in genomic intersections and alignments. Proc Natl Acad Sci U S A 103: 13121–13125. doi:10.1073/pnas.0605735103.
52. Li W (1991) Expansion-modification systems: A model for spatial 1/f spectra. Phys Rev A 43: 5240–5260.
53. Takayasu H, Takayasu M, Provata A, Huber G (1991) Statistical properties of aggregation with injection. J Stat Phys 65: 725–745.
54. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, et al. (2006) In vivo enhancer analysis of human conserved non-coding sequences. Nature 444: 499–502. doi:10.1038/nature05295.
55. Venkatesh B, Kirkness EF, Loh YH, Halpern AL, Lee AP, et al. (2006) Ancient noncoding elements conserved in the human genome. Science 314: 1892. doi:10.1126/science.1130708.
56. Hufton AL, Mathia S, Braun H, Georgi U, Lehrach H, et al. (2009) Deeply conserved chordate noncoding sequences preserve genome synteny but do not drive gene duplicate retention. Genome Res 19: 2036–2051. doi:10.1101/gr.093237.109.
57. Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, et al. (2008) Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nat Genet 40: 158–160. doi:10.1038/ng.2007.55.
58. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842. doi:10.1093/bioinformatics/btq033.
59. Xie X, Kamal M, Lander ES (2006) A family of conserved noncoding elements derived from an ancient transposable element. Proc Natl Acad Sci U S A 103: 11659–11664. doi:10.1073/pnas.0604768103.
60. Clauset A, Shalizi CR, Newman MEJ (2009) Power-Law Distributions in Empirical Data. SIAM Rev 51: 661–703. doi:10.1137/070710111.
61. Adamic LA, Huberman BA (2002) Zipf's law and the Internet. Glottometrics 3: 143–150.
62. Li W (2002) Zipf's law everywhere. Glottometrics 5: 14–21.
63. Stumpf MPH, Porter MA (2012) Critical truths about power laws. Science 335: 665–666. doi:10.1126/science.1216142.
64. Newman MEJ (2005) Power laws, Pareto distributions and Zipf's law. Contemp Phys 46: 323–351. doi:10.1080/00107510500052444.
65. Kumar S, Hedges SB (1998) A molecular timescale for vertebrate evolution. Nature 392: 917–920. doi:10.1038/31927.
66. Viturawong T, Meissner F, Butter F, Mann M (2013) A DNA-Centric Protein Interaction Map of Ultraconserved Elements Reveals Contribution of Transcription Factor Binding Hubs to Conservation. Cell Rep 5: 531–545. doi:10.1016/j.celrep.2013.09.022.
67. De Grassi A, Lanave C, Saccone C (2008) Genome duplication and gene-family evolution: the case of three OXPHOS gene families. Gene 421: 1–6. doi:10.1016/j.gene.2008.05.011.
68. Kehrer-Sawatzki H, Cooper DN (2008) Molecular mechanisms of chromosomal rearrangement during primate evolution. Chromosome Res 16: 41–56. doi:10.1007/s10577-007-1207-1.
69. Kirsch S, Münch C, Jiang Z, Cheng Z, Chen L, et al. (2008) Evolutionary dynamics of segmental duplications from human Y-chromosomal euchromatin/heterochromatin transition regions. Genome Res 18: 1030–1042. doi:10.1101/gr.076711.108.
70. McLysaght A, Enright AJ, Skrabanek L, Wolfe KH (2000) Estimation of synteny conservation and genome compaction between pufferfish (Fugu) and human. Yeast 17: 22–36. doi:10.1002/(SICI)1097-0061(200004)17:1<22::AID-YEA5>3.0.CO;2-S.
71. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, et al. (2002) Recent segmental duplications in the human genome. Science 297: 1003–1007. doi:10.1126/science.1072047.
72. Lynch M (2000) The Evolutionary Fate and Consequences of Duplicate Genes. Science 290: 1151–1155. doi:10.1126/science.290.5494.1151.
73. Gibson TJ, Spring J (2000) Evidence in favour of ancient octaploidy in the vertebrate genome. Biochem Soc Trans 28: 259–264.
74. Sémon M, Wolfe KH (2007) Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. Trends Genet 23: 108–112. doi:10.1016/j.tig.2007.01.003.
75. Kasahara M (2007) The 2R hypothesis: an update. Curr Opin Immunol 19: 547–552. doi:10.1016/j.coi.2007.07.009.
76. Derti A, Roth FP, Church GM, Wu C (2006) Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. Nat Genet 38: 1216–1220. doi:10.1038/ng1888.
77. Sun H, Skogerbø G, Chen R (2006) Conserved distances between vertebrate highly conserved elements. Hum Mol Genet 15: 2911–2922. doi:10.1093/hmg/ddl232.
78. Lee AP, Kerk SY, Tan YY, Brenner S, Venkatesh B (2011) Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes. Mol Biol Evol 28: 1205–1215. doi:10.1093/molbev/msq304.
79. Wang J, Lee AP, Kodzius R, Brenner S, Venkatesh B (2009) Large number of ultraconserved elements were already present in the jawed vertebrate ancestor. Mol Biol Evol 26: 487–490. doi:10.1093/molbev/msn278.
80. Martínez-Mekler G, Alvarez Martínez R, Beltrán del Río M, Mansilla R, Miramontes P, et al. (2009) Universality of rank-ordering distributions in the arts and sciences. PLoS One 4: e4791. doi:10.1371/journal.pone.0004791.
81. Clark AG (2001) The search for meaning in noncoding DNA. Genome Res 11: 1319–1320. doi:10.1101/gr.201601.
82. Martignetti L, Caselle M (2007) Universal power law behaviors in genomic sequences and evolutionary models. Phys Rev E 76: 021902. doi:10.1103/PhysRevE.76.021902.
83. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326: 289–293. doi:10.1126/science.1181369.

Genomics 104 (2014) 79–86

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Classification of selectively constrained DNA elements using feature vectors and rule-based classifiers

Dimitris Polychronopoulos [a,b,1], Emanuel Weitschek [c,d,*,1], Slavica Dimitrieva [e,f], Philipp Bucher [e,f], Giovanni Felici [d], Yannis Almirantis [a]

[a] Institute of Biosciences and Applications, National Center for Scientific Research "Demokritos", 15310 Athens, Greece
[b] Department of Biochemistry and Molecular Biology, Faculty of Biology, National and Kapodistrian University of Athens, 15701 Athens, Greece
[c] Department of Engineering, Roma Tre University, Via della Vasca Navale 79, 00146 Rome, Italy
[d] Institute of Systems Analysis and Computer Science "Antonio Ruberti", National Research Council, Viale Manzoni 30, 00185 Rome, Italy
[e] Swiss Institute for Experimental Cancer Research (ISREC), School of Life Sciences, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
[f] Swiss Institute of Bioinformatics (SIB), CH-1015 Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 9 June 2014
Accepted 15 July 2014
Available online 22 July 2014

Keywords:
Conserved noncoding elements (CNEs)
Ultraconserved elements (UCEs)
Classification
Rule-based classifiers
Feature vector representation
Alignment-free classification

ABSTRACT

Scarce work has been done in the analysis of the composition of conserved non-coding elements (CNEs) that are identified by comparisons of two or more genomes and are found to exist in all metazoan genomes.
Here we present the analysis of CNEs with a methodology that takes into account word occurrence at various lengths scales in the form of feature vector representation and rule based classifiers. We implement our approach on both protein-coding exons and CNEs, originating from human, insect (*Drosophila melanogaster*) and worm (*Caenorhabditis elegans*) genomes, that are either identified in the present study or obtained from the literature. Alignment free feature vector representation of sequences combined with rule-based classification methods leads to successful classification of the different CNEs classes. Biologically meaningful results are derived by comparison with the genomic signatures approach, and classification rates for a variety of functional elements of the genomes along with surrogates are presented.

1. Introduction

1.1. Conserved non-coding elements in metazoan genomes

It is generally believed that sequence conservation across genomes is a key indication for functional relevance. As a consequence, since the early days of comparative genomics several groups have focused on the detection of genomic sequences highly similar between two or more organisms. Bejerano et al. described a set of 481 sequences that display 100% identity between the human, mouse and rat genomes, so-called ultraconserved elements (UCEs) [1]. Surprisingly, the majority of these sequences do not encode for proteins. The UCEs along with several other classes of elements called UltraConserved Non-coding Elements (UCNEs) (>95% identity over > 200 bp between human and chicken) [2] and Long Conserved Non-Coding Sequences (LCNS) (>95% identity over > 500 bp between human and mouse) [3] are among the most conserved sequences described in the biomedical literature with mostly unknown functionality overall. These sequences represent only the tip of the iceberg of a much larger class of conserved non-coding elements (CNEs), whose mean level of conservation frequently exceeds that of protein-coding sequences [4,5].

Conserved non-coding elements can be found in the literature under various definitions, depending on the minimal sequence similarity between species under consideration and the similarity score achieved over a minimal sequence length [6]. Throughout this article, we use the term CNE(s) to describe all such elements despite their specific characterization as UCEs, UCNEs, LCNSs, etc. in the related literature. We use the specific name only when we refer to the corresponding class of elements.

Due to the level of conservation of those elements, their non-random co-localization around developmental genes [7] and their widespread distribution throughout genomes [5,8], many plausible functional roles have been attributed to CNEs (reviewed in [5]). It has been shown, amongst others, that many of the identified CNEs function as regulatory elements that are important in the early stages of vertebrate development and brain formation [9–11].

Although early studies have primarily focused on CNEs in vertebrate genomes, CNEs with similar properties have been also identified in

* Corresponding author at: Department of Engineering, Roma Tre University, Via della Vasca Navale 79, 00146 Rome, Italy.
E-mail addresses: dpolychr@bio.demokritos.gr (D. Polychronopoulos), emanuel@dia.uniroma3.it (E. Weitschek), slavica.dimitrieva@epfl.ch (S. Dimitrieva), philipp.bucher@epfl.ch (P. Bucher), giovanni.felici@iasi.cnr.it (G. Felici), yalmir@bio.demokritos.gr (Y. Almirantis).
[1] Joint first authors.

http://dx.doi.org/10.1016/j.ygeno.2014.07.004
0888-7543/

invertebrate genomes (insects, worms) [12,13] and in plants [14]. This has suggested that they are of very ancient origin and has provoked speculations about the emergence of those elements [5].

When compared to their surrounding genomic environment, CNEs are generally AT-rich, often containing runs of identical nucleotides (homopurines or homopyrimidines, unpublished results). Walter et al. analyzed the base composition of human and fugu CNEs at single nucleotide level and found that they are A + T rich, much more so than the region they reside in, in contrast to their flanking region just outside their boundaries, which exhibits a marked drop in A + T content that forms a unique pattern [15]. But surprisingly, very little is known about the sequence intrinsic properties of CNEs that are important for their function and that segregate them from other classes of functional elements. It is therefore of great interest to further investigate the compositional preferences of constrained regions in greater detail.

1.2. Alignment-based and alignment-free methods for sequence analysis

Similarity of sequences is generally used as an indication of a corresponding similarity in their functionality. Also, similarity studies have been widely used in the phylogenetic reconstruction based on molecular grounds. Most of the current sequence analysis methods are based on alignments, i.e. aligning areas of sequences at several length scales, from single genes to whole genomes. Each alignment is evaluated with a score that depends on the number of same and contiguous characters in the sequences. Optimal methods for sequence alignments rely on dynamic programming techniques, the most widely used optimal sequence alignment algorithms being the ones of Needleman and Wunsch [16] and Smith and Waterman [17]. These algorithms are computationally demanding and their complexity grows exponentially with the length of the sequences. Heuristics have been proposed that solve the sequence alignment problem, e.g. BLAST [18] and FASTA [19]. In order to perform the alignment of multiple sequences more efficiently, several algorithms have been proposed: ClustalW [20], Muscle [21], Mafft [22], and Motalign [23].

Since the first decades of systematic sequencing of protein coding and non-coding genomic regions, it has been noticed that, while alignment of protein coding genomic segments both between organisms and inside genomes reveals a richness of information, it has limited application on the non-coding. This is because the non-coding, in its greatest part is not evolutionary constrained (i.e. conserved due to its functionality in the course of evolutionary time). Nonetheless, alignment methods may be useful when applied in short non-coding DNA stretches. In larger chromosomal regions their use is limited, because in long regions synthesis is not conserved between organisms, which are evolutionarily relatively distant. Alignment is particularly useful in the study of transposable elements, which are found in multiple copies in most organisms, and are marked by variable degrees of homology between them, depending on their age in the genome. A recent application of alignment of whole genomes between two or more species led to the aforementioned discovery of thousands of conserved and highly conserved non-coding genomic elements (CNEs, UCNEs, etc.) through various strategies.

Alignment-free methods are valuable when we want to extract compositional profiles and preferences, and are applicable both in whole-genome comparisons between organisms and in intra-genomic detection of segments which exhibit particular modalities in their composition, often related to their functionality. One classical alignment-free method is based on studying distances between the genomic signatures of sequences, which is briefly presented in the methods section and is used in the present study for comparison with the novel method applied herein. Based on alignment free techniques, there are also various algorithms for locating CpG islands, which are short genomic sequence stretches with no mutual similarity but with several distinct compositional traits. Protein coding segments could be determined by both alignment and alignment-free techniques, as the use of the genetic code and the modalities of the machinery of protein synthesis (mRNA-ribosome binding, tRNA abundances, etc.) endow the protein-coding part of the genome with characteristic compositional biases. Overall, we could say that alignment and alignment-free methods are complementary and that particular components of the genome could be studied by suitable combinations of both.

Alignment free techniques have been proven to be particularly successful in phylogeny and sequence analysis [24]. In alignment free methods the similarity of two sequences is assessed based only on the dictionary of subsequences that appear within the studied sequences, irrespective of their relative position. A promising alignment free method is based on the feature vector representation of a sequence [24-26], and [27] where a sequence is encoded in a vector containing the occurrences of its substrings (k-mers). A k-mer is defined as a subsequence of the original sequence, k characters long. In k-mer frequency analysis every possible k-mer over the nucleotide alphabet {A, C, G, T} is extracted, its occurrences in the original sequence are counted and its frequency is calculated. A vector containing the relative frequency of every k-mer is computed for each sequence in the analyzed data set. Two sequences are rated similar by analyzing the dictionary of their subsequences, without taking care of their relative position.

CNEs, as their name suggests, are identified through pairwise or multiple sequence alignments between two or more genomes. They tend to appear in single copies in the genome [1,28,29] and have been proposed as markers for phylogenomic studies [30]. Herein, we propose an approach that does not rely upon the information content of an alignment between different inter-genomic DNA regions and we apply it to the classification of functional sequences taken from several genomes, which range from invertebrates to vertebrates. We also proceed to several comparisons that test the robustness of our approach by comparing the performance of our proposed method in classifying genomic elements not only between species, but also within the same organism and of different possible functionalities and evolutionary depth. For that purpose, we use a wide collection of vertebrate conserved non-coding elements published in the literature as well as new datasets of human and invertebrate CNEs identified in this study.

2. Methods

2.1. Published datasets

In this study, we consider several published datasets of constrained sequences extracted from the human (*Homo sapiens*), insect (*Drosophila melanogaster*) and worm (*Caenorhabditis elegans*) genomes. We randomly select 1000 elements from each superset for subsequent analysis, given the heterogeneity of the datasets. Only worm elements are studied in their entirety, since this set includes only 1869 elements. The exonic sequences of human, worm and insect genomes are downloaded from UCSC [31]. Apart from exonic sequences, several classes of constrained non-exonic sequences are used as follows (for various metrics, also see the Supplementary Excel available at dmb.iasi.cnr.it/cnes.php):

(a) UltraConserved Non-coding Elements (UCNEs): These are human non-protein coding sequences (hg19 assembly) that display ≥95% sequence identity when compared to the chicken genome and are longer than 200 bp [32].
(b) EU100 non-exonic CNEs (EU100nx CNEs): These are human sequences (hg18 assembly) that are identical over at least 100 bp in at least 3 out of 5 placental mammals (human, mouse, rat, dog and cow) [29]. The whole set is named EU100+ and since we remove elements overlapping exons, we name it EU100nx.
(c) Amniotic and mammalian CNEs: These are elements identified by Kim and Pritchard [33]. Mammalian CNEs are sequences that are

conserved within mammals but not found in chicken or fish, while Amniotic CNEs are conserved in mammals and chicken but not found in fish. LiftOver is used in order to convert the coordinates to the most recent release of the human genome (from hg17 to hg19).

2.2. Identification of sets of CNEs in vertebrates, insects and worms

For the needs of our study, we also identify sets of conserved elements in vertebrates, insects and worms using uniform criteria in terms of the similarity thresholds applied between the compared sequences and of the considered minimum length. We choose to compare whole genomes of *D. melanogaster*/*Drosophila virilis* and *C. elegans*/*Camellia japonica* because the species which form these pairs are distant enough to allow a clear separation between functionally conserved and neutrally evolving genomic sequences, while they are selected so that evolutionary distances within every pair are close [34]. Whole-genome alignments between *D. melanogaster (dm3)*/*D. virilis* (droVir3) and *C. elegans (ce10)*/*C. japonica (caeJap4)* are downloaded from the UCSC Genome Browser [31]. Sequence regions, where the percentage of sequence identity is consistently ≥90% and the length is >60 bp, are considered as conserved elements. The sequence identity is computed in an asymmetric fashion by taking as references *D. melanogaster* and *C. elegans* genomes and counting the number of conserved bases in the target species in a 61 bp sliding window. The number obtained from every window is used to assign a percentage identity value to the base at the center of this window, as described in [34]. In total, 45,345 sequences are identified as conserved between *D. melanogaster* and *D. virilis* and 1869 between *C. elegans* and *C. japonica*. Based on the Ensembl gene annotations for *D. melanogaster* and *C. elegans* genomes, each of the conserved sequences is classified as protein-coding or non-coding (intronic, UTR-associated or intergenic).

We also identify vertebrate CNEs (based on comparisons between human and chicken) that exhibit various degrees of identity (75–80%, 80–85%, 85–90%, 90–95%). To obtain CNEs that have identity between e.g. 75% and 80%, we first extract a set of CNEs that exhibits identity of 75% (using the same methodology as described above) and from this set we remove the CNEs that exhibit identity of 80%. From the remaining sequences, we keep only those that are longer than 199 bp. We repeat the same procedure for the other sets of CNEs, accordingly.

2.3. Surrogate sequences extraction

FASTA sequences are obtained from BED files using the *fastaFromBed* package of BEDTools [35]. Overlapping elements between different datasets are calculated using *intersectBed* package from the same suite of tools. In addition, we apply the EMBOSS suite to estimate fractional GC content of sequences [36]. Given a CNE or another functional element of each collection under study, an analogue of it is extracted from the corresponding genome. Every such element (surrogate sequence) is ensured to be of the same length and GC content (±1%) with its corresponding element in the collection under study. Statistics of the datasets (mean length of sequences and GC content) are available in the Supplementary Excel available at dmb.iasi.cnr.it/cnes.php. As vertebrate and invertebrate CNEs are of different lengths (the ones belonging to the latter category are considerably smaller, see [34]), we make sure that we take elements of same lengths as follows (in all these cases the compared sets include one thousand sequences each): (i) each time, we compute the length of one element from the 'short sequences collection', (ii) then we take one element from the 'long sequences collection' and we trim it retaining only its central part equal to the length of the corresponding 'short sequence', (iii) we repeat this procedure exhaustively (1000 times in all the considered datasets). Thus, we end up with a new dataset for the 'long sequences' (collection of sequences in the form of a BED file), of which members have lengths equal to the members of the 'short sequences collection'.

2.4. The alignment-free sequences classification method LAF

To show that we can distinguish CNEs from other functional sequences and from surrogate sequences based solely on sequence intrinsic properties, we apply an alignment-free method based on

- a feature vector representation of the sequences and
- classification with rule based supervised machine learning techniques.

And we call it LAF (Logic Alignment Free).

The feature vector is a well-established technique for representing biological sequences and for permitting a classification of them. This methodology is described in Kudenko and Hirsch [25], Vinga and Almeida [24] and Xing et al. [27], where sequence representations with feature vectors are introduced and combined with supervised classification methods, e.g., Support Vector Machines, for classifying specimen to species. Another recent work is the one of Kuksa and Pavlovic [26], who apply feature vector representations for DNA Barcode classification.

The feature vector representation of a sequence S is based on the computation of the substrings occurrences of a given length k in the original sequence by applying a sliding window in S. These substrings are called k-mers. The k-mer counts of a sequence are represented in a feature vector, where each component of the vector is associated with the occurrences of a particular k-mer.

The authors of Ref. [24] provide the following formal definition. Let S be a sequence of n characters over an alphabet A, e.g. A = {A,C,G,T}, and let $k \in I$, $k < n$, $k > 0$. If K is a generic subsequence of S of length k, K is called k-mer. Let the set $V = \{k\text{-mer}_1, k\text{-mer}_2, \ldots, k\text{-mer}_r\}$ be all possible k-mers over A, V has size $t = |A|^k$. The k-mers are computed by counting the occurrences of the substrings in S with a sliding window of length k over S, starting at position 1 and ending at position n-k + 1. A feature vector C contains for each k-mer its occurrences (or counts) $C = \{c_{k\text{-mer}1}, c_{k\text{-mer}2}, \ldots, c_{k\text{-mer}t}\}$. The frequencies are then computed accordingly and stored in a vector $F = \{f_{k\text{-mer}1}, f_{k\text{-mer}2}, \ldots, f_{k\text{mer}t}\}$, for a generic k-mer j the frequency is defined as $f_j = c_{k\text{mer}j}/(n\text{-}k + 1)$. For example considering the 4 letters alphabet {A, C, G, T}, the 2-mers, and the sequence ACGACT, the feature vector C is:

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	2	0	0	0	0	1	1	1	0	0	0	0	0	0	0

and the frequencies vector F is:

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	2/5	0	0	0	0	1/5	1/5	1/5	0	0	0	0	0	0	0

The sequences are so represented in a coordinate space, that is mathematically tractable with linear algebra and statistics, e.g., by considering the vector representation of the sequences it is possible to compute different distance measures between two sequences or to give the vector representation as input to a supervised machine learning algorithm, i.e. a classifier.

The supervised machine learning approach is also called classification: any collection of the analyzed sequences must contain an a priori known class label, i.e. every sequence is associated with a given class, e.g. *Vertebrate*, *Invertebrate*, *Amniotic*, and *Mammalian*. Such a collection is called training set. Based on the training set, the method extracts a classification model, e.g., "if-then rules", for distinguishing the different sequences present in the data set. The model can then be evaluated on a test set, that may be composed of unknown sequences or sequences that belong to a known class, in order to verify the classification performances. Our adopted approach combines the feature vector representation with rule based supervised machine learning methods [37,38].

As a technique for biological sequence analysis, rule based classifiers have been proposed in Bertolazzi et al. [39] and in Weitschek et al. [40], in particular for classifying organisms using DNA Barcode sequences and for viruses identification. In these works, the genomic sequences were analyzed with a positional approach: each nucleotide was analyzed independently by referring only to its position. Therefore an alignment of the sequences or an overlapping gene region was necessary: an analysis of the characteristic nucleotides present in a determined position for every class was performed, leading to logic rules of the type, e.g., *if pos90 = A then the sequence belongs to class X*. The alignment was necessary, because a positional analysis is only possible when the sequences come from the same gene regions and are aligned on a reference position.

The output of a rule based classifier is a collection of if-then rules, assigning a sequence to a particular class (species), e.g., *if pos30 = A and pos40 = C then the sequence belongs to the squalus edmundsi species*. The major advantage of rule-based classification is the additional knowledge that is given by the compact human interpretable model (the if-then rules). In this work, the following algorithms are taken into account for performing the rule based classification analysis of the CNEs sequences feature vector representations: RIPPER [41], RIDOR [42] and PART [43].

RIPPER is a classification algorithm that uses a direct method for rules extraction: it infers the rules by processing directly the data. RIPPER is composed of following computational steps:

1) rule growth
2) rule pruning
3) model optimization
4) model selection.

In the first step, the rules are computed in a greedy way by processing the data attributes. The rules are then pruned (simplified) in step 2 according to statistical measures on the training set. In step 3 the rules are optimized by extending them and adding new pruned rules. In the last step the best performing rules are selected and the remaining rules are discarded from the classification model.

The RIDOR classification algorithm also extracts the rules directly from data. It firstly computes a default rule that covers the most represented class (e.g., "*all sequences are Vertebrates*") and then exceptions rules which cover the other classes (e.g., "*except if freq(ACG) > 250 and freq(AGT) < 200*" *these sequences are Invertebrates*).

On the other hand, PART is an indirect method for rules extraction: it processes the data by using the C4.5 tree based classifier [44] generating a pruned decision tree for every iteration. Finally it selects the best classification tree and uses the leaves as logic rules.

According to the previously described methods (feature vector representation and rule based supervised learning) the LAF approach that is adopted in this work performs for every sequence s in a data set, which is associated to a class (e.g. Vertebrate, Invertebrate), the following computational steps (see Fig. 1):

1. The reverse complement of the sequence s is computed.
2. The sequences S and its reverse complement s_r is combined by concatenation in a sequence S.
3. The counts of the k-mers (for k = 3,4,5) are calculated on the sequence S, obtaining the feature vector $C = \{c_{k\text{-mer1}}, c_{k\text{-mer2}}, \ldots, c_{k\text{-mert}}\}$.
   The k-mer counts are extracted with the Jellyfish software [45]. k has been chosen between 3 and 5, based on the following references [46,47], which state the optimality of such lengths as they provide an ideal balance between the length of the subsequences and their number, when the sequences are expressed in the 4-letter (A,C,G,T) alphabet.
4. The frequencies of each k-mer are computed: $f_j = c_{kmerj}/(n\text{-}k + 1)$.

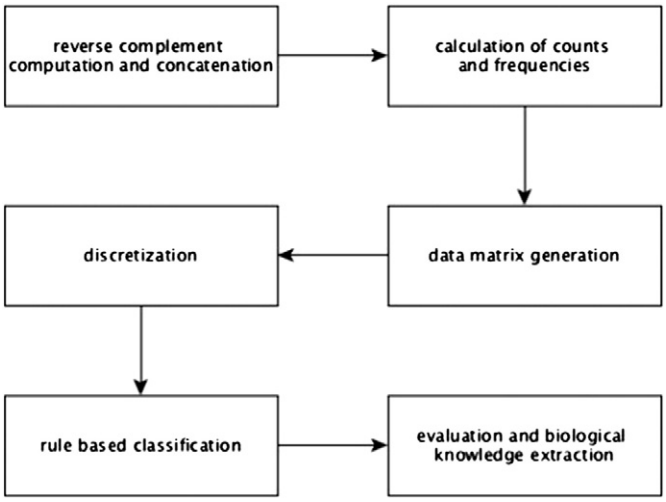


Fig. 1. The LAF method flow chart.

5. A feature frequency vector of each sequence is obtained:

$$F = \{f_{k\text{-mer1}}, f_{k\text{-mer2}}, \ldots, f_{kmert}\}$$

6. The feature vectors are combined in a data matrix, where each row represents a sequence and each column the k-mer frequency. A header with the name of the sequences and their belonging class is present. An example of the data matrix is given in Table 1.
7. The numeric data sets are discretized, i.e. the frequencies are converted from numerical to nominal by the definition of intervals, according to Fayyad & Irani's MDL method [48]; the discretization procedure improves the performance of the rule based algorithms.
8. The data matrix is given as input to three rule based supervised machine learning algorithms: RIPPER, RIDOR, and PART.
9. The classification methods are run in 10 fold cross validation mode to evaluate the performances.
   Cross validation is a standard sampling technique that splits the dataset in a random way in m disjoints sets, and the data mining procedure is run m times with different sets. At a given run n the nth subset is used as test set and the remaining m-1 sets are merged and used as training set for building the model. Each of the m sets contains a random distribution of the data. The cross validation sampling procedure builds m models and each of these models is validated with a different set of data. Classification statistics are computed for every model and the average of these represents an accurate estimation of the data mining procedure performance.
10. The classification models are extracted, e.g.,
    if freq(AAAC) < 0.195 then the organism is Vertebrate;
    if freq(AAAC) > 0.195 then the organism is Invertebrate.
11. The models are evaluated in terms of correct classification rate.

Table 1
Data matrix example of the LAF method.
An example of the data matrix obtained by the application of the LAF method. The data matrix is composed by two heading columns, the first with the identifiers of the sequences and the second with the class labels. The following rows contain the frequency values of the k-mers in the sequences.

	Seq 1	Seq 2	....	Seq N-1	Seq N
	Vertebrate	Vertebrate	....	Invertebrate	Invertebrate
AAA	0.46	0.26	...	0.24	0.54
AAC	0.12	0.16	...	0.23	0.24
AAG	0.12	0.23	...	0.23	0.23
....	...	...	...	....	...

Scripts for filtering, reverse complementing, joining the data sets, calculating the frequencies, discretizing, and classifying have been implemented and are available upon request.

The Weka [48] implementations of the ruled based classification algorithms are adopted for performing the analysis. The experiments have been run under a 64-bit Debian Linux workstation with kernel 2.6.26, 32GB of RAM, and an Intel i7 4-core processor.

2.5. The "genomic signature" method

A different technique for classification of biological sequences is considered for a direct comparison with the results obtained using the LAF approach proposed here. A classical method for quantifying the neighbor preferences in a DNA sequence of an entire genome is the computation of the vector of the odds ratios for dinucleotides [49]. The odds ratio of each dinucleotide is the quantity: $\rho_{ij} = f_{ij}/(f_i\, f_j)$, where f_{ij} and f_i, f_j stand for the frequencies of occurrence in the studied sequence of a dinucleotide and its constituent nucleotides respectively. Thus subscripts *i*, *j* represent any pair of A, G, C and T. This is the ratio of the "observed" dinucleotide frequency over the "expected" one under no neighbor preference, thus it expresses the actual neighbor preferences of the given pair of nucleotides. Before computing the odds ratios for a given sequence, this is concatenated to its reverse complement. Consequently, the relevant ratios are only ten, i.e. four for the self-complementary dinucleotides and six for the mutually complementary couples. Karlin and co-workers found that these quantities differentiate between different genomes, according, approximately, to their evolutionary distance [50]. Thus they have assigned to the vector of these ten "first neighbor preferences" the name of genomic signature [51]. In what follows, we use classification based on genomic signatures for comparison with classification based on the alignment-free method described in the previous section. We also consider RIPPER, PART and RIDOR as classifiers for all the experiments and we also perform the discretization procedure as described previously.

All scripts used throughout this work are available upon request to interested readers. All the used files (in BED format) are available as Supplementary material at dmb.iasi.cnr.it/cnes.php.

3. Results and discussion

In this section, we present a systematic classification analysis of short genomic segments displaying different functionalities and found in the human and other genomes by using a supervised machine learning approach. For every classification task, we adopt the technique of feature vector representation and rule based classification explained in previous sections. Every classification task is performed using feature vector representation with k-mers of length 3, 4 and 5, which are given as input to three different rule based algorithms (RIPPER, RIDOR, PART), adopting a 10-fold cross validation sampling technique. The full list of accuracy results with all parameters is summarized in Supplementary Excel available at dmb.iasi.cnr.it/cnes.php. We omit the percentages of false positives and false negatives, because the errors are equally distributed between the different classes in the datasets. We report here and discuss based on the best result (highest classification rate) among each different rule based algorithm, applied on the feature vectors composed of k-mers length 3, 4, 5. The reason for choosing the overall best for this purpose is that we consider that it captures more directly the potential of k-mers based classification and consequently it might be correlated more meaningfully with several biological characteristics and reveal functional modalities of the studied sequences. In the Supplementary Excel, interested readers are provided with the average classification rates as calculated based on all 26 experiments performed. Based on our results, we suggest k = 3 and PART as the optimal parameters for an ab initio classification experiment (see Supplementary Excel). For all the considered experiments, genomic signatures are also used as an alternative to the k-mers based classification method. Genomic signatures, as characteristic constitutional traces of different genomes, have a long history and represent a classical standard to which we would like to compare the k-mers based approach. In the case of genomic signatures, the best overall result in terms of algorithm used is also considered.

The following classification analyses are presented in order to assess the potential usefulness of our approach in distinguishing among different genomic elements in the same or in different metazoan genomes. The classification rules extracted with the alignment-free sequences classification method (LAF), described in the "Methods" section, are available for download at dmb.iasi.cnr.it/cnes.php. The interested reader can identify the characteristic k-mers for every pairwise experiment and class of the investigated data sets.

3.1. Inter-species comparison of background sequences

In these comparisons between human, worm and insect background sequences we use as representative samples of the different genomes surrogate sets, composed either for exons or for CNEs. Comparisons involving *H. sapiens* yield always the best classification rates using both LAF and Genomic Signatures (GS), see Table 2. This may be understood on the grounds of the high difference of neighbor preferences, mainly in CpG and TpA between *H. sapiens* and the invertebrates, while these preferences are found to be close between

Table 2
Comparison of background genomic sequences from different organisms.
Inter-genomic comparisons between pairs of collections of 1000 background sequences each (non-CNE, non-exonic) from the three studied genomes. These sequences are produced as surrogates for the CNEs and exons studied later. Each sequence has same length and same GC% with one member of the CNE or exon collections. The highest value of LAF or GS accuracy in each classification experiment is shown in bold. For more details on the used methods see in the text and supplementary material.

Experiment no	Description	Average length	Average GC	LAF, accuracy	GS, accuracy
#1	Surrogate for human exons	167.84	0.40	84.21	86.93
	Surrogates for worm exons	169.32	0.52		
#14	Surrogates for human UCNEs	86.09	0.36	81.45	84.30
	Surrogates for insect UCNEs	86.58	0.39		
#20	Surrogates for human exons	169.82	0.51	84.66	87.89
	Surrogates for insect exons	169.32	0.52		
#22	Surrogates for worm exons	213.37	0.42	78.60	70.95
	Surrogates for insect exons	212.86	0.52		
#23	Surrogates for worm UCNEs	83.18	0.43	82.10	84.48
	Surrogates for hUCNEs	82.93	0.36		
#13	Surrogates for worm UCNEs	83.41	0.43	64.70	63.65
	Surrogates for insect UCNEs	86.58	0.39		
Average				79.29	79.70

invertebrate species: *D. melanogaster* (insect) and *C. elegans* (worm). GS is exclusively a quantification of first neighbor preference, while in LAF several other components of the sequence composition are implicitly included alongside, e.g. mono-nucleotide composition and higher order neighbor preference. In accordance with the above description, in cases where human sequences are included in the comparison, GS perform systematically better than LAF, while in the remaining two cases LAF is the best.

Another conclusion that stems from simple inspection of the Table 2 is that in interspecies comparisons between background genomic sequences, the best results are obtained for exonic surrogates (vs. CNE surrogates). This relies on the simple fact that the mean lengths of exons, in all cases, clearly exceed mean lengths of CNEs (and the same holds true for their surrogate sequences, by construction). Higher mean lengths obviously allow better statistics and thus higher classification rates in both LAF & GS. Note that, concerning interspecies CNE comparisons, the sequences of the vertebrate CNE set are always trimmed to the mean length of the shorter invertebrate CNEs set (for details see in the Methods section).

3.2. Comparison of constrained sequences vs. their corresponding surrogates

The following experiments refer to comparisons of constrained sequences against their surrogates (Table 3). Note that surrogates share with the initial sequences the same GC% and length (see the Methods section). As evidenced, comparisons involving invertebrate constrained sequences are not classified as successfully as their human counterparts using LAF. The latter may be understood on the grounds of several particularities of the warm-blooded animals (and often of all vertebrates) especially in their non-functional, non-constrained background genomic fraction. These include a high enrichment in repetitive sequences. Such genomes also present a typical profile of avoided dinucleotides (especially CpG and TpA) that are less avoided in the constrained elements (exons, CNEs), which having functional roles, do not strictly follow the average genomic compositional trends. Note that invertebrate genomes are much less abundant in repeats and less marked by underrepresentation of specific dinucleotides. In the comparisons by means of GS, the same trend is followed, but differences are minor, due to the simpler structure of this genomic metrics. Furthermore, we know from earlier studies that genomic signatures do not perform well in intra-species comparisons because neighbor preferences remain relatively constant within the same genome. Therefore, in all cases listed, LAF performs better than GS.

The last six rows of Table 3 denote comparisons of several CNE sequences collections against their surrogates. In general we notice that among constrained sequences, human CNE sequences vs. surrogates exhibit relatively higher classification rates, if compared with exonic sequences vs. their corresponding surrogates. Although, we are not able at this stage to clearly interpret this finding, it might be related to the hypothesis that CNEs orchestrate highly sophisticated developmental processes including brain formation [11]. The instructions that direct these processes might be hardwired and reflected in their sequences themselves. It is known from the literature that CNEs do serve as transcription factor binding sites and bear several motifs [52,53] that k-mer analysis is possibly sensitive enough to detect and thus increase the found CNE vs. background towards exons vs. background classification rates.

Table 3
Comparison of constrained sequences versus their corresponding surrogates.
Intra-genomic comparisons between pairs of collections of CNEs or exons vs. surrogates. The highest value of LAF or GS accuracy in each classification experiment is shown in bold. Here worm CNEs are studied in their entirety against their surrogates (1869 sequences). For more details on the used methods see in the text and supplementary material.

Experiment no	Description	Average length	Average GC	LAF, accuracy	GS, accuracy
#2	Worm exons	213.37	0.42	65.63	63.75
	Surrogates	213.37	0.42		
#3	Human exons	169.82	0.52	73.31	65.81
	Surrogates	169.82	0.52		
#17	Insect exons	388.82	0.54	63.72	61.95
	Surrogates	381.56	0.54		
#4	Worm UCNEs	82.88	0.43	61.00	57.71
	Surrogates	82.88	0.43		
#5a	Human UCNEs	326.92	0.37	81.15	73.55
	Surrogates	326.92	0.37		
#5b	Human EU100nx CNEs	155.50	0.38	75.95	65.15
	Surrogates	155.50	0.38		
#5c	Amniotic CNEs	289.06	0.38	76.25	66.10
	Surrogates	289.06	0.38		
#5d	Mammalian CNEs	246.49	0.40	73.65	60.00
	Surrogates	246.49	0.40		
#12	Insect UCNEs	86.58	0.39	67.95	66.95
	Surrogates	86.58	0.39		
Average				70.96	64.55

3.3. Intra-species comparison of constrained sequences (functional sequences vs. functional sequences of a different type)

3.3.1. Case of exons vs. UCNEs

We next proceed to the study of the sequence characteristics of elements that are characterized as functional; exons that are known to be under selective constraints and encode for polypeptide chains and CNEs that are mostly of unknown functionality, which is though implied by their high degree of conservation. At a first glance, differences in classification rates between LAF and GS, tabulated in Table 4, may be related to the gradation of GC content between the studied classes. Indeed, it is known from the literature that CNEs are generally AT rich [15], while protein coding genes are relatively GC rich, see also GC content values included in Table 4. However, this may not be entirely true, as classification success in the case of worm exons vs. worm UCNEs is medium (see experiment #6, 68.65%) while CG content difference is minimal. This might indicate that exons and CNEs functional differences are inscribed in their sequence composition. Known such modalities are codon biases and amino-acid frequencies in coding exons, or inscription of several protein binding and other consensus sequences frequent within CNEs.

3.3.2. Case of different classes of CNEs

Here we compare different classes of CNEs which are identified using the same criteria but with gradually increasing similarity thresholds, using a step of 5 percentage points per dataset (see the Methods section). Note that differentiation on the basis of GC composition between these datasets is quite low. However, we still obtain good

Table 4
Comparison of different types of constrained sequences.
Pair-wise intra-genomic comparisons between collections of CNEs, vs. exons, containing 1000 sequences each. The highest value of LAF or GS accuracy in each classification experiment is shown in bold. For more details on the used methods see in the text and supplementary material.

Experiment no	Description	Average length	Average GC	LAF, accuracy	GS, accuracy
#6	Worm exons	82.64	0.42	68.65	61.88
	Worm UCNEs	83.11	0.43		
#7	Human exons	169.82	0.52	82.79	77.66
	Human UCNEs	169.32	0.37		
#18	Insect exons	86.09	0.52	82.80	81.90
	Insect UCNEs	86.58	0.39		
Average				78.08	73.81

Table 5
Comparison of CNEs identified using the same criteria but with different thresholds.
Intra-genomic comparisons between pairs of collections of 1000 human CNE sequences each. These datasets consist of sequences identified between pairwise human/chicken alignments with gradually increasing similarity thresholds. The highest value of LAF or GS accuracy in each classification experiment is shown in bold. For more details on the used methods, see in the text and supplementary material.

Experiment no	Description	Average length	Average GC	LAF, accuracy	GS, accuracy
#24	CNEs 75–80%	248.92	0.39	55.45	53.95
	CNEs 80–85%	248.39	0.38		
#25	CNEs 75–80%	248.92	0.39	57.45	50.00
	CNEs 85–90%	250.63	0.38		
#26	CNEs 75–80%	248.92	0.39	60.95	59.05
	CNEs 90–95%	268.64	0.37		
#27	CNEs 75–80%	248.92	0.39	68.50	63.25
	CNEs 95–100%	319.70	0.37		
#28	CNEs 80–85%	248.39	0.38	50.00	50.00
	CNEs 85–90%	250.63	0.38		
#29	CNEs 80–85%	248.39	0.38	58.90	54.30
	CNEs 90–95%	268.64	0.37		
#30	CNEs 80–85%	248.39	0.38	68.60	61.80
	CNEs 95–100%	319.70	0.37		
#31	CNEs 85–90%	250.63	0.38	58.00	50.00
	CNEs 90–95%	268.64	0.37		
#32	CNEs 85–90%	250.63	0.38	65.65	59.30
	CNEs 95–100%	319.70	0.37		
#33	CNEs 90–95%	268.64	0.37	61.55	50.00
	CNEs 95–100%	319.70	0.37		
#34	CNEs 75–80%	248.92	0.39	76.20	64.10
	Surrogates	248.92	0.39		
#35	CNEs 85–90%	250.63	0.38	78.45	65.25
	Surrogates	250.63	0.38		
#5a	CNEs 95–100%	326.92	0.37	81.15	73.55
	Surrogates	326.92	0.37		
Average				64.68	58.04

classification rates, clearly better than those using genomic signatures. Especially high accuracy values are obtained when we deal with ultraconserved elements (UCNEs, 95–100% similarity between human and chicken, experiments #27, #30, #32, #33, #5a). UCNEs are presumed to form a distinct class, distinguishing themselves from CNEs of the other datasets included in Table 5 comparisons, that are identified from the same organisms but using less stringent criteria of sequence identity. Furthermore, when we compare CNEs 75–80%, CNEs 85–90% and CNEs 95–100% vs. their corresponding surrogate sequences in the human genome, we observe a gradual increase in performance (compare #34 and #35 with #5a). Note here, that we consider for this kind of experiments raw sequences that are not processed in any way (i.e. trimming is not performed). Overall, as we have mentioned earlier, we get the best classification rates when we have UCNEs (Experiment #5a). The information content hardwired in those sequences probably renders them detectable by our method with high efficiency. The genomic signatures approach also performs well in this case, which means that those sequences do probably form a distinct category of ultraconserved sequences from the point of view of first neighbor preferences too.

3.4. Classification based on k-mer analysis and rule based classifiers proves efficient in distinguishing sequences from the same species (intra-species comparisons)

By definition, genomic signatures are widely used for the identification of different species [50]. Using a graphical representation of genomic composition termed Chaos Game Representation (CGR), Deschavanne et al. reported that subsequences of a genome display the main characteristics of the whole genome converging to the genomic signature concept [54]. Interestingly, the authors attributed the variability observed among sequences to base concentrations, runs of complementary bases or purines/pyrimidines and over- or under-expressed words of various lengths with the aim to measure phylogenetic proximity. We categorize the performed experiments as inter-species or intra-species based on whether the sequences under study derive from different or the same organism respectively. Based on the statistics presented in Table 6, it seems that our method performs almost equally well with the GS approach, when dealing with sequences from different species (80.20% vs. 80.19% for the genomic signatures). When it comes to intra-species comparisons, however, the method proposed herein performs fairly better (70.90% vs. 65.83%). This lies in accordance with the fact that k-mer analysis of relatively short sequences combined with logic mining classifiers is a promising method, since it also takes into account the occurrence in the studied sequences of other oligonucleotide stretches such as homopurines, homopyrimidines (unpublished results) or overlapping transcription factor binding sites that are found in the sequences and represent another type of "signature" that distinguishes them from the bulk or from other non-constrained sequences.

4. Conclusions

We present and apply an approach based on k-mer analysis and rule based classification called Logic Alignment Free (LAF) in order to represent and subsequently classify biological sequences of different functionalities and origins. We compare this approach vis-à-vis Karlin's Genomic Signature (GS) method, and present classification rates. To our knowledge, our study is the first one that applies those methodologies on short biological sequences, as up to now GS had only been applied in more extended genomic regions of 50 kb or more [49]. Based on our results, we deduce that LAF performs particularly well when dealing with sequences from the same species surpassing the performance of Genomic Signatures (GS), while in inter-species comparisons, where the potential of GS has already been validated by comparing large genomic regions, the two methods perform equally well. Therefore, LAF analysis of biological sequences could be further used in applications involving sequence analysis such as categorization of different possible functionalities within groups of CNEs or identification of metagenomic components.

Competing interests

The authors declare that they have no competing interests.

Table 6
Overall statistics describing the effectiveness of our method. LAF (Best).
The best combination for each experiment, in terms of k-mer and classifier used, is considered and then averaged, LAF (average of 9): All the different combinations for each experiment of k-mers and rule-based classifiers are averaged, LAF (k = 3, PART): Based on all the experiments performed, the most suitable combination of k and algorithm is used for the analysis in each experiment, GS (without discretization): genomic signatures as calculated without the extra discretization step (described in the Methods section) that we apply when we perform our analysis.

	LAF, accuracy, Best	LAF, accuracy, average of 9	LAF, accuracy, k = 3, PART	GS, accuracy, without discretization	GS, accuracy, Best, with discretization	GS, accuracy, mean of 3, with discretization
Overall	**75.19**	71.33	74.39	70.86	72.45	71.78
Inter-species	**80.20**	75.56	79.81	79.62	80.19	79.53
Intra-Species	**70.90**	67.69	69.74	63.35	65.83	65.14

Authors' contributions

D.P, E.W and Y.A ideated the work, wrote the paper, designed and performed the classification experiments. E.W implemented the scripts and software for feature vector representation and ruled based supervised learning. D.P implemented scripts for genomic analysis and construction of genomic surrogates. S.D generated new CNE datasets. P.B, G.F and Y.A directed research, discussed results and revised the manuscript.

Acknowledgments

E.W and D.P would like to express their gratitude to the organizers of the 2012 European Molecular Biology Organization (EMBO) Practical Course: Bioinformatics and Comparative Genome Analyses in Naples, which enabled this collaboration. This work is partially supported by the Italian PRIN GenData 2020 (2010RTFWBH) and the Flagship Project InterOmics (E.W., PB.05). D.P would also like to thank EMBO for giving him the opportunity to visit the lab of P.B. at EPFL, Switzerland under a short-term fellowship (EMBO ASTF 453-2012).

References

[1] G. Bejerano, M. Pheasant, I. Makunin, S. Stephen, W.J. Kent, J.S. Mattick, et al., Ultraconserved elements in the human genome, Science 304 (2004) 1321–1325.
[2] S. Dimitrieva, P. Bucher, UCNEbase—a database of ultraconserved non-coding elements and genomic regulatory blocks, Nucleic Acids Res. 41 (2013) D101–D109.
[3] Y. Sakuraba, T. Kimura, H. Masuya, H. Noguchi, H. Sezutsu, K.R. Takahasi, et al., Identification and characterization of new long conserved noncoding sequences in vertebrates, Mamm. Genome 19 (2008) 703–712.
[4] E.T. Dermitzakis, A. Reymond, N. Scamuffa, C. Ucla, E. Kirkness, C. Rossier, et al., Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs), Science 302 (2003) 1033–1035.
[5] N. Harmston, A. Baresic, B. Lenhard, The mystery of extreme non-coding conservation, Philos. Trans. R. Soc. Lond. B Biol. Sci. 368 (2013) 20130021.
[6] G. Elgar, T. Vavouri, Tuning in to the signals: noncoding sequence conservation in vertebrate genomes, Trends Genet. 24 (2008) 344–352.
[7] A. Sandelin, P. Bailey, S. Bruce, P.G. Engstrom, J.M. Klos, W.W. Wasserman, et al., Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes, BMC Genomics 5 (2004) 99.
[8] D. Polychronopoulos, D. Sellis, Y. Almirantis, Conserved noncoding elements follow power-law-like distributions in several genomes as a result of genome dynamics, PLoS One 9 (2014) e95437.
[9] L.A. Pennacchio, N. Ahituv, A.M. Moses, S. Prabhakar, M.A. Nobrega, M. Shoukry, et al., In vivo enhancer analysis of human conserved non-coding sequences, Nature 444 (2006) 499–502.
[10] A. Visel, S. Prabhakar, J.A. Akiyama, M. Shoukry, K.D. Lewis, A. Holt, et al., Ultraconservation identifies a small subset of extremely constrained developmental enhancers, Nat. Genet. 40 (2008) 158–160.
[11] M. Matsunami, N. Saitou, Vertebrate paralogous conserved noncoding sequences may be related to gene expressions in brain, Genome Biol. Evol. 5 (2013) 140–150.
[12] E.A. Glazov, M. Pheasant, E.A. McGraw, G. Bejerano, J.S. Mattick, Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing, Genome Res. 15 (2005) 800–808.
[13] T. Vavouri, K. Walter, W.R. Gilks, B. Lehner, G. Elgar, Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans, Genome Biol. 8 (2007) R15.
[14] S. Lockton, B.S. Gaut, Plant conserved non-coding sequences and paralogue evolution, Trends Genet. 21 (2005) 60–65.
[15] K. Walter, I. Abnizova, G. Elgar, W.R. Gilks, Striking nucleotide frequency pattern at the borders of highly conserved vertebrate non-coding sequences, Trends Genet. 21 (2005) 436–440.
[16] S.B. Needleman, C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, J. Mol. Biol. 48 (1970) 443–453.
[17] T.F. Smith, M.S. Waterman, Identification of common molecular subsequences, J. Mol. Biol. 147 (1981) 195–197.
[18] S. Altschul, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[19] W.R. Pearson, Rapid and sensitive sequence comparison with FASTP and FASTA, Methods Enzymol. 183 (1990) 63–98.
[20] J.D. Thompson, T.J. Gibson, D.G. Higgins, Multiple sequence alignment using ClustalW and ClustalX, Curr. Protoc. Bioinforma. 2 (2002) (Unit 2.3).
[21] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res. 32 (2004) 1792–1797.
[22] K. Katoh, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic Acids Res. 30 (2002) 3059–3066.
[23] A. Mokaddeml, M. Elloumi, Motalign: a multiple sequence alignment algorithm based on a new distance and a new score function, 2013 24th Int. Work. Database Expert Syst. Appl., IEEE, 2013, pp. 81–84.
[24] S. Vinga, J. Almeida, Alignment-free sequence comparison—a review, Bioinformatics 19 (2003) 513–523.
[25] D. Kudenko, H. Hirsh, Feature generation for sequence categorization, 1998. 733–738.
[26] P. Kuksa, V. Pavlovic, Efficient alignment-free DNA barcode analytics, BMC Bioinforma. 10 (Suppl. 1) (2009) S9.
[27] Z. Xing, J. Pei, E. Keogh, A brief survey on sequence classification, ACM SIGKDD Explor. Newsl. 12 (2010) 40.
[28] C. McLean, G. Bejerano, Dispensability of mammalian DNA, Genome Res. 18 (2008) 1743–1751.
[29] S. Stephen, M. Pheasant, I.V. Makunin, J.S. Mattick, Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock, Mol. Biol. Evol. 25 (2008) 402–408.
[30] J.E. McCormack, B.C. Faircloth, N.G. Crawford, P.A. Gowaty, R.T. Brumfield, T.C. Glenn, Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis, Genome Res. 22 (2012) 746–754.
[31] W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, et al., The human genome browser at UCSC, Genome Res. 12 (2002) 996–1006.
[32] S. Dimitrieva, P. Bucher, Genomic context analysis reveals dense interaction network between vertebrate ultra-conserved non-coding elements, Bioinformatics 28 (2012) i395–i401.
[33] S.Y. Kim, J.K. Pritchard, Adaptive evolution of conserved noncoding elements in mammals, PLoS Genet. 3 (2007) 1572–1586.
[34] D. Retelska, E. Beaudoing, C. Notredame, C.V. Jongeneel, P. Bucher, Vertebrate conserved non coding DNA regions have a high persistence length and a short persistence time, BMC Genomics 8 (2007) 398.
[35] A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics 26 (2010) 841–842.
[36] P. Rice, I. Longden, A. Bleasby, EMBOSS: the European Molecular Biology Open Software Suite, Trends Genet. 16 (2000) 276–277.
[37] T. Lehr, J. Yuan, D. Zeumer, S. Jayadev, M.D. Ritchie, Rule based classifier for the analysis of gene-gene and gene-environment interactions in genetic association studies, BioData Min. 4 (2011) 4.
[38] E. Weitschek, G. Fiscon, G. Felici, Supervised DNA Barcodes species classification: analysis, comparisons and results, BioData Min. 7 (2014) 4.
[39] P. Bertolazzi, G. Felici, E. Weitschek, Learning to classify species with barcodes, BMC Bioinforma. 10 (Suppl. 1) (2009) S7.
[40] E. Weitschek, A. Lo Presti, G. Drovandi, G. Felici, M. Ciccozzi, M. Ciotti, et al., Human polyomaviruses identification by logic mining techniques, Virol. J. 9 (2012) 58.
[41] W.W. Cohen, Y. Singer, A simple, fast, and effective rule learner, 1999. 335–342.
[42] B.R. Gaines, P. Compton, Induction of ripple-down rules applied to modeling large databases, J. Intell. Inf. Syst. 5 (1995) 211–228.
[43] E. Frank, I.H. Witten, Generating accurate rule sets without global optimization, Proceedings of the Fifteenth International Conference on Machine Learning, 1998, pp. 144–151.
[44] J.R. Quinlan, C4.5: Programs for Machine Learning, 1993.
[45] G. Marçais, C. Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers, Bioinformatics 27 (2011) 764–770.
[46] H. Teeling, J. Waldmann, T. Lombardot, M. Bauer, F.O. Glöckner, TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences, BMC Bioinforma. 5 (2004) 163.
[47] V. Brendel, J.S. Beckmann, E.N. Trifonov, Linguistics of nucleotide sequences: morphology and comparison of vocabularies, J. Biomol. Struct. Dyn. 4 (1986) 11–21.
[48] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software, ACM SIGKDD Explor. Newsl. 11 (2009) 10.
[49] S. Karlin, Global dinucleotide signatures and analysis of genomic heterogeneity, Curr. Opin. Microbiol. 1 (1998) 598–610.
[50] S. Karlin, J. Mrázek, Compositional differences within and between eukaryotic genomes, Proc. Natl. Acad. Sci. U. S. A. 94 (1997) 10227–10232.
[51] S. Karlin, C. Burge, Dinucleotide relative abundance extremes: a genomic signature, Trends Genet. 11 (1995) 283–290.
[52] T. Viturawong, F. Meissner, F. Butter, M. Mann, A DNA-centric protein interaction map of ultraconserved elements reveals contribution of transcription factor binding hubs to conservation, Cell Rep. 5 (2013) 531–545.
[53] X. Xie, T.S. Mikkelsen, A. Gnirke, K. Lindblad-Toh, M. Kellis, E.S. Lander, Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 7145–7150.
[54] P.J. Deschavanne, A. Giron, J. Vilain, G. Fagot, B. Fertil, Genomic signature: characterization and classification of species assessed by chaos game representation of sequences, Mol. Biol. Evol. 16 (1999) 1391–1399.

Analysis and Classification of Constrained DNA Elements with N-gram Graphs and Genomic Signatures

Dimitris Polychronopoulos[1,4], Anastasia Krithara[2], Christoforos Nikolaou[3], Giorgos Paliouras[2], Yannis Almirantis[1], and Giorgos Giannakopoulos[2] ⋆

[1] Institute of Biosciences and Applications, NCSR Demokritos, 15310 Athens, Greece.
[2] Institute of Informatics and Telecommunications, NCSR Demokritos, 15310 Athens, Greece.
[3] Department of Biology, University of Crete, 71409 Heraklion, Greece.
[4] Department of Biochemistry and Molecular Biology, Faculty of Biology, National and Kapodistrian University of Athens, 15701 Athens, Greece.

Abstract. Most common methods for inquiring genomic sequence composition, are based on the bag-of-words approach and thus largely ignore the original sequence structure or the relative positioning of its constituent oligonucleotides. We here present a novel methodology that takes into account both word representation and relative positioning at various lengths scales in the form of n-gram graphs (NGG). We implemented the NGG approach on short vertebrate and invertebrate constrained genomic sequences of various origins and predicted functionalities and were able to efficiently distinguish DNA sequences belonging to the same species (intra-species classification). As an alternative method, we also applied the Genomic Signatures (GS) approach to the same sequences. To our knowledge, this is the first time that GS are applied on short sequences, rather than whole genomes. Together, the presented results suggest that NGG is an efficient method for classifying sequences, originating from a given genome, according to their function.

Keywords: genomic sequence representation, n-gram graphs, conserved non-coding elements, CNEs, UCEs, ultraconserved elements, classification, genomic signatures

1 Introduction

1.1 Constrained Elements in Eukaryote Genomes

High throughput sequencing at a massive scale combined with comparative genome analysis has led to the discovery of a variety of constrained genomic elements. In fact, there are many more selectively constrained noncoding than protein-coding sequences in mammalian genomes. One of the most interesting discoveries that have arisen from comparative genomics among mammalian genomes are the hundreds of such noncoding elements of more than 200bp in length that show absolute conservation among mammalian orders [1]. These only represent the tip of the iceberg of a much larger class of conserved noncoding elements (CNEs), a general class of sequence elements that are significantly more conserved than protein-coding genes and non-coding RNAs (ncRNAs)[2]. In the following analysis, we implement the term "constrained elements" for both protein-coding and conserved non-coding sequence stretches, while we refer to noncoding as strictly not protein-coding.

Conserved non-coding elements can be found in the literature under various definitions, depending on the percentage of identity between two or more organisms and the minimum length. Increasing evidence suggests that CNEs are selectively constrained and not mutational cold-spots

⋆ Corresponding author.

[3] and there is a plethora of studies indicating possible functions of those elements [2]. Among the various reported functional roles of CNEs, enhancers appear to be the most plausible. Nevertheless, the relative abundance, genomic distribution and variable length of CNEs is indicative of various alternatives. CNEs have been shown to bear resemblance to CTCF insulator sites [4], matrix attachment regions [5], while a recent, concise study across 29 mammals revealed constrained elements of smaller sizes that may be directly related to transcriptional regulation as well as to the encoding of functional RNA molecules [6]. CNE existence may be extended even further back in evolutionary time as suggested by recent works including both bony and cartilaginous fish species [7]. While the majority of the analyses have been conducted in mammals, there is growing evidence that CNEs are not a vertebrate innovation and can also be found in invertebrates and plants. Despite the fact that vertebrate and invertebrate CNEs bear no sequence identity, they share common sequence characteristics, indicating a parallel evolution of those sequences in order to perform the same, possibly essential, functions [8].

Among the various attributes of these genomic sequences, DNA composition has been greatly ignored. When compared to non-CNEs and near-promoter sequences, CNEs possess an excess of AT-rich motifs, often containing runs of identical nucleotides. In a recent paper, Walter *et al* have analysed the base composition of human and Fugu CNEs at single nucleotide level [9]. They have found that those elements are A+T rich, much more so than the region they reside in, in contrast to their flanking region just outside their boundaries, which exhibits a marked drop in A+T content that forms a unique pattern. Such compositional extremes are strong indications of functionality as has been shown for gene-dense regions and CpG islands [10,11]. It is therefore of great interest to further investigate the compositional preference of constrained regions in greater detail. To this end, conventional approaches addressing composition through histograms, or bag-of-words approaches tend to overlook the positional information, while probabilistic sequential methods like HMMs are likely to undermine the effect of local sequence boundaries. In the following we describe a novel methodology that is able to address both such aspects with the additional advantage of allowing for similarity measurements between any two sequences.

1.2 Analyzing Sequence Composition Through a Combination of Word-content and Relative Positioning Information

A sequence can easily be considered equivalent to a natural language text, under the assumption that the vocabulary is very limited. Traditionally, natural language processing methods based on n-grams (n-nucleotides correspondingly) have been applied on biological sequences, aiming to support sequence matching [12], indexing [13], analysis of protein sequences [14] and coding and non-coding DNA sequences [15]. The n-gram models of sequences indicate how short subsequences of length n appear in the whole analyzed sequence. Other alternatives of analysis, like Hidden Markov Models or Conditional Random Fields [16], model probabilistically the possible combinations of elements in a sequence.

In this work, we propose the application of the n-gram graph (NGG) representation methodology [17], which manages to capture both local and global characteristics of the analysed sequences. The main idea behind the n-gram graphs is that the neighborhood between sub-sequences in a sequence contains a crucial part of the sequence information. The n-gram graph, as derived from a single sequence, is essentially a histogram of the co-occurrences of symbols. The symbols are considered to co-occur when found within a maximum distance (window) of each other. The size of the window, which is a parameter of the n-gram graph, allows for fuzziness in the representation of co-occurrences within a sequence. The fact that n-gram graphs take into account co-occurrences offers the local descriptiveness. The fact that they act as a histogram of such co-occurrences provides their global representation potential.

As opposed to probabilistic models, the n-gram graphs are deterministic. As opposed to n-gram models, n-gram graphs offer more information, based on the representation of co-occurrences. Overall, they provide a trade-off between expressiveness and generalization.

The n-gram graph framework, also offers a set of important operators. These operators allow combining individual graphs into a model graph (the update operator), and comparing pairs of graphs providing graded similarity measurements (similarity operators). In the sequence composition setting, the representation and set of operators provide one more means of analysis and comparison, one that is lacking from widely-implemented probabilistic models such as HMMs.

Finally, the n-gram graphs can be combined with vector representation of sequences to allow the application of machine learning methods for the classification of sequences. Within this study, both conserved non-coding and protein coding segments are analyzed through a NGG-based approach and an approach based on the method of genomic signatures [18]. Additionally, datasets of CNEs and protein coding segments(coding exons) are studied along with suitably chosen (see Methods) surrogate sequence sets.

2 Methods

2.1 Datasets Retrieval

We consider various published datasets of constrained sequences. Human, worm and insect denote sequences taken from *H.sapiens*, *C.elegans* and *D.melanogaster* genomes. Given that those datasets are heterogeneous in numbers, we randomly select 1000 elements from each set for our subsequent analysis. Only worm CNEs are studied in their entirety, as this dataset is relatively small. The exonic sequences of human, worm and insect genomes were obtained from the UCSC genome repository based on the *RefSeq* annotation referring to the latest genome assemblies (hg19, ce10, dm3) [19]. The datasets described below along with their surrogates are used in 26, in total, pairwise classification experiments, denoted throughout the text as #1, #2, ...; see Supplementary Spreadsheet[5] where further information is provided. Apart from exonic sequences, the following classes of constrained non-exonic sequences are used:

- UltraConserved Noncoding Elements (UCNEs): These are sequences of at least 200bp in length mapped on the human genome (hg19) that display sequence similarity which is greater than or equal to 95% between human and chicken whole genome alignments [20]
- EU100 nonexonic CNEs (EU100nx CNEs): These are sequences mapped on the human genome (hg18) that are identical over 100bp in at least 3 out of 5 placental mammals (human, mouse, rat, dog and cow) [21]. The whole set is named EU100+ and since we remove elements overlapping exons it will be referred to as EU100nx.
- Amniotic and Mammalian CNEs: These are elements identified by Kim and Pritchard[22]. Mammalian CNEs are sequences that are conserved within mammals but not found in chicken or fish, while Amniotic CNEs are conserved in mammals and chicken but not found in fish. LiftOver [23] is used to convert the coordinates to the most recent release of the human genome (from hg17 to hg19).
- Worm and Insect UCNEs: These are elements mapped on ce10 and dm3 genome releases of *C.elegans* and *D.melanogaster* respectively. Worm UCNEs are DNA stretches longer than 60bp that exhibit sequence similarity greater than 90% among *C.elegans* and *C.japonica*, while Insect UCNEs are stretches longer than 60bp that display sequence similarity greater than 90% among *D.melanogaster* and *D.virilis* (unpublished results, Philipp Bucher's group, EPFL)

[5] You can download the spreadsheet from `http://users.iit.demokritos.gr/~ggianna/Publications/alcob2014/`.

2.2 Treatment of Sequences and Extraction of Surrogate Sequences

A useful suite of tools called BEDTools [24] is used in order to extract FASTA sequences from BED files and to calculate overlapping elements. Additionally, we make use of the EMBOSS suite to calculate fractional GC content of sequences [25]. It is known from the literature that vertebrate and invertebrate CNEs are of significantly different lengths (the ones belonging to the latter category are considerably smaller) [26]. We make sure that we take segments of equal lengths as follows: for classification experiments involving CNE sets of vertebrate and invertebrate origin, we truncate the vertebrate ones around their middle point to the length of the shortest invertebrate CNE included in the experiment. For each element of each collection under study (CNE or exon), an analogue of it is extracted from the corresponding genome. Every resulting DNA segment (surrogate sequence) is ensured that is of the same length and GC content (within a 1% deviation limit) with its corresponding element in the collection under study. Statistics of the datasets (mean length and GC content of sequences) are available in the Supplementary Spreadsheet. All custom shell scripts are available upon request to interested readers. All the files (coordinates in BED and sequences in FASTA format) are also available upon request.

2.3 From Sequences to the N-gram Graph Similarity Vector Space

We have followed a set of steps to represent our sequences using n-gram graphs. The idea is that from known (labeled) sequences we form representatives of each class of sequences. Then, we describe all sequences based on their similarity to the representatives of each class. Thus, there are essentially three steps in our application of the n-gram graphs:

- Representation of individual sequences using n-gram graphs.
- Calculation of representative (model) n-gram graphs for each training class used.
- Calculation of similarity between training instances and model graphs.
- Representation of instances using only their similarities, i.e., in a similarity vector space.

In the following paragraphs we elaborate on these steps.

Representation of Sequences The *n-gram graph* is a graph $G =< V^G, E^G, L, W >$, where V^G is the set of vertices, E^G is the set of edges, L is a function assigning a label to each vertex *and to each edge* and W is a function assigning a weight to every edge. The graph has n-grams labeling its vertices $v^G \in V^G$. The edges $e^G \in E^G$ (the superscript G will be omitted where easily assumed) connecting the n-grams indicate proximity of the corresponding vertex n-grams. The weight of the edges can indicate a variety of traits. In our implementation we apply as weight the number of times the two connected n-grams were found to co-occur. It is important to note that in n-gram graphs *each vertex is unique*. To ensure that no duplicate vertices exist, we also require that the labelling function is an one-to-one function. Two vertices are considered equal if and only if their labels are equal.

We repeat that the edges E are assigned weights of $c_{i,j}$ where $c_{i,j}$ is the number of times a given pair S_i, S_j of n-grams happen to be neighbors in a string within some distance D_{win} of each other. The distance d of two n-grams $S^{(i,i+a)}, S^{(j,j+b)}$ is $d = |i - j|$. To create the n-gram graph from a given sequence, a fixed-width window D_{win} of characters (or words) around a given n-gram $N_0 \equiv S^r, r \in \mathbb{N}^*$ is used. All character n-grams within the window are considered to be neighbors of N_0. These neighbors are represented as connected vertices in the text graph. We use a symmetric approach, where a window of length $2 \times D_{\text{win}} + 1$ runs over the text, centered at the beginning of N_0. If the n-gram we are interested in is located at position p_0, then the window will span from $p_0 - D_{\text{win}}$ to $p_0 + D_{\text{win}}$, taking into account *both preceding and succeeding*

characters or words. Each edge $e =< a, b >$ is weighted based on the number of co-occurrences of the neighbors within a window in the text:

$$w(e) = |\{S^{(i,i+n)} = L(a), S^{(j,j+n)} = L(b) : abs(i-j) \leq D_{\text{win}}, i \neq j\}| \quad (1)$$

Creation of Representative Graphs The n-gram graph representation specification indicates how to represent a text using an n-gram graph. However, in sequence analysis it is often required to represent a whole sequence set, i.e. a sequence class. In our applications we have used the *update function U* to represent sets. The update function $U(G_1, G_2, l)$ takes two graphs as input. One graph is considered to be the pre-existing graph G_1 and one that is considered to be the new graph G_2.

The U function takes an additional argument, the *learning factor* $l \in [0, 1]$, which determines the sensitivity of G_1 to the change G_2 brings. The higher the value of learning factor, the higher the impact of the new graph to the existing graph. The definition of the weighting performed in the graph resulting from $U(G_1, G_2, l)$ is:

$$w(e) = w_1(e) + (w_2(e) - w_1(e)) \times l \quad (2)$$

for every edge $e \in E_1 \cup E_2$.

The U function allows creating a representative graph for a set of documents, in analogy to the centroid of a set of vectors. The creation of a set-representative graph G_s, for a given set of graphs $\mathbb{G} = \{G_1, G_2, G_3, ..., G_n\}$ is as follows. We initialize the class graph with the first document of the class $G_s = G_1$. Then, iteratively we update G_s by:

$$G_s = U(G_s, G_i, \frac{1}{i+1}) \text{ for } i \in 2, n. \quad (3)$$

After all iterations the weight of every edge $e \in E_s$ will have a weight of:

$$w_e = \frac{\sum\limits_{i:e \in E_i} w_i}{|i : e \in E_i|} \quad (4)$$

This weight is the average of the weights of edge e over all the graphs G_i, where it appears. This means that graphs where the edge does not appear, *are not taken into account* in this calculation. This essentially means that having a graph with an edge e, where $w(e) = 0$ is different than not having the edge at all.

Measuring Similarity and Vector Space Representation In the n-gram graph framework there are different ways to measure similarity. We choose the Value Similarity function [17]. This measure quantifies the ratio of common edges between two graphs, taking into account the ratio of weights of common edges. In this measure each matching edge e having weight w_e^i in graph G^i contributes $\frac{\text{VR}(e)}{\max(|G^i|,|G^j|)}$ to VS, while not matching edges do not contribute (consider that for an edge $e \notin G^i$ we define $w_e^i = 0$).

The *ValueRatio (VR)* scaling factor is defined as:

$$\text{VR}(e) = \frac{\min(w_e^i, w_e^j)}{\max(w_e^i, w_e^j)} \quad (5)$$

The equation indicates that the *ValueRatio* takes values in $[0, 1]$, and is symmetric. Thus, the full equation for *VS* is:

$$\text{VS}(G_i, G_j) = \frac{\sum_{e \in G_i} \frac{\min(w_i^e, w_j^e)}{\max(w_i^e, w_j^e)}}{\max(|G_i|, |G_j|)} \quad (6)$$

Given the similarity function $\mathrm{VS}(G_i, G_j)$, a sequence instance S with a corresponding graph G_S and class-representative graphs G_i for classes C_i, we can describe S in a similarity vector space. In this space, each dimension reflects the similarity of the instance sequence to a corresponding class. Thus, the vector $\overline{V}(S)$ is of the form

$$\overline{V}(S) =< \mathrm{VS}(G_S, G_1), \mathrm{VS}(G_S, G_2), ..., \mathrm{VS}(G_S, G_n) > \quad (7)$$

where n is the number of classes we have. Thus, the above process allows us to map each sequence in a dataset to a vector space of similarities.

In the following section we discuss another representation of sequences we use in our experiments: genomic signatures.

2.4 The "Genomic Signature" Method

This is a standard methodology for classifying and distinguishing genomes based on the quantification of neighbor preferences in a DNA sequence of an entire genome by computing the vector of the odds ratios for dinucleotides [27]. The odds ratio of each dinucleotide is the quantity: $r_{ij} = f_{ij}/(f_i f_j)$, where f_{ij} and f_i, f_j stand for the frequencies of occurrence in the studied sequence of a dinucleotide and its constituent nucleotides respectively. Therefore, the subscripts i, j represent any pair of A, G, C and T. This is the ratio of the "observed" dinucleotide frequency over the "expected" one under no neighbor effects, thus it expresses the actual neighbor preferences of the given pair of nucleotides. Before computing the odds ratios for a given sequence, this is concatenated to its reverse complement. Consequently, the relevant ratios are only ten, i.e. four for the self-complementary dinucleotides and six for the mutually complementary couples. Karlin and co-workers first proposed that these quantities differentiate between different genomes, according, approximately, to their evolutionary distance. Thus they have assigned to the vector of these ten "first neighbour preferences" the name of genomic signature (GS). It has to be noted that GS filter out mononucleotide composition retaining only the first neighbor preferences.

For a direct comparison with the n-gram graphs (NGG) approach described in the previous section, we use classification based on genomic signatures (GS) and apply that to the same pairs of genomic sequences. When applying classification, in both the NGG and the GS approaches we use the output of the analysis (similarity vectors, or ratio correspondingly) as input vectors to train a well-known, rule-based classifier: the JRIP implementation [28] of RIPPER [29]. We note that the results provided by RIPPER are comparable to those of other state-of-the-art classifiers we tried (e.g., Support Vector Machine based classification). In the following section we report our findings based on experiments using both analysis methods (NGG and GS).

3 Results and Discussion

In this section we describe a systematic analysis of short genomic segments, which display different functionalities and stem from human, worm and insect genomes. For every classification task, we adopt the technique of NGG and GS as explained previously in order to estimate how different modalities could be separated based on sequence composition.

We note that the n-gram graph representation embeds three parameters. The first is the value of m, which defines the minimum length of the n-grams into which a sequence is split. The second is M, which defines the maximum length of the n-grams into which a sequence is split. The third is D which represents the maximum distance within which we consider the n-grams (n-bases) to be neighbors. Keeping $m = M = D$ simplifies the analysis step reducing the required time, while not significantly altering the results in most cases [17].

An estimator described in the the original n-gram graphs' work [17] was used in order to define these parameters. As it was devised for a different, linguistic task (summarization system evaluation) not taking into account the prediction potential of n-grams for a classification task, we performed an approximate optimization based on exhaustive experiments (for $m \in [2, 9]$) using 10-fold cross validation. The tuning was performed using 6 different datasets. In Figure 1 we illustrate the performance of the n-gram graph based classification. Each column corresponds to a parameter value combination. The strong horizontal line in each column indicates the average performance of this combination over the set of 6 datasets. The remaining lines represent the quantile values of the performance. We deduce that there exists a persistent high plateau of performance when $5 \leq m \leq 8$, which shows that the method does not improve significantly after $m = 5$. Another observation is that 2-2-2 performs considerably better than 3-3-3 and almost equally to 4-4-4. This might be related to the importance of first neighbor preferences as reflected in the NGG methodology.

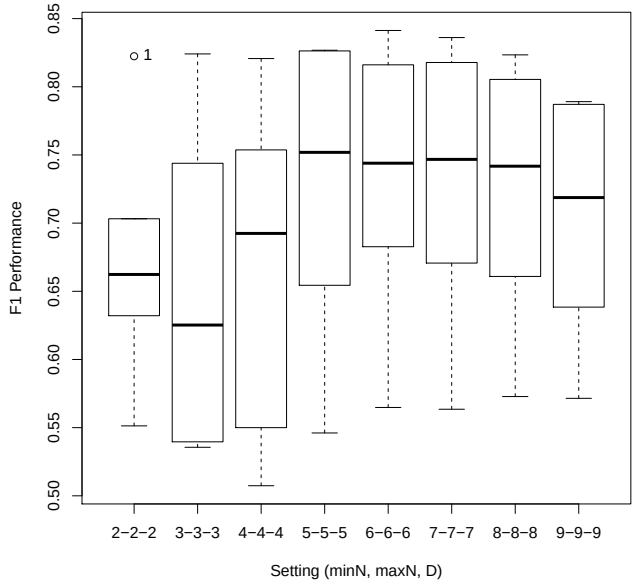


Fig. 1. Boxplots of the performance of n-gram graphs classification over varying parameters

In the tables included throughout the Discussion section, the NGG performance obtained for the optimal combination of parameter values is included; e.g. in Exp #1, Table 1, the value 83, 86 is given for NGG, corresponding to the parameter choice (7,7,7). We propose $(m, M, D) = (5, 5, 5)$ as the best parameter settings for analyzing short genomic sequences, based on all the 26 performed experiments (see Supplementary Spreadsheet, Overall Statistics tab). Genomic signatures have been used as an alternative to the NGG based classification method. To our knowledge, for such short genomic segments of possible functional importance, not only the NGG method but also the genomic signature approach has not been applied before.

3.1 Inter-Species Comparisons of Background Sequences

For comparisons between human, worm and insect background sequences, we use the surrogate sets described in Methods as representative samples of the different genomes. Comparisons involving *H. sapiens* yield always the best classification rates using both n-gram graphs (NGG) and Genomic Signatures (GS), see Table 1. This may be understood on the grounds of the high difference of neighbor preferences, mainly in CpG and TpA between *H.sapiens* and the invertebrates, while these preferences are found to be quite close between invertebrate species: *D.melanogaster* (insect) and *C.elegans* (worm). GS are exclusively a quantification of first neighbor preferences and as a consequence are not influenced by GC content. On the contrary, NGG are able to conceptually incorporate various components/aspects of sequence composition, such as mononucleotide composition and higher order neighbor preferences (a criterion that is provided as a parameter value in our technique by adjusting D).

In cases where human sequences are included in the comparison, GS perform systematically better than NGG. This difference is also statistically significant with a p-value below 0.10, based on a paired t-test. It is also worth mentioning that the two cases with the highest differences in GC content between the sets involved in the classification experiment are: experiment #22 which is the only one where NGG perform better than GS; and experiment #1 where GS perform better than NGG, but with the lowest difference in their performances. I.e. the cases of the highest relative preponderance of NGG are the ones with the highest differences in GC content, as expected due to the relative sensitivities of the two methods to this composition parameter.

Table 1. Inter-species comparisons of background genomic sequences

Exp	Description	Average length	Average GC	NGG	GS
#1	surrogates for human exons	167.837	0.5155	83.86	85.98
	surrogates for worm exons	169.318	0.4049		
#14	surrogates for human UCNEs	86.094	0.3651	79.38	84.05
	surrogates for insect UCNEs	86.582	0.3949		
#20	surrogates for insect exons	169.318	0.5202	80.48	87.49
	surrogates for human exons	169.816	0.5087		
#22	surrogates for worm exons	213.365	0.4194	73.50	70.35
	surrogates for insect exons	212.858	0.5194		
#23	surrogates for human UCNEs	82.932	0.3648	80.35	83.75
	surrogates for worm UCNEs	82.875	0.4297		
#13	surrogates for worm UCNEs	83.407	0.4265	58.79	64
	surrogates for insect UCNEs	86.582	0.3949		
	Average			**76.06**	**79.27**

3.2 Classification Experiments of Constrained DNA Sequences Versus their Background Surrogates (Intra-Species Classification)

The following experiments refer to comparisons of constrained sequences against their surrogates (Table 2). Note that surrogates share the same GC% and length with the initial sequences (see Methods). As evidenced from inspection of Table 2, comparisons involving invertebrate constrained sequences are not classified as successfully as their human counterparts using NGG, see experiments #2 and #17. Only insect UCNEs appear to be resistant to this effect, see experiment

#12. This finding might be understood on the grounds of several particularities of the warm-blooded animals (often of all vertebrates) especially in their non-functional, non-constrained background genomic fraction. These include a high enrichment in transposable elements, microsatellites, homo-purine/homo-pyrimidine tracts and several other characteristic compositional motifs. Such genomes also present a typical profile of avoided dinucleotides (especially CpG and TpA) that are less avoided in the constrained elements (exons, CNEs), which having functional roles, do not strictly follow the average genomic compositional trends. Note that invertebrate genomes are much less abundant in repeats and less marked by under-representation of specific dinucleotides. In the comparisons by means of GS, the same trend is followed but the differences are minor, due to the sensitivity of these quantities solely upon first neighbor preferences. We know from earlier studies that genomic signatures do not perform well in intra-species comparisons because neighbor preferences remain relatively constant within the same genome. Consequently, in most cases listed, NGG perform better than GS (compare averages, 68,76% versus 61,84%). The difference is statistically significant (p-value below 0.10, based on a paired t-test).

Table 2. Classification of constrained DNA sequences versus background surrogates

Exp	Description	Average length	Average GC	NGG	GS
#2	worm exons	213.365	0.4243	57.7	61.05
	surrogates	213.365	0.4239		
#3	human exons	169.816	0.5190	74.3	63.41
	surrogates	169.816	0.5183		
#17	insect exons	388.822	0.5412	52.36	59.45
	surrogates	381.557	0.5389		
#4	worm UCNEs	82.875	0.4309	56.76	55.62
	surrogates	82.875	0.4308		
#5a	human UCNEs	326.923	0.3676	82.63	72.00
	surrogates	326.923	0.3676		
#5b	human EU100nx CNEs	155.499	0.3783	76.43	63.75
	surrogates	155.499	0.3783		
#5c	amniotic CNEs	289.061	0.3756	78.62	63.00
	surrogates	289.061	0.3756		
#5d	mammalian CNEs	246.488	0.4015	75.85	55.65
	surrogates	246.488	0.4018		
#12	insect UCNEs	86.582	0.3949	64.15	62.65
	surrogates	86.582	0.3949		
	Average			**68.76**	**61.84**

The six last rows of Table 2 denote comparisons of several CNE sequences collections against their surrogates. In general we notice that among constrained sequences, human CNE sequences versus surrogates exhibit relatively higher classification rates, if compared with exonic sequences versus their corresponding surrogates. Although, we are not able at this stage to clearly interpret this finding, it might be related to the hypothesis that CNEs orchestrate highly sophisticated developmental processes including brain formation [30]. The instructions that direct these processes might be hardwired and reflected in the sequences themselves. In fact, it is known from the literature that CNEs do serve as transcription factor binding sites and bear several motifs [31] that NGG analysis is possibly sensitive enough to detect and thus, increase the obtained CNE-background vs. exon-background classification rates.

3.3 Genomic Signatures Perform Better in Classifying Genomic Segments of Functional Importance and Different Origin

In the experiments included in Table 3 we study the performance of NGG and GS when they are applied on sets of functional sequences (CNEs or exons) of different genomes. We verify that GS perform better in inter-genomic classification and their performance is slightly improved in Table 3, if compared to Table 1. This means that it is not blurred due to the constrained character of the sequences, i.e. the first neighbour preference characterizing different genomes and filtered by GS is clearly retained in exons and CNEs. Once again, we tested the statistical significance in performance between the NGG and GS methods, based on a paired t-test, and there exists statistically significant difference (with a p-value below 0.05).

On the other hand, the performance of NGG drops. This might be associated with the complex set of compositional traits that are characteristic of the NGG method: nucleotide composition of the sequence, first neighbor preferences and also higher neighbor relationships. Thus, in this context, the inter-genomic differences leading to an efficient NGG based classification of background sequences (Table 1, average = 76.06) are blurred by the presence of common compositional constraints related to function as shown by inspection of Table 3 (average = 74.47). Such an example is the known over-representation of adenine and guanine in the composition of protein coding exons (purine loading).

Table 3. Comparisons of functional sequences between genomes

Exp	Description	Average length	Average GC	NGG	GS
#9	worm exons	169.318	0.3886	74.68	82.21
	human exons	169.816	0.5190		
#10	worm UCNEs	83.113	0.4277	77.77	82.29
	human UCNEs	82.640	0.3728		
#15	worm UCNEs	83.086	0.4285	70.89	74.95
	insect UCNEs	86.582	0.3949		
#16	human UCNEs	86.094	0.3704	82.08	86.70
	insect UCNEs	86.582	0.3949		
#19	insect exons	169.318	0.5148	70.03	81.29
	human exons	169.816	0.5089		
#21	worm exons	213.365	0.4196	71.39	72.35
	insect exons	212.858	0.5093		
	Average			**74.47**	**79.97**

4 Conclusions

The problem of assigning functionality to un-annotated sequences is highly relevant in light of the complexity of mammalian genomes. By performing various comparisons between representative sequences of different origin and functionality, we were able to show the potential of the n-gram graphs (NGG) in effectively discriminating between genomic sequence fragments with expected function against the bulk of the genome. NGG were able to quantify the effect of sequence constraint, thus their implementation in the description of other functional sequences of mammalian genomes (such as regions with structural properties, matrix attachment regions and various non-coding RNA species) may provide valuable insight in the interplay between

composition and function. When it comes to inter-genomic comparisons between sequences of different origin, NGG are less effective compared to the method of Genomic Signatures (GS). The efficiency of the latter is for the first time demonstrated here at this length scale, as up to now GS had only been applied in more extended genomic regions of 50kb or more [32]. Further work is needed for a better exploitation of the presented results, while combinations of the two methods may result in higher classification rates.

Acknowledgments

D.P. would like to thank Professors Georgios Rodakis and Stavros Hamodrakas (Faculty of Biology, University of Athens) for serving as academic advisors. We would also like to express our gratitude to Dr Slavica Dimitrieva and Dr Philipp Bucher for providing unpublished datasets of worm and insect UCNEs

References

1. Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., Haussler, D.: Ultraconserved elements in the human genome. Science **304**(5675) (2004) 1321–1325
2. Harmston, N., Baresic, A., Lenhard, B.: The mystery of extreme non-coding conservation. Philosophical transactions of the Royal Society of London. Series B, Biological sciences **368**(1632) (December 2013) 20130021
3. Drake, J.A., Bird, C., Nemesh, J., Thomas, D.J., Newton-Cheh, C., Reymond, A., Excoffier, L., Attar, H., Antonarakis, S.E., Dermitzakis, E.T., Hirschhorn, J.N.: Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genet **38**(2) (2006) 223–227
4. Xie, X., Mikkelsen, T.S., Gnirke, A., Lindblad-Toh, K., Kellis, M., Lander, E.S.: Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. Proc Natl Acad Sci U S A **104**(17) (2007) 7145–7150
5. Glazko, G.V., Koonin, E.V., Rogozin, I.B., Shabalina, S.A.: A significant fraction of conserved noncoding DNA in human and mouse consists of predicted matrix attachment regions. Trends Genet **19**(3) (2003) 119–124
6. Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., Ward, L.D., Lowe, C.B., Holloway, A.K., Clamp, M., Gnerre, S., Alfoldi, J., Beal, K., Chang, J., Clawson, H., Cuff, J., Di Palma, F., Fitzgerald, S., Flicek, P., Guttman, M., Hubisz, M.J., Jaffe, D.B., Jungreis, I., Kent, W.J., Kostka, D., Lara, M., Martins, A.L., Massingham, T., Moltke, I., Raney, B.J., Rasmussen, M.D., Robinson, J., Stark, A., Vilella, A.J., Wen, J., Xie, X., Zody, M.C., Baldwin, J., Bloom, T., Chin, C.W., Heiman, D., Nicol, R., Nusbaum, C., Young, S., Wilkinson, J., Worley, K.C., Kovar, C.L., Muzny, D.M., Gibbs, R.A., Cree, A., Dihn, H.H., Fowler, G., Jhangiani, S., Joshi, V., Lee, S., Lewis, L.R., Nazareth, L.V., Okwuonu, G., Santibanez, J., Warren, W.C., Mardis, E.R., Weinstock, G.M., Wilson, R.K., Delehaunty, K., Dooling, D., Fronik, C., Fulton, L., Fulton, B., Graves, T., Minx, P., Sodergren, E., Birney, E., Margulies, E.H., Herrero, J., Green, E.D., Haussler, D., Siepel, A., Goldman, N., Pollard, K.S., Pedersen, J.S., Lander, E.S., Kellis, M.: A high-resolution map of human evolutionary constraint using 29 mammals. Nature **478**(7370) (2011) 476–482
7. Lee, A.P., Kerk, S.Y., Tan, Y.Y., Brenner, S., Venkatesh, B.: Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes. Mol Biol Evol **28**(3) (2011) 1205–1215
8. Vavouri, T., Walter, K., Gilks, W.R., Lehner, B., Elgar, G.: Parallel evolution of conserved noncoding elements that target a common set of developmental regulatory genes from worms to humans. Genome Biol **8**(2) (2007) R15
9. Walter, K., Abnizova, I., Elgar, G., Gilks, W.R.: Striking nucleotide frequency pattern at the borders of highly conserved vertebrate non-coding sequences. Trends Genet **21**(8) (2005) 436–440
10. Touchon, M., Arneodo, A., d'Aubenton Carafa, Y., Thermes, C.: Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes. Nucleic acids research **32**(17) (2004) 4969–4978

11. Zhang, L., Kasif, S., Cantor, C.R., Broude, N.E.: Gc/at-content spikes as genomic punctuation marks. Proceedings of the National Academy of Sciences of the United States of America **101**(48) (2004) 16855–16860
12. Kim, J.Y., Shawe-Taylor, J.: Fast string matching using an n-gram algorithm. Software: Practice and Experience **24**(1) (1994) 79–88
13. Kim, M.S., Whang, K.Y., Lee, J.G., Lee, M.J.: n-gram/2l: A space and time efficient two-level n-gram inverted index structure. In: Proceedings of the 31st international conference on Very large data bases, VLDB Endowment (2005) 325–336
14. Ganapathiraju, M., Weisser, D., Rosenfeld, R., Carbonell, J., Reddy, R., Klein-Seetharaman, J.: Comparative n-gram analysis of whole-genome protein sequences. In: Proceedings of the second international conference on Human Language Technology Research, Morgan Kaufmann Publishers Inc. (2002) 76–81
15. Mantegna, R., Buldyrev, S., Goldberger, A., Havlin, S., Peng, C.K., Simons, M., Stanley, H.: Systematic analysis of coding and noncoding dna sequences using methods of statistical linguistics. Physical Review E **52**(3) (1995) 2939
16. Culotta, A., Kulp, D., McCallum, A.: Gene prediction with conditional random fields. University of Massachusetts, Amherst, Tech. Rep. UM-CS-2005-028 (2005)
17. Giannakopoulos, G., Karkaletsis, V., Vouros, G., Stamatopoulos, P.: Summarization system evaluation revisited: N-gram graphs. ACM Trans. Speech Lang. Process. **5**(3) (2008) 139
18. Karlin, S., Mrázek, J.: Compositional differences within and between eukaryotic genomes. Proceedings of the National Academy of Sciences of the United States of America **94**(19) (September 1997) 10227–32
19. Pruitt, K.D., Tatusova, T., Maglott, D.R.: Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research **35**(suppl 1) (2007) D61–D65
20. Dimitrieva, S., Bucher, P.: Genomic context analysis reveals dense interaction network between vertebrate ultraconserved non-coding elements. Bioinformatics **28**(18) (2012) i395–i401
21. Stephen, S., Pheasant, M., Makunin, I.V., Mattick, J.S.: Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. Mol Biol Evol **25**(2) (2008) 402–408
22. Kim, S.Y., Pritchard, J.K.: Adaptive evolution of conserved noncoding elements in mammals. PLoS genetics **3**(9) (September 2007) 1572–86
23. Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., et al.: The ucsc genome browser database. Nucleic acids research **31**(1) (2003) 51–54
24. Quinlan, A.R., Hall, I.M.: BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**(6) (2010) 841–842
25. Rice, P., Longden, I., Bleasby, A.: EMBOSS: the European Molecular Biology Open Software Suite. Trends in genetics : TIG **16**(6) (June 2000) 276–7
26. Retelska, D., Beaudoing, E., Notredame, C., Jongeneel, C.V., Bucher, P.: Vertebrate conserved non coding DNA regions have a high persistence length and a short persistence time. BMC Genomics **8** (2007) 398
27. Karlin, S., Burge, C.: Dinucleotide relative abundance extremes: a genomic signature. Trends in genetics **11**(7) (1995) 283–290
28. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD Explorations Newsletter **11**(1) (2009) 1018
29. Cohen, W.W.: Fast effective rule induction. In: ICML. Volume 95. (1995) 115–123
30. Matsunami, M., Sumiyama, K., Saitou, N.: Evolution of conserved non-coding sequences within the vertebrate Hox clusters through the two-round whole genome duplications revealed by phylogenetic footprinting analysis. J Mol Evol **71**(5-6) (2010) 427–436
31. Viturawong, T., Meissner, F., Butter, F., Mann, M.: A DNA-Centric Protein Interaction Map of Ultraconserved Elements Reveals Contribution of Transcription Factor Binding Hubs to Conservation. Cell reports **5**(2) (October 2013) 531–45
32. Karlin, S.: Global dinucleotide signatures and analysis of genomic heterogeneity. Current opinion in microbiology **1**(5) (1998) 598–610