ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Σύστημα Διαχείρισης Διαχρονικών Δεδομένων για Γονίδια

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΚΩΝΣΤΑΝΤΙΝΟΥ ΖΑΓΓΑΝΑ

Επιβλέπων : Ιωάννης Βασιλείου
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούνιος 2014

Στο Θανάση και τον Ηλία για την πολύτιμη βοήθειά τους.
Στους φίλους μου Παναγιώτη, Δημήτρη και Δημήτρη.
Σε αυτούς που δεν πίστεψαν ποτέ σε 'μενα και σε αυτούς που προσπάθησαν να με αποτρέψουν από τον να ασχοληθώ με το αντικείμενο.

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Σύστημα Διαχείρισης Διαχρονικών Δεδομένων για Γονίδια

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΚΩΝΣΤΑΝΤΙΝΟΥ ΖΑΓΓΑΝΑ

Επιβλέπων : Ιωάννης Βασιλείου
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 12η Ιουνίου 2014.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Ιωάννης Βασιλείου	Κωνσταντίνος Κοντογιάννης	Νεκτάριος Κοζύρης
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούνιος 2014

(Υπογραφή)

.....

ΚΩΝΣΤΑΝΤΙΝΟΣ ΖΑΓΓΑΝΑΣ

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η μελέτη των *βιομορίων* που συμμετέχουν στους μηχανισμούς της ζωής (πχ DNA, πρωτεΐνες, μόρια microRNA κτλ) είναι απαραίτητη για να μπορέσουν οι ερευνητές να κατανοήσουν και να θεραπεύσουν γενετικές ασθένειες που απασχολούν την Ιατρική και τη Βιολογία τους τελευταίους αιώνες. Οι πληροφορίες που σχετίζονται με αυτά τα βιομόρια αποκαλύπτονται μέσω βιολογικών πειραμάτων και καταγράφονται σε βάσεις δεδομένων για να είναι προσβάσιμες στους ερευνητές. Όμως τα βιολογικά πειράματα χρησιμοποιούν μηχανήματα και αναλύσεις που είναι επιρρεπή σε σφάλματα. Ως αποτέλεσμα, οι προαναφερθείσες βάσεις δεδομένων οφείλουν να μεταβάλλονται συνεχώς προκειμένου να είναι ενημερωμένες με τις ακριβέστερες μετρήσεις. Για διάφορους λόγους (πχ για την συγκριτική μελέτη σχετικής βιβλιογραφίας) είναι χρήσιμη η πρόσβαση όχι μόνο στις πιο πρόσφατες εκδόσεις των πληροφοριών αλλά επίσης και στις προηγούμενες καταστάσεις τους. Οι υπάρχουσες βάσεις δεδομένων δεν διευκολύνουν την αναζήτηση αυτών των προηγούμενων καταστάσεων δημιουργώντας πρόβλημα στους ερευνητές.

Στόχος της παρούσας εργασίας είναι (α) η μελέτη, η μοντελοποίηση, η καταγραφή και η οπτικοποίηση των πληροφοριών που σχετίζονται με τα γονίδια του DNA διαχρονικά και (β) ο εμπλουτισμός όλων των διαχρονικών δεδομένων για βιομόρια με πρόσθετες πληροφορίες (πχ ονόματα γονιδίων, εναλλακτικά ονόματα στη βιβλιογραφία, σύμβολα σε άλλες βάσεις γονιδιακών δεδομένων).

Λέξεις Κλειδιά: βιολογία, γονίδια, transcripts, Ensembl, διαδικτυακή εφαρμογή, Yii Framework, βάσεις δεδομένων, διαχρονικά δεδομένα

Η σελίδα αυτή είναι σκόπιμα λευκή.

Abstract

The study of biomolecules that play an active role in the mechanisms of life (e.g. DNA, proteins, microRNA molecules etc) is necessary so that researchers become able to understand the causes of several genetic diseases and find possible treatments for them. These diseases have been a subject of study in the research fields of Medicine and Biology for the last few centuries. The data related to these biomolecules is produced by biological experiments and are stored in databases, to be accessible to researchers.

However, the equipment and methods used during the aforementioned experiments are prone to produce errors. As a result, the aforementioned databases must be constantly updated, in order for them to be updated with the latest, most accurate measurements. For many reasons (e.g. the comparative study of relevant bibliography) it is useful to access not only the latest version of this data but also its previous versions. The current database technology does not facilitate searching for previous versions, thus, creating difficulties for researchers.

The aim of this project is (a) the study, the modeling, the recording and the visualization of temporal data relevant to genes and (b) the enrichment of that temporal biomolecular data with additional information (e.g. gene names, synonyms, symbols recorded in other gene databases).

Keywords: biology, genes, transcripts, Ensembl, internet application, Yii Framework, databases, temporal evolution of data

Η σελίδα αυτή είναι σκόπιμα λευκή.

Πίνακας περιεχομένων

1 *Εισαγωγή*

1.1 Καταγραφή της εξέλιξης των γονιδιακών δεδομένων

Το σύνολο της γενετικής πληροφορίας ενός οργανισμού κωδικοποιείται σε ακολουθίες DNA, που ονομάζονται *γονίδια*. Το κύτταρο «διαβάζει» τη γενετική πληροφορία που κωδικοποιούν τα γονίδια και, με βάση αυτή, παράγει *πρωτεΐνες*, θέτοντας έτσι σε εφαρμογή τους μηχανισμούς της ζωής. Δυσλειτουργίες κατά την παραγωγή πρωτεϊνών μπορούν να δημιουργήσουν προβλήματα στους μηχανισμούς αυτούς. Τέτοιες δυσλειτουργίες αποτελούν την αιτία πολλών γενετικών ασθενειών.

Γονιδιακά δεδομένα λέγονται τα δεδομένα που καταγράφουν πληροφορία σχετική με τα γονίδια διαφόρων οργανισμών καθώς και το πώς αυτά εκφράζονται μέσω της παραγωγής των αντίστοιχων πρωτεϊνών.

Το 1990 ξεκίνησε μια προσπάθεια καταγραφής του ανθρωπίνου γονιδιώματος, η οποία ολοκληρώθηκε το 2003.[2] Ταυτόχρονα ξεκίνησε η έρευνα πάνω στον τομέα της χαρτογράφησης του γενετικού υλικού διαφόρων οργανισμών και κυρίως του ανθρώπου. Μέσω αυτής της χαρτογράφησης επιχειρείται:

1. Η αναγνώριση διαφόρων περιοχών του γονιδιώματος και πώς αυτές χρησιμοποιούνται για τη λειτουργία ενός οργανισμού
2. Η ανακάλυψη της εξελικτικής ιστορίας διαφόρων οργανισμών ανά τους αιώνες μέσω της αναγνώρισης ομοιοτήτων ανάμεσα στο γονιδιακό υλικό διαφόρων «κοντινών» οργανισμών (πχ άνθρωπος και πίθηκος).
3. Η εύρεση διαφορών ανάμεσα στη γονιδιακή πληροφορία διαφόρων ομάδων του ίδιου οργανισμού. Δηλαδή πώς διάφορα γονίδια εκφράζονται σε διαφορετικές ομάδες (πχ φυλετικές ομάδες του ανθρώπου).

4. Η αναγνώριση του μηχανισμού με τον οποίο προκύπτουν διάφορες γονιδιακές ασθένειες που πλήττουν τον άνθρωπο (πχ σύνδρομο Down, Kleinefelter κ.α.) καθώς και η προσπάθεια εύρεσης τρόπων προκειμένου να αντιμετωπιστούν αυτές σε γονιδιακό επίπεδο.

Λόγω όσων αναφέρθηκαν παραπάνω, οι πληροφορίες που σχετίζονται με γονίδια είναι ιδιαίτερα χρήσιμες στους ερευνητές. Έτσι τις τελευταίες δεκαετίες έχουν εμφανιστεί αρκετές διαδικτυακές βάσεις δεδομένων που συγκεντρώνουν τέτοιες πληροφορίες (πχ Ensembl (http://www.ensembl.org), miRBase (http://www.mirbase.org/), NCBI (http://www.ncbi.nlm.nih.gov) και άλλες). Τα δεδομένα που καταγράφονται σε αυτές προκύπτουν από βιολογικά πειράματα, που χρησιμοποιούν μηχανήματα και αναλύσεις επιρρεπείς σε σφάλματα. Ως αποτέλεσμα, το περιεχόμενο των βάσεων αυτών ενημερώνεται συνεχώς με βάση τις νεότερες και ακριβέστερες μετρήσεις. Για διάφορους λόγους (πχ για την συγκριτική μελέτη σχετικής βιβλιογραφίας) είναι χρήσιμη η πρόσβαση όχι μόνο στις πιο πρόσφατες εκδόσεις των πληροφοριών αλλά επίσης και στις προηγούμενες καταστάσεις τους. Όμως οι υπάρχουσες υποδομές δεν διευκολύνουν την αναζήτηση αυτών των προηγούμενων καταστάσεων, δημιουργώντας έτσι πρόβλημα στους ερευνητές.

1.2 Αντικείμενο διπλωματικής.

Στο Ινστιτούτο Πληροφοριακών ΣΥστημάτων του ΕΚ «Αθηνά» και σε συνεργασία με το ΕΚ βιοϊατρικής «Αλέξανδρος Φλέμινγκ» μελετήθηκαν, στο πλαίσιο προηγούμενης διπλωματικής εργασίας, οι βασικές πληροφορίες που σχετίζονται με τα μόρια microRNA διαχρονικά, όπως καταγράφονται από τις διάφορες εκδόσεις της βάσης δεδομένων miRBase. Επιπλέον, αποθηκεύτηκαν αυτές οι πληροφορίες σε ένα βολικό σχήμα βάσης δεδομένων και υλοποιήθηκε ένα λογισμικό οπτικοποίησής τους. Επίσης, χρησιμοποιήθηκε η διαχρονική πληροφορία προκειμένου να διευκολυνθεί η αναζήτηση βιβλιογραφίας για τα μόρια microRNA. Με τον τρόπο αυτό διευθετήθηκε ένα πρώτο κομμάτι του προβλήματος της διαχείρισης των διαχρονικών δεδομένων για βιομόρια.

Η παρούσα εργασία περιλαμβάνει:

- Μελέτη, μοντελοποίηση, καταγραφή και οπτικοποίηση των διαχρονικών δεδομένων για γονίδια τεσσάρων οργανισμών
- Εμπλουτισμός όλων των διαχρονικών δεδομένων για βιομόρια με επιπλέον πληροφορίες από διάφορες διαδικτυακές βάσεις δεδομένων.

1.2.1 Συνεισφορά

Παρότι υπάρχει πληθώρα γενετικών δεδομένων, εντούτοις, υπάρχει ελάχιστη πληροφορία για το πώς αυτά εξελίσσονται στο χρόνο. Τα γονιδιακά δεδομένα παράγονται με μεθόδους που

είναι επιρρεπείς σε σφάλματα, όπως αναφέρθηκε και παραπάνω, με αποτέλεσμα ανά τακτά χρονικά διαστήματα αυτά τα σφάλματα να αναγνωρίζονται και στη συνέχεια να παράγονται καινούριες εκδόσεις των δεδομένων. Οι αλλαγές αυτές έχουν να κάνουν κυρίως με την ακολουθία διαφόρων γονιδίων(διόρθωση λανθασμένων βάσεων, συμπλήρωση κενών στο γονιδίωμα) καθώς και με αλλαγές στην τοποθεσία και το όνομα ενός γονιδίου. Επίσης κάποια γονίδια μπορεί να σταματήσουν σε κάποια χρονική στιγμή να καταγράφουν και να αντικατασταθούν από ένα ή περισσότερα άλλα. Η σημασία αυτής της πληροφορίας είναι αρκετά μεγάλη για έναν ερευνητή-βιολόγο. Μπορεί να ασχολείται για παράδειγμα με ένα συγκεκριμένο γονίδιο και στην επόμενη έκδοση το γονίδιο αυτό να αντικατασταθεί από ένα άλλο με τέτοιο τρόπο που να μην είναι πολύ εμφανές στον ίδιο το πώς συνέβη αυτό. Προκειμένου να συνεχίσει απρόσκοπτα την έρευνά του, είναι αναγκασμένος να αναζητά χωρίς αυτόματο τρόπο την πληροφορία, μια χρονοβόρα διαδικασία, ειδικά όταν πρέπει να συγκρίνει τη δομή βιομορίων. Η χρήση ενός υπολογιστή επιταχύνει πολύ αυτή τη διαδικασία, καθώς ο έλεγχος εκτελείται μία μόνο φορά και η αλλαγή αποθηκεύεται. Παράλληλα, μέσω διεπαφής ο ερευνητής μπορεί να δει αυτές τις αλλαγές με το πάτημα ενός κουμπιού κάτι που καθιστά τη συγκεκριμένη προσέγγιση ελκυστική για την αναζήτηση αλλαγών ακόμα και σε ακολουθίες 400-1000 χαρακτήρων.

Η συνεισφορά της συγκεκριμένης εργασίας στον στόχο που αναφέρεται παραπάνω είναι:

1. Η συλλογή γονιδιακών δεδομένων από τη βάση γονιδιακών δεδομένων Ensembl για διάφορες εκδόσεις αυτής.
2. Η ανάλυση των δεδομένων ανάμεσα σε δύο διαδοχικές εκδόσεις προκειμένου να αναγνωριστούν αλλαγές στα δεδομένα ενός γονιδίου που συνέβησαν καθώς και αποθήκευση των αλλαγών για μελλοντική προβολή τους από σύστημα οπτικοποίησης.
3. Η συλλογή πληροφορίας για το πώς κάποια γονίδια αντικαθιστούν κάποια άλλα, καθώς και πληροφορίας για ονόματα ή σύμβολα διαφόρων γονιδίων από άλλες βάσεις γονιδιακών δεδομένων.
4. Δημιουργία διαδικτυακής διεπαφής (interface) που οπτικοποιεί τις αλλαγές που συμβαίνουν στα γονιδιακά δεδομένα μαζί με όλες τις διάφορες εκδόσεις αυτών των δεδομένων. Η διεπαφή αυτή κάνει εύκολα κατανοητές τις αλλαγές που συνέβησαν στους ερευνητές-βιολόγους, αλλά και κάθε ενδιαφερόμενο καθώς το σύστημα θα είναι ανοιχτό στο ευρύ κοινό.[1]

1.3 Οργάνωση κειμένου

Στο κεφάλαιο 1 γίνεται μία εισαγωγή στο αντικείμενο της εργασίας. Στο Κεφάλαιο 2 τίθεται το θεωρητικό βιολογικό υπόβαθρο προκειμένου να γίνει κατανοητή η μεθοδολογία που

ακολουθήθηκε σχετικά με την παραγωγή δεδομένων. Η διαδικασία παραγωγής αυτών των δεδομένων περιγράφεται στο Κεφάλαιο 3. Στο Κεφάλαιο 4 περιγράφεται η διεπαφή της εφαρμογής που αναπτύχθηκε και οι τεχνολογίες που χρησιμοποιήθηκαν για αυτό το σκοπό. Το Κεφάλαιο 5 περιέχει τον επίλογο ενώ το τελευταίο κεφάλαιο, το κεφάλαιο 6 περιέχει τη βιβλιογραφία.

2 *Θεωρητικό υπόβαθρο και σχετικές εργασίες*

Στο κεφάλαιο αυτό θα γίνει μια σύντομη αναφορά στο θεωρητικό υπόβαθρο σχετικά με το γενετικό υλικό και το γονιδίωμα. Στη συνέχεια θα γίνει αναφορά στις διάφορες βάσεις δεδομένων που παρέχουν πληροφορία για βιομόρια καθώς και σε σχετικές εργασίες που έχουν γίνει στο παρελθόν.

2.1 Βασικές έννοιες από τη βιολογία

Στο σημείο αυτό θεωρείται σκόπιμο να δοθεί ένα βασικό βιολογικό υπόβαθρο προκειμένου να γίνουν κατανοητές οι έννοιες που έχουν να κάνουν με τα βιολογικά δεδομένα που χρησιμοποιήθηκαν στη διπλωματική εργασία.

2.1.1 Κύτταρο και Κυτταρική Μεμβράνη

Κατά τη βιολογία, *κύτταρο* ονομάζεται η βασική δομική και λειτουργική μονάδα που εκδηλώνει το φαινόμενο της ζωής. Έτσι, ως κύτταρο νοείται το μικρότερο δομικό συστατικό της έμβιας ύλης, που αποτελείται από μια συστηματικά οργανωμένη ομάδα μορίων που βρίσκονται σε δυναμική αλληλεπίδραση μεταξύ τους. Το κύτταρο διαθέτει μορφολογική, φυσική και χημική οργάνωση όπως επίσης και ικανότητα αφομοίωσης, ανάπτυξης και αναπαραγωγής. Είναι μια μονάδα της ζωής ανεξάρτητη ως προς την αυτορρύθμιση και την προσαρμοστικότητα του σε σχέση με το περιβάλλον. Με βάση τον υφιστάμενο αριθμό αυτών, οι οργανισμοί διακρίνονται σε μονοκύτταρους και πολυκύτταρους. Ο χώρος εντός του οποίου βιώνουν τα κύτταρα των πολυκύτταρων οργανισμών ονομάζεται μεσοκυττάριο υγρό. Μεγάλες ομάδες ομοειδών, κατά σύσταση και φυσιολογική λειτουργία, κυττάρων,

χαρακτηρίζονται *ιστοί*. Οι ιστοί αποτελούν τη μονάδα δεύτερης τάξης στον ανθρώπινο οργανισμό μετά τα κύτταρα.

Τα κύτταρα παρουσιάζουν μεγάλη ποικιλία μεγεθών και διαστάσεων, αντιπροσωπευτικών της ικανότητάς τους για εξελικτική προσαρμογή και διαφοροποίηση σε διαφορετικά περιβάλλοντα. Η διάμετρός τους ποικίλλει από δέκατα του μικρομέτρου (βακτήρια) έως μερικά εκατοστόμετρα (θαλάσσια φύκη ή αυγά πτηνών). Τα ανθρώπινα κύτταρα είναι της τάξης μεγέθους των 5 μικρομέτρων έως 1.5 χιλιοστομέτρων. Υπολογίζεται ότι το ανθρώπινο σώμα αποτελείται από εκατό τρισεκατομμύρια κύτταρα.

Γενικά τα κύτταρα, προκειμένου να διατηρούν τη λειτουργικότητά τους, υποχρεώνονται να ανταλλάσουν συνεχώς ουσίες με το περιβάλλον τους. Η αμφίδρομη αυτή ανταλλαγή (εισαγωγή χρήσιμων ουσιών και αποβολή αχρήστων) γίνεται μέσω της πλασματικής μεμβράνης που αποτελεί και το όριο μεταξύ έμβιας και άβιας ύλης. Όσο μεγαλύτερη είναι η επιφάνεια της πλασματικής μεμβράνης, τόσο μεγαλύτερη και η δυνατότητα της ανταλλαγής.

Εκτός όμως από την ανταλλαγή ουσιών που καλύπτει πλήρως τις μεταβολικές ανάγκες, παρατηρείται και η ανταλλαγή ουσιών-μηνυμάτων, μέσω των οποίων επικοινωνεί το κύτταρο με το περιβάλλον του και «αντιλαμβάνεται» τις διάφορες μεταβολές. Με βάση τις πληροφορίες των μηνυμάτων αυτόν και υπό τον έλεγχο του γενετικού υλικού, το κύτταρο εναρμονίζει τις λειτουργίες των επιμέρους τμημάτων του. Αυτή η μεταβίβαση όμως των μηνυμάτων μπορεί να πραγματοποιηθεί μόνο αν το κύτταρο έχει σχετικά μικρό όγκο. Έτσι λοιπόν δικαιολογείται ο μικρός όγκος των κυττάρων με τη μεγαλύτερη δυνατή επιφάνεια, προκειμένου να ικανοποιούνται ταυτόχρονα και οι δύο απαραίτητες για την επιβίωση του κυττάρου προϋποθέσεις, που είναι

1. Η μεγάλη επιφάνεια για ανταλλαγές ουσιών και υποδοχή μηνυμάτων και
2. Ο μικρός όγκος για την έγκαιρη μετάβαση των μηνυμάτων στο ενδοκυτταρικό περιβάλλον.[3,4]

2.1.2 Πυρήνας, Προκαρυωτικά και Ευκαρυωτικά Κύτταρα

Τόσο τα φυτικά όσο και τα ζωικά κύτταρα έχουν την ίδια βασική οργάνωση. Σε κάθε κύτταρο υπάρχει μια ογκώδης κεντρική δομή με χαρακτηριστικό σχήμα που ονομάζεται πυρήνας και ένας μεγάλος αριθμός μεμβρανικών διαμερισμάτων. Τα κύτταρα αυτά ονομάζονται ευκαρυωτικά κύτταρα, από τις ελληνικές λέξεις «ευ»(=καλός) και «κάρυον»(=πυρήνας). Εκτός από τα ευκαρυωτικά κύτταρα υπάρχει και μια απλούστερη μορφή κυττάρων που έχουν μια πρωτόγονη μορφή οργάνωσης του πυρήνα και ονομάζονται προκαρυωτικά κύτταρα. Προκαρυωτικά κύτταρα είναι τα βακτήρια, τα οποία είναι πολύ μικρότερα από τα ευκαρυωτικά κύτταρα και ζουν μόνα ή σε χαλαρές αποικίες. Τα βακτήρια υπάρχουν σε μία τεράστια ποικιλία ειδών και έχουν επιτύχει να ζουν ακόμα και στα πλέον

«εχθρικά» περιβάλλοντα. Χωρίς την ύπαρξη βακτηρίων, δεν θα υπήρχε η «ευκαρυωτική ζωή», όπως τουλάχιστον τη γνωρίζουμε σήμερα. Μια μεγάλη ποικιλία βακτηρίων συμβιώνουν αρμονικά με άλλους οργανισμούς επιτελώντας λειτουργίες που οι οργανισμοί αυτοί δεν μπορούν να επιτελέσουν Αρκετά βακτήρια, από την άλλη πλευρά, είναι ζημιογόνα επειδή έχουν την ικανότητα να εισέρχονται σε ανώτερους οργανισμούς και να προκαλούν ασθένειες.

Τα προκαρυωτικά κύτταρα είναι μικρά, περιβάλλονται από πλασματική μεμβράνη και σε ορισμένες ομάδες περικλείονται από κυτταρικό τοίχωμα. Το πυρηνικό υλικό των προκαρυωτικών κυττάρων βρίσκεται ελεύθερο στο κυτταρόπλασμα, δεν περιβάλλεται από μεμβράνη και ονομάζεται νουκλεοειδές. Στα νουκλεοειδή δεν παρατηρείται πυρηνίσκος, ενώ ένας σχετικά μικρός αριθμός πρωτεϊνών είναι συνδεδεμένος με το DNA. Παρά το γεγονός ότι στο κυτταρόπλασμα των πλέων σύνθετων προκαρυωτικών κυττάρων υπάρχουν μικρά μεμβρανικά κυστίδια, εντούτοις, δεν παρατηρούνται οργανωμένα οργανίδια, πχ μιτοχόνδρια ή χλωροπλάστες, όπως συμβαίνει με τα ευκαρυωτικά κύτταρα.

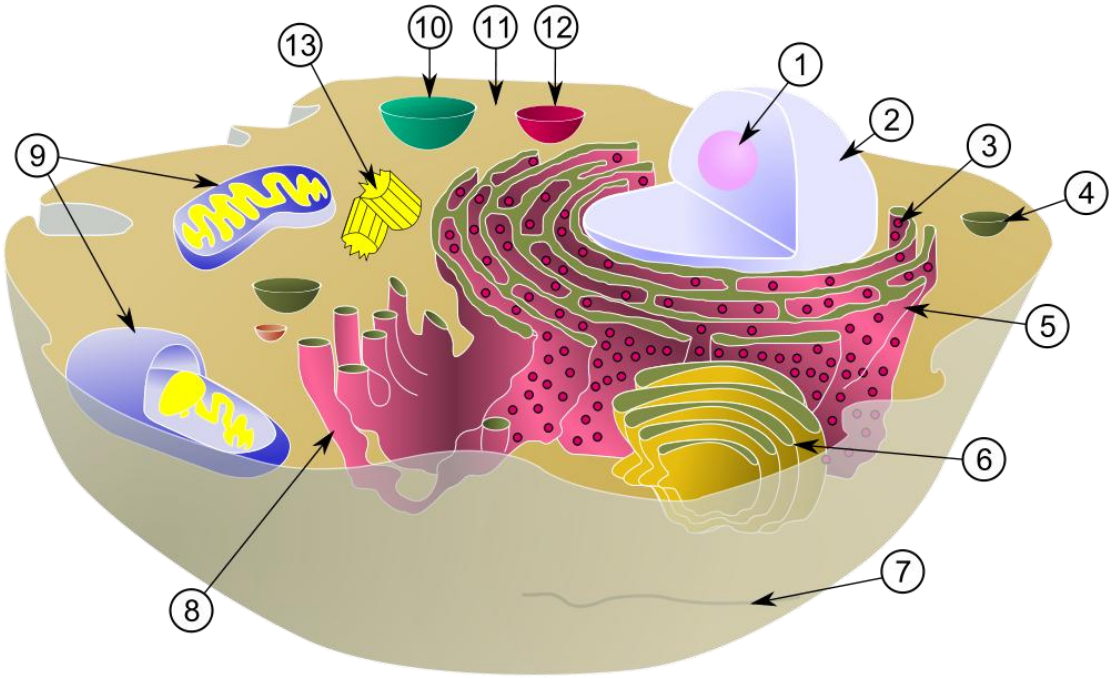
Τα ευκαρυωτικά κύτταρα περιβάλλονται από πλασματική μεμβράνη και παρουσιάζουν μεγάλο βαθμό εξειδικευμένης διαμερισματοποίησης. Η διαμερισματοποίηση αυτή, που επιτυγχάνεται τόσο στα ζώα, όσο και στα φυτά με τις ενδοκυτταρικές μεμβράνες, επιτρέπει τη διεξαγωγή βιοχημικών αντιδράσεων σε συγκεκριμένους χώρους και ταυτόχρονα διατηρεί την απαιτούμενη συγκέντρωση των αντιδρώντων μορίων, παρά το μεγάλο μέγεθος των κυττάρων. Στα ευκαρυωτικά κύτταρα ο πυρήνας διαχωρίζεται από το κυτταρόπλασμα με μια διπλομεμβρανική δομή που ονομάζεται πυρηνικός φάκελος. Στον πυρήνα υπάρχουν τα ινίδια χρωματίνης (που σχηματίζονται από DNA, ιστόνες και μη-ιστονικές πρωτεΐνες) και ένας ή περισσότεροι πυρηνίσκοι. Στο κυτταρόπλασμα υπάρχουν πολυάριθμα ριβοσώματα και μεμβρανικά συστήματα που σχηματίζουν το ενδοπλασματικό δίκτυο, τη συσκευή Golgi, τα μικροσωμάτια, τα μιτοχόνδρια και ένα μεγάλο αριθμό διαφόρων κυστιδίων. Σε όλα τα ευκαρυωτικά κύτταρα υπάρχουν μικροσωληνίσκοι και τα μικροϊνίδια, τα οποία μαζί με τα ενδιάμεσα ινίδια που παρατηρούνται στα ζωικά κύτταρα, αποτελούν τον κυτταρικό σκελετό. Στα φυτικά κύτταρα παρατηρείται επίσης κυτταρικό τοίχωμα, χλωροπλάστες και μεγάλα κενοτόπια. [5]

Ένα ευκαρυωτικό, ζωικό κύτταρο αποτελείται από τα παρακάτω:

1. Πυρηνίσκος
2. Πυρήνας
3. Ριβόσωμα
4. Κυστίδιο
5. Τραχύ ενδοπλασματικό δίκτυο
6. Συσκευή Golgi

7. Κυτταροσκελετός
8. Λείο ενδοπλασματικό δίκτυο
9. Μιτοχόνδρια
10. Κενοτόπιο
11. Κυτταρόπλασμα
12. Λυσόσωμα
13. Κεντριόλια μέσα σε Κεντροσωμάτιο

Ολόκληρο το κύτταρο περιβάλλεται από την κυτταρική ή πλασματική μεμβράνη. Οι δομές αυτές αριθμημένες φαίνονται στην παρακάτω εικόνα [6]:



Εικόνα 1. *Διάγραμμα ενός τυπικού ζωικού ευκαρυωτικού κυττάρου, όπου δείχνονται και οι υποκυττάριες μονάδες.*

2.1.3 Γενετικό υλικό, Γονιδίωμα, Γονίδια, Χρωμοσώματα

Το *δεσοξυριβοζονουκλεϊκό οξύ ή αλλιώς DNA* όπως και το *ριβοζονουκλεϊκό οξύ ή RNA*, είναι ένα μακρομόριο, που αποτελείται από νουκλεοτίδια. Κάθε νουκλεοτίδιο του DNA αποτελείται από μια πεντόζη, τη *δεσοξυριβόζη*, ενωμένη με μια φωσφορική ομάδα και μια αζωτούχο βάση. Στα νουκλεοτίδια του DNA η αζωτούχος βάση μπορεί να είναι μία από τις : *αδενίνη* (A), *θυμίνη* (T), *κυτοσίνη* (C) και *γουανίνη* (G). Στο RNA αντίστοιχα η πεντόζη είναι η *ριβόζη* και η βάση θυμίνη αντικαθίσταται από την *ουρακίλη* (U). Σε κάθε νουκλεοτίδιο η αζωτούχος βάση συνδέεται με τον 1’ άνθρακα της δεσοξυριβόζης και η φωσφορική ομάδα με τον 5’ άνθρακα. Μία πολυνουκλεοτιδική αλυσίδα σχηματίζεται από την ένωση πολλών νουκλεοτιδίων 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 με τον 5’άνθρακα της πεντόζης του επόμενου νουκλεοτιδίου. Ο δεσμός αυτός ονομάζεται 3’-5’ φωσφοδιεστερικός δεσμός. Με τον τρόπο αυτό η πολυνουκλεοτιδική αλυσίδα που δημιουργείται έχει έναν σκελετό, που αποτελείται από επανάληψη των μορίων *φωσφορική ομάδα-πεντόζη-φωσφορική ομάδα-πεντόζη*. Ανεξάρτητα από τον αριθμό των νουκλεοτιδίων από τα οποία αποτελείται η πολυνουκλεοτιδική αλυσίδα, το πρώτο της νουκλεοτίδιο 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. Για το λόγο αυτό αναφέρεται ότι ο προσανατολισμός της πολυνουκλεοτιδικής αλυσίδας είναι 5’→3’.

Το DNA αποτελεί το γενετικό υλικό όλων των κυττάρων και των περισσοτέρων ιών. Κάποιοι ιοί έχουν ως γενετικό υλικό RNA (RNA-ιοί). Συνοπτικά οι λειτουργίες του γενετικού υλικού είναι:

1. Η *αποθήκευση της γενετικής πληροφορίας*. Στο DNA (ή στο RNA των RNA-ιών) περιέχονται οι πληροφορίες που καθορίζουν όλα τα χαρακτηριστικά ενός οργανισμού και οι οποίες οργανώνονται σε λειτουργικές μονάδες, τα *γονίδια*.
2. Η διατήρηση και η μεταβίβαση της γενετικής πληροφορίας από κύτταρο σε κύτταρο και από οργανισμό σε οργανισμό, που εξασφαλίζονται με τον *αυτοδιπλασιασμό του DNA*.
3. Η έκφραση των γενετικών πληροφοριών, που επιτυγχάνεται με τον έλεγχο της σύνθεσης των πρωτεϊνών.

Το γενετικό υλικό ενός κυττάρου αποτελεί το *γονιδίωμά* του. Τα κύτταρα στα οποία το γονιδίωμα υπάρχει σε ένα μόνο αντίγραφο, όπως είναι τα προκαρυωτικά κύτταρα και οι γαμέτες των διπλοειδών οργανισμών ονομάζονται *απλοειδή*. Τα κύτταρα στα οποία το γονιδίωμα υπάρχει σε δύο αντίγραφα, όπως είναι τα σωματικά κύτταρα των ανώτερων ευκαρυωτικών οργανισμών, ονομάζονται διπλοειδή. Στα ευκαρυωτικά κύτταρα το γενετικό υλικό κατανέμεται στον πυρήνα, στα μιτοχόνδρια και στους χλωροπλάστες. Συνήθως ο όρος γονιδίωμα αναφέρεται στο γενετικό υλικό που βρίσκεται στον πυρήνα. Για την περιγραφή του μήκους ή της αλληλουχίας ενός νουκλεϊκού οξέος, χρησιμοποιείται ο όρος αριθμός ή αλληλουχία βάσεων αντίστοιχα. Στην πραγματικότητα, εννοούμε τον αριθμό ή την ακολουθία των νουκλεοτιδίων του νουκλεϊκού οξέως. Η απλούστευση αυτή γίνεται γιατί το μόνο τμήμα του νουκλεοτιδίου που αλλάζει είναι η αζωτούχος βάση.

Η διπλή έλικα του DNA τυλίγεται γύρω από πρωτεΐνες που ονομάζονται *ιστόνες* σχηματίζοντας τα *νουκλεοσώματα*, τα οποία πακετάρονται και δημιουργούν *ινίδια χρωματίνης*. Τα ινίδια χρωματίνης αναδιπλώνονται στον χώρο σχηματίζοντας πολύπλοκες δομές που ονομάζονται *χρωμοσώματα*.[6]

2.1.4 Έκφραση Γενετικής Πληροφορίας – Κεντρικό Δόγμα της Μοριακής Βιολογίας

Το DNA ενός οργανισμού αποτελεί την αποθήκη που περιέχει ακριβείς οδηγίες, οι οποίες καθορίζουν τη δομή και τη λειτουργία του οργανισμού. Ταυτόχρονα περιέχει την πληροφορία για τον αυτοδιπλασιασμό του εξασφαλίζοντας έτσι τη μεταβίβασή των γενετικών οδηγιών από ένα κύτταρο στα θυγατρικά του και από έναν οργανισμό στους απογόνους του.

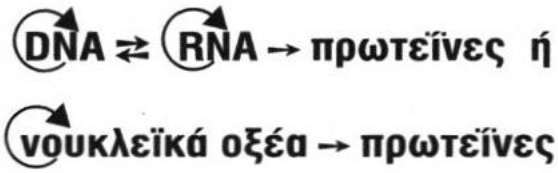
Το πρώτο βήμα για την έκφραση της πληροφορίας που υπάρχει στο DNA είναι η μεταφορά της στο RNA με τη διαδικασία της μεταγραφής. Το RNA μεταφέρει με τη σειρά του, μέσω της διαδικασίας της μετάφρασης, την πληροφορία στις πρωτεΐνες που είναι υπεύθυνες για τη δομή και τη λειτουργία των κυττάρων και κατ' επέκταση και των οργανισμών.

Η σχέση αυτή συνοψίζεται στο ακόλουθο σχήμα, όπου τα βέλη δείχνουν την κατεύθυνση της μεταφοράς της γενετικής πληροφορίας:

Το σχήμα αυτό αποτελεί το *κεντρικό δόγμα της Μοριακής Βιολογίας*, όπως ονομάστηκε από τον F.Crick το 1958. Η γενετική πληροφορία είναι η καθορισμένη σειρά των βάσεων, όπως η πληροφορία μιας γραπτής φράσης είναι η σειρά των γραμμάτων που την αποτελούν. Η πληροφορία υπάρχει σε τμήματα του DNA με συγκεκριμένη ακολουθία, τα γονίδια. Αυτά δια μέσου της *μεταγραφής* και της *μετάφρασης* καθορίζουν τη σειρά των αμινοξέων στην πρωτεΐνη. Οι πορείες της μεταγραφής και της μετάφρασης των γονιδίων αποτελούν τη *γονιδιακή έκφραση*.

Για αρκετό καιρό, οι ερευνητές πίστευαν ότι όλη η ροή της γενετικής πληροφορίας γινόταν προς τη μία μόνο κατεύθυνση, δηλαδή ότι το DNA μεταγραφόταν σε RNA. Σήμερα είναι γνωστό ότι μερικοί ιοί έχουν RNA ως γενετικό υλικό. Ένα ένζυμο που υπάρχει στους ίδιους τους ιούς, η *αντίστροφη μεταγραφάση*, χρησιμοποιεί ως καλούπι το RNA για να συνθέσει DNA. Επιπλέον σε ορισμένους ιούς το RNA έχει την ικανότητα να αυτοδιπλασιάζεται.

Έτσι σήμερα το κεντρικό δόγμα περιγράφεται ως εξής:



Διαπιστώνουμε λοιπόν, ότι η αντιγραφή του DNA διαιωνίζει τη γενετική πληροφορία ενώ η μετάφραση χρησιμοποιεί αυτή την πληροφορία για να κατασκευάσει ένα πολυπεπτίδιο. Η

μεταγραφή καθορίζει ποια γονίδια θα εκφραστούν ,σε ποιους ιστούς(στους πολυκύτταρους ευκαρυωτικούς οργανισμούς) και σε ποια στάδια της ανάπτυξης.

Όλα τα κύτταρα ενός πολυκύτταρου οργανισμού έχουν το ίδιο DNA. Σε κάθε ομάδα κυττάρων όμως εκφράζονται διαφορετικά γονίδια. Στα πρόδρομα ερυθροκύτταρα, για παράδειγμα, εκφράζονται κυρίως τα γονίδια των αντισωμάτων. Τα γονίδια διακρίνονται σε δύο κατηγορίες.

1. Στα γονίδια που μεταγράφονται σε mRNA και μεταφράζονται στη συνέχεια σε πρωτεΐνες και
2. Στα γονίδια που μεταγράφονται και παράγουν άλλα είδη RNA όπως τα tRNA, rRNA, snRNA, microRNA.

Από το σύνολο του ανθρώπινου γονιδιώματος, μόνο ένα μικρό ποσοστό μεταγράφεται σε RNA, δηλαδή αποτελεί τα γονίδια. Υπάρχουν αρκετά είδη μορίων RNA που παράγονται με τη μεταγραφή. Μερικά από αυτά είναι: το αγγελιοφόρο RNA (mRNA), το μεταφορικό RNA (tRNA), το ριβοσωμικό RNA (rRNA), το μικρό πυρηνικό RNA (snRNA) καθώς και το mircoRNA.

Τα τρία πρώτα είδη υπάρχουν και στους προκαρυωτικούς και στους ευκαρυωτικούς οργανισμούς, αλλά το τέταρτο και το πέμπτο υπάρχουν μόνο στους ευκαρυωτικούς. Η λειτουργία κάθε είδους RNA περιγράφεται παρακάτω:

1. *Αγγελιοφόρο RNA (mRNA)*. Τα μόρια αυτά μεταφέρουν την πληροφορία του DNA για την παραγωγή μιας πολυπεπτιδικής αλυσίδας.
2. *Ριβοσωμικό RNA (rRNA)*. Τα μόρια αυτά συνδέονται με πρωτεΐνες και σχηματίζουν το *ριβόσωμα*, ένα «σωματίδιο» απαραίτητο για την πραγματοποίηση της πρωτεϊνοσύνθεσης.
3. *Μεταφορικό RNA (tRNA)*. Κάθε μεταφορικό RNA συνδέεται με ένα συγκεκριμένο αμινοξύ και το μεταφέρει στη θέση της πρωτεϊνοσύνθεσης.
4. *Μικρό πυρηνικό RNA (snRNA)*. Είναι μικρά μόρια RNA, τα οποία συνδέονται με πρωτεΐνες και σχηματίζουν μικρά ριβονουκλεοπρωτεϊνικά σωματίδια. Τα σωματίδια αυτά καταλύουν την «ωρίμανση» του mRNA, μια διαδικασία που γίνεται μόνο στους ευκαρυωτικούς οργανισμούς. [7]
5. *microRNA (miRNA)*. Ανήκουν στην κατηγορία των RNA που δεν κωδικοποιούν πρωτεΐνες, αλλά παρόλα αυτά ρυθμίζουν τη διαδικασία της μετάφρασης. Είναι μικρά σε μήκος και απαντώνται σε φυτά, ζώα και μερικούς ιούς. Λειτουργούν μετά τη μεταγραφή και προκύπτουν από την μετάφραση του DNA. Μέσω σύνδεσης των βάσεων με συμπληρωματικές ακολουθίες πάνω σε διάφορα μόρια mRNA, προκαλούν το μπλοκάρισμα τους γιατί δεν μπορούν να μεταφραστούν πλέον σε πρωτεΐνες από τα ριβοσώματα. Το ανθρώπινο γονιδίωμα κωδικοποιεί πάνω από 1000

miRNA, που μπορούν να «στοχεύσουν» παραπάνω από το 60% των γονιδίων των θηλαστικών. Τέλος βρίσκονται σε μεγάλη αφθονία σε πολλούς τύπους κυττάρων του ανθρώπινου οργανισμού.[7]

2.1.5 Μεταγραφή

Ο μηχανισμός της μεταγραφής είναι ίδιος στους προκαρυωτικούς και ευκαρυωτικούς οργανισμούς. Η μεταγραφή καταλύεται από ένα ένζυμο, την *RNA πολυμεράση* (στους ευκαρυωτικούς οργανισμούς υπάρχουν τρία είδη RNA πολυμερασών).

Η RNA πολυμεράση συνδέεται σε ειδικές περιοχές του DNA, που ονομάζονται *υποκινητές*, με τη βοήθεια πρωτεϊνών που ονομάζονται *μεταγραφικοί παράγοντες*. Οι υποκινητές και οι μεταγραφικοί παράγοντες αποτελούν τα ρυθμιστικά στοιχεία της μεταγραφής του DNA και επιτρέπουν στην RNA πολυμεράση να αρχίσει σωστά τη μεταγραφή. Οι υποκινητές βρίσκονται πάντοτε πριν από την αρχή κάθε γονιδίου.

Κατά την έναρξη της μεταγραφής ενός γονιδίου, η RNA πολυμεράση προσδένεται στον υποκινητή και προκαλεί τοπικό ξετύλιγμα της διπλής έλικας του DNA. Στη συνέχεια, τοποθετεί τα ριβονουκλεοτίδια απέναντι από τα δεσοξυριβονουκλεοτίδια μιας αλυσίδας του DNA σύμφωνα με τον κανόνα της συμπληρωματικότητας των βάσεων, όπως και στην αντιγραφή, με τη διαφορά ότι εδώ, απέναντι από την αδενίνη τοποθετείται το ριβονουκλεοτίδιο που περιέχει ουρακίλη. Η RNA πολυμεράση συνδέει τα ριβονουκλεοτίδια που προστίθενται που προστίθενται το ένα μετά το άλλο, με 3’-5’-φωσφοριεστερικό δεσμό. Η μεταγραφή έχει προσανατολισμό 5’→3’ όπως και η αντιγραφή. Η σύνθεση του RNA σταματά στο τέλος του γονιδίου, όπου ειδικές αλληλουχίες, που ονομάζονται *αλληλουχίες λήξης της μεταγραφής*, επιτρέπουν την απελευθέρωσή του. Το μόριο RNA που συντίθεται είναι συμπληρωματικό προς τη μία αλυσίδα της διπλής έλικας του DNA του γονιδίου. Η αλυσίδα αυτή είναι η μεταγραφόμενη και ονομάζεται μη-κωδική. Η συμπληρωματική αλυσίδα του DNA του γονιδίου ονομάζεται κωδική. Το RNA είναι το κινητό αντίγραφο της πληροφορίας ενός γονιδίου.

Στους προκαρυωτικούς οργανισμούς το mRNA αρχίζει να μεταφράζεται σε πρωτεΐνη πριν ακόμη ολοκληρωθεί η μεταγραφή του. Αυτό είναι δυνατό επειδή δεν υπάρχει πυρηνική μεμβράνη.

Αντίθετα, στους ευκαρυωτικούς οργανισμούς, το mRNA που παράγεται κατά τη μεταγραφή ενός γονιδίου συνήθως δεν είναι έτοιμο να μεταφραστεί, αλλά υφίσταται μια πολύπλοκη διαδικασία ωρίμανσης. Η διαδικασία αυτή είναι ένα από τα πιο ενδιαφέροντα ευρήματα της Μοριακής Βιολογίας, γιατί οδήγησε στο συμπέρασμα ότι τα περισσότερα γονίδια των ευκαρυωτικών οργανισμών (και των ιών που τους προσβάλλουν) είναι *ασυνεχή ή διακεκομμένα*. Δηλαδή η αλληλουχία που μεταφράζεται σε αμινοξέα διακόπτεται από

ενδιάμεσες αλληλουχίες οι οποίες δεν μεταφράζονται σε αμινοξέα. Οι αλληλουχίες που μεταφράζονται σε αμινοξέα ονομάζονται *εξώνια* και οι ενδιάμεσες αλληλουχίες ονομάζονται *εσώνια*.

Όταν ένα γονίδιο που περιέχει εσώνια μεταγράφεται, δημιουργείται το *πρόδρομο mRNA* που περιέχει εξώνια και εσώνια. Το πρόδρομο mRNA μετατρέπεται σε mRNA με τη διαδικασία της *ωρίμανσης*, κατά την οποία τα εσώνια κόβονται από μικρά ριβονουκλεοπρωτεϊνικά σωματίδια που αποτελούνται από *snRNA και από πρωτεΐνες* και λειτουργούν ως ένζυμα : κόβουν τα εσώνια και συρράπτουν τα εξώνια μεταξύ τους. Έτσι σχηματίζεται το *«ώριμο» mRNA*. [8]

Παρόμοια διαδικασία ακολουθείται και για τα miRNAs. Το αρχικό προϊόν της μεταγραφής ονομάζεται *hairpin miRNA*, το οποίο μέσω της ωρίμανσης μετατρέπεται σε *mature miRNA (ώριμο miRNA)*.[9]

2.2 Βάσεις Δεδομένων με Πληροφορίες για γονίδια

Στο συγκεκριμένο κεφάλαιο θα γίνει μια αναφορά στις γονιδιακές βάσεις δεδομένων. Η πλήρης λίστα των βάσεων που ασχολούνται με γονίδια αποτελείται από τις [9]:

1. Bioinformatic Harvester
2. SNPedia
3. CAMERA Πηγή μικροβιακού γονιδιώματος και μετα-γονιδιώματος
4. Corn, η Βάση Δεδομένων Γενετικής και Γονιδιώματος Maize
5. EcoCyc μία βάση δεδομένων που περιγράφει το γονιδίωμα και τη βιοχημική μηχανική του οργανισμού μοντέλο *E. coli K-12*
6. Ensembl παρέχει αυτοματοποιημένες βάσεις δεδομένων γονιδιώματος για τον άνθρωπο, το ποντίκι και άλλα σπονδυλωτά και ευκαρυωτικούς οργανισμούς. Ensembl Genomes παρέχει δεδομένα γονιδιακής κλίμακας για βακτήρια, πρωτόζωα, μύκητες, φυτά και ασπόνδυλα μετάζωα , μέσω ενός ενοποιημένου συνόλου διαδραστικών και προγραμματιστικών διεπαφών (χρησιμοποιώντας ως βάση την πλατφόρμα λογισμικού της Ensembl)
7. PATRIC, the PathoSystems Resource Integration Center
8. Flybase, το γονιδίωμα του οργανισμού-μοντέλο Drosophila Melanogaster
9. MGI Mouse Genome (Jackson Lab.) η ΒΔ γονιδιώματος του ποντικού
10. JGI Genomes του DOE-Joint Genome Institute παρέχει βάσεις δεδομένων διαφόρων γονιδιωμάτων από ευκαρυωτικούς και μικροβιακούς οργανισμούς.
11. National Microbial Pathogen Data Resource. Μία βάση δεδομένων με χειροκίνητη επιμέλεια, η οποία παρέχει γονιδιακή πληροφορία για τα παρακάτω παθογόνα μικρόβια: Campylobacter, Chlamydia, Chlamydophila, Haemophilus, Listeria,

Mycoplasma, Neisseria, Staphylococcus, Streptococcus, Treponema, Ureaplasma, and Vibrio.

12. RegulonDB η RegulonDB είναι ένα μοντέλο της λειτουργίας που έχει να κάνει είτε με την σύνθετη ρύθμιση της έναρξης της μεταγραφής στο κύτταρο E. coli K-12 είτε με ρυθμιστικό δίκτυο του κυττάρου αυτού.
13. Saccharomyces Genome Database, γονιδίωμα του οργανισμού-μοντέλο της μαγιάς.
14. Viral Bioinformatics Resource Center Επιμελημένη Βάση Δεδομένων που παρέχει γονιδιακή πληροφορία για 11 τύπους οικογενειών από ιούς.
15. Η πλατφόρμα SEED για γονιδιακή ανάλυση μικροβιακών οργανισμών περιέχει όλα τα πλήρη και τα περισσότερα τμηματικά γονιδιώματα όλων των μικροβιακών οργανισμών. Η πλατφόρμα χρησιμοποιείται για να γίνει επισημείωση των μικροβιακών γονιδιωμάτων χρησιμοποιώντας υποσυστήματα.
16. Xenbase, το γονιδίωμα των οργανισμών-μοντέλο Xenopus tropicalis και Xenopus laevis
17. Wormbase, το γονιδίωμα του οργανισμού-μοντέλο Caenorhabditis elegans
18. Zebrafish Information Network, γονιδίωμα του ψαριού-μοντέλο Zebrafish.
19. TAIR, Η πηγή πληροφορίας του λουλουδιού Arabidopsis.
20. UCSC Malaria Genome Browser, το γονιδίωμα των οργανισμών που προκαλεί ελονοσία (*Plasmodium falciparumata* και άλλοι)
21. RGD Rat Genome Database: Γονιδιακή και φαινοτυπική πληροφορία για τον αρουραίο Rattus norvegicus
22. INTEGRALL: Βάση δεδομένων αφιερωμένη στα integrons, που είναι βακτηριακά γενετικά στοιχεία που συμμετέχουν στην αντίσταση στα αντιβιοτικά.
23. Fourmidable ant genome database παρέχει αναζήτηση blast για το γονιδίωμα του μερμηγκιού καθώς επίσης παρέχει και «κατέβασμα» ακολουθιών.
24. VectorBase Το κέντρο πόρων βιοπληροφορικής NIAID για ασπόνδυλα στελέχη οργανισμών παθογόνων για τον άνθρωπο.
25. EzGenome, περιεκτικές πληροφορίες για χειροκίνητα επιμελημένα γονιδιακά δεδομένα από προκαρυωτικούς οργανισμούς (αρχαιοβακτηρίων και βακτηρίων).
26. HUGO επίσημη βάση δεδομένων ονοματολογίας των ανθρωπίνων γονιδίων
27. NCBI-UniGene αμερικάνικο εθνικό κέντρο πληροφορίας που αφορά τη Βιοτεχνολογία.

Παρακάτω θα γίνει αναφορά στις βάσεις δεδομένων που χρησιμοποιήθηκαν καθώς και σε ανταγωνιστικές τους. Οι βάσεις από τις οποίες αντλήσαμε τα δεδομένα είναι η Hugo και η Ensembl. Ο αμερικάνικος «ανταγωνιστής» της Ensembl είναι η NCBI. Πρωτού

προχωρήσουμε όμως είναι απαραίτητο να γίνει θα γίνει αναφορά στο Genome Reference Consortium (κοινοπραξία ανθρώπινου γονιδιώματος).

2.2.1 Genome Reference Consortium (GRC)

Ένα *γονιδίωμα αναφοράς* (reference genome) γνωστό και ως *συναρμολόγηση αναφοράς* (reference assembly) είναι μία ψηφιακή βάση δεδομένων ακολουθίας νουκλεϊκών οξέων, η οποία έχει συναρμολογηθεί από τους επιστήμονες, ως ένα αντιπροσωπευτικό παράδειγμα του συνόλου τον γονιδίων κάποιου είδους. Καθότι συνήθως συναρμολογούνται από την ακολουθία του DNA, που προέρχεται από έναν αριθμό δωρητών, τα γονιδιώματα αναφοράς δεν μπορούν να αναπαραστήσουν με ακρίβεια το σύνολο των γονιδίων κάθε μεμονωμένου ατόμου. Αντίθετα, κάθε αναφορά προσφέρει ένα απλοτυπικό μωσαϊκό από διαφορετικές ακολουθίες DNA του κάθε δωρητή.

Για παράδειγμα η GRCh37 ή αλλιώς το ανθρώπινο γονιδίωμα αναφοράς του Genome Reference Consortium (έκδοση 37) προέρχεται από δεκατρείς ανώνυμους εθελοντές από το Buffalo της Νέας Υόρκης. Το σύστημα ομάδων αίματος ΑΒΟ διαφέρει από άνθρωπο σε άνθρωπο, αλλά το γονιδίωμα αναφοράς του ανθρώπου περιέχει μόνο ένα αλληλόμορφο γονίδιο Ο (παρότι και τα άλλα αλληλόμορφα σημειώνονται).

Τα γονιδιώματα αναφοράς του ανθρώπου και του ποντικού, συντηρούνται και βελτιώνονται από την *Κοινοπραξία Γονιδιώματος Αναφοράς* (*Genome Reference Consortium – GRC*). Η GRC είναι μια ομάδα που αποτελείται από λιγότερους από 20 επιστήμονες, που προέρχονται από έναν αριθμό ινστιτούτων έρευνας γονιδιώματος. Αυτά περιλαμβάνουν το *Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής* (European *Bioinformatics Institute - EMBL*), το *Εθνικό Κέντρο Πληροφορίας της Βιοτεχνολογίας* (*National Center for Biotechnology Information – NCBI*), το *Ινστιτούτο Sanger* καθώς και το *πανεπιστήμιο της Ουάσινγκτον στο st.Louis*. Η GRC, συνεχίζει να βελτιώνει τα γονιδιώματα αναφοράς με την οικοδόμηση νέων στοιχίσεων, που περιέχουν λιγότερα κενά και που διορθώνουν λάθος ερμηνείες στην ακολουθία. Από το 2010, το γονιδίωμα αναφοράς του ανθρώπου είναι στην 19^{η} έκδοσή του. Η έκδοση GRCh37 περιέχει περίπου 250 κενά, ενώ η πρώτη έκδοση είχε περίπου 150.000 κενά.[11]

2.2.2 National Centre for Biotechnology Information (NCBI)

Η βάση δεδομένων NCBI ανήκει στο ομοσπονδιακό κράτος των ΗΠΑ. Παρέχει μια βάση δεδομένων για γονίδια διαφόρων οργανισμών χωρίς πρόσβαση σε παλαιότερες εκδόσεις. Παρόλα αυτά, παρέχει ένα αρχείο με το ιστορικό των αλλαγών.

Το εγγενές αναγνωριστικό του γονιδίου που χρησιμοποιείται από τη βάση αυτή είναι το tax_id με βάση την ταξινόμηση που γίνεται από την ίδια την NCBI.

Η πληροφορία παρέχεται σε ένα μεγάλο σύνολο αρχείων τα οποία προσφέρονται για κατέβασμα από έναν FTP server και ανανεώνονται καθημερινά. Διατίθενται τα εξής αρχεία:

- 1 αρχείο (gene_history) με πληροφορία για το ιστορικό γονιδίων που δεν είναι ανανεωμένα,
- 1 αρχείο (gene_info) με το σύνολο του γονιδιώματος που περιέχεται στη βάση (περίπου 1.5 GB)
- 1 αρχείο (gene_neighbors) με πληροφορίες για τα γειτονικά γονίδια ενός δεδομένου γονιδίου
- 1 αρχείο (gene_groups) με πληροφορία για τις ομάδες των γονιδίων
- Πολλά αρχεία που συσχετίζουν τα Gene IDs με τα αναγνωριστικά άλλων βάσεων γονιδιώματος όπως η Ensembl ή άλλες ιατρικές βάσεις όπως η PubMed.

2.2.2.1 Αρχείο gene_info

Υπάρχει ένας φάκελος που περιέχει αρχεία για κάθε οργανισμό ξεχωριστά, με βάση τη δομή του αρχείου gene_info που περιγράφεται παρακάτω.

Το αρχείο gene_info είναι αρχείο κειμένου και περιέχει την πληροφορία δομημένη σε στήλες. Τα γονίδια χωρίζονται μεταξύ τους με μια αλλαγή γραμμής. Η πληροφορίες για κάθε γονίδιο χωρίζονται με χαρακτήρα στήλης (Tab). Ένα δείγμα του αρχείου φαίνεται παρακάτω:

```
#Format: tax_id GeneID Symbol LocusTag Synonyms dbXrefs chromosome
map_location description type_of_gene Symbol_from_nomenclature_authority
Full_name_from_nomenclature_authority Nomenclature_status Other_designations
Modification_date (tab is used as a separator, pound sign - start of a
comment)
```

```
9606   1      A1BG   -      A1B|ABG|GAB|HYST2477
       HGNC:5|MIM:138670|Ensembl:ENSG00000121410|HPRD:00726|Vega:OTTHUMG00000
183507 19     19q13.4      alpha-1-B glycoprotein     protein-coding
       A1BG   alpha-1-B glycoprotein     O      alpha-1B-glycoprotein
       20131006
```

Η πρώτη γραμμή είναι σχόλιο για επεξηγηματικούς σκοπούς και ορίζει το σε ποιο πεδίο αναφέρεται η κάθε στήλη. Τα διαθέσιμα πεδία είναι τα εξής:

- tax_id : το μοναδικό αναγνωριστικό με βάση την ταξινόμηση(taxonomy) της NCBI για το κάθε είδος ή στέλεχος.
- GeneID : το μοναδικό αναγνωριστικό για κάθε γονίδιο.
- Symbol : το σύμβολο του γονιδίου.
- LocusTag : η τιμή LocusTag του γονιδίου.
- Synonyms : συνώνυμες ονομασίες για το γονίδιο. Χωρίζονται με «|».
- dbXrefs : αναγνωριστικά για άλλες βάσεις γονιδιακών δεδομένων. Χωρίζονται με «|».
- chromosome : το χρωμόσωμα στο οποίο βρίσκεται το γονίδιο
- map_location : θέση στο γονιδιακό χάρτη.
- description : περιγραφή του γονιδίου.
- type_of_gene : ο τύπος του γονιδίου με βάση τις επιλογές που βρίσκονται εδώ: http://www.ncbi.nlm.nih.gov/IEB/ToolBox/CPP_DOC/lxr/source/src/objects/entrezgene/entrezgene.asn
- Symbol_from_nomenclature_authority : σύμβολο με βάση την αρχή ονοματολογίας στην οποία ανήκει.
- Full_name_from_nomenclature_authority : πλήρες όνομα με βάση τη αρχή ονοματολογίας στη οποία ανήκει.
- Nomenclature_status : η κατάσταση του ονόματος στην ονοματολογία. Το «O» δείχνει επίσημη κατάσταση ενώ το «I» προσωρινή.
- Other_designations : Άλλες ονομασίες που έχουν αναφερθεί για το γονίδιο.
- Modification_date : Ημερομηνία τελευταίας τροποποίησης στη μορφή ΕΕΕΕΜηΜηΜεΜε.

Αν ένα πεδίο είναι κενό, τότε αυτό υποδηλώνεται με το σύμβολο πλην «-».

Ένα παράδειγμα εγγραφής για το γονιδίωμα του ανθρώπου είναι το εξής

tax_id	GeneID	Symbol	LocusTag	Synonyms
9606	2	A2M	-	A2MD|CPAMD5|FWP007|S863-7

dbXrefs
HGNC:7|MIM:103950|Ensembl:ENSG00000175899|HPRD:00072|Vega:OTTHUMG00000150267

chromosome	map_location	description	type_of_gene
12	12p13.31	alpha-2-macroglobulin	protein-coding

Symbol_from_nomenclature_authority	Full_name_from_nomenclature_authority	Nomenclature_status
A2M	alpha-2-macroglobulin	O

Other_designations
C3 and PZP-like alpha-2-macroglobulin domain-containing protein 5|alpha-2-M

Modification_date
20131006

2.2.2.2 RefSeq

Η βάση δεδομένων *Reference Sequence (RefSeq)* της NCIB είναι μια συλλογή από ταξινομητικά ποικίλες, μη επαναλαμβανόμενες και επαρκώς επισημειωμένες ακολουθίες, που αναπαριστούν μόρια από DNA, RNA και πρωτεΐνες που προκύπτουν με φυσικό τρόπο. Περιλαμβάνονται ακολουθίες από πλασμίδια, οργανίδια, ιούς, αρχαιοβακτήρια, βακτήρια και ευκαρυωτικούς οργανισμούς. Κάθε RefSeq κατασκευάζεται πλήρως από ακολουθιακή πληροφορία που υποβάλλεται στο International Nucleotide Sequence Database Collaboration (INSDC). Παρόμοια με ένα review article, κάθε RefSeq είναι μια σύνθεση από ενσωματωμένη πληροφορία που βρίσκεται σε πολλές πηγές ταυτόχρονα. Τα RefSeqs παρέχουν ένα θεμέλιο για να ενοποιηθούν ακολουθιακά δεδομένα με γενετική και λειτουργική πληροφορία. Δημιουργούνται για να παρέχουν μια τυπική αναφορά σε δεδομένα για πολλαπλούς σκοπούς όπως για επισημειώσεις γονιδιώματος ή για αναφορά στην τοποθεσία που βρίσκεται μια παραλλαγή ακολουθίας σε ιατρικά μητρώα. Η συλλογή είναι διαθέσιμη χωρίς περιορισμούς και μπορεί να ληφθεί από διάφορες πηγές.

Η NCIB διαθέτει FTP server (ftp://ftp.ncbi.nlm.nih.gov/), όπου παρέχει αυτά τα αρχεία, τα οποία είναι σε μορφή FASTA.

Παράδειγμα τέτοιου αρχείου(από το RefSeqGene) είναι το εξής:

refseqgene1.genomic.fna

```
>gi|254939587|ref|NG_012567.1| Homo sapiens NADPH oxidase 1 (NOX1), RefSeqGene on chromosome X
ATTCTGTGATCACCAGCTTATCAAAAGACTTCCTAGTACTCTGATATTGGGAATGGGGGTCCTACCTCACAGACAT
AAGG
GTCCAATCAGCATGGCATATATAATTCTTTAGATAATACATAAATTGTCATCCAGATTATAGATCATTCTTTTATG
AATC
ACAGGATCTCAATGTTGGAGTATATTTAAGGGACATTTAGTTAACCATCTACCTGGTGCTGATATTCCCCTTATAA
AAAG…
```

Το αρχείο αυτό περιέχει 1778 ακολουθίες γονιδίων. Δημιουργήθηκε πρόγραμμα σε C, το οποίο μετρούσε αυτές τις εγγραφές και τύπωνε τις γραμμές που περιέχουν την πληροφορία χωρίς να τυπώνει τις ακολουθίες:

```
>gi|254939587|ref|NG_012567.1| Homo sapiens NADPH oxidase 1 (NOX1), RefSeqGene on chromosome X
>gi|207438642|ref|NG_008645.1| Homo sapiens dopamine beta-hydroxylase (dopamine beta-monooxygenase) (DBH), RefSeqGene on chromosome 9
>gi|224809262|ref|NG_011468.1| Homo sapiens matrix metallopeptidase 9 (gelatinase B, 92kDa gelatinase, 92kDa type IV collagenase) (MMP9), RefSeqGene on chromosome 20
>gi|261245123|ref|NG_013071.1| Homo sapiens FXYD domain containing ion transport regulator 6 (FXYD6), RefSeqGene on chromosome 11
>gi|283806565|ref|NG_016335.1| Homo sapiens ADAM metallopeptidase domain 9 (ADAM9), RefSeqGene on chromosome 8
>gi|322812157|ref|NG_028131.1| Homo sapiens feline leukemia virus subgroup C cellular receptor 1 (FLVCR1), RefSeqGene on chromosome 1
>gi|225637518|ref|NG_011691.1| Homo sapiens solute carrier family 45, member 2 (SLC45A2), RefSeqGene on chromosome 5
>gi|225543278|ref|NG_011648.1| Homo sapiens angiotensin I converting enzyme (peptidyl-dipeptidase A) 1 (ACE), RefSeqGene on chromosome 17
>gi|259013220|ref|NG_012921.1| Homo sapiens 2,3-bisphosphoglycerate mutase (BPGM), RefSeqGene on chromosome 7
>gi|216548416|ref|NG_009108.1| Homo sapiens elongation of very long chain fatty acids (FEN1/Elo2, SUR4/Elo3, yeast)-like 4 (ELOVL4), RefSeqGene on chromosome 6
>gi|301897467|ref|NG_023443.1| Homo sapiens eyes shut homolog (Drosophila) (EYS), RefSeqGene on chromosome 6
```

Το αρχείο αυτό βέβαια είναι ενδέχεται να περιέχει αρκετά πεπαλαιωμένη πληροφορία. Οι καινούριες εγγραφές βρίσκονται στον φάκελο release του FTP Server και περιέχουν το γονιδίωμα των σπονδυλωτών μαζί με τα θηλαστικά.

2.2.2.3 RefSeqGene

Το RefSeqGene είναι ένα υποσύνολο το RefSeq της NCBI και ορίζει γονιδιακές ακολουθίες που μπορούν να χρησιμοποιηθούν ως σταθερές για επαρκώς επισημειωμένα γονίδια. Αυτές οι ακολουθίες, που είναι σημειωμένες με τη λέξη κλειδί RefSeqGene στη βάση δεδομένων νουκλεοτιδίων της NCBI, χρησιμεύουν ως σταθερό θεμέλιο για την αναφορά μεταλλάξεων, τη θεμελίωση συμβάσεων για την αρίθμηση εξωνίων και εσωνίων καθώς και για τον ορισμό συντεταγμένων άλλων διαφοροποιήσεων.

Οι ακολουθίες RefSeq του mRNA και των πρωτεϊνών χρησιμοποιούνται εδώ και καιρό για αυτό το σκοπό, αλλά παρουσιάζουν την προφανή αδυναμία του να μην παρέχουν ακριβείς συντεταγμένες για συνοδευτικές ακολουθίες ή ακολουθίες εσωνίων. Οι ακολουθίες χρωμοσωμάτων RefSeq δεν παρέχουν σχετικές συντεταγμένες με βάση άλλα γονίδια, αλλά αντίθετα προσφέρουν πολύ μεγάλες τιμές απόλυτων συντεταγμένων, που αλλάζουν όταν η ακολουθία ανανεώνεται λόγω ανασυναρμολόγησης. Οι ακολουθίες του RefSeqGene project μπορούν να αντιμετωπίσουν αυτά τα μειονεκτήματα παρέχοντας μια πιο σταθερή ακολουθία για κάθε γονίδιο. Εάν πρέπει να γίνουν αλλαγές σε οποιαδήποτε ακολουθία RefSeqGene, αυτό θα γίνει σε σταθερή έκδοση και θα παρέχονται εργαλεία για τη μετατροπή των συντεταγμένων. Οι ακολουθίες RefSeqGene είναι στοιχισμένες σε χρωμοσώματα αναφοράς. Τρέχουσες καθώς και προηγούμενες χρωμοσωμικές συντεταγμένες είναι διαθέσιμες λόγω αυτής της αναστοίχισης.

Για το λόγο αυτό παρέχεται FTP server με αρχεία σε FASTA format (ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/RefSeqGene/).

2.2.3 Human Genome Organization (Hugo) – Human Genome Nomenclature Committee (HGNC)

Η HGNC (HUGO Nomenclature Committee) ή *Επιτροπή Ονοματολογίας Γονιδίων του HUGO* (Human Genome Organization) ή αλλιώς *Οργανισμός του Ανθρωπίνου Γονιδιώματος,*

είναι μία επιτροπή, που σαν σκοπό της έχει την ανάθεση τυποποιημένων αναγνωριστικών στα ανθρώπινα γονίδια.

Η HGNC αποδέχεται τόσο μία συντόμευση του ονόματος, γνωστή και ως σύμβολο του γονιδίου όσο και ένα μακρύτερο και πιο περιγραφικό όνομα. Κάθε σύμβολο που ανατίθεται, είναι μοναδικό και η επιτροπή εξασφαλίζει ότι σε κάθε γονίδιο θα δίνεται μόνο ένα αποδεκτό σύμβολο. Αυτό επιτρέπει ξεκάθαρη και χωρίς αμφισημίες αναφορά σε γονίδια μέσα στην επιστημονική κοινότητα και βοηθά στην ανάκτηση ηλεκτρονικής πληροφορίας από βάσεις δεδομένων και δημοσιεύσεις.

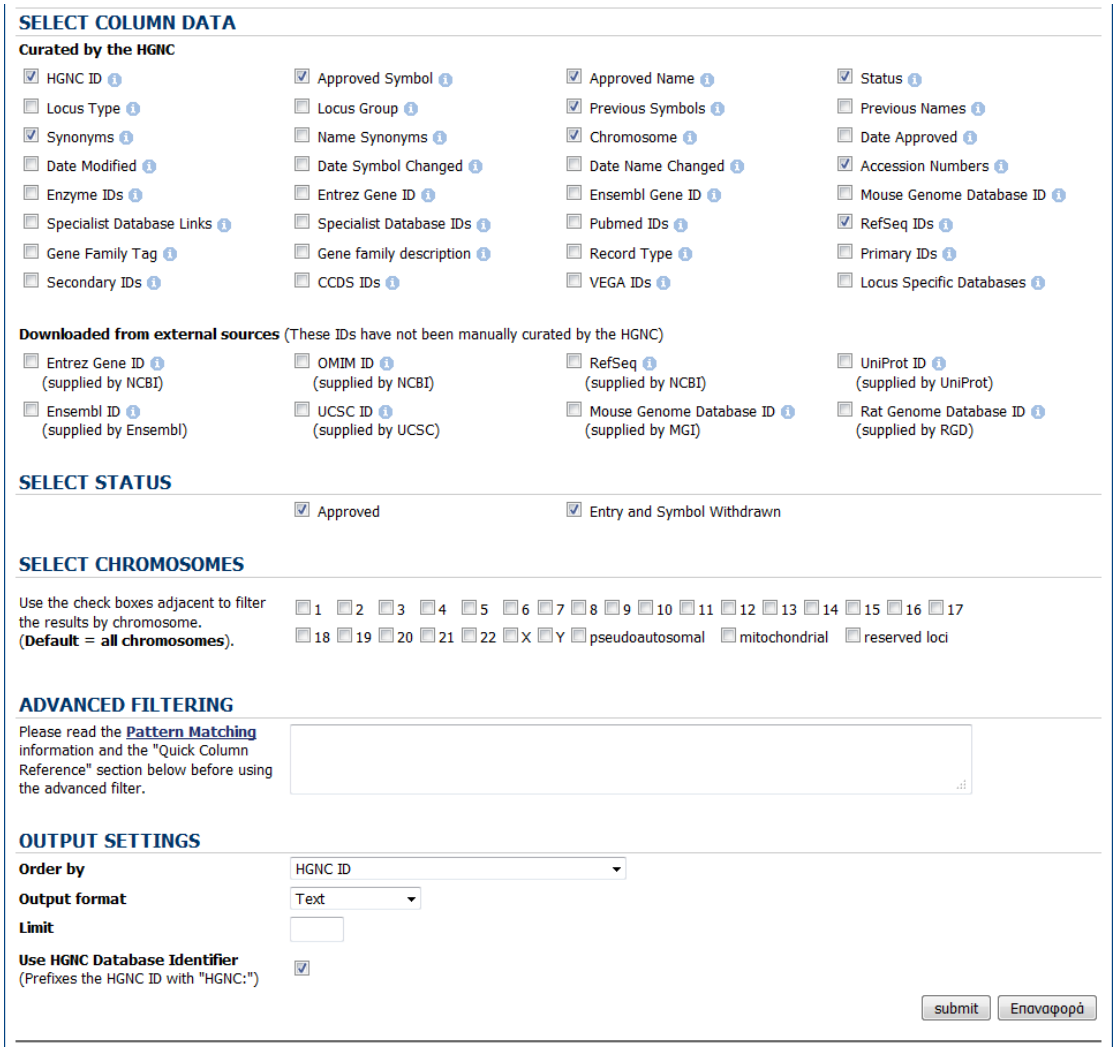
Η HGNC δίνει τη δυνατότητα για κατέβασμα των δεδομένων σε μορφή αρχείου κειμένου, αλλά δεν επιτρέπει την πρόσβαση σε παλαιότερες εκδόσεις. Παρόλα αυτά, διαθέτει προηγουμένως αποδεκτά σύμβολα που αντικαταστάθηκαν από κάποια άλλα.

Επίσης παρέχει πρόσβαση μέσω εξυπηρετητή με REST API σ(Application Programming Interface), το επιστρέφει πληροφορία με βάση κατάλληλα HTTP Requests.

2.2.3.1 Custom Downloads

Παράλληλα, με το αρχείο που περιέχει το σύνολο του γονιδιώματος και όλα τα πεδία της βάσης δεδομένων της HGNC, παρέχεται ένα σημαντικό front-end εργαλείο στο χρήστη, που το επιτρέπει να παράγει μόνο τη χρήσιμη πληροφορία που χρειάζεται. Με αυτόν τον τρόπο ο χρήστης μπορεί να επιλέξει τα πεδία της βάσης που επιθυμεί, ώστε να παραχθεί γρηγορότερα αρχείο μικρότερου μεγέθους. Επιπρόσθετα, γίνεται σύνδεση με άλλες βάσεις γονιδιακών δεδομένων, ώστε να παρέχονται τα αναγνωριστικά άλλων βάσεων για το ίδιο γονίδιο, όπως η Ensembl, η PubMed, η Entrez και άλλες.

Η παρακάτω εικόνα παρουσιάζει αυτό το εργαλείο:



Εικόνα 2. Εργαλείο παραγωγής αρχείων γονιδιακών δεδομένων HGNC

Τα πεδία που παρέχονται περιγράφονται παρακάτω. Μαζί με κάθε πεδίο, δίνεται και ο τύπος δεδομένων του πεδίου αυτού στη βάση δεδομένων, π.χ. varchar(100). Επιπλέον, ο συμβολισμός CD (Comma Delimited) δίπλα σε κάθε πεδίο, δηλώνει ότι μπορεί το πεδίο αυτό να έχει πολλαπλές τιμές χωρισμένες με κόμμα, ενώ ο συμβολισμός QCD δηλώνει πολλαπλές τιμές, περικλειόμενες σε εισαγωγικά (“”) και διαχωρισμένες με κόμμα.

Τα διαθέσιμα πεδία είναι τα εξής:

- **HGNC ID** (int) – Μοναδικό αναγνωριστικό που παρέχεται από την HGNC.
- **Approved Symbol** (varchar(255)) – Το επίσημο σύμβολο που έχει εγκριθεί και δημοσιευθεί από την HGNC. Τα σύμβολα είναι εγκεκριμένα με βάση τους κανόνες ονοματολογίας της HGNC.
- **Approved Name** (text) – Το επίσημο όνομα του γονιδίου που έχει εγκριθεί και δημοσιευθεί από την HGNC. Τα σύμβολα είναι επίσης εγκεκριμένα με βάση τους κανόνες ονοματολογίας της HGNC.

- **Status** (varchar(50)) – Δείχνει εάν το γονίδιο χαρακτηρίζεται ως:
    - *Approved* – Τα γονίδια αυτά έχουν HGNC-εγκεκριμένα γονιδιακά σύμβολα
    - *Entry withdrawn* – αυτά τα προηγουμένως εγκεκριμένα γονίδια, θεωρείται ότι δεν υφίστανται πλέον
    - *Symbol withdrawn* – ένα προηγουμένως εγκεκριμένο σύμβολο που έχει πλέον συγχωνευθεί με μια άλλη εγγραφή
- **Locus Type** (varchar(100)) – Καθορίζει τον τύπο του γονιδιακού τόπου που περιγράφεται από τις παρακάτω ετικέτες:
    1. **gene with protein product** – γονίδια που κωδικοποιούν πρωτεΐνη (η πρωτεΐνη μπορεί να έχει ανακαλυφθεί αλλά να έχει άγνωστη λειτουργία)
    2. **RNA, cluster** – περιοχή που περιέχει ένα σύμπλεγμα από γονίδια που δεν κωδικοποιούν RNA
    3. **RNA, long non-coding** – Γονίδια που δεν κωδικοποιούν πρωτεΐνη αλλά που κωδικοποιούν μακριά non-coding RNA (lncRNAs). Αυτά έχουν μήκος τουλάχιστον 200 νουκλεοτίδια. Οι υποτύποι αυτού περιλαμβάνουν τα these, και antisense.
    4. **RNA, micro** – μη πρωτεϊνικά γονίδια που κωδικοποιούν microRNA (miRNAs)
    5. **RNA, ribosomal** - μη πρωτεϊνικά γονίδια που κωδικοποιούν ριβοσωμικά RNA (rRNAs)
    6. **RNA, small nuclear** - μη πρωτεϊνικά γονίδια που κωδικοποιούν μικρά πυρηνικά RNA (snRNAs)
    7. **RNA, small nucleolar** - non μη πρωτεϊνικά γονίδια που κωδικοποιούν μικρά πυρηνισκικό RNA (snoRNAs)
    8. **RNA, small cytoplasmic** - μη πρωτεϊνικά γονίδια που κωδικοποιούν μικρά κυτταροπλασμικά RNA (scRNAs)
    9. **RNA, transfer** μη πρωτεϊνικά γονίδια που κωδικοποιούν μεταφορικά(transfer) RNA (tRNAs)
    10. **RNA, small misc** - μη πρωτεϊνικά γονίδια που κωδικοποιούν διάφορους τύπους από μικρά ncRNAs, όπως τα vault και τα Y γονίδια
    11. **phenotype only** - χαρτογραφημένοι φαινότυποι
    12. **pseudogene** – γονιδιακές DNA ακολουθίες που είναι παρόμοιες με τα πρωτεϊνικά γονίδια, αλλά δεν κωδικοποιούν πρωτεΐνες

13. **RNA, pseudogene** – ψευδογονίδιο που παράγει μη πρωτεϊνικά RNA
14. **complex locus constituent** – μεταγραφική μονάδα που είναι τμήμα ενός ονοματισμένου πολύπλοκου τόπου
15. **endogenous retrovirus** – ενσωματωμένα ρετροϊικά στοιχεία που μεταδίδονται μέσω του μικροβίων
16. **fragile site** – κληρονομήσιμος τόπος σε ένα χρωμόσωμα που είναι επιρρεπής σε «σπάσιμο» του DNA
17. **immunoglobulin gene** – τμήματα γονιδίων που υφίστανται *σωματικό ανασυνδυασμό* για να σχηματίσουν ανοσοσφαιρίνη ελαφριάς ή βαριάς αλυσίδας. Επιπρόσθετα περιέχει ανοσοφαιρινικά τμήματα γονιδίων με ανοιχτά πλαίσια ανάγνωσης που είτε υφίστανται σωματικό ανασυνδυασμό είτε κωδικοποιούν ένα πεπτίδιο που δεν είναι σίγουρο ότι θα αναδιπλωθεί σωστά.
18. **immunoglobulin pseudogene** – τμήματα ανοσοσφαιρινικών γονιδίων που έχουν απενεργοποιηθεί λόγω μεταλλάξεων μετακίνησης πλαισίου και/η κωδικωνίων λήξης σε ανοιχτό πλαίσιο ανάγνωσης
19. **protocadherin** – τμήματα γονιδίου που συνιστούν τι τρεις πρωτοκαδερίνες(άλφα, βήτα και γάμμα)
20. **readthrough** – μια μεταγραφή που προκύπτει φυσικά και που περιέχει ακολουθία που κωδικοποιείται από δύο ή περισσότερα γονίδια που μπορούν να μεταγραφούν αυτόνομα
21. **region** – μήκη γονιδιακής ακολουθίας που περιέχουν ένα ή περισσότερα γονίδια καθώς και μη γονιδιακές περιοχές που δεν εμπίπτουν σε κάποια άλλη κατηγορία
22. **T cell receptor gene** – τμήματα γονιδίων που υφίστανται σωματικό ανασυνδυασμό για να δημιουργήσουν άλφα, βήτα, γάμμα ή δελτα αλυσίδες από γονίδια υποδοχέων Τ-κυττάρων. Επίσης περιλαμβάνει τμήματα γονιδίων με ανοιχτά πλαίσια ανάγνωσης που είτε δεν μπορούν να συμμετέχουν σε σωματική αναδιάταξη, είτε δεν μπορούν να κωδικοποιήσουν ένα πεπτίδιο για το οποίο μπορούμε να προβλέψουμε αν θα διπλωθεί σωστά. Αυτά αναγνωρίζονται από την εισαγωγή του όρου “non-functional” στο όνομα του γονιδίου.
23. **T cell receptor pseudogene** – τμήματα γονιδίων υποδοχέων Τ κυττάρων που είναι απενεργοποιημένα λόγω μεταλλάξεων μετατόπισης του πλαισίου ή λόγω κωδικονίων τερματισμού σε ένα ανοιχτό πλαίσιο ανάγνωσης

24. **transposable element** – τμήμα επαναλαμβανόμενου DNA που μπορεί να μετακινηθεί ή να ρετρο-μεταφραστεί σε νέα τμήματα στο γονιδίωμα
25. **unknown** – εγγραφές για τις οποίες η τοποθεσία τους είναι αυτή τη στιγμή άγνωστη
26. **virus integration site** – ακολουθία στόχος για την ενσωμάτωση ιικού DNA στο γονιδίωμα

- **Locus Group** (varchar(100)) – Ομαδοποιεί τους τύπους γονιδιακού τόπου σε σχετικά σύνολα. Παρακάτω φαίνεται μια λίστα με τους τόπους σε κάθε ομάδα:
  - **protein-coding gene** – περιέχει τον τύπο τόπου «γονιδίου με παράγωγη πρωτεΐνη»
  - **non-coding RNA** – περιέχει τους παρακάτω τύπους:
    - RNA, cluster
    - RNA, long non-coding
    - RNA, micro
    - RNA, ribosomal
    - RNA, small cytoplasmic
    - RNA, small misc
    - RNA, small nuclear
    - RNA, small nucleolar
    - RNA, transfer
  - **pseudogene** – περιέχει τους παρακάτω τύπους:
    - immunoglobulin pseudogene
    - pseudogene
    - RNA, pseudogene
    - T cell receptor pseudogene
  - **phenotype** – περιέχει τον τύπο τόπου «μόνο φαινότυπος»
  - **other** – περιέχει τους παρακάτω τύπους:
    - endogenous retrovirus
    - fragile site
    - immunoglobulin gene
    - protocadherin
    - readthrough
    - region
    - T cell receptor gene
    - transposable element
    - unknown

    - virus integration site
  - **withdrawn** – περιέχει μόνο τον τύπο τόπου «έχει αποσυρθεί»
- **Previous Symbols** (text) CD – Σύμβολα προηγουμένως εγκεκριμένα από την HGNC για αυτό το γονίδιο
- **Previous Names** (text) QCD - Ονόματα προηγουμένως εγκεκριμένα από την HGNC για αυτό το γονίδιο
- **Synonyms** (text) CD – άλλα συνώνυμα σύμβολα που χρησιμοποιούνται για αναφορά σε αυτό το γονίδιο
- **Name Synonyms** (text) QCD – άλλα ονόματα που χρησιμοποιούνται για αναφορά σε αυτό το γονίδιο
- **Chromosome** (varchar(255)) – Δείχνει την τοποθεσία του γονιδίου ή της περιοχής στο χρωμόσωμα
- **Date Approved** (date) – Ημερομηνία που το σύμβολο και το όνομα εγκρίθηκαν από την HGNC
- **Date Modified** (date) – Αν υπάρχει, τότε είναι η ημερομηνία που μεταβλήθηκε από την HGNC
- **Date Symbol Changed** (date) – Εάν υπάρχει, τότε δείχνει την ημερομηνία που το σύμβολο του γονιδίου τροποποιήθηκε από την HGNC τελευταία φορά. Πολλά γονίδια λαμβάνουν εγκεκριμένα σύμβολα και ονόματα που θεωρούνται προσωρινά ή μη-βέλτιστα μόλις νέα πληροφορία γίνει διαθέσιμη αργότερα. Στην περίπτωση μεμονωμένων γονιδίων, μια αλλαγή στο όνομα (και συνεπώς στο σύμβολο) γίνεται μόνο εάν το αρχικό όνομα είναι αρκετά παραπλανητικό.
- **Date Name Changed** (date) – Εάν υπάρχει, είναι η ημερομηνία που ένα προηγουμένως εγκεκριμένο όνομα τροποποιήθηκε από την HGNC
- **Accession Numbers** (text) CD - Accession numbers για κάθε εγγραφή, που έχουν επιλεγεί από την HGNC
- **Enzyme ID** (text) CD – Οι εγγραφές ενζύμων έχουν αριθμούς Enzyme Commission (EC) συσχετισμένους με αυτές, οι οποίοι δείχνουν τις ιεραρχικά λειτουργικές κλάσεις στις οποίες ανήκουν.
- **Entrez Gene ID** (int) - η Entrez Gene στη βάση NCBI παρέχει επιμελημένες ακολουθίες και περιγραφικές πληροφορίες για γενετικούς τόπους, συμπεριλαμβανομένης της επίσημης ονοματολογίας. Επίσης περιέχει συνώνυμα, accessions numbers κάποιας ακολουθίας, φαινότυπους, αριθμούς EC, αριθμούς MIM, συμπλέγματα UniGene, ομολογία, τοποθεσίες στο χάρτη και σχετικές ιστοσελίδες. Η Entrez Gene έχει αντικαταστήσει την LocusLink.

- **CCDS ID** (text) - Το Consensus CDS (CCDS) project είναι μία συνεργατική προσπάθεια για να αναγνωριστεί ένα βασικό σύνολο από περιοχές που κωδικοποιούν πρωτεΐνες στους ανθρώπους και στα ποντίκια, οι οποίες είναι επισημειωμένες με συνεπή τρόπο και είναι υψηλής ποιότητας. Ο μακροπρόθεσμος στόχος είναι να υποστηριχθεί η σύγκλιση σε ένα πρότυπο σύνολο γονιδιακών επισημειώσεων.
- **VEGA ID** (text) – Περιέχει ένα επιμελημένο VEGA gene ID
- **Locus Specific Databases** (text) – Περιέχει μια λίστα από συνδέσμους σε βάσεις δεδομένων ή εγγραφές σχετικές με το γονίδιο.
- **Mouse Genome Database ID** (varchar(50)) – Αναγνωριστικό της βάσης δεδομένων του γονιδιώματος ποντικού
- **Specialist Database Links** (text) CD – Αυτή η στήλη περιέχει συνδέσμους σε εξειδικευμένες βάσεις δεδομένων που έχουν ενδιαφέρον όσον αφορά το συγκεκριμένο σύμβολο ή γονίδιο.
- **Ensembl Gene ID** (varchar(50)) – Η στήλη αυτή περιέχει ένα χειροκίνητα επιμελημένο Ensembl Gene ID
- **Specialist Database IDs** (text) CD – Η στήλη αυτή περιέχει HTML συνδέσμους στις Βάσεις Δεδομένων προς αναφορά. Περιέχει μόνο το αναγνωριστικό της βάσης. Οι βάσεις που διατίθενται είναι οι:
    1. miRBase the microRNA database
    2. HORDE ID Human Olfactory Receptor Data Exploratorium
    3. CD Human Cell Differentiation Antigens
    4. Rfam RNA families database of alignments and CMs
    5. snoRNABase database of human H/ACA and C/D box snoRNAs
    6. KZNF Gene Catalog Human KZNF Gene Catalog
    7. Intermediate Filament DB Human Intermediate Filament Database
    8. IUPHAR Committee on Receptor Nomenclature and Drug Classification.(mapped)
    9. IMGT/GENE-DB the international ImMunoGeneTics information system for immunoglobulins (mapped)
    10. MEROPS the peptidase database
    11. COSMIC Catalogue Of Somatic Mutations In Cancer
    12. Orphanet portal for rare diseases and orphan drugs
    13. Pseudogene.org database of identified pseudogenes

14. piRNABank database of piwi-interacting RNA clusters
15. HomeoDB a database of homeobox gene diversity
16. Mamit-tRNAdb a compilation of mammalian mitochondrial tRNA genes
17. lncRNAdb a database providing comprehensive annotations of eukaryotic long non-coding RNAs (lncRNAs).
18. BioParadigms SLC tables Provide information on the SLC families and their members, with relevant links to gene databases and reviews in literature.

Τα περισσότερα από αυτά τα αναγνωριστικά έχουν περάσει από χειροκίνητη επιμέλεια, παρόλα αυτά, κάποια από αυτά έχουν ληφθεί από αρχεία που ανανεώνονται ανά τακτά διαστήματα και παρέχονται από τη βάση που ενδιαφέρει.

- **Pubmed IDs** (text) CD – Αναγνωριστικό που συνδέει δημοσιευμένα άρθρα στην σε εγγραφές της Βάσης Δεδομένων PubMed.
- **RefSeq IDs** (varchar(50)) CD – Το αναγνωριστικό RefSeq (Reference Sequence) (RefSeq) για αυτή την εγγραφή, που παρέχεται από την NCBI. Εφόσον δεν θεωρείται σκόπιμο να περάσουν από επιμέλεια όλες οι παραλλαγές ενός γονιδίου, εμφανίζεται μόνο ένα RefSeq ανά γονίδιο. Το RefSeq στοχεύει στο να παρέχει ένα περιεκτικό, ενσωματωμένο σύνολο από ακολουθίες, που περιλαμβάνουν γονιδιακό DNA, μεταφράσιμο γενετικό προϊόν (RNA) και πρωτεϊνικά προϊόντα. Τα αναγνωριστικά RefSeq είναι σχεδιασμένα να προφέρουν μία σταθερή αναφορά για την αναγνώριση και τον χαρακτηρισμό γονιδίων, την ανάλυση μεταλλάξεων, τις μελέτες έκφρασης, την ανακάλυψη πολυμορφισμών και τις συγκριτικές αναλύσεις.
- **Gene Family Tag** (text) CD – Ετικέτα που χρησιμοποιείται για να δηλώσει μία οικογένεια γονιδίων ή ομάδα στην οποία έχει προστεθεί το γονίδιο, σύμφωνα είτε με ομοιότητα ακολουθίας είτε με πληροφορίες από δημοσιεύσεις, επιστήμονες που ασχολούνται με αυτήν την οικογένεια ή από άλλες βάσεις δεδομένων. Οι οικογένειες/ομάδες μπορεί να είναι είτε δομικές, είτε λειτουργικές και συνεπώς ένα γονίδιο μπορεί να ανήκει σε περισσότερες από μία ομάδες ή οικογένειες. Οι ετικέτες αυτές χρησιμοποιούνται για να παραχθούν συγκεκριμένες σελίδες για οικογένειες ή ομάδες στο genenames.org και δεν αντικατοπτρίζουν απαραίτητα την επίσημη ονοματολογία. Κάθε οικογένεια γονιδίων έχει μια συσχετισμένη ετικέτα και περιγραφή. Εάν ένα συγκεκριμένο γονίδιο είναι μέλος σε παραπάνω από μία οικογένειες, τότε οι ετικέτες και οι περιγραφές θα εμφανίζονται με την ίδια σειρά.
- **Gene Family Description** (text) CD – Το όνομα που δίνεται σε μια συγκεκριμένη οικογένεια γονιδίων. Η περιγραφή αυτή έχει και μια αντίστοιχη ετικέτα.

- **Mapped Field Definitions** Τα δεδομένα αυτά, λαμβάνονται από εξωτερικές πηγές και ως τέτοια, δεν έχουν υποστεί αυστηρή επιμέλεια. Πρέπει να αντιμετωπίζονται με κάποια προσοχή.
    - **Mouse Genome Database ID** (mapped data) (varchar(50))
    - **Rat Genome Database ID** (mapped data) (varchar(50)) – Αναγνωριστικό της βάσης γονιδιώματος αρουραίου
    - **Entrez Gene ID** (mapped data) (int)
    - **OMIM ID** (mapped data) (varchar(50)) – Αναγνωριστικό που παρέχεται από την Online Mendelian Inheritance in Man (OMIM) στην NCBI. Η βάση δεδομένων αυτή περιγράφεται ως ένας κατάλογος ανθρωπίνων γονιδίων και γενετικών ανωμαλιών που περιέχουν κειμενική πληροφορία και συνδέσμους στη MEDLINE, εγγραφές ακολουθιών στο σύστημα Entrez και συνδέσμους σε σχετικούς πόρους στην NCBI και αλλού.
    - **RefSeq** (mapped data) (varchar(50))
    - **UniProt ID** (mapped data) (varchar(50)) – Το αναγνωριστικό UniProt παρέχεται από το European Bioinformatics Institute. Η UniProt (βάση δεδομένων για πρωτεΐνες) περιγράφεται ως μια επιμελημένη βάση δεδομένων πρωτεϊνικής ακολουθίας που παρέχει ένα υψηλό επίπεδο επισημειώσεων, ελάχιστη περιττή πληροφορία καθώς και υψηλό επίπεδο ενσωμάτωσης με άλλες βάσεις δεδομένων.
    - **Ensembl Gene ID** (mapped data) (varchar(50)) – Το Ensembl ID λαμβάνεται από την τελευταία έκδοση της βάσης και παρέχεται από την ομάδα της Ensembl.
    - **UCSC** (mapped data) (varchar(50)) - Το UCSC ID λαμβάνεται από την τελευταία έκδοση της UCSC database(http://genome.ucsc.edu/).

2.2.3.2 Δομή των αρχείων της HGNC

Τα αρχεία της βάσης HGNC είναι αρχεία κειμένου με συγκεκριμένη δομή. Όπως και στην NCIB τα αρχεία είναι tab delimited, δηλαδή τα πεδία είναι χωρισμένα με χαρακτήρες '\t'. Κάθε εγγραφή, χωρίζεται από την επόμενή της με ένα χαρακτήρα αλλαγής γραμμής('\n'). Όπου υπάρχει κενή πληροφορία, δεν υπάρχει ο χαρακτήρας '-' όπως στην NCIB αλλά αντίθετα ακολουθεί ο επόμενος χαρακτήρας tab.

Ένα δείγμα του αρχείου με τα πλήρη περιεχόμενα της βάσης φαίνεται παρακάτω:

```
HGNC ID       Approved Symbol      Approved Name Status Locus Type    Locus
Group  Previous Symbols     Previous Names       Synonyms      Name Synonyms
       Chromosome    Date Approved Date Modified Date Symbol Changed Date Name
Changed       Accession Numbers    Enzyme IDs    Entrez Gene ID      Ensembl
Gene ID       Mouse Genome Database ID   Specialist Database Links  Specialist
Database IDs  Pubmed IDs    RefSeq IDs    Gene Family Tag      Gene family
description   Record Type   Primary IDs   Secondary IDs CCDS IDs      VEGA IDs
       Locus Specific Databases   Entrez Gene ID (supplied by NCBI) OMIM ID
(supplied by NCBI)  RefSeq (supplied by NCBI)  UniProt ID (supplied by
UniProt)      Ensembl ID (supplied by Ensembl)  UCSC ID (supplied by UCSC)
       Mouse Genome Database ID (supplied by MGI)     Rat Genome Database ID
(supplied by RGD)
HGNC:5 A1BG   alpha-1-B glycoprotein     Approved      gene with protein
product       protein-coding gene                          19q13.43
       1989-06-30    2013-10-15                            1
       ENSG00000121410            <!--,--> <!--,--> <!--,--> <!--,--> <!--,--
> <!--,--> <!--,--> <!--,--> <!--,--> <a
href="http://merops.sanger.ac.uk/cgi-
bin/merops.cgi?id=I43.950">MEROPS</a><!--,--> <a
href="http://www.sanger.ac.uk/perl/genetics/CGP/cosmic?action=gene&amp;ln=A1
BG">COSMIC</a><!--,--> <!--,--> <!--,--> <!--,--> <!--,--> <!--,--> <!--,-->
       , , , , , , , , , I43.950, A1BG, , , , , , ,   2591067       NM_130786
       IGD    "Immunoglobulin superfamily / Immunoglobulin-like domain
containing"   Standard                   CCDS12976.1   OTTHUMG00000183507
       1      138670 XM_005258393  P04217 ENSG00000121410     uc002qsd.4
       MGI:2152878   RGD:69417
```

Μπορούμε να παρατηρήσουμε ότι σε κάποια από τα πεδία, δίνονται σύνδεσμοι σε μορφή HTML. Πιο συγκεκριμένα, τα πεδία που δίνονται φαίνονται στον παρακάτω πίνακα:

HGNC ID	Approved Symbol	Approved Name	Status	Locus Type
HGNC:5	A1BG	alpha-1-B glycoprotein	Approved	gene with protein product

Locus Group	Previous Symbols	Previous Names	Synonyms	Name Synonyms
protein-coding gene				

Chromosome	Date Approved	Date Modified	Date Symbol Changed	Date Name Changed
19q13.43	1989-06-30	2013-10-15		

Accession Numbers	Enzyme IDs	Entrez Gene ID	Ensembl Gene ID	Mouse Genome Database ID
		1	ENSG00000121410	

Specialist Database Links
http://merops.sanger.ac.uk/cgi-bin/merops.cgi?id=I43.950 **http://www.sanger.ac.uk/perl/genetics/CGP/cosmic?action=gene&ln=A1BG**

Specialist Database IDs
, , , , , , , , I43.950, A1BG, , , , , , ,

Pubmed IDs	RefSeq IDs	Gene Family Tag	Gene family description	Record Type
2591067	NM_130786	IGD	"Immunoglobulin superfamily / Immunoglobulin-like domain containing"	

Primary IDs	Secondary IDs	CCDS IDs	VEGA IDs	Locus Specific Databases
		CCDS12976.1	OTTHUMG00000183507	

Entrez Gene ID (supplied by NCBI)	OMIM ID (supplied by NCBI)	RefSeq (supplied by NCBI)	UniProt ID (supplied by UniProt)	Ensembl ID (supplied by Ensembl)
1	138670	XM_005258393	P04217	ENSG00000121410

UCSC ID (supplied by UCSC)	Genome Database ID (supplied by MGI)	Rat Genome Database ID (supplied by RGD)
uc002qsd.4	MGI:2152878	RGD:69417

2.2.3.3 REST Service

Η διαδικτυακή υπηρεσία REST του genenames.org είναι ένας βολικός και γρήγορος τρόπος για αναζήτηση και λήψη πληροφορίας από την βάση δεδομένων με ένα script/πρόγραμμα. Οι χρήστες μπορούν να κάνουν αίτηση για αποτελέσματα είτε σε μορφή XML είτε σε μορφή JSON, καθιστώντας τα δεδομένα εύκολα για επεξεργασία. Επίσης παρέχει αναλυτικά προγράμματα για αιτήσεις, σε μια πληθώρα γλωσσών προγραμματισμού, προκειμένου να διευκολύνει τους χρήστες. Θα γίνει εκτενέστερη αναφορά σε επόμενο κεφάλαιο στο REST service, καθώς ήταν αποτέλεσε βασικό εργαλείο για τη λήψη πληροφορίας των διαφόρων γονιδίων από την HGNC. [12]

2.2.4 Ensembl

Η βάση δεδομένων Ensembl είναι η ευρωπαϊκή απάντηση στην NCBI. Το project Ensembl ξεκίνησε το 1999, λίγα χρόνια πριν ολοκληρωθεί η χαρτογράφηση του ανθρωπίνου γονιδιώματος. Ακόμα και σε αυτό το πρώιμο στάδιο, ήταν εμφανές ότι η χειροκίνητη επισημείωση ακολουθίας 3 δισεκατομμυρίων ζευγών βάσεων δεν θα μπορούσε να προσφέρει στους ερευνητές έγκαιρη πρόσβαση στα τελευταία δεδομένα. Ο στόχος της Ensembl ήταν λοιπόν, να επισημειώσει *αυτόματα* το γονιδίωμα, να ενσωματώσει αυτή τη επισημείωση μαζί με άλλα διαθέσιμα βιολογικά δεδομένα και να τα δημοσιεύσει στο Διαδίκτυο. Από την έναρξη λειτουργίας της ιστοσελίδας τον Ιούλιο του 2000 έχουν προστεθεί στην Ensembl πολλά άλλα γονιδιώματα και το εύρος της διαθέσιμης πληροφορίας έχει επεκταθεί ώστε να συμπεριλάβει και συγκριτική γονιδιακή πληροφορία, παραλλαγές καθώς και ρυθμιστικά δεδομένα.

Ο αριθμός των ατόμων που ασχολούνται με το project έχει επίσης αυξηθεί. Αυτή τη στιγμή, η ομάδα της Ensembl αποτελείται από 40 με 50 άτομα, χωρισμένα διάφορες υποομάδες. Η ομάδα Genebuild δημιουργεί τα σύνολα γονιδίων για τα διάφορα είδη. Το αποτέλεσμα της δουλειάς τους αποθηκεύεται σε κεντρικές βάσεις δεδομένων, τις οποίες συντηρεί η ομάδα Λογισμικού (IT). Αυτή η ομάδα επίσης είναι υπεύθυνη για την ανάπτυξη και την συντήρηση του εργαλείου εξόρυξης δεδομένων BioMart. Οι ομάδες Compara, Variation και Regulation είναι υπεύθυνες αντίστοιχα για την σύγκριση, την διαφοροποίηση και τη ρύθμιση στα δεδομένα.

Η ομάδα Web εξασφαλίζει ότι όλα τα δεδομένα παρουσιάζονται στην ιστοσελίδα με ξεκάθαρο και φιλικό προς τον χρήστη τρόπο. Τέλος η ομάδα Outreach απαντά σε ερωτήσεις από χρήστες και διοργανώνει συνέδρια ανά τον κόσμο για τη χρήση της Ensembl.
Η Ensembl είναι ένα κοινό Project ανάμεσα στο *Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής* (European Bioinformatics Institute – EBI), ένα εξωτερικό τμήμα του *Ευρωπαϊκού Εργαστηρίου Μοριακής Βιολογίας* (European Molecular Biology Laboratory – EMBL) και του *Wellcome Trust Sanger Institute* (WTSI).

2.2.4.1 Ensembl και GRC

Η Ensembl παράγει μια νέα έκδοση της ιστοσελίδας και των υποκείμενων βάσεων δεδομένων κάθε 2-3 μήνες. Αυτό επιτρέπει τη δημιουργία νέων δεδομένων και αναλύσεων που γίνονται διαθέσιμες ύστερα από ενδελεχείς ποιοτικούς ελέγχους. Κάθε νέα έκδοση μπορεί να περιλαμβάνει νέα ή/και ανανεωμένα δεδομένα, όπως καινούρια είδη, νέες γονιδιακές συναρμολογήσεις, ανανεωμένα σύνολα γονιδίων, νέα δεδομένα ποικιλομορφίας, κατασκευή δέντρων γονιδίων, στοιχίσεων καθώς και υποσημείωση των ρυθμιστικών χαρακτηριστικών. Υπάρχουν επίσης προσθήκες και βελτιώσεις στο web interface και στη δομή της βάσης δεδομένων.

Επιπλέον χρησιμοποιεί την πιο πρόσφατα ανανεωμένη έκδοση του ανθρωπίνου γονιδιώματος που βρίσκεται στους εξυπηρετητές του GRC. Η NCBI και ο genome browser του UCSC (University of California, Santa Cruz) χρησιμοποιούν το ίδιο γονιδίωμα (το UCSC αναφέρεται στο πρόσφατο γονιδίωμα του ανθρώπου ως GRCh37/hg19).

Η GRC παράγει ελάσσονες εκδόσεις της GRCh37, ανά τακτά χρονικά διαστήματα, περίπου κάθε τρεις μήνες. Αυτές οι ενημερώσεις, που είναι γνωστές ως *«μπαλώματα»* (patches) στην έκδοση GRCh37, ενσωματώνονται στην Ensembl.

Εξ ορισμού, οι συντεταγμένες των χρωμοσωμάτων δεν αλλάζουν όταν ανανεώνεται το ανθρώπινο γονιδίωμα στην τελευταία ελάσσονα έκδοση.

Ένας πίνακας με τις εκδόσεις της Ensembl και τις αντίστοιχες εκδόσεις της βάσης της GRC καθώς και βάσεις για τα άλλα είδη, φαίνονται παρακάτω:

<table>
<tr><th rowspan="2">Ημερομηνία</th><th rowspan="2">Έκδοση Ensembl</th><th colspan="2">Έκδοση GRC</th><th colspan="2">Λοιπά Είδη</th></tr>
<tr><th>Homo Sapiens</th><th>Mus Musculus</th><th>Drosophila Melanogaster</th><th>Caenorhabditis Elegans</th></tr>
<tr><td>Μάιος 2009</td><td>54</td><td>NCBI 36</td><td rowspan="8">NCBI m37</td><td rowspan="2">BDGP 5.4</td><td>WS190</td></tr>
<tr><td>Ιούλιος 2009</td><td>55</td><td rowspan="7">GRCh37</td><td rowspan="3">WS200</td></tr>
<tr><td>Σεπτέμβριος 2009</td><td>56</td><td rowspan="3">BDGP 5.13</td></tr>
<tr><td>Μάρτιος 2010</td><td>57</td></tr>
<tr><td>Μάιος 2010</td><td>58</td><td rowspan="4">WS210</td></tr>
<tr><td>Αύγουστος 2010</td><td>59</td><td rowspan="3">BDGP 5.25</td></tr>
<tr><td>Νοέμβριος 2010</td><td>60</td></tr>
</table>

<table>
<tr><td>Φεβρουάριος 2011</td><td>61</td><td rowspan="2"></td><td rowspan="2"></td><td rowspan="4"></td><td rowspan="6">WS220</td></tr>
<tr><td>Απρίλιος 2011</td><td>62</td></tr>
<tr><td>Ιούνιος 2011</td><td>63</td><td rowspan="2">GRCh37.p3</td><td rowspan="6">NCBIM37</td></tr>
<tr><td>Σεπτέμβριος 2011</td><td>64</td></tr>
<tr><td>Δεκέμβριος 2011</td><td>65</td><td>GRCh37.p5</td><td rowspan="12">BDGP 5</td></tr>
<tr><td>Φεβρουάριος 2012</td><td>66</td><td>GRCh37.p6</td></tr>
<tr><td>Μάιος 2012</td><td>67</td><td>GRCh37.p7</td><td rowspan="4">WBcel215</td></tr>
<tr><td>Ιούλιος 2012</td><td>68</td><td rowspan="3">GRCh37.p8</td><td rowspan="2">GRCm38</td></tr>
<tr><td>Οκτώβριος 2012</td><td>69</td></tr>
<tr><td>Ιανουάριος 2013</td><td>70</td><td rowspan="4">GRCm38.p1</td></tr>
<tr><td>Απρίλιος 2013</td><td>71</td><td>GRCh37.p10</td><td rowspan="6">WBcel235</td></tr>
<tr><td>Ιούνιος 2013</td><td>72</td><td>GRCh37.p11</td></tr>
<tr><td>Σεπτέμβριος 2013</td><td>73</td><td>GRCh37.p12</td></tr>
<tr><td>Δεκέμβριος 2013</td><td>74</td><td rowspan="2">GRCh37.p13</td><td rowspan="2">GRCm38.p2</td></tr>
<tr><td>Φεβρουάριος 2014</td><td>75</td></tr>
</table>

2.2.4.2 Μπαλώματα(Patches)

Η GRC εκδίδει ανά διαστήματα *μπαλώματα* (patches) στην πρωτεύουσα γονιδιακή ακολουθία, είτε για να διορθώσει σφάλματα στη ακολουθία, είτε για να δώσει εναλλακτικές ακολουθίες για το ίδιο γονίδιο.

Τα μπαλώματα δίνονται σε μορφή αρχείων FASTA και περιέχουν πληροφορίες για το γονίδιο, καθώς και τη νέα ακολουθία του.

Η Ensembl παίρνει αυτά τα αρχεία, τα επεξεργάζεται και δημιουργεί νέα αρχεία FASTA με χρωμοσωμικό DNA.

Ο αριθμός των αρχείων των μπαλωμάτων δεν είναι σταθερός. Μπορεί να μην υπάρχει κανένα και μπορεί να υπάρχουν πολλαπλά. Αυτό σχετίζεται με το πόσες διορθώσεις ή εναλλακτικές ακολουθίες εκδίδει η GRC σε κάθε minor (patch) έκδοση.
Με κάθε patch αλλάζουν και τα αντίστοιχα transcripts.

2.2.4.3 Εγγενές αναγνωριστικό

Τα εγγενή αναγνωριστικά της Ensembl είναι 4:

1. ENSG : Ensembl Gene ID
2. ENST : Ensembl Transcript ID
3. ENSP : Ensembl Peptide ID
4. ENSE : Ensembl Exon ID

Τα αναγνωριστικά αυτά είναι σταθερά, οπότε και αλλαγή να υπάρξει στην πληροφορία, τα αναγνωριστικά θα παραμείνουν ως έχουν. Ακόμη, τα περισσότερα αναγνωριστικά έχουν μήκος 12 ψηφία.
Επιπρόσθετα, για οργανισμούς εκτός του ανθρώπου, εισάγεται ένα επιπλέον επίθεμα:
πχ για τον βάτραχο C.Savignyi έχουμε ENSCSAVG ID, ενώ για τον Danio Rerio έχουμε ENSDARG ID.

2.2.4.4 Ανάγνωση γονιδιώματος

Αρχικά, τα χρωμοσώματα, σπάνε σε κομμάτια ακολουθίας τα οποία πρέπει να αναγνωριστούν. Όπου αυτά τα μικρότερα κομμάτια μπορούν να ενωθούν, τότε ενώνονται σε κομμάτια συνεχούς ακολουθίας (contiguous sequence – contigs). Τα contigs με τη σειρά τους ενώνονται για να δημιουργήσουν «σκαλωσιές» (scaffoldings). Τέλος οι σκαλωσιές ενώνονται για να δημιουργήσουν τα χρωμοσώματα, αφήνοντας κενά όπου αυτό είναι αναγκαίο.

2.2.4.5 Μετάγραφα (Transcripts)

Είναι κομμάτια RNA που παράγονται από γονίδια με τη διαδικασία της μεταγραφής. Χρησιμεύουν όχι μόνο στην παραγωγή πρωτεϊνών αλλά και σε άλλες λειτουργίες, όπως αναφέρεται σε προηγούμενο κεφάλαιο (2.1.4).
Στην Ensembl δίνονται με τη μορφή cDNA(coding DNA) και ncDNA(non-coding DNA), που είναι το συμπληρωματικό DNA του RNA. Οι ακολουθίες δίνονται στη μορφή FASTA για κάθε μετάγραφο και περιέχουν το αναγνωριστικό του μετάγραφου της Ensembl καθώς και τον κωδικό του γονιδίου από το οποίο παράγεται.

2.2.4.6 FTP Server – FASTA FILES

Η Ensembl προσφέρει έναν FTP server με τεράστιο όγκο δεδομένων. Οι διαθέσιμες εκδόσεις ξεκινούν από την 19 και φτάνουν μέχρι την πιο πρόσφατη που είναι η 75. Προσφέρει μια μεγάλη γκάμα τύπων αρχείων όπως FASTA και EMBL.

Οι ακολουθίες δίνονται συνήθως σε αρχεία FASTA format (.fa).

Το FASTA format είναι μια μορφή καταγραφής βασισμένη σε κείμενο για απεικόνιση είτε ακολουθιών νουκλεοτιδίων είτε ακολουθιών πεπτιδίων, όπου τα νουκλεοτίδια ή τα πεπτίδια αναπαρίστανται με ένα μονό γράμμα.

Μια ακολουθία σε FASTA format ξεκινάει με μια περιγραφή μιας γραμμής, ακολουθούμενη από γραμμές δεδομένων ακολουθίας. Η γραμμή περιγραφής διαχωρίζεται από τα δεδομένα ακολουθίας με το σύμβολο “>” σαν πρώτο χαρακτήρα. Συνίστανται όλες οι γραμμές κειμένου να είναι μικρότερες από 80 χαρακτήρες σε μήκος.[9] Παρόλα αυτά, η επικεφαλίδα μπορεί να έχει μήκος μεγαλύτερο από 80 χαρακτήρες.

Ένα παράδειγμα ακολουθίας σε FASTA format από το αρχείο *Homo_sapiens.GRCh37.73.dna.chromosome.10.fa.gz* είναι το παρακάτω:

```
>10 dna:chromosome chromosome:GRCh37:10:1:135534747:1 REF
 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN…
…NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
 GAATTCCTTGAGGCCTAAATGCATCGGGGTGCTCTGGTTTTGTTGTTGTTATTTCTGAAT
 GACATTTACTTTGGTGCTCTTTATTTTGCGTATTTAAAACTATTAGATCGTGTGATTATA
 TTTGACAGGTCTTAATTGACGCGCTGTTCAGCCCTTTGAGTTCGGTTGAGTTTTGGGTTG
 GAGAATTTTCTTCCACAAGGGATTGTCTTGGATTTTTCTGTTTCTCCCTCAATATCCACC…
```

Τα γράμματα A,T,C,G, αντιστοιχούν στις βάσεις αδενίνη (A), θυμίνη (T), κυτοσίνη (C) και γουανίνη (G). Το γράμμα N δηλώνει άγνωστη βάση σε αυτήν τη θέση.

2.2.4.7 REST API

Η Ensembl, όπως και η HGNC προσφέρει REST Service για πιο γρήγορη αναζήτηση πληροφορίας μέσω προγραμμάτων. Διαθέτει μεγάλη ποικιλία από endpoints που αντιστοιχούν στην πληροφορία που αναζητείται.

Επίσης προσφέρει πρότυπα HTTP αίτησης σε διάφορες γλώσσες, έτσι ώστε οι χρήστες να μπορούν να εξατομικεύσουν το πρόγραμμα και να επεξεργαστούν την πληροφορία όπως αυτοί θεωρούν σκόπιμο. Θα γίνει εκτενέστερη αναφορά σε μεταγενέστερο κεφάλαιο (3.3.3) καθώς η συγκεκριμένη υπηρεσία έπαιξε πολύ σημαντικό ρόλο στη συλλογή πληροφορίας.

2.3 *Αλλαγές στην γονιδιακή πληροφορία.*

Τη στιγμή που γράφεται αυτό το κείμενο, η καταγραφή του γονιδιώματος δεν είναι απόλυτα πλήρης. Σε διάφορα σημεία του γονιδιώματος ενδέχεται να υπάρχουν κενά που πρέπει να πληρωθούν καθώς επίσης και σημεία της ακολουθίας που είναι καταγεγραμμένα με λάθος τρόπο και πρέπει να διορθωθούν. Επιπροσθέτως, υπάρχουν περιοχές του γονιδιώματος με γνωστούς απλοτύπους που δεν μπορούν να περιοριστούν σε μία συγκεκριμένη ακολουθία. Όπου υπάρχουν γνωστές εναλλακτικές ακολουθίες ή διορθώσεις στο κυρίως γονιδίωμα η GRC (Genome Reference Consortium) δημοσιεύει εκδόσεις αυτής της εναλλακτικής ακολουθίας. Υπάρχουν δύο είδη εναλλακτικής ακολουθίας:

1. Οι απλότυποι που περιλαμβάνονται στην τελευταία κύρια έκδοση του γονιδιώματος και
2. Τα «μπαλώματα» (patches) που γίνονται διαθέσιμα σαν ελάσσονες εκδόσεις.

Επιπλέον, όταν η έκδοση του GRC είναι GRCh37.P12, αυτό σημαίνει ότι είμαστε στην έκδοση 37, patch 12.

Υπάρχουν δύο είδη patches:

1. Τα novelty patches, τα οποία συμβολίζουν διαφορετικές εκδόσεις για την ίδια ακολουθία, και που θα διατηρηθούν ως εναλλακτικές ακολουθίες στην επόμενη έκδοση και
2. Τα fix patches, που διορθώνουν κενά ή λανθασμένες ακολουθίες στο γονιδίωμα. Θα ενσωματωθούν στην κύρια ακολουθία μόλις η GRC κάνει διαθέσιμη τη νέα έκδοση.

2.3.1 Αλλαγές στην γονιδιακή πληροφορία της Ensembl

Η πληροφορία που αντλήσαμε από την Ensembl αφορά FASTA αρχεία που περιέχουν cDNA καθώς και ncDNA. Στο επόμενο κεφάλαιο θα γίνει εκτενής αναφορά στη δομή των αρχείων αυτών, καθώς και στα προγράμματα που χρησιμοποιήθηκαν για να εξάγουμε αυτήν την αρχική πληροφορία. Το μόνο που χρειάζεται να αναφέρουμε αυτή τη στιγμή, είναι ότι σε αυτά τα αρχεία κάθε transcript ID είναι συνδεδεμένο με ένα γονίδιο και επίσης δίνονται η συντεταγμένες της τοποθεσίας του καθώς και η ακολουθία του. Επίσης δίνονται οι βιότυποι του γονιδίου και του μετάγραφου. Άλλη πληροφορία για το γονίδιο δεν παρέχεται. Επιπλέον, τα μετάγραφα περιέχουν την ακολουθία του γονιδίου, αφού αφαιρεθούν τα εσώνια. Άρα μπορεί δύο διαφορετικά μετάγραφα να είναι συνδεδεμένα με το ίδιο γονίδιο αλλά να έχουν διαφορετικές ακολουθίες, διότι έχουν αφαιρεθεί διαφορετικά τμήματα κατά τη διαδικασία της ωρίμανσης.

2.3.1.1 Αλλαγές στα αναγνωριστικά της Ensembl (Ensembl IDs)

Στην Ensembl μόλις ένα ID σταματήσει να υπάρχει στα αρχεία FASTA που κατεβάζουμε, τότε αποσύρεται τελείως και ένα ή περισσότερα άλλα παίρνουν τη θέση του. Στην καταγραφή των δεδομένων για μετάγραφα μας ενδιαφέρει να παρέχουμε πληροφορία για τα αναγνωριστικά βιομορίων που αντικατέστησαν το βιομόριο που αποσύρθηκε.

Για να ελεγχθούν οι μηχανισμοί αλλαγών, αρχικά υλοποιήθηκαν κάποια πρότυπα προγράμματα, τα οποία απομονώνοντας τα διακριτά IDs από τα αρχεία όλων των εκδόσεων είχαν τους εξής στόχους:

1. Να διαπιστωθεί η ύπαρξη ή μη , διπλοεγγραφών για τα IDs
2. Να διαπιστωθεί ανά έκδοση ποια IDs διατηρούνται και στην επόμενη και ποια διαγράφονται

Αφού διαπιστώθηκε ότι δεν υφίσταται διπλοεγγραφές για τα IDs, αναπτύχθηκαν προγράμματα για την εύρεση των συγκεκριμένων αλλαγών που παρατηρούνται στα δεδομένα που καταγράφονται από έκδοση σε έκδοση.

2.3.1.2 Αλλαγές στις ακολουθίες

Όπως αναφέρθηκε παραπάνω, η διαδικασίες με την οποίες παράγονται οι ακολουθίες από βιολόγους ενέχουν σφάλματα, με αποτέλεσμα να υπάρχουν κενά ή λανθασμένες ακολουθίες βάσεων.

Αυτό σημαίνει ότι υπάρχουν αλλαγές στην ακολουθία των βάσεων του γονιδίου και κατ' επέκταση σε κάποια από τα μετάγραφα. Τέτοιου είδους πληροφορία ενδιαφέρει τους βιολόγους και κατ' επέκταση πρέπει να καταγραφεί.

2.3.1.3 Αλλαγές στα γονίδια που είναι συνδεδεμένα με τα μετάγραφα

Ένα γονίδιο μπορεί να σταματήσει να καταγράφεται ή να αντικατασταθεί από ένα ή περισσότερα γονίδια. Τότε το ID αποσύρεται και στη θέση εισάγονται νέα IDs. Αυτό όμως δεν σημαίνει απαραίτητα ότι αλλάζει η συσχέτιση γονιδίου και μετάγραφου ή η ακολουθία του μετάγραφου. Ένα γονίδιο μπορεί, για παράδειγμα, να σπάσει σε δύο τμήματα, αλλά το μετάγραφο που καταγράφεται να εμπεριέχεται εξ' ολοκλήρου σε ένα από τα τμήματα και έτσι η συσχέτιση γονιδίου - μετάγραφου να διατηρηθεί. Παράλληλα, μπορεί η ακολουθία του μετάγραφου να μην έχει αλλάξει, παρότι άλλαξε το γονίδιο με το οποίο είναι συσχετισμένο. Σε αυτήν την περίπτωση πρέπει να καταγραφεί σε ποιες εκδόσεις ένα μετάγραφο συνδέεται με κάποιο γονίδιο. Όποτε υπάρχει αλλαγή, αυτή πρέπει με κάποιο τρόπο να καταγραφεί.

2.3.1.4 Αλλαγές στην τοποθεσία ενός μετάγραφου

Μπορεί σε μια καινούρια έκδοση, να ανακαλυφθεί ένα μέρος στην αλυσίδα του DNA το οποίο με παλαιότερη μέθοδο δεν εμφανιζόταν και το οποίο πρέπει να χαρτογραφηθεί. Αυτό σημαίνει ότι εισάγονται νέες «εμβόλιμες» ακολουθίες στο γονιδίωμα ή αφαιρούνται κομμάτια που εξαρχής δεν ήταν σωστά τοποθετημένα. Τέτοιου είδους μετακινήσεις μπορεί να οδηγήσουν και σε μετακίνηση του γονιδίου που παράγει ένα μετάγραφο και κατ' επέκταση να αλλάξουν και οι συντεταγμένες του μετάγραφου. Επιπλέον, αυτό ενδέχεται να συμβεί όταν έχουμε κενά στο γονιδίωμα απροσδιορίστου μήκους. Τέτοιου τύπου αλλαγές πρέπει επίσης να καταγράφονται.

2.3.1.5 Περαιτέρω διερεύνηση

Για να γίνει παραπέρα διερεύνηση για το πόσες και ποιες αλλαγές συμβαίνουν χρησιμοποιήθηκε η γλώσσα προγραμματισμού python. Οι λόγοι γι' αυτή την επιλογή είναι οι εξής:

1. Η ταχύτητα της Python. Ειδικά σε σύγκριση συμβολοσειρών ακόμα και μεγάλου μεγέθους η διαδικασία είναι πολύ γρήγορη και αποδοτική.
2. Το γεγονός ότι διαθέτει Regular Expressions (Κανονικές Εκφράσεις), οι οποίες χρησιμεύουν στην αναγνώριση προτύπων μέσα σε συμβολοσειρές, κάτι που επιταχύνει την ανάγνωση αρχείων και την απομόνωση πληροφορίας από αυτά.

Με χρήση των Regular Expressions, λοιπόν, απομονώσαμε από τα αρχεία της κάθε έκδοσης τα δεδομένα που μας ενδιαφέρουν, ενώ στη συνέχεια αποθηκεύτηκαν, ανά έκδοση σε μια βάση δεδομένων. Προκειμένου να καταγραφούν οι μεταβολές των δεδομένων από έκδοση σε έκδοση, πραγματοποιήθηκαν διαδοχικές συγκρίσεις των δεδομένων των εκδόσεων ανά δύο.

2.3.1.6 Σύγκριση δεδομένων διαδοχικών εκδόσεων

Τα δεδομένα των εκδόσεων που χρειάστηκε να ελέγξουμε έχουν μεγάλο μέγεθος (περιέχουν ακολουθίες με αριθμό χαρακτήρων μεγαλύτερο του 200) και αριθμό εγγραφών(περί τις 100.000-180.000 εγγραφές). Το σχεσιακό πρόγραμμα διαχείρισης βάσεων δεδομένων που χρησιμοποιήθηκε (MySQL 5.5) δεν έκανε χρήση αποδοτικών αλγορίθμων, με αποτέλεσμα να καθιστά τη διαδικασία σύγκρισης αρκετά επίπονη.

Για να επιταχυνθεί αυτή η διαδικασία, για οικονομία χρόνου και πόρων εφευρέθηκε νέος αλγόριθμος σύνδεσης (join) των δεδομένων ανάμεσα σε δύο εκδόσεις ο οποίος χρησιμοποιεί δυαδική αναζήτηση για ανίχνευση εγγραφών με το ίδιο Ensembl transcript ID ανάμεσα σε δύο εκδόσεις και στη συνέχεια συγκρίνει τα δεδομένα τους, καταγράφοντας παράλληλα τις αλλαγές σε αυτά.

Παρόλα, εκτελέστηκαν τα ίδια ερωτήματα στην επόμενη έκδοση της MySQL(5.6) που εισάγει τη χρήση αποδοτικότερων αλγορίθμων με αποτέλεσμα την εκπληκτική μείωση των χρόνων εκτέλεσης.

Οι παραπάνω δοκιμή εκτελέστηκε σε εικονικό μηχάνημα στον Ωκεανό.

Οι δοκιμές σε διαφορετικά μηχανήματα και με διαφορετικές μεθόδους αποτελεί και έναν τρόπο να επιβεβαιώσουμε ότι τα δεδομένα που παράγονται από τη χρήση αυτού του αλγορίθμου που χρησιμοποιήθηκε ήταν όντως σωστά.

2.4 Σχετικές εργασίες

Η παρούσα εργασία στοχεύει στην καταγραφή της εξέλιξης δεδομένων γονιδιακής πληροφορίας. Εργασία σχετική με το αντικείμενο της εξέλιξης βιολογικών δεδομένων πραγματοποιήθηκε στη διπλωματική εργασία του Ηλία Κανέλλου.

Σε αντίθεση με την παρούσα εργασία, σκοπός εκείνης της διπλωματικής εργασίας ήταν η καταγραφή της εξέλιξης των δεδομένων που καταγράφονται συγκεκριμένα για microRNAs στο χρόνο.

Τα αποτελέσματα αυτής της εργασίας ενσωματώθηκαν στο εργαλείο Diana Tools του Ινστιτούτου Πληροφοριακών Συστημάτων που δημιουργήθηκε σε συνεργασία με το Ινστιτούτο Αλέξανδρος Φλέμινγκ. Το όνομα της εφαρμογής που δημιουργήθηκε είναι mirPub.

3 *Κατασκευή της Βάσης Δεδομένων – Back-end της Εφαρμογής*

Στο παρόν κεφάλαιο θα γίνει αναφορά καταρχήν στα αρχεία που χρησιμοποιήθηκαν για την παραγωγή των δεδομένων εξέλιξης. Στη συνέχεια, θα περιγραφεί η δομή της βάσης δεδομένων και πώς αυτή μοντελοποιεί τις αλλαγές που εμφανίζονται διαχρονικά. Τέλος θα γίνει μια αναφορά στο πώς χρησιμοποιήθηκαν τα REST Services της Ensembl και της Hugo για την λήψη πληροφορίας για γονίδια και μετάγραφα. Εδώ πρέπει να σημειωθεί ότι η αναζήτηση αλλαγών γίνεται από την έκδοση 54 της Ensembl και μετά. Ο λόγος για αυτό είναι ότι μέχρι και την έκδοση 54 κάποιοι από τους οργανισμούς που μας ενδιαφέρουν δεν έχουν προστεθεί στη βάση δεδομένων και επίσης η πληροφορία δεν είναι συνεχόμενη, ακόμα και για τον άνθρωπο. Δηλαδή δεν διατίθενται αρχεία FASTA για τους οργανισμούς που μας ενδιαφέρουν σε όλες τις εκδόσεις.

3.1 Αρχεία FASTA cDNA και ncDNA για transcripts

Τα γονιδιακά δεδομένα που απομονώθηκαν από τη ΒΔ Ensembl σχετίζονται με τα μετάγραφα (transcripts) . Η πληροφορία είναι δομημένη σε αρχεία FASTA που περιέχουν πληροφορίες για το κάθε μετάγραφο καθώς και την ακολουθία του. Παράδειγμα εγγραφής ενός τέτοιου αρχείου δίνεται παρακάτω:

```
>ENST00000398344 cdna:known chromosome:GRCh37:22:24313554:24316773:-1
gene:ENSG00000099977 gene_biotype:protein_coding
```

```
transcript_biotype:protein_coding
GTTTGCAGAAGCGGGAGGTACCCTAGGCAGCCAATCGGGGAGCGCCGAGTCTCTGTCCAG
CCAATGAGAAGCCAGGTTGCTGTGGCGCCTCGCCCCTCCTCCCTGGTCCGCGAGCCTTGG
GTACCCCCAGCTTTTCTTCCGCCAGAGCTGTTTCCGTTCCTCTGCCCGCCATGCCGTTCC
TGGAGCTGGACACGAATTTGCCCGCCAACCGAGTGCCCGCGGGGCTGGAGAAACGACTCT
GCGCCGCCGCTGCCTCCATCCTGGGCAAACCTGCGGACCGCGTGAACGTGACGGTACGGC
CGGGCCTGGCCATGGCGCTGAGCGGGTCCACCGAGCCCTGCGCGCAGCTGTCCATCTCCT
CCATCGGCGTAGTGGGCACCGCCGAGGACAACCGCAGCCACAGCGCCCACTTCTTTGAGT
TTCTCACCAAGGAGCTAGCCCTGGGCCAGGACCGGATACTTATCCGCTTTTTCCCCTTGG
AGTCCTGGCAGATTGGCAAGATAGGGACGGTCATGACTTTTTTATGATTGGGCACGGAGG
GATCCAGGGCATCTGTGAACTGGCTGCTTCTTCCAGAGAGATCTCTTGGCAGAGTGAGGG
CCTGGAGATAACCAGCTTTGGATTATCCCGCATGCAACATTCCTGTGATCACATAATCCT
CTTCTTCATCCTCATATGAAATAAATGAAGAGAGCTTCCTCATTCAAACATGA
```

Η επικεφαλίδα του αρχείου αποτελείται από πεδία χωρισμένα με κενά ή άνω και κάτω τελείες και δίνονται ως εξής:

```
      1              2     3          4       5     6     7          8      9
  >ENST00000398344  cdna:known  chromosome:GRCh37:22:24313554:24316773:-1
10 gene:ENSG00000099977 gene_biotype:protein_coding 11
  transcript_biotype:protein_coding 12
```

Τα πεδία έχουν ως εξής:

1. Ensembl ID του transcript αποτελείται από το πρόθεμα ENST(για τον άνθρωπο) και 11 αριθμητικά ψηφία
2. Πεδίο SeqType. Οι διακριτές τιμές του είναι:
    - cdna
    - ncrna
    - havana
    - prj_havana
3. Πεδίο status. Οι διακριτές τιμές που παίρνει το πεδίο είναι:
    - known
    - pseudogene
    - novel
    - putative
    - rRNA
    - Mt_tRNA

- Mt_rRNA
- snRNA
- miRNA
- misc_RNA
- snRNA_pseudogene
- rRNA_pseudogene
- snoRNA_pseudogene
- snoRNA
- scRNA_pseudogene
- Mt_tRNA_pseudogene
- tRNA_pseudogene
- miRNA_pseudogene
- misc_RNA_pseudogene
- scRNA
- protein_coding
- lincRNA
- 3prime_overlapping_ncrna
- ncrna_host
- non_coding
- sense_intronic
- sense_overlapping
- processed_transcript
- antisense
- known
- pseudogene
- novel
- putative
- rRNA
- Mt_tRNA
- Mt_rRNA
- snRNA
- miRNA
- misc_RNA
- snRNA_pseudogene
- rRNA_pseudogene

- snoRNA_pseudogene
- snoRNA
- scRNA_pseudogene
- Mt_tRNA_pseudogene
- tRNA_pseudogene
- miRNA_pseudogene
- misc_RNA_pseudogene
- scRNA
- protein_coding
- lincRNA
- 3prime_overlapping_ncrna
- ncrna_host
- non_coding
- sense_intronic
- sense_overlapping
- processed_transcript
- antisense

4. Πεδίο assembly. Οι διακριτές τιμές που παίρνει είναι οι εξής:
    - supercontig
    - chromosome
5. Έκδοση γονιδιώματος
6. Χρωμόσωμα ή ID του patch στο οποίο βρίσκεται το transcript
7. Συντεταγμένη έναρξης
8. Συντεταγμένη λήξης
9. Αριθμός έλικας
10. Ensembl ID του γονιδίου που κωδικοποιεί το transcript

11. Βιότυπος του γονιδίου. Παίρνει τις εξής διακριτές τιμές:
    - 3prime_overlapping_ncrna
    - ambiguous_orf
    - antisense
    - IG_C_gene
    - IG_C_pseudogene
    - IG_D_gene
    - IG_J_gene

- IG_J_pseudogene
- IG_V_gene
- IG_V_pseudogene
- lincRNA
- ncrna_host
- non_coding
- polymorphic_pseudogene
- processed_pseudogene
- processed_transcript
- protein_coding
- pseudogene
- retained_intron
- sense_intronic
- sense_overlapping
- TR_C_gene
- TR_D_gene
- TR_J_gene
- TR_J_pseudogene
- TR_V_gene
- TR_V_pseudogene

12. Βιότυπος του transcript. Παίρνει τις εξής διακριτές τιμές:
    - 3prime_overlapping_ncrna
    - ambiguous_orf
    - antisense
    - disrupted_domain
    - IG_C_gene
    - IG_C_pseudogene
    - IG_D_gene
    - IG_J_gene
    - IG_J_pseudogene
    - IG_V_gene
    - IG_V_pseudogene
    - lincRNA
    - miRNA
    - misc_RNA

- ncrna_host
- nonsense_mediated_decay
- non_coding
- non_stop_decay
- polymorphic_pseudogene
- processed_pseudogene
- processed_transcript
- protein_coding
- pseudogene
- retained_intron
- retrotransposed
- sense_intronic
- sense_overlapping
- snoRNA
- TEC
- transcribed_processed_pseudogene
- transcribed_unprocessed_pseudogene
- translated_processed_pseudogene
- TR_C_gene
- TR_D_gene
- TR_J_gene
- TR_J_pseudogene
- TR_V_gene
- TR_V_pseudogene
- unitary_pseudogene
- unprocessed_pseudogene

3.2 Μοντελοποίηση της Εξέλιξης των Δεδομένων σε Σχεσιακή Βάση

3.2.1 Αρχική μορφή της βάσης

Για την αποθήκευση των δεδομένων που αναφέρθηκαν παραπάνω σε μορφή κατάλληλη για επεξεργασία, δημιουργήθηκε μια προσωρινή βάση δεδομένων με το παρακάτω σχήμα πίνακα,

το οποίο αντιστοιχεί ακριβώς στα παραπάνω πεδία, με εξαίρεση τα πεδία tid, ens_version και transcript_sequence. Το πρώτο πεδίο εισάγεται ως μοναδικό κλειδί για κάθε εγγραφή, ενώ το δεύτερο περιέχει την έκδοση της Ensembl στην οποία υφίσταται το συγκεκριμένο ID. Το τρίτο αποθηκεύει σε κατάλληλο τύπο δεδομένων την ακολουθία του μετάγραφου:

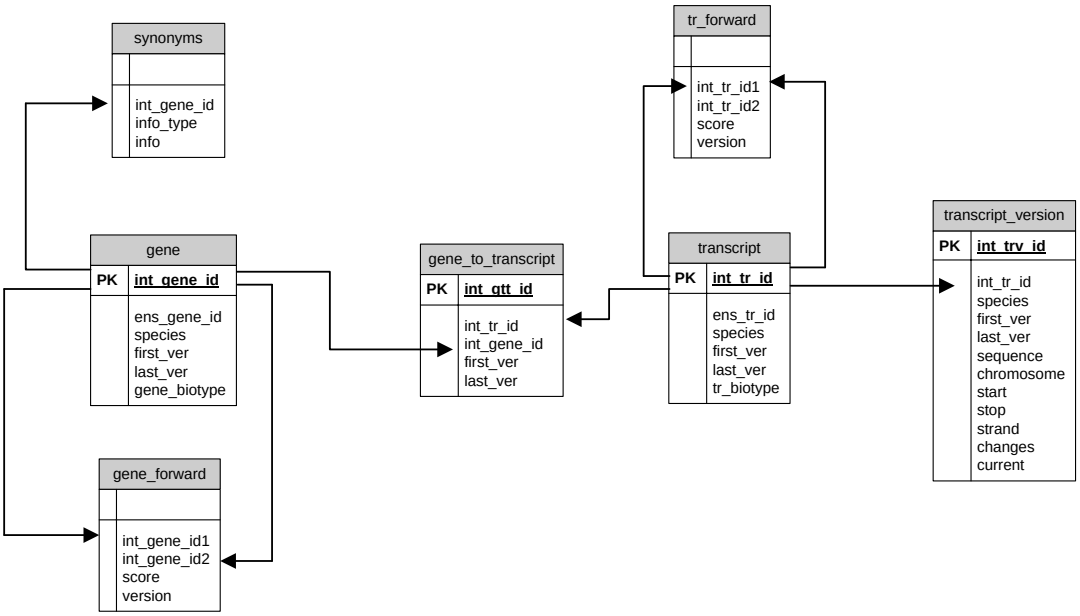
1. ***tid***: bigint, primary key
2. ***ens_version***: int
3. ***ens_transcript_id***: varchar(20)
4. ***seqtype***: varchar(30)
5. ***cdna_typ***e: varchar(40)
6. ***chromosome_version***: varchar(30)
7. ***chromosome***: varchar(30)
8. ***start***: bigint(20)
9. ***stop***: bigint(20)
10. ***strand***: tinyint
11. ***ens_gene_id***: varchar(20)
12. ***gene_biotype***: varchar(40)
13. ***transcript_biotype***: varchar(40)
14. ***transcript_sequence***: longtext

Δημιουργήθηκε ένας πίνακας για καθένα από τα 4 είδη με τα οποία ασχοληθήκαμε (Caenorhabditis Elegans, Drosophila Melanogaster, Homo Sapiens, Mus Musculus).
Το σύνολο των εγγραφών για τους παρακάτω τέσσερις οργανισμούς για τις εκδόσεις 54-75 είναι:

- Caenorhabditis Elegans (cel)
    - Σύνολο εγγραφών: **1.600.044**
    - Μέγεθος δεδομένων στη βάση: **1,5 GB**
- Drosophila Melanogaster (dme)
    - Σύνολο εγγραφών: **549,050**
    - Μέγεθος δεδομένων στη βάση: **1,3 GB**
- Homo Sapiens (hsa)
    - Σύνολο εγγραφών: **3.850.469**
    - Μέγεθος δεδομένων στη βάση: **6,1 GB**
- Mus Musculus (mmu)
    - Σύνολο εγγραφών: **1.702.848**
    - Μέγεθος δεδομένων στη βάση: **5 GB**

3.2.2 Τελική Μορφή της Βάσης – Κωδικοποίηση των Αλλαγών

Η τελική μορφή της βάσης φαίνεται στο παρακάτω διάγραμμα σχεσιακού σχήματος:



Εικόνα 3. Διάγραμμα σχεσιακού σχήματος της τελικής ΒΔ που αναπτύχθηκε

Η δομή του κάθε πίνακα, καθώς και το πώς αυτός κωδικοποιεί την πληροφορία των αλλαγών περιγράφεται παρακάτω.

3.2.2.1 Πίνακας Transcript

Ο πίνακας αυτός έχει 6 πεδία:

1. ***int_tr_id*** : int, unsigned, εσωτερικό ID που είναι και το κύριο κλειδί του πίνακα
2. ***ens_tr_id*** : varchar(20), το Ensembl ID του μετάγραφου
3. ***species*** : enum(‘cel’,’mmu’,’hsa’,’dme’), enumeration για τους 4 οργανισμούς που αναφέρθηκαν νωρίτερα
4. ***first_ver*** : tinyint, unsigned, η πρώτη έκδοση που απαντάται το μετάγραφο
5. ***last_ver*** : tinyint, unsigned, η τελευταία έκδοση που απαντάται το μετάγραφο
6. ***tr_biotype***: varchar(40), ο βιότυπος του transcript

Ο πίνακας αυτός αποθηκεύει τη γενική πληροφορία για το ID κάθε μετάγραφου κατά κύριο λόγο στα πεδία first_ver και last_ver και biotype. Το transcript εμφανίζεται πρώτη φορά στην έκδοση first_ver και αποσύρεται στην last_ver. Μετά την απόσυρσή του υπάρχει πιθανότητα να αντικατασταθεί από ένα ή περισσότερα μετάγραφα.

3.2.2.2 Πίνακας Gene

Ο πίνακας αυτός έχει 6 πεδία:

1. ***int_gene_id*** : int, unsigned, εσωτερικό ID που είναι και το κύριο κλειδί του πίνακα
2. ***ens_gene_id*** : varchar(20), το Ensembl ID του γονιδίου
3. ***species*** : enum('cel','mmu','hsa','dme'), enumeration για τους 4 οργανισμούς που αναφέρθηκαν νωρίτερα
4. ***first_ver*** : tinyint, unsigned, η πρώτη έκδοση που απαντάται το γονίδιο
5. ***last_ver*** : tinyint, unsigned, η τελευταία έκδοση που απαντάται το γονίδιο
6. ***gene_biotype***: varchar(40), ο βιοτύπος του γονιδίου

Όπως και προηγουμένως με τον πίνακα transcript η γενική πληροφορία για την για ένα γονίδιο προέρχεται από τα πεδία first_ver, last_ver και gene_biotype. Μετά την απόσυρσή του συγκεκριμένου ID το γονίδιο είναι πιθανό να αντικατασταθεί από ένα ή περισσότερα gene IDs.

3.2.2.3 Πίνακας Gene_To_Transcript

Ο πίνακας έχει τα παρακάτω 6 πεδία:

1. ***int_gtt_id*** : int, unsigned, εσωτερικό ID για τις εγγραφές του πίνακα που χρησιμεύει ως κύριο κλειδί. Η μόνη του χρήση είναι η γρήγορη ανανέωση των εγγραφών κατά τη διάρκεια κατασκευής των δεδομένων.
2. ***int_gene_id*** : int, unsigned, foreign key εσωτερικό gene ID
3. ***int_tr_id*** : int, unsigned, foreign key εσωτερικό transcript ID
4. ***species*** : enum('cel','mmu','hsa','dme'), enumeration για τους 4 οργανισμούς που αναφέρθηκαν προηγουμένως
5. ***first_ver*** : tinyint, unsigned, η πρώτη έκδοση που απαντάται η συσχέτιση
6. ***last_ver*** : tinyint, unsigned, η τελευταία έκδοση που απαντάται η συσχέτιση

Με βάση τα δεδομένα που αναφέρθηκαν παραπάνω, η σχέση gene και transcript είναι N*M. Με αυτό το δεδομένο και βασιζόμενοι τις αρχές σχεδίασης μιας βάσης δεδομένων χρειαζόμαστε έναν πίνακα συσχέτισης για να την παρουσιάσουμε. Ο πίνακας αυτός είναι ο gene_to_transcript. Ο πίνακας αυτός είναι ίσως από τους πιο σημαντικούς όσον αφορά την παρουσίαση μίας συγκεκριμένης αλλαγής: η αλλαγή του ID του γονιδίου που είναι συσχετισμένο με ένα transcript. Με βάση τα πεδία first_ver και last_ver, καθώς και τα

εσωτερικά IDs μπορούμε να παράγουμε ένα πλήρες ιστορικό της διαχρονικής συσχέτισης transcript-gene. Δηλαδή αν ένα μετάγραφο αλλάξει παράγον γονίδιο στην έκδοση V, τότε το προηγούμενο γονίδιο θα έχει στο πεδίο last_ver την τιμή V-1, ενώ το νέο γονίδιο θα έχει στο πεδίο first_ver την τιμή V. Με αυτόν τον τρόπο δεν χρειάζεται να αποθηκευτεί ρητά η πληροφορία ότι άλλαξε το κωδικοποιόν γονίδιο, κάτι που θα αποτελούσε επανάληψη πληροφορίας.

3.2.2.4 Πίνακας Transcript_Version

Ο πίνακας έχει 12 πεδία:

1. ***int_trv_id*** : int, unsigned, εσωτερικό ID για τις εγγραφές του πίνακα που χρησιμεύει ως κύριο κλειδί.
2. ***int_tr_id*** : int, unsigned, foreign key εσωτερικό transcript ID
3. ***species*** : enum(‘cel’,’mmu’,’hsa’,’dme’), enumeration για τους 4 οργανισμούς που αναφέρθηκαν νωρίτερα
4. ***first_ver*** : tinyint, unsigned, η πρώτη έκδοση που απαντάται η εγγραφή
5. ***last_ver*** : tinyint, unsigned, η τελευταία έκδοση που απαντάται η εγγραφή
6. ***sequence***: longtext, η ακολουθία ενός transcript
7. ***chromosome***: varchar(30) , το χρωμόσωμα στο οποίο ανήκει το transcript
8. ***start*** : bigint, η συντεταγμένη έναρξης του μετάγραφου
9. ***stop*** : bigint, η συντεταγμένη λήξης του μετάγραφου
10. ***strand***: tinyint, ο αριθμός της συμπληρωματικής έλικας του μετάγραφου
11. ***changes***: tinyint, κωδικοποιημένη αποθήκευση της αλλαγής που συνέβη και μετέτρεψε την εγγραφή σε παρωχημένη
12. ***curren***t: tinyint, η εγγραφή είναι η πιο πρόσφατη για το transcript. Χρησιμοποιείται μόνο για την παραγωγή των διαχρονικών δεδομένων.

Ο πίνακας αυτός αποθηκεύει τις διάφορες εκδόσεις ενός transcript διαχρονικά. Για οικονομία χώρου σε σχέση με το αρχικό σχήμα της βάσης δημιουργείται μια νέα εγγραφή μόνο όταν παρατηρείται αλλαγή σε κάποιο από τα στοιχεία του transcript. Τότε αποθηκεύεται η αλλαγή που συνέβη στο πεδίο changes της εγγραφής που μόλις αποσύρθηκε (current → 0).

Τα πεδία chromosome, start, stop και strand αποτελούν την τοποθεσία (Location) ενός transcript. Αν αλλάξει ένα από αυτά, τότε έχουμε αλλαγή τύπου Location Change.

Το πεδίο sequence αποθηκεύει την ακολουθία του transcript. Αν αυτή αλλάξει, τότε έχουμε Sequence Change.

Αν αλλάζει ταυτόχρονα και το Location και το Sequence, τότε έχουμε Sequence & Location Change.

Αυτή η αλλαγή κωδικοποιείται στο πεδίο changes με τον εξής τρόπο:

- Αν έχουμε Sequence Change, τότε κάνουμε 1 το *πρώτο* LSB του πεδίου της προς απόσυρση εγγραφής.
- Αν έχουμε Location Change, τότε κάνουμε 1 το *δεύτερο* LSB του πεδίου της προς απόσυρση εγγραφής. .

Άρα για:

- Sequence Change, changes → changes=1
- Location Change, changes → changes=2
- Sequence & Location Change, changes → changes=3

Τέλος, όπως και παραπάνω, τα πεδία first_ver και last_ver αποθηκεύουν το για ποιες εκδόσεις ήταν ενεργή η εγγραφή. Οι νέες εγγραφές που αντικαθιστούν αυτές που αποσύρονται και περιέχουν τα ανανεωμένα στοιχεία του transcript έχουν την τιμή του πεδίου current ίση με 1.

3.2.2.5 Πίνακας Tr_Forward

Ο πίνακας αυτός περιέχει τα εξής πεδία:

1. ***int_tr_id1*** : int, unsigned, foreign key εσωτερικό transcript ID
2. ***int_tr_id2*** : int, unsigned, foreign key εσωτερικό transcript ID
3. ***score*** : float(10,0), πεδίο που εμφανίζει την πιθανότητα το transcript2 να αντικαθιστά το transcript1.
4. ***version*** : tinyint, η έκδοση στην οποία συμβαίνει η αντικατάσταση.

Ο πίνακας αυτός περιέχει πληροφορία για ενδεχόμενη αντικατάσταση ενός transcript από δύο ή περισσότερα νέα IDs εάν αυτό αποσύρεται σε κάποια έκδοση. Η πιθανότητα αυτή δίνεται στο πεδίο score.

Τα δεδομένα που αποτελούν τις εγγραφές αυτού του πίνακα προέρχονται από το REST Service της Ensembl με Endpoint το archive/:id. Η υπηρεσία αυτή παρέχει πληροφορίες για τα περισσότερα transcripts που βρίσκονται στο αρχείο της Ensembl, δηλαδή σε παλαιότερες εκδόσεις της γονιδιακής βάσης. Κάποια IDs που αποσύρθηκαν πριν πολύ καιρό δεν υπάρχουν στη βάση της Ensembl και άρα δεν επιστρέφουν κάποιο χρήσιμο αποτέλεσμα. Τα υπόλοιπα επιστρέφουν απάντηση σε μορφή JSON ή XML. Αν υπάρχουν *πιθανά ID-αντικαταστάστες* (θα τα αποκαλούνται από εδώ και πέρα *απόγονοι*), τότε αυτά επιστρέφονται στα περιεχόμενα της απάντησης του εξυπηρετητή.. Αλλιώς η λίστα με τους απογόνους του transcript είναι άδεια.

Πρέπει να τονιστεί εδώ ιδιαίτερα, ότι η ίδια η Ensembl δεν παρέχει ακριβή πληροφορία για το κατά πόσο ένα transcript ID διαδέχεται ένα άλλο. Αντίθετα τα ονομάζει «πιθανούς

διαδόχους» δημιουργώντας ουσιαστικά μια οικογένεια από IDs συσχετισμένα μεταξύ τους, που αποτελούν και την ιστορία καθενός από τα IDs που είναι μέλη της.

3.2.2.6 Πίνακας Gene_Forward

Ο πίνακας αυτός περιέχει τα εξής πεδία:

1. ***int_gene_id1*** : int, unsigned, foreign key εσωτερικό gene ID
2. ***int_gene_id2*** : int, unsigned, foreign key εσωτερικό gene ID
3. ***score*** : float(10,0), πεδίο που εμφανίζει την πιθανότητα το gene2 να αντικαθιστά το gene1.
4. ***version*** : tinyint, η έκδοση στην οποία συμβαίνει η αντικατάσταση.

Σε αυτόν τον πίνακα ισχύει για τα gene IDs ό,τι ισχύει στον παραπάνω πίνακα για τα transcript IDs. Δηλαδή αποθηκεύουμε έμμεσα πληροφορία για την «οικογένεια» των gene IDs.

3.2.2.7 Πίνακας Synonyms

Ο πίνακας περιέχει πληροφορία που αφορά επίσημα ονόματα και σύμβολα καθώς και εναλλακτικά σύμβολα και ονόματα, με τα οποία είναι γνωστό ένα γονίδιο στην επιστημονική κοινότητα. Ο πίνακας περιέχει τα παρακάτω πεδία:

1. ***int_gene_id*** : int, unsigned, foreign key εσωτερικό gene ID
2. ***info_type*** : enum(‘alias_name’, ‘alias_symbol’, ‘name’, ‘prev_name’, ’prev_symbol’, ’symbol’), ο τύπος της ακολουθίας που αποθηκεύεται
3. ***info*** : varchar(200), η αποθηκευμένη πληροφορία για το γονίδιο

Η πληροφορίες αυτού του πίνακα προέρχονται από το REST Service της HGNC. Περιέχουν εγκεκριμένα ονόματα, σύμβολα, συνώνυμα καθώς και προηγούμενα σύμβολα και ονόματα. Η αίτηση στο REST service γίνεται με το Ensembl gene ID του κάθε γονιδίου και η απάντηση είναι στη μορφή. Οι τύποι name και symbol παρέχουν το πιο πρόσφατο όνομα και σύμβολο εγκεκριμένο από την HGNC. Οι τύποι alias_name και alias_symbol περιέχουν εναλλακτικά ονόματα και σύμβολα με τα οποία είναι γνωστό ένα γονίδιο. Τέλος, οι τύποι prev_name και prev_symbol περιέχουν ονόματα και σύμβολα που ήταν παλιότερα εγκεκριμένα αλλά έχουν πλέον αποσυρθεί.

3.2.2.8 Συμπεράσματα

Η δομή της βάσης που αναπτύχθηκε μοιάζει στη δομή με τη βάση δεδομένων που δημιουργήθηκε κατά την του συστήματος αλλαγών μορίων microRNA στις εφαρμογές DIANA lab. Είναι αρκετά αποδοτική εφόσον είναι κανονικοποιημένη και άρα δεν έχουμε πλεονασμό στην πληροφορία. Το συγκεκριμένο σχήμα βάσης δεδομένων δίνει τη δυνατότητα

εξαγωγής της πληροφορίας που αφορά τη διαχρονική εξέλιξη των γονιδίων, ακόμα κι αν γίνεται με έμμεσο τρόπο.

3.3 Υλοποίηση Προγραμμάτων για Συγκέντρωση Πληροφορίας

Η συλλογή πληροφορίας και η αναζήτηση αλλαγών έγιναν αποκλειστικά χρησιμοποιώντας γλώσσα Python. Ο λόγος που την κάνει τόσο ελκυστική σε σχέση με άλλες γλώσσες προγραμματισμού είναι κυρίως η ταχύτητα επεξεργασίας, αλλά η μεγάλη ποικιλία σε δομές δεδομένων και βιβλιοθήκες, που καλύπτουν μεγάλο εύρος των αναγκών ενός προγραμματιστή.

3.3.1 Αρχική συλλογή πληροφορίας

Για να γίνει η αρχική συλλογή πληροφορίας, δημιουργήθηκε ένα script σε Python, το οποίο για κάθε έκδοση:

1. Συνδέεται με τον FTP Server της Ensembl και «κατεβάζει» το αντίστοιχο συμπιεσμένο αρχείο fa.gz όταν το αρχείο FASTA για μια δεδομένη έκδοση δεν είναι ήδη αποθηκευμένο. Η λειτουργία αυτή υλοποιείται μέσω της εντολής wget του linux.
2. Αν το αρχείο κατέβει σωστά, τότε το αποσυμπιέζει χρησιμοποιώντας την εντολή gzip του linux.
3. Ανοίγει το αρχείο σε λειτουργία ανάγνωσης.
4. Χρησιμοποιώντας αναγνώριση προτύπων χρησιμοποιώντας κανονικές εκφράσεις (regular expressions) διαβάζει το αρχείο και το αναλύει σε πεδία.
5. Για κάθε ID που αναλύει, γράφει τα δεδομένα σε ένα .sql αρχείο που περιέχει ήδη το σκελετό ενός INSERT query της MySQL χρησιμοποιώντας τη σύνταξη “INSERT INTO table_name (field_list) VALUES (value_list) που επιτρέπει την εισαγωγή πολλαπλών εγγραφών με ένα ερώτημα. Αυτό αυξάνει θεαματικά την απόδοση της MySQL για μεγάλο όγκο δεδομένων.
6. Μόλις τελειώσει με την ανάγνωση του αρχείου FASTA κλείνει το .sql αρχείο και το ανοίγει ξανά σε λειτουργία ανάγνωσης. Εν συνεχεία διαβάζει το αρχείο και κάνει bulk insert στους αντίστοιχους πίνακες της βάσης δεδομένων

Απαραίτητο για να λειτουργήσει το συγκεκριμένο πρόγραμμα, είναι να είναι έχει αυξηθεί το bulk_insert_buffer της MySQL στην τιμή 1GB, διαφορετικά η εκτέλεση του ερωτήματος διακόπτεται μαζί με το πρόγραμμα. Αυτό συμβαίνει λόγω του μεγάλου όγκου πληροφορίας που πρέπει να εισαχθεί στη βάση.

Με αυτόν τον τρόπο δημιουργείται μία δομή πληροφορίας, η οποία επιτρέπει πλέον την εύρεση αλλαγών μεταξύ δύο εκδόσεων.

3.3.2 Εύρεση Αλλαγών Μεταξύ Εκδόσεων

Προκειμένου να εξαχθεί η πληροφορία αυτή από τα ήδη υπάρχοντα δεδομένα, με τρόπο που αποδοτικό, το πρόγραμμα εύρεσης διαχρονικών αλλαγών που δημιουργήθηκε βασίστηκε στον πρότυπο αλγόριθμο σύνδεσης δεδομένων που περιγράφηκε στο προηγούμενο κεφάλαιο. Αρχικά δημιουργήθηκε ένα πρόγραμμα σε Python, το οποίο περνά τα δεδομένα της έκδοσης 54 στο τελικό σχήμα της βάσης.

Στη συνέχεια εκτελείται ένα δεύτερο πρόγραμμα που αναγνωρίζει τις αλλαγές που συνέβησαν για όλες τις διαθέσιμες εκδόσεις που είναι μεταγενέστερες της 54. Δηλαδή για κάθε έκδοση v της Ensembl ξεκινώντας από την έκδοση 55 το πρόγραμμα εκτελεί τα ακόλουθα βήματα:

1. Φέρνει από τη βάση τα δεδομένα της έκδοσης v και v-1 ταξινομημένα.
2. Για κάθε ID στην έκδοση v κάνει δυαδική αναζήτηση στα IDs της έκδοσης v-1.
3. Εάν βρεθεί το σχετικό ID, τότε:
    - Ελέγχει εάν έχει αλλάξει το ID του παράγοντος γονιδίου σε σχέση με την προηγούμενη έκδοση. Αν δεν έχει αλλάξει, τότε ανανεώνει το πεδίο last_ver στους πίνακες gene και gene_to_transcript. Αν έχει αλλάξει, τότε ελέγχεται εάν αυτό υπάρχει ήδη στη βάση από την προηγούμενη αναφορά σε αυτό από κάποιο άλλο transcript. Αν υπάρχει ήδη, τότε ανανεώνεται το last_ver του gene και προστίθεται μια νέα εγγραφή στον πίνακα gene_to_transcript. Διαφορετικά προστίθεται μια νέα εγγραφή στους πίνακες gene και gene_to_transcript.
    - Ελέγχει εάν έχει αλλάξει η ακολουθία (sequence) ή η τοποθεσία του transcript (location). Αν συμβεί αυτό, τότε ανανεώνεται το πεδίο changes της πιο πρόσφατης έκδοσης για το transcript στον πίνακα transcript_version και η εγγραφή αποσύρεται (current→ 0). Προστίθεται μια νέα εγγραφή στον πίνακα transcript_version με current=1 καθώς και τα νέα στοιχεία του transcript. Εάν δεν έχει αλλάξει κανένα στοιχείο, τότε ανανεώνεται το πεδίο last_ver της πιο πρόσφατης εγγραφής.
    - Σε κάθε περίπτωση, εφόσον το transcript ID βρέθηκε, ανανεώνουμε το πεδίο last_ver της εγγραφής στον πίνακα transcript.
4. Εάν το ID της έκδοσης V δεν βρεθεί στην V-1 τότε είναι σίγουρο ότι το ID αυτό δεν υπάρχει σε προηγούμενη έκδοση είναι δεδομένο ότι από τη στιγμή που ένα ID

αποσύρεται, τότε δεν ξαναεμφανίζεται σε επόμενες εκδόσεις. Σε αυτή την περίπτωση:

- Εάν το gene ID υπάρχει ήδη στον πίνακα genes λόγω της αναφοράς σε αυτό από κάποιο άλλο transcript, τότε προστίθεται απλά μια εγγραφή στον πίνακα gene_to_transcript και ανανεώνεται το last_ver του gene. Εάν δεν υπάρχει, τότε προστίθενται νέες εγγραφές στους πίνακες gene και gene_to_transcript.
- Προστίθενται νέες εγγραφές στους πίνακες transcript και transcript_version με βάση τα στοιχεία του νέου ID.

5. Προχωρά στην επόμενη έκδοση (v++).

Το τελευταίο script, βασίζεται στη μοναδικότητα ενός συγκεκριμένου ID σε μία δεδομένη έκδοση της Ensembl. Επίσης, για μεγαλύτερη ταχύτητα, τα δεδομένα γράφονται σε αρχεία κειμένου, τα οποία περιέχουν ένα ερώτημα με βάση τα διαχρονικά δεδομένα που βρέθηκαν κατά την επεξεργασία της πληροφορίας που διαθέτουμε. Δηλαδή για κάθε έκδοση, πρώτα γίνεται η επεξεργασία των αντίστοιχων δεδομένων και στη συνέχεια η αποθήκευση της πληροφορίας που παρήχθη στη βάση. Αυτό μειώνει αρκετά το χρόνο εκτέλεσης, καθότι μειώνεται η επιβάρυνση που προκαλείται από τη δημιουργία σύνδεσης με τη βάση δεδομένων καθώς και την επικοινωνία και τη μεταφορά πληφορορίας. Τα παραπάνω τρία προγράμματα επανεκτελούνται για καθέναν από τους 4 οργανισμούς που μας ενδιαφέρουν.

3.3.3 Λήψη Πληροφορίας Σχετικά με τα Forwards

Για το σκοπό αυτό δημιουργήθηκε πάλι ένα script σε Python, με βάση το πρότυπο χρήσης που δίνεται από το REST Service της Ensembl. Όπως αναφέρθηκε και νωρίτερα το endpoint που χρησιμοποιήθηκε είναι το archive/:id. Το συγκεκριμένο script για κάθε έκδοση v (το v παίρνει τιμές από 54 έως την *προτελευταία έκδοση*) :

1. Φέρνει στη μνήμη τα transcript και gene IDs, των οποίων το πεδίο last_ver είναι ίσο με v.
2. Στη συνέχεια για κάθε ID στη μνήμη εκτελείται ένα HTTP request στον REST Server της Ensembl.
3. Αν η απόκριση του Server ήταν επιτυχής τότε εάν η λίστα με τους πιθανούς απογόνους δεν είναι άδεια, προστίθενται εγγραφές συσχέτισης στον πίνακα tr_forward/gene_forward. Εάν η λίστα είναι άδεια, τότε δεν γίνεται τίποτα
4. Εάν η απόκριση του Server επιστρέψει κωδικό λάθους, τότε το ID μπαίνει σε μια λίστα που περιέχει τα IDs που δεν επέστρεψαν απάντηση.
5. Μόλις τελειώσει η εκτέλεση της συγκεκριμένης έκδοσης και όσο ο αριθμός των IDs που δεν επέστρεψαν απάντηση στη λίστα μειώνεται, επιστρέφουμε στο βήμα 1 και

εκτελούμε ξανά ερωτήματα για τα IDs στη λίστα με τις λάθος απαντήσεις. Το βήμα αυτό είναι απαραίτητο ειδικά σε μηχανήματα με πολύ γρήγορη σύνδεση στο Διαδίκτυο, διότι στέλνουν αιτήσεις HTTP πολύ γρήγορα, με αποτέλεσμα να μην προλαβαίνει να ανταποκριθεί ο εξυπηρετητής και να στέλνει πίσω κωδικό λάθους.

6. Μόλις ο αριθμός αυτός σταθεροποιηθεί, τότε προχωράμε στην επόμενη έκδοση.

Το συγκεκριμένο script επίσης γράφει την πληροφορία σε αρχεία κειμένου πριν τα περάσει στη βάση για κάθε έκδοση. Στη συγκεκριμένη περίπτωση, πέραν της μείωσης του χρόνου εκτέλεσης, έχουμε και ένα επιπρόσθετο πλεονέκτημα : εάν το πρόγραμμα «κολλήσει» σε μια έκδοση ή χαθεί η πρόσβαση στο Διαδίκτυο στο μηχάνημα στο οποίο εκτελείται το πρόγραμμα (πράγμα πολύ πιθανό), τότε είναι σίγουρο ότι οι αλλαγές που έχουν καταχωρηθεί στη βάση αντιστοιχούν μέχρι και την προηγούμενη έκδοση, οπότε μπορούμε απλά να εκτελέσουμε ξανά το πρόγραμμα από την έκδοση στην οποία προέκυψε το πρόβλημα και μετά.

3.3.4 Λήψη Πληροφορίας Σχετικά με τα Ονόματα Γονιδίων

Στο στάδιο αυτό έγινε χρήση του REST Service της HGNC προκειμένου να γίνει λήψη πιο περιγραφικής πληροφορίας για κάποιο γονίδιο από το Ensembl gene ID. Για άλλη μια φορά το πρόγραμμα δημιουργήθηκε σε Python με βάση το πρότυπο σύνδεσης που παρέχεται από την HGNC, που είναι ελαφρώς διαφορετικό από αυτό που χρησιμοποιήθηκε στο προηγούμενο βήμα από την Ensembl. Εδώ, λοιπόν, και με σκοπό εκτός των άλλων, να διατηρηθεί το πλεονέκτημα που υπήρχε και στο προηγούμενο στάδιο, πραγματοποιούνται ερωτήματα για τα δεδομένα ανά έκδοση. Δηλαδή για κάθε έκδοση v ακολουθούνται τα εξής βήματα:

1. Φορτώνονται στη μνήμη τα Ensembl gene IDs που τελειώνουν στην έκδοση v.
2. Για κάθε τέτοιο ID εκτελείται ερώτημα στον server της HGNC.
3. Αν η απόκριση είναι σωστή, τότε γίνεται επεξεργασία και αποθήκευση της χρήσιμης πληροφορίας.
4. Εάν η απόκριση δεν ήταν επιτυχής, το ID μπαίνει σε μια λίστα που επέστρεψαν κωδικό λάθους.
5. Μόλις τα ID τελειώσουν, ξαναεκτελούνται ερωτήματα για τα IDs που επέστρεψαν κωδικό λάθους μέχρι ο αριθμός τους να μείνει σταθερός. Όπως και παραπάνω το βήμα αυτό είναι απαραίτητο εφόσον χρειάζεται να πάρουμε σωστή πληροφορία σε μηχανήματα με πολύ γρήγορη σύνδεση στο Διαδίκτυο.
6. Εάν δεν μειώνεται άλλο ο αριθμός των απαντήσεων με κωδικό λάθους, τότε προχωρά η διαδικασία στην επόμενη έκδοση.

Και σε αυτό το πρόγραμμα και για τους λόγους που αναφέρθηκαν παραπάνω, πρώτα εκτελείται η επεξεργασία των δεδομένων και στη συνέχεια αποθηκεύονται στη βάση ανά έκδοση.

3.3.5 Πρόγραμμα Εύρεσης Αλλαγών για Μελλοντικές Εκδόσεις

Το τελευταίο script δημιουργήθηκε ως συνδυασμός των παραπάνω. Δέχεται σαν παράμετρο ένα σύνδεσμο προς το νεότερο αρχείο που εκδόθηκε. Στη συνέχεια, με χρήση κανονικών εκφράσεων για αναγνώριση προτύπων αποθηκεύει τα δεδομένα της καινούριας έκδοσης στη μνήμη RAM και συγκρίνει τα δεδομένα με αυτά της τελευταία καταγεγραμμένης έκδοσης στη βάση δεδομένων, τα οποία επίσης φορτώνονται στη μνήμη. Βρίσκει τις αλλαγές και τις αποθηκεύει είτε ως ανανεώσεις είτε ως νέες εγγραφές στη βάση δεδομένων, ανανεώνοντας έτσι τα δεδομένα στην πιο πρόσφατη έκδοση.

Στη συνέχεια, πρέπει να εκτελεστούν τα προγράμματα για τα forwards και τα synonyms για όλες τις εκδόσεις γιατί είναι δεδομένα που υπόκεινται σε συχνές αλλαγές και πρέπει να ανανεώνονται για κάθε νέα έκδοση που παράγεται από την Ensembl. Για παράδειγμα μπορεί για το γονίδιο ENSG00000xxxxxx η HGNC να έχει εγκρίνει ένα νέο όνομα στο χρονικό διάστημα από την τελευταία μέχρι τη νέα έκδοση της Ensembl. Επιθυμητό είναι να υπάρχει η τελευταία έκδοση αυτής της πληροφορίας στη βάση μας.

3.4 Ειδικές Περιπτώσεις

Παρότι δεν αντιμετωπίστηκαν προβλήματα κατά την εκτέλεση των παραπάνω προγραμμάτων, εντούτοις τα δεδομένα που χρησιμοποιήσαμε για να παράγουμε την πληροφορία είναι σε μερικές περιπτώσεις ελλιπή. Για παράδειγμα, παρότι εκπρόσωπος της Ensembl μας επιβεβαίωσε ύστερα από επικοινωνία ότι μόλις ένα ID αποσυρθεί δεν επανεμφανίζεται, γεγονός που ενώ ισχύει, παρόλα αυτά κάποια IDs απουσιάζουν από τα αντίστοιχα FASTA αρχεία μόνο ορισμένων εκδόσεων.

Το συγκεκριμένο γεγονός είναι αποτελεί πρόβλημα δεδομένου του τρόπου που λειτουργούν τα προγράμματά που υλοποιήθηκαν, καθώς έχουν ως αποτέλεσμα την εμφάνιση διπλοεγγραφών σε transcripts με διαφορετικό εσωτερικό ID, αλλά ίδιο Ensembl ID και διαφορετικά πεδία first_ver και last_ver. Ευτυχώς Ωστόσο τα δεδομένα αυτά είναι λίγα σε αριθμό (περίπου 20.000 από τα 500.000 συνολικά για όλους τους οργανισμούς – μόνο 14 εγγραφές για τον άνθρωπο) οπότε μπορούν να αντιμετωπιστούν κατάλληλα από το front-end της εφαρμογής, το οποίο μπορεί να προβλέψει τη συγχώνευση τέτοιων εγγραφών. Σαν μελλοντικό πρόβλημα προς αντιμετώπιση, θα μπορούσε να οριστεί η αναγνώριση αυτών των IDs και η εξαγωγή των δεδομένων τους για τις ενδιάμεσες εκδόσεις, με σύνδεση και ερωτήματα στη βάση δεδομένων που διαθέτει η Ensembl προς το κοινό.

4 *Ανάπτυξη Εφαρμογής*

Στο παρόν κεφάλαιο θα γίνει λόγος για την εφαρμογή οπτικοποίησης των δεδομένων που παρήχθησαν με τις μεθόδους που περιγράφηκαν στο κεφάλαιο 3.3.

4.1 Ανάλυση Απαιτήσεων της Εφαρμογής

Η εφαρμογή που αναπτύχθηκε, προορίζεται για τον όσο το δυνατόν συνοπτικότερο και πιο ευανάγνωστο τρόπο παρουσίασης των μεταβολών των δεδομένων σε ειδικευμένους επιστήμονες βιολόγους. Για τον λόγο αυτό, προτιμήθηκε η προβολή των δεδομένων με πίνακες, οι οποίοι στη συνέχεια με τη χρήση εργαλείων CSS μετατράπηκαν σε γράφους – χρονοδιαγράμματα με ενδιάμεσες συνδέσεις. Οι πίνακες παρέχουν έναν εύκολο τρόπο εποπτικής παρουσίασης της πληροφορίας. Παράλληλα, εφόσον η πληροφορία που διαθέτουμε για τα IDs, ειδικότερα σε ότι αφορά σε αντικαταστάσεις IDs από κάποια άλλα, δεν είναι απόλυτα γραμμική, δηλαδή δεν μπορούμε με απόλυτη σιγουριά να αποφανθούμε αν ένα gene ή transcript αντικαθίσταται από ένα άλλο, χρησιμοποιούμε μια δομή που θα επιτρέπει να εμφανιστούν ταυτόχρονα όλες οι πιθανότητες.

Τα χρονοδιαγράμματα αυτά παρουσιάζουν στον οριζόντιο άξονα όλες τις διαθέσιμες εκδόσεις για τις οποίες έχουμε πληροφορία ενώ στον κάθετο περιέχουν κάποια IDs που συνδέονται μεταξύ τους είτε μέσω της σχέσης gene ↔ transcript, είτε μέσω της σχέσης gene ↔ gene είτε μέσω της σχέσης transcript ↔ transcript .

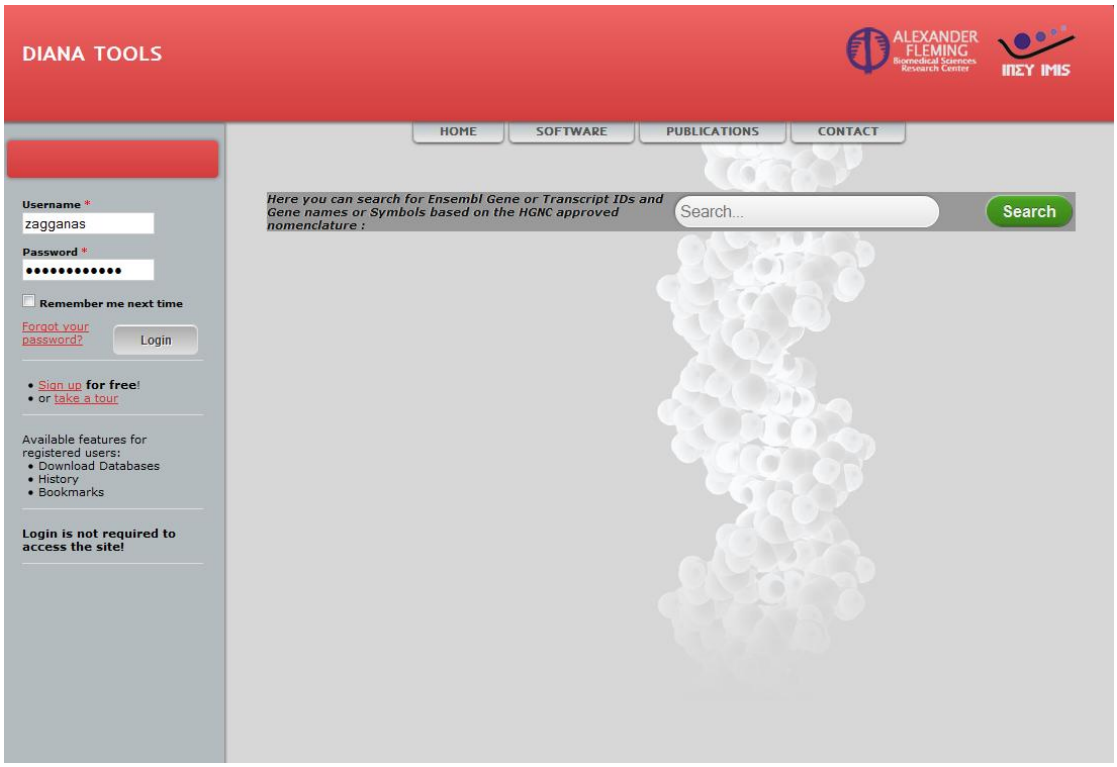
Για κάθε γραμμή του χρονοδιαγράμματος ισχύει ότι υπάρχει κάποιος κόμβος κάτω από μία έκδοση στην περίπτωση που το αντίστοιχο ID εμφανίζεται στη συγκεκριμένη έκδοση. Οι κόμβοι της ίδιας γραμμής συνδέονται μεταξύ τους με οριζόντιες γραμμές που υποδηλώνουν είτε κάποιο είδος αλλαγής είτε τη διατήρηση αναλλοίωτων των δεδομένων. Κάθε είδος αλλαγής σημειώνεται με ξεχωριστό χρώμα. Αυτή η μέθοδος μας επιτρέπει την συνοπτική εμφάνιση μια «οικογένεια» από IDs δείχνοντας σε ποιες εκδόσεις είναι ενεργό το κάθε ID.

Με μια σωστή ταξινόμηση μπορεί να φανούν οι διάδοχοι ενός συγκεκριμένου ID. Εποπτικά φαίνεται σαν να υπάρχει μια «συνέχεια» χωρίς αυτή να δηλωθεί ρητά με βέλη. Η ίδια η Ensembl παρέχει ιστορικό για τα IDs, αλλά δεν παρέχει το ιστορικό των αλλαγών πλην της διαγραφής κάποιου ID.

4.1.1 Σελίδα Αναζήτησης

Η σελίδα αυτή περιλαμβάνει ένα πλαίσιο μέσα στο οποίο βρίσκεται μία μπάρα αναζήτησης. Σε αυτή τη σελίδα ο χρήστης μπορεί να αναζητήσει είτε ένα συγκεκριμένο transcript και gene ID είτε το όνομα ενός γονιδίου με βάση την πληροφορία που διαθέτουμε από την HGNC. Στην τελευταία περίπτωση, εμφανίζεται μια σελίδα με πιθανά αποτελέσματα και προτάσεις ακόμα κι αν υπάρχει ένα μόνο αποτέλεσμα, διότι δεν είναι λειτουργικό για το χρήστη να προωθείται στη σελίδα του ID που βρέθηκε χωρίς πρώτα να το επιλέξει. Η αμφισημία των ονομάτων γονιδίων επιβάλλει το παραπάνω, αφού ο χρήστης μπορεί να εισάγει μόνο μέρος του ονόματος ενός γονιδίου στη μπάρα αναζήτησης.

Για διευκόλυνση του χρήστη, η μπάρα αναζήτησης προβάλλεται παντού, ώστε αν προκύψει ανάγκη νέας αναζήτησης, ο χρήστης να διευκολύνεται, καθώς δεν χρειάζεται να επιστρέψει σε προηγούμενη σελίδα. Τέλος, το λήμμα το οποίο αναζήτησε ο χρήστης εμφανίζεται συνέχεια μέσα στη μπάρα στη σελίδα των αποτελεσμάτων, μέχρις ότου αναζητήσει κάτι καινούριο.

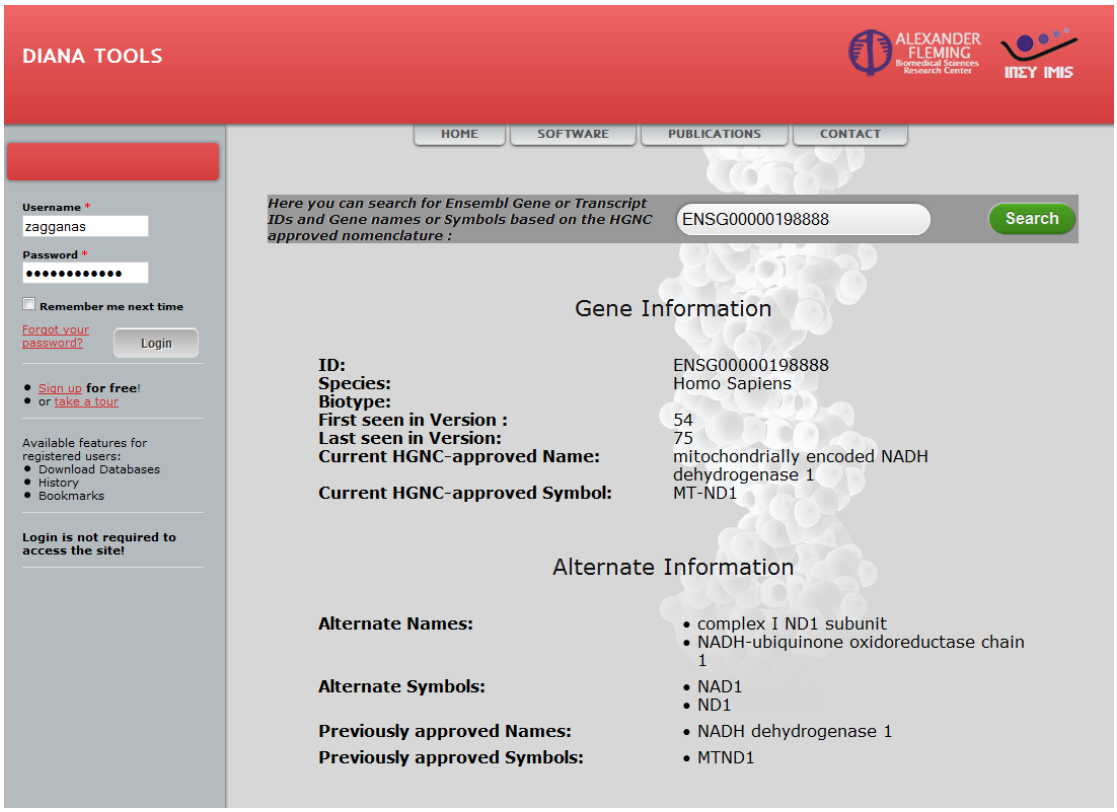


Εικόνα 4. Η σελίδα αναζήτησης ενός Ensembl ID ή ονόματος γονιδίου

4.1.2 Σελίδα αποτελεσμάτων για γονίδια

Η σελίδα αυτή αρχικά παρουσιάζει τις διαθέσιμες πληροφορίες για ένα γονίδιο. Οι πληροφορίες παρουσιάζονται σε δύο πίνακες και δύο γραφήματα.

Στον πρώτο πίνακα παρουσιάζονται πληροφορίες που αφορούν τις εκδόσεις στις οποίες το ID ήταν ενεργό καθώς και τον βιοτύπο του. Επιπρόσθετα, εμφανίζονται, αν υπάρχουν, οι πληροφορίες που αντλήσαμε από την HGNC, όπως το πιο πρόσφατο όνομα και το σύμβολο του γονιδίου. Παράλληλα στον δεύτερο πίνακα εμφανίζεται, αν υπάρχει, πληροφορία για εναλλακτικά ονόματα και σύμβολα καθώς και προηγούμενα σύμβολα και ονόματα. Η πληροφορία αυτή παρέχεται από τον πίνακα της βάσης synonyms. Αυτό φαίνεται στην παρακάτω εικόνα:



Εικόνα 5. Το τμήμα εμφάνισης των πληροφοριών ενός γονιδίου

Στον πρώτο πίνακα – γράφημα εμφανίζουμε τη σχέση gene ↔ trancripts. Δηλαδή παρουσιάζουμε με ποια μετάγραφα είναι συνδεδεμένο ένα γονίδιο σε ποιες εκδόσεις συμβαίνει αυτό. Ο λόγος για μια τέτοια παρουσίαση, είναι γιατί μπορεί σε μια έκδοση να

υπάρχουν περισσότερα από ένα μετάγραφα που συνδέονται με το γονίδιο. Αυτά τα δεδομένα αντλούνται από τον πίνακα της βάσης gene_to_transcript. Θεωρήθηκε όμως σκόπιμο, ο πίνακας αυτός να περιέχει και τις *αλλαγές που συμβαίνουν σε κάθε μετάγραφο*. Αυτές παρουσιάζονται με διαφορετικό χρώμα και με τα παρακάτω σύμβολα:

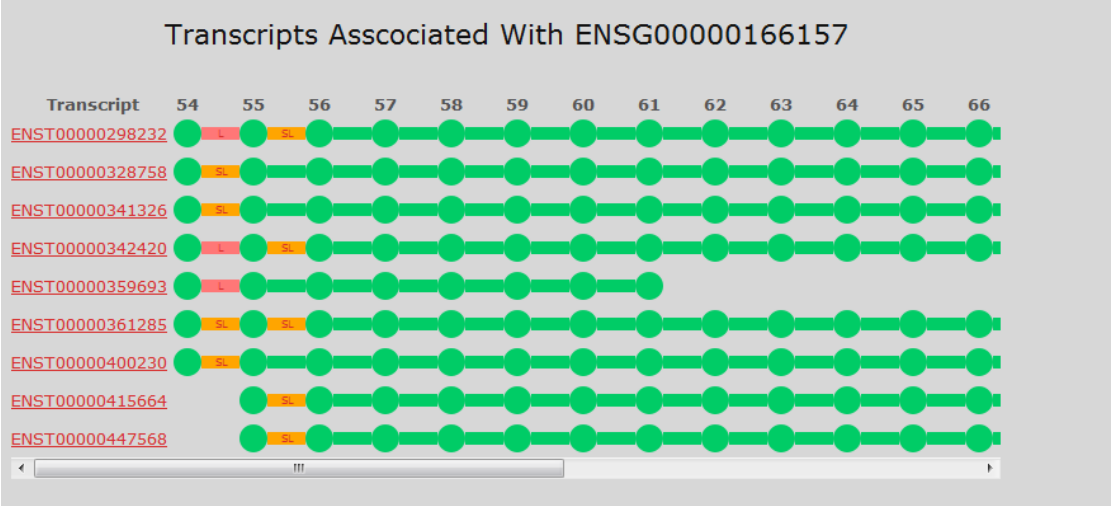
1. **"S"** : Sequence Change (αλλαγή ακολουθίας)
2. **"L"** : Location Change (αλλαγή τοποθεσίας)
3. **"S&L"** : Sequence & Location Change (αλλαγή ακολουθίας και τοποθεσίας ταυτόχρονα)

Εφόσον έχουμε περισσότερη διαθέσιμη πληροφορία για τα transcripts από ότι για τα γονίδια, κάθε κελί στον πίνακα με τα transcripts είναι ταυτόχρονα και σύνδεσμος, ο οποίος ανοίγει νέο παράθυρο (popup), το οποίο παρουσιάζει:

1. Της πληροφορίες του μετάγραφου για τη συγκεκριμένη έκδοση εάν πατηθεί κελί που αφορά έκδοση ή
2. Τις πληροφορίες του μετάγραφου για τις εκδόσεις εκατέρωθεν του συνδέσμου εάν πατήθηκε κελί που περιέχει κάποιο είδος αλλαγής.

Τα δύο αυτά παράθυρα θα περιγραφούν στις παραγράφους (4.1.4, 4.1.5).

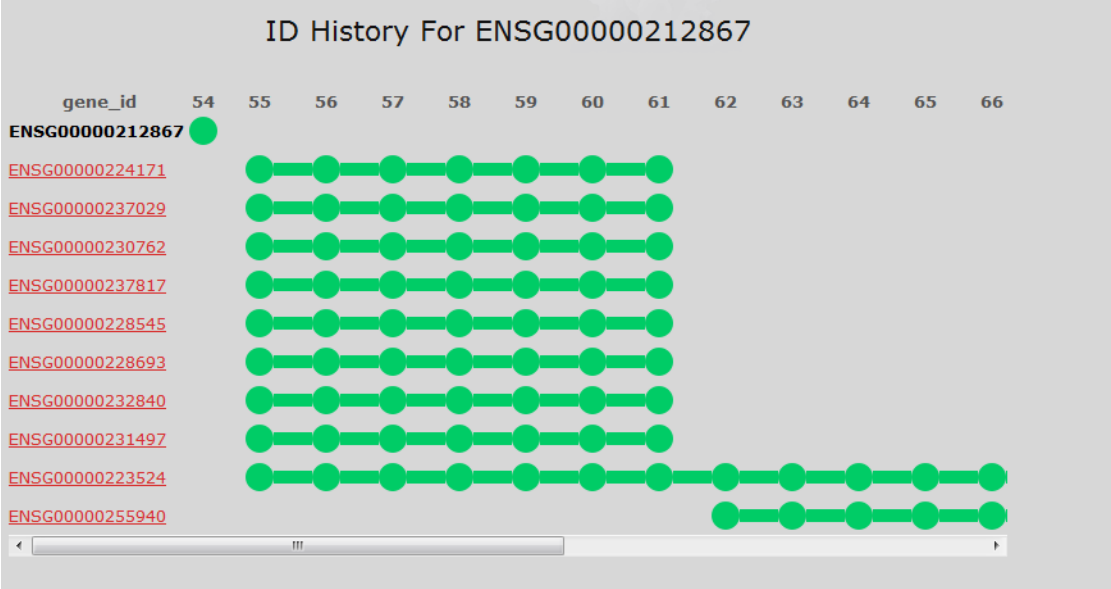
Ένα παράδειγμα αυτού του γραφήματος φαίνεται στην επόμενη εικόνα:



Εικόνα 6. Το τμήμα της σελίδα που εμφανίζει πληροφορίες για τα μετάγραφα που είναι συνδεδεμένα με ένα γονίδιο.

Στο τέλος της σελίδας, θέλουμε να εμφανίζεται η πληροφορία για τη σχέση gene ↔ gene, που ουσιαστικά αποτελούν τους πιθανούς προκατόχους και τους ακόλουθους του συγκεκριμένου Ensembl gene ID, δεδομένου ότι υπάρχει μία τέτοια σχέση. Τα συγκεκριμένα δεδομένα αποθηκεύονται στον πίνακα gene_forward της βάσης δεδομένων και ουσιαστικά

μοντελοποιούν μια οικογένεια από γονίδια. Παράδειγμα του γραφήματος που εμφανίζει αυτή τη σχέση φαίνεται παρακάτω:

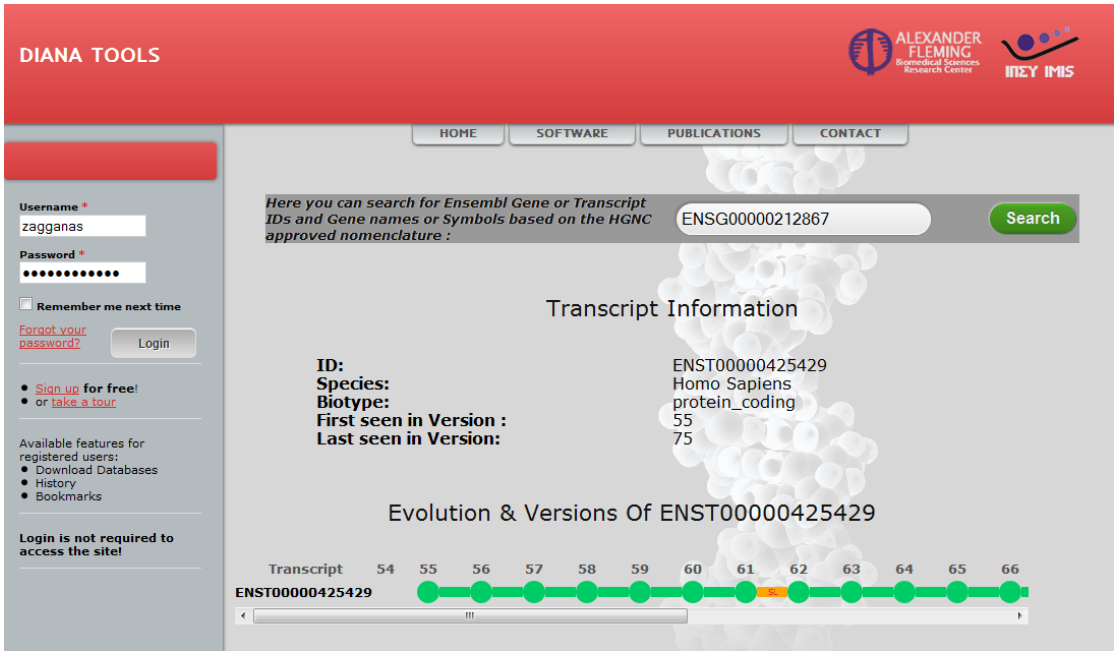


Εικόνα 7. Το τμήμα της σελίδα που εμφανίζει πληροφορίες για την οικογένεια των γονιδίων

4.1.3 Σελίδα αποτελεσμάτων για Transcripts

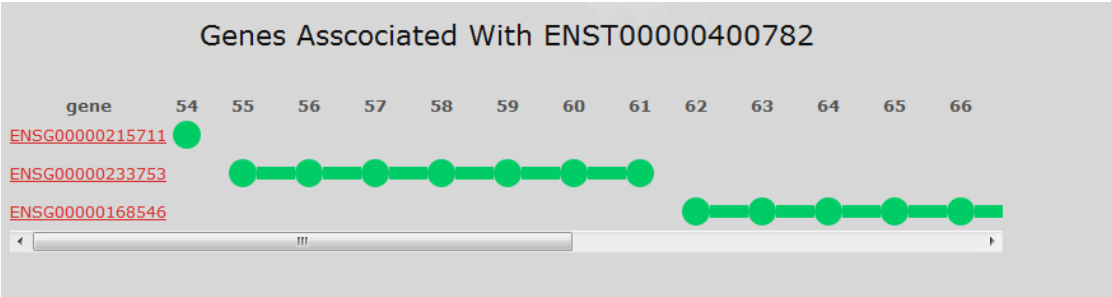
Η σελίδα αυτή, εμφανίζει τη σχέση gene ↔ transcript από την αντίστροφη πλευρά. Δηλαδή αυτό που μας ενδιαφέρει είναι η πληροφορία του transcript καθώς και το με ποια γονίδια είναι συνδεδεμένο στην ιστορία του. Επίσης μας ενδιαφέρει η σχέση transcript ↔ transcript, δηλαδή ποια «οικογένεια» μετάγραφων υπάρχει και για ποιες εκδόσεις.

Αρχικά στη σελίδα εμφανίζονται οι σταθερές πληροφορίες του μετάγραφου, όπως ο βιότυπος του και για ποιες εκδόσεις είναι ενεργό. Στη συνέχεια με ένα πίνακα – γράφημα παρουσιάζουμε τις αλλαγές που έχουν συμβεί στο συγκεκριμένο μετάγραφο διαχρονικά. Τα κελιά του πίνακα περιέχουν συνδέσμους παρόμοιους με αυτούς της σελίδας των γονιδίων, με τη διαφορά ότι ο πίνακας περιέχει μόνο μία γραμμή που παρουσιάζει το transcript το οποίο αναζητήθηκε από το χρήστη. Παράδειγμα του συγκεκριμένου τμήματος φαίνεται στην επόμενη εικόνα:



Εικόνα 8. Το τμήμα της σελίδα που εμφανίζει γενικές πληροφορίες για ένα μετάγραφο καθώς και την διαχρονική του εξέλιξη

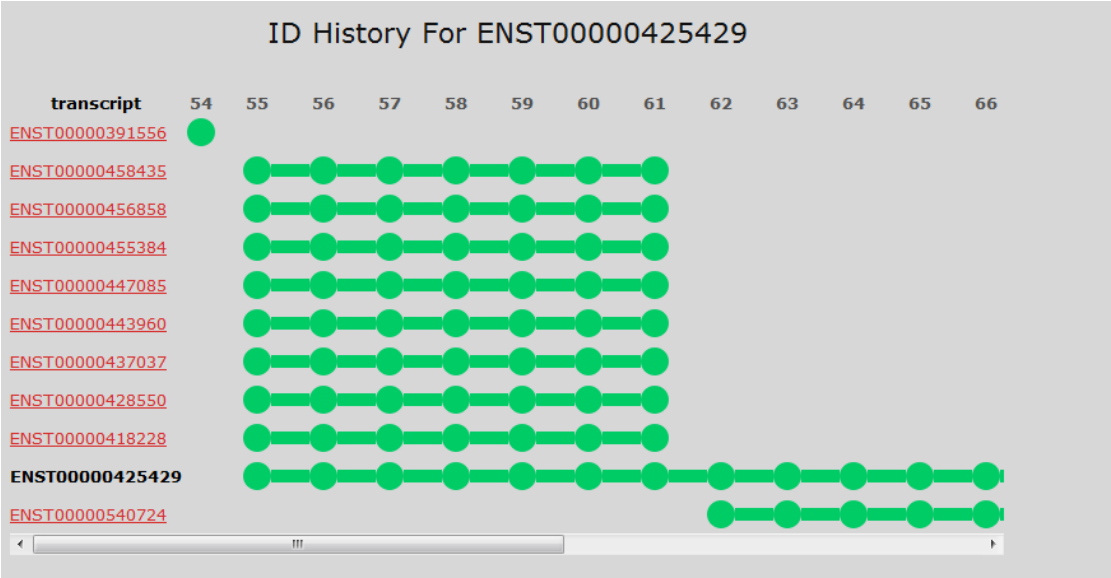
Επόμενο τμήμα πληροφορίας είναι η σχέση transcript ↔genes. Αυτή παρουσιάζεται με έναν πίνακα όπως στη σελίδα των γονιδίων αλλά με την αντίστροφη φορά της σχέσης. Δηλαδή μας ενδιαφέρει για το συγκεκριμένο transcript ποια γονίδια είναι διαχρονικά συνδεδεμένα με αυτό. Ένα χαρακτηριστικό παράδειγμα φαίνεται στην εικόνα 9:



Εικόνα 9. Το τμήμα της σελίδα που εμφανίζει τα γονίδια που είναι διαχρονικά συνδεδεμένα με το μετάγραφο

Στο κάτω μέρος της σελίδας παρουσιάζουμε έναν πίνακα με την πληροφορία για τα την οικογένεια μεταγράφων, δηλαδή τη σχέση transcript ↔ transcript. Αν είναι σωστά ταξινομημένες οι εγγραφές του πίνακα, σχηματίζεται μια πολύ καλή εποπτική εικόνα για τον τρόπο με τον οποίο κάποιο ID αντικαθιστά ένα παλιότερο και αντικαθίσταται από κάποιο νεότερο. Βέβαια, η πληροφορία περιέχει διπλοεγγραφές γιατί η Ensembl δεν μας δίνει την πληροφορία αυτή με πιθανότητα 100%, αλλά μας δίνει transcripts που *πιθανώς* να

αντικατέστησαν κάποια άλλα. Ένα χαρακτηριστικό παράδειγμα ακολουθεί παρακάτω στην εικόνα 10:



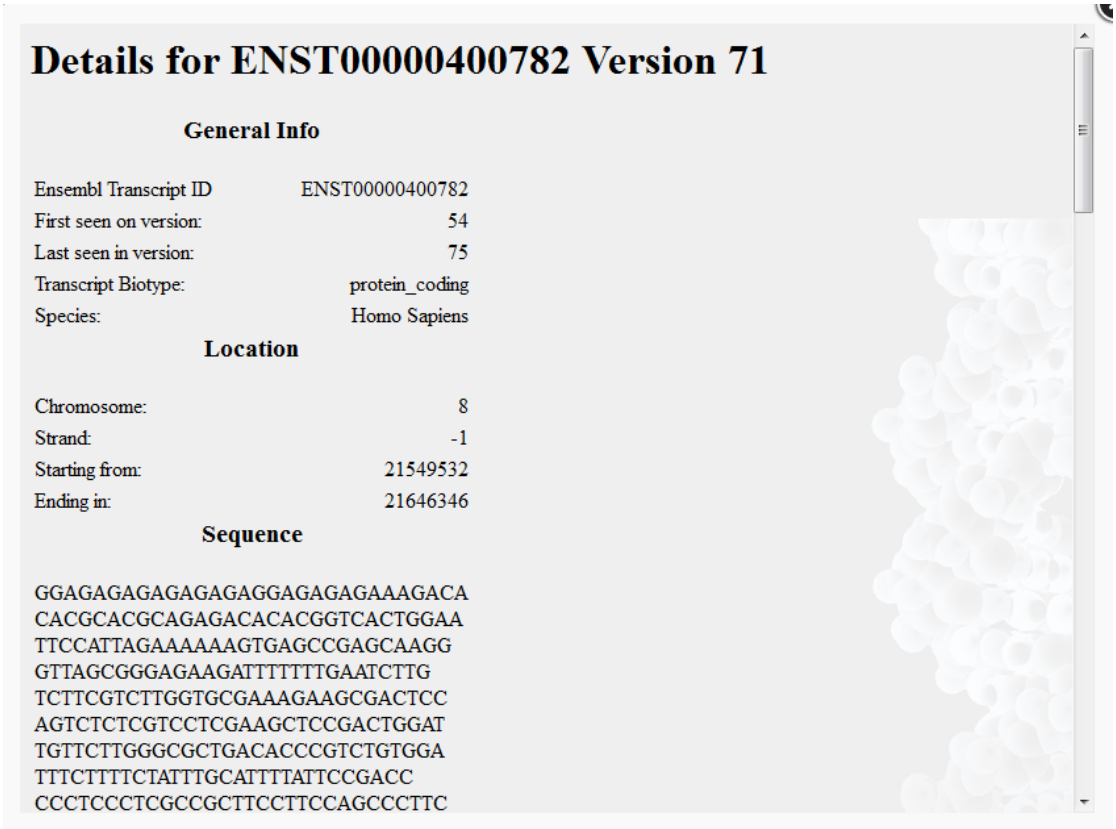
Εικόνα 10. Το τμήμα της σελίδα που εμφανίζει την οικογένεια μεταγράφων

4.1.4 Σελίδα για Λεπτομέρειες σχετικές με Transcripts

Η συγκεκριμένη σελίδα έχει τη μορφή ενός popup. Ανοίγει με κλικ στο σύνδεσμο που περιέχεται στο κελί που αντιστοιχεί στην έκδοση του transcript που θέλουμε να δούμε. Οι πληροφορίες που εμφανίζονται είναι αυτές που περιέχονται στους πίνακες transcript και transcript_version και είναι οι ακόλουθες:

- Οι σταθερές πληροφορίες του μετάγραφου όπως η πρώτη και τελευταία έκδοση στην οποία εμφανίζεται καθώς και ο βιότυπος του.
- Οι πληροφορίες που σχετίζονται με την έκδοση, όπως η τοποθεσία και η ακολουθία του transcript.

Ένα παράδειγμα φαίνεται παρακάτω:



Details for ENST00000400782 Version 71

General Info

Ensembl Transcript ID	ENST00000400782
First seen on version:	54
Last seen in version:	75
Transcript Biotype:	protein_coding
Species:	Homo Sapiens

Location

Chromosome:	8
Strand:	-1
Starting from:	21549532
Ending in:	21646346

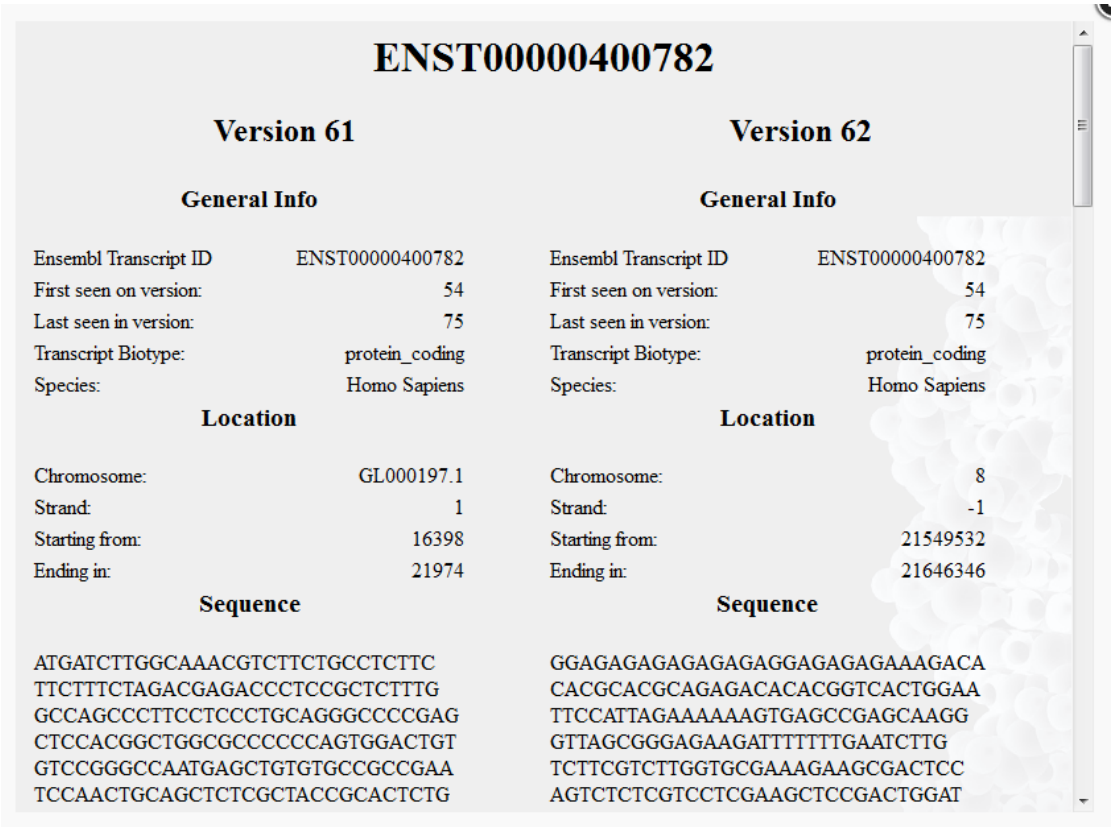
Sequence

GGAGAGAGAGAGAGAGGAGAGAGAAAGACA
CACGCACGCAGAGACACACGGTCACTGGAA
TTCCATTAGAAAAAAGTGAGCCGAGCAAGG
GTTAGCGGGAGAAGATTTTTTTGAATCTTG
TCTTCGTCTTGGTGCGAAAGAAGCGACTCC
AGTCTCTCGTCCTCGAAGCTCCGACTGGAT
TGTTCTTGGGCGCTGACACCCGTCTGTGGA
TTTCTTTTCTATTTGCATTTTATTCCGACC
CCCTCCCTCGCCGCTTCCTTCCAGCCCTTC

Εικόνα 11. Η σελίδα με τις αναλυτικές πληροφορίες ενός μεταγράφου

4.1.5 Σελίδα για Σύγκριση Λεπτομερειών Δύο Διαδοχικών Εκδόσεων Όταν Υπάρχει Αλλαγή

Η σελίδα αυτή εμφανίζεται επίσης σε ένα popup window και παρουσιάζει σε δύο στήλες τις δύο εκδόσεις του γονιδίου ώστε αντιπαραβάλλοντάς τις να φαίνονται οι αλλαγές που συνέβησαν. Σε μελλοντική επέκταση της εφαρμογής θα εισαχθεί και διαφορετικός χρωματισμός των δεδομένων που μεταβάλλονται, για καλύτερη εποπτική σύγκριση και γρηγορότερη αναφορά. Το παράθυρο φαίνεται στην παρακάτω εικόνα:



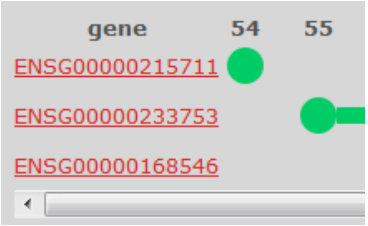
ENST00000400782

Version 61		Version 62	
General Info		**General Info**	
Ensembl Transcript ID	ENST00000400782	Ensembl Transcript ID	ENST00000400782
First seen on version:	54	First seen on version:	54
Last seen in version:	75	Last seen in version:	75
Transcript Biotype:	protein_coding	Transcript Biotype:	protein_coding
Species:	Homo Sapiens	Species:	Homo Sapiens
Location		**Location**	
Chromosome:	GL000197.1	Chromosome:	8
Strand:	1	Strand:	-1
Starting from:	16398	Starting from:	21549532
Ending in:	21974	Ending in:	21646346
Sequence		**Sequence**	
ATGATCTTGGCAAACGTCTTCTGCCTCTTC		GGAGAGAGAGAGAGAGGAGAGAGAAAGACA	
TTCTTTCTAGACGAGACCCTCCGCTCTTTG		CACGCACGCAGAGACACACGGTCACTGGAA	
GCCAGCCCTTCCTCCCTGCAGGGCCCCGAG		TTCCATTAGAAAAAAGTGAGCCGAGCAAGG	
CTCCACGGCTGGCGCCCCCCAGTGGACTGT		GTTAGCGGGAGAAGATTTTTTTGAATCTTG	
GTCCGGGCCAATGAGCTGTGTGCCGCCGAA		TCTTCGTCTTGGTGCGAAAGAAGCGACTCC	
TCCAACTGCAGCTCTCGCTACCGCACTCTG		AGTCTCTCGTCCTCGAAGCTCCGACTGGAT	

Εικόνα 12. Η σελίδα με τις αναλυτικές πληροφορίες δύο εκδόσεων ενός μεταγράφου στις οποίες παρατηρήθηκε αλλαγή

4.1.6 Σύνδεσμοι στους πίνακες

Για να γίνει πιο εύκολη η πλοήγηση στην εφαρμογή, σε κάθε γραμμή του πίνακα που περιλαμβάνει ένα ID (Gene ή Transcript) προστέθηκαν σύνδεσμοι πάνω στο ID, με τους οποίους ο χρήστης μπορεί να μεταβεί στη σελίδα των αποτελεσμάτων αναζήτησης που σχετίζονται με το ID αυτό χωρίς να χρειάζεται να το αναζητήσει ρητά στη μπάρα αναζήτησης. Αυτό παρουσιάζεται παρακάτω:



Εικόνα 13. Σύνδεσμοι πάνω σε IDs

4.2 Τεχνολογίες που Χρησιμοποιήθηκαν

Στο κεφάλαιο αυτό θα περιγραφούν οι τεχνολογίες που χρησιμοποιήθηκαν προκειμένου να λειτουργεί η διεπαφή. Αυτές είναι:

- Okeanos VM Academic Service
- Apache
- MYSQL
- PHP - Yii Framework
- JavaScript - jQuery
- FancyBox

Μια περιγραφή της κάθε τεχνολογίας ακολουθεί παρακάτω

4.2.1 Okeanos VM Service

Ο ωκεανός (https://okeanos.grnet.gr/home/) είναι η υπηρεσία Cloud του ΕΔΕΤ για την Ελληνική Ερευνητική και Ακαδημαϊκή Κοινότητα. Αποτελείται από τις «κυκλάδες» (https://okeanos.grnet.gr/services/cyclades/) καθώς και τον «πίθο+» (https://okeanos.grnet.gr/services/pithos/).

Η υπηρεσία *«κυκλάδες»* επιτρέπει στους χρήστες να δημιουργήσουν εικονικές μηχανές. Οι μηχανές αυτές είναι συνδεδεμένες στο διαδίκτυο και επιτρέπουν εύκολη επιλογή των τεχνικών τους χαρακτηριστικών.

Μέσα σε ελάχιστα λεπτά, ο χρήστης μπορεί να επιλέξει το λειτουργικό σύστημα, την επεξεργαστική ισχύ και το μέγεθος της μνήμης RAM με το οποίο θα χτιστεί η εικονική μηχανή. Οι χρήστες επιπρόσθετα, μπορούν να συνδέσουν τις εικονικές μηχανές τους μέσω τοπικών εικονικών δικτύων.

Η υπηρεσία *«πίθος+»* παρέχει αποθηκευτικό χώρο στο cloud με άμεση πρόσβαση στα δεδομένα από οποιοδήποτε μέρος. Επίσης χρησιμεύσει ως αποθηκευτικός χώρος για τις ανάγκες που μπορεί να έχουν οι εικονικές μηχανές.

Η υπηρεσία είναι διαθέσιμη για όλα τα μέλη της Ερευνητικής και Ακαδημαϊκής κοινότητας. Προπτυχιακοί και μεταπτυχιακοί φοιτητές, υποψήφιοι διδάκτορες, ερευνητές, καθηγητές και διδακτικό προσωπικό δικαιούται πρόσβαση στην υπηρεσία και τις λειτουργίες της.

Οι εικονικές μηχανές που δημιουργούνται έχουν τις ίδιες ικανότητες όπως ένας «φυσικός» υπολογιστής, αλλά προσθέτουν επιπλέον ευκολία και ευελιξία όπως το ότι προσφέρουν εύκολη εγκατάσταση του λειτουργικού συστηματος κατά τη δημιουργία τους και επίσης είναι μόνιμα ενεργοποιημένες.

Για τις ανάγκες της εργασίας δημιουργήθηκε εικονική μηχανή, η οποία έχει δύο πυρήνες, 6GB μνήμης RAM καθώς και 100 GB αποθηκευτικό χώρο. Το λειτουργικό σύστημα της μηχανής είναι Ubuntu Server 14.04 και η σύνδεση γινόταν με SSH μέσω του προγράμματος Putty.

Στην μηχανή εγκαταστάθηκε η τελευταία έκδοση της MySQL (5.6), η οποία παρέχει index join μέσω της λειτουργίας Batched Key Access. Αυτό έγινε για μεγαλύτερη ταχύτητα στα join queries που χρειάζονται για την εφαρμογή, καθώς και για να γίνουν κάποιες δοκιμές σχετικά με τους αλγορίθμους που χρησιμοποιήθηκαν στην εργασία. Εν συγκρίσει με ένα φυσικό μηχάνημα, πάνω στο οποίο έγιναν οι ίδιες σχεδόν δοκιμές, η εικονική μηχανή ήταν εξίσου γρήγορη ή ακόμα και ξεπερνούσε το φυσικό μηχάνημα.

Παράλληλα, το γεγονός ότι η μηχανή έχει στατική διεύθυνση IP, σημαίνει ότι μπορεί πολύ εύκολα να στηθεί ένα website ή μια διαδικτυακή εφαρμογή της.

4.2.2 Apache HTTP Server

Το Project Apache HTTP Server είναι μία συνεργατική προσπάθεια ανάπτυξης λογισμικού, που στοχεύει στο να δημιουργήσει μία ισχυρή, εμπορικού επιπέδου, με πλήρη χαρακτηριστικά και ελεύθερα διαθέσιμη υλοποίηση πηγαίου κώδικα ενός HTTP(web) server. Το project διαχειρίζεται μία ομάδα από εθελοντές από όλο τον κόσμο, που χρησιμοποιούν το Διαδίκτυο και τον Ιστό για να επικοινωνήσουν, να σχεδιάσουν και να αναπτύξουν το πρόγραμμα εξυπηρετητή καθώς και τα σχετική τεκμηρίωση. Το project είναι μέρος του Apache Software Foundation. Επιπροσθέτως, εκατοντάδες χρήστες έχουν συνεισφέρει ιδέες, κώδικα και τεκμηρίωση για το Project.

Το Φεβρουάριο του 1995, το πιο διάσημο πρόγραμμα εξυπηρετητή στο διαδίκτυο ήταν ο υπηρεσία HTTP του public domain, που αναπτύχθηκε από τον Rob McCool στο National Center for Supercomputing Applications (NCSA) στο πανεπιστήμιο Urbana-Champaign του Ιλινόις. Παρόλα αυτά, η ανάπτυξη του httpd σταμάτησε μόλις αυτός έφυγε από το NCSA στα μέσα του 1994. Πολλοί διαχειριστές ανέπτυξαν δικές τους επεκτάσεις και bug fixes, και έτσι έγινε αναγκαίο να δημιουργηθεί μία ενιαία διανομή.

Χρησιμοποιώντας το httpd της NCSA ως βάση, προστέθηκαν όλα τα δημοσιευμένα bug fixes καθώς και όλες οι βελτιώσεις που μπορούσαν να μετά από δοκιμές έγινε διαθέσιμη τον Απρίλιο του 1995 η πρώτη έκδοση του Apache (0.6.2). Την ίδια περίοδο και εντελώς τυχαία, η NCSA ξανάρχισε την ανάπτυξη του δικού της server και μέσω μιας mailing list τα δύο projects άρχισαν να μοιράζονται ιδέες και διορθώσεις.

Μετά από την προσθήκη νέων δυνατοτήτων και νέας τεκμηρίωσης καθώς και μετά από εξαντλητικό beta testing η έκδοση Apache 1.0 έγινε διαθέσιμη την 1^{η} Δεκεμβρίου του 1995.

Σε λιγότερο από ένα χρόνο αφότου δημιουργήθηκε η ομάδα, ο Apache server ξεπέρασε σε αριθμό χρηστών τον httpd server της NCSA και έγινε ο υπ' αριθμόν ένα server στο διαδίκτυο, θέση που διατηρεί μέχρι και σήμερα.

Η τελευταία έκδοση του Apache είναι η 2.0, εκδόθηκε το 2004 και λειτουργεί κάτω από την GPL. Είναι το προεπιλεγμένο πρόγραμμα web server σε όλες τις διανομές Linux καθώς και στο εικονικό μηχάνημα που χρησιμοποιήσαμε. Η συγκεκριμένη έκδοση που τρέχει αυτή τη στιγμή στο VM είναι η 2.4.7.

4.2.3 MySQL

Η MySQL είναι το δεύτερο πιο ευρέως χρησιμοποιούμενο σύστημα διαχείρισης σχεσιακών βάσεων δεδομένων ανοιχτού λογισμικού στον κόσμο (Μάρτιος 2014).

Το project ανάπτυξης της MySQL έχει κάνει διαθέσιμο τον πηγαίο κώδικά της υπό τους όρους της GNU καθώς και υπό τους όρους μιας σειράς εμπορικών συμφωνιών. Μια σειρά από μεγάλους οργανισμούς και ιστοσελίδες χρησιμοποιούν τη MySQL για διαχείριση των δεδομένων τους, όπως το Facebook και η Wikipedia.

Η MySQL δημιουργήθηκε από μια σουηδική εταιρία, την MySQL AB, που ιδρύθηκε από τους David Axmark, Allan Larsson και Michael Widenius. Η πρώτη έκδοση της MySQL εμφανίστηκε στις 23 Μαΐου του 1995. Δημιουργήθηκε αρχικά για προσωπική χρήση με βάση την mSQL και την χαμηλού επιπέδου γλώσσα ISAM. Δημιουργήθηκε μία νέα διεπαφή SQL, διατηρώντας το ίδιο API (Application Programming Interface) με την mSQL έτσι ώστε πολλοί προγραμματιστές να μπορούν να χρησιμοποιήσουν την MySQL αντί για την (εμπορικής άδειας) πρόγονο της, τη mSQL.

Η βασική έκδοση της MySQL προσφέρει μια πληθώρα χαρακτηριστικών, που είναι αντάξια άλλων εμπορικών προγραμμάτων. Μερικά από αυτά είναι:

- Ένα ευρύ υποσύνολο της ANSI SQL 99, καθώς και επεκτάσεις
- Υποστήριξη για πολλαπλές πλατφόρμες
- Stored Procedures, χρησιμοποιώντας μια διαδικαστική γλώσσα που συμμορφώνεται στενά με το SQL/PSM
- Triggers
- Cursors
- Ανανεώσιμα Views
- Information schema
- Αυστηρή λειτουργία (εξασφαλίζει ότι η MySQL δεν περικόπτει ή τροποποιεί πληροφορία προκειμένου να συμμορφώνεται με τον υποκείμενο τύπο δεδομένων, όταν μία μη έγκυρη τιμή εισάγεται σε αυτόν τον τύπο)

- Υποστήριξη Distributed Transaction Processing (DTP). Ως μέρος αυτού παρέχεται commit δύο φάσεων, χρησιμοποιώντας τη μηχανή InnoDB της Oracle.
- Ανεξάρτητες μηχανές αποθήκευσης(MyISAM για ταχύτητα ανάγνωση, InnoDB για συναλλαγές και ακεραιότητα αναφορών, MySQL Archive για την αποθήκευση ιστορικών δεδομένων σε μικρό χώρο)
- Συναλλαγές με τις μηχανές αποθήκευσης InnoDB και NDB Cluster. Επίσης η InnoDB παρέχει savepoints
- Υποστήριξη SSL
- Query Caching
- Sub-SELECTs(πχ ένθετα SELECTs)
- Υποστήριξη για αντιγραφή (πχ Master-Master & Master-Slave) με ένα slave ανά master ή πολλούς slaves ανά master. Αντιγραφή πολλαπλών master παρέχεται στην MySQL Cluster και επίσης μπορεί να προστεθεί σε άλλες μηχανές.
- Ευρετήρια πλήρους κειμένου (αρχικά λειτουργία μόνο της MyISAM). Υποστηρίζεται και από την InnoDB από την έκδοση 5.6 και μετά.
- Ενσωματωμένη βιβλιοθήκη βάσεων δεδομένων
- Υποστήριξη Unicode
- Partitioned tables με «κλάδεμα» των partitions στον optimizer
- Υποστήριξη ACID (Atomicity, Consistency, Isolations, Durability) όταν χρησιμοποιούνται μηχανές με συναλλαγές, όπως η InnoDB και η Cluster
- Shared-nothing clustering χρησιμοποιώντας την MySQL Cluster
- Πολλαπλές μηχανές αποθήκευσης, επιτρέποντας στο χρήστη να επιλέξει αυτήν που είναι πιο αποδοτική για την εφαρμογή του (στην MySQL 5.0 οι μηχανές αποθήκευσης πρέπει να μεταγλωττιστούν πρώτα, ενώ στην 5.1 οι μηχανές αναζήτησης φορτώνονται δυναμικά στο χρόνο εκτέλεσης):
    - Μηχανές αποθήκευσης της MySQL (MyISAM, Falcon, Merge, Memory (heap), Federated, Archive, CSV, Blackhole, Cluster, EXAMPLE, Aria καθώς και InnoDB που είναι και η προεπιλεγμένη μηχανή από την έκδοση 5.5 και μετά)
    - Μηχανές αποθήκευσης ανεπτυγμένες από partners (solidDB, Infobright(προηγουμένως Brighthouse), Kickfire, XtraDB, IBM DB2). Η InnoDB παλιότερα ανήκε σε αυτήν την λίστα, αλλά με τις νέες εξαγορές, η Oracle κατέχει τόσο τον πυρήνα της MySQL όσο και την InnoDB
    - Μηχανές αποθήκευσης ανεπτυγμένες από την κοινότητα (memcache engine, httpd, PBXT, RevisionEngine)
    - Άλλες προσαρμοσμένες μηχανές αποθήκευσης.

- Ομαδοποίηση των commit, συγκεντρώνοντας πολλαπλές συναλλαγές από πολλαπλές συνδέσεις, για να αυξηθεί ο αριθμός των commit ανά δευτερόλεπτο. (Η PostrgreSQL διαθέτει μια ανεπτυγμένη μορφή αυτής της λειτουργικότητας).

Για τις ανάγκες της διπλωματικής χρησιμοποιήθηκαν δύο διαφορετικές εκδόσεις της MySQL. Η 5.5 χρησιμοποιήθηκε κατά τη συλλογή δεδομένων σε ένα φυσικό μηχάνημα ενώ η 5.6 χρησιμοποιήθηκε για την ανάπτυξη της εφαρμογής σε εικονικό μηχάνημα στον ωκεανό.

Για τη συλλογή δεδομένων χρησιμοποιήθηκε η μηχανή αποθήκευσης InnoDB προκειμένου να γίνουν κάποιοι έλεγχοι για την ακεραιότητα των δεδομένων, αλλά η τελική εφαρμογή είναι σε MyISAM, προκειμένου να επιτρέπονται πιο γρήγορες αναγνώσεις.

Παράλληλα, στην έκδοση 5.6 ενεργοποιήθηκε η λειτουργία Batched Key Access (index join) η οποία αυξάνει δραματικά τους χρόνους που χρειάζεται για να γίνει join πάνω σε indexed πεδία.

4.2.4 PHP – Yii Framework

4.2.4.1 PHP

Η PHP είναι μια server-side scripting γλώσσα προγραμματισμού σχεδιασμένη για ανάπτυξη διαδικτυακών εφαρμογών αλλά χρησιμοποιείται και σαν γλώσσα προγραμματισμού γενικού σκοπού. Τον Ιούλιο του 2013 η PHP ήταν εγκατεστημένη σε παραπάνω από 240 εκατομμύρια ιστοτόπους(39% από το δείγμα) και σε 2.1 εκατομμύρια web servers. Δημιουργηθη αρχικά από τον Rasmus Lerdorf το 1994και ενώ το ακρονύμιο «PHP» αρχικά ήταν συντομογραφία του “Personal Home Page”, τώρα είναι πλέον αναδρομική συντομογραφία του “PHP : Hypertext Preprocessor”.

Ο κώδικας PHP μπορεί να αναμειχθεί με κώδικα HTML ή μπορεί να χρησιμοποιηθεί σε συνδυασμό με διάφορες μηχανές δημιουργίας προτύπων ή web frameworks. Η PHP υπόκειται σε επεξεργασία από έναν διερμηνέα που συνήθως υλοποιείται σαν εγγενής μονάδα του web server ή σαν εκτελέσιμο Common Gateway Interface. Αφού διερμηνευθεί και εκτελεστεί ο κώδικας PHP, ο εξυπηρετητής στέλνει το αποτέλεσμα στον πελάτη, συνήθως με τη μορφή ενός μέρους μιας δημιουργημένης ιστοσελίδας (για παράδειγμα η PHP μπορεί να δημιουργήσει τον κώδικα HTML μιας ιστοσελίδας, μια εικόνα ή άλλα δεδομένα). Με την εξέλιξη της PHP δημιουργήθηκε ένα command-line interface για χρήση σε standalone γραφικές εφαρμογές.

4.2.4.2 Yii Framework – Αρχιτεκτονική MVC

4.2.4.2.1 Yii Framework

Το Yii Framework είναι ένα δωρεάν, ανοιχτού κώδικα Web framework για την δημιουργία εφαρμογών, που είναι γραμμένο σε PHP5 και προωθεί καθαρό, DRY(Don't Repeat Yourself) σχεδιασμό που βοηθά στην γρήγορη ανάπτυξη εφαρμογών. Χρησιμεύει στον εξορθολογισμένη ανάπτυξη μιας εφαρμογής και βοηθά τη διασφάλιση της αποδοτικότητας, την επεκτασιμότητα και τη συντηρησιμότητα του τελικού προϊόντος.

Όντας εξαιρετικά βελτιστοποιημένο στην απόδοση, το Yii είναι μια πολύ καλή επιλογή για projects οποιουδήποτε μεγέθους. Παρόλα αυτά δημιουργήθηκε με σκοπό εξελιγμένες εμπορικές εφαρμογές. Ο προγραμματιστής έχει πλήρη έλεγχο στις ρυθμίσεις της εφαρμογής από την αρχή ως το τέλος, έτσι ώστε το η εφαρμογή να συμμορφώνεται με τις συγκεκριμένες εμπορικές προδιαγραφές. Το πακέτο περιέχει τόσο το framework όσο και εργαλεία για τον έλεγχο και την αποσφαλμάτωση της εφαρμογής και έχει ξεκάθαρη και πλήρη τεκμηρίωση.

Το Yii δημιουργήθηκε από τον Qiang Xue, ο οποίος ξεκίνησε το Yii project την 1η Ιανουαρίου του 2008. Ο Qiang είχε προηγουμένως αναπτύξει και συντηρούσε το Prado framework . Τα χρόνια εμπειρίας που απέκτησε, καθώς και σχόλια από προγραμματιστές παγίωσαν την ανάγκη για ένα γρήγορο, ασφαλές και επαγγελματικό framework που είναι ραμμένο στα μέτρα των προδιαγραφών για την ανάπτυξη εφαρμογών στο Web 2.0. Στις 3 Δεκεμβρίου του 2008 ύστερα από ένα χρόνο ανάπτυξης, το Yii 1.0 έγινε διαθέσιμο στο κοινό. Η εκπληκτική του ταχύτητα εν συγκρίσει με άλλα PHP frameworks αποκόμισε αμέσως θετικά σχόλια και η δημοτικότητά του και η υιοθέτησή του από προγραμματιστές αυξάνονται με ταχύ ρυθμό.

Όπως και τα περισσότερα frameworks, το Yii χρησιμοποιεί αμιγή Αντικειμενοστρεφή Προγραμματισμό. Σε αντίθεση όμως με άλλα frameworks, το Yii απαιτεί την έκδοση 5 της PHP. Αυτό είναι σημαντικό, διότι η PHP5 έχει αρκετά βελτιωμένη και ανεπτυγμένη δομή αντικειμένων, εν συγκρίσει με την (παλαιότερη) PHP 4 (πόσο μάλλον την απαρχαιωμένα μοντέλα αντικειμένων που υπήρχαν στην PHP 3).

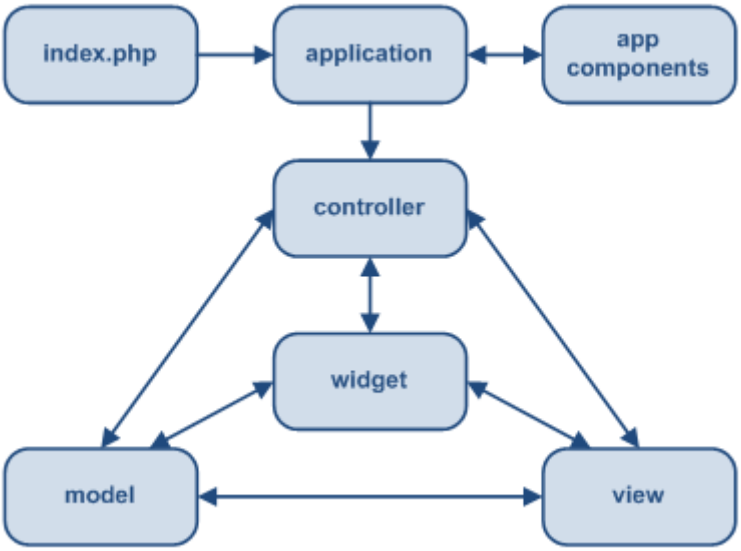
Επιπρόσθετα, το Yii προσφέρει ένα ακόμα επίπεδο ασφαλείας, φροντίζοντας να καλύπτει την εφαρμογή από cross-site scripting και SQL injection.

Το Yii framework είναι η βασική τεχνολογία για τη δημιουργία της εφαρμογής, κυρίως διότι η ιστοσελίδα DIANA TOOLS και οι εφαρμογές της είναι φτιαγμένες με αυτό το framework.

4.2.4.2.2 Αρχιτεκτονική Model-View-Controller

Το Yii χρησιμοποιεί τη de facto αρχιτεκτονική Model-View-Controller (MVC). Η αρχιτεκτονική αυτή βοηθά να γίνει πιο ξεκάθαρη η λειτουργία της κάθε μονάδας και διαχωρίζει την παραγωγή των δεδομένων (Model) από την εμφάνισή τους (View) με τον Controller να παίζει το ρόλο του «μεσάζοντα» ανάμεσα στα Model και View. Το πρότυπο αυτό χρησιμοποιείται ευρέως σε διαδικτυακές εφαρμογές. Ο σκοπός του είναι να διαχωρίζει τη λογική της εφαρμογής από τη διεπαφή με το χρήστη, έτσι ώστε οι αλλαγές που εφαρμόζουν οι προγραμματιστές στο ένα από τα τμήματα να μην επηρεάζουν το άλλο. Στο μοντέλο MVC το μοντέλο αναπαριστά τα δεδομένα καθώς επίσης και τους κανόνες με τους οποίους λειτουργούν. Το κομμάτι view εμπεριέχει στοιχεία της διεπαφής του χρήστη, όπως κείμενο και δεδομένα από φόρμες, ενώ ο ελεγκτής (controller) είναι υπεύθυνος για την επικοινωνία μεταξύ του model και του view.

Το Yii επίσης εισάγει την έννοια ενός κύριου ελεγκτή (controller) που ονομάζεται application και αναπαριστά την επεξεργασία των αιτήσεων των χρηστών της εφαρμογής. Αυτός ευθύνεται για την επεξεργασία των αιτήσεων που γίνονται, και στέλνει τα αιτήματα σε άλλους κατάλληλους ελεγκτές, οι οποίοι εκτελούν τις περαιτέρω ενέργειες. Μια τυπική ροή εργασιών σε μια εφαρμογή αναπτυγμένη με το framework Yii δίνεται στο ακόλουθο σχήμα:



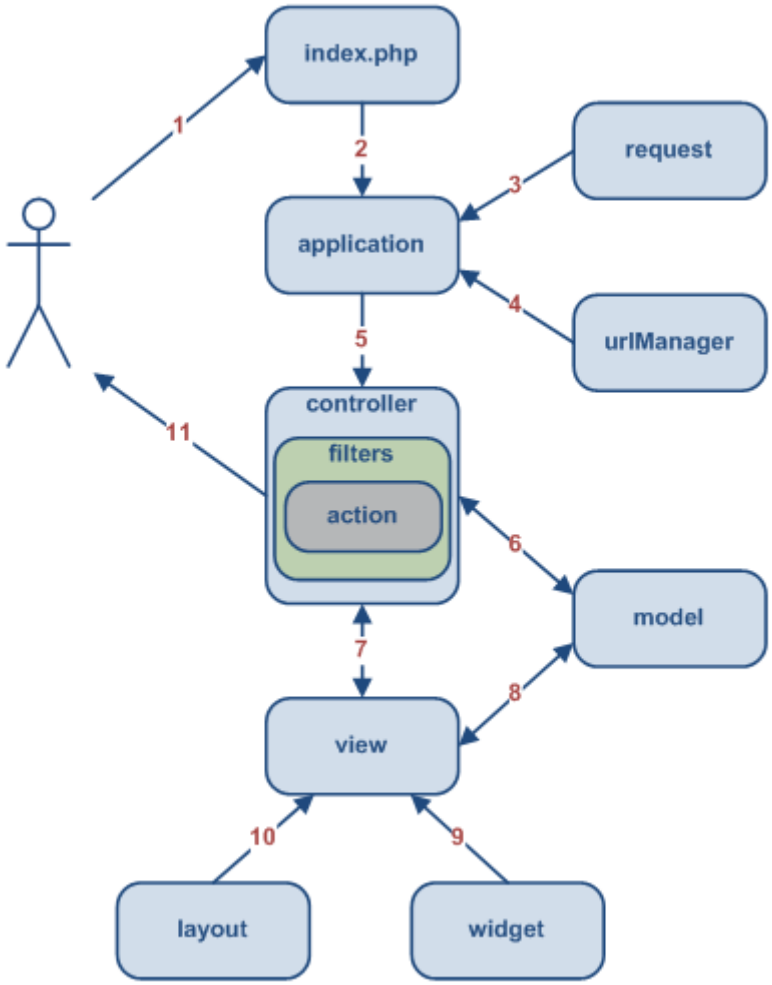
Εικόνα 13. Στατική δομή μίας εφαρμογής στο Yii

Τα βήματα χειρισμού μιας αίτησης συνοψίζονται στα:

1. Μόλις ο χρήστης κάνει μια αίτηση στη διεύθυνση όπου βρίσκεται η εφαρμογή Yii, εκτελείται από τον εξυπηρετητή το bootstrap πρόγραμμα με όνομα index.php.
2. Δημιουργείται ένα αντικείμενο application (controller) και το εκτελεί.

3. Το αντικείμενο application επεξεργάζεται την αίτηση και λαμβάνει τις πληροφορίες της αίτησης από ένα συστατικό που ονομάζεται request.
4. Στη συνέχεια προσδιορίζει την κατάλληλη ενέργεια (action) του ελεγκτή (controller) που αφορά η αίτηση, με τη βοήθεια ενός συστατικού που ονομάζεται urlManager.
5. Δημιουργείται ένα αντικείμενο του κατάλληλου ελεγκτή και καθορίζεται η ενέργεια που πρέπει να πραγματοποιηθεί. Πριν γίνει όμως αυτό, διαβάζει τους κανόνες που καθορίζονται μέσα στον ελεγκτή, προκειμένου να καθορίσει αν επιτρέπεται η πρόσβαση στη συγκεκριμένη ενέργεια(action).
6. Η ενέργεια διαβάζει το κατάλληλο μοντέλο (δεδομένα).
7. Η ενέργεια «προβάλλει» το κατάλληλο αρχείο view (διεπαφή).
8. Το view διαβάζει και προβάλλει τα κατάλληλα δεδομένα από το μοντέλο.
9. Εκτελούνται ορισμένα widgets, εάν αυτά υπάρχουν.
10. Τα αρχεία view προβάλλονται μέσα σε μια διάταξη (layout) η οποία καθορίζεται από την εφαρμογή.
11. Η ενέργεια(action) ολοκληρώνει την δημιουργία της προβολής των αρχείων view και παρουσιάζει τα αποτελέσματα στον χρήστη.

Η διαδικασία που περιγράφηκε παραπάνω παρουσιάζεται στο παρακάτω σχήμα:



Εικόνα 14. Τυπική ροή μίας εφαρμογής του Yii

Στη συνέχεια θα γίνει αναφορά ξεχωριστά στα κύρια στοιχεία μιας εφαρμογής του Yii framework, δηλαδή τα model, view και controller.

4.2.4.2.2.1 Controller

Οι ελεγκτές δημιουργούνται από το αντικείμενο application κάθε φορά που γίνονται αιτήσεις από το χρήστη. Ένας ελεγκτής περιέχει μια σειρά μεθόδων που ονομάζονται ενέργειες (actions), οι οποίες όταν εκτελούνται συνήθως μεταφέρουν στοιχεία της αίτησης σε αντικείμενα μοντέλων. Τα μοντέλα δημιουργούν τα αντικείμενα, επιστρέφουν την πληροφορία στον ελεγκτή και ο ελεγκτής εμφανίζει στη συνέχεια το αρχείο view που απαιτείται.

Κάθε controller έχει και μια προεπιλεγμένη μέθοδο, η οποία εκτελείται όταν καλείται ο ελεγκτής στην περίπτωση που δεν καθορίζεται ρητά κάποιο άλλο action από την αίτηση του χρήστη. Επίσης περιέχει ένα σύνολο κανόνων για το ποιες ενέργειες επιτρέπεται να

εκτελεστούν από κάθε είδος χρήστη (μη εγγεγραμμένος, συνδεδεμένος, διαχειριστής). Στην περίπτωση που η ενέργεια που ζητήθηκε δεν αντιστοιχεί στον κατάλληλο χρήστη ο χρήστης προωθείται στην αρχική σελίδα.

4.2.4.2.2.2 Model

Οι κλάσεις μοντέλων αναπαριστούν ένα αντικείμενο διαχείρισης δεδομένων που μπορεί να είναι είτε μία φόρμα εισαγωγής δεδομένων είτε μια εγγραφή από βάση δεδομένων. Το Yii framework έχει βασικά δύο ειδών model, τα active records και τα form models.

Ένα form model χρησιμοποιείται για να αποθηκεύει δεδομένα που εισάγουν οι χρήστες. Τέτοια δεδομένα συνήθως εξαφανίζονται αφού χρησιμοποιηθούν μια φορά, δηλαδή δεν είναι επαναχρησιμοποιούμενα.

Ένα active record είναι ένα σχεδιαστικό πρότυπο που χρησιμοποιείται για την πρόσβαση σε βάσεις δεδομένων με βάση το αντικειμενοστρεφές μοντέλο. Κάθε active record αναπαριστά μια εγγραφή σε μια βάση και τα πεδία μιας τέτοιας εγγραφής αποτελούν ιδιότητες του αντικειμένου. Επίσης υπάρχει η δυνατότητα ένα μοντέλο να μεικτό, δηλαδή να είναι αναπαράσταση δύο τύπων εγγραφών σε μια βάση δεδομένων, όπως έγινε στα πλαίσια της εργασίας.

4.2.4.2.2.3 View

Τα τμήματα μιας εφαρμογής Yii που αποτελούν το view είναι προγράμματα PHP , τα οποία σχετίζονται κυρίως με την εμφάνιση της διεπιφάνειας χρήστη.[20] Τα αρχεία view δεν πρέπει να περιέχουν περισσότερη PHP στον κώδικά τους από αυτήν που χρειάζεται για να προβάλουν τα δεδομένα. Η πιο πολύπλοκη δουλειά που μπορεί να χρειαστεί να κάνουν είναι να χρησιμοποιήσουν ένα βρόχο για να τυπώσουν τα δεδομένα ενός πίνακα. Άρα πρέπει ακολουθείται το *μοντέλο της ανάποδης πυραμίδας*, δηλαδή *«παχύ» μοντέλο, «μέτριος» controller* και *«λεπτό» view*. Τα views καλούνται από τους controllers με τους οποίους σχετίζονται, μετά από την εκτέλεση της κατάλληλης μεθόδου. Με τον τρόπο αυτό διαχωρίζεται η εμφάνιση των δεδομένων από την εκτέλεση των πράξεων που απαιτείται κατά την επεξεργασία τους.[21]

4.2.5 JavaScript – jQuery

Η JavaScript είναι μια δυναμική γλώσσα προγραμματισμού. Χρησιμοποιείται στην πλειοψηφία των περιπτώσεων ως μέρος των περιηγητών διαδικτύου (web browsers). Η υλοποίηση των τελευταίων επιτρέπει σε client-side scripts να αλληλεπιδρούν με τον χρήστη, να ελέγχουν τον περιηγητή, να επικοινωνούν ασύγχρονα κσαθώς και να αλλάζουν το περιεχόμενο του εγγράφου που παρουσιάζεται. Επίσης η JavaScrpt χρησιμοποιείται για

server-side προγραμματισμό δικτύου, ανάπτυξη παιχνιδιών καθώς και τη δημιουργία εφαρμογών για υπολογιστή και κινητό.

Η JavaScript είναι γλώσσα scripting βασισμένη σε πρωτότυπα με δυναμικό σύστημα τύπων και έχει συναρτήσεις πρώτης τάξης. Η σύνταξή της είναι επηρεασμένη από τη C. Παράλληλα η JavaScript αντιγράφει πολλά ονόματα και συμβάσεις ονομάτων από τη Java, αλλά οι δύο γλώσσες είναι κατά τα άλλα άσχετες μεταξύ τους και έχουν πολύ διαφορετική σημασιολογία.

Η εφαρμογή της JavaScript σε χρήση εκτός ιστοσελίδων – για παράδειγμα σε έγγραφα PDF, εξειδικευμένους περιηγητές για συγκεκριμένους ιστοτόπους και εφαρμογές επιφάνειας εργασίας – είναι αρκετά σημαντική. Νεότερες και πιο γρήγορες εικονικές μηχανές JavaScript καθώς και πλατφόρμες χτισμένες πάνω τους έχουν αυξήσει τη δημοτικότητα της JavaScript για server-side διαδικτυακές εφαρμογές. Από την πλευρά του πελάτη, η JavaScript υλοποιούταν παραδοσιακά σαν μια διερμηνευόμενη γλώσσα αλλά στους πρόσφατους πλοηγούς (2012 και μετά) γίνεται just-in-time μεταγλώττιση.

Η jQuery είναι μια cross-platform βιβλιοθήκη της JavaScript, σχεδιασμένη να απλοποιήσει το client-side scripting της HTML. Έγινε διαθέσιμη τον Ιανουάριο του 2006 στο BarCamp στη Νέα Υόρκη από τον John Resig. Όντας σε χρήση σε παραπάνω από το 80% των πιο επισκέψιμων ιστοσελίδων, η jQuery είναι η πιο διαδεδομένη βιβλιοθήκη της JavaScript σήμερα.

Η jQuery αποτελεί δωρεάν λογισμικό, ανοιχτού κώδικα υπό την άδεια του MIT. Η σύνταξή της είναι σχεδιασμένη ώστε να κάνει πιο εύκολη την πλοήγηση σε ένα έγγραφο, την επιλογή στοιχείων του DOM, τη δημιουργία animations, το χειρισμό γεγονότων και την ανάπτυξη εφαρμογών Ajax. Η jQuery παρέχει επίσης τη δυνατότητα στους προγραμματιστές να δημιουργήσουν plug-ins πάνω στη βιβλιοθήκη της JavaScript. Αυτό επιτρέπει στους προγραμματιστές να δημιουργήσουν αφηρημένα στοιχεία για αλληλεπίδραση χαμηλού επιπέδου και animation, ανεπτυγμένα εφέ και υψηλού επιπέδου εφαρμογές με υποστήριξη θεμάτων. Αυτή η προσέγγιση μονάδων στην βιβλιοθήκη jQuery επιτρέπει τη δημιουργία ισχυρών δυναμικών ιστοσελίδων και διαδικτυακών εφαρμογών.

Η JavaScript και η jQuery χρησιμοποιήθηκαν για να ενσωματωθούν δυναμικές λειτουργίες στην εφαρμογή όπως η απόκρυψη στοιχείων ή η λειτουργία του εργαλείου FancyBox που περιγράφεται παρακάτω.

4.2.6 FancyBox

Το FancyBox είναι μία εναλλακτική στο LightBox. Είναι ένα εργαλείο που προσφέρει έναν όμορφο τρόπο για την προσθήκη δημιουργίας παραθύρου popup για εικόνες, html περιεχόμενο καθώς και πολυμεσικό περιεχόμενο σε ιστοσελίδες. Είναι φτιαγμένο πάνω στην βιβλιοθήκη jQuery της JavaScript και είναι εύκολο να υλοποιηθεί και να τροποποιηθεί. Η

λειτουργία iframe του FancyBox χρησιμοποιήθηκε για τα popup παράθυρα που δείχνουν τις λεπτομέρειες ενός transcript καθώς και τη σύγκριση μεταξύ δύο εκδόσεων ενός transcript.

4.3 Υλοποίηση της εφαρμογής

Η όλη εφαρμογή ουσιαστικά αποτελείται από το κυρίως μέρος της που είναι ένας controller στο yii framework. Ο controller αυτός αποφασίζει ποιο action θα εκτελεστεί με βάση το link στον browser. Φροντίζει για την επικοινωνία ανάμεσα στο back-end που είναι το μοντέλο και στο front-end που είναι τα views. Το μοντέλο αναλαμβάνει την επικοινωνία με τη βάση δεδομένων καθώς και την παραγωγή δομών δεδομένων κατάλληλων ώστε να μπορέσουν να προβληθούν από τα αρχεία view.

4.3.1 Controller

Ο controller είναι η «καρδιά» της εφαρμογής. Αυτός δημιουργήθηκε αρχικά με το εργαλείο Gii που παρέχει το Yii για τη δημιουργία μοντέλων, controllers και views. Ο controller αυτός έχει default actions για CRUD (Create, Read, Update, Delete) λειτουργικότητα. Παρόλα αυτά καμία από αυτές δεν μας χρησίμευε, οπότε απομακρύνθηκαν και στη θέση τους μπήκαν νέες που έχουν να κάνουν με τη λειτουργία της εφαρμογής. Επιπλέον, δημιουργήθηκε άλλη μια συνάρτηση αυτόματα από το Gii. Η συνάρτηση αυτή λέγεται accessRules και περιέχει τους κανόνες με τους οποίους περιγράφεται ποιοι χρήστες της εφαρμογής έχουν δικαίωμα να κάνουν τι (unregistered users, logged-in users, administrators). Εμείς επιτρέπουμε αυτά τα actions να μπορούν να γίνονται από όλους τους unregistered users, καθώς η εφαρμογή θα είναι ανοιχτή στο ευρύ κοινό. Με αυτόν τον τρόπο εξασφαλίζεται ότι δεν θα μπορέσει κάποιος που δεν δικαιούται πρόσβαση σε συγκεκριμένα δεδομένα να τα δει ή να τα τροποποιήσει. Τα actions που προστέθηκαν θα περιγραφούν αναλυτικότερα στη συνέχεια.

Ο τρόπος με τον οποίο καλούμε ένα action ενός controller είναι με τη μορφή controller/action. Δηλαδή, αν για παράδειγμα θέλουμε να καλέσουμε το action view του controller genechange θα ακολουθήσουμε το σύνδεσμο:

http://mydomain.gr/index.php/?r=genechange/view

Το Yii framework επιτρέπει αλλαγή στις ρυθμίσεις ώστε ο παραπάνω σύνδεσμος για ευκολία να γίνει:

http://mydomain.gr/index.php/genechange/view

Παρόλα αυτά όμως, στον συγκεκριμένο ιστότοπο στον οποίο θα φιλοξενηθεί η εφαρμογή, η ρύθμιση αυτή δεν είναι ενεργοποιημένη.

Κάτι τελευταίο που πρέπει να αναφερθεί για τον controller είναι ότι οι συναρτήσεις/actions παίρνουν σαν παραμέτρους τους τις μεταβλητές GET. Δηλαδή αν το action «SomeName» έχει επικεφαλίδα:

```
public function actionSomeName($var1, $var2)
```

τότε ο σύνδεσμος που θα πρέπει να ακολουθήσουμε είναι :

```
http://mydomain.gr/index.php/?r=genechange/view&var1=value1&var2=value2
```

Σε περίπτωση που ξεχάσουμε να βάλουμε έστω και μία από τις δύο μεταβλητές(έστω και με κενή τιμή), το Yii framework θα επιστρέψει HTTP Error Code 400, Invalid Request. Όλες οι παρακάτω συναρτήσεις έχουν πριν το όνομά τους τις λέξεις "public function". Ο controller αποτελεί μια κλάση της PHP που κληρονομεί από την κλάση CController του Yii.

4.3.1.1 actionIndex()

Είναι το action που φορτώνεται είτε έχει δηλωθεί στο σύνδεσμο είτε όχι, δηλαδή φορτώνεται αυτόματα από το Yii σε κάθε περίπτωση και είναι το προεπιλεγμένο action του controller.

Το συγκεκριμένο action λειτουργεί ως εξής:

- Εάν η μεταβλητή $_POST έχει τεθεί, δηλαδή υπάρχει, τότε προώθεί στο action view αυτό που αναζητείται μέσω της $_GET μεταβλητής..
- Αν όχι, τότε φορτώνει απλά το αρχείο με τη φόρμα αναζήτησης στο layout του ιστοτόπου.

4.3.1.2 actionView($id, $searched, $type)

Ίσως το πιο σημαντικό action από όλα, γιατί είναι αυτό που παίρνει τα δεδομένα της αναζήτησης, τα προωθεί στο μοντέλο και ανάλογα με το τι αποφάσισε το μοντέλο ότι αναζητήθηκε, θα προβάλει και τα ανάλογα δεδομένα. Παίρνει τρεις $_GET μεταβλητές ως παραμέτρους (φροντίζει το Yii framework να αντιστοιχίσει τις μεν με τις δε). Στη συνέχεια δημιουργεί ένα νέο μοντέλο που επεξεργάζεται τα δεδομένα τόσο για τα γονίδια όσο και για τα μετάγραφα.

Καλεί τη συνάρτηση του μοντέλου που μόλις δημιούργησε, η οποία αναζητά στη βάση αν υπάρχει εγγραφή με ακριβές transcript ή gene ID και όμοια εγγραφή στα synonyms.

Αν δεν υπάρχει τέτοιο όνομα ή ID τότε γίνεται προώθηση σε κάποια σελίδα που γράφει ότι δεν βρέθηκε αποτέλεσμα και δίνει την επιλογή στο χρήστη να ξαναψάξει.

Δεν χρειάζεται να γίνει έλεγχος εάν περάστηκε ο σωστός αριθμός των παραμέτρων, γιατί το Yii framework θα εμφανίσει μήνυμα σφάλματος 400.

Αν βρεθεί κάτι, διαβάζεται ο τύπος του αποτελέσματος. Αν βρέθηκε κάποιο ID και δεν είναι όνομα γονιδίου (βρέθηκε gene ή transcript), τότε:

- Αν είναι transcript εκτελούνται οι ρουτίνες του μοντέλου για το transcript
- Αν είναι gene εκτελούνται οι ρουτίνες του μοντέλου για το gene.

Στην περίπτωση που έχουμε περισσότερα από ένα αποτελέσματα ή το αποτέλεσμα αποτελεί μέρος ενός ονόματος, τότε θέτουμε ένα flag στο μοντέλο, το οποίο λέει στο view file ότι δεν έχουμε κάνει ακόμα την επιλογή του τελικού ID και ότι χρειάζεται ο χρήστης να επιλέξει ID και μετά θα του δείξουμε τα στοιχεία του αποτελέσματος. Εδώ εισέρχεται και η $_GET μεταβλητή $type. Μόλις ο χρήστης επιλέξει το ID του οποίου τα αποτελέσματα θέλει να εμφανίσει, καλείται πάλι το action του controller και η μεταβλητή γεμίζει με τον τύπο του αποτελέσματος, ώστε να μην χρειαστεί να ξανακάνουμε αναζήτηση, αλλά να τρέξουμε τις κατάλληλες ρουτίνες ανάλογα αν είναι gene ή transcript.

Στο τέλος, φορτώνουμε το κεντρικό αρχείο view.php που καλεί άλλα αρχεία για να εμφανίσει το αποτέλεσμα της αναζήτησης.

4.3.1.1 actionViewTrv($vid,$vers)

Το συγκεκριμένο action καλείται για να δημιουργήσει τη σελίδα που θα φορτωθεί στο popup window για την έκδοση ενός transcript. Στη μεταβλητή $vid μπαίνει η τιμή του εσωτερικού transcript version ID, το οποίο θα φορτώσουμε, και στη μεταβλητή $vers η τιμή της έκδοσης την οποία θέλουμε να δούμε για την επικεφαλίδα.

Το action δημιουργεί ένα νέο μοντέλο TranscriptVersion και στη συνέχεια καλεί τη ρουτίνα με την οποία γεμίζει μια λίστα με τα πεδία του αποτελέσματος. Στη συνέχεια στέλνει τα αποτελέσματα σε ένα view file το οποίο τα παρουσιάζει.

4.3.1.1 actionViewTrvCmp($vid1,$vid2,$vers)

Σε αυτή την περίπτωση γίνεται η σύγκριση δύο διαφορετικών εκδόσεων του ίδιου transcript. Οι μεταβλητές $vid1 και $vid2 περιέχουν το εσωτερικό transcript version ID ενώ η μεταβλητή $vers περιέχει την έκδοση όπως και πριν.

Αυτή τη φορά όμως δημιουργούνται δύο μοντέλα και όχι ένα, καλούνται οι συναρτήσεις που φέρνουν την πληροφορία από τη βάση δεδομένων και τα δύο μοντέλα μαζί με την έκδοση περνιούνται ως παράμετροι σε ένα αρχείο view για να εμφανιστεί η πληροφορία.

4.3.1 Models

Για τη συγκεκριμένη εφαρμογή έγινε η χρήση δύο μοντέλων. Τα μοντέλα αυτά είναι το Gene και το TranscriptVersion. Θα μπορούσαμε να χρησιμοποιήσουμε ένα ακόμα μοντέλο, το Transcript αν θέλαμε να είμαστε πλήρως σωστοί στη λογική του Αντικειμενοστρεφούς προγραμματισμού και την αρχιτεκτονική MVC, αλλά παρόλα αυτά υπήρχε ήδη ένα μοντέλο με το όνομα Transcript από κάποια άλλη εφαρμογή του Diana Tools. Άρα οι διακριτές

λειτουργίες που θα έκανε ένα τέτοιο μοντέλο έχουν περαστεί στο μοντέλο Gene. Οι ρουτίνες του κάθε μοντέλου θα περιγραφούν αναλυτικά παρακάτω.

4.3.1.1 Gene

Το μοντέλο αυτό είναι υπεύθυνο να κάνει την αναζήτηση για το εάν υπάρχει το λήμμα προς αναζήτηση στη βάση δεδομένων. Στη συνέχεια παρέχει ρουτίνες για κάθε τύπο ID, ώστε να επιστρέφονται τα σωστά δεδομένα κάθε φορά. Ουσιαστικά αποτελεί το back-end της εφαρμογής και περιέχει όλο τον κώδικα που δεν επιτρέπεται να περιέχουν τα views. Το μοντέλο είναι μια κλάση της PHP η οποία κληρονομεί το CActiveRecord.

Η ρουτίνα searchID($id) παίρνει ως παράμετρο ένα ID από τον controller και εκτελεί το παρακάτω ερώτημα με βάση το $id

```
(SELECT ens_gene_id as id, ‘’ as info, ‘gene’ as type
FROM gene WHERE ens_gene_id=’some_id’)
UNION
(SELECT ens_tr_id as id, ‘’ as info, ‘trans’ as type
FROM gene WHERE ens_tr_id=’some_id’)
UNION
(SELECT g.ens_gene_id as id, s.info as info, ‘name’ as type
FROM synonyms s
INNER JOIN gene g ON g.int_gene_id=s.gene_id
WHERE s.info LIKE %’some_id’%)
```

Το ερώτημα αυτό επιστρέφει έναν πίνακα με μηδέν ή περισσότερα στοιχεία, ανάλογα με το αποτέλεσμα.

Αν είναι περισσότερα από μηδέν, τότε υπάρχει αποτέλεσμα και ένα flag γίνεται true.

Αν υπάρχουν περισσότερα από ένα ή το αποτέλεσμα που επεστράφη είναι μέρος ενός ονόματος, τότε επιστρέφεται μια λίστα και ένα συγκεκριμένο flag γίνεται true. Αυτό συμβαίνει για να “πούμε” στον controller να φορτώσει τη σελίδα με τα προτεινόμενα αποτελέσματα.

Εάν υπάρχει μόνον ένα αποτέλεσμα και αυτό είναι τύπου gene, τότε θέτουμε το ID που βρέθηκε στην αντίστοιχη μεταβλητή του gene. Εάν το αποτέλεσμα είναι transcript, τότε χρησιμοποιούμε τις αντίστοιχες μεταβλητές του transcript. Στη συνέχεια, ο controller, ανάλογα με τον τύπο του αποτελέσματος, καλεί κάποιες από τις παρακάτω ρουτίνες.

Εάν το αποτέλεσμα είναι *gene*, τότε καλούνται οι

- findGeneInfo
- findGeneForward
- findGeneTranscripts

Στην περίπτωση του *transcript* καλούνται οι παρακάτω τρεις ρουτίνες:

- findTrForward
- findTranscriptGenes
- findTranscriptVersions

Οι ρουτίνες findGeneForward και findTrForward δημιουργούν μια οικογένεια από IDs που σε συμφωνία με τα γραφικά εργαλεία της Ensembl παρουσιάζουν το ιστορικό ενός ID.

Αρχικά φέρνουμε από τη βάση δεδομένων το εσωτερικό ID του αποτελέσματος που αναζητά ο χρήστης. Στη συνέχεια το χρησιμοποιούμε για να φέρουμε τους προγόνους και τους απογόνους με βάση τους πίνακες tr_forward και gene_forward. Η διαδικασία χρησιμοποιεί μια ουρά FIFO για τη λίστα των προς αναζήτηση IDs καθώς και μια λίστα(hash table) που περιέχει τα IDs που έχουν αναζητηθεί ήδη.

Η διαδικασία που ακολουθείται είναι η εξής:

- Φέρε το εσωτερικό ID του αποτελέσματος της αναζήτησης
- Βάλε το στη λίστα προς αναζήτηση
- Η λίστα με τα IDs που έχουν ελεγχθεί είναι άδεια.
- Όσο η λίστα προς έλεγχο είναι γεμάτη:
    - Βγάλε από τη λίστα το πρώτο στοιχείο
    - Αν αυτό είναι στη λίστα με τα ελεγμένα IDs τότε αγνόησέ το και προχώρα στην επόμενη επανάληψη
    - Αλλιώς φέρε τα εσωτερικά IDs των προγόνων από τη βάση και πρόσθεσέ τα στη λίστα προς έλεγχο
    - Φέρε τα εσωτερικά Ids των απογόνων από τη βάση και πρόσθεσέ τα στη λίστα προς έλεγχο
- Μόλις η λίστα προς έλεγχο αδειάσει, δημιούργησε μια λίστα από τα εσωτερικά IDs που βρίσκονται στον πίνακα με τα ελεγμένα στοιχεία, χωρισμένα με κόμματα(,).
- Στη συνέχεια φέρε τα στοιχεία όλων των IDs με το εξής ερώτημα:

```
SELECT * FROM gene/transcript WHERE int_gene/tr_id IN (id-list)
```

Πριν ολοκληρωθεί η ρουτίνα δημιουργεί ένα hash table με τα στοιχεία του κάθε ID προκειμένου να τυπωθούν από το αντίστοιχο αρχείο view. Το βήμα αυτό δεν είναι απολύτως απαραίτητο για τα genes, αλλά στα transcripts υπάρχει το πρόβλημα των διπλοεγγραφών, όπως αναφέρθηκε στο κεφάλαιο 3, διότι τα δεδομένα είναι λάθος. Με αυτόν τον τρόπο

συγχωνεύονται οι εγγραφές αυτές. Παρόλα αυτά για λόγους ομοιότητας η διαδικασία εκτελείται και για τα genes και για τα transcripts.

Οι ρουτίνες findGeneTranscripts και findTranscriptVersions λειτουργούν με παρόμοιο τρόπο. Και οι δύο εκτελούν ένα παρόμοιο ερώτημα, το οποίο είναι:

```
SELECT t.ens_tr_id, t.first_ver as tr_fv, t.last_ver as tr_lv, tv.int_trv_id, tv.first_ver as trv_fv,
tv.last_ver as trv_lv, tv.changes
FROM gene g
INNER JOIN gene_to_transcript gtt ON g.int_gene_id=gtt.int_gene_id
INNER JOIN transcript t ON t.int_tr_id=gtt.int_tr_id
INNER JOIN transcript version tv ON tv.int_tr_id=t.int_tr_id
WHERE g./t.ens_gene/transcript_id='some_id'
```

Το συγκεκριμένο ερώτημα φέρνει πίσω μια σειρά από εγγραφές οι οποίες είναι δύσκολο να τυπωθούν αυτούσιες από ένα αρχείο και ειδικά στην περίπτωση των transcripts μπορεί να έχουμε διαφορετικές εγγραφές για το ίδιο ens_tr_id με διαφορετικές εκδόσεις.

Για το λόγο αυτό δημιουργούμε μια ειδική δομή δισδιάστατου hash table στο οποίο για κάθε ID αποθηκεύουμε τα βασικά στοιχεία του transcript ID, τις εκδόσεις με τις οποίες είναι συνδεδεμένο με το gene(αν πρόκεται για gene) ή τις εκδόσεις για τις οποίες είναι ενεργό (αν πρόκειται για transcript) και σε ένα εσωτερικό hash table αποθηκεύουμε για κάθε έκδοση από αυτές που υπάρχει ποιο είναι το εσωτερικό ID της εγγραφής του πίνακα που περιέχει τις εκδόσεις των transcripts. Άλλο ένα εσωτερικό hash table αποθηκεύει σε ποιες εκδόσεις έχουμε αλλαγές και ποιες.

Με αυτόν τον τρόπο μπορεί το αντίστοιχο αρχείο view να δημιουργήσει έναν πίνακα με τα κελιά που περιέχουν τις ενεργές εκδόσεις καθώς και τις αλλαγές και επίσης σε κάθε κελί μπορεί να μπει ένας σύνδεσμος που θα παρουσιάζει τις αλλαγές που έχουν συμβεί.

Η ρουτίνα findTranscriptGenes εκτελεί το εξής ερώτημα:

```
SELECT g.ens_gene_id, gtt.first_ver, gtt.last_ver
FROM gene g
INNER JOIN gene_to_transcript gtt ON g.int_gene_id=gtt.int_gene_id
INNER JOIN transcript t ON t.int_tr_id=gtt.int_tr_id
WHERE g./t.ens_gene/transcript_id='some_id'
```

Το αποτέλεσμα γυρνάει μια σειρά από εγγραφές οι οποίες μπαίνουν σε ένα απλό μονοδιάστατο hash table για να είναι εύκολα προσβάσιμα από το αρχείο view που θα τυπώσει τον πίνακα.

Η ρουτίνα findGeneInfo() επιστρέφει τα στοιχεία του πίνακα synonyms που αντιστοιχούν στο αντίστοιχο Ensemble gene ID και τοποθετούνται σε ένα κατάλληλο δισδιάστατο hash table για εύκολη πρόσβαση από ένα αρχείο view. Το ερώτημα που εκτελείται είναι :

```
SELECT s.info_type, s.info
FROM gene g
INNER JOIN synonyms s ON g.int_gene_id=s.int_gene_id
WHERE g.ens_gene_id='some_id'
ORDER BY s.info ASC
```

4.3.1.1 TranscriptVersion

Το μοντέλο αυτό έχει ουσιαστικά μία μόνο ρουτίνα, η οποία εκτελεί το παρακάτω ερώτημα στη βάση δεδομένων και φέρνει τα αποτελέσματα σε ένα associative array που μπορεί να προσπελαστεί άμεσα από το αρχείο view το οποίο θα τυπώσει την αντίστοιχη σελίδα του popup παραθύρου.

```
SELECT t.ens_tr_id, t.first_ver as tfv, t.last_ver as tlv, t.tr_biotype,t.species,tv.sequence, tv.chromosome, tv.start, tv.stop, tv.strand
FROM transcript t
INNER JOIN transcript_version tv ON tv.int_tr_id=t.int_tr_id
WHERE tv.int_trv_id='some_id'
```

Έτσι επιστρέφεται μόνο μία εγγραφή και δεν χρειάζεται περαιτέρω επεξεργασία.

4.3.2 Views

Τα views που δημιουργήθηκαν, όπως περιγράφεται στις προδιαγραφές του Yii δεν περιέχουν περισσότερες εντολές από κάποια echo που τυπώνουν HTML ή κάποια foreach loops, τα οποία τυπώνουν πίνακες. Το μόνο που περιέχει κάπως περισσότερο κώδικα είναι το αρχείο view.php που κάνει render άλλα αρχεία, ανάλογα με τον τύπο του ID που θα τυπωθεί.

Τα αρχεία view που δημιουργήθηκαν είναι τα παρακάτω:

- view.php : παρουσιάζει τα αποτελέσματα για ένα ID κάνοντας κλήσεις σε άλλα αρχεία που περιέχουν τις αντίστοιχες μονάδες κάθε σελίδας
- _form.php : περιέχει τη μπάρα αναζήτησης
- _no_id.php: εμφανίζεται όταν το ID που αναζητήθηκε δεν υπάρχει
- search.php: εμφανίζεται ως μέρος του action index και καλεί μόνο το _form.php για να εμφανίσει τη μπάρα αναζήτησης.
- _suggestions.php : εμφανίζει τα αποτελέσματα που επεστράφησαν αν είναι παραπάνω από ένα και προσθέτει σύνδεσμο σε καθένα από αυτά προκειμένου να εμφανιστεί η αντίστοιχη σελίδα με τις πληροφορίες του κάθε ID
- _view_gene_forward.php : εμφανίζει το γράφημα που περιέχει το ιστορικό του gene ID
- _view_gene_trans.php : εμφανίζει το γράφημα που περιέχει τα transcript IDs που είναι συνδεδεμένα με το γονίδιο καθώς και το ιστορικό εξέλιξής τους
- _viw_info.php : Εμφανίζει τις σταθερές πληροφορίες του ID που αναζητήθηκε, όπως ο βιότυπος καθώς και οι εκδόσεις στις οποίες είναι ενεργό το ID. Στην περίπτωση που το ID είναι αναγνωριστικό ενός γονιδίου, τότε προβάλλεται και πίνακας με τα συνώνυμα του γονιδίου
- _view_tr_forward.php : εμφανίζει το γράφημα με το ιστορικό του transcript ID
- _view_tr_genes.php : εμφανίζει τον γράφημα με τα gene IDs που είναι συνδεδεμένα με το transcript
- _view_transcript_versions_tb.php : εμφανίζει τις εκδόσεις ενός *μόνο* transcript καθώς και το ιστορικό εξέλιξής του
- _view_tr_version.php : εμφανίζει τις πληροφορίες μίας έκδοσης ενός transcript
- _view_tr_comparison.php : εμφανίζει τις πληροφορίες δύο εκδόσεων δίπλα-δίπλα, προκειμένου να φανούν οι αλλαγές
- _view_tr_single.php : καλείται από τα δύο παραπάνω αρχεία προκειμένου να τυπώσει εσωτερικά τους τα δεδομένα των εκδόσεων που ζητούνται

5 *Επίλογος*

Στο παρόν κεφάλαιο γίνεται μια ανακεφαλαίωση του αντικειμένου της διπλωματικής εργασίας. Επιπρόσθετα αναφέρονται πιθανές χρήσεις του εργαλείου που δημιουργήθηκε καθώς και μελλοντικές επεκτάσεις που μπορούν να γίνουν.

5.1 Σύνοψη

Η διπλωματική εργασία είχε ως στόχο τη δημιουργία ενός συστήματος καταγραφής και προβολής των αλλαγών που συμβαίνουν σε βιολογικά δεδομένα στην πάροδο του χρόνου. Προκειμένου να επιτευχθεί αυτός χρειάστηκε να μελετηθούν εκτενώς τα αρχεία που περιέχουν πληροφορία για βιολογικά δεδομένα, και κυρίως γονίδια, προκειμένου να επιλεγούν αυτά που μπορούν να προσφέρουν τα κατάλληλα δεδομένα για την εύρεση ιστορικής εξέλιξης. Τελικά ως καταλληλότερη επιλέχθηκε η βάση γονιδιώματος Ensembl, καθώς τα αρχεία που παρείχε είχαν τη μεγαλύτερη ποικιλία διαχρονικής πληροφορίας για γονίδια. Επίσης επιλέχθηκε και η βάση ονοματολογίας HGNC του HUGO, για τη λήψη περιγραφικών δεδομένων όπως το όνομα ή το καθολικό σύμβολο ενός γονιδίου, διότι παρόμοια πληροφορία δεν δινόταν από την Ensembl. Στη συνέχεια δημιουργήθηκε μια αρχική βάση δεδομένων στην οποία αποθηκεύτηκαν τα ακατέργαστα δεδομένα που προέρχονταν από τα αρχεία της Ensembl. Το σχήμα αυτό της βάσης χρησιμοποιήθηκε στη συνέχεια, προκειμένου να καταγραφούν και να αποθηκευτούν οι διαχρονικές αλλαγές στα αναγνωριστικά των μεταγράφων και των γονιδίων με τρόπο αποδοτικό, χωρίς περιττή πληροφορία και με τρόπο εύκολο για προβολή των αποτελεσμάτων.

Η πληροφορία που παρήχθη χρησιμοποιήθηκε για την παραγωγή μιας διεπαφής, μέσω της οποίας ένας ερευνητής βιολόγος μπορεί να αναζητήσει και να προβάλει αποτελέσματα για ένα συγκεκριμένο αναγνωριστικό της Ensembl. Επιπλέον επιτρέπεται η αναζήτηση ενός γονιδίου με όλο ή μέρος του συμβόλου και του ονόματος του. Η διεπαφή αυτή θα

ενσωματωθεί στη σελίδα Diana Tools, η οποία περιέχει ήδη μεγάλη πληθώρα εργαλείων για την ανάλυση μιας ποικιλίας βιομορίων.

Στα πλαίσια της διπλωματικής λοιπόν επετεύχθησαν οι εξής στόχοι:

- Δημιουργία συστήματος το οποίο μοντελοποιεί και καταγράφει τα διαχρονικά δεδομένα για γονίδια τεσσάρων οργανισμών
- Εμπλουτισμός αυτών των διαχρονικών δεδομένων για βιομόρια με επιπλέον πληροφορίες από άλλες διαδικτυακές βάσεις δεδομενων
- Υλοποίηση ενός εργαλείου προβολής της διαχρονικής πληροφορίας για γονίδια
- Δημιουργία συστήματος το οποίο φροντίζει για την ανανέωση της βάσης δεδομένων κάθε φορά που κυκλοφορεί μια νέα έκδοση της βάσης δεδομένων Ensembl.

5.2 Μελλοντικές εργασίες

Η εφαρμογή που δημιουργήθηκε θα ενσωματωθεί στη σουίτα εργαλείων της ιστοσελίδας Diana Tools. Επιπλέον η πληροφορία που συλλέχθηκε έγινε με βάση έναν τύπο νουκλεϊκού οξέως, το RNA και τα αντίστοιχα προϊόντα της μεταγραφής που ονομάζονται *μετάγραφα*. Αυτό ουσιαστικά καθορίζει το πλαίσιο στο οποίο κινήθηκε η εργασία και προσδιορίζει πεδία έρευνας πάνω στα οποία μπορεί να κινηθεί μια μελλοντική εργασία

5.2.1 Παραγωγή διαχρονικών δεδομένων με βάση πληροφορία γονιδίων

Το σύστημα που δημιουργήθηκε χρησιμοποιεί δεδομένα από μετάγραφα που προκύπτουν από γονίδια προκειμένου να παραχθεί διαχρονική πληροφορία για τα γονίδια αυτά. Ένα σημείο που χρειάζεται διερεύνηση είναι αν το κατά πόσο μπορεί να παραχθεί πληροφορία για τα γονίδια με βάση στοιχεία των γονιδίων και όχι των μεταγράφων τους.

Πιθανόν για αυτή τη διαδικασία να χρειαστεί να αντληθούν δεδομένα από πολλαπλές βάσεις δεδομένων και να εξαχθούν ακολουθίες από τα αρχεία τους με βάση τις συντεταγμένες που δίνονται από κάποιες άλλες βάσεις.

5.2.1 Παραγωγή διαχρονικών δεδομένων για βιομόρια πρωτεϊνών

Η ίδια διαδικασία που ακολουθήθηκε για να παραχθούν διαχρονικά δεδομένα για γονίδια μπορεί να ακολουθηθεί και για τα προϊόντα τους, τις πρωτεΐνες. Μάλιστα θα ήταν δόκιμο μια τέτοια πληροφορία να ενσωματωθεί στα δεδομένα που παρήχθησαν με αυτή τη διπλωματική εργασία έτσι ώστε να γίνει μια σύνδεση ανάμεσα σε αναγνωριστικά μεταγράφων και αναγνωριστικά πρωτεϊνών.

6 *Βιβλιογραφία*

[1] Περιγραφή της διπλωματικής εργασίας.

[2] http://unlockinglifescode.org/timeline?tid=4

[3] Θωμόπουλος, Γ. Ν., *Βιολογία Κυττάρου*, University Studio Press, (Θεσσαλονίκη 1990)

[4] Μαργαρίτης, Λ.Χ., *Κυτταρική Βιολογία*, Ιατρικές εκδόσεις Λίτσας, (Αθήνα 1989)

[5] http://www.elearning.aua.gr/elearning/auth/cell_biol/kefalaio1/p4.htm

[6] http://el.wikipedia.org/wiki/%CE%91%CF%81%CF%87%CE%B5%CE%AF%CE%BF:Biological_cell.svg

[7] Βιολογία Γ' Λυκείου Θετικής Κατεύθυνσης, Οργανισμός Εκδόσεως Διδακτικών Βιβλίων, 2014

[8] http://en.wikipedia.org/wiki/MicroRNA (citation an the bottom of the page)

[9] Αναζήτηση σε επιστημονικές βάσεις δεδομένων με βάση την ιστορική εξέλιξη των δεδομένων, διπλωματική εργασία, Ηλίας Κανέλλος, 2012

[10] http://en.wikipedia.org/wiki/List_of_biological_databases#Genome_databases

[11] http://en.wikipedia.org/wiki/Reference_genome

[12] http://www.ncbi.nlm.nih.gov/genome

[13] http://www.genenames.org/

[14] http://www.ensembl.org/index.html

[15] https://okeanos.grnet.gr/support/faq/okeanos-what-is-okeanos/

[16] http://httpd.apache.org/ABOUT_APACHE.html

[17] http://en.wikipedia.org/wiki/Mysql redirect από http://www.mysql.com/about/

[18] http://en.wikipedia.org/wiki/Php

[19] http://www.yiiframework.com/about/

[20] The definitive Yii Guide, http://www.yiiframework.com/doc/guide/

[21] The Yii Book by Larry Ullman, 2014

[22] http://en.wikipedia.org/wiki/Jquery

[23] http://fancyapps.com/fancybox/