Αυτόματη ενημέρωση των DIANA microT εργαλείων σχετικά με τη γενετική βάση δεδομένων Ensembl

Κεσόγλου Μαρίνα

Επιβλέπουσα καθηγήτρια:
Χατζηγεωργίου Άρτεμις

Διπλωματική εργασία

Πανεπιστήμιο Θεσσαλίας
Οκτώβριος 2014

Περίληψη

Η Ensembl είναι ένα project που συνίσταται στη δημιουργία βάσεων δεδομένων για το γονιδίωμα σπονδυλωτών και άλλων ευκαρυωτικών οργανισμών ,καθιστώντας όλη την πληροφορία ελεύθερα προσβάσιμη μέσω διαδικτύου. Η συνεχής εμφάνιση νέων γονιδιακών δεδομένων έχει ως συνέπεια την ανανέωση του project σε περιοδικό κύκλο τριών περίπου μηνών και την δημοσίευση μιας καινούριας έκδοσης της Ensembl .

Οι διαφορές ανάμεσα σε δυο διαδοχικές εκδόσεις μπορεί να αφορούν τη δημιουργία μιας βάσης δεδομένων για ένα νέο οργανισμό ,την ανανέωση της πληροφορίας για έναν ήδη υπάρχοντα οργανισμό καθώς και την προσθήκη/αναμόρφωση των εργαλείων της διαδικτυακής πλατφόρμας του Project.

Σκοπός αυτής της διπλωματικής εργασίας είναι η καταγραφή συγκεκριμένων αλλαγών που σχετίζονται με το γονιδίωμα του ανθρώπου και του ποντικιού από την πρώτη έκδοση της Ensembl μέχρι την τελευταία , καθώς και η αυτόματη ανανέωση αυτού του ιστορικού κατά τη δημοσίευση μιας μελλοντικής έκδοσης. Το εύρος των αλλαγών που καταγράφονται αφορά κυρίως τα αναγνωριστικά των γονιδίων και των μεταγράφων.

Για την επίτευξη του παραπάνω στόχου υλοποιήθηκαν δυο εφαρμογές. Η πρώτη εξυπηρετεί την εύρεση του ιστορικού ενός ID επιστρέφοντας τα IDs στα οποία αντιστοιχεί στην τρέχουσα έκδοση της βάσης και η δεύτερη, την εξόρυξη δεδομένων γι’ αυτά τα IDs από την τελευταία έκδοση της Ensembl .

Ο συνδυασμός των δυο αυτών υλοποιήσεων εξυπηρετεί το ερευνητικό κοινό ,καθιστώντας πιο εύκολη την ανανέωση των εργαλείων που βασίζονται στο project της Ensembl (χρησιμοποιώντας τα stable IDs) καθώς και την εξόρυξη πληροφορίας για κάποιο ID που έχει πιθανώς καταργηθεί ή έχει αλλάξει ανάμεσα στις διαφορετικές εκδόσεις.

Abstract

Ensembl is a project that consists of producing genome databases for vertebrates and other eukaryotic species and making this information freely available online. As a result of Molecular Biology and Technology development, new data are being produced in a daily basis, therefore, a new updated version of Ensembl is released on an approximately three-month cycle .The difference between two Ensembl Releases may refer to building a database for a new species , updating a species already existing database and introducing or retiring features in Ensembl website and API. The purpose of this thesis is to document specific changes that have occured since the first Ensembl Release and are associated with human and mouse genome, such as the automatic update of this record in a future release. The extent of the changes recorded ,is foremost about changes concerning Gene and Transcript Identifiers. Two applications were developed in order to achieve this. The first one is looking into the history of an archive ID, returning a list of current IDs that it has been mapped to ,while the other application is a data mining process ,extracting data concerning these current IDs from the Ensembl Database. The combination of these two implementations is referred to the research community, making it easier to update tools based on the Ensembl Project or extract information for retired or updated IDs.

Πίνακας Περιεχομένων

1.Βασικές Βιολογικές Έννοιες

1.1 Το DNA και το Μοντέλο της Διπλής Έλικας.

Τα κύτταρα διακρίνονται σε προκαρυωτικά και ευκαρυωτικά.

- Τα **προκαρυωτικά** κύτταρα, τα κύτταρα των βακτηρίων, είναι μικρού σχετικά μεγέθους και η πλασματική τους μεμβράνη περιβάλλεται από ένα σκληρό κυτταρικό τοίχωμα. Στο εσωτερικό των προκαρυωτικών κυττάρων δεν υπάρχει οργανωμένος πυρήνας αλλά μια δομή χρωματίνης, το νουκλεοειδές.
- Τα **ευκαρυωτικά** κύτταρα χαρακτηρίζονται από σαφή διαφοροποίηση του κυτταροπλάσματος (κυτταρικά οργανίδια: μιτοχόνδρια, χλωροπλάστες, ενδοπλασματικό δίκτυο κ.ά., κυτταρικός σκελετός). Το γενετικό υλικό των ευκαρυωτικών κυττάρων διαχωρίζεται από το κυτταρόπλασμα με τον πυρηνικό φάκελο. [1]

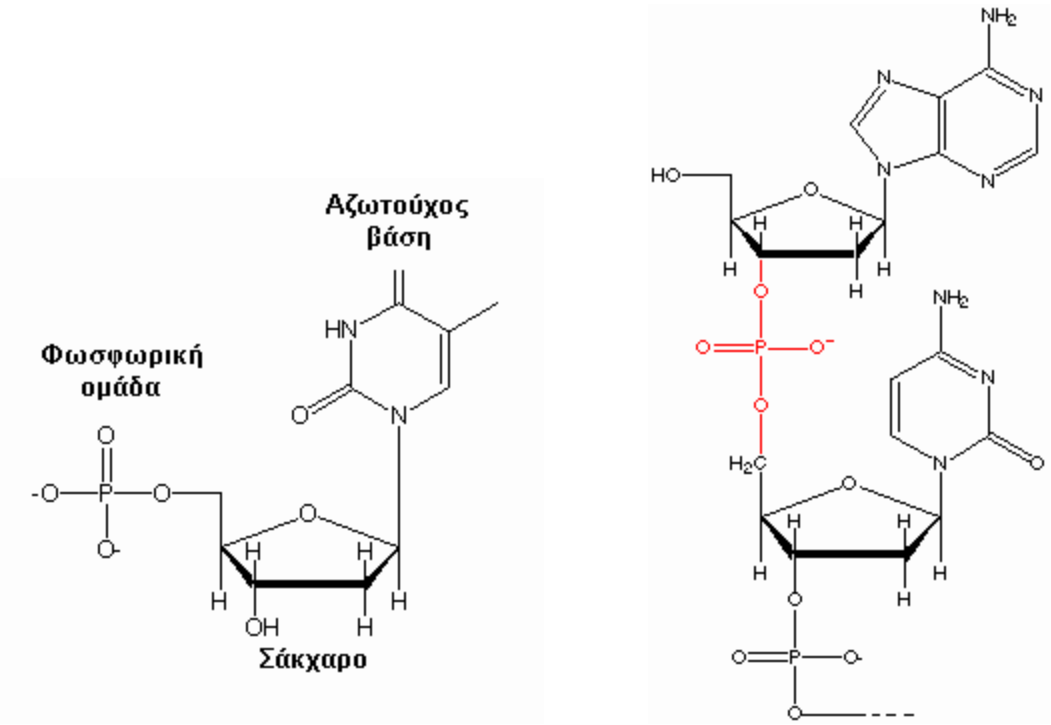
Οι οργανισμοί που έχουν προκαρυωτικά ή ευκαρυωτικά κύτταρα χαρακτηρίζονται ως προκαρυώτες ή ευκαρυώτες αντίστοιχα.

1.1.1 Δομή του DNA

Το δε(σ)οξυριβο(ζο)νουκλεϊ(νι)κό οξύ (Deoxyribonucleic acid - DNA) είναι ένα μακρομόριο που αποτελείται από νουκλεοτίδια , καθένα από τα οποία συνίσταται από μία πεντόζη, τη δεοξυριβόζη, ενωμένη με μία φωσφορική ομάδα και μία αζωτούχο βάση(αδενίνη (A), γουανίνη (G), κυτοσίνη (C) και θυμίνη (T)).

προκαρυώτες	ευκαρυώτες
κυκλικό DNA στο κυτταρόπλασμα	γραμμικό DNA που σχηματίζει χρωμοσώματα στο εσωτερικό του πυρήνα

Πίνακας 1.1: Προκαρυώτες - Ευκαρυώτες : Διαφορά στο DNA



Εικόνα 1.1: Η μορφή του νουκλεοτιδίου και ο 3’-5’ φωσφοδιεστερικός δεσμός

Τα νουκλεοτίδια ενώνονται μεταξύ τους με ομοιοπολικό δεσμό. Ο δεσμός αυτός δημιουργείται μεταξύ του υδροξυλίου του 3' άνθρακα της πεντόζης (του πρώτου νουκλεοτιδίου) και της φωσφορικής ομάδας που είναι συνδεδεμένη στον 5' άνθρακα της πεντόζης (του επόμενου νουκλεοτιδίου). Ο δεσμός αυτός ονομάζεται 3'-5' φωσφοδιεστερικός .Το πρώτο νουκλεοτίδιο μιας πολυνουκλεοτιδικής αλυσίδας έχει πάντα μία ελεύθερη φωσφορική ομάδα συνδεδεμένη στον 5' άνθρακα της πεντόζης του και το τελευταίο νουκλεοτίδιο της έχει ελεύθερο το υδροξύλιο του 3' άνθρακα της πεντόζης του. Για το λόγο αυτό, αναφέρεται ότι ο προσανατολισμός της πολυνουκλεοτιδικής αλυσίδας είναι 5'→3' . (Εικόνα 1.1)

1.1.2 Ιστορικά

Οι βιολόγοι της δεκαετίας του 1940 δυσκολεύονταν να αποδεχτούν το DNA ως γενετικό υλικό , λόγω της φαινομενικά απλής χημείας του. Ήταν γνωστό ότι το DNA είναι ένα μακρύ πολυμερές που αποτελείται από τέσσερις διαφορετικούς τύπους υπομονάδων , οι οποίες μοιάζουν χημικά μεταξύ τους. Στις αρχές τις δεκαετίας του 1950 ,το DNA εξετάστηκε για πρώτη φορά με την τεχνική της περίθλασης ακτίνων Χ ,τα αποτελέσματα της οποίας έδειξαν ότι αποτελείται από δύο κλώνους , τυλιγμένους σε μορφή έλικας.[2]

1.1.3 Το Μοντέλο της Διπλής Έλικας

Η παρατήρηση ότι το DNA είναι δίκλωνο ήταν κρίσιμης σημασίας και ώθησε τους J.D Watson και F.H.C Crick να προτείνουν το μοντέλο της δομής του DNA , γνωστό ως "Μοντέλο της Διπλής Έλικας".[3] Τα βασικά χαρακτηριστικά αυτού του μοντέλου είναι τα εξής :

- Το DNA αποτελείται από δυο πολυνουκλεοτιδικές αλυσίδες σε μορφή δυο αντιτακτών κλώνων που σχηματίζουν δεξιόστροφη διπλή έλικα.

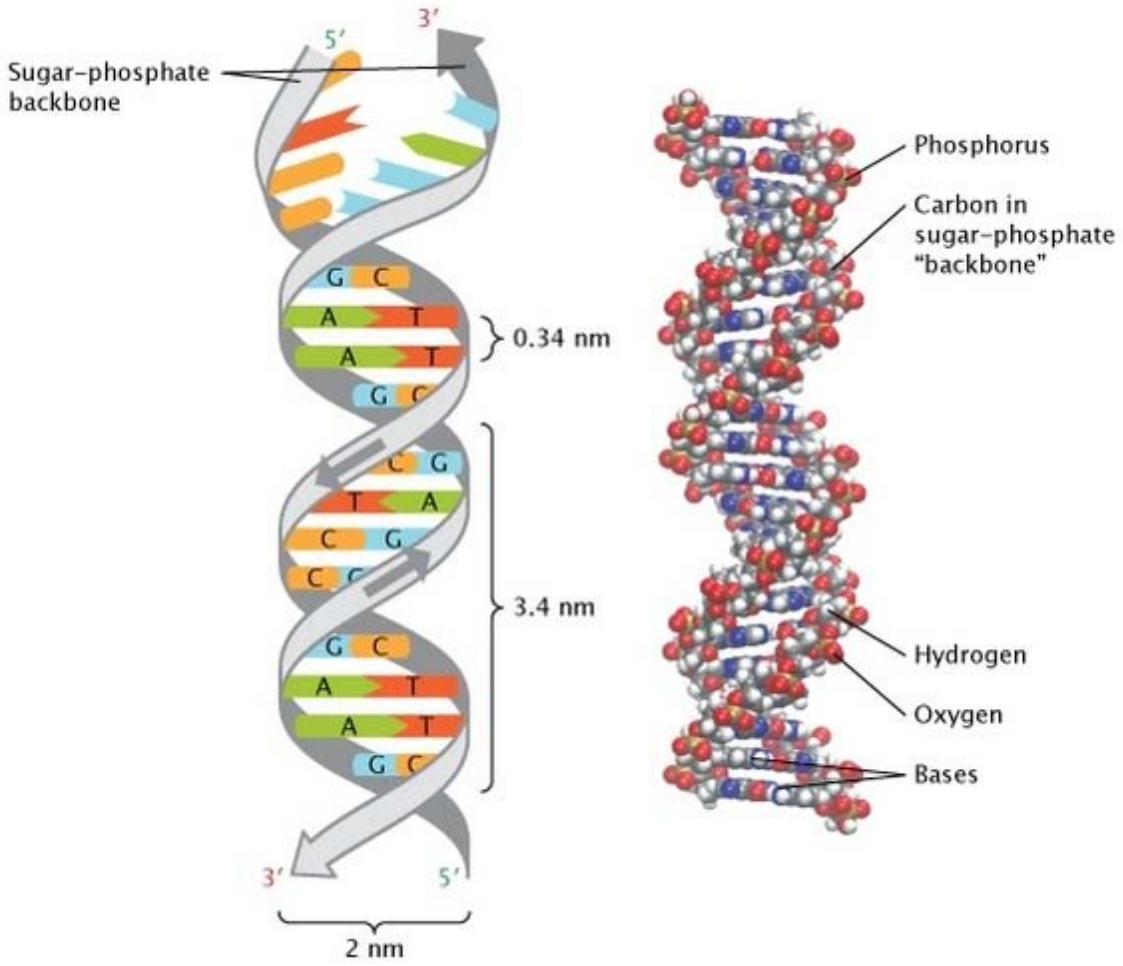
- Η διπλή έλικα έχει σταθερό σκελετό , που αποτελείται από επαναλαμβανόμενα μόρια φωσφορικής ομάδας-δεοξυριβόζης ενωμένων με φωσφοδιεστερικό δεσμό.

- Οι αζωτούχες βάσεις της μιας αλυσίδας συνδέονται με δεσμούς υδρογόνου με τις αζωτούχες βάσεις της απέναντι αλυσίδας με βάση τον κανόνα της συμπληρωματικότητας. Η αδενίνη συνδέεται μόνο με θυμίνη και αντίστροφα, ενώ η κυτοσίνη μόνο με γουανίνη και αντίστροφα. Οι δεσμοί υδρογόνου που αναπτύσσονται μεταξύ των βάσεων σταθεροποιούν τη δευτεροταγή δομή του μορίου.

- Ανάμεσα στην αδενίνη και τη θυμίνη σχηματίζονται δυο δεσμοί υδρογόνου, ενώ ανάμεσα στη γουανίνη και την κυτοσίνη σχηματίζονται τρεις δεσμοί υδρογόνου.

- Οι δύο αλυσίδες ενός μορίου DNA είναι συμπληρωματικές, και αυτό υποδηλώνει ότι η αλληλουχία της μιας καθορίζει την αλληλουχία της άλλης.

- Οι δύο αλυσίδες είναι αντιπαράλληλες, δηλαδή το 3' άκρο της μίας είναι απέναντι από το 5' άκρο της άλλης. [4]



Εικόνα 1.2: Το μοντέλο της διπλής έλικας του DNA.

προκαρυώτες	ευκαρυώτες
μια θέση έναρξης της αντιγραφής	πολλές θέσεις έναρξης της αντιγραφής

Πίνακας 1.2: Προκαρυώτες – Ευκαρυώτες: Διαφορά στην αντιγραφή του DNA

1.2 Κεντρικό Δόγμα της Μοριακής Βιολογίας

1.2.1 Αντιγραφή του DNA (DNA ⇒ DNA)

Ίσως η πιο συναρπαστική πτυχή του "Μοντέλου της Διπλής Έλικας" των Watson και Crick ήταν, όπως εκφράστηκε και από τους ίδιους το ότι *"το ειδικό ζευγάρωμα (ανάμεσα στις βάσεις) που έχουμε υποθέσει αμέσως υποδεικνύει έναν πιθανό μηχανισμό αντιγραφής του γενετικού υλικού"*.Αν η διπλή έλικα χωριστεί σε δυο μονές αλυσίδες,τότε μπορεί να αναπαραχθεί , αφού κάθε μια χρησιμεύει σαν πρότυπο για την κατασκευή της συμπληρωματικής της.[5]

Η διαδικασία της αντιγραφής είναι η εξής :

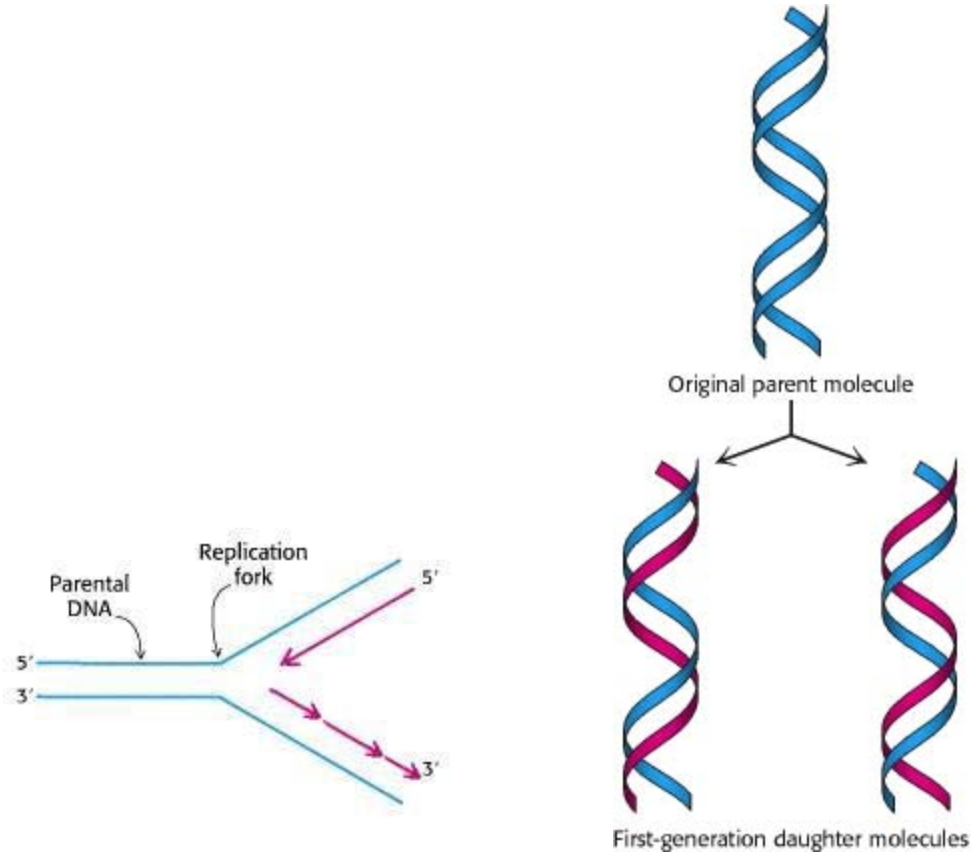
- Οι δυο αλυσίδες είναι ενωμένες μεταξύ τους μέσω των δεσμών υδρογόνου που σχηματίζονται ανάμεσα στις συμπληρωματικές αζωτούχες βάσεις. Στις θέσεις έναρξης της αντιγραφής επεμβαίνουν ειδικά ένζυμα , οι DNA ελικάσες, οι οποίες σπάνε τους δεσμούς υδρογόνου κι αρχίζουν να ξετυλίγουν την έλικα, δημιουργώντας μια "θηλιά". Η θηλιά αυτή επεκτείνεται και προς τις δυο κατευθύνσεις .(Εικόνα 1.3)

- Η αντιγραφή της κάθε αλυσίδας για την κατασκευή της συμπληρωματικής της γίνεται με τη βοήθεια άλλων ενζύμων που ονομάζονται DNA πολυμεράσες. Τα ένζυμα αυτά λειτουργούν προς μία μόνο κατεύθυνση , τοποθετώντας το επόμενο νουκλεοτίδιο στο 3ο άκρο της πεντόζης του τελευταίου νουκλεοτιδίου κάθε αναπτυσσόμενης αλυσίδας. Οπότε η αντιγραφή έχει προσανατολισμό 5′→ 3′.

- Κάθε νέα αλυσίδα έχει προσανατολισμό 5′→ 3′ οπότε και στις δυο διπλές έλικες που παράγονται , οι αλυσίδες είναι μεταξύ τους αντιπαράλληλες. Για να γίνει αυτό , σε κάθε τμήμα DNA που γίνεται η αντιγραφή, η σύνθεση του DNA είναι συνεχής στη μια αλυσίδα και ασυνεχής στην άλλη. Τα κομμάτια της ασυνεχούς αλυσίδας ονομάζονται τμήματα

Okazaki και συνδέονται μεταξύ τους με τη βοήθεια ενός ένζυμου, που ονομάζεται DNA δεσμάση. Το ίδιο ένζυμο συνδέει και όλα τα κομμάτια που προκύπτουν από τις διάφορες *θέσεις έναρξης* αντιγραφής.(Εικόνα 1.4)

Εικόνα 1.3: Ξεδίπλωμα της διπλής έλικας – δημιουργία "θηλιάς"



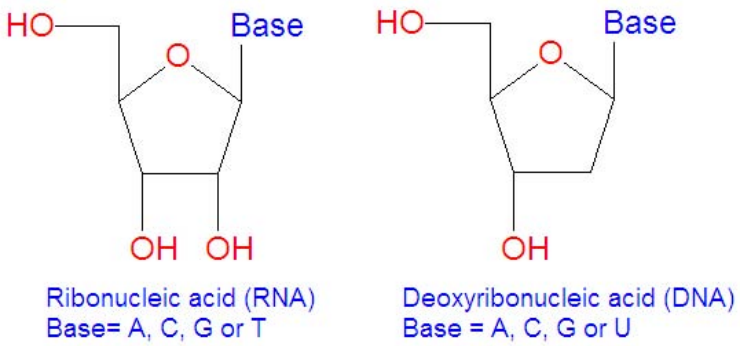
Εικόνα 1.4: Η κατασκευή των θυγατρικών αλυσίδων

Η αντιγραφή του DNA είναι απίστευτα ακριβής, μόνο ένα νουκλεοτίδιο στα 10.000 μπορεί να ενσωματωθεί λάθος. Τα επιδιορθωτικά ένζυμα περιορίζουν τα λάθη σε 1 στα 10^{10}.

1.2.2 Δομή του RNA

Το RNA (Ριβονουκλεικό οξύ) είναι το κινητό αντίγραφο της πληροφορίας ενός γονιδίου. Η δομή του μοιάζει πολύ μ' αυτήν του DNA με μόνες διαφορές τις παρακάτω :

1. στη θέση του σακχάρου περιέχει ριβόζη αντί για δεοξυριβόζη .
2. στη θέση της θυμίνης περιέχει ουρακίλη.
3. είναι μονόκλωνο σε σχέση με το DNA που είναι δίκλωνο, δηλαδή αποτελείται από μία μόνο αλυσίδα.



Εικόνα 1.5: Διαφορές στη σύσταση ανάμεσα σε RNA και DNA

1.2.3 Μεταγραφή του DNA (DNA ⇒ RNA)

Η ποσότητα DNA είναι ίδια σε όλα τα είδη κυττάρων ενός οργανισμού εκτός από τους γαμέτες των ανώτερων οργανισμών , οι οποίοι έχουν τη μισή ποσότητα. Ο γενετικός κώδικας συχνά αναφέρεται ως "σχεδιάγραμμα" γιατί περιέχει όλες τις οδηγίες που απαιτούνται για την επιβίωση ενός κυττάρου. Γνωρίζουμε πλέον , ότι σ' αυτές τις οδηγίες υπάρχει κάτι περισσότερο από μια απλή ακολουθία γραμμάτων του νουκλεοτιδικού κώδικα. Για παράδειγμα ,υπάρχουν άπειρες αποδείξεις ότι ο κώδικας αυτός

είναι η βάση για την παραγωγή μορίων ,όπως το RNA και οι πρωτεΐνες.[6] Η γενετική πληροφορία είναι αποθηκευμένη σε συγκεκριμένες αλληλουχίες νουκλεοτιδίων του DNA που ονομάζονται **γονίδια (genes).** Το σύνολο των γονιδίων ενός οργανισμού αποτελεί το **γονιδίωμά** του. Η πρώτη φάση της έκφρασης των γονιδίων είναι η μετατροπή τους σε RNA , σε μια λειτουργία που λέγεται μεταγραφή.

Η διαδικασία της μεταγραφής έχει ως εξής :

- Τα ρυθμιστικά στοιχεία της μεταγραφής είναι ένα ένζυμο , η RNA πολυμεράση και συγκεκριμένες αλληλουχίες DNA που βρίσκονται πριν από κάθε γονίδιο και ονομάζονται υποκινητές. Στην πρώτη φάση , η RNA πολυμεράση προσδένεται στον υποκινητή και προκαλεί το τοπικό ξετύλιγμα της διπλής έλικας του DNA.

- Η RNA πολυμεράση τοποθετεί απέναντι από κάθε δεοξυριβονουκλεοτίδιο της μιας αλυσίδας DNA , το αντίστοιχο ριβονουκλεοτίδιο ,σύμφωνα με τον κανόνα της συμπληρωματικότητας των βάσεων. Η μόνη διαφορά με την αντιγραφή έγκειται στο ότι απέναντι από το δεοξυριβονουκλεοτίδιο που περιέχει αδενίνη , τοποθετείται το ριβονουκλεοτίδιο που περιέχει ουρακίλη.

- Η RNA πολυμεράση , συνδέει κάθε νέο ριβονουκλεοτίδιο που προσθέτει (με το προηγούμενο) με 3′–5′ φωσφοδιεστερικό δεσμό. Έτσι , η μεταγραφή , όπως και η αντιγραφή έχει προσανατολισμό 5′→ 3′.

- Στο τέλος κάθε γονιδίου υπάρχει μια συγκεκριμένη αλληλουχία , που ονομάζεται “αλληλουχία λήξης” που επιτρέπει την απελευθέρωση του RNA και τον τερματισμό της μεταγραφής.

- Η αλληλουχία RNA που παράγεται ονομάζεται **μετάγραφο (transcript)** .

προκαρυώτες	ευκαρυώτες
η μεταγραφή λαμβάνει χώρα στο πυρηνοειδές	η μεταγραφή λαμβάνει χώρα στο εσωτερικό του πυρήνα
το μεγαλύτερο μέρος του DNA κωδικοποιεί RNA	ελάχιστο μέρος του DNA δείχνει να έχει κωδικοποιητική λειτουργία
τα γονίδια δεν περιέχουν εσώνια	τα γονίδια περιέχουν εσώνια

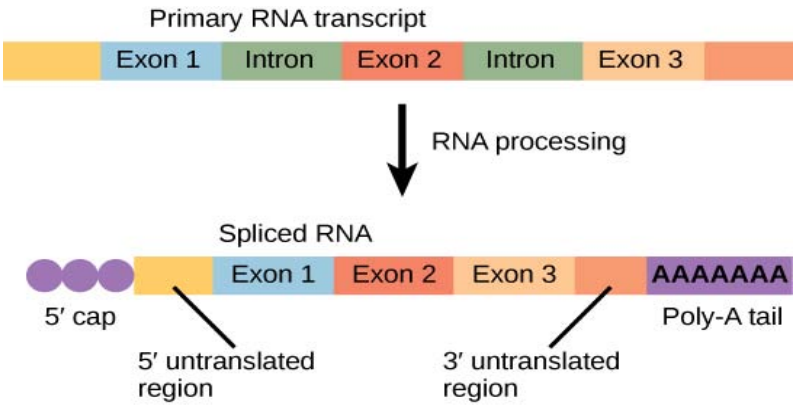
Πίνακας 1.3: Προκαρυώτες - Ευκαρυώτες : Διαφορά στη μεταγραφή του DNA

1.2.4 Κατεργασία του πρόδρομου RNA στους ευκαρυώτες :

- Στους ανώτερους ευκαρυωτικούς οργανισμούς το μέγεθος των γονιδίων ξεπερνά τα 100.000 ζεύγη βάσεων ,από αυτά όμως μόνο 1000 περίπου χρειάζονται για να κωδικοποιήσουν μια πρωτεΐνη. Στην πραγματικότητα ένα γονίδιο αποτελείται από μεγάλες αλληλουχίες βάσεων που δεν κωδικοποιούνται ,τα **εσώνια** (**introns**) ,οι οποίες παρεμβάλλονται ανάμεσα σε άλλες μικρότερες αλληλουχίες που κωδικοποιούνται ,τα **εξώνια** (**exons**).

- Αν το γονίδιο περιέχει εσώνια , το RNA που παράγεται κατά τη μεταγραφή ονομάζεται "πρόδρομο μετάγραφο".

- Το "**πρόδρομο** μετάγραφο" μετατρέπεται σε "**ώριμο**" με τη βοήθεια κάποιων ενζύμων - ριβονουκλεοτιδίων , τα οποία απομακρύνουν τα εσώνια και συρράπτουν τα εξώνια μεταξύ τους.



Εικόνα 1.6 : Μετατροπή από "πρόδρομο" σε "ώριμο" RNA

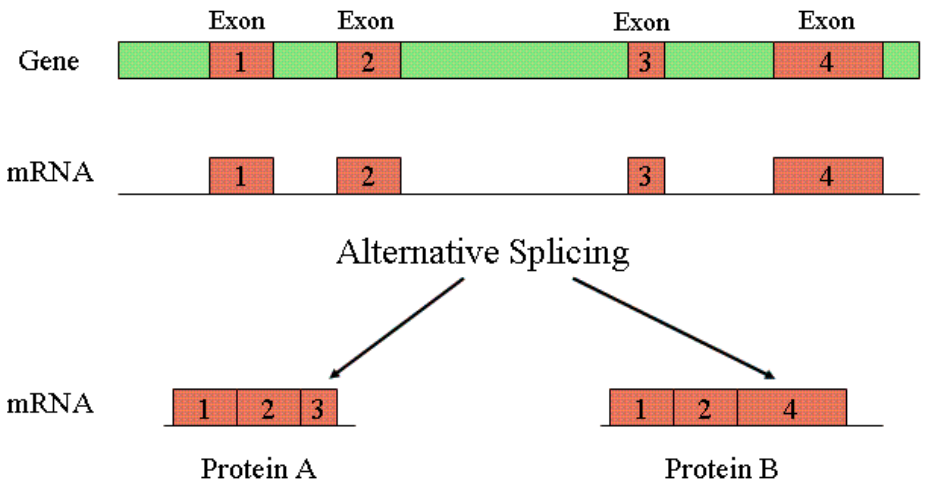
Το RNA που προκύπτει από τη μεταγραφή χωρίζεται σε δυο κατηγορίες :

1. Μη κωδικοποιητικό (**Non coding**) RNA , δηλαδή RNA που δεν κωδικοποιεί πρωτεΐνη. Ο ρόλος αυτών των RNA είναι λειτουργικός . Για παράδειγμα, τα **tRNA** (transfer RNA) και **rRNA** (ribosomal RNA) που έχουν σημαντικό ρόλο στη μετάφραση.

2. Το αγγελιαφόρο RNA (**mRNA**), το οποίο μεταφέρει τη γενετική πληροφορία που είναι αποθηκευμένη στο DNA και κωδικοποιεί μια πρωτεΐνη. [7]

Παρατήρηση

Το "πρόδρομο mRNA" μετατρέπεται σε "ώριμο" με την αφαίρεση των εσωνίων. Το μάτισμα αυτό μπορεί να γίνει με διάφορους εναλλακτικούς τρόπους (alternative splicing) ,έτσι ώστε ένα πρόδρομο mRNA να παράξει περισσότερα από ένα "ώριμα" RNA . Κατ' επέκταση **ένα γονίδιο** μπορεί να έχει περισσότερα από ένα μετάγραφα (transcripts) ,άρα να **κωδικοποιεί παραπάνω από μια πρωτεΐνες**. Ο πιο διαδεδομένος τύπος του alternative splicing είναι η παράλειψη ενός ή περισσότερων εξωνίων (exon skipping) από το μετάγραφο. [8]



Εικόνα 1.7 : Παραγωγή δυο διαφορετικών πρωτεϊνών απ' το ίδιο γονίδιο με exon skipping

1.2.5 Μετάφραση του DNA (RNA ⇒ πρωτεΐνες)

Η μετάφραση του mRNA είναι η διαδικασία κατά την οποία τα ριβονουκλεοτίδια αντιστοιχίζονται σε αμινοξέα που συνδέονται μεταξύ τους και σχηματίζουν μια πολυπεπτιδική αλυσίδα. Τα ριβονουκλεοτίδια κωδικοποιούν ανά τριάδες ένα αμινοξύ. Κάθε τέτοια τριπλέτα ονομάζεται **κωδικόνιο.** Η μετάφραση λαμβάνει χώρα στα ριβοσώματα με τη συμμετοχή των tRNA , καθώς και κάποιων πρωτεϊνών και ενέργειας.

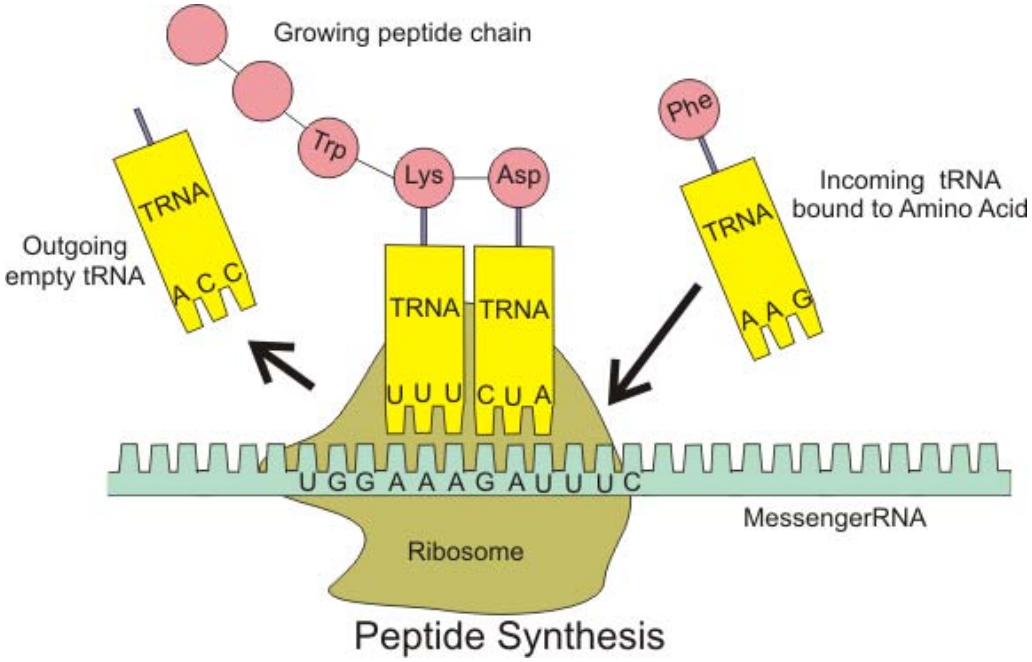
Κάθε ριβόσωμα αποτελείται από δυο υπομονάδες , μια μεγάλη και μια μικρή .Η μικρή υπομονάδα περιέχει μια θέση πρόσδεσης του mRNA. Η μεγάλη υπομονάδα περιέχει δυο θέσεις πρόσδεσης των tRNA. Κάθε μόριο tRNA περιέχει μια θέση σύνδεσης με ένα συγκεκριμένο αμινοξύ ,καθώς και μια αλληλουχία νουκλεοτιδίων , το αντικωδικόνιο ,το οποίο συνδέεται συμπληρωματικά με το αντίστοιχο κωδικόνιο του mRNA.[4]

Η μετάφραση χωρίζεται σε τρεις φάσεις :

1. Έναρξη : Η μετάφραση ξεκινά με τη μικρή υπομονάδα ενός ριβοσώματος, το οποίο περιέχει ένα μόριο tRNA μαζί με το αμινοξύ "μεθειονίνη" .Όταν η μικρή υπομονάδα συναντήσει ένα mRNA ,προσδένεται πάνω του και ξεκινά να το σκανάρει για το κωδικόνιο έναρξης , που είναι το AUG (κωδικοποιεί τη μεθειονίνη).Όταν το βρει , η μεγάλη υπομονάδα ενώνεται στη μικρή σχηματίζοντας ένα ολοκληρωμένο ριβόσωμα και η μετάφραση αρχίζει.
2. Επιμήκυνση : Ένα νέο tRNA μαζί με το αμινοξύ του εισχωρεί στο ριβόσωμα. Αν το αντικωδικόνιό του είναι συμπληρωματικό με το αμέσως επόμενο κωδικόνιο του mRNA , η μεθειονίνη ενώνεται με το αμινοξύ του και το πρώτο tRNA (αυτό που "έφερε" τη μεθειονίνη) αποσυνδέεται από το ριβόσωμα και απελευθερώνεται στο κυτταρόπλασμα. (Αν το κωδικόνιο δεν είναι συμπληρωματικό του κωδικονίου ,απορρίπτεται.) Το ριβόσωμα "διαβάζει" την επόμενη τριπλέτα του mRNA και πλέον ένα άλλο tRNA μπορεί να εισέλθει μαζί με το αμινοξύ του , επαναλαμβάνοντας την ίδια διαδικασία.
3. Τερματισμός : Όταν το ριβόσωμα φτάσει σε ένα από τα τρία κωδικόνια τερματισμού (UAA, UAG και UGA) , στα οποία δεν αντιστοιχεί κανένα tRNA (δεν υπάρχει tRNA με αντικωδικόνιο AUU , AUC ή ACU) ξεκινά ο τερματισμός της μετάφρασης. Το τελευταίο tRNA αποσυνδέεται απ' το ριβόσωμα και η πολυπεπτιδική αλυσίδα απελευθερώνεται. [9]

Παρατήρηση

Ορισμένα κύτταρα χρειάζονται μεγάλες ποσότητες μιας συγκεκριμένης πρωτεΐνης. Για να ικανοποιήσουν αυτήν τους την απαίτηση , έχουν πολλά tRNA που δουλεύουν κατά μήκος ενός mRNA κι έτσι παράγονται πολλά αντίγραφα της πρωτεΐνης.



Εικόνα 1.8 : Η διαδικασία της μετάφρασης.

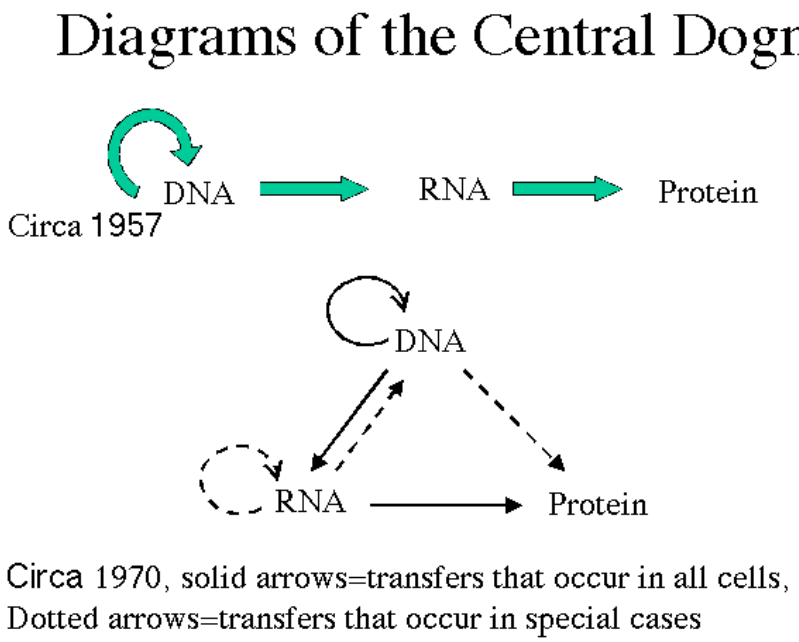
προκαρυώτες	ευκαρυώτες
η μετάφραση αρχίζει πριν τη λήξη της μεταγραφής γιατί δεν υπάρχει κυτταρική μεμβράνη	η μεταγραφή και η ωρίμανση του mRNA γίνονται στον πυρήνα ενώ η μετάφραση στο κυτταρόπλασμα άρα πρέπει να γίνουν σειριακά.

Πίνακας 1.4: Προκαρυώτες – Ευκαρυώτες : Διαφορά στη μετάφραση

1.2.6 Το Κεντρικό Δόγμα

Η κατεύθυνση με την οποία η γενετική πληροφορία, που είναι καταγεγραμμένη στο μόριο του DNA «ρέει» προς τις πρωτεΐνες, ονομάστηκε **Κεντρικό Δόγμα της Βιολογίας** και διατυπώθηκε για πρώτη φορά από τον Francis Crick το 1957.

Έως το 1970 οι επιστήμονες πίστευαν στην καθολικότητα του δόγματος της Μοριακής Βιολογίας. Εκείνη τη χρονιά οι Χ. Τέμιν και Ντ. Μπάλτιμορ ανακάλυψαν ότι ορισμένοι ιοί, οι οποίοι διαθέτουν RNA ως γενετικό υλικό, μπορούν, με πρότυπο αυτό, να συνθέτουν DNA. [10]
Αυτό οδηγεί στην επαναδιατύπωση του κεντρικού δόγματος της Βιολογίας. Έτσι στη σύγχρονη διατύπωση του περιλαμβάνεται και η αμφίδρομη πορεία της γενετικής πληροφορίας από το RNA στο DNA.



Εικόνα 1.9 :Το Κεντρικό Δόγμα της Μοριακής βιολογίας όπως διατυπώθηκε το 1957 και αναδιατυπώθηκε το 1970.

2.Η επιστήμη της Βιοπληροφορικής

2.1 Εισαγωγή στη Βιοπληροφορική

Βιοπληροφορική είναι ο επιστημονικός κλάδος όπου η σύμπραξη της Βιολογίας με την Πληροφορική, την Στατιστική και τα Μαθηματικά εξερευνά νέους τρόπους για την προσέγγιση των βιολογικών προβλημάτων, καθώς και την αντίληψη βασικών αρχών της Βιολογίας. Ηλεκτρονικοί Υπολογιστές χρησιμοποιούνται για τη συλλογή , την αποθήκευση , την ανάλυση και την ενσωμάτωση βιολογικών και γενετικών δεδομένων ,τα οποία μπορεί στη συνέχεια να συντελέσουν στην ανακάλυψη και την ανάπτυξη φαρμάκων. [11]

Η ανάπτυξη της Βιοπληροφορικής ξεκίνησε εξαιτίας της ανάγκης για δημιουργία μεγάλων βάσεων δεδομένων όπως η GenBank , η EMBL και η βάση της Ιαπωνίας, με στόχο την αποθήκευση και την σύγκριση μεγάλων αλληλουχιών DNA του ανθρώπου και άλλων οργανισμών.

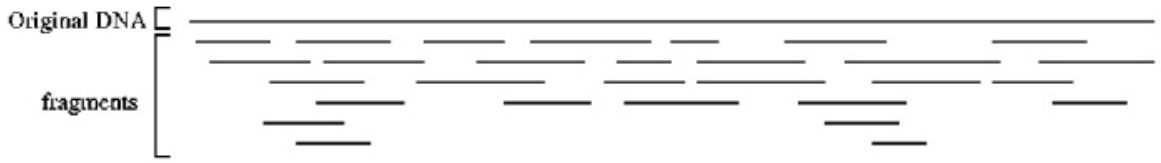
Σήμερα, η Βιοπληροφορική έχει να κάνει με την ανάλυση πρωτεϊνικών δομών, την συλλογή δεδομένων σχετικά με τη λειτουργία των γονιδίων και των πρωτεϊνών , με δεδομένα ασθενών , με προ-κλινικές και κλινικές δοκιμές και με τα μεταβολικά μονοπάτια διάφορων οργανισμών.

Ο κλάδος της Βιοπληροφορικής θεωρείται, παγκοσμίως, ένας από τους πλέον αναπτυσσόμενους. Ουσιαστικά, κατέχει κεντρική θέση στις σύγχρονες εξελίξεις των Επιστημών της Ζωής, με πιο χαρακτηριστικό παράδειγμα ένα από τα μεγαλύτερα ερευνητικά προγράμματα της σύγχρονης επιστήμης , το πρόγραμμα της "Αποκωδικοποίησης" των Γονιδιωμάτων, περιλαμβανομένου και αυτού του Ανθρώπου (Human Genome Project).[12]

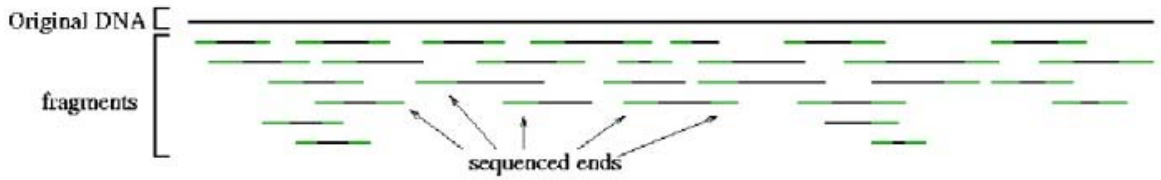
2.2 Genome Sequencing και Genome Assembly

Η διαδικασία μέσω της οποίας γίνεται η χαρτογράφηση της αλληλουχίας του DNA ενός οργανισμού ονομάζεται αλληλούχιση (**sequencing**).

Το 1975 ,ο Frederick Sanger ανέπτυξε τη βασική τεχνολογία αλληλούχισης ,η οποία εξακολουθεί να χρησιμοποιείται ευρέως μέχρι σήμερα. Ενώ αυτή η τεχνολογία βελτιώνεται συνεχώς τα τελευταία 30 χρόνια , ο αριθμός των βάσεων που μπορούν να αποκωδικοποιηθούν ταυτόχρονα, αντιστοιχεί σε 1000-2000 ζεύγη. Ο περιορισμός αυτός είναι πολύ σημαντικός δεδομένου ότι ακόμα και οι απλούστεροι ιοί αποτελούνται από χιλιάδες βάσεις ,τα βακτήρια από εκατομμύρια ,ενώ το γονιδίωμα των θηλαστικών από δισεκατομμύρια ζεύγη βάσεων.
Για να ξεπεραστεί αυτό το εμπόδιο , οι επιστήμονες έχουν αναπτύξει μια τεχνική που λέγεται **"shotgun sequencing"** ,όπου η αλληλουχία DNA ενός οργανισμού "τεμαχίζεται" σε ένα μεγάλο αριθμό μικρών θραυσμάτων (Εικόνα 2.1). Έπειτα , αποκωδικοποιείται η αλληλουχία στα άκρα των θραυσμάτων (Εικόνα 2.2) και τέλος, οι αλληλουχίες που προκύπτουν ενώνονται μεταξύ τους μέσω ενός προγράμματος στον υπολογιστή που ονομάζεται **"Assembler"**(Εικόνα 2.3) . [13]



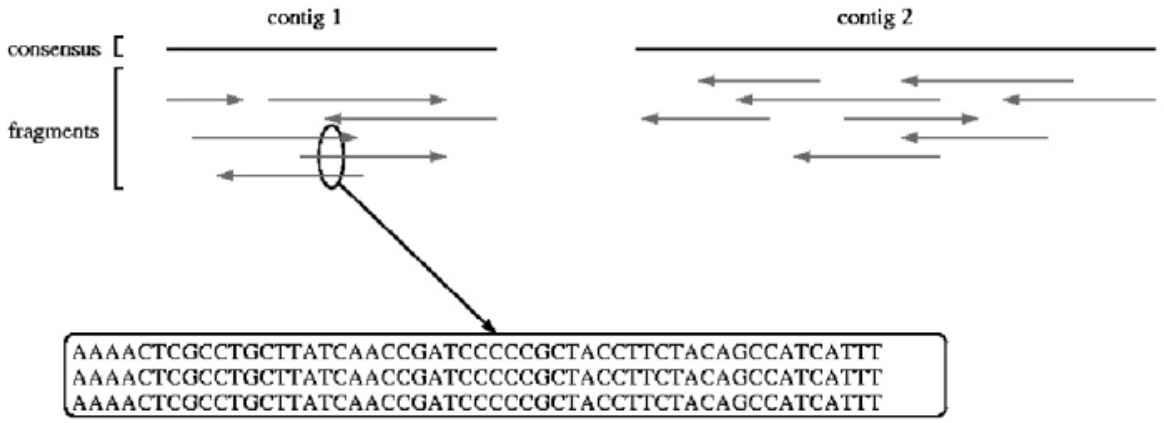
Εικόνα 2.1 :Τεμαχισμός του DNA σε μικρά κομμάτια



Εικόνα 2.2 :αποκωδικοποίηση της αλληλουχίας στα άκρα των μικρών θραυσμάτων

2.2.1 Πως λειτουργεί ο "Assembler"

Ο "Assembler" στηρίζεται στη βασική υπόθεση ότι αν δυο αλληλουχίες DNA που έχουν αποκωδικοποιηθεί (έχει καθοριστεί η σειρά των βάσεών τους) μοιράζονται μια ίδια αλληλουχία βάσεων, προέρχονται από την ίδια θέση στο γονιδίωμα.
Βασιζόμενος σ' αυτές τις επικαλύψεις, ο αλγόριθμος συνενώνει αυτές τις αλληλουχίες με τρόπο παρόμοιο με την κατασκευή ενός παζλ.

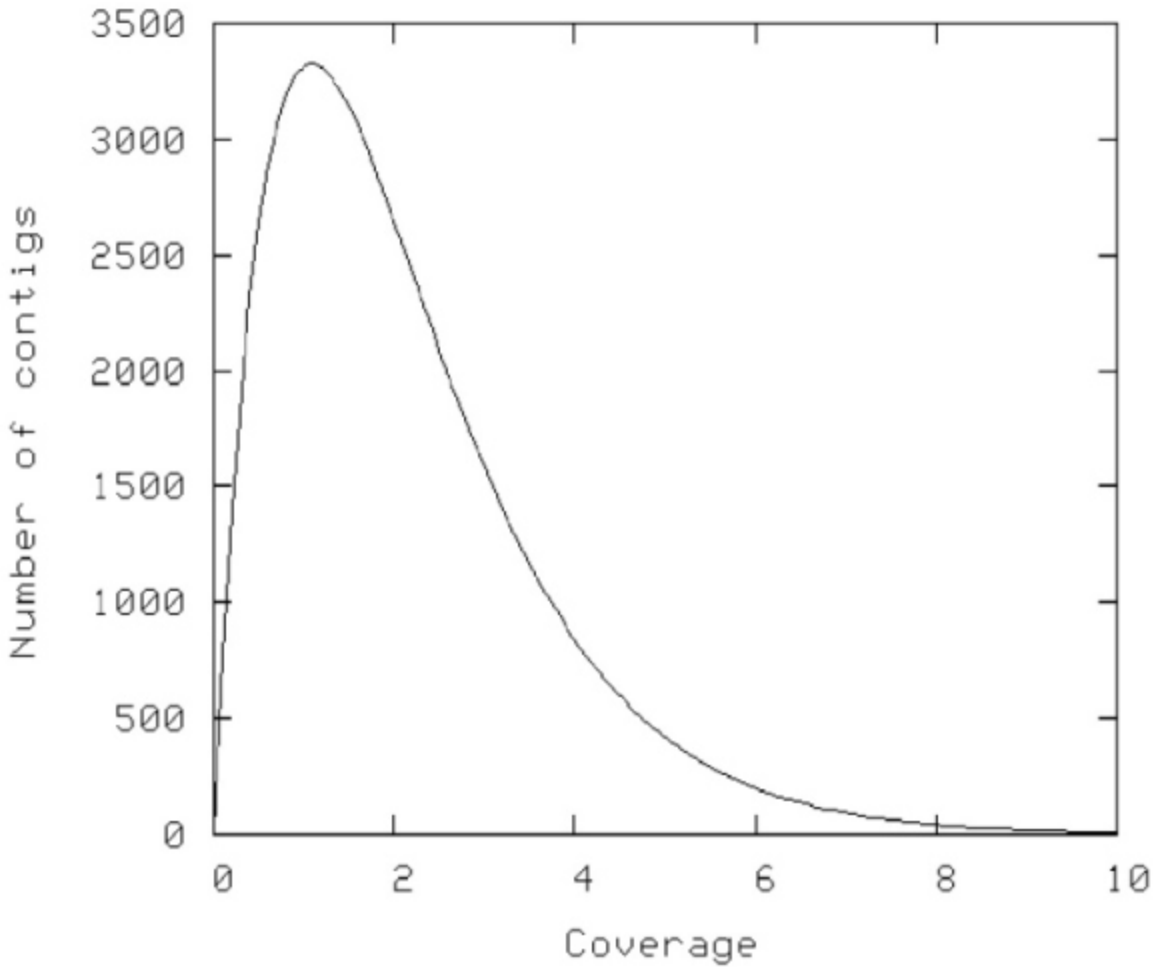


Εικόνα 2.3 : Τα θραύσματα ενώνονται μεταξύ τους βάσει της ομοιότητας στην αλληλουχία των άκρων τους

2.2.2 Assembly statistics

Είναι σημαντικό να σημειωθεί ότι η διαδικασία του "shotgun sequencing" είναι αρκετά "σπάταλη" αφού , λόγω της τυχαιότητας με την οποία το DNA διασπάται σε μικρότερα κομμάτια , η συναρμολόγηση θεωρείται εφικτή μόνο όταν έχουν "συγκεντρωθεί" αρκετές αλληλουχίες ώστε να καλύψουν 8-10 φορές το μήκος του γονιδιώματος. Αυτό το φαινόμενο , μοντελοποιήθηκε μαθηματικά από τους Eric Lander και Michael Waterman το 1988. Αυτοί εξέτασαν τη σχέση ανάμεσα στην υπερδειγματοληψία του γονιδιώματος και στον αριθμό των "συνεχόμενων κομματιών" DNA

(γνωστών και ως **contigs**) που μπορούν να ανακατασκευαστούν από έναν ιδανικό "Assembler" . (Εικόνα 2.4)
Τα contigs που προκύπτουν σχηματίζουν το αρχικό γονιδίωμα στο μεγαλύτερο ποσοστό του.



Εικόνα 2.4 : Γραφική παράσταση της εξίσωσης Lander- Waterman για ένα γονιδίωμα που αποτελείται από 1Mbp = 1.000.000 ζεύγη βάσεων (αν "πάρουμε" δείγμα ίσο με 8-10 φορές το μήκος του, το μεγαλύτερο μέρος του γονιδιώματος θα συναρμολογηθεί από 5 περίπου μεγάλα contigs).

2.2.3 Προκλήσεις κατά την αποκωδικοποίηση του γονιδιώματος

Ιδανικά , ένα πρόγραμμα "συναρμολόγησης" (Assembler) θα έπρεπε να ανακατασκευάζει ένα contig για κάθε χρωμόσωμα. Στις περισσότερες όμως περιπτώσεις παράγονται περισσότερα contigs γεγονός που οφείλεται σε διάφορους παράγοντες.
Οι δυο κύριοι λόγοι είναι οι εξής :

1. Ακόμα και αν η υπερδειγματοληψία πλησιάζει το 8- 10-πλάσιο του μήκους του γονιδιώματος ,υπάρχει πιθανότητα κάποια τμήματα να μη συμπεριληφθούν. Επίσης , η κατανομή των θραυσμάτων (στα οποία τεμαχίζεται το DNA) κατά μήκος του γονιδιώματος δε μπορεί να μοντελοποιηθεί ως μια τέλεια διεργασία Poisson. Έτσι, δημιουργούνται πολλά αντίγραφα για κάθε θραύσμα ώστε να είναι επιτυχημένη η αλληλούχιση.
2. Υπάρχουν αλληλουχίες , οι οποίες επαναλαμβάνονται πολλές φορές μέσα στο γονιδίωμα με αποτέλεσμα να περιορίζουν την απόδοση του Assembler (αναγνωρίζει δυο διαφορετικά αντίγραφα ως ένα και γίνεται λάθος στην assembly).

2.2.4 Τελικό στάδιο και δημοσίευση της assembly

Ο απώτερος σκοπός κάθε "Sequencing Project" είναι να αποκωδικοποιηθούν όλα τα ζεύγη βάσεων όλων των χρωμοσωμάτων . Όπως περιγράφηκε παραπάνω , σπανίως ένας Assembler είναι σε θέση να ανακατασκευάσει ένα μόνο contig για κάθε χρωμόσωμα , γεγονός που οδηγεί στη δημιουργία "κενών" στη συναρμολόγηση του αρχικού γονιδιώματος. Τα κενά αυτά "καλύπτονται" μέσα από εξειδικευμένες μεθόδους και στη συνέχεια εκτελούνται εργαστηριακά πειράματα για να επικυρώσουν την ορθότητα της Assembly που δημιουργήθηκε. Για τους δημοφιλείς οργανισμούς (όπως ο άνθρωπος και το ποντίκι) δημοσιεύονται περιοδικά νέες assemblies στις οποίες διορθώνονται πιθανά λάθη ή καλύπτονται κενά. Η πιο πρόσφατη assembly για τον άνθρωπο είναι η GRCh38 ενώ αντίστοιχα για το ποντίκι , η GRCm38.

2.3 Genome Annotation

Το **Genome Annotation** (σχολιασμός του γονιδιώματος) είναι μια υποενότητα της ανάλυσης του γονιδιώματος, η οποία λίγο πολύ περιλαμβάνει οτιδήποτε μπορεί να γίνει σε μια αλληλουχία DNA μέσω υπολογιστικών μέσων.[14] Ουσιαστικά, είναι η επισύναψη βιολογικών δεδομένων σε αλληλουχίες DNA (coding και non coding) , η βάση της οποίας είναι :

1. η περιγραφή ενός γονιδίου καθώς και των μεταγράφων (transcripts) και των πρωτεϊνών του (proteins)
2. η περιγραφή των 'non coding' αλληλουχιών
3. ο προσδιορισμός της λειτουργικής ιδιότητας κάθε τέτοιου στοιχείου

Το Genome Annotation συνεπάγεται αναγκαστικά κάποιο επίπεδο αυτοματισμού. Το βασικό κομμάτι του Annotation χρησιμοποιεί το **BLAST** (**B**asic **L**ocal **A**lignment **S**earch **T**ool), έναν αλγόριθμο για την σύγκριση βιολογικών δεδομένων , όπως της αλληλουχίας των αμινοξέων δυο (ή περισσότερων) πρωτεϊνών ή η σύγκριση της αλληλουχίας των νουκλεοτιδίων δυο (ή περισσότερων) γονιδίων. Τόσο η εισαγωγή των παραμέτρων όσο και η εξαγωγή των αποτελεσμάτων στο BLAST γίνονται αυτοματοποιημένα (με ειδικό λογισμικό που εκτελεί συγκεκριμένες ρουτίνες και οργανώνει τα αποτελέσματά που επιστρέφουν).

Με εξαίρεση αυτό το κομμάτι του Genome Annotation , όλα τα υπόλοιπα γίνονται ως επί το πλείστον "χειροκίνητα" γιατί τα περισσότερα δεδομένα (όπως πχ η εύρεση της λειτουργία ενός γονιδίου) προϋποθέτουν την ανθρώπινη εμπειρία.

Το Genome Annotation αποτελεί ενεργό τομέα της έρευνας και απασχολεί πολλούς διαφορετικούς επιστημονικούς οργανισμούς. Οι οργανισμοί αυτοί δημοσιεύουν τα αποτελέσματα των προσπαθειών τους σε δημόσια διαθέσιμες βιολογικές βάσεις δεδομένων (προσβάσιμες μέσω διαδικτύου) όπως η Ensembl , η RefSeq , η Entrez Gene, η Uniprot , η ENCODE και άλλες.

2.4 Genome Project

Η χαρτογράφηση (Genome Sequencing and Assembly) του γονιδιώματος σε συνδυασμό με την προσθήκη σχολίων (Genome Annotation) ,συντελούν στην "ολοκλήρωση" ενός Genome Project.

Ουσιαστικά , ένα Genome Project θεωρείται ολοκληρωμένο όταν έχει αποκωδικοποιηθεί κάθε ζεύγος βάσεων στο γονιδίωμα του οργανισμού. Όπως αναφέρθηκε και παραπάνω , αυτό είναι πολύ δύσκολο να επιτευχθεί οπότε κάθε τόσο δημοσιεύονται νέα βελτιωμένα drafts του Genome Project για κάθε οργανισμό. Με το πέρασμα των χρόνων , δημοσιεύονται Genome Projects για όλο και περισσότερους οργανισμούς ,λόγω της μείωσης του κόστους της αποκωδικοποίησης .

2.5 GenBank , DDBJ και EMBL

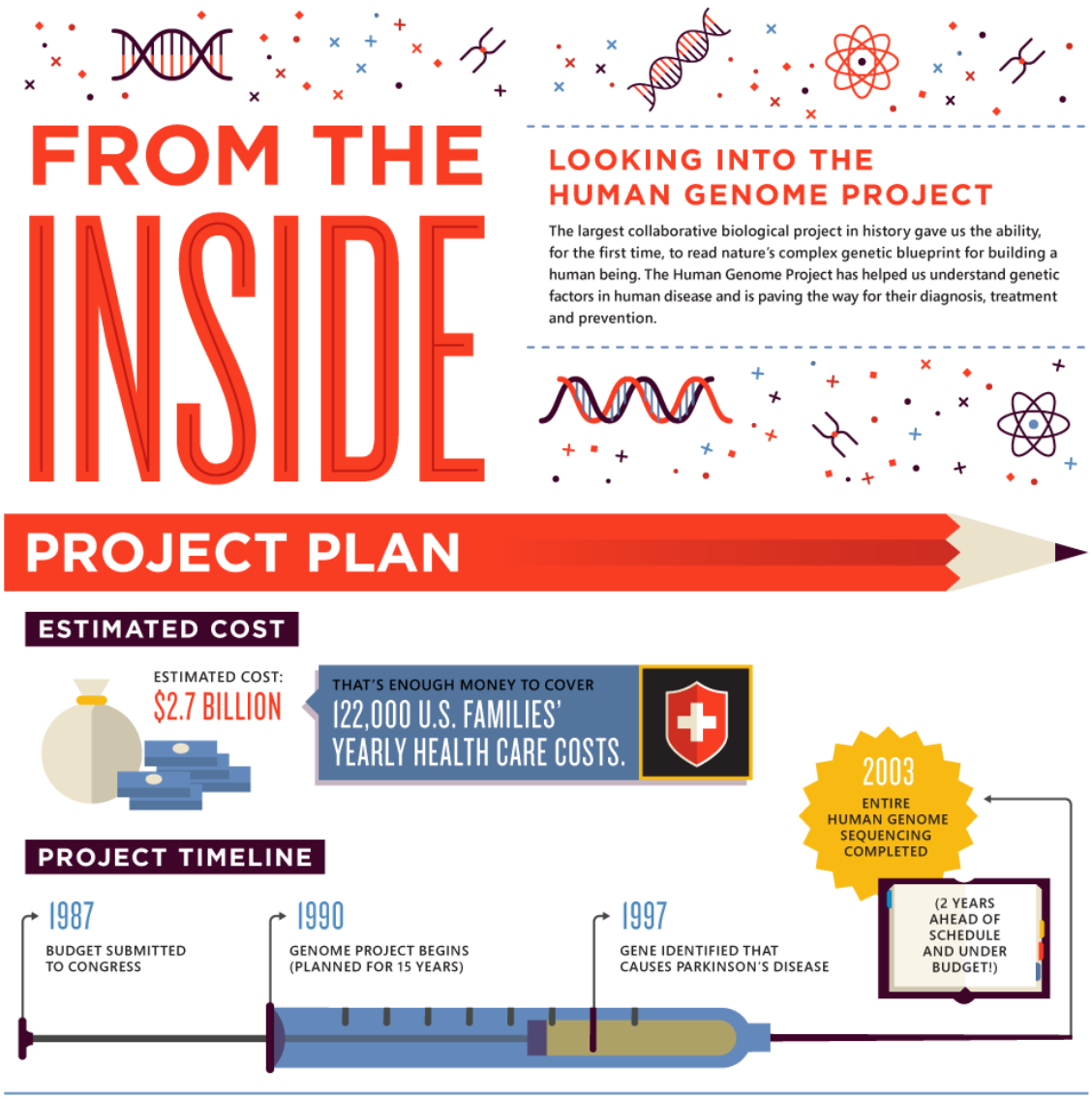
Με την ολοκλήρωση του annotation , το Genome Project κατατίθεται στις τρεις κύριες βάσεις δεδομένων στον κόσμο (GenBank , DDBJ και EMBL).

Η ***GenBank*** στις Ηνωμένες Πολιτείες (*National Center for Biotechnology Information , NCBI*) , μαζί με την DNA DataBank στην Ιαπωνία (*DNA DataBank of Japan,* ***DDBJ***) και το Ευρωπαϊκό Εργαστήριο Μοριακής Βιολογίας (European Molecular Biology Laboratory, ***EMBL***) συνιστούν τη Διεθνή Συνεργασία Βάσεων Δεδομένων για Αλληλουχίες Νουκλεοτιδίων (***International Nucleotide Sequence Database Collaboration***).

Πρόκειται για εκτενείς βάσεις δεδομένων αποτελούμενες από αποκωδικοποιημένες αλληλουχίες νουκλεοτιδίων, οι οποίες αντιστοιχούν σε περισσότερους από 300.000 οργανισμούς. Οι βάσεις αυτές, ανταλλάσσουν δεδομένα σε καθημερινή βάση, έτσι ώστε διασφαλίζεται η διάδοση της γνώσης σε παγκόσμιο επίπεδο.[15]

2.6 Human Genome Project

Το Πρόγραμμα Αποκωδικοποίησης του Ανθρώπινου Γονιδιώματος είναι ένα ερευνητικό πρόγραμμα , με στόχο την πλήρη χαρτογράφηση και κατανόηση του ανθρώπινου γονιδιώματος. Με τον όρο "χαρτογράφηση" ορίζεται ο καθορισμός των ζευγών βάσεων της αλληλουχίας DNA του ανθρώπου.



Εικόνα 2.5.1 :Human Genome Project - Project Plan

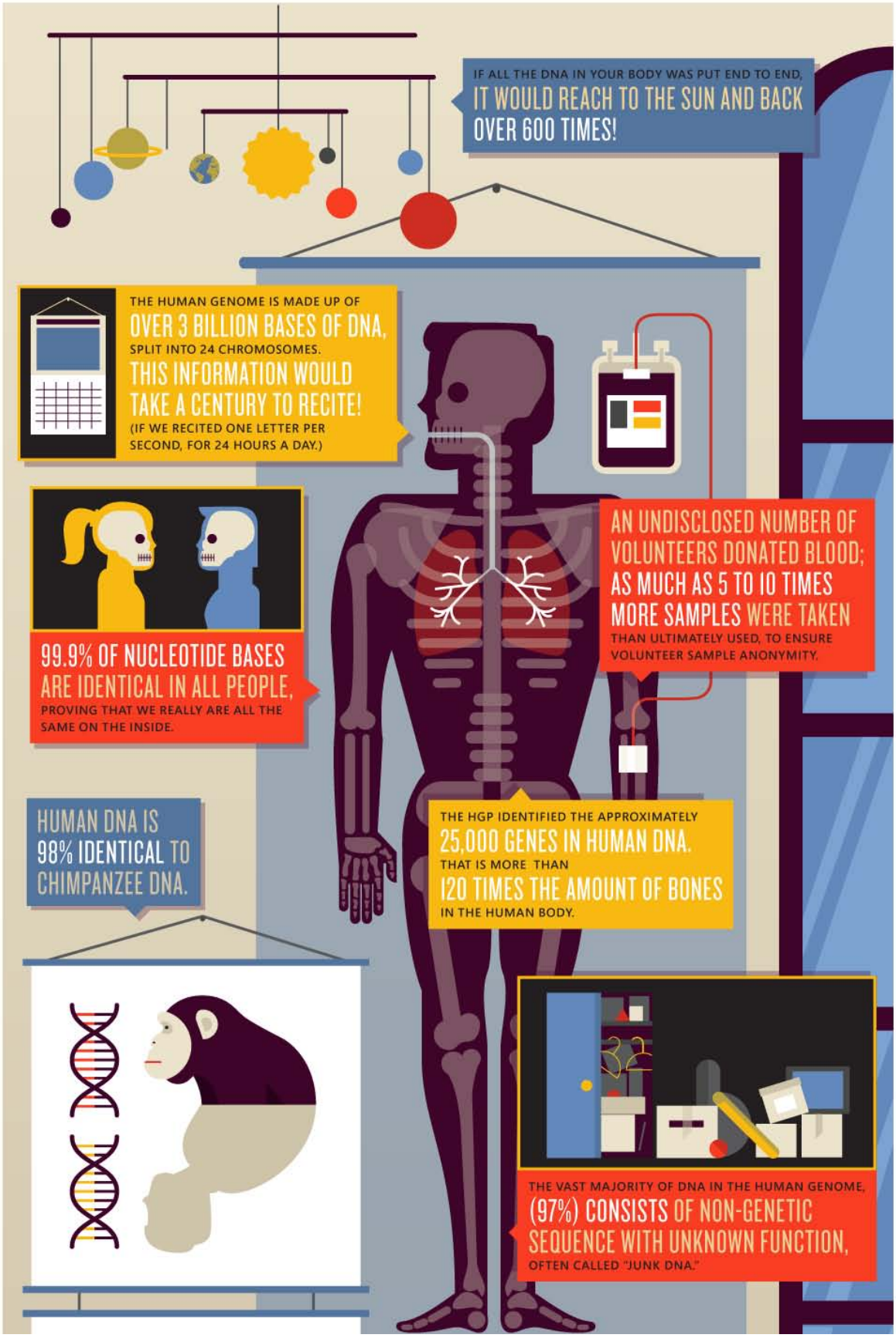
2.6.1 Ιστορικά

Το **Human Gene Project (HGP)** ήταν η φυσική εξέλιξη της ιστορίας της γενετικής έρευνας. Το 1911 ,ο Alfred Sturtevant – τότε ένας προπτυχιακός ερευνητής στο εργαστήριο του Thomas Hunt Morgan – συνειδητοποίησε ότι προκειμένου να διαχειριστεί τα δεδομένα του ,έπρεπε να χαρτογραφήσει τις θέσεις των γονιδίων της μύγας Drosophila, οι μεταλλάξεις της οποίας (από γενιά σε γενιά) παρακολουθούνταν στο εργαστήριο Morgan.
Αν παρομοιάσουμε την καταγραφή του πρώτου γονιδίου από τον Sturtevant με την πρώτη πτήση των αδερφών Wright στο Kitty Hawk τότε με τη σειρά του το HGP είναι μπορεί να συγκριθεί με το πρόγραμμα Apollo , που οδήγησε την ανθρωπότητα στο φεγγάρι.[16]

Το πρόγραμμα ουσιαστικά ξεκίνησε το 1990. Το "International Human Genome Sequencing Consortium" δημοσίευσε το προσχέδιο (90% πλήρες) της χαρτογράφησης του ανθρώπινου γονιδιώματος στο περιοδικό Nature ,τον Φεβρουάριο του 2001 .Ένα αναπάντεχο εύρημα αυτού του προσχεδίου, ήταν ότι ο αριθμός των ανθρώπινων γονιδίων φαίνεται να είναι σημαντικά μικρότερος από τις προηγούμενες εκτιμήσεις, οι οποίες κυμαίνονταν από 50.000 μέχρι και 140,000 γονίδια .Η πλήρης ακολουθία ολοκληρώθηκε και δημοσιεύθηκε τον Απρίλιο του 2003.

Εικόνα 2.5.2 (ακολουθεί) :Το Human Genome Project - Project Details

PROJECT DETAILS



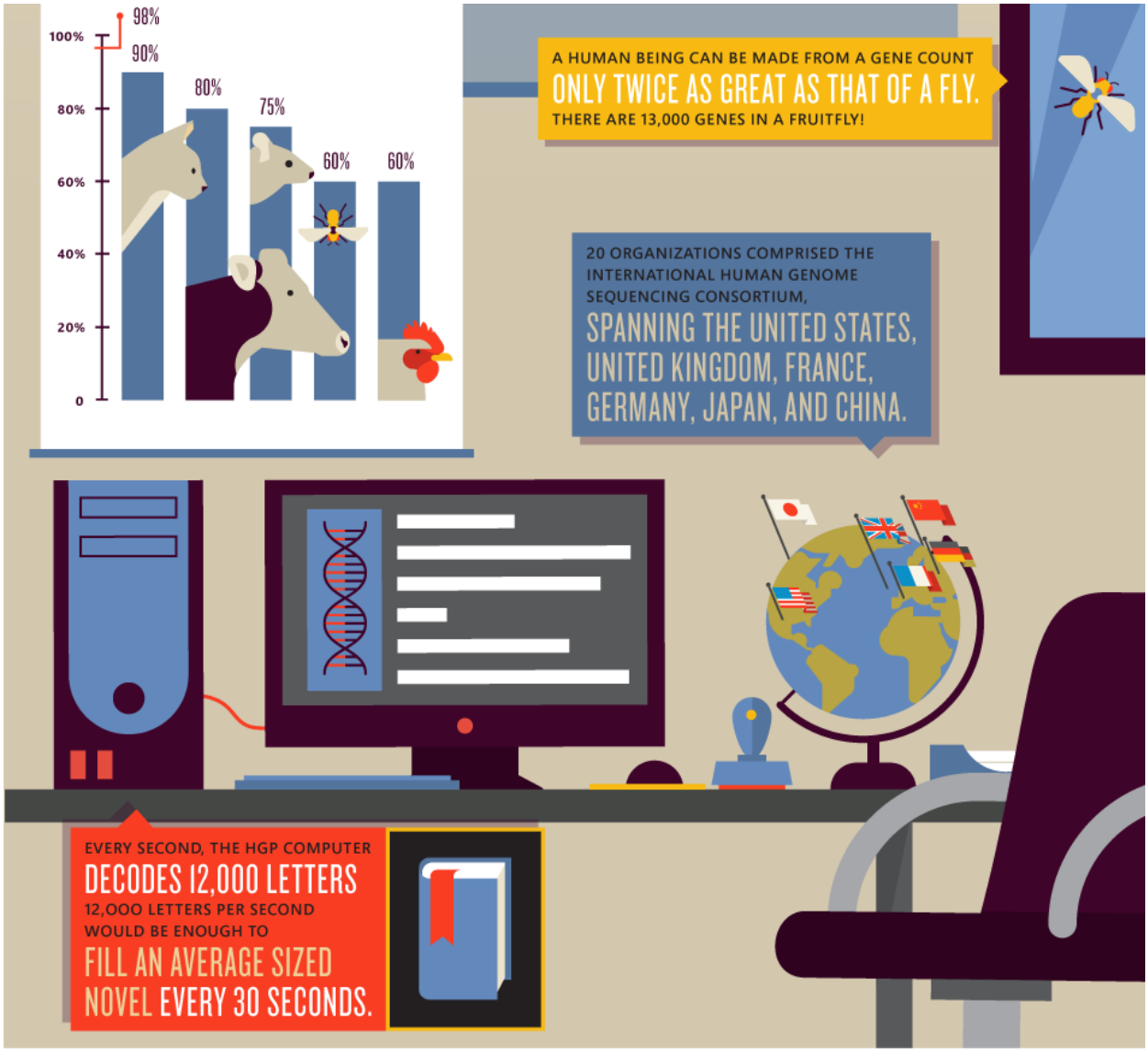
2.6.2 Σύσταση του Project

Το HGP είναι αποτέλεσμα συνεργασίας εθνικών ινστιτούτων και οργανισμών , καθώς και ιδιωτικών εταιριών. Το μεγαλύτερο μέρος του έλαβε χώρα σε πανεπιστημιακά και ερευνητικά ινστιτούτα κυρίως της Αμερικής και της Μεγάλης Βρετανίας.

Οι ερευνητές του HGP έχουν αποκρυπτογραφήσει το ανθρώπινο γονιδίωμα με τους εξής τρόπους :

1. καθορίζοντας τη σειρά ή την "αλληλουχία" όλων των βάσεων του DNA.
2. δημιουργώντας χάρτες που δείχνουν τη θέση κάθε γονιδίου για μεγάλες περιοχές των ανθρώπινων χρωμοσωμάτων .
3. κατασκευάζοντας πολύπλοκους χάρτες διασύνδεσης μέσω των οποίων μπορούν να καταγράφονται διάφορα κληρονομικά χαρακτηριστικά (όπως οι γενετικές ασθένειες) από γενιά σε γενιά.

Σχόλιο : Στα πλαίσια του προγράμματος έγινε χαρτογράφηση του γονιδιώματος και άλλων οργανισμών ,οι οποίοι λειτούργησαν ως μοντέλα για την καλύτερη ερμηνεία της λειτουργίας του ανθρώπινου DNA ,μέσα από την εύρεση των γονιδίων του και την κατανόηση της λειτουργίας τους.



Εικόνα 2.5.3 :To Human Genome Project - Project Details (2)

2.6.3 Αποτελέσματα

Το Human Gene Project αποκάλυψε ότι υπάρχουν περίπου 20.500 ανθρώπινα γονίδια , οι θέσεις των οποίων ,μπορούν πλέον να προσδιοριστούν. Τα αποτελέσματα του προγράμματος , έχουν δώσει στον κόσμο μια πηγή με λεπτομερείς πληροφορίες σχετικά με τη δομή , την οργάνωση και τη λειτουργία ενός ολοκληρωμένου συνόλου ανθρώπινων γονιδίων.

Εικόνα 2.5.4 :To Human Genome Project - Project Outcome

3.Η Ensembl

Η Ensembl είναι ένα project που στόχο έχει τον αυτόματο σχολιασμό (Annotation) των αλληλουχιών του γονιδιώματος ,την ενσωμάτωσή του σε άλλα διαθέσιμα βιολογικά δεδομένα και την ελεύθερη διάθεσή του σε Γενετιστές, Μοριακούς Βιολόγους, Βιοπληροφορικούς και στην ευρύτερη ερευνητική κοινότητα μέσω του διαδικτύου. [17]

3.1 Γενικά στοιχεία

Το Project της Ensembl ξεκίνησε το 1999, μερικά χρόνια πριν τη δημοσίευση του προσχεδίου της χαρτογράφησης του ανθρώπινου γονιδιώματος (2001) .Ακόμα και σ' εκείνο το πρώιμο στάδιο ήταν ξεκάθαρο ότι ο χειροκίνητος σχολιασμός (manual annotation) τριών δισεκατομμυρίων ζευγών βάσεων δε θα ήταν σε θέση να προσφέρει στους ερευνητές άμεση πρόσβαση στα πιο πρόσφατα δεδομένα. Αν και η αφορμή για τη δημιουργία της Ensembl ήταν το Human Genome Project , από το 2000 (που ξεκίνησε η λειτουργία της online πλατφόρμας) μέχρι σήμερα έχουν προστεθεί τα γονιδιώματα πολλών οργανισμών (77 οργανισμών) και το εύρος των διαθέσιμων δεδομένων έχει επεκταθεί έτσι ώστε να περιλαμβάνει και πληροφορίες σχετικά με :

- την σύγκριση του γονιδιώματος διάφορων οργανισμών
- τον πολυμορφισμό
- τον μηχανισμό της ρύθμισης των γονιδίων στα κύτταρα

Η Ensembl (Εικόνα 3.1) είναι ένα κοινό έργο μεταξύ :

1. του Ευρωπαϊκού Ινστιτούτου Βιοπληροφορικής (*European Bioinformatics Institute ,EBI*),που υπάγεται στο Ευρωπαϊκό Εργαστήριο Μοριακής Βιολογίας (EMBL)
2. του Wellcome Trust Sanger Institute (WTSI)

Και τα δυο αυτά ιδρύματα βρίσκονται στο Wellcome Trust Genome Campus στο Hinxton , νότια του Cambridge ,στο Ηνωμένο Βασίλειο.

3.2 Σύσταση της ομάδας της Ensembl

Ο αριθμός των ατόμων που συμμετέχουν στο έργο αυξάνεται σταθερά. Σήμερα ,το ανθρώπινο δυναμικό της Ensembl, με επικεφαλή τον Paul Flicek , αποτελείται από 40-50 άτομα, τα οποία χωρίζονται στις εξής ομάδες :

- **Genebuild team,** η οποία δημιουργεί τα gene sets για τους διάφορους οργανισμούς
- **Software team,** που επεξεργάζεται τα gene sets και φροντίζει για την ανάπτυξη λογισμικού και την ενημέρωση του Biomart (εργαλείο εξόρυξης δεδομένων)
- **Compara team,** που ασχολείται με τα δεδομένα σύγκρισης των γονιδιωμάτων
- **Variation team,** η οποία είναι υπεύθυνη για τον πολυμορφισμό
- **Regulation team,** που επεξεργάζεται την πληροφορία σχετικά με τη ρύθμιση των γονιδίων
- **Web team,** που εξασφαλίζει ότι όλα τα δεδομένα παρουσιάζονται στην ιστοσελίδα με σαφή και φιλικό προς το χρήστη τρόπο
- **Outreach team,** η οποία απαντά στα ερωτήματα των χρηστών και οργανώνει εκπαιδευτικά σεμινάρια σχετικά με την Ensembl σε όλο τον κόσμο. [18]

Εικόνα 3.1 : Τα logos της Ensembl , του Εθνικού Ινστιτούτου Βιοπληροφορικής και του Wellcome Trust Sanger Institute

3.3 Ensembl Genome Annotation

3.3.1 Genome Assemblies in Ensembl

Το Project της Ensembl δεν περιλαμβάνει την παραγωγή των Genome Assemblies ,αλλά παρέχει σχολιασμό σε Assemblies που έχουν κατατεθεί στη Διεθνή Συνεργασία Βάσεων Δεδομένων για Αλληλουχίες Νουκλεοτιδίων (INSDC: GenBank , EMBL , DDBJ).Για κάποιους οργανισμούς υπάρχουν περισσότερες από μια Assemblies. Σ' αυτήν την περίπτωση , η Ensembl , το NCBI και το UCSC (Science Department ,University of California) αποφασίζουν από κοινού για ποια Assembly θα γίνει ο σχολιασμός.[18]

3.3.2 Αξιοπιστία

Το Genome Annotation που παρέχεται από την Ensembl περιλαμβάνει αυτόματο σχολιασμό ,δηλαδή καθορισμό των transcripts σε όλο το μήκος του γονιδιώματος. Για επιλεγμένα είδη (όπως ο άνθρωπος , το ποντίκι , το γουρούνι και άλλα) μπορεί να περιλαμβάνεται και manual annotation. Όλα τα transcripts της Ensembl βασίζονται σε πειραματικά αποδεδειγμένα στοιχεία και η αυτοματοποίηση του σχολιασμού βασίζεται σε mRNA και πρωτεϊνικές αλληλουχίες που έχουν καταχωρηθεί σε δημόσιες βάσεις δεδομένων. Τα transcripts που καταχωρούνται χειροκίνητα παράγονται από το HAVANA group στο Wellcomme Trust Sanger Institute . [18]

3.3.3 Gene και Transcript IDs

Σε κάθε Ensembl γονίδιο (με ειδικό αναγνωριστικό ENSG.. -Ensembl gene ID) αντιστοιχίζονται τα μετάγραφά του (ENST.. -Ensembl transcript ID), οι κωδικές αλληλουχίες των οποίων είναι επικαλυπτόμενες. Τα transcripts που ανήκουν στο ίδιο gene ID διαφέρουν στο μάτισμα, δηλαδή έχουν διαφορετικά εξώνια κι εσώνια (αν και προέρχονται από την ίδια αλληλουχία DNA) ,επομένως παράγουν διαφορετικές πρωτεΐνες. Αυτό είναι απόρροια

του alternative splicing. Μπορεί να υπάρχουν transcripts, τα οποία έχουν επικάλυψη στη μη κωδική περιοχή τους (UnTranslated Region, UTR). Αυτά ταξινομούνται σε διαφορετικά gene IDs .Αφού οριστούν οι αλληλουχίες των γονιδίων και των μεταγράφων, εκχωρούνται τα αντίστοιχα ονόματα (gene και transcript names).[18]

3.3.4 Genebuild

Για τη δημιουργία και ολοκλήρωση ενός Ensembl Genebuild (πρόβλεψη και καθορισμός των γονιδίων) χρειάζονται τουλάχιστον 4 μήνες. Αν και η διαδικασία προσαρμόζεται για κάθε οργανισμό (ανάλογα με τα δεδομένα που είναι διαθέσιμα) , τα βασικά βήματα που ακολουθούνται για την κατασκευή ενός genebuild είναι ίδια. Τα δεδομένα που υπάρχουν διαθέσιμα εξαρχής από άλλες βάσεις δεδομένων (EMBL ,UniprotKB ,RefSeq) είναι κάποιες πρωτεΐνες (που έχουν βρεθεί πειραματικά) καθώς και cDNA (complementary DNA), μόρια που προκύπτουν από ώριμο mRNA (το mRNA μεταγράφεται σε cDNA .Το cDNA αντιστοιχεί στην αρχική αλληλουχία DNA που παρήγαγε το mRNA χωρίς όμως τα εσώνια).

Προεργασία
Η Assembly του οργανισμού ενσωματώνεται στη βάση δεδομένων της Ensembl και εκτελούνται διάφορες αναλύσεις κατά μήκος του γονιδιώματος όπως τη συγκάλυψη των επαναλήψεων (masking of repeats).

Βήμα 1ο
Το πρώτο στάδιο κατασκευής του genebuild ονομάζεται Targetted Stage και περιλαμβάνει την “στοίχιση” ,κατά μήκος του γονιδιώματος, των πρωτεϊνών που αφορούν μόνο τον συγκεκριμένο οργανισμό και την κατασκευή μιας transcript δομής για κάθε πρωτεΐνη.

Βήμα 2ο
Το δεύτερο στάδιο ονομάζεται Similarity Stage και περιλαμβάνει τη δημιουργία transcript δομών με την χρήση πρωτεϊνών από στενά συγγενικά είδη ώστε να καλυφθούν τα κενά που δημιουργεί η απουσία των Targetted transcript structures. Για κάποιους οργανισμούς υπάρχουν πολύ λίγες καταχωρημένες πρωτεΐνες ,οπότε για τη δημιουργία του genebuild βασιζόμαστε σε πρωτεΐνες συγγενικών τους οργανισμών.

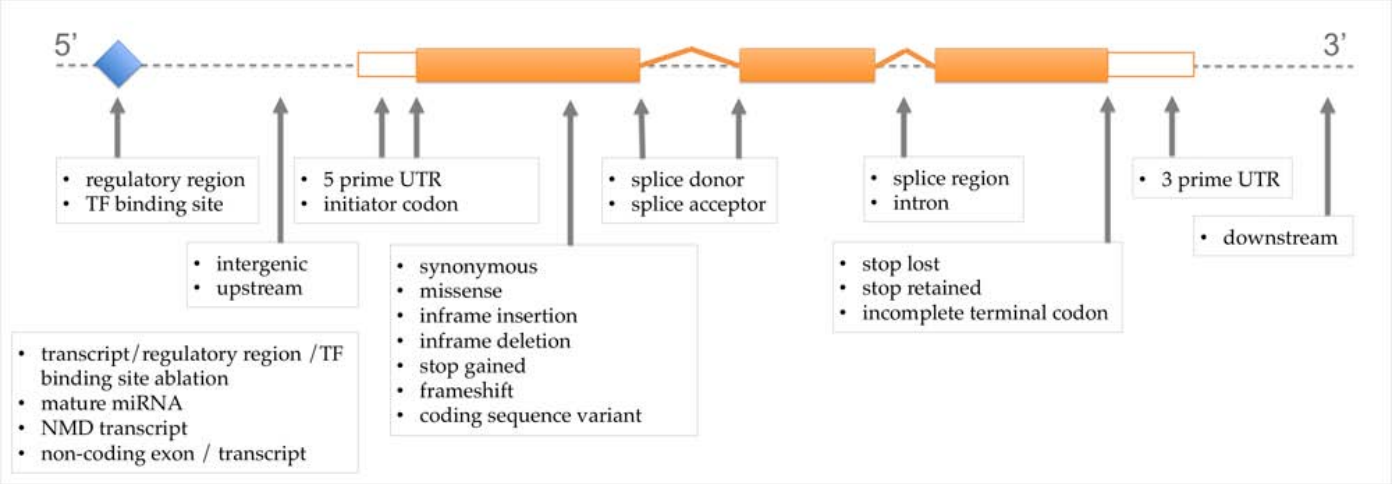
Βήμα 3ο
Σ' αυτό το σημείο γίνεται "στοίχιση" των μορίων cDNA και EST(υποαλληλουχίες cDNA) που προέρχονται μόνο από τον συγκεκριμένο οργανισμό. Όταν υπάρχει αλληλεπικάλυψη ανάμεσα σε μια αλληλουχία cDNA και ένα μεταγραφο που έχει καταχωρηθεί σ' ένα από τα προηγούμενα βήματα ,τα άκρα του cDNA που δεν επικαλύπτονται θα "βλέπουν" στις 3′ και 5′ UTR (μη κωδικές περιοχές) του transcript . Για παράδειγμα :

cDNA : CGATTTG**XXX**…….**XXX**AAAGGCT ,όπου XXX...XXXX η κωδική
transcript: AAUUGGU**XXX**…….**XXX**GGCUTCU περιοχή
5 prime UTR 3 prime UTR

Όταν είναι διαθέσιμες , χρησιμοποιούνται και αλληλουχίες RNA-seq ώστε να χτιστούν τα transcript models.

Βήμα 4ο
Η τελευταία ομάδα των gene predictions λαμβάνεται με τη συγχώνευση πανομοιότυπων transcripts που προήλθαν από διαφορετικές πρωτεϊνικές αλληλουχίες. Αυτό υποδηλώνει alternative splicing και τα transcripts αντιστοιχίζονται σε ένα γονίδιο.

Μετα-επεξεργασία
Όταν ληφθεί το 'τελικό' σύνολο των γονιδίων ,υπόκειται σε μερικές ακόμα διαδικασίες όπως ο σχολιασμός των μη κωδικών RNA περιοχών ,η πρόβλεψη των ψευδογονιδίων και η διασταύρωση με εξωτερικές πηγές. [18]



Εικόνα 3.2 : Διάγραμμα που δείχνει την τοποθεσία κάθε όρου πάνω στο transcript model.

3.3.5 GENCODE

Τα Genebuilds του ανθρώπου και του ποντικιού περιλαμβάνουν και κάποια πρόσθετα βήματα για να παράγουν τα GENCODE gene sets. Όταν είναι διαθέσιμη και η χειροκίνητη επιμέλεια (manual curation) για ένα transcript , συγκρίνονται τα transcript models της HAVANA και της Ensembl και αν είναι πανομοιότυπα, συγχωνεύονται .Αυτός ο συνδυασμός των gene sets της HAVANA και της Ensembl είναι το default gene set του GENCODE project.

3.3.6 Ονόματα και εξωτερικές πηγές

Τα περισσότερα ανθρώπινα γονίδια έχουν και το αντίστοιχο HGNC σύμβολο από την HUGO Επιτροπή Ονοματολογίας Γονιδίων. Αυτά που προέρχονται από το Genebuild (αυτόματο gene prediction) παίρνουν αυτόματα ένα HGNC symbol ενώ όσα προέρχονται από χειροκίνητο annotation αντιστοιχίζονται χειροκίνητα με ένα σύμβολο από την Havana/Vega .Στα transcripts, τα οποία δε μπορούν να συνδεθούν με κάποιο HGNC σύμβολο, αποδίδεται ένα

"clone- based" αναγνωριστικό είτε από την Ensembl είτε από την Havana/Vega .

Κάθε transcript name ακολουθείται από έναν αριθμό. Αν ο αριθμός ξεκινά με '0' (001,002 κτλ), τότε έχει τεθεί από την Havana/Vega. Αν ο αριθμός ξεκινάει με '2' (201,202 κτλ), τότε έχει τεθεί αυτόματα από την Ensembl.[18]

Gene: CXorf40B ENSG00000197021

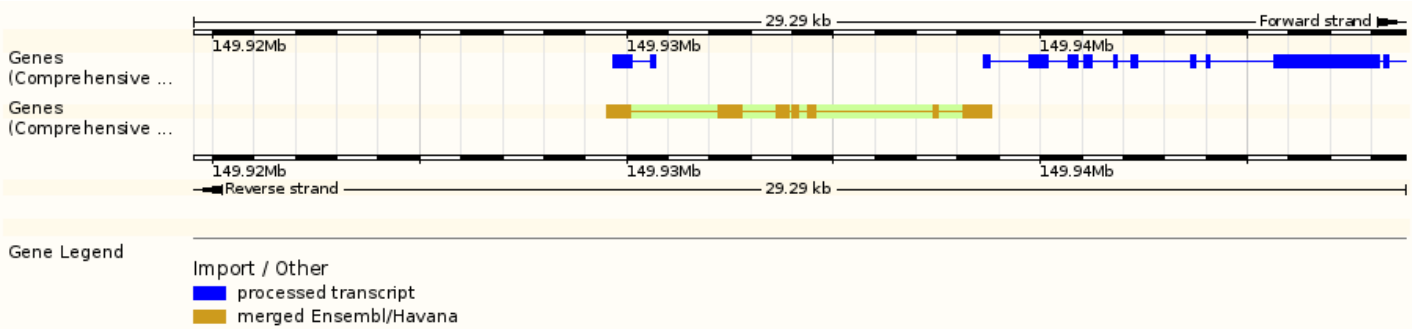
Description	chromosome X open reading frame 40B [Source:HGNC Symbol;Acc:HGNC:17402]
Synonyms	
Location	Chromosome X: 149,929,527-149,938,811 reverse strand.
INSDC coordinates	chromosome:GRCh38:CM000685.2:149929527:149938811:1
Transcripts	This gene has 7 transcripts (splice variants) Hide transcript table

Show/hide columns (1 hidden) Filter

Name	Transcript ID	Length	Protein	Biotype	CCDS	RefSeq	Flags
CXorf40B-001	ENST00000370406	1653 bp	158 aa (view)	Protein coding	CCDS35426	-	GENCODE basic
CXorf40B-006	ENST00000370404	1316 bp	158 aa (view)	Protein coding	CCDS35426	NM_001013845 NP_001013867	GENCODE basic
CXorf40B-002	ENST00000355203	1067 bp	158 aa (view)	Protein coding	CCDS35426	-	GENCODE basic
CXorf40B-007	ENST00000462691	1440 bp	146 aa (view)	Protein coding	-	-	GENCODE basic
CXorf40B-004	ENST00000370409	1129 bp	147 aa (view)	Protein coding	-	-	CDS 3' incomplete
CXorf40B-005	ENST00000483447	808 bp	45 aa (view)	Protein coding	-	-	CDS 3' incomplete
CXorf40B-003	ENST00000497550	591 bp	No protein product	Processed transcript	-	-	

Εικόνα 3.3

Στην Εικόνα 3.3 παρουσιάζεται το γονίδιο με Ensembl ID ENSG00000197021 και gene name CXorf40B (HGNC Symbol) με την περιγραφή του , την τοποθεσία του , τις συντεταγμένες του στο INSDC καθώς και ο πίνακας με τα transcripts του και διάφορες πληροφορίες γι' αυτά (Ensembl IDs , μήκος , ποια πρωτεΐνη παράγουν- αν παράγουν κτλ)



Εικόνα 3.4

Στην Εικόνα 3.4 βλέπουμε σχηματισμένες με μαύρο χρώμα τις δυο αλυσίδες του DNA (forward και reversed strand) και τρία γονίδια(δύο στην forward αλυσίδα και ένα στη reversed).Τα κουτάκια που είναι γεμισμένα με χρώμα συμβολίζουν τα εξώνια ενώ οι γραμμές που τα ενώνουν ,τα εσώνια. Το μπλε χρώμα σημαίνει ότι το γονίδιο έχει αναλυθεί από την HAVANA (manual annotation) ενώ το χρυσό ότι έχουν συγχωνευτεί οι αναλύσεις από την Ensembl και την HAVANA (merged - περιέχει και automatic και manual annotation).Επιλέγοντας κάποιο γονίδιο , ο χρήστης μπορεί να δει πληροφορίες όπως το που ακριβώς βρίσκεται (πχ Chromosome X: 149,929,527-149,938,811) ή το gene type του (protein coding , antisense κτλ).

3.4 Comparative Genomics by Ensembl

Η Ensembl επικεντρώνεται σε δυο βασικούς τομείς της σύγκρισης γονιδιωμάτων :

1. στη δημιουργία Γονιδιακών Δέντρων (Gene Trees) που βασίζονται σε αντιπροσωπευτικές πρωτεΐνες για κάθε είδος. Εδώ γίνεται σύγκριση της αλληλουχίας των γονιδίων ανάμεσα στους οργανισμός και δημιουργούνται τα gene trees.
2. στην ευθυγράμμιση αλληλουχιών DNA για να βρεθεί κατά πόσο μια αλληλουχία (ένα γονίδιο) διατηρείται ανάμεσα σε διαφορετικούς οργανισμούς και σε ποιους.

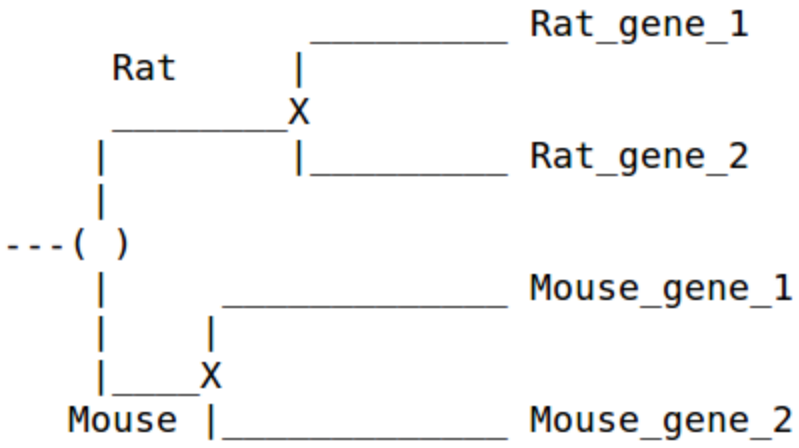
Η παραπάνω ανάλυση γίνεται για όλους τους οργανισμούς που περιέχονται στη βάση δεδομένων της Ensembl.

3.4.1 Homology (Ομολογία)

Στη Βιολογία , ο όρος "Ομολογία" αναφέρεται στην ύπαρξη ενός κοινού προγόνου ανάμεσα σε ένα ζεύγος γονιδίων , πρωτεϊνών ή άλλων δομών. [19] Ένα παράδειγμα ομολογίας είναι τα φτερά των πουλιών και τα φτερά της νυχτερίδας.

Η Ομολογία χωρίζεται σε δυο υποκατηγορίες:

- όταν είναι αποτέλεσμα διπλασιασμού ενός γονιδίου και τα δυο αντίγραφα που προκύπτουν εξελίσσονται παράλληλα στην ιστορία ενός οργανισμού πρόκειται για **παράλογα (paralogous) γονίδια.** (βρίσκονται στον ίδιο οργανισμό)
- όταν είναι αποτέλεσμα ειδογένεσης (εξελικτική διαδικασία με την οποία προκύπτουν νέα βιολογικά είδη) πρόκειται για **ορθόλογα γονίδια (orthologous).** (βρίσκονται σε διαφορετικούς οργανισμούς) [20]



Εικόνα 3.5

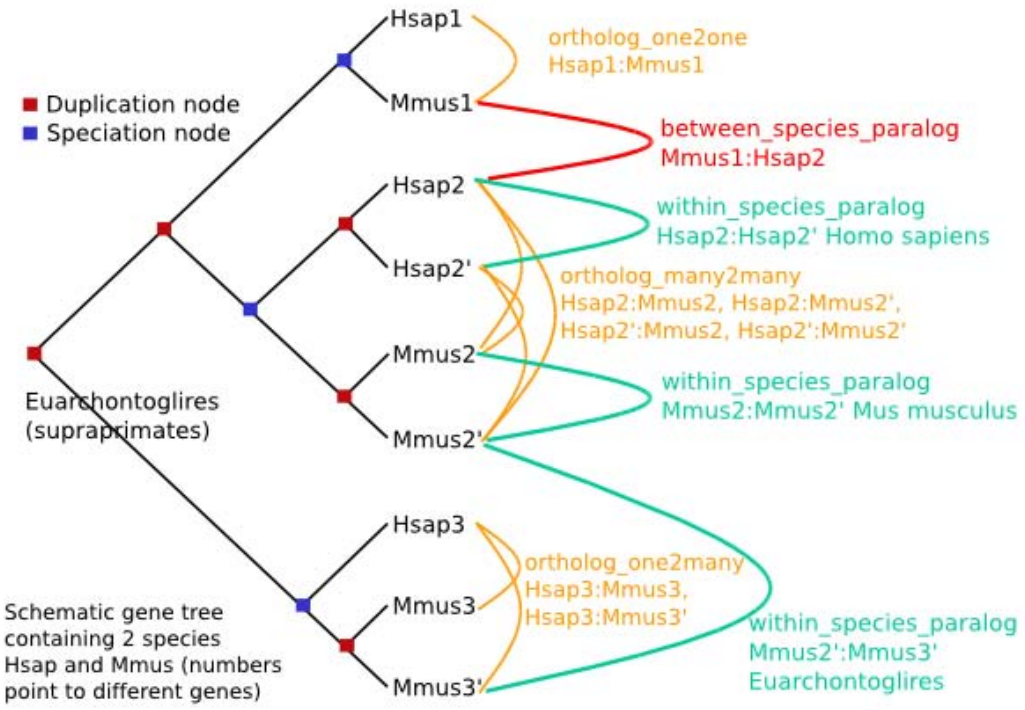
Στην παραπάνω εικόνα η παρένθεση () συμβολίζει ειδογένεση ενώ το 'X' συμβολίζει διπλασιασμό γονιδίου. Τα γονίδια Mouse_gene_1 και Mouse_gene_2 είναι μεταξύ τους παράλογα. Το ίδιο ισχύει και για τα Rat_gene_1 και Rat_gene_2. Αντίστοιχα , τα δυο γονίδια του ποντικιού είναι ορθόλογα με τα δυο γονίδια του αρουραίου. [21]

3.4.2 Πρωτεϊνικά Δέντρα στην Ensembl

Για την σύγκριση των γονιδιωμάτων των διάφορων οργανισμών δημιουργούνται τα φυλογενετικά δέντρα. Τα δέντρα αυτά αναπαριστούν κάθε τεκμηριωμένη εξελικτική σχέση μεταξύ των ειδών με βάση τις ομοιότητες και τις διαφορές στα φυσικά και στα γενετικά τους χαρακτηριστικά. Φυλογενετικά δέντρα δημιουργούνται τόσο για όλα τα γονίδια που κωδικοποιούν πρωτεΐνη όσο και για τα ncRNA (non coding RNA) γονίδια (υπολογίζονται σε δυο διαφορετικά pipelines).

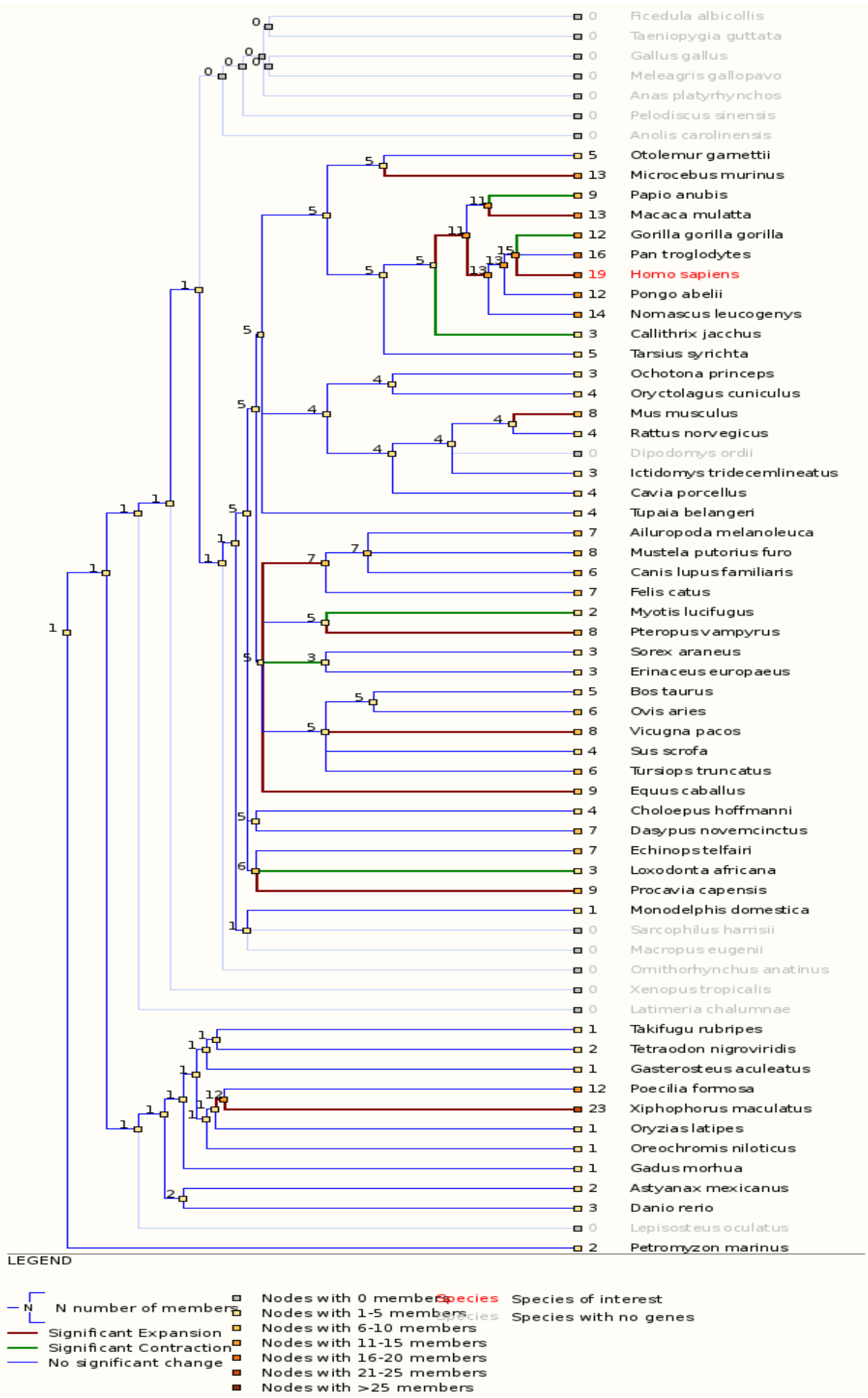
Η πρόβλεψη των ορθόλογων και παράλογων γονιδίων γίνεται μέσα από ένα pipeline όπου τα φυλογενετικά δέντρα παίζουν μεγάλο

ρόλο. Στόχος είναι η αναπαράσταση της ιστορίας της εξέλιξης κάθε οικογένειας γονιδίων, δηλαδή γονιδίων που προέρχονται από έναν κοινό πρόγονο. [18]



Εικόνα 3.6: Αναπαράσταση των σχέσεων των γονιδίων μέσω των Gene και Species Trees. Παράλογα και ορθόλογα γονίδια και η αντιστοιχία τους (1:1 , 1:N, N:N).Οι μπλε και κόκκινοι κόμβοι παριστάνουν την ειδογένεση και τον διπλασιασμό γονιδίου αντίστοιχα.

Στην παρακάτω εικόνα (Εικόνα 3.7) παρατηρούμε ένα Species Tree (Gain/Loss Tree) για το γονίδιο ZNF235 (ENSG00000159917).Η κόκκινη ακμή συμβολίζει τη σημαντική επέκταση ενός γονιδίου (αύξηση των μελών της οικογένειας-gene family), η πράσινη τη σημαντική "συστολή" ενός γονιδίου (μείωση των μελών της οικογένειας) ενώ η έλλειψη σημαντικών αλλαγών συμβολίζεται με μπλε ακμή. Οι κόμβοι κάθε υπάρχοντα οργανισμού ή των προγόνων του επισημαίνονται με τον αριθμό των μελών της οικογένειας .


Ficedula albicollis
Taeniopygia guttata
Gallus gallus
Meleagris gallopavo
Anas platyrhynchos
Pelodiscus sinensis
Anolis carolinensis
Otolemur garnettii
Microcebus murinus
Papio anubis
Macaca mulatta
Gorilla gorilla gorilla
Pan troglodytes
Homo sapiens
Pongo abelii
Nomascus leucogenys
Callithrix jacchus
Tarsius syrichta
Ochotona princeps
Oryctolagus cuniculus
Mus musculus
Rattus norvegicus
Dipodomys ordii
Ictidomys tridecemlineatus
Cavia porcellus
Tupaia belangeri
Ailuropoda melanoleuca
Mustela putorius furo
Canis lupus familiaris
Felis catus
Myotis lucifugus
Pteropus vampyrus
Sorex araneus
Erinaceus europaeus
Bos taurus
Ovis aries
Vicugna pacos
Sus scrofa
Tursiops truncatus
Equus caballus
Choloepus hoffmanni
Dasypus novemcinctus
Echinops telfairi
Loxodonta africana
Procavia capensis
Monodelphis domestica
Sarcophilus harrisii
Macropus eugenii
Ornithorhynchus anatinus
Xenopus tropicalis
Latimeria chalumnae
Takifugu rubripes
Tetraodon nigroviridis
Gasterosteus aculeatus
Poecilia formosa
Xiphophorus maculatus
Oryzias latipes
Oreochromis niloticus
Gadus morhua
Astyanax mexicanus
Danio rerio
Lepisosteus oculatus
Petromyzon marinus
LEGEND
N number of members
Significant Expansion
Significant Contraction
No significant change
Nodes with 0 members
Nodes with 1-5 members
Nodes with 6-10 members
Nodes with 11-15 members
Nodes with 16-20 members
Nodes with 21-25 members
Nodes with >25 members
Species Species of interest
Species Species with no genes

3.5 Πολυμορφισμός (Variation Data)

Στη βάση δεδομένων της Ensembl για τον πολυμορφισμό αποθηκεύονται περιοχές του γονιδιώματος οι οποίες εμφανίζονται σε διάφορες παραλλαγές (variants) στον ίδιο οργανισμό καθώς και πληροφορίες για τον φαινότυπο και τις ασθένειες που σχετίζεται μ'αυτές τις παραλλαγές.[18]

3.5.1 Είδη πολυμορφισμού

Οι βασικοί τύποι των παραλλαγών (variants) για τους διάφορους οργανισμούς είναι οι εξής :

- πολυμορφισμός σε ένα μόνο νουκλεοτίδιο (Single Noucleotide Polymorphism)
- προσθήκη ή/και διαγραφή μιας μικρής αλληλουχίας νουκλεοτιδίων
- παραλλαγές σε μεγαλύτερες αλληλουχίες νουκλεοτιδίων ,οι οποίες καταχωρούνται ως "δομικές" (structural variants).

Οι πρώτες δυο κατηγορίες παραθέτονται μαζί με παραδείγματα στον παρακάτω πίνακα της Εικόνας 3.8 :

Sequence variants

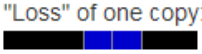
Type	Description	Example (Reference / Alternative)	
SNP	Single Nucleotide Polymorphism	Ref: ...TTG**A**CGTA...	Alt: ...TTG**G**CGTA...
Insertion	Insertion of one or several nucleotides	Ref: ...TTGACGTA...	Alt: ...TTGA**TG**CGTA...
Deletion	Deletion of one or several nucleotides	Ref: ...TTG**AC**GTA...	Alt: ...TTGGTA...
Indel	An insertion and a deletion, affecting 2 or more nucleotides	Ref: ...TTG**A**CGTA...	Alt: ...TTG**GCT**CGTA...
Substitution	A sequence alteration where the length of the change in the variant is the same as that of the reference.	Ref: ...TTG**AC**GTA...	Alt: ...TTG**TA**GTA...

Εικόνα 3.8: πολυμορφισμοί σε ένα νουκλεοτίδιο ή σε μια μικρή αλληλουχία νουκλεοτιδίων

Οι δομικές παραλλαγές εμφανίζονται στην Εικόνα 3.9 και περιλαμβάνουν την αύξηση/μείωση του αριθμού τον αντιγράφων μιας

αλληλουχίας (CNV) , την αναστροφή της (inversion) καθώς και την μετατόπιση της σε μια νέα θέση (translocation).

Structural variants

Type	Description	Example (Reference / Alternative)	
CNV	Copy Number Variation: increases or decreases the copy number of a given region	Reference: 	"Gain" of one copy:  "Loss" of one copy: 
Inversion	A continuous nucleotide sequence is inverted in the same position	Reference: 	Alternative: 
Translocation	A region of nucleotide sequence that has translocated to a new position	Reference: 	Alternative: 

Εικόνα 3.9: πολυμορφισμοί σε μεγαλύτερες αλληλουχίες νουκλεοτιδίων - Structural Variants

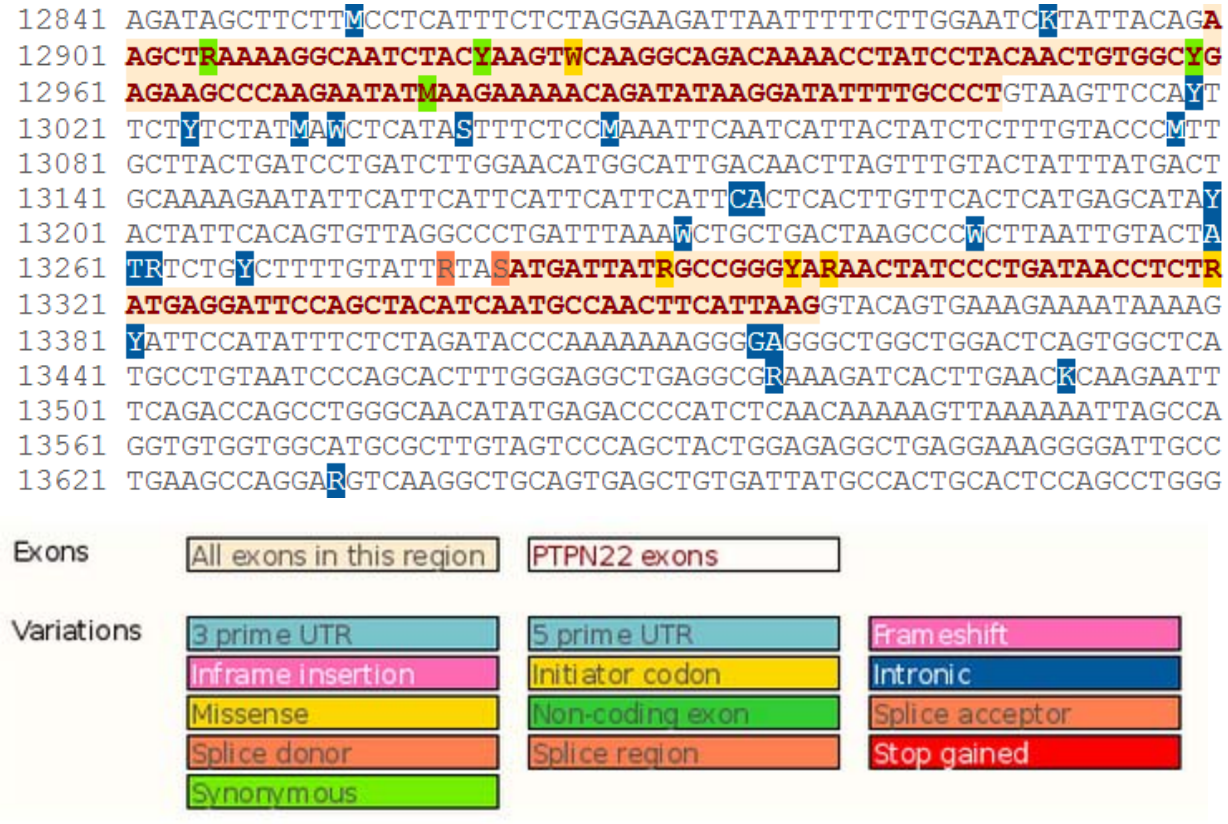
Για κάθε πολυμορφισμό παρέχεται πληροφορία σχετικά με τα variants από τα οποία αποτελείται, τα γονίδια/μετάγραφα τα οποία επηρεάζει αλλά και τους πληθυσμούς στους οποίους εμφανίζεται πιο συχνά. Ενώ προβλέπονται όλες οι επιδράσεις που προκαλούν οι παραλλαγές των μεταγράφων και των ρυθμιστικών παραγόντων για κάθε είδος της Ensembl, ο χρήστης μπορεί να αναλύσει τα δικά του δεδομένα χρησιμοποιώντας το εργαλείο **Variant Effect Predictor.**

3.5.2 Variant Effect Predictor

Ο Variant Effect Predictor καθορίζει τις συνέπειες των παραλλαγών που θέτει σαν ορίσματα ο χρήστης (SNPs, προσθήκη/διαγραφή αλληλουχίας, CNVs και δομικές παραλλαγές) στα γονίδια ,στα μετάγραφα ,σε πρωτεϊνικές αλληλουχίες καθώς και στους ρυθμιστικούς παράγοντες .Ο χρήστης εισάγει τις **συντεταγμένες των παραλλαγών** και τις **αλλαγές των νουκλεοτιδίων** και το εργαλείο του επιστρέφει :

- τα γονίδια και τα μετάγραφα που επηρεάζονται από αυτές τις παραλλαγές

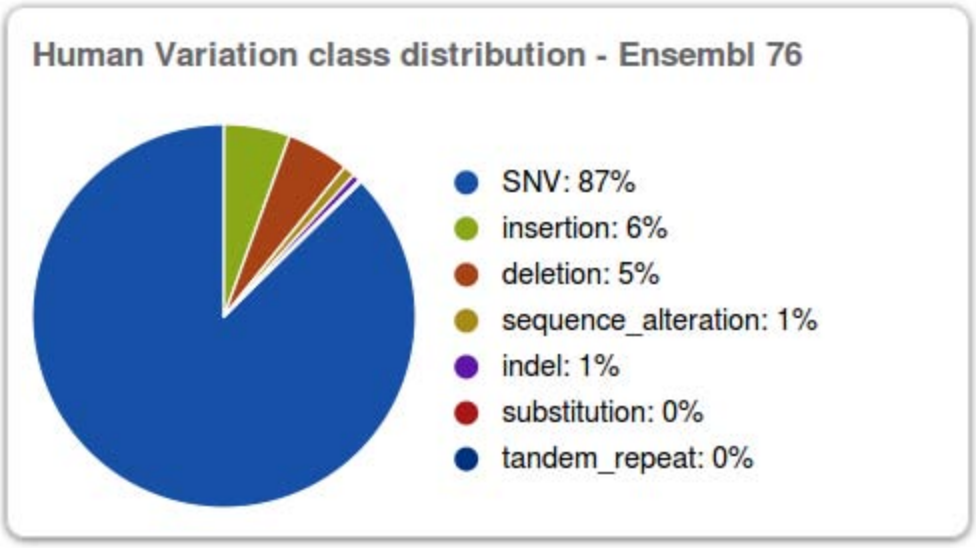
- την θέση των παραλλαγών (πχ αν "πέφτουν" πάνω σ' ένα μετάγραφο , αν είναι σε coding περιοχή ,αν βρίσκονται πάνω σε non-coding RNA ή σε ρυθμιστική περιοχή)
- τις συνέπειες των παραλλαγών στην πρωτεϊνική αλληλουχία (πχ δημιουργία ή διαγραφή ενός κωδικονίου λήξης της μετάφρασης , παραγωγή λάθος αμινοξέους)
- γνωστές παραλλαγές που ταιριάζουν με τις παραλλαγές εισόδου.



Εικόνα 3.10:Τμήμα της αλληλουχίας ενός γονιδίου που δείχνει τις θέσεις των SNPs.

Στην παραπάνω εικόνα οι μαρκαρισμένες περιοχές αποτελούν τα εξώνια, ενώ οι αμαρκάριστες τα εσώνια. Το μπλε χρώμα συμβολίζει ότι ο SNP βρίσκεται πάνω σ' ένα εσώνιο, το πράσινο ότι βρίσκεται σε ένα εξώνιο που δεν κωδικοποιεί πρωτεΐνη ,το λαχανί

ότι το SNP έχει σαν συνέπεια την παραγωγή του ίδιου αμινοξέους ,ενώ τα πορτοκαλί ότι βρίσκεται σε Splice donor ή Splice region (αλληλουχίες που ενώνουν τα εσώνια με τα εξώνια ,βλ. Εικόνα 3.2).Τέλος το γαλάζιο σημαίνει ότι ο SNP βρίσκεται σε 3′ ή 5′ μη κωδική περιοχή , το κίτρινο ότι παράγεται διαφορετικό αμινοξύ από το προβλεπόμενο και το κόκκινο ότι ο SNP δημιουργεί ένα κωδικόνιο λήξης οπότε προκαλεί διακοπή της μετάφρασης της πρωτεΐνης. [22]



Εικόνα 3.11: Κατανομή των ειδών πολυμορφισμού στο ανθρώπινο γονιδίωμα σύμφωνα με την Ensembl 76.

3.6 Δεδομένα Γονιδιακής Ρύθμισης (Regulation Data)

Η Ensembl παρέχει πληροφορίες που περιγράφουν τους μηχανισμούς της ρύθμισης των γονιδίων στα κύτταρα του ανθρώπου και του ποντικιού. Συγκεκριμένα, επικεντρώνεται στην ανάλυση των μεταγραφικών και μετα-μεταγραφικών παραγόντων, δηλαδή των ειδικών πρωτεϊνών που συμμετέχουν στην ενεργοποίηση ή στην καταστολή της γονιδιακής έκφρασης.

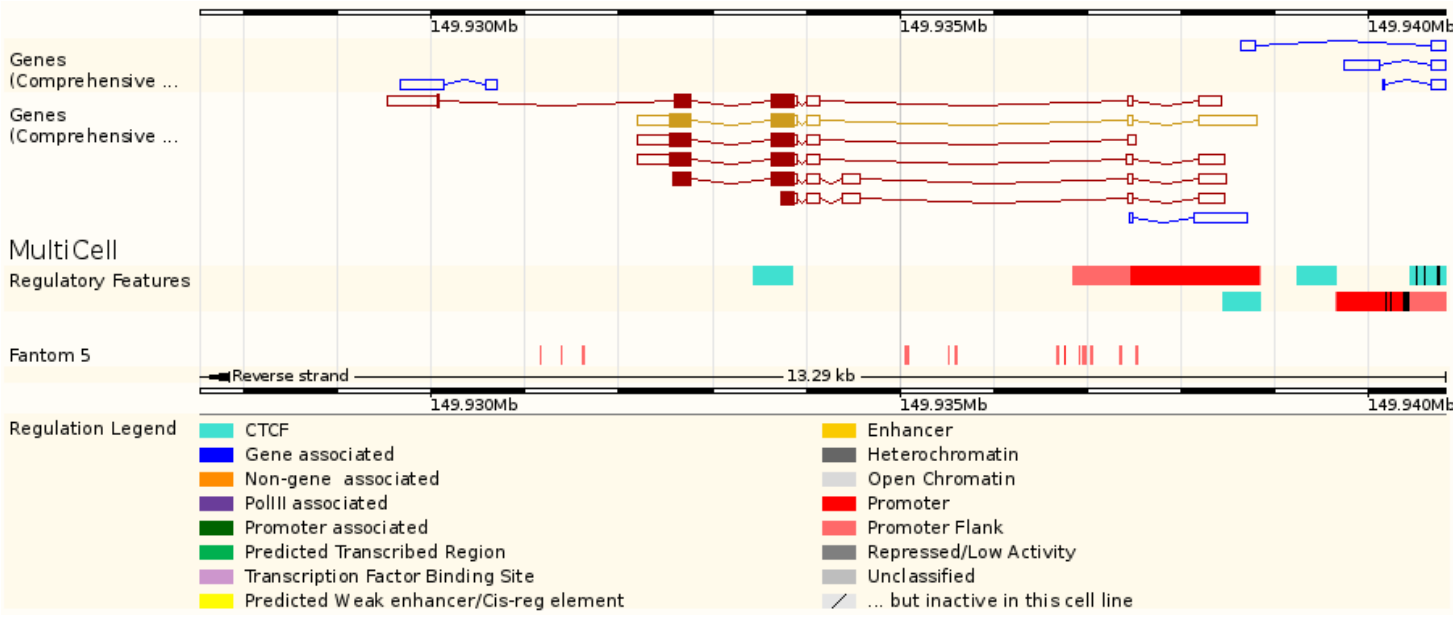
3.6.1 Regulatory Features

Σε μια βάση δεδομένων της Ensembl, διατηρούνται όλες οι περιοχές (στο γονιδίωμα του ανθρώπου και του ποντικιού) που θα μπορούσαν να εμπλέκονται στη ρύθμιση της μεταγραφής των γονιδίων. Οι περιοχές αυτές, που ονομάζονται ρυθμιστικά στοιχεία, προκύπτουν από κάποια πειραματικά σύνολα δεδομένων, όπως

- η μελέτη των περιοχών *ανοιχτής χρωματίνης* με την μεθοδολογία Faire-Seq και DNase1-Seq,
- η μελέτη της τροποποίησης των ιστόνων και της πρόσδεσης των μεταγραφικών παραγόντων με τη μέθοδο ChIP-Seq και
- η εύρεση των μοτίβων πρόσδεσης των μεταγραφικών παραγόντων. [18]

Βάσει αυτών των δεδομένων και μέσω μιας διαδικασίας που ονομάζεται Regulatory Build ,υπολογίζονται τα ρυθμιστικά στοιχεία, ανεξάρτητα από το είδος και την λειτουργία των κυττάρων. Τα στοιχεία αυτά χωρίζονται στις εξής κατηγορίες :

- πιθανοί υποκινητές (Predicted promoters)
- πιθανοί ενισχυτές (Predicted enhancers)
- θέσεις πρόσδεσης CTCF
- θέσεις πρόσδεσης μεταγραφικών παραγόντων
- περιοχές ανοιχτής χρωματίνης



Εικόνα 3.12: Γονίδια και Ρυθμιστικοί παράγοντες του γονιδίου CXorf40B

3.6.2 Other Regulatory Data

Στη βάση δεδομένων της Ensembl Regulation αποθηκεύονται και άλλα δεδομένα , τα οποία εισάγονται κατευθείαν από εξωτερικές πηγές. Αυτά είναι :

- προβλέψεις στόχων Micro RNA για τον άνθρωπο και το ποντίκι χρησιμοποιώντας το DIANA TarBase
- πειραματικά επικυρωμένοι ενισχυτές στο ανθρώπινο γονιδίωμα χρησιμοποιώντας τον VISTA Browser.[18]

3.7 Ensembl Releases

Η Ensembl παράγει/δημοσιεύει μια καινούρια έκδοση (release) της ιστοσελίδας καθώς και των υποκείμενων βάσεων δεδομένων της κάθε δυο με τρεις μήνες , κάνοντας διαθέσιμα νέα δεδομένα και αναλύσεις μετά από αυστηρό έλεγχο.

3.7.1 Γενικές διαφορές ανάμεσα στις Ensembl Releases

Μια καινούρια έκδοση μπορεί να περιλαμβάνει νέα ή/και ανανεωμένα δεδομένα όπως :

- νέους οργανισμούς
- νέες assemblies για ήδη υπάρχοντες οργανισμούς
- ανανεωμένα gene sets
- νέα δεδομένα πολυμορφισμού (new variation data)
- κατασκευή νέων γονιδιακών δέντρων (new gene trees)
- νέα alignments
- νέες ομολογίες (new homologies)
- ανανεωμένο σχολιασμό ρυθμιστικών στοιχείων (regulation features)
- βελτιώσεις και προσθήκες στο web-interface και στην ιεραρχία της βάσης δεδομένων της Ensembl.

3.7.2 Παλιότερες Ensembl Releases (Archive)

Η βασική ιστοσελίδα της Ensembl (www.ensembl.org) ανανεώνεται με τα τελευταία δεδομένα σε μηνιαία βάση

Οι χρήστες μπορούν να έχουν πρόσβαση στις ιστοσελίδες των παλιότερων εκδόσεων χρησιμοποιώντας ένα stable link που υπάρχει για κάθε έκδοση και το οποίο διαρκεί για περίπου 5 χρόνια (Για παράδειγμα αυτή τη στιγμή υπάρχει ένα stable link για κάθε έκδοση της Ensembl από το 2009 και μετά , *Ensembl 54: May 2009*).

Το όφελος της πρόσβασης στα Archive Releases είναι ότι οι χρήστες μπορούν να παρακολουθούν τις αλλαγές ανάμεσα στις διαφορετικές εκδόσεις. Για παράδειγμα ,αν ο χρήστης έχει προσκομίσει δεδομένα από την Ensembl 71 (Απρίλιος του 2011) μπορεί να ανατρέξει στο αντίστοιχο site της και να βρει αυτά τα δεδομένα. Για τον λόγο αυτό ,είναι χρήσιμο να σημειώνεται και ο αριθμός της έκδοσης της Ensembl όταν χρησιμοποιούνται οι βάσεις δεδομένων της.

Συμπληρωματικά , στο site της Ensembl παρέχεται και ένας πίνακας με τις assemblies κάθε οργανισμού που χρησιμοποιήθηκαν σε κάθε release [23].

3.7.3 Αλλαγές σε Stable IDs και gene names

Το σύστημα του Ensembl Annotation χρησιμοποιεί ένα σύνολο σταθερών αναγνωριστικών (stable IDs) ,καθένα από τα οποία έχει πρόθεμα βασισμένο στο επιστημονικό όνομα του οργανισμού [24] (πχ ENS για τον άνθρωπο , ENSMU για το ποντίκι) καθώς και στο είδος του στοιχείου που αντιπροσωπεύει (πχ ENSG για γονίδιο , ENST για transcript , ENSP για πρωτεΐνη).Το πρόθεμα ακολουθούν κάποιοι αριθμοί (πχ ENSG00000139618).

Η Ensembl έχει ως στόχο να διατηρεί σταθερά τα stable IDs για τα γονίδια (ENSG),τα μετάγραφα (ENST), τις πρωτεΐνες (ENSP) και τα εξώνια (ENSE), όσο αυτό είναι εφικτό. Κάποιες όμως αλλαγές στην assembly της αλληλουχίας του γονιδιώματος ή η ανανέωση του annotation μπορεί να αλλάξει δραματικά το γονιδιακό μοντέλο. Σ' αυτές τις περιπτώσεις ,το παλιό σύνολο των stable IDs αποσύρεται και ορίζεται ένα καινούριο.

Στα stable IDs εκχωρούνται versions έτσι ώστε για "μικρές" αλλαγές να μη χρειάζεται να αλλάξει το stable ID του στοιχείου, απλώς να αυξηθεί η version κατά ένα (πχ ENSG00000139618.1 → ENSG00000139618.2).

Τα κριτήρια για αύξηση της version ενός stable ID είναι τα εξής [25]:

1. Για εξώνια:
   - καμία αλλαγή στη θέση ή στην αλληλουχία → καμία αύξηση
     old → θέση: 1000-2000, αλληλουχία : "ATGG...GTA"
     new → θέση: 1000-2000, αλληλουχία : "ATGG...GTA"

   - αλλαγή στη θέση , ίδια αλληλουχία → καμία αύξηση
     old → θέση: 1000-2000, αλληλουχία : "ATGG...GTA"
     new → θέση: 1100-2100, αλληλουχία : "ATGG...GTA"

   - αλλαγή στη θέση , μικρή αλλαγή αλληλουχία → αύξηση
     old → θέση: 1000-2000, αλληλουχία : "**A**TGG...GTA"
     new → θέση: 1100-2100, αλληλουχία : "**T**TGG...GTA"

   - καμία αλλαγή στη θέση , μικρή αλλαγή αλληλουχία → αύξηση
     old → θέση: 1000-2000, αλληλουχία : "**A**TGG...GTA"
     new → θέση: 1000-2000, αλληλουχία : "**T**TGG...GTA"

2. Για transcripts:
   - καμία αλλαγή στα εξώνια → καμία αύξηση
     old → exon1 (θέση: 100-199, αλληλουχία : "ATG...GTA")
           exon2 (θέση: 300-399, αλληλουχία : "ACT...TAA")
     new → exon1 (θέση: 100-199, αλληλουχία : "ATG...GTA")
           exon2 (θέση: 300-399, αλληλουχία : "ACT...TAA")

   - μικρή αλλαγή στα εξώνια → αύξηση
     old → exon1 (θέση: 100-199, αλληλουχία : "ATG...GTA")
           exon2 (θέση: 300-399, αλληλουχία : "AC**T**...TAA")
     new → exon1 (θέση: 100-199, αλληλουχία : "ATG...GTA")
           exon2 (θέση: 300-399, αλληλουχία : "AC**C**...TAA")

   Στη δεύτερη περίπτωση , αντί για αύξηση του version θα μπορούσε απλώς να προστεθεί η πληροφορία διόρθωσης της βάσης 'C' σε 'T' στο 2ο εξώνιο του νέου transcript και η version να παραμείνει η ίδια (seq_edit : "302>T" , όπου 302 η θέση της "λανθασμένης" βάσεως).

3. Για translations:
   - καμία αλλαγή στην αλληλουχία των αμινοξέων → καμία αύξηση
     old → αλληλουχία : "MVTK"
     new → αλληλουχία : "MVTK"

- αλλαγή στην αλληλουχία των transcripts αλλά όχι στην αλληλουχία των αμινοξέων (λόγω συνωνυμίας) → καμία αύξηση
  old → αλλ. αμινοξέων : "MVTK" ,transcript : "ATGGTCACA**AAG**"
  new → αλλ. αμινοξέων : "MVTK" ,transcript : "ATGGTCACA**AAA**"

- αλλαγή στην αλληλουχία των transcripts και στην αλληλουχία των αμινοξέων → αύξηση
  old → αλλ. αμινοξέων : "MVT**K**" ,transcript : "ATGGTCACA**AAG**"
  new → αλλ. αμινοξέων : "MVT**N**" ,transcript : "ATGGTCACA**AAC**"

Στην τρίτη περίπτωση , αντί για αύξηση του version θα μπορούσε απλώς να προστεθεί η πληροφορία διόρθωσης του αμινοξέους 'N' με 'K' στο νέο translation και η version να παραμείνει η ίδια (seq_edit : "4>K" , όπου 4 η θέση του "λανθασμένου" αμινοξέους.

4. Για γονίδια:
   - καμία αλλαγή στα transcripts του → καμία αύξηση
     old → transcr_stable_id1: ENST1.1 ,transcr_stable_id2: ENST2.1
     new → transcr_stable_id1: ENST1.1 ,transcr_stable_id2: ENS2.1

   - αλλαγή της version κάποιου transcript → αύξηση
     old → transcr_stable_id1: ENST1.1 ,transcr_stable_id2: ENST2.**1**
     new → transcr_stable_id1: ENST1.1 ,transcr_stable_id2: ENS2.**2**

   - αλλαγή του stable id κάποιου transcript → αύξηση
     old → transcr_stable_id1: **ENST1**.1 ,transcr_stable_id2: ENST2.1
     new → transcr_stable_id1: **ENST3**.1 ,transcr_stable_id2: ENST2.1

Υπάρχουν όμως και αλλαγές ,οι οποίες έχουν πολύ μεγαλύτερη επίπτωση σε ένα γονίδιο ή σε ένα μετάγραφο ,δηλαδή μπορούν να αλλάξουν σε μεγάλο βαθμό την αλληλουχία των νουκλεοτιδίων.

Οι αλλαγές αυτές οφείλονται κυρίως στις δυο παρακάτω περιπτώσεις :

- αν ένα γονίδιο/μετάγραφο χωριστεί σε δυο ή περισσότερα (**split event**)
- αν ένα ή περισσότερα γονίδια/μετάγραφα συγχωνευτούν σε ένα (**merge event**).

Έτσι, ένα stable ID μπορεί να αποσυρθεί ή να αντιστοιχεί πλέον σε κάποιο άλλο γονίδιο/transcript. Επίσης ,μετά από ένα split event δημιουργούνται νέα IDs που αντιστοιχούν στα νέα γονίδια/transcripts.

Σχετικά με τα gene names ,όπως αναφέρθηκε και στην ενότητα 3.3.6, στα γονίδια που προέρχονται από αυτόματο annotation αντιστοιχίζονται αυτόματα HGNC symbols. Τα gene names των υπόλοιπων γονιδίων προέρχονται από τη βάση δεδομένων Havana/Vega. Όταν ένα γονίδιο αποσύρεται (λόγω ενός split event) , αποσύρεται μαζί και το gene name του. Για λόγους ανάκτησης δεδομένων , η Ensembl αποθηκεύει το gene name που αποσύρεται , ως συνώνυμο των gene names των "διαδόχων" του γονιδίου που χωρίστηκε (split event).Εναλλακτικά, μπορεί κάποιος να ανατρέξει στα site των αντίστοιχων βάσεων δεδομένων για να βρει το ιστορικό κάποιου gene name που δε χρησιμοποιείται πια.

3.7.4 ID History Converter

Ο ID History Converter είναι ένα εργαλείο της Ensembl , το οποίο δέχεται σαν είσοδο μια λίστα από Ensembl stable IDs (Γονιδίων, Μεταγράφων και Πρωτεϊνών) και επιστρέφει το ιστορικό τους.

Συγκεκριμένα , επιστρέφει για το κάθε ID :

- πότε (σε ποια Release της Ensembl) άλλαξε version (πχ από ENSG00000187888.1 σε ENSG00000187888.2)
- πότε άλλαξε το ίδιο το ID (πχ από ENSG00000161853.3 σε ENSG00000237900.1,)
- ένα mapping score που συμβολίζει τη σχέση ανάμεσα στο παλιό και στο νέο ID.

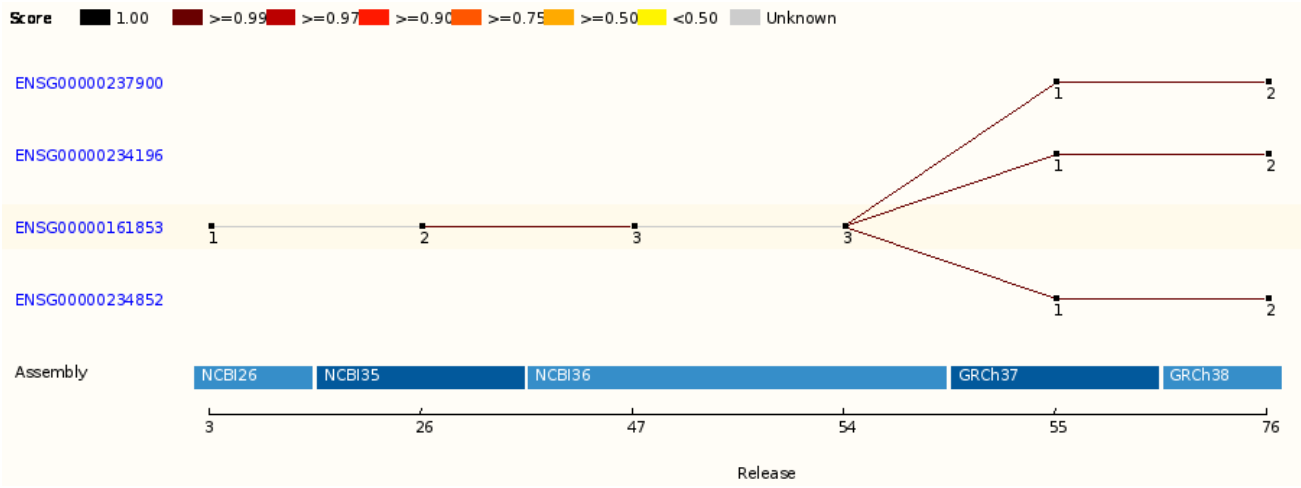
Για παράδειγμα , με είσοδο το Gene ID **ENSG00000161853,** η έξοδος είναι :

Old stable ID	New stable ID	Release	Mapping score
ENSG00000161853.1	ENSG00000161853.2	26	0
ENSG00000161853.2	ENSG00000161853.3	47	0.997334
ENSG00000161853.3	<retired>	55	0
ENSG00000161853.3	ENSG00000234196.1	55	0.997326
ENSG00000161853.3	ENSG00000234852.1	55	0.997326
ENSG00000161853.3	ENSG00000237900.1	55	0.997326

Από τον παραπάνω πίνακα συμπεραίνουμε ότι :

1) η version του ID άλλαξε από 1 σε 2 στην έκδοση 26 και από 2 σε 3 στην 47.
2) Στην έκδοση 55, το ID αποσύρθηκε και δημιουργήθηκαν 3 νέα stable IDs, γεγονός που υποδηλώνει ότι το γονίδιο με ID ENSG00000161853 χωρίστηκε σε 3 νέα γονίδια, τα ENSG00000234196, ENSG00000234852 και ENSG00000237900 (split event).

Αντίστοιχα ,μπορούμε να συμπεράνουμε και τα merge events.



Εικόνα 3.13 : Γραφική αναπαράσταση του split event του γονιδίου.

4.Στόχος και Υλοποίηση

4.1 Αναγκαιότητα της διπλωματικής εργασίας

Ενώ με τον ID History Converter μπορεί κάθε χρήστης να ανακτήσει πληροφορίες σχετικές με το ιστορικό ενός stable ID , δεν υπάρχει κάποιο πιο άμεσο εργαλείο για την έυρεση των "απογόνων" (αν υπάρχουν) του ID στην τρέχουσα έκδοση της Ensembl. Μετά από συζήτηση με το DIANA LAB (βλέπε κεφάλαιο 4.1.1), κρίθηκε αναγκαία η ανάπτυξη μιας εφαρμογής η οποία θα δέχεται μια λίστα με παλιά (και όχι μόνο) stable IDs και θα επιστρέφει την αντιστοίχισή τους (mapping) στην τρέχουσα Ensembl Release.

4.1.1 DIANA LAB (DNA Intelligent Analysis)

Τα υπολογιστικά μοντέλα πρόβλεψης αλληλουχιών DNA έχουν πλέον καθοριστικό ρόλο στα σύγχρονα βιολογικά συστήματα. Η έρευνα του DIANA lab επικεντρώνεται στην ανάπτυξη αλγορίθμων, βάσεων δεδομένων και εργαλείων για την ερμηνεία και την αρχειοθέτηση γονιδιακών δεδομένων , στο πλαίσιο μιας συστηματικής ανάλυσης.

Το ενδιαφέρον του εργαστηρίου επικεντρώνεται στην ανάλυση του microRNA(miRNA) καθώς και σε γονίδια που κωδικοποιούν πρωτεΐνη. Τα microRNA, όπως έχει αποδειχθεί τα τελευταία χρόνια, υπάρχουν σε άφθονη ποσότητα στο γονιδίωμα των θηλαστικών και αποτελούν κομβικό στοιχείο στη ρύθμιση της ανάπτυξής τους.

Οι δραστηριότητες του DIANA lab κυμαίνονται από την ανάλυση της ρύθμισης της γονιδιακής έκφρασης μέσω δεδομένων αλληλουχιών , τον σχολιασμό των στόχων microRNA και των ρυθμιστικών στοιχείων μέχρι την ερμηνεία του ρόλου του microRNA σε διάφορες ασθένειες. [26]

4.1.2 Αλλαγή Ensembl stable ID – Προβλήματα που δημιουργούνται

Το DIANA lab, όπως και άλλα ερευνητικά κέντρα, στηρίζουν τα εργαλεία τους πάνω στο annotation της Ensembl. Αν λοιπόν κάποιο εργαλείο χρησιμοποιεί ένα σύνολο από stable IDs τα οποία έχουν ληφθεί από μια παλιότερη έκδοση της Ensembl και κάποια από αυτά έχουν καταργηθεί ή αντιπροσωπεύουν άλλα στοιχεία στην τρέχουσα έκδοση , η βάση δεδομένων των εργαλείων πρέπει να ανανεωθεί συμπεριλαμβάνοντας την καινούρια πληροφορία.
Ένα παράδειγμα:
Κάποιο εργαλείο, το οποίο αναπτύχθηκε το 2007, έχει βασιστεί σε δεδομένα της έκδοσης 47 της Ensembl και χρησιμοποιεί το Ensembl ID **ENSG00000161853.** Το ID αυτό, δεν υπάρχει στην τρέχουσα έκδοση της Ensembl άρα πρέπει να ανανεωθεί και να αντικατασταθεί με τους απογόνους του ,δηλαδή τα IDs **ENSG00000234196, ENSG00000234852** και **ENSG00000237900** (Εικόνα 3.13). Αυτό μπορεί να γίνει με δυο τρόπους:

1. Ο χρήστης πρέπει να ανατρέξει στο site της Ensembl, να αναζητήσει το ENSG00000161853 και να βρει τους απογόνους του ή
2. να εκτελέσει το εργαλείο ID History Converter.

Και στις δυο περιπτώσεις , θα πρέπει να το αντικαταστήσει χειροκίνητα (manually). Αυτό δεν είναι καθόλου λειτουργικό καθώς τα εργαλεία αυτού του τύπου χειρίζονται έναν πολύ μεγάλο αριθμό από stable IDs .

4.1.3 Αντιμετώπιση του προβλήματος

Στην περίπτωση αυτή, δεν εξυπηρετεί η ανάκτηση του ιστορικού του stable ID αλλά η αυτόματη αντικατάστασή του με τους απογόνους τους στην τρέχουσα Ensembl έκδοση. Το κλειδί για τη λύση του προβλήματος είναι η ανάπτυξη ενός αυτόματου εργαλείου, στο οποίο ο χρήστης θα εισάγει το σύνολο των stable IDs του (IDs προερχόμενα από οποιαδήποτε έκδοση της Ensembl) και θα του επιστρέφεται μια λίστα στην οποία τα stable IDs που έχουν αποσυρθεί, θα έχουν αντικατασταθεί με τους απογόνους τους.

4.2 Υλοποίηση των εφαρμογών

Οι εφαρμογές μοντελοποιήθηκαν και αναπτύχθηκαν στα πλαίσια των αναγκών του DIANA lab ,απευθύνονται όμως στο ευρύ ερευνητικό κοινό που χρησιμοποιεί τη βάση δεδομένων της Ensembl.

4.2.1 Προεργασία

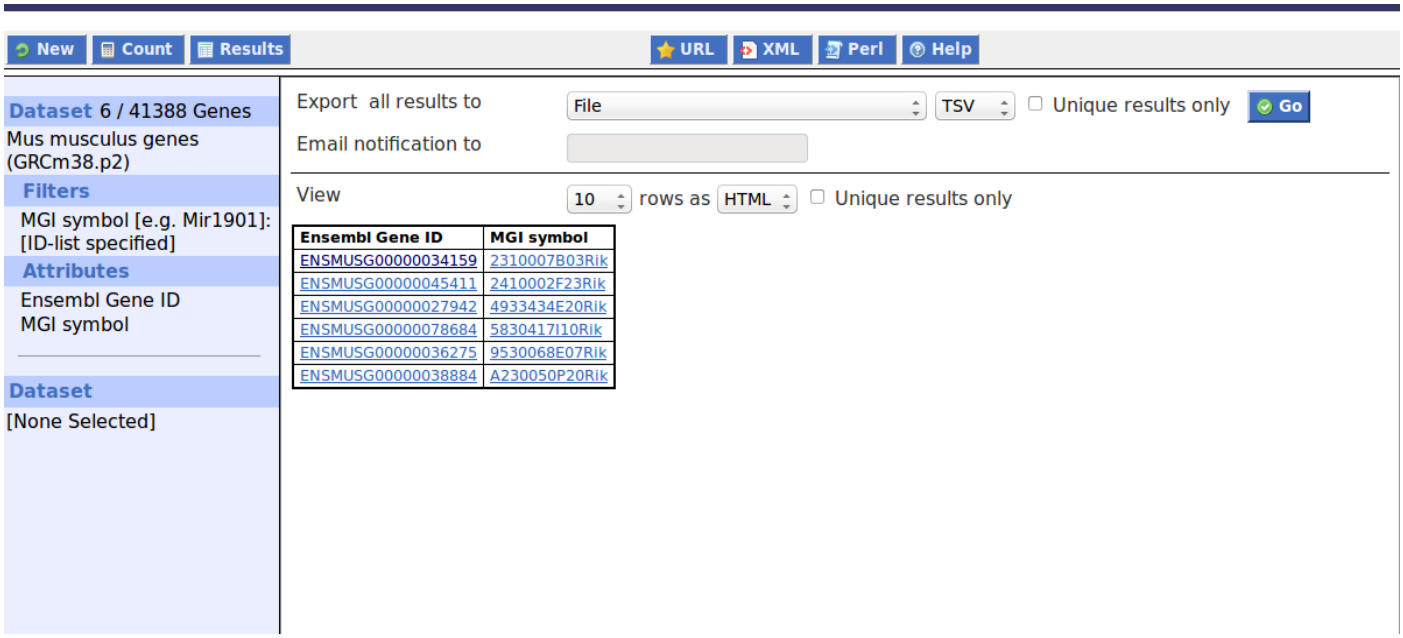
Στην database του DIANA lab υπήρχαν γονίδια , τα οποία είχαν καταχωρηθεί χειροκίνητα στις πρώτες εκδόσεις. Για κάποια από τα γονίδια αυτά υπήρχαν αποθηκευμένα τα Ensembl IDs τους. Η πληροφορία που υπήρχε όμως για κάποια άλλα, προερχόταν από άλλες βάσεις δεδομένων όπως για παράδειγμα από την RefSeq , από την HGNC , από την MGI και άλλες. Το αρχικό στάδιο λοιπόν της εργασίας ήταν η μετατροπή των IDs που προέρχονταν από εξωτερικές πηγές σε Ensembl stable IDs. Αυτό πραγματοποιήθηκε κυρίως με τη χρήση του εργαλείου **BioMart**.

Biomart
Το BioMart είναι ένα web-based εργαλείο μέσω του οποίου μπορεί κάποιος χωρίς προγραμματιστικές γνώσεις να εξάγει δεδομένα (sequences ή πληροφορίες) για τα γονίδια ,τα transcripts και γι’ άλλα βιολογικά στοιχεία.
Ανάμεσα στις υπόλοιπες δυνατότητές του , υποστηρίζει το mapping (αντιστοίχιση) των IDs μιας βάσης δεδομένων σε μια άλλη. Τα βήματα που εφαρμόστηκαν για κάθε ξεχωριστή βάση ήταν τα εξής:

1. Ορισμός του Dataset : Επιλογή της βάσης δεδομένων και του οργανισμού στον οποίο απευθυνόμαστε (πχ Ensembl Genes 76 -Homo Sapiens genes)

2. Ορισμός του Filter (ορίσματα : λίστα με τα ids της εξωτερικής βάσης δεδομένων και το όνομα της πχ RefSeq_ids)

3. Ορισμός των Attributes (η βάση δεδομένων στην οποία θέλουμε να αντιστοιχίσουμε τα IDs : Ensembl_ids).



Εικόνα 4.1 : Αποτελέσματα του mapping μιας λίστας από MGI symbols σε Ensembl IDs για το ποντίκι.

4.2.2 Ανάπτυξη πρώτης εφαρμογής

Αφού ανακτήθηκε το Ensembl ID για κάθε γονίδιο/transcript του DIANA lab, σειρά είχε η ανάπτυξη της πρώτης εφαρμογής. Πρόκειται για ένα script που δέχεται σαν όρισμα μια λίστα με Ensembl gene/transcript IDs, τα οποία μπορεί να προέρχονται από διάφορες εκδόσεις της Ensembl και τα μετατρέπει στην πιο πρόσφατη έκδοση της βάσης. Μαζί με τα gene IDs αλλάζουν και τα αντίστοιχα transcript ids ή/και τα gene names.

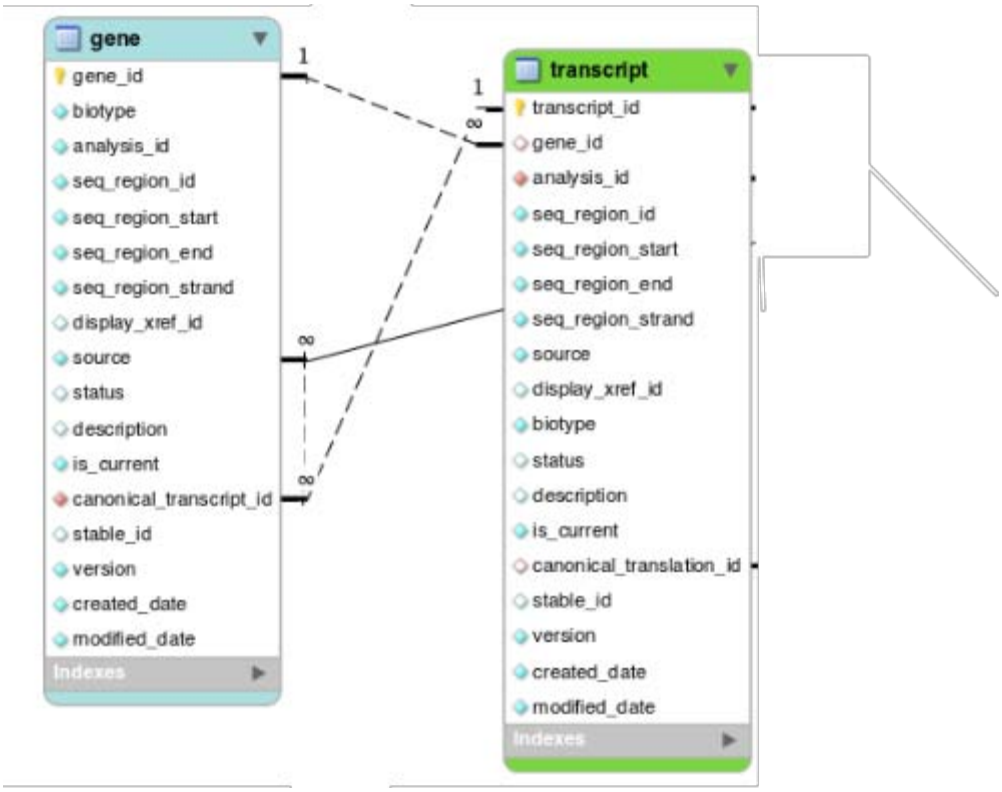
Ensembl APIs

Για την αποθήκευση της πληροφορίας της, η Ensembl χρησιμοποιεί σχεσιακές MySQL βάσεις δεδομένων ,δηλαδή συλλογές δεδομένων ,οργανωμένες σε συσχετισμένους πίνακες που παρέχουν ταυτόχρονα ένα μηχανισμό για ανάγνωση, εγγραφή και τροποποίηση. Ένα ολοκληρωμένο σύνολο από Διεπαφές Προγραμματισμού Εφαρμογών (Application Programme Interfaces, APIs) χρησιμεύουν ως ένα

ενδιάμεσο στρώμα μεταξύ των υποκείμενων σχημάτων της βάσης δεδομένων και των εφαρμογών.
Τα APIs στοχεύουν στο να αποδώσουν περιληπτικά τη δομή της βάσης δεδομένων, παρέχοντας αποτελεσματική πρόσβαση υψηλού επιπέδου στους πίνακες δεδομένων και απομονώνοντας τις εφαρμογές από τυχόν αλλαγές στη δομή της βάσης .Το API της Ensembl είναι γραμμένο σε Perl .

API Tables

Όπως αναφέρεται και παραπάνω , όλη η πληροφορία της Ensembl είναι αποθηκευμένη σε πίνακες(tables) [27]. Οι πίνακες αποτελούνται από διακριτά πεδία και ο συσχετισμός δυο πινάκων γίνεται με τη χρήση ενός κοινού πεδίου.



Εικόνα 4.2 : Gene και Transcript tables και ο συσχετισμός τους.

Στην Εικόνα 4.2 απεικονίζονται οι πίνακες για τα Γονίδια και τα Transcripts.

Το Gene table περιέχει για κάθε γονίδιο το gene_id του (το αναγνωριστικό του στον συγκεκριμένο πίνακα), τη version, την αρχή, το τέλος και το ID της αλληλουχίας του(seq_region_start/seq_region_end/seq_region_id) ,το όνομα του γονιδίου από την HGNC ή από την Vega/Havana (display_xref_id), πληροφορία για το αν το γονίδιο βρίσκεται στην τρέχουσα Ensembl Release (is_current) ,το Ensembl stable ID του ,έναν δείκτη στα transcripts του (canonical transcript ID) καθώς και άλλες πληροφορίες.

Το transcript table περιέχει περίπου τις ίδιες πληροφορίες για ένα transcript. Επιπροσθέτως , περιέχει έναν δείκτη στα translations του transcript, καθώς και το gene id του γονιδίου στο οποίο ανήκει.

Οι δυο πίνακες είναι συσχετισμένοι μεταξύ τους ,αφού έχουν ένα κοινό πεδίο (gene_id). Δηλαδή, αν γνωρίζουμε το transcript , μπορούμε από το transcript table να ανακτήσουμε το gene_id του γονιδίου στο οποίο ανήκει και να το χρησιμοποιήσουμε για να πάρουμε πληροφορίες από το gene table.

Ensembl Core API Documentation

Το API χρησιμοποιεί μια αντικειμενοστραφή προσέγγιση για την μοντελοποίηση των πραγματικών βιολογικών αντικειμένων (όπως των γονιδίων και των transcripts), κάνοντας πιο ξεκάθαρη την ανάπτυξη εφαρμογών (scripts) για την ανάκτηση και ανάλυση δεδομένων από τον χρήστη.

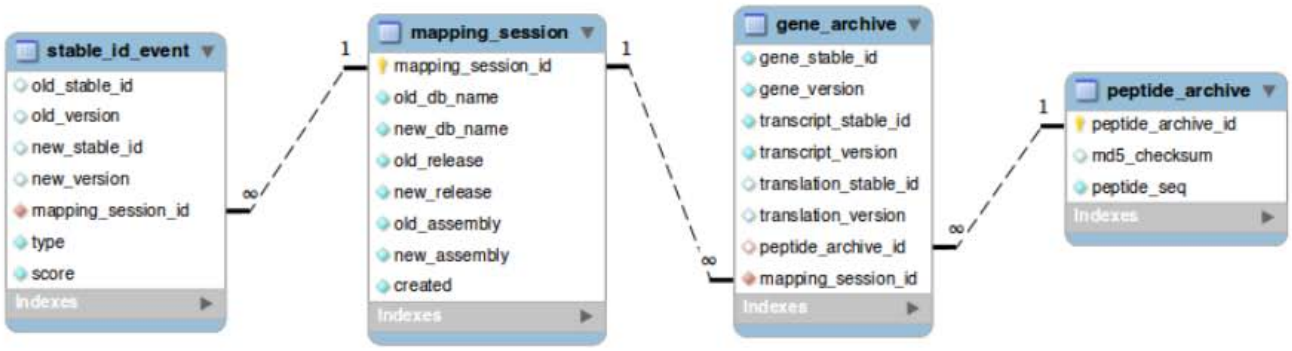
Το Ensembl Core API συνοδεύεται από ένα επαρκές σύνολο τεκμηρίωσης του κώδικα (Doxygen Perl Documentation) σε μορφή standard Perl POD (Plain Old Documentation) [28].Το documentation δίνει μια περιγραφή και παραθέτει το αρχείο κώδικα (.pm) για κάθε κλάση της Ensembl αναλύοντας διεξοδικά κάθε μέθοδο που ανήκει στην κλάση. Επίσης, αναλύεται η ιεραρχία ανάμεσα στις κλάσεις και παρέχονται γράφοι εξάρτησης.

ArchiveStableId Class Reference

Η βασική κλάση που χρησιμοποιήθηκε για την ανάπτυξη της εφαρμογής είναι η ArchiveStableId Class. Τα αντικείμενα αυτής της κλάσης είναι η κύρια μονάδα για την ανάκτηση πληροφοριών σχετικά με IDs που έχουν αποσυρθεί από τον πυρήνα της βάσης δεδομένων της Ensembl.

Για τη δημιουργία αντικειμένων αυτής της κλάσης χρησιμοποιείται ο ArchiveStableIdAdaptor, ο οποίος έχει πρόσβαση στα εξής tables του πυρήνα:

- stable_id_event
- mapping_session
- peptite_archive
- gene_archive



Εικόνα 4.3 : Τα tables που χρησιμοποιεί ο ArchiveStableIdAdaptor και ο συσχετισμός τους.

Μέσω των παραπάνω tables, αν είναι γνωστό ένα παλιό stable_id (old_stable_id) μπορούμε να ανακτήσουμε πληροφορίες για το γονίδιο ή transcript στο οποίο ανήκει (gene_stable_id/ transcript_stable_id).

Η συνάρτηση-κλειδί της κλάσης ArchiveStableId που χρησιμοποιείται είναι η “get_all_successors”. Η συνάρτηση αυτή, επιστρέφει μια λίστα από ArchiveStableIds στα οποία έχει αντιστοιχηθεί (mapping) το αντικείμενο το οποίο την καλεί.

Άλλη μια καθοριστική συνάρτηση για την εφαρμογή είναι η “is_current”. Αυτή είναι μια boolean συνάρτηση που επιστρέφει αν το stable id του αντικειμένου που την καλεί υπάρχει στην τρέχουσα έκδοση της Ensembl.

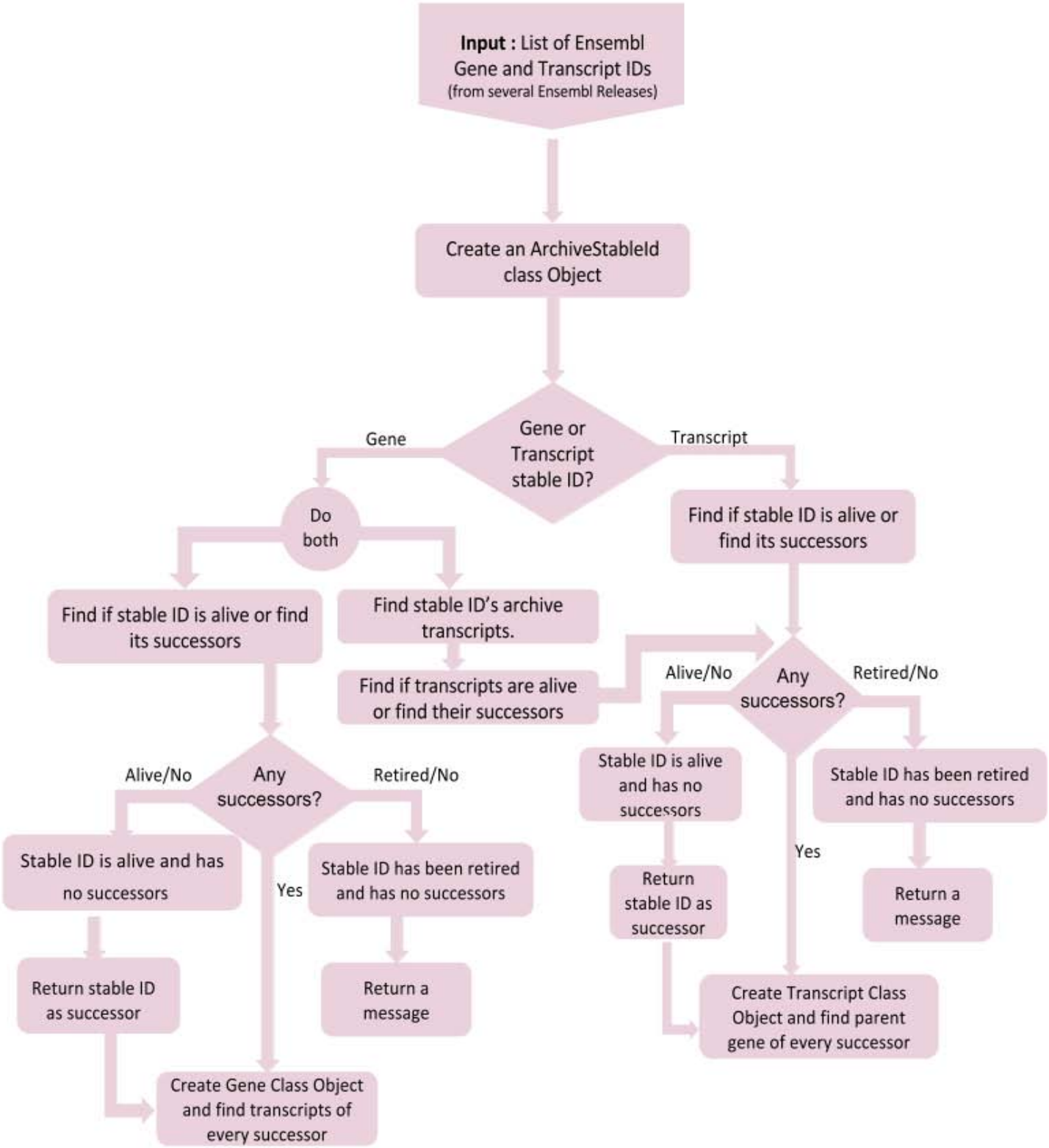
Στην επόμενη σελίδα(Εικόνα 4.4) ακολουθεί ένα διάγραμμα ροής που αποτυπώνει τον αλγόριθμο αυτής της εφαρμογής. Όπως βλέπουμε, η διαδικασία που ακολουθείται για ένα stable Gene ID και για ένα stable Transcript ID είναι παρόμοια. Η μόνη διαφορά έγκειται στο γεγονός ότι όταν πρόκειται για ένα γονίδιο ,η εφαρμογή βρίσκει:

1. τους απογόνους του stable ID
2. τα transcripts που του αντιστοιχούσαν
3. τα transcripts των απογόνων του
4. τους απογόνους των transcripts που του αντιστοιχούσαν,

ενώ όταν πρόκειται για ένα μετάγραφο επιστρέφει :

1. τους απογόνους του stable ID, δηλαδή τα transcripts που αντιστοιχούν στην τρέχουσα έκδοση
2. σε ποιο γονίδιο ανήκουν αυτά τα transcripts.

Εικόνα 4.4 : Διάγραμμα ροής αλγορίθμου


Input : List of Ensembl Gene and Transcript IDs (from several Ensembl Releases)
Create an ArchiveStableId class Object
Gene or Transcript stable ID?
Gene
Transcript
Do both
Find if stable ID is alive or find its successors
Find stable ID's archive transcripts.
Find if transcripts are alive or find their successors
Find if stable ID is alive or find its successors
Alive/No
Any successors?
Retired/No
Stable ID is alive and has no successors
Stable ID has been retired and has no successors
Return stable ID as successor
Return a message
Yes
Create Gene Class Object and find transcripts of every successor
Alive/No
Any successors?
Retired/No
Stable ID is alive and has no successors
Stable ID has been retired and has no successors
Return stable ID as successor
Return a message
Yes
Create Transcript Class Object and find parent gene of every successor

Εικόνα 4.5 : Μέρος του κώδικα

```
# get an ArchiveStableIdAdaptor from the Registry
        my $arch_adaptor = $registry->get_adaptor($species, 'Core', 'ArchiveStableId');
        my $gene_adaptor = $registry->get_adaptor($species, 'Core', 'Gene');
        my $transcript_adaptor = $registry->get_adaptor($species , 'Core', 'Transcript');


my @all_successors;
sub findSuccessors{
        my $self = shift;
        my $succ = $arch_adaptor->fetch_by_stable_id($self->stable_id)->get_all_successors;
        for my $a_id (@$succ){
                findSuccessors($a_id);
        }
        if ($arch_adaptor->fetch_by_stable_id($self->stable_id)->is_current){
                push @all_successors, $self->stable_id ;
        }

}

sub findTranscripts{
        my $self = shift;
        my $gene = $gene_adaptor->fetch_by_stable_id($self);
        my $transcripts = $gene->get_all_Transcripts();
        if(@$transcripts){
                print "\nTranscripts of successor ". $self. ":\n";
                print $out2 "\n#Transcripts of successor ". $self. ":\n";
                for my $tr (@$transcripts) {
                        print $tr->stable_id . "\n";
                        print $out2 $tr->stable_id . "\n";
                }
        }
        return @$transcripts;

}




sub archiveTranscriptSuccessors{
        my $self = shift;
        my $arch_trans = $self->get_all_transcript_archive_ids;
        if(@$arch_trans){
                print "Archive Transcripts of Gene ID ".$self->stable_id." are\n" ;
                print $out3 "#Archive Transcripts of Gene ID ".$self->stable_id." are\n" ;
                foreach my $trans (@$arch_trans) {
                        print "->".$trans->stable_id ."\n" ;
                        print "#".$trans->stable_id ."\n" ;
                }
        print "--------------------------------------------------------------------------------\n";
        }
        return $arch_trans;
}
```

Η προηγούμενη εικόνα απεικονίζει μέρος του κώδικα της πρώτης εφαρμογής. Συγκεκριμένα , φαίνεται η δημιουργία των τριών adaptors ώστε να εξασφαλιστεί η χρήση των αντίστοιχων κλάσεων(Gene,Transcript,ArchiveStableId).

Ακολουθούν τρεις συναρτήσεις :

1. **η findSuccessors** για την εύρεση των απογόνων
2. **findTranscripts** στην περίπτωση που πρόκειται για gene ID, όπου μαζί με τους απογόνους επιστρέφονται και τα transcripts τους
3. **ArchiveTranscriptSuccessor** στην περίπτωση του gene ID, όπου αφού βρεθούν τα transcripts που αντιστοιχούσαν στο ID, βρίσκονται οι απόγονοι του κάθε Transcript και σε ποιο γονίδιο ανήκουν.

Η εφαρμογή εξάγει τα αποτελέσματά της σε τρία αρχεία:

- στο πρώτο αρχείο εξάγει τους απογόνους
- στο δεύτερο αρχείο τα transcripts των απογόνων*
- στο τρίτο τους απόγονους των archive transcripts*

*όταν πρόκειται για gene ID.

4.2.3 Ανάπτυξη δεύτερης εφαρμογής

Η δεύτερη εφαρμογή δέχεται σαν είσοδο τα τρία αρχεία που εξάγει η πρώτη. Αφού ανακτήσουμε όλα τα IDs που μας ενδιαφέρουν με την πρώτη εφαρμογή, χρησιμοποιούμε την δεύτερη για την εξόρυξη πληροφοριών σχετικά με αυτά.

Συγκεκριμένα για όλα τα Gene Ensembl IDs ,η εφαρμογή μας επιστρέφει τις εξής πληροφορίες :

1. το όνομα του γονιδίου στο οποίο αντιστοιχεί το ID
2. την εξωτερική βάση από την οποία πήρε το όνομά του (HGNC /Vega) το γονίδιο
3. την version του ID
4. το χρωμόσωμα στο οποίο βρίσκεται το γονίδιο
5. ο βιότυπός του (πχ αν είναι protein coding)
6. η αρχή και το τέλος του γονιδίου
7. το μήκος του
8. σε ποια από τις δυο αλυσίδες βρίσκεται
9. μια λίστα με τα transcripts του.

Για τα Transcript Ensembl IDs επιστρέφει :

1. το όνομα ου μεταγράφου στο οποίο αντιστοιχεί το ID
2. την εξωτερική βάση από την οποία πήρε το όνομά του (HGNC /Vega)
3. την version του ID
4. το γονίδιο στο οποίο ανήκει το transcript
5. το χρωμόσωμα στο οποίο βρίσκεται
6. ο βιότυπός του
7. η αρχή και το τέλος του
8. το μήκος του (χωρίς τα εσώνια)
9. μια λίστα με τα εξώνιά του.

```
if (substr($stable_id ,3,1) eq 'G' ){
        my $gene = $gene_adaptor->fetch_by_stable_id($stable_id);
        print "Info for Gene ID ". $stable_id. ":\n\n" ;
        print "Gene name   : ".$gene->external_name()."\n";
        print "External DB : " .$gene->external_db . "\n";
        print "ID version  : " .$gene->version . "\n";
        print "Chromosome  : " .$gene->slice->seq_region_name. "\n";
        print "Biotype     : " .$gene->biotype . "\n";
        print "Start       : " .$gene->seq_region_start . "\n";
        print "End         : " .$gene->seq_region_end . "\n";
        print "Gene Length : " .$gene->length . "\n";
        print "Strand      : " .$gene->seq_region_strand . "\n";
        my $transcripts = $gene->get_all_Transcripts();
        if(@$transcripts){
                print "\nTranscripts : \n". $stable_id. ":\n";
                for my $tr (@$transcripts) {
                        print $tr->stable_id . "\n";
                }
        }
        print "\n-------------------------------------------------------------------------------\n";
}
```

Εικόνα 4.6 : Απόσπασμα του κώδικα της δεύτερης εφαρμογής

Επίλογος

Τα τελευταία χρόνια ,η σημαντική εξέλιξη του τομέα της Μοριακής Βιολογίας σε συνδυασμό με την εξέλιξη της τεχνολογίας που χρησιμοποιείται , έχουν ως αποτέλεσμα την εκθετική αύξηση των πληροφοριών που παράγονται από τη βιολογική κοινότητα. Το γεγονός αυτό κάνει επιτακτική την ανάγκη για ανάπτυξη νέων εργαλείων και εφαρμογών που θα διαχειρίζονται και θα αναλύουν όλα αυτά τα βιολογικά δεδομένα. Στόχος είναι αυτή η επεξεργασία να γίνεται όσο το δυνατόν πιο αυτοματοποιημένα .

Το χρονικό διάστημα ανάμεσα σε δυο διαδοχικές εκδόσεις της Ensembl έχει πλέον μειωθεί στους δυο μήνες (περίπου) οπότε είναι πολύ έντονη η ανάγκη για αυτόματη ενημέρωση των Βάσεων Δεδομένων που βασίζονται σ' αυτήν. Οι δυο εφαρμογές που αναλύθηκαν στη συγκεκριμένη διπλωματική εργασία ανήκουν σ'αυτήν την κατηγορία , δηλαδή στα εργαλεία που διευκολύνουν την επεξεργασία της βιολογικής πληροφορίας της Ensembl από το ερευνητικό κοινό.

Further Considerations

Γενικά : Η μελλοντική έρευνα στον τομέα της Βιοπληροφορικής πρέπει να εστιάσει στην έλλειψη υπολογιστικής ισχύος , η οποία δεν μπορεί να καλύψει τη βιολογική πολυπλοκότητα.

Όσον αφορά το αντικείμενο αυτής της διπλωματικής εργασίας, κάποια μελλοντική δουλειά θα μπορούσε να εστιάσει στην αξιοποίηση του ιστορικού που εξάγει η πρώτη εφαρμογή για την αυτόματη ενημέρωση κάποιου συγκεκριμένου εργαλείου ή στην αύξηση του εύρους της εφαρμογής ώστε να καλύπτει εκτός από γονίδια και μετάγραφα , πρωτεΐνες και ρυθμιστικούς παράγοντες.

Βιβλιογραφία

[1]http://kpe-kastor.kas.sch.gr/leaf/texts/cell-types.htm

[2] http://www.ncbi.nlm.nih.gov/books/NBK26821/

[3] http://www.nature.com/nature/dna50/watsoncrick.pdf

[4] http://ebooks.edu.gr/modules/ebook/show.php/DSGL-C112/479/3165,12730/

[5]http://www.ncbi.nlm.nih.gov/books/NBK21202/

[6] http://www.nature.com/scitable/topicpage/dna-transcription-426

[7] http://www.ncbi.nlm.nih.gov/books/NBK21603/

[8] http://www.nature.com/nrg/journal/v11/n5/box/nrg2776_BX1.html

[9]http://www.nobelprize.org/educational/medicine/dna/b/translation/translation_process.html

[10] http://www.nature.com/nature/focus/crick/pdf/crick227.pdf

[11] http://www.bioplanet.com/what-is-bioinformatics/

[12] http://bioinformatics.biol.uoa.gr/msc/gr/general.html#Anchor-49575

[13] http://www.cbcb.umd.edu/research/assembly_primer

[14] http://www.ncbi.nlm.nih.gov/books/NBK20253/

[15] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2808980/

[16] http://www.genome.gov/12011238

[17] https://www.sanger.ac.uk/resources/databases/ensembl.html

[18] http://www.ensembl.org

[19] http://www.britannica.com/EBchecked/topic/270557/homology

[20] http://lcbb4.epfl.ch/reading/homology/orthology/1970-fitchDistinguishingHomologousAnalogousProteins.pdf

[21] http://www.icp.ucl.ac.be/~opperd/private/orthol.html

[22] http://www.ensembl.org/info/genome/variation/data_description.html

[23] http://www.ensembl.org/info/website/archives/assembly.html

[24] http://www.ensembl.org/info/genome/stable_ids/index.html

[25] http://www.ensembl.org/info/genome/stable_ids/versions.html

[26] http://diana.cslab.ece.ntua.gr/

[27] http://www.ensembl.org/info/docs/api/core/core_schema.html

[28] http://www.ensembl.org/info/docs/Doxygen/core-api/index.html

Πηγές Εικόνων

Εικόνα 1.1:
http://www.daviddarling.info/encyclopedia/N/nucleotide.html

Εικόνα 1.2:
http://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397

Εικόνες 1.3 , 1.4:
http://www.ncbi.nlm.nih.gov/books/NBK21202/

Εικόνα 1.5:
http://en.citizendium.org/wiki/File:RNA_base_vs_DNA_base.jpg

Εικόνα 1.6:
http://cnx.org/contents/0f5c6adf-8fa4-4623-9bc1-89283c5b6087@1

Εικόνα 1.7:
http://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/alternative_splicing.html

Εικόνα 1.8: http://en.wikipedia.org/wiki/Ribosome

Εικόνα 1.9: http://web.uconn.edu/mcb201/lecture01.html

Εικόνες 2.1, 2.2 ,2.3, 2.4:
http://www.cbcb.umd.edu/research/assembly_primer

Εικόνα 2.5:
http://greatprojectscampaign.com/human-genome-project.html

Εικόνα 3.1:http://www.ensembl.org/index.html

Εικόνα 3.2:
http://www.ensembl.org/info/genome/variation/predicted_data.html

Εικόνα 3.3:
http://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000197021;r=X:149929527-149938811

Εικόνα 3.4:
http://www.ensembl.org/Homo_sapiens/Share/4d03629c76bca96cca36cdeb0cc3a3d4156684586

Εικόνα 3.5:
http://www.icp.ucl.ac.be/~opperd/private/orthol.html

Εικόνα 3.6:
http://www.ensembl.org/info/genome/compara/homology_method.html#homology_types

Εικόνα 3.7:
http://www.ensembl.org/Homo_sapiens/Gene/SpeciesTree?db=core;g=ENSG00000159917;r=19:44782947-44813601;t=ENST00000291182

Εικόνες 3.8, 3.9:
http://www.ensembl.org/info/genome/variation/index.html

Εικόνα 3.10: http://www.ensembl.org (μορφοποίηση αποτελεσμάτων)

Εικόνα 3.11:
http://www.ensembl.org/info/genome/variation/data_description.html

Εικόνα 3.12:
http://www.ensembl.org/Homo_sapiens/Gene/Regulation?db=core;g=ENSG00000197021;r=X:149929527-149938811

Εικόνα 3.13: http://www.ensembl.org (μορφοποίηση αποτελεσμάτων)

Εικόνα 4.1: http://www.ensembl.org/biomart/martview

Εικόνα 4.2:
http://uswest.ensembl.org/info/docs/api/core/fundamental_tables_core.pdf

Εικόνα 4.3:
http://uswest.ensembl.org/info/docs/api/core/id_mapping_core.pdf

Εικόνες 4.4, 4.5, 4.6: -> προσωπική επεξεργασία