Πτυχιακή Εργασία Υπεύθυνος Καθηγητής: Χριστόφορος Νικολάου

Ο νόμος του Menzerath στις οικογένειες γονιδίων του ανθρώπου στο επίπεδο γονιδίου – εξονίου και η σχέση του με την εξελικτική τους ιστορία

Σάββας Παραγκαμιάν (Α.Μ 1894)

Προπτυχιακός Φοιτητής, Εργαστήριο Υπολογιστικής Γονιδιωματικής, Τμήμα Βιολογίας, Πανεπιστήμιο Κρήτης, 71409 Ηράκλειο Κρήτης, Ελλάδα
s.paragamian@gmail.com, +306934008102

Παραδόθηκε στις 14.07.2015

Περίληψη

Το γονιδίωμα είναι ένα πολύπλοκο σύστημα μακριά από την ισορροπία. Σε αυτά τα συστήματα συχνά αναδύονται αυθόρμητα ιδιότητες μέσω των αλληλεπιδράσεων των μονάδων που τα αποτελούν. Τέτοια παραδείγματα αυτοοργάνωσης έχουν φανεί σε συστήματα που τα μεγέθη τους ακολουθούν κατανομή νόμου δύναμης. Ο νόμος του Zipf είναι η πιο συχνή περίπτωση νόμου δύναμης και έχει ισχύ σε πληθώρα φυσικών και ανθρωπογενών συστημάτων. Ο νόμος Menzerath – Altmann είναι μια άλλη περίπτωση νόμου δύναμης η οποία παρατηρήθηκε στις φυσικές γλώσσες του ανθρώπου και αναφέρει ότι όσο μεγαλύτερο είναι το σύνολο τόσο μικρότερα είναι τα μέρη του. Αρχικά εφαρμόστηκε στις ανθρώπινες γλώσσες στα επίπεδα λέξεων – συλλαβών και πρόσφατα εφαρμόστηκε στο γονιδίωμα στα επίπεδα γονιδίων – εξονίων. Στο γονιδίωμα του ανθρώπου φάνηκε ότι όσο μεγαλύτερο είναι ένα γονίδιο σε αριθμό εξονίων τόσο τα εξόνιά του τείνουν να μικραίνουν σε αριθμό νουκλεοτιδίων. Εμβαθύνοντας στο ίδιο πλαίσιο φάνηκε ότι η ισχύς του νόμου του Menzerath εξασθενεί όσο αυξάνεται η μεταγραφική πολυπλοκότητα, με την εμφάνιση εναλλακτικών εξονίων, στα εξωτερικά εξόνια και τέλος με την αύξηση της συντήρησης της αλληλουχίας. Εδώ εστιάσαμε στον έλεγχο υπακοής των οικογενειών γονιδίων του ανθρώπου στο νόμο του Menzerath. Παρατηρήσαμε μεγάλη ποικιλία κατανομών της σχέσης μεγέθους εξονίων με τον αριθμό των εξονίων των μεταγράφων. Εμφανίστηκαν οικογένειες που ακολουθούσαν το νόμο του Menzerath, άλλες που δεν εμφάνιζαν ισχυρή συσχέτιση αλλά και άλλες που είχαν αντίστροφη κατανομή. Αυτή η διαφοροποίηση πιθανότατα οφείλεται στη εξελικτική τους ιστορία καθώς παρατηρήσαμε ότι οι οικογένειες που εμφανίζουν μετριασμένο ισοζύγιο εξάπλωσης – κυρίως μέσω διπλασιασμού - και συντήρησης ακολουθούν το νόμο του Menzerath. Ενώ σε αυτές που η συντήρηση υπερισχύει της εξάπλωσης υπάρχει χαμηλή υπακοή. Αντίθετα στις οικογένειες που εμφανίζεται μεγάλη ποικιλότητα λειτουργιών και αυξημένη δράση του διπλασιασμού ενώ παράλληλα υπάρχει μικρή συντήρηση, το μέγεθος των εξονίων αυξάνεται με τον αριθμό των εξονίων των μεταγράφων. Τέλος φάνηκε ότι τα μετάγραφα μερικών οικογενειών ήταν ομαδοποιημένα σε ξεχωριστές ομάδες ανάλογα με τον αριθμό εξονίων τους. Αυτές οι οικογένειες αποτελούνται από πρωτεΐνες των οποίων οι υπομονάδες έχουν διαφορετικό πρόγονο οπότε η κατανομή μεταγράφων σε ομάδες οφείλεται στη διαφορετική έκφραση των υπομονάδων μέσω του εναλλακτικού ματίσματος.

Λέξεις κλειδιά: *γονιδιωματική πολυπλοκότητα, νόμοι δύναμης, νόμος Menzerath-Altmann, γονιδιακές οικογένειες, εξέλιξη*

1 Εισαγωγή

1.1 Γονιδιωματική πολυπλοκότητα

Πολύπλοκο ονομάζεται το σύστημα το οποίο δεν ακολουθεί έναν καθολικό νόμο (λόγω εξαιρέσεων), η περιγραφή του είναι μεγάλη και η αναπαραγωγή του είναι χρονοβόρα (Almirantis, Arndt, Li, & Provata, 2014). Όλα τα παραπάνω ισχύουν για το γονιδίωμα του οποίου η πολυπλοκότητα εμφανίζεται σε πολλά επίπεδα (αριθμός γονιδίων – μεταγράφων - εξονίων, επαναλαμβανόμενες αλληλουχίες, οργάνωση στο χώρο κτλ) και αυξάνεται από τους προκαρυωτικούς προς στους πολυκύτταρους ευκαρυωτικούς οργανισμούς. Οι (Michael Lynch & Conery, 2003) αναφέρουν ότι η αύξηση της πολυπλοκότητας του γονιδιώματος έχει αρνητική συσχέτιση με το δραστικό μέγεθος του πληθυσμού του οργανισμού. Αυτό, προτείνουν ότι οφείλεται στην τυχαία γενετική παρέκκλιση -η οποία δρα όσο μικραίνει ο πληθυσμός- γιατί δεν απορρίπτει αλλαγές (που βραχυπρόθεσμα στην πλειονότητά τους είναι βλαβερές) όπως γίνεται ταχύτερα μέσω της επιλογής σε είδη με μεγάλο δραστικό μέγεθος πληθυσμού. Έτσι υποστηρίζουν ότι επιτράπηκε η αύξηση σε αριθμό και σε μήκος των μη κωδικών αλληλουχιών, ο διπλασιασμός γονιδίων και άλλων λειτουργικών μονάδων και η εμφάνιση μεταθετών γενετικών στοιχείων · τα οποία ευθύνονται για τη ανάδυση νέων λειτουργιών και χαρακτηριστικών του γονιδιώματος. Χαρακτηριστικό παράδειγμα αποτελεί η εμφάνιση εξονιών και ιντρονίων στους ανώτερους ευκαρυωτικούς δίνοντας τη δυνατότητα στα γονίδια να έχουν διαφορετικές μορφές μέσω του εναλλακτικού ματίσματος. Το γονιδίωμα, όπως και πληθώρα άλλων φυσικών συστημάτων, εμφανίζει πολυπλοκότητα ως αποτέλεσμα της κατάστασής του μακριά από την ισορροπία (Prigogine & Antoniou, 2000).

1.2 Νόμοι Δύναμης σε Πολύπλοκα Συστήματα

Κατανομή νόμου – δύναμης (power law) ακολουθούν μια πληθώρα συστημάτων τόσο στη φύση όσο και σε ανθρωπογενή συστήματα. Μια ποσότητα λέγεται ότι ακολουθεί κατανομή νόμου-δύναμης όταν μεταβάλλεται σε συνάρτηση της δύναμης μιας άλλης ποσότητας. Δηλαδή όταν ισχύει:

$$Z = YX^{b} \quad (1)$$

όπου Y και b είναι σταθερές. Διαφορετικές τιμές του b αντιστοιχούν σε διαφορετικές εκφάνσεις κατανομών νόμου-δύναμης οι οποίες κάποιες φορές αντιστοιχούν σε επιμέρους νόμους. Οι αλλομετρικοί νόμοι κλίμακας (allometry scaling laws) αφορούν τη σχέση μάζας του οργανισμού (Z) με διάφορους άλλους (X) παράγοντες του οργανισμού (όπως ακτίνα αορτής, διάρκεια ζωής ατόμου, ρυθμοί κυτταρικού μεταβολισμού κ.α). Η κάθε επιμέρους σχέση παραμένει η ίδια για όλους τους οργανισμούς που μελετάται (π.χ θηλαστικά) και το δικό της συγκεκριμένο εκθέτη b που την χαρακτηρίζει. Οι ρυθμοί κυτταρικού μεταβολισμού κλιμακώνονται ως $X^{3/4}$ σε σχέση με τη μάζα του σώματος ενώ η ακτίνα της αορτής ως $X^{3/8}$, για όλα τα θηλαστικά (West, 1997). Επίσης ο νόμος του Zipf αναφέρει ότι η συχνότητα εμφάνισης (Z) μια τιμής της μεταβλητής (X) είναι αντίστροφη της δύναμης της τιμής (X^{b}), άρα όσο πιο μεγάλη είναι η τιμή τόσο μικρότερη πιθανότητα να εντοπιστεί στον πληθυσμό (Newman, 2005). Ο νόμος του Zipf ισχύει σε φαινόμενα όπως ένταση σεισμών, κατανομή πληθυσμών σε πόλεις, συχνότητα λέξεων σε κείμενα και άλλα. Είναι τόσο διαδεδομένος, που στη βιβλιογραφία συχνά ταυτίζεται με την ευρύτερη κατηγορία νόμου – δύναμης. Στη φύση ο εκθέτης του νόμου συνήθως κυμαίνεται στο διάστημα $-3<b<-2$ (Corominas-Murtra & Solé, 2010; Newman, 2005) (περισσότερα για το νόμο του Zipf στο 1.3.2). Γενικά, κατανομές νόμου – δύναμης έχουν εντοπιστεί σε αξιοσημείωτη ποικιλομορφία και ποσότητα συστημάτων γεγονός που έχει οδηγήσει στην αναζήτηση κάποιων κοινών μηχανισμών και ιδιοτήτων που ευθύνονται για την εμφάνισή τους.

Οι κατανομές νόμου-δύναμης έχουν κάποια κοινά χαρακτηριστικά. Είναι ανεξάρτητες κλίμακας, έτσι εκτείνονται σε πολλές τάξεις μεγέθους σχηματίζοντας “βαριές” (μακριές) ουρές. Η εμφάνιση “βαριών” ουρών σημαίνει ότι πολύ ακραία φαινόμενα έχουν τη δυνατότητα να συμβούν, σε αντίθεση με την κανονική κατανομή (Corominas-Murtra & Solé, 2010; Newman, 2005). Η αντοχή, των συστημάτων που ακολουθούν κατανομή νόμου δύναμης, σε διακυμάνσεις είναι εντυπωσιακή κάτι που οι (Corominas-Murtra & Solé, 2010) αποδίδουν στο γεγονός ότι είναι πολύπλοκα, μακριά από την ισορροπία συστήματα, στα οποία εμφανίζεται δυναμική ισορροπία μεταξύ θετικών και αρνητικών αναδράσεων (feedbacks). Γενικά τα περισσότερα από αυτά χαρακτηρίζονται ως ανοιχτά, πολύπλοκα, στοχαστικά συστήματα που βρίσκονται μακριά από την ισορροπία και μεταβάλλονται μη αντιστρέψιμα στο χρόνο (Barabasi & Albert, 1999;

Corominas-Murtra & Solé, 2010).

Τα αίτια εμφάνισης κατανομών νόμου – δύναμης ποικίλουν και πιθανότατα απαιτείται συνδυασμός διαφορετικών αιτίων για την ανάδυσή τους. Εδώ θα αναφέρουμε δύο από τα μοντέλα που οδηγούν σε power law, αυτά είναι η διαδικασία Yule και η αυτοοργανωμένη "κρισιμότητα" (self-organized criticality). Η διαδικασία του Yule αναφέρει ότι η πιθανότητα της εμφάνισης χ είναι ανάλογη με την πιθανότητα που έχει ήδη η μεταβλητή, ένας μηχανισμός που είναι γνωστός και ως "το φαινόμενο του Ματθαίου" ή "ο πλούσιος γίνεται πλουσιότερος" (the rich gets richer). Παραδείγματα τέτοιων μηχανισμών είναι: ο αριθμός των αναφορών σε επιστημονικά άρθρα και στις πωλήσεις βιβλίων, αλλά και η κατανομή των εισοδημάτων και ο πληθυσμός των πόλεων (Newman, 2005). Η αυτοοργανωμένη κρισιμότητα έχει χρησιμοποιηθεί για την ερμηνεία βιολογικής εξέλιξης, έντασης σεισμών, φωτιές δασών και άλλα. Αφορά στην ικανότητα του συστήματος λόγω των διαφορετικών κλιμάκων κάποιας παραμέτρου του να αυτοοργανώνεται γύρω από ένα κρίσιμο σημείο (Bak & Tang, 1987; Newman, 2005). Αυτό το σημείο αποτελεί έναν ελκυστή στον οποίο οδηγούνται οι παράμετροι του συστήματος όταν βρίσκονται μακριά από την ισορροπία. Οι (Sole & Goodwin, 2008; G Theraulaz & Bonabeau, 1995; Guy Theraulaz & Bonabeau, 1995) ορίζουν την αυτοοργάνωση ως "το σύνολο των μηχανισμών όπου οι δομές εμφανίζονται σε συνολικό επίπεδο του συστήματος μέσω αλληλεπιδράσεων μεταξύ των τμημάτων του συστήματος". Τα τμήματα του συστήματος δεν έχουν επίγνωση των συνολικών δομών που αναδύονται μέσω των αλληλεπιδράσεών τους (G Theraulaz & Bonabeau, 1995). Η αυτοοργάνωση, ως ιδιότητα των πολύπλοκων συστημάτων, πιστεύεται ότι διαδραματίζει πολύ σημαντικό ρόλο στην αυθόρμητη παραγωγή τάξης στα βιολογικά συστήματα πάνω στην οποία δρα η φυσική επιλογή (Kauffman, 1991).

1.3. Γλωσσολογικές Αναλογίες στο γονιδιωματικό "κείμενο"

1.3.1 Γλωσσολογικοί Νόμοι

Ο (Chomsky, 1957) όρισε τη γλώσσα ως "το σύνολο (πεπερασμένο ή άπειρο) των προτάσεων, που η κάθε μια έχει πεπερασμένο μέγεθος και αποτελείται από πεπερασμένο αριθμό στοιχείων". Κάθε γλώσσα έχει τη δική της γραμματική η οποία αποτελεί το μηχανισμό παραγωγής μόνο των συντακτικά ορθών φράσεων (B. Searls, 1992). Η ποσοτική μελέτη των γλωσσών οδηγεί στην βαθύτερη κατανόηση των δυναμικών της γραμματικής τους (E. G. Altmann & Gerlach, 2014). Έχουν περιγραφεί αρκετοί νόμοι που αναδεικνύουν ιδιότητες ανεξάρτητες κλίμακας όπως ο νόμος του Zipf, ο νόμος των Menzerath - Altmann και ο νόμος του Heaps. Οι ιδιότητες και οι κανόνες που αναδύονται από τη μελέτη της δυναμικής και της εξέλιξης των ανθρώπινων γλωσσών έχουν αποτελέσει πηγή έμπνευσης σε πολλά διαφορετικά επιστημονικά πεδία με τη χρήση αναλογιών (D. B. Searls, 2002). Η χρήση αναλογιών – μεταφορών είναι συνηθισμένη για τον άνθρωπο καθώς γίνεται καθημερινά τόσο στην επικοινωνία όσο και στις σκέψεις του. Όμως για τη σωστή χρήση των αναλογιών ο (Paton, 1996) αναφέρει ότι θα πρέπει να αντικατοπτρίζουν τις γενικές αρχές των εννοιών που προσδιορίζουν. Στη βιολογία, εκτός από τη χρήση φυσικής και χημείας, φαίνεται να είναι απαραίτητη η χρήση αναλογιών από τη γλωσσολογία και τη σημειολογία για την ερμηνεία φαινομένων (Ouzounis & Mazière, 2006). Τέλος, μεταφορές παρμένες από τη γλωσσολογία αποτελούν σημαντικά εργαλεία για τη μελέτη της γενωμικής πολυπλοκότητας (Ouzounis & Mazière, 2006; B. Searls, 1992).

1.3.2 Ο Νόμος του Heaps

Ο νόμος του Heaps ή αλλιώς νόμος του Herdan αναφέρει ότι το σύνολο των μοναδικών λέξεων ενός κειμένου αυξάνεται με τη δύναμη του μεγέθους του κειμένου (σε αριθμό λέξεων) (Heaps, 1978). Η γενικευμένη μορφή του νόμου αφορά τη σχέση μεταξύ χαρακτηριστικών (που αφορούν τα μοναδικά στοιχεία) και οντοτήτων που περιέχονται αυτά. Δηλαδή το σύνολο των χαρακτηριστικών (Z, όπως λέξεις) αυξάνεται όσο αυξάνονται οι οντότητές τους (X, μέγεθος κειμένου) (Egghe, 2007). Αυτό εκφράζεται με τη σχέση:

$$Z = YX^b \qquad (2)$$

όπου Y και b είναι σταθερά, με b ανήκει στο διάστημα 0<b<1. Στη γονιδιωματική ο νόμος του Heaps έχει εφαρμοστεί στο επίπεδο του γονιδιώματος των μικροβίων. Χρησιμοποιώντας την αναλογία χαρακτηριστικά – γονίδια και οντότητες – γονιδιώματα έχει δειχθεί ότι ισχύει ο νόμος του Heaps για κάποιους οργανισμούς μικροβίων (*Baccilus cereus, Haemophilus influenzae, E. Coli κ.α*). Πιο συγκεκριμένα παρουσιάστηκε ότι ο

απόλυτος αριθμός γονιδίων του οργανισμού αυξάνεται ακολουθώντας τη δύναμη του αριθμού των γονιδιωμάτων των διαφορετικών στελεχών του οργανισμού (Tettelin, Riley, Cattuto, & Medini, 2008). Αυτό σημαίνει ότι για να εντοπιστεί η ποικιλότητα του οργανισμού χρειάζεται αλληλούχιση μεγάλου αριθμού γονιδιωμάτων διαφορετικών στελεχών. Αντίστροφη ερμηνεία είναι ότι ο ρυθμός εντοπισμού νέων γονιδίων μειώνεται ως δύναμη του αριθμού γονιδιωμάτων των διαφορετικών στελεχών του οργανισμού. Αυτή η τελευταία διαπίστωση είναι ισοδύναμη με το νόμο του Zipf που θα σχολιάσουμε παρακάτω. Στην βιβλιογραφία έχει φανεί ότι οι δύο νόμοι συνδέονται μεταξύ τους. Οι (Lu, Zhang, & Zhou, 2010) αναφέρουν ότι ο νόμος του Zipf είναι βασικότερος από τον αντίστοιχο του Heaps στα συστήματα που συνυπάρχουν, κάτι που δεν έχει γίνει τελείως αποδεκτό από την επιστημονική κοινότητα.

1.3.3 Ο Νόμος του Zipf

Μια μεταβλητή λέγεται ότι ακολουθεί το νόμο του Zipf ή την κατανομή Pareto όταν η πιθανότητα εμφάνισης μιας ποσότητάς της μεταβάλλεται με αντίστροφο τρόπο από την δύναμη αυτής της ποσότητας (Newman, 2005). Οι δυο ονομασίες είναι ταυτόσημες, η μόνη διαφορά είναι ότι οι 2 επιστήμονες (από τους οποίους πήρε το όνομα η κατανομή) αναπαριστούσαν με αντίστροφο τρόπο τα δεδομένα σε διάγραμμα (ο Pareto έβαζε το f(x) στον οριζόντιο άξονα) (Newman, 2005). Ο νόμος διατυπώθηκε όταν παρατηρήθηκε ότι η συχνότητα εμφάνισης μιας λέξης στις ανθρώπινες γλώσσες είναι αντιστρόφως ανάλογη με την κατάταξή της με βάση την εμφάνισή της (Li, 1992). Στην απλή της μορφή η μαθηματική διατύπωση του νόμου είναι:

$$f(x) \propto x^{-a} \quad (3)$$

όπου ισχύει α>0 και για συγκεκριμένο διάστημα του χ (στην ουρά της κατανομής), με εξαιρέσεις (Clauset, Shalizi, & Newman, 2009). Στα φυσικά συστήματα η τιμή του κυμαίνεται στο διάστημα 2<α<3 συνήθως ενώ στις φυσικές γλώσσες είναι α=1.

Ο νόμος του Zipf είναι ευρέως διαδεδομένος σε μια τεράστια ποικιλία φαινομένων. Φαινόμενα όπως διακυμάνσεις χρηματιστηρίου, κατανομή πληθυσμού πόλεων, συχνότητα οικογενειακών ονομάτων, εντάσεις σεισμών, αριθμός ειδών σε βιολογικά τάξα, αριθμός αναφορών επιστημονικών εργασιών και πολλά άλλα υπακούν στο νόμο αυτό. Οι (Tsonis & Tsonis, 2002) διερεύνησαν το νόμο του Zipf σε 4 γονιδιώματα, στη ζύμη, στο νηματώδη *C. Elegans*, στη φρουτόμυγα *D.melanogaster*, και στον άνθρωπο. Η αναλογία που χρησιμοποίησαν είναι το γονιδίωμα ως γλώσσα και οι τομείς των πρωτεϊνών ως λέξεις. Στήριξαν την αναλογία που επέλεξαν αναφέροντας ότι οι τομείς αποτελούν ένα πολύ σημαντικό μέρος της λειτουργίας των πρωτεϊνών και λόγω της διαφορετικής δράσης του διπλασιασμού, εξαιτίας ποικίλων παραγόντων, εμφανίζονται σε διαφορετικές συχνότητες όπως και οι λέξεις. Οπότε δείχθηκε ότι η συχνότητα παρουσίας ενός τομέα πρωτεΐνης (πιο συγκεκριμένα της αλληλουχίας DNA που τον κωδικοποιεί) στο γονιδίωμα είναι αντιστρόφως ανάλογη με την κατάταξή του με βάση τη συχνότητα εμφάνισης (Tsonis & Tsonis, 2002).

Η περιγραφή γονιδιώματος, και άλλων πολύπλοκων συστημάτων με υψηλή ποικιλότητα, μπορεί να γίνει με τη χρήση πολύπλοκων δικτύων. Αυτά τα δίκτυα οι (Barabasi & Albert, 1999) τα χαρακτήρισαν ως δίκτυα ανεξάρτητα κλίμακας καθώς η πιθανότητα μιας κορυφής του δικτύου f(x) να συνδέεται με x άλλες κορυφές ακολουθεί το νόμο του Zipf με τον εκθέτη α να εξαρτάται από το είδος του δικτύου (μαθηματικός τύπος 3). Αυτή η ιδιότητα των δικτύων στηρίζεται στην ικανότητά τους να αυτοοργανώνονται. Βασιζόμενοι σε αυτή τη διαπίστωση οι (Rzhetsky & Gomez, 2001) έδειξαν ότι αυτό ισχύει για τη συχνότητα των πρωτεϊνικών υπομονάδων ανά γονιδίωμα. Οπότε φάνηκε ότι υπάρχουν υπομονάδες πρωτεϊνών (άρα και αλληλουχιών DNA) με πολύ μικρή συχνότητα εμφάνισης στο γονιδίωμα που πολλές φορές είναι μοναδικές του είδους. Τέλος, επιστρέφοντας στις βάσεις του νόμου του Zipf οι (Piantadosi, Tily, & Gibson, 2011) προτείνουν ότι χρειάζεται αναβάθμιση καθώς μέχρι τώρα υπολογίζεται το μήκος λέξεων σε σχέση με τη συχνότητα εμφάνισης. Αναφέρουν λοιπόν ότι το περιεχόμενο της πληροφορίας μιας λέξης είναι πιο αντιπροσωπευτικό μέτρο πρόβλεψης του μεγέθους της λέξης από τον υπολογισμό της συχνότητάς της. Άρα ισχυρίζονται ότι οι λέξεις που είναι πιο προβλέψιμες (περιέχουν μικρότερη πληροφορία) και όχι οι πιο συχνές τείνουν να είναι μικρότερες.

1.3.4 Ο Νόμος των Menzerath Altmann

Στα τέλη του 19ου αιώνα ο γλωσσολόγος Antoine Gregoire μελετώντας τις φωνητικές ιδιότητες της γαλλικής γλώσσας παρατήρησε ότι το φωνήεν “a” προφέρεται πιο κοφτά στις λέξεις με περισσότερες

συλλαβές (Cramer, 2005). Έπειτα, το 1928, ο Menzerath δημοσίευσε ότι όσο μεγαλύτερη είναι μια λέξη σε αριθμό συλλαβών τόσο μικρότερες τείνουν να είναι οι συλλαβές που την αποτελούν σε αριθμό γραμμάτων. Αυτή η διαπίστωση επαληθεύτηκε για τη γερμανική γλώσσα (Menzerath, 1954) και αργότερα για τις περισσότερες φυσικές γλώσσες. Ο νόμος του Menzerath έχει βρεθεί ότι ισχύει επίσης σε μουσικά κείμενα και στο γονιδίωμα (όπως θα δούμε παρακάτω).

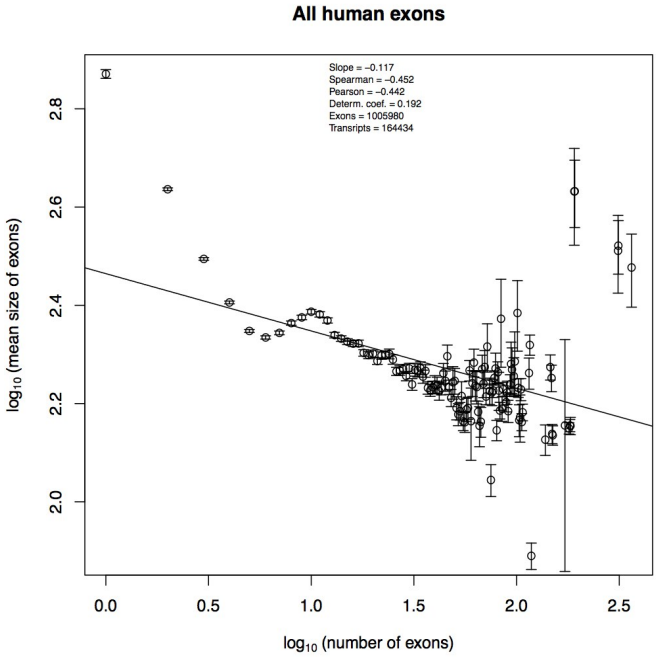
Στη γενικευμένη του μορφή διατυπώθηκε ως εξής: “όσο μεγαλύτερο το σύνολο τόσο μικρότερα τα μέρη του” (G. Altmann, 1980). Ο νόμος του Menzerath ακολουθεί κατανομή νόμου – δύναμης και αφορά 2 μεταβλητές που εκτείνονται σε 3 επίπεδα και εκφράστηκε μαθηματικά από τον (G. Altmann, 1980) γι' αυτό στη βιβλιογραφία αναφέρεται και ως νόμος Menzerath – Altmann. Ο γενικός τύπος του νόμου είναι :

$$Z = \alpha X^{b} e^{cX} \quad (4)$$

όπου b<-1, c και α είναι σταθερά. Σύμφωνα με τον τύπο ισχύει ότι το σύνολο (λέξη) αποτελείται από X μέρη (αριθμός συλλαβών) και κάθε μέρος αποτελείται από Z αριθμό τμημάτων (γράμματα), οπότε ο νόμος αναφέρει ότι όσο περισσότερα μέρη (μεγάλο X – λέξη με πολλές συλλαβές) αποτελείται το σύνολο (λέξη) τόσο λιγότερα τα τμήματα που αποτελείται το Z (μικρό Z - λίγα γράμματα κάθε συλλαβή) (Li, 2012).

Στη γονιδιωματική η πρώτη εφαρμογή του νόμου του Menzerath αφορούσε την αναλογία: γονιδίωμα ως λέξεις, χρωμοσώματα ως συλλαβές και αριθμός νουκλεοτιδίων για το μέγεθος “συλλαβών”. Αυτό έγινε για είδη διαφορετικών τάξων (έντομα, θηλαστικά, μύκητες κτλ) (Ferrer-i-cancho & Forns, 2010). Αργότερα ο (Solé, 2010) σχολίασε ότι στην προαναφερθείσα εργασία η αναλογία που χρησιμοποιήθηκε δεν έχει βιολογική αξία και ότι η σχέση χρωμοσώματος – αριθμού βάσεων είναι εξ' ορισμού αντίστροφη καθώς το σταθμισμένο γινόμενό τους είναι το μέγεθος του γονιδιώματος (σταθερά). Σε απάντησή τους στην κριτική, οι (Baixeries, Hernández-fernández, & Ferrer-i-cancho, 2012; Baixeries, Hernández-fernández, Forns, & Ferrer-i-cancho, 2013; Ferrer-i-cancho, Hernández-fernández, & Baixeries, 2014) αναφέρουν ότι η ισχύς του γενικού τύπου του νόμου του Menzerath χρειάζεται διεύρυνση υποστηρίζοντας ότι η σχέση είναι ο νόμος του Menzerath με b=-1, c=0. Αναγνώρισαν ότι η κριτική για την εμφάνιση συσχέτισης του αριθμού χρωμοσωμάτων (X) με το μέγεθος του γονιδιώματος (Y) είναι σωστή αλλά τη χαρακτήρισαν συντηρητική. Τέλος, επιχειρηματολογούν ότι αυτή η ειδική περίπτωση του νόμου δεν αποτελεί αναπόφευκτη ιδιότητα του συστήματος καθώς οι μεταβλητές Y (μέγεθος χρωμοσώματος) και X (αριθμός χρωμοσωμάτων οργανισμού) στην πλειονότητα των περιπτώσεων δεν είναι ανεξάρτητες και επομένως δεν εμφανίζεται η αμφιλεγόμενη συσχέτισή τους. Η εξάρτηση εμφανίζεται λόγω των περιορισμών της δομής των χρωμοσωμάτων γιατί έχουν κεντρομερές και τελομερή (Almirantis et al., 2014).

Στον χώρο της πρωτεομικής έγινε έρευνα για την υπακοή στο νόμο του Menzerath για τη σχέση πλήθους πρωτεϊνών ανάλογα με το μήκος τους σε αμινοξέα για 10 διαφορετικούς οργανισμούς (Eroglu, 2013). Ο συγγραφέας χρησιμοποίησε την αναλογία λέξη – σύνολο πρωτεϊνών, συλλαβή – πρωτεΐνη και γράμμα – αμινοξύ και υποστηρίζει, για κάθε οργανισμό, ότι όσο αυξάνεται το μέγεθος των πρωτεϊνών σε μήκος τόσο λιγότερο είναι το πλήθος τους. Ο (Li, 2012)σε εργασία του για το ανθρώπινο γονιδίωμα χρησιμοποίησε τις αναλογίες γονίδιο – λέξη, εξόνιο – συλλαβή, νουκλεοτίδια – γράμματα. Ο Li έδειξε ότι υπάρχει αρνητική συσχέτιση μεταξύ του αριθμού των εξονίων των γονιδίων με το μέσο μήκος των εξονίων τους, δηλαδή όσο μεγαλύτερο είναι το γονίδιο σε αριθμό εξονίων τόσο τα εξόνιά του μικραίνουν σε αριθμό νουκλεοτιδίων. Αυτή η σχέση είναι γραμμική σε διπλή λογαριθμική κλίμακα γεγονός ανάλογο με το νόμο του Menzerath. Επίσης έδειξε ότι η σχέση αυτή είναι ανεξάρτητη από το μέγεθος των γονιδίων αποκλίνοντας την δυνατότητα αμφισβήτησης όπως έγινε στην περίπτωση των (Ferrer-i-cancho & Forns, 2010).



Εικόνα 1: Η σχέση σε διπλή λογαριθμική κλιμακα του μέσου μεγέθους εξονίων με τον αριθμό των εξονίων για όλα τα εξόνια του ανθρώπου. Κάθε σημείο αντιπροσωπεύει το σύνολο των μεταγράφων που έχουν τον ίδιο αριθμό εξονίων. Τα error bars παρουσιάζουν το τυπικό σφάλμα της μέσης τιμής.

Εν συνεχεία της δουλειάς του (Li, 2012)ο (Nikolaou, 2014) έδειξε την υπακοή του νόμου του Menzerath και στο γονιδίωμα του ποντικού (*Mus musculus*). Όμως διερευνώντας βαθύτερα δείχθηκε ότι όσο αυξάνουν τα εναλλακτικά μετάγραφα των γονιδίων τόσο τείνουν να μην υπακούν στο νόμο κάτι που ερμηνεύτηκε ως αποτέλεσμα της αύξησης των περιορισμών της πολύπλοκης ρύθμισης ματίσματος. Παρατηρήθηκε επίσης, ότι τα εξόνια που είναι κοινά (συστατικά) σε όλα τα μετάγραφα ακολουθούν με μεγαλύτερη πιστότητα το νόμο του Menzerath σε σχέση με αυτά που είναι εναλλακτικά. Άλλος ένας παράγοντας που εξετάστηκε είναι αν διαφέρει η υπακοή στο νόμο του Menzerath στα εσωτερικά σε σχέση με τα εξωτερικά (πρώτο-τελευταίο) εξόνια. Σε αυτήν την περίπτωση δείχθηκε μεγαλύτερη υπακοή στα γονίδια που βρίσκονται εσωτερικά όμως με μικρές διαφορές.

Το πλαίσιο πάνω στο οποίο συνδέθηκαν όλα τα παραπάνω ήταν η συντήρηση αλληλουχίας. Μετά από αναλύσεις δείχθηκε ότι τα γονίδια με υψηλή συντήρηση αλληλουχίας δεν ακολουθούν το νόμο του Menzerath. Επιπλέον εφαρμόζοντας τις προηγούμενες συσχετίσεις (εσωτερικά-εξωτερικά εξόνια, μεταγραφική πολυπλοκότητα, εναλλακτικά – κοινά εξόνια μεταγράφων) φάνηκε ότι υπάρχει περισσότερη συντήρηση αλληλουχίας

1. όσο αυξάνεται η μεταγραφική πολυπλοκότητα
2. στα εναλλακτικά εξόνια μεταγράφων
3. στα εξωτερικά εξόνια

στα δεδομένα δηλαδή που δεν εμφανίζουν ή εμφανίζουν μικρότερη υπακοή στο νόμο του Menzerath. Κλείνοντας ο (Nikolaou, 2014) προτείνει ένα εξελικτικό μοντέλο που αναφέρει ότι όσο αυξάνονται οι περιορισμοί (εναλλακτικά μετάγραφα, συντήρηση αλληλουχίας) τόσο μειώνεται η υπακοή στο νόμο ενώ αντίθετα οι νέες δομές (χωρίς τέτοιους περιορισμούς) είναι δεκτικές σε αλλαγές και αυξάνεται το μέγεθός τους.

Η διερεύνηση υπακοής στον νόμο του Menzerath σε διαφορετικές ομάδες γονιδίων (gene sets) μπορεί να βοηθήσει στην καλύτερη κατανόηση της δυναμικής των γονιδίων και του γονιδιώματος. Τα γονίδια ή/και τα προϊόντα τους μπορούν να κατηγοριοποιηθούν σε πολλές διαφορετικές ομάδες αναλόγως με τα χρησιμοποιούμενα κριτήρια. Οι (Li, Freudenberg, & Oswald, 2015) αναφέρουν 5 γενικές κατηγορίες ομάδων με βάση : την ομολογία (όπως οι οικογένειες γονιδίων), τη φυσική θέση (χρωμόσωμα και συντεταγμένες για τα γονίδια και κυτταρική τοπολογία για τις πρωτεΐνες), τις αλληλεπιδράσεις (κυρίως δίκτυα αλληλεπιδράσεων πρωτεϊνών), το φαινότυπο (γονίδια που συσχετίζονται με ασθένειες) και τέλος τη

βιολογική διεργασία. Η γενικότητα της βιολογικής διεργασίας ως έννοια έχει οδηγήσει στη δημιουργία πολλών διαφορετικών επιμέρους κριτηρίων όπως πρωτεϊνική λειτουργία και βιολογικά μονοπάτια. Πρώτη εφαρμογή υπακοής στο νόμο Menzerath για ομάδες γονιδίων έγινε για 2 ομάδες με βάση το βιοχημικό μονοπάτι, και δείχθηκε ότι τα γονίδια των μεταγραφικών παραγόντων ακολουθούν το νόμο σε αντίθεση με τα γονίδια του μεταβολισμού (Nikolaou, 2014). Και πάλι φαίνεται ότι η ομάδα που έχει πολλούς περιορισμούς (γονίδια μεταβολισμού) δεν ακολουθεί το νόμο του Menzerath.

2 Στόχος

Στην παρούσα εργασία στηριζόμαστε σε 2 μεταφορές, την έννοια της οικογένειας και την έννοια της λέξης. Οικογένεια γονιδίων είναι το σύνολο των ομόλογων γονιδίων που έχουν σημαντική ομοιότητα στην αλληλουχία και πιθανότατα έχουν παρόμοιες λειτουργίες (Demuth, De Bie, Stajich, Cristianini, & Hahn, 2006; Huynen & van Nimwegen, 1998). Η ομολογία των γονιδίων φαίνεται ότι προκύπτει από τον διπλασιασμό ενός αρχικού γονιδίου (M Lynch & Conery, 2000). Η 2η μεταφορά αφορά τα γονίδια ως λέξεις, τη βασική μονάδα μετάδοσης πληροφορίας της γλώσσας. Τα γονίδια αποτελούνται από εξόνια όπως οι λέξεις αποτελούνται από συλλαβές και κάθε εξόνιο αποτελείται από νουκλεοτίδια όπως οι συλλαβές από γράμματα του αλφάβητου (Li, 2012; P Copland, 2005).

Στόχος της παρούσας εργασίας είναι να εξεταστεί η σχέση μέσου μεγέθους εξονίων με το πλήθος των εξονίων των μεταγράφων για κάθε οικογένεια γονιδίων και να ερμηνευθεί η πιθανή υπακοή (ή όχι) στο νόμο του Menzerath. Πιο συγκεκριμένα θα εφαρμοστεί η αναλογία μετάγραφο – λέξη, εξόνιο – συλλαβή και νουκλεοτίδιο – γράμμα για τις οικογένειες γονιδίων του ανθρώπου (*Homo sapiens*) ώστε να βρεθούν εκείνες που έχουν κατανομή παρόμοια με το νόμο του Menzerath. Έγινε έλεγχος αν η κλίση της κατανομής είναι αρνητική σε διπλή λογαριθμική κλίμακα και επίσης αν εμφανίζονται βαριές ουρές.
Τέλος, θα προσπαθήσουμε να αντλήσουμε χρήσιμα συμπεράσματα για τον τρόπο ή τους τρόπους με τους οποίους τυχόν εξελίσσονται διαφορετικές οικογένειες πρωτεϊνών με βάση τα πρότυπα υπακοής (ή μη υπακοής) στον νόμο του Menzerath.

3 Δεδομένα και Μέθοδοι

3.1 Πηγή δεδομένων

Από το σύστημα διερεύνησης γονιδιώματος Ensembl (http://www.ensembl.org) αντλήσαμε από την βάση δεδομένων Ensembl Genes 79 δεδομένα του annotation GRCh38.p2 για το ανθρώπινο γονιδίωμα (*Homo sapiens*) (Cunningham et al., 2014). Η άντληση των δεδομένων έγινε με το εργαλείο αναζήτησης δεδομένων BioMart (http://www.ensembl.org/biomart/martview/) τον Απρίλιο του 2015. Κατεβάσαμε 2 tsv αρχεία κειμένου, το πρώτο περιείχε: αριθμό χρωμοσώματος, αρχή εξονίου (βάσεις), τέλος εξονίου (βάσεις), ταυτότητα εξονίου, ταυτότητα μεταγράφου, ταυτότητα γονιδίου. Το 2ο αρχείο περιείχε τις ταυτότητες γονιδίων, μεταγράφων, εξονίων και οικογενείων και τις περιγραφές οικογενειών. Έπειτα ενώσαμε τα 2 αρχεία ώστε για κάθε εξόνιο να έχουμε τις συντεταγμένες του και όλα τα αποδιδόμενα σε αυτά λειτουργικά και γονιδιωματικά χαρακτηριστικά (εξόνιο, μετάγραφο, γονίδιο, οικογένεια) (βλέπε 3.3).

3.2 Περιγραφή δεδομένων

Στην έκδοση GRCh38.p2 της Ensembl περιέχονται 1005981 μοναδικά ζεύγη εξονίου – μεταγράφου. Λόγω εναλλακτικού ματίσματος 1 εξόνιο μπορεί να ανήκει σε διαφορετικά μετάγραφα, κάθε γονίδιο έχει πολλά διαφορετικά μετάγραφα και επίσης 1 μετάγραφο μπορεί να ανήκει σε πολλές διαφορετικές οικογένειες. Καταλαβαίνουμε λοιπόν ότι η βασική μονάδα πληροφορίας δεν είναι το εξόνιο ή το μετάγραφο αλλά το ζεύγος τους. Και επιπλέον επειδή κάθε μετάγραφο οδηγεί σε διαφορετικό πρωτεϊνικό προϊόν η εφαρμογή του νόμου του Menzerath έγινε για μετάγραφα – εξόνια (Nikolaou, 2014) αντί για γονίδια – εξόνια. Από το σύνολο των ζευγών μεταγράφων - εξονίων το 71% έχει καταταχθεί σε οικογένειες γονιδίων οι οποίες απαριθμούνται σε 24161 (Πίνακας 1) στον άνθρωπο. Άρα κάθε γραμμή του αρχείου περιέχει πλέον : αριθμό χρωμοσώματος, συντεταγμένες εξονίου, ταυτότητα εξονίου, ταυτότητα μεταγράφου, ταυτότητα γονιδίου και ταυτότητα οικογένειας για το ζεύγος εξονίου – μεταγράφου. Επίσης κάθε οικογένεια συνοδεύεται και από την περιγραφή της.

Κρατώντας μόνο τα εξόνια που ανήκουν σε μία τουλάχιστον οικογένεια παρατηρούμε ότι το μεγαλύτερο εξόνιο είναι 33287 βάσεις και τα μικρότερα έχουν λιγότερες από 10 βάσεις, δηλαδή το μέγεθος των εξονίων εκτείνεται σε 4 τάξεις μεγέθους. Παρατηρήσαμε ότι υπάρχουν 14 εξόνια οικογενειών (15 σε όλο το γονιδίωμα) με 0 βάσεις (συντεταγμένες αρχής και τέλους ήταν ίσες) γεγονός που προκύπτει από προβλήματα του annotation. Τα μηδενικά εξόνια δεν βρέθηκαν στις οικογένειες που επιλέξαμε να αναλύσουμε σε αυτήν την εργασία οπότε δεν μας απασχόλησαν. Αντίστοιχα για τα μετάγραφα, το μεγαλύτερο έχει 363 εξόνια ενώ υπάρχουν 470 μετάγραφα με 1 εξόνιο. Οπότε το μέγεθος των μεταγράφων σε αριθμό εξονίων κυμαίνεται σε 3 τάξεις μεγέθους (Πίνακας 2).

Οι όροι οικογένεια γονιδίου και οικογένεια πρωτεΐνης χρησιμοποιούνται ως συνώνυμοι στη βιβλιογραφία. Οικογένειες γονιδίων ονομάζονται οι ομάδες γονιδίων που σχηματίστηκαν με βάση την ομοιότητα της αλληλουχίας τους η οποία οφείλεται στην κοινή καταγωγή τους. Στην πλειονότητα των περιπτώσεων τα μέλη μιας οικογένειας μοιράζονται παρόμοια τρισδιάστατη δομή και συμμετέχουν στα ίδια βιοχημικά μονοπάτια (Chothia, 1992). Οπότε η κατηγοριοποίηση ενός γονιδίου, που είναι γνωστή μόνο η αλληλουχία του, σε οικογένεια μπορεί να προσφέρει και άλλες χρήσιμες πληροφορίες (Enright, Dongen, & Ouzounis, 2002). Το Ensembl Project χρησιμοποίησε των αλγόριθμο TRIBE-MCL για την κατηγοριοποίηση των γονιδίων σε οικογένειες. Ο TRIBE-MCL αναπτύχθηκε από τους (Enright et al., 2002), βασίζεται στον αλγόριθμο Markov clustering και αποτελεί την πιο εκλεπτυσμένη κατηγοριοποίηση γονιδίων σε οικογένειες. Παλαιότερα γινόταν χρήση BLAST και άλλων αλγορίθμων για τη σύγκριση και κατάταξη σε ομάδες ανάλογα με την ομοιότητα αλληλουχίας. Αυτές οι μέθοδοι όμως, αναφέρουν οι (Enright et al., 2002), οδηγούν σε εσφαλμένη κατηγοριοποίηση πολλών πρωτεϊνών γιατί δεν λαμβάνουν υπόψιν τους ότι οι πρωτεΐνες αποτελούνται από λειτουργικές υπομονάδες (domains). Έτσι υπάρχει πρόβλημα για τους εξής λόγους:

1. η εμφάνιση κοινής υπομονάδας (άρα κοινή αλληλουχία) δεν σημαίνει απαραίτητα ότι οι πρωτεΐνες συμμετέχουν σε κοινό βιολογικό μονοπάτι
2. επίσης υπάρχουν πολλές μικρές υπομονάδες που είναι ευρέως διαδεδομένες σε πρωτεΐνες χωρίς όμως αυτό να σημαίνει ότι οι πρωτεΐνες τους προέρχονται από κοινό πρόγονο

Τέλος, οι περισσότεροι αλγόριθμοι, μέχρι τη δημιουργία του TRIBE-MCL, είχαν μικρή απόδοση και δεν μπορούσαν να εφαρμοστούν σε ολόκληρα γονιδιώματα.

Το μέγεθος κάθε οικογένειας καθορίζεται από την δυναμική ισορροπία του διπλασιασμού γονιδίων και της εξαφάνισης γονιδίων (Demuth et al., 2006). Οι (Kunin, Cases, Enright, Lorenzo, & Ouzounis, 2003) αναφέρουν ότι έχουν περιγραφεί περισσότερες από 56000 (στον άνθρωπο 24161) οικογένειες σε σύνολο 83 είδη οργανισμών. Μάλιστα αναφέρουν ότι το γεγονός πως όσο αυξάνονται οι πρωτεΐνες που περιγράφονται τόσο αυξάνεται ο αριθμός των οικογενειών αποτελεί δείκτη ότι υπάρχουν πολλές πρωτεΐνες και ιδιότητές τους που δεν έχουν περιγραφεί ακόμα.

Πίνακας 1 *Περιεχόμενα δεδομένων του ανθρώπινου γονιδιώματος (Ensembl Genes 79, GRCh38.p2)*

Τύπος	**Σύνολο στοιχείων**
Μοναδικά εξόνια	574884
Μετάγραφα	164434
Γονίδια	45414
Οικογένειες	24161
Ζεύγη εξονίου - μεταγράφου	1005980
Εξόνια που ανήκουν σε οικογένειες	713919 (71%)
Αταξινόμητα εξόνια	292061 (29%)

3.3 Επεξεργασία δεδομένων – Υπολογισμοί

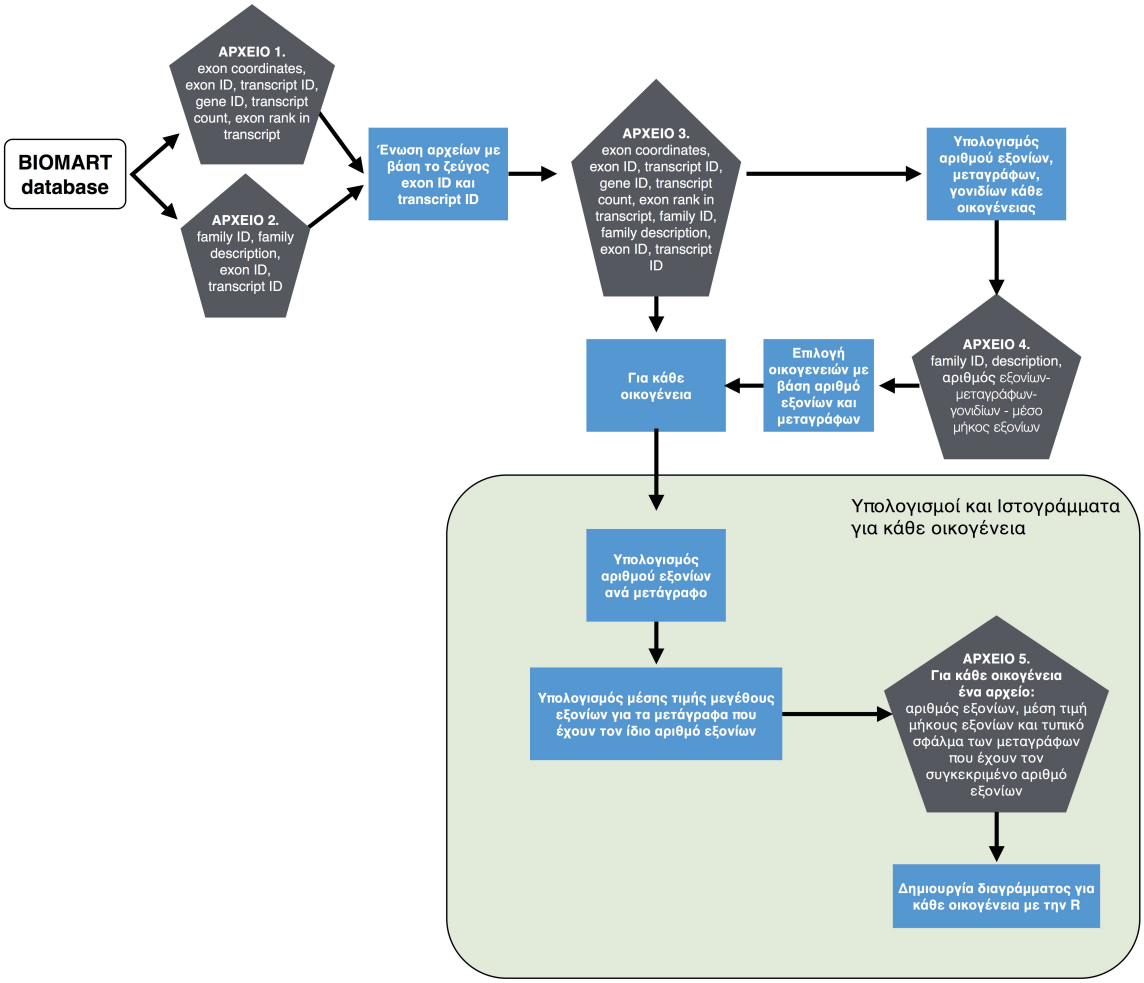
Η επεξεργασία των δεδομένων έγινε με τη γλώσσα προγραμματισμού Perl και τα διαγράμματα έγιναν με το στατιστικό περιβάλλον της R ενώ επίσης χρησιμοποιήθηκε και η γλώσσα shell για μερικούς χειρισμούς των αρχείων κειμένου. Δημιουργήθηκαν 3 διαφορετικοί κώδικες (Perl scripts) που αφορούσαν α) την ένωση των 2 αρχείων από την Ensembl, β) τους υπολογισμούς εξονίων, μεταγράφων και γονιδίων για κάθε οικογένεια και γ) τους υπολογισμούς -για κάθε οικογένεια- της μέσης τιμής του μεγέθους των εξονίων των μεταγράφων τα οποία έχουν ίδιο αριθμό εξονίων (Εικόνα 2). Στον τελευταίο κώδικα της Perl ενσωματώσαμε τον κώδικα της R για τη δημιουργία ιστογραμμάτων. Αυτό έγινε μετά από εγκατάσταση του πακέτου Statistics::R από το CPAN (http://search.cpan.org/~fangly/Statistics-R/lib/Statistics/R.pm). Οι κώδικες παρατίθενται στο παράρτημα.

α) Όπως αναφέραμε παραπάνω από την Ensembl πήραμε 2 αρχεία. Ο πρώτος στόχος ήταν να ενωθούν αυτά τα αρχεία ώστε για κάθε εξόνιο να έχουμε συγκεντρωμένες όλες τις πληροφορίες που χρειαζόμαστε. Η συγχώνευση αρχείων έγινε με βάση μια κοινή στήλη στα 2 tsv αρχεία, την ταυτότητα εξονίων. Ο πιο ασφαλής τρόπος είναι να γίνει συγχώνευση με βάση το ζεύγος εξόνιο και μετάγραφο καθώς αυτά ορίζουν τη μοναδικότητα των γραμμών των αρχείων. Χρειάζεται προσοχή γιατί μπορεί μερικές γραμμές να είναι πολλαπλές φορές στο αρχείο κάτι που πιθανότατα συμβαίνει επειδή τα 2 αρχεία έχουν άνισο αριθμό γραμμών ή γιατί έτσι ήταν τα δεδομένα από την Ensembl. Οπότε είναι καλό να ελεγχθεί η μοναδικότητα των γραμμών, κάτι που γίνεται εύκολα μέσω του shell.

β) Για λόγους διερεύνησης των οικογενειών φτιάξαμε ένα νέο αρχείο που περιέχει τα δεδομένα κάθε οικογένειας γονιδίου. Για κάθε οικογένεια υπολογίσαμε τον αριθμό των εξονίων, των μεταγράφων και των γονιδίων καθώς και τη μέση τιμή μεγέθους των εξονίων. Έτσι μπορέσαμε να αντλήσουμε στατιστικά στοιχεία για τις οικογένειες αλλά το σημαντικότερο ήταν ότι με βάση αυτό το αρχείο μπορούμε να εφαρμόσουμε σε συγκεκριμένες οικογένειες με τα δικά μας κριτήρια το νόμο του Menzerath.

γ) Σε αυτό το Perl script γίνονται οι υπολογισμοί, το data binning και τα ιστογράμματα για κάθε οικογένεια

γονιδίων. Εξετάζουμε την υπακοή στο νόμο του Menzerath στο επίπεδο μεταγράφων – εξονίων. Το μέγεθος των μεταγράφων μετριέται σε πλήθος εξονίων και το μέγεθος των εξονίων σε αριθμό νουκλεοτιδίων. Άρα για κάθε οικογένεια κάνουμε ένα ιστόγραμμα όπου ο y άξονας αντιστοιχεί στο μέσο μήκος εξονίων και ο x άξονα αντιστοιχεί στον αριθμό εξονίων.



Εικόνα 2: Σε αυτό το διάγραμμα ροής φαίνεται η αλληλουχία σκέψων, τα αρχεία που χρησιμοποιήθηκαν και οι αναλύσεις που έγιναν στα δεδομένα για την δημιουργία ιστογραμμάτων των οικογενειών γονιδίων. Οι κώδικες που δημιουργήθηκαν παρατίθονται στο Παράρτημα 6.2.

Αρχικά κάναμε υπολογισμούς και διάγραμμα για όλα τα εξόνια του ανθρώπου. Έπειτα μέσω του αρχείου των δεδομένων των οικογενειών επιλέγουμε σε ποιες οικογένειες θέλουμε να κάνουμε ιστόγραμμα με βάση αριθμό εξονίων, μεταγράφων ή/και γονιδίων. Τα κριτήρια που θέσαμε για την επιλογή των οικογενειών είναι να έχουν περισσότερα από 100 εξόνια και 9 μετάγραφα. Έτσι από το σύνολο των 24161 οικογενειών κάναμε υπολογισμούς και διαγράμματα για 1261. Έπειτα για κάθε οικογένεια που πληροί τις προϋποθέσεις κάνουμε τα εξής:

1. Data binning. Γίνεται για το σύνολο των μεταγράφων(k) που έχουν τον ίδιο αριθμό εξονίων(n). Το κάνουμε για να μειώσουμε τα μικρά σφάλματα που προκύπτουν ιδιαίτερα στην ουρά των power law κατανομών (Newman, 2005).
   Για τα δεδομένα κάθε οικογένειας, υπολογίζουμε από το συγχωνευμένο αρχείο τον αριθμό των εξονίων κάθε μεταγράφου. Έπειτα συλλέγουμε όλα τα μετάγραφα(k) με τον ίδιο αριθμό εξονίων(n) για να κάνουμε υπολογισμούς τα οποία αποτελούν 1 bin. Πλέον κάθε σημείο στο διάγραμμα αντιστοιχεί στο σύνολο των μεταγράφων με τον ίδιο αριθμό εξονίων της οικογένειας.
2. Υπολογισμός μέσης τιμής μεγέθους εξονίων. Για κάθε bin, σύνολο μεταγράφων με ίδιο αριθμό εξονίων, αθροίζεται το μήκος όλων των εξονίων και διαιρείται προς το πλήθος τους. Κάθε bin έχει k

μετάγραφα και κάθε μετάγραφο του ίδιου bin έχει n εξόνια άρα το πλήθος των εξονίων είναι n*k.

$$mean\ exon\ size = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n}\left(exon_{end} - exon_{start}\right)}{kn} \quad (5)$$

3. Υπολογισμός τυπικής απόκλισης και τυπικού σφάλματος. Για κάθε bin για να υπολογίσουμε το τυπικό σφάλμα πρέπει να υπολογίσουμε πρώτα την τυπική απόκλιση. Άρα υπολογίζουμε την τυπική απόκλιση για όλα τα εξόνια που περιέχονται σε κάθε bin.

$$stddev = \sqrt{\frac{\sum_{j=1}^{k}\sum_{i=1}^{n}\left(exon_{size} - mean\ exon\ size\right)^2}{kn - 1}} \quad (6)$$

Όμοια με την μέση τιμή, το σύνολο των εξονίων για ένα bin είναι το σύνολο των μεταγράφων(k) του bin επί τον αριθμό εξονίων(n) που έχουν. Όπως φαίνεται στον Πίνακα 2 υπάρχουν 470 οικογένειες με 1 εξόνιο, για αυτές ο υπολογισμός της τυπικής απόκλισης πρέπει να έχει παρονομαστή το kn=1 γιατί αλλιώς μηδενίζεται. Κάτι τέτοιο δεν μας επηρεάζει γιατί χρησιμοποιούμε οικογένειες με περισσότερα από 100 εξόνια.
Ακολούθως υπολογίζουμε το τυπικό σφάλμα σύμφωνα με τον τύπο:

$$stderror = \frac{stddev}{\sqrt{kn}} \quad (7)$$

Το τυπικό σφάλμα το χρειαζόμαστε για να δούμε τη διασπορά της κατανομής της μέσης τιμής. Στα ιστογράμματα χρησιμοποιήθηκε στα error bars για να φανεί το 68% της κατανομής των μεγεθών των εξονίων των μεταγράφων που έχουν ίδιο αριθμό εξονίων.

4. Μετά τους υπολογισμούς για κάθε οικογένεια υπάρχουν τα δεδομένα: bin με n εξόνια, μέση τιμή μεγέθους εξονίων, τυπικό σφάλμα. Άρα υπάρχουν οι μεταβλητές για να φτιαχτεί το διάγραμμα. Μέσο του Statistics::R πακέτου ο κώδικας της R ενσωματώνεται στην Perl. Έτσι φτιάχνουμε 1 διάγραμμα για κάθε οικογένεια που στον y άξονα αντιστοιχεί η μέση τιμή των εξονίων και στον x ο αριθμός των εξονίων. Οι τιμές μέσης τιμής και αριθμού εξονίων λογαριθμίζονται με δεκαδικό λογάριθμο οπότε το ιστόγραμμα είναι σε διπλή λογαριθμική κλίμακα. Επίσης φτιάχτηκε ευθεία ελαχίστων τετραγώνων και υπολογίστηκε ο συντελεστής συσχέτισης Pearson και ο συντελεστής συσχέτισης με βάση την κατάταξη Spearman.

3.4 Αναγνώριση υπακοής στο νόμο του Menzerath

Το ιστόγραμμα κάθε οικογένειας εξετάστηκε αν ακολουθεί το νόμο του Menzerath με εμπειρικό τρόπο. Με τη χρήση της διπλής λογαριθμικής κλίμακας φέρνουμε κοντά δεδομένα που βρίσκονται σε διαφορετική τάξη μεγέθους. Επίσης οι (Stumpf & Porter, 2012) αναφέρουν ότι “ένδειξη ότι τα δεδομένα ακολουθούν κατανομή νόμου δύναμης είναι όταν μεταβλητές x, y εμφανίζουν σχεδόν γραμμική σχέση σε διπλή λογαριθμική κλίμακα για 2 τάξεις μεγέθους το λιγότερο”. Παρατηρώντας αν η κλίση της ευθείας μιας γραμμικής παλινδρόμησης είναι αρνητική και η συσχέτιση του Pearson είναι επίσης αρνητική σημαίνει ότι υπάρχει η τάση προς τη συμφωνία με το νόμο του Menzerath. Χρησιμοποιούμε το συντελεστή συσχέτισης του Pearson γιατί αποτελεί μέτρο της ισχύος της γραμμικής σχέσης μεταξύ των 2 μεταβλητών (μέσο μέγεθος εξονίων, αριθμός εξονίων) (Hauke & Kossowski, 2011), και το συντελεστή προσδιορισμού (coefficient of determination, R^2) γιατί αποτελεί μέτρο της προσαρμογής της ευθείας μιας γραμμικής παλινδρόμησης στα δεδομένα (Γναρδέλλης, 2003).

Στην παρούσα εργασία δεν εφαρμόστηκε ακριβής στατιστική ανάλυση των δεδομένων ώστε να αποκλειστούν άλλες πιθανές κατανομές αντί για νόμου δύναμης (π.χ Log Normal, εκθετική κτλ) ((Clauset et al., 2009; Virkar & Clauset, 2014). Άρα δεν μπορεί να επιβεβαιωθει πλήρως ότι μια οικογένεια ακολουθεί το νόμο του Menzerath και να προσδιορίσει τον εκθέτη του νόμου για την κάθε οικογένεια. Όμως αυτό που μπορεί να φανεί είναι ποιες οικογένειες εμφανίζουν “βαριές” ουρές και αν υπάρχει συσχέτιση μεταξύ μέσου μεγέθους εξονίων και αριθμού εξονίων των μεταγράφων.

4 Αποτελέσματα και συζήτηση

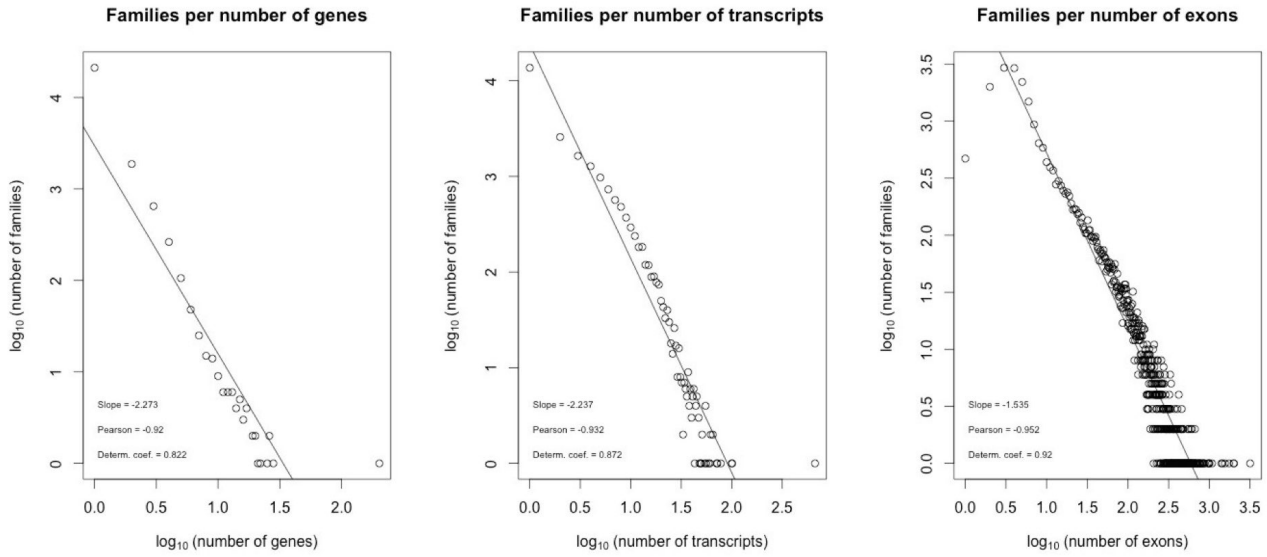
4.1 Επιλογή οικογενειών για ανάλυση

Από τις 24161 οικογένειες γονιδίων του ανθρώπου το 56% έχουν μόνο 1 μετάγραφο ενώ μόλις 0.01% έχει περισσότερα από 100 μετάγραφα (Πίνακας 2). Στα ιστογράμματα (Εικόνα 3) φαίνεται η αρνητική γραμμική πορεία σε διπλή λογαριθμική κλίμακα της σχέσης αριθμού οικογενειών με τον αριθμό εξονίων, μεταγράφων και γονιδίων αντίστοιχα. Η κατανομή που ακολουθούν τα δεδομένα μοιάζουν με power law κάτι που έρχεται σε συμφωνία με την εργασία των (Huynen & van Nimwegen, 1998) για 9 μονοκύτταρους οργανισμούς και 2 ιούς.

Όπως αναφέραμε και στην ενότητα 3.3 επιλέξαμε τις οικογένειες που έχουν περισσότερα από 100 εξόνια και 9 μετάγραφα. Αυτό έγινε γιατί αλλιώς θα είχαμε ανεπαρκή αριθμό δεδομένων για να μπορέσουμε να βγάλουμε κάποιο συμπέρασμα. Το κριτήριο που θέσαμε πληρούσαν συνολικά 1261 οικογένειες, μόλις το 5% του συνόλου των οικογενειών. Έπειτα με οπτικό έλεγχο (ενότητα 3.4) επιλέξαμε τα διαγράμματα που πληρούσαν αυτά τα κριτήρια καθώς και άλλα που αξίζει να σχολιαστούν. Γενικά, για τους παράγοντες μέσο μέγεθος εξονίου και αριθμό εξονίων μεταγράφου, ισχύει ότι 134 οικογένειες έχουν μικρότερο συντελεστή συσχέτισης Pearson από -0.5 ενώ 431 οικογένειες μεγαλύτερο από 0.5. Επίσης σε 87 οικογένειες η κλίση της ευθείας από τη γραμμική παλινδρόμηση ήταν μικρότερη από -0.5.

Πίνακας 2 *Περιεχόμενα οικογενειών γονιδίων. Εξόνια, μετάγραφα και γονίδια μπορεί να ανήκουν σε περισσότερες από 1 οικογένεια*

Σύνολο μοναδικών εξονίων	Σύνολο εξονίων	Σύνολο μεταγράφων	Σύνολο γονιδίων	Διάμεσος αριθμού εξονίων οικογενειών	Μέσος αριθμός εξονίων ανά οικογένεια	Μεγαλύτερο εξόνιο (σε βάσεις)	Μεγαλύτερο μετάγραφο (σε εξόνια)
354798	713919	80642	30076	7	29.5	33287 (nt)	363



Εικόνα 3: Ιστογράμματα της συχνότητας των οικογενειών γονιδίων του ανθρώπου με βάση τον αριθμό: γονιδίων, μεταγράφων και εξονίων που περιέχουν. Η κατανομή είναι νόμου δύναμης και πιο συγκεκριμένα οι οικογένειες γονιδίων ακολουθούν το νόμο του Zipf.

Πίνακας 3 *Ποσοστά εξονίων και μεταγράφων των οικογενειών γονιδίων*

Εξόνια στις οικογένειες	**Ποσότητα**	**Μετάγραφα στις οικογένειες**	**Ποσότητα (επί τοις %)**
Οικογένειες με 1 εξόνιο	470 (1.95%)	Οικογένειες με 1 μετάγραφο	13677 (56.61%)
Οικογένειες με [2-10] εξόνια	14137 (58.51%)	Οικογένειες με [2-10] μετάγραφα	8913 (36.89%)
Οικογένειες με [11-100] εξόνια	7835 (32.43%)	Οικογένειες με [11-100] μετάγραφα	1567 (6.48%)
Οικογένειες με [101-999] εξόνια	1709 (7.07%)	Οικογένειες με [101] μετάγραφα	3 (0.01%)
Οικογένειες με [1000 -] εξόνια	9 (0.04%)		

4.2 Διάκριση οικογενειών σε κατηγορίες με βάση τα διαγράμματα

Οι 1261 οικογένειες γονιδίων για τις οποίες δημιουργήσαμε διαγράμματα παρουσιάζουν μεγάλη ποικιλία κατανομών του μέσου μεγέθους εξονίων σε σχέση με τον αριθμό εξονίων των μεταγράφων. Η ποικιλία αυτή συνοψίζεται σε πέντε (5) γενικές κατηγορίες και παρουσιάζεται μέσω 23 αντιπροσωπευτικών οικογενειών γονιδίων (ενότητες 4.3 και 6.1).

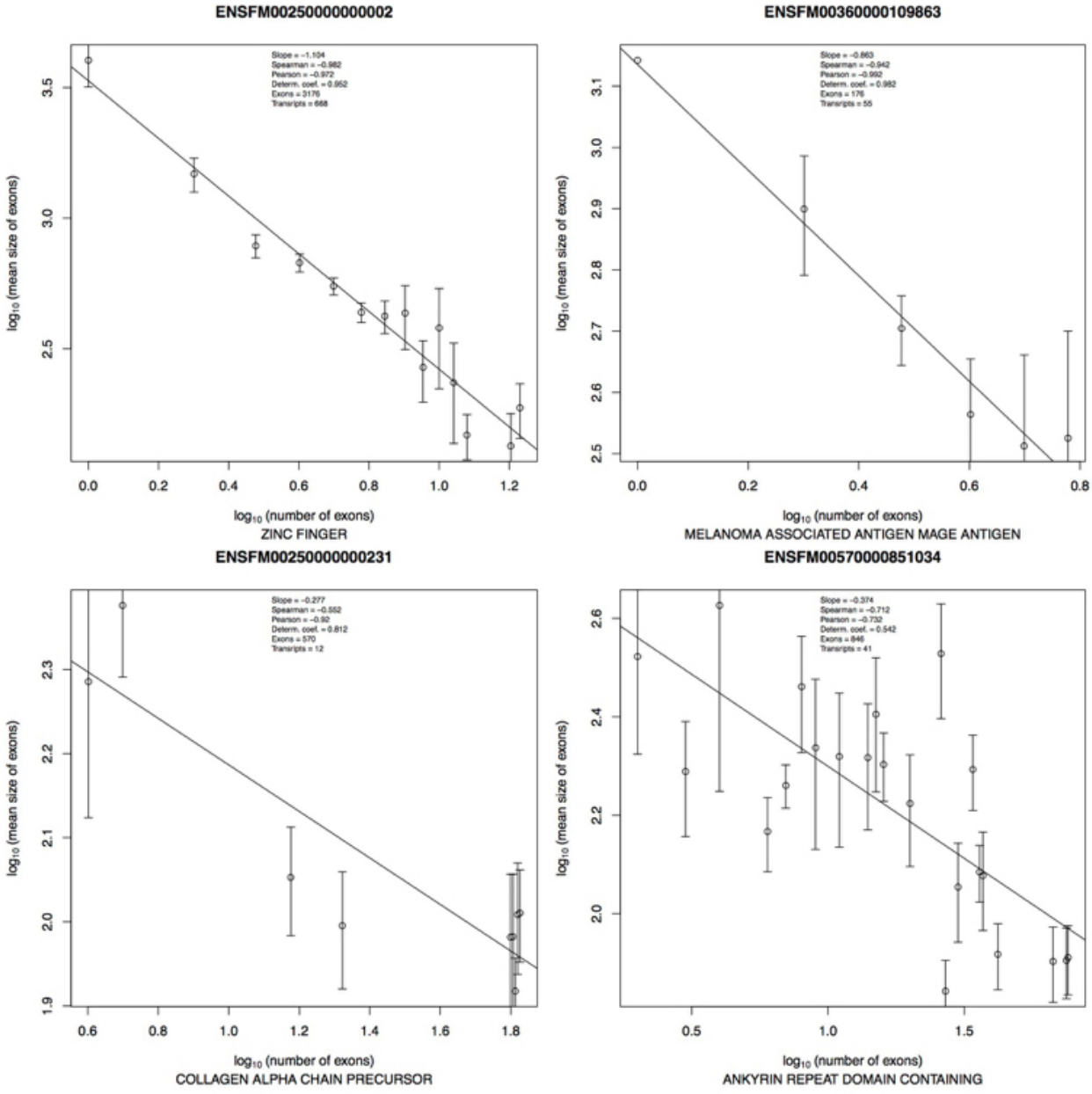
1. Παρουσιάζονται 10 οικογένειες γονιδίων που ακολουθούν το νόμο του Menzerath. Δηλαδή υπάρχει γραμμική αρνητική συσχέτιση σε διπλή λογαριθμική κλίμακα του μέσου μεγέθους των εξονίων με το πλήθος τους στα μετάγραφα.
2. Παρουσιάζονται συνολικά 7 οικογένειες γονιδίων που έχουν αντίστροφη συμπεριφορά από το νόμο του Menzerath. Δηλαδή εμφανίζεται θετική συσχέτιση σε διπλή λογαριθμική κλίμακα, άρα όσο αυξάνεται ο αριθμός εξονίων των μεταγράφων τόσο αυξάνεται το μέσο μέγεθος των εξονίων.
3. Παρουσιάζονται 3 οικογένειες γονιδίων που το μέγεθος εξονίων δεν έχει ισχυρή συσχέτιση με τον αριθμό εξονίων των μεταγράφων αλλά η κατανομή παραμένει γραμμική σε διπλή λογαριθμική κλίμακα.
4. Παρουσιάζονται 3 οικογένειες γονιδίων που εμφανίζουν 2 ή περισσότερες ανεξάρτητες γραμμικές συσχετίσεις στη διπλή λογαριθμική κλίμακα μεταξύ του μεγέθους των εξονίων και τον αριθμό των εξονίων των μεταγράφων.
5. Ακόμα υπάρχουν οικογένειες γονιδίων που δεν εμφανίζουν κάποια συσχέτιση αλλά ούτε και γραμμική κατανομή σε διπλή λογαριθμική κλίμακα μεταξύ του μεγέθους των εξονίων και τον αριθμό των εξονίων των μεταγράφων.

4.3 Χαρακτηριστικές περιπτώσεις

Ακολουθία με το νόμο του Menzerath είχαν 3 διαφορετικές οικογένειες των zinc fingers. Η οικογένεια ENSFM00250000000002 είναι η μεγαλύτερη οικογένεια σε αριθμό γονιδίων, μεταγράφων και εξονίων του ανθρώπου και είναι εκείνη που έχει την μεγαλύτερη υπακοή στο νόμο (Εικόνα 4). Τα zinc fingers είναι πρωτεϊνικά μοτίβα που έχουν την ικανότητα να προσδένονται στο DNA (3 νουκλεοτίδια μήκος κάθε μοτίβο) και εμφανίζονται σε περίπου τους μισούς μεταγραφικούς παράγοντες του ανθρώπου. Εμφανίζονται σε ομαδοποιημένα επαναλαμβανόμενα μοτίβα στις πρωτεΐνες. Στις ανθρώπινες πρωτεΐνες ο αριθμός τους κυμαίνεται από 4 έως 30 μοτίβα οπότε αναγνωρίζουν από 12 έως 90 νουκλεοτίδια του DNA ανάλογα με τον αριθμό τους. Οι (Emerson & Thomas, 2009) αναφέρουν ότι τα zinc fingers εμφανίζονται στους

ευκαρυωτικούς οργανισμούς όμως η οικογένειά τους εξαπλώνεται όλο και περισσότερο όσο προχωράμε από τα μετάζωα, στα τετράποδα, στα θηλαστικά και τέλος στα ανθρωποειδή. Αναφέρουν ακόμα ότι η μεγέθυνση των οικογενειών οφείλεται στο διπλασιασμό γονιδίων, η οποία ακολουθείται με διαγραφή γονιδίων και αναδιαρθρώσεις των ομαδοποιημένων μοτίβων.

Ένα ακόμα πρωτεϊνικό μοτίβο φαίνεται να ακολουθεί το νόμο του Menzerath, το μοτίβο Ankyrin repeat (Εικόνα 4). Είναι και αυτό ένα από τα συχνότερα μοτίβα και σχετίζεται με τις αλληλεπιδράσεις μεταξύ πρωτεϊνών. Έχει βρεθεί σε όλα τα βασίλεια οργανισμών και κυρίως σε πρωτεΐνες που σχετίζονται με διακυτταρική σηματοδότηση, μεταγραφή, ανοσολογική απόκριση και άλλα. Αποτελείται από 33 αμινοξέα τα οποία εμφανίζουν αρκετή συντήρηση και όπως κάθε μοτίβο εμφανίζεται σε επαναλήψεις. Και αυτή η οικογένεια έχει δεχθεί αυξήσεις αλλά και μειώσεις στα δομικά της στοιχεία (Mosavi & Cammett, 2004).

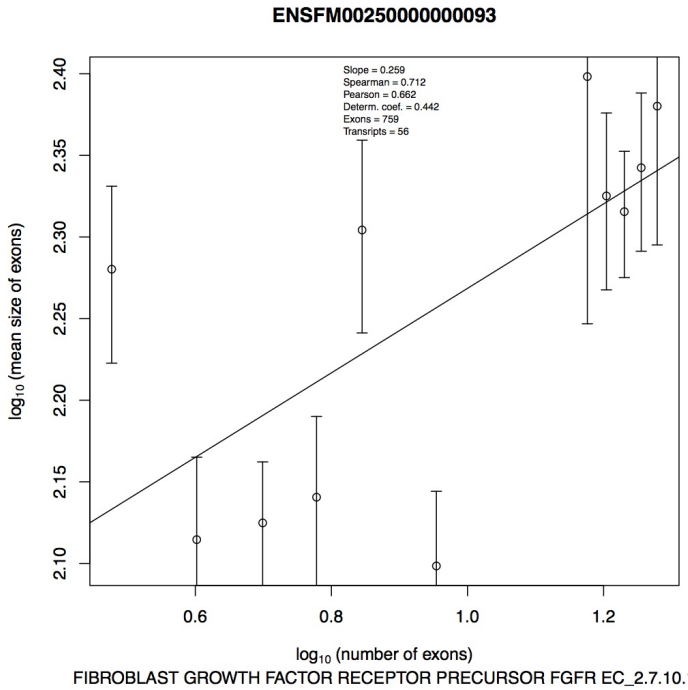


Εικόνα 4: Οι οικογένειες που εμφανίζουν πιστότητα στο νόμο του Menzerath. Φαίνεται η αρνητική σχέση σε διπλή λογαριθμική κλίμακα του μέσου μεγέθους εξονίων με των αριθμό των εξονίων των μεταγράφων. Κάθε σημείο αντιπροσωπεύει το σύνολο των μεταγράφων που έχουν τον ίδιο αριθμό εξονίων. Τα error bars παρουσιάζουν το τυπικό σφάλμα της μέσης τιμής.

Η οικογένεια του πρόδρομου μόριου της α αλυσίδας του κολλαγόνου υπακούει επίσης στο νόμο του Menzerath (Εικόνα 4). Το κολλαγόνο είναι η πιο άφθονη πρωτεΐνη των θηλαστικών και αποτελείται από τριμερές αλυσίδων α. Οι αρκετές ισομορφές των αλυσίδων α καθώς επίσης και οι μεταμεταφραστικές τροποποιήσεις αυξάνουν την ποικιλομορφία του κολλαγόνου. Κάθε αλυσίδα αποτελείται από επαναλήψεις τριών αμινοξέων Γλυκίνη - X – Y, όπου συνήθως X και Y είναι προλίνη και 4 – υδροξυπρολίνη αντίστοιχα. Το αρχέγονο γονίδιο αναφέρεται ότι είχε ένα εξόνιο απο 54 βάσεις και έπειτα μετά από διπλασιασμούς σχηματίστηκαν εξόνια και διαφορετικού μήκος λόγω άνισου ανασυνδυασμού. Ακολούθησαν 2 διπλασιασμοί γονιδιώματος (στην αρχή των χορδωτών και στο διαχωρισμό των κυκλοστομάτων και των γναθοστομάτων) (Exposito, Valcourt, Cluzel, & Lethias, 2010). Έπειτα τα γονίδια διατηρήθηκαν και δεν υπήρχαν άλλα φαινόμενα διπλασιασμού όμως συσσώρευσαν μεταλλάξεις τόσες που χάθηκαν τα ίχνη των αρχικών διπλασιασμών (Slatter, Farndale, & Slatter, 2015).

Η οικογένεια melanoma associated antigen MAGE antigen (ENSFM00360000109863) ακολούθησε διαφορετική εξελικτική ιστορία από τις προηγούμενες οικογένειες αλλά με παρόμοια αποτελέσματα (Εικόνα 4). Ο μηχανισμός retroposition φαίνεται να συνέβη στο αρχέγονο γονίδιο της οικογένειας δημιουργώντας έτσι αντίγραφα χωρίς ιντρόνια. Τα ιντρόνια χάνονται με αυτόν τον μηχανισμό γιατί το ώριμο mRNA μετατρέπεται σε cDNA το οποίο έπειτα εισάγεται στο DNA. Αυτός ο μηχανισμός είναι σπάνιος καθώς προϋποθέτει πως το γονίδιο εκφράζεται στα γαμετικά κύτταρα. Επίσης έδρασε και ο διπλασιασμός σε υποκατηγόρια γονιδίων κάτι που οδήγησε σε πολύ γρήγορη εξέλιξη. Όμως μερικές πρωτεΐνες της οικογένειας έχουν παραμείνει ανέπαφες (Chomez et al., 2001) οπότε και σε αυτήν την περίπτωση εμφανίζεται η δυναμική σχέση μεταξύ συντήρησης και αλλαγής κάτι που φαίνεται απαραίτητο για την εμφάνιση υπακοής στο νόμο του Menzerath.

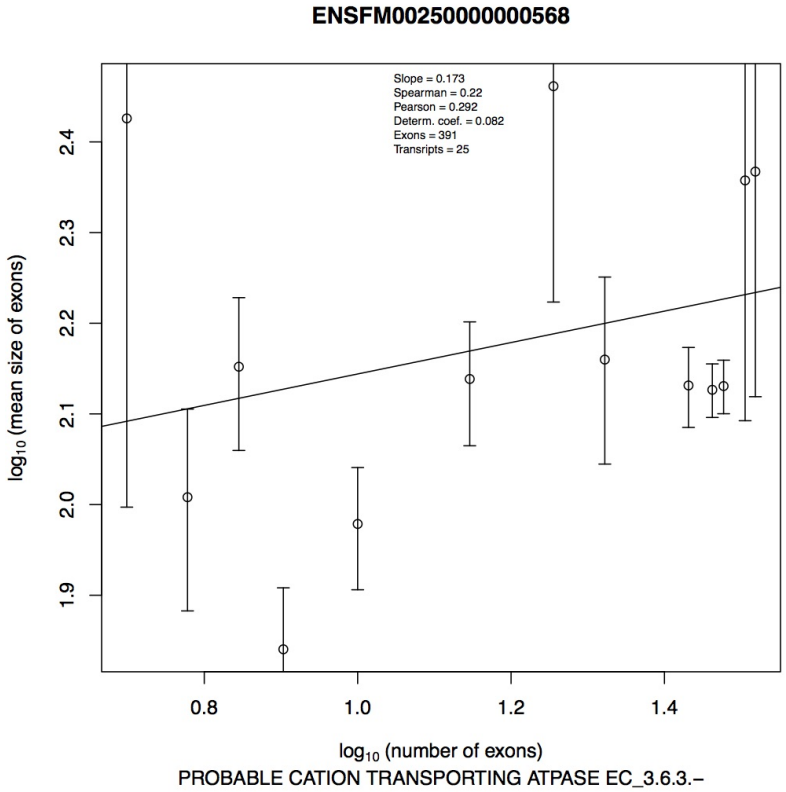
Σε μερικές οικογένειες παρατηρήθηκε η ταυτόχρονη αύξηση του αριθμού εξονίων και του μεγέθους τους στα μετάγραφα. Τέτοια περίπτωση αποτελεί και η περίπτωση των υποδοχέων των αυξητικών παραγόντων των ινοβλαστών – fibroblast growth factor receptors (FGFRs) (Εικόνα 5). Τα γονίδια της οικογένειας αυτής είναι μεμβρανικοί υποδοχείς εξειδικευμένοι για του αυξητικούς παράγοντες ινοβλαστών. Οι 2 αυτές οικογένειες γονιδίων έχουν συνεξελιχθεί και μαζί πέρασαν δύο φάσεις εξάπλωσης, η πρώτη οφειλόταν σε διπλασιασμό γονιδίων στην αρχή της εμφάνισης των μεταζώων ενώ η δεύτερη οφειλόταν σε διπλασιασμό γονιδιώματος που συνέβη στην αρχή των σπονδυλωτών. Τα γονίδια της FGFRs οικογένειας εμφανίζουν μεγάλη ποικιλότητα λειτουργιών σε αναπτυξιακές και φυσιολογικές διεργασίες λόγω κυρίως του



Εικόνα 5: Παρόλο που είναι μεγάλα τα error bars είναι εμφανής η θετική συσχέτιση μέσου μεγέθους εξονίων και αριθμού εξονίων των μεταγράφων. Κάθε σημείο αντιπροσωπεύει το σύνολο των μεταγράφων που έχουν τον ίδιο αριθμό εξονίων. Τα error bars παρουσιάζουν το τυπικό σφάλμα της μέσης τιμής.

εναλλακτικού ματίσματος. Τα γονίδια αυξητικών παραγόντων ινοβλαστών είναι χωρισμένα σε 18 οικογένειες οπότε δεν μπορέσαμε να εξετάσουμε αν η συνεξέλιξη τους με τα FGFRs αντικατοπτρίζεται στη σχέση μεγέθους εξονίων με αριθμό γονιδίων μεταγράφων (Itoh, 2007).

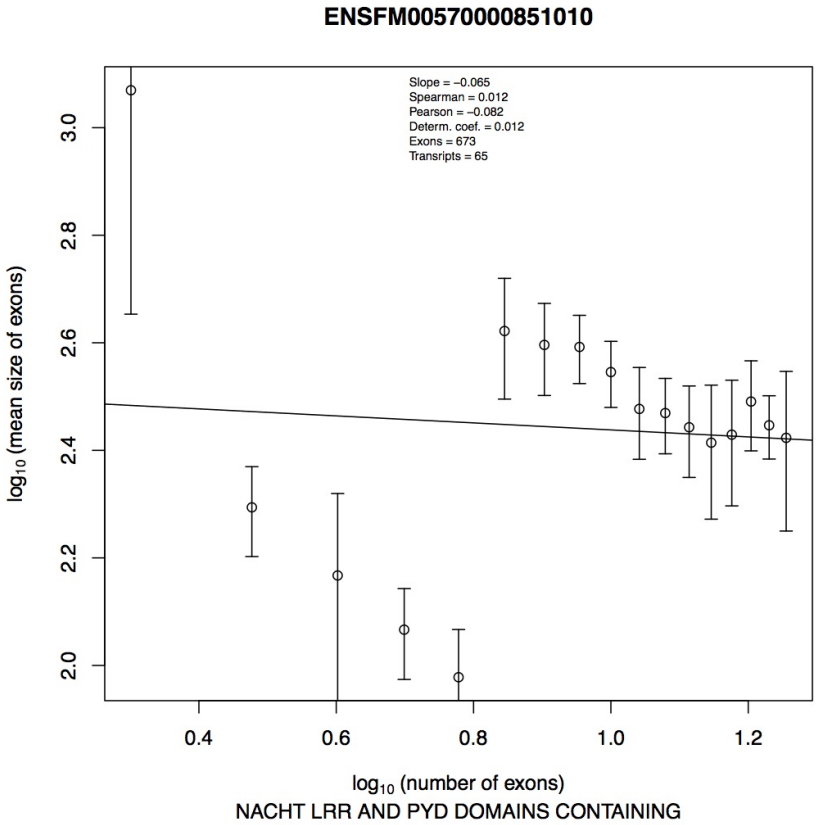
Στην κατηγορία που δεν υπάρχει υπακοή στο νόμο του Menzerath χαρακτηριστικό παράδειγμα αποτελεί η οικογένεια P-type ATPases (probable cation transporting ATPase). Αυτή η οικογένεια παρουσιάζει μικρή θετική γραμμική συσχέτιση σε διπλή λογαριθμική κλίμακα μεταξύ μεγέθους εξονίων και αριθμού εξονίων των μεταγράφων (Εικόνα 6). Τα γονίδιά της είναι αντλίες κατιόντων τα οποία έχουν μοναδική αλληλουχία στα γονιδιώματα όλων των οργανισμών. Δηλαδή οι αντλίες αυτές δεν μοιράζονται πρωτεϊνικούς τομείς με άλλα ένζυμα. Επίσης έχει δειχθεί ότι προήλθαν από διπλασιασμούς ενός αρχέγονου γονιδίου. Αυτοί οι διπλασιασμοί αναφέρεται ότι συνέβησαν πριν 2 δισεκατομμύρια χρόνια, δηλαδή πριν την εμφάνιση των ευκαρυωτών. Άρα τα γονίδια είναι πάρα πολύ διατηρημένα στο χρόνο, 2 από τα 3 τμήματά τους παρουσιάζουν υψηλή συντήρηση αλληλουχία (Fagan & Saier, 1994).



Εικόνα 6: Σε αυτή την περίπτωση δεν μπορούμε να στηρίξουμε ότι υπάρχει κάποια συγκεκριμένη συσχέτιση γιατί παρόλο που ο συντελεστής Pearson είναι θετικός γιατί τα error bars είναι σχετικά μεγάλα. Κάθε σημείο αντιπροσωπεύει το σύνολο των μεταγράφων που έχουν τον ίδιο αριθμό εξονίων. Τα error bars παρουσιάζουν το τυπικό σφάλμα της μέσης τιμής.

Η οικογένεια NACHT, LRR AND PYD DOMAINS CONTAINING (ENSFM00570000851010) εμφανίζει ενδιαφέρουσα κατανομή (Εικόνα 7) την οποία εμφανίζουν και άλλες οικογένειες (Παράρτημα 6.1). Παρατηρείται ότι υπάρχουν 2 διαφορετικές κατανομές με βάση το μέγεθος εξονίου σε σχέση με αριθμό μεταγράφων. Αυτό σημαίνει ότι υπάρχουν 2 διακριτές ομάδες μεταγράφων που διαφέρουν σε αριθμό εξονίων, δίνοντας την αίσθηση ότι προέρχονται από διαφορετικές οικογένειες. Η συγκεκριμένη οικογένεια περιέχει πρωτεΐνες που έχουν 3 διαφορετικούς τομείς και διαδραματίζουν σημαντικό ρόλο στην έμφυτη ανοσία. Η λειτουργία τους εμπλέκεται στη θανάτωση μικροβίων εισβολών στον οργανισμό αλλά δεν έχουν προσδιοριστεί τα βιοχημικά μονοπάτια (Tschopp, Martinon, & Burns, 2003). Η σύνδεση των τομέων NACHT και PYD αναφέρεται ότι έγινε άπαξ μέσω του domain shuffling ενώ έπειτα ακολούθησαν πολλά φαινόμενα διπλασιασμού (Proell, Riedl, Fritz, Rojas, & Schwarzenbacher, 2008). Με περισσότερη διερεύνηση στα μετάγραφα και τα εξόνια που περιέχουν θα μπορέσει κανείς να δει ποιοι τομείς απουσιάζουν

από τις διαφορετικές ομάδες μεταγράφων που σχηματίζονται στο Εικόνα 7.



Εικόνα 7: Φαίνεται ότι υπάρχουν 2 ξεχωριστές ομάδες μεταγράφων. Αυτό οφείλεται στο εναλλακτικό μάτισμα των μεταγράφων. Από ότι φαίνεται στις πρωτεΐνες που έχουν διαφορετικές υπομονάδες υπάρχουν περιπτώσεις που δεν εκφράζεται κάποια. Φαίνεται επίσης η ξεχωριστή αλλά ίδια τάση, που έχουν τα μετάγραφα, να μειώνεται το μέσο μέγεθος εξονίων όσο αυξάνονται τα εξόνια. Κάθε σημείο αντιπροσωπεύει το σύνολο των μεταγράφων που έχουν τον ίδιο αριθμό εξονίων. Τα error bars παρουσιάζουν το τυπικό σφάλμα της μέσης τιμής.

Οι οικογένειες που αναφέρθηκαν αποτελούν ένα μικρό μέρος των οικογενειών που χωρίστηκαν σε κατηγορίες στην προηγούμενη ενότητα οπότε θα πρέπει να αναλυθούν και οι υπόλοιπες (ενότητα 6.1) για περαιτέρω τεκμηρίωση.

4.4 Συζήτηση

Η υπακοή στο νόμο του Menzerath, όπως εφαρμόστηκε εδώ, δεν αποτελεί τον κανόνα αλλά μια ειδική κατάσταση στις οικογένειες γονιδίων. Αυτή σκιαγραφεί την στοχαστική συσχέτιση μεταξύ αριθμού εξονίων και μεγέθους τους σε κάθε μετάγραφο. Ο (Nikolaou, 2014) έδειξε ότι όταν αυξάνουν παράγοντες όπως συντήρηση αλληλουχίας και πολύπλοκη μεταγραφική ρύθμιση - λόγω αυξημένων εναλλακτικών μεταγράφων - ελαττώνεται η υπακοή στο νόμο του Menzerath. Στη γλωσσολογία ισχύει ότι οι λέξεις είναι φορείς πληροφορίας που έχουν πεπερασμένο μέγεθος. Η πληροφορία που περιέχουν βρίσκεται στους συνδυασμούς δομικών στοιχείων, των συλλαβών. Η αλληλουχία αυτών των δομικών στοιχείων καθορίζει την πληροφορία της συνολικής δομής, δηλαδή της λέξης. Όμως επειδή οι λέξεις έχουν πεπερασμένο μέγεθος όσο αυξάνεται η πληροφορία που περιέχουν εκείνη θα πρέπει να περιοριστεί σε συγκεκριμένο χώρο (Cramer, 2005). Δηλαδή το μήνυμα της πληροφορίας μπορεί να περιοριστεί χωρίς όμως να χαθεί η πληροφορία. Στις ανθρώπινες γλώσσες υπάρχει αυτή η δυνατότητα καθώς η πληροφορία που περιέχουν οι λέξεις μπορεί να αποδοθεί με λιγότερα γράμματα γιατί περιέχουν πλεονάζοντα στοιχεία (Ruelle, 1993). Η σχέση που προσδιορίζει ο νόμος του Menzerath (μεταξύ του μεγέθους συνόλου και του μεγέθους των μερών που το αποτελούν) δεν συμπεριλαμβάνει το βέλος του χρόνου. Δηλαδή, όπως και οι νόμοι των Zipf και Heaps, αποτελεί μια στιγμιαία απεικόνιση χωρίς να συμπεριλαμβάνει τη δυναμική του συστήματος

(Petersen, Tenenbaum, Havlin, & Stanley, 2012). Όμως η “συμπεριφορά” των συστημάτων ανεξάρτητων από την κλίμακα την οποία περιγράφουν αυτοί οι νόμοι φαίνεται να οφείλεται στην εξελικτική τους ιστορία (Barabasi & Albert, 1999).

Ο διαχωρισμός των γονιδίων σε οικογένειες γίνεται με βάση την ομοιότητα της αλληλουχίας τους. Η ύπαρξη ομοιότητας αλληλουχιών μεταξύ γονιδίων προμηνύει την κοινή τους καταγωγή από κάποιο αρχέγονο γονίδιο. Πολύ σημαντικός μηχανισμός για την αύξηση του αριθμού γονιδίων είναι ο διπλασιασμός τους (Zhang, 2003). Το διπλασιασμένο γονίδιο συνήθως απορρίπτεται από τη φυσική επιλογή όμως υπάρχουν φορές που ο διπλασιασμός οδηγεί σε καινοτομίες στο γονιδίωμα. Οι καινοτομίες μπορεί να είναι είτε νέες λειτουργίες που αποκτά το ένα αντίγραφο είτε μέσω μεταλλάξεων τα αντίγραφα εξειδικεύονται σε επιμέρους λειτουργίες που είχε το προγονικό γονίδιο (M Lynch & Conery, 2000). Η δημιουργική διάσταση του διπλασιασμού ενισχύεται όταν η δράση της φυσικής επιλογής περιορίζεται. Αυτό συμβαίνει γιατί τα προϊόντα ενός διπλασιασμού μπορεί βραχυπρόθεσμα να μειώνουν την προσαρμοστικότητα του οργανισμού αλλά μακροπρόθεσμα σε συνδυασμό με μεταλλάξεις να αναδυθούν πλεονεκτήματα. Όμως όταν η φυσική επιλογή είναι κυρίαρχη στην εξέλιξη ενός οργανισμού χάνονται οι μακροπρόθεσμες προοπτικές καθώς οι φαινομενικά βλαβερές αλλαγές απορρίπτονται άμεσα. Αντίθετα με την τυχαία γενετική παρέκκλιση διατίθεται περισσότερος χρόνος σε αλλαγές όπως ο διπλασιασμός γονιδίων. Αυτό συμβαίνει σε οργανισμούς με μικρό δραστικό μέγεθος πληθυσμού καθώς ενισχύεται η γενετική παρέκκλιση (Michael Lynch & Conery, 2003).

Διπλασιασμός δεν συμβαίνει μόνο στα γονίδια. Έχει παρατηρηθεί ότι υπάρχει αρνητική συσχέτιση μεταξύ αριθμού γονιδίων των οικογενειών, δηλαδή εμφάνισης διπλασιασμένων γονιδίων, με τον αριθμό των εναλλακτικών μεταγράφων (Kopelman, Lancet, & Yanai, 2005). Ο βαθμός της έκφρασης και το μέγεθος του γονιδίου επηρεάζουν την επιλογή της διακλάδωσης μεταξύ του διπλασιασμού του ή την αύξηση εναλλακτικών μεταγράφων του (Grishkevich & Yanai, 2014). Πιο συγκεκριμένα όσο μεγαλύτερη είναι η έκφραση ή/και το μέγεθος ενός γονιδίου τόσο περιορίζεται η πιθανότητα διπλασιασμού του ενώ μεγαλώνει η πιθανότητα αύξησης εναλλακτικών μεταγράφων του. Η αύξηση του μεγέθους του γονιδίου οδηγεί σε μεγαλύτερη αστάθεια του διπλασιασμού δημιουργώντας έτσι κατακερματισμένα προϊόντα που συνήθως απορρίπτει η φυσική επιλογή. Οι (Kopelman et al., 2005) κατέληξαν στο συμπέρασμα ότι τα αντίγραφα και τα εναλλακτικά μετάγραφα είναι αποτέλεσμα του ίδιου μηχανισμού, του διπλασιασμού, ο οποίος επιδρά σε διαφορετικό ιεραρχικό επίπεδο. Άρα αναφέρουν ότι η αύξηση των εναλλακτικών μεταγράφων είναι αποτέλεσμα διπλασιασμού εξονίων. Όμως προχωρώντας παραπέρα ο διπλασιασμός φαίνεται να συμβαίνει σε πολλά διαφορετικά επίπεδα της ζωής, από τα εξόνια, τα γονίδια,τα χρωμοσώματα,τα γονιδιώματα σε ολόκληρα τα κύτταρα και ακόμα -σε μια πιο πολύπλοκη μορφή- στην αναπαραγωγή πολυκύτταρων οργανισμών. Ο διπλασιασμός μονάδων ενός δικτύου - συστήματος φαίνεται να είναι ιδιότητα ανεξάρτητη κλίμακας που συνδράμει στην ανάδυση νέων χαρακτηριστικών των επιμέρους μονάδων αλλά και ολόκληρου του δικτύου – συστήματος.

Το διπλασιασμό γονιδίων αντισταθμίζει η διαγραφή γονιδίων. Η διαγραφή γονιδίων γίνεται αρχικά με τη μετατροπή τους σε ψευδογονίδια μέσω συσσώρευση μεταλλάξεων και έπειτα ακολουθεί - σε κάποιες περιπτώσεις - η απαλοιφή της αλληλουχίας τους. Η δυναμική του μεγέθους των οικογενειών γονιδίων περιγράφεται καλύτερα μέσω τυχαίων διεργασιών δημιουργίας νέων γονιδίων και απαλοιφή άλλων (Demuth et al., 2006; Demuth & Hahn, 2009). Συχνά λόγω φυσικής επιλογής το φαινόμενο του διπλασιασμού ακολουθεί η διαγραφή. Επίσης έχει παρατηρηθεί ότι με το πέρασμα του χρόνου το μέγεθος των γονιδίων τείνει να αυξάνεται εξαιτίας κυρίως των μεταθετών στοιχείων. Η αύξηση του μεγέθους των γονιδίων συνεπάγεται και την αύξηση των εξονίων τους (Roux & Robinson-Rechavi, 2011). Επιπροσθέτως όσο αυξάνεται το μέγεθος ενός γονιδίου αυξάνονται και τα εναλλακτικά μετάγραφά του.

Συμπερασματικά, ο νόμος του Menzerath φαίνεται να αντικατοπτρίζει ένα συγκεκριμένο εύρος της δυναμικής σχέσης μεταξύ της αύξησης των μερών του συνόλου με τον ταυτόχρονο περιορισμό του μεγέθους τους. Συστήματα που έχουν αυξημένους περιορισμούς δεν επιτρέπουν την αύξηση των μερών τους ενώ άλλα συστήματα ενδεχομένως να μην περιορίζουν τα μέρη όσο αυτά αυξάνονται σε αριθμό. Αυτές οι 3 διαφορετικές δυναμικές φάνηκαν στις οικογένειες που σχολιάστηκαν προηγουμένως. Αρχικά οι οικογένειες που εμφάνισαν πολλά φαινόμενα διπλασιασμού (επαναλήψεις αλληλουχιών zinc fingers) αλλά ταυτόχρονα και εξειδίκευση - η οποία διατηρείται μέσω περιορισμών – παρουσίασαν υπακοή στο νόμο του Menzerath.

Έπειτα η πολύ συντηρημένη οικογένεια αντλιών κατιόντων φαίνεται να έχει φτάσει σε σταθερή κατάσταση όπου τα μετάγραφα έχουν πολλά εξόνια με μικρές διαφορές μεγέθους παρουσιάζοντας έτσι ασθενή υπακοή στο νόμο. Οι 2 αυτές περιπτώσεις έρχονται σε συμφωνία με το εξελικτικό μοντέλο του (N(Nikolaou, 2014) για το νόμο του Menzerah.

Οι οικογένειες γονιδίων όμως περιείχαν μια ακόμα περίπτωση που ήταν αντίστροφη με το νόμο του Menzerath. Οι 2 οικογένειες που σχολιάστηκαν (ενότητα 4.2) φάνηκε ότι εξαπλώθηκαν γρήγορα μέσω διπλασιασμών αποκτώντας παράλληλα μεγάλη ποικιλότητα. Η ποικιλότητα λειτουργιών ίσως είναι ο λόγος που δεν περιορίζεται το μέγεθος των εξονίων όταν αυτά αυξάνονται σε αριθμό στα μετάγραφα. Η σχέση ποικιλότητας και περιορισμών πιθανόν να έχουν αρνητική συσχέτιση. Τέλος φάνηκε η διαφορετική έκφραση τομέων πρωτεϊνών μιας οικογένειας γονιδίων σε κάποιες περιπτώσεις μπορεί να εντοπιστεί μέσω της ομαδοποίησης εναλλακτικών μεταγράφων με βάση τον αριθμό εξονίων.

Η συσχέτιση μεγέθους μονάδας με αριθμό μονάδων του συνόλου είναι σημαντική καθώς φαίνεται ότι η ιστορία και οι δυναμικές κάποιων συστημάτων αντανακλούν σε στατιστικά μέτρα. Παρόλο που περιγράφηκαν κάποιες τάσεις στις οικογένειες που ομαδοποιούνται ανάλογα με τη σχέση αριθμού εξονίων με το μέγεθός τους στα μετάγραφα απαιτούνται περισσότερα δεδομένα και αναλύσεις για πληρέστερη τεκμηρίωση. Πιο συγκεκριμένα χρειάζεται επιμέρους ανάλυση της συντήρησης αλληλουχίας και του βαθμού έκφρασης γονιδίων καθώς και τον προσδιορισμό σημαντικών εξελικτικών γεγονότων (πρωτογενής έρευνα ή/και εκτενέστερη αναζήτηση στη βιβλιογραφία) των μελών κάθε οικογένειας γονιδίων για περαιτέρω έλεγχο των προηγούμενων ερμηνειών. Επίσης η διερεύνηση της σχέσης ποικιλότητας λειτουργιών με την συντήρηση – περιορισμό στο επίπεδο γονιδιώματος ίσως διαφώτιζε περισσότερο τις αιτίες εμφάνισης οικογενειών που έχουν αντίστροφη συσχέτιση από το νόμο του Menzerath. Ακόμα το γεγονός ότι η ENSEMBL κατηγοριοποιεί τα γονίδια σε τεράστιο αριθμό οικογενειών οδήγησε στην απόρριψη του 95% των οικογενειών γιατί οι περισσότερες έχουν ανεπαρκή αριθμό γονιδίων για ανάλυση (Πίνακας 3). Οπότε σε μελλοντικές μελέτες ίσως να ήταν καλύτερο να χρησιμοποιηθούν μεγαλύτερες ομάδες γονιδίων. Η διερεύνηση και σε άλλες ομάδες γονιδίων με βάση τη βιολογική διεργασία, την τοπολογία ή/ και το βιοχημικό μονοπάτι θα βοηθούσε στην καλύτερη κατανόηση των μηχανισμών που οδηγούν στη δυναμική σχέση μεταξύ αριθμού εξονίων και μεγέθους τους σε κάθε μετάγραφο (ή γονίδιο). Αυτά θα βοηθήσουν στην εμβάθυνση των γνώσεών μας για τις δυναμικές που αντιπροσωπεύει ο νόμος του Menzerath στο γονιδίωμα κάτι που ίσως να αποτελέσει εργαλείο χαρακτηρισμού δυναμικών σχέσεων και σε άλλα συστήματα.

5 Βιβλιογραφία

Almirantis, Y., Arndt, P., Li, W., & Provata, A. (2014). Editorial: Complexity in genomes. *Computational Biology and Chemistry*, *53 Pt A*(14), 1–4. doi:10.1016/j.compbiolchem.2014.08.003

Altmann, E. G., & Gerlach, M. (2014). Statistical laws in linguistics. In *Flow Machines Workshop: Creativity and Universality in Language* (pp. 1–12). Paris. Retrieved from http://arxiv.org/abs/1502.03296

Altmann, G. (1980). Prolegomena to Menzerath ' s law. *Glottometrika*, *2*(1), 1–10.

Baixeries, J., Hernández-fernández, A., & Ferrer-i-cancho, R. (2012). Random models of Menzerath – Altmann law in genomes. *BioSystems*, *107*(3), 167–173. doi:10.1016/j.biosystems.2011.11.010

Baixeries, J., Hernández-fernández, A., Forns, N., & Ferrer-i-cancho, R. (2013). The Parameters of Menzerath-Altmann. *Journal of Quantitative Linguistics*, *20*(2), 94–104.

Bak, P., & Tang, C. (1987). Self organized criticality/ an explanation of 1/f noise. *Physical Review Letters*, *59*(4), 381–384.

Barabasi, A., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, *286*(October), 509–512.

Chomez, P., Backer, O. De, Bertrand, M., Plaen, E. De, Boon, T., & Lucas, S. (2001). An Overview of the MAGE Gene Family with the Identification of All Human Members of the Family. *Cancer Research*, *61*, 5544–5551.

Chomsky, N. (1957). *Syntactic Structures* (First Edit.). Berlin: Mouton de Gruyter.

Chothia, C. (1992). One thosand families for the molecular biologist. *Nature*, *357*(18 June), 543–544.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-Law Distributions in Empirical Data. *SIAM Review*, *51*(4), 661–703. doi:10.1137/070710111

Corominas-Murtra, B., & Solé, R. V. (2010). Universality of Zipf's law. *Physical Review E*, *82*(1), 011102. doi:10.1103/PhysRevE.82.011102

Cramer, I. (2005). The Parameters of the Altmann-Menzerath Law. *Journal of Quantitative Linguistics*, *12*(1), 41–52. doi:10.1080/09296170500055301

Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., … Flicek, P. (2014). Ensembl 2015. *Nucleic Acids Research*, *43*(D1), D662–D669. doi:10.1093/nar/gku1010

Demuth, J. P., De Bie, T., Stajich, J. E., Cristianini, N., & Hahn, M. W. (2006). The evolution of mammalian gene families. *PloS One*, *1*(1), e85. doi:10.1371/journal.pone.0000085

Demuth, J. P., & Hahn, M. W. (2009). The life and death of gene families. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, *31*(1), 29–39. doi:10.1002/bies.080085

Egghe, L. (2007). Untangling Herdan's Law and Heaps' Law: Mathematical and Informetric Arguments. *AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*, *58*(5), 702–709. doi:10.1002/asi

Emerson, R. O., & Thomas, J. H. (2009). Adaptive evolution in zinc finger transcription factors. *PLoS Genetics*, *5*(1). doi:10.1371/journal.pgen.1000325

Enright, A. J., Dongen, S. Van, & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families, *30*(7), 1575–1584.

Eroglu, S. (2013). Language-Like Behavior of Protein Length Distribution in Proteomes. *Complexity*, 1–13. doi:10.1002/cplx

Exposito, J. Y., Valcourt, U., Cluzel, C., & Lethias, C. (2010). The fibrillar collagen family. *International Journal of Molecular Sciences*, *11*(2), 407–426. doi:10.3390/ijms11020407

Fagan, M. J., & Saier, M. H. (1994). P-type ATPases of eukaryotes and bacteria: sequence analyses and construction of phylogenetic trees. *Journal of Molecular Evolution*, *38*(1), 57–99. doi:10.1007/BF00175496

Ferrer-i-cancho, R., & Forns, N. (2010). The Self-Organization of Genomes, *15*(5), 34–36. doi:10.1002/cplx

Ferrer-i-cancho, R., Hernández-fernández, A., & Baixeries, J. (2014). When is Menzerath-Altmann law mathematically trivial ? A new approach, *13*(6), 633–644. doi:10.1515/sagmb-2013-0034

Grishkevich, V., & Yanai, I. (2014). Gene length and expression level shape genomic novelties. *Genome Research*, *24*(9), 1497–503. doi:10.1101/gr.169722.113

Hauke, J., & Kossowski, T. (2011). Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data. *Quaestiones Geographicae*, *30*(2), 87–93. doi:10.2478/v10117-011-0021-1

Heaps, H. S. (1978). *Information Retrieval: Computational and Theoretical Aspects*. Orlando, FL, USA: Academic Press, Inc.

Huynen, M. a., & van Nimwegen, E. (1998). The frequency distribution of gene family sizes in complete genomes. *Molecular Biology and Evolution*, *15*(5), 583–589. doi:10.1093/oxfordjournals.molbev.a025959

Itoh, N. (2007). The Fgf families in humans, mice, and zebrafish: their evolutional processes and roles in development, metabolism, and disease. *Biological & Pharmaceutical Bulletin*, *30*(10), 1819–1825. doi:10.1248/bpb.30.1819

Kauffman, S. A. (1991). Antichaos and Adaptation. *Scientific American*, (August), 78–84.

Kopelman, N. M., Lancet, D., & Yanai, I. (2005). Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. *Nature Genetics*, *37*(6), 588–9. doi:10.1038/ng1575

Kunin, V., Cases, I., Enright, A. J., Lorenzo, V. De, & Ouzounis, C. A. (2003). Myriads of protein families , and still counting, 1–2.

Li, W. (1992). Random Texts Exhibit Zipfs-Law-Like Word Frequency Distribution Wentian. *IEEE TRANSACTIONS ON INFORMATION THEORY*, *38*(6), 1842–1845.

Li, W. (2012). Menzerath's Law at the Gene-Exon Level in the Human Genome, *17*(4), 49–53. doi:10.1002/cplx

Li, W., Freudenberg, J., & Oswald, M. (2015). Principles for Organizing Gene-Sets. *Unpublisted Work*, 1–29.

Lu, L., Zhang, Z.-K., & Zhou, T. (2010). Zipf's Law Leads to Heaps' Law: Analyzing Their Relation in Finite-Size Systems. *PloS One*, *5*(12), 1–11. doi:10.1371/Citation

Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science (New*

York, N.Y.), *290*(5494), 1151–1155. doi:10.1126/science.290.5494.1151

Lynch, M., & Conery, J. S. (2003). The origins of genome complexity. *Science (New York, N.Y.)*, *302*(5649), 1401–4. doi:10.1126/science.1089370

Menzerath, P. (1954). *Die Architektonik des deutschen Wortschatzes.* F. D{ü}mmler. Retrieved from https://books.google.gr/books?id=z2UuAAAAIAAJ

Mosavi, L., & Cammett, T. (2004). The ankyrin repeat as molecular architecture for protein recognition. *Protein Science*, *13*, 1435–1448. doi:10.1110/ps.03554604.ity

Newman, M. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, *46*(5), 323–351. doi:10.1080/00107510500052444

Nikolaou, C. (2014). Menzerath-Altmann law in mammalian exons reflects the dynamics of gene structure evolution. *Computational Biology and Chemistry*, *53 Pt A*, 134–43. doi:10.1016/j.compbiolchem.2014.08.018

Ouzounis, C., & Mazière, P. (2006). Maps, books and other metaphors for systems biology. *Bio Systems*, *85*(1), 6–10. doi:10.1016/j.biosystems.2006.02.007

P Copland. (2005). The book of life. *J Med Ethics*, *52*, 278–279. doi:10.1093/jts/os-XIII.52.580

Pappas, C. T., Bliss, K. T., Zieseniss, A., & Gregorio, C. C. (2012). The Nebulin Family: an Actin Support Group. *Trends in Cell Biology*, *21*(1), 29–37. doi:10.1016/j.tcb.2010.09.005.The

Paton, R. (1996). Metaphors, models and bioinformation. *Biosystems*, *38*(2-3), 155–162. doi:10.1016/0303-2647(95)01586-8

Petersen, A. M., Tenenbaum, J., Havlin, S., & Stanley, H. E. (2012). Statistical laws governing fluctuations in word use from word birth to word death. *Scientific Reports*, *2*, 313. doi:10.1038/srep00313

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529. doi:10.1073/pnas.1012551108/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1012551108

Prigogine, I., & Antoniou, I. (2000). Science, Evolution and Complexity. In *Genetics in Europe - Open days 2000 (GEOD 2000)* (pp. 21–36). Brussels.

Proell, M., Riedl, S. J., Fritz, J. H., Rojas, A. M., & Schwarzenbacher, R. (2008). The Nod-Like Receptor (NLR) family: A tale of similarities and differences. *PLoS ONE*, *3*(4), 1–11. doi:10.1371/journal.pone.0002119

Roux, J., & Robinson-Rechavi, M. (2011). Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication. *Genome Research*, *21*(3), 357–363. doi:10.1101/gr.113803.110

Ruelle, D. (1993). *Chance And Chaos*. Princeton University Press.

Rzhetsky, a, & Gomez, S. M. (2001). Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. *Bioinformatics (Oxford, England)*, *17*(10), 988–996. doi:10.1093/bioinformatics/17.10.988

Searls, B. (1992). The Linguistics of DNA. *American Scientist*, *80*(6), 579–591.

Searls, D. B. (2002). The language of genes. *Nature*, *420*, 211–217.

Slatter, D. A., Farndale, R. W., & Slatter, D. A. (2015). Structural constraints on the evolution of the collagen fibril : convergence on a 1014-residue COL domain. *Open Biology*, *5*.

Sole, R., & Goodwin, B. (2008). *Signs Of Life How Complexity Pervades Biology: How Complexity Pervades Biology*. Basic Books. Retrieved from https://books.google.gr/books?id=d9TMFMlwx1QC

Solé, R. V. (2010). Genome size, self-organisation and DNA ’ s Dark Matter. *Complexity*, *16*(1), 20–23. doi:10.1002/cplx

Stumpf, M. P. H., & Porter, M. a. (2012). Critical Truths About Power Laws. *Science*, *335*(6069), 665–666. doi:10.1126/science.1216142

Tettelin, H., Riley, D., Cattuto, C., & Medini, D. (2008). Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology*, *11*(5), 472–477. doi:10.1016/j.mib.2008.09.006

Theraulaz, G., & Bonabeau, E. (1995). Coordination in Distributed Building, *269*(August), 686–688.

Theraulaz, G., & Bonabeau, E. (1995). Modelling the Collective Building of Complex Architectures in Social Insects with Lattice Swarms. *J. Theor. Biol.*, (177), 381–400. doi:10.1006/jtbi.1995.0255

Tschopp, J., Martinon, F., & Burns, K. (2003). NALPs: a novel protein family involved in inflammation. *Nature Reviews. Molecular Cell Biology*, *4*(2), 95–104. doi:10.1038/nrm1019

Tsonis, P. a., & Tsonis, A. a. (2002). Linguistic Features in Eukaryotic Genomes. *Complexity*, *7*(4), 13–15. doi:10.1002/cplx.10035

Virkar, Y., & Clauset, A. (2014). Power-law distributions in binned empirical data. *The Annals of Applied Statistics*, *8*(1), 89–119. doi:10.1214/13-AOAS710

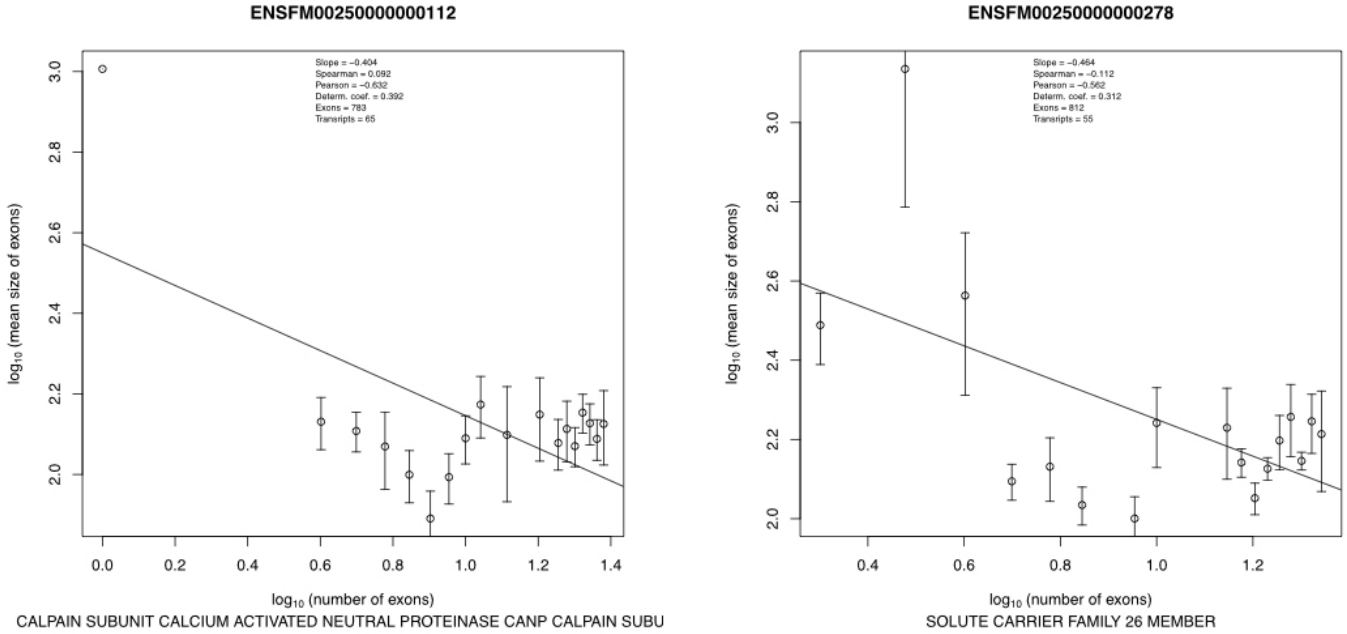
West, G. B. (1997). A General Model for the Origin of Allometric Scaling Laws in Biology. *Science*, *276*(5309), 122–126. doi:10.1126/science.276.5309.122

Zhang, J. (2003). Evolution by gene duplication: An update. *Trends in Ecology and Evolution*, *18*(6), 292–298. doi:10.1016/S0169-5347(03)00033-8

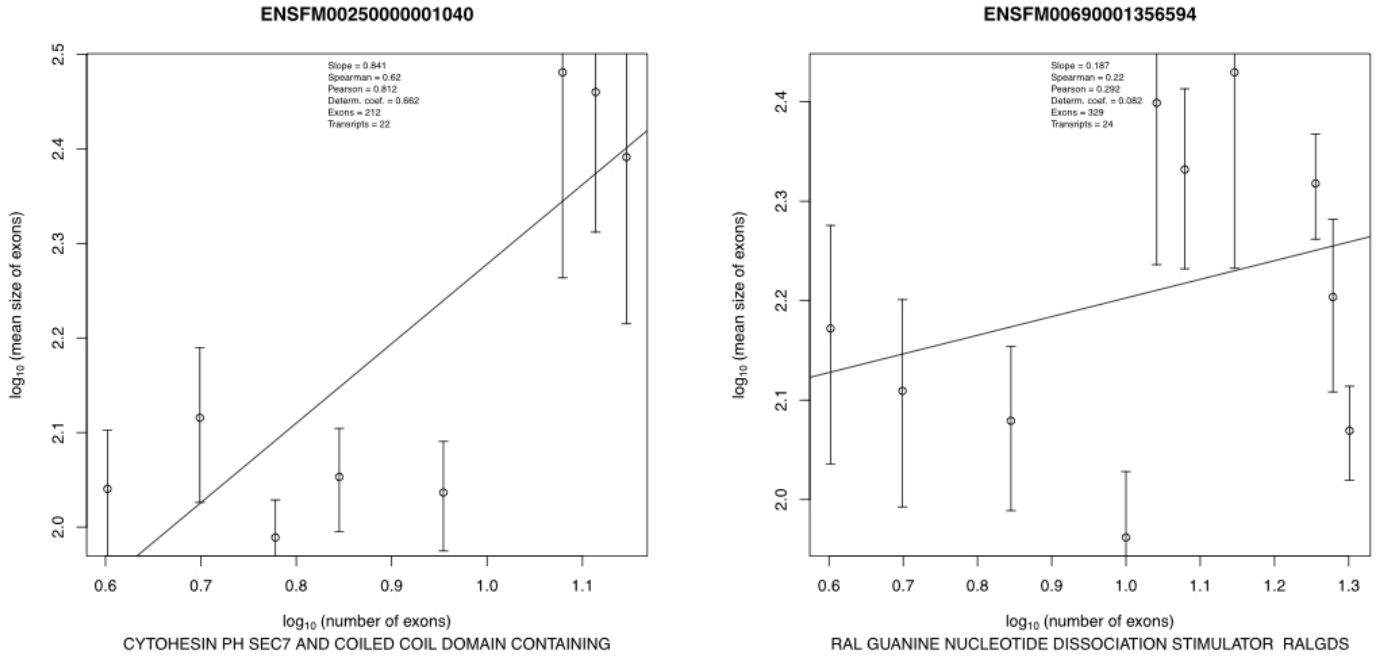
Γναρδέλλης, Χ. (2003). *Εφαρμοσμένη Στατιστική*. Αθήνα: Εκδόσει Παπαζήση.

6 Παραρτήματα

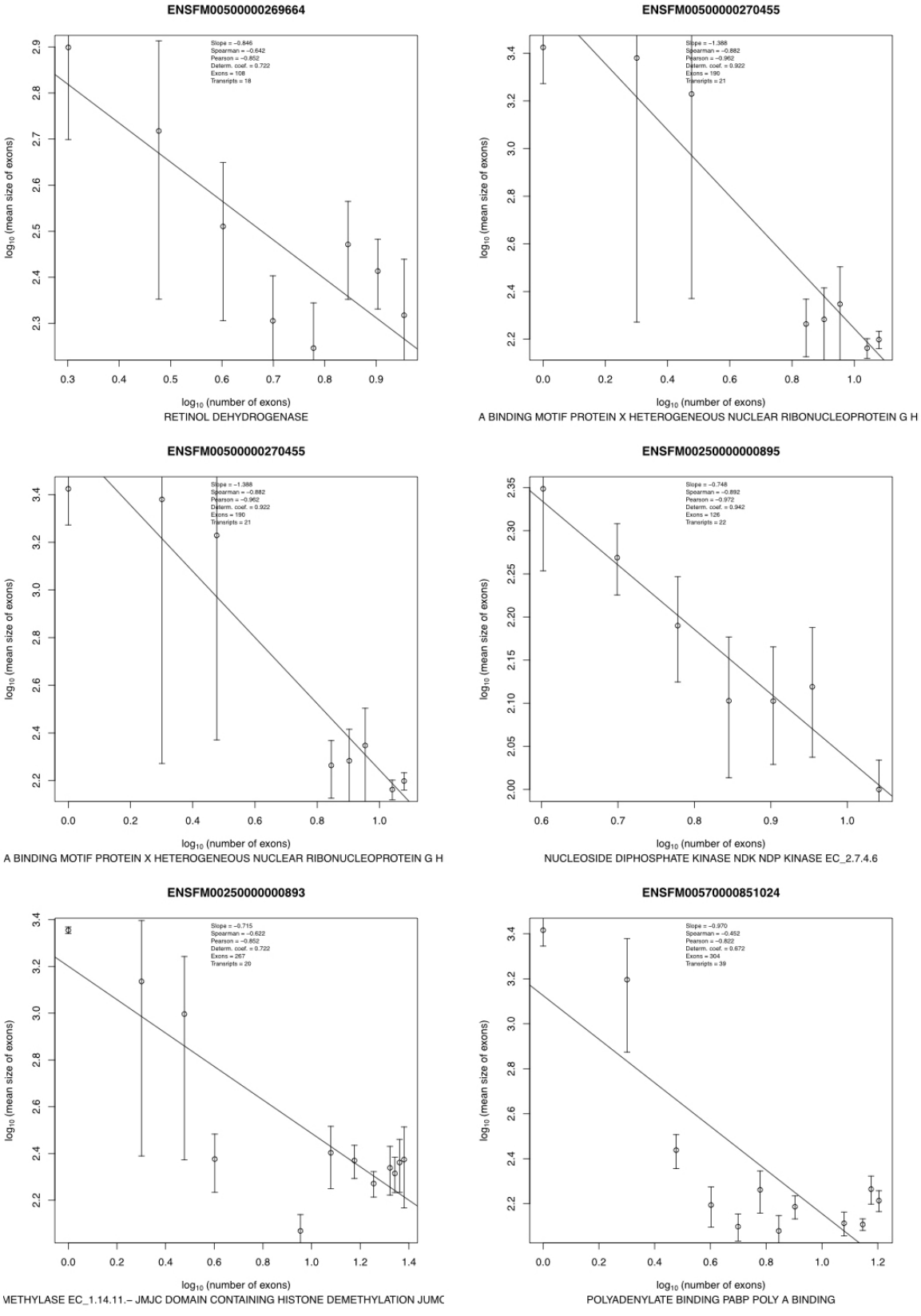
6.1 Παράρτημα Α (διαγράμματα)



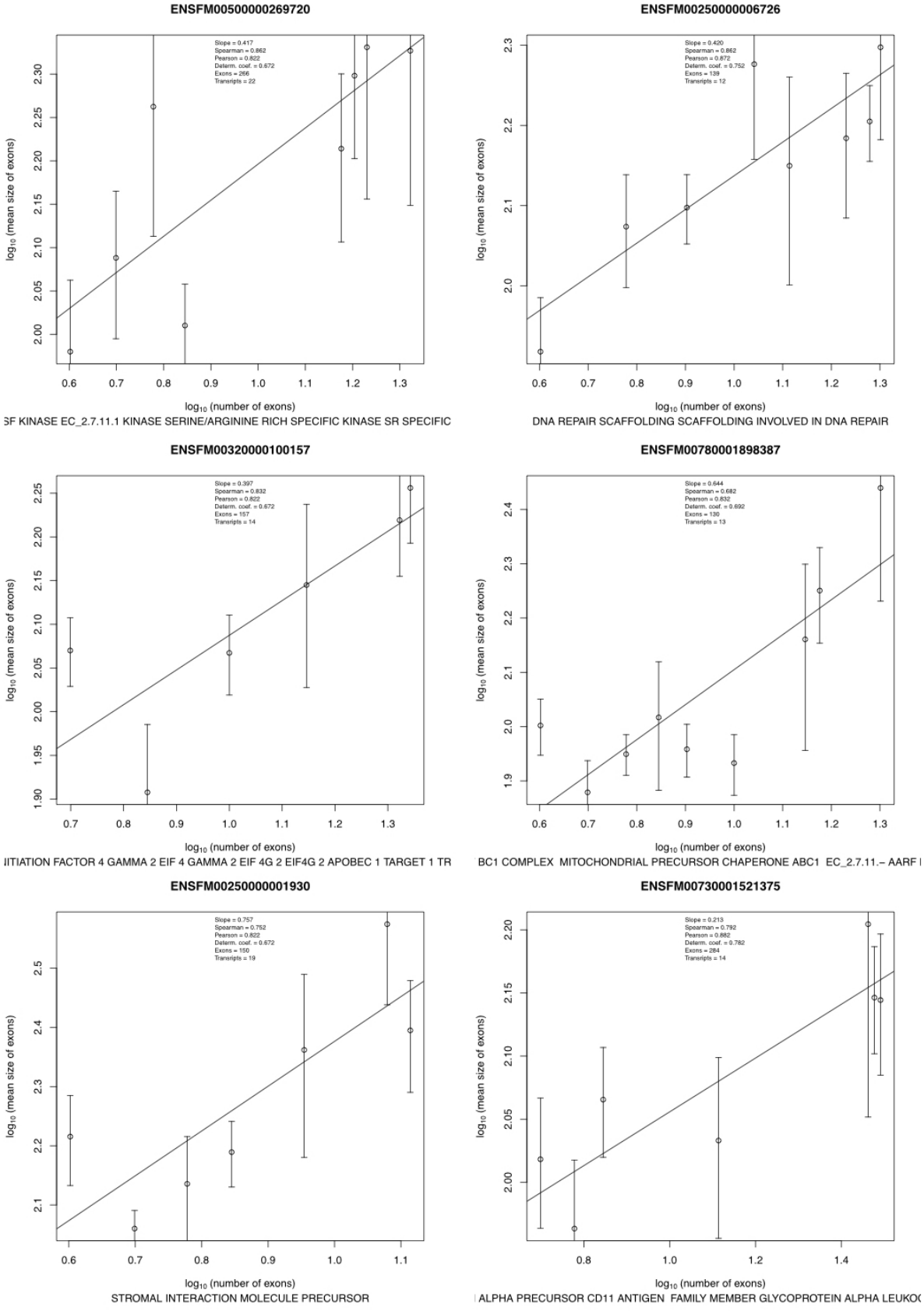
Εικόνα 8: Οικογένειες γονιδίων που δεν διακρίνεται ισχυρή συσχέτιση του μέσου μεγέθους των εξονίων με τον αριθμό τους. Παρόλα αυτά υπάρχει γραμμικότητα σε διπλή λογαριθμική κλίμακα. Κάθε σημείο αντιπροσωπεύει το σύνολο των μεταγράφων που έχουν τον ίδιο αριθμό εξονίων. Τα error bars παρουσιάζουν το τυπικό σφάλμα της μέσης τιμής.



Εικόνα 9: Οικογένειες γονιδίων που εμφανίζουν ξεχωριστές ομάδες μεταγράφων στην κατανομή τους με βάση τον αριθμό εξονίων. Κάθε ομάδα έχει τη δική της ξεχωριστή κατανομή άρα ίσως και δική της ιστορία. Αυτό ισχύει στις πρωτεΐνες που αποτελούνται από διαφορετικές υπομονάδες καθώς αυτές ενώθηκαν κάποια στιγμή στην εξέλιξη σχηματίζοντας ένα υβρίδιο. Κάθε σημείο αντιπροσωπεύει το σύνολο των μεταγράφων που έχουν τον ίδιο αριθμό εξονίων. Τα error bars παρουσιάζουν το τυπικό σφάλμα της μέσης τιμής.



Εικόνα 10: Οικογένειες γονιδίων οι οποίες ακολουθούν την κατανομή του νόμου του Menzerath στα επίπεδα μεταγράφου – εξονίου. Κάθε σημείο αντιπροσωπεύει το σύνολο των μεταγράφων που έχουν τον ίδιο αριθμό εξονίων. Τα error bars παρουσιάζουν το τυπικό σφάλμα της μέσης τιμής.



Εικόνα 11: Οικογένειες γονιδίων στις οποίες το μέσο μέγεθος εξονίων αυξάνει όσο αυξάνονται τα εξόνια των μεταγράφων. Κάθε σημείο αντιπροσωπεύει το σύνολο των μεταγράφων που έχουν τον ίδιο αριθμό εξονίων. Τα error bars παρουσιάζουν το τυπικό σφάλμα της μέσης τιμής.

6.2 Παράρτημα Β (κώδικας)

α) Συγχώνευση 2 αρχείων κειμένου με 1 κοινή στήλη στην Perl

```
#! /usr/bin/perl
use warnings;
use strict;

open EXA1, "</biomart Data exons ids april 2015.txt";
open EXA2, "</uniq_biomart_families.txt";
open OUT, ">merge_biomart_exon_family.txt";

my %hash = ();
my @coordinates = <EXA1>;
my @ids = <EXA2>;

foreach (@coordinates)
{
   chomp;
   my ($char, $start, $end, $rank, $tr_count, $exon_id, $transcript_id, $gene_id)=split("\t", $_);
   my $value = "$gene_id\t$char\t$start\t$end\t$rank\t$tr_count";
   my $pair = "$exon_id\t$transcript_id"; #monadikotita grammis.
   $hash{$pair} = $value; # topothetish stoixeion sto hash me kleidia stili 0. An kapoio stoixeio kleidiou iparxei 1< fores to kleidi apodidei
perissoteres times.
}
foreach (@ids)
{
   chomp;
   my ($gene_id2, $transcript_id2, $exon_id2, $family_id, $family_discription)=split("\t", $_);

   my $pair = "$exon_id2\t$transcript_id2";
   if (exists $hash{$pair})
   {
      print OUT $exon_id2, "\t",$transcript_id2, "\t",$hash{$pair},"\t", $family_id, "\t", $family_discription, "\n";
   }
   else
   {
      print "$gene_id2\n";
   }

}
#-----------------------------------------TELOS---------------------------------------#
```

β) Υπολογισμοί εξονίων, μεταγράφων, μέσου μέγεθους εξονίων για κάθε οικογένεια στην Perl

```
# script gia upologismo exonion, metagrafon, gonidion, kai mesou megethous exonion gia kathe oikogeneia tou anthropou. Dedomena apo
ENSEMBL. Gia ton antropo to script thelei peripou 90 lepta gia na trexei.

#! /usr/bin/perl

use strict;
use warnings;

open DATA, "</final_merge_biomart_exon_transcript_6_2015.txt";
open OUT, ">3_6_15_exons_transcripts_genes_per_family.txt";
my @rawdata = <DATA>;
my %families=();
my @fam_data=();

foreach (@rawdata) #creation of hash with keys the family IDs.
{
   chomp;
   my @array = split("\t", $_);
   if (scalar(@array) ==10)
   {
      $families{$array[8]} = $array[9];
   }

}

my $i=0;
for my $family (sort keys(%families)) # in this loop there is the calculations and plot export. for each family an array is created which contains all the
lines that include this family ID.
```

```
{
   foreach (@rawdata)
   {
      chomp;
      if ($_=~ /$family/)
      {
         push(@fam_data, $_); # gia kathe grammi tou arxeiou topothetountai sto arxeio mono autes pou exoun to kleidi (family ID). Etsi gemizei to
array me dedomena 1 mono oikogeneias. Sto telos ths loupas adeiazetai to array kai xanagemizei me to kainourgio kleidi.Stis grammes autou tou
array ginontai oles oi katametriseis tis oikogeneias
      }
   }

   ### pairno ta fam_data gia na ginou ta epomena
   my %sumsize=();
   my $countexon=0;
   my %avg=();
   my %standsum=();
   my %stderror=();
   my %binsums=();
   my %count_transcripts=();
   my %count_genes=();
   my %fam_number_transcripts=();
   my $exon_count_verification;
   my $total_exon_size=0;

   ### thelo na metrithoun ola ta exonia!! kathe transcript 1 fora.

   foreach (@fam_data) #mpainoun ta dedomena. meta upologismos mean megethous ton exonion kai arithmos tous gia kathe transcript ID
   {
      chomp;
      my @array = split("\t", $_);

      $total_exon_size += ($array[5] - $array[4]); # megethos(value) kathe exoniou(key)
      $countexon++; #  arithmos pou emfanizontai ta exonia
      $count_transcripts{$array[1]}++; # arithmos pou emfanizetai kathe transcripts(value)=exons pou vriskontai sta dedomena pou trexei h loupa
      $count_genes{$array[2]}++; #katametrisi gonidion

   }
   for my $transcript (sort keys(%count_transcripts))
   {
      $exon_count_verification += $count_transcripts{$transcript};
   }

   my $avg= $total_exon_size/$countexon;

   print OUT $family,"\t", $countexon, "\t", scalar(keys(%count_transcripts)),"\t",scalar(keys(%count_genes)),"\t", $avg,"\t", $families{$family},"\n";

   my $leftfam = scalar(keys(%families))-$i; #katametrisi poson oikogeneion apomenoun.
   $i++;
   print $leftfam,"\n";
   @fam_data=(); # adeiasma tou array.IMPORTANT!!!!!!!!!!
}
#-----------------------------------------TELOS-----------------------------------#
```

γ) Υπολογισμοί (Perl) και δημιουργία ιστογράμματος (R programming) για κάθε οικογένεια

```
#! /usr/bin/perl

use strict;
use warnings;
use Statistics::R;

open DATA, "</Users/mac/CG2_Lab/workspace/uniq_biomart_merged.txt";
open FAM, "</Users/mac/CG2_Lab/workspace/code_and_results/more100exons10transcripts.txt";
my @fam_selection = <FAM>;
my @rawdata = <DATA>;
my %families=();
my %families_description=();
my %family_exons=();
my %family_transcripts=();
my %transcript_number=();
my @fam_data=();
our $family;
```

```
foreach (@fam_selection) #creation of hash with keys the family IDs.
{
   chomp;
   my @array = split("\t", $_);
   $families{$array[0]} = $array[1];
   $transcript_number{$array[0]} = $array[2];
   $families_description{$array[0]} = $array[4];
   $family_exons{$array[0]}=$array[1];
   $family_transcripts{$array[0]}=$array[2];
   }

for $family (sort keys(%families)) # in this loop there is the calculations and plot export. for each family an array is created which contains all the lines that include this family ID.
{
   #print $_,"\n";
   if (($families{$family} > 100) && ($transcript_number{$family} >= 10) ) # epilogh oikogeneion me vasi ton arithmo exon pou exoun KAI ton arithmo exonion.
   {

      foreach my $line (@rawdata)
      {
         chomp($line);
         if ($line =~ /$family/)
         {
            push(@fam_data, $line); # the array with lines that include the key(family ID). With each iteration the array is empted and new lines are imported for the specific key. This array is the source for the rest of the program for the calculations and plotting
            #print "$_%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%\n";
            #print "nai\n";
         }
      }

      MEAN_SIZE_STDERROR(@fam_data);

      R_CALC_PLOT($family, \%families_description, \%family_exons, \%family_transcripts);
   }
   @fam_data=(); # ADEIASMA ARRAY!!!!
}

exit;
##########-----------------------------------------END OF PROGRAM------------------------------##############

sub R_CALC_PLOT # regression analysis and plot in R for every family
{
   ####----------------------------------enter R for plots------------------------------#######

   open STATISTIC, ">>family_slope_spearman.txt";
   my ($family_name, $family_description, $family_exons, $family_transcripts) = @_;
   my $R = Statistics::R->new();

   $R -> run (q'mydata <- read.table("bashfileforR.txt", header = TRUE)'); # use of temporary file in R
   $R->run(qq`pdf("$family.pdf")`);
   $R->run(
   q 'fit <- lm(log10(mydata$mean_size_of_exons)~log10(mydata$exon_count))', # regression analysis
   q 'plot(log10(mydata$exon_count),log10(mydata$mean_size_of_exons), pch=1, main="", xlab="", ylab="" ,sub="")', # dimiourgia plot me dipli logarithmiki klimaka ton metavlitvn
   qq 'title(main = "$family_name",sub = "${$family_description}{$family_name}", xlab=bquote(log[10]~"(number of exons)"), ylab=bquote(log[10]~"(mean size of exons)"))', # titloi kai axones
   q'arrows(log10(mydata$exon_count),log10(mydata$mean_size_of_exons-mydata$standard_error), log10(mydata$exon_count), log10(mydata$mean_size_of_exons+mydata$standard_error), length=0.05, angle=90, code=3)', # topothetish error bars. einai ena mikro hack pou xrisimopoiountai arrows.
   q'abline(coef=coef(fit))', # eutheia grammikis pallidromisis
   q'calcslope <- coef(fit)[2]', # klisi eutheias
   q'calcspearman <-cor(log10(mydata$exon_count),log10(mydata$mean_size_of_exons), method = "spearman")', # upologismos Spearman coefficient
   q'calcpearson <-cor(log10(mydata$exon_count),log10(mydata$mean_size_of_exons), method = "pearson")', #upologismos Pearson coefficient
   q'slope <- paste0("Slope = ",format(round((calcslope), 3), nsmall = 3))', # ektiposi
      q'spearman <- paste0("Spearman = ", round(calcspearman,2), nsmall=2)', # ektiposi
      q'pearson <- paste0("Pearson = ", round(calcpearson,2), nsmall=2)',# ektiposi
      qq'exons <- paste0("Exons = ","${$family_exons}{$family_name}")',# ektiposi
      qq'transcripts <- paste0("Transripts = ", "${$family_transcripts}{$family_name}")',# ektiposi
      q'legend("top", legend=c(slope, spearman,pearson, exons, transcripts),col = "black", cex = .6)');# topothetish ton prohgoumenon ektiposeon sto upomnhma tou plot

   my $spearman = $R -> get('calcspearman'); #krataei metavliti
   my $pearson = $R -> get('calcpearson');
   my $slope = $R -> get('calcslope');
```

```
    $R->run(q`dev.off()`);

    $R->stop();

    print STATISTIC $family, "\t", $spearman, "\t", $pearson, "\t", $slope, "\n"; # ektiposi gia kathe oikogeneia ta statistika pou tin xaraktirizoun

    }

sub MEAN_SIZE_STDERROR
{
    #----------------EXON MEAN SIZE PER NUMBER OF EXONS calculations from raw data ----------------------------###
    ### pairno ta fam_data gia na ginou ta epomena
    my %sumsize=();
    my %countexon=();
    my %avg=();
    my %standsum=();
    my %stderror=();
    my %binsums=();
    my %count_transcripts=();
    my %families=();
    my @fam_data=@_;

    foreach (@fam_data) #mpainoun ta dedomena. meta upologismos mean megethous ton exonion kai arithmos tous gia kathe transcript ID
    {
        chomp;
        my @array = split("\t", $_);
        my $exonsize = ($array[5] - $array[4]); # megethos kathe exoniou
        $sumsize{$array[1]} += $exonsize; # athrisma megethous exonion
        $countexon{$array[1]}++; # arithmos exonion ana transcript
        #printf $array[0],"\t", $array[1],"\t", $array[2],"\t", $array[3], "\n";
    }

    for (sort keys (%countexon))
    {
        $binsums{$countexon{$_}} += $sumsize{$_}; #athrisma ton megethon ton exonion olon ton transcripts pou exount ton idio arithmo exonion
        $count_transcripts{$countexon{$_}}++; #sunolo transcripts pou exoun ton idio arithmo exons.

    }

    for (sort keys %binsums)
    {
        $avg{$_} = $binsums{$_}/($_*$count_transcripts{$_}); #upologismos mean megethous ton exonion gia ola ta transcript ID pou exoun ton idio
arutmo exons.
    }

    foreach (@fam_data){ # gia kathe grammi tou arxeiou (ola ta exons) upologizetai to arhroisma toy standard deviation
        chomp;
        my @array = split("\t", $_);
        my $exonsize = ($array[5] - $array[4]);
        $standsum{$countexon{$array[1]}} += ($exonsize - $avg{$countexon{$array[1]}})**2; # dhmiourgia tou athrismatos gia ton tipo tou standard
deviation. gia kathe timi transcript dinetai i timi tou hash countexon i opoia apotelei kleidi tou hash standsum.
    }

    open OUT, ">bashfileforR.txt";

    print OUT "exon_count\tmean_size_of_exons\tstandard_error\tstandard_deviation\tnumber_of_transcripts\ttotal_exons\n";

    for (sort keys(%standsum)) # here the calculations of mean and standard error for the size of exons per exon number are printed in a file. This file is
overwritten for each family ID.
    {
        my $stddev;
        if ($_*$count_transcripts{$_} == 1){
            $stddev = ($standsum{$_}/(($_*$count_transcripts{$_})))**0.5; # to evala LEIPEI to -1 !!!!!!!!!!!!!!!!!!!!! An se mia oikogeneia uparxei 1
mono transcript me 1 exonio tote to sunolo ton exonion gia bin=1 einai 1.
        }
        else {
            $stddev = ($standsum{$_}/(($_*$count_transcripts{$_})-1))**0.5; #
        }
        $stderror{$_} = $stddev/($_*$count_transcripts{$_})**0.5;
        my $exons = $_*$count_transcripts{$_};
        print OUT "$_\t$avg{$_}\t$stderror{$_}\t$stddev\t$count_transcripts{$_}\t$exons\n";
    }
}
```