Ημερολόγιο Πτυχιακής

Thursday, 19March, 2015 12:35 PM

1. Πώς θα ενωθούν 2 αρχεία κειμένου με στήλες που είναι χωρισμένες με \t όταν έχουν μια κοινή στήλη;;
2. Πώς θα χωριστούν τα Εξώνια με βάση την οικογένεια;;
   a. Μπορεί ένα εξώνιο να ανήκει σε 2 οικογένειες;;
3. Μέσος όρος μεγέθους εξωνίων ανά μετάγραφο
   a. Sum(End - start)/n εξωνίων
   b. Ποιο είναι το n;; Το μέγιστο του rank ανά μετάγραφο.
4. Δημιουργία διαγράμματος στην R
5. Πώς ομαδοποιώ δεδομένα με βάση μια τιμή μιας στήλης;;

Thursday, 19March, 2015
12:25 PM

Δεδομένα από biomart:

1. Ένα gene ID μπορεί να ανήκει σε περισσότερες από 1 οικογένειες;;
   a. Πώς θα γίνει έλεγχος;
2. Πώς είναι κατασκευασμένες οι οικογένειες της ensembl;;;
3. Πώς δημιουργήθηκε η βάση δεδομένων με όλες τις αλληλουχίες και τις συντεταγμένες;;
4. Πόσα γονίδια είναι αντιστοιχημένα σε οικογένειες; Πόσα όχι και γιατί;;
5. Τα Gene ID & Exon ID στα αρχεία δεν είναι μοναδικά. Γιατί όχι στο Exon ID;;

23/3/15

Στις στήλες των αρχείων δεν υπάρχει τίποτα μοναδιαίο.. Υπάρχει πρόβλημα με το hash???? Χρειάζομαι hash για να ενώσω τα 2 αρχεία. .= κρατάω κάθε τιμή που πάει σε key και προσθέτω την άλλη σαν επιπλέον value.

Keys in a hash must be unique. If keys in file1 are unique, use file1 to create the hash. If keys are not unique in either file, you have to use a more complicated data structure: hash of arrays, i.e. store several values at each unique key
http://docstore.mik.ua/orelly/perl2/prog/ch09_02.htm

21/4/15

http://stackoverflow.com/questions/10154424/making-arrays-from-tab-delimited-text-file-column
Εκτύπωση των στοιχείων για κάθε κλειδί hash, θα μου χρειαστεί για το σε πόσες οικογένειες ανήκει κάθε γονίδιο.

Ένωση αρχείων: http://stackoverflow.com/questions/12377892/merging-two-files-based-on-first-column-and-returns-multiple-values-for-each-key

http://stackoverflow.com/questions/18861782/perl-multiple-keys-and-merging-two-files?rq=1

Εκτύπωση hash of arrays http://docstore.mik.ua/orelly/perl2/prog/ch09_02.htm

23/4/15

Κλειδί κάθε οικογένεια	Τα μετάγραφα που αντιστοιχούν	Και για κάθε μετάγραφο το μέσο μήκος εξωνίων και αριθμό εξωνίων

Πόσες οικογένειες έχω;
Πόσα μετάγραφα ανά οικογένεια;

27/4/15
Ένωση των αρχείων από τη biomart με τη χρήση της R. Υπήρχε πρόβλημα με την το perl script
Καθώς το αρχείο εμφανιζόταν να έχει 20γίγα μέγεθος.. Δεν ξέρω γιατί.
ο κώδικας

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

open EXA1, "< $ARGV[0]";
open EXA2, "< $ARGV[1]";
open OUT, "> /Users/mac/Desktop/output.txt";
```

```
#@array1 = <EXA1>;
#@array2 = <EXA2>;
%hash = ();

while (<EXA1>)
{
    chomp;
    my ($char, $start, $end, $rank, $tr_count, $exon_id, $transcript_id,
$gene_id)=split("\t", $_);
    $value = "$char\t$start\t$end\t$rank\t$tr_count\t$exon_id\t
$transcript_id";

    push @{$hash{$gene_id}}, $value; # topothetish stoixeion sto hash me
kleidia stili 0. An kapoio stoixeio kleidiou iparxei 1< fores to kleidi
apodidei perissoteres times.
}
while (<EXA2>)
{
    chomp;
    my ($gene_id2, $transcript_id2, $exon_id2, $family_id,
$family_discription)=split("\t", $_);

    if (exists $hash{$gene_id2})
    {
        foreach my $value1 (@{$hash{$gene_id2}})
        {
            print OUT $value1,"\t",$transcript_id2, "\t", $family_id, "\t",
$family_discription, "\n";
        }

            }
    else
    {
        print "$gene_id2\n";
    }

}
```

Έτσι χρησιμοποίησα την R.

merge(biomart.Data.exons.ids.april.2015,biomart.families.and.ids.2015)

R programming

write.table(filesmerge, file="merge2.txt", sep="\t", quote=FALSE, append = FALSE, row.names=FALSE)

Όνομα αρχείου:

Επίσης χρησιμοποιήθηκαν 2 αρχεία που περιέχουν

1. Όλες τις οικογένειες και τα transcripts που αντιστοιχούν σε κάθε οικογένεια. Με χρήση Perl
    a. Το αρχείο που σχηματίστηκε έχει στην πρώτη στήλη την οικογένεια και έπειτα σε tab separated στήλες όλα τα transcripts ανήκουν στην οικογένεια αυτή

οικογένεια αυτή.

b.

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

open (PAME, "<$ARGV[0]") or die "Can't open 'PAME': $!";
open OUT, "> $ARGV[1]";

foreach (<PAME>)
{
    chomp;
    my ($geneid, $transcript, $exon, $proteinid,
$fam_discription) = split("\t", $_);
    $hash{$proteinid} .= $transcript."\t";

}
foreach $key (keys (%hash))
{
    $hash{$key} =~ s/\t$//g ;
    print OUT $key, "\t", $hash{$key}, "\n";
}
```

```
ENSFM00610000967374     ENST00000525132
        ENST00000525132     ENST00000525132
        ENST00000525132     ENST00000525132
        ENST00000525132     ENST00000525132
        ENST00000525132     ENST00000525132
        ENST00000525132
ENSFM00580000910271     ENST00000471390
        ENST00000471390     ENST00000471390
        ENST00000471390     ENST00000471390
        ENST00000471390     ENST00000471390
        ENST00000471390     ENST00000471390
        ENST00000471390
```

2. Όλα τα γονίδια και τις οικογένειες που ανήκει κάθε γονίδιο. Ομοίως με πριν

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

open (PAME, "<$ARGV[0]") or die "Can't open 'PAME': $!";
open OUT, "> $ARGV[1]";

foreach (<PAME>)
{
    chomp;
    my ($geneid, $transcript, $exon, $proteinid,
$fam_discription) = split("\t", $_);
    $hash{$geneid} .= $proteinid."\t";

}
foreach $key (keys (%hash))
```

```
foreach $key (keys (%hash))
{
    $hash{$key} =~ s/\t$//g ;
    print OUT $key, "\t", $hash{$key}, "\n";
}
```

Πόσες οικογένειες ανήκει κάθε γονίδιο;

Από το αρχείο που προέκυψε αρκεί να μετρήσουμε πόσες στήλες υπάρχουν ανά γραμμή (γονίδιο) -1 (που είναι η στήλη που ανήκει το γονίδιο).

Στην awk: awk '{FS="\t"}; {print $1, "\t", NF-1}'

Το κάνουμε αυτό και για τα 2 αρχεία.

Για τα διαγράμματα

Για κάθε οικογένεια

Για κάθε transcript

Αριθμός εξονίων

Μέγεθος κάθε εξονίου (END-START)

Μέσο μέγεθος εξονίων (Sum(μέγεθος εξονίων)/αριθμός εξονίων στο transcript)

28/4/15

Το transcript count αφορά τον αριθμό των διαφορετικών transcrpts που υπάρχουν για ένα συγκεκριμένο gene ID.

Μέγεθος εξονίου στην R.
http://stackoverflow.com/questions/23888037/subtracting-two-columns-to-give-a-new-column-in-r

Για την ανάλυση και την δημιουργία διαγράμματος Αριθμός εξονίων/μέσο μέγεθος εξονίων για κάθε transcript της κάθε οικογένειας. Θέλουμε αρχείο:

Protein_fam	Transcript_ID	Mean_exon_size	Number_of_exons	Fam_discription

ΠΡΟΣΟΧΗ! Κάποιες γραμμές είναι στο αρχείο είναι διπλές, ολόκληρες γραμμές, αυτό δημιουργεί πρόβλημα στον υπολογισμό μέσης τιμής μεγέθους εξονίων. Γιατί αν υπολογίσουμε το μέγεθος κάθε εξονίου και μετά για το συγκεκριμένο transcript id θέλουμε το μέσο μέγεθος αυτό που κάνουμε είναι να αθροίζουμε όλα τα μεγέθη που αντιστοιχούν στο συγκεκριμένο ID και διαιρούμε με το πλήθος τους. Οι επαναλαμβανόμενες γραμμές εμφανίστηκαν μετά τη συγχώνευση αρχείων.

Αρχείο με εξόνια	Αρχείο με οικογένειες	Αρχείο συγχώνευσης
1171176	1171395	1171395

Μοναδικά στο συγχώνευσης		
1005982		

	Exons (not unique from ensemble)	families	Merged with R	Merged with Perl(exon based)	Families with unique results from ensembl	
Wc	1171176	1171395	1171395	1171395	1005982	
dublicates	0	165238	165238	231552	0	
Unique values		1005982		1005982		
Sum of values of occurrence		1171235				

Τα μοναδικά του families είναι ίδιο με τα μοναδικά του merged.
4/5/15

Script στο bash για να κάνω τα εξής για 1 αρχείο :

1. Αριθμός γραμμών
2. Εύρεση των φορών που υπάρχει κάθε γραμμή
3. Αριθμός μοναδικών γραμμών
4. Έλεγχος εάν το άθροισμα των φορών που εμφανίζεται η κάθε γραμμή είναι ίσο με τον αριθμό των γραμμών (κανονικά πρέπει να είναι)

Calculate Mean
http://www.perlmonks.org/?node_id=851634

```
use strict;
use warnings;
my %data;
while (<DATA>) {
   chomp;
   my ($id, $value) = split /,/;
   $data{$id}{sum} += $value;
   $data{$id}{count}++;
}
for my $id (sort keys %data) {
   my $avg = $data{$id}{sum}/$data{$id}{count};
   print "$id: sum=$data{$id}{sum} avg=$avg\n";
}
```

__DATA__
A,10
A,11
A,12
A,13
B,15
B,16
C,17
D,18
prints:
A: sum=46 avg=11.5
B: sum=31 avg=15.5
C: sum=17 avg=17
D: sum=18 avg=18

ΠΡΟΣΟΧΗ. Για να χρησιμοποιηθεί αυτός ο κώδικας πρέπει σε κάθε transcript_ID να υπάρχουν διαφορετικά exon_Ids, δηλαδή να μην υπάρχουν διπλασιασμένες τιμές!!!
Αν υπάρχουν γραμμές με ίδιο exon & tanscript ID σημαίνει ότι είναι ίδιες οι γραμμές καθώς το exon id είναι η πιο συγκεκριμένη πληροφορία. Με το uniq στο unix θα φύγει.

7/5/15
Έτοιμο το mean και standard error για κάθε transcript ιδ. Τώρα πρέπει να γίνει data binning ώστε να κάνω ομάδες με κριτήριο τον αριθμό εξονίων.

8/5/15

Στόχος να έχω 1 script perl που να παίρνει τα δεδομένα 1 αρχείο και μετά από επιλογή οικογένειας (ή άλλο τμήμα δεδομένων) από το χρήστη να γίνεται εξαγωγή γραφήματος από την R.
Άρα χρειάζεται στο perl script να υπάρχουν οι εντολές της R για να βγει το τελικό διάγραμμα.
Χρειάζεται 1 εντολή για να εισάγεται ο κώδικας της R αλλά θα πρέπει να παίρνει η R τα δεδομένα από την Perl.

12/5/15

Υπάρχουν 2 τρόποι να ελέγχω την R από την perl.

1. Γράφω ένα αρχείο με κώδικα της R και μέσα στον κώδικα της perl καλώ το bash με τη system .

system("R CMD BATCH test.R");

Explanation: Execution of R in Batch mode, see R CMD BATCH --help

Usage: R CMD BATCH [options] infile [outfile]

Run R non-interactively with input from infile and place output (stdout and stderr) to another file. If not given, the name of the output file is the one of the input file, with a possible '.R' extension stripped, and '.Rout' appended.).

Το πρόβλημα είναι: πώς θα μεταφέρω ένα object από την perl στην R;;; Ίσως πρώτα να μεταφερθεί στο bash και από κει στην R.

2. Install module Statistics::R οπότε θα γραφτεί ο κώδικας της r μέσα στο script της περλ.
   http://search.cpan.org/~fangly/Statistics-R/lib/Statistics/R.pm

   Οδηγίες χρήσης http://search.cpan.org/~fangly/Statistics-R/lib/Statistics/R.pm

Εγκατάσταση εργαλείων της perl, (perlblew, cpamn, local::lib) -> http://blog.jambura.com/2013/02/19/setup-homebrew-perlbrew-ruby-rvm-perl-cpanm-nginx-in-mountain-lion/

Γραφήματα με R : http://stackoverflow.com/questions/13032777/scatter-plot-with-error-bars

http://www.cookbook-r.com/Graphs/Plotting_means_and_error_bars_(ggplot2)/

http://egret.psychol.cam.ac.uk/statistics/R/graphs2.html

17/5/15

Statistics::R docs recommend quoting with q`...` (if you aren't interpolating Perl variables) or qq`...` (if you are interpolating). Those are backticks, by the way (but, really, any character not used in typical R code should suffice). The reason for this is

really, any character not used in typical R code should suffice). The reason for this is that if you have single or double quotes in your R commands, you won't run into troubles. Another thing to watch out for is if you are wrapping your R code in double quotes because you need to interpolate Perl variables and are using $ in your R code to specify a column of data, for example, you need to escape it

Πρόβλημα με τα plots. Όταν μπαίνει το log10 στο plot τα arrows δεν σχηματίζονται.

Αυτό λειτουργεί:

plot(data$exon_count,data$mean_size_of_exons, pch=1, col="red", col.axis="orange", xlab="log10(number of exons)", ylab="log10(mean size of exons)")
arrows(data$exon_count,data$mean_size_of_exons-data$standard_error, data $exon_count, data$mean_size_of_exons+data$standard_error, length=0.05, angle=90, code=3)
Μόλις γίνει log10(data$exon_count) τα αρροους δεν μπαίνουν.

Αυτό γίνεται γιατί δεν φαίνονται στην κλίμακα μετά το log.

Έτσι δουλεύει:

plot(log10(data$exon_count),log10(data$mean_size_of_exons), pch=1, col="red", col.axis="orange", xlab="log10(number of exons)", ylab="log10(mean size of exons)")
arrows(log10(data$exon_count),log10(data$mean_size_of_exons-data $standard_error), log10(data$exon_count), log10(data$mean_size_of_exons+data $standard_error), length=0.05, angle=90, code=3)

Επόμενο βήμα: μια λούπα που θα βγάζει κάθε οικογένεια σαν μεταβλητή (π.χ κλειδί hash) και μετά σε 2η λούπα θα εντοπίζεται σε ποιες γραμμές βρίσκεται αυτή η μεταβλητή στο αρχείο με τα δεδομένα και θα τοποθετεί ολόκληρες τις γραμμές σε array με στόχο αυτό να χρησιμοποιηθεί για υπολογισμούς και να γίνει γράφημα στην R μόνο με τις γραμμές που αντιστοιχούν στο συγκεκριμένο κλειδί.

18/5/15

Ενώθηκαν όλα τα κομμάτια του προγράμματος (υπολογισμοί, εντολές r, εμφάνιση δεδομένων κάθε οικογένειας). Το πρόβλημα βρίσκεται στο 3 κομμάτι.
Διαίρεση με 0....

```
awk '{FS="\t"}{family[$9] = $2;times[$9]++}END{for (i in family) {print i
"\t" times[i]}}' biomart\ exons\ families\ merged\ 2015.txt | sort -nk2
```

21/5/15
Δημιουργία αρχείου με υπολογισμούς για κάθε οικογένεια

Οικογένεια	Αριθμός εξονίων	Αριθμός μεταγράφων	Μέσο μέγεθος εξονίων	Περιγραφή οικογένειας

Υπολογίζουμε όλα τα εξόνια που έχει η οικογένεια! Όλα, ΟΧΙ μόνο τα μοναδικά. Το ζεύγος εξόνιο-μετάγραφο είναι μοναδικό.

Έπειτα θα χρησιμοποιηθεί το αρχείο αυτό για να επιλέξουμε οικογένειες που θα κάνουμε διάγραμμα.
Πρόγραμμα με subroutines.
Δημιουργία plots.

27/5/15

Έτοιμο το πρόγραμμα με subroutines.
Για το κλειδί (όνομα οικογένειας) το έκανα global variable με το our $family (http://stackoverflow.com/questions/2450055/why-cant-my-perl-subroutine-see-the-value-for-the-variable-in-the-foreach-loop)

2 υπορουτίνες: πράξεις και R plot

Πρέπει να ξανατρέξω τους υπολογισμούς για κάθε οικογένεια γιατί έχασα το αρχείο.

Βγάζω διαγράμματα. Τώρα πρέπει
1. Να κάνω cummulative analysis
2. Να βάλω στην R να κάνει lm (linear model)

28/5/15

1. Να βγάλω εικόνες για οικογένειες με περισσότερα από 100 εξόνια && 10 μετάγραφα
2. Να φτιάξω flow chart με τον αλγόριθμο για να φαίνονται οι πράξεις και να ξεκαθαρίσει το φαινόμενο inception.
3. Να γράψω μεθοδολογία

~~Μάλλον υπάρχει πρόβλημα με τις πράξεις~~
29/5/15
Τελικά δεν ήταν λάθος στις πράξεις.

Σύμφωνα με τον παρακάτω τύπο

$$stddev = \sqrt{\frac{\sum_{j=1}^{k}\sum_{i=1}^{n}\left(x_i - x_{mean}\right)}{kn - 1}}$$

Ο τύπος υπολογίζει την τυπική απόκλιση του μεγέθους των εξονίων (αριθμό βάσεων) για το σύνολο των μεταγράφων (k) που έχουν n αριθμό εξονίων. Στο πρόγραμμα συλλέγονται τα εξόνια κάθε οικογένειας.

Xi είναι κάθε εξόνιο,
n είναι το σύνολο των εξονίων κάθε μεταγράφου,
K είναι το σύνολο των μεταγράφων με n αριθμό εξονίων κάθε οικογένειας.

Οπότε στην περίπτωση που σε μια οικογένεια υπάρχει k=1 μοναδικό μετάγραφο με n=1 μόνο εξόνιο ο παρονομαστής γίνεται μηδέν επομένως πρέπει να φτιαχτούν 2 κατηγορίες με if statement για αν n=1 && k=1 ο παρονομαστής του stddev να μην έχει το -1.

Αφού έγινε αυτό επιλέχθηκαν για plot οι οικογένειες που έχουν περισσότερα από 100 εξώνια σε περισσότερα ή ίσα με 10 μετάγραφα.

```
$families{$family} > 100) && ($transcript_number{$family} >= 10)
 Αυτά αντιστοιχούν σε 1261 οικογένειες άρα 1261 διαγράμματα.
Το πρόγραμμα τρέχει σε 11 λεπτά.
```

Στα διαγράμματα μπήκε τίτλος το μητρώο της οικογένειας και υπότιτλος η περιγραφή της οικογένειας. Για να γίνει αυτό φτιάχτηκε ένα hash το οποίο έχει κλειδί τ μητρώο της οικογένειας και σαν τιμή την περιγραφή της. Το hash μπήκε στην υπορουτίνα του προγράμματος με reference \%hash. Και όταν μπήκε στην υπορουτίνα έγινε derefference ${$hash}{key}.
Η περιγραφή κόβεται σε μερικά διαγράμματα γιατί είναι μεγάλη. Πρέπει να βρω τρόπο να τη χωράει.
πρέπει

1. Να φτιάξω flow chart με τον αλγόριθμο για να φαίνονται οι πράξεις και να ξεκαθαρίσει το φαινόμενο inception.
2. Να γράψω μεθοδολογία

.. . . .

31/5/15
Για τα διαγράμματα της R.
http://stackoverflow.com/questions/7367138/text-wrap-for-plot-titles
http://stackoverflow.com/questions/12093927/wrapping-legend-text-to-fit-the-plot-window

Ίσως να σπάσω την περιγραφή οικογενειών και να χωρέσει.

1/6/15

Αν θέλουμε μεγαλύτερα data sets μπορούμε να χρησιμοποιήσουμε GO ή superfamilies.

2/6/15

Πίνακας ανάλυσης δεδομένων
Εύρεση ζεύγους εξονίων, μεταγράφων. Κάθε ζεύγος είναι μοναδικό

```
awk '{$uniq=$1"\t"$2;exon[$uniq]++}END{for (i in exon) print
i"\t"exon[i]}' uniq_biomart_merged.txt | wc
```

http://tuxgraphics.org/~guido/scripts/awk-one-liner.html

Οικογένειες γονιδίων

1. Πόσες είναι
2. Πόσα εξόνια έχουν
3. Πόσα μετάγραφα
4. Πόσα γονίδια
5. Πόσα εξόνια κατά μέσο όρο
6. Πόσα μετάγραφα κατά μέσο όρο
7. Τοις % οικογενειών με 1 εξόνιο
8. Τοις % 2-10, 10-100, 100 - 1000, 1000 < εξόνια

Το αρχείο only_data_with_families περιέχει όλες τις γραμμές που έχουν οικογένειες. Σύνολο γραμμών `713919`. Όσα και τα εξόνια των οικογενειών από το αρχείο `correct_exons_per_family.txt`

Πώς κάνουμε binning δεδομένων στην R????

3/6/15

```
Savass-MacBook-Pro:workspace mac$ awk '{FS="\t"}{if ($9) print}'
merge_biomart_exon_transcript_6_2015.txt | wc
  831425 14794757 130489459
Savass-MacBook-Pro:workspace mac$ awk '{FS="\t"}{if ($9) print}'
uniq_biomart_merged.txt | wc
  713919 12660761 111751596
```

Πρόβλημα με το ενωμένο αρχείο!!!! Έχω χάσει 120000 δεδομένα που έχουν οικογένειες.
Κανένα πρόβλημα τελικά.

Το ξαναέκανα με το ζεύγος exon ID, transcript ID και βγήκε ακριβώς το ίδιο.

10/6/15

Έχουμε 15 εξόνια (14 ανήκουν σε οικογένειες) με μήκος 0 nt
Αυτό είναι πρόβλημα του annotation

17/6/15

Να επιλέξω τις υπόλοιπες οικογένειες και να κάνω πίνακα με χαρακτηριστικά τους

19/6/15

```
x <- binning_exons_per_family$number.of.exons
y <- binning_exons_per_family$number.of.families
linear <-lm(log10(y)~log10(x))
plot(log10(x), log10(y), pch=19, xlab="log10(exon_count)",
ylab="log10(mean_size_of_exons)")
abline(coef=coef(linear))
correlation <-cor(log10(x),log10(y), method = "spearman")

slope <- paste0("Slope = ",format(round((coef(linear)[2]), 4), nsmall = 4))
spearman <- paste0("Spearman coef= ", round(correlation,2), nsmall=2)

legend("topright", legend=c(slope, spearman),col = "black", cex = .6)
```

Έβαλα να κάνει γραμμικό μοντέλο (ελάχιστα τετράγωνα) για όλες τις οικογένειες

```
$R->run(qq`pdf("$family.pdf")`);
```

```
    $R->run(
        q 'fit <- lm(log10(mydata$mean_size_of_exons)~log10(mydata
$exon_count))',
        q 'plot(log10(mydata$exon_count),log10(mydata$mean_size_of_exons),
pch=1, main="", xlab="", ylab="" ,sub="")',
        qq 'title(main = "$family_name",sub = "${$family_description}
{$family_name}", xlab="log10(number of exons)", ylab="log10(mean size of
exons)")',
        q'arrows(log10(mydata$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)',
        q'abline(coef=coef(fit))',
        q'slope <- format(round((coef(fit)[2]), 4), nsmall = 4)',
        q'legend("topright", paste0("Slope = ",slope), col = "black", cex
= .6)');

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

    $R->stop();
```

Φτιάχτηκε αρχείο με κάθε οικογένεια και το slope και spearman της. Ίσως να φανεί χρήσιμο για το τι υπάρχει

Spearman
Min. 1st Qu. Median Mean 3rd Qu. Max.
-12.0000 -0.0320 0.3200 0.2964 0.5820 12.0000 γιατί πάει στο 12;;;;;;;

Standard dev 1.615025

Slope
Min. 1st Qu. Median Mean 3rd Qu. Max.
-6.0540 -0.0390 0.1390 0.1578 0.3620 10.8200

Standard dev 0.6003278

21/6/15

Στο παράρτημα να εισαχθεί ο κώδικας. Θέλουν διορθώσεις τα σχόλια. Ελληνικά ή αγγλικά;;; Μάλλον Ελληνικά

28/6/15
Έτοιμο το διάγραμμα με όλα τα εξόνια του ανθρώπου.

Επιλέχθησαν οι οικογένειες που έχουν ενδιαφέρον. Σύνολο 43 οικογένειες

Επιλεχθήσαν οι οικογένειες που έχουν ενδιαφέρον. Σύνολο 43 οικογένειες.

Ακολουθούν Menzerath	21
Αντίστροφες του Menzerath	13
Άλλες με ενδιαφέρον	9

Ίσως θα πρέπει να εξεταστεί κάποιο μέτρο με βάση τον διπλασιασμό.... Ώστε να γίνει μια σύγκριση μεταξύ τους και να δούμε συσχέτιση διπλασιασμού και υπακοής στο νόμο του Menzerath...

Γενικά για να ερμηνεύσουμε ή έστω να βρούμε κάποια συσχέτιση πρέπει να δούμε αν υπάρχει κάποιο χαρακτηριστικό που μοιράζονται οι οικογένειες που ακολουθούν το νόμο. Πιθανό ότι δέχονται πολλούς διπλασιασμούς.....

Να διαβάσω paper Νικολαου, Ισραηλ, Lynch(2), kopelman

30/6/15
Pearson or Spearman
http://stats.stackexchange.com/questions/8071/how-to-choose-between-pearson-and-spearman-correlation

Ο Pearson για το πόσο είναι γραμμική η συσχέτιση. Χρησιμοποιεί όλα τα δεδομένα. Ο spearman χρησιμοποιεί rank.

Κλίση πληθυσμιακής ευθείας παλινδρόμησης ή κλίση γραμμικής παρεμβολής;;;